Enrique Castillo,

José Manuel Gutiérrez, y

Ali S. Hadi

Sistemas Expertos y Modelos de Redes Probabilísticas

Con más de 150 ejemplos y 250 figuras

Enrique Castillo
Universidad de Cantabria
39005 Santander, España
E-mail: castie@ccaix3.unican.es

José Manuel Gutiérrez
Universidad de Cantabria
39005 Santander, España
E-mail: gutierjm@ccaix3.unican.es

Ali S. Hadi
Universidad de Cornell
358 Ives Hall
Ithaca, NY 14853-3901, USA
E-mail: ali-hadi@cornell.edu

A todo el pueblo de la desaparecida Yugoslavia con la esperanza de que vivan juntos en paz y sean amigos, como lo son los autores de este libro, a pesar de sus diferencias en religiones, lenguas y nacionalidades.

Prefacio

En las dos últimas décadas se ha producido un notable desarrollo en el área de la inteligencia artificial y, en particular, en la de los sistemas expertos. Debido a su carácter multidisciplinar, muchos de los resultados obtenidos en esta disciplina han sido publicados en diversas revistas de numerosos campos: ciencias de la computación, ingeniería, matemáticas, estadística, etc. Este libro trata de reunir, organizar y presentar estos resultados de forma clara y progresiva. Se ha tratado de mantener la información actualizada al máximo, de tal forma que incluso algunos de los conceptos presentados en el libro no han sido publicados previamente (por ejemplo, algunos resultados de los Capítulos 7, 11 y 12).

Este libro está destinado a estudiantes e investigadores de áreas teóricas y aplicadas de disciplinas tales como ciencias de la computación, ingeniería, medicina, matemáticas, economía y ciencias sociales. Dado este carácter multidisciplinar, se han intentado mantener al mínimo los conocimientos previos necesarios para leer este libro. Así, sólo se requieren algunas nociones básicas de estadística y probabilidad y estar familiarizado con los conceptos elementales del álgebra lineal (ver, por ejemplo, Hadi (1996)). En algunas ocasiones los conceptos se ilustran utilizando algunos programas de *Mathematica*. Para un completo entendimiento de estos programas, se requiere cierto conocimiento del programa *Mathematica* (ver Castillo y otros (1993)).

Este libro puede ser utilizado como libro de consulta, o como libro de texto en cursos de postgrado o en últimos cursos de carrera. El libro contiene numerosos ejemplos ilustrativos y ejercicios al final de cada capítulo. También se han desarrollado varios programas que implementan los algoritmos y metodologías presentadas. Estos programas, junto con los manuales de usuario correspondientes, pueden obtenerse de la dirección World Wide Web (WWW) http://ccaix3.unican.es/~AIGroup. Creemos que pueden ayudar a los lectores a entender los conceptos presentados y a aplicar esta metodología a sus respectivos ámbitos profesionales y de estudio. Por ejemplo, estos programas han sido utilizados para resolver algunos de los ejemplos y ejercicios del libro, así como para analizar varios ejemplos prácticos reales (Capítulo 12). Finalmente, al final del libro se incluye una extensa bibliografía para consultas adicionales.

Aunque en el libro se presentan tanto la teoría como las aplicaciones prácticas de esta disciplina, se ha puesto un interés especial en las aplicaciones prácticas. Por eso, muchos de los teoremas presentados se incluyen sin demostración, refiriéndose a la bibliografía para aquellos lectores interesados. Así mismo, se introducen numerosos ejemplos para ilustrar cada uno de los conceptos presentados.

Este libro está organizado de la forma siguiente. El Capítulo 1 presenta una introducción al área de la inteligencia artificial y los sistemas expertos que, entre otras cosas, analiza algunos ejemplos ilustrativos, describe los componentes de un sistema experto, así como las etapas necesarias para su desarrollo, y analiza la relación de los sistemas expertos con otras áreas de la inteligencia artificial. Los Capítulos 2 y 3 describen los dos tipos principales de sistemas expertos: los sistemas expertos basados en reglas y los basados en probabilidad. Aunque estos dos tipos de sistemas se introducen de forma separada, en el Capítulo 3 se muestra que los sistemas expertos basados en reglas se pueden considerar como un tipo particular de sistemas expertos probabilísticos. También se muestra que dos de las componentes más complejas e importantes de un sistema experto son el "subsistema de control de coherencia" y el "motor de inferencia" y estos son, quizás, los dos componentes más débiles de los sistemas expertos desarrollados hasta la fecha. En los Capítulos 5−10 se muestra la forma de implementar estos componentes de forma eficiente.

A partir del Capítulo 5 se requieren algunos conceptos de la teoría de grafos, ya que éstos serán la base para construir las redes probabilísticas. Los conceptos necesarios en este libro, que son un prerequisito básico para entender los capítulos siguientes, se presentan en el Capítulo 4. Los Capítulos 5−7 analizan el problema de la construcción de modelos probabilísticos, desde varias perspectivas. En particular, los conceptos de dependencia e independencia condicional, necesarios para definir la estructura de las redes probabilísticas, se introducen y analizan con detalle en el Capítulo 5. El Capítulo 6 presenta los dos modelos más importantes de redes probabilísticas, las redes de Markov y las redes Bayesianas, que se definen a partir de una estructura gráfica no dirigida y dirigida, respectivamente. El Capítulo 7 presenta algunas extensiones de los modelos gráficos para definir modelos probabilísticos más generales a partir de multigrafos, conjuntos de relaciones de independencia condicional, modelos multifactorizados y modelos definidos condicionalmente.

Los Capítulos 8 y 9 presentan los métodos de propagación exacta y aproximada más importantes, respectivamente. El Capítulo 10 analiza la propagación simbólica que es uno de los avances más recientes de la propagación en redes probabilísticas. El Capítulo 11 está dedicado al problema del aprendizaje; en concreto, al problema del aprendizaje de redes Bayesianas a partir de un conjunto de datos (una base de datos, etc.). Finalmente, el Capítulo 12 ilustra la aplicación de todos los conceptos presentados en el libro mediante su aplicación a ejemplos reales.

Muchos de nuestros colegas y estudiantes han leido versiones anteriores de este libro y nos han proporcionado muchas sugerencias que han ayudado a mejorar notablemente distintas partes del mismo. En particular, agradecemos la inestimable ayuda de (en orden alfabético): Noha Adly, Remco Bouckaert, Federico Ceballos, Jong Wang Chow, Javier Díez, Dan

Geiger, Joseph Halpern, Judea Pearl, Julius Reiner, José María Sarabia, Milan Studený, y Jana Zvárová.

Enrique Castillo
Jose Manuel Gutiérrez
Ali S. Hadi

Tabla de Contenidos

Capítulo 1
Introducción

1.1 Introducción

No hace mucho tiempo, se creía que algunos problemas como la demostración de teoremas, el reconocimiento de la voz y el de patrones, ciertos juegos (como el ajedrez o las damas), y sistemas altamente complejos de tipo determinista o estocástico, debían ser resueltos por personas, dado que su formulación y resolución requieren ciertas habilidades que sólo se encuentran en los seres humanos (por ejemplo, la habilidad de pensar, observar, memorizar, aprender, ver, oler, etc.). Sin embargo, el trabajo realizado en las tres últimas décadas por investigadores procedentes de varios campos, muestra que muchos de estos problemas pueden ser formulados y resueltos por máquinas.

El amplio campo que se conoce como *inteligencia artificial* (IA) trata de estos problemas, que en un principio parecían imposibles, intratables y difíciles de formular utilizando ordenadores. A. Barr y E. A. Feigenbaum, dos de los pioneros de la investigación en IA, definen ésta como sigue: (véase Barr y Feigenbaum (1981), página 4):

> *La Inteligencia Artificial es la parte de la Ciencia que se ocupa del diseño de sistemas de computación inteligentes, es decir, sistemas que exhiben las características que asociamos a la inteligencia en el comportamiento humano que se refiere a la*

> *comprensión del lenguaje, el aprendizaje, el razonamiento, la resolución de problemas, etc.*

Hoy en día, el campo de la IA engloba varias subáreas tales como los sistemas expertos, la demostración automática de teoremas, el juego automático, el reconocimiento de la voz y de patrones, el procesamiento del lenguaje natural, la visión artificial, la robótica, las redes neuronales, etc. Este libro está dedicado a los *sistemas expertos*. Aunque los sistemas expertos constituyen una de las áreas de investigación en el campo de la IA, la mayor parte de las restantes áreas, si no todas, disponen de una componente de sistemas expertos formando parte de ellas.

Este capítulo presenta una introducción a los sistemas expertos. Se comienza con algunas definiciones de sistemas expertos en la Sección 1.2. La Sección 1.3 da algunos ejemplos que sirven para motivar los sistemas expertos en varios campos de aplicación. Estos ejemplos muestran la importancia y la amplia aplicabilidad de los sistemas expertos en la práctica. Algunas de las razones para utilizar los sistemas expertos se indican en la Sección 1.4. Los principales tipos de sistemas expertos se presentan en la Sección 1.5. La Sección 1.6 discute y analiza la estructura de los sistemas expertos y sus principales componentes. Las diferentes etapas necesarias para el diseño, desarrollo e implementación de los sistemas expertos se analizan en la Sección 1.7. Finalmente, la Sección 1.8 se dedica a mencionar algunas de las restantes áreas de investigación de la IA y suministran al lector interesado algunas de las referencias más importantes, revistas, y direcciones de acceso (WWW).

1.2 ¿Qué es un Sistema Experto?

En la literatura existente se pueden encontrar muchas definiciones de sistema experto. Por ejemplo, Stevens (1984), página 40, da la definición siguiente:

> *Los sistemas expertos son máquinas que piensan y razonan como un experto lo haría en una cierta especialidad o campo. Por ejemplo, un sistema experto en diagnóstico médico requeriría como datos los síntomas del paciente, los resultados de análisis clínicos y otros hechos relevantes, y, utilizando éstos, buscaría en una base de datos la información necesaria para poder identificar la correspondiente enfermedad.* [...] *Un Sistema Experto de verdad, no sólo realiza las funciones tradicionales de manejar grandes cantidades de datos, sino que también manipula esos datos de forma tal que el resultado sea inteligible y tenga significado para responder a preguntas incluso no completamente especificadas.*

Aunque la anterior es todavía una definición razonable de un sistema experto, han surgido desde entonces otras definiciones, debido al rápido desarrollo de la tecnología (ver, por ejemplo, Castillo y Álvarez (1991) y Durkin (1994)). El sentido de estas definiciones puede resumirse como sigue:

Definición 1.1 Sistema Experto. *Un sistema experto puede definirse como un sistema informático (hardware y software) que simula a los expertos humanos en un área de especialización dada.*

Como tal, un sistema experto debería ser capaz de procesar y memorizar información, aprender y razonar en situaciones deterministas e inciertas, comunicar con los hombres y/u otros sistemas expertos, tomar decisiones apropiadas, y explicar por qué se han tomado tales decisiones. Se puede pensar también en un sistema experto como un *consultor* que puede suministrar ayuda a (o en algunos casos sustituir completamente) los expertos humanos con un grado razonable de fiabilidad.

Durante la última década se han desarrollado muy rápidamente numerosas aplicaciones de sistemas expertos a muchos campos (ver, por ejemplo, Quinlan (1987, 1989)). Durkin (1994) examina unos 2,500 sistemas expertos y los clasifica por criterios, tales como áreas de aplicación, tareas realizadas, etc. Tal como puede verse en la Figura 1.1, la economía, la industria y la medicina continúan siendo los campos dominantes entre aquellos en los que se utilizan los sistemas expertos. La sección siguiente muestra algunos ejemplos que motivan la aplicación de los sistemas expertos en algunos de estos campos.

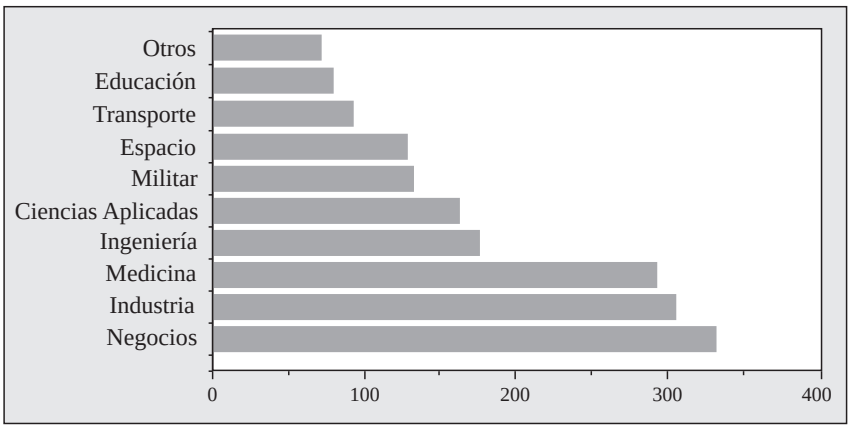


FIGURA 1.1. Campos de aplicación de los sistemas expertos. Adaptado de Durkin (1994) y Castillo, Gutiérrez, y Hadi (1995a).

1.3 Ejemplos Ilustrativos

Los sistemas expertos tienen muchas aplicaciones. En esta seección se dan unos pocos ejemplos ilustrativos del tipo de problemas que pueden resolverse mediante sistemas expertos. Otros ejemplos prácticos se dan a lo largo del libro.

Ejemplo 1.1 Transacciones bancarias. No hace mucho, para hacer una transacción bancaria, tal como depositar o sacar dinero de una cuenta, uno tenía que visitar el banco en horas de oficina. Hoy en día, esas y otras muchas transacciones pueden realizarse en cualquier momento del día o de la noche usando los cajeros automáticos que son ejemplos sencillos de sistemas expertos. De hecho, se pueden realizar estas transacciones desde casa comunicándose con el sistema experto mediante la línea telefónica. ■

Ejemplo 1.2 Control de tráfico. El control de tráfico es una de las aplicaciones más importantes de los sistemas expertos. No hace mucho tiempo, el flujo de tráfico en las calles de una ciudad se controlaba mediante guardias de tráfico que controlaban el mismo en las intersecciones. Hoy se utilizan sistemas expertos que operan automáticamente los semáforos y regulan el flujo del tráfico en las calles de una ciudad y en los ferrocarriles. En la Sección 2.6 y en los ejercicios del Capítulo 2 se dan ejemplos de estos sistemas. ■

Ejemplo 1.3 Problemas de planificación. Los sistemas expertos pueden utilizarse también para resolver problemas complicados de planificación de forma que se optimicen ciertos objetivos como, por ejemplo, la organización y asignación de aulas para la realización de exámenes finales en una gran universidad, de forma tal que se logren los objetivos siguientes:

- Eliminar las coincidencias de asignación simultánea de aulas: Sólo se puede relizar un examen en cada aula al mismo tiempo.
- Asientos suficientes: Un aula asignada para un examen debe tener al menos dos asientos por estudiante.
- Minimizar los conflictos temporales: Minimizar el número de alumnos que tienen exámenes coincidentes.
- Eliminar la sobrecarga de trabajo: Ningún alumno debe tener más de dos exámenes en un periodo de 24 horas.
- Minimizar el número de exámenes realizados durante las tardes.

Otros ejemplos de problemas de planificación que pueden ser resueltos mediante sistemas expertos son la planificación de doctores y enfermeras en un gran hospital, la planificación en una gran factoría, y la planificación de autobuses para las horas de congestión o de días festivos. ■

Ejemplo 1.4 Diagnóstico médico. Una de las aplicaciones más importantes de los sistemas expertos tiene lugar en el campo médico, donde éstos pueden ser utilizados para contestar a las siguientes preguntas:

1. ¿Cómo se puede recoger, organizar, almacenar, poner al día y recuperar la información médica (por ejemplo, registros de pacientes) de una forma eficiente y rápida? Por ejemplo, supóngase que un doctor en un centro médico está interesado en conocer información sobre una cierta enfermedad (E) y tres síntomas asociados (S_1, S_2, y S_3). Se puede utilizar un sistema experto para buscar en la base de datos, extraer y organizar la información deseada. Esta información puede resumirse en tablas tales como la dada en la Tabla 1.1 o en gráficos como el de la Figura 1.2.
2. ¿Cómo se puede aprender de la experiencia? Es decir, cómo se actualiza el conocimiento de los doctores en medicina cuando el número de pacientes que éstos tratan aumenta?
3. Supuesto que un paciente presenta un conjunto de síntomas, ¿cómo se decide qué enfermedad es la que más probablemente tiene el paciente?
4. ¿Cuáles son las relaciones entre un conjunto (normalmente no observable) de enfermedades y un conjunto (observable) de síntomas? En otras palabras, ¿qué modelos pueden utilizarse para describir las relaciones entre los síntomas y las enfermedades?
5. Dado que el conjunto de síntomas conocidos no es suficiente para diagnosticar la enfermedad con cierto grado de certeza, ¿qué información adicional debe ser obtenida (por ejemplo, ¿qué síntomas adicionales deben ser identificados? o ¿qué pruebas médicas deben realizarse?).
6. ¿Cuál es el valor de cada una de éstas piezas de información? En otras palabras, ¿cuál es la contribución de cada uno de los síntomas adicionales o pruebas a la toma de decisión? ■

Ejemplo 1.5 Agentes secretos. Alberto, Luisa, Carmen, y Tomás son agentes secretos, cada uno está en uno de los cuatro países: Egipto, Francia, Japón y España. No se sabe dónde está cada uno de ellos. Por tanto, se ha pedido información y se han recibido los cuatro telegramas siguientes:

- Desde Francia: Luisa está en España.
- Desde España: Alberto está en Francia.
- Desde Egipto: Carmen está en Egipto.
- Desde Japón: Carmen está en Francia.

No se sabe quién es el que ha mandado cada uno de los mensajes, pero se sabe que Tomás miente (¿un agente doble?) y que los demás agentes dicen la verdad.

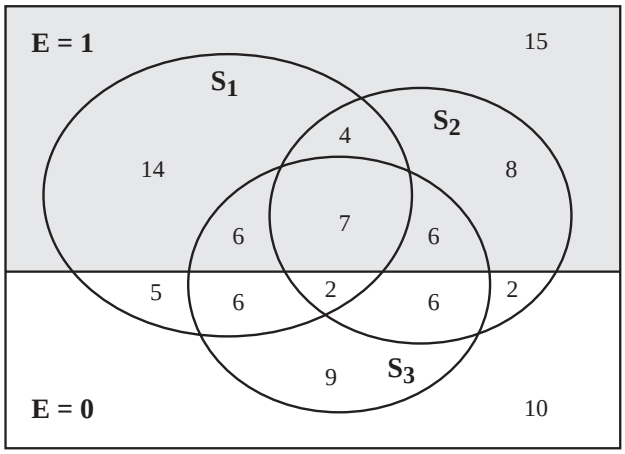


FIGURA 1.2. Una representación gráfica de la distribución de frcuencias de una enfermedad (D) y tres síntomas binarios (S_1, S_2, y S_3) en una base de datos médica.

E	S_1	S_2	S_3	Frecuencia
1	1	1	1	7
1	1	1	0	4
1	1	0	1	6
1	1	0	0	14
1	0	1	1	6
1	0	1	0	8
1	0	0	1	0
1	0	0	0	15
0	1	1	1	2
0	1	1	0	0
0	1	0	1	6
0	1	0	0	5
0	0	1	1	6
0	0	1	0	2
0	0	0	1	9
0	0	0	0	10

TABLA 1.1. Una representación tabular de la distribución de frecuencias de una enfermedad (D) y tres síntomas binarios (S_1, S_2, y S_3) en una base de datos médica (1 representa la presencia y 0 representa la ausencia de la enfermedad o el síntoma indicado).

La cuestión que se desea responder es: ¿Qué agente está en cada país? Aunque se trata de un problema de lógica, que contiene afirmaciones muy simples, su solución no es fácil de obtener a simple vista. En la Sección 2.4.2 se muestra la forma de resolver automáticamente este problema utilizando un conjunto de reglas. ■

1.4 ¿Por Qué los Sistemas Expertos?

El desarrollo o la adquisición de un sistema experto es generalmente caro, pero el mantenimiento y el coste marginal de su uso repetido es relativamente bajo. Por otra parte, la ganancia en términos monetarios, tiempo, y precisión resultantes del uso de los sistemas expertos son muy altas, y la amortización es muy rápida. Sin embargo, antes de desarrollar o adquirir un sistema experto debe realizarse un análisis de factibilidad y de coste-beneficio. Hay varias razones para utilizar sistemas expertos. Las más importantes son:

1. Con la ayuda de un sistema experto, personal con poca experiencia puede resolver problemas que requieren un conocimiento de experto. Esto es también importante en casos en los que hay pocos expertos humanos. Además, el número de personas con acceso al conocimiento aumenta con el uso de sistemas expertos.

2. El conocimiento de varios expertos humanos puede combinarse, lo que da lugar a sistemas expertos más fiables, ya que se obtiene un sistema experto que combina la sabiduría colectiva de varios expertos humanos en lugar de la de uno solo.

3. Los sistemas expertos pueden responder a preguntas y resolver problemas mucho más rápidamente que un experto humano. Por ello, los sistemas son muy valiosos en casos en los que el tiempo de respuesta es crítico.

4. En algunos casos, la complejidad del problema impide al experto humano resolverlo. En otros casos la solución de los expertos humanos no es fiable. Debido a la capacidad de los ordenadores de procesar un elevadísimo número de operaciones complejas de forma rápida y aproximada, los sistemas expertos suministran respuestas rápidas y fiables en situaciones en las que los expertos humanos no pueden.

5. Los sistemas expertos pueden ser utilizados para realizar operaciones monótonas, aburridas e inconfortables para los humanos. En verdad, los sistemas expertos pueden ser la única solución viable en una situación en la que la tarea a realizar desborda al ser humano (por ejemplo, un avión o una cápsula espacial dirigida por un sistema experto).

6. Se pueden obtener enormes ahorros mediante el uso de sistemas expertos.

El uso de los sistemas expertos se recomienda especialmente en las situaciones siguientes:

- Cuando el conocimiento es difícil de adquirir o se basa en reglas que sólo pueden ser aprendidas de la experiencia.
- Cuando la mejora continua del conocimiento es esencial y/o cuando el problema está sujeto a reglas o códigos cambiantes.
- Cuando los expertos humanos son caros o difíciles de encontrar.
- Cuando el conocimiento de los usuarios sobre el tema es limitado.

1.5 Tipos de Sistemas Expertos

Los problemas con los que pueden tratar los sistemas expertos pueden clasificarse en dos tipos: problemas esencialmente deterministas y problemas esencialmente estocásticos. Por ejemplo, aunque el ejemplo 1.1 (transacciones bancariass) y el Ejemplo 1.2 (control de tráfico) pueden contener algunos elementos de incertidumbre, son esencialmente problemas deterministas. Por otra parte, en el campo médico (ver Ejemplo 1.4) las relaciones entre síntomas y enfermedades se conocen sólo con un cierto grado de certeza (la presencia de un conjunto de síntomas no siempre implica la presencia de una enfermedad). Estos tipos de problemas pueden también incluir algunos elementos deterministas, pero se trata fundamentalmente de problemas estocásticos.

Consecuentemente, los sistemas expertos pueden clasificarse en dos tipos principales según la naturaleza de problemas para los que están diseñados: deterministas y estocásticos.

Los problemas de tipo determinista pueden ser formulados usando un conjunto de reglas que relacionen varios objetos bien definidos. Los sistemas expertos que tratan problemas deterministas son conocidos como *sistemas basados en reglas*, porque sacan sus conclusiones basándose en un conjunto de reglas utilizando un mecanismo de *razonamiento lógico*. El Capítulo 2 se dedica a los sistemas expertos basados en reglas.

En situaciones inciertas, es necesario introducir algunos medios para tratar la incertidumbre. Por ejemplo, algunos sistemas expertos usan la misma estructura de los sistemas basados en reglas, pero introducen una medida asociada a la incertidumbre de las reglas y a la de sus premisas. En este caso se pueden utilizar algunas fórmulas de propagación para calcular la incertidumbre asociada a las conclusiones. Durante las últimas décadas han sido propuestas algunas medidas de incertidumbre. Algunos ejemplos

de estas medidas son *los factores de certeza*, usados en las conchas para generar sistemas expertos tales como el sistema experto MYCIN (véase Buchanan y Shortliffe (1984)); la *lógica difusa* (véase, por ejemplo, Zadeh (1983) y Buckley, Siler, y Tucker (1986)); y la *teoría de la evidencia* de Dempster y Shafer (véase Shafer (1976)).

Otra medida intuitiva de incertidumbre es la *probabilidad*, en la que la *distribución conjunta* de un conjunto de variables se usa para describir las relaciones de dependencia entre ellas, y se sacan conclusiones usando fórmulas muy conocidas de la teoría de la probabilidad. Este es el caso del sistema experto PROSPECTOR (véase Duda, Gaschnig, y Hart (1980)), que utiliza el teorema de Bayes para la exploración de mineral.

Los sistemas expertos que utilizan la probabilidad como medida de incertidumbre se conocen como sistemas expertos *probabilísticos* y la estrategia de razonamiento que usan se conoce como *razonamiento probabilístico*, o *inferencia probabilística*.. Este libro está dedicado a los sistemas expertos de tipo probabilístico. Otros libros que sirven para introducirse de forma general en otras medidas de incertidumbre son Buchanan y Shortliffe (1984), Waterman (1985), Pearl (1988), Jackson (1990), Neapolitan (1990), Castillo y Álvarez (1991), Durkin (1994) y Jensen (1996).

En los comienzos de los sistemas expertos de tipo probabilístico surgieron varios obstáculos, debido a las dificultades encontradas para definir la distribución de probabilidad conjunta de las variables. Ello ha ralentizado su desarrollo. Con la introducción de los *modelos de redes probabilísticas*, estos obstáculos se han superado y los sistemas expertos probabilísticos han vuelto de forma espectacular durante las dos últimas décadas. Estos modelos, que incluyen las redes de Markov y las Bayesianas, se basan en una representación gráfica de las relaciones entre las variables. Esta representación conduce no sólo a formas más eficientes de definir la distribución conjunta de probabilidad sino también a una propagación de incertidumbre muy eficiente, que permite sacar conclusiones. Ejemplos de tales conchas para el desarrollo de sistemas expertos son el sistema HUGIN (véase Andersen y otros (1989)) y *X-pert Nets*,[1] que ha sido desarrollado por los autores de este libro.

1.6 Componentes de un Sistema Experto

Las definiciones de sistemas expertos dadas en la Sección 1.2 se entienden mejor cuando se examinan las principales componentes de los sistemas expertos. Estas componentes se muestran esquemáticamente en la Figura 1.3 y se explican seguidamente.

[1] Ésta y otras conchas para sistemas expertos pueden obtenerse de la dirección WWW http://ccaix3.unican.es/~AIGroup.

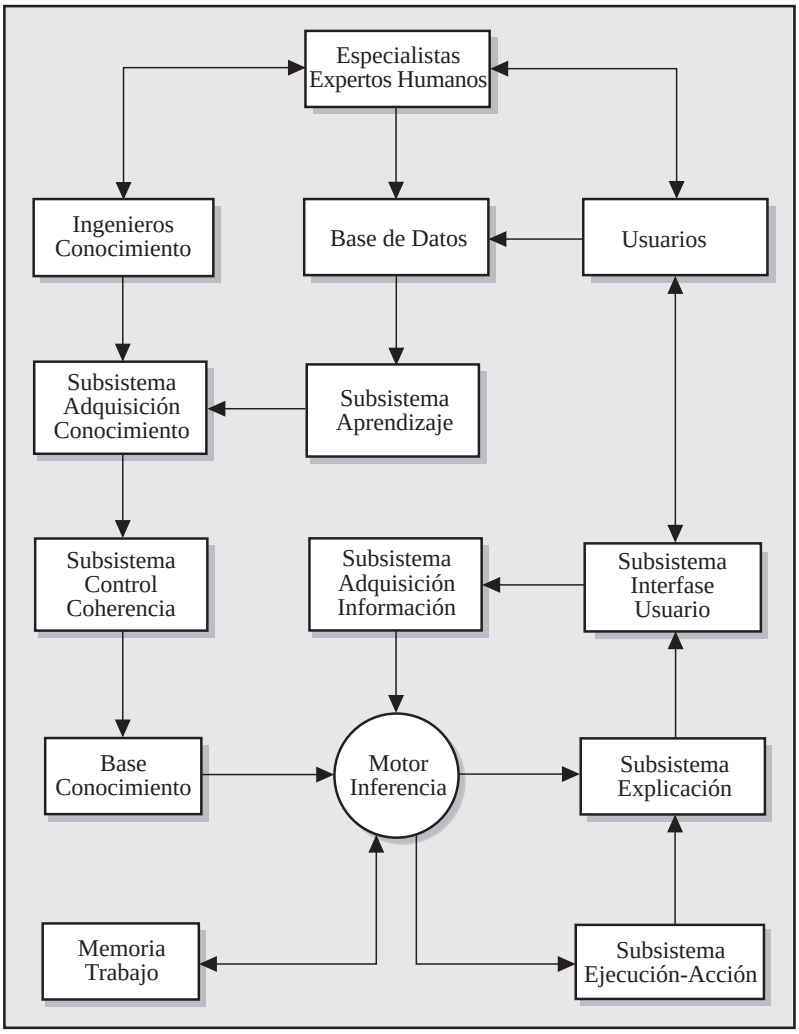


FIGURA 1.3. Componentes típicos de un sistema experto. Las flechas representan el flujo de la información.

1.6.1 La Componente Humana

Un sistema experto es generalmente el resultado de la colaboración de uno o varios *expertos humanos especialistas en el tema de estudio* y los *ingenieros del conocimiento*, con los *usuarios* en mente. Los expertos humanos suministran el conocimiento básico en el tema de interés, y los ingenieros del conocimiento trasladan este conocimiento a un lenguaje, que el sistema experto pueda entender. La colaboración de los expertos humanos, los ingenieros del conocimiento y los usuarios es, quizás, el elemento más

importante en el desarrollo de un sistema experto. Esta etapa requiere una enorme dedicación y un gran esfuerzo debido a los diferentes lenguajes que hablan las distintas partes y a las diferentes experiencias que tienen.

1.6.2 La Base de Conocimiento

Los especialistas son responsables de suministrar a los ingenieros del conocimiento una base de conocimiento ordenada y estructurada, y un conjunto de relaciones bien definidas y explicadas. Esta forma estructurada de pensar requiere que los expertos humanos repiensen, reorganicen, y reestructuren la base de conocimiento y, como resultado, el especialista se convierte en un mejor conocedor de su propio campo de especialidad.

Hay que diferenciar entre *datos* y *conocimiento*. El conocimiento se refiere a afirmaciones de validez general tales como reglas, distribuciones de probabilidad, etc. Los datos se refieren a la información relacionada con una aplicación particular. Por ejemplo, en diagnóstico médico, los síntomas, las enfermedades y las relaciones entre ellos, forman parte del conocimiento, mientras los síntomas particulares de un paciente dado forman parte de los datos. Mientras el conocimiento es permanente, los datos son efímeros, es decir, no forman parte de la componente permanente de un sistema y son destruídos después de usarlos. El conocimiento se almacena en la base de conocimiento y los datos se almacenan en la *memoria de trabajo*. Todos los procedimientos de los diferentes sistemas y subsistemas que son de carácter transitorio se almacenan también en la memoria de trabajo.

1.6.3 Subsistema de Adquisición de Conocimiento

El subsistema de adquisición de conocimiento controla el flujo del nuevo conocimiento que fluye del experto humano a la base de datos. El sistema determina qué nuevo conocimiento se necesita, o si el conocimiento recibido es en realidad nuevo, es decir, si debe incluirse en la base de datos y, en caso necesario, incorpora estos conocimientos a la misma.

1.6.4 Control de la Coherencia

El subsistema de control de la coherencia ha aparecido en los sistemas expertos muy recientemente. Sin embargo, es una componente esencial de un sistema experto. Este susbsistema controla la consistencia de la base de datos y evita que unidades de conocimiento inconsistentes entren en la misma. En situaciones complejas incluso un experto humano puede formular afirmaciones inconsistentes. Por ello, sin un subsistema de control de la coherencia, unidades de conocimiento contradictorio pueden formar parte de la base de conocimiento, dando lugar a un comportamiento insatisfactorio del sistema. Es también bastante común, especialmente en sistemas con

mecanismos de propagación de incertidumbre, que se llegue a conclusiones absurdas o en conflicto como, por ejemplo, situaciones en las que el sistema genera probabilidades mayores que la unidad o negativas. Por ello, el subsistema de control de la coherencia comprueba e informa a los expertos de las inconsistencias. Por otra parte, cuando se solicita información de los expertos humanos, éste subsistema informa sobre las restricciones que ésta debe cumplir para ser coherente con la existente en la base de conocimiento. De esta forma, ayuda a los expertos humanos a dar información fiable.

1.6.5 El Motor de Inferencia

El motor de inferencia es el corazón de todo sistema experto. El cometido principal de esta componente es el de sacar conclusiones aplicando el conocimiento a los datos. Por ejemplo, en diagnóstico médico, los síntomas de un paciente (datos) son analizados a la luz de los síntomas y las enfermedades y de sus relaciones (conocimiento).

Las conclusiones del motor de inferencia pueden estar basadas en *conocimiento determinista* o *conocimiento probabilístico*. Como puede esperarse, el tratamiento de situaciones de incertidumbre (probabilísticas) puede ser considerablemente más difícil que el tratamiento de situaciones ciertas (deterministas). En muchos casos, algunos hechos (datos) no se conocen con absoluta certeza. Por ejemplo, piénsese en un paciente que no está seguro de sus síntomas. Puede darse el caso de tener que trabajar con conocimiento de tipo no determinista, es decir, de casos en los que se dispone sólo de información aleatoria o difusa. El motor de inferencia es también responsable de la propagación de este conocimiento incierto. De hecho, en los sistemas expertos basados en probabilidad, la propagación de incertidumbre es la tarea principal del motor de inferencia, que permite sacar conclusiones bajo incertidumbre. Esta tarea es tan compleja que da lugar a que ésta sea probablemente la componente más débil de casi todos los sistemas expertos existentes. Por esta razón, la mayor parte de este libro se dedica al análisis y resolución del problema de la propagación de incertidumbre.

1.6.6 El Subsistema de Adquisición de Conocimiento

Si el conocimiento inicial es muy limitado y no se pueden sacar conclusiones, el motor de inferencia utiliza el *subsistema de adquisición de conocimiento* para obtener el conocimiento necesario y continuar con el proceso de inferencia hasta que se hayan sacado conclusiones. En algunos casos, el usuario puede suministrar la información requerida para éste y otros objetivos. De ello resulta la necesidad de una *interfase de usuario* y de una comprobación de la consistencia de la información suministrada por el usuario antes de introducirla en la memoria de trabajo.

1.6.7 Interfase de Usuario

La interfase de usuario es el enlace entre el sistema experto y el usuario. Por ello, para que un sistema experto sea una herramienta efectiva, debe incorporar mecanismos eficientes para mostrar y obtener información de forma fácil y agradable. Un ejemplo de la información que tiene que ser mostrada tras el trabajo del motor de inferencia, es el de las conclusiones, las razones que expliquen tales conclusiones y una explicación de las acciones iniciadas por el sistema experto. Por otra parte, cuando el motor de inferencia no puede concluir debido, por ejemplo, a la ausencia de información, la interfase de usuario es un vehículo para obtener la información necesaria del usuario. Consecuentemente, una implementación inadecuada de la interfase de usuario que no facilite este proceso minaría notablemente la calidad de un sistema experto. Otra razón de la importancia de la interfase de usuario es que los usuarios evalúan comúnmente los sistemas expertos y otros sistemas por la calidad de dicha interfase más que por la del sistema experto mismo, aunque no se debería juzgar la calidad de un libro por su portada. Los lectores que estén interesados en el diseño de una interfase de usuario pueden consultar los libros de Shneiderman (1987) y Brown y Cunningham (1989).

1.6.8 El Subsistema de Ejecución de Órdenes

El *subsistema de ejecución de órdenes* es la componente que permite al sistema experto iniciar acciones. Estas acciones se basan en las conclusiones sacadas por el motor de inferencia. Como ejemplos, un sistema experto diseñado para analizar el tráfico ferroviario puede decidir retrasar o parar ciertos trenes para optimizar el tráfico global, o un sistema para controlar una central nuclear puede abrir o cerrar ciertas válvulas, mover barras, etc., para evitar un accidente. La explicación de las razones por las que se inician estas acciones pueden darse al usuario mediante el *subsistema de explicación.*

1.6.9 El Subsistema de Explicación

El usuario puede pedir una explicación de las conclusiones sacadas o de las acciones iniciadas por el sistema experto. Por ello, es necesario un subsistema que explique el proceso seguido por el motor de inferencia o por el subsistema de ejecución. Por ejemplo, si un cajero automático decide rechazar la palabra clave (una acción), la máquina puede mostrar un mensaje (una explicación) como la siguiente:

> *¡Lo siento!, su palabra clave es todavía incorrecta tras tres intentos.*
> *Retenemos su tarjeta de crédito, para garantizar su seguridad.*
> *Por favor, póngase en contacto con su banco en horas de oficina.*

En muchos dominios de aplicaciones, es necesaria la explicación de las conclusiones debido a los riesgos asociados con las acciones a ejecutar. Por ejemplo, en el campo del diagnóstico médico, los doctores son responsable últimos de los diagnósticos, independientemente de las herramientas técnicas utilizadas para sacar conclusiones. En estas situaciones, sin un subsistema de explicación, los doctores pueden no ser capaces de explicar a sus pacientes las razones de su diagnóstico.

1.6.10 El Subsistema de Aprendizaje

Una de las principales características de un sistema experto es su capacidad para aprender. Diferenciaremos entre aprendizaje estructural y aprendizaje paramétrico. Por *aprendizaje estructural* nos referimos a algunos aspectos relacionados con la estructura del conocimiento (reglas, distribuciones de probabilidad, etc.). Por ello, el descubrimiento de nuevos síntomas relevantes para una enfermedad o la inclusión de una nueva regla en la base de conocimiento son ejemplos de aprendizaje estructural. Por *aprendizaje paramétrico* nos referimos a estimar los parámetros necesarios para construir la base de conocimiento. Por ello, la estimación de frecuencias o probabilidades asociadas a síntomas o enfermedades es un ejemplo de aprendizaje paramétrico.

Otra característica de los sistemas expertos es su habilidad para obtener *experiencia* a partir de los *datos* disponibles. Estos datos pueden ser obtenidos por expertos y no expertos y pueden utilizarse por el *subsistema de adquisición de conocimiento* y por el *subsistema de aprendizaje*.

De las componentes antes mencionadas puede verse que los sistemas expertos pueden realizar varias tareas. Estas tareas incluyen, pero no se limitan a, las siguientes:

- Adquisición de conocimiento y la verificación de su coherencia; por lo que el sistema experto puede ayudar a los expertos humanos a dar conocimiento coherente.
- Almacenar (memorizar) conocimiento.
- Preguntar cuándo se requiere nuevo conocimiento.
- Aprender de la base de conocimiento y de los datos disponibles.
- Realizar inferencia y razonamiento en situaciones deterministas y de incertidumbre.
- Explicar conclusiones o acciones tomadas.
- Comunicar con los expertos y no expertos humanos y con otros sistemas expertos.

1.7 Desarrollo de un Sistema Experto

Weiss y Kulikowski (1984) sugieren las etapas siguientes para el diseño e implementación de un sistema experto (ver también Hayes-Roth, Waterman, y Lenat (1983), Luger y Stubblefield (1989), y la Figura 1.4):

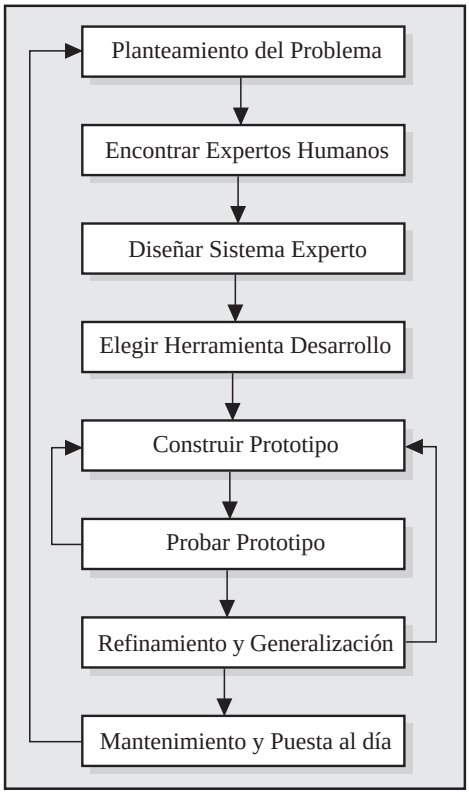


FIGURA 1.4. Etapas en el dasarrollo de un sistema experto.

1. **Planteamiento del problema**. La primera etapa en cualquier proyecto es normalmente la definición del problema a resolver. Puesto que el objetivo principal de un sistema experto es responder a preguntas y resolver problemas, esta etapa es quizás la más importante en el desarrollo de un sistema experto. Si el sistema está mal definido, se espera que el sistema suministre respuestas erróneas.

2. **Encontrar expertos humanos que puedan resolver el problema**. En algunos casos, sin embargo, las bases de datos pueden jugar el papel del experto humano.

3. **Diseño de un sistema experto**. Esta etapa incluye el diseño de estructuras para almacenar el conocimiento, el motor de inferencia, el subsistema de explicación, la interfase de usuario, etc.

4. **Elección de la herramienta de desarrollo, concha, o lenguaje de programación**. Debe decidirse si realizar un sistema experto a medida, o utilizar una concha, una herramienta, o un lenguaje de programación. Si existiera una concha satisfaciendo todos los requerimientos del diseño, ésta debería ser la elección, no sólo por razones de tipo financiero sino también por razones de fiabilidad. Las conchas y herramientas comerciales están sujetas a controles de calidad, a los que otros programas no lo están.

5. **Desarrollo y prueba de un prototipo**. Si el prototipo no pasa las pruebas requeridas, las etapas anteriores (con las modificaciones apropiadas) deben ser repetidas hasta que se obtenga un prototipo satisfactorio.

6. **Refinamiento y generalización**. En esta etapa se corrigen los fallos y se incluyen nuevas posibilidades no incorporadas en el diseño inicial.

7. **Mantenimiento y puesta al día**. En esta etapa el usuario plantea problemas o defectos del prototipo, corrige errores, actualiza el producto con nuevos avances, etc.

Todas estas etapas influyen en la calidad del sistema experto resultante, que siempre debe ser evaluado en función de las aportaciones de los usuarios. Para el lector interesado en estos temas recomendamos la lectura de los trabajos de O'Keefe, Balci y Smith (1987), Chandrasekaran (1988) y Preece (1990).

1.8 Otras Áreas de la IA

En esta sección se da una breve descripción panorámica del ámbito y dominio de algunas áreas de la IA distintas de la de los sistemas expertos. Puesto que este libro está dedicado exclusivamente a sistemas expertos, se dan algunas referencias para que el lector interesado pueda acceder a otras áreas de la IA. Debe tenerse en mente que ésta no es una lista exhaustiva de todas las áreas de la IA y que la IA es un campo que se desarrolla muy rápidamente, y emergen continuamente nuevas ramas para tratar las nuevas situaciones de esta ciencia que no para de crecer.

Hay varios libros que dan una visión general de la mayoría de los temas incluídos en la IA. El multivolumen *Handbook of Artificial Intelligence* editado por Barr y Feigenbaum (1981, 1982) (volúmenes 1 y 2) y por Cohen y Feigenbaum (1982) (volumen 3), y la *Encyclopedia of Artificial Intelligence*,

editado por Shapiro (1987) contienen discusiones detalladas de varios de los temas de la IA. Hay otros muchos libros que cubren las áreas de IA. Por mencionar unos pocos, citamos a: Charniak y McDermott (1985), Rich y Knight (1991), Winston (1992), Ginsberg (1993), Russell y Norvig (1995).

Como consecuencia de la intensa investigación realizada en el área de la IA, hay también un número creciente de revistas que publican artículos en los distintos campos de la IA y temas relacionados con ella. Algunas de estas revistas son: *Applied Artificial Intelligence*, *Applied Intelligence*, *Artificial Intelligence*, *Artificial Intelligence Magazine*, *International Journal of Intelligent Systems*.

Por otra parte, revistas tales como *Artificial Intelligence in Medicine*, *Biocybernetics and Biomedical Engineering*, *Cybernetics and Systems*, *Fuzzy Sets and Systems*, *IEEE Expert*, *IEEE Transactions on Systems, Man and Cybernetics*, *International Journal for Artificial Intelligence in Engineering*, *International Journal of Approximate Reasoning*, *International Journal of Computer Vision*, *International Journal of Expert Systems*, *Machine Learning*, *Networks*, *Neural Networks*, y *Pattern Recognition Letters* se especializan en un tema o en un cierto dominio de aplicaciones. [2]

1.8.1 Representación del Conocimiento

Hay muchas fuentes de información o conocimiento relacionadas con la IA. El campo de la representación del conocimiento se refiere a los mecanismos para representar y manipular esta información. Los esquemas de representación resultantes deberían permitir una búsqueda o una operación eficiente de los mecanismos de inferencia. Por ejemplo, en algunos casos la información puede ser representada mediante objetos (o variables) y por reglas lógicas (que expresan relaciones entre los objetos). Por ello, esta representación puede manipularse usando análisis lógico. Éste es el mecanismo de representación del conocimiento utilizado, por ejemplo, en los sistemas expertos basados en reglas (Capítulo 2). Para tener una visión general de las diferentes metodologías de representación del conocimiento véase, por ejemplo, Bachman, Levesque, y Reiter (1991), Bench-Capon (1990), y los *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning* (KR-89, 91, 92, y 94) publicados por Morgan y Kaufmann Publishers.

[2] Una lista que contiene la mayoría de revistas en el campo de la IA se puede obtener en la dirección WWW "http://ai.iit.nrc.ca/ai_journals.html"; véase también "http://www.bus.orst.edu/faculty/brownc/aies/journals.htm."

1.8.2 Planificación

Cuando se trata con problemas complejos, es importante dividir las tareas en partes más pequeñas que sean más fáciles de manejar. Los métodos de planificación analizan diferentes estrategias para descomponer un problema dado, resolver cada una de sus partes, y llegar a una solución final. La interacción entre las partes dependerá del grado de descomponibilidad del problema. Por otra parte, el comienzo de la computación paralela, capaz de realizar varias tareas simultáneamente, da lugar a nuevos problemas que requieren estrategias especiales de planificación. En esta situación, el objetivo consiste en dividir las tareas de forma adecuada para resolver muchas partes simultáneamente. El trabajo editado por Allen, Hendler, y Tate (1990), da una descripción general de este campo. Por otra parte, la colección de artículos editada por Bond y Gasser (1988) está dedicada al *razonamiento paralelo*, también conocido como *razonamiento distribuido.*

1.8.3 Demostración Automática de Teoremas

La capacidad de hacer deducciones lógicas fue considerada durante mucho tiempo como una posibilidad reservada a la mente humana. La investigación desarrollada en los años 1960 en el área de la demostración automática de teoremas ha mostrado que esta tarea puede ser realizada por máquinas programables. Tales máquinas son capaces no sólo de modificar el conocimiento existente, sino también de obtener conclusiones nuevas. En primer lugar, *los demostradores de teoremas* han sido utilizados en varios campos de las matemáticas, tales como la Lógica, la Geometría, etc. El campo de la Matemática constituye un área natural para esta metodología por la existencia de mecanismos de deducción y de una extensa base de conocimiento. Sin embargo, los demostradores de teoremas, pueden ser adaptados para resolver problemas de otras áreas de conocimiento con estas dos mismas características. Una introducción general a este tema se da en Wos y otros (1984) y Bundy (1983), que contiene el código en Prolog de un demostrador de teoremas muy simple. Referencias más recientes son las de Newborn (1994), Almulla (1995) y las incluídas en ellos.

1.8.4 Los Juegos Automatizados

Los juegos automatizados constituyen un ejemplo de una de las más antiguas y fascinantes áreas de la IA (véase, por ejemplo, Newell, Shaw, y Simon (1963)). Los juegos por computador (tales como el ajedrez, backgammon, y los de cartas) han visto un desarrollo masivo en los últimos años. Por ejemplo, los programas de juegos de ajedrez son capaces de competir e incluso vencer a bien conocidos maestros. El juego automático requiere un estudio teórico profundo y tiene varias aplicaciones en otras áreas tales como *métodos de búsqueda*, *optimización*, etc. Una buena discusión

de este campo, que incluye además referencias históricas de interés, puede encontrarse en Levy (1988).

1.8.5 Reconocimiento de patrones

El reconocimiento de patrones trata de diferentes técnicas de clasificación para identificar los subgrupos, o conglomerados, con características comunes en cada grupo. El grado de asociación de cualquiera de los objetos con cada uno de los grupos suministra un modo de sacar conclusiones. Por ello, los algoritmos desarrollados en este área son herramientas útiles para tratar con varios problemas de muchos campos tales como reconocimiento de imágenes, reconocimiento de señales, diagnosis de fallos de equipos, control de procesos, etc. Para una introducción general a este campo véase Sing-Tze (1984) y Niemann (1990) y Patrick y Fattu (1984) para una discusión de tipo estadístico.

1.8.6 Reconocimiento de la Voz

La voz es, con mucho, el medio de comunicación más usado por el hombre. El reconocimiento de voz trata del problema de procesar el lenguaje hablado y capturar los diferentes elementos semánticos que forman la conversación. Los problemas asociados con las diferentes pronunciaciones y tonos de voz son los principales obtáculos que esta disciplina tiene que afrontar. Una introducción general al problema del reconocimiento de voz se da en Rabiner y Juang (1993).

1.8.7 Procesamiento de Lenguaje Natural

Un objetivo del procesamiento del lenguaje natural consiste en extraer tanta información como sea posible de un texto escrito. Con el uso creciente de los ordenadores en el tratamiento de la información, el lenguaje escrito está jugando un papel muy importante como medio de comunicación. Puesto que el reconocimiento de la voz es inherentemente un problema más difícil, resulta necesario disponer de un procesado eficiente del lenguaje escrito. El problema inverso del procesamiento del lenguaje es el de la *generación de lenguaje*, es decir, dotar a los computadores de capacidades para generar sentencias de lenguaje natural en vez de mensajes de tipo telegráfico. La combinación de estas dos tareas permitiría, por ejemplo, la posibilidad de traducir textos escritos en diferentes idiomas, lo que se conoce como *traducción asistida por ordenador*. Una referencia clásica a este campo es Schank y Abelson (1977). También se pueden encontrar descripciones interesantes de procesamiento del lenguaje natural en Allen (1995) y McKeown (1985).

1.8.8 Visión Artificial

Uno de los objetivos de la visión artificial es la posibilidad de usar ordenadores para localizar y reconocer automáticamente objetos en tres dimensiones. Muchas otras áreas de la IA tales como la representación del conocimiento, el reconocimiento de patrones, y las redes neuronales juegan un papel esencial en la visión artificial. Los muy significativos avances técnicos producidos durante la última década has sido aplicados a varios sistemas comerciales utilizados en fabricación, inspección, tareas de guía, etc. Para una introducción general a este área, véase Fischler y Firschein (1987) y Shapiro y Rosenfeld (1992).

1.8.9 Robótica

La robótica es una de las áreas de la IA más populares. Los robots combinan elementos mecánicos, sensores, y ordenadores que les permiten tratar con objetos reales y realizar muchas tareas de forma precisa, rápida y cómoda. Por ello, se puede pensar en los robots como ordenadores que interaccionan con el mundo real. Una revisión general de la robótica se presenta en McKerrow (1991), mientras Jones y Flynn (1993) tratan el tema de las aplicaciones prácticas.

1.8.10 Redes Neuronales

Las redes neuronales se crearon con el objetivo de reproducir de forma básica las funciones elementales del cerebro humano. Las arquitecturas en red con un gran número de conexiones entre varias capas de procesadores fueron introducidas para reproducir la estructura del cerebro humano. La información contenida en una red neuronal se codifica en la estructura de la red y en los pesos de las conexiones. Por tanto, en una situación particular, los pesos de las conexiones tienen que modificarse para reproducir la salida deseada. Esta tarea de aprendizaje se consigue mediante una técnica de aprender por analogía, es decir, el modelo se entrena para reproducir las salidas de un conjunto de señales de entrenamiento con el objetivo de codificar de esta forma la estructura del fenómeno. La aparición de ordenadores rápidos en los que pudieran simularse redes grandes y complejas, y el descubrimiento de potentes algoritmos de aprendizaje han sido las causas que han posibilitado el desarrollo rápido de este área de conocimiento. Para una introducción ver, por ejemplo, Freeman y Skapura (1991) y Lisboa (1992).

1.9 Conclusiones

A partir de la breve descripción de las variadas áreas de la IA mostradas en este capítulo, se puede ver que éstas están interrelacionadas. Por ejemplo, la robótica utiliza otras áreas de la IA tales como la visión automática y el reconocimiento de patrones o de la voz. El área de la IA, como un todo, es altamente interdisciplinar. Por ejemplo, los sistemas expertos requieren varios conceptos de la ciencia del computador, la lógica matemática, la teoría de grafos, la teoría de la probabilidad y la estadística. Por ello, el trabajo en este campo requiere la colaboración de muchos investigadores en diferentes áreas de especialización.

Capítulo 2
Sistemas Basados en Reglas

2.1 Introducción

En nuestra vida diaria encontramos muchas situaciones complejas gobernadas por reglas deterministas: sistemas de control de tráfico, sistemas de seguridad, transacciones bancarias, etc. Los sistemas basados en reglas son una herramienta eficiente para tratar estos problemas. Las reglas deterministas constituyen la más sencilla de las metodologías utilizadas en sistemas expertos. La base de conocimiento contiene el conjunto de reglas que definen el problema, y el motor de inferencia saca las conclusiones aplicando la lógica clásica a estas reglas. Una introducción general a los sistemas expertos basados en reglas, puede encontrarse , por ejemplo, en Buchanan y Shortliffe (1984), Castillo y Álvarez (1991), Durkin (1994), Hayes-Roth (1985), Waterman (1985), y también en el trabajo editado por García y Chien (1991). El libro de Pedersen (1989) muestra un enfoque práctico e incluye varios algoritmos.

Este capítulo presenta los conceptos básicos que forman parte de los sistemas expertos basados en reglas. No se pretende realizar una descripción detallada de estos sistemas, para la que hay libros mucho más completos que éste, sino sólo introducir al lector, de forma simple e intuitiva, en esta metodología. La intención de este capítulo es mostrar cómo los sistemas probabilísticos pueden considerarse como una generalización de los sistemas basados en reglas. La Sección 2.2 describe la base de conocimiento de los sistemas expertos basados en reglas y da una definición y ejemplos de reglas, que constituyen el corazón de la base de conocimiento. Seguidamente, se

Objeto	Conjunto de valores posibles
Nota	{0, 1, ..., 10}
Calificación	{sobresaliente, notable, aprobado, suspenso}
Puesto	{0, 1, ..., 100}
Admitir	{sí, pendiente, no}
Notificar	{sí, no}

TABLA 2.1. Un ejemplo de objetos con sus posibles valores.

discute cómo opera el motor de inferencia (Sección 2.3), cómo trabaja el subsistema de control de la coherencia (Sección 2.4), y cómo se explican las conclusiones sacadas por el motor de inferencia (Sección 2.5). La Sección 2.6, muestra un ejemplo de aplicación. Finalmente, la Sección 2.7 muestra algunas de las limitaciones de los sistemas expertos basados en reglas.

2.2 La Base de Conocimiento

En los sistemas basados en reglas intervienen dos elementos importantes: la base de conocimiento y los datos. Los datos están formados por la evidencia o los hechos conocidos en una situación particular. Este elemento es dinámico, es decir, puede cambiar de una aplicación a otra. Por esta razón, no es de naturaleza permanente y se almacena en la memoria de trabajo.

En situaciones deterministas, las relaciones entre un conjunto de objetos pueden ser representadas mediante un conjunto de reglas. El conocimiento se almacena en la base de conocimiento y consiste en un conjunto de objetos y un conjunto de reglas que gobiernan las relaciones entre esos objetos. La información almacenada en la base de conocimiento es de naturaleza permanente y estática, es decir, no cambia de una aplicación a otra, a menos que se incorporen al sistema experto elementos de aprendizaje.

Para dar una idea intuitiva de lo que es una regla, supóngase que se tiene un conjunto de *objetos* y, por simplicidad, que cada objeto puede tener uno y sólo uno de un conjunto de posibles valores. Ejemplos de objetos con sus posibles valores se dan en la Tabla 2.1.

Seguidamente se dan unos pocos ejemplos de reglas:

Regla 1: Si nota > 9, entonces calificación = sobresaliente.
Regla 2: Si puesto < 20 o nota > 7, entonces Admitir = sí y Notificar = sí.

Cada una de las reglas anteriores relaciona dos o más objetos y está formada por las partes siguientes:

- La *premisa* de la regla, que es la *expresión lógica* entre las palabras clave *si* y *entonces*. La premisa puede contener una o más afirmaciones

objeto-valor conectadas con operadores lógicos *y*, *o*, o *no*. Por ejemplo, la premisa de la Regla 1 consta de una única afirmación objeto-valor, mientras que las premisas de la Regla 2 constan de dos afirmaciones objeto-valor conectadas por un operador lógico.

- La *conclusión* de la regla, que es la expresión lógica tras la palabra clave *entonces*.

Los ejemplos anteriores facilitan la definición siguiente de regla.

Definición 2.1 Regla. *Una regla es una afirmación lógica que relaciona dos o más objetos e incluye dos partes, la premisa y la conclusión. Cada una de estas partes consiste en una expresión lógica con una o más afirmaciones objeto-valor conectadas mediante los operadores lógicos* y, o, *o* no.

Una regla se escribe normalmente como "Si *premisa*, entonces *conclusión*". En general, ambas, la premisa y la conclusión de una regla, pueden contener afirmaciones múltiples objeto-valor. Una expresión lógica que contiene sólo una afirmación objeto-valor se denomina *expresión lógica simple*; en caso contrario, la expresión se dice *expresión lógica compuesta*. Por ejemplo, las expresiones lógicas en ambas, premisa y conclusión de la Regla 1, son simples, mientras que las expresiones lógicas de las premisas y la conclusión de la Regla 2 es compuesta. Correspondientemente, una regla que contiene solamente expresiones lógicas simples se denomina una *regla simple*; en otro caso, se llama *regla compuesta*. Por ejemplo, la Regla 1 es simple, mientras que la Reglas 2 es compuesta.

Ejemplo 2.1 Cajero Automático. Como ejemplo de problema determinista que puede ser formulado usando un conjunto de reglas, considérese una situación en la que un usuario (por ejemplo, un cliente) desea sacar dinero de su cuenta corriente mediante un cajero automático (CA). En cuanto el usuario introduce la tarjeta en el CA, la máquina la lee y la verifica. Si la tarjeta no es verificada con éxito (por ejemplo, porque no es legible), el CA devuelve la tarjeta al usuario con el mensaje de error correspondiente. En otro caso, el CA pide al usuario su número de identificación personal (NIP). Si el número fuese incorrecto, se dan tres oportunidades de corregirlo. Si el NIP es correcto, el CA pregunta al usuario cuánto dinero desea sacar. Para que el pago se autorice, la cantidad solicitada no debe exceder de una cierta cantidad límite diaria, además de haber suficiente dinero en su cuenta.

En este caso se tienen siete objetos, y cada objeto puede tomar uno y sólo un valor de entre sus posibles valores. La Tabla 2.2 muestra estos objetos y sus posibles valores.

La Figura 2.1 muestra siete reglas que gobiernan la estrategia que el CA debe seguir cuando un usuario intenta sacar dinero de su cuenta. En la Regla 1, por ejemplo, la premisa consiste en seis afirmaciones objeto-valor conectadas mediante el operador lógico *y*, lo que indica que la premisa

Objeto	Conjunto de posibles valores
Tarjeta	{verificada, no verificada}
Fecha	{expirada, no expirada}
NIP	{correcto, incorrecto}
Intentos	{excedidos, no excedidos}
Balance	{suficiente, insuficiente}
Límite	{excedido, no excedido}
Pago	{autorizado, no autorizado}

TABLA 2.2. Objetos y posibles valores para el ejemplo del cajero automático.

es cierta si las seis afirmaciones lo son. Por ello, la Regla 1 relaciona el objeto *Pago* (en la conclusión) con los demás objetos. Según la Regla 1, la acción que debe iniciar el CA es dar el dinero al usuario si la tarjeta se ha verificado correctamente, la fecha no ha expirado, el NIP es correcto, el número de intentos para dar el NIP correcto no se ha excedido y la cantidad solicitada no excede ni la cantidad disponible ni el límite máximo diario. Las expresiones lógicas en cada una de las restantes reglas de la Figura 2.1 constan de una sola afirmación. Nótese que la Regla 1 indica cuándo debe permitirse el pago, y las restantes cuándo debe rechazarse. ∎

Ejemplo 2.2 Gente famosa. Supóngase que se dispone de una base de datos consistente en N individuos. Para cada individuo, la base de datos contiene cuatro atributos: nombre, sexo, nacionalidad y profesión. Supóngase que la base de datos muestra sólo si una persona es americana, política y/o si es mujer. Cada uno estos atributos es binario (toma sólo dos valores posibles). En este caso, la base de datos puede contener, como mucho, $2^3 = 8$ conjuntos disjuntos. Estos conjuntos se muestran en la Figura 2.2. La figura muestra también el nombre de una persona en cada subconjunto. La Tabla 2.3 da un ejemplo de una base de datos que contiene $N = 8$ personas famosas. En este caso se tienen cuatro objetos: $Nombre$, $Americano$, $Político$, y $Mujer$. El primer objeto puede tomar uno de N posibles valores (los nombres de cada persona) y cada uno de los tres últimos objetos pueden tomar el valor *sí* o el valor *no*.

A partir de la Tabla 2.3 se pueden construir reglas para identificar a cada persona, resultando un total de ocho reglas. Por ejemplo, la regla siguiente corresponde al presidente Clinton:

- Regla 1: Si $Nombre = Clinton$, entonces $Americano =$ *sí* y $Político =$ *sí* y $Mujer = no$.

Las restantes siete reglas pueden construirse de forma análoga. ∎

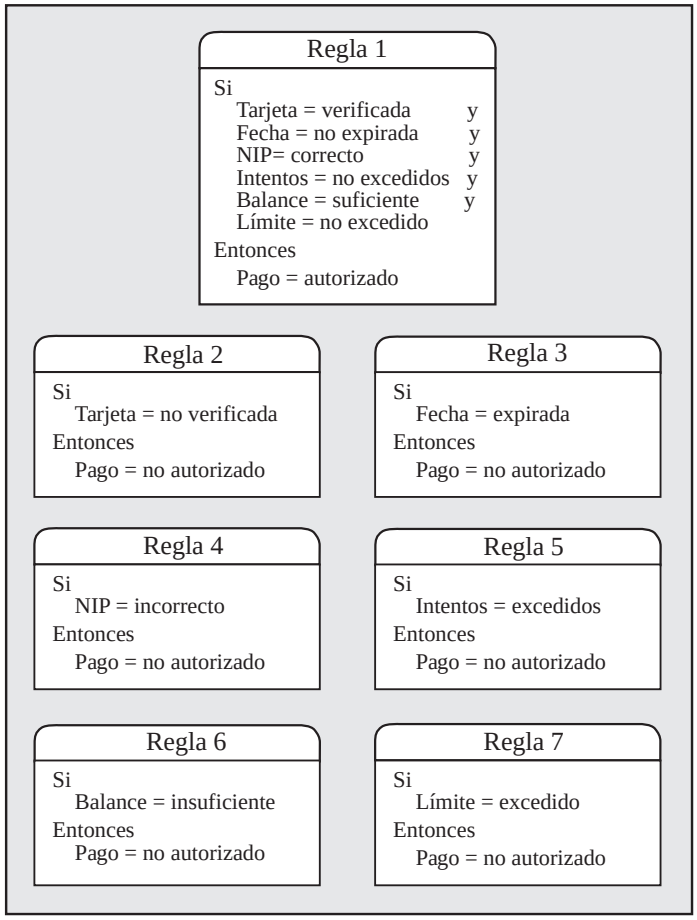


FIGURA 2.1. Ejemplos de reglas para sacar dinero de un cajero automático.

Nombre	Americano	Político	Mujer
Barbara Jordan	sí	sí	sí
Bill Clinton	sí	sí	no
Barbara Walters	sí	no	sí
Mohammed Ali	sí	no	no
Margaret Thatcher	no	sí	sí
Anwar El-Sadat	no	sí	no
Marie Curie	no	no	sí
Pablo Picasso	no	no	no

TABLA 2.3. Una base de datos mostrando cuatro objetos y sus valores correspondientes para el ejemplo de las personas famosas.

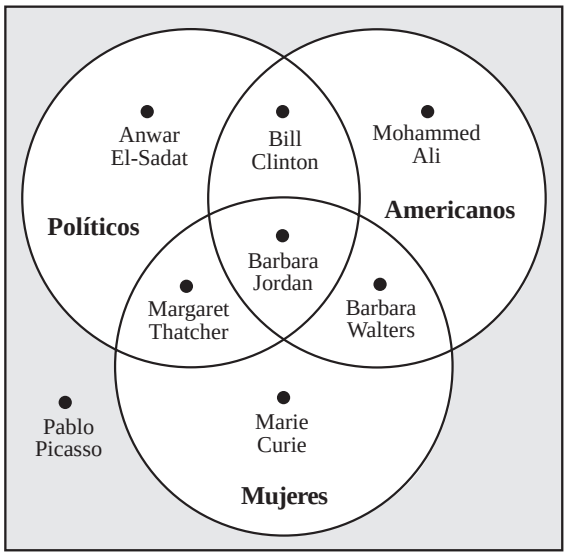


FIGURA 2.2. Un ejemplo de una base de datos con tres atributos binarios que dividen la población en ocho conjuntos disjuntos.

Los Ejemplos 2.1 y 2.2 se utilizarán posteriormente en este capítulo para ilustrar varios conceptos relacionados con los sistemas expertos basados en reglas.

Algunos sistemas imponen ciertas restricciones a las reglas. Por ejemplo:

- No permitir en la premisa el operador lógico *o*, y
- Limitar las conclusiones a expresiones lógicas simples.

Hay buenas razones para imponer estas restricciones. En primer lugar, las reglas que satisfacen estas restricciones son fáciles de tratar a la hora de escribir un programa de ordenador. En segundo lugar, las dos restricciones anteriores no dan lugar a una pérdida de generalidad, puesto que reglas mucho más generales pueden ser reemplazadas por conjuntos de reglas de esta forma. A esto se le llama *sustitución de reglas*. Por tanto, el conjunto de reglas especificado inicialmente por el experto humano puede requerir una sustitución posterior por un conjunto de reglas equivalente para satisfacer estas restricciones.

La Tabla 2.4 da ejemplos de sustitución de reglas. Nótese que cada regla de la primera columna puede ser sustituída por el correspondiente conjunto de reglas de la segunda columna y que todas las reglas de ésta satisfacen las condiciones anteriores. Por ejemplo, la primera regla compuesta de la Tabla 2.4:

- Regla 1: Si A o B, entonces C,

puede ser reemplazada por las dos reglas simples:

Regla	Reglas Equivalentes
Si A *o* B, entonces C	Si A, entonces C Si B, entonces C
Si $\overline{A \text{ o } B}$, entonces C	Si $\bar{A}$ *y* $\bar{B}$, entonces C
Si $\overline{A \text{ y } B}$, entonces C	Si $\bar{A}$, entonces C Si $\bar{B}$, entonces C
Si $(A \text{ o } B)$ *y* C, entonces D	Si A *y* C, entonces D Si B *y* C, entonces D
Si $\overline{(A \text{ o } B)}$ *y* C, entonces D	Si $\bar{A}$ *y* $\bar{B}$ *y* C, entonces D
Si $\overline{A \text{ y } B}$ *y* C, entonces D	Si $\bar{A}$ *y* C, entonces D Si $\bar{B}$ *y* C, entonces D
Si A, entonces B *y* C	Si A, entonces B Si A, entonces C
Si A, entonces B *o* C	Si A *y* $\bar{B}$, entonces C Si A *y* $\bar{C}$, entonces B
Si A, entonces $\overline{B \text{ y } C}$	Si A *y* B, entonces $\bar{C}$ Si A *y* C, entonces $\bar{B}$
Si A, entonces $\overline{B \text{ o } C}$	Si A, entonces $\bar{B}$ Si A, entonces $\bar{C}$

TABLA 2.4. Ejemplos de sustitución de reglas: Las reglas en la primera columna son equivalentes a las reglas de la segunda columna. Nótese que en los seis primeros ejemplos las sustituciones se aplican a la premisa y en los cuatro últimos, a la conclusión.

- Regla 1a: Si A, entonces C.
- Regla 1b: Si B, entonces C.

Como ejemplo adicional, la Tabla 2.5 muestra que

- Regla 2: Si $\overline{A \text{ o } B}$, entonces C,

puede ser reemplazada por la regla

- Regla 2: Si $\bar{A}$ *y* $\bar{B}$, entonces C,

donde $\bar{A}$ significa *no* A. La Tabla 2.5 se llama *tabla de verdad*.

A	B	$\bar{A}$	$\bar{B}$	$\overline{A \ o \ B}$	$\bar{A} \ y \ \bar{B}$
C	C	F	F	F	F
C	F	F	C	F	F
F	C	C	F	F	F
F	F	C	C	C	C

TABLA 2.5. Una tabla de verdad mostrando que las expresiones lógicas $\overline{A \ o \ B}$ y $\bar{A} \ y \ \bar{B}$ son equivalentes. Los símbolos C y F se utilizan para cierto y falso, respectivamente.

2.3 El Motor de Inferencia

Tal como se ha mencionado en la sección anterior, hay dos tipos de elementos: los datos (hechos o evidencia) y el conocimiento (el conjunto de reglas almacenado en la base de conocimiento). El motor de inferencia usa ambos para obtener nuevas conclusiones o hechos. Por ejemplo, si la premisa de una regla es cierta, entonces la conclusión de la regla debe ser también cierta. Los datos iniciales se incrementan incorporando las nuevas conclusiones. Por ello, tanto los hechos iniciales o datos de partida como las conclusiones derivadas de ellos forman parte de los hechos o datos de que se dispone en un instante dado.

Las conclusiones pueden clasificarse en dos tipos: *simples* y *compuestas*. Las conclusiones simples son las que resultan de una regla simple. Las conclusiones compuestas son las que resultan de más de una regla. Para obtener conclusiones, los expertos utilizan diferentes tipos de reglas y estrategias de inferencia y control (véase, por ejemplo, Castillo y Álvarez (1991), Durkin (1994), Shapiro (1987), Waterman (1985)). En el resto de esta sección se discuten las reglas de inferencia

- Modus Ponens,
- Modus Tollens,
- Resolución,

y las estrategias de inferencia

- Encadenamiento de reglas,
- Encadenamiento de reglas orientado a un objetivo,
- Compilación de reglas,

que son utilizadas por el motor de inferencia para obtener conclusiones simples y compuestas. Las dos primeras reglas de inferencia se usan para obtener conclusiones simples y el resto de reglas y estrategias para obtener conclusiones compuestas.

Nótese, sin embargo, que ninguna de las estrategias anteriores, si se implementan solas, conduce a todas las conclusiones posibles. Por ello, deben implementarse varias reglas y estrategias en el sistema experto para que el motor de inferencia sea capaz de obtener tantas conclusiones como sea posible.

2.3.1 Modus Ponens y Modus Tollens

El *Modus Ponens* es quizás la regla de inferencia más comúnmente utilizada. Se utiliza para obtener conclusiones simples. En ella, se examina la premisa de la regla, y si es cierta, la conclusión pasa a formar parte del conocimiento. Como ilustración, supóngase que se tiene la regla, "Si A es cierto, entonces B es cierto" y que se sabe además que "A es cierto." Entonces, tal como muestra la Figura 2.3, la regla Modus Ponens concluye que "B es cierto." Esta regla de inferencia, que parece trivial, debido a su familiaridad, es la base de un gran número de sistemas expertos.

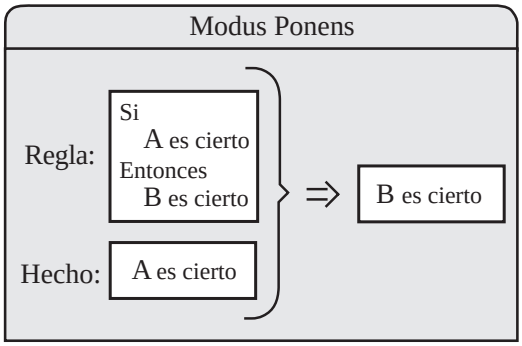


FIGURA 2.3. Una ilustración de la regla de inferencia Modus Ponens.

La regla de inferencia *Modus Tollens* se utiliza también para obtener conclusiones simples. En este caso se examina la conclusión y si es falsa, se concluye que la premisa también es falsa. Por ejemplo, supóngase de nuevo que se tiene la regla, "Si A es cierto, entonces B es cierto" pero se sabe que "B es falso." Entonces, utilizando la regla Modus Ponens no se puede obtener ninguna conclusión, pero, tal como se muestra en la Figura 2.4, la regla Modus Tollens concluye que "A es falso." Aunque muy simple y con muchas aplicaciones útiles, la regla Modus Tollens es menos utilizada que la Modus Ponens.

Por ello, la regla Modus Ponens se mueve hacia adelante, es decir, de la premisa a la conclusión de una regla, mientras que la regla Modus Tollens se mueve hacia atrás, es decir, de la conclusión a la premisa. Las dos reglas de inferencia no deben ser vistas como alternativas sino como complementarias. La regla Modus Ponens necesita información de los objetos

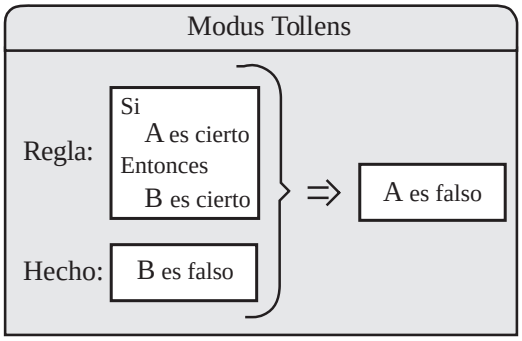


FIGURA 2.4. Una ilustración de la regla Modus Tollens.

de la premisa para concluir, mientras que la regla Modus Tollens necesita información sobre los objetos de la conclusión. De hecho, para un motor de inferencia que solamente utiliza Modus Ponens, la incorporación de la regla de inferencia Modus Tollens puede ser considerada como una expansión de la base de conocimiento mediante la adición de reglas, tal como ilustra el ejemplo que sigue.

Ejemplo 2.3 La regla Modus Tollens equivale a una expansión de la base de conocimiento. Supóngase que la base de conocimiento consiste sólo en la Regla 1, que se muestra en la Figura 2.5. Se puede utilizar la regla de inferencia Modus Tollens para "invertir" la Regla 1 y obtener alguna conclusión cuando se tiene información sobre los objetos de su conclusión. Entonces, aplicar la regla Modus Tollens a la regla "Si A, entonces B" es equivalente a aplicar la regla Modus Ponens a la regla "Si $\bar{B}$, entonces $\bar{A}$." En este caso de Regla 1, utilizando la equivalencia

$$\overline{A = C \; y \; B = C} \Leftrightarrow \overline{A = F} \; o \; \overline{B = F},$$

se obtiene la Regla 1b, que se muestra en la Figura 2.6. Por ello, utilizar ambas, las reglas Modus Ponens y Modus Tollens cuando la base de conocimiento contiene sólo la Regla 1, es equivalente a usar la regla Modus Ponens cuando la base de conocimiento contiene ambas, la Regla 1 y la Regla 1b. ■

Por otra parte, el rendimiento del motor de inferencia depende del conjunto de reglas en su base de conocimiento. Hay situaciones en las que el motor de inferencia puede concluir utilizando un conjunto de reglas, pero no puede, utilizando otro (aunque éstos sean lógicamente equivalentes). A continuación se da un ejemplo ilustrativo.

Ejemplo 2.4 Inferencia con dos conjuntos equivalentes de reglas. Supóngase de nuevo que se tienen dos motores de inferencia: El motor E_1, cuya base de conocimiento contiene las siete reglas de la Figura 2.1, y el

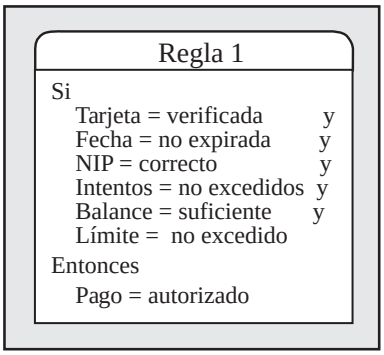


FIGURA 2.5. Regla 1 tomada de la Figura 2.1.

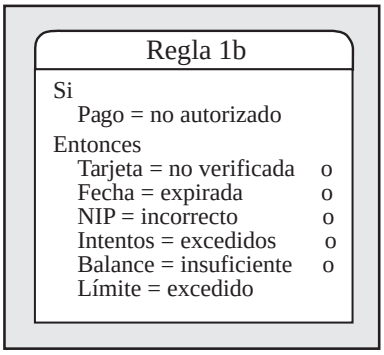


FIGURA 2.6. La Regla 1b puede obtenerse de la Regla 1 utilizando la regla de inferencia Modus Tollens.

motor E_2, cuya base de conocimiento contiene las siete reglas de la Figura 2.7. Nótese que los dos conjuntos de reglas son lógicamente equivalentes. Supóngase además que se sabe que el valor de NIP es *incorrecto*. Si ambos E_1 y E_2 utilizan sólo la regla de inferencia Modus Ponens, entonces E_1 será capaz de concluir que $Pago = no\ autorizado$ (por la Regla 4), pero E_2 no concluirá. Por ello, algunas de las conclusiones lógicamente derivables pueden no ser obtenidas usando sólo la regla de inferencia Modus Ponens. Por otra parte, si ambos motores usan la regla Modus Tollens, entonces ambos concluirán. ■

2.3.2 El Mecanismo de Resolución

Las reglas de inferencia Modus Ponens y Modus Tollens pueden ser utilizadas para obtener conclusiones simples. Por otra parte, las conclusiones compuestas, que se basan en dos o más reglas, se obtienen usando el lla-

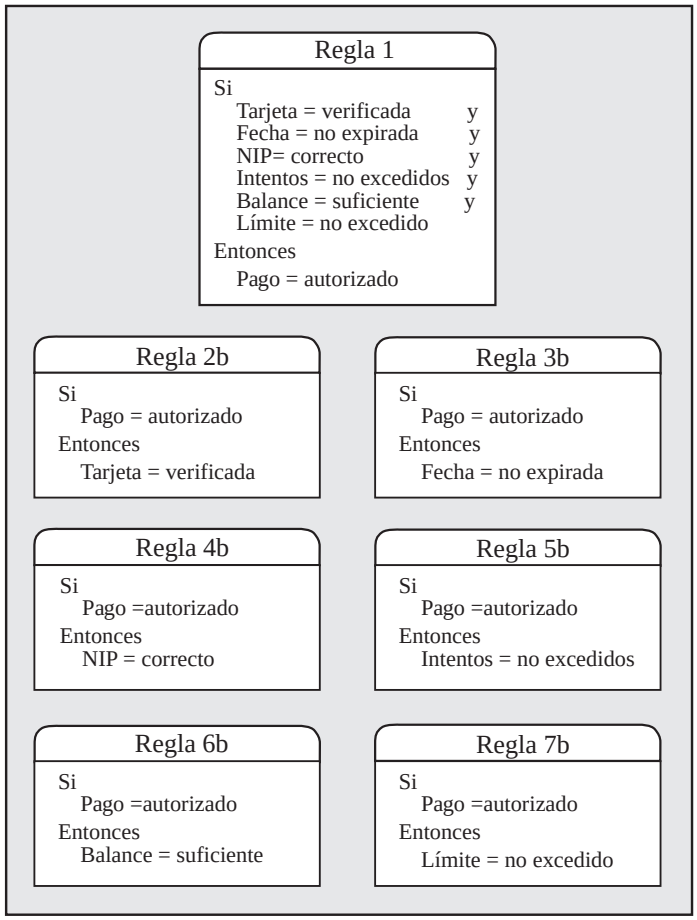


FIGURA 2.7. Un conjunto de reglas lógicamente equivalentes al conjunto de reglas de la Figura 2.1).

mado *mecanismo de resolución*. Esta regla de inferencia consiste en las etapas siguientes:

1. Las Reglas son sustituidas por expresiones lógicas equivalentes.
2. Estas expresiones lógicas se combinan en otra expresión lógica.
3. Esta última expresión se utiliza para obtener la conclusión.

Estas etapas involucran conceptos tales como la combinación y simplificación de expresiones lógicas, que se ilustran de un modo intuitivo en los ejemplos que siguen. Para un tratamiento detallado de esta regla de inferencia el lector puede consultar alguno de los libros específicos citados en la introducción de este capítulo.

A	B	$\bar{A}$	Si A, entonces B	$\bar{A}$ o B
C	C	F	C	C
C	F	F	F	F
F	C	C	C	C
F	F	C	C	C

TABLA 2.6. Una tabla de verdad mostrando que la regla "Si A es cierto, entonces B es cierto" es equivalente a la expresión lógica "A es falso o B es cierto."

Ejemplo 2.5 Mecanismo de resolución 1. Supóngase que se tienen las dos reglas:

- Regla 1: Si A es cierto, entonces B es cierto.
- Regla 2: Si B es cierto, entonces C es cierto.

La primera etapa en el mecanismo de resolución consiste en sustituir cada una de las dos reglas por expresiones lógicas equivalentes. Esto se hace como sigue (véase la Figura 2.8):

- La Regla 1 es equivalente a la expresión lógica: "A es falso o B es cierto." Una prueba de esta equivalencia se muestra an la tabla de verdad que se muestra en la Tabla 2.6.
- Similarmente, la Regla 2 es equivalente a la expresión lógica: "B es falso o C es cierto."

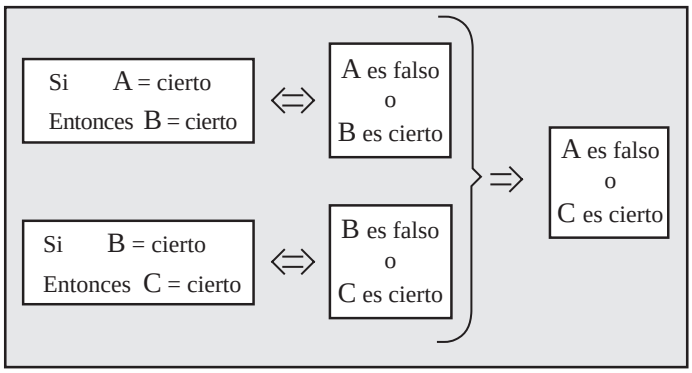


FIGURA 2.8. Un ejemplo ilustrando la regla de inferencia correspondiente al mecanismo de resolución.

La segunda etapa consiste en combinar las dos expresiones anteriores en una, tal como sigue: Las expresiones lógicas "A es falso o B es cierto" y "B es falso o C es cierto" implican la expresión "A es falso o C es

A	B	C	$\bar{A}$ o B	$\bar{B}$ o C	($\bar{A}$ o B) y ($\bar{B}$ o C)	$\bar{A}$ o C
C	C	C	C	C	C	C
C	C	F	C	F	F	F
C	F	C	F	C	F	C
C	F	F	F	C	F	F
F	C	C	C	C	C	C
F	C	F	C	F	F	C
F	F	C	C	C	C	C
F	F	F	C	C	C	C

TABLA 2.7. Una tabla de verdad que muestra que las expresiones lógicas "A es falso o B es cierto" y "B es falso o C es cierto" implican la expresión lógica "A es falso o C es cierto."

cierto." Una prueba de esta equivalencia se muestra en la Tabla 2.7. Esta última expresión se utiliza seguidamente en la tercera etapa para obtener la conclusión. Las etapas anteriores se ilustran en la Figura 2.8. ∎

Ejemplo 2.6 Mecanismo de resolución 2. Considérese de nuevo el ejemplo del CA con el objeto añadido *Explicar*, que puede tomar los valores $\{sí, no\}$, indicando si se necesita explicar las acciones del CA. Apliquemos ahora el mecanismo de resolución a la evidencia $NIP = incorrecto$ y a las reglas siguientes:

- Si $NIP = incorrecto$ entonces $Pago = no\ autorizado$.
- Si $Pago = no\ autorizado$ entonces $Explicar = sí$.

Tal como se ilustra en la Figura 2.9, la regla de inferencia correspondiente al mecanismo de resolución conduce a la conclusión $Explicar = sí$. En efecto, siguiendo los pasos indicados, se tiene

1. Las dos reglas se sustituyen por las expresiones equivalentes:
   - $NIP = correcto$ o $Pago = no\ autorizado$
   - $Pago = autorizado$ o $Explicar = sí$
2. Las dos expresiones anteriores se combinan de la forma indicada para dar la expresión $NIP = correcto$ o $Explicar = sí$, y
3. Esta última expresión se combina con la evidencia $NIP = incorrecto$, y se obtiene la conclusión compuesta, $Explicar = sí$. ∎

Es importante señalar que la regla de inferencia correspondiente al mecanismo de resolución no siempre conduce a conclusiones, pues, de hecho,

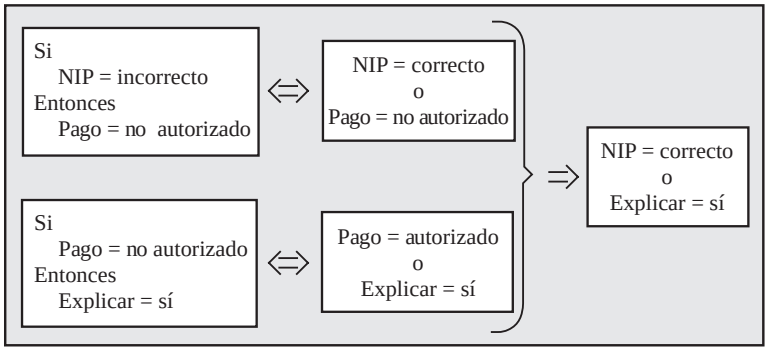


FIGURA 2.9. La regla de inferencia del mecanismo de resolución aplicada al ejemplo del CA.

puede no conocerse la verdad o falsedad de ciertas expresiones. Si esto ocurre, el sistema experto, o más precisamente, su motor de inferencia, debe decidir entre:

- Abandonar la regla, dada la imposibilidad de obtener conclusiones, o
- Preguntar al usuario, mediante el subsistema de demanda de información, sobre la verdad o falsedad de una o varias expresiones para poder continuar el proceso de inferencia hasta que se obtenga una conclusión.

2.3.3 Encadenamiento de Reglas

Una de las estrategias de inferencia más utilizadas para obtener conclusiones compuestas es el llamado *encadenamiento de reglas*. Esta estrategia puede utilizarse cuando las premisas de ciertas reglas coinciden con las conclusiones de otras. Cuando se encadenan las reglas, los hechos pueden utilizarse para dar lugar a nuevos hechos. Esto se repite sucesivamente hasta que no pueden obtenerse más conclusiones. El tiempo que consume este proceso hasta su terminación depende, por una parte, de los hechos conocidos, y, por otra, de las reglas que se activan. La estrategia de encadenamiento de reglas se da en el algoritmo siguiente:

Algoritmo 2.1 Encadenamiento de reglas.

- **Datos:** Una base de conocimiento (objetos y reglas) y algunos hechos iniciales.
- **Resultado:** El conjunto de hechos derivados lógicamente de ellos.

1. Asignar a los objetos sus valores conocidos tales como los dan los hechos conocidos o la evidencia

2. Ejecutar cada regla de la base de conocimiento y concluir nuevos hechos si es posible.

3. Repetir la Etapa 2 hasta que no puedan ser obtenidos nuevos hechos. ■

Este algoritmo puede ser implementado de muchas formas. Una de ellas comienza con las reglas cuyas premisas tienen valores conocidos. Estas reglas deben concluir y sus conclusiones dan lugar a nuevos hechos. Estos nuevos hechos se añaden al conjunto de hechos conocidos, y el proceso continúa hasta que no pueden obtenerse nuevos hechos. Este proceso se ilustra, a continuación, con dos ejemplos.

Ejemplo 2.7 Encadenamiento de Reglas 1. La Figura 2.10 muestra un ejemplo de seis reglas que relacionan 13 objetos, del A al M. Las relaciones entre estos objetos implicadas por las seis reglas pueden representarse gráficamente, tal como se muestra en la Figura 2.11, donde cada objeto se representa por un nodo. Las aristas representan la conexión entre los objetos de la premisa de la regla y el objeto de su conclusión. Nótese que las premisas de algunas reglas coinciden con las conclusiones de otras reglas. Por ejemplo, las conclusiones de las Reglas 1 y 2 (objetos C y G) son las premisas de la Regla 4.

Supóngase que se sabe que los objetos A, B, D, E, F, H e I son *ciertos* y los restantes objetos son de valor desconocido. La Figura 2.12 distingue entre objetos con valor conocido (los hechos) y objetos con valores desconocidos. En este caso, el algoritmo de encadenamiento de reglas procede como sigue:

- La Regla 1 concluye que $C = cierto$.
- La Regla 2 concluye que $G = cierto$.
- La Regla 3 concluye que $J = cierto$.
- La Regla 4 concluye que $K = cierto$.
- La Regla 5 concluye que $L = cierto$.
- La Regla 6 concluye que $M = cierto$.

Puesto que no pueden obtenerse más conclusiones, el proceso se detiene. Este proceso se ilustra en la Figura 2.12, donde los números en el interior de los nodos indican el orden en el que se concluyen los hechos. ■

Ejemplo 2.8 Encadenamiento de reglas 2. Considérense de nuevo las seis reglas de la Figura 2.10 y supóngase ahora que se dan los hechos $H = cierto$, $I = cierto$, $K = cierto$ y $M = falso$. Esto se ilustra en la Figura 2.13, donde los objetos con valores conocidos (los hechos) aparecen

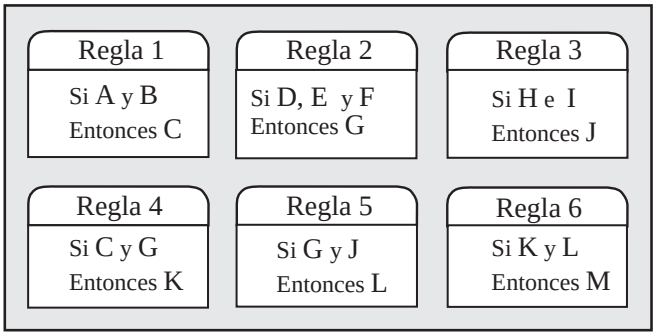


FIGURA 2.10. Un ejemplo de un conjunto de seis reglas relacionando 13 objetos.

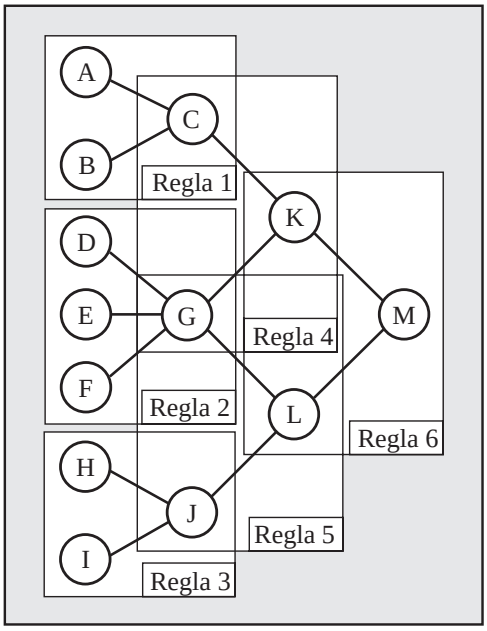


FIGURA 2.11. Una representación gráfica de las relaciones entre las seis reglas de la Figura 2.10.

sombreados y la variable objetivo se muestra rodeada por una circunferencia. Supóngase, en primer lugar, que el motor de inferencia usa las dos reglas de inferencia Modus Ponens y Modus Tollens. Entonces, aplicando el Algoritmo 2.1, se obtiene

1. La Regla 3 concluye que $J = cierto$ (Modus Ponens).

2. La Regla 6 concluye (Modus Tollens) que $K = falso$ o $L = falso$, pero, puesto que $K = cierto$, deberá ser $L = falso$.

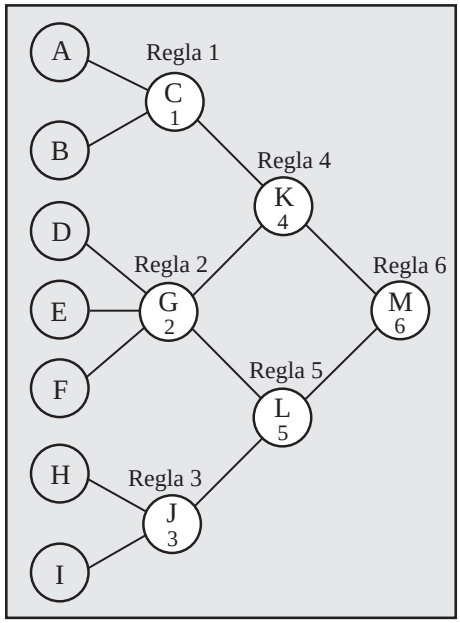


FIGURA 2.12. Un ejemplo que ilustra la estrategia de encadenamiento de reglas. Los nodos con valores conocidos aparecen sombreados y los números en su interior indican el orden en el que se concluyen los hechos.

3. La Regla 5 concluye (Modus Tollens) que $G = falso$ o $J = falso$, pero, puesto que $J = cierto$, deberá ser $G = falso$.

En consecuencia, se obtiene la conclusión $G = falso$. Sin embargo, si el motor de inferencia sólo utiliza la regla de inferencia Modus Ponens, el algoritmo se detendrá en la Etapa 1, y no se concluirá nada para el objeto G. Este es otro ejemplo que ilustra la utilidad de la regla de inferencia Modus Tollens. ■

Nótese que la estrategia de encadenamiento de reglas diferencia claramente entre la memoria de trabajo y la base de conocimiento. La memoria de trabajo contiene datos que surgen durante el periodo de consulta. Las premisas de las reglas se comparan con los contenidos de la memoria de trabajo y cuando se obtienen nuevas conclusiones son pasadas también a la memoria de trabajo.

2.3.4 Encadenamiento de Reglas Orientado a un Objetivo

El algoritmo de encadenamiento de reglas orientado a un objetivo requiere del usuario seleccionar, en primer lugar, una variable o nodo objetivo; entonces el algoritmo navega a través de las reglas en búsqueda de una conclusión para el nodo objetivo. Si no se obtiene ninguna conclusión con la

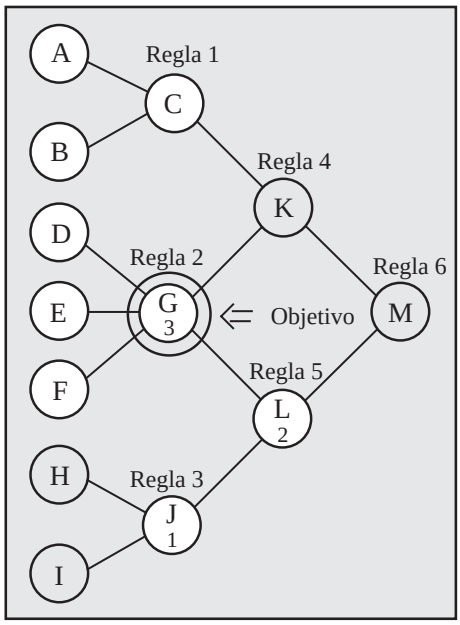


FIGURA 2.13. Otro ejemplo que ilustra el algoritmo de encadenamiento de reglas. Los nodos con valores conocidos aparecen sombreados, la variable objetivo se muestra rodeada por una circunferencia, y los números del interior de los nodos indican el orden en el que se concluyen los hechos.

información existente, entonces el algoritmo fuerza a preguntar al usuario en busca de nueva información sobre los elementos que son relevantes para obtener información sobre el objetivo.

Algunos autores llaman a los algoritmos de encadenamiento y de encadenamiento orientado a un objetivo *encadenamiento hacia adelante* y *encadenamiento hacia atrás*, respectivamente. Pero esta terminología puede ser confusa, puesto que ambos algoritmos pueden, en realidad, utilizar las dos reglas de inferencia Modus Ponens (hacia adelante) y Modus Tollens (hacia atrás).

El algoritmo de encadenamiento de reglas orientado a un objetivo se describe a continuación.

Algoritmo 2.2 Encadenamiento de reglas orientado a un objetivo.

- **Datos:** Una base de conocimiento (objetos y reglas), algunos hechos iniciales, y un nodo o variable objetivo.
- **Resultado:** El valor del nodo o variable objetivo.

1. Asigna a los objetos sus valores conocidos tales como están dados en los hechos de partida, si es que existe alguno. Marcar todos los objetos

cuyo valor ha sido asignado. Si el nodo objetivo está marcado, ir a la Etapa 7; en otro caso:

(a) Designar como objetivo *inicial* el objetivo *en curso*.

(b) Marcar el objetivo en curso.

(c) Sea $ObjetivosPrevios = \phi$, donde ϕ es el conjunto vacío.

(d) Designar todas las reglas como activas (ejecutables).

(e) Ir a la Etapa 2.

2. Encontrar una regla activa que incluya el objetivo en curso y ninguno de los objetos en $ObjetivosPrevios$. Si se encuentra una regla, ir a la Etapa 3; en otro caso, ir a la Etapa 5.

3. Ejecutar la regla referente al objetivo en curso. Si concluye, asignar el valor concluido al objetivo en curso, e ir a la Etapa 6; en otro caso, ir a la Etapa 4.

4. Si todos los objetos de la regla están marcados, declarar la regla como *inactiva* e ir a la Etapa 2; en otro caso:

   (a) Añadir el objetivo en curso a $ObjetivosPrevios$.

   (b) Designar uno de los objetos no marcados en la regla como el objetivo en curso.

   (c) Marcar el objetivo en curso.

   (d) Ir a la Etapa 2.

5. Si el objetivo en curso es el mismo que el objetivo inicial, ir a la Etapa 7; en otro caso, preguntar al usuario por el valor del objetivo en curso. Si no se da un valor, ir a la Etapa 6; en otro caso asignar al objeto el valor dado e ir a la Etapa 6.

6. Si el objetivo en curso es el mismo que el objetivo inicial, ir a la Etapa 7; en otro caso, designar el objetivo previo como objetivo en curso, eliminarlo de $ObjetivosPrevios$, e ir a la Etapa 2.

7. Devolver el valor del objetivo en curso si es conocido. ■

A continuación se ilustra el encadenamiento de reglas orientado a un objetivo mediante algunos ejemplos.

Ejemplo 2.9 Encadenamiento de reglas orientado a un objetivo. Considérense las seis reglas de las Figuras 2.10 y 2.11. Supóngase que se selecciona el nodo M como nodo objetivo y que se sabe que los objetos D, E, F y L son ciertos. Estos nodos están sombreados en la Figura 2.14. Las etapas del algoritmo de encadenamiento de reglas orientado a un objetivo se ilustran en la Figura 2.14, donde el número en el interior de un nodo indica el orden en el que se visita cada nodo. Estas etapas son:

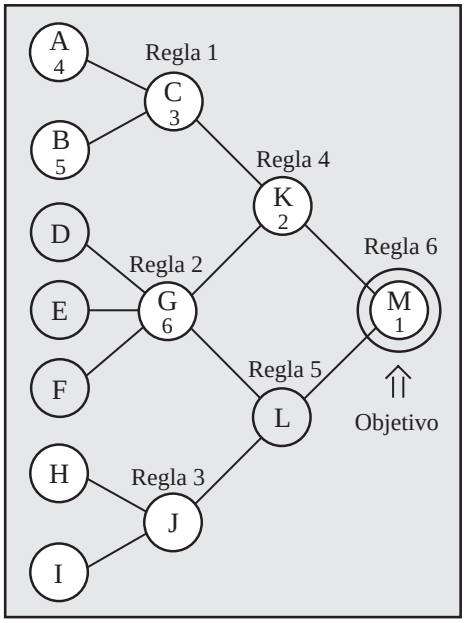


FIGURA 2.14. Un ejemplo que ilustra el algoritmo de encadenamiento de reglas orientado a un objetivo. Los nodos cuyo valor es conocido se han sombreado, el nodo objetivo se ha rodeado por una circunferencia, y el número en el interior de un nodo indica el orden en el que se visita cada nodo.

- Etapa 1: Se asigna el valor *cierto* a los objetos D, E, F y L y se marcan. Puesto que el nodo objetivo M no está marcado, entonces
  - Se designa el objeto M como objeto *en curso*.
  - Se marca el objeto M. Por tanto, se tiene $ObjetosMarcados = \{D, E, F, L, M\}$.
  - $ObjetivosPrevios = \phi$.
  - Las seis reglas están activas. Por tanto, se tiene $ReglasActivas = \{1, 2, 3, 4, 5, 6\}$.
  - Se va a la Etapa 2.
- Etapa 2. Se busca una regla que incluya el objetivo en curso M. Se encuentra la Regla 6, por lo que se va a la Etapa 3.
- Etapa 3. La Regla 6 no puede concluir puesto que el valor del objeto K es desconcoido. Así que se va a la Etapa 4.
- Etapa 4. El objeto K no está marcado. Entonces
  - $ObjetivosPrevios = \{M\}$.

  - Se elige el objeto K como objetivo en curso.
  - El objeto K está marcado. Por tanto se tiene, $ObjetosMarcados = \{D, E, F, L, M, K\}$.
  - Se va a la Etapa 2.

- Etapa 2. Se busca una regla que incluya el objetivo en curso K pero no el anterior M. Se encuentra la Regla 4, y se continúa con la Etapa 3.

- Etapa 3. La Regla 4 no puede concluir puesto que se desconocen los valores de los objetos C y G. Por ello, se continúa con la Etapa 4.

- Etapa 4. Los objetos C y G no están marcados. Entonces
  - $ObjetivosPrevios = \{M, K\}$.
  - Se elige uno de los objetos no marcados C o G como el nuevo objetivo en curso. Supóngase que se elige C.
  - Se marca el objeto C.
    Por tanto, se tiene $ObjetosMarcados = \{D, E, F, L, M, K, C\}$.
  - Se continúa con la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo en curso C pero no los objetos previos $\{M, K\}$. Se encuentra la Regla 1, así que se va a la Etapa 3.

- Etapa 3. La Regla 1 no puede concluir puesto que se desconocen los valores de los objetos A y B. Por tanto se continúa con la Etapa 4.

- Etapa 4. Los objetos A y B no están marcados. Entonces
  - $ObjetivosPrevios = \{M, K, C\}$.
  - Se elige uno de los objetos no marcados A y B como nuevo objetivo en curso. Supóngase que se elige A.
  - Se marca el objeto A.
    Por ello, $ObjetosMarcados = \{D, E, F, L, M, K, C, A\}$.
  - Se continúa con la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo en curso A pero no los objetivos previos $\{M, K, C\}$. No se encuentra ninguna regla que satisfaga estas condiciones, así que se pasa a la Etapa 5.

- Etapa 5. Puesto que el objetivo en curso A es diferente del inicial M, se pregunta al usuario por el valor del objeto A. Supóngase que A toma el valor cierto, entonces se hace $A = cierto$ y se sigue con la Etapa 6.

- Etapa 6. El objetivo en curso A no coincide con el previo M. Por tanto, el objeto C se designa como objetivo en curso y se elimina de la lista $ObjetivosPrevios$. Por ello, $ObjetivosPrevios = \{M, K\}$ y se continúa con la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo C pero no los anteriores $\{M, K\}$. Se encuentra la Regla 1, por lo que se va a la Etapa 3.

- Etapa 3. La Regla 1 no puede concluir porque el valor del objeto B es desconocido. Así que se va a la Etapa 4.

- Etapa 4. El objeto B no está marcado. Entonces
  - $ObjetivosPrevios = \{M, K, C\}$.
  - Se elige como objetivo en curso el único objeto no marcado, B.
  - Se marca el objeto B.
    Por ello, $ObjetosMarcados = \{D, E, F, L, M, K, C, A, B\}$.
  - Se va a la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo B pero no los objetivos previos $\{M, K, C\}$. Como no se encuentra ninguna regla, se va a la Etapa 5.

- Etapa 5. Puesto que el objetivo en curso B no coincide con el inicial M, se pregunta al usuario el valor del objetivo en curso B. Supóngase que se da un valor cierto a B, entonces se hace $B = cierto$ y se va a la Etapa 6.

- Etapa 6. Como el objetivo en curso B no coincide con el inicial M, se designa el objetivo previo C como objetivo en curso y se elimina de $ObjetivosPrevios$. Por ello, $ObjetivosPrevios = \{M, K\}$ y se va a la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo en curso C pero no los anteriores $\{M, K\}$. Se encuentra la Regla 1, por lo que se va a la Etapa 3.

- Etapa 3. Puesto que $A = cierto$ y $B = cierto$, entonces $C = cierto$ por la Regla 1. Ahora se va a la Etapa 6.

- Etapa 6. El objetivo en curso C no coincide con el inicial M. Entonces, se designa el objetivo previo K como objetivo en curso y se elimina de $ObjetivosPrevios$. Por ello, $ObjetivosPrevios = \{M\}$ y se va a la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo en curso K pero no los anteriores $\{M\}$. Se encuentra la Regla 4, por lo que se va a la Etapa 3.

- Etapa 3. La Regla 4 no puede concluir puesto que el valor del objeto G es desconocido. Por tanto, se va a la Etapa 4.

- Etapa 4. El objeto G no está marcado. Entonces
  - $ObjetivosPrevios = \{M, K\}$.
  - El único objeto no marcado G se elige como objetivo en curso.
  - Se marca el objeto G.
    Por ello, $ObjetosMarcados = \{D, E, F, L, M, K, C, A, B, G\}$.
  - Se va a la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo en curso G pero no los anteriores $\{M, K\}$. Se encuentra la Regla 2, por lo que se va a la Etapa 3.

- Etapa 3. Puesto que $D = cierto$, $E = cierto$ y $F = cierto$, entonces $G = cierto$ por la Regla 2. Ahora se va a la Etapa 6.

- Etapa 6. El objetivo en curso G no coincide con el inicial M. Entonces, se designa el objetivo previo K como objetivo en curso y se elimina de $ObjetivosPrevios$. Por ello, $ObjetivosPrevios = \{M\}$ y se va a la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo en curso K pero no los anteriores $\{M\}$. Se encuentra la Regla 4, por lo que se va a la Etapa 3.

- Etapa 3. Puesto que $C = cierto$ y $G = cierto$, entonces $K = cierto$ por la Regla 4. Seguidamente se va a la Etapa 6.

- Etapa 6. El objetivo en curso K no coincide con el inicial M. Entonces, se designa el objetivo previo M como objetivo en curso y se elimina de $ObjetivosPrevios$. Por ello, $ObjetivosPrevios = \phi$ y se va a la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo en curso M. Se encuentra la Regla 6, por lo que se va a la Etapa 3.

- Etapa 3. Puesto que $K = cierto$ y $L = cierto$, entonces $M = cierto$ por la Regla 6. Ahora se va a la Etapa 6.

- Etapa 6. El objetivo en curso M coincide con el inicial. En consecuencia, se va a la Etapa 7.

- Etapa 7. El algoritmo devuelve el valor $M = cierto$. ∎

Nótese que a pesar de que los objetos H, I y J tienen valores desconocidos, el algoritmo orientado a un objetivo ha sido capaz de concluir el valor del objetivo M. La razón de este resultado está en que el conocimiento del objeto L convierte al conocimiento de los objetos H, I y J es irrelevante para el conocimiento del objeto M. ■

Las estrategias de encadenamiento de reglas se utilizan en problemas en los que algunos hechos (por ejemplo, síntomas) se dan por conocidos y se buscan algunas conclusiones (por ejemplo, enfermedades). Por el contrario, las estrategias de encadenamiento de reglas orientadas a un objetivo se utilizan en problemas en los que se dan algunos objetivos (enfermedades) y se buscan los hechos (síntomas) para que éstas sean posibles.

Ejemplo 2.10 Encadenamiento de reglas orientado a un objetivo sin Modus Tollens. Considérense las seis reglas de las Figuras 2.10 y 2.11. Supóngase que se selecciona el nodo J como objetivo y que se dan los siguientes hechos: $G = cierto$ y $L = falso$. Esto se ilustra en la Figura 2.15, donde los objetos con valores conocidos (los hechos) aparecen sombreados y el objetivo rodeado por una circunferencia. Supóngase, en primer lugar, que el motor de inferencia utiliza sólo la regla de inferencia Modus Ponens. En este caso, las etapas del Algoritmo 2.2 son como sigue:

- Etapa 1: Se marcan los objetos G y L y se les asignan los valores $G = cierto$ y $L = falso$. Puesto que el objetivo J no está marcado, entonces
  - Se designa el objeto J como objetivo *en curso*.
  - Se marca el objeto J. Por ello, $ObjetosMarcados = \{G, L, J\}$.
  - $ObjetivosPrevios = \phi$.
  - Todas las reglas están activas. Por tanto, se tiene $ReglasActivas = \{1, 2, 3, 4, 5, 6\}$.
  - Se va a la Etapa 2.
- Etapa 2. Se busca una regla activa que incluya el objetivo en curso J. Puesto que sólo se utiliza la regla de inferencia Modus Ponens, se encuentra la Regla 3 (es la única regla en la que el objetivo en curso J forma parte de su conclusión). Por tanto, se va a la Etapa 3.
- Etapa 3. La Regla 3 no puede concluir puesto que los valores de los objetos H e I son desconocidos. Por tanto, se va a la Etapa 4.
- Etapa 4. Los objetos H y I no están marcados. Entonces
  - $ObjetivosPrevios = \{J\}$.
  - Se elige uno de los objetos no marcados H o I como objetivo en curso. Supóngase que se elige H.

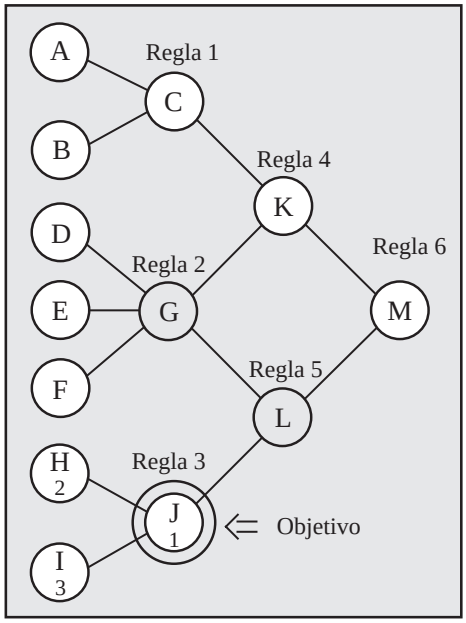


FIGURA 2.15. Un ejemplo que ilustra el encadenamiento de reglas orientado a un objetivo sin Modus Tollens. Los objetos con valores conocidos (los hechos) aparecen sombreados, el objetivo se muestra rodeado por una circunferencia, y el número en el interior de cada nodo indica el orden en el que se visita.

  - El objeto H está marcado. Por tanto, se tiene $ObjetosMarcados = \{G, L, J, H\}$.
  - Se va a la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo en curso H pero no el previo J. No se encuentra tal regla, por lo que se va a la Etapa 5.

- Etapa 5. Puesto que el objetivo en curso H es diferente del objetivo inicial J, se pregunta al usuario el valor del objetivo en curso H. Supóngase que no se da un valor para dicho objeto, entonces se va a la Etapa 6.

- Etapa 6. El objetivo en curso H no coincide con el objetivo inicial J. Entonces, se designa el objetivo previo J como objetivo en curso y se elimina de $ObjetivosPrevios$. Por ello, $ObjetivosPrevios = \phi$ y se va a la Etapa 2.

- Etapa 2. Se busca una regla activa que incluya el objetivo en curso J. Se encuentra la Regla 3 y se va a la Etapa 3.

- Etapa 3. La Regla 3 no puede concluir puesto que se desconocen los valores de los objetos H e I. Por tanto, se va a la Etapa 4.
- Etapa 4. El objeto I no está marcado. Entonces
  - $ObjetivosPrevios = \{J\}$.
  - Se elige el único objeto no marcado I como el objetivo en curso.
  - Se marca el objeto I. Por ello, $ObjetosMarcados = \{G, L, J, H, I\}$.
  - Se va a la Etapa 2.
- Etapa 2. Se busca una regla activa que incluya el objetivo en curso I pero no el previo J. Como no se encuentra tal regla, se va a la Etapa 5.
- Etapa 5. Puesto que el objetivo en curso I es diferente del objetivo inicial J, se pregunta al usuario el valor del objetivo en curso I. Supóngase que no se da un valor para el objeto I, entonces, se va a la Etapa 6.
- Etapa 6. El objetivo en curso I no es el mismo que el inicial J. Por tanto, se designa el objetivo previo J como objetivo en curso y se elimina de la lista $ObjetivosPrevios$. Por ello, se hace $ObjetivosPrevios = \phi$ y se vuelve a la Etapa 2.
- Etapa 2. Se busca una regla activa que incluya el objetivo en curso J. Se encuentra la Regla 3, por lo que se va a la Etapa 3.
- Etapa 3. La Regla 3 no puede concluir puesto que no se conocen los valores de los objetos H e I. Se va a la Etapa 4.
- Etapa 4. Todos los objetos de la Regla 3 están marcados, por lo que la Regla 3 se declara inactiva. Por ello, se hace $ReglasActivas = \{1, 2, 4, 5, 6\}$. Se continúa en la Etapa 2.
- Etapa 2. Se busca una regla activa que incluya el objetivo en curso J. Puesto que la Regla 3 se ha declarado inactiva y no se utiliza la regla de inferencia Modus Tollens, no se encuentra ninguna regla y se va a la Etapa 5.
- Etapa 5. Puesto que el objetivo en curso J coincide con el inicial, se va a la Etapa 7.
- Etapa 7. El motor de inferencia no puede concluir un valor para el objetivo J. ∎

Ejemplo 2.11 Encadenamiento de reglas orientado a un objetivo con Modus Tollens. En el ejemplo 2.10, el motor de inferencia utiliza sólo la regla de inferencia Modus Ponens. Considérense las mismas hipótesis

del Ejemplo 2.10 excepto que ahora el motor de inferencia utiliza ambas reglas de inferencia, la Modus Ponens y la Modus Tollens. Las etapas del Algoritmo 2.2 en este caso son como sigue:

- Etapa 1: Se marcan los objetos G y L y se les asignan los valores $G = cierto$ y $L = falso$. Puesto que el objetivo J no está marcado, entonces
  - Se designa el objeto J como objetivo *en curso*.
  - Se marca el objeto J. Por ello, se hace $ObjetosMarcados = \{G, L, J\}$.
  - $ObjetivosPrevios = \phi$.
  - Las seis reglas son activas. Por ello, se hace $ReglasActivas = \{1, 2, 3, 4, 5, 6\}$.
  - Se va a la Etapa 2.
- Etapa 2. Se busca una regla activa que incluya el objetivo en curso J. Puesto que se utilizan ambas reglas de inferencia, las Reglas 3 y 5 incluyen el objeto J. Supóngase que se elige la Regla 5. Se va a la Etapa 3. (Si se elige la Regla 3, el algoritmo tardará más tiempo en terminar.)
- Etapa 3. Puesto que se usa la regla de inferencia Modus Tollens, la Regla 5 concluye que $J = falso$ (puesto que $G = cierto$ y $L = falso$). Por tanto, se va a la Etapa 6.
- Etapa 6. El objetivo en curso J coincide con el inicial, por lo que se va a la Etapa 7.
- Etapa 7. Se devuelve $J = falso$. ■

2.3.5 Compilación de reglas

Otra forma de tratar con reglas encadenadas consiste en comenzar con un conjunto de datos (información) y tratar de alcanzar algunos objetivos. Esto se conoce con el nombre de *compilación de reglas*. Cuando ambos, datos y objetivos, se han determinado previamente, las reglas pueden ser compiladas, es decir, pueden escribirse los objetivos en función de los datos para obtener las llamadas *ecuaciones objetivo*. La compilación de reglas se explica mejor con un ejemplo.

Ejemplo 2.12 Compilación de reglas. Considérese el conjunto de seis reglas de la Figura 2.11 y supóngase que son conocidos los valores de los objetos A, B, D, E, F, H, e I y que los restantes objetos, C, G, J, K, L y M, son objetivos. Denotemos por $\wedge$ el operador lógico *y*; entonces, utilizando las seis reglas, pueden obtenerse las siguientes ecuaciones objetivo:

- La Regla 1 implica $C = A \wedge B$.
- La Regla 2 implica $G = D \wedge E \wedge F$.
- La Regla 3 implica $J = H \wedge I$.
- La Regla 4 implica $K = C \wedge G = (A \wedge B) \wedge (D \wedge E \wedge F)$.
- La Regla 5 implica $L = G \wedge J = (D \wedge E \wedge F) \wedge (H \wedge I)$.
- La Regla 6 implica $M = K \wedge L = A \wedge B \wedge D \wedge E \wedge F \wedge H \wedge I$.

Las tres primeras ecuaciones son equivalentes a las tres primeras reglas. Las tres ecuaciones objetivo son, respectivamente, equivalentes a las reglas siguientes:

- Regla 4a: *Si A y B y D y E y F, entonces K.*
- Regla 5a: *Si D y E y F y H e I, entonces L.*
- Regla 6a: *Si A y B y D y E y F y H e I, entonces M.*

Por ello, si, por ejemplo, cada uno de los objetos $\{A, B, D, E, F, H, I\}$ toma el valor cierto, entonces se obtiene de forma inmediata, a partir de las Reglas 4a, 5a y 6a, que los objetos $\{K, L, M\}$ deben ser ciertos. ■

2.4 Control de la Coherencia

En situaciones complejas, incluso verdaderos expertos pueden dar información inconsistente (por ejemplo, reglas inconsistentes y/o combinaciones de hechos no factibles). Por ello, es muy importante controlar la coherencia del conocimiento tanto durante la construcción de la base de conocimiento como durante los procesos de adquisición de datos y razonamiento. Si la base de conocimiento contiene información inconsistente (por ejemplo, reglas y/o hechos), es muy probable que el sistema experto se comporte de forma poco satisfactoria y obtenga conclusiones absurdas.

El objetivo del control de la coherencia consiste en

1. Ayudar al usuario a no dar hechos inconsistentes, por ejemplo, dándole al usuario las restricciones que debe satisfacer la información demandada.
2. Evitar que entre en la base de conocimiento cualquier tipo de conocimiento inconsistente o contradictorio.

El control de la coherencia debe hacerse controlando la coherencia de las reglas y la de los hechos.

Objetos		Conclusiones		Conclusiones
A	B	Regla 1	Regla 2	contradictorias
C	C	$B = C$	$B = F$	Sí
C	F	$B = C$	$B = F$	Sí
F	C	–	–	No
F	F	–	–	No

TABLA 2.8. Una tabla de verdad que muestra que las Reglas 1 y 2 son coherentes.

2.4.1 Coherencia de Reglas

Definición 2.2 Reglas coherentes. *Un conjunto de reglas se denomina coherente si existe, al menos, un conjunto de valores de todos los objetos que producen conclusiones no contradictorias.*

En consecuencia, un conjunto coherente de reglas no tiene por qué producir conclusiones no contradictorias para todos los posibles conjuntos de valores de los objetos. Es decir, es suficiente que exista un conjunto de valores que conduzcan a conclusiones no contradictorias.

Ejemplo 2.13 Conjunto de reglas incoherentes. Considérense las cuatro reglas siguientes, que relacionan dos objetos A y B binarios $\{C, F\}$:

- Regla 1: Si $A = C$, entonces $B = C$.
- Regla 2: Si $A = C$, entonces $B = F$.
- Regla 3: Si $A = F$, entonces $B = C$.
- Regla 4: Si $A = F$, entonces $B = F$.

Entonces, pueden obtenerse las siguientes conclusiones:

1. Las Reglas 1−2 son coherentes puesto que, tal como se muestra en la Tabla 2.8, para $A = F$, no producen conclusiones.
2. Las Reglas 1−3 son coherentes puesto que para $A = F$ y $B = C$, producen una conclusión ($B = C$) (véase la Tabla 2.9).
3. Las Reglas 1−4 son incoheremtes porque producen conclusiones contradictorias para todos los posibles valores de A y B, tal como se ve en la Tabla 2.10. ∎

Nótese que un conjunto de reglas puede ser coherente, aunque algunos conjuntos de valores puedan producir conclusiones inconsistentes. Estos conjuntos de valores se llaman *valores no factibles*. Por ejemplo, las Reglas 1−2 son coherentes, aunque producen conclusiones inconsistentes en todos los casos en que $A = C$. En consecuencia, el subsistema de control de

Objetos		Conclusiones			Conclusiones
A	B	Regla 1	Regla 2	Regla 3	contradictorias
C	C	$B = C$	$B = F$	–	Sí
C	F	$B = C$	$B = F$	–	Sí
F	C	–	–	$B = C$	No
F	F	–	–	$B = C$	Sí

TABLA 2.9. Una tabla de verdad que muestra que las Reglas 1−3 son coherentes.

Objetos		Conclusiones				Conclusiones
A	B	Regla 1	Regla 2	Regla 3	Regla 4	contradictorias
C	C	$B = C$	$B = F$	–	–	Sí
C	F	$B = C$	$B = F$	–	–	Sí
F	C	–	–	$B = C$	$B = F$	Sí
F	F	–	–	$B = C$	$B = F$	Sí

TABLA 2.10. Una tabla de verdad que muestra que las Reglas 1−4 son incoherentes.

coherencia eliminará automáticamente el valor C de la lista de posibles valores del objeto A, permitiendo de esta forma al usuario seleccionar sólo valores factibles de los objetos.

Definición 2.3 Valores no factibles. *Se dice que un valor a para el objeto A no es factible si las conclusiones obtenidas al hacer $A = a$ contradicen cualquier combinación de valores del resto de los objetos.*

Por ello, cualquier valor no factible debe ser eliminado de la lista de valores posibles de su correspondiente objeto para eliminar la posibilidad de que el motor de inferencia pueda obtener conclusiones inconsistentes.

Ejemplo 2.14 Valores no factibles. Considérese el conjunto de las cuatro reglas del Ejemplo 2.13. En este caso, el motor de inferencia concluirá lo siguiente:

1. Las dos primeras reglas implican que $A \neq C$, puesto que $A = C$ siempre conduce a conclusiones inconsistentes. Por tanto, el valor $A = C$ deberá ser eliminado automáticamente de la lista de valores factibles de A. Dado que A es binario, entonces resulta $A = F$ (el único valor posible).

2. Las tres primeras reglas implican que $A = F$ y $B = C$. Por tanto, el valor $B = F$ deberá ser eliminado automáticamente de la lista de valores factibles de B.

3. Las primeras cuatro reglas implican que $A \neq C$, $A \neq F$, $B \neq C$ y $B \neq F$. Por tanto, los valores $\{C, F\}$ son eliminados de las listas de valores de A y B, con lo que las listas de valores factibles de todos los objetos están vacías, lo que implica que las cuatro reglas son incoherentes. ∎

Nótese que es suficiente realizar la comprobación de la coherencia de las reglas sólo una vez, tras ser introducida cada regla, y que todos los valores no factibles pueden ser eliminados de sus correspondientes listas, nada más ser detectados.

El conjunto de reglas que forman el conocimiento debe ser coherente; en otro caso, el sistema podrá obtener conclusiones erróneas. Por ello, antes de añadir una regla a la base de conocimiento, hay que comprobar la consistencia de esta regla con el resto de ellas, incluídas en la base de conocimiento. Si la regla fuese consistente con el resto de reglas, se añadiría a la base de conocimiento; en caso contrario, se devolvería al experto humano para su corrección.

Ejemplo 2.15 Coherencia de reglas. Supóngase que se tienen los cuatro objetos: $A \in \{0, 1\}, B \in \{0, 1\}$, $C \in \{0, 1, 2\}$ y $D \in \{0, 1\}$. Considérense las cuatro reglas:

- Regla 1: Si $A = 0$ y $B = 0$, entonces $C = 0$.
- Regla 2: Si $A = 0$ y $D = 0$, entonces $C = 1$.
- Regla 3: Si $A = 0$ y $B = 0$, entonces $C = 1$.
- Regla 4: Si $A = 0$, entonces $B = 0$.
- Regla 5: Si $B = 0$, entonces $A = 1$.

Supóngase ahora que se desea añadir las tres últimas reglas a una base de conocimiento que contiene las dos primeras reglas. Entonces, las Reglas 1 y 3 son inconsistentes, puesto que tienen la misma premisa pero diferentes conclusiones. Por tanto, la Regla 3 debe ser rechazada y el experto humano informado de la razón del rechazo. El experto humano corregirá la regla en cuestión y/o las reglas existentes si fueran incorrectas. La Regla 4 entrará en la base de conocimiento, puesto que es consistente con las Reglas 1 y 2. La Regla 5 es inconsistente con la Regla 4. Por ello, la consistencia de ambas reglas debe ser comprobada antes de pasar a formar parte de la base de conocimiento. ∎

2.4.2 Coherencia de hechos

Los datos o evidencias suministrados por los usuarios deben ser también consistentes en sí y con el conjunto de reglas de la base de datos. Por ello, el sistema no debe aceptar hechos que contradigan el conjunto de reglas y/o

el conjunto de hechos existente en cada instante del proceso. Por ejemplo, con una base de conocimiento que contenga las dos primeras reglas del Ejemplo 2.15, el sistema no debe aceptar el conjunto de hechos $A = 0$, $B = 0$ y $C = 1$ puesto que contradicen la Regla 1.

El sistema debe también comprobar si existe o no, una solución factible e informar al usuario en consecuencia. Si en el ejemplo anterior se trata de dar la información $A = 0, B = 0$ y $D = 0$, el sistema debe detectar que no existe ningún valor de C que sea consistente con la base de conocimiento. Nótese que antes de conocer los valores de los objetos, existe una solución factible. Por ejemplo, $A = 0, B = 0, C = 0$ y $D = 1$ (estos hechos no contradicen la base de conocimiento). Por ello, la inconsistencia surge de que los hechos y las reglas sean inconsistentes.

La coherencia de los hechos puede lograrse mediante las estrategias siguientes:

1. Eliminar todos los valores no factibles (los que contradicen el conjunto de reglas y/o hechos) de los objetos una vez detectados. Cuando se pregunte al usuario por información sobre los valores de un conjunto de objetos, el sistema experto debería aceptar sólo los valores de cada objeto que sean consistentes con las reglas y con el conocimiento previo. Considérese, por ejemplo, la base de conocimiento del Ejemplo 2.15 y supóngase que al sistema experto se le ha dado la información $A = 0$ y $C = 1$; entonces el sistema debe saber que $B \neq 0$. Por ello, este valor debe ser eliminado de la lista de posibles valores del objeto B.

2. El motor de inferencia debe comprobar que los hechos conocidos no contradicen el conjunto de reglas. En la situación anterior, por ejemplo, el sistema no debe aceptar el conjunto de hechos $A = 1$, $B = 1$ y $C = 2$. Si el sistema no elimina los valores no factibles, entonces el usuario podrá dar evidencias contradictorias tales como $Pago = autorizado$ y $NIP = incorrecto$ en el Ejemplo 2.1 del CA . Por ello, tan pronto como se de la primera evidencia, $Pago = autorizado$, el sistema debe seleccionar sólo los valores del NIP que no conduzcan a conclusiones contradictorias.

3. Suministrar al usuario una lista de objetos a los que no se ha asignado valores previamente.

4. Para cada uno de los objetos, mostrar y aceptar sólo sus valores factibles.

5. Actualizar continuamente la base de conocimiento, es decir, tan pronto como se dé un hecho o se obtenga una conclusión, y eliminar los valores no factibles. El motor de inferencia obtiene todas las conclusiones posibles examinando, y posiblemente concluyendo, las reglas tan pronto como una simple unidad de información llega al sistema.

Nótese que dar varias unidades de información simultáneamente puede conducir a inconsistencias en la base de datos. Por ejemplo, dado $A = 0$, no se puede dar la información combinada $B = 0$ y $C = 1$. En este caso, el orden de la información afecta a los posibles valores futuros de los objetos que conducen a compatibilidad, es decir, tras dar $A = 0$ se puede dar $B = 0$ ó $C = 1$, pero estas dos opciones imponen restricciones diferentes a los posibles futuros valores de los restantes objetos.

La actualización continua de la base de conocimiento es muy importante puesto que no actualizar implica la posibilidad de que evidencias contradictorias puedan convivir en la base de conocimiento. Por ello, el conocimiento debe ser actualizado inmediatamente tras la incorporación de cada hecho.

Por ello, tanto la eliminación automática de valores no factibles como la actualización continua del conocimiento aseguran la coherencia de la base de conocimiento. El ejemplo siguiente ilustra la aplicación de esta técnica al problema de los agentes secretos presentado en el Capítulo 1.

Ejemplo 2.16 Los Agentes Secretos. En este ejemplo se retoma el problema de los agentes secretos introducido en el Ejemplo 1.5, en el que cada uno de los cuatro agentes secretos, Alberto, Luisa, Carmen y Tomás, está en uno de los cuatro paises: Egipto, Francia, Japón y España. Se han recibido los siguientes telegramas de los agentes:

- De Francia: Luisa está en España.
- De España: Alberto está en Francia.
- De Egipto: Carmen está en Egipto.
- De Japón: Carmen está en Francia.

El problema radica en que no se sabe quién ha enviado cada uno de los mensajes, pero es conocido que Tomás miente (¿es un agente doble?) y que los demás agentes dicen la verdad. El misterio que trata de escudriñarse es el de responder a la pregunta ¿quién está en cada país?

Seguidamente se diseña un sistema experto para resolver este problema. Se tienen cuatro objetos: Alberto, Luisa, Carmen y Tomás. Cada objeto puede tomar uno de cuatro valores: Egipto, Francia, Japón o España. Puesto que Tomás es el único que miente, se considera que un telegrama suyo es siempre falso. Esto da lugar a dos reglas por cada mensaje:

1. El mensaje de Francia (Luisa está en España) da lugar a:
    - Regla 1: Si Tomás está en Francia, entonces Luisa no está en España.

- Regla 2: Si Tomás no está en Francia, entonces Luisa está en España.

2. El mensaje de España (Alberto está en Francia) da lugar a:
   - Regla 3: Si Tomás está en España, entonces Alberto no está en Francia.
   - Regla 4: Si Tomás no está en España, entonces Alberto está en Francia.
3. El mensaje de Egipto (Carmen está en Egipto) da lugar a:
   - Regla 5: Si Tomás está en Egipto, entonces Carmen no está en Egipto.
   - Regla 6: Si Tomás no está en Egipto, entonces Carmen está en Egipto.
4. El mensaje de Japón (Carmen está en Francia) da lugar a:
   - Regla 7: Si Tomás está en Japón, entonces Carmen no está en Francia.
   - Regla 8: Si Tomás no está en Japón, entonces Carmen está en Francia.

Utilizando sólo estas ocho reglas, se intentará ahora averiguar el valor que toma el objeto Tomás:

1. Tomás está posiblemente en Egipto. Si Tomás está en Egipto, se obtienen las conclusiones siguientes:
   - Luisa está en España, por la Regla 2.
   - Alberto está en Francia, por la Regla 4.
   - Carmen no está en Egipto, por la Regla 5.
   - Carmen está en Francia, por la Regla 8.

   Se ve que con esta hipótesis se llega a la conclusión de que tanto Alberto como Carmen están en Francia, lo que contradice la información de que sólo un agente puede estar en cada país (pero el conjunto de las ocho reglas anteriores no contiene esta información). Por tanto, se concluye que Egipto es un valor imposible para el objeto Tomás, es decir, Tomás no puede estar en Egipto.
2. Tomás está posiblemente en Japón. Si Tomás está Japón, se obtienen las conclusiones siguientes:
   - Luisa está en España, por la Regla 2.
   - Alberto está en Francia, por la Regla 4.

- Carmen está en Egipto, por la Regla 6.

En este caso no hay una contradicción, lo que significa que Japón es un valor posible para el objeto Tomás.

Con las ocho reglas anteriores, el motor de inferencia no puede concluir en qué país está cada uno de los agentes, puesto que las reglas no contienen la información "sólo un agente puede estar en cada país." Seguidamente se considera esta situación y se obtiene un conjunto de reglas adicionales que tienen en cuenta esta información.

Puesto que cada país puede estar ocupado por exactamente un agente, supóngase que un agente está en un país dado. Entonces, se necesitan tres reglas para garantizar que ninguno de los restantes agentes está en ese mismo país. Dado que se tienen cuatro agentes, resultan un total de 12 reglas (3 reglas $\times$ 4 agentes). Sin embargo, si se utiliza la regla de inferencia Modus Tollens, sólo son necesarias seis reglas, pues las restantes resultan redundantes. Por ejemplo, para Egipto se tienen las reglas:

- Regla 9: Si Alberto está en Egipto, entonces Luisa no está en Egipto.
- Regla 10: Si Alberto está en Egipto, entonces Carmen no está en Egipto.
- Regla 11: Si Alberto está en Egipto, entonces Tomás no está en Egipto.
- Regla 12: Si Luisa está en Egipto, entonces Carmen no está en Egipto.
- Regla 13: Si Luisa está en Egipto, entonces Tomás no está en Egipto.
- Regla 14: Si Carmen está en Egipto, entonces Tomás no está en Egipto.

Nótese que existen un conjunto de seis reglas equivalentes a las anteriores. Por ejemplo, la regla:

- Regla 14a: Si Tomás está en Egipto, entonces Carmen no está en Egipto,

es equivalente a (Modus Tollens) la Regla 14. Por tanto, se necesitan sólo seis reglas por país.

Los conjuntos de seis reglas para cada uno de los restantes países se generan de forma similar. Por tanto, se tienen un total de 24 reglas adicionales que representan el hecho de que exactamente un agente puede estar en cada país. ■

2.5 Explicando Conclusiones

Tal como se ha indicado en el Capítulo 1, las conclusiones no bastan para satisfacer al usuario de un sistema experto. Normalmente, los usuarios esperan que el sistema les dé algún tipo de explicación que indique el por qué de las conclusiones. Durante el proceso realizado por el motor de inferencia, las reglas activas (las que han concluído) forman la base del mecanismo de explicación, que es regulado por el subsistema de explicación.

En los sistemas expertos basados en reglas, es fácil dar explicaciones de las conclusiones obtenidas. El motor de inferencia obtiene conclusiones basándose en un conjunto de reglas y, por tanto, conoce de qué regla procede cada conclusión. Por ello, el sistema puede dar al usuario la lista de hechos concluidos junto con las reglas que se han utilizado para obtenerlos.

Ejemplo 2.17 Explicando conclusiones. Considérense las seis reglas de las Figuras 2.10 y 2.11. Como en el Ejemplo 2.7, supóngase que se sabe que los objetos A, B, D, E, F, H, y I son *ciertos* y que los restantes objetos toman valores desconocidos. Entonces, aplicando el Algoritmo 2.1 y examinando las reglas que han sido ejecutadas, el sistema experto puede suministrar la explicación siguiente a las conclusiones obtenidas:

1. Hechos dados:

$$A = cierto, \quad B = cierto, \quad D = cierto, \quad E = cierto,$$
$$F = cierto, \quad H = cierto, \quad I = cierto.$$

2. Conclusiones y explicaciones:

   - $C = cierto$, basada en la Regla 1.
   - $G = cierto$, basada en la Regla 2.
   - $J = cierto$, basada en la Regla 3.
   - $K = cierto$, basada en la Regla 4.
   - $L = cierto$, basada en la Regla 5.
   - $M = cierto$, basada en la Regla 6. ∎

2.6 Ejemplo de Aplicación

Los sistemas de control de tráfico actuales son necesariamente complejos. En esta sección se muestra un ejemplo para ilustrar cómo puede diseñarse un sistema experto basado en reglas para resolver un problema de control de tráfico muy simple. La Figura 2.16 muestra un trazado de ferrocarril en el que varios trenes pueden circular en las dos direcciones. Hay cinco vías, $S_1, \ldots, S_5$, y 14 señales de tráfico, ocho en la parte superior del diagrama,

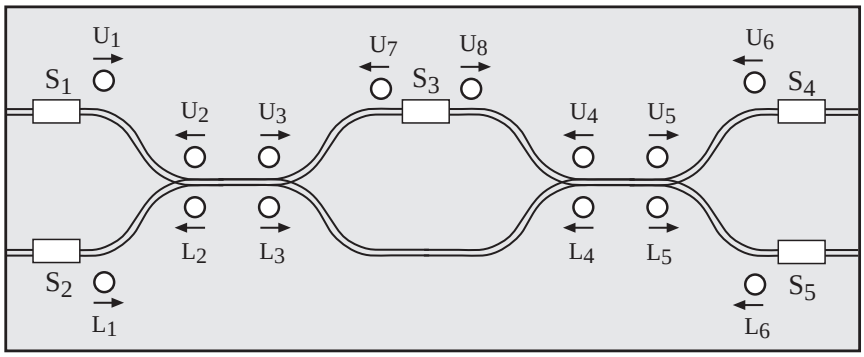


FIGURA 2.16. Trazado de ferrocarril con cinco vías.

Objeto	Valor
U_1 a U_8	$\{verde, rojo\}$
L_1 a L_6	$\{verde, rojo\}$
S_1 a S_5	$\{libre, ocupada\}$

TABLA 2.11. Objetos y sus correspondientes valores para el ejemplo del control de tráfico ferroviario.

$U_1, \ldots, U_8$, y seis en la parte inferior, $L_1, \ldots, L_6$. Todos los objetos y sus posibles valores se muestran en la Tabla 2.11.

El objetivo de este sistema es diseñar un conjunto de reglas que eviten la colisión de los trenes. Estas reglas pueden obtenerse como sigue:

1. Si la señal de tráfico U_1 está *verde*, entonces puede permitirse la salida de un tren que esté en la vía S_1 y no debe permitirse la salida de los trenes de la vía S_2, por lo que L_1 tiene que estar en *rojo*. Lo mismo es cierto para las vías S_4 y S_5. Esto da las dos primeras reglas de la Tabla 2.12. Nótese que si el motor de inferencia utiliza la regla de inferencia Modus Tollens, estas reglas garantizan también que cuando las señales de la parte baja de las vías estén en verde, las señales de sus correspondientes partes superiores estén en rojo. Es decir, las dos primeras reglas de la Tabla 2.12 implican las dos reglas siguientes:

   - Regla 1a: Si $L_1 = verde$, entonces $U_1 = rojo$.
   - Regla 2a: Si $L_6 = verde$, entonces $U_6 = rojo$.

2. Si la vía S_1 está ocupada, entonces la señal U_2 debe estar en *rojo* para evitar que un tren entre en la vía ocupada. Similarmente, para las demás vías. Esto da lugar a las seis reglas adicionales (Reglas 3−8) en la Tabla 2.12.

3. Si ambas señales U_3 y L_3 están en rojo, entonces ningún tren puede salir de la vía S_1. La misma condición vale para las señales U_5 y L_5. Por ello, se tiene la regla

   - Regla 9: Si ($U_3 = rojo$ y $L_3 = rojo$) o ($U_5 = rojo$ o $L_5 = rojo$), entonces $U_1 = rojo$.

   Las cinco reglas asociadas a las otras cinco vías pueden ser obtenidas de forma análoga. Todas las reglas se muestran en la Tabla 2.12 como las Reglas 9−14.

4. Para evitar la colisión de los trenes procedentes de las vías $S_1 - S_2$ y $S_4 - S_5$, son necesarias las reglas siguientes:

   - Regla 15: Si $U_3 = verde$, entonces $U_4 = rojo$
   - Regla 16: Si $L_3 = verde$, entonces $L_4 = rojo$.

5. Para evitar que las señales de la parte alta y sus correspondientes señales de la parte baja estén simultáneamente en verde, es necesario incluir las Reglas 17−20 de la Tabla 2.12.

6. Finalmente, para evitar la colisión de un tren de la vía S_3 con un tren de las otras cuatro vías, se imponen las dos últimas reglas de la Tabla 2.12.

Para mantener la coherencia de los hechos, es necesario actualizar automáticamente el conocimiento tan pronto como se conozca un nuevo hecho o conclusión.

Seguidamente se considera un ejemplo para ilustrar el comportamiento de un sistema experto cuya base de conocimiento consta de los objetos de la Tabla 2.11 y el conjunto de reglas de la Tabla 2.12.

Ejemplo 2.18 Control de tráfico ferroviario. En este ejemplo se usará la concha *X-pert Reglas*. En primer lugar, se necesita escribir un fichero que contenga una descripción de la base de conocimiento anterior. Este fichero es leído por la concha *X-pert Reglas*. Puesto que *X-pert Reglas* no permite el uso del operador *o* en la premisa de las reglas, es necesario reemplazar las Reglas 9−12 por el conjunto siguiente de reglas equivalente (ver las equivalencias de la Tabla 2.4):

- Regla 9a: Si ($U_3 = rojo$ y $L_3 = rojo$), entonces $U_1 = rojo$.
- Regla 9b: Si ($U_5 = rojo$ y $L_5 = rojo$), entonces $U_1 = rojo$.
- Regla 10a: Si ($U_3 = rojo$ y $L_3 = rojo$), entonces $L_1 = rojo$.
- Regla 10b: Si ($U_5 = rojo$ y $L_5 = rojo$), entonces $L_1 = rojo$.
- Regla 11a: Si ($U_2 = rojo$ y $L_2 = rojo$), entonces $U_6 = rojo$.

Regla	Premisa	Conclusión
Regla 1	$U_1 = verde$	$L_1 = rojo$
Regla 2	$U_6 = verde$	$L_6 = rojo$
Regla 3	$S_1 = ocupada$	$U_2 = rojo$
Regla 4	$S_2 = ocupada$	$L_2 = rojo$
Regla 5	$S_3 = ocupada$	$U_3 = rojo$
Regla 6	$S_3 = ocupada$	$U_4 = rojo$
Regla 7	$S_4 = ocupada$	$U_5 = rojo$
Regla 8	$S_5 = ocupada$	$L_5 = rojo$
Regla 9	($U_3 = rojo$ y $L_3 = rojo$) o ($U_5 = rojo$ y $L_5 = rojo$)	$U_1 = rojo$
Regla 10	($U_3 = rojo$ y $L_3 = rojo$) o ($U_5 = rojo$ y $L_5 = rojo$)	$L_1 = rojo$
Regla 11	($U_2 = rojo$ y $L_2 = rojo$) o ($U_4 = rojo$ y $L_4 = rojo$)	$U_6 = rojo$
Regla 12	($U_2 = rojo$ y $L_2 = rojo$) o ($U_4 = rojo$ y $L_4 = rojo$)	$L_6 = rojo$
Regla 13	$U_2 = rojo$ y $L_2 = rojo$	$U_7 = rojo$
Regla 14	$U_5 = rojo$ y $L_5 = rojo$	$U_8 = rojo$
Regla 15	$U_3 = verde$	$U_4 = rojo$
Regla 16	$L_3 = verde$	$L_4 = rojo$
Regla 17	$U_2 = verde$	$L_2 = rojo$
Regla 18	$U_3 = verde$	$L_3 = rojo$
Regla 19	$U_4 = verde$	$L_4 = rojo$
Regla 20	$U_5 = verde$	$L_5 = rojo$
Regla 21	$U_1 = verde$ o $L_1 = verde$	$U_7 = rojo$
Regla 22	$U_6 = verde$ o $L_6 = verde$	$U_8 = rojo$

TABLA 2.12. Reglas para el ejemplo del control de tráfico ferroviario.

- Regla 11b: Si ($U_4 = rojo$ y $L_4 = rojo$), entonces $U_6 = rojo$.
- Regla 12a: Si ($U_2 = rojo$ y $L_2 = rojo$), entonces $L_6 = rojo$.
- Regla 12b: Si ($U_4 = rojo$ y $L_4 = rojo$), entonces $L_6 = rojo$.

Por ello, se crea el fichero de texto "TrafficControl.txt," que contiene todos los objetos y las 26 reglas.[1] Supóngase que inicialmente se tienen trenes

[1]El fichero "TrafficControl.txt" con la base de conocimiento y la concha para construir sistemas expertos *X-pert Reglas* puede obtenerse de la dirección de World Wide Web http://ccaix3.unican.es/~AIGroup.

esperando en las vías S_1, S_2 y S_3 tal como se indica en la Figura 2.17. Lo que sigue muestra una sesión interactiva utilizando *X-pert Reglas* y tras leer el fichero de texto "TrafficControl.txt". En primer lugar se especifican los hechos: $S_1 = ocupada$, $S_3 = ocupada$ y $S_5 = ocupada$. Entonces se obtienen los siguientes hechos (dados) y las conclusiones (hechos concluidos):

1. Hechos:
   - $S_1 = ocupada$.
   - $S_3 = ocupada$.
   - $S_5 = ocupada$.
2. Conclusiones:
   - $U_2 = rojo$ (basada en la Regla 3).
   - $U_3 = rojo$ (basada en la Regla 5).
   - $U_4 = rojo$ (basada en la Regla 6).
   - $L_5 = rojo$ (basada en la Regla 8).

Por ello, se ponen en rojo cuatro señales para evitar la colisión de los trenes que esperan en las vías. El resto de los objetos toman valores desconocidos. La Figura 2.17 muestra esta información en forma gráfica.

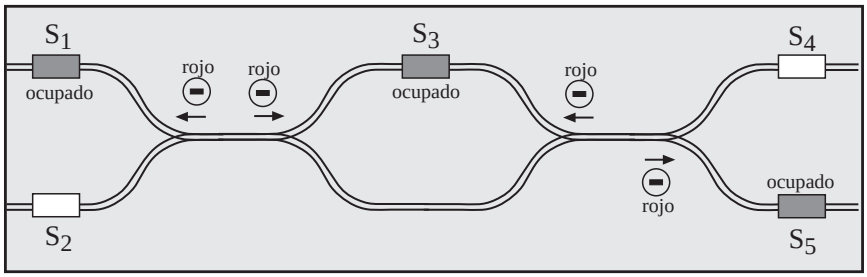


FIGURA 2.17. Nuevas conclusiones resultantes de los hechos $S_1 = S_3 = S_5 = ocupada$.

Supóngase ahora que se desea permitir al tren en la vía S_1 salir en dirección Este. Para ello, se hace $U_1 = verde$. Entonces se tienen los siguientes hechos y conclusiones:

1. Hechos:
   - $S_1 = ocupada$.
   - $S_3 = ocupada$.
   - $S_5 = ocupada$.
   - $U_1 = verde$.

2. Conclusiones:

   - $U_2 = rojo$ (basada en la Regla 3).
   - $U_3 = rojo$ (basada en la Regla 5).
   - $U_4 = rojo$ (basada en la Regla 6).
   - $L_5 = rojo$ (basada en la Regla 8).
   - $L_1 = rojo$ (basada en la Regla 1).
   - $U_7 = rojo$ (basada en la Regla 21).
   - $L_3 = verde$ (basada en la Regla 9a).
   - $U_5 = verde$ (basada en la Regla 9b).
   - $L_4 = rojo$ (basada en la Regla 16).
   - $S_4 \neq ocupada$ (basada en la Regla 7).
   - $S_4 = free$ (es el único valor posible).
   - $U_6 = rojo$ (basada en la Regla 11b).
   - L_6=rojo (basada en la Regla 12b).

La Figura 2.18 muestra las conclusiones resultantes. Nótese que el tren que está en la vía S_1 puede ahora partir y dirigirse a la vía S_4. Este camino se muestra en la Figura 2.18. ∎

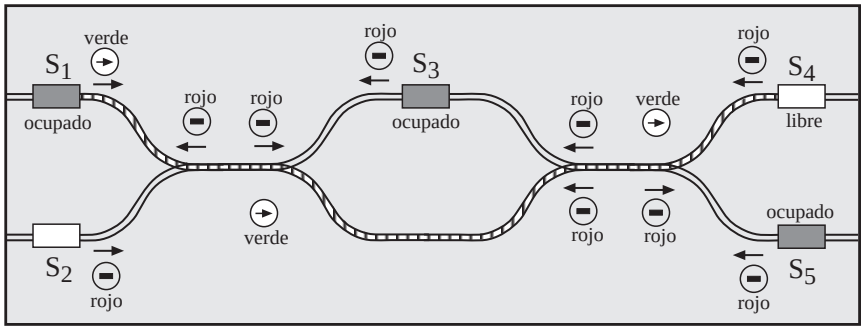


FIGURA 2.18. Nuevas conclusiones resultantes de los hechos $S_1 = S_3 = S_5 = ocupada$ and $U_1 = verde$.

2.7 Introduciendo Incertidumbre

Los sistemas basados en reglas descritos en este capítulo pueden aplicarse sólo a situaciones deterministas. Sin embargo, hay muchos casos prácticos que implican incertidumbre. Por ejemplo, en el Ejemplo 1.4 del diagnóstico

médico, la presencia de algunos síntomas no siempre implica la existencia de una enfermedad dada, incluso aunque haya una fuerte evidencia sobre la existencia de esa enfermedad. Por ello, es útil extender la lógica clásica para incorporar incertidumbre. Esto ha sido realizado mediante la introducción de varias medidas para tratar la incertidumbre. Castillo y Álvarez (1990, 1991) describen la aplicación de estas medidas para mejorar los sistemas expertos basados en reglas. Por otra parte, Johnson y Keravnou (1988) describen algunos prototipos de sistemas expertos basados en lógicas inciertas. El Capítulo 3 describe en detalle los *sistemas expertos basados en probabilidad*, que incorporan la incertidumbre.

Ejercicios

2.1 En el Ejemplo 2.3, se usan dos objetos binarios A y B y se da un ejemplo en el que la regla de inferencia Modus Tollens expande la base de conocimiento. Utilizando objetos no binarios, dar un ejemplo similar. Por ejemplo, cuando A y B puedan tomar los valores $\{0, 1, 2\}$.

2.2 Mostrar que los dos conjuntos de reglas de las Figuras 2.1 y 2.7 son lógicamente equivalentes.

2.3 En algún momento del Ejemplo 2.11, se ha buscado una regla activa que incluyera el objeto en curso J. Se encontraron las Reglas 3 y 5, y se eligió la Regla 5. Completar las etapas del algoritmo si se hubiera elegido la Regla 3 en vez de la Regla 5.

2.4 Considérese una intersección de dos calles de doble sentido, tal como se indica en la Figura 2.19, en la que se muestran los giros permitidos. Sean $T_1 - T_3$, $R_1 - R_3$, $B_1 - B_3$ y $L_1 - L_3$ los semáforos asociados a dichos carriles. Definir un conjunto de reglas que regulen la intersección de forma que no pueda producirse ninguna colisión.

2.5 Considérese la línea ferroviaria con seis vías dada en la Figura 2.20. Completar el conjunto de reglas dado en la Sección 2.6 para incorporar la nueva vía S_6.

2.6 Considérese la línea ferroviaria de la Figura 2.21 con cuatro vías $\{S_1, \ldots, S_4\}$. Diseñar un sistema de control de tráfico ferroviario para controlar las señales. Obtener un conjunto de reglas que garanticen la imposibilidad de colisiones de trenes.

2.7 Supóngase que se tienen las seis reglas del Ejemplo 2.7. Siguiendo el proceso dado en los Ejemplos 2.9 y 2.10, aplicar un algoritmo de encadenamiento orientado a un objetivo para concluir un valor para los objetivos dados en las Figuras 2.22(a) y 2.22(b). Los objetos que

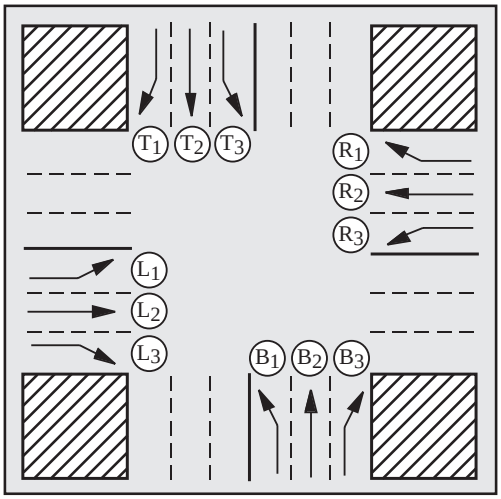


FIGURA 2.19. Intersección mostrando los giros permitidos.

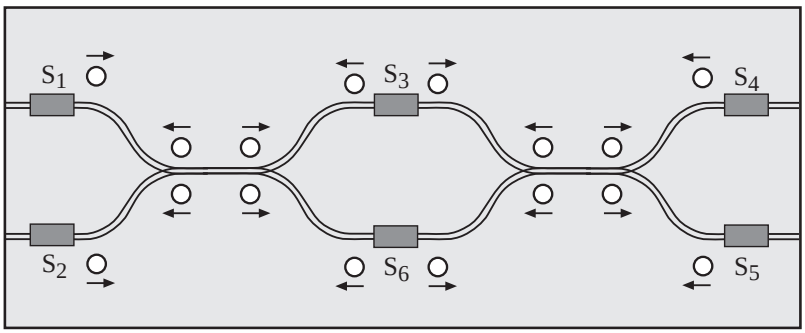


FIGURA 2.20. Una línea ferroviaria con seis vías.

se indican en gris son objetos con valores asignados. Los correspondientes valores se muestran próximos a los objetos. ¿Cuáles serían las conclusiones con un motor de inferencia que sólo incluya la regla de inferencia Modus Ponens?

2.8 Diseñar un sistema experto basado en reglas que sirva para jugar al "Tres en Raya". Por turno, dos jugadores ponen una de sus piezas en un tablero de 9 cuadrados (3×3) (véase la Figura 2.23). Gana el jugador que consiga poner sus 3 piezas en columna (Figura 2.23(a)), en fila (Figura 2.23(b)), o en diagonal (Figura 2.23(c)). Considérense las estrategias siguientes:

- Estrategia defensiva: Definir las reglas para evitar que el contrario gane.

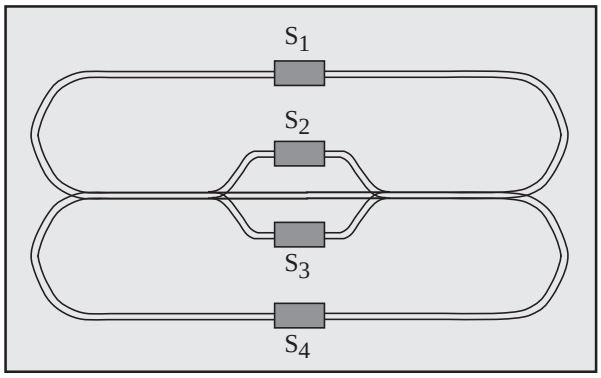


FIGURA 2.21. Una línea ferroviaria con cuatro vías.

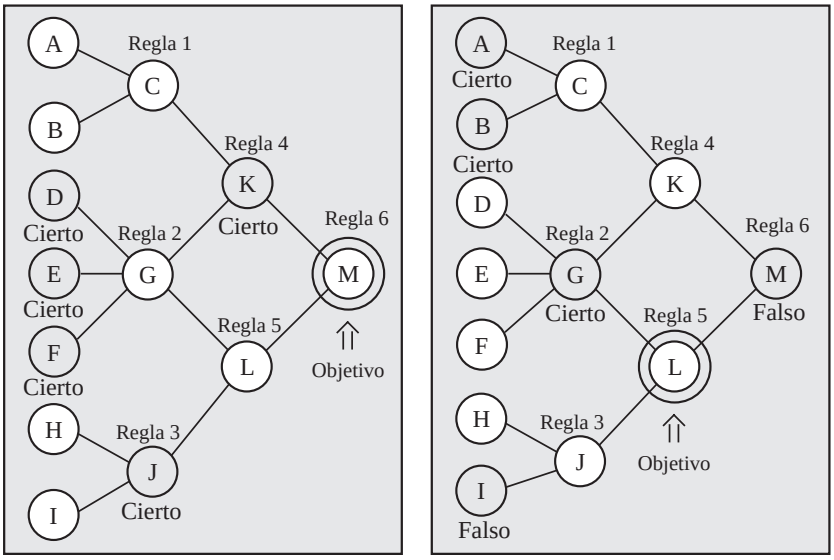


FIGURA 2.22. Hechos iniciales y objetivos para un algoritmo de encadenamiento de reglas orientado a un objetivo.

- Estrategia atacante: Añadir el conjunto de reglas que definan la estrategia para ganar.

2.9 Diseñar un sistema experto basado en reglas para clasificar animales o plantas basándose en un conjunto mínimo de características. Síganse las etapas siguientes:

- Decidir el conjunto de animales o plantas a clasificar.
- Elegir las características diferenciales.

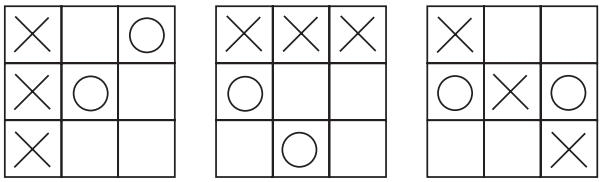


FIGURA 2.23. Tableros del juego del "Tres en Raya": Tres ejemplos en los que el jugador "X" es el ganador.

- Definir las reglas necesarias para identificar cada animal o planta.
- Eliminar las características innecesarias.
- Escribir las reglas en el sistema.
- Comprobar exhaustivamente el sistema experto.
- Rediseñar el sistema a la vista de lo anterior.

2.10 En el ejemplo de los agentes secretos del Ejemplo 2.16, ¿qué conclusiones pueden sacarse utilizando sólo las ocho primeras reglas cuando (a) se da Francia como valor posible del objeto Tomás y (b) se da España como valor posible para el mismo objeto?.

Capítulo 3
Sistemas Expertos Basados en Probabilidad

3.1 Introducción

Los sistemas expertos basados en reglas descritos en el capítulo anterior, no tienen en cuenta ningún tipo de incertidumbre, puesto que los objetos y las reglas son tratados por ellas de forma determinista. Sin embargo, en la mayor parte de las aplicaciones, la incertidumbre es lo común y no la excepción. Por ejemplo, una pregunta típica en diagnóstico médico es: dado que el paciente presenta un conjunto de síntomas, ¿cuál de las enfermedades posibles es la que tiene el paciente? Esta situación implica un cierto grado de incertidumbre puesto que:

- Los hechos o datos pueden no ser conocidos con exactitud. Por ejemplo, un paciente puede no estar seguro de haber tenido fiebre la noche pasada. Por ello, hay un cierto grado de incertidumbre en la información asociada a cada paciente (subjetividad, imprecisión, ausencia de información, errores, datos ausentes, etc.).

- El conocimiento no es determinista. Por ejemplo, las relaciones entre las enfermedades y los síntomas no son deterministas, puesto que un mismo conjunto de síntomas puede estar asociado a diferentes enfermedades. De hecho, no es extraño encontrar dos pacientes con los mismos síntomas pero diferentes enfermedades.

Por ello, es clara la necesidad de contar con sistemas expertos que traten situaciones de incertidumbre. Este capítulo describe un tipo de sistema

experto que trata este tipo de situaciones de forma efectiva. Éstos son los *sistemas expertos basados en probabilidad*.

En los primeros sistemas expertos, se eligió la probabilidad como medida para tratar la incertidumbre (véase Cheeseman (1985) o Castillo y Álvarez (1991)). Pero, desgraciadamente, muy pronto se encontraron algunos problemas, debidos al uso incorrecto de algunas hipótesis de independencia, utilizadas para reducir la complejidad de los cálculos. Como resultado, en las primeras etapas de los sistemas expertos, la probabilidad fue considerada como una medida de incertidumbre poco práctica. La mayoría de las críticas a los métodos probabilísticos se basaban en el altísimo número de parámetros necesarios, la imposibilidad de una asignación o estimación precisa de los mismos, o las hipótesis poco realistas de independencia.

Consecuentemente, en la literatura de la época, surgieron medidas alternativas a la probabilidad, como los factores de certeza, las credibilidades, las plausibilidades, las necesidades o las posibilidades, para tratar la incertidumbre (véase, por ejemplo, Shafer (1976), Zadeh (1983), Buchanan y Shortliffe (1984), Yager y otros (1987), y Almond (1995)).

Sin embargo, con la aparición de las redes probabilísticas (principalmente las redes Bayesianas y Markovianas, que se presentan en el capítulo 6), la probabilidad ha resurgido de forma espectacular, y es, hoy en día, la más intuitiva y la más aceptada de las medidas de incertidumbre. Lindley (1987), por ejemplo, dice:

> *"La única descripción satisfactoria de la incertidumbre es la probabilidad. Esto quiere decir que toda afirmación incierta debe estar en forma de una probabilidad, que varias incertidumbres deben ser combinadas usando las reglas de la probabilidad, y que el cálculo de probabilidades es adecuado para manejar situaciones que implican incertidumbre. En particular, las descripciones alternativas de la incertidumbre son innecesarias."*

Este capítulo introduce los sistemas expertos de tipo probabilístico, que se basan en la probabilidad como una medida de incertidumbre. Se describen en detalle sus principales componentes (por ejemplo, la base de conocimiento, el motor de inferencia, el sistema de control de coherencia, etc.) y se comparan con los sistemas expertos basados en reglas. En la Sección 3.2 se da una introducción breve a los conceptos de la teoría de la probabilidad, que se necesitan para entender el material de éste y otros capítulos. La Sección 3.3 define y discute las reglas generalizadas como un intento de extender los sistemas expertos basados en reglas para tratar situaciones de incertidumbre. Manteniendo el mismo tratamiento de los sistemas basados en reglas, se examina la estructura de la base del conocimiento, el motor de inferencia, y el sistema de control de la coherencia de los sistemas expertos basados en probabilidad. En particular, la Sección 3.4 ilustra este tipo de sistemas expertos mediante un ejemplo. La Sección 3.5 describe la base de conocimiento y presenta varios modelos para describir las relaciones entre

un conjunto de variables de interés. En la Sección 3.6, se discute el motor de inferencia. El problema del control de la coherencia se presenta en la Sección 3.7. Finalmente, en la Sección 3.8 se termina con una comparación de los sistemas basados en reglas y los sistemas basados en probabilidad.

3.2 Algunos Conceptos Básicos de la Teoría de la Probabilidad

En esta sección se introduce el siguiente material básico que será utilizado posteriormente:

- Medida de probabilidad.
- Distribuciones de probabilidad.
- Dependencia e independencia.
- Teorema de Bayes.
- Tipos de errores.

Los lectores que estén familiarizados con estos conceptos pueden omitir esta sección e ir directamente a la Sección 3.3. Por otra parte, el material presentado en esta sección es un mínimo necesario. Para repasar más conceptos y resultados, el lector interesado puede consultar cualquiera de los libros clásicos de teoría de la probabilidad y estadística, por ejemplo, DeGroot (1987), Durrett (1991), Hogg (1993), y Billingsley (1995).

3.2.1 Medida de Probabilidad

Para medir la incertidumbre se parte de un marco de discernimiento dado S, en el que se incluyen todos los posibles resultados de un cierto experimento como conjunto exhaustivo y mútuamente exclusivo. El conjunto S se conoce como *espacio muestral*. Una vez definido este conjunto, el objetivo consiste en asignar a todo subconjunto de S un número real que mida el grado de incertidumbre sobre su realización. Para obtener medidas con significado físico claro y práctico, se imponen ciertas condiciones o propiedades intuitivas adicionales que definen una clase de medidas que se conocen como *medidas de probabilidad*.

Definición 3.1 Medida de Probabilidad. *Una función p que proyecta los subconjuntos $A \subseteq S$ en el intervalo $[0, 1]$ se llama medida de probabilidad si satisface los siguientes axiomas:*

- **Axioma 1 (Normalización):** $p(S) = 1$.

- **Axioma 2 (Aditividad):** *Para cualquier sucesión infinita,* $A_1, A_2, \ldots$, *de subconjuntos disjuntos de* S*, se cumple la igualdad*

$$p\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} p\left(A_i\right). \tag{3.1}$$

El Axioma 1 establece que, independientemente de nuestro grado de certeza, ocurrirá un elemento del conjunto universal S (es decir, el conjunto S es exhaustivo). El Axioma 2 es una fórmula de agregación que se usa para calcular la probabilidad de la unión de subconjuntos disjuntos. Establece que la incertidumbre de un cierto subconjunto es la suma de las incertidumbres de sus partes (disjuntas). Nótese que esta propiedad también se cumple para sucesiones finitas.

De los axiomas anteriores pueden deducirse propiedades muy interesantes de la probabilidad. Por ejemplo:

- **Propiedad 1 (Normalización):** $p(\phi) = 0$.
- **Propiedad 2 (Monotonicidad):** Si $A \subseteq B \subseteq S$, entonces $p(A) \leq p(B)$.
- **Propiedad 3 (Continuidad-Consistencia):** Para toda sucesión creciente $A_1 \subseteq A_2 \subseteq \ldots$ o decreciente $A_1 \supseteq A_2 \supseteq \ldots$ de subconjuntos de S se tiene

$$\lim_{i \to \infty} p(A_i) = p(\lim_{i \to \infty} A_i).$$

- **Propiedad 4 (Inclusión-Exclusión):** Dado cualquier par de subconjuntos A y B de S, se cumple siempre la siguiente igualdad:

$$p(A \cup B) = p(A) + p(B) - p(A \cap B). \tag{3.2}$$

La Propiedad 1 establece que la evidencia asociada a una ausencia completa de información es cero. La Propiedad 2 muestra que la evidencia de la pertenencia de un elemento a un conjunto debe ser al menos la evidencia de cualquiera de sus subconjuntos. En otras palabras, la evidencia de que un elemento pertenezca a un conjunto dado A no debe decrecer con la adición de elementos a A.

La Propiedad 3 puede ser considerada como una propiedad de consistencia o continuidad. Si se eligen dos sucesiones de conjuntos que convergen al mismo subconjunto de S, se debe obtener la misma evidencia o incertidumbre. La Propiedad 4 establece que las probabilidades de los conjuntos $A, B, A \cap B$, y $A \cup B$ no son independientes, sino que están relacionadas por (3.2).

Un ejemplo clásico que ilustra estos axiomas es el del lanzamiento de un dado no trucado. Aquí el espacio muestral es $S = \{1, 2, 3, 4, 5, 6\}$, es decir, el conjunto de los posibles resultados del lanzamiento. Sea $p(A)$ la probabilidad de que ocurra el suceso A. Entonces, por ejemplo, se tiene $p(S) = 1$, $p(\{1\}) = 1/6$, $p(\{3\}) = 1/6$, y $p(\{1, 3\}) = p(\{1\}) + p(\{3\}) = 1/3$.

3.2.2 Distribuciones de Probabilidad

Sea $\{X_1, \ldots, X_n\}$ un conjunto de variables aleatorias discretas y $\{x_1, \ldots, x_n\}$ el conjunto de sus posibles realizaciones. Nótese que las variables aleatorias se denotan con mayúsculas y que sus realizaciones se denotan con minúsculas. Por ejemplo, si X_i es una variable binaria, entonces x_i puede ser 1 ó 0. Los resultados que siguen son también válidos si las variables son continuas, pero en este caso los símbolos de suma deben sustituirse por integrales.

Sea $p(x_1, \ldots, x_n)$ la *función de probabilidad conjunta*[1] de las variables de X, es decir,

$$p(x_1, \ldots, x_n) = p(X_1 = x_1, \ldots, X_n = x_n). \tag{3.3}$$

Entonces, la *función de probabilidad marginal* de la i-ésima variable se obtiene mediante la fórmula

$$p(x_i) = p(X_i = x_i) = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n} p(x_1, \ldots, x_n). \tag{3.4}$$

El conocimiento de la ocurrencia de un suceso puede modificar las probabilidades de otros sucesos. Por ejemplo, la probabilidad de que un paciente tenga una enfermedad dada puede cambiar tras el conocimiento de los resultados de un análisis de sangre. Por ello, cada vez que se dispone de nueva información, las probabilidades de los sucesos pueden, y suelen, cambiar. Esto conduce al concepto de *probabilidad condicional*.

Definición 3.2 Probabilidad condicional. *Sean X e Y dos conjuntos disjuntos de variables tales que $p(y) > 0$. Entonces, la probabilidad condicional (función de probabilidad condicionada) de X dado $Y = y$ viene dada por*

$$p(X = x|Y = y) = p(x|y) = \frac{p(x, y)}{p(y)}. \tag{3.5}$$

La ecuación (3.5) implica que la función de probabilidad conjunta de X e Y puede escribirse como

$$p(x, y) = p(y)p(x|y). \tag{3.6}$$

Se obtiene un caso particular de (3.5) cuando X es una única variable e Y es un subconjunto de variables. En este caso, (3.5) se convierte en

$$p(x_i|x_1, \ldots, x_k) = \frac{p(x_i, x_1, \ldots, x_k)}{p(x_1, \ldots, x_k)}$$

[1] Cuando las variables son discretas, $p(x_1, \ldots, x_n)$ se llama *función de probabilidad*, y cuando las variables son continuas, se llama *función de densidad*. Por simplicidad, nos referiremos a ambas como *función de probabilidad conjunta* de las variables.

$$= \frac{p(x_i, x_1, \ldots, x_k)}{\sum_{x_i} p(x_i, x_1, \ldots, x_k)}, \tag{3.7}$$

que es la función de probabilidad de la i-ésima variable, X_i, dado el subconjunto de variables $\{X_1, \ldots, X_k\}$. La suma del denominador de (3.7) se extiende a todos los valores posibles de X_i. Nótese que ambas, las fórmulas de la probabilidad marginal en (3.4) y de la probabilidad condicional en (3.5) siguen siendo válidas si la variable X_i se reemplaza por un subconjunto de variables siempre que los conjuntos de variables sean disjuntos. Nótese también que si el conjunto $\{X_1, \ldots, X_k\}$ en (3.5) se sustituye por el conjunto vacío ϕ, entonces (3.5) se convierte en $p(x_i)$. Por ello, puede pensarse de la probabilidad marginal como un caso particular de probabilidad condicional.

3.2.3 Dependencia e Independencia

Definición 3.3 Independencia de dos variables. *Sean X e Y dos subconjuntos disjuntos del conjunto de variables aleatorias $\{X_1, \ldots, X_n\}$. Entonces se dice que X es independiente de Y si y solamente si*

$$p(x|y) = p(x), \tag{3.8}$$

para todos los valores posibles x e y de X e Y; en otro caso, X se dice dependiente de Y.

Nótese que si x e y son valores posibles de X e Y, entonces $p(x) > 0$ y $p(y) > 0$. Por ello, la condición $p(y) > 0$ es natural en el sentido de que no puede observarse $Y = y$ si no se satisface la condición.

La ecuación (3.8) significa que si X es independiente de Y, entonces nuestro conocimiento de Y no afecta nuestro conocimiento sobre X, es decir, Y no tiene información sobre X. También, si X es independiente de Y, pueden combinarse (3.6) y (3.8) para obtener $p(x, y)/p(y) = p(x)$, que implica

$$p(x, y) = p(x)p(y). \tag{3.9}$$

La ecuación (3.9) indica que si X es independiente de Y, entonces la función de probabilidad conjunta de X e Y es igual al producto de sus marginales. En realidad, (3.9) es una definición de independencia equivalente a la (3.8).

Una propiedad importante de la relación de independencia es su *simetría*, es decir, si X es independiente de Y, entonces Y es independiente de X. Esto ocurre porque

$$p(y|x) = \frac{p(x, y)}{p(x)} = \frac{p(x)p(y)}{p(x)} = p(y). \tag{3.10}$$

Por la propiedad de simetría se dice que X e Y son *independientes* o *mútuamente independientes*. La implicación práctica de la simetría es que si el conocimiento de Y es relevante (irrelevante) para X, entonces el conocimiento de X es relevante (irrelevante) para Y.

Los conceptos de dependencia e independencia de dos variables aleatorias pueden ser extendidos al caso de más de dos variables aleatorias como sigue:

Definición 3.4 Independencia de un conjunto de variables. *Las variables aleatorias $\{X_1, \ldots, X_m\}$ se dice que son independientes si y sólo si*

$$p(x_1, \ldots, x_m) = \prod_{i=1}^{m} p(x_i), \tag{3.11}$$

para todos los valores posibles $x_1, \ldots, x_m$ de $X_1, \ldots, X_m$. En otro caso, se dice que son dependientes.

En otras palabras, $\{X_1, \ldots, X_m\}$ se dicen independientes si y sólo si su función de probabilidad conjunta es igual al producto de sus funciones de probabilidad marginal. Nótese que (3.11) es una generalización de (3.9).

Nótese también que si $X_1, \ldots, X_m$ son condicionalmente independientes dado otro subconjunto $Y_1, \ldots, Y_n$, entonces

$$p(x_1, \ldots, x_m | y_1, \ldots, y_n) = \prod_{i=1}^{m} p(x_i | y_1, \ldots, y_n). \tag{3.12}$$

Una implicación importante de la independencia es que no es rentable obtener información sobre variables independientes, pues es irrelevante. Es decir, independencia significa irrelevancia.

Ejemplo 3.1 Cuatro variables. Considérense las siguientes características (variables y sus posibles valores) de las personas de una población dada:

- $Sexo = \{hombre, mujer\}$
- $Fumador = \{sí\ (f), no\ (\bar{f})\}$
- $Estado\ Civil = \{casado\ (c), no\ casado\ (\bar{c})\}$
- $Embarazo = \{sí\ (e), no\ (\bar{e})\}$

La función de probabilidad conjunta de estas cuatro variables se da en la Tabla 3.1. Por ello, por ejemplo, el 50% de las personas de una población son mujeres, y el

$$\frac{0.01 + 0.04 + 0.01 + 0.10}{(0.01 + 0.04 + 0.01 + 0.10) + (0.00 + 0.02 + 0.00 + 0.07)} = 64\%$$

de los fumadores son mujeres.

		hombre		mujer	
		f	$\bar{f}$	f	$\bar{f}$
c	e	0.00	0.00	0.01	0.05
	$\bar{e}$	0.02	0.18	0.04	0.10
$\bar{c}$	e	0.00	0.00	0.01	0.01
	$\bar{e}$	0.07	0.23	0.10	0.18

TABLA 3.1. La función de probabilidad conjunta de las cuatro variables: *Sexo* $(hombre, mujer)$, *Fumador* $(f, \bar{f})$, *Estado Civil* $(c, \bar{c})$ y *Embarazo* $(e, \bar{e})$.

Sea A una persona elegida al azar de la población. Sin conocer si la persona es fumadora, la probabilidad de que se trate de una mujer es $p(A = mujer) = 0.50$. Pero si se sabe que la persona es fumadora, esta probabilidad cambia de 0.50 a $p(A = mujer|A = f) = 0.64$. Por tanto, se tiene $p(A = mujer|A = f) \neq p(A = mujer)$; por lo que las variables *Sexo* y *Fumador* son dependientes.

Supóngase ahora que se sabe que la persona está embarazada. Entonces resulta

$$p(A = mujer|A = e) = 1 \neq p(A = mujer) = 0.50;$$

por lo que, las variables *Sexo* y *Embarazo* son dependientes. Por ello, las dos variables *Fumador* y *Embarazo* contienen información relevante sobre la variable *Sexo*. Sin embargo, el suceso "la persona está embarazada" contiene mucha más información sobre *Sexo* que el suceso "la persona es fumadora." Esto puede medirse por el cociente

$$\frac{p(A = mujer|A = e)}{p(A = mujer|A = f)} = \frac{1}{0.64} > 1.$$

Por otra parte, la variable *Estado Civil* no contiene información relevante sobre la variable *Sexo* y viceversa. Esto puede verse en la Tabla 3.1, en la que las probabilidades conjuntas coinciden con el producto de las marginales para todos los valores posibles de las dos variables. Por ello, las variables *Sexo* y *Estado Civil* son independientes. ∎

Ejemplo 3.2 Distribuciones de probabilidad. Considérese la función de probabilidad conjunta de las tres variables binarias X, Y y Z dadas en la Tabla 3.2. Entonces se tiene:

- Las funciones de probabilidad marginal de X, Y y Z se muestran en la Tabla 3.3. Por ejemplo, la función de probabilidad marginal de X se calcula mediante

$$p(X = 0) = \sum_{y=0}^{1} \sum_{z=0}^{1} p(0, y, z) = 0.12 + 0.18 + 0.04 + 0.16 = 0.5,$$

x	y	z	$p(x,y,z)$
0	0	0	0.12
0	0	1	0.18
0	1	0	0.04
0	1	1	0.16
1	0	0	0.09
1	0	1	0.21
1	1	0	0.02
1	1	1	0.18

TABLA 3.2. Función de probabilidad conjunta de tres variables binarias.

$$p(X=1) = \sum_{y=0}^{1}\sum_{z=0}^{1} p(1,y,z) = 0.09 + 0.21 + 0.02 + 0.18 = 0.5.$$

- Las funciones de probabilidad conjunta de las parejas se dan en la Tabla 3.4. Por ejemplo, la función de probabilidad conjunta de X e Y es

$$\begin{aligned}
p(X=0,Y=0) &= \sum_{z=0}^{1} p(0,0,z) = 0.12 + 0.18 = 0.3,\\
p(X=0,Y=1) &= \sum_{z=0}^{1} p(0,1,z) = 0.04 + 0.16 = 0.2,\\
p(X=1,Y=0) &= \sum_{z=0}^{1} p(1,0,z) = 0.09 + 0.21 = 0.3,\\
p(X=1,Y=1) &= \sum_{z=0}^{1} p(1,1,z) = 0.18 + 0.02 = 0.2.
\end{aligned}$$

- Las funciones de probabilidad condicional de una variable dada la otra se muestran en la Tabla 3.5. Por ejemplo, la probabilidad condicional de X dada Y es

$$\begin{aligned}
p(X=0|Y=0) &= \frac{p(X=0,Y=0)}{p(Y=0)} = \frac{0.3}{0.6} = 0.5,\\
p(X=0|Y=1) &= \frac{p(X=0,Y=1)}{p(Y=1)} = \frac{0.2}{0.4} = 0.5,\\
p(X=1|Y=0) &= \frac{p(X=1,Y=0)}{p(Y=0)} = \frac{0.3}{0.6} = 0.5,
\end{aligned}$$

x	$p(x)$
0	0.5
1	0.5

y	$p(y)$
0	0.6
1	0.4

z	$p(z)$
0	0.27
1	0.73

TABLA 3.3. Funciones de probabilidad marginal.

x	y	$p(x,y)$
0	0	0.3
0	1	0.2
1	0	0.3
1	1	0.2

x	z	$p(x,z)$
0	0	0.16
0	1	0.34
1	0	0.11
1	1	0.39

y	z	$p(y,z)$
0	0	0.21
0	1	0.39
1	0	0.06
1	1	0.34

TABLA 3.4. Funciones de probabilidad conjunta por pares.

y	x	$p(x\|y)$
0	0	0.5
0	1	0.5
1	0	0.5
1	1	0.5

z	x	$p(x\|z)$
0	0	16/27
0	1	11/27
1	0	34/73
1	1	39/73

z	y	$p(y\|z)$
0	0	21/27
0	1	6/27
1	0	39/73
1	1	34/73

TABLA 3.5. Funciones de probabilidad condicional de una variable dada la otra.

$$p(X=1|Y=1) = \frac{p(X=1,Y=1)}{p(Y=1)} = \frac{0.2}{0.4} = 0.5.$$

De los resultados anteriores se ve que $p(x,y) = p(x)p(y)$ para todos los valores posibles de x e y, por tanto, X e Y son independientes. Nótese que esta independencia puede comprobarse también con la definición alternativa de independencia $p(x|y) = p(x)$. Sin embargo, se hace notar que $p(x,z) \neq p(x)p(z)$ para algunos valores (en este caso todos) de x y z. Por tanto, X y Z son dependientes. Similarmente, se puede demostrar que Y y Z son dependientes. ∎

Los conceptos de dependencia e independencia se refieren a dos subconjuntos de variables. Seguidamente, se generaliza el concepto de independencia cuando hay implicados más de dos conjuntos.

Definición 3.5 Dependencia e independencia condicional. *Sean X, Y y Z tres conjuntos disjuntos de variables, entonces X se dice condicionalmente independiente de Y dado Z, si y sólo si*

$$p(x|z,y) = p(x|z), \tag{3.13}$$

para todos los valores posibles de x, y y z de X, Y y Z; En otro caso X e Y se dicen condicionalmente dependientes dado Z.

Cuando X e Y son condicionalmente independientes dado Z, se escribe $I(X, Y|Z)$. La relación $I(X, Y|Z)$ se denomina *relación de independencia condicional*. Similarmente, cuando X e Y son condicionalmente dependientes dado Z, se escribe $D(X, Y|Z)$, que se conoce como una *relación de dependencia condicional*. A veces escribimos $I(X, Y|Z)_p$ o $D(X, Y|Z)_p$ para indicar que la relación se deriva, o es implicada, por el modelo probabilístico asociado a la probabilidad p (la función de probabilidad conjunta).

La definición de independencia condicional lleva en sí la idea de que una vez que es conocida Z, el conocimiento de Y no altera la probabilidad de X. En otras palabras, si Z ya es conocida, el conocimiento de Y no añade información alguna sobre X.

Una definición alternativa, pero equivalente, de independencia condicional es

$$p(x, y|z) = p(x|z)p(y|z). \tag{3.14}$$

La equivalencia de (3.13) y (3.14) puede demostrarse de forma similar a la de (3.8) y (3.9).

Nótese que la independencia (incondicional) puede ser tratada como un caso particular de la independencia condicional. Por ejemplo, se puede escribir $I(X, Y|\phi)$, para indicar que X e Y son incondicionalmente independientes, donde ϕ es el conjunto vacío. Nótese, sin embargo, que X e Y pueden ser independientes incondicionalmente pero condicionalmente dependientes dado Z, es decir, la relación de independencia condicional $I(X, Y|\phi)$ y la de dependencia condicional $D(X, Y|Z)$ pueden satisfacerse simultáneamente.

Ejemplo 3.3 Dependencia e independencia condicional. Considérese la función de probabilidad conjunta de las tres variables binarias X, Y y Z de la Tabla 3.2. En el Ejemplo 3.2 se determina si cualesquiera dos variables son independientes (incondicionalmente). Se tienen las siguientes relaciones de independencia condicional:

$$I(X, Y|\phi), \quad D(X, Z|\phi) \text{ y } D(Y, Z|\phi).$$

Por ejemplo, para determinar si X e Y son independientes, se necesita comprobar si $p(x, y) = p(x)p(y)$ para todos los valores posibles de x e y.

También se puede determinar si cualesquiera dos variables son condicionalmente independientes dada una tercera variable. Por ejemplo, para comprobar si X e Y son condicionalmente independientes dado Z, es necesario comprobar si $p(x|y, z) = p(x, y, z)/p(y, z) = p(x|z)$ para todos los

y	z	x	$p(x\|y,z)$
0	0	0	$12/21 \approx 0.571$
0	0	1	$9/21 \approx 0.429$
0	1	0	$18/39 \approx 0.462$
0	1	1	$21/39 \approx 0.538$
1	0	0	$4/6 \approx 0.667$
1	0	1	$2/6 \approx 0.333$
1	1	0	$16/34 \approx 0.471$
1	1	1	$18/34 \approx 0.529$

z	x	$p(x\|z)$
0	0	$16/27 \approx 0.593$
0	1	$11/27 \approx 0.407$
1	0	$34/73 \approx 0.466$
1	1	$39/73 \approx 0.534$

TABLA 3.6. Funciones de probabilidad obtenidas de la función de probabilidad conjunta de la Tabla 3.2.

valores posibles de x, y y z. Para ello, se calculan las probabilidades

$$p(x|y,z) = \frac{p(x,y,z)}{p(y,z)},$$

$$p(x|z) = \frac{p(x,z)}{p(z)},$$

cuyos valores se muestran en la Tabla 3.6. En esta tabla puede verse que $p(x|y,z) \neq p(x|z)$ y, por tanto, $D(X,Y|Z)$. Por ello, la función de probabilidad conjunta de la Tabla 3.2 implica que X e Y son incondicionalmente independientes, $I(X,Y|\phi)$, aunque son condicionalmente dependientes dado Z, $D(X,Y|Z)$. ∎

3.2.4 Teorema de Bayes

Una conocida fórmula de la teoría de la probabilidad puede obtenerse como sigue. Utilizando (3.3) y (3.5), se obtiene

$$\begin{aligned} p(x_i|x_1,\ldots,x_k) &= \frac{p(x_i,x_1,\ldots,x_k)}{\sum_{x_i} p(x_i,x_1,\ldots,x_k)} \\ &= \frac{p(x_i)p(x_1,\ldots,x_k|x_i)}{\sum_{x_i} p(x_i)p(x_1,\ldots,x_k|x_i)}. \end{aligned} \tag{3.15}$$

La ecuación (3.15) se conoce como *Teorema de Bayes*.

Para ilustrar el uso del teorema de Bayes, supóngase que un paciente puede estar sano (no tiene enfermedad alguna) o tiene una de $m-1$ enfermedades posibles $\{E_1,\ldots,E_{m-1}\}$. Por simplicidad de notación, sea E una variable aleatoria que puede tomar uno de m posibles valores, $\{e_1,\ldots,e_m\}$,

donde $E = e_i$ significa que el paciente tiene la enfermedad E_i, y $E = e_m$ significa que el paciente no tiene ninguna enfermedad. Supóngase también que se tienen n síntomas $\{S_1, \ldots, S_n\}$. Ahora, dado que el paciente tiene un conjunto de síntomas $\{s_1, \ldots, s_k\}$, se desea calcular la probabilidad de que el paciente tenga la enfermedad E_i, es decir, $E = e_i$. Entonces, aplicando el teorema de Bayes, se obtiene

$$p(e_i|s_1, \ldots, s_k) = \frac{p(e_i)p(s_1, \ldots, s_k|e_i)}{\sum_{e_i} p(e_i)p(s_1, \ldots, s_k|e_i)}. \tag{3.16}$$

Conviene hacer los siguientes comentarios sobre la fórmula (3.16):

- La probabilidad $p(e_i)$ se llama probabilidad *marginal*, *prior*, *"a priori"* o *inicial* de la enfermedad $E = e_i$ puesto que puede ser obtenida *antes* de conocer los síntomas.
- La probabilidad $p(e_i|s_1, \ldots, s_k)$ es la probabilidad *posterior*, *"a posteriori"* o *condicional* de la enfermedad $E = e_i$, puesto que se calcula *después* de conocer los síntomas $S_1 = s_1, \ldots, S_k = s_k$.
- La probabilidad $p(s_1, \ldots, s_k|e_i)$ se conoce por el nombre de *verosimilitud* de que un paciente con la enfermedad $E = e_i$ tenga los síntomas $S_1 = s_1, \ldots, S_k = s_k$.

Por ello, se puede utilizar el teorema de Bayes para actualizar la probabilidad "a posteriori" usando ambas, la probabilidad "a priori" y la verosimilitud, tal como se muestra en el ejemplo que sigue:

Ejemplo 3.4 Adenocarcinoma gástrico. Un centro médico tiene una base de datos consistente en las historias clínicas de $N = 1,000$ pacientes. Estas historias clínicas se resumen gráficamente en la Figura 3.1. Hay 700 pacientes (la región sombreada) que tienen la enfermedad *adenocarcinoma gástrico* (G), y 300 que no la tienen (se considera *estar sano* como otro valor posible de la enfermedad). Tres síntomas, *dolor* (D), *pérdida de peso* (P) y *vómitos* (V), se considera que están ligados a esta enfermedad. Por tanto, cuando un paciente nuevo llega al centro médico, hay una probabilidad $700/1,000 = 70\%$ de que el paciente tenga adenocarcinoma gástrico. Esta es la probabilidad inicial, o "a priori", puesto que se calcula con la información inicial, es decir, antes de conocer información alguna sobre el paciente.

Por simplicidad de notación, se utiliza g para indicar que la enfermedad está presente y $\bar{g}$ para indicar que la enfermedad está ausente. Notaciones similares se utilizan para los síntomas. Por tanto, pueden hacerse las afirmaciones siguientes:

- probabilidad "a priori": 440 de 1,000 pacientes vomitan. Por ello, $p(v) = card(v)/N = 440/1,000 = 0.44$, donde $card(v)$ denota el número de pacientes de la base de datos que vomitan. Esto significa que el 44% de los pacientes vomitan.

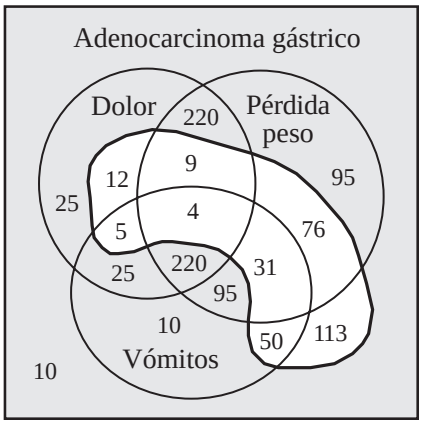


FIGURA 3.1. Pacientes de un centro médico clasificados por una enfermedad (adenocarcinoma gástrico) y tres síntomas (dolor, vómitos y pérdida de peso).

- Verosimilitud: El 50% de los pacientes que tienen la enfermedad vomitan, puesto que $p(v|g) = card(v,g)/card(g) = 350/700 = 0.5$, mientras que sólo 30% de los pacientes que no tienen la enfermedad vomitan, puesto que $p(v|\bar{g}) = card(v,\bar{g})/card(\bar{g}) = 90/300 = 0.3$.

- Verosimilitud: El 45% de los pacientes que tienen la enfermedad vomitan y pierden peso, $p(v,p|g) = card(v,p,g)/card(g) = 315/700 = 0.45$, mientras que sólo el 12% de los que no tienen la enfermedad vomitan y pierden peso, $p(v,p|\bar{g}) = card(v,p,\bar{g})/card(\bar{g}) = 35/300 \approx 0.12$.

Puesto que la probabilidad inicial de que el paciente tenga adenocarcinoma gástrico, $p(g) = 0.7$, no es suficientemente alta para hacer un diagnóstico (nótese que tomar una decisión ahora implica una probabilidad 0.3 de equivocarse), el doctor decide examinar al paciente para obtener más información. Supóngase que los resultados del examen muestran que el paciente tiene los síntomas vómitos ($V = v$) y pérdida de peso ($P = p$). Ahora, dada la evidencia (el paciente tiene esos síntomas), ¿cuál es la probabilidad de que el paciente tenga la enfermedad? Esta probabilidad "a posteriori" puede ser obtenida de la probabilidad "a priori" y de las verosimilitudes, aplicando el teorema de Bayes en dos etapas, como sigue:

- Tras observar que $V = v$ la probabilidad "a posteriori" es

$$\begin{aligned} p(g|v) &= \frac{p(g)p(v|g)}{p(g)p(v|g) + p(\bar{g})p(v|\bar{g})} \\ &= \frac{0.7 \times 0.5}{(0.7 \times 0.5) + (0.3 \times 0.3)} = 0.795. \end{aligned}$$

- Tras observar que $V = v$ y $P = p$ la probabilidad "a posteriori" es

$$\begin{aligned} p(g|v,p) &= \frac{p(g)p(v,p|g)}{p(g)p(v,p|g) + p(\bar{g})p(v,p|\bar{g})} \\ &= \frac{0.7 \times 0.45}{(0.7 \times 0.45) + (0.3 \times 0.12)} = 0.9. \end{aligned} \tag{3.17}$$

Nótese que cuando se aplica el teorema de Bayes sucesivamente, la probabilidad "a posteriori" calculada en una etapa dada es la misma que la probabilidad "a priori" en la etapa siguiente. Por ejemplo, la probabilidad "a posteriori", que se ha calculado en la primera etapa anterior, puede ser usada como probabilidad "a priori" en la segunda etapa, es decir,

$$\begin{aligned} p(g|v,p) &= \frac{p(g|v)p(p|g,v)}{p(g|v)p(p|g,v) + p(\bar{g}|v)p(p|\bar{g},v)} \\ &= \frac{0.795 \times 0.9}{(0.795 \times 0.9) + (0.205 \times 0.389)} = 0.9, \end{aligned}$$

que da la misma respuesta que en (3.17). Nótese también que la probabilidad cambia tras observar las evidencias. La probabilidad de tener la enfermedad era inicialmente 0.7, después aumentó a 0.795, y luego a 0.9 tras observar la evidencia acumulada $V = v$ y $P = p$, respectivamente. Al final de la última etapa, el paciente tiene una probabilidad 0.9 de tener la enfermedad. Esta probabilidad puede ser suficientemente alta (comparada con la probabilidad "a priori" 0.7) para que el doctor diagnostique que el paciente tiene la enfermedad. Sin embargo, sería conveniente observar nuevas evidencias antes de hacer este diagnóstico. ∎

3.2.5 Tipos de Errores

Los síntomas son observables, pero las enfermedades no lo son. Pero, puesto que las enfermedades y los síntomas están relacionados, los médicos utilizan los síntomas para hacer el diagnóstico de las enfermedades. Una dificultad que surge con este tratamiento del problema es que las relaciones entre síntomas y enfermedades no son perfectas. Por ejemplo, los mismos síntomas pueden ser causados por diferentes enfermedades. Estudiando estas relaciones entre síntomas y enfermedades, los médicos pueden aumentar su conocimiento y experiencia, y, por tanto, pueden llegar a ser capaces de diagnosticar enfermedades con un mayor grado de certeza.

Sin embargo, debería ser reconocido que cuando se toman decisiones en ambiente de incertidumbre, estas decisiones pueden ser incorrectas. En situaciones de incertidumbre pueden cometerse dos tipos de errores:

- Una decisión positiva falsa, también conocida como error de tipo I, y

Decisión médica	Estado de la naturaleza	
	Sí	No
Sí	Decisión correcta	Decisión incorrecta (Tipo I)
No	Decisión incorrecta (Tipo II)	Decisión correcta

TABLA 3.7. El doctor está sometido a la posibilidad de cometer uno de los dos errores dependiendo del verdadero estado de la naturaleza.

- Una decisión negativa falsa, también conocida como error de tipo II.

En un caso de diagnóstico médico, por ejemplo, los posibles errores son:

- *Error de Tipo I:* Un paciente no tiene la enfermedad pero el doctor concluye que la tiene.
- *Error de Tipo II:* Un paciente tiene la enfermedad pero el doctor concluye que no la tiene.

Estos tipos de errores se ilustran en la Tabla 3.7. En la realidad (el verdadero estado de la naturaleza), un paciente puede tener o no tener la enfermedad. El doctor tiene que tomar la decisión de si el paciente tiene o no, la enfermedad. Esta decisión es correcta si coincide con el verdadero estado de la naturaleza; en otro caso, la decisión es incorrecta. Por ello, cuando se diagnostica, el doctor está sometido a la posibilidad de cometer uno de los dos errores anteriores dependiendo del verdadero estado de la naturaleza.

Sin embargo, en algunas situaciones las consecuencias de un error pueden ser mucho más graves que las consecuencias del otro. Por ejemplo, si la enfermedad sospechada es cáncer, se puede argüir que el error de Tipo II es más serio que el error de Tipo I. Es cierto que si el paciente no tiene la enfermedad pero el doctor concluye que la tiene, el paciente sufrirá psicológicamente y posiblemente físicamente (debido al efecto del tratamiento o la operación quirúrgica). Por otra parte, si el paciente realmente tiene la enfermedad y el doctor concluye que no la tiene, este error puede conducir a la muerte del paciente.

Idealmente, al doctor le gustaría mantener las probabilidades de cometer esos errores reducidas a un mínimo, pero los riesgos relativos asociados a los dos tipos de errores deben tomarse en consideración cuando se hace un diagnóstico. Como ilustración, supóngase que un nuevo paciente con una enfermedad desconocida viene al centro médico. Tras el examen por un doctor, se determina que el paciente tiene k síntomas, $s_1, s_2, \ldots, s_k$. La pregunta que ambos, doctor y paciente, quieren responder consiste en saber, dados esos síntomas, ¿cuál de las enfermedades es más probable que tenga el paciente?. La respuesta a esta pregunta puede obtenerse sin más que calcular las probabilidades "a posteriori" de $E = e$ para cada

una de las enfermedades $e = e_i$ dados los síntomas $s_1, s_2, \ldots, s_k$, es decir, $p(e_i|s_1, s_2, \ldots, s_k)$. Estas probabilidades pueden calcularse usando (3.16). Por ello, dado que el paciente tiene los síntomas $s_1, s_2, \ldots, s_k$, el doctor puede concluir que la enfermedad más probable del paciente es la que maximice la probabilidad, es decir, $max_i\{p(e_i|s_1, s_2, \ldots, s_k)\}$. Si el valor de $max_i\{p(e_i|s_1, s_2, \ldots, s_k)\}$ está cercano a la unidad, el doctor puede decidir que el paciente tiene la enfermedad correspondiente. En otro caso, es necesario un examen adicional o la identificación de nuevos síntomas.

La ecuación (3.16) puede utilizarse para calcular la nueva probabilidad condicional para cada enfermedad dados todos los síntomas acumulados (información), tal como se ha hecho en el Ejemplo 3.4. Este proceso debe repetirse, añadiendo más evidencia, hasta que la probabilidad $max_i\{p(e_i|s_1, s_2, \ldots, s_k)\}$ sea cercana a la unidad. Cuando esto ocurra, el médico podrá tomar una decisión y terminar el proceso de diagnosis. El criterio de decidir lo que se entiende por *cercana a la unidad* le corresponde al doctor, dependiendo de los riesgos asociados a decisiones erróneas.

Por tanto, es necesario medir las consecuencias de nuestras decisiones. Una forma de hacer ésto es utilizando las llamadas *funciones de utilidad*. Una función de utilidad asigna un valor a toda posible decisión. Sea X la variable aleatoria cuya función de probabilidad es $p(x)$ y sea $u(x)$ el valor asignado por la función de utilidad a la decisión x. Entonces el valor esperado de esta utilidad es

$$E[u] = \sum_x u(x)p(x).$$

Se pueden asignar diferentes funciones de utilidad $u_i(x); i = 1, \ldots, q$ a decisiones diferentes y decidir en favor de la decisión que maximiza la utilidad.

3.3 Reglas Generalizadas

La medida de probabilidad descrita en la Sección 3.2.1 puede utilizarse para medir la incertidumbre y para extender los sistemas basados en reglas a situaciones de incertidumbre. Una forma de introducir la incertidumbre en los sistemas basados en reglas consiste en utilizar *reglas generalizadas*. Por ejemplo, dada la regla determinista

- Regla 1: Si A es cierta, entonces B es cierta,

se puede introducir incertidumbre asociando una probabilidad a esta afirmación

- Regla 2: Si A es cierta, entonces la probabilidad de que B sea cierta es $p(b) = \theta$,

donde $0 \leq \theta \leq 1$ es una medida de la incertidumbre de B. Claramente, La Regla 1 es un caso especial de la Regla 2 puesto que se obtiene de la Regla 2 haciendo $\theta = 1$ (certeza). Pero cuando $0 < \theta < 1$ (incertidumbre), La Regla 1 ya no es apropiada. Por tanto, se puede pensar en la Regla 2 como una regla generalizada. Por ello, el valor de θ determina el nivel de implicación como sigue (véase la Figura 3.2):

- *Implicación fuerte* ($\theta = 1$): En la lógica clásica, la que se ha utilizado hasta aquí en los sistemas expertos basados en reglas (Modus Ponens y Modus Tollens), si la premisa de una regla es cierta, su conclusión debe ser también cierta. Por ello, dada la regla

  Si A es cierta, entonces B es cierta,

  se puede decir que A implica B con probabilidad 1. Esto se ilustra en la Figura 3.2(a).

- *Implicación débil* ($0 < \theta < 1$): La regla anterior puede ser vista en un sentido generalizado cuando A implica B sólo en algunas ocasiones. En este caso, se dice que A implica B con probabilidad $p(B = \textit{cierto} | A = \textit{cierto})$, como se muestra en la Figura 3.2(b).

- *No implicación* ($\theta = 0$): El caso en que A no implica B puede considerarse como que A implica B con probabilidad 0. Esto se ilustra en la Figura 3.2(c).

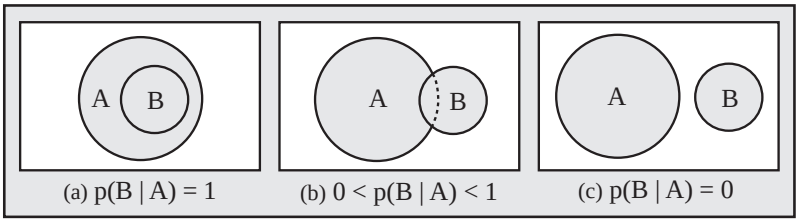


FIGURA 3.2. Ejemplos de implicaciones inciertas: A implica B con probabilidad 1 (a), A implica B con probabilidad θ, donde $0 < \theta < 1$ (b), y A implica B con probabilidad 0 (c).

El uso de reglas generalizadas requiere utilizar medidas de incertidumbre para ambos, objetos y reglas, junto con fórmulas de agregación para combinar la incertidumbre de los objetos en las premisas con la de las reglas para obtener la incertidumbre de los objetos en las conclusiones. Nótese que ahora toda afirmación (hecho) debe estar acompañado por una medida de incertidumbre y que cuando se combinan varios hechos inciertos, deben darse las conclusiones con sus correspondientes medidas de incertidumbre.

Uno de los primeros sistemas expertos que utilizó la probabilidad como medida de incertidumbre fué el PROSPECTOR, un sistema experto para

exploración de mineral (Duda, Hart y Nilsson (1976), Duda, Gaschnig y Hart (1980)). Además de las reglas que forman la base de conocimiento, se asocian probabilidades "a priori" a los objetos del modelo, y probabilidades condicionales a las reglas. Por ello, cuando se observa nueva evidencia, debe utilizarse algún método de propagación de probabilidades para actualizar éstas.

Por ello, se puede tratar la incertidumbre utilizando las reglas generalizadas o esquemas similares. Sin embargo, estos modelos también tienen problemas. Cuando se combinan hechos inciertos, deben darse las conclusiones con sus correspondientes medidas de incertidumbre. Para propagar las incertidumbres de la evidencia observada, son necesarias hipótesis de independencia condicional que pueden no estar justificadas (véase Neapolitan (1990), Capítulo 4). Este es el caso, por ejemplo, de los métodos de la razón de verosimilitud (Duda, Hart y Nilsson (1976)) desarrollados para propagar probabilidades en el sistema experto PROSPECTOR, o los métodos de los factores de certeza utilizados en el sistema MYCIN (véase Buchanan y Shortliffe (1984)).

Una forma alternativa de utilizar la medida de probabilidad consiste en describir las relaciones entre los objetos (variables) mediante una función de probabilidad conjunta. A los sistemas expertos que utilizan las funciones de probabilidad conjunta de las variables como base para hacer la inferencia, se les conoce como *sistemas expertos de tipo probabilístico*. En el resto de este capítulo, se introducen los sistemas expertos probabilísticos, se describe sus componentes, y se comparan con los sistemas expertos basados en reglas.

3.4 Introduciendo los Sistemas Expertos Basados en Probabilidad

El núcleo de los sistemas expertos basados en reglas es el conjunto de reglas que describen las relaciones entre los objetos (variables). En los sistemas expertos probabilísticos las relaciones entre las variables se describen mediante su función de probabilidad conjunta. Por ello, la función de probabilidad conjunta forma parte de lo que se llama conocimiento. Para facilitar la discusión se utiliza un ejemplo de diagnóstico médico (síntomas y enfermedades), pero los conceptos descritos se aplican a otros muchos campos de aplicación. De hecho, el diagnóstico médico es una de las áreas en la que los sistemas expertos han encontrado mayor número de aplicaciones (véase la Sección 1.2), y como se verá en la sección siguiente, algunos modelos de sistemas expertos probabilísticos fueron desarrollados para resolver problemas con la estructura "síntomas-enfermedad".

Ejemplo 3.5 Diagnóstico médico. Supóngase que se dispone de una base de datos con información sobre N pacientes y que un paciente puede tener una y sólo una de m enfermedades, $e_1, \ldots, e_m$, tal como muestra la

Figura 3.3 para $m = 5$ enfermedades. Supóngase también que un paciente puede tener ninguno, uno, o más de n síntomas $S_1, \ldots, S_n$, como indica la Figura 3.4 para $n = 3$ síntomas. Por simplicidad, supóngase que la variable aleatoria enfermedad, E, toma como valores las enfermedades $e_1, \ldots, e_m$. Supóngase también que los síntomas son variables binarias, de forma que cada una toma el valor 1, si está presente, o el valor 0, si está ausente. Nótese que cualquier variable aleatoria en el conjunto $\{E, S_1, \ldots, S_n\}$ define una partición del conjunto universal de pacientes en una clase disjunta y exhaustiva de conjuntos. Entonces, combinando las enfermedades y los síntomas, cada paciente puede clasificarse en una y sólo una región tal como se muestra en la Figura 3.5, que proviene de superponer las Figuras 3.3 y 3.4. Por ejemplo, el círculo negro de la Figura 3.5 representa un paciente que tiene la enfermedad e_4 y los tres síntomas: S_1, S_2 y S_3.

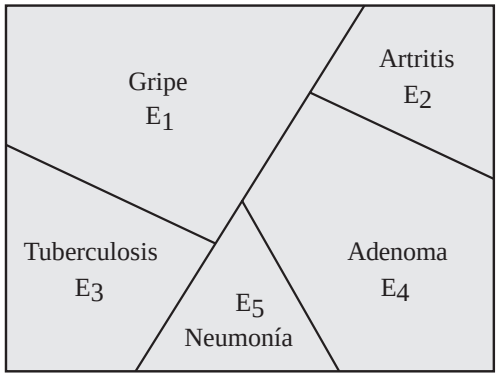


FIGURA 3.3. Una representación gráfica de una población de pacientes clasificados por cinco enfermedades mútuamente exclusivas $e_1 - e_5$.

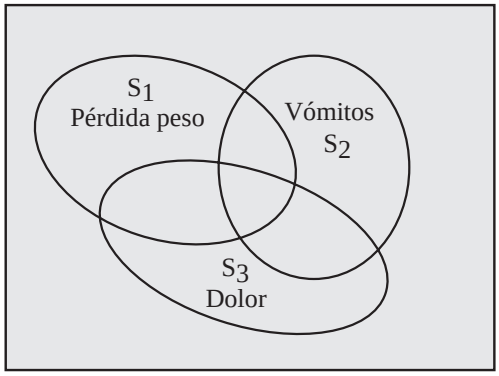


FIGURA 3.4. Una representación gráfica de una población de pacientes clasificados por tres síntomas $S_1 - S_3$.

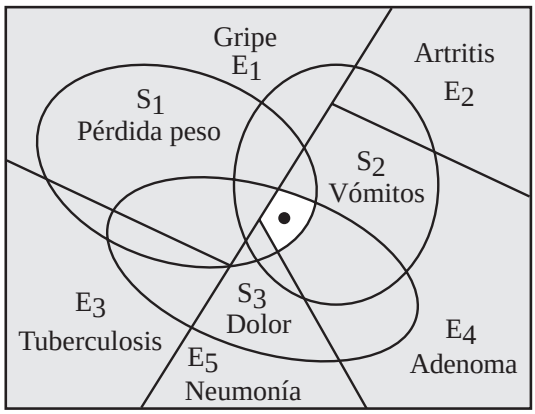


FIGURA 3.5. Una representación gráfica de una población de pacientes clasificados por cinco enfermedades mútuamente exclusivas $e_1 - e_5$ y tres síntomas $S_1 - S_3$.

En este ejemplo, los objetos o variables son la enfermedad E y los síntomas $S_1, \ldots, S_n$. La función de probabilidad conjunta de las variables $(E, S_1, \ldots, S_n)$ está dada por las frecuencias, es decir, el número de pacientes que hay en cada una de las regiones del diagrama de la Figura 3.5. Continuando con la notación introducida en la Sección 3.2.2, una variable se representa mediante una letra mayúscula, mientras que la letra minúscula correspondiente representa uno de sus valores posibles (realizaciones). En este ejemplo, la enfermedad D se supone que toma m valores posibles y los síntomas se suponen binarios. En otras palabras, los posibles valores de E son $e_1, \ldots, e_m$, y los valores posibles de la variable S_j son 1 (presente) ó 0 (ausente). ∎

Las probabilidades asociadas a la enfermedad E pueden ser estimadas mediante

$$p(E = e) \approx card(E = e)/N, \tag{3.18}$$

donde N es el número total de pacientes de la base de datos y $card(E = e)$ es el número de pacientes con $E = e$. Por ejemplo,

- Enfermedad e_1 presente: $p(E = e_1) \approx card(E = e_1)/N$,
- Enfermedad e_1 ausente: $p(E \neq e_1) \approx card(E \neq e_1)/N$.

Un problema que surge con frecuencia en diagnóstico médico es que sólo se observan un subconjunto de síntomas, y basándose en los síntomas observados, se desea diagnosticar con un grado de certeza razonable la enfermedad que dan lugar a los síntomas. En otras palabras, se necesita abordar la cuestión siguiente: Dado que un paciente presenta un subconjunto de k síntomas $S_1 = s_1, \ldots, S_k = s_k$, ¿cuál es la enfermedad que tiene el

Enfermedad e	$p(e\|s_1,\ldots,s_k)$
e_1	0.2
e_2	0.1
e_3	0.8 ← más probable
e_4	0.4
e_5	0.0 ← menos probable
e_6	0.7
⋮	⋮

TABLA 3.8. Probabilidades condicionales de todas las enfermedades e_i, dado el conjunto de síntomas $S_1 = s_1, \ldots, S_k = s_k$.

paciente con mayor probabilidad? Por ello, el problema consiste en calcular la probabilidad de que el paciente tenga la enfermedad e_i, dado el conjunto de valores $s_1, \ldots, s_k$ de los síntomas $S_1, \ldots, S_k$. En otras palabras, para $i = 1, \ldots, m$, se desean calcular las probabilidades condicionales $p(E = e_i | S_1 = s_1, \ldots, S_k = s_k)$. Se puede pensar en éste como un problema de clasificación generalizado: Un paciente puede ser clasificado en uno o más grupos (enfermedades). Por ejemplo, se pueden obtener las probabilities que se muestran en la Tabla 3.8.

Los sistemas expertos probabilísticos pueden utilizarse para resolver éstos y otros problemas. Por ejemplo:

1. Los sistemas expertos pueden memorizar información. Uno puede almacenar y recuperar información de la base de datos. Un ejemplo de tal base de datos se da en la Tabla 3.9, donde se supone que las enfermedades y los síntomas son variables categóricas (binarias o multinomiales). Por ejemplo, la Tabla 3.10 puede representar la información de una base de datos con diez pacientes para el problema de diagnóstico con cinco enfermedades binarias y tres síntomas, también binarios, introducidos en el Ejemplo 3.5.

2. Los sistemas expertos pueden contar o calcular las frecuencias absolutas y relativas de cualquier subconjunto de variables a partir de la base de datos. Estas frecuencias pueden utilizarse para calcular las probabilidades condicionales $p(e_i|s_1, \ldots, s_k)$ aplicando la bien conocida fórmula para la probabilidad condicional [2]

$$p(e_i|s_1, \ldots, s_k) = \frac{p(e_i, s_1, \ldots, s_k)}{p(s_1, \ldots, s_k)}. \tag{3.19}$$

[2]Por simplicidad notacional se escribe $p(E = e_i|S_1 = s_1, \ldots, S_k = s_k)$ en la forma $p(e_i|s_1, \ldots, s_k)$.

	Enfermedad	Síntomas		
Paciente	e	s_1	...	s_n
1	e_m	1	...	1
2	e_1	0	...	0
3	e_3	1	...	0
⋮	⋱	⋮	⋱	⋮
N	e_m	1	...	1

TABLA 3.9. Un ejemplo de una base de datos con N pacientes y sus correspondientes enfermedades y síntomas.

	Enfermedad	Síntomas		
Paciente	E	S_1	S_2	S_3
1	e_5	1	1	1
2	e_2	1	0	1
3	e_3	1	1	0
4	e_5	0	0	1
5	e_3	0	1	0
6	e_1	1	1	0
7	e_1	1	1	1
8	e_3	1	0	0
9	e_1	1	1	1
10	e_5	1	0	1

TABLA 3.10. Un ejemplo de una base de datos con 10 pacientes para el problema del diagnóstico médico del Ejemplo 3.5.

Esta probabilidad puede ser estimada mediante

$$\frac{card(e_i, s_1, \ldots, s_k)}{card(s_1, \ldots, s_k)}, \tag{3.20}$$

donde $card(e_i, s_1, \ldots, s_k)$ es la frecuencia de aparición en la base de datos de los pacientes que tienen los valores indicados de las variables. Por ejemplo, dada la base de datos con diez pacientes de la Tabla 3.10, se pueden calcular las frecuencias asociadas a cualquier combinación de valores de síntomas y enfermedades sin más que contar el número de casos de la base de datos que coinciden con la evidencia. Por ejemplo, $card(E \neq e_1 | S_1 = 1, S_2 = 1) = 2$ puesto que hay dos pacientes (los pacientes 1 y 3) que no presentan la enfermedad e_1 pero muestran los síntomas S_1 y S_2. Similarmente, $card(E = e_1 | S_1 = 1, S_2 = 1) = 3$, $card(S_1 = 1, S_2 = 1) = 5$, etc. Entonces, ésta puede calcularse usando (3.3), las probabilidades condicionales asociadas a una enfermedad

dada y un conjunto de síntomas. Por ejemplo:

$$p(E \neq e_1|S_1 = 1, S_2 = 1) \approx \frac{card(E \neq e_1|S_1 = 1, S_2 = 1)}{card(S_1 = 1, S_2 = 1)} = \frac{2}{5} = 0.4,$$
$$p(E = e_1|S_1 = 1, S_2 = 1) \approx \frac{card(E = e_1|S_1 = 1, S_2 = 1)}{card(S_1 = 1, S_2 = 1)} = \frac{3}{5} = 0.6.$$

3. Los sistemas expertos pueden aprender de la experiencia. Tan pronto como un nuevo paciente es examinado y diagnosticado, se añade la nueva información a la base de datos y se cambian las frecuencias como corresponda. Por ejemplo, si un nuevo paciente que presenta los síntomas $S_1 = 1, S_2 = 1$ y $S_3 = 0$ se sabe que tiene la enfermedad e_1, se puede actualizar la base de datos con esta nueva información sin más que incluir un caso más en la base de datos de la Tabla 3.10.

4. Los sistemas expertos pueden tomar (o ayudar a los expertos humanos a tomar) decisiones tales como:

   - ¿se tiene suficiente información como para diagnosticar la enfermedad?
   - ¿se necesitan nuevas pruebas clínicas? y si la respuesta es positiva, ¿qué prueba o pruebas suministran la máxima información sobre la enfermedad que se sospecha tiene el paciente?

En las tres secciones siguientes se describen las tres componentes principales de los sistemas expertos probabilísticos.

3.5 La Base de Conocimiento

Tal como se ha visto en el Capítulo 2, la base de conocimiento de un sistema experto basado en reglas consta del conjunto de objetos (variables) y del conjunto de reglas. La base de conocimiento de un sistema experto probabilístico consiste en un conjunto de variables, $\{X_1, \ldots, X_n\}$, y una función de probabilidad conjunta definida sobre ellas, $p(x_1, \ldots, x_n)$. Por ello, para construir la base de conocimiento de un sistema experto probabilístico, se necesita definir la función de probabilidad conjunta de las variables.

El modelo más general posible se basa en especificar directamente la función de probabilidad conjunta; es decir, asignar un valor numérico (parámetro) a cada una de las posibles combinaciones de valores de las variables. Desgraciadamente, la especificación directa de la función de probabilidad conjunta implica un gran número de parámetros. Por ejemplo, con n variables binarias, la función de probabilidad conjunta más general tiene 2^n parámetros (las probabilidades $p(x_1, \ldots, x_n)$ para toda posible realización $\{x_1, \ldots, x_n\}$ de las variables), un número tan grande que no hay ordenador

en el mundo capaz de almacenarlo incluso para un valor de n tan pequeño como 50. Esta fue una de las primeras críticas al uso de la probabilidad en los sistemas expertos. Sin embargo, en la mayor parte de las situaciones prácticas, muchos subconjuntos de variables pueden ser independientes o condicionalmente independientes. En tales casos, se pueden obtener simplificaciones del modelo más general teniendo en cuenta la estructura de independencia de las variables. Esto suele dar lugar a una reducción importante del número de parámetros. En esta sección se discuten los siguientes ejemplos de tales simplificaciones:

1. El Modelo de Síntomas Dependientes (MSD).
2. El Modelo de Síntomas Independientes (MSI).
3. El Modelo de Síntomas Relevantes Independientes (MSRI).
4. El Modelo de Síntomas Relevantes Dependientes (MSRD).

Sin embargo, estos cuatro modelos son modelos *ad hoc* que se aplican principalmente en el campo médico (véase Castillo y Álvarez (1991)). Modelos probabilísticos más generales y potentes (por ejemplo, modelos de redes de Markov, modelos de redes Bayesianas, y modelos especificados condicionalmente) se presentan en los Capítulos 6 y 7. Estos modelos pueden utilizarse en el campo médico y también en otros campos.

Para introducir los modelos anteriores se considera el problema del diagnóstico médico introducido en la Sección 3.2.4, en la que se tenían n síntomas $S_1, \ldots, S_n$, y una variable aleatoria E, que podía tomar uno de m valores posibles $e_1, \ldots, e_m$. En este problema se desea diagnosticar la presencia de una enfermedad dado un conjunto de síntomas $s_1, \ldots, s_k$. Para ello se tiene la función de probabilidad conjunta de la enfermedad y los síntomas $p(e, s_1, \ldots, s_n)$.

Como se ha indicado con anterioridad, la forma más general de esta función de probabilidad conjunta depende de un número muy grande de parámetros. Para reducir el número de parámetros, se pueden imponer algunas hipótesis (restricciones) entre ellos. Los modelos presentados en las subsecciones siguientes son ejemplos de tales restricciones. En todos ellos, se supone que las enfermedades son independientes de los síntomas.

3.5.1 El Modelo de Síntomas Dependientes

En este modelo, se supone que los síntomas son dependientes pero que las enfermedades son independientes entre sí, dados los síntomas. El MSD se ilustra en la Figura 3.6, donde todo síntoma se conecta con los demás síntomas y con todo valor posible de E (indicando dependencia).

Entonces la función de probabilidad conjunta para el MSD puede escribirse como

$$p(e_i, s_1, \ldots, s_n) = p(s_1, \ldots, s_n)p(e_i|s_1, \ldots, s_n). \tag{3.21}$$

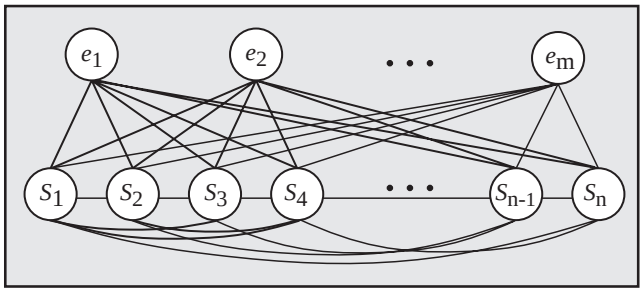


FIGURA 3.6. Una ilustración gráfica del modelo de síntomas dependientes.

Nótese que esta ecuación se obtiene utilizando (3.6) con $X = \{E\}$ e $Y = \{S_1, \ldots, S_n\}$. Ahora, $p(e_i|s_1, \ldots, s_n)$ puede expresarse como

$$p(e_i|s_1, \ldots, s_n) = \frac{p(e_i, s_1, \ldots, s_n)}{p(s_1, \ldots, s_n)} \tag{3.22}$$

$$= \frac{p(e_i)p(s_1, \ldots, s_n|e_i)}{p(s_1, \ldots, s_n)} \tag{3.23}$$

$$\propto p(e_i)p(s_1, \ldots, s_n|e_i). \tag{3.24}$$

La primera de las ecuaciones anteriores se deduce de (3.3), y la segunda se obtiene aplicando (3.6). La proporcionalidad se sigue de que $p(s_1, \ldots, s_n)$ es una constante de normalización.

Nótese que (3.24) sólo incluye probabilidades "a priori" y verosimilitudes (probabilidades condicionales de los síntomas para cada una de las enfermedades) cuyos valores pueden estimarse a partir de la información objetiva dada por las frecuencias de enfermedades y síntomas en la población. La ecuación (3.24) muestra que los parámetros necesarios para la base de datos del MSD son:

- Las probabilidades marginales $p(e_i)$, para todos los valores posibles de E.
- Las verosimilitudes $p(s_1, \ldots, s_n|e_i)$, para todas las combinaciones posibles de síntomas y enfermedades.

Por ejemplo, para m enfermedades y n síntomas binarios, la función de probabilidad marginal de E, $p(e_i)$, depende de $m - 1$ parámetros (puesto que los m parámetros deben sumar uno). Por ello, se necesita especificar $m - 1$ parámetros para la función de probabilidad marginal de E. Con respecto a las verosimilitudes $p(s_1, \ldots, s_n|e_i)$, se necesita especificar $(2^n - 1)$ parámetros[3] para cada valor posible de E, un total de $m(2^n - 1)$ parámetros.

[3]Nótese que para n síntomas binarios hay 2^n parámetros (un parámetro para cada combinación posible de síntomas). Sin embargo, estos parámetros deben

e	$p(e)$
$\bar{g}$	0.3
g	0.7

			$p(d,v,p\|e)$	
d	v	p	$E=g$	$E=\bar{g}$
0	0	0	0.014	0.377
0	0	1	0.136	0.253
0	1	0	0.014	0.167
0	1	1	0.136	0.103
1	0	0	0.036	0.040
1	0	1	0.314	0.030
1	1	0	0.036	0.017
1	1	1	0.314	0.013

TABLA 3.11. Funciones de probabilidad requeridas para la especificación del MSD.

Por ello, el MSD requiere un total de $m - 1 + m(2^n - 1) = m2^n - 1$ parámetros.

Ejemplo 3.6 El Modelo se síntomas dependientes. Para ilustrar el MSD considérense los datos del Ejemplo 3.4, que se dan en la Figura 3.1. En este caso, la única enfermedad de interés es el adenocarcinoma gástrico. Por ello, la variable E toma dos posibles valores, g (cuando un paciente tiene adenocarcinoma gástrico) y $\bar{g}$ (cuando un paciente no tiene adenocarcinoma gástrico). Hay tres síntomas binarios, D, V y P. Es conveniente a veces utilizar los números 1 y 0 para indicar la presencia y la ausencia del síntoma, respectivamente. Para definir el MSD, se necesita conocer la función de probabilidad marginal $p(e_i)$ y las funciones de probabilidad condicional de los síntomas dada la enfermedad, $p(d, v, p|e_i)$. Estas funciones de probabilidad se extraen de la Figura 3.1 y están tabuladas en la Tabla 3.11.

Utilizando (3.24) y la función de probabilidad de la Tabla 3.11, se puede calcular la probabilidad de la enfermedad dada cualquier combinación de síntomas. Estas probabilidades están dadas en la Tabla 3.12. Por ejemplo, la función de probabilidad condicionada de la enfermedad dado que estén presentes los tres síntomas se calcula como sigue:

$$p(\bar{g}|d, v, p) \propto p(\bar{g})p(d, v, p|\bar{g}) = 0.3 \times 0.013 = 0.0039,$$
$$p(g|d, v, p) \propto p(g)p(d, v, p|g) = 0.7 \times 0.314 = 0.2198.$$

Dividiendo ahora por la constante de normalización $0.2198 + 0.0039 = 0.2237$, se obtiene

$$p(\bar{g}|d, v, p) = 0.0039/0.2237 = 0.02,$$
$$p(g|d, v, p) = 0.2198/0.2237 = 0.98,$$

sumar uno; en consecuencia, se tienen sólo $2^n - 1$ parámetros libres para cada valor posible de E.

d	v	p	$E = g$	$E = \bar{g}$
0	0	0	0.08	0.92
0	0	1	0.56	0.44
0	1	0	0.17	0.83
0	1	1	0.75	0.25
1	0	0	0.68	0.32
1	0	1	0.96	0.04
1	1	0	0.83	0.17
1	1	1	0.98	0.02

TABLA 3.12. La función de probabilidad condicionada $p(e|d, v, p)$ con $e = \bar{g}$ y $e = g$, para el modelo MSD.

que se dan en la última fila de la Tabla 3.12. ∎

El principal problema del MSD es que requiere un número muy alto de parámetros. Claramente, especificar las frecuencias para todas esas combinaciones es muy difícil y se hace imposible al crecer los números de las enfermedades y los síntomas. Por ejemplo, con 100 enfermedades y 200 síntomas (que no es una situación irreal), el número de frecuencias (parámetros) necesarios es mayor que 10^{62}, tan grande que no hay ningún ordenador capaz de almacenarla.

La discusión anterior supone síntomas binarios (síntomas con sólo dos posibles opciones, tales como fiebre, no fiebre; dolor, no dolor; etc.). Las dificultades se incrementan notablemente en el MSD en casos en los que se tengan síntomas con múltiples (más de dos) opciones o niveles, tales como fiebre alta, fiebre media, fiebre baja y no fiebre.

3.5.2 El Modelo de Síntomas Independientes

Debido a la imposibilidad de trabajar con el modelo anterior en muchos casos prácticos, resulta necesario proceder a la simplificación del modelo. Una simplificación posible consiste en suponer que, para una enfermedad dada, los síntomas son condicionalmente independientes entre sí. El modelo resultante se denomina *modelo de síntomas independientes* (MSI). El MSI se ilustra en la Figura 3.7, donde los síntomas no están ligados, para indicar la independencia.

Puesto que los síntomas se suponen condicionalmente independientes dada la enfermedad, se tiene

$$p(s_1, \ldots, s_n|e_i) = \prod_{j=1}^{n} p(s_j|e_i). \tag{3.25}$$

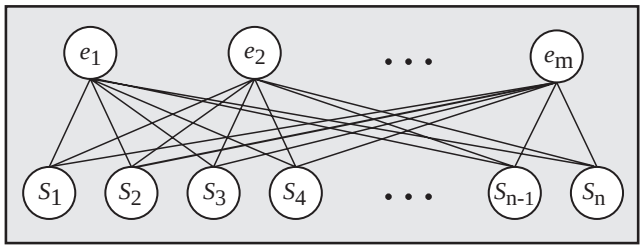


FIGURA 3.7. Una ilustración gráfica del modelo de síntomas independientes.

Por ello, se puede escribir la función de probabilidad conjunta de la enfermedad E dados los síntomas $s_1, \ldots, s_n$ como

$$\begin{aligned} p(e_i|s_1, \ldots, s_n) &= \frac{p(e_i)p(s_1, \ldots, s_n|e_i)}{p(s_1, \ldots, s_n)} \\ &= \frac{p(e_i) \prod_{j=1}^{n} p(s_j|e_i)}{p(s_1, \ldots, s_n)} \qquad (3.26) \\ &\propto p(e_i) \prod_{j=1}^{n} p(s_j|e_i). \qquad (3.27) \end{aligned}$$

Sustituyendo (3.26) en (3.21), se obtiene el MSI.

La ecuación (3.26) muestra cómo la hipótesis de independencia modifica las probabilidades de todas las enfermedades cuando se conocen nuevos síntomas. Por ello, la probabilidad inicial de la enfermedad e_i es $p(e_i)$, pero tras conocer los síntomas s_j, para $j = 1, \ldots, k$, resulta proporcional a $p(s_j|e_i)$. Nótese que cada nuevo síntoma conduce a un nuevo factor. Nótese también que $p(s_1, \ldots, s_n)$, en el denominador de (3.26), es una constante de normalización que no es necesario calcular directamente.

A partir de (3.27), puede verse que los parámetros necesarios para la base de conocimiento del MSI son

- las probabilidades marginales $p(e_i)$, para todos los valores posibles de la enfermedad E.
- Las probabilidades condicionales $p(s_j|e_i)$, para todos los valores posibles del síntoma S_j y la enfermedad E.

Por ello, con las hipótesis de independencia de los síntomas, el número de parámetros se reduce considerablemente. Con m enfermedades posibles y n síntomas binarios, el número total de parámetros es $m(n + 1) - 1$. Por ejemplo, con $m = 100$ enfermedades y $n = 200$ síntomas, se tienen 20,099 parámetros en el MSI en vez de más de 10^{62} parámetros para el MSD.

Ejemplo 3.7 El Modelo de síntomas independientes. Para ilustrar el MSI, se utilizan los historiales clínicos de dos centros médicos, cada uno

Centro Médico 1						
		g		$\bar{g}$		
		d	$\bar{d}$	d	$\bar{d}$	Total
v	p	220	95	4	31	350
	$\bar{p}$	25	10	5	50	90
$\bar{v}$	p	220	95	9	76	400
	$\bar{p}$	25	10	12	113	160
Total		490	210	30	270	1000

Centro Médico 2						
		g		$\bar{g}$		
		d	$\bar{d}$	d	$\bar{d}$	Total
v	p	140	210	0	0	350
	$\bar{p}$	0	0	30	60	90
$\bar{v}$	p	280	0	0	120	400
	$\bar{p}$	70	0	0	90	160
Total		490	210	30	270	1000

TABLA 3.13. Números de pacientes clasificados por una enfermedad G y tres síntomas, D, V y P en dos centros médicos.

e	$p(e)$
$\bar{g}$	0.3
g	0.7

e	d	$p(d\|e)$
$\bar{g}$	0	0.9
$\bar{g}$	1	0.1
g	0	0.3
g	1	0.7

e	v	$p(v\|e)$
$\bar{g}$	0	0.7
$\bar{g}$	1	0.3
g	0	0.5
g	1	0.5

e	p	$p(p\|e)$
$\bar{g}$	0	0.6
$\bar{g}$	1	0.4
g	0	0.1
g	1	0.9

TABLA 3.14. Probabilidades requeridas para la especificación del MSI.

de ellos consta de $N = 1000$ pacientes; dos valores de la enfermedad (g y $\bar{g}$); y tres síntomas, D, V y P. Los datos se resumen en la Tabla 3.13. Nótese que los datos del Centro Médico 1 son los mismos que los de la Figura 3.1, pero dados ahora en forma tabular, en vez de forma gráfica.

Para especificar el MSI, se necesita la probabilidad marginal, $p(e_i)$, de la enfermedad y las probabilidades condicionales de cada síntoma dada cada enfermedad, $p(d|e_i), p(v|e_i)$ y $p(p|e_i)$. Estas probabilidades se extraen de la Tabla 3.13 y se dan en la Tabla 3.14. Nótese que sólo 7 parámetros son libres. Un aspecto interesante de los dos conjuntos de datos es que aunque son muy diferentes, conducen a idénticas probabilidades, como se muestra en la Tabla 3.14.

En la Tabla 3.15 se da la probabilidad condicional de E dadas varias combinaciones de los síntomas para los dos centros médicos. Nótese que

			Centro Médico 1		Centro Médico 2	
d	v	p	Valor Real	MSI	Valor Real	MSI
0	0	0	0.08	0.08	0.00	0.08
0	0	1	0.56	0.56	0.00	0.56
0	1	0	0.17	0.18	0.00	0.18
0	1	1	0.75	0.74	1.00	0.74
1	0	0	0.68	0.66	1.00	0.66
1	0	1	0.96	0.96	1.00	0.96
1	1	0	0.83	0.82	0.00	0.82
1	1	1	0.98	0.98	1.00	0.98

TABLA 3.15. La probabilidad condicional $p(g|d, v, p)$ para los datos de la Tabla 3.13. Los valores verdaderos se calculan utilizando la definición de probabilidad condicional en (3.3). Los valores correspondientes al MSI se calculan aplicando la fórmula del MSI dada en (3.27). Nótese que $p(\bar{g}|d, v, p) = 1 - p(g|d, v, p)$.

$p(\bar{g}|d, v, p) = 1 - p(g|d, v, p)$. Los valores exactos se calculan directamente de la Tabla 3.13 utilizando la definición de probabilidad condicional dada en (3.3). Los valores de las columnas etiquetadas MSI se calculan aplicando la fórmula para el MSI en (3.27). Por ejemplo, para el Centro Médico 1, el valor de $p(g|d, v, p)$ se calcula mediante

$$p(g|d, v, p) = \frac{p(g, d, v, p)}{p(d, v, p)} = \frac{220}{220 + 4} = 0.98.$$

El valor de $p(g|d, v, p)$ según el MSI se calcula usando (3.27) como sigue:

$$p(g|d, v, p) \propto p(g)p(d|g)p(v|g)p(p|g) = 0.7 \times 0.7 \times 0.5 \times 0.9 = 0.2205,$$
$$p(\bar{g}|d, v, p) \propto p(\bar{g})p(d|\bar{g})p(v|\bar{g})p(p|\bar{g}) = 0.3 \times 0.1 \times 0.3 \times 0.4 = 0.0036.$$

Dividiendo 0.2205 por la constante de normalización $0.2205 + 0.0036 = 0.2241$, se obtiene $p(g|d, v, p) = 0.2205/0.2241 = 0.98$ y $p(\bar{g}|d, v, p) = 0.0036/0.2241 = 0.02$.

Una comparación entre las probabilidades verdaderas y las correspondientes al MSI de la Tabla 3.15 muestra que los dos conjuntos de probabilidades son parecidos para el Centro Médico 1, pero discrepan notablemente para el Centro Médico 2. Por ejemplo, para el Centro Médico 2 el valor real de $p(g|d, v, \bar{p})$ es 0, mientras que el correspondiente al MSI es 0.82. Esto es una prueba de que el MSI falla al tratar de describir la probabilidad de los datos del Centro Médico 2. Nótese que se tienen dos conjuntos de datos con las mismas probabilidades "a priori" y las mismas verosimilitudes; sin embargo, el MSI es apropiado para reproducir uno de ellos y no, para el otro. De este ejemplo puede concluirse que las probabilidades "a priori" y las verosimilitudes no son suficientes para especificar un modelo probabilístico. ∎

El Ejemplo 3.7 ilustra el hecho de que el correcto comportamiento de un sistema experto probabilístico se basa en la especificación correcta de la función de probabilidad conjunta. Por tanto, debe ponerse especial cuidado en la selección del modelo probabilístico a utilizar en un caso dado.

Aunque la hipótesis de independencia da lugar a una gran reducción del número de parámetros, el número de parámetros en el MSI es todavía muy alto para ser práctico. Por tanto, se necesita simplificar el modelo aún más.

3.5.3 Modelo de Síntomas Relevantes Independientes

Se puede conseguir una redución aún mayor del número de parámetros suponiendo que cada enfermedad tiene un número reducido de síntomas relevantes. En consecuencia, para cada valor e_i de la enfermedad E se seleccionan algunos síntomas relevantes $S_1, \ldots, S_r$ (relativamente pocos frente al total de síntomas) y los restantes síntomas se suponen independientes para ese valor de E. El MSRI se ilustra en la Figura 3.8. Nótese que para e_1, el conjunto de síntomas relevantes es $\{S_1, S_2\}$; para e_2, el conjunto de síntomas relevantes es $\{S_2, S_3, S_4\}$; y así sucesivamente.

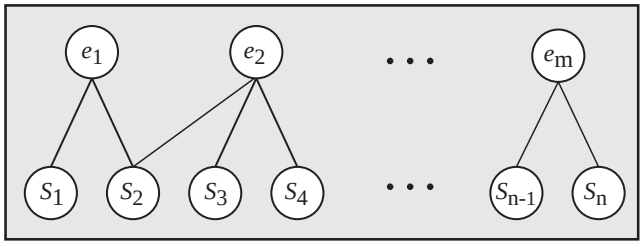


FIGURA 3.8. Una ilustración gráfica del modelo de síntomas relevantes independientes.

Por simplicidad de notación, supóngase que $S_1, \ldots, S_{r_i}$ son relevantes para la enfermedad e_i y que los restantes síntomas $S_{r_i+1}, \ldots, S_n$ son irrelevantes. Según el MSRI, $p(s_j|e_i)$ se supone idéntica para todos los síntomas que son irrelevantes para la enfermedad e_i. Entonces la función de probabilidad conjunta de la enfermedad e_i dados los síntomas $s_1, \ldots, s_n$ puede escribirse como sigue

$$\begin{aligned} p(e_i|s_1, \ldots, s_n) &= \frac{p(e_i)p(s_1, \ldots, s_n|e_i)}{p(s_1, \ldots, s_n)} \\ &= \frac{p(e_i) \prod_{j=1}^{r_i} p(s_j|e_i) \prod_{j=r_i+1}^{n} p(s_j|e_i)}{p(s_1, \ldots, s_n)} \\ &= \frac{p(e_i) \prod_{j=1}^{r_i} p(s_j|e_i) \prod_{j=r_i+1}^{n} p_j}{p(s_1, \ldots, s_n)} \end{aligned} \tag{3.28}$$

$$\propto p(e_i) \prod_{j=1}^{r_i} p(s_j|e_i) \prod_{j=r_i+1}^{n} p_j, \tag{3.29}$$

donde $p_j = p(s_j|e_i)$, que es la misma para todas las enfermedades para la que S_j es irrelevante. Sustituyendo (3.28) en (3.21), se obtiene el MSRI.

De (3.29), se deduce que es necesario almacenar las probabilidades siguientes en la base de conocimiento del MSRI:

- Las probabilidades marginales $p(e_i)$, para todos los valores posibles de la enfermedad E.
- Las probabilidades condicionales $p(s_j|e_i)$, para cada valor posible de E y cada uno de sus correspondientes síntomas relevantes.
- Las probabilidades p_j, para cada valor posible de E que tiene al menos un síntoma irrelevante. (Esto implica que $p_j = p(s_j|e_i)$ es idéntica para todos los síntomas irrelevantes para e_i.)

La ecuación (3.28) implica que en la base de conocimiento se necesita almacenar las probabilidades de todos los síntomas relevantes para cada enfermedad, y la misma probabilidad para todos los síntomas irrelevantes para cada valor de E. Por ello, si se tienen m posibles enfermedades y n síntomas binarios, el número de parámetros en el MSRI es

$$m - 1 + n - a + \sum_{i=1}^{m} r_i, \tag{3.30}$$

donde r_i es el número de síntomas relevantes para la enfermedad e_i y a es el número de síntomas que son relevantes para todas las enfermedades. El número de parámetros se reduce significativamente cuando r_i es mucho menor que n. Por ejemplo, con 100 enfermedades y 200 síntomas, si $r_i = 10$ para todas las enfermedades,[4] el número de parámetros en el MSRI se reduce de 20,099 para el MSI a 1,299 para el MSRI.

Nótese que se puede obtener el MSRI a partir del MSI, sin más que imponer algunas restricciones adicionales en los parámetros del MSI, puesto que en el MSRI las probabilidades $p(s_j|e_i)$ deben ser las mismas para todos los síntomas irrelevantes para las enfermedades e_i. El número de restricciones es

$$a - n + \sum_{j=1}^{n} n_j,$$

donde n_j es el número de enfermedades para las que S_j es irrelevante. Por ello, el número de parámetros en el MSRI coincide con el número de

[4]Nótese que $r_i = 10$ para todas las enfermedades implica que $a = 0$, es decir, toda enfermedad tiene al menos un síntoma irrelevante.

parámetros en el MSI, $(m(n+1)-1)$, menos el número de restricciones. En total, se tiene

$$m(n+1)-1+n-a-\sum_{j=1}^{n} n_j, \tag{3.31}$$

que es la misma fórmula obtenida en (3.30).

3.5.4 El Modelo de Síntomas Relevantes Dependientes

Aunque el MSRI reduce el número de parámetros considerablemente, desgraciadamente, es poco realista, ya que los síntomas asociados a ciertas enfermedades suelen producirse en grupos o síndromes. Por ello, puede ser poco razonable suponer que los síntomas relevantes son independientes. El *modelo de síntomas relevantes dependientes* (MSRD) evita este inconveniente. El MSRD es el mismo que el MSRI pero sin obligar a los síntomas relevantes a ser independientes, dada la correspondiente enfermedad. De esta forma, se supone que sólo los síntomas irrelevantes son independientes pero los síntomas relevantes pueden ser dependientes. Por ello, se puede pensar en el MSRD como una solución de compromiso entre el MSD y el MSRI. El MSRD se ilustra en la Figura 3.9, donde los síntomas relevantes para cada enfermedad están conectados, indicando la dependencia.

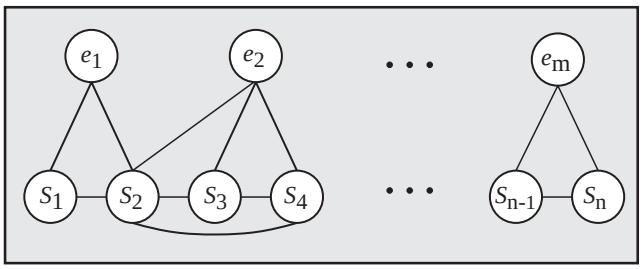


FIGURA 3.9. Una ilustración gráfica del modelo de síntomas relevantes dependientes.

Supóngase que $S_1,\ldots,S_{r_i}$ son relevantes para la enfermedad e_i y que los restantes síntomas $S_{r_i+1},\ldots,S_n$ son irrelevantes. Entonces según el MSRD, la función de probabilidad conjunta de e_i dados los síntomas $s_1,\ldots,s_n$ puede escribirse como

$$\begin{aligned} p(e_i|s_1,\ldots,s_n) &= \frac{p(e_i)p(s_1,\ldots,s_{r_i}|e_i)\prod_{j=r_i+1}^{n} p(s_j|e_i)}{p(s_1,\ldots,s_n)} \\ &= \frac{p(e_i)p(s_1,\ldots,s_{r_i}|e_i)\prod_{j=r_i+1}^{n} p_j}{p(s_1,\ldots,s_n)} \end{aligned} \tag{3.32}$$

Modelo	Número de parámetros	
	Fórmula	Valor
MSD	$m2^n - 1$	$> 10^{62}$
MSI	$m(n+1) - 1$	20,099
MSRI	$m(r+1) + n - 1$	1,299
MSRD	$m2^r + n - 1$	102,599

TABLA 3.16. Números de parámetros necesarios para especificar cuatro modelos en el caso de $m = 100$ enfermedades binarias, $n = 200$ síntomas binarios, y $r = 10$ síntomas relevantes por enfermedad.

$$\propto p(e_i)p(s_1, \ldots, s_{r_i}|e_i) \prod_{j=r_i+1}^{n} p_j, \tag{3.33}$$

donde $p_j = p(s_j|e_i)$, que es la misma para todas las enfermedades para las que S_j es irrelevante. Sustituyendo (3.32) en (3.21), se obtiene el MSRD. Para este modelo, es necesario almacenar las siguientes probabilidades en la base de datos:

- Las probabilidades marginales $p(e_i)$, para todos las posibles valores de la enfermedad E.
- Las probabilidades condicionales $p(s_1, \ldots, s_{r_i}|e_i)$, para todos los posibles valores de la enfermedad E y sus síntomas relevantes $S_1, \ldots, S_{r_i}$.
- Las probabilidades p_j, para cada valor posible de E que tenga al menos un síntoma irrelevante. (Como en el MSRI, esto implica que $p_j = p(s_j|e_i)$ coincide para todos los síntomas irrelevantes para e_i.)

En consecuencia, para m enfermedades binarias y n síntomas binarios, el número total de parámetros en el MSRD es

$$m - 1 + n - a + \sum_{i=1}^{m}(2^{r_i} - 1) = n - 1 - a + \sum_{i=1}^{m} 2^{r_i}. \tag{3.34}$$

Nótese que cuando $r_i = r$ para todos los valores e_i, entonces (3.34) resulta $m2^r + n - 1$. Nótese también que si todos los síntomas son relevantes para todas las enfermedades ($a = n$ y $r_i = n$ para todo e_i), el MSRD se convierte en el MSD. La Tabla 3.16 muestra una comparación de los números de parámetros necesarios para especificar los modelos discutidos en esta sección en el caso de $m = 100$ enfermedades binarias, $n = 200$ síntomas binarios, y $r = 10$ síntomas relevantes por enfermedad.

En el MSRD el número de parámetros es muy reducido comparado con el MSD, y eso a pesar de que es un modelo realista, puesto que considera las dependencias entre los síntomas más importantes (relevantes) para cada

enfermedad. Sin embargo, debido a la hipótesis de dependencia, el número de parámetros del MSRD es mayor que el número de parámetros en los MSI y MSRI.

Se puede conseguir una reducción adicional sin más que dividir el conjunto de síntomas relevantes en subconjuntos (bloques) que se suponen mútuamente independientes, pero los síntomas en cada bloque se consideran dependientes.

3.5.5 Conclusiones

En esta sección se han discutido cuatro modelos *ad hoc* para describir las relaciones existentes entre un conjunto de variables. El conjunto de parámetros necesario para definir la base de conocimiento depende del modelo elegido. Cada uno de estos modelos tiene sus propias ventajas e inconvenientes. Sin embargo, estos cuatro modelos sólo se aplican en situaciones particulares. En los Capítulos 6 y 7, se introducen modelos probabilísticos más generales, tales como los modelos de redes de Markov, los modelos de redes Bayesianas, los modelos especificados por listas de relaciones de independencia, y los modelos especificados condicionalmente.

Sin embargo, sea cual sea el modelo elegido, la base de conocimiento debe contener el conjunto de variables de interés y el mínimo de parámetros (probabilidades o frecuencias) necesarios para especificar la función de probabilidad conjunta de las variables.

3.6 El Motor de Inferencia

Hay dos tipos de conocimiento en los sistemas expertos probabilísticos:

1. El *conocimiento*, que está formado por el conjunto de variables y el conjunto de probabilidades asociadas necesarias para construir su función de probabilidad conjunta. Este tipo de conocimiento se almacena en la base de conocimiento.

2. Los datos, que consisten en un conjunto de valores de algunas variables (por ejemplo, síntomas) conocidas por el usuario. A esta información se la conoce con el nombre de *evidencia* y se almacena en la memoria de trabajo.

El motor de inferencia utiliza ambos, el conocimiento y los datos para responder a ciertas cuestiones hechas por el usuario. Ejemplos de tales preguntas son:

- **Cuestión 1:** Antes de que sea examinado por un doctor, ¿cuál es la enfermedad más probable para el paciente? Aquí, no hay evidencia disponible. El paciente todavía no ha sido examinado y el conjunto de

síntomas que presenta el paciente es vacío (queda por determinar). El problema consiste en calcular la probabilidad marginal (inicial) de E,

$$p(E = e_i), \quad i = 1, \ldots, m.$$

- **Cuestión 2:** Dado que el paciente presenta un subconjunto de síntomas $S_1 = s_1, \ldots, S_k = s_k$, ¿qué enfermedad tiene el paciente con mayor probabilidad? El conjunto evidencial en este caso consiste en el conjunto de valores $s_1, \ldots, s_k$. El problema consiste en calcular la función de probabilidad conjunta para cada enfermedad e_i dada la evidencia $s_1, \ldots, s_k$:

$$p(E = e_i | s_1, \ldots, s_k), \quad i = 1, \ldots, m.$$

La probabilidad marginal de E, $p(E = e_i)$, se conoce también como *probabilidad "a priori"* puesto que se calcula antes de conocer la evidencia. La probabilidad condicional de e_i dada una realización del conjunto de síntomas $p(e_i|s_1, \ldots, s_k)$ se conoce como la *probabilidad "a posteriori"* puesto que se calcula tras conocer la evidencia. Nótese que la probabilidad marginal ("a priori") puede interpretarse como un caso especial de probabilidad "a posteriori", en la que el conjunto de síntomas observado es el conjunto vacío, ϕ.

Una de las tareas del motor de inferencia en los sistemas expertos probabilísticos consiste en calcular las probabilidades condicionales de diferentes enfermedades cuando se conocen nuevos síntomas o datos. El motor de inferencia es responsable de actualizar las probabilidades condicionales:

$$p(e_i|s_1, \ldots, s_k) = \frac{p(e_i, s_1, \ldots, s_k)}{p(s_1, \ldots, s_k)}; \; i = 1, \ldots, m, \tag{3.35}$$

para todos los posibles valores de los síntomas, y de decidir cuáles tienen probabilidades condicionales altas. Normalmente se selecciona un número reducido y se muestran al usuario (por ejemplo, a médicos y pacientes) para observarlas y obtener las conclusiones pertinentes.

En (3.35), el papel del término $p(s_1, \ldots, s_k)$ consiste en actuar como una constante de normalización. Por tanto, una decisión basada en el máximo de $p(e_i|s_1, \ldots, s_k)$ coincide con la basada en el máximo de $p(e_i, s_1, \ldots, s_k)$. Por ello, los cocientes

$$R_i = \frac{p(e_i, s_1, \ldots, s_k)}{\max\limits_i p(e_i, s_1, \ldots, s_k)}; \; i = 1, \ldots, m, \tag{3.36}$$

suministran información sobre la importancia relativa de las diferentes enfermedades.

Nótese que el teorema de Bayes se utiliza para calcular con facilidad las probabilidades "a posteriori" cuando se tienen unas pocas enfermedades y síntomas. Pero cuando el número de variables (enfermedades y/o síntomas) es alto, que es lo que sucede normalmente en la práctica, se necesitan

métodos y modelos más eficientes para calcular ambas, las probabilidades "a priori" y las probabilidades "a posteriori". Estos métodos, que se conocen como métodos de *propagación de evidencia o incertidumbre*, se presentan en los Capítulos 8, 9 y 10.

3.7 Control de la Coherencia

Uno de los problemas más serios de los sistemas expertos es la presencia de incoherencias en su base de conocimiento y/o en su memoria de trabajo. Hay varias razones para ello. Por ejemplo,

1. Los expertos humanos pueden suministrar conocimiento incoherente.
2. El usuario puede suministrar datos incoherentes.
3. El motor de inferencia no actualiza los hechos (véase la Sección 2.4.2).
4. No hay un subsistema para controlar la coherencia que evite que llegue conocimiento inconsistente a la base de conocimiento y/o la memoria de trabajo.

Seguidamente se dan algunos ejemplos para ilustrar la importancia del mantenimiento de un conocimiento coherente en los sistemas expertos.

Ejemplo 3.8 Restricciones para dos variables. Supóngase que se tienen sólo dos variables binarias, E y S. Tal como se ha indicado en la sección anterior, las probabilidades necesarias para la base de conocimiento de cualquiera de los métodos anteriores son $p(e)$, $p(s)$, $p(s|e)$. Por ello, el sistema experto comienza preguntando al usuario por los valores de $p(d)$ y $p(s)$. Estos valores deben satisfacer las restricciones triviales $0 \leq p(e) \leq 1$ y $0 \leq p(s) \leq 1$. Una vez que se han definido $p(e)$ y $p(s)$, el sistema pregunta al usuario los valores de $p(s|e)$. El sistema debería informar al usuario sobre las restricciones que deben satisfacer estos valores. Por ejemplo, dando sus respectivas cotas inferior y superior. En algunos casos, algunos valores son redundantes y el sistema experto debería asignar automáticamente los valores apropiados sin preguntar al usuario. Por ejemplo,

$$p(s|E=0) + p(s|E=1) = p(s), \text{ para todo } s.$$

Por ello, se tiene

$$p(s|E=1) = p(s) - p(s|E=0). \tag{3.37}$$

Por tanto, tan pronto como se conoce $p(s)$, el sistema experto no necesita preguntar al usuario los valores de $p(s|e)$, puesto que sólo dos de ellos son necesarios: $p(S=0|E=0)$ y $p(S=1|E=0)$. Por otra parte, estas dos

probabilidades deben sumar uno. Por tanto, sólo una de estas probabilidades es suficiente para definir los parámetros correspondientes de la base de datos. ∎

Además de las relaciones entre las diferentes probabilidades que intervienen en la definición de la función de probabilidad conjunta, hay también otras condiciones que deben satisfacer las probabilidades para ser consistentes. Por tanto, el subsistema de control de la coherencia debe ser capaz de informar al usuario de las restricciones a que deben someterse las nuevas unidades de información. El ejemplo que sigue ilustra esta idea.

Ejemplo 3.9 Restricciones para dos conjuntos. Supóngase que se tienen sólo dos conjuntos A y B. Las probabilidades que intervienen en la definición de la base de conocimiento de un sistema experto probabilístico son $p(A)$, $p(B)$, $p(A \cup B)$ y $p(A \cap B)$. Estas probabilidades deben satisfacer las restricciones siguientes:

$$\begin{array}{l} 0 \leq p(A) \leq 1, \\ 0 \leq p(A) \leq 1, \\ max\{\, 0, p(A) + p(B) - 1 \,\} \leq p(A \cap B) \leq min\{\, p(A), p(B) \,\}, \\ max\{\, p(A), p(B) \,\} \leq p(A \cup B) \leq min\{\, 1, p(A) + p(B) \,\} \end{array} \tag{3.38}$$

La restricción $p(A) + p(B) - 1 \leq p(A \cap B)$ se obtiene como sigue:

$$\begin{aligned} p(A \cap B) &= p(\overline{\overline{A \cap B}}) = p(\overline{\overline{A} \cup \overline{B}}) = 1 - p(\overline{A} \cup \overline{B}) \\ &\geq 1 - (1 - p(A) + 1 - p(B)) = p(A) + p(B) - 1. \end{aligned}$$

Por ello, el sistema experto comienza preguntando al usuario los valores de $p(A)$ y $p(B)$. Estos valores deben satisfacer las dos primeras restricciones en (3.38). Una vez que $p(A)$ y $p(B)$ ya han sido especificadas y comprobadas, el subsistema de adquisición de conocimiento pregunta los valores de $p(A \cap B)$ o de $p(A \cup B)$; el sistema debe informar al usuario de las cotas inferior y superior de estas probabilidades dadas en las dos últimas restricciones de (3.38). En otro caso, podrían darse valores fuera de los intervalos de coherencia. En tal caso se violarían los axiomas de la probabilidad y el sistema podría generar conclusiones erróneas. Supóngase que $p(A \cap B)$ ha sido dada y comprobada; entonces se asignará automáticamente a $p(A \cup B)$ el valor

$$p(A \cup B) = p(A) + p(B) - p(A \cap B), \tag{3.39}$$

de acuerdo con (3.2). ∎

El lector puede imaginar la complejidad del conjunto de restricciones que resultan a medida que aumenta el número de subconjuntos. Por tanto, el riesgo de que el usuario viole las restricciones aumenta con el número de variables. En estas situaciones es importante disponer de un sistema capaz de controlar la coherencia del conocimiento (Smith (1961)).

En algunos modelos probabilísticos (por ejemplo, en los de redes Bayesianas presentados en el Capítulo 6), el control de la coherencia no es un problema, puesto que los modelos son coherentes por construcción. Sin embargo, en otros modelos probabilísticos debe controlarse la coherencia.

En algunos modelos probabilísticos el control de la coherencia es una necesidad, no un lujo. El subsistema de control de coherencia impide que el conocimiento incoherente entre en la base de conocimiento y/o la memoria de trabajo. Un método para comprobar la consistencia de un modelo probabilístico se describe en el Capítulo 7.

3.8 Comparando los dos Tipos de Sistemas Expertos

Se concluye este capítulo con una breve comparación de los sistemas expertos basados en reglas con los sistema expertos basados en probabilidad. Se discuten sus analogías y diferencias, y sus ventajas y desventajas. La Tabla 3.17 muestra un resumen de algunas componentes de cada tipo de sistema experto y de la estructura (lógica o probabilística) en la que se basa.

1. **Base de Conocimiento:**

   El conocimiento de un sistema experto basado en reglas consiste en los objetos y el conjunto de reglas. El conocimiento de un sistema experto basado en probabilidad consiste en el espacio de probabilidad, que incluye las variables, sus posibles valores, y su función de probabilidad conjunta. Por otra parte, los datos de ambos sistemas consisten en la evidencia asociada a los casos a analizar.

   La base de conocimiento en los sistemas expertos basados en reglas es fácil de implementar, puesto que sólo es necesario utilizar elementos simples, tales como objetos, conjuntos de valores, premisas, conclusiones y reglas. Sin embargo, el conocimiento que puede ser almacenado es limitado cuando se compara con el de los sistemas expertos basados en probabilidad. Un inconveniente de los sistemas expertos probabilísticos es el alto número de parámetros que manejan, lo que hace que sea difícil su especificación y definición.

2. **Motor de Inferencia:**

   En los sistemas expertos basados en reglas las conclusiones se obtienen de los hechos aplicando las diferentes estrategias de inferencia, tales como Modus Ponens, Modus Tollens y encadenamiento de reglas. Por ello, el motor de inferencia es rápido y fácil de implementar. En los sistemas expertos basados en probabilidad, el motor de inferencia es más complicado que en el caso de los sistemas expertos

basados en reglas. El motor de inferencia de un sistema experto probabilístico se basa en la evaluación de las probabilidades condicionales utilizando uno o varios métodos propuestos por los diferentes tipos de sistemas expertos probabilístico. (véanse los Capítulos 8 y 9). El grado de dificultad depende del modelo seleccionado y varía desde baja, para los modelos de independencia, a alta, para los modelos de dependencia generales.

3. **Subsistema de Explicación:**

   La explicación es fácil en el caso de los sistemas expertos basados en reglas, ya que se sabe qué reglas han sido utilizadas para concluir en cada momento. El motor de inferencia sabe qué reglas se han utilizado en el encadenamiento y han contribuído a obtener conclusiones y qué reglas se han utilizado sin éxito.

   En el caso de los sistemas expertos basados en probabilidad, la información sobre qué variables influyen en otras está codificada en la función de probabilidad conjunta. Por ello, la explicación se basa en los valores relativos de las probabilidades condicionales que miden los grados de dependencia. Una comparación de las probabilidades condicionales para diferentes conjuntos de evidencia permite analizar sus efectos en las conclusiones.

4. **Subsistema de Aprendizaje:**

   En los sistemas expertos basados en reglas, el aprendizaje consiste en incorporar nuevos objetos, nuevos conjuntos de valores factibles para los objetos, nuevas reglas o modificaciones de los objetos existentes, de los conjuntos de valores posibles, o de las reglas. En los sistemas expertos probabilísticos, el aprendizaje consiste en incorporar o modificar la estructura del espacio de probabilidad: variables, conjunto de posibles valores, o los parámetros (valores de las probabilidades).

Ejercicios

3.1 Utilizar la función de probabilidad conjunta de la Tabla 3.2 para calcular las funciones de probabilidad condicional siguientes, para todos los valores de x, y y z:

(a) $p(x|y,z)$.

(b) $p(y|x,z)$.

(c) $p(z|x,y)$.

3.2 Construir una función de probabilidad conjunta de tres variables X, Y y Z de la que pueda concluirse que X e Y son independientes, X y Z

	Basados en Reglas	Probabilísticos
Base de Conocimiento	Objetos, reglas Hechos	Variables, FPC Hechos
Motor de Inferencia	Estrategias de inferencia Encadenamiento de reglas	probabilidad condicional métodos de evaluación
Subsistema de Explicación	Basado en reglas activas	Basado en probabilidad condicional
Aprendizaje	Cambio en objetos y reglas	Cambio en modelo probabilístico

TABLA 3.17. Una comparación entre los sistemas expertos basados en reglas y los basados en probabilidad.

son dependientes, e Y y Z son dependientes. Seguidamente, utilícese dicha función de probabilidad conjunta para calcular las funciones de probabilidad siguientes, para todos los valores de x, y, y z:

(a) $p(y|x)$.

(b) $p(x|y)$.

(c) $p(x|y,z)$.

3.3 Considérese la función de probabilidad conjunta de la Tabla 3.2.

(a) Generar todas las posibles relaciones de independencia condicional que incluyan a las variables X, Y y Z: $\{I(X,Y|\phi),\ldots\}$.

(b) Comprobar cuáles de estas relaciones están implicadas por la función de probabilidad conjunta de la Tabla 3.2.

3.4 En el Ejemplo 3.4 se aplicó el teorema de Bayes para mostrar que tras observar las evidencias $V = v$ y $P = p$, la probabilidad "a posteriori" del adenocarcinoma gástrico es 0.9. Completar el problema de diagnóstico calculando las probabilidades "a posteriori" con la información adicional $D = d$. Con esta información adicional, ¿cuál es la probabilidad de que el diagnóstico sea incorrecto? ¿cómo cambia esta probabilidad cuando $D = \bar{d}$?

3.5 Utilizar los datos del Ejemplo 3.4 para calcular las probabilidades "a posteriori" del adenocarcinoma gástrico

(a) Utilizando el Teorema de Bayes.

(b) Usando la definición de probabilidad condicional y la Figura 3.1.

Considérense los casos dados por los siguientes conjuntos evidenciales:

$Enfermedad$	$SíntomasRelevantes$
E_1	S_1, S_2, S_5
E_2	S_2, S_3, S_5
E_3	S_3, S_4, S_5

TABLA 3.18. Enfermedades y sus correspondientes síntomas relevantes.

(a) $V = \bar{v}$ y $P = p$.

(b) $V = v$ y $P = \bar{p}$.

(c) $V = \bar{v}$ y $P = \bar{p}$.

(d) $V = \bar{v}$, $P = \bar{p}$ y $D = \bar{d}$.

3.6 Mostrar que las dos fórmulas en (3.30) y (3.31) para el número de parámetros en el MSRI son iguales.

3.7 Dada una población de pacientes clasificados por cinco enfermedades $E_1, \ldots, E_5$ mútuamente exclusivas y tres síntomas binarios S_1, S_2 y S_3, hacer hipótesis apropiadas para cada uno de los cuatro modelos de sistemas expertos probabilísticos dados en la Sección 3.5. Seguidamente, determinar los números de parámetros necesarios en cada modelo.

3.8 Dada una población de pacientes clasificados por tres enfermedades E_1, E_2 y E_3 mútuamente exclusivas y cinco síntomas binarios $S_1, \ldots, S_5$. Indicar qué parámetros son necesarios para especificar cada uno de los modelos siguientes:

(a) El MSI.

(b) El MSRI, dados los síntomas relevantes para cada una de las tres enfermedades que se muestran en la Tabla 3.18.

(c) El MSRD, dados los síntomas relevantes para cada una de las tres enfermedades que se muestran en la Tabla 3.18.

3.9 Considérese el problema del diagnóstico médico descrito en el Ejemplo 3.5 y supóngase que se desea construir un sistema experto probabilístico con el MSI para el problema de diagnosis. Escribir un programa de ordenador que haga lo siguiente:

(a) Leer las probabilidades "a priori" $p(e_i),\ i = 1, \ldots, m$, de un fichero de texto.

(b) Leer las verosimilitudes $p(s_j|e_i),\ i = 1, \ldots, d;\ j = 1, \ldots, n$, de un fichero de texto.

		Experto Humano				
Orden	Datos	1	2	3	4	5
1	$p(a)$	0.8	0.8	0.5	0.5	0.6
2	$p(b)$	0.7	0.7	0.6	0.6	0.5
3	$p(c)$	0.5	0.5	0.6	0.7	0.4
4	$p(a,b)$	0.6	0.2	0.3	0.4	0.3
5	$p(a,c)$	0.4	0.2	0.3	0.2	0.2
6	$p(b,c)$	0.2	0.3	0.4	0.4	0.2
7	$p(a,b,c)$	0.1	0.2	0.2	0.2	0.1

TABLA 3.19. Cinco conjuntos de probabilidades suministradas por cinco expertos humanos diferentes.

(c) Actualizar las probabilidades de las enfermedades tras conocer ciertos síntomas, usando la función de probabilidad conjunta del modelo resultante (3.26) y el teorema de Bayes (3.16).

3.10 Se ha preguntado a cinco expertos humanos diferentes el valor de las siguientes probabilidades en el orden indicado:

$$p(a),\ p(b),\ p(c),\ p(a,b),\ p(a,c),\ p(b,c),\ \text{y}\ p(a,b,c).$$

Los datos se dan en la Tabla 3.19. Usando los resultados de la Sección 3.7, determinar si la información dada por los expertos es coherente.

3.11 En el Ejemplo 3.9, se han dado las restricciones necesarias para controlar la coherencia en los casos en que se tienen dos conjuntos A y B. Ahora, considérense tres conjuntos A, B y C y sea

$$\begin{aligned}
L_1 &= \max\{\,0, p(A)+p(B)-1\,\},\\
U_1 &= \min\{\,p(A), p(B)\,\},\\
L_2 &= \max\{\,0, p(A)+p(C)-1\,\},\\
U_2 &= \min\{\,p(A), p(C)\,\},\\
L_3 &= \max\{\,0, p(A\cap B)+p(A\cap C)-p(A), p(B)+p(C)-1,\\
&\quad p(A)+p(B)+p(C)-1-p(A\cap B)-p(A\cap C)\,\},\\
U_3 &= \min\{\,p(C), p(B), p(C)-p(A\cap C)+p(A\cap B),\\
&\quad p(B)-p(A\cap B)+p(A\cap C)\,\},\\
L_4 &= \max\{\,0, p(A\cap B)+p(A\cap C)-p(A), p(A\cap B)+p(B\cap C),\\
&\quad -p(B), p(A\cap C)+p(B\cap C)-p(C)\,\},\\
U_4 &= \min\{\,p(A\cap B), p(A\cap C), p(B\cap C),\\
&\quad p(A)+p(B)+p(C)-p(A\cap B)-p(A\cap C)-p(B\cap C)-1\,\}.
\end{aligned}$$

Mostrar que en este caso se necesitan las siguientes restricciones para obtener probabilidades coherentes:

(a) $0 \leq p(A) \leq 1$, (b) $0 \leq p(B) \leq 1$,
(c) $0 \leq p(C) \leq 1$, (d) $L_1 \leq p(A \cap B) \leq U_1$,
(e) $L_2 \leq p(A \cap C) \leq U_2$, (f) $L_3 \leq p(B \cap C) \leq U_3$,
(g) $L_4 \leq p(A \cap B \cap C) \leq U_4$.

3.12 Supóngase que se quieren clasificar cuatro objetos: cometa, pájaro, avión, y hombre, basándose en las siguientes características (variables) binarias: Vuela (si el objeto vuela), Motor (si el objeto tiene), y Sangre (si el objeto tiene sangre). También pueden identificarse otros objetos mediante estas variables.

(a) Diseñar un sistema experto probabilístico para resolver este problema de clasificación.

(b) ¿Cuál de los sistemas expertos es el más eficiente en este caso?

3.13 Diseñar un sistema experto probabilístico para ayudar a los alumnos a elegir la carrera universitaria. Proceder como sigue:

(a) Seleccionar un conjunto de $m = 10$ carreras $\{X_1, \ldots, X_{10}\}$.

(b) Seleccionar un conjunto de $n = 5$ indicadores apropiados (habilidades o capacidades) $\{Y_1, \ldots Y_5\}$ que puedan ser utilizadas para seleccionar la carrera.

(c) Estimar las probabilidades "a priori" $p(x_i), i = 1, \ldots, 10$ por la proporción de estudiantes en cada una de las carreras.

(d) Especificar las verosimilitudes $p(y_j|x_i)$ para cada carrera X_i y cada indicador Y_j eligiendo valores razonables. Nótese que las probabilidades "a priori" y las verosimilitudes constituyen el conocimiento.

(e) Utilícese el teorema de Bayes y las fórmulas de este capítulo para diseñar el motor de inferencia.

3.14 Diseñar un sistema experto basado en reglas para ayudar a los alumnos a elegir la carrera universitaria. Proceder como sigue:

(a) Seleccionar un conjunto de $m = 10$ carreras $\{X_1, \ldots, X_{10}\}$.

(b) Seleccionar un conjunto de $n = 5$ indicadores apropiados (habilidades o capacidades) $\{Y_1, \ldots Y_5\}$ que puedan ser utilizadas para seleccionar la carrera.

(c) Elegir un conjunto razonable de reglas que relacionen las capacidades/habilidades y las carreras.

(d) Utilizar las estrategias de inferencia y, en particular, el encadenamiento de reglas para diseñar el motor de inferencia.

Comparar este sistema experto con el del ejercicio anterior. ¿Cuál de ellos es el más eficiente en este caso?

Capítulo 4
Algunos Conceptos sobre Grafos

4.1 Introducción

En este capítulo se presentan algunos conceptos sobre la teoría de grafos que son necesarios en el resto del libro. Como ya se ha podido observar en los capítulos anteriores, los grafos son herramientas muy útiles para definir sistemas expertos y otros modelos utilizados en el área de la inteligencia artificial. Muchos de los resultados teóricos de la teoría de grafos pueden ser utilizados para analizar diversos aspectos de estos campos. Los lectores familiarizados con los conceptos elementales de la teoría de grafos pueden saltar directamente a la Sección 5. Por otra parte, aquellos lectores que deseen profundizar en los conceptos introducidos, u obtener las demostraciones de algunos de los resultados aquí presentados, pueden consultar libros específicos de este tema como, por ejemplo, Harary (1969), Berge (1973), Bondy y Murty (1976), Golumbic (1980), Liu (1985), Ross y Wright (1988), y Biggs (1989).

Este capítulo está estructurado de la siguiente forma. En la Sección 4.2 se introducen algunos conceptos básicos y definiciones. Las Secciones 4.3 y 4.4 presentan los dos tipos básicos de grafos utilizados en este libro, los grafos *no dirigidos* y *dirigidos*, respectivamente, así como sus características principales. Un tipo especial de grafos con múltiples aplicaciones, los grafos *triangulados*, se analiza en la Sección 4.5. Por otra parte, la Sección 4.6 introduce el concepto de grafos de aglomerados (grafos de *conglomerados*, grafos de *unión*, y grafos de *familias*), que se construyen agrupando conjuntos de nodos con ciertas características comunes en un grafo dado. En

la Sección 4.7 se presentan, desde un punto de vista teórico y algorítmico, distintas formas de representación de un grafo (representación *simbólica*, *gráfica*, y *numérica*). Finalmente, en la Sección 4.8 se introducen diversos algoritmos para el análisis de la estructura topológica de un grafo.

4.2 Conceptos Básicos y Definiciones

Supóngase un conjunto de objetos $X = \{X_1, X_2, \ldots, X_n\}$ que pueden relacionarse entre sí. El conjunto X puede ser representado gráficamente por una colección de *nodos* o *vértices*, asociando un nodo a cada elemento de X. Estos nodos pueden conectarse por *aristas*, indicando las relaciones existentes entre los mismos. Una arista entre los nodos X_i y X_j se denotará mediante L_{ij}. Así mismo, el conjunto de todas las aristas se denotará por $L = \{L_{ij} \mid X_i \text{ y } X_j \text{ están conectados}\}$. Por tanto, un grafo puede definirse de forma intuitiva mediante el conjunto de nodos, X, y las relaciones entre los mismos, L. En el siguiente ejemplo se ilustra esta idea intuitiva. A continuación se introduce una definición formal.

Ejemplo 4.1 Grafos. La Figura 4.1 es un ejemplo de un grafo compuesto de seis nodos $X = \{A, B, \ldots, G\}$ y de un conjunto de seis aristas,

$$L = \{L_{AB}, L_{AC}, L_{BD}, L_{CE}, L_{DF}, L_{DG}\}.$$

Los nodos están representados por círculos y las aristas por líneas que unen los nodos correspondientes. ■

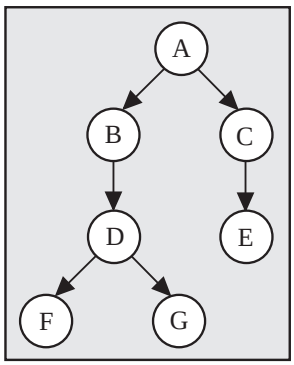


FIGURA 4.1. Ejemplo de un grafo o red.

Definición 4.1 Grafo o Red. *Un grafo es un par de conjuntos* $G = (X, L)$*, donde* $X = \{X_1, X_2, \ldots, X_n\}$ *es un conjunto finito de elementos*

(nodos), y L es un conjunto de aristas, es decir, un subconjunto de pares ordenados de elementos distintos de X. Los términos grafo y red se emplearán como sinónimos en este libro.

El concepto de grafo puede definirse de forma más general. Por ejemplo, puede permitirse que dos nodos estén conectados por más de una arista, o incluso que un nodo esté conectado consigo mismo. Sin embargo, en el campo de los sistemas expertos, los grafos se utilizan para representar un conjunto de variables proposicionales (nodos), y unas relaciones de dependencia entre ellas (aristas). Por tanto, no es necesario que dos nodos estén unidos por más de una arista, o que una arista una un nodo consigo mismo.

Las aristas de un grafo pueden ser *dirigidas* o *no dirigidas*, dependiendo de si se considera o no, el orden de los nodos. En la práctica, esta distinción dependerá de la importancia del orden en que se relacionen los objetos.

Definición 4.2 Arista dirigida. *Dado un grafo $G = (X, L)$, si $L_{ij} \in L$ y $L_{ji} \notin L$, la arista L_{ij} entre los nodos X_i y X_j se denomina dirigida y se denota mediante $X_i \rightarrow X_j$.*

Definición 4.3 Arista no dirigida. *Dado un grafo $G = (X, L)$, si $L_{ij} \in L$ y $L_{ji} \in L$, la arista L_{ij} se denomina no dirigida y se denota mediante $X_i - X_j$ o $X_j - X_i$.*

Definición 4.4 Grafo dirigido y no dirigido. *Un grafo en el cual todas las aristas son dirigidas se denomina grafo dirigido, y un grafo en el que todas sus aristas son no dirigidas se denomina no dirigido.*

Por tanto, en un grafo dirigido es importante el orden del par de nodos que definen cada arista, mientras que en un grafo no dirigido, el orden carece de importancia.

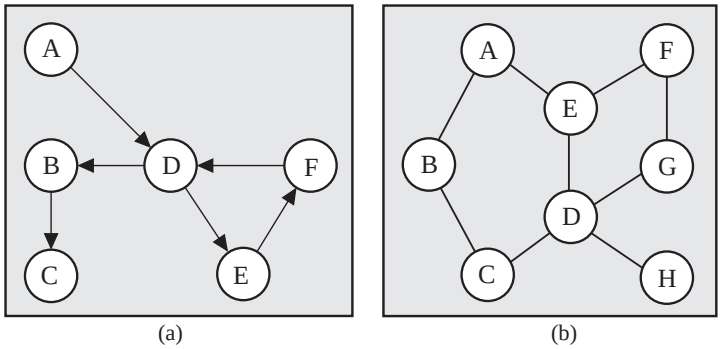


FIGURA 4.2. Ejemplos de un grafo dirigido (a), y uno no dirigido (b).

Ejemplo 4.2 Grafos dirigidos y no dirigidos. En las Figuras 4.2(a) y 4.2(b) se muestran ejemplos de un grafo dirigido y de un grafo no dirigido,

respectivamente. El grafo de la Figura 4.2(a) está definido por:

$$X = \{A, B, C, D, E, F\},$$
$$L = \{A \to D, B \to C, D \to B, F \to D, D \to E, E \to F\},$$

mientras que para el grafo de la Figura 4.2(b) se tiene

$$X = \{A, B, C, D, E, F, G, H\},$$
$$L = \{A - B, B - C, C - D, D - E, E - A, E - F, F - G, G - D, D - H\}.$$

■

Definición 4.5 Conjunto adyacente. *Dado un grafo $G = (X, L)$ y un nodo X_i, el conjunto adyacente del nodo X_i es el conjunto de nodos que son directamente alcanzables desde X_i, es decir, $Ady(X_i) = \{X_j \in X \mid L_{ij} \in L\}$.*

Esta definición proporciona una descripción alternativa de un grafo mediante un conjunto de nodos, X, y los conjuntos adyacentes de cada uno de los nodos en X; es decir, el grafo (X, L) puede ser representado de forma equivalente mediante (X, Ady), donde $X = \{X_1, \ldots, X_n\}$ es el conjunto de nodos y $Ady = \{Ady(X_1), \ldots, Ady(X_n)\}$ es la lista de conjuntos adyacentes. Como se verá más adelante, en la Sección 4.8, esta forma de representación de un grafo es muy conveniente desde un punto de vista computacional.

Ejemplo 4.3 Conjuntos adyacentes. El grafo dirigido dado en la Figura 4.2(a) tiene asociados los siguientes conjuntos de nodos adyacentes:

$$Ady(A) = \{D\}, \quad Ady(B) = \{C\}, \quad Ady(C) = \phi,$$
$$Ady(D) = \{B, E\}, \quad Ady(E) = \{F\}, \quad Ady(F) = \{D\}.$$

Por otra parte, los conjuntos adyacentes del grafo no dirigido de la Figura 4.2(b) son:

$$Ady(A) = \{B, E\}, \quad Ady(B) = \{A, C\},$$
$$Ady(C) = \{B, D\}, \quad Ady(D) = \{C, E, G, H\},$$
$$Ady(E) = \{A, D, F\}, \quad Ady(F) = \{E, G\},$$
$$Ady(G) = \{D, F\}, \quad Ady(H) = \{D\}.$$

Por tanto, los grafos mostrados en la Figura 4.2 pueden ser definidos de forma equivalente por (X, L) o por (X, Ady). ■

El conjunto adyacente de un nodo X_i contiene los nodos que son directamente alcanzables desde X_i. Por tanto, comenzando en un nodo dado y pasando de forma sucesiva a uno de sus nodos adyacentes, se puede formar un *camino* a través del grafo. Como se verá más adelante, el concepto de camino entre dos nodos juega un papel central en la teoría de grafos.

Definición 4.6 Camino entre dos nodos. *Un camino del nodo X_i al nodo X_j es un sucesión de nodos $(X_{i_1}, \ldots, X_{i_r})$, comenzando en $X_{i_1} = X_i$ y finalizando en $X_{i_r} = X_j$, de forma que existe una arista del nodo X_{i_k} al nodo $X_{i_{k+1}}$, $k = 1, \ldots, r-1$, es decir,*

$$X_{i_{k+1}} \in Ady(X_{i_k}), \;\; k = 1, \ldots, r-1.$$

La longitud del camino, $(r-1)$, se define como el número de aristas que contiene.

En el caso de grafos no dirigidos, un camino $(X_{i_1}, \ldots, X_{i_r})$ puede representarse mediante $X_{i_1} - \ldots - X_{i_r}$, indicando el carácter no dirigido de las aristas. De modo similar, otra forma de representar un camino en un grafo dirigido es mediante $X_{i_1} \rightarrow \ldots \rightarrow X_{i_r}$.

Ejemplo 4.4 Caminos. Considérese el grafo dirigido dado en la Figura 4.2(a). Existe un único camino de longitud 2 de D a F en este grafo, $D \rightarrow E \rightarrow F$. Por otra parte, existe un camino de A a B de longitud 2, $A \rightarrow D \rightarrow B$, y otro de longitud 5, $A \rightarrow D \rightarrow E \rightarrow F \rightarrow D \rightarrow B$. Obsérvese que, por el contrario, no existe ningún camino de B a A. Por otra parte, existe al menos un camino entre cada par de nodos del grafo no dirigido de la Figura 4.2(b). Por ejemplo, algunos de los caminos entre A a H son

$A - E - D - H$, de longitud 3,

$A - B - C - D - H$, de longitud 4, y

$A - E - F - G - D - H$, de longitud 5. ∎

Nótese que en un grafo dirigido han de tenerse en cuenta las direcciones de las aristas para formar un camino. Por ejemplo, en el grafo dirigido de la Figura 4.2(a) existe un camino de A a C ($A \rightarrow D \rightarrow B \rightarrow C$), pero no existe ningún camino que una los nodos en sentido inverso.

Definición 4.7 Camino cerrado. *Un camino $(X_{i_1}, \ldots, X_{i_r})$ se dice que es cerrado si el nodo inicial coincide con el final, es decir, $X_{i_1} = X_{i_r}$.*

Ejemplo 4.5 Caminos cerrados. El camino $D \rightarrow G \rightarrow F \rightarrow D$ en el grafo dirigido de la Figura 4.3(a) es un camino cerrado. El grafo no dirigido dado en la Figura 4.3(b) contiene varios caminos cerrados como, por ejemplo, el camino $A - B - C - D - E - A$. ∎

Si un camino contiene un nodo más de una vez, entonces el camino contiene un subcamino cerrado. Por ejemplo, en el grafo de la Figura 4.3(b), el camino $C - D - E - F - G - D - H$ contiene dos veces el nodo D. Por tanto, este camino ha de contener un subcamino cerrado: $D-E-F-G-D$. Eliminando este camino cerrado, se puede hallar un camino más corto entre los nodos extremos, $C - D - H$.

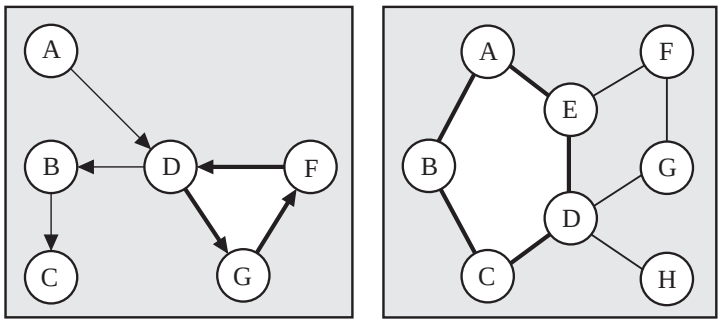


FIGURA 4.3. Ejemplos de caminos cerrados en un grafo dirigido (a) y en un grafo no dirigido (b).

4.3 Características de los Grafos no Dirigidos

En esta sección se presentan algunas características propias de los grafos no dirigidos. Un estudio similar para el caso de los grafos dirigidos se presentará en la Sección 4.4.

4.3.1 Definiciones y Conceptos Básicos

Definición 4.8 Grafo completo. *Un grafo no dirigido se denomina completo si contiene una arista entre cada par de nodos.*

Por tanto, existe un único grafo completo de n nodos. Este grafo se denota por K_n. Por ejemplo, la Figura 4.4 muestra una representación gráfica de K_5.

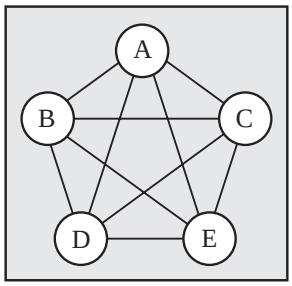


FIGURA 4.4. Grafo completo de cinco nodos.

Definición 4.9 Conjunto completo. *Un subconjunto de nodos S de un grafo G se denomina completo si existe una arista en G para cada par de nodos en S.*

Una consecuencia inmediata de esta definición es que cualquier par de nodos adyacentes en un grafo forma un conjunto completo. Por ejemplo, el grafo de la Figura 4.3(b) no contiene conjuntos completos con más de dos nodos. Por el contrario, el grafo mostrado en la Figura 4.5(a) contiene dos subconjuntos completos de tres nodos: $\{D, E, G\}$ y $\{E, F, G\}$.

Los conjuntos completos maximales de un grafo desempeñan un paper fundamental en la caracterización de su estructura topológica.

Definición 4.10 Conglomerado. *Un conjunto completo de nodos C se denomina un conglomerado si no es subconjunto propio de otro conjunto completo, es decir, si es maximal.*

Ejemplo 4.6 Conglomerados. El grafo mostrado en la Figura 4.5(a) contiene los siguientes conglomerados: $C_1 = \{A, B\}$, $C_2 = \{B, C\}$, $C_3 = \{C, D\}$, $C_4 = \{D, H\}$, $C_5 = \{D, E, G\}$, $C_6 = \{E, F, G\}$ y $C_7 = \{A, E\}$. Sin embargo, si se añade alguna arista al grafo, alguno de estos conglomerados ya no será un conjunto maximal y el conjunto de conglomerados del nuevo grafo será distinto. Por ejemplo, en el grafo de la Figura 4.5(b), obtenido añadiendo tres aristas al grafo de la Figura 4.5(a), los conjuntos C_1, C_2, C_3 y C_7 ya no son completos. El nuevo grafo contiene solamente cinco conglomerados: $C_1 = \{A, B, D, E\}$, $C_2 = \{B, C, D\}$, $C_3 = \{D, H\}$, $C_4 = \{D, E, G\}$, y $C_5 = \{E, F, G\}$. ∎

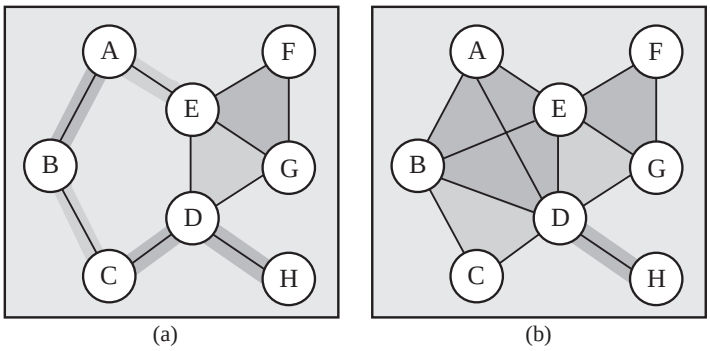


FIGURA 4.5. Ejemplo de los conglomerados asociados a dos grafos distintos.

Definición 4.11 Bucle. *Un bucle es un camino cerrado en un grafo no dirigido.*

Ejemplo 4.7 Bucle. Considérese el grafo no dirigido mostrado en la Figura 4.5(b). El camino cerrado $A - B - C - D - E - A$ es un bucle de longitud 5. Obsérvese que si en un bucle se reemplaza un camino entre dos nodos por un camino alternativo, se obtiene un nuevo bucle. Por ejemplo, si se reemplaza la arista $D-E$ por el camino $D-G-F-E$ en el bucle anterior, se obtiene un nuevo bucle de longitud 7: $A-B-C-D-G-F-E-A$. ∎

Definición 4.12 Vecinos de un nodo. *El conjunto de nodos adyacentes a un nodo X_i en un grafo no dirigido se denomina conjunto de vecinos de X_i, $Vec(X_i) = \{X_j \mid X_j \in Ady(X_i)\}$.*

Nótese que en el caso de grafos no dirigidos, el conjunto de nodos adyacentes a un nodo dado coincide con el conjunto de vecinos de dicho nodo. Por ejemplo, los nodos sombreados, $\{A, D, F\}$, en la Figura 4.6 son los vecinos del nodo E.

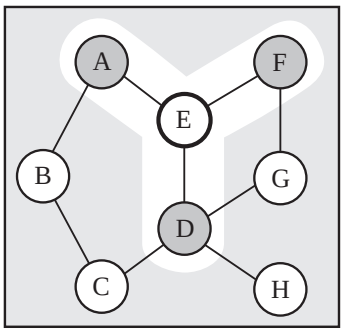


FIGURA 4.6. Conjunto de vecinos del nodo E.

Definición 4.13 Frontera de un conjunto de nodos. *La unión de los conjuntos de vecinos de los nodos de un conjunto dado, S, excluyendo los nodos de S, se denomina la frontera de S y se denota por $Frn(S)$.*

$$Frn(S) = \left(\bigcup_{X_i \in S} Vec(X_i) \right) \setminus S,$$

donde $X \setminus S$ es el conjunto de nodos de X excluyendo los de S.

Por ejemplo, los nodos sombreados en la Figura 4.7, $\{A, C, F, G, H\}$, son la frontera del conjunto $\{D, E\}$.

En el caso de que S contenga un único nodo, la frontera se reduce al conjunto de vecinos.

4.3.2 Tipos de Grafos no Dirigidos

En muchas situaciones prácticas es importante conocer si existe un camino entre un par de nodos dados. Por ejemplo, en el campo de los sistemas expertos, los grafos se utilizan para representar relaciones de dependencia entre las variables que componen el sistema. En estos casos, es muy útil conocer el número de posibles caminos entre dos nodos, a efectos de entender la estructura de dependencia contenida en el grafo. Desde este punto de vista, una clasificación útil de los grafos debe tener en cuenta el número de caminos distintos existentes entre cada par de nodos.

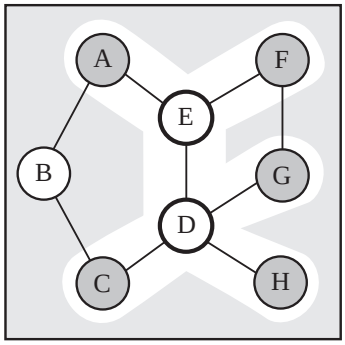


FIGURA 4.7. Frontera del conjunto $\{D, E\}$.

Definición 4.14 Grafos conexos no dirigidos. *Un grafo no dirigido se denomina conexo si existe al menos un camino entre cada par de nodos. En caso contrario, el grafo se denomina inconexo.*

Por ejemplo, el grafo de la Figura 4.7 es un grafo conexo. Sin embargo, el grafo representado en la Figura 4.8 es inconexo pues, por ejemplo, no existe ningún camino entre los nodos A y F. Obsérvese que el grafo mostrado en la Figura 4.8(a) parece conexo a primera vista, pues las aristas se cruzan ocultando este hecho. Esta característica se refleja de forma más directa en la representación gráfica de la Figura 4.8(b). El problema de la representación gráfica de un grafo se analiza en detalle en la Sección 4.7.

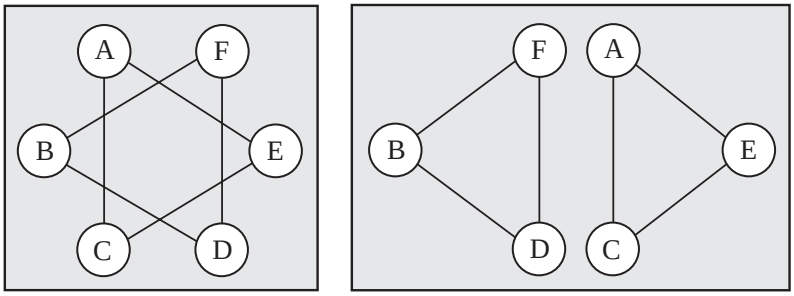


FIGURA 4.8. Dos representaciones distintas del mismo grafo inconexo.

Un grafo inconexo puede dividirse en un conjunto de grafos conexos llamados *componentes conexas*. Por ejemplo, el grafo inconexo anterior contiene las componentes conexas $\{A, C, E\}$ y $\{B, D, F\}$. Este hecho hace que, en la práctica, se suponga que los grafos son conexos pues, en caso contrario, podría argumentarse sobre cada una de las componentes conexas del grafo de forma análoga. En la Sección 4.8 se desarrollará un algoritmo

para determinar si un grafo es conexo, y calcular sus componentes conexas, caso de no serlo.

La complejidad topológica de un grafo aumenta con el número de caminos distintos entre dos nodos. Por tanto, además de considerar la existencia de un camino entre dos nodos, se ha de considerar también el número de caminos posibles.

Definición 4.15 Árbol. *Un grafo conexo no dirigido se denomina un árbol si existe un único camino entre cada par de nodos.*

De la definición anterior se deduce que un árbol es un grafo conexo, pero si se elimina una cualquiera de sus aristas, el grafo se vuelve inconexo. De forma similar, se puede deducir que un árbol no contiene bucles, pero si se añade una arista cualquiera al grafo se forma un bucle.

La Figura 4.9(a) muestra un ejemplo de un árbol. Obsérvese que la eliminación de una cualquiera de sus aristas divide al grafo en dos partes inconexas. Por otra parte, si se añade al grafo una arista cualquiera, como se indica en la Figura 4.9(b), se creará un bucle en el grafo y éste ya no será un árbol.

Definición 4.16 Grafo múltiplemente conexo. *Un grafo conexo se denomina múltiplemente conexo si contiene al menos un par de nodos que estén unidos por más de un camino o, equivalentemente, si contiene al menos un bucle.*

Nótese que si un grafo contiene dos caminos distintos entre un par de nodos, éstos pueden combinarse para formar un bucle. Por tanto, las dos definiciones anteriores son efectivamente equivalentes. Por ejemplo, el grafo de la Figura 4.9(b) es múltiplemente conexo, pues existen los caminos $D - E - G - J$ y $D - F - H - J$ que unen los nodos D y J. Estos dos caminos forman el bucle $D - E - G - J - H - F - D$.

Los distintos tipos de grafos no dirigidos introducidos en esta sección se muestran de forma esquemática en la Figura 4.10.

4.4 Características de los Grafos Dirigidos

En esta sección se describen las principales características de los grafos dirigidos.

4.4.1 Definiciones y Conceptos Básicos

Definición 4.17 Padre e hijo. *Cuando existe una arista dirigida, $X_i \rightarrow X_j$, del nodo X_i al nodo X_j, entonces se dice que el nodo X_i es un padre del nodo X_j, y que el nodo X_j es un hijo de X_i.*

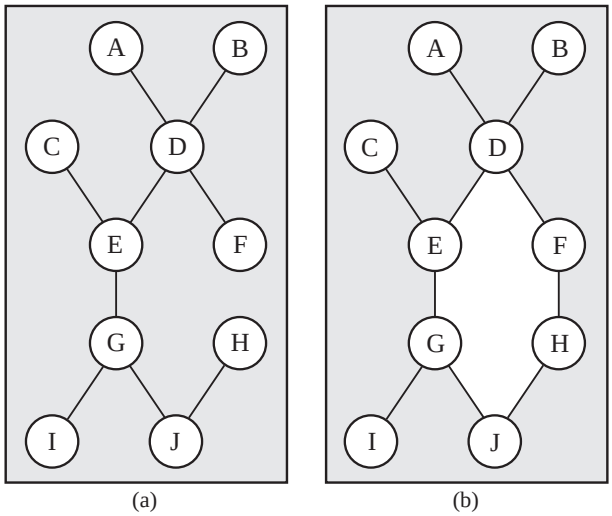


FIGURA 4.9. Ejemplo de un árbol (a) y de un grafo múltiplemente conexo (b).

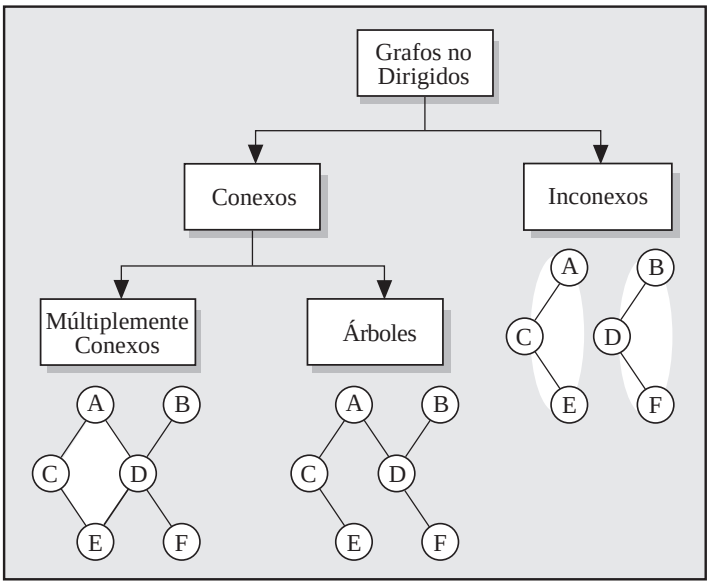


FIGURA 4.10. Tipos de grafos no dirigidos.

El conjunto de los padres de un nodo X_i se denota mediante Π_{X_i} o simplemente Π_i. Por ejemplo, los nodos C y D son los padres del nodo E en el grafo de la Figura 4.11. En un grafo dirigido, el conjunto de hijos de un nodo coincide con el conjunto de nodos adyacentes.

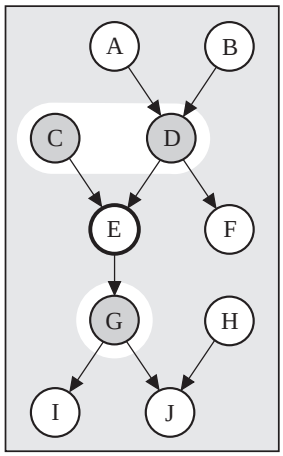


FIGURA 4.11. Padres e hijos del nodo E.

Definición 4.18 Familia de un nodo. *El conjunto formado por un nodo y sus padres se denomina la familia del nodo.*

Por ejemplo, las distintas zonas sombreadas en el grafo de la Figura 4.12 muestran las distintas familias asociadas a este grafo. En este ejemplo se pueden observar familias con uno, dos y tres nodos. Las familias de un grafo jugarán un papel muy importante en los capítulos posteriores, pues la estructura de dependencias codificada en un grafo dirigido podrá trasladarse a una función de probabilidad definiendo distribuciones de probabilidad locales sobre cada familia del grafo.

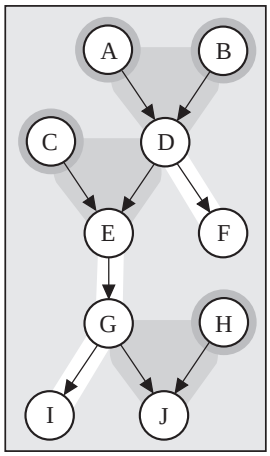


FIGURA 4.12. Familias asociadas a los nodos de un grafo.

Definición 4.19 Ascendientes de un nodo. *Un nodo X_j se denomina ascendiente del nodo X_i si existe un camino de X_j a X_i.*

Definición 4.20 Conjunto ancestral. *Un conjunto de nodos S se denomina un conjunto ancestral si contiene los ascendientes de todos sus nodos.*

Definición 4.21 Descendientes de un nodo. *Un nodo X_j se denomina descendiente del nodo X_i si existe un camino de X_i a X_j.*

La Figura 4.13 muestra los conjuntos de ascendientes y descendientes del nodo E.

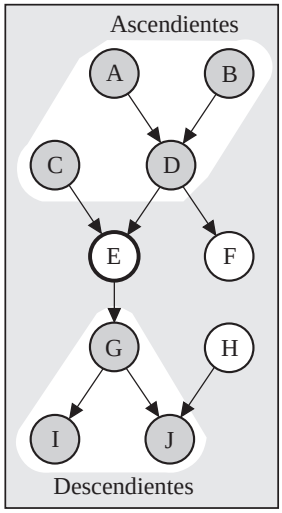


FIGURA 4.13. Ascendientes y descendientes del nodo E.

Hasta ahora se han analizado distintos atributos de los nodos de un grafo referidos a su relación de dependencia con el resto de los nodos (padres, hijos, familia, etc.). En ocasiones esta estructura de dependencia, u otras propiedades topológicas del grafo, pueden plasmarse de forma global en una ordenación de los nodos $X = \{X_1, \ldots, X_n\}$.

Definición 4.22 Ordenación. *Dado un conjunto $X = \{X_1, \ldots, X_n\}$ de nodos, una ordenación, α, es una biyección que asigna un número del conjunto $\{1, \ldots, n\}$ a cada nodo:*

$$\alpha : \{1, \ldots, n\} \longrightarrow \{X_1, \ldots, X_n\}.$$

Por tanto, $\alpha(i)$ denota el i-ésimo nodo de la numeración. Una numeración puede representarse mediante la sucesión ordenada de nodos $(\alpha(1), \ldots, \alpha(n))$.

Una numeración de los nodos que muestra la estructura de ascendientes-descendientes de forma global es la *numeración ancestral*.

Definición 4.23 Numeración ancestral. *Una numeración de los nodos de un grafo dirigido se denomina ancestral si el número correspondiente a cada nodo es menor que los correspondientes a sus hijos.*

Por ejemplo, las dos numeraciones mostradas en las Figuras 4.14 son dos numeraciones ancestrales distintas del mismo grafo. Por tanto, este tipo de numeración no es necesariamente única. Por otra parte, existen grafos dirigidos que no admiten ninguna numeración ancestral. Este problema se analiza en detalle, desde un punto de vista teórico y algorítmico, en la Sección 4.7.1.

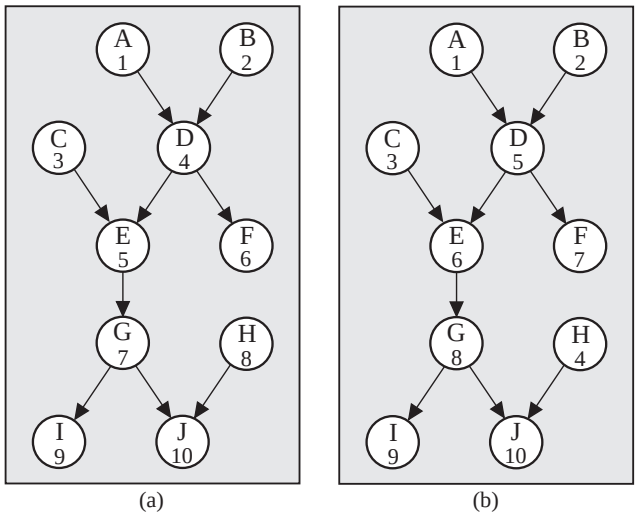


FIGURA 4.14. Dos numeraciones ancestrales del mismo grafo.

Un grafo dirigido puede convertirse de forma sencilla a un grafo no dirigido, sin más que eliminar la direccionalidad de sus aristas.[1]

Definición 4.24 Grafo no dirigido asociado a un grafo dirigido. *Dado un grafo dirigido, el grafo no dirigido obtenido al reemplazar cada arista dirigida del grafo por la correspondiente arista no dirigida se denomina el grafo no dirigido asociado.*

[1]Obsérvese que el problema inverso es más complejo pues existen dos alternativas para orientar una arista $X_i - X_j$: $X_i \rightarrow X_j$ o $X_j \rightarrow X_i$. Por tanto, se pueden definir varios grafos dirigidos asociados a un mismo grafo no dirigido (una discusión más detallada de este problema se presenta en Ross y Wright (1988)).

Por ejemplo, el grafo de la Figura 4.15(b) es el grafo no dirigido asociado al grafo dirigido de la Figura 4.15(a).

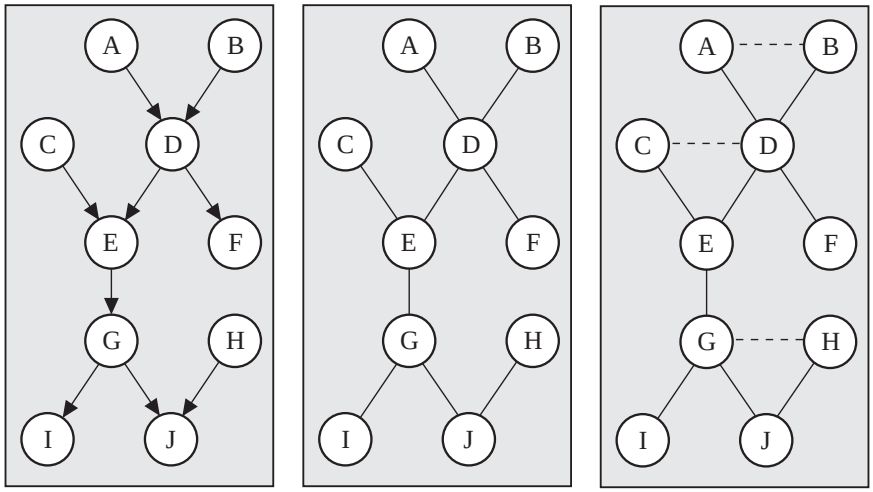


FIGURA 4.15. Ejemplo de un grafo dirigido (a), el grafo no dirigido asociado (b), y el grafo moralizado (c).

Definición 4.25 Grafo moral. *El grafo no dirigido asociado al grafo dirigido que se obtiene al añadir una arista entre cada par de nodos con algún hijo común en un grafo no dirigido, se denomina el grafo moral asociado a dicho grafo.*

Por ejemplo, el la Figura 4.15(c) muestra el grafo moral correspondiente al grafo de la Figura 4.15(a). Cada par de nodos (A, B), (C, D) y (G, H) tienen un hijo común en este grafo. Por tanto, el grafo moral asociado se forma añadiendo las tres aristas indicadas con línea discontinua y eliminando la direccionalidad de todas las aristas.

Los caminos cerrados reciben dos nombres distintos en un grafo dirididdo, según se tenga en cuenta o no la direccionalidad de las aristas. Cuando un camino cerrado está definido en el grafo dirigido original se denomina un *ciclo*; en cambio, cuando se define sobre el grafo no dirigido asociado, se denomina bucle (ver Sección 4.3).

Definición 4.26 Ciclo. *Un ciclo es un camino cerrado en un grafo dirigido.*

Ejemplo 4.8 Bucles y ciclos. La Figura 4.16(a) muestra un grafo dirigido que contiene un sólo ciclo: $D \rightarrow G \rightarrow F \rightarrow D$. Sin embargo, el grafo no dirigido asociado contiene dos bucles: $D-G-F-D$ y $A-B-D-A$. ∎

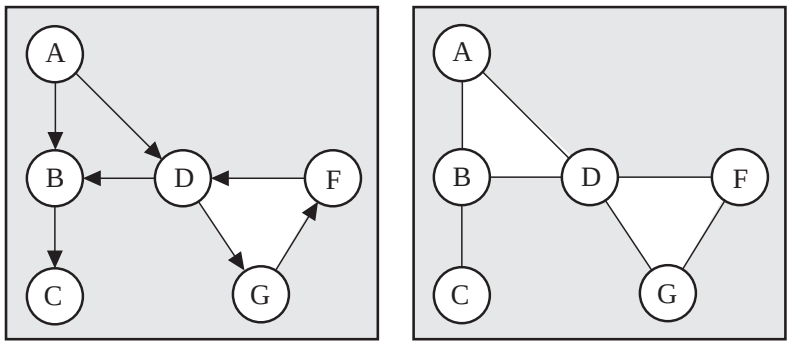


FIGURA 4.16. Bucles y ciclos de un grafo dirigido.

4.4.2 Tipos de Grafos Dirigidos

Definición 4.27 Grafos dirigidos conexos. *Un grafo dirigido se denomina conexo si el grafo no dirigido asociado es conexo; en caso contrario se denomina inconexo.*

Definición 4.28 Árboles y grafos múltiplemente conexos. *Un grafo dirigido conexo se denomina árbol si el grafo no dirigido asociado es un árbol; en caso contrario se denomina múltiplemente conexo.*

Definición 4.29 Grafos cíclicos y acíclicos. *Un grafo dirigido se denomina cíclico si contiene al menos un ciclo; en caso contrario se denomina grafo dirigido acíclico.*

Los grafos dirigidos acíclicos jugarán un papel muy importante en capítulos posteriores, pues serán la base para construir los modelos probabilísticos conocidos como *Redes Bayesianas*.

Dentro de los grafos dirigidos, los árboles suelen clasificarse en dos tipos, dependiendo del número de aristas que convergen en un mismo nodo.

Definición 4.30 Grafos simples y poliárboles. *Un árbol dirigido se denomina un árbol simple si cada nodo tiene como máximo un padre; en caso contrario se denomina un poliárbol.*

La Figura 4.17 muestra un ejemplo de un árbol simple y un ejemplo de un poliárbol. La Figura 4.18 muestra un grafo cíclico y uno múltiplemente conexo. La Figura 4.19 muestra de modo esquemático estos tipos de grafos dirigidos.

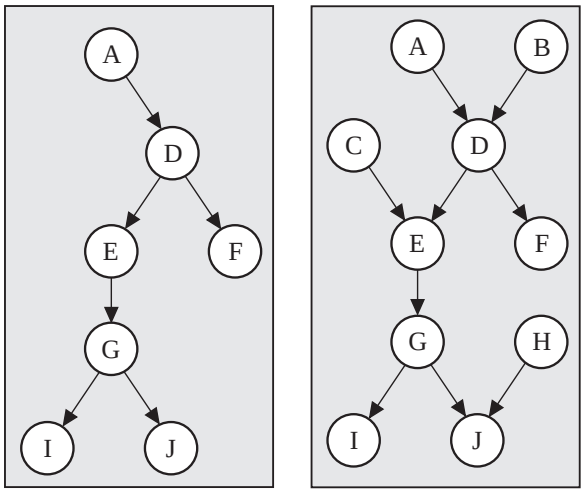


FIGURA 4.17. Ejemplos de grafos dirigidos: árbol simple (a) y poliárbol (b).

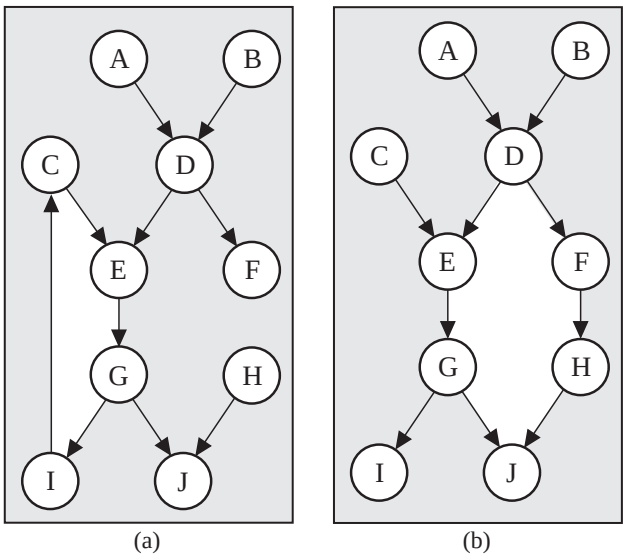


FIGURA 4.18. Ejemplos de grafos dirigidos: grafo cíclico (a) y múltiplemente conexo (b).

4.5 Grafos Triangulados

Los grafos triangulados son un tipo especial de grafos no dirigidos que tienen muchas aplicaciones prácticas interesantes en varios campos. Por ejemplo, en el Capítulo 6 se verá que este tipo de grafos constituyen la es-

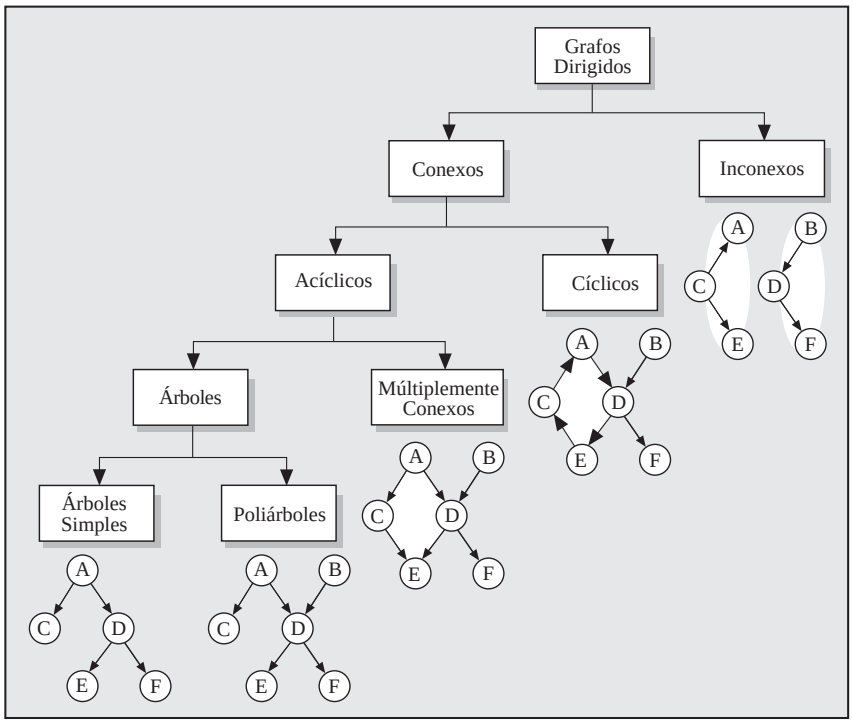


FIGURA 4.19. Tipos de grafos dirigidos.

tructura gráfica del tipo de modelos probabilísticos conocidos como *modelos descomponibles* (Lauritzen, Speed y Vijayan (1984)). Los grafos triangulados también reciben el nombre de *circuitos rígidos* (Dirac (1961)) y *grafos cordales* (Gavril (1972, 1974)).

Esta sección introduce los grafos triangulados, así como una serie de algoritmos para comprobar si un grafo es triangulado y cómo triangularlo en caso de que no lo sea.

Definición 4.31 Cuerda de un bucle. *Una cuerda es una arista que une dos nodos de un bucle y que no pertenece al bucle.*

Por ejemplo, en el grafo de la Figura 4.20, la arista $E - G$ es una cuerda del bucle $E - F - G - D - E$. Obsérvese que la cuerda divide el bucle en dos bucles menores: $E - F - G - E$ y $E - G - D - E$. Por otra parte, el bucle $A - B - C - D - E - A$ no contiene ninguna cuerda.

Dada su estructura, los bucles de longitud 3 son los únicos que no pueden poseer cuerdas. Por ello, estos son los menores elementos en los que puede descomponerse un bucle mediante la incorporación de cuerdas en el grafo. Los bucles de longitud 3 se denominan *triángulos*.

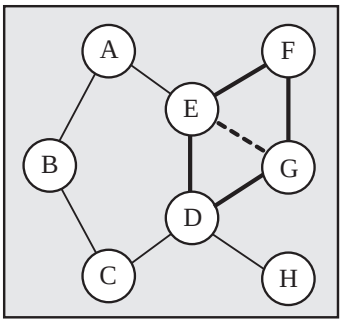


FIGURA 4.20. Ejemplo de un bucle con una cuerda.

Definición 4.32 Grafo triangulado. *Un grafo no dirigido se denomina triangulado, o cordal, si cada bucle de longitud mayor o igual que cuatro contiene al menos una cuerda.*

Ejemplo 4.9 Grafo triangulado. La Figura 4.21(a) muestra un grafo triangulado. El grafo contiene dos bucles de longitud cuatro, $A-B-E-C-A$ y $B-C-E-D-B$, y un bucle de longitud cinco, $A-B-D-E-C-A$, y cada uno de ellos tiene al menos una cuerda.

Por otra parte, el grafo de la Figura 4.21(b) no es triangulado, pues contiene al bucle $A-B-C-D-E-A$, que no posee ninguna cuerda. ■

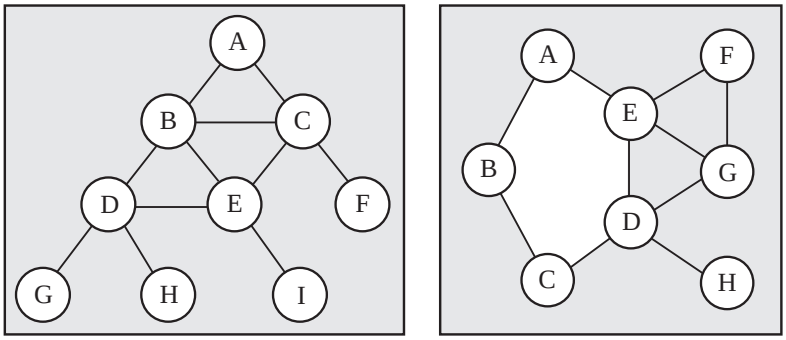


FIGURA 4.21. Ejemplo de grafo triangulado (a) y no triangulado (b).

Si un grafo no es triangulado, es posible convertirlo en triangulado añadiendo cuerdas que dividan los bucles. Este proceso se denomina *rellenado* o *triangulación*. Es importante destacar que triangular un grafo no consiste en dividirlo en triángulos. Por ejemplo, el grafo de la Figura 4.21(a) es triangulado y, por tanto, no necesita la adición de aristas extra, como aquellas que se indican mediante líneas de puntos en la Figura 4.22.

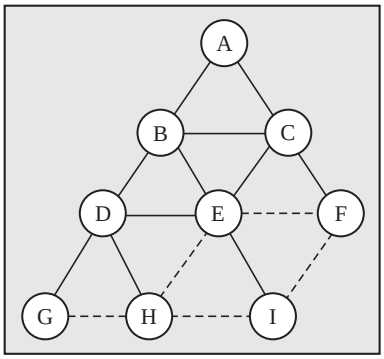


FIGURA 4.22. Triangular no significa dividir en triángulos.

Puesto que un bucle puede romperse de varias formas distintas con una cuerda, existen varias formas distintas de triangular un grafo. Por ejemplo, los dos grafos mostrados en la Figura 4.23 corresponden a dos triangulaciones distintas asociadas con el grafo de la Figura 4.21(b).

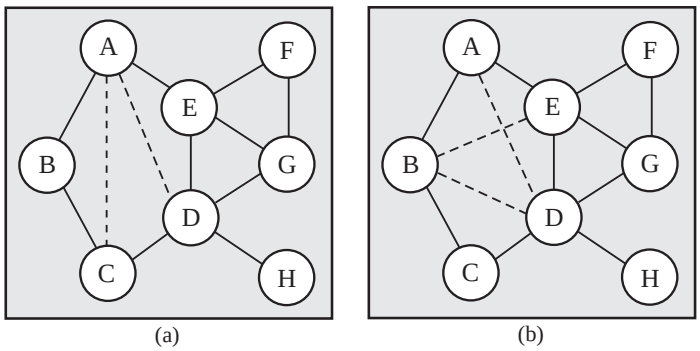


FIGURA 4.23. Dos triangulaciones distintas del mismo grafo. Las líneas de puntos representan las cuerdas añadidas.

Con objeto de preservar lo máximo posible la topología original del grafo en el proceso de triangulación, es importante añadir el mínimo número de aristas posible. En este sentido, una triangulación se dice *minimal* si contiene un número mínimo de cuerdas por debajo del cual no es posible triangular el grafo original. Nótese que la triangulación de la Figura 4.23(a) es minimal. En cambio la triangulación mostrada en la Figura 4.23(b) no es minimal, pues puede eliminarse la arista $A - D$ o la $B - E$ y el grafo resultante sigue siendo triangulado. El problema de calcular una triangulación minimal es NP-complejo[2] (Yannakakis (1981)). Dada la complejidad

[2]Una introducción a la complejidad de algoritmos y problemas NP-complejos puede consultarse en Garey y Johnson (1979).

de este problema, han sido desarrollados varios algoritmos de ejecución en tiempo lineal para triangular un grafo (Rose, Tarjan y Leuker (1976), Tarjan y Yannakakis (1984)); sin embargo, ninguno de ellos garantiza que la triangulación resultante sea minimal. A continuación se introduce un algoritmo simple llamado *algoritmo de búsqueda de máxima cardinalidad* (ver Tarjan y Yannakakis (1984)). Antes son necesarias algunas definiciones.

Definición 4.33 Numeración perfecta. *Una numeración de los nodos de un grafo, α, se denomina perfecta si el subconjunto de nodos*

$$Frn(\alpha(i)) \cap \{\alpha(1), \ldots, \alpha(i-1)\}$$

es completo para $i = 2, \ldots, n$.

Ejemplo 4.10 Numeración perfecta. La Figura 4.24(a) muestra una numeración de los nodos del grafo: $\alpha(1) = A$, $\alpha(2) = B$, $\alpha(3) = C$, $\alpha(4) = E$, etc. A continuación se comprueba que se verifican las condiciones de numeración perfecta:

- Para $i = 2$, $Frn(\alpha(2)) \cap \{\alpha(1)\} = Frn(B) \cap \{A\} = \{A, C, D, E\} \cap \{A\} = \{A\}$, que es trivialmente un conjunto completo.
- Para $i = 3$, $Frn(\alpha(3)) \cap \{\alpha(1), \alpha(2)\} = \{A, B, E, F\} \cap \{A, B\} = \{A, B\}$ es completo, pues la arista $A - B$ está contenida en el grafo.
- Para $i = 4$, $Frn(\alpha(4)) \cap \{\alpha(1), \alpha(2), \alpha(3)\} = \{B, C, D, I\} \cap \{A, B, C\} = \{B, C\}$ también es completo.

De forma análoga se puede comprobar que la condición también se cumple para $i = 5, \ldots, 9$. Por tanto, α es una numeración perfecta. ∎

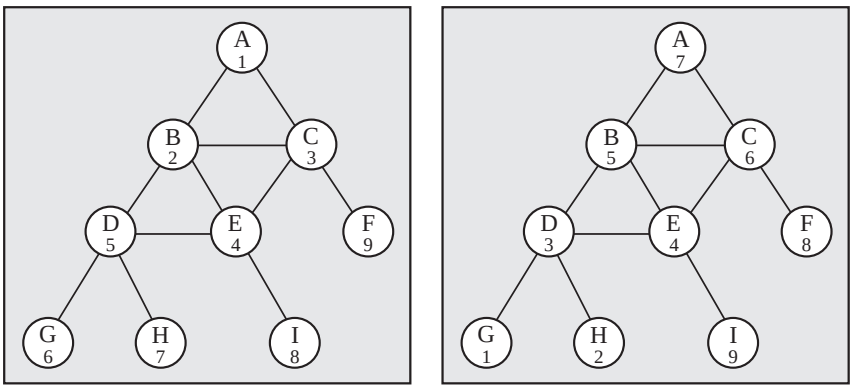


FIGURA 4.24. Dos numeraciones perfectas de los nodos.

Nótese que la numeración perfecta no es necesariamente única. Por ejemplo, la Figura 4.24(b) muestra otra numeración perfecta para el mismo grafo. Por otra parte, también existen grafos que no admiten ninguna numeración perfecta. Por ejemplo, el grafo de la Figura 4.21(b) no admite numeración perfecta; la presencia de bucles sin cuerdas hace imposible la numeración perfecta de los nodos.

Tarjan y Yannakakis (1984) desarrollaron un algoritmo, rápido y conceptualmente sencillo, para comprobar si un grafo no dirigido es triangulado. Este algoritmo, que se conoce como *algoritmo de búsqueda de cardinalidad máxima*, (en inglés, *maximum cardinality search*), se basa en la búsqueda de una numeración perfecta de los nodos del grafo. Este algoritmo está basado en el siguiente teorema que relaciona los conceptos de numeración perfecta y grafo triangulado (ver Fulkerson y Gross (1965), y Golumbic (1980)).

Teorema 4.1 Triangulación y numeración perfecta. *Un grafo no dirigido admite una numeración perfecta si y sólo si es triangulado.*

El algoritmo de máxima cardinalidad genera una numeración de los nodos del grafo que será perfecta sólo en caso de que el grafo esté triangulado.

Algoritmo 4.1 Búsqueda de máxima cardinalidad.

- **Datos:** Un grafo no dirigido $G = (X, L)$ y un nodo inicial X_i.
- **Resultado:** Una numeración α de los nodos de X.

1. *Iniciación:* Asignar el primer número al nodo inicial, es decir, $\alpha(1) = X_i$.
2. Repetir la etapa siguiente con $i = 2, \ldots, n$.
3. *Iteración i:* En la i-ésima etapa de iteración, se asigna el número i a un nodo que no haya sido numerado previamente y que tenga el máximo número de vecinos numerados. Los empates se resuelven de forma arbitraria. ■

La Figura 4.25 muestra el pseudocódigo para el algoritmo de máxima cardinalidad. El siguiente teorema permite reconocer si un grafo es triangulado utilizando el algoritmo de máxima cardinalidad (ver Tarjan (1983) y Tarjan y Yannakakis (1984)).

Teorema 4.2 Numeración de máxima cardinalidad. *Cualquier numeración de los nodos de un grafo triangulado obtenida aplicando el algoritmo de máxima cardinalidad es una numeración perfecta.*

Por tanto, cuando la numeración generada por el Algoritmo 4.1 no sea perfecta, significará que el grafo no será triangulado. De esta forma, se puede

Búsqueda de Máxima Cardinalidad

Datos: Un grafo $G = (X, L)$ y un nodo inicial X_i
Resultado: Una numeración α de los nodos en X

Etapa Inicial:

$\alpha(1) \leftarrow X_i$
$Numerados \leftarrow \{X_i\}$

Etapa Iterativa:

for i = 2 **to** n
    $X_k \leftarrow$ elige un nodo X_k en $X \setminus Numerados$
        con máximo $|Vec(X_k) \cap Numerados|$
    $\alpha(i) \leftarrow X_k$
    añade X_k a $Numerados$

FIGURA 4.25. Pseudocódigo para el algoritmo de máxima cardinalidad.

modificar fácilmente el Algoritmo 4.1 para comprobar si un grafo es triangulado. Cuando el grafo no sea triangulado, entonces el propio algoritmo añade las aristas necesarias para triangularlo. Los lectores interesados en los aspectos computacionales de este algoritmo pueden consultar las referencias Tarjan y Yannakakis (1984) o Neapolitan (1990). Una eficiente implementación de este algoritmo se ejecutará en tiempo lineal en el tamaño de la red, es decir $o(n+l)$, donde n es el número de nodos y l es el número de aristas del grafo. Con el fin de ilustrar el funcionamiento del algoritmo, pero sin cuidar la eficiencia de su implementación, se introduce el siguiente algoritmo:

Algoritmo 4.2 Triangulación de máxima cardinalidad.

- **Datos:** Un grafo no dirigido $G = (X, L)$ y un nodo inicial X_i.
- **Resultado:** Un conjunto de nuevas aristas L', tal que, $G' = (X, L \cup L')$ sea triangulado.

Etapa de Iniciación:

1. Inicialmente la nueva lista de aristas es vacía, $L' = \phi$.
2. Sea $i = 1$ y asígnese el primer número de la numeración al nodo inicial X_i, es decir, $\alpha(1) = X_i$.

Etapa de Iteración:

3. Se asigna el número i a un nodo X_k no numerado con máximo número de vecinos numerados, $\alpha(i) = X_k$.

4. Si $Vec(X_k) \cap \{\alpha(1), \ldots, \alpha(i-1)\}$ no es un conjunto completo, añadir a L' las aristas necesarias para completar el conjunto y volver a la Etapa 2; en caso contrario, ir a la Etapa 5.

5. Si $i = n$, el algoritmo finaliza; en caso contrario, asginar $i = i + 1$ e ir a la Etapa 3. ■

Utilizando el Teorema 4.2 puede demostrarse que cuando un grafo es triangulado, el conjunto de nuevas aristas L' necesarias para triangularlo obtenidas con el Algoritmo 4.2 es vacío; en caso contrario, el conjunto L' contiene las aristas necesarias para triangular el grafo.

Ejemplo 4.11 Triangulación de máxima cardinalidad. El grafo de la Figura 4.26 no es un grafo triangulado. El Algoritmo 4.2 permite construir una triangulación del grafo. Por ejemplo, eligiendo el nodo C como el nodo inicial para el algoritmo se tiene:

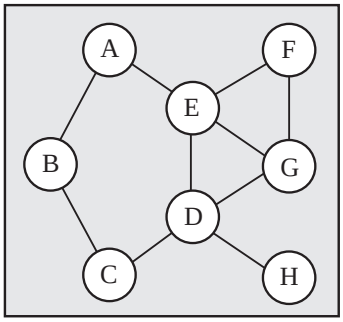


FIGURA 4.26. Grafo no dirigido y no triangulado.

- Etapa 1: $L' = \phi$.
- Etapa 2: Sean $i = 1$ y $\alpha(1) = C$.
- Etapa 3: Los nodos B y D son los únicos que tienen un vecino numerado. Deshaciendo el empate de forma arbitraria, se elige el nodo D y se numera con el número 2, es decir, $\alpha(2) = D$.
- Etapa 4: Nótese que, en este caso, los vecinos previamente numerados forman un conjunto completo. Por tanto, no es necesario añadir ninguna arista a L' y el algoritmo continúa.
- Etapa 5: Puesto que $i \neq n$, se incrementa en una unidad el contador i y se va a la Etapa 3.
- Etapas $3 - 5$: Siguiendo un proceso similar, los nodos B y E se numeran como 3 y 4, respectivamente.

- Etapas $3-4$: Los nodos con número máximo de vecinos numerados son A y G. El empate se deshace eligiendo A. Sin embargo, como puede verse en la Figura 4.27(a), el conjunto de vecinos numerados de A, $\{B, E\}$, no es un conjunto completo. Por tanto, ha de añadirse la arista $B-E$ (ver Figura 4.27(b)) a L' y comenzar de nuevo con la Etapa 2. Nótese que, ahora, $L' = \{B-E\}$.

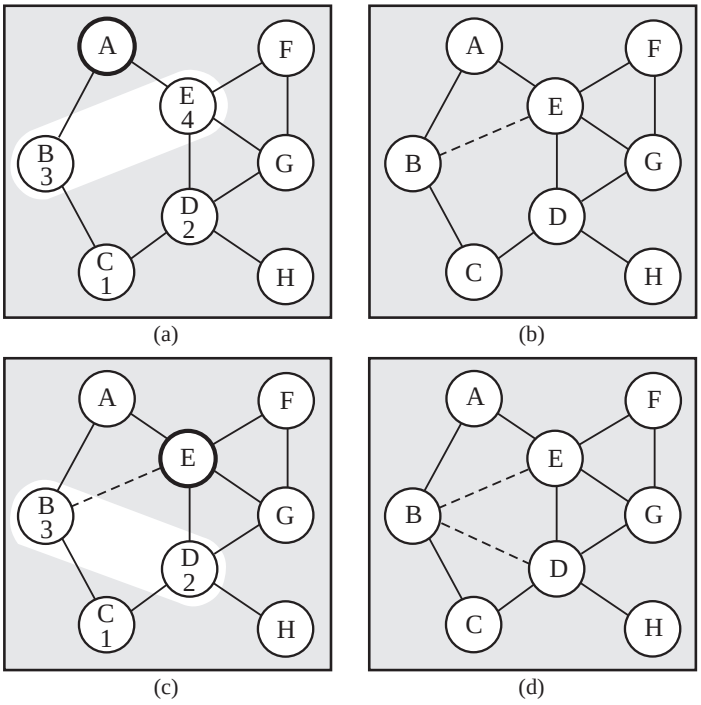


FIGURA 4.27. Numeración perfecta de los nodos utilizando el algoritmo de máxima cardinalidad.

- Etapas $2-5$: Los nodos C, D y B se numeran 1, 2 y 3, respectivamente.

- Etapas $3-4$: El nodo E posee el máximo número de vecinos numerados, $\{B, D\}$, pero este conjunto no es completo (ver Figura 4.27(c)). Por tanto, se añade la arista $B-D$ a L' y se comienza de nuevo con la Etapa 2. Ahora, $L' = \{B-E, B-D\}$ (ver Figura 4.27(d)).

- Etapas $2-5$: Los nodos C, D, B, E, A, G, F y H se numeran sucesivamente de 1 a 8. El grafo resultante $G' = (X, L \cup L')$ es un grafo triangulado y la numeración final mostrada en la Figura 4.28 es una numeración perfecta. ■

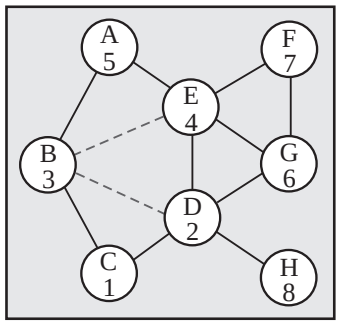


FIGURA 4.28. Numeración perfecta de los nodos utilizando el algoritmo de máxima cardinalidad.

Nótese que, dependiendo de la elección del nodo inicial y de cómo se deshacen los empates, es posible obtener varias triangulaciones del mismo grafo. Por ejemplo, el algoritmo de máxima cardinalidad puede producir las dos numeraciones perfectas mostradas en la Figura 4.29.

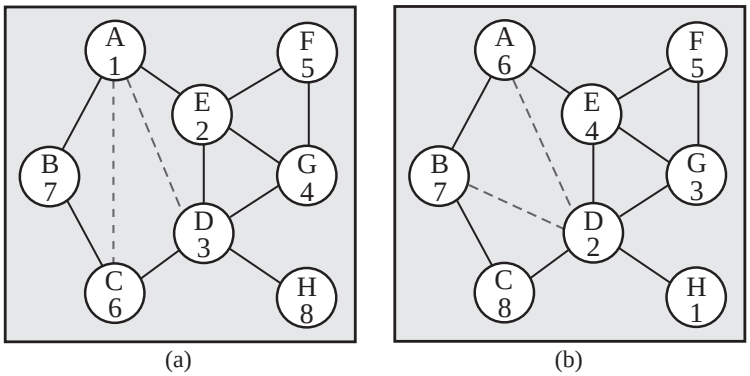


FIGURA 4.29. Dos numeraciones perfectas distintas del grafo de la Figura 4.26.

Una propiedad interesante de los grafos triangulados, que resulta especialmente útil cuando se trabaja con las denominadas redes de Markov (Capítulos 6 y 8), se conoce como la *propiedad de intersección dinámica* (en inglés, *running intersection property*).

Definición 4.34 Propiedad de intersección dinámica. *Una numeración de los conglomerados de un grafo no dirigido* $(C_1, \ldots, C_m)$ *se dice que satisface la propiedad de intersección dinámica, si el conjunto* $C_i \cap (C_1 \cup \ldots \cup C_{i-1})$ *está contenido en, al menos, uno de los conglomerados* $\{C_1, \ldots, C_{i-1}\}$*, para todo* $i = 1, \ldots, m$*.*

Esta propiedad establece que los conglomerados de un grafo pueden ser ordenados de tal forma que el conjunto de los nodos comunes a un con-

glomerado dado y a todos los conglomerados anteriores esté contenido en alguno de los conglomerados anteriores. Una sucesión de conglomerados que satisface la propiedad de intersección dinámica se denomina una *cadena de conglomerados*. Se puede dar el caso de grafos no dirigidos que no poseen ninguna cadena de conglomerados y de grafos que poseen más de una. El siguiente teorema caracteriza los grafos que poseen, al menos, una cadena de conglomerados.

Teorema 4.3 Cadena de conglomerados. *Una grafo no dirigido tiene asociada una cadena de conglomerados si y sólo si es triangulado.*

A continuación se introduce un algoritmo para construir una cadena de conglomerados a partir de un grafo no dirigido. Este algoritmo está basado en el algoritmo de máxima cardinalidad y supone que el grafo es triangulado. En caso contrario, el grafo puede ser previamente triangulado utilizando el Algoritmo 4.2.

Algoritmo 4.3 Generación de una cadena de conglomerados.

- **Datos:** Un grafo triangulado no dirigido $G = (X, L)$.
- **Resultado:** Una cadena de conglomerados $(C_1, \ldots, C_m)$ asociada a G.

1. *Iniciación:* Elegir cualquier nodo como nodo inicial y utilizar el Algoritmo 4.1 para obtener una numeración perfecta de los nodos, $X_1, \ldots, X_n$.
2. Determinar los conglomerados del grafo, C.
3. Asignar a cada conglomerado el máximo de los números (correspondientes a la numeración perfecta) de sus nodos.
4. Ordenar los conglomerados, $(C_1, \ldots, C_m)$, en orden ascendente de acuerdo a los números asignados (deshacer empates de forma arbitraria). ∎

Ejemplo 4.12 Generación de una cadena de conglomerados. En este ejemplo se aplica el Algoritmo 4.3 para generar una cadena de conglomerados asociada al grafo triangulado dado en la Figura 4.30(a). En primer lugar se utiliza el Algoritmo 4.1 para obtener una numeración perfecta de los nodos. La Figura 4.30(b) muestra los números obtenidos tomando el nodo A como nodo inicial. Los conglomerados del grafo son: $C_1 = \{A, B, C\}$, $C_2 = \{B, C, E\}$, $C_3 = \{B, D, E\}$, $C_4 = \{C, F\}$, $C_5 = \{D, G\}$, $C_6 = \{D, H\}$ y $C_7 = \{E, I\}$. A continuación, se asigna a cada conglomerado el mayor de los números que contenga. Por ejemplo, para el caso del conglomerado C_1, el mayor número perfecto asociado a los nodos A, B y C es tres, que corresponde al nodo C. Por tanto, se asigna el

número 3 al conglomerado C_1. El número correspondiente al conglomerado C_2 es 4 (que corresponde al nodo E), y así sucesivamente. Obsérvese que los conglomerados ya se encuentran ordenados de forma ascendente en la ordenación natural. El conglomerado C_1 es el que tiene el número perfecto más bajo, después el C_2, y así sucesivamente. Por tanto, $(C_1, \ldots, C_7)$ es una cadena de conglomerados para el grafo de la Figura 4.30(a). ∎

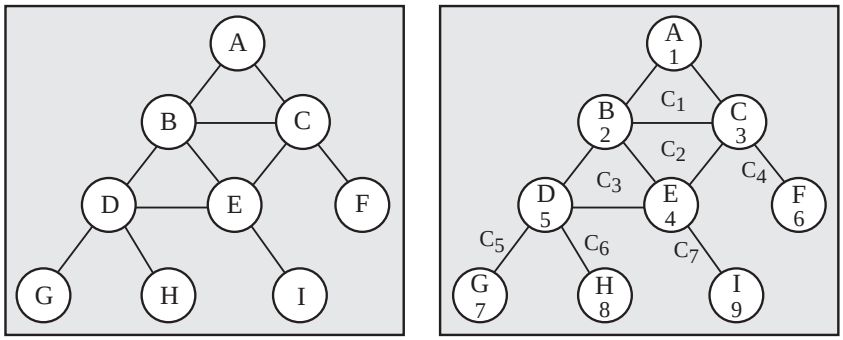


FIGURA 4.30. Un grafo triangulado (a) y una numeración perfecta de sus nodos necesaria para construir una cadena de conglomerados (b).

4.6 Grafos de Aglomerados

Los grafos de aglomerados se forman agrupando nodos con ciertas características comunes de un grafo dado. Este proceso permite obtener nuevos grafos con estructuras topológicas más simples que retienen ciertas propiedades del grafo original. En los Capítulos 6 y 8 se analizarán varias aplicaciones de los grafos de aglomerados.

Definición 4.35 Aglomerado. *Un conjunto de nodos de un grafo se denomina un aglomerado.*

Definición 4.36 Grafo de aglomerados de un grafo dado. *Supongamos un grafo $G = (X, L)$ y un conjunto de aglomerados de X, $C = \{C_1, \ldots, C_m\}$, tal que $X = C_1 \cup \ldots \cup C_m$. El grafo $G' = (C, L')$ se denomina un grafo de aglomerados de G si las aristas contenidas en L' sólo unen aglomerados que contengan algún nodo común, es decir, $(C_i, C_j) \in L' \Rightarrow C_i \cap C_j \neq \phi$.*

Un análisis detallado de las propiedades de los grafos de aglomerados se presenta en los libros Beeri y otros (1983) y Jensen (1988) (y las referencias incluidas en ellos).

Los aglomerados de un grafo no son en general conjuntos arbitrarios, pues se desea preservar al máximo posible la estructura topológica del grafo original. En este capítulo se considerarán tipos especiales de grafos de aglomerados que satisfagan ciertas propiedades deseables.

Definición 4.37 Grafo de conglomerados. *Un grafo de aglomerados se denomina un grafo de conglomerados asociado a un grafo no dirigido G si sus aglomerados son los conglomerados de G.*

Por ejemplo, el grafo de aglomerados mostrado en la Figura 4.31 es un grafo de conglomerados asociado al grafo de la Figura 4.30(a).

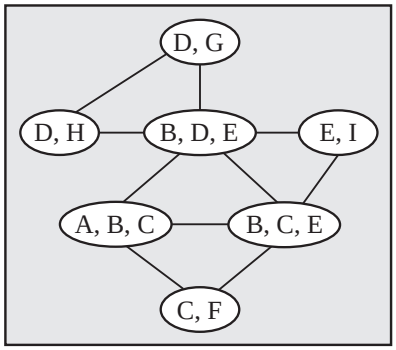


FIGURA 4.31. Grafo de conglomerados asociado al grafo de la Figura 4.30(a).

Definición 4.38 Grafo de unión. *Un grafo de conglomerados asociado a un grafo no dirigido se denomina un grafo de unión si contiene todos las aristas posibles que unan conglomerados con algún nodo común.*

Nótese que el grafo de unión asociado a un grafo dado es único. Por ejemplo, el grafo de conglomerados de la Figura 4.31 es el grafo de unión asociado al grafo de la Figura 4.30(a).

Los grafos de unión tienen la propiedad de que los conglomerados con nodos comunes forman un conjunto completo. Por esta razón, los grafos de unión suelen contener numerosas aristas. Por tanto, sería interesante obtener algún grafo de estructura más simple (por ejemplo, un árbol) que retuviese la propiedad de conectar los conglomerados que tengan elementos comunes.

Definición 4.39 Árbol de unión. *Un grafo de conglomerados se denomina un árbol de unión si es un árbol y todo nodo que pertenezca a dos conglomerados también pertenezca a todos los conglomerados contenidos en el camino que los une.*

Nótese que en un árbol de unión existe un único camino entre cada par de conglomerados con un nodo común.

Ejemplo 4.13 Árbol de unión. El árbol de conglomerados de la Figura 4.32(b) es un árbol de unión que se ha obtenido eliminando cuatro aristas del grafo de unión dado en la Figura 4.32(a). Se puede comprobar fácilmente que todos los conglomerados contenidos en el camino que une dos conglomerados contiene también los nodos comunes a éstos. Por ejemplo, los conglomerados $\{D, H\}$ y $\{B, D, E\}$ tienen un nodo común, D, que también pertenece al resto de los conglomerados en el camino que los une, $\{D, H\} - \{D, G\} - \{B, D, E\}$. ∎

En el Capítulo 8 se analizarán diversos métodos de propagación de incertidumbre que utilizan un árbol de unión para simplificar los cálculos necesarios para actualizar las probabilidades.

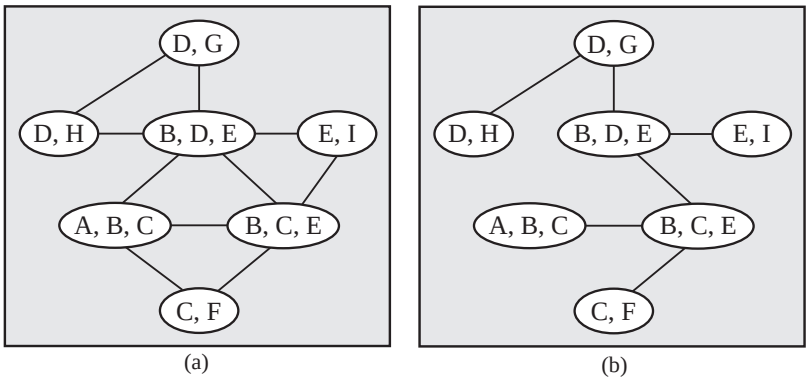


FIGURA 4.32. Un grafo de unión (a) y un árbol de unión asociado a él (b).

El teorema siguiente indica cuándo es posible convertir un grafo de unión en un árbol de unión eliminando algunas de sus aristas (ver Jensen (1988)).

Teorema 4.4 Árbol de unión. *Un grafo no dirigido posee un árbol de unión si y sólo si es triangulado.*

Ejemplo 4.14 Grafo sin árbol de unión. La Figura 4.33(a) muestra un grafo no triangulado y el grafo de unión asociado cuyos conglomerados son $C_1 = \{A, B\}, C_2 = \{B, D\}, C_3 = \{C, D\}$ y $C_4 = \{A, C\}$. Obsérvese que en esta situación es imposible construir un árbol de unión a partir de este grafo, pues se trata de un grafo no triangulado. Por ejemplo, si se eliminase del grafo la arista $C_1 - C_4$, el árbol resultante no sería un árbol de unión pues, por ejemplo, el nodo A está contenido en C_1 y C_4 pero no, en los dos conglomerados restantes del camino $C_1 - C_2 - C_3 - C_4$. ∎

En la Sección 4.5 se introdujo la propiedad de intersección dinámica para conglomerados. Esta propiedad permite ordenar los conglomerados de un grafo triangulado obteniendo una *cadena de conglomerados*. El algoritmo

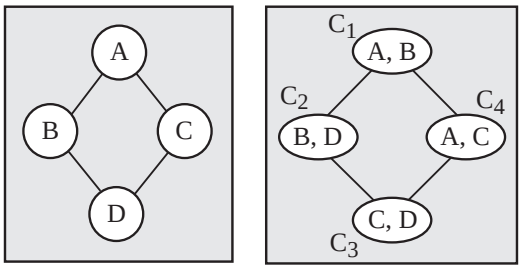


FIGURA 4.33. Grafo no triangulado (a) y grafo de unión asociado (b).

siguiente permite construir un árbol de unión asociado a un grafo triangulado. La idea básica consiste en organizar una cadena de conglomerados en una estructura de árbol.

Algoritmo 4.4 Generando un árbol de unión.

- **Datos:** Un grafo triangulado no dirigido $G = (X, L)$.
- **Resultado:** Un árbol de unión $G' = (C, L')$ asociado a G.

1. *Iniciación:* Utilizar el Algoritmo 4.3 para obtener una cadena de conglomerados del grafo G, $(C_1, \ldots, C_m)$.
2. Para cada conglomerado $C_i \in C$, escoger un conglomerado C_k en $\{C_1, \ldots, C_{i-1}\}$ con el máximo número de nodos comunes y añadir la arista $C_i - C_k$ a L' (inicialmente vacía). Los empates se deshacen de forma arbitraria. ∎

Ejemplo 4.15 Generación de un árbol de unión. A continuación se aplica el Algoritmo 4.4 para generar un árbol de unión asociado al grafo triangulado dado en la Figura 4.30(a). En el Ejemplo 4.12, se obtuvo la cadena de conglomerados $C_1 = \{A, B, C\}$, $C_2 = \{B, C, E\}$, $C_3 = \{B, D, E\}$, $C_4 = \{C, F\}$, $C_5 = \{D, G\}$, $C_6 = \{D, H\}$ y $C_7 = \{E, I\}$. Para generar un árbol de unión se procede a añadir las aristas necesarias a un conjunto L', inicialmente vacío, de la siguiente forma:

- Los conglomerados C_2 y C_3 tienen el máximo número de nodos en común con el conglomerado C_7. Deshaciendo el empate arbitrariamente, se elige el conglomerado C_3 y se añade la arista $C_7 - C_3$ a L'.
- C_3 y C_5 tienen el máximo número de nodos coincidentes con C_6. Se elige arbitrariamente uno de ellos, por ejemplo C_5 y se añade la arista $C_6 - C_5$ a L'.

- De entre los conglomerados de $\{C_1, C_2, C_3, C_4\}$, el conglomerado C_3 es el que tiene más elementos en común con C_5. Por tanto, se añade la arista $C_5 - C_3$ a L'.
- Procediendo de forma similar, se añaden las aristas $C_4 - C_2$, $C_3 - C_2$ y $C_2 - C_1$.

El árbol de unión resultante se muestra en la Figura 4.34. Dado que muchos empates se deciden de forma arbitraria, el algoritmo podría generar varios árboles de unión distintos para un mismo grafo no dirigido. ■

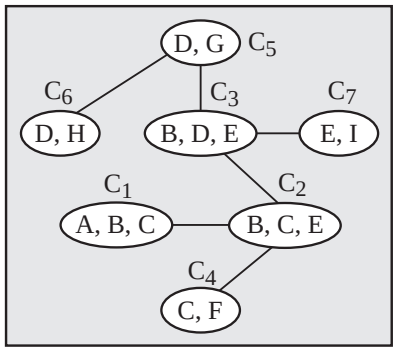


FIGURA 4.34. Un árbol de unión asociado al grafo de la Figura 4.30(a).

Hasta ahora se ha tratado el problema de la construcción de grafos de aglomerados asociados a grafos no dirigidos. Sin embargo, este concepto también puede ser aplicado a los grafos dirigidos trabajando indirectamente con el grafo no dirigido asociado. Como ya se verá en el Capítulo 8, las familias de nodos en un grafo dirigido juegan un papel importante en los mecanismos de propagación de evidencia. Existe también un tipo de redes probabilísticas que se definen mediante funciones locales de probabilidad definidas en las familias de los nodos (ver Sección 6.4.2). Por tanto, estamos interesados en el desarrollo de grafos de aglomerados tales que todas las familias del grafo dirigido original estén contenidas en, al menos, un aglomerado. Se tiene la siguiente definición.

Definición 4.40 Árbol de familias. *Un árbol de familias de un grafo dirigido D, es un árbol de unión de algún grafo no dirigido G asociado a D, en el cual la familia de cada nodo está contenida en al menos un conglomerado.*

El proceso de moralización de un grafo dirigido garantiza que la familia de cualquier nodo estará contenida en al menos un conglomerado del grafo no dirigido resultante. Por tanto, aplicando el Algoritmo 4.4 a cualquier versión triangulada del grafo moral se obtendrá un árbol de familias del grafo dirigido original.

Algoritmo 4.5 Generando un árbol de familias.

- **Datos:** Un grafo dirigido $D = (X, L)$.
- **Resultado:** Un árbol de familias $G' = (C, L')$ asociado a D.

1. Moralizar el grafo dirigido.
2. Triangular el grafo no dirigido resultante utilizando el Algoritmo 4.2.
3. Aplicar Algoritmo 4.4 para calcular un árbol de unión del grafo resultante. ∎

Ejemplo 4.16 Generando un árbol de familias. Considérese el grafo dirigido dado en la Figura 4.35(a), en la que las familias de los nodos se indican con sombras de distinta intensidad: $\{A\}$, $\{B\}$, $\{C\}$, $\{A, B, D\}$, $\{C, D, E\}$, $\{D, F\}$, $\{E, G\}$, $\{F, H\}$, $\{G, I\}$ y $\{G, H, J\}$. Aplicando el Algoritmo 4.5 se obtiene un árbol de familias asociado a este grafo:

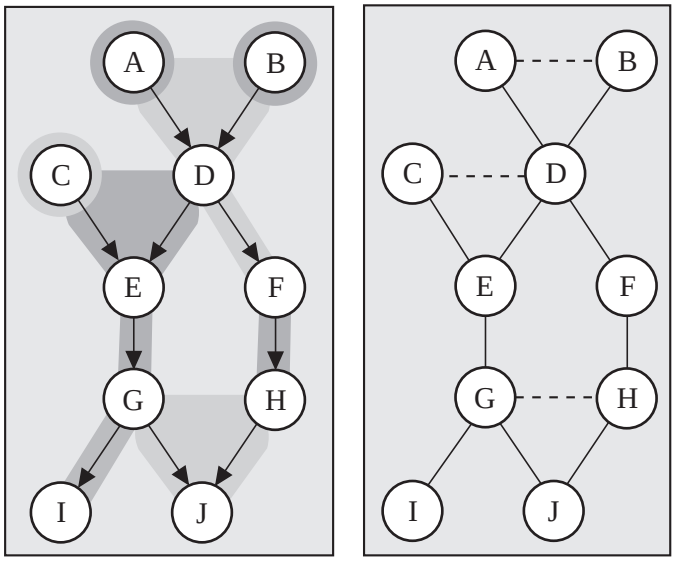


FIGURA 4.35. Un grafo dirigido (con las familias indicadas con distintos sombreados) (a) y el grafo moral correspondiente (b).

- Para construir el grafo moral correspondiente al grafo dirigido de la Figura 4.35(a), es necesario añadir las tres aristas mostradas en la Figura 4.35(b) con línea discontinua.

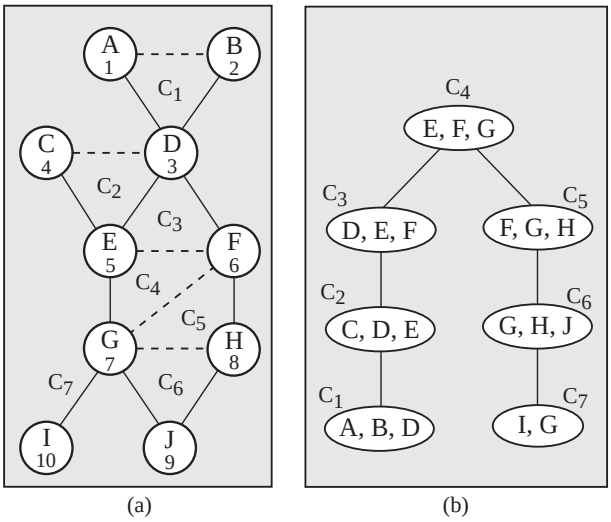


FIGURA 4.36. Numeración perfecta, ordenación de conglomerados para una triangulación del grafo, y árbol de unión asociado al grafo resultante (b).

- El grafo moral puede triangularse utilizando el Algoritmo 4.2. El grafo triangulado resultante, que contiene dos aristas adicionales, se muestra en la Figura 4.36(a).

- Finalmente, el árbol de familias asociado puede obtenerse aplicando a este grafo el Algoritmo 4.4 de la misma forma que en el Ejemplo 4.15. El árbol de familia resultante se muestra en la Figura 4.36(b). Puede comprobarse fácilmente que todas las familias del grafo dirigido de la Figura 4.35(a) están contenidas en al menos un conglomerado del árbol de familias. ■

4.7 Representación de Grafos

Un grafo puede ser representado de varias formas equivalentes que pongan de manifiesto en mayor o menor medida determinadas características. Las formas de representación más comunes son:

- Simbólicamente, como un par (X, L), donde X es un conjunto de variables y L es un conjunto de aristas entre pares de variables. Como ya se ha comentado anteriormente, una representación simbólica equivalente a la anterior viene dada por (X, Ady), donde Ady es la clase de los conjuntos de adyacencia de los nodos.

- Gráficamente, por medio de un diagrama formado por un conjunto de nodos (uno para cada variable) y un conjunto de líneas o flechas (una para cada arista del conjunto L).

- Numéricamente, utilizando ciertos tipos de matrices.

Cada una de estas representaciones presenta ventajas e inconvenientes. Por ejemplo, la representación simbólica es conceptualmente simple (cada grafo puede ser representado por un par de conjuntos), pero no proporciona información directa sobre la topología del grafo. La representación gráfica permite observar globalmente las distintas relaciones que existen entre las variables, pero tiene la desventaja de volverse extremadamente compleja cuando el número de aristas entre los nodos es muy elevado. Por último, la representación numérica permite obtener características de los grafos por simples manipulaciones algebraicas, pero tiene la desventaja de ser muy abstracta.

En esta sección se presentan dos métodos de representación gráfica de grafos y se analizan sus ventajas e inconvenientes (Sección 4.7.1). También se muestra la forma de representar la estructura de adyacencia de un grafo por medio de ciertos tipos de matrices que caracterizan algunas propiedades topológicas del grafo (Sección 4.7.2).

4.7.1 Representación Gráfica de un Grafo

Un grafo está formado por un conjunto de nodos y un conjunto de aristas. Esta sección se dedica al problema de representar gráficamente los nodos y aristas del grafo, por ejemplo en un hoja de papel o en la pantalla de un ordenador. El principal obstáculo con el que nos encontramos al intentar abordar este problema es que un grafo puede ser representado de muchas formas distintas. Sin embargo, algunas de estas representaciones son mejores que otras en términos de sencillez, capacidad para mostrar las principales características del grafo, etc. Estas representaciones permiten analizar visualmente ciertas propiedades topológicas del grafo de forma sencilla. Por ejemplo, el tipo de grafos que pueden ser dibujados en el plano sin que sus aristas se crucen se conocen como *grafos planos* y tiene numerosas aplicaciones interesantes. Los libros de Preparata y Shamos (1985) y Tamassia y Tollis (1995) ofrecen una descripción de los problemas asociados con la representación de grafos. En este libro se considerará que una representación gráfica es buena si cumple los siguientes requisitos:

1. Puede ser construida de forma sencilla y rápida utilizando algún algoritmo.

2. Las características topológicas del grafo podrán ser analizadas mediante la representación gráfica. Por ejemplo, la Figura 4.37 muestra dos representaciones distintas del mismo grafo; sin embargo, es más

fácil comprobar que el grafo es múltiplemente conexo a partir de la representación dada en la Figura 4.37(b) que a partir de la dada en la Figura 4.37(a). Otro ejemplo se tiene en las dos representaciones gráficas mostradas en la Figura 4.8, que corresponden a un grafo inconexo. Esta propiedad topológica se puede comprobar más fácilmente en el diagrama de la Figura 4.8(b) que en el de la Figura 4.8(a).

3. La representación será simple, teniendo un número mínimo de cortes de aristas.

En esta sección se presentan dos formas sistemáticas de representación gráfica:

1. La representación *circular* y
2. La representación *multinivel*.

Representación Circular de un Grafo

Una de las formas más sencillas de representar gráficamente un grafo es dibujar los nodos sobre una circunferencia a distancias iguales. Por ejemplo, la Figura 4.37(a) muestra la representación circular de un grafo. Esta representación tiene una propiedad importante: garantiza que no puede haber más de dos nodos alineados. Por tanto, si las aristas del grafo se representan mediante líneas rectas, esta representación garantiza que no habrá aristas ocultas entre los nodos. Así, la representación circular es la representación óptima para grafos con un gran número de aristas.

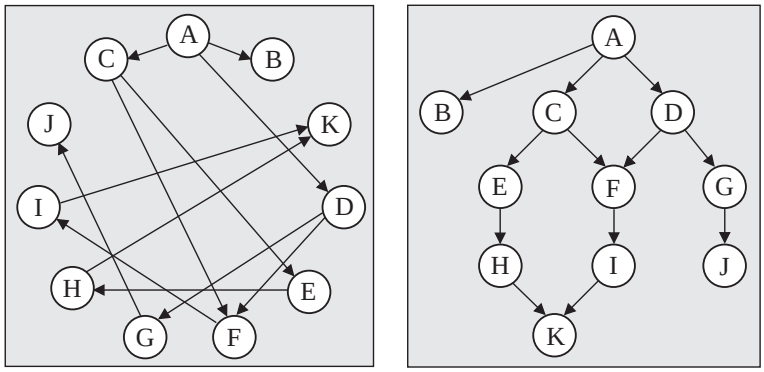


FIGURA 4.37. Dos representaciones gráficas del mismo grafo: circular (a) y multinivel (b).

La representación circular presenta las siguientes ventajas:

- Es fácil de construir.

- Todas las aristas son transparentes en el diagrama.
- Es la más conveniente para grafos completos o casi completos.

La principal desventaja de esta representación es que pueden existir numerosos cortes entre las aristas, complicando el diagrama. Por ejemplo, a pesar de que los dos diagramas de la Figura 4.37 representan el mismo grafo, en el diagrama (b) no existe ningún corte entre las aristas.

Representación Multinivel de un Grafo

La idea básica de la representación multinivel es organizar los nodos en distintos niveles, o capas, de tal forma que no existan aristas entre nodos del mismo nivel y que todo nodo en un nivel esté conectado con algún nodo del nivel previo. Así se podrá lograr una representación clara del grafo situando los nodos en niveles horizontales (como en la Figura 4.38) o verticales (como en la Figura 4.39). Para desarrollar esta idea en detalle, son necesarias algunas definiciones previas.

Definición 4.41 Subconjunto totalmente inconexo. *Dado un grafo (X, L), un subconjunto de nodos $S \subset X$ se denomina totalmente inconexo si no existe ninguna arista entre los nodos de S, es decir, si $(X_i, X_j) \in L \Rightarrow X_i \notin S$ o $X_j \notin S$.*

Definición 4.42 Representación multinivel. *Una representación multinivel de un grafo no dirigido (X, L) es una partición*

$$X = \bigcup_{k=1}^{m} S_k, \tag{4.1}$$

donde los niveles S_i, $i = 1, \ldots, m$, son subconjuntos disjuntos y totalmente inconexos de X tales que

$$\text{si } X_i \in S_k \Rightarrow \exists\, X_j \in S_{k-1} \text{ con } X_i \in Ady(X_j).$$

Es decir, no existen aristas entre nodos del mismo nivel y los nodos de un nivel son adyacentes a, al menos, un nodo del nivel anterior.

Nótese que los nodos que forman el primer nivel sólo tienen que satisfacer la propiedad de ser un subconjunto totalmente inconexo. Así, la elección del primer nivel es bastante arbitraria y, por tanto, un grafo puede tener varias representaciones multinivel. Los nodos del primer nivel se denominan *nodos raíz*. Por ejemplo, los grafos de las Figuras 4.38(a) y (b) muestran dos representaciones multinivel del mismo grafo. Los niveles asociados a la representación de la Figura 4.38(a) son

$$\{\{A\}, \{B, C\}, \{D, E, F\}, \{G, H, I\}\},$$

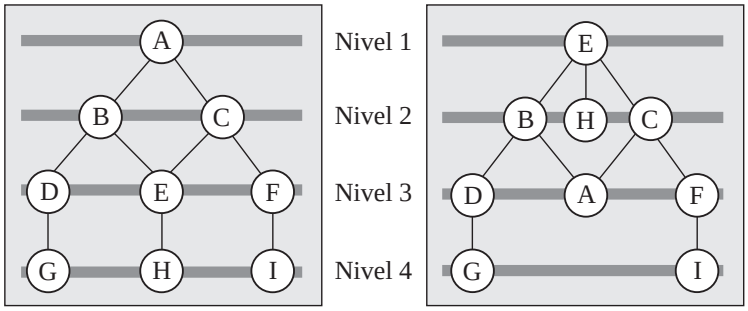


FIGURA 4.38. Dos representaciones multinivel del mismo grafo. Las líneas horizontales sombreadas indican los niveles.

y los niveles asociados con la Figura 4.38(b) son

$$\{\{E\}, \{B, H, C\}, \{D, A, F\}, \{G, I\}\}\,.$$

Obsérvese que estas dos representaciones contienen un único nodo raíz. La Figura 4.39 muestra una representación diferente con dos nodos raíz $\{D, H\}$.

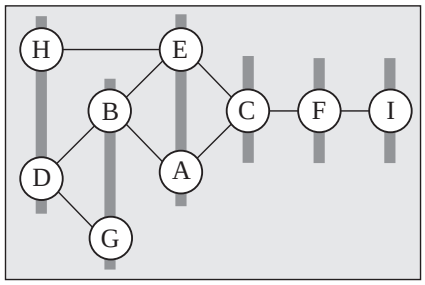


FIGURA 4.39. Representación multinivel vertical del grafo de la Figura 4.38.

Algunas ventajas de la representación multinivel son

- Es muy conveniente para árboles, o grafos con pocas aristas.
- Muestra la estructura ancestral del grafo a través de los distintos niveles de la representación.

Siempre es posible obtener una representación multinivel de un grafo no dirigido eligiendo como conjunto de nodos raíz cualquier subconjunto totalmente inconexo de nodos del grafo. El segundo nivel de la representación está formado por algunos de los nodos adyacentes a este conjunto de nodos, y así sucesivamente. Por ejemplo, en las representaciones mostradas en las Figuras 4.38 los únicos nodos raíz son los nodos A y E, respectivamente.

Sin embargo, en la Figura 4.39 el conjunto de nodos raíz es $\{D, H\}$. Esta es la idea básica del siguiente algoritmo.

Algoritmo 4.6 Representación multinivel.

- **Datos:** Un grafo (X, Ady) de n nodos y un conjunto de nodos raíz R.
- **Resultado:** Una representación multinivel S del grafo.

1. *Iniciación:* $Asignados = R$. $Nivel(1) = R$. $Nivel(k) = \phi$ para $k = 2, \ldots, n$. Tomar $j = 1$.
2. Si $Nivel(j) = \phi$, devolver $S = \{Nivel(1), \ldots, Nivel(j-1)\}$, que es una representación multinivel del grafo y terminar; en caso contrario, hacer $NivelActual = Nivel(j)$ e ir a la Etapa 3.
3. Seleccionar $X_k \in NivelActual$:
   (a) Añadir los elementos de $Ady(X_k) \setminus Asignados$ a $Nivel(j+1)$ y a $Asignados$.
   (b) Añadir los elementos de $Ady(X_k) \cap Nivel(j)$ a $Nivel(j+1)$ y eliminar estos elementos de $Nivel(j)$ y de $NivelActual$ e ir a la Etapa 4.
4. Eliminar X_k de $NivelActual$. Si $NivelActual = \phi$, tomar $j = j+1$ e ir a la Etapa 2; en caso contrario, ir a la Etapa 3. ∎

La Etapa 3(a) en el algoritmo anterior añade al nivel actual todos los vecinos no asignados de los nodos del nivel previo, mientras que la Etapa 3(b) elimina nodos vecinos del mismo nivel.

Nótese que si el conjunto de nodos raíz R no fuese totalmente inconexo, el algoritmo eliminaría de forma automática algunos de los nodos de este nivel hasta obtener un subconjunto totalmente inconexo.

Ejemplo 4.17 Representación multinivel. En este ejemplo se aplica el Algoritmo 4.6 para obtener una representación multinivel del grafo dado en la Figura 4.39. Este grafo tiene asociados los conjuntos de adyacencia:

$$\begin{array}{lll} Ady(A) = \{B, C\}, & Ady(B) = \{A, D, E\}, & Ady(C) = \{A, E, F\}, \\ Ady(D) = \{B, G\}, & Ady(E) = \{B, C, H\}, & Ady(F) = \{C, I\}, \\ Ady(G) = \{D\}, & Ady(H) = \{E\}, & Ady(I) = \{F\}. \end{array}$$

Considérese el conjunto de nodos raíz $\{D, H\}$. La Tabla 4.1 muestra el proceso que resulta de aplicar el Algoritmo 4.6. Las filas de la tabla muestran el estado de las variables correspondientes al final de cada etapa. Como resultado del algoritmo se obtiene la representación multinivel:

$$\begin{array}{ll} \text{Nivel(1)=\{D,H\}}, & \text{Nivel(2)=\{B,G\}}, \\ \text{Nivel(3) =\{A,E\}}, & \text{Nivel(4)=\{C\}}, \\ \text{Nivel(5) =\{F\}}, & \text{Nivel(6)=\{I\}}. \end{array}$$

Una representación gráfica de esta partición se muestra en la Figura 4.39.

Aplicando el Algoritmo 4.6 al grafo de la Figura 4.38 con los nodos raíz A y E se obtienen las representaciones multinivel mostradas en las Figuras 4.38(a) y 4.38(b), respectivamente. ∎

En el caso de grafos dirigidos puede considerarse el grafo no dirigido asociado para construir una representación multinivel. Sin embargo, este proceso no tendría en cuenta la direccionalidad de las aristas. A continuación se muestra que, en el caso de grafos dirigidos acíclicos, se puede dar un carácter dirigido a la representación multinivel en el sentido de que todas las aristas están orientadas en la misma dirección.

Definición 4.43 Representación multinivel dirigida. *Una representación multinivel de un grafo dirigido* (X, L) *es una partición*

$$X = \bigcup_{k=1}^{m} S_k, \tag{4.2}$$

donde los niveles $S_i,\ i = 1, \ldots, m$, *son subconjuntos disjuntos totalmente inconexos de* X *tales que*

$$X_i \in S_k \text{ y } (X_i, X_j) \in L \ \Rightarrow\ X_j \in S_r \text{ con } r > k,$$

es decir, todos los padres de un nodo dado han de estar en niveles anteriores al nivel del nodo.

Por ejemplo, la Figura 4.40 muestra dos representaciones multinivel distintas del grafo de la Figura 4.37. Los niveles (subconjuntos $S_k,\ k = 1, \ldots, 5$) se distinguen en la figura mediante sombras. Los niveles de las Figuras 4.40(a) y (b) son

$$\{\{A\}, \{B, C, D\}, \{E, F, G\}, \{H, I, J\}, \{K\}\}$$

y

$$\{\{A, B\}, \{C, D\}, \{E, F\}, \{H, I, G\}, \{K, J\}\},$$

respectivamente. Es fácil comprobar que ambas particiones cumplen (4.2).

Una ventaja de las representaciones multinivel dirigidas es que los diagramas resultantes son fáciles de interpretar, ya que todas las aristas están orientadas en la misma dirección. La Figura 4.40, muestra un diagrama multinivel en el que todas las aristas están orientadas en la dirección arriba-abajo.

A diferencia de los grafos no dirigidos, no todos los grafos dirigidos admiten una representación multinivel. El siguiente teorema caracteriza la clase de grafos dirigidos que admiten una representación multinivel dirigida.

Teorema 4.5 Representación multinivel dirigida. *Un grafo dirigido* (X, L) *admite una representación multinivel dirigida si y sólo si* (X, L) *es un grafo dirigido acíclico.*

Etapa	X_k	j	$Nivel(j)$	$Nivel(j+1)$	$NivelActual$
1	–	1	$\{D, H\}$	ϕ	ϕ
2	–	1	$\{D, H\}$	ϕ	$\{D, H\}$
3(a)	D	1	$\{D, H\}$	$\{B, G\}$	$\{D, H\}$
3(b)	D	1	$\{D, H\}$	$\{B, G\}$	$\{D, H\}$
4	D	1	$\{D, H\}$	$\{B, G\}$	$\{H\}$
3(a)	H	1	$\{D, H\}$	$\{B, G, E\}$	$\{H\}$
3(b)	H	1	$\{D, H\}$	$\{B, G, E\}$	$\{H\}$
4	H	2	$\{B, G, E\}$	ϕ	ϕ
2	H	2	$\{B, G, E\}$	ϕ	$\{B, G, E\}$
3(a)	B	2	$\{B, G, E\}$	$\{A\}$	$\{B, G, E\}$
3(b)	B	2	$\{B, G\}$	$\{A, E\}$	$\{B, G\}$
4	B	2	$\{B, G\}$	$\{A, E\}$	$\{G\}$
3(a)	G	2	$\{B, G\}$	$\{A, E\}$	$\{G\}$
3(b)	G	2	$\{B, G\}$	$\{A, E\}$	$\{G\}$
4	G	3	$\{A, E\}$	ϕ	ϕ
2	G	3	$\{A, E\}$	ϕ	$\{A, E\}$
3(a)	A	3	$\{A, E\}$	$\{C\}$	$\{A, E\}$
3(b)	A	3	$\{A, E\}$	$\{C\}$	$\{A, E\}$
4	A	3	$\{A, E\}$	$\{C\}$	$\{E\}$
3(a)	E	3	$\{A, E\}$	$\{C\}$	$\{E\}$
3(b)	E	3	$\{A, E\}$	$\{C\}$	$\{E\}$
4	E	4	$\{C\}$	ϕ	ϕ
2	E	4	$\{C\}$	ϕ	$\{C\}$
3(a)	C	4	$\{C\}$	$\{F\}$	$\{C\}$
3(b)	C	4	$\{C\}$	$\{F\}$	$\{C\}$
4	C	5	$\{F\}$	ϕ	ϕ
2	C	5	$\{F\}$	ϕ	$\{F\}$
3(a)	F	5	$\{F\}$	$\{I\}$	$\{F\}$
3(b)	F	5	$\{F\}$	$\{I\}$	$\{F\}$
4	F	6	$\{I\}$	ϕ	ϕ
2	F	6	$\{I\}$	ϕ	$\{I\}$
3(a)	I	6	$\{I\}$	ϕ	$\{I\}$
3(b)	I	6	$\{I\}$	ϕ	$\{I\}$
4	I	7	$\{I\}$	ϕ	ϕ

TABLA 4.1. Etapas de la construcción de una representación multinivel utilizando el Algoritmo 4.6.

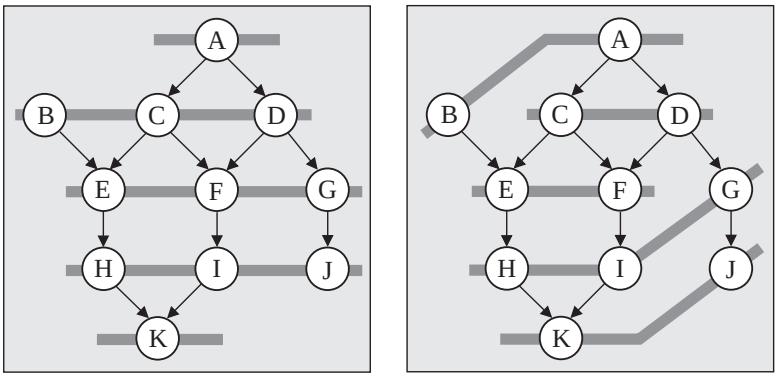


FIGURA 4.40. Dos representaciones multinivel de un grafo dirigido.

Utilizando el conjunto de nodos sin padres $\{X_i \mid \Pi_{X_i} = \phi\}$ de un grafo dirigido acíclico como conjunto de nodos raíz R, el Algoritmo 4.6 calculará una representación multinivel del grafo, como la dada en (4.2).

El siguiente teorema muestra una relación interesante entre representaciones multinivel dirigidas y numeraciones ancestrales.

Teorema 4.6 Numeración ancestral. *Si $\{S_1, \ldots, S_m\}$ es una representación multinivel de un grafo dirigido acíclico (X, L), entonces cualquier numeración de los nodos α que satisfaga $\alpha(X_i) > \alpha(X_j)$ para $X_i \in S_k$ y $X_j \in S_r$ con $k > r$ es una numeración ancestral de los nodos.*

Por tanto, el Algoritmo 4.6 permite obtener una numeración ancestral para un grafo dirigido acíclico. Además, este tipo de grafos es el único que posee este tipo de numeración. La comprobación de este teorema se deja como ejercicio para el lector.

En el resto de esta sección se desarrolla un algoritmo para dividir cualquier grafo dirigido acíclico de la forma mostrada en (4.2). Para ello, se necesitan las siguientes definiciones.

Definición 4.44 Profundidad ascendente. *La profundidad ascendente de un nodo X_i en un grafo dirigido acíclico, $PA(X_i)$, es la longitud máxima de los caminos del grafo que terminan en el nodo X_i.*

Definición 4.45 Profundidad descendente. *La profundidad descendente de un nodo X_i en un grafo dirigido acíclico, $PD(X_i)$, es la longitud máxima de los caminos del grafo que comienzan en el nodo X_i.*

Para calcular la profundidad ascendente de un nodo basta con conocer la profundidad ascendente de sus padres. Análogamente, la profundidad descendente de un nodo está determinada por la profundidad descendente de sus hijos. La profundidad descendente crece siguiendo el sentido de o-

rientación de las aristas, mientras que la profundidad ascendente crece en el sentido contrario.

Los conceptos de profundidad descendente y ascendente satisfacen las siguientes propiedades:

- $0 \leq PA(X_i) \leq n-1$ y $0 \leq PD(X_i) \leq n-1$, donde n es el número de nodos.
- Si X_i no tiene padres, entonces $PA(X_i) = 0$.
- Si X_i no tiene hijos, entonces $PD(X_i) = 0$.

Un grafo puede ser dividido en niveles calculando la profundidad de todos los nodos del grafo (cada nivel estará formado por los nodos con la misma profundidad).

Definición 4.46 Niveles de profundidad de un grafo. *Dado un grafo dirigido, el k-ésimo nivel de profundidad ascendente, NA_k, es el subconjunto de nodos $\{X_i \in X \mid PA(X_i) = k\}$. De forma similar, el k-ésimo nivel de profundidad descendente, ND_k, es el subconjunto $\{X_i \in X \mid PD(X_i) = k\}$.*

El número de niveles no vacíos de un grafo puede ser calculado en función de la longitud máxima de sus caminos. Sea m la longitud del camino más largo del grafo, entonces $N_k = \phi$, $\forall k > m$ y $N_k \neq \phi$, $\forall k \leq m$. Así, se tiene un número finito de niveles que definen una partición del grafo como la indicada en la Definición 4.43. Este resultado es confirmado por el teorema siguiente.

Teorema 4.7 Niveles de profundidad y representación multinivel dirigida. *Para cualquier grafo dirigido acíclico, los conjuntos $\{ND_k : k = 0, \ldots, m\}$ y $\{NA_k : k = 0, \ldots, m\}$ son dos representaciones multinivel dirigidas que cumplen (4.2).*

Por ejemplo, la Figura 4.41 muestra los niveles de profundidad ascendente y descendente asociados al grafo dirigido de la Figura 4.40.

A continuación se introduce un algoritmo para calcular los niveles de profundidad de un grafo dirigido. Dado el carácter iterativo del algoritmo, éste también permite comprobar si un grafo es acíclico. El algoritmo se ilustra para el caso de la profundidad ascendente. Basta una pequeña modificación del algoritmo para obtener profundidades descendentes. Esta modificación se deja como ejercicio al lector.

Algoritmo 4.7 Profundidad ascendente para grafos dirigidos.

- **Datos:** Un grafo (X, Ady) de n nodos.
- **Resultado:** Los niveles de profundidad ascendente $\{NA_0, \ldots, NA_k\}$ del grafo.

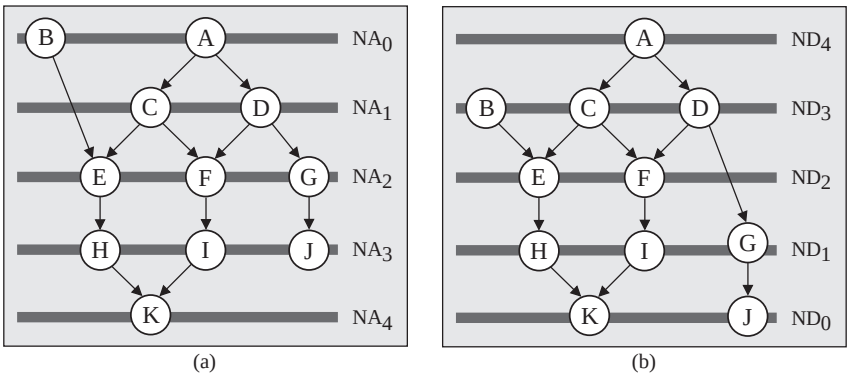


FIGURA 4.41. Niveles de profundidad ascendente (a) y descendente (b) de un grafo dirigido.

1. *Iniciación:* Definir $PA(X_i) = 0$ para todos los nodos X_i que no posean padres. Si todos los nodos del grafo tienen algún padre, el algoritmo finaliza pues el grafo contiene algún ciclo. En caso contrario, tomar $profundidad = 1$ y continuar con la Etapa 2.

2. Si $profundidad \leq n$, ir a la Etapa 3; en caso contrario, el algoritmo finaliza (el grafo contiene ciclos).

3. Seleccionar un nodo X_i con $PA(X_i) = profundidad - 1$. Asignar a todos los nodos X_j adyacentes a X_i la profundidad $PA(X_j) = profundidad$. Repetir este proceso con todos los nodos en el nivel $profundidad - 1$, e ir a la Etapa 4.

4. Si ningún nodo tiene profundidad $profundidad$, entonces el algoritmo finaliza y las profundidades de todos los nodos han sido calculadas. En caso contrario, incrementar $profundidad$ en una unidad y volver a la Etapa 2. ■

La Figura 4.42 muestra el pseudocódigo para el algoritmo de profundidad ascendente. Si el algoritmo no finaliza antes de la Etapa n, o finaliza en la Etapa inicial, entonces el grafo contiene algún ciclo. En caso contrario, el grafo es acíclico, y el algoritmo obtiene las profundidades de los nodos. En este caso, el grafo contiene tantos niveles como etapas realizadas.

4.7.2 Representación Numérica de Grafos

Un grafo también puede ser representado numéricamente utilizando determinados tipos de matrices. La siguiente representación permite calcular de forma sencilla diversas características topológicas de un grafo.

Algoritmo de Niveles de Profundidad Ascendente

Datos: Un grafo dirigido (X, Ady) de n nodos.

Resultado: Los niveles de profundidad ascendente asociados.

Etapa Inicial:

para $k = 1$ **hasta** n **hacer**
  si nodo X_k no tiene padres **entonces** $PA(X_k) \leftarrow 0$
si $PA(X_k) \neq 0,\ k = 1, \ldots, n$
  Terminar. El grafo contiene algún ciclo.

Etapa de Iteración i-ésima:

si $i > n$ **entonces**
  Terminar. El grafo contiene algún ciclo.
en otro caso
  para todo X_k tal que $PA(X_k) = i - 1$ **hacer**
    $PA(X_r) \leftarrow i$, para todo $X_r \in Ady(X_k)$
  si $PA(X_k) \neq i,\ k = 1, \ldots, n$
    Terminar. Todas las profundiades calculadas.
  en otro caso
    ir a la etapa iterativa $i + 1$-ésima

FIGURA 4.42. Pseudocódigo del algoritmo de profundidad ascendente para grafos dirigidos.

Definición 4.47 Matriz de adyacencia. *Sea $G = (X, L)$ un grafo de n nodos y sea $A = (a_{ij})$ una matriz $n \times n$, donde*

$$a_{ij} = \begin{cases} 1, & \textit{si } L_{i\,j} \in L, \\ 0, & \textit{en caso contrario.} \end{cases}$$

La matriz A se denomina matriz de adyacencia del grafo G.

Mediante sencillas manipulaciones algebráicas de la matriz de adyacencia se pueden obtener algunas características del grafo como, por ejemplo, el número de caminos distintos que unen dos nodos, comprobar si el grafo es conexo, etc.

La Figura 4.43, muestra el proceso de construcción de la matriz de adyacencia de un grafo dado. Cuando $a_{ij} = 0$, entonces no existe ninguna arista del nodo X_i al nodo X_j. En cambio, $a_{ij} = 1$ indica que el nodo X_i está conectado al nodo X_j, o que los nodos son adyacentes, de ahí el nombre de esta matriz.

La matriz A contiene toda la información topológica del grafo asociado; por tanto, esta matriz caracteriza al grafo. Notar que:

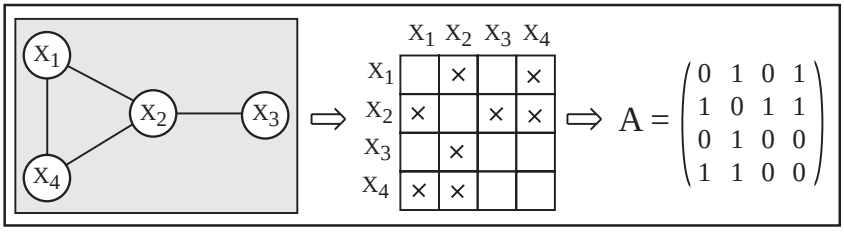


FIGURA 4.43. Proceso de construcción de la matriz de adyacencia de un grafo.

- La matriz de adyacencia de un grafo no dirigido es simétrica.
- Dado que $L_{i\,i} \notin L$ para todos los valores de i, los elementos diagonales de A son nulos.
- La matriz de adyacencia de un grafo no dirigido completo debe contener un uno en todos los elementos no diagonales.

La matriz de adyacencia permite comprobar si existe algún camino entre cada para de nodos. También puede calcularse la longitud de todos los caminos que unan cada par de nodos. El teorema siguiente muestra cómo se puede utilizar la matriz de adyacencia para esta tarea.

Teorema 4.8 Potencias de la matriz de adyacencia. *Sea A^r la r-ésima potencia de la matriz de adyacencia asociada con el grafo $G = (X, L)$. Entonces, el ij-ésimo elemento de A^r da el número de caminos de longitud r del nodo X_i al nodo X_j.*

Demostración: La demostración de este teorema puede ser fácilmente obtenida, por inducción, de la forma siguiente: El teorema se cumple para $r = 1$, ya que $a_{ij} = 1$ si existe un camino de longitud 1 (una arista) entre los nodos i y j, y $a_{ij} = 0$ en caso contrario.

Suponiendo que el resultado es cierto para A^r, para A^{r+1} se tiene que

$$A^{r+1} = A^r A \Leftrightarrow a_{ij}^{r+1} = \sum_{k=1}^{n} a_{ik}^r a_{kj},$$

es decir, si hay a_{ik}^r caminos de longitud r del nodo X_i al nodo X_k y existe una arista del nodo X_k al nodo X_j ($a_{kj} = 1$), entonces se tienen a_{ik}^r caminos de longitud $(r + 1)$. ∎

El Teorema 4.8 implica

- El elemento ij-ésimo de A^r es cero si y sólo si no existe ningún camino de longitud r de X_i a X_j.
- Calculando las potencias sucesivas de la matriz de adyacencia de un grafo dado $A, A^2, A^3, \ldots$, se pueden calcular directamente el número de caminos de longitud $1, 2, 3, \ldots$ que unen cada par de nodos.

Estas propiedades se ilustran en el ejemplo siguiente.

Ejemplo 4.18 Potencias de la matriz de adyacencia. Las primeras tres potencias de la matriz de adyacencia del grafo de la Figura 4.43 son

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix}, \quad A^2 = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 3 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}, \quad A^3 = \begin{pmatrix} 2 & 4 & 1 & 3 \\ 4 & 2 & 3 & 4 \\ 1 & 3 & 0 & 1 \\ 3 & 4 & 1 & 2 \end{pmatrix},$$

de las cuales puede deducirse que, por ejemplo, sólo existe un camino de longitud 3 del nodo X_1 al nodo X_3 ($a^3_{13} = 1$). La Figura 4.43 muestra este camino, $X_1 - X_4 - X_2 - X_3$. ∎

La matriz de adyacencia también puede ser utilizada para comprobar si un grafo es conexo o inconexo. Para ello, se introduce la siguiente matriz asociada a un grafo.

Definición 4.48 Matriz de alcanzabilidad. *La matriz de alcanzabilidad, $T = (t_{ij})$, de un grafo G se define como*

$$t_{ij} = \begin{cases} 1, & \textit{si existe algun camino del nodo } X_i \textit{ al nodo } X_j \\ 0, & \textit{en caso contrario.} \end{cases}$$

La matriz de alcanzabilidad está claramente relacionada con las potencias de la matriz de adyacencia. El siguiente resultado da una cota del número máximo de potencias de esta matriz que es necesario conocer para poder calcular la matriz de alcanzabilidad.

Teorema 4.9 Acotación a la longitud de un camino. *Dado un grafo con n nodos, si existe un camino del nodo X_i al nodo X_j, entonces también existe un camino de longitud menor que n de X_i a X_j.*

La demostración del teorema anterior se deja como ejercicio al lector. Por tanto, la matriz de alcanzabilidad puede ser obtenida a partir de un número finito de potencias de la matriz de adyacencia, $A, A^2, A^3, \ldots, A^{n-1}$. El número de potencias necesario es $n - 1$. De hecho, se tiene

$$t_{ij} = \begin{cases} 0, & \textit{si } a^k_{ij} = 0, \forall\, k < n \\ 1, & \textit{en caso contrario.} \end{cases} \tag{4.3}$$

En un grafo conexo, todos los elementos de la matriz de alcanzabilidad han de ser distintos de cero. Por tanto, la propiedad de conexión de un grafo se puede analizar a través de su matriz de alcanzabilidad. Además, en caso de que el grafo no sea conexo, la estructura de esta matriz permite identificar las componentes conexas del grafo.

Ejemplo 4.19 Matriz de alcanzabilidad. Dado el grafo de la Figura 4.44, es posible calcular su matriz de alcanzabilidad obteniendo las primeras

$n = 5$ potencias de su matriz de adyacencia. La matriz de adyacencia de este grafo es

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}.$$

Calculando las cinco primeras potencias, se obtiene la matriz de alcanzabilidad asociada utilizando (4.3):

$$T = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

A partir de la estructura de T, pueden distinguirse dos componentes conexas: $\{X_1, X_2, X_3\}$ y $\{X_4, X_5, X_6\}$. Esta conclusión podría ser difícil de obtener por medio de una reprentación gráfica. Por tanto, en grafos complejos, las matrices de adyacencia y alcanzabilidad son herramientas útiles para investigar la estructura topológica del grafo. ∎

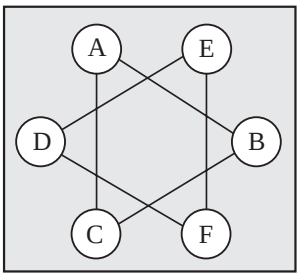


FIGURA 4.44. Ejemplo de un grafo inconexo.

4.8 Algunos Algoritmos para Grafos

En las secciones anteriores se han introducido varias propiedades y conceptos para analizar las características de los grafos. En esta sección se introducen algunos algoritmos útiles para comprobar si un grafo posee alguna de esas propiedades. Más concretamente, dado un grafo, estamos interesados en

1. Obtener un camino entre dos nodos.

2. Comprobar si el grafo es conexo y hallar sus componentes conexas.

3. Identificar si el grafo contiene bucles o ciclos.

No es nuestra intención mostrar los algoritmos óptimos para resolver cada uno de estos problemas, sino mostrar, con la ayuda de ejemplos ilustrativos, las ideas básicas que subyacen a estos métodos. Los lectores interesados en el diseño de algoritmos eficientes pueden consultar los libros de Cormen, Leiserson y Rivest (1990) y Golumbic (1980). Una descripción más detallada de algoritmos para grafos puede encontrarse en Gibbons (1985), McHugh (1990) y Skiena (1990).

4.8.1 Métodos de Búsqueda

Muchos algoritmos para grafos necesitan un mecanismo de búsqueda para explorar los nodos y aristas de un grafo. Por ejemplo, entre otras cosas, los algoritmos de búsqueda pueden ser utilizados para obtener un camino entre dos nodos, o para buscar un bucle o ciclo en un grafo. Estos métodos son la base para la construcción de los algoritmos introducidos en esta sección.

La exploración de un grafo comienza en un nodo inicial y consiste en la definición de un criterio para moverse hacia adelante y hacia atrás a través de las aristas del grafo, pasando de un nodo a un nodo vecino en cada etapa. Por tanto, la diferencia entre los distintos métodos de búsqueda radica en el criterio elegido para moverse de un nodo a otro. Por ejemplo, la Figura 4.45 muestra una búsqueda exhaustiva de un grafo comenzando en el nodo A. Obsérvese que, siguiendo la secuencia indicada en esta figura, y pasando de un nodo a un nodo vecino en cada etapa, se pueden visitar todos los nodos del grafo en un orden predeterminado: A, B, D, G, H, C, E, I, F, J y K. Obsérvese también que cualquier arista del grafo es recorrida como máximo dos veces: una en la dirección de avance (líneas continuas) para alcanzar nuevos nodos y una en la dirección de retroceso (líneas discontinuas) volviendo hacia atrás, a algún nodo ya visitado.

En la literatura han sido propuestas numerosas técnicas de búsqueda heurísticas (ver, por ejemplo, Rich y Knight (1991)). En esta sección se analizan dos de las técnicas más utilizadas para explorar un grafo:

- **Método de búsqueda en profundidad:** En cada etapa del método de búsqueda en profundidad se visita alguno de los vecinos no visitados del nodo actual (ver Figura 4.46(a), donde los números indican el orden en que se visitan los nodos). En caso de que el nodo actual no tenga ningún vecino no visitado, el algoritmo vuelve atrás al nodo visitado anteriormente y el proceso de búsqueda continua hasta que todos los nodos han sido visitados.

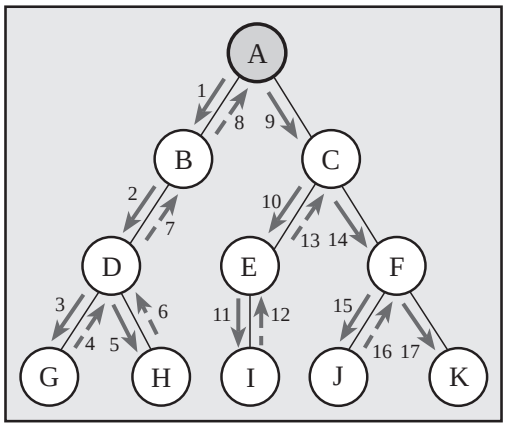


FIGURA 4.45. Ejemplo de un proceso de búsqueda.

- **Método de búsqueda en anchura:** El método de búsqueda en anchura visita los nodos del grafo capa a capa, comenzando en un nodo inicial y visitando, en la primera etapa todos los vecinos del nodo inicial. Después, se selecciona alguno de estos vecinos como nuevo nodo y se repite el proceso (ver Figura 4.46(b), donde los números indican el orden en que se visitan los nodos).

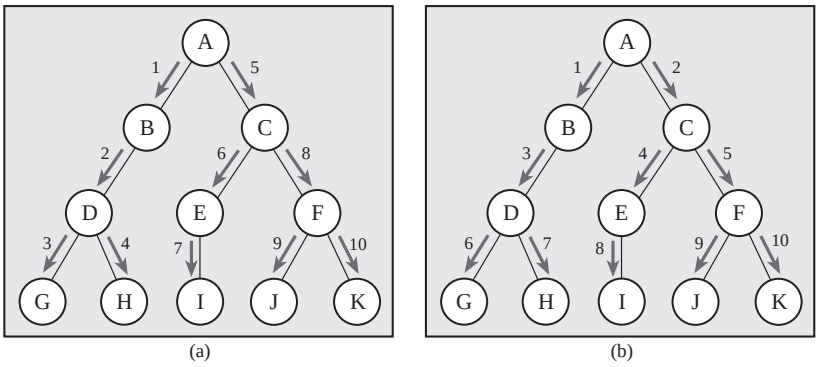


FIGURA 4.46. Ilustración del método de búsqueda en profundidad (a) y de búsqueda en anchura (b). Los números indican el orden en que se visitan los nodos

En las secciones siguientes se desarrollan varios algoritmos basados en estos métodos.

4.8.2 Algoritmos de Búsqueda de Caminos

Dado un grafo $G = (X, L)$, se trata de encontrar un camino del nodo X_i al nodo X_j, en caso de que exista. En esta sección se introducen dos algoritmos de búsqueda de caminos basados en las dos estrategias anteriores. Para este propósito es más conveniente y eficiente utilizar la representación de un grafo por medio de los conjuntos de adyacencia (ver Definición 4.5). El grafo no digido de la Figura 4.47(a) puede ser representado por (X, L), donde X es el conjunto de nodos $\{A, B, C, D, E, F, G\}$ y L es el conjunto de aristas $\{L_1, \ldots, L_8\}$. Sin embargo, desde un punto de vista computacional, la representación del grafo por medio de sus conjuntos de adyacencia es más adecuada:

$$\begin{array}{lll} Ady(A) = \{B, C, D\}, & Ady(B) = \{A, E\}, & Ady(C) = \{A, F\}, \\ Ady(D) = \{A, F\}, & Ady(E) = \{B, G\}, & Ady(F) = \{C, D, G\}, \\ Ady(G) = \{E, F\}. & & \end{array} \tag{4.4}$$

Por tanto, $G = (X, L)$ puede ser representado también mediante $G = (X, Ady)$, donde Ady son los conjuntos de adyacencia dados en (4.4). Esta representación es más eficiente para los métodos de búsqueda pues evita tener que comprobar todas las aristas del grafo para elegir el siguiente nodo del proceso.

El grafo dirigido de la Figura 4.47(b) tiene los conjuntos siguientes de adyacencia:

$$\begin{array}{lll} Ady(A) = \{B, C, D\}, & Ady(B) = \{E\}, & Ady(C) = \{F\}, \\ Ady(D) = \{F\}, & Ady(E) = \{G\}, & Ady(F) = \{G\}, \\ Ady(G) = \phi. & & \end{array} \tag{4.5}$$

Otra propiedad importante de los conjuntos de adyacencia es que proporcionan una representación independiente del carácter dirigido o no dirigido del grafo. Por ejemplo, si nos diesen el grafo dirigido de la Figura 4.47(b) y se quisiese realizar alguna operación de carácter no dirigido (obtener bucles, caminos no dirigidos, etc.), bastaría con considerar los conjuntos de adyacencia correspondientes al grafo no dirigido asociado (4.4).

Basándose en las dos técnicas de búsqueda descritas anteriormente, es posible definir de forma sencilla los siguientes algoritmos de búsqueda de caminos: *búsqueda de caminos en profundidad* y *búsqueda de caminos en anchura*.

Algoritmo 4.8 Búsqueda de caminos en profundidad

- **Datos:** Un grafo arbitrario (X, Ady) y dos nodos X_i y X_j.
- **Resultado:** Un camino $Camino = \{X_{i_1}, \ldots, X_{i_r}\}$ del nodo $X_i = X_{i_1}$ al nodo $X_j = X_{i_r}$. Si no existe tal camino, entonces $Camino = \phi$.

1. *Iniciación:* Tomar $X_k = X_i$, $Camino = \{X_i\}$ y $Visitados = \{X_i\}$.

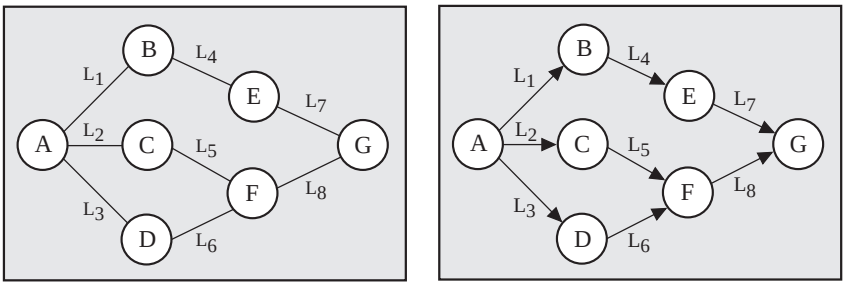


FIGURA 4.47. Ejemplo de un grafo no dirigido (a) y dirigido (b).

2. *Iteración:* Si todos los nodos de $Ady(X_k)$ han sido ya visitados, o si $Ady(X_k) = \phi$, ir a la Etapa 4; en caso contrario, ir a la Etapa 3.

3. *Etapa de avance:* Elegir un nodo $X_r \in Ady(X_k)$, tal que $X_r \notin Visitados$, y añadir X_r a $Camino$ y a $Visitados$. Si $X_r = X_j$, entonces el algorimo finaliza con resultado $Camino$; en caso contrario, tomar $X_k = X_r$ e ir a la Etapa 2.

4. *Etapa de retroceso:* Si $X_k = X_i$, el algoritmo finaliza pues no hay ningún camino de X_i a X_j; en caso contrario, eliminar X_k de $Camino$, asignar a X_k el último nodo de $Camino$, e ir a la Etapa 2. ∎

La Figura 4.48 muestra el pseudocódigo para este algoritmo. En cada etapa del algoritmo se actualizan las listas:

- *Camino*, que contiene un camino de X_i al último nodo visitado.
- *Visitados*, que contiene los nodos que ya han sido visitados.

Ejemplo 4.20 Búsqueda de caminos en profundidad. Dado el grafo no dirigido de la Figura 4.49(a), se desea obtener un camino entre los nodos A y F. En la Tabla 4.2 se recoge el resultado de aplicar el Algoritmo 4.8 a este grafo. Esta tabla muestra los valores del nodo actual X_k, sus nodos adyacentes aún no visitados $Ady(X_k) \setminus Visitados$, y las listas $Camino$ y $Visitados$ al final de la etapa indicada. Obsérvese que en este caso no se realiza ninguna etapa de retroceso para obtener el camino $A-B-E-G-F$ (ver Figura 4.49(a)).

Por otra parte, si se considera el grafo dirigido de la Figura 4.49(b), el algoritmo de búsqueda de caminos en profundidad realiza diversas etapas de avance y retroceso hasta encontrar el camino $A \rightarrow C \rightarrow F$. La Tabla 4.3 recoge las etapas de este ejemplo. En este caso, el proceso de búsqueda llega en algún momento al nodo G, pero $Ady(G) = \phi$. Por tanto, el algoritmo vuelve al nodo anterior para poder continuar el proceso de búsqueda (ver Figura 4.49(b)). ∎

Algoritmo de Búsqueda de Caminos en Profundidad

Datos: Un grafo (X, Ady) y dos nodos X_i y X_j.
Resultado: Un camino de X_i a X_j, o ϕ si no existe ningún camino.

Etapa Inicial:

$X_k \leftarrow X_i$
$Camino \leftarrow \{X_i\}$
$Visitados \leftarrow \{X_i\}$

Etapa de Iteración:

si existe $X_r \in Ady(X_k) \setminus Visitados$, **entonces**
    añadir X_r a $Visitados$ y a $Camino$
    si $X_r = X_j$, **entonces**
        Terminar. Se ha hallado un camino.
    en otro caso
        Tomar $X_k \leftarrow X_r$
        repetir la etapa de iteración.
en otro caso
    si $X_k = X_i$, **entonces**
        Terminar. No existe camino entre los nodos.
    en otro caso
        eliminar X_k de $Camino$
        $X_k \leftarrow$ último nodo en $Camino$
        repetir la etapa de iteración.

FIGURA 4.48. Pseudocódigo para el algoritmo de búsqueda de caminos en profundidad.

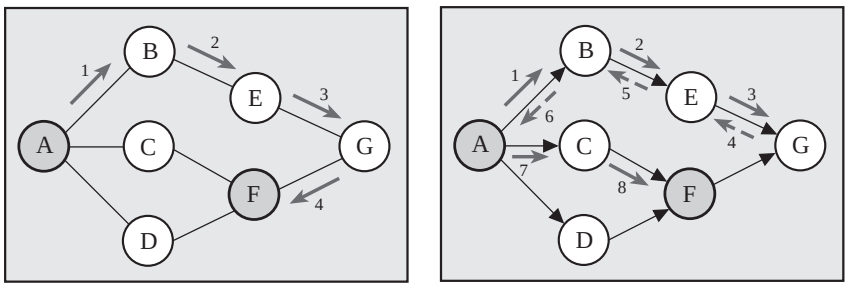


FIGURA 4.49. Etapas del algoritmo de búsqueda de caminos en profundidad para hallar un camino entre los nodos A y F en un grafo no dirigido (a) y en un grafo dirigido (b).

Etapas	X_k	$Ady(X_k) \setminus Visitados$	Visitados	Camino
1	A	$\{B, C, D\}$	$\{A\}$	$\{A\}$
2, 3	B	$\{E\}$	$\{A, B\}$	$\{A, B\}$
2, 3	E	$\{G\}$	$\{A, B, E\}$	$\{A, B, E\}$
2, 3	G	$\{F\}$	$\{A, B, E, G\}$	$\{A, B, E, G\}$
2, 3	F	$\{C, D, G\}$	$\{A, B, E, G, F\}$	$\{A, B, E, G, F\}$

TABLA 4.2. Etapas del algoritmo de búsqueda de caminos en profundidad para hallar un camino entre los nodos A y F en el grafo de la Figura 4.47(a).

Etapas	X_k	$Ady(X_k) \setminus Visitados$	Visitados	Camino
1	A	$\{B, C, D\}$	$\{A\}$	$\{A\}$
2, 3	B	$\{E\}$	$\{A, B\}$	$\{A, B\}$
2, 3	E	$\{G\}$	$\{A, B, E\}$	$\{A, B, E\}$
2, 3	G	ϕ	$\{A, B, E, G\}$	$\{A, B, E, G\}$
2, 4	E	ϕ	$\{A, B, E, G\}$	$\{A, B, E\}$
2, 4	B	ϕ	$\{A, B, E, G\}$	$\{A, B\}$
2, 4	A	$\{C, D\}$	$\{A, B, E, G\}$	$\{A\}$
2, 3	C	$\{F\}$	$\{A, B, E, G, C\}$	$\{A, C\}$
2, 3	F	ϕ	$\{A, B, E, G, C, F\}$	$\{A, C, F\}$

TABLA 4.3. Etapas del algoritmo de búsqueda de caminos en profundidad para hallar un camino entre los nodos A y F en el grafo de la Figura 4.47(b).

Dada la forma en que está implementado el Algoritmo 4.8, siempre se obtienen caminos simples, es decir, caminos que no contienen dos veces el mismo nodo (no contienen bucles ni ciclos). Sin embargo, como se verá en la Sección 4.8.4, el algoritmo anterior puede modificarse fácilmente para hallar bucles y ciclos. Por otra parte, el camino encontrado por este algoritmo no es, generalmente, el camino más corto entre los nodos.

A continuación se considera la estrategia de *búsqueda en anchura* describiendo un algoritmo para comprobar si existe un camino entre un par de nodos dados. La obtención del camino concreto se deja como ejercicio al lector.

Algoritmo 4.9 Búsqueda de caminos en anchura.

- **Datos:** Un grafo arbitrario (X, Ady) y dos nodos X_i y X_j.
- **Resultado:** Existencia o no de un camino entre X_i y X_j.

1. *Iniciación:* Definir $Visitados = \phi$ y $Cola = \{X_i\}$.
2. *Iteración:* Seleccionar el primer nodo, X_k, en la lista $Cola$, eliminarlo de esta lista y añadirlo a $Visitados$.

Algoritmo de Búsqueda de Caminos en Anchura

Datos: Un grafo (X, Ady) y dos nodos X_i y X_j.
Resultado: Existencia de un camino de X_i a X_j.

Etapa Inicial:

$Cola \leftarrow \{X_i\}$
$Visitados \leftarrow \phi$

Etapa de Iteración:

$X_k \leftarrow$ primer nodo en $Cola$
Eliminar X_k de $Cola$
Añadir X_k a $Visitados$
si $X_k = X_j$, **entonces**
    Terminar (existe un camino de X_i a X_j).
en otro caso
    $S \leftarrow Ady(X_k) \setminus Visitados$
    si $S \neq \phi$ **entonces**
        Añadir S al comienzo de $Cola$
        Repetir la etapa de iteración.
    en otro caso
    si $Cola = \phi$, **entonces**
        Terminar (no existe ningún camino de X_i a X_j).
    en otro caso
        Repetir la etapa de iteración.

FIGURA 4.50. Pseudocódigo del algoritmo de búsqueda de caminos en anchura.

3. Si $X_k = X_j$, entonces existe un camino entre X_i y X_j y el algoritmo finaliza. En caso contrario, si todos los vecinos de X_k han sido visitados previamente, ir a la Etapa 4; en caso contrario, ir a la Etapa 5.

4. Si $Cola = \phi$, entonces no existe ningún camino entre X_i y X_j y el algoritmo finaliza. En caso contrario, ir a la Etapa 2.

5. Añadir a la lista $Cola$ todos los nodos no visitados de $Ady(X_k)$ e ir a la Etapa 2. ∎

La Figura 4.50 muestra el pseudocódigo del algoritmo de búsqueda de caminos en anchura. En cada etapa del algoritmo, se actualizan las siguientes listas:

- *Visitados*, que contiene los nodos que ya han sido visitados.

Etapas	Nodo X_k	$Visitados$	$Cola$
1	–	ϕ	$\{A\}$
2	A	$\{A\}$	$\{\}$
3,5	A	$\{A\}$	$\{B, C, D\}$
2	B	$\{A, B\}$	$\{C, D\}$
3,5	B	$\{A, B\}$	$\{C, D, E\}$
2	C	$\{A, B, C\}$	$\{D, E\}$
3,5	C	$\{A, B, C\}$	$\{D, E, F\}$
2	D	$\{A, B, C, D\}$	$\{E, F\}$
3,4	D	$\{A, B, C, D\}$	$\{E, F\}$
2	E	$\{A, B, C, D, E\}$	$\{F\}$
3,5	E	$\{A, B, C, D, E\}$	$\{F, G\}$
2	F	$\{A, B, C, D, E, F\}$	$\{G\}$

TABLA 4.4. Etapas del algoritmo de búsqueda de caminos en anchura para comprobar si existe un camino entre los nodos A y F.

- *Cola*, que contiene los nodos en cola pendientes de visitar.

Si durante el proceso de ejecución de este algoritmo se alcanza el nodo X_j, entonces se habrá hallado un camino. En caso contrario, después de la búsqueda exhaustiva de un camino, el algoritmo concluye que tal camino no existe.

Ejemplo 4.21 Búsqueda de caminos en anchura. En este ejemplo se utiliza el algoritmo de búsqueda en anchura para comprobar si existe algún camino entre los nodos A y F en los grafos no dirigido y dirigido de la Figura 4.51. En este caso, el algoritmo sigue las mismas etapas en ambos grafos. La Tabla 4.4 muestra los valores de las variables que intervienen en cada etapa de este proceso. El algoritmo finaliza en la Etapa 3, concluyendo que existe un camino entre los nodos. ■

La complejidad de los algoritmos anteriores es lineal en el número de aristas y nodos del grafo. La eficiencia de cada uno de estos algoritmos dependerá de la topología particular que se tenga en cada caso. En general, el algoritmo de búsqueda en profundidad es más eficiente que el de búsqueda en anchura cuando los caminos que unen los nodos inicial y final son largos (ver Figura 4.52); en cambio, la situación es la contraria si los nodos están unidos por caminos cortos.

4.8.3 Comprobando la Conexión de un Grafo

Los métodos de búsqueda descritos anteriormente también pueden utilizarse para comprobar si un grafo es conexo. La idea es realizar una búsqueda exhaustiva de los nodos del grafo, obteniendo el conjunto S de

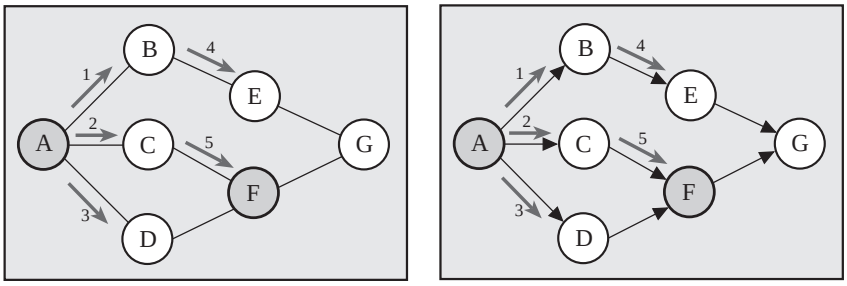


FIGURA 4.51. Etapas del algoritmo de búsqueda de caminos en anchura para comprobar si existe un camino entre los nodos A y F.

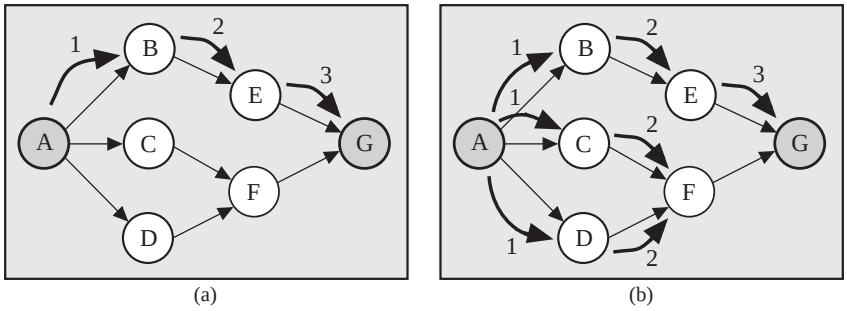


FIGURA 4.52. Búsqueda de un camino entre los nodos A y G con el algoritmo de búsqueda en profundidad (a) y en anchura (b).

nodos que son alcanzables desde un nodo inicial. Si el grafo es conexo, entonces el conjunto S contendrá todos los nodos del grafo; en caso contrario, el subconjunto de nodos S sólo contendrá la componente conexa del grafo que contiene al nodo inicial.

Los Algoritmos 4.8 y 4.9 pueden ser utilizados para realizar una búsqueda exhaustiva considerando el mismo nodo inical y final, es decir, el conjunto $Visitados$ resultante de la ejecución de estos algoritmos contendrá la componente conexa correspondiente a X_i.

Algoritmo 4.10 Búsqueda de componentes conexas.

- **Datos:** Un grafo (X, Ady).
- **Resultado:** El conjunto de componentes conexas C de (X, Ady).

1. *Iniciación:* Definir $Visitados = \phi$, $C = \phi$.
2. Si $X \setminus Visitados = \phi$, finalizar y devolver C; en caso contrario, elegir un nodo de $X_i \in X \setminus Visitados$ e ir a la Etapa 3.

3. Utilizar el Algoritmo 4.8 ó 4.9 para realizar una búsqueda exhaustiva del grafo (X, Ady) comenzando en el nodo X_i y obtener el conjunto S de nodos visitados.

4. Añadir S a C. Añadir a $Visitados$ todos los nodos en S. Ir a la Etapa 2. ∎

Si el conjunto C contiene una sóla componente conexa, entonces el grafo es conexo; en caso contrario, el grafo es inconexo y C contiene todas las componentes conexas del grafo.

Ejemplo 4.22 Búsqueda de Componentes Conexas. En la Sección 4.3.2 se ha visto que el grafo no dirigido dado en la Figura 4.53(a) es inconexo. Utilizando el Algoritmo 4.10 se pueden calcular sus componentes conexas.

- Inicialmente se considera $Visitados = \phi$ y $C = \phi$.
- $X \setminus Visitados = X = \{A, B, C, D, E, F\}$. Se elige el primero de estos nodos como nodo inicial $X_k = A$ para la primera búsqueda exhaustiva.
- Se utiliza el Algoritmo 4.8 con $X_i = X_j = A$, obteniéndose el conjunto de nodos visitados $S = C_1 = \{A, C, E\}$.
- Por tanto, se tiene $C = \{C_1\}$ y $Visitados = \{A, C, E\}$.
- $X \setminus Visitados = \{B, D, F\}$. Se toma $X_k = B$.
- Utilizando de nuevo el Algoritmo 4.8 con $X_i = X_j = B$, se obtiene el conjunto de nodos visitados $C_2 = \{B, D, F\}$.
- Ahora se tiene $Visitados = \{A, C, E, B, D, F\}$, $C = \{C_1, C_2\}$.
- Dado que $X \setminus Visitados = \phi$, el algoritmo finaliza obteniendo C.

Entre las componentes conexas están los subconjuntos $C_1 = \{A, C, E\}$ y $C_2 = \{B, D, F\}$. Por tanto, el grafo de la Figura 4.53(a) es inconexo y contiene las dos componentes conexas, C_1 y C_2, tal y como se muestra en la Figura 4.53(b). ∎

4.8.4 Búsqueda de Bucles y Ciclos

Como ya se mencionó al final del Ejemplo 4.20, los algoritmos de búsqueda de caminos pueden modificarse fácilmente para hallar bucles o ciclos en un grafo. En esta sección se muestan las modificaciones necesarias para adaptar el algoritmo de búsqueda en profundidad para esta tarea. Dado que el objetivo de este algoritmo es hallar un camino cerrado (un bucle o

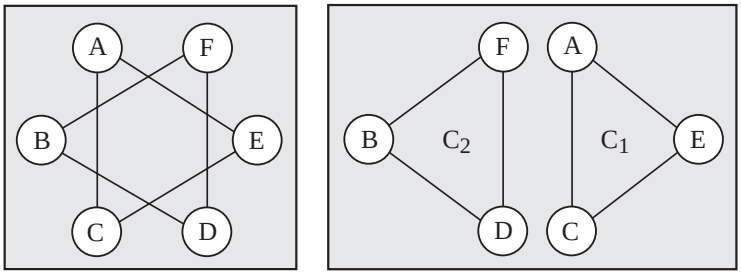


FIGURA 4.53. Grafo no dirigido inconexo (a) y sus componentes conexas (b).

un ciclo), se puede utilizar el Algoritmo 4.8 comprobando en cada etapa si hay algún nodo contenido en el camino que también esté contenido en la lista de nodos adyacentes del nodo actual. Los caminos cerrados resultantes serán bucles (si el grafo es no dirigido) o ciclos (si el grafo es dirigido). El algoritmo selecciona un nodo inicial arbitrario y busca de forma exhaustiva un camino cerrado en el grafo.

Algoritmo 4.11 Búsqueda de caminos cerrados en profundidad.

- **Datos:** Un grafo (X, Ady).
- **Resultado:** Un camino cerrado, *Camino*. Si el grafo no contiene ningún camino cerrado, entonces $Camino = \phi$.

1. *Iniciación:* Definir $Camino = \phi$ y $Visitados = \phi$.
2. Si existe algún nodo $X_i \in X \setminus Visitados$, ir a la Etapa 3; en caso contrario, el algoritmo finaliza (no existe ningún camino cerrado en el grafo).
3. Añadir X_i a $Visitados$ y tomar $Camino = \{X_i\}$, tomar $X_k = X_i$ y $Previo = X_i$.
4. *Iteración:* Si existe algún nodo $X_r \in Ady(X_k) \cap Camino$, con $X_r \neq Previo$, entonces añadir X_r a $Camino$ y finalizar (se ha encontrado un camino cerrado); en caso contrario, ir a la Etapa 5.
5. Si todos los nodos de $Ady(X_k)$ han sido ya visitados, o $Ady(X_k) = \phi$, ir a la Etapa 7; en caso contrario, ir a la Etapa 6.
6. *Etapa de Avance:* Elegir algún nodo $X_r \in Ady(X_k)$, tal que $X_r \notin Visitados$. Definir $Previo = X_k$, añadir X_r a $Camino$ y $Visitados$, tomar $X_k = X_r$, e ir a la Etapa 4.
7. *Etapa de Retroceso:* Eliminar X_k de $Camino$. Si $X_k = X_i$, ir a la Etapa 2; en caso contrario, asignar a X_k el último nodo en $Camino$, e ir a la Etapa 5. ∎

El algoritmo anterior considera un nodo arbitrario del grafo, X_i, como nodo incial. Si no se encuentra ningún camino cerrado (el algoritmo vuelve al nodo original), entonces se comprueba si todos los nodos han sido visitados concluyéndose, en ese caso, que no existe ningún camino cerrado en el grafo; en caso contrario, el algoritmo elige alguno de los nodos no visitados como nodo inicial y repite el proceso. La forma en que este algoritmo está concebido hace que no sólo sea válido para grafos dirigidos y no dirigidos, sino también para grafos conexos e inconexos. El siguiente ejemplo ilustra la aplicación de este algoritmo.

Ejemplo 4.23 Búsqueda de Bucles y Ciclos. Considérese el grafo no dirigido dado en la Figura 4.54(a) que contiene dos bucles, $A-B-D-A$ y $D-G-F-D$. Supóngase que se aplica el Algoritmo 4.11 comenzando en el nodo A. La Tabla 4.5 muestra las etapas seguidas por el algoritmo. Esta tabla muestra, en cada etapa del algoritmo, el nodo X_k, el nodo *Previo* asociado, el *Camino* actual, el conjunto $Ady(X_k) \cap Camino$, que se utiliza para indicar si existe algún camino cerrado, y el conjunto $Visitados$ que contiene los nodos que han sido visitados. Las etapas seguidas por el algoritmo se ilustran en la Figura 4.54(a).

Estas etapas se resumen de la siguiente forma, donde se utiliza el orden alfabético para seleccionar los nodos: Inicialmente se viaja de A a B y de B a C. Al alcanzar el nodo C ya no es posible avanzar, luego se vuelve un paso atrás, al nodo B y se viaja al único vecino aún no visitado, D. El conjunto $Ady(D) \cap Camino$ contiene al nodo A, que no es el nodo *Previo* a D. Por tanto, se ha encontrado el bucle $A-B-D-A$. Si se ignorase el nodo A se podría continuar viajando, buscando un bucle distinto. De esta forma se pueden obtener todos los bucles contenidos en el grafo. Por ejemplo, si en la Etapa 5 se eligiese el nodo G o F en lugar del nodo A, se obtendría un bucle distinto: $D-G-F-D$.

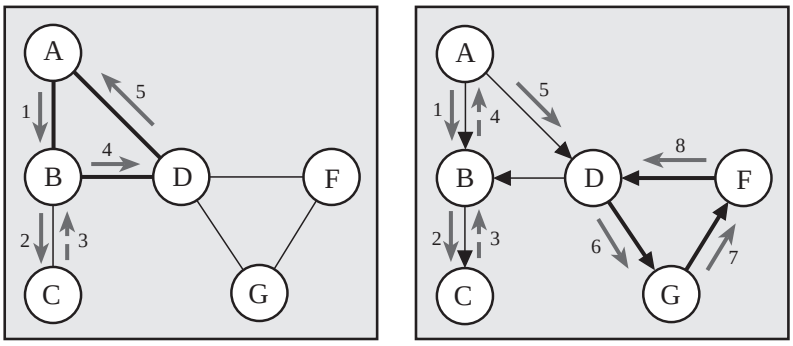


FIGURA 4.54. Etapas del algoritmo de búsqueda de caminos cerrados en profundidad para un grafo no dirigido (a) y un grafo dirigido (b).

Etapa	X_k	*Previo*	*Camino*	*Visitados*
1	–	–	ϕ	ϕ
2, 3	A	A	$\{A\}$	$\{A\}$
4, 5, 6	B	A	$\{A, B\}$	$\{A, B\}$
4, 5, 6	C	B	$\{A, B, C\}$	$\{A, B, C\}$
4, 5, 7	B	B	$\{A, B\}$	$\{A, B, C\}$
5, 6	D	B	$\{A, B, D\}$	$\{A, B, C, D\}$
4	A	–	$\{A, B, D, A\}$	$\{A, B, C, D\}$

TABLA 4.5. Etapas del Algoritmo 4.11 para buscar bucles en el grafo no dirigido de la Figura 4.54(a).

Etapa	X_k	*Previo*	*Camino*	*Visitados*
1	–	–	ϕ	ϕ
2, 3	A	A	$\{A\}$	$\{A\}$
4, 5, 6	B	A	$\{A, B\}$	$\{A, B\}$
4, 5, 6	C	B	$\{A, B, C\}$	$\{A, B, C\}$
4, 5, 7	B	B	$\{A, B\}$	$\{A, B, C\}$
5, 7	A	B	$\{A\}$	$\{A, B, C\}$
5, 6	D	A	$\{A, D\}$	$\{A, B, C, D\}$
4, 5, 6	G	D	$\{A, D, G\}$	$\{A, B, C, D, G\}$
4, 5, 6	F	G	$\{A, D, G, F\}$	$\{A, B, C, D, G, F\}$
4	F	G	$\{A, D, G, F, D\}$	$\{A, B, C, D, G, F\}$

TABLA 4.6. Etapas del Algoritmo 4.11 para hallar algún ciclo en el grafo de la Figura 4.54(b).

Considérese ahora el grafo dirigido cíclico de la Figura 4.54(b). Procediendo de la misma forma que en el caso anterior, y comenzando en el nodo A, el algoritmo realiza las etapas indicadas en la Tabla 4.6 y la Figura 4.54(b). En este caso, el algoritmo termina hallando el ciclo $D \to G \to F \to D$. Obsérvese que el grafo de la Figura 4.54(a) es el grafo no dirigido asociado a este grafo dirigido. Por tanto, sin más que cambiar los conjuntos de adyacencia se pueden obtener los ciclos de un grafo dirigido, o los bucles del grafo no dirigido asociado. ■

Ejercicios

4.1 Dado un grafo conexo $G = (X, L)$ y una cualquiera de sus aristas $L_{ij} \in L$, demostrar que las siguientes afirmaciones son equivalentes:

(a) El grafo $(X, L \setminus \{L_{ij}\})$ es conexo.

(b) La arista L_{ij} está contenida en algún bucle del grafo.

4.2 Dado un grafo no dirigido G, demostrar que las siguientes afirmaciones son equivalentes:

(a) Existe un único camino entre cada par de nodos en G.

(b) G es conexo, pero al eliminar una cualquiera de sus aristas se vuelve inconexo.

(c) G no tiene bucles, pero al añadir una arista cualquiera se forma un bucle.

Por tanto, estas condiciones proporcionan tres definiciones alternativas de árbol.

4.3 Demostrar que cualquier grafo que contenga un número igual o mayor de aristas que de nodos contiene al menos un bucle.

4.4 Considérese el grafo de la Figura 4.55:

(a) Encontrar los conjuntos de nodos ascendentes y descendentes del nodo C.

(b) ¿Cuál es la frontera del conjunto $\{B, C\}$ en el grafo no dirigido asociado?

(c) Repetir los cálculos anteriores en el grafo que resulta de invertir las aristas $A \to C$ y $C \to E$ en el grafo de la Figura 4.55.

(d) ¿Qué puede decirse sobre los conjuntos de nodos ascendentes y descendentes asociados a un nodo contenido en el ciclo $A \to B \to E \to C \to A$?

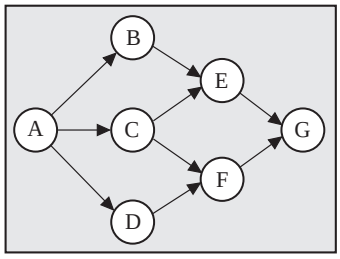


FIGURA 4.55. Ejemplo de un grafo dirigido.

4.5 Demostrar que un grafo que contiene un bucle de longitud 4 o mayor sin ninguna cuerda no posee una numeración perfecta. Es decir, probar que las numeraciones perfectas son exclusivas de grafos triangulados.

4.6 Completar el Ejemplo 4.10, comprobando que las dos numeraciones mostradas en la Figura 4.24 son numeraciones perfectas del grafo dado.

4.7 Triangular el grafo de la Figura 4.55 utilizando el algoritmo de triangulación por máxima cardinalidad (Algoritmo 4.2). ¿Cuantas triangulaciones distintas posee el grafo?. ¿Cuál de ellas es la mejor? Construir un árbol de familias del grafo dirigido original y un árbol de unión del grafo triangulado resultante.

4.8 Repetir el ejercicio anterior considerando el grafo de la Figura 4.56.

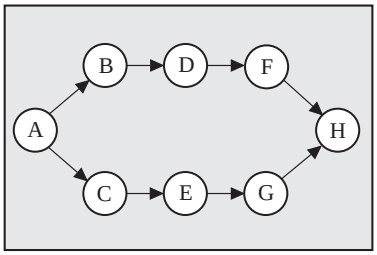


FIGURA 4.56. Un grafo dirigido.

4.9 Triangular el grafo de la Figura 4.57 utilizando el algoritmo de triangulación por máxima cardinalidad eligiendo el nodo F como nodo inicial. Seguir las mismas etapas que en el Ejemplo 4.11.

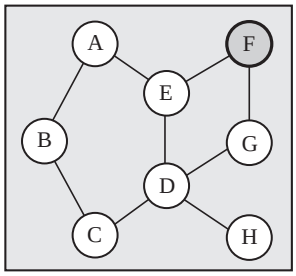


FIGURA 4.57. Un grafo no dirigido y no triangulado.

4.10 Probar que los grafos dirigidos acíclicos son el único tipo de grafos dirigidos que poseen una numeración ancestral.

4.11 Dado un grafo con la siguiente matriz de adyacencia:

$$A = \begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

- ¿Se trata de un grafo dirigido o no dirigido?
- Dibujar el grafo.
- ¿Es conexo?
- ¿Cuantos caminos de longitud 3 existen entre cada par de nodos distintos?

4.12 Probar el Teorema 4.9.

4.13 ¿Qué se puede decir acerca de un grafo cuya matriz de alcanzabilidad tiene ceros en todos los elementos de la diagonal?

4.14 Calcular el número de caminos de longitud m que unen dos nodos en un grafo completo de n nodos, K_n (ver Definición 4.8).

4.15 Modificar el algoritmo de búsqueda de caminos en anchura (Algoritmo 4.9) para hallar un camino entre dos nodos.

4.16 Aplicar el algoritmo de búsqueda de caminos en profundidad (Algoritmo 4.8) para encontrar un camino del nodo A a G en el grafo de la Figura 4.55. Proceder eligiendo los nodos en orden alfabético. ¿Qué pasaría si se eliminase del grafo la arista $E \rightarrow G$?

4.17 Utilizar el algoritmo de búsqueda de caminos en profundidad (Algoritmo 4.11) para encontrar bucles en el grafo de la Figura 4.55. Construir la tabla de nodos *Visitados* y la lista *Camino* en cada etapa del algoritmo. ¿Qué pasaría si se eliminase del grafo la arista $A \rightarrow B$?

4.18 Escribir y ejecutar un programa para cada uno de los siguientes algoritmos:

- Búsqueda de máxima cardinalidad (Algoritmo 4.1).
- Triangulación por máxima cardinalidad (Algoritmo 4.2).
- Árbol de unión por máxima cardinalidad (Algoritmo 4.4).
- Representación multinivel (Algoritmo 4.6).
- Búsqueda de caminos en profundidad (Algoritmo 4.8).
- Búsqueda de caminos en anchura (Algoritmo 4.9).
- Búsqueda de componentes conexas de un grafo (Algoritmo 4.10).
- Búsqueda de bucles en profundidad (Algoritmo 4.11).

Capítulo 5
Construcción de Modelos Probabilísticos

5.1 Introducción

En el Capítulo 3 se ha visto que la base de conocimiento de un sistema experto probabilístico esta formada por un conjunto de variables y un modelo probabilístico (una función de probabilidad conjunta) que describa las relaciones entre ellas. Por tanto, el funcionamiento del sistema experto depende de la correcta definición de la función de probabilidad conjunta que define el modelo probabilístico. Con el fin de que el proceso de definición del modelo sea lo más preciso posible, es conveniente seguir los siguientes pasos:

1. **Planteamiento del problema.** Como ya se mencionó en el Capítulo 1, el primer paso en el desarrollo de un sistema experto es la definición del problema a resolver. Por ejemplo, el problema del diagnóstico médico es un ejemplo clásico en el campo de los sistemas expertos: Dado que un paciente presenta una serie de síntomas, ¿cuál es la enfermedad más probable en esa situación?. La definición del problema es un paso crucial en el desarrollo del modelo, pues un mal planteamiento inicial tendrá consecuencias fatales para el modelo desarrollado.

2. **Selección de variables.** Una vez que el problema ha sido definido, el siguiente paso consiste en seleccionar un conjunto de variables que sean relevantes para su definición (esta tarea debe ser realizada por expertos en el problema a analizar). Por ejemplo, las variables relevantes para el problema de diagnóstico médico son las enfermedades

y sus correspondientes síntomas. Las variables relevantes para la definición de un modelo han de ser cuidadosamente seleccionadas a fin de eliminar posibles redundancias. Por ejemplo, en un problema de diagnóstico médico habrán de elegirse aquellos síntomas que mejor discriminen el conjunto de enfermedades dado.

3. **Adquisición de información relevante.** Una vez que se ha realizado el planteamiento inicial del problema, el siguiente paso consiste en la adquisición y análisis de toda la información (datos) que sea relevante para la definición del modelo. La información puede ser cuantitativa o cualitativa, obtenida de un experto, o de una base de datos. Esta información deberá ser cuidadosamente analizada utilizando técnicas de diseño experimental apropiadas. Es importante contar en esta etapa con la ayuda de especialistas en Estadística, pues el uso de métodos estadísticos permite mejorar la calidad de los datos y confirmar la validez de los métodos empleados para la obtención de las conclusiones.

4. **Construcción del modelo probabilístico.** Una vez que se conoce un conjunto de variables relevantes para el problema a analizar, y que se ha adquirido suficiente información para su definición, el siguiente paso consiste en la definición de una función de probabilidad conjunta que describa las relaciones entre las variables. Éste es, quizás, el paso más crítico y difícil en el desarrollo de un sistema experto:

   (a) Es crítico porque la bondad de los resultados del sistema experto dependerá de la precisión con que se haya definido la función de probabilidad conjunta, es decir, la calidad de los resultados no podrá superar a la calidad del modelo. Por tanto, una incorrecta definición del modelo probabilístico redundará en un sistema experto que dará conclusiones erróneas y/o contradictorias.

   (b) La estructura de la función de probabilidad conjunta (es decir, la estructura de dependencia e independencia entre las variables) no suele ser conocida en la práctica. Por tanto, habrá de ser inferida del conjunto de datos obtenidos previamente. Por tanto, la calidad del modelo tampoco podrá superar la calidad de los datos relevantes disponibles.

   (c) La estructura del modelo probabilístico puede depender de un número muy elevado de parámetros que complican su definición (ver Sección 3.5). Cuanto mayor sea el número de parámetros más complicada será la asignación de valores numéricos concretos en el proceso de definición del modelo. En cualquier caso, esta asignación habrá de ser realizada por un experto, o estimada a partir de la información disponible.

Los dos próximos capítulos están dedicados a la construcción de modelos probabilísticos (funciones de probabilidad conjunta) que definen la base de conocimiento de este tipo de sistemas expertos. Para ello, existen distintas metodologías

- Modelos definidos gráficamente.
- Modelos definidos por un conjunto de relaciones de independencia condicional.

Estas dos metodologías se analizan en los Capítulos 6 y 7, respectivamente. En este capítulo se introducen los conceptos necesarios. En la Sección 5.2 se describen algunos criterios de *separación gráfica* que permiten obtener las relaciones de independencia condicional asociadas a un grafo. Se recuerda al lector que una relación de independencia condicional, o simplemente una *independencia*, denotada por $I(X, Y|Z)$, significa que "X e Y son condicionalmente independientes dado Z", donde X, Y y Z son subconjuntos disjuntos de un conjunto de variables $\{X_1, \ldots, X_n\}$ (ver Sección 3.2.3). Cuando la relación de independencia es obtenida mediante un criterio de separación gráfico se emplea, de forma equivalente, la terminología "X e Y están separados por Z". En la Sección 5.3 se introducen varias propiedades de la independencia condicional. Dada una lista inicial de relaciones de independencia, estas propiedades permiten obtener independencias adicionales que estarán contenidas en el modelo probabilístico. La Sección 5.5 analiza distintas formas de factorizar una función de probabilidad conjunta mediante un producto de funciones de probabilidad condicionada. Finalmente, en la Sección 5.6 se describen los pasos necesarios para la construcción de un modelo probabilístico.

5.2 Criterios de Separación Gráfica

Los grafos son herramientas muy potentes para describir de forma intuitiva las relaciones de dependencia e independencia existentes en un conjunto de variables $\{X_1, \ldots, X_n\}$. Por tanto, una forma de definir un modelo probabilístico es partir de un grafo que describa las relaciones existentes entre las variables (este grafo puede venir dado, por ejemplo, por un experto en el tema). Este planteamiento motiva el siguiente problema:

- **Problema 5.1.** ¿Pueden representarse las estructuras de dependencia e independencia definidas por un grafo (dirigido o no dirigido) de forma equivalente por un conjunto de relaciones de independencia condicional? En caso afirmativo, ¿cómo se puede obtener este conjunto?

La respuesta al problema anterior es afirmativa, y una forma de obtener este conjunto de independencias es utilizar un criterio de separación gráfica

para comprobar cuáles, de entre todas las posibles relaciones de independencia condiconal, son satisfechas por el grafo. Los criterios de separación gráfica son las reglas para entender cómo pueden codificarse dependencias e independencias en un grafo. Estos criterios dependen del tipo de grafo (dirigido o no dirigido) que se esté considerando.

5.2.1 Separación en Grafos no Dirigidos

En muchas situaciones prácticas, las relaciones existentes entre un conjunto de variables $\{X_1, \ldots, X_n\}$ pueden ser representadas por un grafo no dirigido G. Como ya se mencionó en el Capítulo 4, cada variable puede ser representada por un nodo del grafo. Si dos variables son dependendientes, esta relación puede representarse por un camino que conecte estos nodos. Por otra parte, si dos variables son independientes, entonces no deberá existir ningún camino que una estos nodos. De esta forma, el concepto de dependencia entre variables puede relacionarse con el concepto de conexión entre nodos.

De forma similar, si la dependencia entre las variables X e Y es indirecta, a través de una tercera variable Z (es decir, si X e Y son condicionalmente dependientes dada Z), el nodo Z se representará de forma que no intersecte todos los caminos entre X y Y, es decir, Z no es un *conjunto de corte* (en inglés, *cutset*) de X e Y. Esta correspondencia entre dependencia condicional y separación en grafos no dirigidos constituye la base de la teoría de los *campos de Markov* (Isham (1981), Lauritzen (1982), Wermuth y Lauritzen (1983)), y ha sido caracterizada axiomáticamente de formas diversas (Pearl y Paz (1987)).

Para representar relaciones de independencia condicional por medio de grafos no dirigidos se necesita definir de forma precisa un criterio de separación apropiado, basándose en las ideas anteriormente expuestas. Este criterio se conoce como *criterio de U-separación*. A continuación se da una definición de este criterio y un algoritmo que permite su aplicación.

Definición 5.1 U-separación. *Sean X, Y y Z tres conjunto disjuntos de nodos de un grafo no dirigido G. Se dice que Z separa X e Y si y sólo si cada camino entre nodos de X y nodos de Y contiene algún nodo de Z. Cuando Z separe X e Y en G, y se denotará $I(X, Y|Z)_G$ para indicar que esta relación de independencia se deriva de un grafo G; en caso contrario, se denotará por $D(X, Y|Z)_G$, para indicar que X e Y son condicionalmente dependientes dada Z, en el grafo G.*

Se dice que X es *gráficamente independiente* de Y dada Z si Z separa X e Y. Por tanto, el criterio de U-separación permite obtener la lista de relaciones de independencia asociadas a un grafo no dirigido. Este criterio da la solución al Problema 5.1 para grafos no dirigidos. El caso de grafos dirigidos se analizará en la Sección 5.2.2.

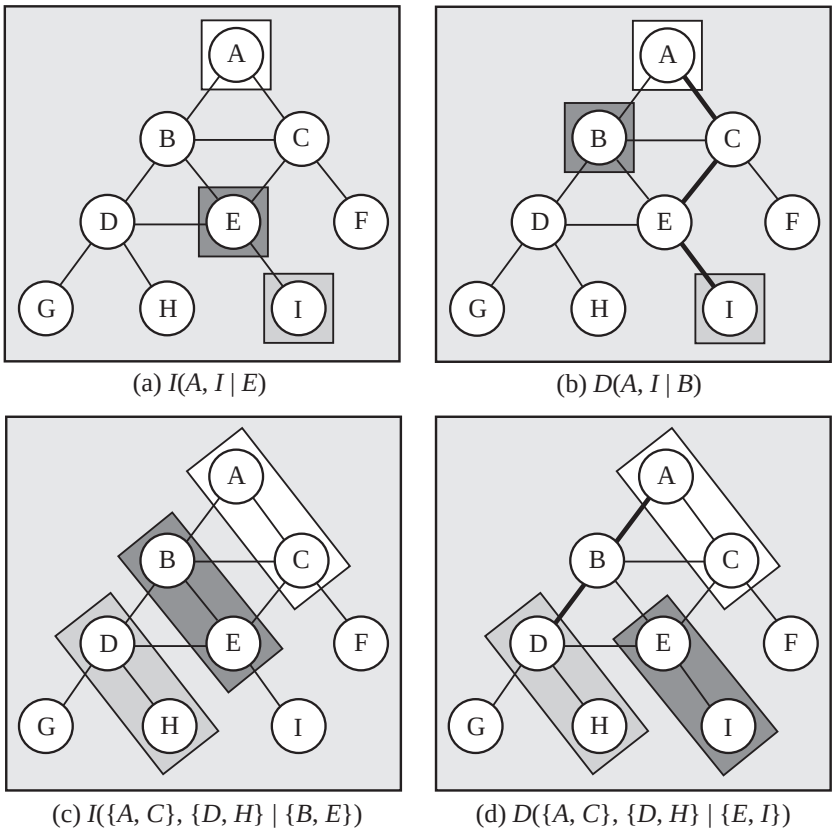
Ejemplo 5.1 U-separación. La Figura 5.1 ilustra cuatro casos distintos del concepto de U-separación En todos los casos, los tres conjuntos de interés están contenidos en cajas para su diferenciación: la caja asociada con el primer conjunto no está sombreada, la segunda tiene un sombreado claro, y la tercera (la asociada con el conjunto separador) muestra un sombreado oscuro.

- En la Figura 5.1(a), las variables A e I son condicionalmente independientes dada E, pues cada camino entre A e I contiene al nodo E. Por tanto, $I(A, I|E)_G$.
- En la Figura 5.1(b), los nodos A e I son condicionalmente dependientes dada B. En este caso, existe un camino, $(A - C - E - I)$, que no contiene al nodo B.
- En la Figura 5.1(c), los subconjuntos $\{A, C\}$ y $\{D, H\}$ son condicionalmente independientes dado el conjunto $\{B, E\}$, pues cada camino entre los dos conjuntos contiene, o bien a B, o bien a E. Por tanto, se tiene
$$I(\{A, C\}, \{D, H\}|\{B, E\})_G.$$
- Finalmente, en la Figura 5.1(d), los subconjuntos $\{A, C\}$ y $\{D, H\}$ son condicionalmente dependientes dado $\{E, I\}$, pues el camino $(A - B - D\}$ no contiene ninguna de las variables E e I. Por tanto,
$$D(\{A, C\}, \{D, H\}|\{E, I\})_G.$$
Siguiendo un proceso análogo, se puede comprobar si el grafo satisface cualquier otra relación de independencia. ■

5.2.2 Separación en Grafos Dirigidos

Para comprobar si un grafo dirigido verifica una relación de independencia dada, es necesario introducir otro criterio de separación, conocido como *criterio de D-separación*. Con el fin de dar una idea intuitiva de este concepto, considérese el siguiente ejemplo en el que intervienen seis variables relacionadas de la forma que se muestra en la Figura 5.2:

- L: Situación laboral.
- G: Ganancias por inversiones.
- E: Situación económica.
- S: Salud.
- D: Donaciones.



(a) $I(A, I \mid E)$ (b) $D(A, I \mid B)$

(c) $I(\{A, C\}, \{D, H\} \mid \{B, E\})$ (d) $D(\{A, C\}, \{D, H\} \mid \{E, I\})$

FIGURA 5.1. Ejemplo de ilustración del concepto de U-separación.

- F: Felicidad.

El grafo de la Figura 5.2 muestra que la situación laboral y las ganancias, fruto de inversiones, son causas directas de la situación económica de una persona. Por otra parte, la situación económica y la salud influyen en la felicidad. Finalmente, la situación económica determina las donaciones que realizada la persona. Dada esta situación, sería lógico pensar, por ejemplo, que la salud y la situación económica fuesen incondicionalmente independientes, pero condicionalmente dependientes una vez se tiene información sobre el estado de felicidad de la persona (un incremento de nuestra confianza en una variable disminuiría nuestra confianza en la otra). Para detectar las independencias definidas por este grafo, se necesita introducir un criterio de separación apropiado para grafos dirigidos, el concepto de D-separación; ver Pearl (1988) y Geiger, Verma y Pearl (1990a).

Definición 5.2 Nodo de aristas convergentes en un camino. *Dado un grafo dirigido y un camino no dirigido ($\ldots - U - A - V - \ldots$), el nodo*

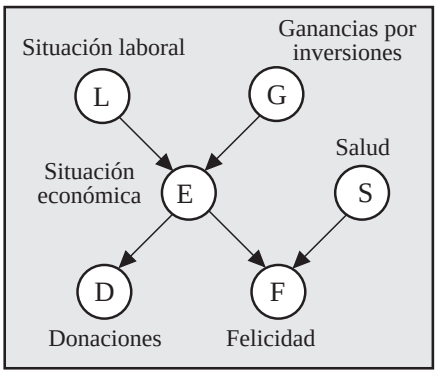


FIGURA 5.2. Un grafo dirigido ilustrando el concepto de D-separación.

A se denomina un nodo de aristas convergentes en este camino si las dos aristas del camino convergen a este nodo en el grafo dirigido, es decir, si el grafo dirigido contiene las aristas $U \rightarrow A$ *y* $V \rightarrow A$*).*

Ejemplo 5.2 Nodo de aristas convergentes. El nodo F es el único nodo de aristas convergentes en el camino no dirigido $L - E - F - S$ del grafo de la Figura 5.2. Obsérvese que aunque el nodo E posee dos aristas convergentes, no es un nodo de aristas convergentes en el camino, pues la arista $G \rightarrow E$ no está contenida en el camino. Sin embargo, el nodo E es un nodo de aristas convergentes en el camino no dirigido $L - E - G$. ■

Definición 5.3 D-Separación. *Sean* X*,* Y *y* Z *tres subconjuntos disjuntos de nodos en un grafo dirigido acíclico* D*; entonces se dice que* Z *D-separa* X *e* Y *si y sólo si a lo largo de todo camino no dirigido entre cualquier nodo de* X *y cualquier nodo de* Y *existe un nodo intermedio* A *tal que, o bien*

1. *A es un nodo de aristas convergentes en el camino y ni A ni sus descendientes están en* Z*, o bien*

2. *A no es un nodo de aristas convergentes en el camino y A está en* Z*.*

Cuando Z *D-separa* X *e* Y *en* D*, se escribe* $I(X, Y|Z)_D$ *para indicar que la relación de independencia viene dada por el grafo* D*; en caso contrario, se escribe* $D(X, Y|Z)_D$ *para indicar que* X *e* Y *son condicionalmente dependientes dado* Z *en el grafo* D*.*

Por tanto, si se puede encontrar un nodo en algún camino no dirigido que no cumpla las dos condiciones anteriores, entonces $D(X, Y|Z)_D$; en caso contrario, $I(X, Y|Z)_D$. Estas condiciones reflejan la idea de que las causas (padres) de cualquier mecanismo causal resultan dependientes una vez que se dispone de información del efecto que producen (un hijo). Por ejemplo, en el grafo dirigido de la Figura 5.2, la situación laboral y las ganancias fruto

de inversiones son incondicionalmente independientes, es decir, $I(L, G|\phi)_D$. Sin embargo, si se dispone de alguna información de la situación económica, entonces L y G se vuelven dependientes, $D(L, G|E)_D$, porque existe una relación entre la creencia que se tiene en las dos causas.

Ejemplo 5.3 *D*-**separación.** Considérese el grafo dirigido mostrado en la Figura 5.2. A partir de este grafo, se pueden derivar las relaciones de independencia siguientes:

- Caso (a). Independencia incondicional, $I(L, G|\phi)_D$: Los nodos L y G son incondicionalmente independientes pues están D-separados por ϕ. Tal y como puede observarse en la Figura 5.3(a), el único camino no dirigido, $L - E - G$, entre los nodos L y G contiene al nodo de aristas convergentes E, y ni él ni ninguno de sus descendientes están contenidos en ϕ.

- Caso (b). Dependencia condicional, $D(L, S|F)_D$: Los nodos L y S son condicionalmente dependientes dado F. En la Figura 5.3(b) puede verse que el único camino no dirigido entre L y S, $L - E - F - S$, contiene a los nodos E y F, y ninguno de estos nodos cumple las condiciones de la D-separación. Por tanto, L y S no están D-separados por F.

- Caso (c). Independencia condicional, $I(D, F|\{L, E\})_D$: Los nodos D y F son condicionalmente independientes dado $\{L, E\}$, pues el único camino no dirigido $D - E - F$ entre los nodos D y F contiene un sólo nodo intermedio, E, que no es un nodo de aristas convergentes, pero está contenido en $\{L, E\}$ (ver Figura 5.3(c)).

- Caso (d). Dependencia condicional, $D(D, \{S, F\}|L)_D$: El nodo D y el conjunto de nodos $\{S, F\}$ son condicionalmente dependientes dado L (ver Figura 5.3(d)). Obsérvese que el camino no dirigido $D - E - F$ entre D y F contiene al nodo E, que no es un nodo de aristas convergentes en este camino, pero no está contenido en $\{L\}$. ∎

El concepto de D-separación permite representar estructuras de dependencia e independencia en grafos dirigidos y, de esta forma, proporciona una solución al Problema 5.1. A continuación se introduce una definición alternativa de D-separación que es más fácil de aplicar en la práctica que la Definición 5.3.

Definición 5.4 D-Separación. *Sean X, Y y Z tres subconjuntos disjuntos en un grafo dirigido acíclico D, entonces se dice que Z D-separa a X e Y si y sólo si Z separa X e Y en el grafo moral del menor subconjunto ancestral*[1] *que contenga a los nodos de X, Y y Z.*

[1] Recuérdese que un conjunto ancestral es un conjunto de nodos que contiene los ascendientes de todos sus nodos (Definición 4.20).

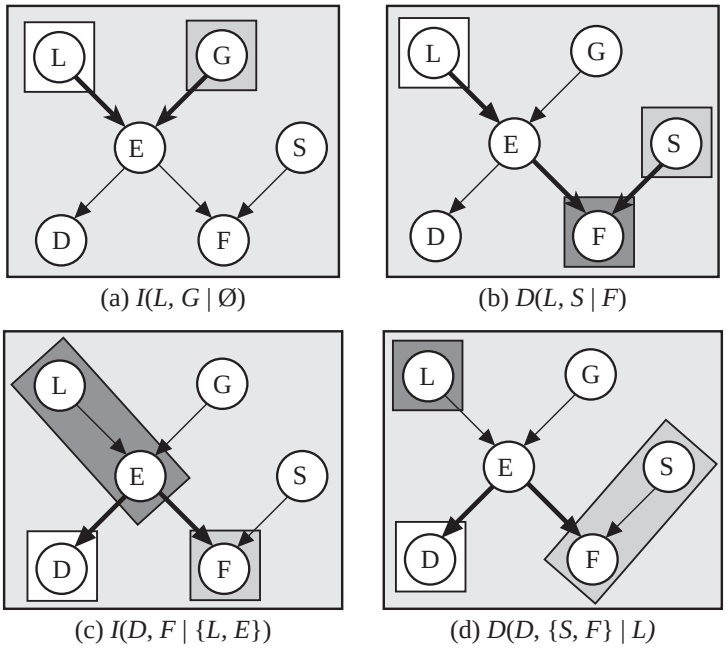


FIGURA 5.3. Ejemplos de ilustración del criterio de D-separación utilizando la Definición 5.3.

Esta definición alternativa fué propuesta por Lauritzen y otros (1990) que mostraron la equivalencia de la Definición 5.3 y la Definición 5.4, que ellos denominaron originalmente como *A-separación*.

La idea de moralizar el grafo, utilizada en esta definición, refleja la primera de las dos condiciones de la Definición 5.3. Si existiese un nodo de aristas convergentes A en un camino entre los nodos X e Y, tal que A o alguno de sus descendientes estuviese en Z, entonces A también estaría contenido en el menor conjunto ancestral que contuviera a X, Y y Z. Por tanto, puesto que A es un nodo de aristas convergentes, incluso en el caso de que A estuviera en Z, el proceso de moralización garantizaría la existencia de un camino no dirigido entre X e Y no interceptado por Z en el grafo moralizado correspondiente. Esta definición alternativa sugiere el siguiente algoritmo para la D-separación:

Algoritmo 5.1 D-Separación.

- **Datos:** Un grafo dirigido acíclico, D, y tres subconjuntos disjuntos de nodos X, Y y Z.

- **Resultado:** Comprobación de la relación de independencia $I(X, Y|Z)$ en D.

1. Obtener el menor subgrafo que contenga a X, Y, Z y sus subconjuntos de ascendientes.
2. Moralizar el grafo obtenido.
3. Utilizar el criterio de U-separación para comprobar si Z separa a X de Y. ∎

Ejemplo 5.4 D-separación. Considérese de nuevo el grafo dirigido de la Figura 5.2 y supóngase que se quieren comprobar, utilizando el Algoritmo 5.1, las mismas relaciones de independencia analizadas en el Ejemplo 5.3. La Figura 5.4 representa los cuatro casos, indicando con línea discontinua aquellas aristas que son eliminadas al construir el subgrafo ancestral.

- Caso (a). Independencia incondicional, $I(L, G|\phi)_D$: No existe ningún camino que conecte los nodos L y G en el grafo moral del menor subgrafo ancestral que contenga a L, G y ϕ (ver Figura 5.4(a)). Por tanto, $I(L, G|\phi)_D$.
- Caso (b). Dependencia condicional, $D(L, S|F)_D$: La Figura 5.4(b) muestra que existe un camino, $L - E - S$, que no contiene ningún nodo en $\{F\}$ y que conecta los nodos L y S en el grafo moral del menor subgrafo ancestral que contiene a L, S y F. Por tanto, $D(L, S|F)_D$.
- Caso (c). Independencia condicional, $I(D, F|\{L, E\})_D$: Existen dos caminos entre D y F, $D - E - F$ y $D - E - S - F$, en el grafo moral del menor subgrafo ancestral que contiene a D, L, E y F (ver Figura 5.4(c)). Ambos caminos contienen al nodo E, que está contenido en el conjunto $\{L, E\}$. Por tanto, $I(D, F|\{L, E\})_D$.
- Caso (d). Dependencia condicional, $D(D, \{S, F\}|L)_D$: La Figura 5.4 (d), muestra el camino $D - E - F$ que conecta D y $\{S, F\}$ en el grafo moral del menor subgrafo ancestral de $\{D, S, F, L\}$. Sin embargo, este camino no contiene al nodo L. Por tanto, $D(D, \{S, F\}|L)_D$. ∎

5.3 Algunas Propiedades de la Independencia Condicional

Hasta ahora se han introducido tres modelos distintos para definir relaciones de independencia condicional: modelos probabilísticos, modelos gráficos no dirigidos, y modelos gráficos dirigidos. En esta sección se analizan algunas propiedades de la independencia condicional que cumplen algunos de estos modelos. Estas propiedades permiten obtener nuevas relaciones de independencia a partir de un conjunto inicial de relaciones de independencia, dado por uno de estos modelos. Por ejemplo, dada la función de probabilidad conjunta $p(x_1, \ldots, x_n)$ de un conjunto de variables $\{X_1, \ldots, X_n\}$,

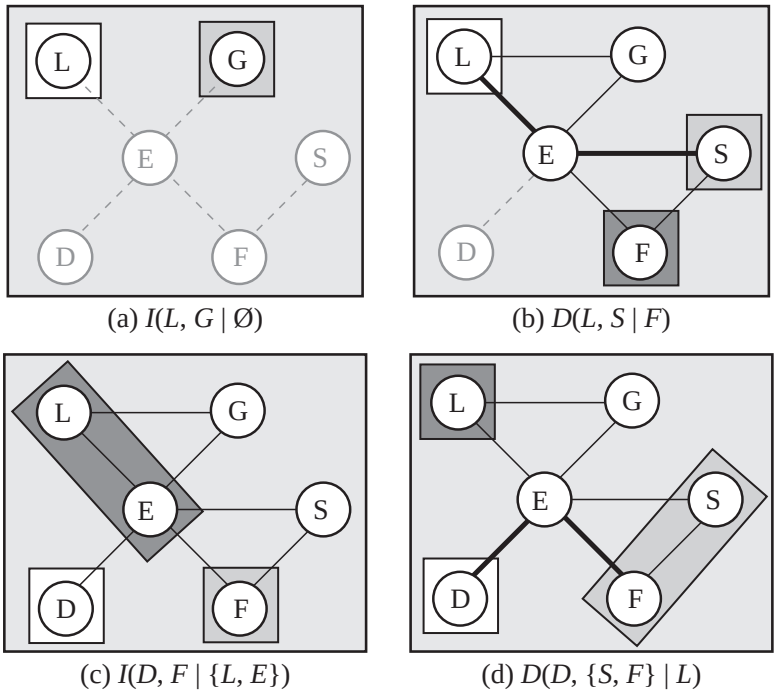


FIGURA 5.4. Ejemplos de ilustración del criterio de D-separación utilizando la Definición 5.4.

se puede obtener el conjunto completo de relaciones de independencia asociado a este modelo probabilístico comprobando cuáles de todas las posibles independencias en $\{X_1, \ldots, X_n\}$ son verificadas por la función $p(x_1, \ldots, x_n)$. Sin embargo, en la práctica, esta función es a menudo desconocida y, por tanto, sólo se dispone de un conjunto de relaciones de independencia que describen las relaciones entre las variables. Este conjunto se denomina *lista inicial de independencias*.

Definición 5.5 Lista inicial. *Una lista inicial de independencias L es un conjunto de relaciones de independencia de la forma $I(X, Y|Z)$, donde X, Y y Z son tres subconjuntos disjuntos de $\{X_1, \ldots, X_n\}$, lo cual significa que X e Y son condicionalmente independientes dado Z.*

Una vez que se dispone de una lista inicial de independencias, es necesario conocer si esta lista implica otras independencias que no estén contenidas en el modelo inicial, pero que tengan que ser satisfechas para que el modelo cumpla una serie de propiedades de independencia condicional conocidas. Esto motiva el siguiente problema:

- **Problema 5.2:** Dada una lista inicial de independencias L, ¿cómo pueden obtenerse nuevas independencias a partir de L utilizando ciertas propiedades de independencia condicional?

En esta sección se introduce un algoritmo para obtener las independencias derivadas de una lista inicial. También se verá que para que una lista de independencias sea compatible con los axiomas de la probabilidad, es necesario que cumpla una serie de propiedades conocidas que permitirán obtener nuevas independencias del modelo. Estas independencias adicionales se denominan *independencias derivadas* y, en caso de que existan, habrán de ser confirmadas por los expertos para que el modelo sea consistente con la realidad. El conjunto completo de independencias (iniciales y derivadas) describe las relaciones existentes entre las variables. Los modelos de dependencia resultantes son conocidos como *modelos definidos por una lista inicial*, y se describen en el Capítulo 7.

A continuación se introducen algunas propiedades de la independencia condicional. Cada uno de los modelos anteriormente descritos (probabilístico, gráfico no dirigido y gráfico dirigido) verifica algunas de estas propiedades, lo que permitirá caracterizarlos parcial o totalmente. Con el fin de ilustrar estas propiedades de forma gráfica, se han utilizado los modelos gráficos no dirigidos mostrados en las Figuras 5.5 y 5.6. En estas figuras cada uno de los tres subconjuntos que intervienen en cada relación de independencia (por ejemplo, $I(X, Y|Z)$) está contenido en un rectángulo. Para distinguir entre los tres subconjuntos, el rectángulo correspondiente al primero de ellos no está sombreado, el correspondiente al segundo muestra una sombra clara, y el correspondiente al tercero (separador), una sombra oscura.

Primeramente se introducen cuatro propiedades que, como se muestra en el apéndice de este capítulo, son satisfechas por cualquier modelo probabilístico. Un análisis más amplio de estas propiedades puede obtenerse, por ejemplo, en Lauritzen (1974) y Dawid (1979, 1980). En el Capítulo 6 (Teoremas 6.1 y 6.8), se describen las propiedades que son satisfechas por los modelos gráficos no dirigidos y dirigidos, respectivamente.

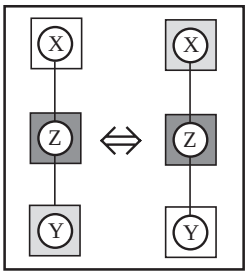
1. **Simetría:** Si X es condicionalmente independiente de Y dada Z, entonces Y es condicionalmente independiente de X dada Z, es decir,

$$I(X, Y|Z) \Leftrightarrow I(Y, X|Z). \tag{5.1}$$

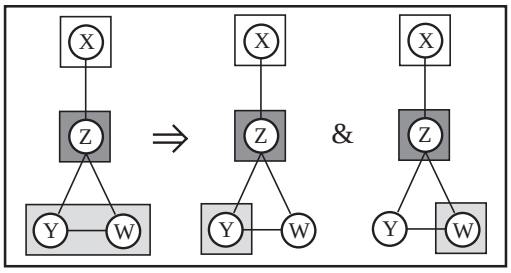
   La Figura 5.5(a) ilustra esta propiedad.

2. **Descomposición:** Si X es condicionalmente independiente de $Y \cup W$ dada Z, entonces X es condicionalmente independiente de Y dada Z, y X es condicionalmente independiente de W dada Z, es decir,

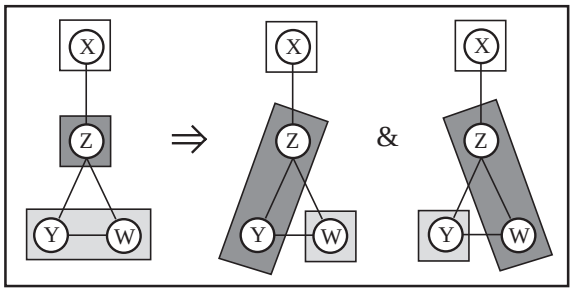
$$I(X, Y \cup W|Z) \Rightarrow I(X, Y|Z) \text{ y } I(X, W|Z), \tag{5.2}$$



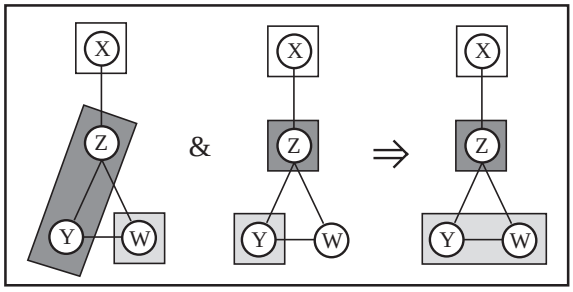
(a) Simetría



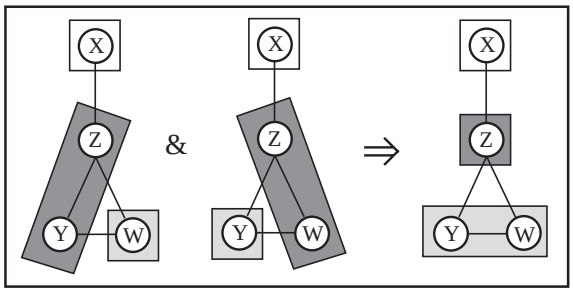
(b) Descomposición



(c) Unión débil



(d) Contracción



(e) Intersección

FIGURA 5.5. Ilustración gráfica de algunas propiedades de independencia condicional: (a) Simetría, (b) Descomposición, (c) Unión débil, (d) Contracción, e (e) Intersección. El conjunto separador se indica con un rectángulo con sombra oscura, y los otros dos subconjuntos con rectángulos sin sombra, y con sombra clara, respectivamente.

Obsérvese que Y y W no tienen por qué ser necesariamente disjuntos.

Esta propiedad se ilustra en la Figura 5.5(b). La implicación recíproca de (5.2) se conoce como *propiedad de composición*. Sin embargo, esta propiedad no se cumple en todos los modelos probabilísticos, como indica el ejemplo siguiente.

Ejemplo 5.5 Violación de la propiedad de composición. Considérese el conjunto de variables binarias $\{X, Y, Z, W\}$. En la Tabla 5.1 se muestran dos funciones de probabilidad distintas para este conjunto de variables. Estas funciones han sido obtenidas fijando valores numéricos para algunos de los parámetros (los indicados con dos cifras decimales en la Tabla 5.1) y calculando los valores restantes para que la función de probabilidad p_1 viole la propiedad de composición, y para que la función p_2 cumpla esta propiedad.

Es fácil comprobar que $p_1(x, y, z, w)$ cumple las relaciones de independencia $I(X, Y|Z)$ y $I(X, W|Z)$ pero, en cambio, no cumple $I(X, Y \cup W|Z)$, lo que prueba que no satisface la propiedad de composición. Puede comprobarse que no existe ninguna combinación de valores de las variables (x, y, z, w) que cumpla la igualdad

$$p(x|y, w, z) = p(x|z).$$

Por el contrario, la función de probabilidad conjunta $p_2(x, y, z, w)$ verifica $I(X, Y|Z)$, $I(X, W|Z)$ y $I(X, Y \cup W|Z)$. Por tanto, esta función de probabilidad cumple la propiedad de composición mientras que $p_1(x, y, z, w)$ no la cumple. ■

3. **Unión Débil:**

$$I(X, Y \cup W|Z) \Rightarrow I(X, W|Z \cup Y) \text{ y } I(X, Y|Z \cup W). \tag{5.3}$$

La Figura 5.5(c) ilustra gráficamente esta propiedad, que refleja el hecho de que el conocimiento de información irrelevante Y no puede hacer que otra información irrelevante W se convierta en relevante.

4. **Contracción:** Si W es irrelevante para X después de conocer alguna información irrelevante Y, entonces W debe haber sido irrelevante antes de conocer Y, es decir,

$$I(X, W|Z \cup Y) \text{ y } I(X, Y|Z) \Rightarrow I(X, Y \cup W|Z). \tag{5.4}$$

La Figura 5.5(d) ilustra gráficamente esta propiedad.

Las propiedades de *unión débil* y *contracción* caracterizan el hecho de que la información irrelevante no debe alterar la relevancia de

x	y	z	w	$p_1(x,y,z,w)$	$p_2(x,y,z,w)$
0	0	0	0	0.012105300	0.0037500
0	0	0	1	0.005263160	0.0050000
0	0	1	0	0.000971795	0.1312200
0	0	1	1	0.024838000	0.1574640
0	1	0	0	0.01	0.0087500
0	1	0	1	0.02	0.01
0	1	1	0	0.03	0.2361960
0	1	1	1	0.04	0.02
1	0	0	0	0.05	0.03
1	0	0	1	0.06	0.04
1	0	1	0	0.07	0.05
1	0	1	1	0.08	0.06
1	1	0	0	0.09	0.07
1	1	0	1	0.10	0.08
1	1	1	0	0.11	0.09
1	1	1	1	0.296822000	0.0076208

TABLA 5.1. Ejemplos de dos funciones de probabilidad conjunta. $p_2(x,y,z,w)$ verifica la propiedad de composición; sin embargo $p_1(x,y,z,w)$ no la verifica.

otra información en el modelo. En otras palabras, la información relevante permanece relevante y la información irrelevante permanece irrelavante.

Cualquier modelo probabilístico cumple las cuatro propiedades anteriores; sin embargo, como se muestra en el apéndice de este capítulo, la propiedad siguiente sólo se cumple si la función de probabilidad es no extrema.

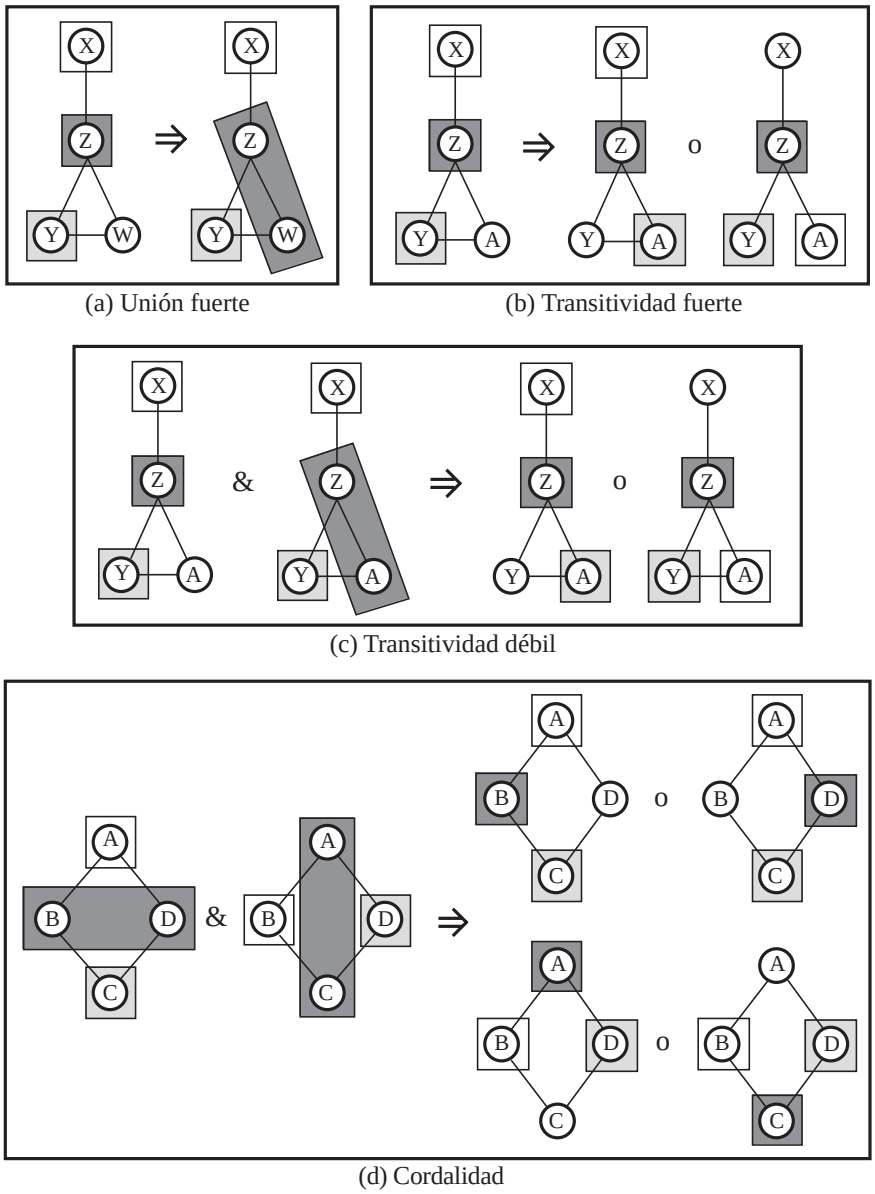
5. **Intersección:**

$$I(X,W|Z\cup Y) \text{ y } I(X,Y|Z\cup W) \Rightarrow I(X,Y\cup W|Z).$$

Esta propiedad se ilustra gráficamente en la Figura 5.5(e) y establece que, a menos que Y afecte a X cuando W es conocida, o que W afecte a X cuando Y es conocida, entonces ni W ni Y, ni su combinación, pueden afectar a X.

Las cuatro propiedades siguientes no son satisfechas, en general, por los modelos probabilísticos pero, como se verá en el Capítulo 7, permitirán caracterizar los modelos gráficos de dependencia.

6. **Unión Fuerte:** Si X es condicionalmente independiente de Y dado Z, entonces X también es condicionalmente independiente de Y dado



(a) Unión fuerte

(b) Transitividad fuerte

(c) Transitividad débil

(d) Cordalidad

FIGURA 5.6. Ilustración gráfica de algunas propiedades de independencia condicional: (a) Unión fuerte, (b) Transitividad fuerte, (c) Transitividad débil, y (d) Cordalidad. El conjunto separador se indica con un rectángulo con sombra oscura, y los otros dos subconjuntos con rectángulos sin sombra, y con sombra clara, respectivamente.

$Z \cup W$, es decir,

$$I(X, Y|Z) \Rightarrow I(X, Y|Z \cup W). \tag{5.5}$$

Esta propiedad se ilustra gráficamente por medio del grafo no dirigido de la Figura 5.6(a). El ejemplo siguiente muestra que, por el contrario, los modelos gráficos dirigidos cumplen esta propiedad.

Ejemplo 5.6 Violación de la propiedad de unión fuerte. Considérese el grafo dirigido acíclico dado en la Figura 5.7(a). Utilizando el criterio de D-separación se puede concluir que el grafo cumple la relación de independencia $I(X, Y|Z)$ (pues existe un único camino entre X e Y en el grafo moral del menor subgrafo ancestral que contiene a X, Y y Z, y este camino contiene al nodo Z). Sin embargo, si se añade el nodo W al conjunto separador, entonces los nodos X e Y resultan dependientes (ver Figura 5.7(b)). Este hecho es debido a que existe un camino entre X e Y que no contiene al nodo Z en el grafo moral del menor subgrafo ancestral que contiene a $\{X, Y, W, Z\}$. Por tanto, se tiene la relación $D(X, Y|\{Z, W\})$, que muestra que los modelos gráficos dirigidos no verifican la propiedad de unión fuerte. ■

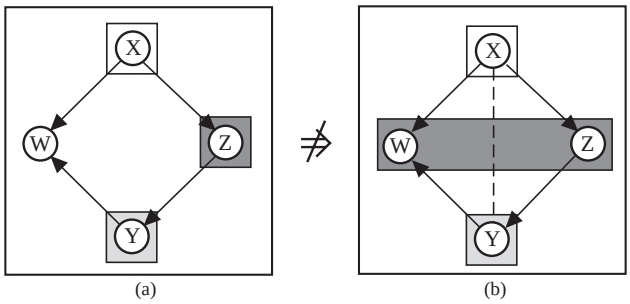


FIGURA 5.7. Ilustración gráfica de que los grafos dirigidos no verifican la propiedad de unión fuerte.

7. **Transitividad Fuerte:** Si X es condicionalmente independiente de A dado Z, y A es condicionalmente dependiente de Y dado Z, entonces X es condicionalmente dependiente de Y dado Z, es decir,

$$D(X, A|Z) \text{ y } D(A, Y|Z) \Rightarrow D(X, Y|Z),$$

o, de forma equivalente,

$$I(X, Y|Z) \Rightarrow I(X, A|Z) \text{ o } I(A, Y|Z), \tag{5.6}$$

donde A es una única variable.

La propiedad de transitividad fuerte afirma que dos variables han de ser dependientes si existe otra variable A que dependa de ambas (ver Figura 5.6(b)).

8. **Transitividad Débil:** Si X y A son condicionalmente dependientes dado Z, e Y y A son condicionalmente dependientes dado Z, entonces X e Y son condicionalmente dependientes dado Z, o X e Y son condicionalmente dependientes dado $Z \cup A$, es decir,

$$D(X, A|Z) \text{ y } D(A, Y|Z) \Rightarrow D(X, Y|Z) \text{ o } D(X, Y|Z \cup A),$$

o, de forma equivalente,

$$I(X, Y|Z) \text{ y } I(X, Y|Z \cup A) \Rightarrow I(X, A|Z) \text{ o } I(A, Y|Z), \qquad (5.7)$$

donde A es una única variable. La Figura 5.6(c) ilustra esta propiedad.

9. **Cordalidad:** Si A y C son condicionalmente dependientes dado B, y A y C son condicionalmente dependientes dado D, entonces A y C son condicionalmente dependientes dado $B \cup D$, o B y D son condicionalmente dependientes dado $A \cup C$, es decir,

$$D(A, C|B) \text{ y } D(A, C|D) \Rightarrow D(A, C|B \cup D) \text{ o } D(B, D|A \cup C),$$

o, de forma equivalente,

$$I(A, C|B \cup D) \text{ y } I(B, D|A \cup C) \Rightarrow I(A, C|B) \text{ o } I(A, C|D), \qquad (5.8)$$

donde A, B, C y D son conjuntos de una única variable. Esta propiedad se ilustra en la Figura 5.6(d).

Antes de concluir esta sección, se muestran las siguientes implicaciones entre las propiedades descritas:

1. Unión fuerte (UF) implica unión débil:

$$I(X, Y \cup W|Z) \overset{UF}{\Rightarrow} I(X, Y \cup W|Z \cup W) \Rightarrow I(X, Y|Z \cup W).$$

2. Transitividad fuerte implica transitividad débil.

3. Unión fuerte e intersección (IN) implican contracción:

$$I(X, Y|Z) \overset{UF}{\Rightarrow} \left.\begin{array}{l} I(X, Y|Z \cup W) \\ I(X, W|Z \cup Y) \end{array}\right\} \overset{IN}{\Rightarrow} I(X, Y \cup W|Z).$$

4. Unión fuerte e intersección también implican composición:

$$\left.\begin{array}{ll} I(X, Y|Z) & \overset{UF}{\Rightarrow} \; I(X, Y|Z \cup W) \\ I(X, W|Z) & \overset{UF}{\Rightarrow} \; I(X, W|Z \cup Y) \end{array}\right\} \overset{IN}{\Rightarrow} I(X, Y \cup W|Z).$$

Las propiedades anteriores se utilizarán en la sección siguiente para concluir independencias adicionales a partir de algunas listas de independencias particulares, que verifican ciertas propiedades y permitirán caracterizar las estructuras de dependencia e independencia contenidas en los modelos probabilísticos y gráficos.

5.4 Modelos de Dependencia

Ahora que ya han sido introducidas algunas propiedades de la independencia condicional es posible analizar el Problema 5.2:

- **Pregunta 5.2:** Dada una lista inicial de relaciones de independencia L, ¿cómo pueden obtenerse nuevas independencias a partir de L utilizando ciertas propiedades de independencia condicional?

Obsérvese que, hasta ahora, no se ha requerido que las listas de relaciones de independencia cumpliesen ninguna condición (sólo que los subconjuntos que componen cada relación sean disjuntos). Cuando se impone alguna condición a estos modelos como, por ejemplo, que cumplan un cierto conjunto de propiedades de independencia condicional, se obtienen algunos tipos especiales de listas de independencias, algunos de los cuales se describen a continuación.

Definición 5.6 Grafoide. *Un grafoide es un conjunto de relaciones de independencia que es cerrado con respecto a las propiedades de simetría, descomposición, unión débil, contracción e intersección.*

Definición 5.7 Semigrafoide. *Un semigrafoide es un conjunto de relaciones de independencia que es cerrado con respecto a las propiedades de simetría, descomposición, unión débil y contracción.*

Por tanto, un grafoide debe satisfacer las cinco primeras propiedades, mientras que un semigrafoide debe satisfacer sólo las cuatro primeras (ver Pearl y Paz (1987) y Geiger (1990)).

Dada una lista inicial de independencias, un grafo, o una función de probabilidad conjunta, siempre es posible determinar qué relaciones de independencia se cumplen en el modelo y, por tanto, determinar su estructura cualitativa. Por tanto, estos tipos de modelos definen clases particulares de los denominados *modelos de dependencia*.

Definición 5.8 Modelo de Dependencia. *Cualquier modelo M de un conjunto de variables $\{X_1, \ldots, X_n\}$ mediante el cual se pueda determinar si la relación $I(X, Y|Z)$ es o no cierta, para todas las posibles ternas de subconjuntos X, Y y Z, se denomina modelo de dependencia.*

Definición 5.9 Modelo de dependencia probabilístico. *Un modelo de dependencia M se denomina probabilístico si contiene todas las relaciones de independencia dadas por una función de probabilidad conjunta $p(x_1, \ldots, x_n)$.*

Definición 5.10 Modelo de dependencia probabilístico no extremo. *Un modelo de dependencia probabilístico no extremo es un modelo de dependencia probabilístico obtenido de una función de probabilidad no extrema, o positiva; es decir, $p(x_1, \ldots, x_n)$ toma valores en el intervalo abierto $(0, 1)$.*

Dado que todas las funciones de probabilidad satisfacen las cuatro primeras propiedades de independencia condicional, todos los modelos de dependencia probabilísticos son semigrafoides. Por otra parte, dado que sólo las funciones de probabilidad no extremas satisfacen la propiedad de intersección, sólo los modelos de dependencia probabilísticos no extremos son grafoides.

Definición 5.11 Modelo de dependencia compatible con una probabilidad. *Un modelo de dependencia M se dice compatible con una función de probabilidad $p(x_1, \ldots, x_n)$ si todas las relaciones de independencia derivadas M son también satisfechas por $p(x_1, \ldots, x_n)$.*

Obsérvese que un modelo de dependencia compatible con una probabilidad es aquel que puede obtenerse de una función de probabilidad conjunta $p(x_1, \ldots, x_n)$, pero sin necesidad de ser completo, es decir, no tienen por qué contener todas las relaciones de independencia que pueden obtenerse de $p(x_1, \ldots, x_n)$.

Dado que toda función de probabilidad cumple las cuatro primeras propiedades de la independencia condicional, si un modelo de dependencia M es compatible con una función de probabilidad $p(x_1, \ldots, x_n)$, entonces el menor semigrafoide generado por M también debe ser compatible con $p(x_1, \ldots, x_n)$. Por tanto, un problema interesante desde el punto de vista práctico es calcular el menor semigrafoide generado por un modelo de dependencia M. El siguiente algoritmo puede ser utilizado con este fin:

Algoritmo 5.2 Generando un grafoide mínimo.

- **Datos:** Un modelo de dependencia inicial M.
- **Resultado:** El mínimo grafoide que contiene a M.

1. Generar nuevas relaciones de independencia aplicando las propiedades de simetría, descomposición, unión débil, contracción e intersección a las relaciones del modelo M. El conjunto resultante es el grafoide buscado. ∎

El algoritmo anterior también puede ser utilizado para generar un semigrafoide; para ello basta con no utilizar la propiedad de intersección. El ejemplo siguiente ilustra este algoritmo.

Ejemplo 5.7 Generando grafoides. Supóngase que se tiene un conjunto de cuatro variables $\{X_1, X_2, X_3, X_4\}$ y que se da la siguiente lista de relaciones de independencia:

$$M = \{I(X_1, X_2|X_3),\ I(X_1, X_4|X_2),\ I(X_1, X_4|\{X_2, X_3\})\}. \tag{5.9}$$

La Tabla 5.2 muestra las relaciones de independencia iniciales, y las relaciones derivadas necesarias para completar el modelo hasta convertirlo en un semigrafoide y un grafoide, respectivamente. Las nuevas relaciones de independencia son generadas utilizando un programa de ordenador llamado *X-pert Maps*,[2] que implementa el Algoritmo 5.2. La Tabla 5.2 también muestra las relaciones de independencia que se utilizan para obtener las nuevas independencias. ■

Por tanto, las cinco primeras propiedades pueden ser utilizadas para aumentar un modelo de dependencia M compatible con una función de probabilidad $p(x_1, \ldots, x_n)$. Tanto el modelo inicial como el completado son compatibles con $p(x_1, \ldots, x_n)$. Esto motiva el siguiente problema:

- **Pregunta 5.3.** ¿Constituyen las cuatro propiedades descritas anteriormente una caracterización completa de los modelos probabilísticos?

Pearl y Paz (1987) (ver Pearl, (1988) p. 88) conjeturaron que las primeras cuatro propiedades (simetría, descomposición, unión débil, y contracción) eran completas. Sin embargo, esta conjetura fue refutada por Studený (1989) encontrando, primeramente, un propiedad que no puede derivarse de las cuatro anteriores y mostrando, después, que no existe ningún conjunto completo de propiedades que caractericen los modelos probabilísticos (Studený (1992)).

Como se verá en los capítulos siguientes, la estructura cualitativa de un modelo probabilístico puede ser representada mediante un modelo de dependencia que permitirá obtener una factorización de la función de probabilidad. En la sección siguiente se introducen algunos conceptos sobre factorizaciones de una función de probabilidad.

5.5 Factorizaciones de una Función de Probabilidad

Cualquier función de probabilidad de un conjunto de variables aleatorias puede ser definida por medio de funciones de probabilidad condicionada más sencillas formando una factorización. En esta sección se analizan distintas formas de factorizar una función de probabilidad.

---

[2] El programa *X-Pert Maps* puede obtenerse en la dirección WWW http://ccaix3.unican.es/~AIGroup.

Lista inicial
$M = \{I(X_1, X_2\|X_3),\ I(X_1, X_4\|X_2),\ I(X_1, X_4\|X_2X_3)\}$

RIC adicionales para Semigrafoide		
Propiedad	RIC Derivadas	Derivada de
Simetría	$I(X_2, X_1\|X_3)$	$I(X_1, X_2\|X_3)$
Simetría	$I(X_4, X_1\|X_2)$	$I(X_1, X_4\|X_2)$
Simetría	$I(X_4, X_1\|X_2X_3)$	$I(X_1, X_4\|X_2X_3)$
Contracción	$I(X_1, X_2X_4\|X_3)$	$I(X_1, X_2\|X_3)$ y $I(X_1, X_4\|X_2X_3)$
Simetría	$I(X_2X_4, X_1\|X_3)$	$I(X_1, X_2X_4\|X_3)$
Unión Débil	$I(X_1, X_2\|X_3X_4)$	$I(X_1, X_2X_4\|X_3)$
Simetría	$I(X_2, X_1\|X_3X_4)$	$I(X_1, X_2\|X_3X_4)$
Descomposición	$I(X_1, X_4\|X_3)$	$I(X_1, X_2X_4\|X_3)$
Simetría	$I(X_4, X_1\|X_3)$	$I(X_1, X_4\|X_3)$

RIC adicionales para Grafoide		
Propiedad	RIC Derivadas	Derivada de
Intersección	$I(X_1, X_2X_4\|\phi)$	$I(X_1, X_2\|X_3X_4)$ y $I(X_1, X_4\|X_2)$
Simetría	$I(X_2X_4, X_1\|\phi)$	$I(X_1, X_2X_4\|\phi)$
Descomposición	$I(X_1, X_2\|\phi)$	$I(X_1, X_2X_4\|\phi)$
Simetría	$I(X_2, X_1\|\phi)$	$I(X_1, X_2\|\phi)$
Unión Débil	$I(X_1, X_2\|X_4)$	$I(X_1, X_2X_4\|\phi)$
Simetría	$I(X_2, X_1\|X_4)$	$I(X_1, X_2\|X_4)$
Descomposición	$I(X_1, X_4\|\phi)$	$I(X_1, X_2X_4\|\phi)$
Simetría	$I(X_4, X_1\|\phi)$	$I(X_1, X_4\|\phi)$

TABLA 5.2. Mínimos semigrafoide y grafoide generados por la lista inicial M de relaciones de independencia condicional (RIC) en (5.9), obtenidos utilizando el Algoritmo 5.2.

Definición 5.12 Factorización mediante funciones potenciales. *Sean $C_1, \ldots, C_m$ subconjuntos de un conjunto de variables $X = \{X_1, \ldots, X_n\}$. Si la función de probabilidad conjunta de X puede ser escrita como producto de m funciones no negativas Ψ_i $(i = 1, \ldots, m)$, es decir,*

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{m} \Psi_i(c_i), \tag{5.10}$$

donde c_i es una realización de C_i, entonces se dice que (5.10) es una factorización de la función de probabilidad. Las funciones Ψ_i se denominan factores potenciales de la función de probabilidad.

En el Capítulo 6 se verán ejemplos importantes de este tipo de factorización. Obsérvese que los conjuntos $C_1, \ldots, C_m$ no son necesariamente disjuntos

y que las funciones Ψ_i no son necesariamente funciones de probabilidad. Cuando se exige que las funciones Ψ_i sean funciones de probabilidad, se obtienen factorizaciones particulares, algunas de las cuales se comentan a continuación.

Sea $\{Y_1, \ldots, Y_m\}$ una partición (subconjuntos disjuntos dos a dos cuya unión es el conjunto total) del conjunto $\{X_1, \ldots, X_n\}$. Un tipo importante de factorizaciones se obtiene aplicando la fórmula siguiente, conocida como *regla de la cadena*.

Definición 5.13 Regla de la cadena. *Cualquier función de probabilidad de un conjunto de variables $\{X_1, \ldots, X_n\}$ puede ser expresada como el producto de m funciones de probabilidad condicionada de la forma*

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{m} p(y_i|b_i), \tag{5.11}$$

o, de modo equivalente,

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{m} p(y_i|a_i), \tag{5.12}$$

donde $B_i = \{Y_1, \ldots, Y_{i-1}\}$ es el conjunto de variables anteriores a Y_i y $A_i = \{Y_{i+1}, \ldots, Y_n\}$ es el conjunto de variables posteriores a Y_i. Obsérvese que a_i y b_i son realizaciones de A_i y B_i, respectivamente.

Cuando los conjuntos Y_i están formados por una única variable, entonces se tiene $m = n$ y el conjunto $\{Y_1, \ldots, Y_n\}$ es simplemente una permutación de $\{X_1, \ldots, X_n\}$. En este caso, (5.11) y (5.12) se denominan reglas canónicas de la cadena y se tiene

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(y_i|b_i) \tag{5.13}$$

y

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(y_i|a_i), \tag{5.14}$$

respectivamente.

Ejemplo 5.8 Regla de la cadena. Considérese el conjunto de variables $\{X_1, \ldots, X_4\}$ y la partición $Y_1 = \{X_1\}$, $Y_2 = \{X_2\}$, $Y_3 = \{X_3\}$, $Y_4 = \{X_4\}$. Entonces (5.13) y (5.14) proporcionan la siguientes factorizaciones equivalentes de la función de probabilidad:

$$p(x_1, \ldots, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3) \tag{5.15}$$

y

$$p(x_1, \ldots, x_4) = p(x_1|x_2, x_3, x_4)p(x_2|x_3, x_4)p(x_3|x_4)p(x_4). \tag{5.16}$$

Por tanto, la función de probabilidad puede expresarse como el producto de cuatro funciones de probabilidad condicionada. Nótese que existen varias formas de aplicar la regla de la cadena a una misma función de probabilidad (considerando distintas particiones), lo que origina distintas factorizaciones. Por ejemplo, a continuación se muestran dos factorizaciones equivalentes obtenidas aplicando la regla de la cadena a distintas particiones de $\{X_1, \ldots, X_4\}$:

- La partición $Y_1 = \{X_1\}$, $Y_2 = \{X_2, X_3\}$, y $Y_3 = \{X_4\}$ da lugar a

$$p(x_1, \ldots, x_4) = p(x_1)p(x_2, x_3|x_1)p(x_4|x_1, x_2, x_3).$$

- La partición $Y_1 = \{X_1, X_4\}$ y $Y_2 = \{X_2, X_3\}$ produce la factorización

$$p(x_1, \ldots, x_4) = p(x_1, x_4)p(x_2, x_3|x_1, x_4).$$

■

En la Sección 3.5 se ha visto que el número de parámetros que definen un modelo probabilístico puede ser reducido imponiendo ciertas restricciones. Por ejemplo, los distintos modelos presentados en la Sección 3.5 fueron obtenidos suponiendo ciertas relaciones de independencia condicional para el modelo. Con el fin de ilustrar la forma en la que la inclusión de una relación de independencia en un modelo probabilístico da lugar a una reducción de parámetros en el modelo, es conveniente escribir la función de probabilidad conjunta como producto de funciones de probabilidad condicionada utilizando, por ejemplo, la regla de la cadena. Este hecho se ilustra en el siguiente ejemplo.

Ejemplo 5.9 Restricciones dadas por independencias. Considérese el conjunto de variables dado en el Ejemplo 5.8 y supóngase que un experto propone las dos siguientes relaciones de independencia:

$$I(X_3, X_1|X_2) \text{ y } I(X_4, \{X_1, X_3\}|X_2). \tag{5.17}$$

A fin de incluir estas relaciones en el modelo probabilístico, interesa calcular las restricciones que deben cumplir los parámetros del modelo para satisfacer estas condiciones de independencia. La primera de estas relaciones implica

$$p(x_3|x_1, x_2) = p(x_3|x_2), \tag{5.18}$$

mientras que la segunda implica

$$p(x_4|x_1, x_2, x_3) = p(x_4|x_2). \tag{5.19}$$

Obsérvese que la forma general del modelo probabilístico no es una forma conveniente para calcular las restricciones entre los parámetros, dadas por (5.18) y (5.19). Sin embargo, si se sustituyen estas dos igualdades en la

factorización del modelo probabilístico (5.15), se obtiene la siguiente estructura

$$p(x_1, \ldots, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_2). \tag{5.20}$$

Suponiendo que las variables son binarias, la función de probabilidad en (5.15) depende de $2^4 - 1 = 15$ parámetros libres.[3] Por otra parte, la función de probabilidad en (5.20) depende de siete parámetros ($p(x_1)$ depende de un parámetro, y cada una de las restantes funciones de probabilidad condicionada depende de dos parámetros). Por tanto, las dos relaciones de independencia dadas en (5.17) dan lugar a una reducción de 8 parámetros en el modelo probabilístico. ∎

Definición 5.14 Función de probabilidad condicionada canónica. *Sea $U_i \subset X = \{X_1, \ldots, X_n\}$. Una función de probabilidad condicionada $p(x_i|u_i)$ se dice canónica si X_i está formada por una única variable que no está contenida en U_i.*

El siguiente teorema, probado por Gelman y Speed (1993), garantiza que cada conjunto de funciones de probabilidad condicionada, dado en forma no canónica, tiene asociado un conjunto canónico equivalente.

Teorema 5.1 Existencia de formas canónicas. *Considérese el conjunto de variables $X = \{X_1, \ldots, X_n\}$ y supónganse las funciones de probabilidad marginales y condicionadas $P = \{p(u_1|v_1), \ldots, p(u_m|v_m)\}$, donde U_i y V_i son subconjuntos disjuntos de X, tal que $U_i \neq \phi$ y V_i puede ser vacío (para el caso de funciones marginales). Entonces, a partir de P puede obtenerse un conjunto equivalente en el que los nuevos conjuntos U_i contienen una única variable de X.*

Demostración: Aplicando la regla de la cadena a $p(u_i|v_i)$ pueden obtenerse tantas nuevas funciones condicionadas canónicas como variables contenga el conjunto U_i, es decir, el conjunto

$$\{p(x_j|c_{ij}, v_i) \;\; \forall X_j \in U_i\}, \tag{5.21}$$

donde $C_{ij} = \{X_r \,|\, X_r \subset U_i,\ r < j\}$. ∎

El algoritmo siguiente convierte un conjunto dado de funciones condicionadas P en una representación canónica equivalente.

Algoritmo 5.3 Forma canónica.

- **Datos:** Un conjunto $P = \{p(u_i|v_i),\ i = 1, \ldots, m\}$ de m funciones de probabilidad condicionada, donde U_i y V_i son subconjuntos disjuntos de X.

[3] Realmente existen 16 parámetros, pero la suma de todos ha de ser 1. Por tanto, existen únicamente 15 parámetros libres.

- **Resultado:** Un conjunto equivalente P^* en forma canónica.

1. Iniciación: Considerar $P^* = \phi$ e $i = 1$.
2. Asignar $j = 1$, $S_i = U_i \cup V_i$ y $L = card(U_i)$.
3. Eliminar de S_i una de las variables contenidas en U_i, por ejemplo X_ℓ, y añadir $p(x_\ell|s_i)$ a P^*.
4. Si $j < L$, incrementar el índice j en una unidad e ir a la Etapa 3; en caso contrario, ir a la Etapa 5.
5. Si $i < m$, incrementar el índice i en una unidad e ir a la Etapa 2; en caso contrario, devolver P^* como resultado. ∎

Ejemplo 5.10 Supóngase el conjunto de variables $X = \{A, B, C, D\}$ y el conjunto de funciones de probabilidad $P = \{p(a, b|c), p(a, c, d|b)\}$. Utilizando la notación del Algoritmo 5.3, los conjuntos U_i y V_i son

$$\begin{array}{ll} U_1 = \{A, B\}, & V_1 = \{C\}, \\ U_2 = \{A, C, D\}, & V_2 = \{B\}. \end{array}$$

Para convertir las dos funciones de probabilidad condicionada de P en sus formas canónicas correspondientes, se utiliza el Algoritmo 5.3 obteniéndose

$$\begin{array}{l} p(a, b|c) = p(a|b, c)p(b|c), \\ p(a, c, d|b) = p(a|c, d, b)p(c|d, b)p(d|b). \end{array} \tag{5.22}$$

Por tanto, se obtiene la representación canónica

$$P^* = \{p(a|b, c)p(b|c); p(a|c, d, b)p(c|d, b)p(d|b)\}. \tag{5.23}$$

La Figura 5.8 muestra un programa de *Mathematica* para convertir el conjunto dado P en forma canónica. Dada una lista de pares $\{U, V\}$, el programa devuelve la lista canónica asociada. Por ejemplo, dadas las funciones de probabilidad en (5.22) los siguientes comandos de *Mathematica* permiten obtener la forma canónica correspondiente mostrada en (5.23):

```
In:=Canonical[List[{{A,B},{C}},{{A,C,D},{B}}]]

Out:=List[{{A},{B,C}},{{B},{C}},{{A},{C,D,B}},
          {{C},{D,B}},{{D},{B}}]
```

∎

Definición 5.15 Probabilidad condicionada canónica estándar. *Sea $\{Y_1, \ldots, Y_n\}$ una permutación del conjunto $X = \{X_1, \ldots, X_n\}$. Una función de probabilidad condicionada $p(y_i|s_i)$ se dice que es una función de probabilidad condicionada en forma canónica estándar si Y_i está formado por una única variable y S_i contiene todas las variables anteriores a Y_i, o todas las variables posteriores a Y_i, es decir, o bien $S_i = \{Y_1, ..., Y_{i-1}\}$, o bien, $S_i = \{Y_{i+1}, ..., Y_n\}$.*

```
Canonical[P_List]:= Module[{U,V,S,l,PCan},
    PCan={};
    Do[U=P[[i,1]]; (* Primer elemento del par i-ésimo *)
        V=P[[i,2]];
        S=Join[U,V];
        l=Length[U];
        Do[S=Drop[S,1]; (* Elimina el último elemento *)
            AppendTo[PCan,{{U[[j]]},S}]
        ,{j,1,l}]
    ,{i,1,Length[P]}];
    Return[PCan]
]
```

FIGURA 5.8. Programa de *Mathematica* para convertir un conjunto dado P de funciones de probabilidad condicionada a forma canónica.

Por ejemplo, dada la permutación $Y = \{Y_1, Y_2, Y_3, Y_4\}$, las funciones de probabilidad $p(y_1)$ y $p(y_3|y_1, y_2)$ son probabilidades condicionadas en forma canónica estándar; sin embargo, $p(y_2|y_1, y_3)$ y $p(y_1|y_3, y_4)$ son canónicas pero no están en forma estándar.

Definición 5.16 Representación canónica estándar de una probabilidad. *Sea $\{Y_1, \ldots, Y_n\}$ una permutación del conjunto de variables $X = \{X_1, \ldots, X_n\}$. Entonces la función de probabilidad $p(x)$ puede expresarse como el producto de n funciones de probabilidad condicionada en forma canónica estándar de la forma siguiente*

$$p(x) = \prod_{i=1}^{n} p(y_i|b_i), \tag{5.24}$$

donde $B_i = \{Y_1, \ldots, Y_{i-1}\}$ o, de forma equivalente,

$$p(x) = \prod_{i=1}^{n} p(y_i|a_i), \tag{5.25}$$

donde $A_i = \{Y_{i+1}, \ldots, Y_n\}$. Las ecuaciones (5.24) y (5.25) se denominan representaciones canónicas estándar de la probabilidad. Los términos $p(y_i|b_i)$ y $p(y_i|a_i)$ se denominan componentes canónicas estándar.

Por ejemplo, (5.24) y (5.25) corresponden a dos representaciones canónicas estándar de $p(x_1, \ldots, x_4)$. Las formas canónicas estándar no son únicas, al igual que las formas canónicas, pues pueden obtenerse distintas representaciones aplicando la regla de la cadena a distintas permutaciones de X.

Las consecuencias prácticas de la existencia de una representación canónica para cualquier conjunto P de funciones de probabilidad condicionada son

1. Cualquier conjunto no canónico de funciones de probabilidad condicionada P puede ser expresado en forma canónica de forma equivalente.

2. Cualquier función de probabilidad puede ser factorizada, utilizando la regla de la cadena, como un producto de funciones de probabilidad condicionada en forma canónica estándar.

3. Sólo es necesario considerar funciones de probabilidad condicionada de una única variable para definir la función de probabilidad de un conjunto de variables.

   Las principales ventajas de este tipo de representaciones son las siguientes:

   - La definición de un modelo probabilístico se simplifica enormemente al tratar con funciones de probabilidad condicionada de una única variable (dado un conjunto de variables). Este proceso es más sencillo que la especificación directa de una función de probabilidad pues, generalmente, las funciones de probabilidad condicionada dependen de muchas menos variables que la función de probabilidad conjunta.
   - La programación de algoritmos también se simplifica ya que sólo es necesario considerar un único modelo genérico para las funciones de probabilidad condicionada.

4. Las formas canónicas estándar permiten identificar fácilmente aquellos conjuntos de funciones de probabilidad condicionada que son consistentes con algún modelo probabilístico. También permiten determinar cuándo es único el modelo probabilístico definido (ver el Capítulo 7).

5.6 Construcción de un Modelo Probabilístico

El problema de construir una función de probabilidad para un conjunto de variables puede simplificarse notablemente considerando una factorización de la probabilidad como producto de funciones de probabilidad condicionada más sencillas. El grado de simplificación dependerá de la estructura de independencia (incondicional o condicional) existente entre las variables del modelo. Por tanto, para encontrar una factorización apropiada del modelo probabilístico, primero se necesita conocer su estructura de independencia. Esta estructura de independencia (modelo de dependencia) caracteriza la

estructura cualitativa de las relaciones entre las variables. Por ejemplo, se necesita definir qué variables son independientes y/o condicionalmente independientes de otras y cuáles no. La estructura de independencia y, por tanto, la factorización asociada al modelo probabilístico, puede ser obtenida de varias formas:

1. **Modelos definidos gráficamente:** Como se ha visto en las secciones anteriores, las relaciones existentes entre las variables de un conjunto pueden ser descritas mediante un grafo. Posteriormente, utilizando un criterio de separación apropiado, se puede obtener el conjunto de relaciones de independencia asociado. Estos modelos de dependencia se conocen como *modelos definidos gráficamente*, y tienen como ejemplos más importantes a las *redes de Markov*, y las *redes Bayesianas*, que se analizan en detalle en los Capítulos 6 y 7. Las tareas de comprobar la validez de un grafo, entender sus implicaciones, y modificarlo de forma apropiada han de ser realizadas partiendo de la comprensión de las relaciones de dependencia e independencia existentes en el conjunto de variables.

2. **Modelos definidos por listas de independencias:** Los grafos son herramientas muy útiles para definir la estructura de independencia de un modelo probabilístico. El problema de los modelos gráficos es que no todas las funciones de probabilidad pueden ser representadas mediante estos modelos (ver Sección 6.2). Una descripción alternativa a los modelos gráficos consiste en utilizar directamente un conjunto M de relaciones de independencia que describan las relaciones entre las variables. Este conjunto puede ser definido por un experto a partir de sus opiniones sobre las relaciones entre las variables del modelo. Cada una de las independencias del conjunto indica qué variables contienen información relevante sobre otras y cuándo el conocimiento de algunas variables hace que otras sean irrelevantes para un conjunto de variables dado. Este conjunto inicial de independencias puede ser completado incluyendo aquellas otras que cumplan una serie de propiedades de independencia condicional. El conjunto resultante puede ser finalmente utilizado para obtener una factorización de la función de probabilidad del modelo. Los modelos resultantes se conocen como *modelos definidos por listas de relaciones de independencia*. El Capítulo 7 presenta un análisis detallado de estos modelos.

3. **Modelos definidos condicionalmente:** Como alternativa a los modelos gráficos y los modelos dados por listas de relaciones de independencia, la estructura cualitativa de un modelo probabilístico puede venir dada por un conjunto de funciones de probabilidad marginales y condicionadas

$$P = \{p_1(u_1|v_1), \ldots, p_m(u_m|v_m)\}.$$

Sin embargo, las funciones de este conjunto no pueden definirse libremente, sino que han de satisfacer ciertas relaciones para ser compatibles y definir un único modelo probabilístico. En el Capítulo 7 se analiza detalladamente la forma de comprobar la compatibilidad, la unicidad, y de obtener la función de probabilidad asociada a un conjunto de probabilidades marginales y condicionadas.

Una ventaja de utilizar modelos gráficos, o modelos definidos por listas de independencias, para construir un modelo probabilístico es que éstos modelos definen una factorización de la función de probabilidad como producto de funciones de probabilidad condicionada que determinan la estructura cualitativa del modelo probabilístico. Normalmente, estas funciones condicionadas contienen un número menor de variables que la función de probabilidad conjunta y, por tanto, el proceso de definición del modelo probabilístico es más sencillo. Esta técnica de romper ("de dividir y conquistar") la función de probabilidad como producto de funciones condicionadas más sencillas se analiza en los Capítulos 6 y 7.

Una vez que se conoce la estructura cualitativa del modelo probabilístico (la factorización de la función de probabilidad), la estructura cuantitativa de un modelo particular se define mediante la asignación de valores numéricos a los parámetros asociados a las funciones de probabilidad condicionada que intervienen en la factorización del modelo. Estos valores han de ser definidos por algún experto, o estimados a partir de un conjunto de datos.

Por tanto, si la estructura cualitativa del modelo es desconocida, que es el caso habitual en la práctica, entonces tanto la estructura cualitativa, como la cuantitativa (los parámetros) han de ser estimadas a partir del conjunto de datos disponible (una base de datos, etc.). Este problema, que se conoce como *aprendizaje*, se trata en detalle en el Capítulo 11.

Como resumen de todo lo anterior, la construcción de un modelo probabilístico puede ser realizada en dos etapas:

1. Factorizar la función de probabilidad mediante un producto de funciones de probabilidad condicionada. Esta factorización puede obtenerse de tres formas distintas:
   (a) Utilizando grafos (ver Capítulo 6).
   (b) Utilizando listas de relaciones de independencia (ver Capítulo 7).
   (c) A partir de un conjunto de funciones de probabilidad condicionada (Capítulo 7).
2. Estimar los parámetros de cada una de las funciones de probabilidad condicionada resultantes.

Este proceso se ilustra de modo esquemático en la Figura 5.9. En este diagrama, una línea continua de un rectángulo A a un rectángulo B significa

que cada miembro de A es también un miembro de B, mientras que una línea discontinua significa que algunos, pero no necesariamente todos, los miembros de A son miembros de B. El camino más simple para definir un modelo probabilístico es comenzar con un grafo que se supone describe la estructura de dependencia e independencia de las variables. A continuación, el grafo puede utilizarse para construir una factorización de la función de probabilidad de las variables. De forma alternativa, también puede comenzarse con una lista de relaciones de independencia y, a partir de ella, obtener una factorización de la función de probabilidad. La factorización obtenida determina los parámetros necesarios para definir el modelo probabilístico. Una vez que estos parámetros han sido definidos, o estimados a partir de un conjunto de datos, la función de probabilidad que define el modelo probabilístico vendrá dada como el producto de las funciones de probabilidad condicionada resultantes.

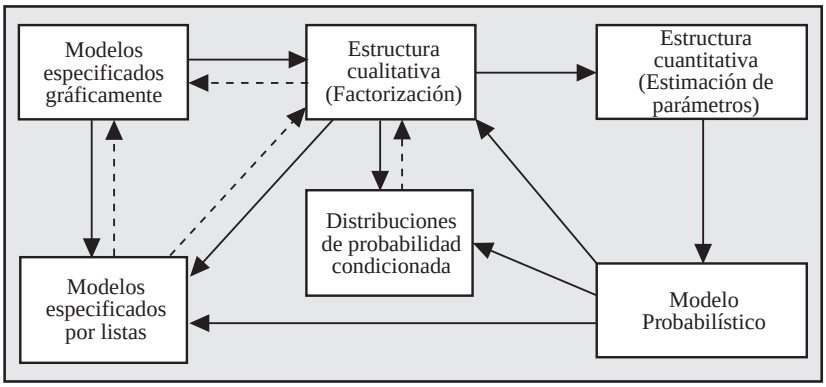


FIGURA 5.9. Diagrama mostrando las formas alternativas de definir un modelo probabilístico.

Por otra parte, si se conoce la función de probabilidad que define un modelo probabilístico (que no es el caso habitual en la práctica), se puede seguir el camino inverso y obtener varias factorizaciones distintas (utilizando la regla de la cadena definida en la Sección 5.5). También se puede obtener la lista de independencias correspondiente al modelo comprobando cuáles de todas las posibles relaciones de independencia de las variables son verificadas por la función de probabilidad. A partir del conjunto de independencias obtenido, también puede construirse una factorización de la familia paramétrica que contiene a la función de probabilidad dada.

Este proceso de construcción de modelos probabilísticos plantea los siguientes problemas.

- **Problema 5.4:** ¿Puede representarse cualquier lista de relaciones de independencia mediante un grafo de forma que las independencias que se deriven del grafo coincidan con las de la lista dada?

Aunque un grafo puede ser representado de forma equivalente por una lista de relaciones de independencia, el recíproco no siempre es cierto. Por esta razón, la Figura 5.9 muestra una arista continua que va del rectángulo que representa a los modelos definidos gráficamente al rectángulo que representa a los modelos definidos por listas de relaciones de independencia, y una arista discontinua en la dirección opuesta. El Capítulo 6 analiza en mayor detalle este hecho, tanto en el caso de grafos dirigidos, como en el caso de grafos no dirigidos.

- **Problema 5.5:** ¿Cómo puede obtenerse la función de probabilidad que contiene las independencias asociadas a un grafo dirigido o no dirigido?
- **Problema 5.6:** ¿Cómo puede obtenerse la función de probabilidad que contiene las independencias de una lista de relaciones de independencia?

Estos dos problemas se analizan en los Capítulos 6 y 7.

Desgraciadamente, los grafos no siempre pueden reproducir las independencias condicionales contenidas en una lista arbitraria de relaciones de independencia, o en un modelo probabilístico. Por tanto, es importante caracterizar las clases de modelos probabilísticos que pueden representarse mediante grafos. Esto plantea los siguientes problemas:

- **Problema 5.7:** ¿Cuál es la clase de modelos probabilísticos que pueden representarse por medio de grafos?
- **Problema 5.8:** ¿Qué listas de relaciones de independencia pueden ser representadas por medio de grafos?
- **Problema 5.9:** ¿Cuál es el conjunto de funciones de probabilidad condicionadas necesario para definir un modelo probabilístico y cuáles son los parámetros necesarios para cuantificarlo?

Estos problemas se analizan en detalle en los Capítulos 6 y 7. En estos capítulos se verá que, aunque todo grafo define una estructura cualitativa de un modelo probabilístico (a través de una factorización), no todas las estructuras cualitativas pueden ser representadas por medio de grafos. Por tanto, la Figura 5.9 muestra una arista sólida que va de los modelos definidos gráficamente a los modelos factorizados, y una arista discontinua en la dirección opuesta. De forma similar, se verá que todo modelo probabilístico define una lista de relaciones de independencia, pero no cualquier lista de independencias define un modelo probabilístico. Este hecho se ilustra en la Figura 5.9 con las correspondientes aristas continua y discontinua.

De la discusión anterior, y de la Figura 5.9, puede concluirse que existen tres formas fundamentales de construir un modelo probabilístico:

- Grafo → Modelos factorizados → Estimación de parámetros → Modelo probabilístico.
- Listas de relaciones de independencia → Modelos factorizados → Estimación de parámetros → Modelo probabilístico.
- Conjunto de funciones condicionadas → Modelos factorizados → Estimación de parámetros → Modelo probabilístico.

En los Capítulos 6 y 7 se verá que la forma más sencilla es comenzar con un grafo, pero que la forma más general es a partir de una lista de relaciones de independencia.

Apéndice al Capítulo 5

En este apéndice se demuestran algunas de las propiedades de independencia condicional que cumplen las funciones de probabilidad. Se demuestra que cualquier función de probabilidad verifica las cuatro primeras propiedades, pero que sólo las probabilidades no extremas verifican la última.

5.7.1 Demostración de la Propiedad de Simetría

Dado que la función de probabilidad $p(x, y, z)$ cumple $I(X, Y|Z)$, se tiene

$$p(x|y, z) = p(x|z) \Leftrightarrow p(x, y|z) = p(x|z)p(y|z). \tag{5.26}$$

Veamos ahora que también se cumple $I(Y, X|Z)$. Suponiendo que $p(x, z) > 0$, se tiene

$$p(y|x, z) = \frac{p(x, y|z)}{p(x|z)} = \frac{p(x|z)p(y|z)}{p(x|z)} = p(y|z) \Rightarrow I(Y, X|Z),$$

donde la segunda igualdad se ha obtenido a partir de (5.26). ∎

5.7.2 Demostración de la Propiedad de Descomposición.

Dado que la función de probabilidad $p(x, y, z)$ cumple $I(X, Y \cup W|Z)$, se tiene

$$p(x|z, y, w) = p(x|z). \tag{5.27}$$

Veamos primero que también se cumple $I(X, Y|Z)$. Se tiene

$$\begin{aligned} p(x|z, y) &= \sum_v p(x, v|z, y) \\ &= \sum_v p(x|z, y, v)p(v|z, y), \end{aligned}$$

donde $V = W \setminus Y$ es el conjunto W excluyendo los elementos de Y. Aplicando (5.27) se tiene

$$\begin{aligned} p(x|z,y) &= \sum_v p(x|z)p(v|z,y) \\ &= p(x|z)\sum_v p(v|z,y) \\ &= p(x|z). \end{aligned}$$

La última igualdad se obtiene de

$$\sum_v p(v|z,y) = 1,$$

es decir, la suma de las probabilidades para todos los valores posibles de una variable ha de ser uno. Por tanto, $p(x|z,y) = p(x|z)$, y así, $I(X,Y|Z)$. Se puede demostrar, de forma similar, que la relación de independencia $I(X,W|Z)$ también se cumple. ∎

5.7.3 Demostración de la Propiedad de Unión Débil

Dado que la función de probabilidad $p(x,y,z)$ cumple $I(X,Y \cup W|Z)$, se tiene

$$p(x|z,y,w) = p(x|z). \tag{5.28}$$

Primero se muestra que esta relación de independencia implica $I(X,W|Z \cup Y)$. Si se aplica la propiedad de descomposición a $I(X,Y \cup W|Z)$, se tiene $I(X,Y|Z)$, es decir,

$$p(x|z,y) = p(x|z). \tag{5.29}$$

Aplicando (5.28) y (5.29) resulta

$$p(x|z,y,w) = p(x|z) = p(x|z,y),$$

lo cual implica $I(X,W|Z \cup Y)$. De forma similar puede obtenerse $I(X,Y|Z \cup W)$. ∎

5.7.4 Demostración de la Propiedad de Contracción

Dado que la función de probabilidad $p(x,y,z)$ cumple $I(X,W|Z \cup Y)$ en (5.4), se tiene que

$$p(x|z,y,w) = p(x|z,y). \tag{5.30}$$

De forma similar, si se satisface $I(X,Y|Z)$, entonces

$$p(x|z,y) = p(x|z). \tag{5.31}$$

A partir de (5.30) y (5.31) resulta

$$p(x|z,y,w) = p(x|z,y) = p(x|z).$$

Por tanto, también se cumple $I(X,Y \cup W|Z)$. ∎

5.7.5 Demostración de la Propiedad de Intersección

Dado que la función de probabilidad no extrema $p(x, y, z)$ cumple $I(X, W|Z \cup Y)$, se tiene

$$p(x|z, y, w) = p(x|z, y). \tag{5.32}$$

De forma similar, si se cumple $I(X, Y|Z \cup W)$, entonces

$$p(x|z, y, w) = p(x|z, w). \tag{5.33}$$

Las ecuaciones (5.32) y (5.33) implican

$$p(x|z, y, w) = p(x|z, y) = p(x|z, w),$$

que, dado que la probabilidad es no extrema, implica $p(x|z, y, w) = p(x|z)$. Por tanto, también se verifica $I(X, Y \cup W|Z)$. ∎

Ejercicios

5.1 Considérese el grafo no dirigido de la Figura 5.10. Comprobar cúales de las siguientes relaciones de independencia son ciertas utilizando el criterio de U-separación:

(a) $I(F, H|\phi)$.
(b) $I(F, H|D)$.
(c) $I(A, G|\{D, E\})$.
(d) $I(C, \{B, G\}|D)$.
(e) $I(\{A, B\}, \{F, G\}|\{C, D\})$.
(f) $I(\{C, F\}, \{G, E\}|\{A, D\})$.

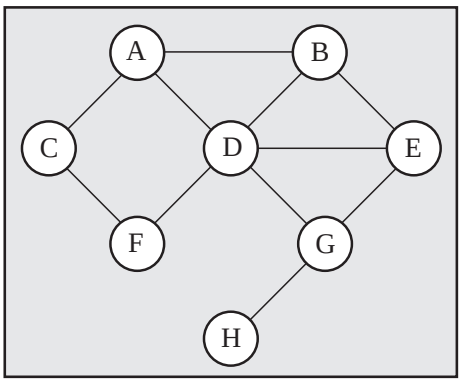


FIGURA 5.10. Grafo no dirigido.

5.2 Considérese el grafo dirigido de la Figura 5.11. Comprobar cúales de las siguientes relaciones de independencia son ciertas utilizando el criterio de D-separación dado en la Definición 5.3:

(a) $I(E, G|\phi)$.

(b) $I(C, D|\phi)$.

(c) $I(C, D|G)$.

(d) $I(B, C|A)$.

(e) $I(\{C, D\}, E|\phi)$.

(f) $I(F, \{E, H\}|A)$.

(g) $I(\{A, C\}, \{H, E\}|D)$.

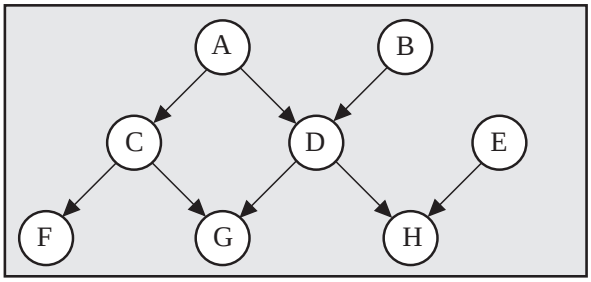


FIGURA 5.11. Grafo dirigido.

5.3 Repetir el ejercicio anterior utilizando el criterio de D-separación dado en la Definición 5.4.

5.4 Considérese el conjunto de cuatro variables $\{X, Y, Z, W\}$, relacionadas mediante

$$I(X, Y|\phi) \text{ y } I(X, Z|\{Y, W\}).$$

Encontrar el conjunto mínimo de relaciones de independencia generado por las dos relaciones de independencia anteriores y que además cumpla:

(a) La propiedad de simetría.

(b) Las propiedades de simetría y descomposición.

(c) Las propiedades de semigrafoide.

(d) Las propiedades de grafoide.

5.5 Repetir el ejercicio anterior considerando las siguientes relaciones de independencia:

$$I(X, W|\{Y, Z\}) \text{ y } I(Y, Z|\{X, W\}).$$

x	y	z	w	$p_1(x,y,z,w)$	$p_2(x,y,z,w)$
0	0	0	0	p_5p_8/p_{13}	$(-p_{13}p_4 + p_{12}p_5 + p_4p_8 + p_5p_8)/a$
0	0	0	1	p_5p_9/p_{13}	$(p_{13}p_4 - p_{12}p_5 + p_4p_9 + p_5p_9)/a$
0	0	1	0	$p_{10}p_7/p_{15}$	$(p_{10}p_6 - p_{15}p_6 + p_{10}p_7 + p_{14}p_7)/b$
0	0	1	1	$p_{11}p_7/p_{15}$	$(p_{11}p_6 + p_{15}p_6 + p_{11}p_7 - p_{14}p_7)/b$
0	1	0	0	$p_{12}p_5/p_{13}$	p_4
0	1	0	1	p_5	p_5
0	1	1	0	$p_{14}p_7/p_{15}$	p_6
0	1	1	1	p_7	p_7
1	0	0	0	p_8	p_8
1	0	0	1	p_9	p_9
1	0	1	0	p_{10}	p_{10}
1	0	1	1	p_{11}	p_{11}
1	1	0	0	p_{12}	p_{12}
1	1	0	1	p_{13}	p_{13}
1	1	1	0	p_{14}	p_{14}
1	1	1	1	p_{15}	p_{15}

TABLA 5.3. Dos familias paramétricas de funciones de probabilidad, donde $a = p_{12} + p_{13}$ y $b = p_{14} + p_{15}$.

5.6 Obtener el conjunto de todas las posibles relaciones de independencia condicional para un conjunto de tres variables.

5.7 Encontrar el conjunto de relaciones de independencia correspondiente a la función de probabilidad

$$p(x,y,z) = 0.3^{x+y}0.7^{2-x-y}\left(\frac{x+y}{2}\right)^z\left(1-\frac{x+y}{2}\right)^{1-z},$$

donde $x, y, z \in \{0, 1\}$.

5.8 Dado el conjunto de cuatro variables $\{X, Y, Z, W\}$ y la familia paramétrica de funciones de probabilidad $p_1(x, y, z, w)$ descrita en la Tabla 5.3,

(a) Probar que esta familia satisface la relación de independencia $I(X, Y \cup W|Z)$.

(b) ¿Es ésta la familia de funciones de probabilidad más general que cumple esta propiedad?

5.9 Dado el conjunto de cuatro variables $\{X, Y, Z, W\}$ y la familia paramétrica de funciones de probabilidad $p_2(x, y, z, w)$ descrita en la Tabla 5.3,

(a) Probar que esta familia satisface $I(X, Y|Z)$ y $I(X, W|Z)$.

(b) ¿Es ésta la familia de funciones de probabilidad más general que cumple estas propiedades?

(c) ¿Es suficiente suponer que $p_6 = p_{14}p_7/p_{15}$ y $p_4 = p_{12}p_5/p_{13}$ para que la familia anterior satisfaga $I(X, Y \cup W|Z)$?

5.10 Expresar en forma factorizada la función de probabilidad del Ejemplo 5.8 considerando las siguientes particiones del conjunto de variables:

(a) $Y_1 = \{X_1, X_3\}$, $Y_2 = \{X_2, X_4\}$.

(b) $Y_1 = \{X_4\}$, $Y_2 = \{X_2\}$, $Y_3 = \{X_1, X_3\}$.

(c) $Y_1 = \{X_2\}$, $Y_2 = \{X_1, X_3, X_4\}$.

5.11 Considérese el conjunto de cuatro variables dado en el Ejemplo 5.9 y supóngase que X_1 es una variable ternaria y que las otras tres variables son binarias.

(a) ¿Cuál es el número máximo de parámetros libres de la función de probabilidad?

(b) ¿Cuántos parámetros libres definen las funciones de probabilidad que cumplen las relaciones de independencia en (5.17)?

5.12 Repetir el ejercicio anterior suponiendo que las tres variables son ahora ternarias.

5.13 Considérese de nuevo el conjunto de cuatro variables dado en el Ejemplo 5.9. Escribir la forma factorizada asociada a cada uno de los siguientes casos y calcular el número de parámetros libres en cada uno de los modelos resultantes

(a) La función de probabilidad que cumple $I(X_1, X_4|\{X_2, X_3\})$.

(b) La función de probabilidad que satisface las condiciones de independencia $I(X_2, X_3|X_1)$, $I(X_3, X_4|X_1)$, y $I(X_2, X_4|X_1)$.

5.14 Encontrar la lista de relaciones de independencia asociada a la función de probabilidad dada en la Tabla 3.2.

5.15 Supóngase que una función de probabilidad de cuatro variables $\{X, Y, Z, W\}$ puede ser factorizada como

$$p(x, y, z, w) = p(x)p(y|x)p(z|x)p(w|y, z).$$

Comprobar cuáles de las siguientes relaciones de independencia se cumplen:

(a) $I(X, W|Y)$.

(b) $I(X, W|Z)$.

(c) $I(X, W|Y, Z)$.

(d) $I(Y, Z|X, W)$.

Capítulo 6
Modelos Definidos Gráficamente

6.1 Introducción

En el Capítulo 3 se ha visto que el funcionamiento de un sistema experto probabilístico depende de la correcta definición del correspondiente modelo, que está caracterizado por la función de probabilidad conjunta de las variables. También se ha visto que la estuctura general de una función de probabilidad conjunta involucra un excesivo número de parámetros. Por esta razón, en la Sección 3.5 se presentaron algunos modelos probabilísticos simplificados, que eran obtenidos imponiendo ciertas hipótesis de independencia globales sobre las variables. Sin embargo, estos modelos son restrictivos y solamente aplicables a problemas del tipo "enfermedades-síntomas". En este capítulo se desarrolla la forma de obtener modelos probabilísticos más generales por medio de grafos. La idea básica consiste en utilizar grafos (no dirigidos o dirigidos) para construir un modelo de dependencia que represente la estructura cualitativa del modelo probabilístico. De esta forma, los modelos resultantes son generales, pues se crean a partir de un modelo de dependencia "arbitrario", y no de uno impuesto inicialmente.

Antes de comenzar, es necesaria cierta notación y aclarar la terminología. El término modelo probabilístico se refiere a la estructura cualitativa y cuantitativa dada por una función de probabilidad. Por tanto, los términos *modelo probabilístico* y *función de probabilidad* se utilizarán de forma sinónima. Los términos *modelo de dependencia* y *modelo de independencia* se refieren exclusivamente a la estrucutura cualitativa de las relaciones existentes en el conjunto de variables. Estos modelos permiten comprobar qué

conjuntos de variables son incondicionalmente o condicionalmente dependientes o independientes. Cada modelo probabilístico tiene asociado un modelo de dependencia M, que puede ser obtenido generando todas las relaciones de independencia condicional posibles para un conjunto de variables dado, y comprobando cuáles de ellas se satisfacen para la función de probabilidad. Por ejemplo, si X, Y y Z son tres subconjuntos disjuntos y $p(x|y,z) = p(x|z)$, para cada combinación de valores de x, y y z, entonces se verifica la relación de independencia $I(X,Y|Z)$ y se puede concluir que X e Y son condicionalmente independientes dado Z. Por otra parte, si $p(x|y,z) \neq p(x|z)$ para algunos valores x, y, z, entonces X e Y son condicionalmente dependientes dado Z. Por tanto, una función de probabilidad contiene una descripción completa (cuantitativa y cualitativa) de las relaciones entre las variables, mientras que el modelo de dependencia M asociado sólo contiene una descripción cualitativa. Por tanto, el término modelo de dependencia probabilístico se refiere únicamente a un modelo de dependencia asociado a una función de probabilidad.

Por otra parte, un modelo de dependencia puede ser definido de forma alternativa mediante un grafo (dirigido o no dirigido), una lista de relaciones de independencia, o un conjunto de funciones de probabilidad condicionada. Estas tres alternativas determinan tres metodologías diferentes para construir un modelo de dependencia:

- Modelos definidos gráficamente.
- Modelos definidos por listas de independencias.
- Modelos definidos condicionalmente.

Estas tres metodologías son más generales que los modelos presentados en la Sección 3.5 y pueden ser aplicadas, no sólo a problemas de diagnóstico médico (problemas tipo "síntoma-enfermedad"), sino también a problemas más generales. Estas metodologías requieren ciertos conceptos previos (criterios de separación gráfica, propiedades de independencia condicional, etc.), ya tratados en las Secciones 5.2 y 5.3. Este capítulo está dedicado a los modelos definidos gráficamente o, de forma más precisa, a los modelos definidos a partir de un único grafo. El problema de los modelos descritos por un conjunto de grafos se analizará en el Capítulo 7.

En el Capítulo 4 se ha visto que un conjunto de variables $X_1, \ldots, X_n$ y sus relaciones pueden ser representados mediante un grafo, asociando cada variable a un nodo y cada relación entre variables a una arista entre los nodos correspondientes. Por tanto, los términos nodo y variable se utilizan de forma sinónima. En algunas ocasiones, el orden de las variables (es decir, la dirección de las aristas) es importante en el grafo (grafos dirigidos) y en otras no (grafo no dirigido). Las representaciones gráficas tienen la ventaja de mostrar explícitamente las relaciones entre las variables y conservar estas relaciones de forma cualitiativa (es decir, para cualquier valor numérico de

los parámetros). Los modelos gráficos son también más intuitivos y fáciles de entender.

En el Capítulo 5 se analizaron dos criterios gráficos de separación distintos para obtener las relaciones de independencia definidas por los grafos dirigidos y los no dirigidos. Según esta distinción, los modelos definidos gráficamente pueden ser clasificados en dos grupos, dependiendo del tipo de grafo que se utilice:

- Modelos de dependencia definidos por grafos no dirigidos, analizados en la Sección 6.3.

- Modelos de dependencia definidos por grafos dirigidos, analizados en la Sección 6.4.

Aunque existe un tercer tipo de modelos gráficos que pueden ser representados por grafos mixtos (grafos que contienen aristas dirigidas y no dirigidas), este capítulo está dedicado a los modelos definidos por grafos dirigidos y no dirigidos. El lector interesado en el análisis de modelos definidos por grafos mixtos puede consultar Lauritzen y Wermuth (1989) y Frydenberg (1990).

Se ha utilizado el término *dependencia* en las definiciones anteriores para enfatizar que un grafo sólo puede definir la estructura cualititativa del modelo. Una vez que se conoce esta estructura cualitativa, puede construirse una factorización de la función de probabilidad e identificarse el conjunto de parámetros que definen el modelo. Los valores numéricos de los parámetros pueden ser dados por un experto, o estimados a partir de un conjunto de datos disponibles (ver Sección 5.6). El conjunto de funciones de probabilidad condicionada junto con los valores de los parámetros asignados se conoce como la *estructura completa* del modelo.

En este capítulo se analiza también la capacidad de los grafos dirigidos y no dirigidos para captar ciertos tipos de estructuras de dependencia propias de los modelos probabilísticos, o de los modelos de dependencia en general. En la Sección 6.2 se introducen algunas definiciones y problemas a analizar. Las Secciones 6.3 y 6.4 analizan los modelos de dependencia definidos por grafos no dirigidos y dirigidos, respectivamente. La Sección 6.5 define y caracteriza las clases de modelos gráficos equivalentes. La Sección 6.6 analiza la capacidad de los grafos dirigidos y no dirigidos para representar modelos de dependencia.

6.2 Algunas Definiciones y Problemas

El objetivo de este capítulo es representar un modelo de dependencia probabilístico mediante un grafo. Por tanto, es importante conocer si los grafos permiten representar cualquier tipo de modelo de dependencia.

Definición 6.1 Mapa perfecto. *Un grafo G se dice que es un mapa perfecto de un modelo de dependencia M si cada relación de independencia obtenida de G también puede ser obtenida de M y viceversa, es decir,*

$$I(X, Y|Z)_M \Leftrightarrow I(X, Y|Z)_G \Leftrightarrow Z \text{ separa } X \text{ de } Y.$$

Dependiendo del carácter dirigido o no dirigido del grafo G, los mapas perfectos se denominan mapas perfectos dirigidos o no dirigidos, respectivamente.

Dado que es necesario tratar con dos tipos de grafos distintos, es necesario reformular el problema anterior de la forma siguiente:

- **Problema 6.1:** ¿Puede representarse mediante un mapa perfecto dirigido o no dirigido cualquier modelo de dependencia?

Desafortunadamente, no todo modelo de dependencia tiene asociado un mapa perfecto. Los dos ejemplos siguientes muestran dos modelos de dependencia que no poseen un mapa perfecto. En las Secciones 6.3 y 6.4 pueden encontrarse más ejemplos.

Ejemplo 6.1 Modelo sin mapa perfecto no dirigido. Considérese el conjunto de tres variables $\{X, Y, Z\}$ que están relacionadas por el siguiente modelo de dependencia

$$M = \{I(X, Y|\phi), I(Y, X|\phi)\}, \tag{6.1}$$

que sólo contiene una relación de independencia y su relación simétrica. Supóngase que se quiere representar este modelo por medio de un grafo no dirigido. En general, para un conjunto de n nodos podrían construirse los $2^{n(n-1)/2}$ grafos no dirigidos distintos (ver Whittaker (1990)). La Figura 6.1 muestra los ocho grafos para el caso de tres variables. Estos grafos están ordenados en filas que contienen grafos con el mismo número de aristas. Así, la figura (a) corresponde al grafo totalmente inconexo (un grafo sin ninguna arista), cada uno de los tres grafos en (b)−(d) contiene una única arista, cada uno de los grafos en (e)−(g) contiene dos aristas, y el último grafo es el grafo completo (un grafo con una arista entre cada par de nodos). La segunda columna de la Tabla 6.1 muestra algunas relaciones de independencia implicadas por cada uno de los grafos, y que no están contenidas en M. El lector puede comprobar fácilmente estas relaciones utilizando el criterio de U-separación descrito en la Sección 5.2. La última columna de la tabla muestra cuando las relaciones de independencia de M están contenidas en el grafo G. Como puede verse en la Tabla 6.1, en cada grafo G se puede encontrar una relación de independencia que no está contenida en M y/o viceversa. Por tanto, ninguno de los grafos de la Figura 6.1 es un mapa perfecto de M en (6.1). Puesto que este conjunto de grafos es exhaustivo, el modelo de dependencia M no posee ningún mapa perfecto no dirigido. ∎

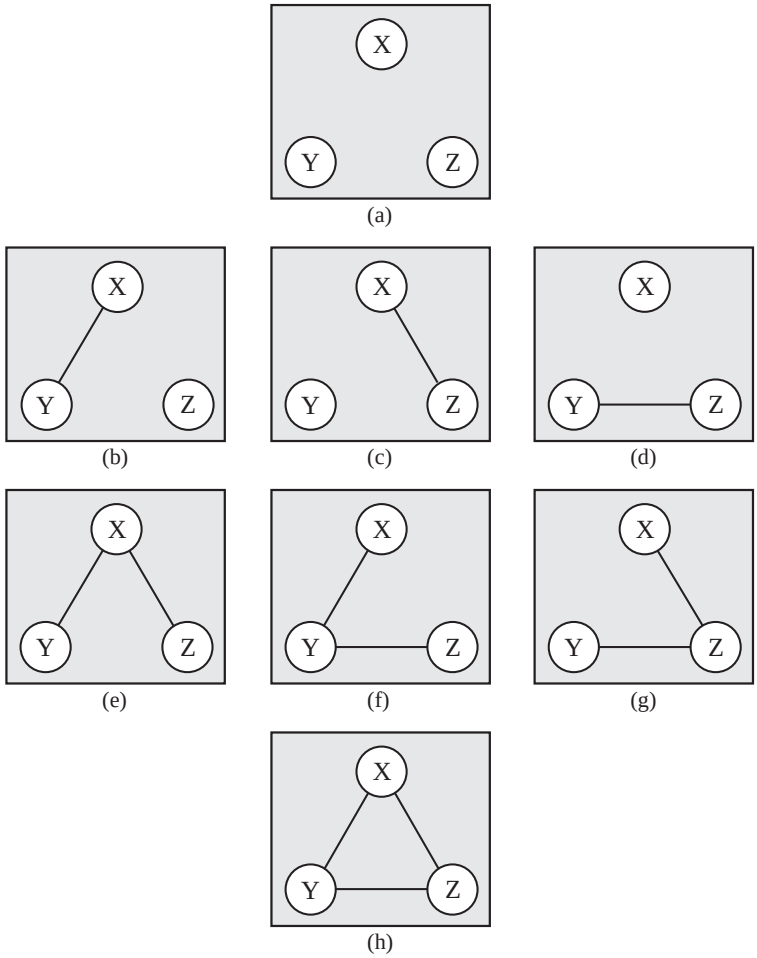


FIGURA 6.1. Ocho posibles grafos no dirigidos con tres variables.

El modelo de dependencia M del Ejemplo 6.1 tiene un mapa perfecto dirigido, a pesar de que no posee ningún mapa perfecto no dirigido. Se deja como ejercicio para el lector demostrar que el grafo dirigido mostrado en la Figura 6.2 es un mapa perfecto dirigido de M. En este caso, los grafos dirigidos son más potentes que los no dirigidos. Sin embargo, no todo modelo de dependencia posee un mapa perfecto dirigido. El ejemplo siguiente muestra uno de estos modelos.

Ejemplo 6.2 Modelo sin mapa perfecto dirigido. Considérese el conjunto de tres variables $\{X, Y, Z\}$ y el modelo de dependencia

$$M = \{I(X, Y|Z), I(Y, Z|X), I(Y, X|Z), I(Z, Y|X)\}. \tag{6.2}$$

Grafo G	Independencia en G pero no en M	Independencia en M pero no en G
(a)	$I(X, Z\|\phi)$	ϕ
(b)	$I(X, Z\|\phi)$	$I(X, Y\|\phi)$
(c)	$I(Y, Z\|\phi)$	ϕ
(d)	$I(X, Z\|\phi)$	ϕ
(e)	$I(Y, Z\|X)$	$I(X, Y\|\phi)$
(f)	$I(X, Z\|Y)$	$I(X, Y\|\phi)$
(g)	$I(X, Y\|Z)$	$I(X, Y\|\phi)$
(h)	ϕ	$I(X, Y\|\phi)$

TABLA 6.1. Algunas relaciones de independencia contenidas en G en la Figura 6.1 pero no en el modelo de dependencia M en (6.1).

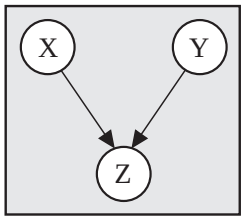


FIGURA 6.2. Mapa perfecto dirigido del modelo de dependencia M en (6.1).

No existe ningún grafo dirigido acíclico D que sea mapa perfecto del modelo de dependencia M. ■

En los casos en los que no existe un mapa perfecto, es necesario asegurarse de que el modelo gráfico que se utilice no posea ninguna independencia que no esté contenida en el modelo, y que el número de independencias del modelo que no sean reproducidas por el grafo sea mínimo. Esto motiva las siguientes definiciones.

Definición 6.2 Mapa de independencia. *Un grafo G se dice que es un mapa de independencia (I-mapa) de un modelo de dependencia M si*

$$I(X, Y|Z)_G \Rightarrow I(X, Y|Z)_M,$$

es decir, si todas las relaciones de dependencia derivadas de G son verificadas por M.

Obsérvese que un I-mapa G de un modelo de dependencia M incluye algunas de las independencias de M, pero no necesariamente todas. Entonces, se tiene

$$I(X, Y|Z)_G \;\Rightarrow\; I(X, Y|Z)_M,$$

lo cual implica

$$D(X, Y|Z)_M \;\Rightarrow\; D(X, Y|Z)_G.$$

Por tanto, todas las dependencias de M están representadas en G. Por ejemplo, solamente el grafo completo de la Figura 6.1(h) es un I-mapa del modelo de dependencia dado en (6.1). Cada uno de los grafos restantes implica algunas independencias que no son propias de M (ver Tabla 6.1). En general, un grafo completo es siempre un I-mapa trivial de cualquier modelo de dependencia.

Definición 6.3 Mapa de dependencia. *Un grafo G se dice que es un mapa de dependencia (D-mapa) de un modelo de dependencia M si*

$$D(X, Y|Z)_G \Rightarrow D(X, Y|Z)_M,$$

es decir, todas las relaciones de dependencia derivadas de G son verificadas por M.

Si G es un D-mapa de M, se tiene

$$D(X, Y|Z)_G \;\Rightarrow\; D(X, Y|Z)_M,$$

lo cual implica

$$I(X, Y|Z)_M \;\Rightarrow\; I(X, Y|Z)_G,$$

es decir, todas la independencias de M están representadas en G.

Obsérvese que un D-mapa de un modelo de dependencia M sólo incluye algunas de las dependencias de M. Por ejemplo, el grafo totalmente inconexo de la Figura 6.1(a) es un D-mapa trivial, aunque inútil, del modelo de dependencia dado en (6.1). Los grafos de las Figuras 6.1(c) y (d) son también D-mapas del modelo de dependencia.

Por tanto, cada modelo de dependencia tiene asociados un I-mapa y un D-mapa triviales. Por ejemplo, cualquier grafo totalmente inconexo es un D-mapa trivial y cualquier grafo completo es un I-mapa trivial de cualquier modelo de dependencia. De esta forma, para que un grafo sea un mapa perfecto de un modelo, ha de ser simultáneamente un I-mapa y un D-mapa de ese modelo.

Definición 6.4 I-mapa minimal. *Se dice que un grafo G es un I-mapa minimal de un modelo de dependencia M si es un I-mapa de M, pero pierde esta propiedad cuando se elimina una cualquiera de sus aristas.*

A pesar de que los modelos de dependencia y las representaciones gráficas tienen numerosas aplicaciones más allá de la probabilidad, el interés principal de este libro es la construcción de modelos probabilísticos y, por tanto, estamos interesados en conocer la relación existente entre las representaciones gráficas y las funciones de probabilidad, es decir, la relación existente entre las nociones formales de dependencia probabilística y la estructura

topológica de un grafo. Una razón importante para representar la estructura de dependencia de un modelo mediante un grafo es que comprobar la conexión de un conjunto de variables en un grafo (utilizando alguno de los criterios gráficos de separación introducidos en el Capítulo 5) es más fácil que comprobar la independencia condicional de un conjunto de variables utilizando las fórmulas de la Probabilidad dadas en la Sección 3.2. Un D-mapa garantiza que todos los nodos que estén conectados en el grafo serán por tanto dependientes; sin embargo, el grafo puede ocasionalmente representar desconectados algunos conjuntos de variables dependientes. Por el contrario, un I-mapa garantiza que los nodos separados en el grafo siempre corresponden a variables independientes, pero no garantiza que todos los nodos conectados sean dependientes. Como ya se mencionó anteriormente, los grafos totalmente inconexos son D-mapas triviales, mientras que los grafos completos son I-mapas triviales.

El problema de definir un modelo gráfico asociado a un modelo de dependencia dado no es un problema trivial. Cuando se trata con algún modelo donde la noción de vecindad o conexión es explícita (por ejemplo, relaciones familiares, circuitos electrónicos, redes de comunicación, etc.) se suelen tener pocos problemas para definir un grafo que represente las características principales del modelo. Sin embargo, cuando se trata con relaciones conceptuales como asociación o relevancia, es a menudo difícil distinguir entre vecinos directos e indirectos. En estos casos, la tarea de construir una representación gráfica se vuelve más difícil. Un ejemplo claro de este problema es la noción de independencia condicional en probabilidad. Dada una función de probabilidad de tres variables X, Y y Z, es fácil comprobar si X e Y son independientes dada Z; sin embargo, la función de probabilidad no proporciona ninguna información sobre cuál de estas variables es la causa y cuál es el efecto.

En el Capítulo 4 se han introducido algunos conceptos elementales sobre la teoría de grafos y se ha visto que los nodos de un grafo representan variables y las aristas representan dependencias locales entre variables conceptualmente relacionadas. Por tanto, las aristas de un grafo permiten representar relaciones cualitativas y la topología del grafo muestra estas relaciones de forma explícita y las conserva tras la asignación numérica de los parámetros. En este capítulo se analiza la forma de representar algunos modelos probabilísticos por medio de grafos dirigidos y no dirigidos.

Dado que no todo modelo probabilístico puede ser representado por un mapa perfecto, se presentan los siguientes problemas:

- **Problema 6.2:** ¿Cuáles son los modelos de dependencia y, en particular, los modelos de dependencia probabilísticos que pueden ser representados por un mapa perfecto?

- **Problema 6.3:** ¿Cuáles son los modelos de dependencia probabilísticos que poseen un único I-mapa minimal?

- **Problema 6.4:** Si un modelo probabilístico posee un único I-mapa minimal ¿cómo se puede obtener este I-mapa?
- **Problema 6.5:** Dado un grafo G, ¿existe algún modelo probabilístico P tal que G sea un I-mapa minimal de P?. En caso afirmativo, ¿cómo se puede construir?

En la Sección 6.3 se analizarán estos problemas para el caso de grafos no dirigidos y en la Sección 6.4, para el caso de grafos dirigidos. Obsérvese que el Problema 5.7, "¿cuál es la clase de modelos probabilísticos que puede ser representada por grafos?" se ha dividido ahora en dos partes: Problemas 6.2 y 6.3.

6.3 Modelos de Dependencia Gráficos no Dirigidos

En esta sección se analiza la forma de definir modelos de dependencia utilizando grafos no dirigidos. Nuestro objetivo es encontrar un grafo que reproduzca tantas independencias asociadas a un modelo probabilístico como sea posible. Se comienza con el problema de representar estos modelos por medio de mapas perfectos e I-mapas y, a continuación, se introduce un clase importante de modelos probabilísticos definidos por grafos no dirigidos. Estos modelos se conocen por *redes de Markov*.

6.3.1 De Modelos a Grafos no Dirigidos

En esta sección se analiza el problema de representar modelos probabilísticos utilizando grafos no dirigidos, es decir, se desea encontrar el grafo correspondiente a un modelo de dependencia probabilístico. Como ya se ha visto en el Ejemplo 6.1, no todos lo modelos probabilísticos de dependencia pueden ser representados por mapas perfectos no dirigidos. Pearl y Paz (1987) probaron el siguiente teorema que caracteriza los modelos de dependencia que pueden ser representados mediante mapas perfectos no dirigidos. El teorema se refiere no sólo a modelos de dependencia probabilísticos, sino a modelos de dependencia en general.

Teorema 6.1 Modelos con mapa perfecto no dirigido. *Una condición necesaria y suficiente para que un modelo de dependencia M tenga un mapa perfecto no dirigido es que satisfaga las siguientes propiedades:*

- **Simetría:**
$$I(X, Y|Z)_M \Leftrightarrow I(Y, X|Z)_M.$$
- **Descomposición:**
$$I(X, Y \cup W|Z)_M \Rightarrow I(X, Y|Z)_M \text{ y } I(X, W|Z)_M.$$

- **Intersección:**

$$I(X, W|Z \cup Y)_M \text{ y } I(X, Y|Z \cup W)_M \Rightarrow I(X, Y \cup W|Z)_M.$$

- **Unión fuerte:**

$$I(X, Y|Z)_M \Rightarrow I(X, Y|Z \cup W)_M.$$

- **Transitividad fuerte:**

$$I(X, Y|Z)_M \Rightarrow I(X, A|Z)_M \text{ o } I(Y, A|Z)_M,$$

donde A es un conjunto formado por un único nodo que no esté contenido en $\{X, Y, Z\}$.

Por tanto, la respuesta al Problema 6.2 para el caso de grafos no dirigidos es que solamente los modelos de dependencia que satisfagan estas propiedades tienen un mapa perfecto no dirigido, en el sentido de que las dependencias e independencias correspondientes al modelo y al mapa perfecto son las mismas. El caso de grafos dirigidos es analizado en la Sección 6.4.1.

Obsérvese que, en general, los grafoides y los semigrafoides no tienen mapas perfectos no dirigidos, pues los semigrafoides solamente han de cumplir las propiedades de simetría y descomposición y los grafoides sólo han de verificar las propiedades de simetría, descomposición, e intersección. Por ejemplo, el modelo de dependencia del Ejemplo 6.1 es un grafoide, pero no tiene asociado un mapa perfecto, pues viola las propiedades de unión y transitividad fuerte.

Los modelos probabilísticos de dependencia también pueden violar las dos últimas propiedades y, por tanto, no todo modelo probabilístico puede representarse mediante un mapa perfecto no dirigido. Los siguientes ejemplos ilustran este hecho.

Ejemplo 6.3 Violación de la unión fuerte y la transitividad fuerte. La propiedad de unión fuerte afirma que si X e Y son independientes dado Z, entonces también son independientes dado un conjunto mayor $Z \cup W$:

$$I(X, Y|Z) \Rightarrow I(X, Y|Z \cup W).$$

Por ejemplo, considerando $Z = \phi$, esta propiedad implica: $I(X, Y|\phi) \Rightarrow I(X, Y|W)$, que afirma que si X e Y son incondicionalmente independientes, entonces también deben ser condicionalmente independientes dado otro subconjunto de variables cualquiera W. Esta afirmación no es siempre cierta para modelos probabilísticos. Por ejemplo para la familia de funciones de probabilidad dada por la factorización

$$p(x, y, z) = p(x)p(y)p(z|x, y),$$

se tiene $I(X, Y|\phi)$, pero $I(X, Y|Z)$ no es cierto, en general. Por tanto, $p(x, y, z)$ viola la propiedad de unión fuerte y, por esta razón, no puede ser representada por un mapa perfecto no dirigido, tal y como se ha visto en el Ejemplo 6.1. Además, esta familia de funciones de probabilidad también viola la propiedad de transitividad fuerte. Según esta propiedad, y considerando $Z = \phi$, se tiene

$$I(X, Y|\phi) \Rightarrow I(X, A|\phi) \text{ o } I(Y, A|\phi),$$

donde A es un conjunto formado por un único nodo distinto de $\{X, Y\}$. En este caso $A = Z$. Sin embargo, aunque se cumple $I(X, Y|\phi)$, la familia de probabilidades anterior no satisface ni $I(X, Z|\phi)$ ni $I(X, Z|\phi)$. Se deja como ejercicio para el lector hallar una combinación de valores numéricos para los parámetros asociados a las funciones de probabilidad condicionada de la familia que permitan obtener una función de probabilidad $p(x, y, z)$ que viole ambas propiedades. Obsérvese que estos parámetros no pueden ser elegidos arbitrariamente pues alguna elección específica de los parámetros puede hacer que las variables X y Z, o Y y Z sean independientes y, por tanto, las propiedades anteriores no serían violadas. ∎

El ejemplo siguiente ilustra la violación de las propiedades de unión fuerte y transitividad fuerte utilizando una función de probabilidad de tipo continuo. Este ejemplo requiere la propiedad siguiente de la función de distribución normal multivariada (ver Anderson (1984), Johnson y Wichern (1988), o Rencher (1995)).

Teorema 6.2 Distribución normal multivariada. *Sean X e Y dos conjuntos de variables aleatorias con función de distribución normal multivariada cuyo vector de medias y matriz de covarianzas son*

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix},$$

donde μ_X y Σ_{XX} son el vector de medias y la matriz de covarianzas de X, μ_Y y Σ_{YY} son el vector de medias y la matriz de covarianzas de Y, y Σ_{XY} es la matriz de covarianzas de X e Y. Entonces, la función de probabilidad condicionada de X dada $Y = y$ es una función normal multivariada con vector de medias $\mu_{X|Y=y}$ y matriz de covarianzas

$$\mu_{X|Y=y} = \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_y), \tag{6.3}$$

$$\Sigma_{X|Y=y} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}. \tag{6.4}$$

Obsérvese que la media condicionada $\mu_{X|Y=y}$ depende del valor y, pero no así la varianza condicionada $\Sigma_{X|Y=y}$.

Ejemplo 6.4 Violación de la unión fuerte y la transitividad fuerte. Supóngase que las variables (X_1, X_2, X_3) están distribuidas de forma nor-

mal con

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} 1 & 0 & 1/4 \\ 0 & 1 & 1/2 \\ 1/4 & 1/2 & 1 \end{pmatrix}. \tag{6.5}$$

La propiedad de unión fuerte implica que si los conjuntos de variables X e Y son incondicionalmente independientes, entonces, también han de ser condicionalmente independientes dado otro subconjunto W. Es decir $I(X, Y|\phi) \Rightarrow I(X, Y|W)$. En este ejemplo, las únicas variables incondicionalmente independientes son X_1 y X_2, ya que $\Sigma_{X_1X_2} = \Sigma_{X_2X_1} = 0$. Por tanto, se tiene $I(X_1, X_2|\phi)$. Sin embargo, X_1 y X_2 no son condicionalmente independientes dada X_3, es decir, no se verifica $I(X_1, X_2|X_3)$. Para comprobar esta última afirmación, se utiliza (6.4) para calcular

$$\begin{aligned} \Sigma_{X_1|X_3} &= \Sigma_{X_1X_1} - \Sigma_{X_1X_3}\Sigma^{-1}_{X_3X_3}\Sigma_{X_3X_1} \\ &= 1 - \frac{1}{4} \times 1 \times \frac{1}{4} = \frac{15}{16}, \end{aligned} \tag{6.6}$$

$$\begin{aligned} \Sigma_{X_1|X_2,X_3} &= \Sigma_{X_1X_1} - \Sigma_{X_1(X_2X_3)}\Sigma^{-1}_{(X_2X_3)(X_2X_3)}\Sigma_{(X_2X_3)X_1} \\ &= 1 - \begin{pmatrix} 0 & \frac{1}{4} \end{pmatrix} \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \frac{1}{4} \end{pmatrix} = \frac{11}{12}. \end{aligned} \tag{6.7}$$

De (6.6) y (6.7) se obtiene que las funciones de distribución normales de las variables $(X_1|X_3)$ y $(X_1|X_2, X_3)$ son distintas; por tanto, la distribución normal cuya matriz de covarianzas está dada en (6.5) viola la propiedad de unión fuerte y, por tanto, no puede ser representada por un mapa perfecto no dirigido.

De forma similar, para demostrar la violación de la propiedad de transitividad fuerte, se considera $Z = \phi$, obteniéndose

$$I(X, Y|\phi) \Rightarrow I(X, A|\phi) \text{ o } I(Y, A|\phi),$$

donde A es un conjunto de una única variable que no está contenida en $\{X, Y\}$. Esta propiedad no se verifica en el modelo probabilístico normal dado en (6.5). Para comprobar esta afirmación, se toma $X = X_1$, $Y = X_2$, y $A = X_3$. Se conoce que X_1 y X_2 son incondicionalmente independientes, pero cada una de ellas depende de X_3. Las ecuaciones (6.6) y (6.5) muestran que X_1 y X_3 no son independientes, pues $\Sigma_{X_1|X_3} \neq \Sigma_{X_1X_1}$. Por otra parte, utilizando (6.4), se tiene

$$\Sigma_{X_2|X_3} = \Sigma_{X_2X_2} - \Sigma_{X_2X_3}\Sigma^{-1}_{X_3X_3}\Sigma_{X_3X_2} = 1 - \frac{1}{2}\frac{1}{2} = \frac{3}{4} \neq \Sigma_{X_2X_2},$$

que muestra que X_2 no es independiente de X_3. Por tanto, el modelo probabilístico normal multivariado dado en (6.5) no satisface la propiedad de transitividad fuerte. ∎

En los casos en los que es imposible construir un mapa perfecto, se trata de contruir un I-mapa del modelo dado. A partir de la Definición 6.2, se sigue que todo modelo probabilístico posee un I-mapa pero, para que éste represente el mayor número posible de independencias de M, ha de ser un I-mapa minimal. Sin embargo, obsérvese que un modelo de dependencia probabilístico puede no tener un único I-mapa minimal. El siguiente teorema, debido a Pearl y Paz (1987) (ver también Verma y Pearl (1990)), muestra las condiciones que ha de satisfacer un modelo de dependencia para tener asociado un único I-mapa no dirigido minimal. Este teorema muestra también la forma de construirlo.

Teorema 6.3 I-mapa no dirigido minimal. *Todo modelo de dependencia M de un conjunto de variables $X = \{X_1, \ldots, X_n\}$ que satisfaga las propiedades de simetría, descomposición, e intersección tiene un único I-mapa no dirigido minimal que se obtiene eliminando del grafo completo toda arista (X_i, X_j) que satisfaga $I(X_i, X_j | X \setminus \{X_i, X_j\})_M$, donde $X \setminus \{X_i, X_j\}$ denota el conjunto de variables en X excluyendo aquellas en X_i y X_j.*

Obsérvese que los modelos probabilísticos no extremos satisfacen las tres propiedades exigidas en el teorema. Por tanto, cada función de probabilidad no extrema tiene asociado un único I-mapa no dirigido minimal. El Teorema 6.3 muestra la solución del Problema 6.3 para grafos no dirigidos: "¿cuáles son los modelos de dependencia probabilísticos que poseen un único I-mapa minimal?" El caso de grafos dirigidos se analizará en la Sección 6.4.1.

Obsérvese que los grafoides satisfacen las propiedades de simetría, descomposición, e intersección. Por tanto, una característica importante de los grafoides es que poseen I-mapas no dirigidos minimales únicos, y permiten, por tanto, su construcción a través de independencias locales. Conectando cada variable X_i en X con cualquier otro subconjunto de variables que haga que X_i sea condicionalmente independiente del resto de las varibles, se obtendrá un grafo que será un I-mapa del grafoide. Esta construcción local no está garantizada para el caso de semigrafoides.

Ejemplo 6.5 I-mapa minimal no dirigido (I). considérese el conjunto de cuatro variables $\{X_1, X_2, X_3, X_4\}$ que están relacionadas por el modelo de dependencia:

$$\begin{aligned} M = \{&I(X_1, X_2|X_3), \quad I(X_1, X_4|X_2), \quad I(X_1, X_4|\{X_2, X_3\}), \\ &I(X_2, X_1|X_3), \quad I(X_4, X_1|X_2), \quad I(X_4, X_1|\{X_2, X_3\})\}, \end{aligned} \tag{6.8}$$

al igual que en el Ejemplo 5.7. Este modelo de dependencia cumple las tres propiedades requeridas en el Teorema 6.3. Por tanto, puede obtenerse el I-mapa no dirigido minimal asociado sin más que comprobar qué independencias de la forma $I(X_i, X_j | X \setminus \{X_i, X_j\})_M$ se cumplen en M. Todas las posibles relaciones de independencia de esta forma para un conjunto de

cuatro variables son

$$I(X_1, X_2|\{X_3, X_4\}), \quad I(X_1, X_3|\{X_2, X_4\}), \quad I(X_1, X_4|\{X_2, X_3\}),$$
$$I(X_2, X_3|\{X_1, X_4\}), \quad I(X_2, X_4|\{X_1, X_3\}), \quad I(X_3, X_4|\{X_1, X_2\}).$$

La única relación de independencia de esta lista que se cumple en M es $I(X_1, X_4|\{X_2, X_3\})$. Por tanto, para obtener el I-mapa no dirigido minimal de M únicamente ha de eliminarse la arista $(X_1 - X_4)$ del grafo completo de la Figura 6.3(a). El grafo resultante se muestra en la Figura 6.3(b).

Considérese ahora el nuevo modelo de dependencia M' creado añadiendo la relación de independencia $I(X_1, X_2|\{X_3, X_4\})$ al modelo M dado en (6.8), es decir,

$$M' = M \cup \{I(X_1, X_2|\{X_3, X_4\})\}. \tag{6.9}$$

Si se aplica a M' el procedimiento anterior para construir un grafo no dirigido (ver Teorema 6.3) se obtendrá el grafo dado en la Figura 6.3(c). Sin embargo, este grafo no es un I-mapa de M'. Por ejemplo, la independencia $I(X_1, X_4|X_3)_G$ se cumple en el grafo, pero no está contenida en M'. La razón de esta discordancia es que M' no cumple las condiciones del Teorema 6.3. Por ejemplo, si se aplica la propiedad de intersección a $I(X_1, X_2|\{X_3, X_4\})$ e $I(X_1, X_4|\{X_2, X_3\})$ se obtiene $I(X_1, \{X_2, X_4\}|X_3)$; si después se aplica la propiedad de descomposición a $I(X_1, \{X_2, X_4\}|X_3)$ se tienen $I(X_1, X_2|X_3)$ y $I(X_1, X_4|X_3)$, que no están contenidas en M'. De forma similar, si se aplica la propiedad de intersección a $I(X_1, X_4|X_2)$ y $I(X_1, X_2|\{X_3, X_4\})$ se obtiene $I(X_1, \{X_2, X_4\}|\phi)$ y aplicando la propiedad de descomposición se tienen $I(X_1, X_2|\phi)$ y $I(X_1, X_4|\phi)$. Por tanto, el modelo de dependencia

$$M \cup C, \tag{6.10}$$

tiene el grafo dado en la Figura 6.3(c) como único I-mapa minimal, donde C es el conjunto que contiene las siguientes relaciones de independencia y sus relaciones simétricas:

$$I(X_1, X_2|\{X_3, X_4\}), \quad I(X_1, \{X_2, X_4\}|X_3), \quad I(X_1, X_4|X_3),$$
$$I(X_1, \{X_2, X_4\}|\phi) \quad I(X_1, X_2|\phi), \quad I(X_1, X_4|\phi).$$

■

A partir del Teorema 6.3 se deduce que toda función de probabilidad no extrema tiene asociado un único I-mapa no dirigido minimal obtenido eliminando del grafo completo toda arista L_{ij} entre nodos X_i y X_j tales que $I(X_i, X_j|X \setminus \{X_i, X_j\})_P$. Obsérvese también que $I(X_i, X_j|X \setminus \{X_i, X_j\})_P$ es equivalente a

$$p(x_i|x \setminus \{x_i, x_j\}) = p(x_i|x \setminus x_i), \tag{6.11}$$

que implica

$$\frac{p(x \setminus x_j)}{p(x \setminus \{x_i, x_j\})} = \frac{p(x)}{p(x \setminus x_i)}. \tag{6.12}$$

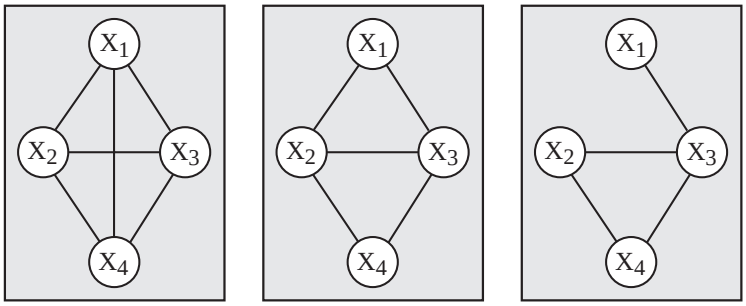


FIGURA 6.3. Grafo completo de cuatro nodos (a), I-mapa no dirigido minimal para el modelo de dependencia M dado por (6.8) (b), e I-mapa no dirigido minimal para el modelo en (6.10) (c).

Esto sugiere el siguiente algoritmo para resolver el Problema 6.4 en el caso de grafos no dirigidos: "Si un modelo probabilístico posee un único I-mapa minimal, ¿cómo se puede obtener este I-mapa?". En la Sección 6.4.1 se analiza el caso de grafos dirigidos.

Algoritmo 6.1 I-Mapa minimal no dirigido.

- **Datos:** Un conjunto de variables $X = \{X_1, \ldots, X_n\}$ y una función de probabilidad no extrema $p(x)$.
- **Resultados:** El I-mapa minimal no dirigido correspondiente a $p(x)$.

1. Considérese un grafo completo de n nodos, en el cual existe una arista entre cada par de nodos.
2. Para cada par de nodos (X_i, X_j) calcular

$$p(x \setminus x_i) = \sum_{x_i} p(x),$$
$$p(x \setminus x_j) = \sum_{x_j} p(x),$$
$$p(x \setminus \{x_i, x_j\}) = \sum_{x_j} p(x \setminus x_i).$$

   Entonces, si

$$p(x)p(x \setminus \{x_i, x_j\}) = p(x \setminus x_i)p(x \setminus x_j),$$

   eliminar la arista L_{ij} entre los nodos X_i y X_j. ∎

El ejemplo siguiente ilustra este algoritmo.

Ejemplo 6.6 I-mapa minimal no dirigido (II). Considérese el conjunto de variables binarias $X = \{X_1, \ldots, X_7\}$ y una función de probabilidad definida mediante la factorización

$$p(x) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_3)p(x_6|x_4)p(x_7|x_4), \quad (6.13)$$

que depende de 15 parámetros $\theta_1, \ldots, \theta_{15}$. Se desea construir el I-mapa minimal no dirigido asociado a $p(x)$. La Figura 6.4 muestra un programa de ordenador que implementa el Algoritmo 6.1 para este caso. Este programa está escrito en *Mathematica* (ver Wolfram (1991)) pero puede ser implementado de forma similar en cualquier otro programa de cálculo simbólico. El programa comienza definiendo la función de probabilidad en (6.13). Con este fin se introducen las funciones $PA[i]$, $PB[i, j]$, etc., que se definen simbólicamente utilizando los parámetros $p1, \ldots, p15$. A continuación, se define la función de probabilidad conjunta como el producto de estas funciones. Las funciones $P1[i]$ y $P2[i, j]$ son funciones auxiliares definidas para obtener funciones de probabilidad marginales de una variable X_i, o de dos variables X_i y X_j, respectivamente. La última parte del programa comprueba si se satisface la condición (6.12) para todas las posibles combinaciones de nodos X_i y X_j. Después de la ejecución del programa, se deduce que las aristas siguientes pueden ser eliminadas del grafo:

$$\begin{array}{cccccccc} L_{14}, & L_{15}, & L_{16}, & L_{17}, & L_{23}, & L_{25}, & L_{26}, \\ L_{27}, & L_{36}, & L_{37}, & L_{45}, & L_{56}, & L_{57}, & L_{67}. \end{array}$$

Por tanto, comenzando con el grafo completo, en el que cada par de nodos está unido por una arista, se obtiene el grafo de la Figura 6.5, que es el I-mapa minimal no dirigido correspondiente al modelo probabilístico dado en (6.13). ∎

Ejemplo 6.7 I-mapa minimal no dirigido (III). Supóngase ahora que una función de probabilidad de cinco variables viene dada por la factorización

$$p(x) = \psi_1(x_1, x_2, x_3)\psi_2(x_1, x_3, x_4)\psi_3(x_1, x_4, x_5), \quad (6.14)$$

donde ψ_1, ψ_2, y ψ_3 son funciones positivas (factores potenciales) indeterminadas. La Figura 6.6 muestra un programa de *Mathematica* para hallar el I-mapa minimal no dirigido correspondiente. La Figura 6.7 muestra el I-mapa obtenido tras la ejecución del programa. Puede comprobarse que el I-mapa minimal no dirigido asociado a la función de probabilidad $p(x)$ en (6.14) se obtiene eliminando las aristas L_{24}, L_{25} y L_{35} del grafo completo. ∎

6.3.2 De Grafos no Dirigidos a Modelos probabilísticos

En las secciones anteriores se ha supuesto que se conoce el modelo probabilístico $p(x)$, o el correspondiente modelo de dependencia M. Por tanto,

```
T={p1,1-p1}; n=1;
Do[PA[i1]=T[[n]];n++,{i1,0,1}];
T={p2,p3,1-p2,1-p3}; n=1;
Do[PB[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];
T={p4,p5,1-p4,1-p5}; n=1;
Do[PC[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];
T={p6,p7,p8,p9,1-p6,1-p7,1-p8,1-p9}; n=1;
Do[PD[i1,i2,i3]=T[[n]];n++,{i1,0,1},{i2,0,1},{i3,0,1}];
T={p10,p11,1-p10,1-p11}; n=1;
Do[PE[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];
T={p12,p13,1-p12,1-p13}; n=1;
Do[PF[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];
T={p14,p15,1-p14,1-p15}; n=1;
Do[PG[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];
P[x1_,x2_,x3_,x4_,x5_,x6_,x7_]=PA[x1]*PB[x2,x1]*PC[x3,x1]*
    PD[x4,x2,x3]*PE[x5,x3]*PF[x6,x4]*PG[x7,x4];
P1=Sum[P[x[1],x[2],x[3],x[4],x[5],x[6],x[7]],
{x[#1],0,1}]&;
P2=Sum[P[x[1],x[2],x[3],x[4],x[5],x[6],x[7]],
    {x[#1],0,1},{x[#2],0,1}]& ;
Do[
    Do[
        a=Simplify[P[x[1],x[2],x[3],x[4],x[5],x[6],x[7]]*
            P2[i,j]-P1[i]*P1[j]];
        If[a==0,Print["Eliminar arista ",i,"--",j]],
    {j,i+1,7}],
{i,1,7}]
```

FIGURA 6.4. Programa de *Mathematica* para encontrar el I-mapa minimal no dirigido correspondiente a la función de probabilidad dada en (6.13).

siempre es posible obtener un I-mapa no dirigido que contenga tantas independencias de M como sea posible. Sin embargo, en la práctica no se suele conocer la función de probabilidad $p(x)$ ni el modelo M. Por tanto, la forma real de construir un modelo probabilístico consiste en los pasos siguientes:

1. Construir un grafo no dirigido G que defina la estructura de dependencia de un conjunto de variables X.

2. Encontrar una función de probabilidad $p(x)$ para la cual G sea un I-mapa.

La construcción del grafo no dirigido es una tarea que ha de ser realizada por un experto en el tema de estudio, o inferida de un conjunto de datos

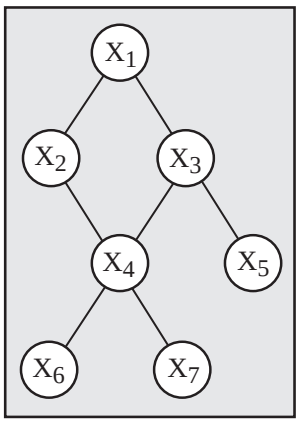


FIGURA 6.5. I-mapa minimal no dirigido asociado a la función de probabilidad dada en (6.13).

```
P[x1_,x2_,x3_,x4_,x5_]=f1[x1,x2,x3]*f2[x1,x3,x4]*
f3[x1,x4,x5]
P1=Sum[P[x[1],x[2],x[3],x[4],x[5]],{x[#1],0,1}]&;
P2=Sum[P[x[1],x[2],x[3],x[4],x[5]],
    {x[#1],0,1},{x[#2],0,1}]&;
Do[
    Do[
        a=Simplify[P[x[1],x[2],x[3],x[4],x[5]]*
            P2[i,j]-P1[i]*P1[j]];
        If[a==0,Print["Eliminar arista ",i,"--",j]],
    {j,i+1,5}],
{i,1,5}]
```

FIGURA 6.6. Programa de *Mathematica* para encontrar el I-mapa minimal no dirigido correspondiente a la función de probabilidad dada en (6.14).

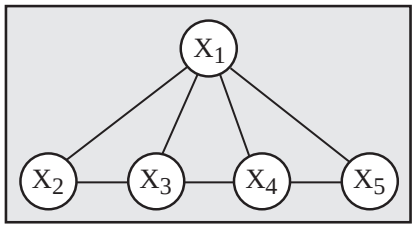


FIGURA 6.7. I-mapa minimal no dirigido asociado a la función de probabilidad dada en (6.14).

mediante alguna técnica de aprendizaje. El resto de la sección se dedica a construir una función de probabilidad que tenga a un grafo dado como I-mapa. Primero se necesitan algunas definiciones.

Definición 6.5 Probabilidad factorizada por un grafo no dirigido. *Una función de probabilidad $p(x)$ se dice que está factorizada por un grafo no dirigido G si puede escribirse de la forma*

$$p(x) = \prod_{i=1}^{m} \psi_i(c_i), \tag{6.15}$$

donde $\psi_i(c_i), i = 1, \ldots, m$, son funciones no negativas; C_i, $i = 1, \ldots, m$, son los conglomerados de G; y c_i es una realización de C_i. Las funciones ψ_i se llaman factores potenciales de la función de probabilidad, y el par $(\{C_1, \ldots, C_m\}, \{\psi_1, \ldots, \psi_m\})$ se denomina representación potencial.

Esta definición ilustra la idea de obtener el modelo probabilístico asociado a un grafo por medio de una factorización dada por la topología del grafo. La factorización resultante contendrá las independencias locales contenidas en el grafo.

Los teoremas siguientes relacionan los conceptos de I-mapa y factorización dada por un grafo no dirigido (ver Lauritzen y otros (1990)).

Teorema 6.4 Implicaciones de las factorizaciones. *Dada una función de probabilidad arbitraria $p(x)$ y un grafo no dirigido G, si $p(x)$ se puede factorizar según G, entonces G es un I-mapa de $p(x)$.*

Por tanto, cualquier independencia obtenida del grafo será también una independencia del modelo probabilístico $p(x)$. El teorema anterior implica varias propiedades locales como, por ejemplo, la *propiedad local de Markov*, que dan información sobre la estructura de independencia local de un modelo probabilístico. Por ejemplo la propiedad local de Markov afirma que para cualquier nodo $X_i \in X$ se tiene

$$I(X_i, X \setminus (\{X_i\} \cup Frn(X_i)) | Frn(X_i))_G,$$

y, por tanto,

$$p(x_i | x \setminus x_i) = p(x_i | Frn(X_i)),$$

donde $Frn(X_i)$ representa la frontera de X_i en el grafo G (ver Definición 4.13).

Teorema 6.5 Factorización de probabilidades no extremas. *Dada una función de probabilidad no extrema $p(x)$ y un grafo G, las dos condiciones siguientes son equivalentes:*

- *$p(x)$ factoriza según G.*
- *G es un I-mapa de $p(x)$.*

Dado que toda función de probabilidad no extrema $p(x)$ posee un único I-mapa minimal no dirigido, siempre se puede factorizar $p(x)$ según su I-mapa minimal.

Teorema 6.6 Factorización según un I-mapa minimal. *Toda función de probabilidad no extrema factoriza según su I-mapa minimal no dirigido asociado.*

Dado un grafo no dirigido G, el siguiente algoritmo sugiere un procedimiento para construir una función de probabilidad factorizada a partir de este grafo (ver, por ejemplo, Isham (1981) o Lauritzen (1982)).

Algoritmo 6.2 Modelo probabilístico de un grafo no dirigido.

- **Datos:** Un grafo no dirigido G.
- **Resultados:** Una función de probabilidad $p(x)$ que tiene a G como I-mapa.

1. Identificar todos los conglomerados $\{C_1, \ldots, C_m\}$ del grafo.
2. Asignar a cada conglomerado C_i una función no negativa $\psi_i(c_i)$ (el factor potencial).
3. Construir el producto de todos los factores potenciales.
4. Normalizar la función obtenida:

$$p(x_1, \ldots, x_n) = \frac{\prod_{i=1}^{m} \psi_i(c_i)}{\sum_{x_1,\ldots,x_n} \prod_{i=1}^{m} \psi_i(c_i)}. \tag{6.16}$$

■

Los resultados anteriores garantizan que el grafo no dirigido G será un I-mapa del modelo probabilístico resultante. Sin embargo, las funciones potenciales que definen el modelo probabilístico no tienen un sentido físico claro y la asignación de valores numéricos para definir un modelo concreto no es una tarea sencilla. A continuación se muestra que los grafos triangulados permiten obtener una factorización de la función de probabilidad como producto de funciones de probabilidad condicionada. Los modelos resultantes se conocen como *modelos descomponibles*.

Definición 6.6 Modelo probabilístico descomponible. *Un modelo probabilístico se dice descomponible si tiene un I-mapa minimal que es triangulado (o cordal).*

La propiedad de intersección dinámica (ver Definición 4.34) permite obtener una factorización de la función de probabilidad a partir de un grafo triangulado. Sean $\{C_1, \ldots, C_m\}$ los conglomerados ordenados de forma que cumplan la propiedad de intersección dinámica.[1] Sean

$$S_i = C_i \cap (C_1 \cup \ldots \cup C_{i-1}),\ i = 2, \ldots, m \tag{6.17}$$

los conjuntos separadores. Dado que $S_i \subset C_i$, se definen los conjuntos residuales como

$$R_i = C_i \setminus S_i. \tag{6.18}$$

En esta situación, la propiedad de intersección dinámica garantiza que los conjuntos separadores S_i separan los conjuntos residuales R_i de los conjuntos $(C_1 \cup \ldots \cup C_{i-1}) \setminus S_i$ en el grafo no dirigido. Dado que el conjunto residual R_i contiene todos los elementos de C_i que no están en $C_1 \cup \ldots \cup C_{i-1}$, también se tiene $I(R_i, R_1 \cup \ldots \cup R_{i-1} | S_i)$. Este hecho permite factorizar la función de probabilidad aplicando la regla de la cadena a la partición dada por los conjuntos residuales (ver Pearl (1988) y Lauritzen y Spiegelhalter (1988)):

$$\begin{aligned} p(x_1, \ldots, x_n) &= \prod_{i=1}^{m} p(r_i | r_1, \ldots, r_{i-1}) \\ &= \prod_{i=1}^{m} p(r_i | s_i), \end{aligned} \tag{6.19}$$

donde m es el número de conglomerados. Obsérvese que (6.19) muestra una factorización de la función de probabilidad mediante funciones de probabilidad condicionada. Por tanto, el anterior es un procedimiento práctico para obtener la función de probabilidad asociada a un grafo no dirigido triangulado. Esto sugiere los siguientes teorema y algoritmo.

Teorema 6.7 Modelos descomponibles. *Si $p(x)$ es descomponible según G, entonces puede escribirse como el producto de las funciones de probabilidad condicionada de los residuos de los conglomerados de G, dados los correspondientes conjuntos separadores.*

Algoritmo 6.3 Factorización de un modelo descomponible.

- **Datos:** Un grafo no dirigido triangulado G.
- **Resultados:** Una factorización de la función de probabilidad $p(x)$ para la cual G es un I-mapa.

1. Identificar todos los conglomerados del grafo.

[1] Los conglomerados pueden ser ordenados utilizando el Algoritmo 4.3.

2. Utilizar el Algoritmo 4.3 para ordenar los conglomerados $\{C_1, \ldots, C_m\}$ de forma que satisfagan la propiedad de intersección dinámica.

3. Calcular los conjuntos separadores $S_i = C_i \cap (C_1 \cup \ldots \cup C_{i-1})$ y los residuos $R_i = C_i \setminus S_i$.

4. Obtener $p(x)$ como
$$p(x) = \prod_{i=1}^{m} p(r_i|s_i).$$

■

La ecuación (6.19) indica que, en el caso de modelos descomponibles, las funciones potenciales en (6.16) pueden ser definidas como $\psi_i(c_i) = p(r_i|s_i)$, $i = 1, \ldots, m$. Obsérvese que éste es uno de los ejemplos de factorización a los que nos referíamos en la Sección 5.5. Una ventaja de los modelos descomponibles es que los factores potenciales resultantes son fácilmente interpretables, ya que las funciones potenciales pueden interpretarse como funciones de probabilidad condicionada. Otra ventaja importante es que la estructura local dada por la factorización permite calcular probabilidades marginales y condicionadas de modo eficiente (ver Capítulo 8). Este resultado da una solución al problema siguiente para el caso de grafos no dirigidos

- **Problema 6.5:** Dado un grafo G, ¿existe algún modelo probabilístico P tal que G sea un I-mapa minimal de P?. En caso afirmativo, ¿Cómo se puede construir?

La Sección 6.4.2 analiza el caso de los grafos dirigidos.

Ejemplo 6.8 Factorización mediante un grafo triangulado. Dado el grafo no dirigido triangulado de la Figura 6.8(a), puede aplicarse el Algoritmo 6.3 para obtener una factorización de la función de probabilidad descomponible asociada. La Figura 6.8(a) muestra los conglomerados de este grafo, que pueden ser ordenados para que cumplan la propiedad de intersección dinámica de la forma siguiente: $C_1 = \{X_1, X_2, X_3\}$, $C_2 = \{X_2, X_3, X_4\}$, $C_3 = \{X_3, X_5\}$, $C_4 = \{X_4, X_6\}$, $C_5 = \{X_4, X_7\}$. Para probar ésto, basta ver que $C_1 \cap C_2 = \{X_2, X_3\} \subset C_1$, $C_3 \cap (C_1 \cup C_2) = \{X_3\}$, que está contenido en C_1 y en C_2, y así sucesivamente. Los árboles de unión proporcionan una interpretacion gráfica de esta propiedad. Por ejemplo, el Algoritmo 4.4 utiliza la propiedad de intersección dinámica para construir un árbol de unión asociado a un grafo triangulado uniendo cada conglomerado con otro conglomerado cualquiera que contenga su conjunto separador. Por ejemplo, el grafo de la Figura 6.8(b) muestra uno de los posibles árboles de unión correspondientes al grafo no dirigido de la Figura 6.8(a). La Figura 6.9 muestra los conjuntos separadores S_2, S_3, S_4, y S_5 para cada uno de los conglomerados del árbol de unión de la Figura 6.8(b).

Utilizando los separadores S_i y los residuos R_i, se puede obtener una factorización de $p(x)$ de forma sencilla. La Tabla 6.2 muestra los conjuntos S_i y R_i asociados al conglomerado C_i. A partir de esta tabla se tiene

$$\begin{aligned} p(x) &= \prod_{i=1}^{5} p(r_i|s_i) \\ &= p(x_1,x_2,x_3)p(x_4|x_2,x_3)p(x_5|x_3)p(x_6|x_4)p(x_7|x_4), \quad (6.20) \end{aligned}$$

que es la función de probabilidad $p(x)$ que tiene al grafo no dirigido de la Figura 6.8(a) como I-mapa minimal. ∎

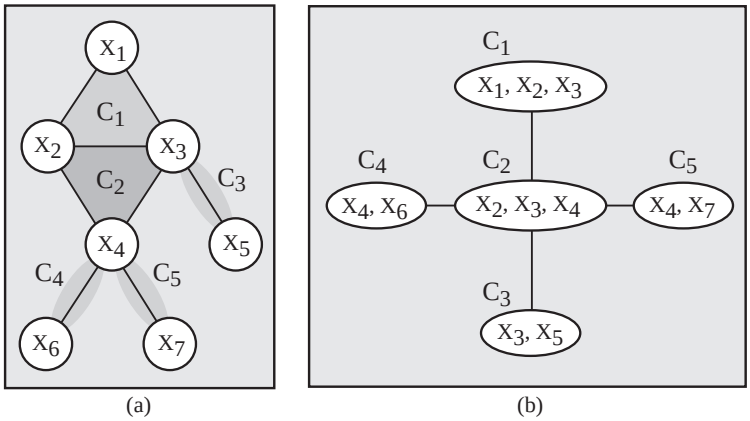


FIGURA 6.8. Un grafo triangulado y los conglomerados asociados (a), y uno de sus árboles de unión (b).

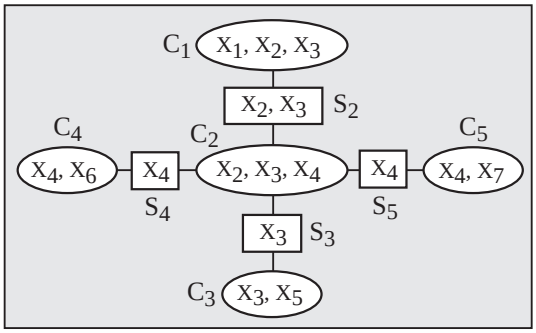


FIGURA 6.9. Conjuntos separadores correspondientes a los conglomerados de la Figura 6.8(a).

i	Conglomerado C_i	Separador S_i	Residuo R_i
1	X_1, X_2, X_3	ϕ	X_1, X_2, X_3
2	X_2, X_3, X_4	X_2, X_3	X_4
3	X_3, X_5	X_3	X_5
4	X_4, X_6	X_4	X_6
5	X_4, X_7	X_4	X_7

TABLA 6.2. Separadores y residuos correspondientes a los conglomerados de la Figura 6.9.

Obsérvese que si un I-mapa minimal G de un modelo probabilístico $p(x)$ no es triangulado, entonces es posible factorizar $p(x)$ de la forma (6.19) según alguna de las triangulaciones del I-mapa. En este caso, se perderá alguna relación de independencia del modelo en el proceso de triangulación. Por tanto, para obtener una función de probabilidad descomponible a partir de un grafo G, primero se necesita triangular el grafo, en caso de que no sea triangulado. En esta situación, alguna de las relaciones de independencia contenidas en el grafo no triangulado original no estarán contenidas en el modelo probabilístico resultante, a no ser que se creen asignando valores numéricos apropiados a las funciones de probabilidad condicionada resultantes.

Ejemplo 6.9 Factorización mediante un grafo triangulado. Supóngase que el grafo de la Figura 6.7 describe las relaciones existentes en un conjunto de cinco variables $X = \{X_1, \ldots, X_5\}$. Este grafo es triangulado y contiene tres conglomerados:

$$C_1 = \{X_1, X_2, X_3\}, \quad C_2 = \{X_1, X_3, X_4\}, \quad C_3 = \{X_1, X_4, X_5\}. \quad (6.21)$$

La ordenación de los conglomerados (C_1, C_2, C_3) cumple la propiedad de intersección dinámica. Los separadores de estos conglomerados son $S_1 = \phi$, $S_2 = \{X_1, X_3\}$ y $S_3 = \{X_1, X_4\}$. Por tanto, el modelo probabilístico que tiene a este grafo por I-mapa minimal no dirigido puede expresarse como

$$p(x) = p(x_1, x_2, x_3)p(x_4|x_1, x_3)p(x_5|x_1, x_4). \quad (6.22)$$

Dado que el grafo es triangulado, la función de probabilidad correspondiente es descomponible. Obsérvese que (6.22) tiene la misma estructura que la función de probabilidad dada en (6.14), que fue utilizada para construir este grafo. Se tiene:

$$\psi_1(x_1, x_2, x_3) = p(x_1, x_2, x_3),$$
$$\psi_2(x_1, x_3, x_4) = p(x_4|x_1, x_3),$$
$$\psi_3(x_1, x_4, x_5) = p(x_5|x_1, x_4).$$

Por tanto, en este caso ha sido posible recuperar la función de probabilidad a partir de su I-mapa no dirigido. ∎

i	Conglomerado C_i	Separador S_i	Residuo R_i
1	X_1, X_2, X_3	ϕ	X_1, X_2, X_3
2	X_1, X_3, X_4	X_1, X_3	X_4
3	X_1, X_4, X_5	X_1, X_4	X_5

TABLA 6.3. Separadores y residuos correspondientes a los conglomerados de la Figura (6.21).

Dado que los grafos de los Ejemplos 6.8 y 6.9 son triangulados, tienen asociados modelos probabilísticos descomponibles. El siguiente ejemplo muestra una factorización a partir de un grafo no triangulado.

Ejemplo 6.10 Modelo probabilístico de un grafo no triangulado. Considérese el grafo no triangulado de la Figura 6.5. Una de las posibles triangulaciones se muestra en la Figura 6.8(a). En el Ejemplo 6.8 se mostró que la función de probabilidad descomponible según este grafo podía ser factorizada como

$$p(x_1, \ldots, x_7) = p(x_1, x_2, x_3)p(x_4|x_2, x_3)p(x_5|x_3)p(x_6|x_4)p(x_7|x_4). \quad (6.23)$$

Obsérvese que la función de probabilidad dada en (6.23) está factorizada según el grafo de la Figura 6.8(a). Por tanto, el grafo original de la Figura 6.5 no es un I-mapa de la función de probabilidad dada en (6.23), a no ser que se impongan algunas restricciones sobre los parámetros para asegurar que las independencias del grafo perdidas en el proceso de triangulación se sigan cumpliendo en la función de probabilidad. Comparando la factorización $p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)$ en (6.23) con la dada en (6.13) se puede ver que el modelo contiene relaciones de dependencia que no están contenidas en el modelo original. Por tanto, la restricción que ha de imponerse a (6.23) para que tenga al grafo triangulado de la Figura 6.5 como I-mapa minimal es: $p(x_3|x_1, x_2) = p(x_3|x_1)$. ∎

6.3.3 Redes de Markov

En las secciones anteriores se ha analizado la relación existente entre grafos no dirigidos y modelos de dependencia. En esta sección se presenta una clase importante de modelos de dependencia associados a I-mapas no dirigidos. Esta clase se conoce como redes de Markov.

Definición 6.7 Red de Markov. *Una red de Markov es un par (G, Ψ) donde G es un grafo no dirigido y $\Psi = \{\psi_1(c_1), \ldots, \psi_m(c_m)\}$ es un conjunto de funciones potenciales definidas en los conglomerados $C_1, \ldots, C_m$ de G (ver Definición 6.5) que definen una función de probabilidad $p(x)$ por medio*

de

$$p(x) = \prod_{i=1}^{n} \psi_i(c_i). \tag{6.24}$$

Si el grafo no dirigido G es triangulado, entonces $p(x)$ también puede ser factorizada, utilizando las funciones de probabilidad condicionada $P = \{p(r_1|s_1), \ldots, p(r_m|s_m)\}$, de la forma siguiente

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{m} p(r_i|s_i), \tag{6.25}$$

donde R_i y S_i son los residuos y separadores de los conglomerados definidos en (6.17) y (6.18). En este caso, la red de Markov viene dada por (G, P). El grafo G es un I-mapa no dirigido de $p(x)$.

Por tanto, una red de Markov puede ser utilizada para definir la estructura cualitativa de un modelo probabilístico mediante la factorización de la función de probabilidad correspondiente a través de funciones potenciales o funciones de probabilidad condicionada. La estructura cuantitativa del modelo corresponderá a los valores numéricos concretos asignados a las funciones que aparezcan en la factorización.

Ejemplo 6.11 Red de Markov. En este ejemplo se construye una red de Markov utilizando el grafo no dirigido triangulado G dado en la Figura 6.10(a). La Figura 6.10(b) muestra los conglomerados de este grafo:

$$\begin{array}{ll} C_1 = \{A, B, C\}, & C_2 = \{B, C, E\}, \\ C_3 = \{B, D\}, & C_4 = \{C, F\}. \end{array} \tag{6.26}$$

Aplicando (6.24), se obtiene la siguiente factorización asociada al grafo:

$$\begin{aligned} p(a, b, c, d, e, f) &= \psi_1(c_1)\psi_2(c_2)\psi_3(c_3)\psi_4(c_4) \\ &= \psi_1(a, b, c)\psi_2(b, c, e)\psi_3(b, d)\psi_4(c, f). \end{aligned} \tag{6.27}$$

Por tanto, la red de Markov está definida por el grafo G y el conjunto de funciones potenciales $\Psi = \{\psi_1(a, b, c), \psi_2(b, c, e), \psi_3(b, d), \psi_4(c, f)\}$.

Por otra parte, como el grafo dado en la Figura 6.10(a) es triangulado, puede obtenerse una factorización alternativa de la función de probabilidad por medio de las funciones de probabilidad condicionada dadas en (6.25). Para obtener esta factorización es necesario ordenar los conglomerados de forma que cumplan la propiedad de intersección dinámica. Puede comprobarse fácilmente que la ordenación (C_1, C_2, C_3, C_4) en (6.26) cumple esta propiedad. La Tabla 6.4 muestra los separadores y residuos correspondientes a esta ordenación de los conglomerados (ver Figura 6.11). A partir de esta tabla y de la ecuación (6.25), se tiene:

$$p(a, b, c, d, e, f) = \prod_{i=1}^{4} p(r_i|s_i)$$

i	Conglomerado C_i	Separador S_i	Residuo R_i
1	A, B, C	ϕ	A, B, C
2	B, C, E	B, C	E
3	B, D	B	D
4	C, F	C	F

TABLA 6.4. Separadores y residuos correspondientes a los conglomerados del grafo de la Figura 6.10(a).

$$= \quad p(a, b, c)p(e|b, c)p(d|b)p(f|c). \tag{6.28}$$

Por tanto, otra forma de obtener una red de Markov asociada al grafo de la Figura 6.10(a) es por medio de las funciones de probabilidad $P = \{p(a, b, c), p(e|b, c), p(d|b), p(f|c)\}$. La Tabla 6.5 muestra una asignación de valores numéricos para estas funciones de probabilidad. Obsérvese que, en este caso, cada una de las funciones potenciales en (6.27) puede ser definida por medio de la correspondiente función de probabilidad condicionada en (6.28). Por tanto, (G, Ψ) y (G, P) son dos representaciones equivalentes de la misma red de Markov. ∎

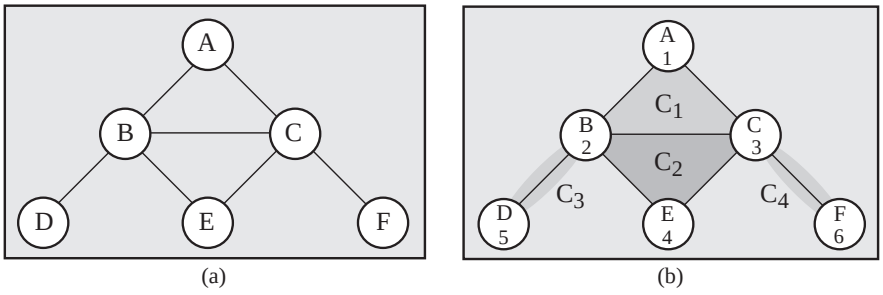


FIGURA 6.10. Grafo no dirigido triangulado (a) y sus conglomerados (b).

6.4 Modelos de Dependencia en Gráficos Dirigidos

La principal deficiencia de los grafos no dirigidos es su incapacidad para representar relaciones de independencia no transitivas; en estos modelos, dos variables independientes estarán conectadas en el grafo siempre que exista alguna otra variable que dependa de ambas. Por tanto, numerosos modelos de dependencia útiles desde un punto de vista práctico no pueden

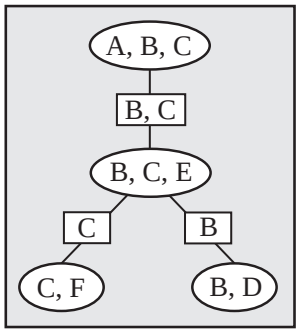


FIGURA 6.11. Árbol de unión con conjuntos separadores.

a	b	c	$p(a,b,c)$
0	0	0	0.024
0	0	1	0.096
0	1	0	0.036
0	1	1	0.144
1	0	0	0.035
1	0	1	0.035
1	1	0	0.315
1	1	1	0.315

e	b	c	$p(e\|b,c)$
0	0	0	0.4
0	0	1	0.6
0	1	0	0.5
0	1	1	0.5
1	0	0	0.7
1	0	1	0.3
1	1	0	0.2
1	1	1	0.8

f	c	$p(f\|c)$
0	0	0.1
0	1	0.9
1	0	0.4
1	1	0.6

b	d	$p(d\|b)$
0	0	0.3
0	1	0.7
1	0	0.2
1	1	0.8

TABLA 6.5. Ejemplo de asignación numérica (estructura cualitativa) de las funciones de probabilidad condicionada que factorizan la función de probabilidad en (6.28).

ser representados por grafos no dirigidos. En el Ejemplo 6.1 se mostró un modelo de dependencia muy simple, $M = \{I(X, Y|\phi)\}$, que no puede ser representado por un grafo no dirigido, ya que no cumple la propiedad de transitividad: se tiene $I(X, Y|\phi)$, pero $D(X, Z|\phi)$ y $D(Y, Z|\phi)$. Los modelos basados en grafos dirigidos permiten solventar esta deficiencia utilizando la direccionalidad de las aristas del grafo, que permiten distinguir dependencias en varios contextos. Por ejemplo, el grafo dirigido mostrado en la Figura 6.2 cumple $I(X, Y|\phi)$, $D(X, Z|\phi)$, y $D(Y, Z|\phi)$. Por tanto, el grafo es una representación perfecta del modelo no transitivo anterior.

Otro ejemplo de este fenómeno lo constituye el grafo mostrado en la Figura 6.12, donde la *felicidad* está determinada por la *salud*, el *dinero* y el *amor*. La disposición convergente de las aristas en el grafo significa que la salud, el dinero y el amor son varibles incondicionalmente independientes, pero podrían resultar dependientes si se dispone de información sobre la felicidad.

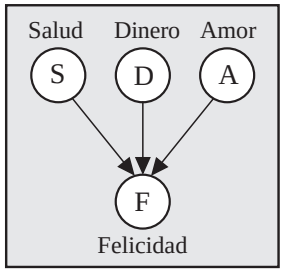


FIGURA 6.12. Modelo de dependencia no transitivo.

Una vez que se ha comprobado que los grafos no dirigidos no proporcionan una metodología general para representar modelos de dependencia, en esta sección se analizan las representaciones gráficas por medio de grafos dirigidos acíclicos. El tratamiento de grafos dirigidos sigue unas pautas análogas al tratamiento realizado en la Sección 6.3 para grafos no dirigidos. Se comienza analizando el problema de la representación de un modelo probabilístico por medio de un grafo dirigido. A continuación se presenta un tipo muy importante de modelos probabilísticos conocidos por redes Bayesianas, que se construyen a partir de grafos dirigidos acíclicos.

6.4.1 De Modelos a Grafos Dirigidos

En esta sección se analiza el problema de la representación de un modelo de dependencia por medio de un grafo dirigido. En primer lugar es conveniente recordar que el criterio gráfico de D-separación sólo es válido para la clase de grafos dirigidos acíclicos. Por tanto, los modelos gráficos dirigidos habrán de definirse siempre sobre uno de estos grafos. Por tanto, cuando se hable de un grafo dirigido en este contexto, se estará suponiendo implícitamente que es un grafo dirigido acíclico. Para el caso de grafos no dirigidos, el Teorema 6.1 da una caracterización completa del tipo de modelo que puede representarse de forma perfecta. El teorema siguiente da una condición necesaria para que un modelo de dependencia tenga una representación perfecta por medio de un grafo dirigido (ver, por ejemplo, Pearl (1988)).

Teorema 6.8 Condición necesaria para la existencia de un mapa perfecto dirigido. *Una condición necesaria para que un modelo de de-*

pendencia M posea un mapa perfecto dirigido es que cumpla las siguientes propiedades:

- **Simetría:**
$$I(X, Y|Z)_M \Leftrightarrow I(Y, X|Z)_M.$$
- **Composición-Descomposición:**
$$I(X, Y \cup W|Z)_M \Leftrightarrow I(X, Y|Z)_M \text{ y } I(X, W|Z)_M.$$
- **Intersección:**
$$I(X, W|Z \cup Y)_M \text{ y } I(X, Y|Z \cup W)_M \Rightarrow I(X, Y \cup W|Z)_M.$$
- **Unión débil:**
$$I(X, Y \cup Z|W)_M \Rightarrow I(X, Y|W \cup Z)_M.$$
- **Transitividad débil:**
$$I(X, Y|Z)_M \text{ y } I(X, Y|Z \cup A)_M \Rightarrow I(X, A|Z)_M \text{ o } I(Y, A|Z)_M,$$
donde A es una variable que no está contenida en $\{X, Y, Z\}$.
- **Contracción:**
$$I(X, Y|Z \cup W)_M \text{ y } I(X, W|Z)_M \Rightarrow I(X, Y \cup W|Z)_M.$$
- **Cordalidad:**
$$I(A, B|C \cup D)_M \text{ y } I(C, D|A \cup B)_M \Rightarrow I(A, B|C)_M \text{ o } I(A, B|D)_M,$$
donde A, B, C y D son conjuntos de una única variable.

El Ejemplo 6.2 muestra un modelo de dependencia M que no posee ningún mapa perfecto dirigido. Puede verse que, por ejemplo, el modelo no cumple la propiedad de intersección y, por tanto, no satisface las condiciones necesarias para tener un mapa perfecto dirigido.

Como ya se ha mencionado anteriormente, el Teorema 6.8 sólo proporciona una condición necesaria para que un modelo de dependencia tenga un mapa perfecto dirigido, pero esta condición no constituye una caracterización completa, pues existen modelos de dependencia que satisfacen esta condición y, sin embargo, no poseen un mapa perfecto dirigido. El ejemplo siguiente, sugerido por Milam Studený, muestra uno de estos modelos.

Ejemplo 6.12 Modelo de dependencia sin mapa perfecto que cumple el Teorema 6.8. El modelo de dependencia definido como

$$M = \{I(X, Y|Z), I(Y, X|Z), I(X, Y|W), I(Y, X|W)\} \tag{6.29}$$

cumple las siete propiedades del Teorema 6.8, pero no posee ningún mapa perfecto dirigido. ∎

Desafortunadamente, los modelos probabilísticos pueden violar las propiedades de transitividad débil, composición (ver Ejemplo 5.5), y cordalidad. Por tanto, no todo modelo probabilístico puede ser representado por un mapa perfecto dirigido. Sin embargo, como se mostrará en la Sección 6.6, la violación de la propiedad de cordalidad no es un problema grave pues puede solucionarse añadiendo nodos auxiliares al grafo.

El Teorema 6.8 proporciona una respuesta particular al Problema 6.2 para el caso de grafos dirigidos: "¿Cuáles son los modelos de dependencia y, en particular, los modelos probabilísticos de dependencia que pueden ser representados por un mapa perfecto?". El problema de si los mapas perfectos dirigidos admiten una caracterización completa mediante un conjunto finito de propiedades es un problema aún no resuelto (ver Geiger (1987)).

En aquellos casos en los que no es posible construir un mapa perfecto dirigido, la siguiente alternativa consiste en construir un I-mapa. Recordemos que un grafo dirigido D se dice que es un I-mapa de un modelo de dependencia M si $I(X, Y|Z)_D \Rightarrow I(X, Y|Z)_M$, es decir, todas las relaciones de independencia derivadas de D son ciertas en M. Un I-mapa de un modelo de dependencia se dice minimal si todas las independencias que contiene son independencias reales del modelo al que representa, pero al eliminar una cualquiera de sus aristas se incluye alguna independencia externa al modelo.

El teorema siguiente, que es equivalente al Teorema 6.3 para grafos no dirigidos, muestra una caracterización (condiciones necesarias y suficientes) para que un modelo de dependencia M tenga un I-mapa minimal dirigido (ver Verma y Pearl (1990) y Lauritzen y otros (1990)).

Teorema 6.9 I-mapa minimal dirigido de un modelo de dependencia. *Todo modelo de dependencia M de un conjunto de variables $X = \{X_1, \ldots, X_n\}$ que sea un semigrafoide, es decir, que cumpla las propiedades de simetría, descomposición, unión débil y contracción tiene un I-mapa minimal dirigido. Este I-mapa puede construirse considerando una ordenación arbitraria de las variables $(Y_1, \ldots, Y_n)$ y designando como conjunto de padres de cada nodo Y_i cualquier conjunto minimal de ascendientes (nodos previos en la ordenación) Π_i que cumplan*

$$I(Y_i, B_i \setminus \Pi_i | \Pi_i)_M, \tag{6.30}$$

donde $\Pi_i \subseteq B_i = \{Y_1, \ldots, Y_{i-1}\}$.

Ejemplo 6.13 I-mapa minimal dirigido de un modelo de dependencia. Supóngase que se tiene el modelo de dependencia

$$M = \{I(A, C|B), I(C, A|B)\}$$

definido en el conjunto de variables binarias $\{A, B, C\}$. Este modelo satisface las cuatro propiedades necesarias para ser un semigrafoide (ver Teorema 6.9). Por tanto, se podrá construir un I-mapa minimal de este modelo

considerando cualquier ordenación de las variables y calculando los conjuntos de padres Π_i que cumplen (6.30). La Figura 6.13 muestra todos los I-mapas minimales posibles asociados a las distintas ordenaciones de los nodos. Por ejemplo, dada la ordenación (A, B, C), se obtienen los siguientes conjuntos de padres:

- Para el nodo A, $\Pi_A = \phi$, ya que no tiene ningún ascendiente.
- Para el nodo B, $\Pi_B = \{A\}$, ya que su único ascendiente es A, y $I(B, A|\phi)$ no se cumple en M.
- Para el nodo C, $\Pi_C = \{B\}$, puesto que M cumple $I(C, A|B)$.

El I-mapa minimal resultante se muestra en la Figura 6.13(a). Obsérvese que dos ordenaciones distintas de las variables pueden dar lugar al mismo grafo. Por ejemplo, el grafo de la Figura 6.13(c) es un I-mapa minimal asociado a las ordenaciones (B, A, C) y (B, C, A).

Los grafos mostrados en las Figuras 6.13(a), (c) y (e) son mapas perfectos dirigidos de M, mientras que los grafos mostrados en las Figuras 6.13(b) y (d) son solamente I-mapas minimales de M. ∎

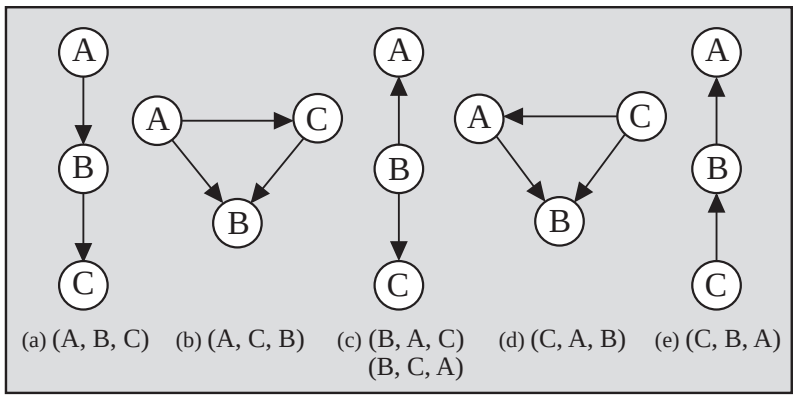


FIGURA 6.13. I-mapas minimales dirigidos asociados al modelo de dependencia M definido en el Ejemplo 6.13.

Cualquier función de probabilidad cumple las cuatro propiedades exigidas en el Teorema 6.9 (cualquier modelo de dependencia probabilístico es un semigrafoide). El teorema siguiente muestra un procedimiento para encontrar un I-mapa minimal para una función de probabilidad dada.

Teorema 6.10 I-mapa minimal dirigido para un modelo probabilístico. *Dada una permutación (una ordenación)* $Y = \{Y_1, \ldots, Y_n\}$ *de un conjunto de variables* $X = \{X_1, \ldots, X_n\}$ *y una funcion de probabilidad*

$p(x)$ de X, el grafo dirigido acíclico creado asignando como padres de cada nodo Y_i cualquier conjunto minimal de ascendientes Π_i que cumpla

$$p(y_i|b_i) = p(y_i|\pi_i), \tag{6.31}$$

para todos los valores π_i de las variables $\Pi_i \subseteq B_i = \{Y_1, \ldots, Y_{i-1}\}$, es un I-mapa minimal dirigido de $p(x)$.

En general, los subconjuntos minimales de ascendientes de la definición anterior no son únicos. Por tanto, una misma permutación de las variables puede dar lugar a varios I-mapas minimales distintos. El teorema siguiente muestra las condiciones necesarias para la unicidad de estos conjuntos y, por tanto, la unicidad del I-mapa minimal.

Teorema 6.11 I-mapa minimal de una función de probabilidad no extrema. *Si la función de probabilidad $p(x)$ es no extrema, entonces los conjuntos de padres $\{\Pi_1, \ldots, \Pi_n\}$ que cumplen (6.31) son únicos y, por tanto, el I-mapa minimal asociado también es único.*

Así, los Teoremas 6.10 y 6.11 proporcionan la respuesta al Problema 6.3 para el caso de grafos dirigidos: "¿Cuáles son los modelos de dependencia probabilísticos que poseen un único I-mapa minimal?". Estos teoremas también sugieren el siguiente algoritmo para construir el I-mapa asociado a una función de probabilidad. Este algoritmo proporciona una solución al Problema 6.4: "Si un modelo probabilístico posee un único I-mapa minimal, ¿Cómo se puede obtener este I-mapa?".

Algoritmo 6.4 I-Mapa minimal de una función de probabilidad.

- **Datos:** Un conjunto de variables X y una función de probabilidad $p(x)$.
- **Resultados:** Un I-mapa minimal D correspondiente a la función de probabilidad $p(x)$.

1. Ordenar las variables de X de forma arbitraria $(X_1, \ldots, X_n)$.
2. Para cada variable X_i, obtener un conjunto minimal de ascendientes Π_i que haga que X_i sea independendiente de $\{X_1, \ldots, X_{i-1}\} \setminus \Pi_i$.
3. Construir el grafo dirigido D que resulta de incluir una arista dirigida de cada variable de Π_i a la variable X_i. ■

El grafo dirigido acíclico resultante es un I-mapa minimal dirigido de $p(x)$ en el sentido que no puede eliminarse ninguna arista del grafo sin destruir su carácter de I-mapa. El ejemplo siguiente ilustra este algoritmo.

Ejemplo 6.14 I-mapa minimal de una distribución normal. Considérese una distribución normal multivariada de un conjunto de variables (X_1, X_2, X_3, X_4) definida por el vector de medias y la matriz de covarianzas siguientes

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} 1 & 1/2 & 1/8 & 1/4 \\ 1/2 & 1 & 1/4 & 1/2 \\ 1/8 & 1/4 & 1 & 0 \\ 1/4 & 1/2 & 0 & 1 \end{pmatrix}.$$

De la matriz de covarianzas, puede deducirse que el único par de variables independientes es (X_3, X_4) ($\sigma_{34} = \sigma_{43} = 0$). A continuación se aplica el Algoritmo 6.4 para obtener los I-mapas minimales correspondientes a dos ordenaciones distintas de las variables:

$$(X_1, X_2, X_3, X_4) \quad \text{y} \quad (X_4, X_3, X_2, X_1).$$

El proceso indicado en la Etapa 2 del algoritmo necesita conocer si se cumplen ciertas relaciones de independencia. Con este fin, se utiliza el Teorema 6.2 para calcular las funciones de probabilidad condicionada necesarias. La Tabla 6.6 muestra las medias y varianzas de las variables normales $X_i|\pi_i$ que aparecen en el proceso de construcción del I-mapa. La Figura 6.14 muestra un programa en *Mathematica* que calcula las medias y varianzas de las funciones de probabilidad condicionada asociadas a una distribución normal (ver Teorema 6.2).

Suponiendo que la media es cero, la Tabla 6.6 muestra las probabilidades condicionadas necesarias para el proceso de construcción del I-mapa. Por ejemplo, la media y la varianza condicionadas de $(X_4|X_1, X_2, X_3)$, que es la primera variable en la tabla, pueden obtenerse de la forma siguiente

```
In:=M={0,0,0,0};
    V={{1,1/2,1/8,1/4},{1/2,1,1/4,1/2},
       {1/8,1/4,1,0},{1/4,1/2,0,1}};
    CondMedVar[4,{3,2,1},M,V]

Out:=Media = 2 (4 x2 - x3)/15
     Varianza = 11/15
```

Para el orden de las variables (X_1, X_2, X_3, X_4), se tiene que

$$p(x_2|x_1) \neq p(x_2), \quad p(x_3|x_2) \neq p(x_3), \quad p(x_4|x_3, x_2, x_1) = p(x_4|x_3, x_2).$$

Esta información se muestra en las dos primeras columnas de la Tabla 6.7. De forma similar, dado el orden (X_4, X_3, X_2, X_1), se tiene

$$p(x_3|x_4) = p(x_3),$$

$$p(x_2|x_3, x_4) = p(x_2|x_3) \ \text{ o } \ p(x_2|x_3, x_4) = p(x_2|x_4),$$

$$p(x_1|x_2, x_3, x_4) = p(x_1|x_2),$$

```
CondMedVar[i_,CondVar_,M_,V_]:=
  Module[{
   Listvar={x1,x2,x3,x4,x5,x6,x7,x8,x9,x10},
   dim=Length[M],n=Length[CondVar]},
   w11=Array[v11,1];
   w21=Array[v21,{n,1}];
   w12=Array[v12,{1,n}];
   w22=Array[v22,{n,n}];
   wchi=Array[chi,n];
   wz=Array[variab,n];
   v11[1]=V[[i]][[i]];
   weta={M[[i]]};
   Do[
       v21[k1,1]=V[[i]][[CondVar[[k1]]]];
       chi[k1]=M[[CondVar[[k1]]]];
       variab[k1]=Listvar[[CondVar[[k1]]]],
   {k1,1,n}];
   Do[
       v22[k1,k2]=V[[CondVar[[k1]]]][[CondVar[[k2]]]],
   {k1,1,n},{k2,1,n}];
   w12=Transpose[w21];
   waux=w12.Inverse[w22];
   Mean=Simplify[weta+waux.(wz-wchi)];
   wVar=Simplify[w11-waux.w21];
   Print["Media = ",Mean];
   Print["Varianza = ",wVar]
  ]
```

FIGURA 6.14. Programa en *Mathematica* para obtener la media y la varianza de una función de probabilidad condicionada asociada a la distribución normal del Ejemplo 6.14.

que se muestra en las dos últimas columnas de la Tabla 6.7. Los I-mapas resultantes se muestran en la Figura 6.15. ■

6.4.2 De Grafos Dirigidos a Modelos probabilísticos

En esta sección se analiza la forma de construir una función de probabilidad a partir de un grafo dirigido. Cuando son conocidos el modelo probabilístico o el modelo de dependencia asociado M, siempre es posible obtener un I-mapa dirigido que reproduzca tantas independencias de M como sea posible. Sin embargo, en la práctica no se suele conocer la función de pro-

X_i	π_i	Media de $X_i\|\pi_i$	Varianza de $X_i\|\pi_i$
X_4	$\{x_3, x_2, x_1\}$	$2(4x_2 - x_3)/15$	$11/15$
X_4	$\{x_2, x_1\}$	$x_2/2$	$3/4$
X_4	$\{x_3, x_1\}$	$2(8x_1 - x_3)/63$	$59/63$
X_4	$\{x_3, x_2\}$	$2(4x_2 - x_3)/15$	$11/15$
X_4	$\{x_1\}$	$x_1/4$	$15/16$
X_4	$\{x_2\}$	$x_2/2$	$3/4$
X_4	$\{x_3\}$	0	1
X_3	$\{x_2, x_1\}$	$x_2/4$	$15/16$
X_3	$\{x_1\}$	$x_1/8$	$63/64$
X_3	$\{x_2\}$	$x_2/4$	$15/16$
X_2	$\{x_1\}$	$x_1/2$	$3/4$
X_1	$\{x_2, x_3, x_4\}$	$x_2/2$	$3/4$
X_1	$\{x_3, x_4\}$	$x_3/8 + x_4/4$	$59/64$
X_1	$\{x_2, x_4\}$	$x_2/2$	$3/4$
X_1	$\{x_2, x_3\}$	$x_2/2$	$3/4$
X_1	$\{x_4\}$	$x_4/4$	$15/16$
X_1	$\{x_3\}$	$x_3/8$	$63/64$
X_1	$\{x_2\}$	$x_2/2$	$3/4$
X_2	$\{x_3, x_4\}$	$x_3/4 + x_4/2$	$11/16$
X_2	$\{x_4\}$	$x_4/2$	$3/4$
X_2	$\{x_3\}$	$x_3/4$	$15/16$
X_3	$\{x_4\}$	0	1

TABLA 6.6. Medias y varianzas condicionadas de las variables normales $(X_i|\pi_i)$ del Ejemplo 6.14.

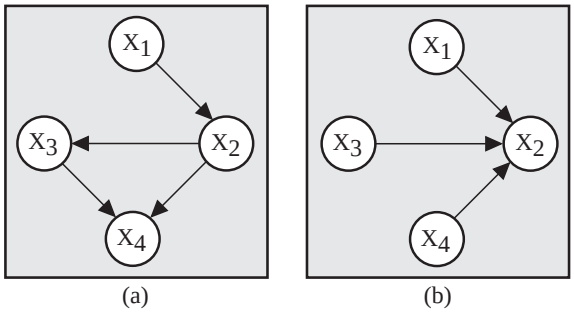


FIGURA 6.15. I-mapas minimales asociados a las ordenaciones de las variables (X_1, X_2, X_3, X_4) y (X_4, X_3, X_2, X_1).

Ordenación (X_1, X_2, X_3, X_4)		Ordenación (X_4, X_3, X_2, X_1)	
X_i	Π_i	X_i	Π_i
X_1	ϕ	X_4	ϕ
X_2	$\{X_1\}$	X_3	ϕ
X_3	$\{X_2\}$	X_2	$\{X_3, X_4\}$
X_4	$\{X_3, X_2\}$	X_1	$\{X_2\}$

TABLA 6.7. Conjuntos minimales de ascendientes que hacen que X_i sea independiente del resto de ascendientes para las dos ordenaciones indicadas de las variables.

babilidad $p(x)$ ni el modelo M. Por tanto, la forma real de construir un modelo probabilístico consiste en las siguientes etapas:

1. Construir un grafo dirigido D que describa la estructura de dependencia entre las variables de X.
2. Encontrar una función de probabilidad $p(x)$ para la cual D sea un I-mapa.

La construcción del grafo no dirigido es una tarea que ha de realizala un experto, o inferida de un conjunto de datos mediante alguna técnica de aprendizaje. En esta sección se analiza el segundo problema, es decir, obtener el modelo probabilístico asociado a un grafo.

Definición 6.8 Factorización recursiva según un grafo dirigido acíclico. *Una función de probabilidad se dice que admite una factorización recursiva según un grafo dirigido acíclico D, si la función de probabilidad se puede expresar como*

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i), \tag{6.32}$$

donde $p(x_i|\pi_i)$ es la función de probabilidad de X_i condicionada a sus padres Π_i.

Teorema 6.12 Factorización recursiva. *Sea D un grafo dirigido acíclico y $p(x)$ un modelo probabilístico de X. Entonces las siguientes condiciones son equivalentes:*

1. *$p(x)$ admite una factorización recursiva según D.*
2. *D es un I-mapa de $p(x)$.*

Por tanto, dado un grafo dirigido D, puede construirse una función de probabiilidad condicionada dadas en (6.32). En esta situación, el Teorema 6.12 permite concluir que D es un I-mapa del modelo probabilístico P resultante (ver Pearl, (1988)). Este proceso se ilustra en el ejemplo siguiente.

Ejemplo 6.15 Factorización según un grafo dirigido. Considérense los dos grafos dirigidos mostrados en la Figura 6.15. Aplicando la definición anterior, se pueden construir dos factorizaciones recursivas de la función de probabilidad asociada a estos grafos. A partir del grafo de la Figura 6.15(a), se tiene

$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= p(x_1|\pi_1)p(x_2|\pi_2)p(x_3|\pi_3)p(x_4|\pi_4) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_2, x_3), \end{aligned}$$

mientras que el grafo de la Figura 6.15(b) sugiere la factorización:

$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= p(x_1|\pi_1)p(x_2|\pi_2)p(x_3|\pi_3)p(x_4|\pi_4) \\ &= p(x_1)p(x_2|x_1, x_3, x_4)p(x_3)p(x_4). \end{aligned}$$

■

6.4.3 Modelos Causales

A pesar de que no todo modelo probabilístico puede ser representado mediante un mapa perfecto dirigido, la clase de modelos probabilísticos que poseen un mapa perfecto dirigido no es demasiado restrictiva. En esta sección se dan las condiciones necesarias para caracterizar estos modelos. Antes son necesarias unas definiciones previas.

Definición 6.9 Lista causal de relaciones de independencia. *Sea* $Y = \{Y_1, \ldots, Y_n\}$ *una permutación de* $X = \{X_1, \ldots, X_n\}$. *Una lista causal de relaciones de independencia es un conjunto que contiene n relaciones de independencia de la forma*

$$I(Y_i, B_i \setminus \Pi_i | \Pi_i), \tag{6.33}$$

una para cada variable, donde $B_i = \{Y_1, \ldots, Y_{i-1}\}$ *es el conjunto de ascendientes del nodo* Y_i, *y* Π_i *es un subconjunto de* B_i *que hace que* Y_i *sea condicionalmente independiente del resto de sus ascendiente,* $B_i \setminus \Pi_i$.

Cuando las variables están ordenadas de forma que una causa siempre precede a su efecto (es decir, los padres preceden a los hijos), entonces el conjunto minimal de ascendientes de una variable X_i que la separa del resto de sus ascendientes se denomina conjunto de *causas directas* de X_i. Este hecho aporta una interpretación a la denominación de *lista causal* utilizada para estos modelos. Se puede obtener una representación gráfica de una lista causal construyendo un grafo dirigido cuyas aristas unan cada causa directa X_j con el efecto correspondiente X_i, $X_j \rightarrow X_i$. El modelo de dependencia asociado al grafo resultante puede obtenerse completando la lista causal inicial utilizando las propiedades de semigrafoide.

Definición 6.10 Modelo causal.[2] *Un modelo causal es un modelo de dependencia probabilístico generado por una lista causal.*

Dado que lo modelos causales están asociados a un modelo probabilístico, estos modelos cumplen las propiedades de semigrafoide. Por tanto, todo modelo causal tiene asociado un I-mapa minimal que viene dado por un grafo dirigido en el que los conjuntos de padres Π_i están definidos por la condición $I(Y_i, B_i \setminus \Pi_i | \Pi_i)$, donde B_i es el conjunto de ascendientes del nodo X_i dado el orden de las variables impuesto por la lista causal. La lista causal permite construir una factorización de la función de probabilidad considerando el conjunto de funciones de probabilidad condicionada:

$$p(y_i|b_i) = p(y_i|\pi_i), \quad i = 1, \ldots, n. \tag{6.34}$$

Cada una de estas probabilidades condicionadas está definida por una de las independencias que forman la lista causal. Obsérvese que todas estas funciones son componentes canónicas estándar (ver Sección 5.5). Por tanto, cada relación de independencia de una lista causal da lugar a una componente canónica estándar del modelo probabilístico. Así, las listas causales permiten definir modelos probabilísticos de un modo muy sencillo a través de la relación anterior entre la representación gráfica y las componentes canónicas estándar de la función de probabilidad. Estas representaciones gráficas, y los modelos probabilísticos asociados se conocen como *redes Bayesianas.*

Ejemplo 6.16 Generación de listas causales. Considérese la lista causal que describe las relaciones de dependencia de un conjunto de cuatro variables de la forma indicada en la Tabla 6.8. Esta lista causal puede ser representada por el grafo dirigido acíclico dado en la Figura 6.16. Tanto la lista causal de la Tabla 6.8 como el grafo dirigido de la Figura 6.16 sugieren la siguiente factorización canónica del modelo probabilístico:

$$\begin{array}{rcl} p(x) & = & p(x_1)p(x_2|x_1)p(x_3|x_1,x_2)p(x_4|x_1,x_2,x_3) \\ & = & p(x_1)p(x_2)p(x_3|x_1,x_2)p(x_4|x_1,x_3). \end{array} \tag{6.35}$$

La segunda igualdad en (6.35) es consecuencia de la primera, aplicando

$$I(X_2, X_1|\phi) \Leftrightarrow p(x_2|x_1) = p(x_2)$$

y

$$I(X_4, X_2|X_1, X_3) \Leftrightarrow p(x_4|x_1, x_2, x_3) = p(x_4|x_1, x_3).$$

Por tanto, para definir la función de probabilidad de las cuatro variables, sólo es necesario definir las cuatro funciones de probabilidad obtenidas, que

[2] Algunos autores utilizan el término *modelo causal* para referirse a los modelos de dependencia M que tienen un mapa perfecto dirigido.

Nodo	B_i	Π_i	$I(X_i, B_i \setminus \Pi_i \mid \Pi_i)$
X_1	ϕ	ϕ	$I(X_1, \phi \mid \phi)$
X_2	X_1	ϕ	$I(X_2, X_1 \mid \phi)$
X_3	X_1, X_2	X_1, X_2	$I(X_3, \phi \mid X_1, X_2)$
X_4	X_1, X_2, X_3	X_1, X_3	$I(X_4, X_2 \mid X_1, X_3)$

TABLA 6.8. Ejemplo de una lista causal.

dependen de un conjunto menor de variables:

$$\begin{array}{ll} p(x_1), & p(x_2), \\ p(x_3|x_1, x_2), & p(x_4|x_1, x_3). \end{array}$$

■

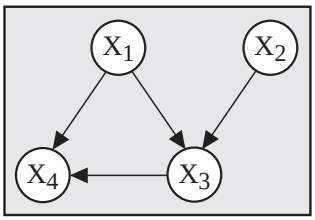


FIGURA 6.16. Grafo dirigido acíclico asociado a la lista causal definida en la Tabla 6.8.

Los siguientes teoremas ilustran las propiedades más importantes de las listas causales (ver Verma y Pearl (1990) y Geiger y Pearl (1990)).

Teorema 6.13 Clausura de una lista causal. *Sea D un grafo dirigido acíclico definido por una lista causal M. Entonces el conjunto de relaciones de independencia verificados por el grafo coincide con la clausura de M bajo las propiedades de simetría, descomposición, unión débil y contracción (propiedades de semigrafoide).*

Teorema 6.14 Completitud de las listas causales. *Sea D un grafo dirigido acíclico definido por una lista causal M. Entonces cada relación de independencia contenida en el mínimo semigrafoide generado por M también se cumple en D (utilizando el criterio de D-separación).*

El Teorema 6.13 garantiza que todas las relaciones de independencia contenidas en el grafo dirigido se pueden obtener de M utilizando las propiedades de semigrafoide. Por otra parte, el Teorema 6.14 garantiza que el grafo dirigido contiene todas las independencias que pueden ser obtenidas de M utilizando las propiedades de semigrafoide. Por tanto, las propiedades de simetría, descomposición, unión débil y contracción constituyen un

conjunto completo capaz de obtener cualquier consecuencia válida de una lista causal. Por tanto, el grafo dirigido obtenido de una lista causal es un mapa perfecto del semigrafoide generado por la lista causal.

Los grafos dirigidos son, por tanto, herramientas convenientes e intuitivas para representar relaciones de independencia condicional. Los nodos del grafo representan a las variables del problema a analizar y su topología está determinada por una lista causal que contiene una sola relación de independencia para cada variable. La lista causal asigna un conjunto de padres a cada variable X_i de forma que X_i sea condicionalmente independiente de todos sus ascendientes, dado su conjunto de padres (en algún orden establecido para las variables). Esta lista causal determina de forma recursiva la relación de cada variable con sus ascendientes. Los modelos de dependencia generados por una lista causal se denominan *modelos causales*.

6.4.4 Redes Bayesianas

En las secciones anteriores se ha analizado la relación existente entre los grafos dirigidos y los modelos de dependencia. En esta sección se presenta un tipo importante de modelos de dependencia associados a I-mapas dirigidos. Esta clase se conoce como redes Bayesianas.

Definición 6.11 Red Bayesiana. *Una red Bayesiana es un par (D, P), donde D es un grafo dirigido acíclico, $P = \{p(x_1|\pi_1), \ldots, p(x_n|\pi_n)\}$ es un conjunto de n funciones de probabilidad condicionada, una para cada variable, y Π_i es el conjunto de padres del nodo X_i en D. El conjunto P define una función de probabilidad asociada mediante la factorización*

$$p(x) = \prod_{i=1}^{n} p(x_i|\pi_i). \tag{6.36}$$

El grafo dirigido acíclico D es un I-mapa minimal de $p(x)$.

El Teorema 6.12 muestra que cualquier relación de independencia que se obtenga del grafo utilizando el criterio de D-separación también estará contenida en el modelo probabilístico correspondiente.

Obsérvese que en el caso de redes Bayesianas, la factorización de la función de probabilidad se obtiene de forma sencilla a partir del grafo dirigido considerando un conjunto de funciones de probabilidad condicionada que involucran a cada nodo con sus padres. Por otra parte, la factorización de una función de probabilidad asociada a una red de Markov requiere varios pasos previos como: construir los conglomerados del grafo, ordenar este conjunto de modo que satisfaga la propiedad de intersección dinámica, encontrar los separadores y residuos, etc. Por tanto, las redes Bayesianas ofrecen una forma más sencilla e intuitiva de construir modelos probabilísticos que las redes de Markov.

Existen diversos tipos de redes Bayesianas dependiendo del tipo de variables que contenga el grafo (discretas, continuas, o ambas) y del tipo de distribución que se considere para cada variable. Dos tipos importantes de redes Bayesianas son las *redes multinomiales* y las *redes normales* o *Gausianas*. Estos tipos se describen a continuación

Redes Bayesianas Multinomiales

En una red Bayesiana multinomial se supone que todas las variables son discretas, es decir, que cada variable puede tomar únicamente un conjunto finito de valores. En este tipo de redes también se supone que la función de probabilidad condicionada asociada a cada variable es una función de probabilidad de tipo multinomial. Este tipo de funciones de probabilidad pueden ser definidas, ya sea de forma paramétrica o numérica, por medio de tablas que asignan valores numéricos a las diferentes combinaciones de las variables involucradas. A continuación se muestra un ejemplo de una red Bayesiana multinomial.

Ejemplo 6.17 Red Bayesiana multinomial. Considérese el grafo dirigido mostrado en la Figura 6.17 y supóngase que todas las variables del conjunto $\{A, B, C, D, E, F, G\}$ son binarias, es decir, solamente pueden tomar dos valores posibles (por ejemplo, 0 ó 1). Este grafo dirigido define una red Bayesiana por medio de las funciones de probabilidad condicionada (6.36) que definen la siguiente factorización de la función de probabilidad

$$p(a, b, c, d, e, f, g) = p(a)p(b)p(c|a)p(d|a, b)p(e)p(f|d)p(g|d, e). \qquad (6.37)$$

En este caso, las funciones de probabildad condicionada son tablas de probabilidades para las diferentes combinaciones de valores de las variables. La Tabla 6.9 muestra un ejemplo de los valores numéricos necesarios para definir el conjunto de funciones de probabilidad condicionada dadas en (6.37). Por tanto, el grafo dirigido acíclico de la Figura 6.17 junto con el conjunto de probabilidades condicionadas dado en la Tabla 6.9 definen una red Bayesiana multinomial. ∎

Redes Bayesianas Gausianas

En una red Bayesiana normal o Gausiana, se supone que las variables del conjunto X están distribuidas por medio de una distribución normal $N(\mu, \Sigma)$ dada por la función de densidad

$$f(x) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{-1/2(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \qquad (6.38)$$

donde μ es el vector n-dimensional de medias, Σ es la matriz $n \times n$ de covarianzas, $|\Sigma|$ es el determinante de Σ, y μ^T denota la traspuesta de μ.

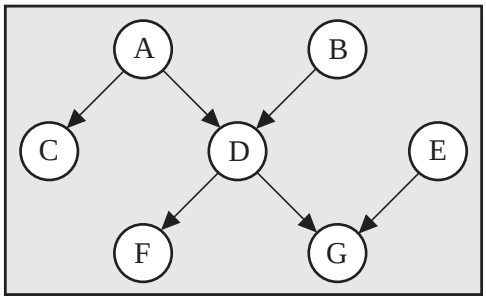


FIGURA 6.17. Grafo dirigido utilizado para construir la red Bayesiana del Ejemplo 6.17.

a	$p(a)$
0	0.3
1	0.7

b	$p(b)$
0	0.6
1	0.4

e	$p(e)$
0	0.1
1	0.9

c	a	$p(c\|a)$
0	0	0.25
0	1	0.50
1	0	0.75
1	1	0.50

f	d	$p(f\|d)$
0	0	0.80
0	1	0.30
1	0	0.20
1	1	0.70

d	a	b	$p(d\|a,b)$
0	0	0	0.40
0	0	1	0.45
0	1	0	0.60
0	1	1	0.30
1	0	0	0.60
1	0	1	0.55
1	1	0	0.40
1	1	1	0.70

g	d	e	$p(g\|d,e)$
0	0	0	0.90
0	0	1	0.70
0	1	0	0.25
0	1	1	0.15
1	0	0	0.10
1	0	1	0.30
1	1	0	0.75
1	1	1	0.85

TABLA 6.9. Funciones de probabilidad condicionada correspondientes al grafo dirigido de la Figura 6.17.

La función de densidad de las variables en una red Bayesiana Gausiana está definida por (6.36) mediante el producto de un conjunto de funciones de probabilidad condicionada dadas por

$$f(x_i|\pi_i) \sim N\left(\mu_i + \sum_{j=1}^{i-1} \beta_{ij}(x_j - \mu_j),\ v_i\right), \tag{6.39}$$

donde β_{ij} es el coeficiente de regresión de X_j en la regresión de X_i sobre los padres, Π_i, de X_i y

$$v_i = \Sigma_i - \Sigma_{i\Pi_i}\Sigma_{\Pi_i}^{-1}\Sigma_{i\Pi_i}^T$$

es la varianza condicionada de X_i, dado $\Pi_i = \pi_i$, donde Σ_i es la varianza incondicional de X_i, $\Sigma_{i\Pi_i}$ son las covarianzas entre X_i y las variables de Π_i, y Σ_{Π_i} es la matriz de covarianzas de Π_i. Obsérvese que β_{ij} mide el grado de relación existente entre las variables X_i y X_j. Si $\beta_{ij} = 0$, entonces X_j no será un padre de X_i.

Mientras que la media condicionada $\mu_{x_i|\pi_i}$ depende de los valores de los padres π_i, la varianza condicionada no depende de esos valores. Por tanto, el conjunto de funciones de probabilidad condicionada que define una red Bayesiana normal está determinado por los parámetros $\{\mu_1, \ldots, \mu_n\}$, $\{v_1, \ldots, v_n\}$, y $\{\beta_{ij} \,|\, j < i\}$, tal y como se muestra en (6.39).

Una función de probabilidad normal puede definirse de forma alternativa mediante su vector de medias μ y su matriz de precisión $W = \Sigma^{-1}$. Shachter y Kenley (1989) describen la transformación general para pasar de $\{v_1, \ldots, v_n\}$ y $\{\beta_{ij} : j < i\}$ a W. Esta transformación viene dada por la siguiente fórmula recursiva, en la cual $W(i)$ representa la matriz superior izquierda $i \times i$ de W y β_i representa el vector columna $\{\beta_{ij} : j < i\}$:

$$W(i+1) = \begin{pmatrix} W(i) + \dfrac{\beta_{i+1}\beta_{i+1}^T}{v_{i+1}} & \dfrac{-\beta_{i+1}}{v_{i+1}} \\ \dfrac{-\beta_{i+1}^T}{v_{i+1}} & \dfrac{1}{v_{i+1}} \end{pmatrix}, \tag{6.40}$$

donde $W(1) = 1/v_1$.

Por tanto, se tienen dos representaciones alternativas de la función de probabilidad de una red Bayesiana normal. El ejemplo siguiente ilustra la forma de construir una red Bayesiana normal.

Ejemplo 6.18 Red Bayesiana normal. Considérese el grafo dirigido acíclico mostrado en la Figura 6.18 y supóngase que las cuatro variables del conjunto $\{A, B, C, D\}$, están distribuidas de forma normal, es decir, $f(a, b, c, d) \sim N(\mu, \Sigma)$. El conjunto de funciones de probabilidad condicionada dado en la factorización (6.36) define la red Bayesiana normal

$$f(a, b, c, d) = f(a)f(b)f(c|a)f(d|a, b), \tag{6.41}$$

donde

$$\begin{aligned} &f(a) \sim N\left(\mu_A, v_A\right), \\ &f(b) \sim N\left(\mu_B, v_B\right), \\ &f(c) \sim N\left(\mu_C + \beta_{CA}(a - \mu_A), v_C\right), \\ &f(d) \sim N\left(\mu_D + \beta_{DA}(a - \mu_A) + \beta_{DB}(b - \mu_B), v_D\right). \end{aligned} \tag{6.42}$$

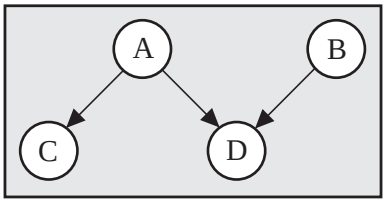


FIGURA 6.18. Grafo dirigido utilizado para construir la red Bayesiana del Ejemplo 6.18.

Este conjunto de funciones de probabilidad condicionada constituye una de las dos descripciones equivalentes de la red Bayesiana. Los parámetros asociados a esta representación son $\{\mu_A, \mu_B, \mu_C, \mu_D\}$, $\{v_A, v_B, v_C, v_D\}$, y $\{\beta_{CA}, \beta_{DA}, \beta_{DB}\}$.

Una representación alternativa puede obtenerse utilizando la fórmula recursiva (6.40). En este caso, después de cuatro iteraciones se obtiene la matriz

$$W = \begin{pmatrix} \dfrac{1}{v_A} + \dfrac{\beta_{CA}^2}{v_C} + \dfrac{\beta_{DA}^2}{v_D} & \dfrac{\beta_{DA}\,\beta_{DB}}{v_D} & -\dfrac{\beta_{CA}}{v_C} & -\dfrac{\beta_{DA}}{v_D} \\ \dfrac{\beta_{DA}\,\beta_{DB}}{v_D} & \dfrac{1}{v_B} + \dfrac{\beta_{DB}^2}{v_D} & 0 & -\dfrac{\beta_{DB}}{v_D} \\ -\dfrac{\beta_{CA}}{v_C} & 0 & \dfrac{1}{v_C} & 0 \\ -\dfrac{\beta_{DA}}{v_D} & -\dfrac{\beta_{DB}}{v_D} & 0 & \dfrac{1}{v_D} \end{pmatrix}.$$

La matriz de covarianzas de la función de probabilidad se obtiene invirtiendo la matriz anterior:

$$\Sigma = \begin{pmatrix} v_A & 0 & \beta_{CA}\,v_A & \beta_{DA}\,v_A \\ 0 & v_B & 0 & \beta_{DB}\,v_B \\ \beta_{CA}\,v_A & 0 & \beta_{CA}^2\,v_A + v_C & \beta_{CA}\,\beta_{DA}\,v_A \\ \beta_{DA}\,v_A & \beta_{DB}\,v_B & \beta_{CA}\,\beta_{DA}\,v_A & \beta_{DA}^2\,v_A + \beta_{DB}^2\,v_B + v_D \end{pmatrix}.$$

Obsérvese que, hasta ahora, todos los parámetros han sido escritos en forma simbólica. Por tanto, la estructura cualitativa de la red Bayesiana se definirá asignando valores numéricos a estos parámetros. Por ejemplo, considerando nulo el valor de las medias, uno el de las varianzas, $\beta_{CA} = 1$, $\beta_{DA} = 0.2$ y $\beta_{DB} = 0.8$, se tiene la siguiente matriz de covarianzas

$$\Sigma = \begin{pmatrix} 1.0 & 0.0 & 1.0 & 0.20 \\ 0.0 & 1.0 & 0.0 & 0.80 \\ 1.0 & 0.0 & 2.0 & 0.20 \\ 0.2 & 0.8 & 0.2 & 1.68 \end{pmatrix}. \tag{6.43}$$

Esta matriz y el vector de medias definen una red Bayesiana normal asociada al grafo dirigido de la Figura 6.18. ∎

6.5 Modelos Gráficos Equivalentes

Los modelos gráficos basados en grafos no dirigidos no son redundantes en el sentido de que grafos distintos siempre determinan modelos de dependencia distintos. Sin embargo, en el caso de grafos dirigidos, un mismo modelo de dependencia puede tener asociados varios grafos distintos. En esta sección se caracterizan las clases de grafos dirigidos que determinan un mismo modelo de dependencia. Para ello se introduce la siguiente definición de *modelos gráficos equivalentes* (Verma y Pearl (1991)).

Definición 6.12 Modelos gráficos equivalentes. *Dos modelos gráficos se dicen equivalentes si los grafos correspondientes determinan el mismo modelo de dependencia, es decir, si tienen asociado el mismo conjunto de relaciones de independencia.*

Dos redes de Markov (grafos no dirigidos) son equivalentes si y sólo si tienen asociado el mismo grafo. Sin embargo, en el caso de redes Bayesianas es posible cambiar la dirección de alguna de las aristas que componen el grafo dirigido dejando inalterado el modelo de dependencia asociado. Según ésto, dos redes Bayesianas construidas a partir de grafos diferentes pueden ser equivalentes. Por tanto, un problema interesante consiste en caracterizar las clases de redes Bayesianas que tienen asociado el mismo modelo de dependencia:

- **Problema 6.6:** ¿Cómo puede comprobarse si dos redes Bayesianas son equivalentes?

Para analizar este problema es necesario introducir algunas definiciones.

Definición 6.13 V-estructura. *Una terna de nodos (X, Z, Y) en una red Bayesiana se dice que es una v-estructura si las aristas desde los nodos X e Y convergen al nodo Z y, además, no existe ninguna arista entre los nodos X e Y. El nodo Z en una v-estructura se suele denominar como un nodo de aristas convergentes no acopladas en el camino no dirigido $X - Z - Y$.*

Las v-estructuras representan las relaciones de independencia no transitivas de una red Bayesiana. Dada una v-estructura (X, Z, Y), se tiene que X e Y son incondicionalmente independientes, pero son condicionalmente dependientes dada Z. Por ejemplo, los nodos (X, Z, Y) en el grafo dirigido de la Figura 6.2 representan una v-estructura que permite representar mediante un grafo dirigido el modelo de dependencia no transitivo del Ejemplo 6.1. En otro ejemplo, el grafo dirigido de la Figura 6.12 tiene tres v-estructuras distintas: (S, F, D), (D, F, A) y (S, F, A).

El teorema siguiente proporciona una solución al Problema 6.6 (Verma y Pearl (1991)).

Teorema 6.15 Redes Bayesianas equivalentes. *Dos redes Bayesianas se dicen equivalentes si tienen asociadas: (a) los mismos grafos no dirigidos y (b) las mismas v-estructuras.*

Los ejemplos siguientes ilustran este teorema.

Ejemplo 6.19 Redes Bayesianas equivalentes. Considérense los seis grafos dirigidos de la Figura 6.19. Los grafos (a)−(c) tienen asociado el mismo modelo de dependencia, ya que tienen asociado el mismo grafo no dirigido y no tienen v-estructuras. Los grafos (a)−(d) tienen asociado el mismo grafo no dirigido, pero el grafo (d) posee la v-estructura (X, Z, Y). Por tanto, el grafo (d) no es equivalente a ninguno de los grafos (a)−(c).

Los grafos (e) y (f) son equivalentes ya que tienen asociado el mismo grafo no dirigido y no poseen v-estructuras. Obsérvese que la terna (X, Z, Y) no es una v-estructura en el grafo (e), pues los nodos X e Y están conectados por una arista.

Por tanto, los seis grafos de la Figura 6.19 definen solamente tres clases de dependencia distintas: $\{(a), (b), (c)\}$, $\{(d)\}$ y $\{(e), (f)\}$, donde las letras indican los grafos correspondientes a cada clase en la Figura 6.19. ∎

Ejemplo 6.20 Todos los grafos completos son equivantes. Como ya se había visto anteriormente, un grafo completo contiene una arista entre cada par de nodos. Dado que todos los grafos dirigidos completos con n nodos tienen asociado el mismo grafo no dirigido (el grafo no dirigido completo de n nodos) y que no existe ninguna v-estructura en un grafo dirigido completo, se tiene que todos los grafos dirigidos completos de n nodos son equivalentes. Para ilustrar este resultado, obsérvese que existen 2^n grafos no dirigidos completos con n nodos. Por ejemplo, la Figura 6.20 muestra los $2^3 = 8$ grafos completos distintos con $n = 3$ nodos. Dado que sólo estamos considerando grafos dirigidos acíclicos, los dos últimos grafos de esta figura son excluidos, pues ambos contienen un ciclo. El número de grafos dirigidos acíclicos distintos con n nodos es $n!$, un grafo para cada orden de las variables. Por tanto, para $n = 3$ se tienen $3! = 6$ grafos dirigidos acíclicos distintos. Los primeros seis grafos de la Figura 6.20 corresponden a las siguientes ordenaciones de las variables.

$$\begin{array}{ll} (X, Y, Z), & (Y, X, Z), \\ (Z, Y, X), & (Z, X, Y), \\ (X, Z, Y), & (Y, Z, X). \end{array}$$

A los grafos (a)−(f) les corresponde el mismo modelo de dependencia, ya que tienen asociado el mismo grafo no dirigido y no contienen v-estructuras. El modelo de dependencia correspondiente a esta clase es el modelo vacío, es decir, el que no contiene ninguna relación de independencia entre las variables (modelo general de dependencia). ∎

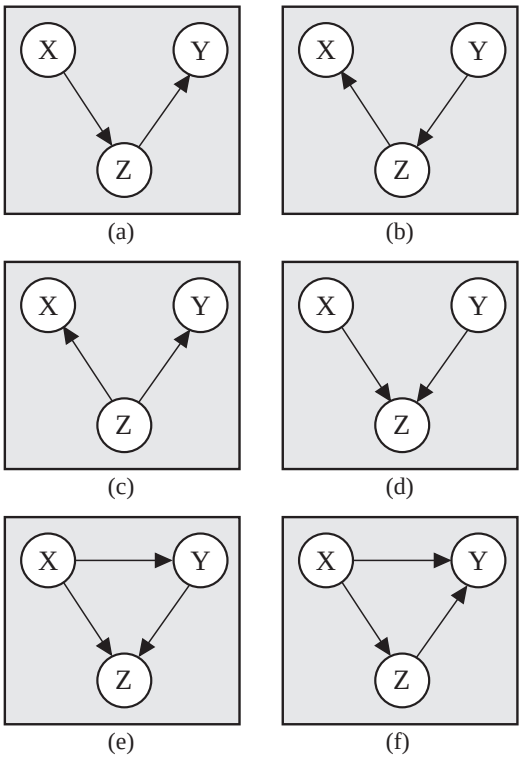


FIGURA 6.19. Grafos dirigidos distintos con tres nodos.

Una consecuencia importante de este concepto de equivalencia es que en una red Bayesiana se puede cambiar la dirección de algunas aristas sin alterar la estructura cualitativa del modelo. Este hecho motiva la siguiente definición de aristas *reversibles* e *irreversibles*.

Definición 6.14 Aristas reversibles e irreversibles. *Sea $D = (X, L)$ un grafo dirigido acíclico. Se dice que una arista $L_{ij} = (X_i \rightarrow X_j) \in L$ es irreversible si $L_{ij} \in L'$ para cualquier grafo dirigido acíclico $D' = (X, L')$ que sea equivalente a D. Cualquier arista que no sea irreversible se denomina reversible.*

Si las aristas de una red Bayesiana tienen alguna interpretación causal, el hecho de cambiar la orientación de alguna de las aristas influirá necesariamente en las relaciones causa-efecto entre las variables (aunque no altere el modelo de dependencia asociado).

Los siguientes algoritmos, debidos a Chickering (1995b), introducen primeramente una ordenación de las aristas del grafo dirigido para, seguidamente, clasificar cada arista como irreversible o como reversible.

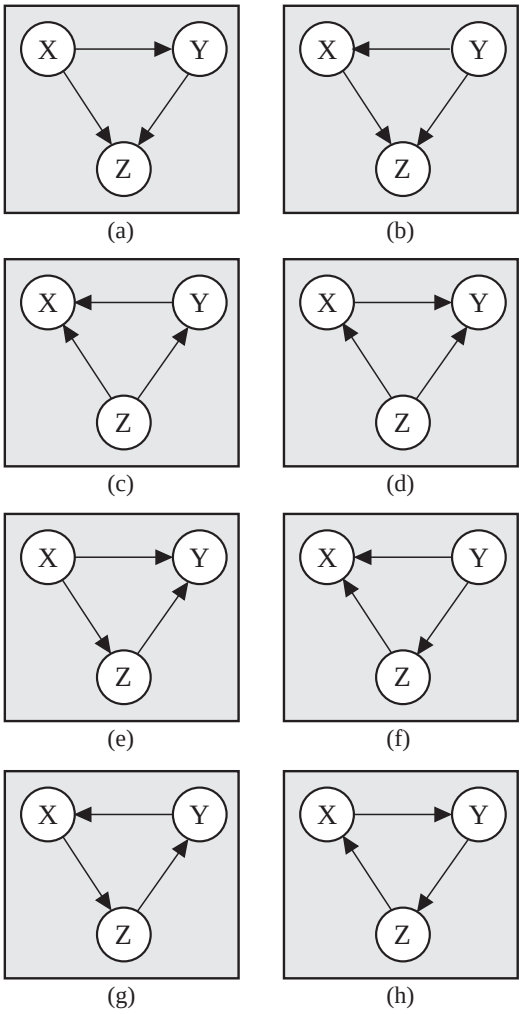


FIGURA 6.20. Conjunto de todos los grafos completos dirigidos de tres nodos. Los grafos (a)−(f) son grafos dirigidos acíclicos. Los grafos (g) y (h) son cíclicos.

Algoritmo 6.5 Ordenación de las aristas de un grafo dirigido.

- **Datos:** Un grafo dirigido acíclico $D = (X, L)$.
- **Resultados:** Una ordenación de las aristas en L.

1. Calcular una ordenación ancestral de los nodos de D (Algoritmo 4.6).
2. Asignar $i \leftarrow 1$.
3. Mientras exista alguna arista no ordenada en L:

(a) Obtener el primer nodo en la ordenación Y que tenga una arista no ordenada convergente.

(b) Obtener el último nodo en la ordenación X para el cual la arista $L_{X\,Y}$ no esté ordenada.

(c) Etiquetar la arista $L_{X\,Y}$ con el número i que le corresponda en la ordenación.

(d) $i \leftarrow i + 1$. ∎

Algoritmo 6.6 Búsqueda de las aristas irreversibles de un grafo dirigido.

- **Datos:** Un grafo dirigido acíclico $D = (X, L)$.
- **Resultados:** Clasificación de las ariastas de L como *irreversibles* o *reversibles*.

1. Ordenar las aristas de L utilizando el Algoritmo 6.5.
2. Etiquetar cada arista de L como *desconocida*.
3. Mientras exista alguna arista en L etiquetada como *desconocida*:
   (a) Calcular la primera arista en la ordenación $L_{X\,Y}$ que esté etiquetada como *desconocida*.
   (b) Para cada arista $L_{W\,X}$ etiquetada *irreversible*, si W no es un padre de Y, entonces etiquetar la arista $L_{X\,Y}$ y toda arista convergente a Y como *irreversible* e ir a la Etapa 3; en caso contrario, etiquetar la arista $L_{W\,Y}$ como *irreversible*.
   (c) Si existe una arista $L_{Z\,Y}$ tal que $Z \neq X$ y Z no es un padre de X, entonces etiquetar la arista $L_{X\,Y}$ y todas las aristas *desconocidas* convergentes a Y como *irreversibles*; en caso contrario, etiquetar la arista $L_{X\,Y}$ y todas las aristas *desconocidas* convergentes a Y como *reversibles*. ∎

Ejemplo 6.21 Aristas reversibles e irreversibles. En la Figura 6.21 se muestra un ejemplo de aplicación de los algoritmos anteriores.

Primeramente se ilustra el uso del Algoritmo 6.5.

- **Etapa 1**: Se realiza una ordenación ancestral de los nodos, tal y como se muestra en la Figura 6.21.
- **Etapa 2**: Se toma $i = 1$.
- **Etapa 3**:
  - (a): El primer nodo de la ordenación que tiene una arista no ordenada convergente es $Y = C$.

- (b): El último nodo de la ordenación para el que la arista L_{XC} no está ordenada es $X = A$.
- (c): Se etiqueta la arista L_{AC} con el número 1.
- (d): Se considera $i = 2$.
- (a): El primer nodo de la ordenación que tiene una arista no ordenada convergente es $Y = D$.
- (b): El último nodo de la ordenación para el que la arista L_{XD} no está ordenada es $X = B$.
- (c): Se etiqueta la arista L_{BD} con el número 2.
- (d): Se considera $i = 3$ y se procede de forma análoga hasta numerar todas las aristas.

La Tabla 6.10 muestra detalladamente todas los pasos de la Etapa 3 en el algoritmo anterior.

A continuación, se ilustra la aplicación del Algoritmo 6.6.

- **Etapa 1**: La primera etapa consiste en la ordenación de las aristas realizada anteriormente.
- **Etapa 2**: Se etiqueta cada arista como *desconocida*.
- **Etapa 3**:
  - (a): La primera arista de la ordenación etiquetada como *desconocida* es L_{AC}.
  - (b): No existe ninguna arista L_{WA} etiquetada *irreversible*.
  - (c): No existe ninguna arista L_{ZC} tal que $Z \neq A$, por tanto, se etiqueta L_{AC} como *reversible*.
  - (a): La primera arista de la ordenación etiquetada como *desconocida* es L_{BD}.
  - (b): No existe ninguna arista L_{WB} etiquetada *irreversible*.
  - (c): La arista L_{AD} verifica $A \neq B$ y A no es un padre de B, por tanto, se etiquetan las aristas L_{BD} y L_{AD} como *irreversibles*.
  - (a): La primera arista de la ordenación etiquetada como *desconocida* es L_{DF}.
  - (b): Las aristas L_{AD} y L_{BD} de la forma L_{WD}, están etiquetadas como *irreversibles*. Dado que A y B no son padres de F, se etiqueta la arista L_{DF} como *irreversible*. No existen más aristas convergentes a F.
  - (a): L_{EG} es la primera arista de la ordenación etiquetada como *desconocida*.
  - (b): No existe ninguna arista L_{WE} etiquetada como *irreversible*.

(a) Y	(b) X	(c) L_{XY}	(d) i
C	A	$Orden(L_{AC}) = 1$	2
D	B	$Orden(L_{BD}) = 2$	3
D	A	$Orden(L_{AD}) = 3$	4
F	D	$Orden(L_{DF}) = 4$	5
G	E	$Orden(L_{EG}) = 5$	6
G	D	$Orden(L_{DG}) = 6$	7

TABLA 6.10. Detalles de la Etapa 3 del Algoritmo 6.5 para el grafo dirigido del Ejemplo 6.21.

(a) L_{XY}	(b)	(c)
$A \rightarrow C$	–	$L_{AC} \leftarrow Reversible$
$B \rightarrow D$	–	$L_{AD} \leftarrow Irreversible$ $L_{BD} \leftarrow Irreversible$
$D \rightarrow F$	$L_{DF} \leftarrow Irreversible$	–
$E \rightarrow G$		$L_{EG} \leftarrow Irreversible$ $L_{DG} \leftarrow Irreversible$

TABLA 6.11. Detalles de la Etapa 3 del Algoritmo 6.5 para el grafo dirigido del Ejemplo 6.21.

– (c): Existe una arista L_{DG} tal que $D \neq E$ y D no es un padre de E, por tanto, las aristas L_{EG} y L_{DG} se etiquetan como *irreversibles*.

La Tabla 6.11 muestra detalladamente los pasos a seguir en la Etapa 3 del Algoritmo 6.6. ∎

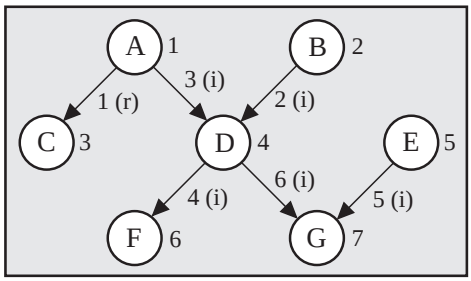


FIGURA 6.21. Grafo dirigido acíclico con las aristas clasificadas como reversibles (r) o irreversibles (i).

El concepto de equivalencia anterior se refiere exclusivamente a la estructura cualitativa de la red Bayesiana, es decir, al modelo de dependencia. La siguiente definición introduce el concepto de equivalencia cuantitativa.

Definición 6.15 Redes Bayesianas probabilísticamente equivalentes. *Dos redes Bayesianas* (D_1, P_1) *y* (D_2, P_2) *se dicen probabilísticamente equivalentes si definen el mismo modelo probabilístico (cualitativo y cuantitativo), es decir, si* D_1 *y* D_2 *son equivalentes y* $P_1 = P_2$.

Obsérvese que dos redes Bayesianas probabilísticamente equivalentes han de tener estructuras gráficas equivalentes, pero el recíproco no tiene por qué ser cierto.

Utilizando el concepto de equivalencia probabilística se puede dividir el conjunto de todas las redes Bayesianas de n variables en clases de equivalencia. Esta posibilidad será particularmente útil en el Capítulo 11 dedicado al aprendizaje de redes Bayesianas.

6.6 Expresividad de los Modelos Gráficos

En las secciones previas se ha analizado en detalle la forma de construir modelos probabilísticos utilizando grafos dirigidos y no dirigidos. Se ha visto que no todo modelo probabilístico puede ser representado por un mapa perfecto dirigido o no dirigido. Para los casos en los que no es posible hallar un mapa perfecto, se introdujo el concepto de I-mapa. Esto motivó la aparición de dos importantes clases de modelos que podían ser representados por I-mapas minimales, las redes de Markov (grafos no dirigidos) y las redes Bayesianas (grafos dirigidos). El siguiente problema se refiere a la capacidad de representación de estos dos tipos de modelos:

- **Problema 6.7:** ¿Todo modelo de dependencia que puede representarse por un tipo de grafos puede ser representado también por el otro tipo?

La respuesta a la pregunta anterior es que, generalmente, no. Por ejemplo, en el Ejemplo 6.1 se mostraba un caso de un modelo de dependencia que puede ser representado por un mapa perfecto dirigido, pero que no posee ningún mapa perfecto no dirigido. El siguiente ejemplo muestra el caso de un modelo que puede ser representado por un mapa perfecto no dirigido, pero que no tiene ningún mapa perfecto dirigido.

Ejemplo 6.22 Modelos obtenidos de grafos no cordales. El grafo no dirigido mostrado en la Figura 6.22(a) define el modelo de dependencia $M = \{I(X, Y|\{W, Z\}), I(W, Z|\{X, Y\})\}$. Por tanto, el grafo no dirigido es un mapa perfecto del modelo M. Sin embargo, no es posible encontrar ningún mapa perfecto dirigido que reproduzca este par de relaciones de

independencia. Obsérvese que todo grafo dirigido que pueda construirse a partir de este grafo contendrá al menos un nodo con aristas convergentes. En consecuencia, contendrá una v-estructura que inducirá algunas dependencias adicionales en el modelo asociado al grafo dirigido. Por ejemplo, dado el grafo dirigido de la Figura 6.22(b), donde las líneas discontinuas indican las aristas necesarias para moralizar el grafo (ver Definición 5.4), se puede obtener la primera relación de independencia, pero no la segunda. En este caso, el grafo contiene la v-estructura (W, Y, Z), que hace que W y Z sean condicionalmente dependientes dada Y. De forma similar, a partir del grafo de la Figura 6.22(c), se puede obtener la segunda relación de independencia, pero no la primera. ■

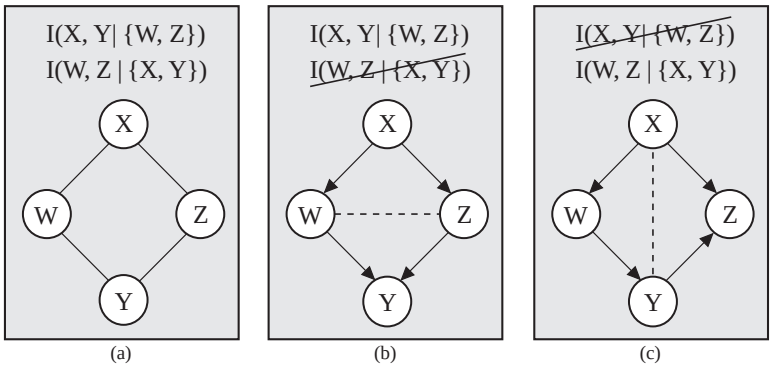


FIGURA 6.22. Modelo de dependencia no dirigido y dos grafos dirigidos que representan parcialmente el modelo.

En general, ninguno de los dos tipos de grafos tiene un poder de representación mayor que el otro (ver Ur y Paz (1994)). Sin embargo, como se analizará brevemente a continuación, la inclusión de nodos auxiliares hace que cualquier modelo representable por medio de grafos no dirigidos pueda ser también representado mediante grafos dirigidos.

Las propiedades que caracterizan los mapas perfectos no dirigidos y dirigidos han sido presentadas en los Teoremas 6.1 y 6.8, respectivamente. Estas propiedades están relacionadas de la forma siguiente. La *unión fuerte* implica la *unión débil*; la *transitividad fuerte* implica la *transitividad débil*; y la *unión fuerte* y la *intersección* implican la *composición* y la *contracción* (ver Sección 5.3). Por tanto, cada modelo de dependencia gráfico no dirigido satisface las seis primeras propiedades del Teorema 6.8. Sin embargo, un modelo gráfico no dirigido no satisface, en general, la propiedad de cordalidad (no tiene por qué ser triangulado). En este caso, el modelo no podrá ser representado por un grafo dirigido. Entonces, para que un modelo de dependencia pueda ser representado tanto por un grafo no dirigido, como por un grafo dirigido, es necesario que sea un modelo cordal, o descomponible. Además, si el grafo es cordal (triangulado), todas las dependencias que

puedan ser representadas por el grafo no dirigido, también tienen que poder ser representadas por un grafo dirigido, como se muestra en los teoremas siguientes (ver Pearl y Verma (1987) y Pearl (1988)).

Teorema 6.16 Intersección de modelos gráficos dirigidos y no dirigidos. *La intersección de los modelos de dependecia basados en grafos no dirigidos y los basados en grafos dirigidos está contenida en la clase de modelos de depencia representables por grafos cordales, o triangulados (modelos descomponibles).*

Ejemplo 6.23 Modelos descomponibles. La Figura 6.23(a) muestra un grafo no dirigido cordal. El modelo de dependencia correspondiente contiene una sentencia de independencia condicional, $I(X, Y|\{W, Z\})$. Por tanto, la red de Markov asociada puede ser factorizada como

$$\begin{aligned} p(x, y, z, w) &= \psi_1(x, w, z)\psi_2(y, w, z) \\ &= p(x, w, z)p(y|w, z). \end{aligned} \tag{6.44}$$

Por otra parte, el grafo dirigido acíclico mostrado en la Figura 6.23(b) define la siguiente factorización de una red Bayesiana:

$$\begin{aligned} p(x, y, z, w) &= p(x)p(w|x)p(z|x, w)p(y|w, z) \\ &= p(x, w, z)p(y|w, z). \end{aligned} \tag{6.45}$$

Por tanto, los grafos (no dirigido y dirigido) de las Figuras 6.23(a) y (b), respectivamente, definen el mismo modelo probabilístico, como puede deducirse de (6.44) y (6.45). ∎

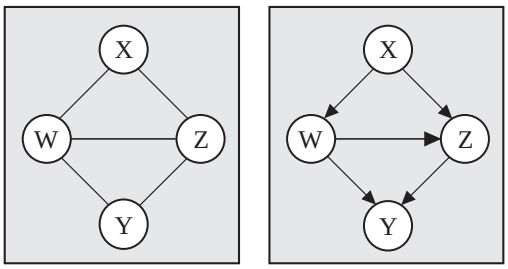


FIGURA 6.23. Grafo no dirigido cordal (a) y un grafo dirigido acíclico asociado (b).

Cuando un modelo de dependencia está basado en un grafo no dirigido que no es cordal, puede obtenerse un grafo dirigido acíclico equivalente con la ayuda de algunos nodos auxiliares (ver Pearl (1988)). Por tanto, cualquier modelo de dependencia representable por un grafo no dirigido también puede ser representado por un grafo dirigido.

Teorema 6.17 Nodos auxiliares. *Cada modelo de dependencia asociado a un grafo no dirigido se puede expresar de forma equivalente por un grafo dirigido acíclico mediante la inclusión de algunos nodos auxiliares.*

Ejemplo 6.24 Nodos auxiliares. Considérese el grafo no dirigido no cordal dado en la Figura 6.22(a). El Ejemplo 6.22 muestra que no existe ningún grafo dirigido acíclido que pueda generar el mismo modelo de dependencia. Sin embargo, si se añade el nodo auxiliar A (ver Figura 6.24) se puede obtener un grafo dirigido acíclico que contiene las mismas relaciones de independencia que el grafo no dirigido original. Por tanto, mediante la asignación de valores a nodos auxiliares se puede obtener el mismo modelo probabilístico representado por el grafo no dirigido. ∎

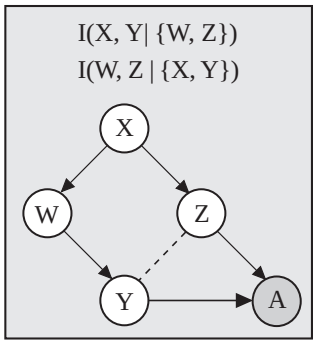


FIGURA 6.24. Ilustración de cómo los grafos dirigidos puden completarse con nodos auxiliares para representar los mismos modelos de dependencia que pueden ser representados por grafos no dirigidos.

De la discusión anterior puede concluirse que las redes Bayesianas proporcionan un metodología intuitiva, sencilla y general para definir modelos probabilísticos para sistemas expertos.

Ejercicios

6.1 Comprobar que el modelo de dependencia M sobre $\{X, Y, Z\}$ dado en el Ejemplo 6.2 no posee ningún mapa perfecto dirigido. Considérense las siguientes etapas:

- Construir las 18 relaciones de independencia condicional distintas con tres variables.
- Construir todos los grafos dirigidos acíclicos con tres variables siguiendo los pasos del Ejemplo 6.1 para el caso de grafos no dirigidos.

- Mostrar que cualquiera de los grafos obtenidos contiene alguna relación de independencia que no se verifica en M o viceversa.

Utilizando todos los posibles grafos dirigidos acíclicos de tres variables, obtener todos los modelos de dependencia con tres variables que poseen un mapa perfecto dirigido.

6.2 Considérese el conjunto de cuatro variables $\{X_1, X_2, X_3, X_4\}$ y el siguiente grafoide:

$$\{I(X_3, X_1|\phi), I(X_2, X_3|X_1), I(X_2, X_3|\phi), I(X_2, X_3|X_1X_4),$$
$$I(X_2, X_4|X_1), I(X_2, X_4|X_1X_3), I(X_3, X_2|X_1), I(X_3, X_2|X_1X_4),$$
$$I(X_4, X_2|X_1), I(X_4, X_2|X_1X_3), I(X_3, X_1X_2|\phi), I(X_3, X_1|X_2),$$
$$I(X_3, X_2|\phi), I(X_1X_2, X_3|\phi), I(X_1, X_3|X_2), I(X_3X_4, X_2|X_1)\}.$$

Obtener el I-mapa minimal no dirigido utilizando el Teorema 6.3.

6.3 Probar que la función de probabilidad dada en el Ejemplo 6.4 no cumple la propiedad de unión fuerte.

6.4 Escribir un programa de ordenador similar al de la Figura 6.4 que calcule un I-mapa minimal no dirigido para la función de probabilidad

$$p(x) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)p(x_4|x_1)p(x_5|x_3, x_4). \qquad (6.46)$$

6.5 Considérese la función de probabilidad en (6.46) y el grafo dado en la Figura 6.25:

(a) Obtener los conglomerados del grafo.

(b) Factorizar la función de probabilidad conjunta de las variables.

(c) ¿Es el grafo un I-mapa minimal no dirigido de (6.46)?

(d) ¿Es descomponible el modelo probabilístico definido por (6.46)?

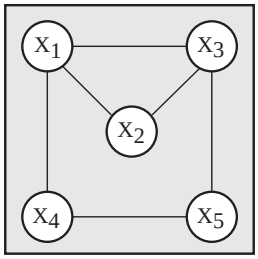


FIGURA 6.25. Grafo no dirigido con cinco nodos.

6.6 Construir la factorización de la función de probabilidad sugerida por cada uno de los grafos no dirigidos mostrados en la Figura 6.1.

6.7 Construir la factorización de la función de probabilidad sugerida por cada uno de los grafos no dirigidos mostrados en la Figura 6.3.

6.8 Comprobar que el grafo dirigido mostrado en la Figura 6.2 es un mapa perfecto de M en el Ejemplo 6.1.

6.9 ¿Es el grafo de la Figura 6.26 un mapa perfecto de la función de probabilidad

$$p(x, y, z, u) = p_1(x)p_2(y)p_3(z|x, y)p_4(w|y, z)?$$

Considérense los casos siguientes:

(a) $p_1(x)$ es Binomial(1, 0.3) y $p_2(y)$ es Binomial(1, 0.7).

(b) $p_3(z|x, y)$ es Binomial(1, $(x + y)/2$) y $p_4(w|z, y)$ es Binomial(1, $(y + z)/2$).

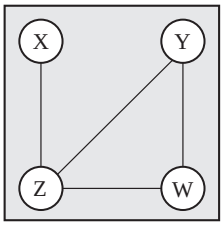


FIGURA 6.26. Grafo dirigido con cuatro nodos.

6.10 Dado el conjunto de variables $(X_1, X_2, X_3, X_4, X_5)$ normalmente distribuidas con vector de medias y matriz de covarianzas

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} 1 & 0.3 & 0 & 0.4 & 0 \\ 0.3 & 1 & 0 & 0.2 & 0 \\ 0 & 0 & 1 & 0 & 0.1 \\ 0.4 & 0.2 & 0 & 1 & 0 \\ 0 & 0 & 0.1 & 0 & 1 \end{pmatrix},$$

utilizar el Teorema 6.10 para encontrar el grafo dirigido acíclico asociado a

(a) la ordenación $(X_1, X_2, X_3, X_4, X_5)$.

(b) la ordenación $(X_5, X_4, X_1, X_3, X_2)$.

6.11 Considérese el conjunto de variables (X_1, X_2, X_3, X_4) normalmente distribuidas con vector de medias y matriz de covarianzas

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} 1 & 1/2 & 1/8 & 1/4 \\ 1/2 & 1 & 1/4 & 1/2 \\ 1/8 & 1/4 & 1 & 1/4 \\ 1/4 & 1/2 & 1/4 & 1 \end{pmatrix}.$$

Utilizar el Algoritmo 6.4 para obtener un I-mapa utilizando dos ordenaciones distintas, (X_1, X_2, X_3, X_4) y (X_4, X_3, X_2, X_1). Comprobar si los grafos de la Figura 6.27 son correctos.

Nota: En la Tabla 6.12 se muestran las medias y varianzas condicionadas de las variables $X_i|\pi_i$ que aparecen en el proceso de construcción del I-mapa.

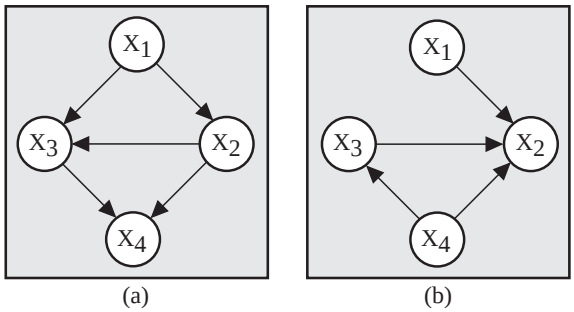


FIGURA 6.27. Grafos asociados con dos ordenaciones distintas de los nodos: (X_1, X_2, X_3, X_4) (a) y (X_4, X_3, X_2, X_1) (b).

6.12 Considerando la situación dada en el Ejemplo 6.12, demostrar:

(a) El modelo M cumple las siete propiedades requeridas por el Teorema 6.8.

(b) Las relaciones de independencia de M se pueden obtener del grafo D mostrado en la Figura 6.28 utilizando el criterio de D-separación.

(c) D cumple $I(\{X, Z\}, Y|W)$ y $I(X, \{W, Y\}|Z)$, pero M no. Por tanto D no es un mapa perfecto de M.

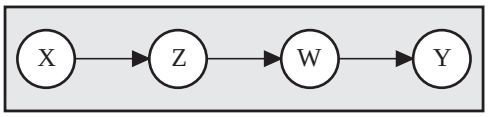


FIGURA 6.28. Grafo dirigido que incluye todas las relaciones de independencia en (6.29), pero que posee algunas independencias no contenidas en M.

6.13 Construir la factorización de la función de probabilidad sugerida por cada uno de los grafos dirigidos mostrados en la Figura 6.13.

6.14 Dado el conjunto de variables $\{X, Y, Z, W\}$ y el modelo de dependencia

$$M = \{I(Y, X|\phi), I(Z, Y|X), I(W, X|\{Y, Z\})\},$$

X_i	π_i	Media de $X_i\|\pi_i$	Varianza de $X_i\|\pi_i$.
X_1	$\{x_2\}$	$x_2/2$	3/4
X_1	$\{x_3\}$	$x_3/8$	63/64
X_1	$\{x_4\}$	$x_4/4$	15/16
X_1	$\{x_2, x_3\}$	$x_2/2$	3/4
X_1	$\{x_2, x_4\}$	$x_2/2$	3/4
X_1	$\{x_3, x_4\}$	$(2x_3 + 7x_4)/30$	14/15
X_1	$\{x_2, x_3, x_4\}$	$x_2/2$	3/4
X_2	$\{x_3\}$	$x_3/4$	15/16
X_2	$\{x_4\}$	$x_4/2$	3/4
X_2	$\{x_3, x_4\}$	$(2x_3 + 7x_4)/15$	11/15
X_3	$\{x_4\}$	$x_4/4$	15/16
X_4	$\{x_3\}$	$x_3/4$	15/16
X_4	$\{x_2\}$	$x_2/2$	3/4
X_4	$\{x_1\}$	$x_1/4$	15/16
X_4	$\{x_3, x_2\}$	$(7x_2 + 2x_3)/15$	11/15
X_4	$\{x_3, x_1\}$	$2(x_1 + x_3)/9$	3/4
X_4	$\{x_2, x_1\}$	$x_2/2$	3/4
X_4	$\{x_3, x_2, x_1\}$	$(7x_2 + 2x_3)/15$	11/15
X_3	$\{x_2\}$	$x_2/4$	15/16
X_3	$\{x_1\}$	$x_1/8$	63/64
X_3	$\{x_2, x_1\}$	$x_2/2$	3/4
X_2	$\{x_1\}$	$x_1/2$	3/4

TABLA 6.12. Medias y varianzas condicionadas de las variables normales $X_i|\pi_i$ en el ejercicio anterior.

elegir una ordenación de las variables y obtener la lista causal generada por M.

6.15 Dado el conjunto de variables $\{X_1, X_2, X_3, X_4, X_5\}$ y la lista causal

$$\begin{array}{ll} I(I(X_2, X_1|\phi), & I(X_3, X_1|X_2), \\ I(X_4, X_1|\{X_2, X_3\}), & I(X_5, X_2|\{X_1, X_3, X_4\}), \end{array}$$

(a) Calcular el conjunto mínimo de relaciones de independencia adicionales para que sea un semigrafoide.

(b) ¿Es un grafoide el conjunto obtenido?

6.16 Generar las factorizaciones asociadas a los grafos dirigido y no dirigido de la Figura 6.23 y comprobar que son las mismas que las utilizadas en el Ejemplo 6.23.

Capítulo 7
Extensiones de los Modelos Gráficos

7.1 Introducción

En el Capítulo 6 se han introducido los modelos gráficos de dependencia, definidos por medio de grafos dirigidos y no dirigidos, y se ha visto cómo estos modelos permiten definir de forma sencilla la estructura cualitativa de un modelo probabilístico. La principal deficiencia de estos modelos es que no todo modelo probalístico se puede definir de forma perfecta mediante un grafo. Por tanto, los modelos gráficos han de entenderse, en general, como mapas de independencia (I-mapas) de los modelos que se desean representar. Esto significa que todas las relaciones de independencia condicional verificadas por el grafo serán independencias reales del modelo, aunque algunas de las independencias del modelo podrán escapar a la representación gráfica. El siguiente ejemplo ilustra esta deficiencia de los modelos gráficos mediante un sencillo ejemplo.

Ejemplo 7.1 Modelo de dependencia sin mapa perfecto dirigido. Considérese el conjunto de variables $\{X, Y, Z\}$ que están relacionadas por las siguientes relaciones de independencia:

$$M = \{I(X, Y|Z), I(Y, X|Z), I(Y, Z|X), I(Z, Y|X)\}. \tag{7.1}$$

El modelo M está formado por dos relaciones de independencia y sus relaciones simétricas. Aunque este modelo es muy simple, no existe ningún grafo dirigido que sea un mapa perfecto de M. Por ejemplo, utilizando el criterio de D-separación (ver Sección 5.2.2), se puede ver que el grafo

dirigido acíclico mostrado en la Figura 7.1(a) implica solamente las dos primeras relaciones de independencia, mientras que el grafo de la Figura 7.1(b) verifica sólo las dos últimas. Por tanto, ninguno de los grafos es una representación perfecta del modelo y, por tanto, sólo se puede pensar en ellos como I-mapas del modelo de dependencia M. Por tanto, es imposible definir un modelo probabilístico que tenga la estructura de dependencia dada en M utilizando un único grafo. ∎

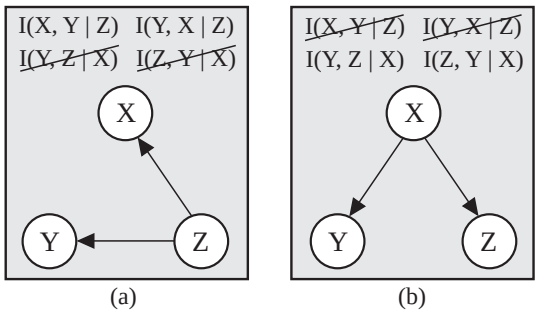


FIGURA 7.1. Ejemplo ilustrando el hecho de que el modelo de dependencia dado en (7.1) no puede ser representado por un único grafo.

En este capítulo se introducen algunos métodos para extender la capacidad de representación de los modelos gráficos y abarcar una clase mayor de modelos de dependencia. Estos modelos incluyen:

1. Modelos definidos por un conjunto de grafos (multigrafos).
2. Modelos definidos por una lista de relaciones de independencia.
3. Modelos definidos por una combinación de grafos y relaciones de independencia.
4. Modelos definidos por un conjunto de funciones de probabilidad condicionada.

A pesar de que estos modelos ofrecen cuatro alternativas distintas para construir modelos de dependencia más generales, existen ciertas similitudes entre ellos. Por ejemplo, utilizando un criterio de separación gráfica adecuado (ver Sección 5.2), se puede obtener la lista de relaciones de independencia que se derivan de un grafo dado. Por tanto, un modelo gráfico se puede convertir en un conjunto equivalente de relaciones de independencia, es decir, los modelos anteriores 1 y 3 se pueden reducir al modelo 2. Por otra parte, se verá que tanto los modelos definidos por multigrafos, como los definidos por listas de relaciones de independencia, definen una serie de factorizaciones de la función de probabilidad por medio de un conjunto de funciones de probabilidad condicionada. Por tanto, los modelos

1−3 se pueden reducir al modelo 4, que proporciona las nociones básicas para entender el resto de los modelos que generalizan las representaciones gráficas.

En este capítulo se analizan estos modelos y sus relaciones. En las Secciones 7.2 y 7.3 se trata el problema de los modelos definidos por multigrafos y por listas de relaciones de independencia, respectivamente. La Sección 7.4 introduce los modelos multifactorizados. En las Secciones 7.5 y 7.6 se muestran dos ejemplos concretos de estos modelos (uno discreto y otro continuo). Los modelos definidos por un conjunto de funciones de probabilidad condicionada se introducen en la Sección 7.7. Finalmente, las Secciones 7.7.1 y 7.7.2 discuten los problemas de existencia y unicidad que aparecen en estos modelos.

7.2 Modelos Definidos por Multigrafos

7.2.1 Definición y Ejemplo

Dado que un único grafo no permite representar cualquier modelo de dependencia, los modelos gráficos se pueden generalizar considerando un conjunto de grafos, en lugar de un único grafo. Los modelos restultantes se denominan *modelos definidos por multigrafos*. Por ejemplo, Geiger (1987) analizó el problema de representar un modelo de dependencia mediante un conjunto de grafos dirigidos acíclicos. De forma similar, Paz (1987) y Shachter (1990b) analizaron las propiedades de las representaciones basadas en un conjunto de grafos no dirigidos. Aunque estos modelos permiten definir una clase más amplia de modelos de dependencia que los modelos basados en un único grafo, Verma (1987) demostró que puede ser necesario un número exponencial de grafos para representar un modelo de dependencia arbitrario de forma perfecta. Por tanto, desde un punto de vista práctico, los modelos basados en multigrafos sólo pueden ser utilizados para extender la capacidad de representación de los modelos gráficos simples. Por tanto, incluso utilizando un conjunto de grafos, alguna independencia del modelo puede escapar a la representación gráfica. En consecuencia, los multigrafos serán, en general, I-mapas mejorados del modelo de dependencia dado. Obsérvese que el término *multigrafo* denota el conjunto (la unión) de las relaciones de independencia implicadas por un conjunto de grafos. Por tanto, los modelos definidos por multigrafos son, en efecto, equivalentes a los modelos definidos por una lista de relaciones de independencia. Estos modelos se analizarán en detalle en la Sección 7.3.

Esta idea sencilla de combinar varios grafos ofrece una extensión importante de los modelos gráficos basados en un único grafo.

Definición 7.1 Modelos definidos por multigrafos. *Considérese el conjunto de variables* $X = \{X_1, \ldots, X_n\}$. *Un modelo definido por un multi-*

grafo en X es un conjunto de redes Bayesianas y de Markov compatibles

$$\{(G^\ell, P^\ell),\ \ell = 1, \ldots, m\}, \tag{7.2}$$

definidias sobre cada uno de los grafos G^ℓ del multigrafo que definen una serie de factorizaciones P^ℓ del correspondiente modelo probabilístico. La compatibilidad requiere que la función de probabilidad conjunta definida por todas las redes en (7.2) sea idéntica, es decir,

$$p(x) = \prod_{i=1}^{n} p^\ell(x_i^\ell | s_i^\ell), \quad \ell = 1, \ldots, m. \tag{7.3}$$

El conjunto de redes Bayesianas y de Markov en (7.2) define la estructura de dependencia del modelo (dada por el multigrafo $\{G^1, \ldots, G^m\}$) y el modelo probabilístico resultante (dado por los conjuntos de factorizaciones). El modelo probabilístico resultante tiene asociada una estructura de dependencia más general que los modelos simples definidos por cada uno de los grafos.

Ejemplo 7.2 Modelo de multired Bayesiana. Sean D^1 y D^2 los grafos dirigidos acíclicos dados en las Figuras 7.1(a) y (b), respectivamente. Cada uno de estos grafos es un I-mapa dirigido del modelo de dependencia M dado en (7.1). El multigrafo $\{D^1, D^2\}$ implica el conjunto de independencias siguiente:

$$M = \{I(X, Y|Z), I(Y, X|Z), I(Y, Z|X), I(Z, Y|X)\}, \tag{7.4}$$

que es el mismo modelo M dado en (7.1). Obsérvese que D_1 implica las dos primeras independencias y D^2 implica las dos segundas. Estos dos grafos, y las correspondientes factorizaciones, definen una multired Bayesiana. En este caso se tiene $m = 2$, y (7.2) resulta

$$\{(D^1, P^1), (D^2, P^2)\},$$

donde

$$P^1 = \{p^1(x|z), p^1(y|z), p^1(z)\},$$
$$P^2 = \{p^2(x), p^2(y|x), p^2(z|x)\}.$$

Para que el modelo sea compatible, tal y como se muestra en (7.3), las dos funciones de probabilidad P^1 y P^2 deben ser idénticas, es decir,

$$p(x, y, z) = p^1(x|z)p^1(y|z)p^1(z) = p^2(x)p^2(y|x)p^2(z|x), \tag{7.5}$$

El problema de la consistencia, es decir, hallar las condiciones para que se cumpla (7.3), se analiza en una sección posterior utilizando el concepto de modelo multifactorizado. ∎

A continuación se analizan los siguientes problemas relacionados con los modelos definidos por multigrafos:

- **Problema 7.1:** ¿Cómo se interpretan gráficamente las independencias del modelo?
- **Problema 7.2:** ¿Se puede reducir el número de grafos que componen el multigrafo sin alterar el modelo de dependencia que define?
- **Problema 7.3:** ¿Cómo se puede obtener el modelo probabilístico asociado al modelo de dependencia?

Estos problemas son tratados en las secciones siguientes.

7.2.2 Interpretación de Independencias en un Multigrafo

El primer problema relacionado con los modelos definidos por multigrafos es la interpretación gráfica de sus independencias. Las redes Bayesianas y de Markov son I-mapas de un cierto modelo de dependencia asociado al modelo probabilístico correspondiente. Entonces, todas las independencias condicionales contenidas en el grafo también son independencias del modelo correspondiente. Por tanto, será cierta en un multigrafo una relación de independencia cualquiera si es cierta en alguno de los grafos que componen el multigrafo; en caso contrario será falsa. Por tanto, el criterio gráfico de separación para multigrafos consiste en la aplicación del criterio de U-separación en los grafos no dirigidos que compongan el multigrafo y el criterio de D-separación en los dirigidos.

7.2.3 Reducción del Conjunto de Grafos

El segundo problema de estos modelos es el de la redundancia en un multigrafo. En algunos casos, todas las independencias implicadas por un grafo del modelo pueden ser obtenidas a partir de los demás grafos. Por ejemplo, Shachter (1990b) introdujo algunas transformaciones gráficas que permiten simplificar la estructura de los grafos eliminando independencias redundantes. En algunos casos, el conjunto de grafos puede ser reducido a un conjunto menor que es una representación más simple y eficiente del modelo.

Definición 7.2 Grafos redundantes. *Dados dos grafos G_1 y G_2, se dice que G_1 es redundante dado G_2 si el conjunto de relaciones de independencia contenidas en G_1 está contenido en G_2.*

Como puede verse en el teorema siguiente, el problema de la redundancia en grafos no dirigidos es fácil de resolver.

Teorema 7.1 Redundancia en multigrafos no dirigidos. *Dados dos grafos no dirigidos* $G^1 = (X, L^1)$ *y* $G^2 = (X, L^2)$ *con el mismo conjunto de variables* X*, entonces* G_1 *es redundante dado* G_2 *si* $L^1 \subset L^2$*.*

Ejemplo 7.3 Redundancia en multigrafos no dirigidos. Sean G^1 y G^2 los grafos no dirigidos mostrados en las Figuras 7.2(a) y (b), respectivamente. Se puede comprobar fácilmente que el grafo G^1 es redundante dado G^2 ya que $L^1 = \{L_{12}, L_{13}, L_{34}, L_{35}\}$ es un subconjunto de $L^2 = \{L_{12}, L_{13}, L_{34}, L_{35}, L_{24}\}$. Por tanto, el multigrafo formado por los dos grafos define el mismo modelo de dependencia que el modelo gráfico formado por G_1. ∎

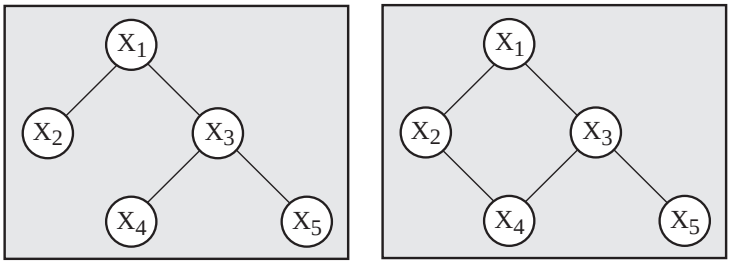


FIGURA 7.2. Dos grafos no dirigidos donde (b) es redundante dado (a).

El problema de la redundancia en grafos dirigidos no es trivial. El ejemplo siguiente ilustra este hecho.

Ejemplo 7.4 Reduciendo un conjunto de grafos dirigidos. Considérese el multigrafo formado por los tres grafos dirigidos D^1, D^2 y D^3 dados en las Figuras 7.3(a)−(c), respectivamente. En este caso, todas las independencias que implica el grafo D^2 pueden ser obtenidas de D^1. Este hecho puede comprobarse de la forma siguiente. Si se incluye la arista L_{24} en D^1, entonces cualquier independencia derivada del nuevo grafo también podrá ser derivada del grafo original (la inclusión de aristas no incluye nuevas independencias). Por otra parte, se puede invertir la dirección de las aristas L_{13} y L_{35} simultáneamente sin modificar el modelo de dependencia asociado al grafo. Por tanto, si en D^1 se añade la arista L_{24} y se invierte la dirección de L_{13} y L_{35}, se obtiene el grafo D^2. Así, todas las independencias del grafo D^2 están contenidas en D^1 y, por tanto, D_2 es redundante dado D_1, es decir, el modelo definido por el multigrafo $\{D^1, D^2\}$ es equivalente al modelo definido únicamente por D^1.

Por otra parte, D^1 y D^3 no son redundantes entre sí, pues D^1 contiene la independencia $I(X_2, X_4|X_1)$, que no es verificada por D^3, y D^3 implica $I(X_1, X_2|X_3)$, que no puede obtenerse de D^1. ∎

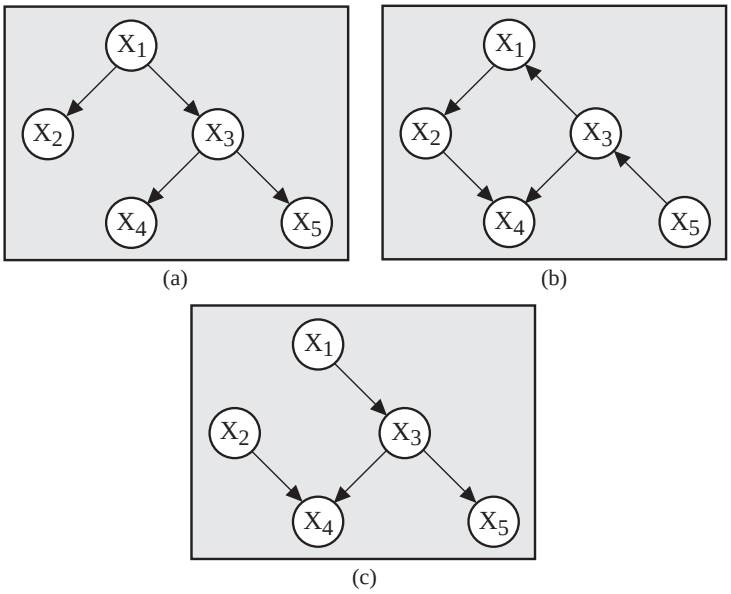


FIGURA 7.3. Tres grafos dirigidos acíclicos que definen un multigrafo.

El teorema siguiente muestra las condiciones para que dos grafos dirigidos sean redundantes.

Teorema 7.2 Redundancia en multigrafos dirigidos. *Sean D^1 y D^2 dos grafos dirigidos acíclicos sobre el mismo conjunto de variables X, y sean G^1 y G^2 los grafos no dirigidos asociados respectivos. Entonces, D^2 es redundante dado D_1 si (a) G^2 está contenido en G^1, (b) cada v-estructura de D^1 está también contenida en D^2, y (c) cada v-estructura (X_i, X_j, X_k) de D^2 está también contenida en D^1 siempre que G^1 contenga el camino $X_i - X_j - X_k$.*

El ejemplo siguiente ilustra este teorema.

Ejemplo 7.5 Redundancia en multigrafos dirigidos. En el Ejemplo 7.4 se vio mediante una serie de transformaciones topológicas de los grafos que D^2 es redundante dado D^1. Esta conclusión puede obtenerse directamente aplicando el Teorema 7.2. En la Figura 7.3 puede verse que cada arista del grafo G^1 (el grafo no dirigido asociado a D^1) también está contenida en G^2 (el grafo no dirigido asociado a D^2). Por tanto, G^2 es redundante dado G_1, por lo que se cumple la primera condición del Teorema 7.2. Dado que D^1 no tiene v-estructuras, la segunda condición también se cumple. Finalmente, D^2 contiene la v-estructura (X_2, X_4, X_3), pero G^1 no contiene el camino $X_2 - X_4 - X_3$. Por tanto, D^2 es redundante dado D^1. ■

7.2.4 Compatibilidad de Multigrafos.

El Problema 7.3 se refiere a la existencia de una función de probabilidad $p(x)$ que cumpla (7.3). Dado que cada grafo proporciona una factorización distinta de $p(x)$, el problema de compatibilidad se reduce a encontrar el modelo probabilístico dado por un conjunto de factorizaciones.

Ejemplo 7.6 Compatibilidad de multigrafos. Considérese de nuevo el problema introducido en el Ejemplo 7.1 con los dos grafos D^1 y D^2 dados en las Figuras 7.1(a) y (b), respectivamente. La red Bayesiana asociada a D^1 implica la factorización:

$$p(x, y, z) = p^1(z)p^1(x|z)p^1(y|z), \tag{7.6}$$

mientras que la correspondinete a D^2 implica

$$p(x, y, z) = p^2(x)p^2(y|x)p^2(z|x), \tag{7.7}$$

donde los superíndices denotan las distintas factorizaciones. Las redes Bayesianas $\{(D^1, P^1), (D^2, P^2)\}$ definen una multired Bayesiana. Obsérvese que al combinar las independencias contenidas en ambos grafos en un mismo modelo probabilístico, éstas pueden implicar en el modelo alguna otra independencia adicional inducida por las propiedades de la independencia condicional (ver Capítulo 5). Por tanto, un multigrafo no será, en general, un mapa perfecto del modelo probabilístico resultante. Por ejemplo, el multigrafo definido por D^1 y D^2 define el modelo probabilístico

$$M = \{I(X, Y|Z), I(Y, X|Z), I(Y, Z|X), I(Z, Y|X)\}. \tag{7.8}$$

Sin embargo, aplicando la propiedad de intersección (que satisfacen los modelos probabilísticos no extremos), se obtiene la independencia adicional $I(Y, \{X, Z\}|\phi)$ que, aplicando la propiedad de descomposición, permite obtener a su vez $I(X, Y|\phi)$ e $I(Y, Z|\phi)$. Por tanto, la familia de funciones de probabilidad compatibles con el multigrafo formado por los dos grafos dados en la Figura 7.1, contiene las independencias siguientes:

$$M_1 = \{I(X, Y|Z), I(Y, Z|X), I(Y, \{X, Z\}|\phi), I(X, Y|\phi), I(Y, Z|\phi)\}, \tag{7.9}$$

así como las correspondientes independencias simétricas. Comparando M en (7.8) y M_1 en (7.9), puede verse que el multigrafo original es solamente un I-mapa del modelo probabilístico.

Las nuevas independencias de M_1 nos permiten reescribir las factorizaciones en (7.6) y (7.7) como

$$p(x, y, z) = p^1(z)p^1(x|z)p^1(y|z) = p^1(z)p^1(x|z)p^1(y) \tag{7.10}$$

y

$$p(x, y, z) = p^2(x)p^2(y|x)p^2(z|x) = p^2(x)p^2(y)p^2(z|x), \tag{7.11}$$

que son dos factorizaciones equivalentes de la misma familia de funciones de probabilidad. Estas factorizaciones están asociadas a los grafos dados en la Figura 7.4, que son dos mapas perfectos equivalentes del modelo de dependencia M_1 en (7.9), pero no son mapas perfectos del multigrafo original en (7.8). Por tanto, el modelo probabilístico compatible con ambas factorizaciones está determinado por (7.10) ó (7.11). Obsérvese que los dos grafos de la Figura 7.4 han sido obtenidos eliminando las aristas $Z \rightarrow Y$ y $X \rightarrow Y$ de los grafos de la Figura 7.1. Por tanto, existe un grafo que contiene todas las independencias del multigrafo y que permite obtener directamente una factorización del modelo probabilístico compatible con ambos modelos.

En este caso el problema de compatibilidad ha sido fácil de resolver. Sin embargo, en general, este problema es complicado y requiere técnicas generales para su tratamiento. En una sección posterior se analizará este problema en el marco de los modelos multifactorizados. ∎

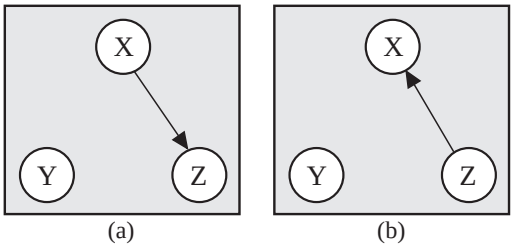


FIGURA 7.4. Dos mapas perfectos del modelo de dependencia en (7.9).

Ejemplo 7.7 Compatibilidad de multigrafos. Considérese el multigrafo dado por los grafos D^1 y D^2 mostrados en las Figuras 7.5(a) y (b), respectivamente. La red Bayesiana definida por D^1 implica:

$$p(x_1, x_2, x_3, x_4) = p^1(x_1)p^1(x_2|x_1)p^1(x_3|x_1)p^1(x_4|x_2, x_3), \tag{7.12}$$

mientras que la definida por D^2 implica

$$p(x_1, x_2, x_3, x_4) = p^2(x_1)p^2(x_2|x_1)p^2(x_4|x_2)p^2(x_3|x_1, x_4). \tag{7.13}$$

Obsérvese que las funciones de probabilidad condicionada en (7.12) y (7.13) están definidas siguiendo las numeraciones ancestrales de las variables implicadas por los grafos correspondientes de la Figura 7.5. A diferencia de lo ocurrido en el ejemplo 7.6, el problema de la compatiblidad de la multired Bayesiana $\{(D^1, P^1), (D^2, P^2)\}$ no es un problema trivial y será resuelto más adelante utilizando las técnicas de los modelos multifactorizados. ∎

Los modelos definidos por multigrafos son un tipo especial de la clase de modelos más general conocida como *modelos multifactorizados* que son analizados en la Sección 7.4.

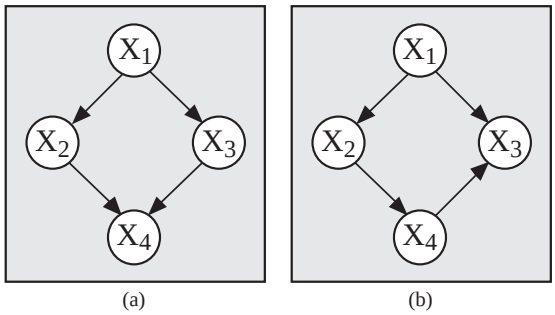


FIGURA 7.5. Dos grafos dirigidos acíclicos que definen un multigrafo.

7.3 Modelos Definidos por Listas de Independencias

Como ya se ha mencionado en el Capítulo 5, las listas de independencias constituyen una alternativa a los modelos gráficos para la construcción de modelos probabilísticos. Esta lista puede venir dada directamente por un experto en el tema a analizar, y representa las relaciones existentes entre las variables del modelo. En esta sección se analiza la relación entre una relación de independencia en un modelo probabilístico y una factorización de la función de probabilidad correspondiente. Esta relación puede resumirse del modo siguiente:

- Siempre se puede encontrar una factorización que contiene una relación de independencia dada.
- Una factorización puede implicar una o más relaciones de independencia.

Ejemplo 7.8 De una relación de independencia a una factorización. Considérese el conjunto de variables $\{X_1, X_2, X_3, X_4\}$ y supóngase que cumplen la relación de independencia $I(X_1, X_2|X_3)$. La función de probabilidad correspondiente puede escribirse como

$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= p(x_2, x_3)p(x_1|x_2, x_3)p(x_4|x_1, x_2, x_3) \\ &= p(x_2, x_3)p(x_1|x_3)p(x_4|x_1, x_2, x_3). \end{aligned} \tag{7.14}$$

Donde la primera igualdad se ha obtenido considerando la partición de las variables $\{\{X_2, X_3\}, X_1, X_4\}$ y aplicando la regla de la cadena a la función de probabilidad $p(x)$, y la segunda igualdad se ha obtenido utilizando la relación de independencia $I(X_1, X_2|X_3)$, que implica $p(x_1|x_2, x_3) = p(x_1|x_3)$. Por tanto, cualquier función de probabilidad que factorice según (7.14) contiene, al menos, la relación de independencia $I(X_1, X_2|X_3)$. Obsérvese que la función de probabilidad podría contener también otras relaciones de independencia derivadas de los axiomas de la probabilidad (por

ejemplo, la relación de independencia simétrica $I(X_2, X_1|X_3)$). Por tanto, la lista de independencias formada por una única relación de independencia es un I-mapa del modelo probabilístico resultante. ∎

Existen listas de independencia que contienen varias relaciones de independencia y que pueden definir una única factorización de forma colectiva. Un ejemplo de ello lo constituyen las listas causales. Dado el conjunto de variables $X = \{X_1, \ldots, X_n\}$, una lista causal definida sobre X es un conjunto de relaciones de independencia de la forma $\{I(Y_1, B_1 \setminus S_1|S_1), \ldots, I(Y_n, B_n \setminus S_n|S_n)\}$, donde $(Y_1, \ldots, Y_n)$ es una permutación de $\{X_1, \ldots, X_n\}$ y $S_i \subset B_i = \{Y_1, \ldots, Y_{i-1}\}$. Esta lista define la siguiente factorización de la función de probabilidad

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i|s_i), \tag{7.15}$$

que incluye todas las relaciones de independencia de la lista causal.

Ejemplo 7.9 De una factorización a una lista de relaciones de independencia. Considérese el conjunto de variables $\{X_1, X_2, X_3, X_4\}$. Aplicando la regla de la cadena, cualquier función de probabilidad de las variables puede expresarse como

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3). \tag{7.16}$$

Esta factorización no implica ninguna relación de independencia pues es una factorización canónica estándar (ver Sección 5.5) y, por tanto, no contiene ninguna independencia entre las variables.

Por otra parte, considérese la factorización

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3). \tag{7.17}$$

Las factorizaciones (7.16) y (7.17) defienen la misma ordenación ancestral de las variables (X_1, X_2, X_3, X_4). Por tanto, se pueden obtener las relaciones de independencia correspondientes a este segundo modelo comparando las funciones de probabilidad condicionada con aquellas contenidas en la factorización general (7.16). Las dos primeras funciones de probabilidad condicionada, $p(x_1)$ y $p(x_2|x_1)$, no implican ninguna relación de independencia pues se hayan contenidas en las dos factorizaciones. Para la tercera función se tiene $p(x_3|x_1, x_2) = p(x_3|x_1)$, que implica la relación de independencia $I(X_2, X_3|X_1)$. Finalmente, $p(x_4|x_1, x_2, x_3) = p(x_4|x_2, x_3)$, que implica $I(X_1, X_4|X_2, X_3)$. Por tanto, la factorización (7.17) implica la lista de relaciones de independencia:

$$M_1 = \{I(X_2, X_3|X_1), I(X_1, X_4|X_2, X_3)\}. \tag{7.18}$$

Obsérvese que a partir de esta lista pueden obtenerse otras relaciones de independencia aplicando las propiedades de la independencia condicional (utilizando, por ejemplo, las propiedades de semigrafoide).

Como ejemplo final, supóngase que un modelo probabilístico está definido por medio de la factorización

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_4|x_2)p(x_3|x_1, x_4), \tag{7.19}$$

que implica la ordenación ancestral de las variables (X_1, X_2, X_4, X_3). Considerando esta ordenación y aplicando la regla de la cadena, se tiene

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_4|x_1, x_2)p(x_3|x_1, x_2, x_4). \tag{7.20}$$

Comparando (7.19) y (7.20) se obtienen las relaciones de independencia siguientes:

$$\begin{aligned} p(x_1) = p(x_1) &\Leftrightarrow \text{sin relaciones de independencia,} \\ p(x_2|x_1) = p(x_2|x_1) &\Leftrightarrow \text{sin relaciones de independencia,} \\ p(x_4|x_1, x_2) = p(x_4|x_2) &\Leftrightarrow I(X_1, X_4|X_2), \\ p(x_3|x_1, x_2, x_4) = p(x_3|x_1, x_4) &\Leftrightarrow I(X_2, X_3|\{X_1, X_4\}). \end{aligned}$$

Por tanto, la factorización (7.19) implica la lista siguiente de relaciones de independencia:

$$M_2 = \{I(X_1, X_4|X_2), I(X_2, X_3|\{X_1, X_4\})\}. \tag{7.21}$$

Obsérvese que esta lista puede completarse utilizando las propiedades de la independencia condicional. ∎

Dado un conjunto de variables $\{X_1, \ldots, X_n\}$, una factorización obtenida aplicando la regla de la cadena canónica a una permutación $(Y_1, \ldots, Y_n)$ de las variables $\{X_1, \ldots, X_n\}$ (ver Definition 5.13)

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i|s_i), \tag{7.22}$$

donde $S_i \subset B_i$=$\{Y_1, \ldots, Y_{i-1}\}$, define la lista causal:

$$\{I(Y_1, B_1 \setminus S_1|S_1), \ldots, I(Y_n, B_n \setminus S_n|S_n)\}. \tag{7.23}$$

Los ejemplos anteriores muestran que toda relación de independencia implica una factorización de la función de probabilidad. Por tanto, dada una lista de relaciones de independencia, se puede obtener un conjunto equivalente de factorizaciones. En ocasiones este conjunto puede ser reducido a una única factorización equivalente. Los siguientes ejemplos ilustran este hecho.

Ejemplo 7.10 Conjunto reducible de factorizaciones. La lista de relaciones de independencia

$$M_1 = \{I(X_2, X_3|X_1), I(X_1, X_4|\{X_2, X_3\})\} \tag{7.24}$$

es equivalente al conjunto de factorizaciones

$$p(x_1, x_2, x_3, x_4) = p^1(x_1, x_2)p^1(x_3|x_1)p^1(x_4|x_1, x_2, x_3)$$

y

$$p(x_1, x_2, x_3, x_4) = p^2(x_1, x_2, x_3)p^2(x_4|x_2, x_3),$$

una factorización para cada una de las relaciones de independencia de M_1, donde los superíndices representan el número de la relación de independencia asociada a cada factorización. Sin embargo, este conjunto es equivalente a una única factorización

$$p(x_1, x_2, x_3, x_4) = p(x_1, x_2)p(x_3|x_1)p(x_4|x_2, x_3), \tag{7.25}$$

ya que

$$\begin{aligned} p(x_1, x_2, x_3, x_4) &= p(x_1, x_2)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3) \\ &= p(x_1, x_2)p(x_3|x_1)p(x_4|x_2, x_3). \end{aligned}$$

La primera de las igualdades se ha obtenido aplicando la regla de la cadena, y la segunda se ha obtenido aplicando las dos relaciones de independencia en M_1. ∎

Ejemplo 7.11 Conjunto irreducible de factorizaciones. Considérense las listas de relaciones de independencia M_1 y M_2, donde M_1 está definida en (7.24) y

$$M_2 = \{I(X_1, X_4|X_2), I(X_2, X_3|\{X_1, X_4\})\}. \tag{7.26}$$

En el Ejemplo 7.10 se ha visto que M_1 da lugar a la factorización (7.25). De forma similar, M_2 implica la factorización

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_4|x_2)p(x_3|x_1, x_4). \tag{7.27}$$

Obsérvese que las factorizaciones (7.25) y (7.27) coinciden con las factorizaciones (7.12) y (7.13) obtenidas a partir de los grafos D^1 y D^2 mostrados en las Figuras 7.5(a) y (b), respectivamente. Esta coincidencia ilustra el hecho de que un modelo de dependencia puede ser descrito de forma indistinta por un grafo o por una lista de relaciones de independencia.

Supóngase que se desea construir un modelo que contenga las independencias de M_1 y M_2, o equivalentemente, que contenga las factorizaciones (7.25) y (7.27). Estas factorizaciones no pueden ser reducidas a una única factorización, a menos que se consideren ciertas restricciones para los parámetros que las componen. Por tanto, se tiene de nuevo un problema de compatibilidad que requiere hallar una función de probabilidad $p(x)$ definida por un conjunto de factorizaciones. ∎

Cuando una lista de relaciones de independencia es equivalente a una única factorización, los parámetros asociados a las funciones de probabilidad condicionada que definen la factorización pueden ser definidos de forma independiente, es decir, sin restricciones. Este es el caso, por ejemplo, de una lista causal, que siempre implica una única factorización de la función de probabilidad. Sin embargo, cuando las factorizaciones no se pueden reducir a una única factorización sin imponer restricciones sobre los parámetros, entonces es necesario resolver el mismo problema de compatibilidad que surge en la definición del modelo probabilístico asociado a un multigrafo. Es decir, es necesario hallar las restricciones que tienen que satisfacer los parámetros de una factorización para que la función de probabilidad resultante pueda ser factorizada en la forma indicada por las factorizaciones que componen el modelo. La sección siguiente analiza el problema de la compatibilidad.

7.4 Modelos probabilísticos Multifactorizados

En las dos últimas secciones hemos visto que la definición de una función de probabilidad mediante multigrafos y listas de relaciones de independencia se reduce a hallar la función de probabilidad compatible con un conjunto dado de factorizaciones. Por tanto, estos dos modelos son casos especiales de un tipo de modelos más generales conocido como modelos probabilísticos multifactorizados.

Definición 7.3 Modelos probabilísticos multifactorizados. *Un modelo probabilístico multifactorizado sobre un conjunto de variables* $X = \{X_1, \ldots, X_n\}$*, es un conjunto de factorizaciones compatibles obtenidas aplicando la regla de la cadena*

$$P = \{P^\ell, \ell = 1, \ldots, m\}, \tag{7.28}$$

donde $P^\ell = \{p^\ell(y_1^\ell|s_1^\ell), \ldots, p^\ell(y_n^\ell|s_n^\ell)\}$ *con* $S_i^\ell \subset B_i^\ell = \{Y_1^\ell, \ldots, Y_{i-1}^\ell\}$*, y* $(Y_1^\ell, \ldots, Y_n^\ell)$ *es una permutación de* $(X_1, \ldots, X_n)$*. Este conjunto define una función de probabilidad* $p(x)$ *compatible con todas las factorizaciones, es decir,*

$$p(x) = \prod_{i=1}^{n} p^\ell(y_i^\ell|s_i^\ell), \quad \ell = 1, \ldots, m. \tag{7.29}$$

Por ejemplo, las factorizaciones (7.12) y (7.13) definen un modelo probabilístico multifactorizado. Esta definición plantea el siguiente problema:

- **Problema 7.4:** ¿Cuáles son las condiciones que tienen que cumplir los conjuntos de funciones de probabilidad condicionada P^ℓ para definir la misma función de probabilidad?

Este problema se conoce por *problema de consistencia* o *problema de compatibilidad*. La Sección 7.5 analiza este problema para el caso de variables multinomiales (discretas), mientras que la Sección 7.6 analiza el caso de variables normales (continuas).

7.5 Modelos Multinomiales Multifactorizados

Antes de analizar el Problema 7.4 para el caso de variables discretas, es necesario analizar la estructura algebraica que define una factorización de una función de probabilidad.

7.5.1 Estructura Paramétrica de una Función de Probabilidad

Considérese el conjunto de variables discretas $\{X_1, \ldots, X_n\}$, donde la variable X_i puede tomar los valores $\{0, \ldots, r_i\}$. Dado que las funciones de probabilidad condicionada $p^\ell(y_i^\ell | s_i^\ell)$, que definen las factorizaciones de la función de probabilidad, pueden ser consideradas como familias paramétricas, una representación apropiada de los parámetros del modelo probabilístico asociado a la factorización ℓ-ésima viene dada por

$$\theta_{ijs}^\ell = p(Y_i^\ell = j | S_i^\ell = s), \quad j \in \{0, \ldots, r_i^\ell\}, \tag{7.30}$$

donde s es una realización de S_i^ℓ. Por tanto, el primer subíndice de θ_{ijs}^ℓ se refiere al número del nodo, el segundo subíndice se refiere al estado del nodo y los subíndices restantes se refieren a la realización de S_i^ℓ. Dado que los parámetros están asociados a probabilidades, han de satisfacer las igualdades

$$\sum_{j=0}^{r_i^\ell} \theta_{ijs}^\ell = 1, \quad \ell = 1, \ldots, m,$$

para cada i y s. Por tanto, uno de los parámetros puede escribirse como uno menos la suma de los restantes. Por ejemplo, $\theta_{ir_is}^\ell$ es

$$\theta_{ir_is}^\ell = 1 - \sum_{j=0}^{r_i^\ell - 1} \theta_{ijs}^\ell, \quad \ell = 1, \ldots, m. \tag{7.31}$$

El conjunto de parámetros θ_{ijs}^ℓ se denota por Θ^ℓ.

Ejemplo 7.12 Estructura paramétrica de una función de probabilidad. Considérese el modelo probabilístico multifactorizado definido por las factorizaciones (7.12) y (7.13) asociadas a los grafos dirigidos acíclicos mostrados en la Figura 7.5 (ver Ejemplo 7.7). Obsérvese que estas factorizaciones coinciden con las dadas en (7.25) y (7.27), obtenidas a partir de las

Variable	Θ^1	Θ^2
X_1	$\theta^1_{10} = p^1(\bar{x}_1)$	$\theta^2_{10} = p^2(\bar{x}_1)$
X_2	$\theta^1_{200} = p^1(\bar{x}_2\|\bar{x}_1)$ $\theta^1_{201} = p^1(\bar{x}_2\|x_1)$	$\theta^2_{200} = p^2(\bar{x}_2\|\bar{x}_1)$ $\theta^2_{201} = p^2(\bar{x}_2\|x_1)$
X_3	$\theta^1_{300} = p^1(\bar{x}_3\|\bar{x}_1)$ $\theta^1_{301} = p^1(\bar{x}_3\|x_1)$	$\theta^2_{3000} = p^2(\bar{x}_3\|\bar{x}_1, \bar{x}_4)$ $\theta^2_{3001} = p^2(\bar{x}_3\|\bar{x}_1, x_4)$ $\theta^2_{3010} = p^2(\bar{x}_3\|x_1, \bar{x}_4)$ $\theta^2_{3011} = p^2(\bar{x}_3\|x_1, x_4)$
X_4	$\theta^1_{4000} = p^1(\bar{x}_4\|\bar{x}_2, \bar{x}_3)$ $\theta^1_{4001} = p^1(\bar{x}_4\|\bar{x}_2, x_3)$ $\theta^1_{4010} = p^1(\bar{x}_4\|x_2, \bar{x}_3)$ $\theta^1_{4011} = p^1(\bar{x}_4\|x_2, x_3)$	$\theta^2_{400} = p^2(\bar{x}_4\|\bar{x}_2)$ $\theta^2_{401} = p^2(\bar{x}_4\|x_2)$

TABLA 7.1. Conjuntos de parámetros, Θ^1 y Θ^2, asociados a las dos factorizaciones (7.12) y (7.13), respectivamente.

listas de relaciones de independencia M_1 y M_2 en los Ejemplos 7.10 y 7.11. Se pueden utilizar dos conjuntos distintos de parámetros para representar el modelo probabilístico asociado a estas factorizaciones. Por ejemplo, si todas las variables son binarias, entonces cada una de estas factorizaciones tiene nueve parámetros libres, como muestra la Tabla 7.1, donde $\bar{x}_i$ y x_i denotan $X_i = 0$ y $X_i = 1$, respectivamente. Estos dos conjuntos de parámetros libres son

$$\Theta^1 = \{\theta^1_{10}, \theta^1_{200}, \theta^1_{201}, \theta^1_{300}, \theta^1_{301}, \theta^1_{4000}, \theta^1_{4001}, \theta^1_{4010}, \theta^1_{4011}\},$$
$$\Theta^2 = \{\theta^2_{10}, \theta^2_{200}, \theta^2_{201}, \theta^2_{3000}, \theta^2_{3001}, \theta^2_{3010}, \theta^2_{3011}, \theta^2_{400}, \theta^2_{401}\}.$$

Cada factorización contiene 18 parámetros, pero 9 de ellos están relacionados con los otros 9 (mostrados en la Tabla 7.1), mediante la relación $\theta^\ell_{i0s} + \theta^\ell_{i1s} = 1$, para $\ell = 1, 2$ e $i = 1, \ldots, 4$. ■

La estructura algebraica de las probabilidades marginales y condicionadas como funciones de los parámetros proporciona una información muy valiosa en muchas situaciones (ver Castillo, Gutiérrez y Hadi (1995c, 1996c)). Se comienza analizando la estructura de las probabilidades marginales asociadas al modelo probabilístico; a continuación se analiza el caso de las probabilidades condicionadas. Para simplificar la notación se considerará

la estructura paramétrica de un modelo probabilístico genérico definido por la factorización

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|s_i). \tag{7.32}$$

Teorema 7.3 *La probabilidad de cualquier realización de las variables* $\{x_1, \ldots, x_n\}$ *es un monomio en los parámetros que definen el modelo probabilístico de grado menor o igual que el número de variables. Sin embargo, es un polinomio de primer grado en cada uno de los parámetros.*

Demostración: Aplicando (7.32) se tiene que la probabilidad de una realización $(x_1, \ldots, x_n)$, es

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|s_i) = \prod_{i=1}^{n} \theta_{ix_is_i}.$$

Obsérvese que todos los parámetros que intervienen en el producto anterior están asociados a variables distintas. Por tanto $p(x_1, \ldots, x_n)$ es un monomio de grado menor o igual que el número de variables. Obsérvese también que $p(x_1, \ldots, x_n)$ puede resultar un polinomio si sólo se considera el conjunto de parámetros libres (ver (7.31)). Para ello sólo se necesita reemplazar los parámetros $\theta_{ir_is_i}$ por

$$\theta_{ir_is_i} = 1 - \prod_{j=0}^{r_i-1} \theta_{ijs_i}.$$

Esta substitución crea tantos monomios nuevos como cardinalidad tenga la variable X_i, pero cada uno de los monomios resultantes sigue siendo de primer grado en cada uno de los parámetros. ∎

El corolario siguiente determina la estructura algebraica de las probabilidades marginales asociadas a un modelo probabilístico.

Corolario 7.1 *La probabilidad marginal de cualquier conjunto de nodos* $Y \subset X$ *es un polinomio en los parámetros que definen el modelo probabilístico de grado menor o igual que el número de variables. Sin embargo, es un polinomio de primer grado en cada uno de los parámetros.*

Demostración: Con objeto de simplificar la notación, supóngase que $Y = \{X_1, \ldots, X_r\}$. Entonces $p(y)$ es la suma de las probabilidades de un conjunto de realizaciones:

$$\begin{aligned} p(y) &= p(x_1, \ldots, x_r) \\ &= \sum_{x_{r+1}, \ldots, x_n} p(x_1, \ldots, x_r, x_{r+1}, \ldots, x_n) \\ &= \sum_{x_{r+1}, \ldots, x_n} \prod_{i=1}^{n} \theta_{ix_is_i}. \end{aligned}$$

Por tanto, las probabilidades marginales de cualquier conjunto de nodos son también polinomios de grado uno en cada uno de los parámetros. ∎

Ejemplo 7.13 Estructura de las probabilidades marginales. Dada la factorización (7.12) con el conjunto asociado de parámetros Θ^1 mostrado en la Tabla 7.1, se puede calcular la probabilidad marginal de un conjunto de nodos utilizando la definición de probabilidad marginal dada en (3.4). Por ejemplo, las probabilidades marginales del nodo X_2 son

$$\begin{aligned} p(X_2 = 0) &= \sum_{x_1, x_3, x_4} p(x_1, 0, x_3, x_4) \\ &= \theta^1_{10}\,\theta^1_{200} + \theta^1_{201} - \theta^1_{10}\,\theta^1_{201}. \end{aligned}$$

y

$$\begin{aligned} p(X_2 = 1) &= \sum_{x_1, x_3, x_4} p(x_1, 1, x_3, x_4) \\ &= 1 - \theta^1_{10}\,\theta^1_{200} - \theta^1_{201} + \theta^1_{10}\,\theta^1_{201}. \end{aligned}$$

Estas expresiones son polinomios de grado dos (que es menor que el número de variables) en los parámetros mostrados en la Tabla 7.1. ∎

Corolario 7.2 *La función de probabilidad condicionada de un conjunto de nodos Y, dada la evidencia $E = e$, es una función racional de los parámetros. Además, el polinomio del denominador depende sólo de la evidencia.*

Demostración: Se tiene

$$p(y|e) = \frac{p(y, e)}{p(e)}. \tag{7.33}$$

Aplicando el Corolario 7.1, el numerador y el denominador de la función racional son polinomios de primer grado en cada uno de los parámetros, ya que son probabilidades marginales de un conjunto de variables. ∎

Obsérvese que en la ecuación (7.33) el polinomio del denominador es el mismo para cuaquier función de probabilidad $p(y|e)$, dada la evidencia $E = e$. Por tanto, en la práctica, será conveniente calcular y almacenar únicamente el polinomio del numerador, y obtener el denominador normalizando las probabilidades obtenidas.

Ejemplo 7.14 Estructura de las probabilidades condicionadas. Considérese de nuevo la factorización (7.12). La probabilidad condicionada de

X_2 dada la evidencia $X_3 = 0$ se puede obtener como

$$\begin{aligned} p(X_2 = 0|X_3 = 0) &= \frac{\sum\limits_{x_1,x_4} p(x_1, 0, 0, x_4)}{\sum\limits_{x_1,x_2,x_4} p(x_1, x_2, 0, x_4)} \\ &= \frac{\theta_{10}\,\theta_{200}\,\theta_{300} + \theta_{201}\,\theta_{301} - \theta_{10}\,\theta_{201}\,\theta_{301}}{\theta_{10}\,\theta_{300} + \theta_{301} - \theta_{10}\,\theta_{301}}, \end{aligned} \tag{7.34}$$

y

$$p(X_2 = 1|X_3 = 0) = \frac{\sum\limits_{x_1,x_4} p(x_1, 1, 0, x_4)}{\sum\limits_{x_1,x_2,x_4} p(x_1, x_2, 0, x_4)}$$

$$= \frac{\theta_{10}\,\theta_{300} - \theta_{10}\,\theta_{200}\,\theta_{300} + \theta_{301} - \theta_{10}\,\theta_{301} - \theta_{201}\,\theta_{301} + \theta_{10}\,\theta_{201}\,\theta_{301}}{\theta_{10}\,\theta_{300} + \theta_{301} - \theta_{10}\,\theta_{301}},$$

que son funciones racionales en los parámetros. Obsérvese que las expresiones anteriores han sido obtenidas aplicando directamente las fórmulas de la probabilidad correspondientes (el método de fuerza bruta). En el Capítulo 10 se presentarán algunos métodos más eficientes para calcular estas funciones de probabilidad en forma simbólica (propagación simbólica). ■

7.5.2 El Problema de la Compatibilidad

El análisis de la estructura paramétrica de las probabilidades, introducido en la sección anterior, permite resolver el problema de la compatibilidad de los modelos multifactorizados, es decir, permite obtener la familia de funciones de probabilidad compatible con el conjunto de factorizaciones dado en (7.29). Obsérvese que siempre existe una solución trivial para este problema, ya que el modelo de independencia total cumple todas las relaciones de independencia posibles. Sin embargo, se está interesado en obtener una función de probabilidad que cumpla las relaciones de independencia necesarias, pero que incluya el mínimo número posible de independencias adicionales.

La idea del método propuesto por Castillo, Gutiérrez y Hadi (1996b) es la de elegir una de las factorizaciones, por ejemplo P^1, y designarla como la *factorización de referencia* de la función de probabilidad. Los parámetros asociados, Θ^1, también se denominan *parámetros de referencia*. Una vez que la factorización de referencia ha sido fijada, el problema de la compatibilidad puede ser resuelto calculando las restricciones sobre los parámetros de referencia para que la función de probabilidad pueda ser factorizada según el resto de factorizaciones. Por tanto, se imponen secuencialmente a

la factorización de referencia las restricciones siguientes, dadas por el resto de las factorizaciones. Para cada P^ℓ, con $\ell = 2, \ldots, m$, se tiene

$$p^1(y_i^\ell | s_i^\ell) = p^1(y_i^\ell | b_i^\ell),\ i = 1, \ldots, n, \tag{7.35}$$

donde $B_i = \{Y_1^\ell, \ldots, Y_{i-1}^\ell\}$ y $S_i^\ell \subset B_i^\ell$. Obsérvese que las ecuaciones dadas en (7.35) determinan las restricciones necesarias para que se cumplan las relaciones de independencia $I(Y_i^\ell, B_i^\ell \setminus S_i^\ell | S_i^\ell)$, $i = 1, \ldots, n$, es decir, las independencias contenidas en la lista causal asociada a la factorización P^ℓ.

Las igualdades en (7.35) determinan el sistema de ecuaciones que permite resolver el problema de la compatibilidad. Cada una de las funciones de probabilidad condicionada en (7.35) es un cociente de polinomios (ver Corolario 7.2). Por tanto, el sistema (7.35) es un sistema polinomial de ecuaciones que puede ser resuelto de forma simultánea o secuencial, es decir hallando directamente el conjunto de soluciones del sistema, o resolviendo ecuación por ecuación utilizando las soluciones parciales anteriores para resolver cada nueva ecuación. Este método iterativo permite comprobar en cada una de las etapas si es necesario resolver la ecuación correspondiente, o si ésta es redundante, dado que la relación de independencia asociada está contenida en el modelo definido por las relaciones de independencia asociadas a las ecuaciones de las etapas anteriores.

Este método se describe en el algoritmo siguiente.

Algoritmo 7.1 Compatibilidad de modelos multifactorizados.

- **Datos:** Un modelo multifactorizado $\{\{p^\ell(y_1^\ell | s_1^\ell), \ldots, p^\ell(y_n^\ell | s_n^\ell)\}, \ell = 1, \ldots, m\}$, donde los parámetros de la primera factorización, Θ^1, se consideran los parámetros de referencia.
- **Resultados:** El conjunto de restricciones que tienen que cumplir los parámetros de referencia Θ^1 para que P^1 defina la función de probabilidad del modelo multifactorizado.

1. Considerar $\ell \leftarrow 2$ y *Ecuaciones* $= \phi$.
2. Para $i \leftarrow 1, \ldots, n$ **hacer**:

   Para cada valor j de Y_i^ℓ y cada realización s de S_i^ℓ :

   - Generar todas las realizaciones posibles de $B_i^\ell \setminus S_i^\ell$: $\{z_1, \ldots, z_k\}$.
   - Añadir las ecuaciones $\theta_{ijs}^\ell = p(y_i^\ell | z_1 \cup s) = \ldots = p(y_i^\ell | z_k \cup s)$ a la lista *Ecuaciones*.
3. Si $\ell = m$ ir a la Etapa 4. En caso contrario, asignar $\ell \leftarrow \ell + 1$ e ir a la Etapa 2.
4. Calcular simbólicamente las funciones de probabilidad condicionada que aparecen en *Ecuaciones*, utilizando los parámetros de referencia

Θ^1. Resolver el sistema de ecuaciones polinomiales resultante encontrando un sistema de ecuaciones simplificado que sea lógicamente equivalente al anterior y que proporcione las restricciones entre los parámetros de referencia.

5. Devolver las ecuaciones resultantes. ■

Obsérvese que en la Etapa 2 se añaden $card(S_i^\ell)$ ecuaciones al sistema y que cada una de estas ecuaciones contiene un total de $|B_i^\ell \setminus S_i^\ell|$ términos. Cada una de las ecuaciones contiene un único parámetro que no es de referencia y varios que sí lo son y que están asociados con $p^1(y_i^\ell|z_1 \cup s)$. Entonces, el sistema de ecuaciones resultante determina las restricciones de los parámetros de referencia y su relación con el resto de los parámetros.

El sistema de ecuaciones que se obtiene como resultado del algoritmo anterior puede ser resuelto directamete (obteniendo un sistema reducido lógicamente equivalente) utilizando un programa de cálculo simbólico como *Mathematica* (Wolfram (1991), Castillo y otros (1993)) o Maple (ver Char y otros (1991) y Abell y Braselton (1994)).

El Algoritmo 7.1 proporciona una solución para el problema de compatibilidad que surge en los modelos definidos por multigrafos (Sección 7.2) y los modelos definidos por listas de relaciones de independencia (Sección 7.3). Los ejemplos siguientes ilustran la aplicación de este algoritmo.

Ejemplo 7.15 Resolviendo un problema de compatibilidad. La función de probabilidad de la multired Bayesiana definida por las dos redes Bayesianas dadas en la Figura 7.5 puede ser factorizada como (7.12) y (7.13):

$$p(x_1, x_2, x_3, x_4) = p^1(x_1)p^1(x_2|x_1)p^1(x_3|x_1)p^1(x_4|x_2, x_3), \quad (7.36)$$

$$p(x_1, x_2, x_3, x_4) = p^2(x_1)p^2(x_2|x_1)p^2(x_4|x_2)p^2(x_3|x_1, x_4). \quad (7.37)$$

Los parámetros asociados a ambas factorizaciones se muestran en la Tabla 7.1. Para que el modelo sea consistente, las funciones de probabilidad en (7.36) y (7.37) deben coincidir. Por tanto un problema importante asociado a estos modelos es el de obtener las condiciones para que (7.36) y (7.37) definan las mismas funciones de probabilidad. En otras palabras, calcular las restricciones para los conjuntos de parámetros Θ^1 y Θ^2, dados en la Tabla 7.1, para que las dos factorizaciones definan el mismo modelo probabilístico.

Para resolver este problema, se selecciona una de las redes Bayesianas de la Figura 7.1 como la red de referencia, y se calculan las condiciones para que la otra red defina la misma función de probabilidad. Para ello, se aplica el Algoritmo 7.1 para resolver este problema de compatibilidad. En este caso $m = 2$. Supóngase que se selecciona (7.36) como la factorización de

referencia. Obsérvese que las ordenaciones ancestrales[1] implicadas por las factorizaciones son (X_1, X_2, X_3, X_4) y (X_1, X_2, X_4, X_3), respectivamente. Entonces, el Algoritmo 7.1 procede mediante las siguientes etapas:

Etapa 1: Asignar $\ell = 2$ y *Ecuaciones* $= \phi$. La ordenación ancestral de la segunda factorización implica la permutación:

$$(Y_1, Y_2, Y_3, Y_4) = (X_1, X_2, X_4, X_3).$$

Etapa 2: Para $i = 1$, se tiene $Y_1 = X_1$ y se considera la función de probabilidad $p^2(x_1)$. En este caso se tiene $B_1^2 = S_1^2 = \phi$. Por tanto, no se genera ninguna ecuación. Para $i = 2$, se considera $Y_2 = X_2$ y $p^2(x_2|x_1)$. Ahora se tiene $B_2^2 = S_2^2 = \{X_1\}$ que, de nuevo, no implica ninguna ecuación.

Para $i = 3$, se considera $Y_3 = X_4$ y $p^2(x_4|x_2)$. Se tiene $B_3^2 = \{X_1, X_2\}$, pero $S_3^2 = \{X_2\}$. Por tanto, para cada realización x_1 de X_1, se tiene

$$\theta^2_{40x_1} = p(X_4 = 0|x_1, X_2 = 0) = p(X_4 = 0|x_1, X_2 = 1),\ x_1 = 0, 1,$$

que implica

$$\begin{array}{rcl} \theta^2_{400} & = & p(X_4 = 0|0, 0) = p(X_4 = 0|0, 1), \\ \theta^2_{401} & = & p(X_4 = 0|1, 0) = p(X_4 = 0|1, 1). \end{array} \tag{7.38}$$

Estas ecuaciones son añadidas a la lista *Ecuaciones*.

Para $i = 4$, se considera $Y_4 = X_3$ y $p^2(x_3|x_1, x_4)$. En este caso, se tiene $B_4^2 = \{X_1, X_2, X_4\}$ y $S_4^2 = \{X_1, X_4\}$. Por tanto, para cada realización x_2 de X_2 se tiene

$$\theta^2_{3000} = p(X_3 = 0|X_1 = 0, x_2, X_4 = 0),$$

$$\theta^2_{3001} = p(X_3 = 0|X_1 = 0, x_2, X_4 = 1),$$

$$\theta^2_{3010} = p(X_3 = 0|X_1 = 1, x_2, X_4 = 0),$$

$$\theta^2_{3011} = p(X_3 = 0|X_1 = 1, x_2, X_4 = 1).$$

Estas relaciones implican las ecuaciones:

$$\begin{array}{l} \theta^2_{3000} = p(X_3 = 0|0, 0, 0) = p(X_3 = 0|0, 1, 0), \\ \theta^2_{3001} = p(X_3 = 0|0, 0, 1) = p(X_3 = 0|0, 1, 1), \\ \theta^2_{3010} = p(X_3 = 0|1, 0, 0) = p(X_3 = 0|1, 1, 0), \\ \theta^2_{3011} = p(X_3 = 0|1, 0, 1) = p(X_3 = 0|1, 1, 1), \end{array} \tag{7.39}$$

[1] El Algoritmo 4.6 proporciona un procedimiento automático para generar una ordenación ancestral de un grafo dirigido acíclico.

que se añaden a la lista *Ecuaciones*. Dado que $i = 4 = n$, la Etapa 2 finaliza.

Etapa 3: Dado que $\ell = 2 = m$, se pasa a la Etapa 4.

Etapa 4: Después de calcular simbólicamente las funciones de probabilidad condicionada en (7.38) y (7.39) se obtiene el sistema de ecuaciones siguiente:

$$\begin{aligned}
\theta^2_{400} &= \theta^1_{300}\theta^1_{4000} + \theta^1_{4001} - \theta^1_{300}\theta^1_{4001} \\
&= \theta^1_{301}\theta^1_{4000} + \theta^1_{4001}(1 - \theta^1_{301}), \\
\theta^2_{401} &= \theta^1_{300}\theta^1_{4010} + \theta^1_{4011}(1 - \theta^1_{300}) \\
&= \theta^1_{301}\theta^1_{4010} + \theta^1_{4011}(1 - \theta^1_{301}), \\
\theta^2_{3000} &= \frac{\theta^1_{300}\theta^1_{4000}}{\theta^1_{300}\theta^1_{4000}+\theta^1_{4001}(1-\theta^1_{300})} = \frac{\theta^1_{300}\theta^1_{4010}}{\theta^1_{300}\theta^1_{4010}+\theta^1_{4011}(1-\theta^1_{300})}, \\
\theta^2_{3001} &= \frac{\theta^1_{300}\big(1-\theta^1_{4000}\big)}{1-\theta^1_{300}\theta^1_{4000}+\theta^1_{4001}(\theta^1_{300}-1)} = \frac{\theta^1_{300}\big(1-\theta^1_{4010}\big)}{1-\theta^1_{300}\theta^1_{4010}+\theta^1_{4011}(\theta^1_{300}-1)}, \\
\theta^2_{3010} &= \frac{\theta^1_{301}\theta^1_{4000}}{\theta^1_{301}\theta^1_{4000}+\theta^1_{4001}(1-\theta^1_{301})} = \frac{\theta^1_{301}\theta^1_{4010}}{\theta^1_{301}\theta^1_{4010}+\theta^1_{4011}(1-\theta^1_{301})}, \\
\theta^2_{3011} &= \frac{\theta^1_{301}\big(1-\theta^1_{4000}\big)}{1-\theta^1_{301}\theta^1_{4000}+\theta^1_{4001}(\theta^1_{301}-1)} = \frac{\theta^1_{301}\big(1-\theta^1_{4010}\big)}{1-\theta^1_{301}\theta^1_{4010}+(\theta^1_{301}-1)}.
\end{aligned} \tag{7.40}$$

Obsérvese que las dos primeras ecuaciones se han obtenido de (7.38) y las cuatro últimas de (7.39). Al resolver este sistema de ecuaciones en los parámetros Θ^1, se obtienen las siguientes soluciones:

$$\begin{array}{ll}
\text{Solución 1}: & \{\theta^1_{300} = 0,\ \ \theta^1_{301} = 0\}, \\
\text{Solución 2}: & \{\theta^1_{300} = 1,\ \ \theta^1_{301} = 1\}, \\
\text{Solución 3}: & \{\theta^1_{4000} = \theta^1_{4001},\ \ \theta^1_{4010} = \theta^1_{4011}\}.
\end{array} \tag{7.41}$$

Por tanto, la familia de funciones de probabilidad que cumple las relaciones de independencia condicional implicadas por los dos grafos de la Figura 7.5 está determinada por los parámetros de la Tabla 7.1 (factorización (7.12)) con uno de los tres conjuntos de restricciones dados en (7.41). Obsérvese, que las dos primeras soluciones implican funciones de probabilidad extremas.

La Figura 7.6(a) muestra un ejemplo numérico de una función de probabilidad, definida mediante la factorización (7.36), que satisface las restricciones dadas en (7.41). Por tanto, esta función de probabilidad contiene todas la relaciones de independencia implicadas por los dos grafos de la Figura 7.5. Una vez que se han hallado los parámetros de referencia, también se puede obtener el resto de los parámetros utilizando las relaciones entre ellos dadas en (7.40). Los valores numéricos de los parámetros

de Θ_2 que definen el mismo modelo probabilístico que los parámetros de referencia dados en la Figura 7.6(a) se muestran en la Figura 7.6(b). Por tanto, ambas redes Bayesianas definen la misma función de probabilidad que incluye las independencias contenidas en ambos grafos.

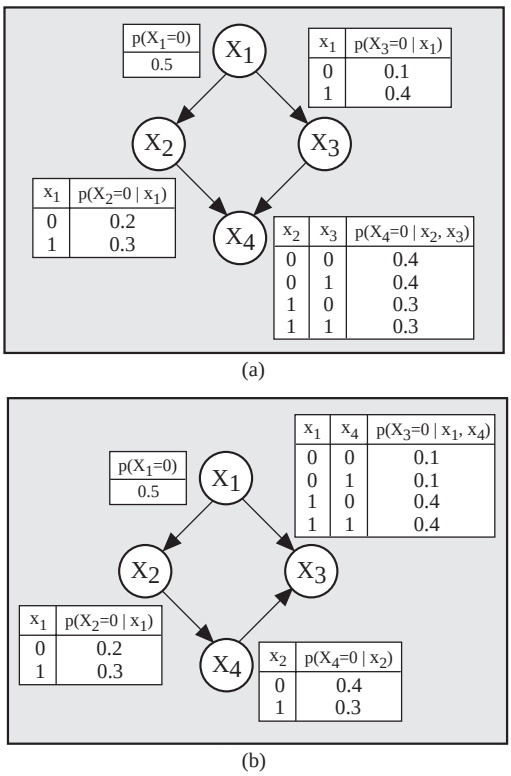


FIGURA 7.6. Dos factorizaciones distintas de la misma función de probabilidad, que contiene todas las independencias dadas por las dos redes Bayesianas del Ejemplo 7.7.

En el ejemplo anterior, se ha elegido la factorización (7.36) como factorización de referencia y se obtuvieron las restricciones para los parámetros asociados de forma que la función de probabilidad (7.37) coincida con (7.36). De forma similar, se puede aplicar el algoritmo utilizando (7.37) como factorización de referencia y hallar las restricciones sobre los parámetros para que la función de probabilidad (7.36) coincida con (7.37). El lector puede comprobar que, en este caso, se obtienen las siguientes

restricciones:

$$\begin{aligned} &\text{Solución 1}: \quad \{\theta^2_{3010} = \theta^2_{3011}, \theta^2_{3000} = \theta^2_{3001}\}, \\ &\text{Solución 2}: \quad \{\theta^2_{400} = \theta^2_{401}, \theta^2_{3010} = \theta^2_{3000}, \theta^2_{3011} = \theta^2_{3001}\}. \end{aligned} \tag{7.42}$$

La solución del problema de compatibilidad viene dada por las ecuaciones (7.41) y (7.42), que proporcionan las restricciones sobre los parámetros Θ^1 y Θ^2, respectivamente. ■

En algunos casos, las factorizaciones que componen un modelo multifactorizado pueden tener varias relaciones de independencia comunes. En esos casos, el Algoritmo 7.1 puede ser mejorado considerando una sóla vez cada independencia, es decir, reduciendo el conjunto de ecuaciones. Esta idea se ilustra en el ejemplo siguiente.

Ejemplo 7.16 Mejorando el método de compatibilidad. Considérese el modelo multifactorizado asociado a las dos redes Bayesianas (D^1, P^1) y (D^2, P^2) dadas en las Figuras 7.7(a) y (b), respectivamente. Entonces, la función de probabilidad de $X = \{X_1, \ldots, X_7\}$ puede ser factorizada como

$$p(x) = p^1(x_1)p^1(x_2|x_1)p^1(x_3|x_1)p^1(x_4|x_2, x_3)p^1(x_5|x_3)p^1(x_6|x_4)p^1(x_7|x_4), \tag{7.43}$$

y

$$p(x) = p^2(x_2)p^2(x_1|x_2)p^2(x_3)p^2(x_4|x_2, x_3)p^2(x_7|x_4)p^2(x_5|x_7)p^2(x_6|x_4). \tag{7.44}$$

Obsérvese que las funciones de probabilidad condicionada en (7.43) y (7.44) están determinadas por las ordenaciones ancestrales de las variables implicadas por los grafos de la Figura 7.7. La Tabla 7.2 muestra los parámetros asociados a estas factorizaciones. Ahora se desea calcular la función de probabilidad que satisface ambas factorizaciones.

Se selecciona (7.43) como la factorización de referencia. La factorización (7.44) implica las siguientes relaciones entre los parámetros:

$$\begin{gathered} p^2(x_2) = p(x_2) = p^1(x_2), \\ p^2(x_1|x_2) = p(x_1|x_2) = p^1(x_1|x_2), \\ p^2(x_3) = p(x_3|x_1, x_2) = p^1(x_3|x_1), \\ p^2(x_4|x_2, x_3) = p(x_4|x_1, x_2, x_3) = p^1(x_4|x_2, x_3), \\ p^2(x_7|x_4) = p(x_7|x_1, x_2, x_3, x_4) = p^1(x_7|x_4), \\ p^2(x_5|x_7) = p(x_5|x_1, x_2, x_3, x_4, x_7) = p^1(x_5|x_3), \\ p^2(x_6|x_4) = p(x_6|x_1, x_2, x_3, x_4, x_7, x_5) = p^1(x_6|x_4). \end{gathered}$$

Por tanto, se tiene el sistema de ecuaciones

$$\theta^2_{20} = \theta^1_{10}(\theta^1_{200} - \theta^1_{201}) + \theta^1_{201},$$

$$\theta^2_{100} = \frac{\theta^1_{10}\theta^1_{200}}{\theta^1_{10}\theta^1_{200} + \theta^1_{201} - \theta^1_{10}\theta^1_{201}}, \quad \theta^2_{101} = \frac{\theta^1_{10}(1-\theta^1_{200})}{1 - \theta^1_{10}\theta^1_{200} - \theta^1_{201} + \theta^1_{10}\theta^1_{201}},$$

$$\theta^2_{30} = \theta^1_{300} = \theta^1_{301},$$

$$\begin{gathered}\theta^2_{4000} = \theta^1_{4000},\ \theta^2_{4001} = \theta^1_{4001},\ \theta^2_{4010} = \theta^1_{4010},\ \theta^2_{4011} = \theta^1_{4011}, \\ \theta^2_{700} = \theta^1_{700},\ \theta^2_{701} = \theta^1_{701}.\end{gathered} \tag{7.45}$$

$$\theta^2_{500} = \theta^1_{500} = \theta^1_{501}, \quad \theta^2_{501} = \theta^1_{500} = \theta^1_{501},$$

$$\theta^2_{600} = \theta^1_{600},\ \theta^2_{601} = \theta^1_{601}.$$

Si se eliminan los parámetros de Θ^2 de las ecuaciones anteriores se obtiene la solución:

$$\theta^1_{300} = \theta^1_{301} \quad \text{y} \quad \theta^1_{500} = \theta^1_{501}. \tag{7.46}$$

Por tanto, la familia de funciones de probabilidad que cumplen las relaciones de independencia implicadas por los dos grafos de la Figura 7.7 está caracterizada por los parámetros de la Tabla 7.2 (Θ^1) con las restricciones (7.46).

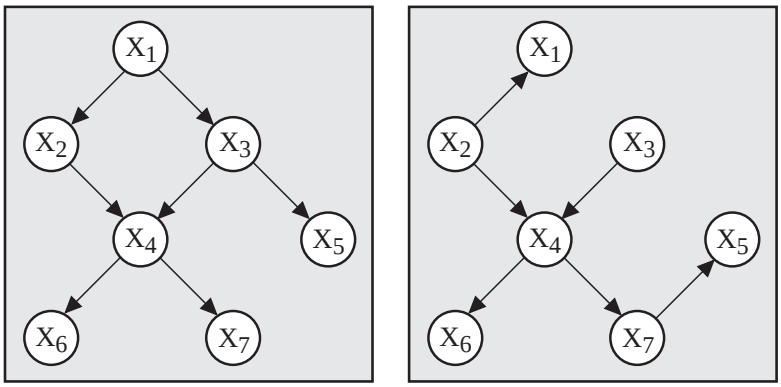


FIGURA 7.7. Dos grafos dirigidos acíclicos que definen una multired Bayesiana.

Por otra parte, si se eliminan los parámetros de Θ^1 de este sistema, se obtiene

$$\theta^2_{500} = \theta^2_{501}. \tag{7.47}$$

Por tanto, la familia de funciones de probabilidad que cumplen las relaciones de independencia implicadas por los dos grafos de la Figura 7.7

Variable	Θ^1	Θ^2
X_1	$\theta^1_{10} = p^1(\bar{x}_1)$	$\theta^2_{100} = p^2(\bar{x}_1\|\bar{x}_2)$
		$\theta^2_{101} = p^2(\bar{x}_1\|x_2)$
X_2	$\theta^1_{200} = p^1(\bar{x}_2\|\bar{x}_1)$	$\theta^2_{20} = p^2(\bar{x}_2)$
	$\theta^1_{201} = p^1(\bar{x}_2\|x_1)$	
X_3	$\theta^1_{300} = p^1(\bar{x}_3\|\bar{x}_1)$	$\theta^2_{30} = p^2(\bar{x}_3)$
	$\theta^1_{301} = p^1(\bar{x}_3\|x_1)$	
X_4	$\theta^1_{4000} = p^1(\bar{x}_4\|\bar{x}_2\bar{x}_3)$	$\theta^2_{4000} = p^2(\bar{x}_4\|\bar{x}_2\bar{x}_3)$
	$\theta^1_{4001} = p^1(\bar{x}_4\|\bar{x}_2 x_3)$	$\theta^2_{4001} = p^2(\bar{x}_4\|\bar{x}_2 x_3)$
	$\theta^1_{4010} = p^1(\bar{x}_4\|x_2\bar{x}_3)$	$\theta^2_{4010} = p^2(\bar{x}_4\|x_2\bar{x}_3)$
	$\theta^1_{4011} = p^1(\bar{x}_4\|x_2 x_3)$	$\theta^2_{4011} = p^2(\bar{x}_4\|x_2 x_3)$
X_5	$\theta^1_{500} = p^1(\bar{x}_5\|\bar{x}_3)$	$\theta^2_{500} = p^2(\bar{x}_5\|\bar{x}_7)$
	$\theta^1_{501} = p^1(\bar{x}_5\|x_3)$	$\theta^2_{501} = p^2(\bar{x}_5\|x_7)$
X_6	$\theta^1_{600} = p^1(\bar{x}_6\|\bar{x}_4)$	$\theta^2_{600} = p^2(\bar{x}_6\|\bar{x}_4)$
	$\theta^1_{601} = p^1(\bar{x}_6\|x_4)$	$\theta^2_{601} = p^2(\bar{x}_6\|x_4)$
X_7	$\theta^1_{700} = p^1(\bar{x}_7\|\bar{x}_4)$	$\theta^2_{700} = p^2(\bar{x}_7\|\bar{x}_4)$
	$\theta^1_{701} = p^1(\bar{x}_7\|x_4)$	$\theta^2_{601} = p^2(\bar{x}_6\|x_4)$

TABLA 7.2. Conjuntos de parámetros Θ^1 y Θ^2 asociados a las dos factorizaciones en (7.43) y (7.44), respectivamente.

también está caracterizada a través de los parámetros de la Tabla 7.2 (Θ^2) con las restricciones (7.47).

El método anterior puede ser mejorado sustancialmente considerando sólo las ecuaciones correspondientes a relaciones de independencia que no puedan ser obtenidas utilizando la información previa. En este caso, se puede comprobar si las relaciones de independencia asociadas a la factorización (7.44) son satisfechas por la función de probabilidad definida por la factorización (7.43), es decir, se puede comprobar si las relaciones de independencia se cumplen en el grafo de la Figura 7.7(a).

La factorización (7.44) implica la ordenación ancestral de los nodos

$$\{X_2, X_1, X_3, X_4, X_7, X_5, X_6\}$$

y las relaciones de independencia siguientes:

$$I(X_2, \phi|\phi),$$
$$I(X_1, \phi|X_2),$$
$$I(X_3, \{X_1, X_2\}|\phi),$$
$$I(X_4, X_1|\{X_2, X_3\}),$$
$$I(X_7, \{X_1, X_2, X_3\}|X_4),$$
$$I(X_5, \{X_1, X_2, X_3, X_4\}|X_7),$$
$$I(X_6, \{X_1, X_2, X_3, X_5, X_7\}|X_4).$$

Las únicas relaciones de independencia que no se pueden obtener del grafo de la Figura 7.7(a) son $I(X_3, \{X_1, X_2\}|\phi)$ y $I(X_5, \{X_1, X_2, X_3, X_4\}|X_7)$. Por tanto, se tiene el sistema de ecuaciones:

$$\theta^2_{30} = \theta^1_{300} = \theta^1_{301}, \;\; \theta^2_{500} = \theta^1_{500} = \theta^1_{501}, \;\; \text{y} \;\; \theta^2_{501} = \theta^1_{500} = \theta^1_{501},$$

que es bastante más sencillo que el sistema (7.45). Obsérvese que, ahora, las soluciones dadas en (7.46) y (7.47) se pueden obtener trivialmente del sistema reducido. ∎

La mejora introducida en el ejemplo anterior puede ser fácilmente incorporada en el Algoritmo 7.1.

7.6 Modelos Normales Multifactorizados

En esta Sección se analiza el caso de los modelos probabilísticos normales definidos por un conjunto de factorizaciones distintas. En este caso la función de probabilidad conjunta sobre $X = \{X_1, \ldots, X_n\}$ es una distribución normal definida por $N(\mu, \Sigma)$, donde μ es el vector de medias y Σ es la matriz de covarianzas. Por tanto, el conjunto de parámetros que definen estos modelos consiste en n medias $\{\mu_i; i = 1, \ldots, n\}$ y las $n(n+1)/2$ varianzas y covarianzas $\{\sigma_{ij};\ i, j = 1, \ldots, n\}$. La matriz de covarianzas Σ es independiente del vector de medias μ, que es un conjunto de parámetros de localización. Por tanto, para el propósito de determinar relaciones de independencia, los únicos parámetros relevantes serán las varianzas y covarianzas del modelo.

El problema de compatibilidad asociado al conjunto de factorizaciones de un modelo normal multifactorizado se reduce al problema de encontrar la matriz de covarianzas de la variable aleatoria multidimensional que sea compatible con las factorizaciones dadas, o con las relaciones de independencia implicadas por ellas. De manera similar al caso de los modelos multinomiales multifactorizados analizados en la Sección 7.5, se pueden designar como parámetros de referencia a los parámetros asociados a la matriz de

covarianzas de la primera factorización. También se puede comenzar, de forma alternativa, con la matriz de covarianzas de una función de probabilidad general (completamente parametrizada) y calcular las restricciones de éstos parámetros para que esta función cumpla todas las independencias implicadas por todas las factorizaciones del modelo $P^\ell, \ell = 1, \ldots, m$ (ver Definición 7.3), es decir, se ha de tener

$$I(X_i^\ell, B_i^\ell \setminus S_i^\ell | S_i^\ell) \Leftrightarrow p(x_i^\ell | b_i^\ell) = p(x_i^\ell | s_i^\ell);\ i = 1, \ldots, n, \tag{7.48}$$

donde $p(.)$ denota la función de densidad normal completa y x_i^ℓ, b_i^ℓ, y s_i^ℓ son realizaciones de X_i^ℓ, B_i^ℓ, y S_i^ℓ, respectivamente. Obsérvese que esta notación supone, sin pérdida de generalidad, que $(X_1^\ell, \ldots, X_n^\ell)$ están dadas siguiendo una ordenación ancestral para cada ℓ.

El sistema de ecuaciones (7.48) proporciona una serie de restricciones sobre los parámetros que definen el modelo probabilístico que es compatible con todas las factorizaciones dadas. El teorema siguiente proporciona un método sencillo para calcular estas restricciones.

Teorema 7.4 Independencia condicional en modelos normales. *Sea $X = \{X_1, \ldots, X_n\}$ un conjunto de variables aleatorias y $\{V, Y, Z\}$ una partición de X. Supóngase que las variables siguen una distribución normal*

$$N\left(\begin{bmatrix} \mu_V \\ \mu_Y \\ \mu_Z \end{bmatrix}; \begin{bmatrix} \Sigma_{VV} & \Sigma_{VY} & \Sigma_{VZ} \\ \Sigma_{YV} & \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZV} & \Sigma_{ZY} & \Sigma_{ZZ} \end{bmatrix}\right), \tag{7.49}$$

donde se ha utilizado la descomposición en bloques asociada a la partición (V, Y, Z) y se ha supuesto que la matriz de covarianzas correspondiente a (V, Y) es no singular. Entonces, una condición necesaria y suficiente para que se cumpla la relación de independencia condicional $I(V, Y|Z)$, es decir, para que se cumpla $p(v|y, z) = p(v|z)$ es que $\Sigma_{VY} = \Sigma_{VZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$.

Una demostración de este teorema así como resultados relacionados adicionales se pueden encontrar en Whittaker (1990), Johnson y Wichern (1988), y Rencher (1995).

Corolario 7.3 Independencia Condicional a través de la Matriz de Precisión. *Sea X una variable aleatoria distribuida de forma normal y sea $\{V, Y, Z\}$ una partición de X como la indicada en el Teorema 7.4. Sea $W = \Sigma^{-1}$ la matriz de precisión del modelo, es decir, la inversa de la matriz de covarianzas Σ. Entonces, se cumple $I(V, Y|Z)$ si y sólo si el bloque W_{VY} de la matriz W es la matriz nula.*

El teorema siguiente muestra que, para variables aleatorias normales, los términos dependencia y correlación son equivalentes, así como los términos dependencia condicional y correlación parcial.

Teorema 7.5 Independencia condicional y correlación parcial. *Sea* (V, Y, Z) *una variable aleatoria distribuida de forma normal. Entonces* V *e* Y *no están correlacionados dado* Z *si y sólo si* $I(V, Y|Z)$.

A continuación se introducen algunos ejemplos ilustrativos de aplicación.

Ejemplo 7.17 Modelo normal dado por una lista de relaciones de independencia. Considérese un vector de variables aleatorias normales $X = \{X_1, X_2, X_3\}$ con la matriz de covarianzas no singular

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}. \tag{7.50}$$

A continuación se calculan las restricciones que han de imponerse a los parámetros a fin de que la función de probabilidad conjunta resultante satisfaga las relaciones de independencia del modelo Dada la lista de relaciones de independencia

$$M = \{I(X_1, X_2|X_3), I(X_1, X_3|X_2), I(X_2, X_3|X_1)\}, \tag{7.51}$$

Para ello se comienza con el modelo probabilístico general de cuatro variables dado por Σ en (7.50), y se calculan las restricciones impuestas por la primera relación de independencia en M $I(X_1, X_2|X_3)$. Denotando $(V, Y, Z) = (X_1, X_2, X_3)$, el Teorema 7.4 da la restricción

$$\sigma_{12} = \frac{\sigma_{13}\sigma_{32}}{\sigma_{33}}. \tag{7.52}$$

Obsérvese que dado que Σ se ha supuesto no singular, entonces $\sigma_{ii} > 0$, para $i = 1, 2, 3$. Por tanto, la matriz de covarianzas que cumple la primera relación de independencia $I(X_1, X_2|X_3)$ tendrá la estructura

$$\Sigma = \begin{pmatrix} \sigma_{11} & \frac{\sigma_{13}\sigma_{23}}{\sigma_{33}} & \sigma_{13} \\ \frac{\sigma_{13}\sigma_{23}}{\sigma_{33}} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix}.$$

De manera similar, las dos restantes relaciones de independencia en M implican las restricciones:

$$\sigma_{13} = \frac{\sigma_{12}\sigma_{23}}{\sigma_{22}}, \tag{7.53}$$

$$\sigma_{23} = \frac{\sigma_{21}\sigma_{13}}{\sigma_{11}}. \tag{7.54}$$

Obsérvese que, dada la simetría de Σ, se tiene $\sigma_{ij} = \sigma_{ji}$. Por tanto, se tienen seis parámetros distintos sujetos a las tres restricciones dadas en

(7.52), (7.53) y (7.54). Si se calculan los parámetros asociados a covarianzas en función de aquellos asociados a varianzas se obtienen las siguientes soluciones para la matriz Σ:

$$\begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{pmatrix}, \quad \begin{pmatrix} \sigma_{11} & -\delta_{12} & -\delta_{13} \\ -\delta_{12} & \sigma_{22} & \delta_{23} \\ -\delta_{13} & \delta_{23} & \sigma_{33} \end{pmatrix},$$
$$\begin{pmatrix} \sigma_{11} & \delta_{12} & \delta_{13} \\ \delta_{12} & \sigma_{22} & \delta_{23} \\ \delta_{13} & \delta_{23} & \sigma_{33} \end{pmatrix}, \quad \begin{pmatrix} \sigma_{11} & -\delta_{12} & \delta_{13} \\ -\delta_{12} & \sigma_{22} & -\delta_{23} \\ \delta_{13} & -\delta_{23} & \sigma_{33} \end{pmatrix}, \tag{7.55}$$
$$\begin{pmatrix} \sigma_{11} & \delta_{12} & -\delta_{13} \\ \delta_{12} & \sigma_{22} & -\delta_{23} \\ -\delta_{13} & -\delta_{23} & \sigma_{33} \end{pmatrix},$$

donde $\delta_{ij} = \sqrt{\sigma_{ii}\sigma_{jj}}$.

Por otra parte, si se despejan las varianzas en función de las covarianzas, se tienen las siguientes soluciones:

$$\begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{pmatrix}, \quad \begin{pmatrix} \dfrac{\sigma_{12}\sigma_{13}}{\sigma_{23}} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \dfrac{\sigma_{12}\sigma_{23}}{\sigma_{13}} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \dfrac{\sigma_{13}\sigma_{23}}{\sigma_{12}} \end{pmatrix}. \tag{7.56}$$

Además, habrá que imponer las restricciones obvias para que estas matrices sean definidas no negativas.

Obsérvese que las primeras soluciones en (7.55) y (7.56) coinciden, y que cada una de las cuatro soluciones restantes en (7.55) cumple la segunda solución en (7.56). Además, si se sustituyen dos cualesquiera de las tres ecuaciones (7.52), (7.53) y (7.54) en la tercera ecuación, se obtiene

$$\sigma_{11}\sigma_{22}\sigma_{33} = \sigma_{12}\sigma_{13}\sigma_{23}, \tag{7.57}$$

supuesto que las covarianzas son distintas de cero. Por tanto, los elementos en la matriz de covarianzas son tales que el producto de las varianzas coincide con el producto de las covarianzas. Como puede verse en (7.55) y (7.56) esta propiedad se satisface en las distintas soluciones del problema. Por tanto, se concluye que el modelo probabilístico normal definido por la lista de relaciones de independencia M en (7.51) puede ser definido por una cualquiera de estas matrices de covarianza. ■

En el Ejemplo 7.17, se han calculado las restricciones que han de imponerse a un modelo normal para que contenga las independencias dadas en una lista de relaciones de independencia. En el ejemplo siguiente, se obtienen las restricciones que es necesario imponer a los parámetros del modelo para que éste sea compatible con un conjunto de factorizaciones dado.

Ejemplo 7.18 Modelo normal multifactorizado. Considérese un conjunto de cuatro variables normales $\{X_1, X_2, X_3, X_4\}$ cuya función de probabilidad conjunta satisface las dos factorizaciones siguientes:

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3), \tag{7.58}$$
$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_4|x_2)p(x_3|x_1, x_4). \tag{7.59}$$

Obsérvese que estas dos factorizaciones son las dadas en (7.17) y (7.19). En el Ejemplo 7.9 se ha visto que estas factorizaciones tienen asociadas las dos listas de relaciones de independencia

$$\begin{array}{rcl} M_1 & = & \{I(X_2, X_3|X_1), I(X_1, X_4|X_2, X_3)\}y \\ M_2 & = & \{I(X_1, X_4|X_2), I(X_2, X_3|X_1, X_4)\}, \end{array} \tag{7.60}$$

dadas en (7.24) y (7.26), respectivamente. Combinando M_1 y M_2, se obtiene

$$M = \{I(X_2, X_3|X_1), I(X_1, X_4|X_2, X_3), I(X_1, X_4|X_2), I(X_2, X_3|X_1, X_4)\}.$$

Por tanto, el modelo multifactorizado definido por (7.58) y (7.59) puede ser obtenido a partir de la lista M. Para obtener este modelo puede seguirse el mismo procedimiento empleado en el Ejemplo 7.17. De esta forma se podrán obtener las restricciones que impone M en la matriz de covarianzas Σ.

Aplicando el Teorema 7.4, se obtienen las siguientes restricciones, correspondientes a las cuatro relaciones de independencia dadas en M:

$$\begin{array}{c} \sigma_{23} = \dfrac{\sigma_{21}\sigma_{13}}{\sigma_{11}}, \\ \sigma_{14} = \begin{pmatrix} \sigma_{12} & \sigma_{13} \end{pmatrix} \begin{pmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{24} \\ \sigma_{34} \end{pmatrix}, \\ \sigma_{14} = \dfrac{\sigma_{12}\sigma_{24}}{\sigma_{22}}, \\ \sigma_{23} = \begin{pmatrix} \sigma_{21} & \sigma_{24} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{14} \\ \sigma_{41} & \sigma_{44} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{13} \\ \sigma_{43} \end{pmatrix}. \end{array} \tag{7.61}$$

Resolviendo este sistema de ecuaciones, utilizando, por ejemplo, un programa de cálculo simbólico como *Mathematica*, se obtiene la siguiente matriz de covarianzas:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \dfrac{\sigma_{11}\sigma_{34}}{\sigma_{13}} \\ \sigma_{12} & \sigma_{22} & \dfrac{\sigma_{12}\sigma_{13}}{\sigma_{11}} & \dfrac{\sigma_{11}\sigma_{22}\sigma_{34}}{\sigma_{12}\sigma_{13}} \\ \sigma_{13} & \dfrac{\sigma_{12}\sigma_{13}}{\sigma_{11}} & \sigma_{33} & \sigma_{34} \\ \dfrac{\sigma_{11}\sigma_{34}}{\sigma_{13}} & \dfrac{\sigma_{11}\sigma_{22}\sigma_{34}}{\sigma_{12}\sigma_{13}} & \sigma_{34} & \sigma_{44} \end{pmatrix}. \tag{7.62}$$

a la que hay que añadir las condiciones para que sea definida no negativa.

Por tanto, el modelo probabilístico normal compatible con las dos factorizaciones dadas en (7.58) y (7.59) está caracterizado por la matriz de covarianzas en (7.62). ∎

En los Ejemplos 7.17 y 7.18, se ilustra la forma de obtener un modelo probabilístico normal por medio de una lista de relaciones de independencia o de un conjunto de factorizaciones, respectivamente. En el ejemplo siguiente se ilustra cómo construir un modelo normal a partir de un multigrafo.

Ejemplo 7.19 Modelo normal definido por un multigrafo. Considérese el modelo multifactorizado introducido en el Ejemplo 7.16 a través de las dos redes Bayesianas (D^1, P^1) y (D^2, P^2) de las Figuras 7.7(a) y (b), respectivamente. Supóngase que las variables están distribuidas de forma normal. En este ejemplo se calcula la matriz de covarianzas del modelo probabilístico normal definido por las dos redes Bayesianas. Aplicando el criterio de D-separación a D^1 y D^2 (ver Definición 5.4), se pueden obtener las siguientes relaciones de independencia condicional:

$$M_1 = \left\{ \begin{array}{c} I(X_7, \{X_1, X_2, X_3\}|X_4), I(X_4, X_1|\{X_2, X_3\}), \\ I(X_3, X_2|X_1), I(X_5, \{X_1, X_2, X_3, X_4, X_7\}|X_3) \\ I(X_6, \{X_1, X_2, X_3, X_5, X_7\}|X_4) \end{array} \right\}, \quad (7.63)$$

$$M_2 = \left\{ \begin{array}{c} I(X_3, \{X_1, X_2\}|\phi), I(X_7, \{X_1, X_2, X_3\}|X_4), \\ I(X_4, X_1|\{X_2, X_3\}), I(X_5, \{X_1, X_2, X_3, X_4\}|X_7) \\ I(X_6, \{X_1, X_2, X_3, X_5, X_7\}|X_4) \end{array} \right\}, \quad (7.64)$$

Obsérvese que también es posible obtener M_1 y M_2 a partir de las factorizaciones dadas en (7.43) y (7.44), respectivamente.

En primer lugar, se calcula la matriz de covarianzas del modelo probabilístico normal definido por la segunda red Bayesiana D^2:

- La relación de independencia $I(X_3, \{X_1, X_2\}|\phi)$ implica:

$$\begin{pmatrix} \sigma_{31} \\ \sigma_{32} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad (7.65)$$

- $I(X_1, X_4|\{X_2, X_3\})$ implica:

$$\sigma_{14} = \begin{pmatrix} \sigma_{42} & \sigma_{43} \end{pmatrix} \begin{pmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{24} \\ \sigma_{34} \end{pmatrix}, \quad (7.66)$$

- $I(X_7, \{X_1, X_2, X_3\}|X_4)$ implica:

$$\begin{pmatrix} \sigma_{17} \\ \sigma_{27} \\ \sigma_{37} \end{pmatrix} = \frac{\sigma_{47}}{\sigma_{44}} \begin{pmatrix} \sigma_{14} \\ \sigma_{24} \\ \sigma_{34} \end{pmatrix}, \quad (7.67)$$

- $I(X_5, \{X_1, X_2, X_3, X_4\}|X_7)$ implica:

$$\begin{pmatrix} \sigma_{51} \\ \sigma_{52} \\ \sigma_{53} \\ \sigma_{54} \end{pmatrix} = \frac{\sigma_{57}}{\sigma_{77}} \begin{pmatrix} \sigma_{71} \\ \sigma_{72} \\ \sigma_{73} \\ \sigma_{74} \end{pmatrix}, \tag{7.68}$$

- $I(X_6, \{X_1, X_2, X_3, X_5, X_7\}|X_4)$ implica:

$$\begin{pmatrix} \sigma_{61} \\ \sigma_{62} \\ \sigma_{63} \\ \sigma_{65} \\ \sigma_{67} \end{pmatrix} = \frac{\sigma_{64}}{\sigma_{44}} \begin{pmatrix} \sigma_{14} \\ \sigma_{24} \\ \sigma_{34} \\ \sigma_{54} \\ \sigma_{74} \end{pmatrix}. \tag{7.69}$$

Resolviendo el sistema de ecuaciones (7.65)−(7.69), y considerando la simetría de la matriz de covarianza, $\sigma_{ij} = \sigma_{ji}$, se tiene

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & 0 & \alpha & \sigma_{15} & \delta & \beta \\ \sigma_{12} & \sigma_{22} & 0 & \frac{\sigma_{25}\,\sigma_{44}}{\sigma_{45}} & \sigma_{25} & \frac{\sigma_{25}\,\sigma_{46}}{\sigma_{45}} & \frac{\sigma_{25}\,\sigma_{47}}{\sigma_{45}} \\ 0 & 0 & \sigma_{33} & \frac{\sigma_{35}\,\sigma_{44}}{\sigma_{45}} & \sigma_{35} & \frac{\sigma_{35}\,\sigma_{46}}{\sigma_{45}} & \frac{\sigma_{35}\,\sigma_{47}}{\sigma_{45}} \\ \alpha & \frac{\sigma_{25}\,\sigma_{44}}{\sigma_{45}} & \frac{\sigma_{35}\,\sigma_{44}}{\sigma_{45}} & \sigma_{44} & \sigma_{45} & \sigma_{46} & \sigma_{47} \\ \sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_{55} & \frac{\sigma_{45}\,\sigma_{46}}{\sigma_{44}} & \sigma_{57} \\ \delta & \frac{\sigma_{25}\,\sigma_{46}}{\sigma_{45}} & \frac{\sigma_{35}\,\sigma_{46}}{\sigma_{45}} & \sigma_{46} & \frac{\sigma_{45}\,\sigma_{46}}{\sigma_{44}} & \sigma_{66} & \frac{\sigma_{46}\,\sigma_{47}}{\sigma_{44}} \\ \beta & \frac{\sigma_{25}\,\sigma_{47}}{\sigma_{45}} & \frac{\sigma_{35}\,\sigma_{47}}{\sigma_{45}} & \sigma_{47} & \sigma_{57} & \frac{\sigma_{46}\,\sigma_{47}}{\sigma_{44}} & \frac{\sigma_{47}\,\sigma_{57}}{\sigma_{45}} \end{pmatrix}, \tag{7.70}$$

donde

$$\alpha = \frac{\sigma_{12}\,\sigma_{25}\,\sigma_{44}}{\sigma_{22}\,\sigma_{45}}, \quad \beta = \frac{\sigma_{12}\,\sigma_{25}\,\sigma_{47}}{\sigma_{22}\,\sigma_{45}}, \quad \delta = \frac{\sigma_{12}\,\sigma_{25}\,\sigma_{46}}{\sigma_{22}\,\sigma_{45}}.$$

Además, habrá que imponer las restricciones para que estas matrices sean definidas no negativas.

Por tanto, la matriz de covarianzas del modelo probabilístico normal que tiene al grafo dirigido acíclico de la Figura 7.7(b) como I-mapa debe de cumplir (7.70).

A continuación se calcula la matriz de covarianzas que caracteriza al modelo probabilístico definido por los grafos dirigidos de la Figura 7.7. La unión de los modelos M_1 y M_2 solamente origina una nueva relación de independencia $I(X_5, \{X_1, X_2, X_4, X_7\}|X_3)$. Esta independencia implica las siguientes restricciones:

$$\begin{pmatrix} \sigma_{51} \\ \sigma_{52} \\ \sigma_{54} \\ \sigma_{57} \end{pmatrix} = \frac{\sigma_{53}}{\sigma_{33}} \begin{pmatrix} \sigma_{31} \\ \sigma_{32} \\ \sigma_{34} \\ \sigma_{37} \end{pmatrix}. \tag{7.71}$$

Resolviendo el sistema resultante, se tiene

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & 0 & 0 & 0 & 0 & 0 \\ \sigma_{12} & \sigma_{22} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_{33} & \rho & \sigma_{35} & -\frac{\sqrt{\sigma_{33}}\,\sigma_{46}}{\sqrt{\sigma_{44}}} & -\frac{\sqrt{\sigma_{33}}\,\sigma_{47}}{\sqrt{\sigma_{44}}} \\ 0 & 0 & \rho & \sigma_{44} & -\frac{\sigma_{35}\sqrt{\sigma_{44}}}{\sqrt{\sigma_{33}}} & \sigma_{46} & \sigma_{47} \\ 0 & 0 & \sigma_{35} & \tau & \sigma_{55} & -\frac{\sigma_{35}\,\sigma_{46}}{\sqrt{\sigma_{33}\sigma_{44}}} & -\frac{\sigma_{35}\,\sigma_{47}}{\sqrt{\sigma_{33}\sigma_{44}}} \\ 0 & 0 & -\frac{\sqrt{\sigma_{33}}\,\sigma_{46}}{\sqrt{\sigma_{44}}} & \sigma_{46} & -\frac{\sigma_{35}\,\sigma_{46}}{\sqrt{\sigma_{33}\sigma_{44}}} & \sigma_{66} & \frac{\sigma_{46}\,\sigma_{47}}{\sigma_{44}} \\ 0 & 0 & -\frac{\sqrt{\sigma_{33}}\,\sigma_{47}}{\sqrt{\sigma_{44}}} & \sigma_{47} & -\frac{\sigma_{35}\,\sigma_{47}}{\sqrt{\sigma_{33}\sigma_{44}}} & \frac{\sigma_{46}\,\sigma_{47}}{\sigma_{44}} & \frac{{\sigma_{47}}^2}{\sigma_{44}} \end{pmatrix},$$

donde

$$\tau = -\frac{\sigma_{35}\sqrt{\sigma_{44}}}{\sqrt{\sigma_{33}}}, \quad \rho = -\sqrt{\sigma_{33}\sigma_{44}}.$$

Además, hay que imponer las restricciones para que estas matrices sean definidas no negativas.

Por tanto, la matriz de covarianzas de la multired Bayesiana formado por los dos grafos dados en la Figura 7.7 ha de tener la estructura anterior. ∎

7.7 Modelos probabilísticos definidos Condicionalmente

En las secciones anteriores se han analizado los modelos multifactorizados, que permiten resolver el problema de compatibilidad de los modelos basados en multigrafos y los modelos basados en una lista de relaciones de independencia. En esta sección se trata el problema de la definición de un modelo probabilístico mediante un conjunto de funciones de probabilidad condicionada. Los modelos definidos de esta forma se denominan *modelos probabilísticos definidos condicionalmente*.

Definición 7.4 Modelos definidos condicionalmente. *Considérese un conjunto de variables $X = \{X_1, \ldots, X_n\}$. Un modelo probabilístico definido condicionalmente consiste en un conjunto de probabilidades marginales y condicionadas de la forma*

$$P = \{p(u_i|v_i);\ i = 1, \ldots, m\}, \tag{7.72}$$

que define una única función de probabilidad de X, donde U_i y V_i son subconjuntos disjuntos de X y $U_i \neq \phi$.

Por tanto, los modelos probabilísticos asociados a redes de Markov descomponibles y redes Bayesianas (analizados en el Capítulo 6) y cualquier otro

modelo probabilístico definido por una factorización son casos especiales de este tipo de modelos. En las secciones anteriores también se ha visto que las listas de relaciones de independencia y los grafos determinan ciertas factorizaciones de las funciones de probabilidad correspondientes mediante un producto de funciones de probabilidad condicionada. Por tanto, los modelos probabilísticos definidos condicionalmente son una generalización de los modelos multifactorizados, y pueden ser utilizados como el marco para analizar los problemas que subyacen a todos los modelos anteriores.

El problema principal de los modelos multifactorizados era la compatibilidad de las distintas factorizaciones que definen el modelo. En esta sección se analiza el problema de la compatibilidad y otros problemas relacionados asociados a los modelos definidos condicionalmente. En particular, se discutirán los siguientes problemas:

- **Problema 7.5. Unicidad:**
  ¿Define una única función de probabilidad un modelo definido condicionalmente? En otras palabras, ¿implica el conjunto de funciones de probabilidad condicionada suficientes restricciones para definir a lo más una función de probabilidad?

- **Problema 7.6. Consistencia o Compatibilidad:**
  ¿Es compatible con alguna función de probabilidad de las variables un modelo definido condicionalmente?

- **Problema 7.7. Parsimonia:**
  Si la respuesta al Problema 7.5 es afirmativa, ¿puede eliminarse alguna de las funciones de probabilidad condicionada del modelo sin pérdida de información?

- **Problema 7.8. Reducción:**
  Si la respuesta al Problema 7.6 es afirmativa, ¿puede reducirse al mínimo este conjunto (por ejemplo, eliminando alguna de las variables condicionantes)?

Los problemas anteriores sugieren que la construcción de modelos probabilísticos generales debe ser realizada cuidadosamente para eliminar inconsistencias, minimizar las redundancias del modelo e incrementar la posibilidad de obtener resultados precisos. Estos problemas se analizan en esta sección, que también discute las formas de definir modelos definidos condicionalmente. En particular se verá que:

1. Cualquier función de probabilidad con buenas propiedades (para asegurar la existencia de funciones de probabilidad condicionada) puede ser representada por un modelo definido condicionalmente.

2. Dado un conjunto de funciones de probabilidad condicionada que defina una función de probabilidad conjunta, siempre es posible obtener

un conjunto equivalente que de lugar a una factorización de la función de probabilidad en la forma dada por la regla de la cadena.

3. Los modelos definidos por redes Bayesianas son siempre consistentes.

Dado que cualquier conjunto de funciones de probabilidad condicionada puede ser escrito en forma canónica (ver Sección 5.5), se supondrá, sin pérdida de generalidad, que las funciones que integran un modelo definido condicionalmente están dadas en forma canónica. Cada una de estas funciones puede estar definida numéricamente o como una familia paramétrica. El problema es que puede no existir ninguna función de probabilidad que sea compatible con todas estas funciones condicionadas o, por el contrario, pueden existir una o varias funciones compatibles. Por tanto, ha de ser verificada la unicidad (Sección 7.7.1) y compatibilidad (Sección 7.7.2) del modelo resultante.

7.7.1 Comprobando la Unicidad

En primer lugar se analiza el problema de la unicidad (Problema 7.5), es decir, determinar si el conjunto de funciones condicionadas proporciona suficientes restricciones para definir a lo máximo un único modelo probabilístico. El teorema siguiente, que es una modificación del teorema introducido por Gelman y Speed (1993) da una condición suficiente para este problema.

Teorema 7.6 Unicidad. *Sea* $Y = \{Y_1, \ldots, Y_n\}$ *una permutación de* $X = \{X_1, \ldots, X_n\}$, $B_i = \{Y_1, \ldots, Y_{i-1}\}$ *y* $A_i = \{Y_{i+1}, \ldots, Y_n\}$. *Todo conjunto de funciones de probabilidad condicionada canónicas que contiene una sucesión de funciones de probabilidad de la forma*

$$p(y_i|s_i, a_i), \quad i = 1, \ldots, n, \tag{7.73}$$

o equivalentemente

$$p(y_i|b_i, u_i), \quad i = 1, \ldots, n, \tag{7.74}$$

define como mucho una función de probabilidad sobre X, *donde* $S_i \subseteq B_i$ *y* $U_i \subseteq A_i$. *Si todos los conjuntos* $S_i = \phi$ *o* $U_i = \phi, i = 1, \ldots, n$, *entonces la función de probabilidad conjunta existe y es única.*

Ejemplo 7.20 Unicidad pero posible inconsistencia. Considérese el conjunto de variables $X = \{X_1, X_2, X_3, X_4\}$ y el conjunto de funciones de probabilidad condicionada

$$\{p(x_4|x_1, x_2, x_3),\ p(x_3|x_1, x_2, x_4),\ p(x_2|x_1), p(x_1|x_4)\}, \tag{7.75}$$

que está asociado a la permutación $(Y_1, Y_2, Y_3, Y_4) = (X_4, X_3, X_2, X_1)$. En este caso, se tiene

$$\begin{array}{ll} S_1 = \phi, & A_1 = \{X_3, X_2, X_1\}, \\ S_2 = \{X_4\}, & A_2 = \{X_2, X_1\}, \\ S_3 = \phi, & A_3 = \{X_1\}, \\ S_4 = \{X_4\}, & A_4 = \phi. \end{array}$$

Por tanto, el conjunto de funciones dado en (7.75) cumple las condiciones del Teorema 7.6. Así, este conjunto es, o bien incompatible, o bien compatible con una única función de probabilidad.

Dado que $S_2 \neq \phi$ y $S_4 \neq \phi$, ha de comprobarse la consistencia de las funciones $p(x_3|x_1, x_2, x_4)$ y $p(x_1|x_4)$. En otras palabras, sólo algunas funciones particulares de la forma $p(x_3|x_1, x_2, x_4)$ y $p(x_1|x_4)$ definen una función de probabilidad consistente con las restantes funciones del modelo. ■

El Teorema 7.6 tiene importantes implicaciones prácticas en la definición de modelos probabilísticos pues:

1. Existe un conjunto mínimo de funciones de probabilidad condicionada necesarias para definir un único modelo probabilístico. Esto ocurre cuando el conjunto viene dado en forma canónica estándar. Por tanto, el número de aristas del grafo dirigido acíclico asociado puede reducirse al mínimo, pero la eliminación de alguna de las aristas en este modelo mínimo destruye la unicidad en la definición del modelo.

2. Para obtener una función de probabilidad por medio de un modelo definido condicionalmente se han de seguir las siguientes etapas:

   - **Etapa 1.** Ordenar el conjunto de variables.
   - **Etapa 2.** Definir un conjunto de funciones de probabilidad condicionada que contenga funciones de la forma (7.73) ó (7.74).
   - **Etapa 3.** Comprobar que las funciones dadas son consistentes. Este problema se analiza a continuación.

7.7.2 Comprobando la Consistencia

Dado que se está analizando el problema de definir un modelo probabilístico único y consistente por medio de un conjunto de funciones de probabilidad condicionada, se supondrá que este conjunto satisface las condicones de unicidad dadas en el teorema anterior. El paso siguiente es comprobar si este conjunto es compatible o si, por el contrario, existe alguna incompatibilidad entre las funciones que lo componen. El teorema siguiente (Arnold, Castillo, y Sarabia (1992, 1996)) proporciona un método iterativo para comprobar la consistencia de este conjunto de funciones.

Teorema 7.7 Consistencia. *Sea* $Y = \{Y_1, \ldots, Y_n\}$ *una permutación de un conjunto de variables* $X = \{X_1, \ldots, X_n\}$. *Dado un conjunto de funciones de probabildad condicionada canónicas de la forma*

$$\{p(y_1|s_1, a_1), \ldots, p(y_n|s_n, a_n)\}, \tag{7.76}$$

o

$$\{p(y_1|b_1, u_1), \ldots, p(y_n|b_n, u_n)\}, \tag{7.77}$$

donde $U_i \subseteq A_i = \{Y_{i+1}, \ldots, Y_n\}$ *y* $S_i \subseteq B_i = \{Y_1, \ldots, Y_{i-1}\}$ *para todo* $i = 1, \ldots, n$, *entonces una condición necesaria y suficiente para que el conjunto (7.76) sea compatible con una función de probabilidad de* X *es que, o bien* $S_i = \phi$, *o bien*

$$R_i = p(y_i|a_i) = \frac{p(y_i|s_i, a_i)/p(s_i|y_i, a_i)}{\sum\limits_{y_i} p(y_i|s_i, a_i)/p(s_i|y_i, a_i)} \tag{7.78}$$

es independiente de S_i, *para* $i = 1, \ldots, n$. *De forma equivalente, una condición necesaria y suficiente para que el conjunto (7.77) sea compatible con una función de probabilidad de* X *es que, o bien* $U_i = \phi$, *o bien*

$$T_i = p(y_i|b_i) = \frac{p(y_i|b_i, u_i)/p(u_i|y_i, b_i)}{\sum\limits_{y_i} [p(y_i|b_i, u_i)/p(u_i|y_i, b_i)]} \tag{7.79}$$

es independente de U_i, *para* $i = 1, \ldots, n$. *Obsérvese que la suma de los denominadores (7.78) y (7.79) ha de ser reemplazada por una integral, en el caso de que las variables sean continuas.*

Corolario 7.4 *El conjunto de funciones de probabilidad condicionada en (7.76) es compatible con una única función de probabilidad de* Y *si cada función* $p(y_i|s_i, a_i)$, $i = 1, \ldots, n$, *es de la forma*

$$p(y_i|s_i, a_i) = \frac{p(y_i|a_i) \sum\limits_{b_i \setminus s_i} \prod\limits_{j=1}^{i-1} p(y_j|a_j)}{\sum\limits_{b_i \cup \{y_i\} \setminus s_i} \prod\limits_{j=1}^{i} p(y_j|a_j)}, \tag{7.80}$$

donde $p(y_k|a_k)$, $k = 1, \ldots, i-1$, *está dado por (7.78), y* $p(y_i|a_i)$ *es una función de probabildad arbitraria.*

Un corolario similar al anterior podría escribirse en términos de (7.77). Obsérvese que una vez que se ha definido el conjunto $\{p(y_k|s_k, a_k);\ k = 1, \ldots, i-1\}$, la función $p(y_i|s_i, a_i)$ está determinada por $p(y_i|a_i)$, es decir, en esta etapa del proceso de construcción del modelo se es completamente libre para elegir la función $p(y_i|a_i)$ pero no $p(y_i|s_i, a_i)$. Obsérvese también que reemplazando la función $p(y_i|s_i, a_i)$ por la función $p(y_i|a_i)$ obtenida de

(7.78), el modelo resultante define la misma función de probabilidad. Esto implica que una vez que se ha definido una ordenación $(Y_1, \ldots, Y_n)$ para las variables, siempre es posible reemplazar una función de la forma $p(y_i|s_i, a_i)$ por otra de la forma $p(y_i|a_i)$ sin modificar el modelo probabilístico asociado. Se puede obtener $p(y_i|a_i)$ reescribiendo (7.80) en la forma

$$p(y_i|a_i) = \frac{p(y_i|s_i, a_i) \sum\limits_{b_i \cup \{y_i\} \setminus s_i} \prod\limits_{j=1}^{i} p(y_j|a_j)}{\sum\limits_{b_i \setminus s_i} \prod\limits_{j=1}^{i-1} p(y_j|a_j)}, \tag{7.81}$$

y dado que la función $p(y_1|a_1)$ ha de estar contenida en el conjunto de funciones de probabilidad condicionada, se puede calcular $p(y_i|a_i)$ a partir de cada $p(y_i|s_i, a_i)$ y las componentes canónicas estándar previas $p(y_j|a_j), j = 1, \ldots, i-1$. Por tanto, se concluye que (7.81) permite calcular el conjunto de funciones en forma canónica estándar equivalente a un conjunto de funciones de probabilidad condicionada dado.

Los Teoremas 7.6 y 7.7 implican los corolarios siguientes:

Corolario 7.5 *El conjunto de funciones de probabilidad condicionada*

$$\{p(x_i|b_i); i = 1, \ldots, n\} \text{ y } \{p(x_i|a_i); i = 1, \ldots, n\}, \tag{7.82}$$

donde $b_i = \{y_1, \ldots, y_{i-1}\}$ *y* $a_i = \{y_{i+1}, \ldots, y_n\}$*, es consistente y no puede reducirse sin destruir la unicidad del modelo.*

El Teorema 7.7 muestra que las factorizaciones asociadas a redes Bayesianas son consistentes, pues cumplen la primera condición del teorema, es decir, $S_i = \phi$, $i = 1, \ldots, n$.

El Teorema 7.7 sugiere el siguiente algoritmo iterativo para comprobar la consistencia de un conjunto de funciones de probabilidad condicionada dado. El algoritmo procede analizando una de las funciones en cada etapa y construyendo una forma canónica con $S_i = \phi$, $i = 1, \ldots, n$. El diagrama de flujo mostrado en la Figura 7.8 muestra una versión del algoritmo adaptada para las funciones de la forma (7.76). Sin embargo, una sencilla modificación de este algoritmo permite adaptarlo a la estructura de (7.77).

Algoritmo 7.2 Comprobando la compatibilidad.

- **Datos:** Un conjunto P de funciones de probabilidad condicionada en forma canónica que cumplen las condiciones de unidad.
- **Resultados:** Cierto o falso, dependiendo de si las funciones en P son consistentes.

1. En cualquier caso, ha de definirse la primera función $p(y_1|a_1)$. En caso contrario, el conjunto dado no cumple las condiciones de unidad. Por tanto, en la etapa inicial se comienza con $p(y_1|a_1)$.

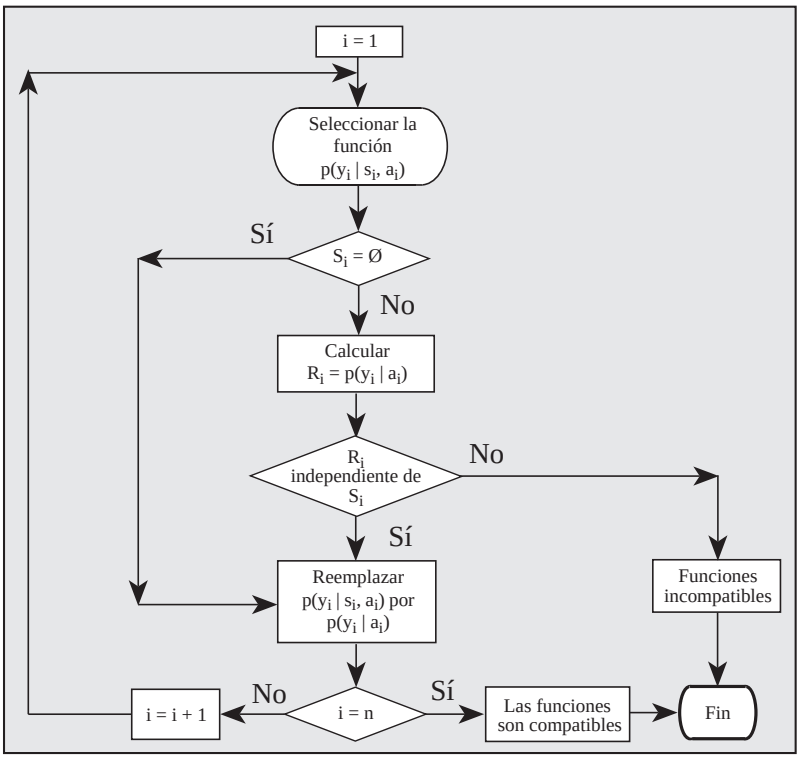


FIGURA 7.8. Diagrama de flujo para comprobar la consistencia de un conjunto de funciones de probabilidad condicionada que cumplen las condiciones de unidad.

2. En la etapa i-ésima se define $p(y_i|a_i)$ o $p(y_i|s_i, a_i)$, donde el conjunto $S_i \subset B_i$. Si $S_i = \phi$ ir a la Etapa 5; en caso contrario, calcular $p(s_i|y_i, a_i)$ marginalizando $p(b_i|y_i, a_i)$ sobre todas las variables de B_i distintas de aquellas en S_i, es decir, utilizando

$$p(s_i|y_i, a_i) = \sum_{b_i \setminus s_i} p(b_i|y_i, a_i). \tag{7.83}$$

3. Calcular la componente canónica estándar $R_i = p(y_i|a_i)$ basada en la información previa utilizando (7.78).

4. Si R_i es independiente de las variables del conjunto S_i, entonces ir a la Etapa 5; en caso contrario, la función $p(y_i|s_i, a_i)$ dada es incompatible con las funciones anteriores.

5. Calcular $p(b_i, y_i|a_i) = p(b_i|y_i, a_i)p(y_i|a_i)$.

6. Repetir las Etapas 2 a la 5 hasta que todas las funciones hayan sido analizadas. ∎

Por tanto, dado el conjunto de funciones de probabilidad condicionada P se puede determinar si este conjunto es consistente utilizando el Algoritmo 7.2 ó el Teorema 7.7. Esto proporciona una solución para el Problema 7.6.

Ejemplo 7.21 Conjunto consistente. Se supone que el siguiente conjunto de funciones de probabilidad condicionada es dado por un experto para definir un modelo probabilístico de las variables, $X = \{X_1, X_2, X_3, X_4\}$:

$$P_1 = \{p(x_4|x_3, x_2, x_1), p(x_3|x_2, x_1), p(x_2|x_1), p(x_1)\}. \tag{7.84}$$

Las funciones que componen el conjunto P_1 permiten factorizar la función de probabilidad utilizando la regla de la cadena de la forma

$$p(x_1, \ldots, x_n) = p(x_4|x_3, x_2, x_1)p(x_3|x_2, x_1)p(x_2|x_1)p(x_1).$$

Por tanto, este conjunto es consistente. La Tabla 7.3 muestra los parámetros que definen la función de probabilidad asociada, donde $\bar{x}_i$ y x_i denotan $X_i = 0$ y $X_i = 1$, respectivamente. Estos parámetros pueden definirse de forma arbitraria considerando valores en el intervalo $[0, 1]$.

Eligiendo la permutación $(Y_1, Y_2, Y_3, Y_4\} = \{X_4, X_3, X_2, X_1\}$, se puede verificar fácilmente que P_1 cumple las condiciones de unicidad. Por tanto, puede utilizarse el Algoritmo 7.2 para comprobar su compatiblidad:

- Para $i = 1$, el primer conjunto $p(x_4|x_3, x_2, x_1)$ es siempre compatible pues $A_1 = X_3, X_2, X_1$ y $S_1 = \phi$.

- Para $i = 2$, la segunda función $p(x_3|x_2, x_1)$ es compatible pues $A_2 = \{X_2, X_1\}$ y $S_2 = \phi$. En este caso se tiene $B_2 = \{Y_1\} = \{X_4\}$. Por tanto

$$p(b_2, y_2|a_2) = p(x_4, x_3|x_2, x_1) = p(x_4|x_3, x_2, x_1)p(x_3|x_2, x_1).$$

- Para $i = 3$, la función $p(x_2|x_1)$ es compatible pues $A_3 = \{X_1\}$ y $S_3 = \phi$. En este caso se tiene $B_3 = \{Y_1, Y_2\} = \{X_4, X_3\}$. Por tanto

$$p(b_3, y_3|a_3) = p(x_4, x_3, x_2|x_1) = p(x_4, x_3|x_2, x_1)p(x_2|x_1).$$

- Para $i = 4$, la siguiente función, $p(x_1)$, también es compatible pues $S_4 = A_4 = \phi$. En este caso se tiene $B_4 = \{Y_1, Y_2, Y_3\} = \{X_4, X_3, X_2\}$. Por tanto

$$p(b_4, y_4|a_4) = p(x_4, x_3, x_2, x_1) = p(x_4, x_3, x_2|x_1)p(x_1).$$

En consecuencia, el conjunto P_1 es consistente. ∎

Variable	Parámetros libres
X_1	$\theta_{10} = p(\bar{x}_1)$
X_2	$\theta_{200} = p(\bar{x}_2\|\bar{x}_1)$ $\theta_{201} = p(\bar{x}_2\|x_1)$
X_3	$\theta_{3000} = p(\bar{x}_3\|\bar{x}_1, \bar{x}_2)$ $\theta_{3001} = p(\bar{x}_3\|\bar{x}_1, x_2)$ $\theta_{3010} = p(\bar{x}_3\|x_1, \bar{x}_2)$ $\theta_{3011} = p(\bar{x}_3\|x_1, x_2)$
X_4	$\theta_{40000} = p(\bar{x}_4\|\bar{x}_1, \bar{x}_2, \bar{x}_3)$ $\theta_{40001} = p(\bar{x}_4\|\bar{x}_1, \bar{x}_2, x_3)$ $\theta_{40010} = p(\bar{x}_4\|\bar{x}_1, x_2, \bar{x}_3)$ $\theta_{40011} = p(\bar{x}_4\|\bar{x}_1, x_2, x_3)$ $\theta_{40100} = p(\bar{x}_4\|x_1, \bar{x}_2, \bar{x}_3)$ $\theta_{40101} = p(\bar{x}_4\|x_1, \bar{x}_2, x_3)$ $\theta_{40110} = p(\bar{x}_4\|x_1, x_2, \bar{x}_3)$ $\theta_{40111} = p(\bar{x}_4\|x_1, x_2, x_3)$

TABLA 7.3. Conjunto de parámetros correspondientes a las funciones de probabilidad condicionada del Ejemplo 7.21.

Ejemplo 7.22 Conjunto inconsistente. Considérese el siguiente conjunto de funciones de probabilidad condicionada definidas en el conjunto de cuatro variables $X = \{X_1, X_2, X_3, X_4\}$:

$$P_2 = \{p(x_4|x_3, x_2, x_1), p(x_3|x_4, x_2, x_1), p(x_2|x_1), p(x_1)\}.$$

De la misma forma que en el ejemplo anterior, se aplica el Algoritmo 7.2 para comprobar la consistencia de este conjunto:

- La primera función, $p(x_4|x_3, x_2, x_1)$, puede darse sin ninguna restricción.

- En cambio, ha de comprobarse la consistencia de la segunda condición, $p(x_3|x_4, x_2, x_1)$, pues $S_2 = \{X_4\} \neq \phi$. La ecuación (7.80) determina la estructura de esta función para ser consistente con la primera función dada. Se tiene

$$p(x_3|x_4, x_2, x_1) = \frac{p(x_3|x_2, x_1)}{\sum\limits_{x_3} p(x_3|x_2, x_1)p(x_4|x_3, x_2, x_1)}, \tag{7.85}$$

Utilizando los parámetros dados en la Tabla 7.3, (7.85) resulta

$$
\begin{aligned}
p(\bar{x_3}|\bar{x_4},\bar{x_2},\bar{x_1}) &= \frac{\theta_{3000}}{\theta_{3000}\theta_{40000}+\theta_{40001}-\theta_{3000}\theta_{40001}},\\
p(\bar{x_3}|x_4,\bar{x_2},x_1) &= \frac{\theta_{3010}}{1-\theta_{3010}\theta_{40100}-\theta_{40101}+\theta_{3010}\theta_{40101}},\\
p(\bar{x_3}|\bar{x_4},x_2,\bar{x_1}) &= \frac{\theta_{3001}}{\theta_{3001}\theta_{40010}+\theta_{40011}-\theta_{3001}\theta_{40011}},\\
p(\bar{x_3}|x_4,x_2,x_1) &= \frac{\theta_{3011}}{1-\theta_{3011}\theta_{40110}-\theta_{40111}+\theta_{3011}\theta_{40111}},\\
p(\bar{x_3}|x_4,\bar{x_2},\bar{x_1}) &= \frac{\theta_{3000}}{1-\theta_{3000}\theta_{40000}-\theta_{40001}+\theta_{3000}\theta_{40001}},\\
p(\bar{x_3}|\bar{x_4},\bar{x_2},x_1) &= \frac{\theta_{3010}}{\theta_{3010}\theta_{40100}+\theta_{40101}-\theta_{3010}\theta_{40101}},\\
p(\bar{x_3}|x_4,x_2,\bar{x_1}) &= \frac{\theta_{3001}}{1-\theta_{3001}\theta_{40010}-\theta_{40011}+\theta_{3001}\theta_{40011}},\\
p(\bar{x_3}|\bar{x_4},x_2,x_1) &= \frac{\theta_{3011}}{\theta_{3011}\theta_{40110}+\theta_{40111}-\theta_{3011}\theta_{40111}}.
\end{aligned}
\tag{7.86}
$$

- La tercera y la cuarta de las funciones de probabilidad, $p(x_2|x_1)$ y $p(x_1)$, son consistentes, pues $S_3 = S_4 = \phi$.

Por tanto, cualquier combinación de valores de los parámetros de las funciones en P_2 que no cumpla las condiciones en (7.86) define un modelo no consistente. ∎

Ejemplo 7.23 Conjunto inconsistente. Considérese el siguiente conjunto de funciones de probabilidad condicionada sobre el conjunto X:

$$P_3 = \{p(x_4|x_2,x_1,x_3,x_5), p(x_2|x_1,x_3,x_5), p(x_1|x_4,x_3,x_5), p(x_3|x_2,x_5), p(x_5)\}.$$

Eligiendo la permutación $\{Y_1,Y_2,Y_3,Y_4,Y_5\} = \{X_4,X_2,X_1,X_3,X_5\}$, el conjunto P_3 cumple las condiciones de unicidad. Aplicando el Algoritmo 7.2 a este conjunto se tiene:

- Para $i = 1$, la primera función, $p(x_4|x_2,x_1,x_3,x_5)$, es siempre compatible pues $A_1 = \{X_2,X_1,X_3,X_5\}$ y $S_1 = \phi$.
- Para $i = 2$, la función, $p(x_2|x_1,x_3,x_5)$, también es compatible, dado que $A_2 = \{X_1,X_3,X_5\}$ y $S_2 = \phi$. Se tiene $B_2 = \{Y_1\} = \{X_4\}$; por tanto

  $$p(b_2,y_2|a_2) = p(x_4,x_2|x_1,x_3,x_5) = p(x_4|x_2,x_1,x_3,x_5)p(x_2|x_1,x_3,x_5).$$

- Para $i = 3$, ha de comprobarse la consistencia de la tercera función, $p(x_1|x_4, x_3, x_5)$, dado que $A_3 = \{X_3, X_5\}$ y $S_3 = \{X_4\} \neq \phi$. En este caso se tiene $B_3 = \{Y_1, Y_2\} = \{X_4, X_2\}$, y se necesita calcular $p(s_3|y_3, a_3) = p(x_4|x_1, x_3, x_5)$ utilizando (7.83). Se obtiene

$$p(x_4|x_1, x_3, x_5) = \sum_{x_2} p(x_4, x_2|x_1, x_3, x_5).$$

  También se necesita calcular R_3 utilizando (7.78):

$$R_3 = p(y_3|a_3) = p(x_1|x_3, x_5) = \frac{p(x_1|x_4, x_3, x_5)/p(x_4|x_1, x_3, x_5)}{\sum_{x_1}(p(x_1|x_4, x_3, x_5)/p(x_4|x_1, x_3, x_5))}.$$

  Entonces, si R_3 no depende de X_4, la función $p(x_1|x_4, x_3, x_5)$ es compatible con las funciones de probabilidad condicionada anteriores. En caso contrario, el conjunto es incompatible. Para ilustrar el resto del proceso, supóngase que este conjunto es compatible. En ese caso, la función $p(x_1|x_4, x_3, x_5)$ se reemplaza por R_3 y se calcula

$$p(b_3, y_3|a_3) = p(x_4, x_2, x_1|x_3, x_5) = p(x_4, x_2|x_1, x_3, x_5)p(x_1|x_3, x_5).$$

- Para $i = 4$, ha de comprobarse la consistencia del cuarto conjunto, $p(x_3|x_2, x_5)$, dado que se tiene $A_4 = \{X_5\}$ y $S_4 = \{X_2\} \neq \phi$. En este caso, resulta $B_4 = \{Y_1, Y_2, Y_3\} = \{X_4, X_2, X_1\}$, y se necesita calcular $p(s_4|y_4, a_4) = p(x_2|x_3, x_5)$ utilizando (7.83). Se obtiene

$$p(x_2|x_3, x_5) = \sum_{x_1, x_4} p(x_4, x_2, x_1|x_3, x_5).$$

  Igualmente, utilizando (7.78) el valor siguiente de R_4 es

$$R_4 = p(y_4|a_4) = p(x_3|x_5) = \frac{p(x_3|x_2, x_5)/p(x_2|x_3, x_5)}{\sum_{x_3}(p(x_3|x_2, x_5)/p(x_2|x_3, x_5)}.$$

  Si R_4 no depende de la variable X_2, la función $p(x_3|x_2, x_5)$ es compatible con las funciones anteriores. En caso contrario, el conjunto de funciones es incompatible. Al igual que en el caso anterior, supóngase que esta función es compatible. Entonces, la función $p(x_3|x_2, x_5)$ se reemplaza por R_4. También se calcula

$$p(b_4, y_4|a_4) = p(x_4, x_2, x_1, x_3|x_5) = p(x_4, x_2, x_1|x_3, x_5)p(x_3|x_5).$$

- Finalmente, para $i = 5$, la última función, $p(x_5)$, es compatible pues $A_5 = \phi$ y $S_5 = \phi$.

Por tanto, si R_3 depende de X_4 o R_4 depende de X_2, entonces P_3 es inconsistente; en caso contrario el modelo es consistente. ∎

En los ejemplos anteriores se pueden constatar las siguientes aplicaciones prácticas del Teorema 7.7:

1. Cualquier función de probabilidad con buenas propiedades (para asegurar la existencia de funciones de probabilidad condicionada) puede ser representada por un modelo definido condicionalmente.

2. Si se incrementan las componentes canónicas estándar con información adicional, entonces será necesario comprobar la compatibilidad del conjunto resultante.

3. Si el conjunto de funciones de probabilidad condicionada está dado en forma estándar, entonces estas funciones pueden definirse de forma arbitraria, es decir, no están restringidas por ninguna condición distinta de los axiomas propios de la probabilidad.

4. Cualquier función de probabilidad condicionada de la forma $p(x_i|s_i, a_i)$, con $S_i \neq \phi$, puede ser reemplazada por una función en forma canónica estándar $p(x_i|a_i)$ sin modificar la función de probabilidad de las variables. La función en forma canónica estándar puede obtenerse utilizando (7.78).

5. Los Teoremas 7.6 y 7.7 y los Algoritmos 7.1 y 7.3 (definidos en la sección siguiente) proporcionan una forma de construir un I-mapa del modelo definido condicionalmente.

Los Teoremas 7.6 y 7.7 proporcionan una solución para los Problemas 7.5 y 7.6. Por tanto, cuando se define un modelo probabilístico mediante un conjunto de funciones de probabilidad condicionada es preferible definir solamente el conjunto mínimo de funciones necesarias para la unicidad del modelo. Cualquier otra información será redundante y podrá dar lugar a inconsistencias en el modelo. Por tanto, se tiene una solución al Problema 7.7. Además, dado un conjunto consistente de funciones de probabilidad condicionada P que define un único modelo probabilístico, se puede reemplazar este conjunto por otro equivalente P' dado en forma canónica estándar. Por tanto, se tiene una solución al Problema 7.8.

7.7.3 Definición de un Modelo Definido Condicionalmente

El Teorema 7.7 supone que el conjunto de funciones de probabilidad condicionada dado cumple las hipótesis de unicidad. Por ello, la unicidad del conjunto habrá de comprobarse antes que la compatibilidad. De esta forma, inicialmente se convierte el conjunto a forma canónica, después se comprueba su unicidad utilizando el Teorema 7.6 y, finalmente, se comprueba su consistencia utilizando el Teorema 7.7.

Cuando $S_i = \phi$ o $U_i = \phi$ para todo i, la consistencia está garantizada y se dice que la forma canónica es un forma canónica estándar y los términos

$p(y_i|a_i)$, o $p(y_i|b_i)$, se denominan componentes canónicas estándar. En caso contrario, ha de comprobarse la compatibilidad del conjunto.

El algoritmo siguiente permite determinar si un conjunto P cumple las condiciones de unicidad y permite obtener los subconjuntos cuya compatibilidad ha de comprobarse.

Algoritmo 7.3 Definición de un modelo definido condicionalmente.

- **Datos:** Un conjunto X de n variables y un conjunto canónico de funciones de probabilidad condicionada $P = \{p(x_i|s_i);\ i = 1, \ldots, m\}$.
- **Resultados:** La lista Q_1 de subconjuntos de P que cumplen la condición de unicidad y el conjunto C_1 de subconjuntos de P cuya compatibilidad ha de comprobarse.

El algoritmo consiste en un procedimiento, *Compatible*, que inicia el conjunto $Soluciones = \phi$, ejecuta el procedimiento recursivo *CompatibleAux* (con X, P, y las listas vacías C_2 y Q_2 como argumentos), e imprime las *Soluciones*:

1. Definir $i \leftarrow 1$, y $m \leftarrow$ número de funciones en P.
2. Definir $P_1 \leftarrow P$, $V \leftarrow X$, $C_1 \leftarrow C_2$ y $Q_1 \leftarrow Q_2$.
3. Si $p(x_i|s_i) \in P_1$ es tal que $V \cup S_i \supset V$, hacer lo siguiente:
   - Eliminar la función $p(x_i|s_i)$ de P_1 y añadirla a Q_1.
   - Si $V \cup S_i \neq V$ añadir $p(x_i|s_i)$ a C_1.
   - Eliminar de P_1 cualquier función $p(x_r|s_r) \in P_1$ tal que $X_r = X_i$ y añadirla a C_1.
   - Si $P \neq \phi$ ejecutar el Algoritmo 7.3 de forma recursiva, con argumentos $V \setminus X_i$ y P_1, y añadir a C la lista C_1 resultante y a Q la lista Q_1 resultante; en caso contrario, añadir su resultado a Soluciones.
   - Ir a la Etapa 4.

   En caso contrario, ir a la Etapa 4.
4. Si $i < m$, considerar $i = i + 1$ y repetir la Etapa 3; en caso contrario, devolver C_1 y Q_1. ∎

El Algoritmo 7.3 considera un conjunto de funciones de la forma (7.73). Sin embargo, puede ser fácilmente modificado para tratar funciones de la forma (7.74).

La Figura 7.9 muestra un programa recursivo de *Mathematica* que implementa el Algoritmo 7.3. La función $Compatible[X, P]$ tiene dos argumentos, X y P, donde X es la lista de variables y P es la lista de funciones

de probabilidad condicionada. Por ejemplo, para ejecutar el programa con $X = \{A, B, C\}$ y

$$P = \{p(a), p(b|a), p(c|a, b)\} \tag{7.87}$$

se necesita ejecutar las siguientes sentencias de *Mathematica*:

```
X={A,B,C};
P=List[{{A},{}},{{B},{A}},{{C},{A,B}}];
Compatible[X,P];
```

La primera de las sentencias define la lista de variables, la segunda define la lista P, y la tercera llama a la función $Compatible[X, P]$, con los argumentos X y P.

La función $Compatible[X, P]$ genera dos listas como resultado, Q y C. La lista Q contiene todos los conjuntos posibles de funciones de P que definen una única función de probabilidad conjunta. El número de subconjuntos de C es igual al de Q. Para cada conjunto en Q, el conjunto correspondiente en C es el formado por las funciones cuya consistencia ha de ser comprobada para que el modelo sea consistente.

Cuando el conjunto P defina una única función de probabilidad, el conjunto Q contendrá un único subconjunto, que coincide con P, y el conjunto C será vacío. Cuando el conjunto P sea consistente pero no defina un único modelo probabilístico, tanto Q como C serán conjuntos vacíos. Este programa se ilustra con los siguientes ejemplos.

Ejemplo 7.24 Conjunto verificando unicidad y compatibilidad. Considérese el conjunto de variables $\{A, B, C\}$, y el conjunto de funciones de probabilidad condicionada (7.87). Es sencillo ver que el conjunto P cumple las hipótesis del Corolario 7.5 y, por tanto, define una única función de probabilidad que está dada por

$$p(a, b, c) = p(a)p(b|a)p(c|a, b). \tag{7.88}$$

Este resultado puede obtenerse utilizando el programa de la Figura 7.9. Para ejecutar el programa primero han de definirse los conjuntos X y P y, a continuación, ejecutar la función $Compatible[X, P]$. Esta función produce como resultado las listas: $Q = P$ y $C = \phi$. ∎

Ejemplo 7.25 Unicidad pero posible inconsistencia. Considérese el conjunto de variables $\{A, B, C\}$, y el de funciones de probabilidad condicionada

$$P = \{p(a|b, c), p(b|c), p(b|c, a), p(c|a, b)\}. \tag{7.89}$$

El programa de *Mathematica* mostrado en la Figura 7.9 permite obtener los subconjuntos que definen una única función de probabilidad conjunta de las tres variables. Para ello, es necesario ejecutar las siguientes sentencias:

```
X={A,B,C};
```

```
Remov[CM_,j_]:=Join[Take[CM,j-1],
   Take[CM,{j+1,Length[CM]}]]

Compatible[X_,P_]:=Module[{},
   Soluciones={};CompatibleAux[X,P,{},{}];
   Soluciones
]
CompatibleAux[X_,P_,C2_,Q2_]:=
Module[{Xi,V,i,Q1,C1},
Do[
   P1=P;V=X;C1=C2;Q1=Q2;
   Uni=Union[P1[[i,1]],P1[[i,2]]];
   If[Uni==Union[V,Uni],AppendTo[Q1,P1[[i]]];
      Xi=P1[[i,1]];If[Uni !=V, AppendTo[C1,P1[[i]]]];
      P1=Remov[P1,i];
      Do[
         If[Xi==P1[[k,1]],
           AppendTo[C1,P1[[k]]];P1=Remov[P1,k],
           True
         ],
      {k,Length[P1],1,-1}];
      If[P1!={},
         Res=CompatibleAux[Complement[V,Xi],P1,C1,Q1];
         C1=Union[C1,Res[[1]]];Q1=Union[Q1,Res[[2]]],
         AppendTo[Soluciones,{Q1,C1}]
      ]
      ],
{i,1,Length[P]}];
Return[{C1,Q1}]]
```

FIGURA 7.9. Programa de *Mathematica* para comprobar si un conjunto P cumple las condiciones de unicidad y compatibilidad.

```
P=List[{{A},{B,C}},{{B},{C}},{{B},{C,A}},{{C},{A,B}}];
Compatible[X,P];
```

La Tabla 7.4 muestra el resultado del programa. En este caso se han obtenido nueve listas distintas en Q (conjuntos de funciones que definen una única función de probabilidad) y C (conjuntos de funciones cuya consistencia ha de ser comprobada). Por ejemplo, ha de comprobarse la consistencia de las funciones $p(b|c,a)$ y $p(c|a,b)$ del primer conjunto de la lista C con las correspondientes funciones del primer conjunto de Q. Si estas funciones son consistentes, entonces el conjunto $\{p(a|b,c),(b|c),(c|a,b)\}$ define

Conjunto	Lista Q	Lista C
1	$p(a\|b,c)$, $p(b\|c)$, $p(c\|a,b)$	$p(b\|c,a)$, $p(c\|a,b)$
2	$p(a\|b,c)$, $p(b\|c,a)$, $p(c\|a,b)$	$p(b\|c,a)$, $p(b\|c)$, $p(c\|a,b)$
3	$p(a\|b,c)$, $p(c\|a,b)$, $p(b\|c)$	$p(c\|a,b)$, $p(b\|c)$, $p(b\|c,a)$
4	$p(a\|b,c)$, $p(c\|a,b)$, $p(b\|c,a)$	$p(c\|a,b)$, $p(b\|c,a)$, $p(b\|c)$
5	$p(b\|c,a)$, $p(a\|b,c)$, $p(c\|a,b)$	$p(b\|c)$, $p(a\|b,c)$, $p(c\|a,b)$
6	$p(b\|c,a)$, $p(c\|a,b)$, $p(a\|b,c)$	$p(b\|c)$, $p(c\|a,b)$, $p(a\|b,c)$
7	$p(c\|a,b)$, $p(a\|b,c)$, $p(b\|c)$	$p(a\|b,c)$, $p(b\|c)$, $p(b\|c,a)$
8	$p(c\|a,b)$, $p(a\|b,c)$, $p(b\|c,a)$	$p(a\|b,c)$, $p(b\|c,a)$, $p(b\|c)$
9	$p(c\|a,b)$, $p(b\|c,a)$, $p(a\|b,c)$	$p(b\|c,a)$, $p(b\|c)$, $p(a\|b,c)$

TABLA 7.4. Resultados del programa de *Mathematica* de la Figura 7.9 aplicado al conjunto P en (7.89). La segunda columna muestra los conjuntos de funciones posibles que cumplen las hipótesis de unicidad. La tercera muestra los conjuntos cuya compatibilidad ha de comprobarse.

un único modelo probabilístico que puede expresarse como

$$p(a,b,c) = p(a|b,c)p(b|c)p(c),$$

donde $p(c)$ ha de ser calculado a partir de (7.81). ∎

Ejercicios

7.1 Utilizar el Algoritmo 6.6 para demostrar que las aristas L_{13} y L_{35} en la Figura 7.3(a) son reversibles.

7.2 Hallar la condición para que se cumpla (7.5).

7.3 Verificar que al utilizar (7.37) como factorización de referencia, el Algoritmo 7.1 obtiene las soluciones dadas en (7.42).

7.4 Utilizar el criterio de D-separación de la Definición 5.4 para verificar:

(a) Que cada relación de independencia en (7.63) es implicada por el grafo dirigido acíclico de la Figura 7.7(a).

(b) Que cada relación de independencia en (7.64) es implicada por el grafo dirigido acíclico de la Figura 7.7(b).

7.5 (a) Comprobar que la factorización de la función de probabilidad en (7.43) implica la lista de relaciones de independencia dada en (7.63).

(b) Comprobar que la factorización de la función de probabilidad en (7.44) implica la lista de relaciones de independencia dada en (7.64).

7.6 Considérese el conjunto de variables $\{X_1, X_2, X_3, X_4\}$ distribuidas de forma normal cuya función de probabilidad satisface las relaciones de independencia

$$I(X_2, X_3|X_1) \; y \; I(X_1, X_4|X_2, X_3).$$

Comprobar que la estructura de la matriz de covarianzas asociada está dada por

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \alpha \\ \sigma_{12} & \sigma_{22} & \frac{\sigma_{12}\sigma_{13}}{\sigma_{11}} & \sigma_{24} \\ \sigma_{13} & \frac{\sigma_{12}\sigma_{13}}{\sigma_{11}} & \sigma_{33} & \sigma_{34} \\ \alpha & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{pmatrix},$$

donde

$$\alpha = \frac{\sigma_{11}(\sigma_{12}\sigma_{24}(\sigma_{11}\sigma_{33} - \sigma_{13}^2) + \sigma_{13}\sigma_{34}(\sigma_{11}\sigma_{22} - \sigma_{12}^2))}{\sigma_{11}^2\sigma_{22}\sigma_{33} - \sigma_{12}^2\sigma_{13}^2}.$$

7.7 Suponer que las variables $\{X_1, \ldots, X_7\}$ están distribuidas de forma normal y su función de probabilidad satisface las independencias

$$I(X_3, \{X_1, X_2\}|\phi), \; I(X_1, X_4|\{X_2, X_3\}), \; y \; I(X_7, \{X_1, X_2, X_3\}|X_4).$$

Comprobar que la estructura de la matriz de covarianzas asociada está dada por

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & 0 & \frac{\sigma_{12}\,\sigma_{24}}{\sigma_{22}} & \sigma_{15} & \sigma_{16} & \frac{\sigma_{12}\,\sigma_{24}\,\sigma_{47}}{\sigma_{22}\,\sigma_{44}} \\ \sigma_{12} & \sigma_{22} & 0 & \sigma_{24} & \sigma_{25} & \sigma_{26} & \frac{\sigma_{24}\,\sigma_{47}}{\sigma_{44}} \\ 0 & 0 & \sigma_{33} & \sigma_{34} & \sigma_{35} & \sigma_{36} & \frac{\sigma_{34}\,\sigma_{47}}{\sigma_{44}} \\ \frac{\sigma_{12}\,\sigma_{24}}{\sigma_{22}} & \sigma_{24} & \sigma_{34} & \sigma_{44} & \sigma_{45} & \sigma_{46} & \sigma_{47} \\ \sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_{55} & \sigma_{56} & \sigma_{57} \\ \sigma_{16} & \sigma_{26} & \sigma_{36} & \sigma_{46} & \sigma_{56} & \sigma_{66} & \sigma_{67} \\ \frac{\sigma_{12}\,\sigma_{24}\,\sigma_{47}}{\sigma_{22}\,\sigma_{44}} & \frac{\sigma_{24}\,\sigma_{47}}{\sigma_{44}} & \frac{\sigma_{34}\,\sigma_{47}}{\sigma_{44}} & \sigma_{47} & \sigma_{57} & \sigma_{67} & \sigma_{77} \end{pmatrix}.$$

7.8 Considérese el conjunto de variables $X = \{X_1, X_2, X_3, X_4, X_5\}$ y las relaciones de independencia siguientes:

- $p(x_1, x_2, x_3|x_4, x_5), p(x_4, x_5|x_1, x_2, x_3)$ y
- $p(x_1, x_3|x_2), p(x_2, x_4, x_5|x_1)$.

Utilizando los métodos descritos en este capítulo

(a) Reescribir el conjunto anterior en forma canónica.

(b) ¿Satisface cada conjunto las condiciones de unicidad?

(c) ¿Satisface cada conjunto las condiciones de compatibilidad?

(d) Encontrar un conjunto sencillo de funciones de probabilidad condicionada en forma canónica que cumpla las condiciones de unicidad y compatibilidad y construir la red Bayesiana asociada.

7.9 Dado el conjunto $X = \{X_1, X_2, X_3, X_4\}$, encontrar las condiciones para que el siguiente conjunto de funciones de probabilidad condicionada sea compatible:

$$\begin{array}{ll} p(x_1|x_2), & p(x_2|x_3), \\ p(x_3|x_2, x_1) = p(x_3|x_2), & p(x_4|x_3, x_2, x_1) = p(x_4|x_3). \end{array}$$

7.10 ¿Cuales son las condiciones para que las variables aleatorias discretas X_1 y X_2 sean compatibles?

7.11 Encontrar la familia más general de variables aleatorias normales $X = \{X_1, X_2, X_3, X_4\}$ que cumpla las siguientes relaciones de independencia:

$$\{I(x_1, x_2|\{x_3, x_4\}), I(x_3, x_1|x_2), I(x_4, x_2|x_3)\}.$$

7.12 Encontrar la familia más general de modelos definidos por el multigrafo de la Figura 7.10.

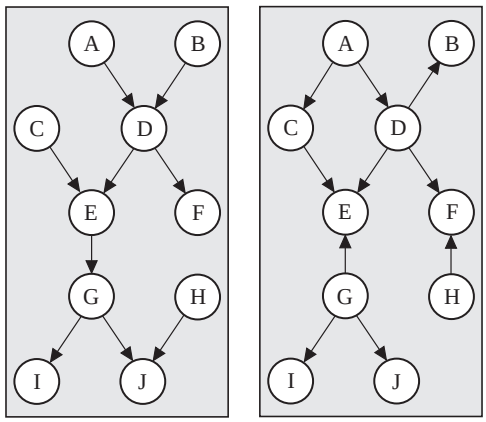


FIGURA 7.10. Ejemplo de multigrafo.

7.13 Dada la función de probabilidad $p(a, b, c, d) = p(a)p(b|a)p(c|a)p(d|b)$, encontrar

(a) La función compatible más general de la forma $p(c|a,b)$.

(b) La función compatible más general de la forma $p(c|a,b)$ cuando se reemplaza $p(d|b)$ por $p(d|a,b,c)$ en la factorización anterior.

7.14 Supóngase el conjunto de funciones de probabilidad condicionada dado en forma canónica (7.74). Modificar el Algoritmo 7.3 para tratar este caso.

7.15 Utilizar el programa de la Figura 7.9 para comprobar si el conjunto de funciones de probabilidad condicionada (7.75) cumple las hipótesis de unicidad.

7.16 Utilizando el Algoritmo 7.2, encontrar las restricciones para que el conjunto (7.75) cumpla la consistencia y, por tanto, defina un único modelo probabilístico.

7.17 Escribir un corolario similar al Corolario 7.4 para el conjunto de funciones de probabilidad (7.77).

7.18 Dibujar un diagrama de flujo similar al mostrado en la Figura 7.8 para el caso en el que el conjunto de funciones tenga la estructura dada en (7.77). Escribir el algoritmo correspondiente, similar al Algoritmo 7.2.

7.19 Dos expertos dan de forma independiente los siguientes conjuntos de funciones de probabilidad condicionada para definir un modelo probabilístico para un conjunto de tres variables binarias $X = \{X_1, X_2, X_3\}$:

$$\{p(x_1),\ \ p(x_2|x_1,x_3),\ \ p(x_3|x_1,x_2)\}.$$

El primer experto da los valores mostrados en la Tabla 7.5, mientras que el segundo da los valores correspondientes de la Tabla 7.6. Determinar cuál de los dos conjuntos define un único modelo probabilístico.

x_1	$p(x_1)$
0	0.3
1	0.7

x_1	x_3	x_2	$p(x_2\|x_1,x_3)$
0	0	0	0.40
0	0	1	0.60
0	1	0	0.40
0	1	1	0.60
1	0	0	0.20
1	0	1	0.80
1	1	0	0.20
1	1	1	0.80

x_1	x_2	x_3	$p(x_3\|x_1,x_2)$
0	0	0	0.90
0	0	1	0.10
0	1	0	0.70
0	1	1	0.30
1	0	0	0.50
1	0	1	0.50
1	1	0	0.60
1	1	1	0.40

TABLA 7.5. Valores numéricos asociados al primer conjunto de funciones condicionadas.

x_1	$p(x_1)$
0	0.3
1	0.7

x_1	x_3	x_2	$p(x_2\|x_1,x_3)$
0	0	0	0.30
0	0	1	0.70
0	1	0	0.40
0	1	1	0.60
1	0	0	0.10
1	0	1	0.90
1	1	0	0.50
1	1	1	0.50

x_1	x_2	x_3	$p(x_3\|x_1,x_2)$
0	0	0	0.90
0	0	1	0.10
0	1	0	0.70
0	1	1	0.30
1	0	0	0.50
1	0	1	0.50
1	1	0	0.60
1	1	1	0.40

TABLA 7.6. Valores numéricos asociados al segundo conjunto de funciones condicionadas.

Capítulo 8
Propagación Exacta en Redes Probabilísticas

8.1 Introducción

En los capítulos anteriores se han analizado diversas formas de definir una base de conocimiento coherente para un sistema experto probabilístico. Ésta está formada por la función de probabilidad conjunta de las variables que componen el modelo. Una vez definida, una de las tareas más importantes de un sistema experto consiste en obtener conclusiones a medida que se va conociendo nueva información, o *evidencia*. Por ejemplo, en el área médica, la principal tarea de los sistemas expertos consiste en obtener un diagnóstico para un determinado paciente que presenta ciertos síntomas (evidencia). El mecanismo para obtener conclusiones a partir de la evidencia se conoce como *propagación de evidencia*[1] o, simplemente, *propagación*. Esta tarea consiste en actualizar las probabilidades de las variables en función de la evidencia. En el caso del diagnóstico médico, se trata de conocer las probabilidades de cada una de las enfermedades, dados los síntomas observados en el paciente.

Existen tres tipos distintos de algoritmos de propagación: exactos, aproximados y simbólicos. Un algoritmo de propagación se denomina exacto si calcula las probabilidades de los nodos sin otro error que el resultante del redondeo producido por las limitaciones de cálculo del ordenador. En este

[1] Algunos autores se refieren a la propagación de evidencia utilizando otra terminología como *propagación de incertidumbre*, *inferencia probabilística*, etc.

capítulo se analizan detalladamente algunos de los métodos de propagación exacta más importantes.

Los *algoritmos de propagación aproximada* utilizan distintas técnicas de simulación para obtener valores aproximados de las probabilidades. Estos algoritmos se utilizan en aquellos casos en los que los algoritmos exactos no son aplicables, o son computacionalmente costosos y se analizan en detalle en el Capítulo 9. Finalmente, un *algoritmo de propagación simbólica* puede operar no sólo con parámetros numéricos, sino también con parámetros simbólicos, obteniendo las probabilidades en forma simbólica, es decir, en función de los parámetros. El Capítulo 10 analiza este problema e introduce algunos métodos de propagación simbólica.

Algunos de los métodos exactos que se describen en este capítulo son aplicables tanto a redes de Markov como a redes Bayesianas. Sin embargo, otros métodos son solamente válidos para redes Bayesianas, ya que aprovechan la representación de la función de probabilidad propia de estos modelos. De la misma forma, algunos de estos métodos de propagación son propios de modelos discretos, mientras que otros pueden ser aplicados a modelos discretos y continuos. La Sección 8.2 introduce el problema de la propagación de evidencia y analiza algunos de sus aspectos computacionales asociados. En la Sección 8.3 se presenta un algoritmo eficiente de propagación para un tipo simple de modelos probabilísticos: las redes Bayesianas con estructura de poliárbol. Este algoritmo ilustra las ideas básicas que se aplican posteriormente en otros algoritmos de propagación. Las Secciones 8.4−8.7 presentan otros métodos de propagación para redes más complejas (redes múltiplemente conexas). La Sección 8.5 presenta el método de propagación por *condicionamiento*, en la Sección 8.6 se analiza el método de propagación por *agrupamiento*, y en la Sección 8.7 se presenta el método de propagación en *árboles de conglomerados*. En la Sección 8.8 se analiza el problema de la propagación orientada a un objetivo. Finalmente, la Sección 8.9 analiza el caso de modelos continuos, mostrando un algoritmo para propagar evidencia en redes Bayesianas Gausianas.

8.2 Propagación de Evidencia

La propagación de evidencia es una de las tareas más importantes de un sistema experto, pues permite obtener conclusiones cuando se dispone de nueva información (síntomas, etc.). Supóngase un conjunto de variables discretas $X = \{X_1, \ldots, X_n\}$ y una función de probabilidad $p(x)$, en X. Cuando no se dispone de ninguna información, es decir, cuando no existe evidencia, el proceso de propagación consiste en calcular las probabilidades marginales $p(X_i = x_i)$, también denotadas por $p(x_i)$, para cada $X_i \in X$. Estas probabilidades proporcionan información "a priori" sobre los distintos valores que pueden tomar las variables.

Cuando se dispone de cierta evidencia, es decir, cuando se conoce un conjunto de variables $E \subset X$ que tienen asociadas los valores $X_i = e_i$, para $X_i \in E$, el proceso de propagación debe tener en cuenta estos valores para calcular las nuevas probabilidades de los nodos.

Definición 8.1 Evidencia. *Un subconjunto de variables $E \subset X$ cuyos valores son conocidos, $E = e$, en una situación dada, se conoce como conjunto de evidencia, o simplemente evidencia.*

En esta situación, la propagación de evidencia consiste en calcular las funciones de probabilidad condicionada $p(x_i|e)$ para cada variable $X_i \notin E$, dada la evidencia $E = e$. Estas funciones de probabilidad condicionada miden el efecto producido por la evidencia en cada variable. Cuando no se dispone de evidencia ($E = \phi$), las funciones condicionadas $p(x_i|e)$ son simplemente las funciones de probabilidad marginal $p(x_i)$.

Un forma de calcular las probabilidades $p(x_i|e)$ consiste en utilizar la fórmula (3.5), que implica

$$p(x_i|e) = \frac{p(x_i, e)}{p(e)} \propto p(x_i, e), \tag{8.1}$$

donde $1/p(e)$ es una constante de proporcionalidad. Por tanto, se puede obtener $p(x_i|e)$, calculando y normalizando las probabilidades marginales $p(x_i, e)$. De esta forma se tiene

$$p(x_i, e) = \sum_{x \setminus \{x_i, e\}} p_e(x_1, \ldots, x_n), \tag{8.2}$$

donde $p_e(x_1, \ldots, x_n)$ es la función de probabilidad obtenida sustituyendo en $p(x_1, \ldots, x_n)$ las variables con evidencia, E, por sus valores e. Por tanto, para calcular $p(x_i, e)$, ha de sumarse $p_e(x_1, \ldots, x_n)$ para todas las posibles combinaciones de valores de las variables que no estén contenidas en E, excepto la variable X_i. Cuando no se dispone de evidencia, la ecuación (8.2) se reduce a

$$p(x_i) = \sum_{x \setminus x_i} p(x_1, \ldots, x_n). \tag{8.3}$$

Debido al elevado número de combinaciones de valores que involucra (8.3), este método de "fuerza bruta" resulta altamente ineficiente, incluso en redes con un número reducido de variables. Por ejemplo, en el caso de variables binarias, la ecuación (8.3) requiere la suma de 2^{n-1} probabilidades distintas. En la Figura 8.1 se muestra el tiempo de computación necesario para calcular $p(x_i)$ en un ordenador personal. Esta figura muestra que el tiempo de computación crece de forma exponencial con el número de variables del modelo, n. Puede observarse que este método es ineficiente incluso para modelos con sólo unas decenas de variables.

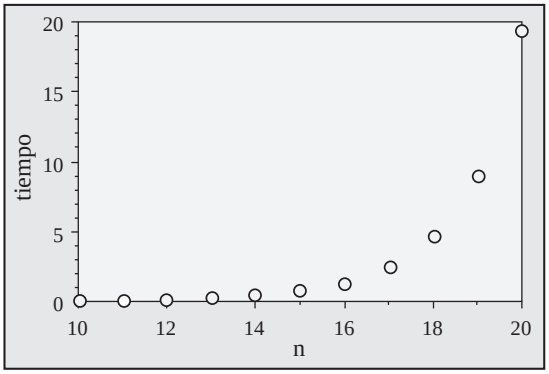


FIGURA 8.1. Tiempo de computación (minutos) necesario para calcular $p(x_i)$ utilizando (8.3) para modelos probabilísticos de n variables.

El problema de las ecuaciones (8.2) y (8.3) es que no tienen en cuenta la estructura de independencia contenida en la función de probabilidad $p(x)$. El número de cálculos necesarios en el proceso de propagación puede ser reducido de forma importante, teniendo en cuenta las distintas relaciones de independencia entre las variables de la función de probabilidad $p(x)$. En el ejemplo siguiente se ilustra este hecho.

Ejemplo 8.1 Utilizando la estructura de independencia. Considérese el modelo probabilístico formado por el conjunto de variables $X = \{A, \ldots, G\}$ y una función de probabilidad $p(x)$ que puede ser factorizada según el grafo dirigido acíclico mostrado en la Figura 8.2 (ver Sección 6.4.4)

$$p(x) = \prod_{i=1}^{n} p(x_i|\pi_i) = p(a)p(b)p(c|a)p(d|a,b)p(e)p(f|d)p(g|d,e), \qquad (8.4)$$

donde π_i es una realización de Π_i, el conjunto de los padres del nodo X_i. Supóngase que desean calcularse las probabilidades marginales de los nodos, es decir, las probabilidades iniciales cuando no se conoce ninguna evidencia. En ese caso, el método más sencillo para obtener $p(x_i)$ es marginalizar la función de probabilidad utilizando (8.3). Por ejemplo, las probabilidades iniciales de la variable D se pueden obtener mediante

$$p(d) = \sum_{x \setminus d} p(x) = \sum_{a,b,c,e,f,g} p(a,b,c,d,e,f,g). \qquad (8.5)$$

Considerando el caso más simple, es decir, suponiendo que todas las variables son binarias, el sumatorio anterior contendría $2^6 = 64$ términos distintos.

Una forma más eficiente de calcular esas probabilidades es utilizar la estructura de independencia contenida en la función de probabilidad conjunta $p(x)$. Esta estructura se pone de manifiesto en la factorización (8.4), que

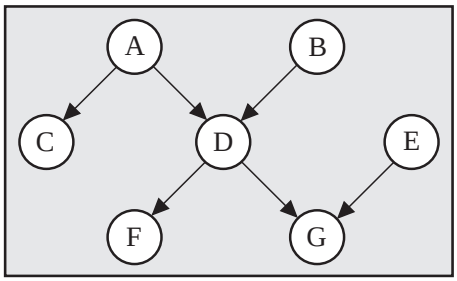


FIGURA 8.2. Un grafo dirigido acíclico.

permite simplificar tanto el proceso de definición del modelo probabilístico, como el proceso de propagación de evidencia. El número de operaciones en (8.5) puede ser reducido agrupando los términos dentro del sumatorio de la forma siguiente:

$$\begin{aligned} p(d) &= \sum_{a,b,c,e,f,g} p(a)p(b)p(c|a)p(d|a,b)p(e)p(f|d)p(g|d,e) \\ &= \left(\sum_{a,b,c} p(a)p(b)p(c|a)p(d|a,b)\right)\left(\sum_{e,f,g} p(e)p(g|d,e)p(f|d)\right), \end{aligned} \tag{8.6}$$

donde cada uno de los dos sumatorios puede ser calculado de forma independiente. Por tanto, el problema original de marginalizar una función de probabilidad de seis variables se reduce a marginalizar dos funciones que dependen sólo de tres variables. Dado que el tiempo de computación necesario para calcular cada sumatorio es exponencial en el número de variables, la complejidad de los cálculos se simplifica enormemente. En este ejemplo, se ha reducido el número de términos de cada sumatorio (de 64 a $2^3 + 2^3 = 16$) y el número de factores de cada uno de los términos (de 7 a 4 y de 7 a 3, respectivamente). Puede obtenerse una reducción adicional utilizando de nuevo la estructura dada por la factorización para reordenar los términos dentro de los sumatorios de (8.6) de la forma siguiente:

$$\sum_a \left[p(a) \sum_c \left[p(c|a) \sum_b p(b)p(d|a,b) \right] \right] \sum_e \left[p(e) \sum_f \left[p(f|d) \sum_g p(g|d,e) \right] \right],$$

reduciendo el número de términos que aparecen dentro de cada sumatorio. ∎

El ejemplo anterior ilustra las simplificaciones que pueden obtenerse utilizando la estructura de independencia contenida en el modelo probabilístico. En este capítulo se describen varios algoritmos de propagación que tienen en cuenta dicha estructura. En la Sección 8.3 se describe un algoritmo muy

eficiente para redes con estructura de poliárbol. A continuación, se presentan otros métodos más generales para tratar el problema de propagación en redes probabilísticas con estructura arbitraria.

8.3 Propagación en Poliárboles

El poliárbol es uno de los modelos gráficos más simples para construir redes Bayesianas. En esta sección se presenta un algoritmo de propagación para este tipo de modelos probabilísticos (ver Kim y Pearl (1983) y Pearl (1986b)). La característica principal de este algoritmo es que su complejidad es lineal en el tamaño de la red (es decir en el número de nodos y aristas que la componen), a diferencia del método de fuerza bruta que requiere un número exponencial de operaciones para realizar la propagación (ver Figura 8.1).

Como ya se ha visto en el Capítulo 4, en un poliárbol dos nodos cualesquiera están unidos por un único camino, lo cual implica que cada nodo divide al poliárbol en dos poliárboles inconexos: uno que contiene a sus padres y a los nodos a los que está conectado a pasando por sus padres, y otro que incluye sus hijos y a los nodos a los que está conectado pasando por sus hijos. Por ejemplo, el nodo D divide al poliárbol de la Figura 8.2 en dos poliárboles inconexos, el primero de los cuales, $\{A, B, C\}$, incluye a sus padres y a los nodos que son accesibles desde D a través de sus padres, y el segundo, $\{E, F, G\}$, que incluye a sus hijos y a los nodos que son accesibles desde D a través de sus hijos. Este hecho se muestra en la Figura 8.3, en la cual también puede comprobarse que el nodo D separa a estos dos conjuntos, es decir, que se verifica gráficamente la relación de independencia $I(\{A, B, C\}, \{E, F, G\}|D)$.

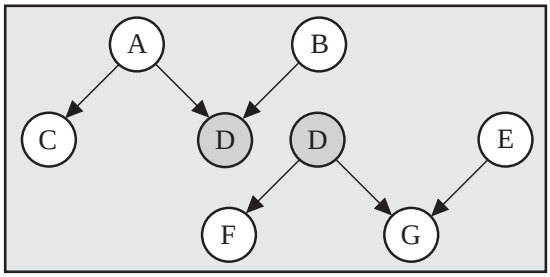


FIGURA 8.3. El nodo D divide al poliárbol en dos poliárboles inconexos.

El proceso de propagación puede realizarse en este tipo de grafos de un modo eficiente combinando la información procedente de los distintos subgrafos mediante el envío de mensajes (cálculos locales) de un subgrafo a otro.

Supóngase que se conoce una cierta evidencia $E = e$ y que se quieren calcular las probabilidades $p(x_i|e)$ para todos los valores x_i de un nodo cualquiera X_i que no esté contenido en E. Para facilitar el cálculo de estas probabilidades, el conjunto de evidencia E se puede descomponer en dos subconjuntos disjuntos, cada uno de los cuales está contenido en uno de los poliárboles separados por el nodo X_i en el poliárbol original. Por tanto, E se puede descomponer como:

- E_i^+, que es el subconjunto de E accesible desde X_i a través de sus padres.
- E_i^-, que es el subconjunto de E accesible desde X_i a través de sus hijos.

Por tanto, se tiene $E = E_i^+ \cup E_i^-$. En algunos casos se utilizará $E_{X_i}^+$ en lugar de E_i^+. Aplicando (8.1) se tiene

$$p(x_i|e) = p(x_i|e_i^-, e_i^+) = \frac{1}{p(e_i^-, e_i^+)} p(e_i^-, e_i^+|x_i)p(x_i).$$

Dado que X_i separa E_i^- de E_i^+ en el poliárbol, es decir, dado que se cumple la relación de independencia $I(E_i^-, E_i^+|X_i)$, entonces se tiene

$$\begin{aligned} p(x_i|e) &= \frac{1}{p(e_i^-, e_i^+)}\, p(e_i^-|x_i)\, p(e_i^+|x_i)p(x_i) \\ &= \frac{1}{p(e_i^-, e_i^+)}\, p(e_i^-|x_i)\, p(x_i, e_i^+) \\ &= k\; p(e_i^-|x_i)\; p(x_i, e_i^+) \\ &= k\; \lambda_i(x_i)\; \rho_i(x_i), \end{aligned}$$

donde $k = 1/p(e_i^-, e_i^+)$ es una constante de normalización,

$$\lambda_i(x_i) = p(e_i^-|x_i), \tag{8.7}$$

que tiene en cuenta la evidencia procedente de los hijos de X_i, y

$$\rho_i(x_i) = p(x_i, e_i^+), \tag{8.8}$$

que tiene en cuenta la evidencia procedente de los padres de X_i. Por tanto, la función de probabilidad condicionada sin normalizar viene dada por

$$\beta_i(x_i) = \lambda_i(x_i)\rho_i(x_i). \tag{8.9}$$

Entonces

$$p(x_i|e) = k\; \beta_i(x_i) \propto \beta_i(x_i). \tag{8.10}$$

Para calcular las funciones $\lambda_i(x_i)$, $\rho_i(x_i)$ y $\beta_i(x_i)$, se considera la situación siguiente, en la que un nodo arbitrario X_i tiene p padres y c hijos. Para simplificar la notación, los padres se representan mediante $U = \{U_1, \ldots, U_p\}$ y los hijos mediante $Y = \{Y_1, \ldots, Y_c\}$, tal como se ilustra en la Figura 8.4.

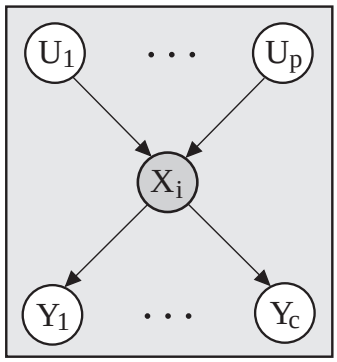


FIGURA 8.4. Los padres e hijos de un nodo arbitrario X_i.

Teniendo en cuenta la estructura del poliárbol, el conjunto de evidencia E_i^+ puede ser descompuesto en p subconjuntos disjuntos, uno para cada padre de X_i:

$$E_i^+ = \{E_{U_1 X_i}^+, \ldots, E_{U_p X_i}^+\}, \tag{8.11}$$

donde la evidencia $E_{U_j X_i}^+$ es el subconjunto de E_i^+ contenido en el subgrafo asociado al nodo U_j cuando se elimina la arista $U_j \to X_i$. De forma similar, la evidencia E_i^- también puede ser dividida en c subconjuntos disjuntos, uno asociado a cada hijo de X_i:

$$E_i^- = \{E_{X_i Y_1}^-, \ldots, E_{X_i Y_c}^-\}, \tag{8.12}$$

donde $E_{X_i Y_j}^-$ es el subconjunto de E_i^- contenido en el subgrafo asociado al nodo Y_j cuando se elimina la arista $X_i \to Y_j$. La Figura 8.5 muestra los distintos subconjuntos de evidencia asociados al nodo X_i.

Sea $u = \{u_1, \ldots, u_p\}$ una realización de los padres del nodo X_i. Entonces, la función $\rho_i(x_i)$ puede ser calculada de la forma siguiente:

$$\begin{aligned}
\rho_i(x_i) &= p(x_i, e_i^+) = \sum_u p(x_i, u \cup e_i^+) \\
&= \sum_u p(x_i | u \cup e_i^+) p(u \cup e_i^+) \\
&= \sum_u p(x_i | u \cup e_i^+) p(u \cup e_{U_1 X_i}^+ \cup \ldots \cup e_{U_p X_i}^+).
\end{aligned}$$

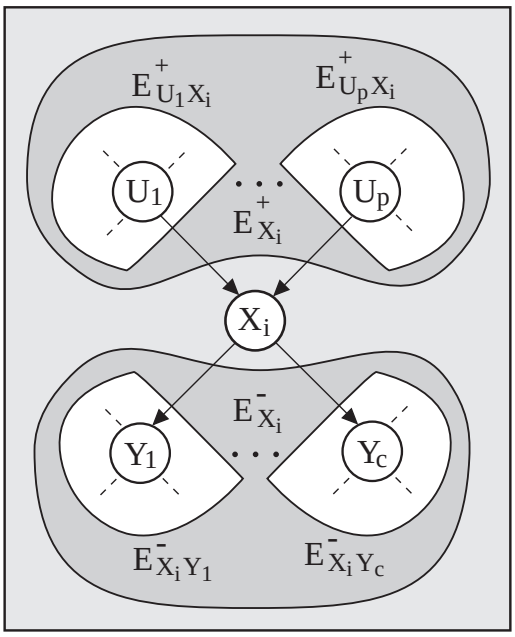


FIGURA 8.5. División del conjunto E en subconjuntos asociados a los padres e hijos de un nodo arbitrario X_i.

Dado que $\{U_j, E^+_{U_j\,X_i}\}$ es incondicionalmente independiente de $\{U_k, E^+_{U_k\,X_i}\}$ para $j \neq k$, se tiene

$$\begin{aligned}\rho_i(x_i) &= \sum_u p(x_i|u \cup e_i^+) \prod_{j=1}^{p} p(u_j \cup e^+_{U_j\,X_i}) \\ &= \sum_u p(x_i|u \cup e_i^+) \prod_{j=1}^{p} \rho_{U_j\,X_i}(u_j), \end{aligned} \tag{8.13}$$

donde

$$\rho_{U_j\,X_i}(u_j) = p(u_j \cup e^+_{U_j\,X_i}) \tag{8.14}$$

es el mensaje ρ que el nodo U_j envía a su hijo X_i. Este mensaje sólo depende de la información contenida en el subgrafo asociado al nodo U_j cuando se elimina la arista $U_j \rightarrow X_i$. Obsérvese que si U_j es una variable con evidencia, $u_j = e_j$, entonces el mensaje correspondiente, $\rho_{U_j\,X_i}(u_j)$, es la función trivial

$$\rho_{U_j\,X_i}(u_j) = \begin{cases} 1, & \text{si } u_j = e_j, \\ 0, & \text{si } u_j \neq e_j. \end{cases} \tag{8.15}$$

La función $\lambda_i(x_i)$ puede ser calculada de forma análoga:

$$\lambda_i(x_i) = p(e_i^-|x_i) = p(e^-_{X_i\,Y_1}, \ldots, e^-_{X_i\,Y_s}|x_i).$$

Dado que X_i D-separa $E^-_{X_i\,Y_j}$ de $E^-_{X_i\,Y_k}$ para $j \neq k$, entonces se tiene[2]

$$\lambda_i(x_i) = \prod_{j=1}^{c} \lambda_{Y_j\,X_i}(x_i), \tag{8.16}$$

donde

$$\lambda_{Y_j\,X_i}(x_i) = p(e^-_{X_i\,Y_j}|x_i) \tag{8.17}$$

es el mensaje λ que el nodo Y_j envía a su padre X_i.

A partir de (8.13) puede verse que un nodo X_i puede calcular su función ρ, $\rho_i(x_i)$, una vez que haya recibido los mensajes ρ de todos sus padres. De forma similar, a partir de (8.16) puede observarse que la función $\lambda_i(x_i)$ puede ser calculada una vez que X_i haya recibido los mensajes λ de todos sus hijos. La Figura 8.6 muestra los distintos mensajes asociados a un nodo arbitrario X_i.

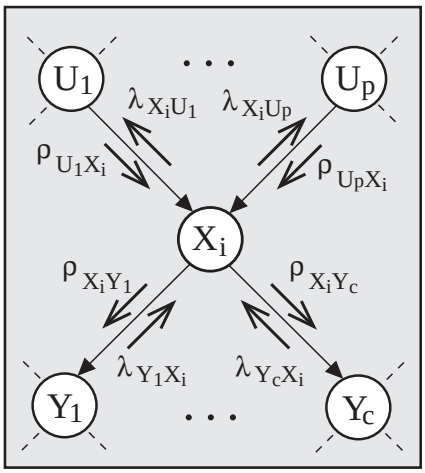


FIGURA 8.6. Mensajes ρ y λ enviados y recibidos por un nodo X_i.

Substituyendo (8.13) y (8.16) en (8.10), se tiene

$$p(x_i|e) \propto \beta_i(x_i) = \left(\sum_u p(x_i|u \cup e_i^+) \prod_{j=1}^{p} \rho_{U_j\,X_i}(u_j)\right) \left(\prod_{j=1}^{c} \lambda_{Y_j\,X_i}(x_i)\right).$$

A continuación se calculan los distintos mensajes que aparecen en la fórmula anterior. Considerando de nuevo un nodo arbitrario X_i y uno

[2]Recuérdese que se dice que un nodo Z D-separa a X de Y, si y sólo si Z separa X e Y en el grafo moral del menor subconjunto ancestral que contenga a X, Y y Z (Definición 5.4). Recuérdese también que un conjunto se dice ancestral si contiene a los ascendientes de todos sus nodos (ver Definición 4.20).

cualquiera de sus hijos, Y_j, y utilizando la igualdad

$$E^+_{X_i\,Y_j} = E^+_i \bigcup_{k\neq j} E^-_{X_i\,Y_k},$$

se tiene

$$\begin{aligned}
\rho_{X_i\,Y_j}(x_i) &= p(x_i \cup e^+_{X_i\,Y_j}) \\
&= p(x_i \cup e^+_i \bigcup_{k\neq j} e^-_{X_i\,Y_k}) \\
&= p(e^+_i | x_i \bigcup_{k\neq j} e^-_{X_i\,Y_k}) p(x_i \bigcup_{k\neq j} e^-_{X_i\,Y_k}) \\
&= p(e^+_i|x_i) p(\bigcup_{k\neq j} e^-_{X_i\,Y_k}|x_i) p(x_i) \\
&\propto p(x_i|e^+_i) \prod_{k\neq j} p(e^-_{X_i\,Y_k}|x_i) \\
&\propto \rho_i(x_i) \prod_{k\neq j} \lambda_{Y_k\,X_i}(x_i). \qquad (8.18)
\end{aligned}$$

Obsérvese que si X_i es un nodo con evidencia, entonces (8.18) también es válido si se considera la función ρ siguiente para este nodo: $\rho_i(x_i) = 1$ si $x_i = e_i$, y $\rho_i(x_i) = 0$ si $x_i \neq e_i$. En este caso, el valor de $\rho_{X_i\,Y_j}(x_i)$ obtenido a partir de (8.18) es el mismo que se obtiene de (8.15). Este hecho hace más sencilla la implementación de este método de propagación.

Por otra parte, para calcular el mensaje $\lambda_{Y_j\,X_i}(x_i)$, se considera el conjunto de todos los padres de Y_j distintos de X_i, $V = \{V_1, \ldots, V_q\}$. Por tanto, el nodo Y_j tiene $q+1$ padres, como se muestra en la Figura 8.7. Entonces,

$$e^-_{X_i\,Y_j} = e^-_{Y_j} \cup e^+_{V\,Y_j},$$

donde $e^+_{V\,Y_j}$ representa la evidencia obtenida a través de todos los padres de Y_j, excepto de X_i. Por tanto, se tiene

$$\begin{aligned}
\lambda_{Y_j\,X_i}(x_i) &= p(e^-_{X_i\,Y_j}|x_i) = \sum_{y_j,v} p(y_j, v, e^-_{X_i\,Y_j}|x_i) \\
&= \sum_{y_j,v} p(y_j, v, e^-_{Y_j}, e^+_{V\,Y_j}|x_i) \\
&= \sum_{y_j,v} p(e^-_{Y_j}|y_j, v, e^+_{V\,Y_j}, x_i) p(y_j|v, e^+_{V\,Y_j}, x_i) p(v, e^+_{V\,Y_j}|x_i) \\
&= \sum_{y_j} p(e^-_{Y_j}|y_j) \sum_{v} p(y_j|v, x_i) p(v, e^+_{V\,Y_j}), \qquad (8.19)
\end{aligned}$$

donde la última igualdad se ha obtenido considerando las relaciones de independencia existentes entre los distintos conjuntos de evidencia. Por

tanto, (8.19) puede escribirse como

$$\lambda_{Y_j\, X_i}(x_i) = \sum_{y_j} \lambda_{Y_j}(y_j) \sum_{v_1,\ldots,v_q} p(y_j|\pi_{Y_i}) \prod_{k=1}^{q} \rho_{V_k\, Y_j}(v_k). \tag{8.20}$$

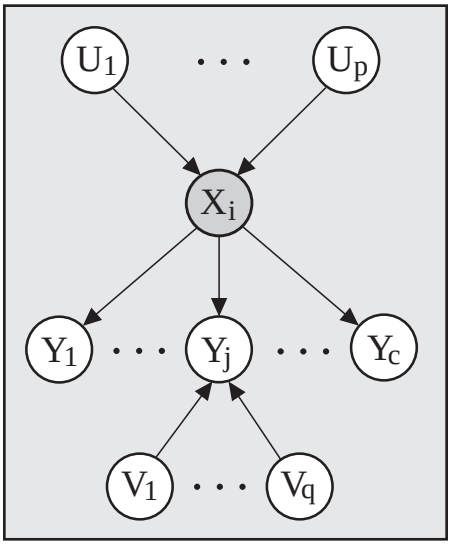


FIGURA 8.7. Conjunto de padres de un hijo, Y_j, del nodo X_i.

A partir de las ecuaciones (8.13), (8.16), (8.18) y (8.20) se puede concluir lo siguiente:

- La ecuación (8.13) muestra que la función $\rho_i(x_i)$ puede ser calculada tan pronto como el nodo X_i haya recibido los mensajes ρ de todos sus padres.
- La ecuación (8.16) muestra que la función $\lambda_i(x_i)$ puede ser calculada tan pronto como el nodo X_i haya recibido los mensajes λ de todos sus hijos.
- La ecuación (8.18) muestra que el nodo X_i puede enviar el mensaje $\rho_{X_i\, Y_j}(x_i)$ a su hijo Y_j tan pronto como haya calculado su función $\rho_i(x_i)$ y haya recibido los mensajes λ del resto de sus hijos.
- La ecuación (8.20) muestra que el nodo X_i puede enviar el mensaje $\lambda_{X_i\, U_j}(u_j)$ a su padre U_j tan pronto como haya calculado su función $\lambda_i(x_i)$ y haya recibido los mensajes ρ del resto de sus padres.

La discusión anterior sugiere el algoritmo iterativo siguiente para calcular $p(x_i|e)$ para todos los nodos $X_i \notin E$.

Algoritmo 8.1 Propagación en Poliárboles.

- **Datos:** Una red Bayesiana (D, P) definida sobre un conjunto de variables X a partir de un poliárbol D y un conjunto de nodos evidenciales E que toman los valores $E = e$.
- **Resultados:** Las funciones de probabilidad condicionada $p(x_i|e)$ para cada nodo $X_i \notin E$.

Etapa de Iniciación:

1. Asignar a todos los nodos evidenciales, $X_i \in E$, las funciones
   - $\rho_i(x_i) = 1$ si $x_i = e_i$, o $\rho_i(x_i) = 0$ si $x_i \neq e_i$.
   - $\lambda_i(x_i) = 1$ si $x_i = e_i$, o $\lambda_i(x_i) = 0$ si $x_i \neq e_i$.

   (El efecto de esta asignación es reducir los valores posibles de los nodos evidenciales X_i, eliminando todos aquellos valores que la contradicen.)
2. Asignar a todos los nodos $X_i \notin E$ que no tengan padres la función $\rho_i(x_i) = p(x_i)$.
3. Asignar a todos los nodos $X_i \notin E$ que no tengan hijos la función $\lambda_i(x_i) = 1$, para todo x_i.

Etapa Iterativa:

4. Para cada nodo $X_i \notin E$:
   (a) Si X_i ha recibido los mensajes ρ de todos sus padres, calcular $\rho_i(x_i)$ utilizando (8.13).
   (b) Si X_i ha recibido los mensajes λ de todos sus hijos, calcular $\lambda_i(x_i)$ utilizando (8.16).
   (c) Si ya se ha calculado $\rho_i(x_i)$, entonces, para cada hijo Y_j de X_i tal que X_i haya recibido los mensajes λ del resto de sus hijos, calcular y enviar el mensaje $\rho_{X_i Y_j}(x_i)$ utilizando (8.18). Por tanto, si X_i ha recibido los mensajes λ de todos sus hijos, entonces, puede ya enviar todos los mensajes ρ correspondientes.
   (d) Si ya se ha calculado $\lambda_i(x_i)$, entonces, para cada padre U_j de X_i tal que X_i haya recibido los mensajes ρ del resto de sus padres, calcular y enviar el mensaje $\lambda_{X_i U_j}(u_i)$ utilizando (8.20). Análogamente al caso anterior, si X_i ha recibido los mensajes ρ de todos sus padres, entonces, ya puede enviar todos los mensajes λ correspondientes.
5. Repetir el Paso 4 tantas veces como sea necesario hasta que se calculen las funciones ρ y λ de todos los nodos $X_i \notin E$, es decir, hasta que no se produzca ningún nuevo mensaje en una iteración completa.

6. Para cada nodo $X_i \notin E$, calcular $\beta_i(x_i)$ utilizando (8.9). Estas son las probabilidades no normalizadas correspondientes a $p(x_i|e)$.

7. Para cada nodo $X_i \notin E$, calcular $p(x_i|e)$ normalizando la función $\beta_i(x_i)$, es decir, $p(x_i|e) = \beta_i(x_i)/k$, donde $k = \sum_{x_i} \beta_i(x_i)$. ■

Obsérvese que durante el proceso de propagación, las funciones ρ y λ de cada nodo se calculan en distintas etapas de la iteración. Por tanto, si sólo se está interesado en una variable objetivo X_i, el algoritmo puede detenerse una vez que se conozcan las funciones $\rho_i(x_i)$ y $\lambda_i(x_i)$. En la Sección 8.8 se muestra cómo puede simplificarse el proceso de propagación en esta situación.

La estructura de envío de mensajes utilizada en el Algoritmo 8.1 hace posible su implementación distribuida (paralela), en la que distintos procesadores realizan diferentes tareas simultáneas cuya combinación permite resolver el problema. Supóngase que se asocia a cada nodo de la red su propio procesador. El procesador de un nodo arbitrario, X_i, necesita conocer la siguiente información para calcular $p(x_i|e)$:

- Dos listas: una formada por los padres de X_i y otra por los hijos. Esta información es independiente de la evidencia E.

- La función de probabilidad condicionada $p(x_i|\pi_{X_i})$, que también es independiente de la evidencia E. Si X_i no tiene padres, entonces $p(x_i|\pi_{X_i}) = p(x_i)$, que es la función de probabilidad marginal de X_i.

- La función $\rho_i(x_i)$, que se calcula por el procesador correspondiente al nodo X_i utilizando (8.13).

- La función $\lambda_i(x_i)$, que se calcula por el procesador asociado al nodo X_i utilizando (8.16).

- El mensaje $\rho_{U_j\,X_i}(u_j)$, recibido de cada uno de los padres, U_j, del nodo X_i. Este mensaje se calcula por el procesador asociado al nodo U_j, utilizando (8.18).

- El mensaje $\lambda_{Y_j\,X_i}(x_i)$, recibido de cada uno de los hijos, Y_j del nodo X_i. Este mensaje se calcula por el procesador asociado al nodo Y_j, utilizando (8.20).

Una vez que el procesador del nodo X_i ha recibido la información anterior, puede calcular las probabilidades no normalizadas $\beta_i(x_i)$ utilizando (8.9). Finalmente, normalizando estos valores se obtiene la función de probabilidad condicionada $p(x_i|e)$.

Por otra parte, cada procesador tiene que calcular los siguientes mensajes para enviar a sus vecinos:

- El mensaje $\rho_{X_i\,Y_j}(x_i)$, que es enviado a cada hijo Y_j del nodo X_i. Este mensaje se calcula por el procesador asociado al nodo X_i, utilizando (8.18).
- El mensaje $\lambda_{X_i\,U_j}(u_j)$, que es enviado a cada padre U_j del nodo X_i. Este mensaje se calcula por el procesador asociado al nodo X_i, utilizando (8.20).

La Figura 8.8 muestra los cálculos realizados por el procesador de un nodo arbitrario, X_i, así como los mensajes recibidos y enviados desde el nodo. Esta figura ilustra las operaciones básicas necesarias para una implementación paralela de este algoritmo (este problema se trata en detalle en Díez y Mira (1994)).

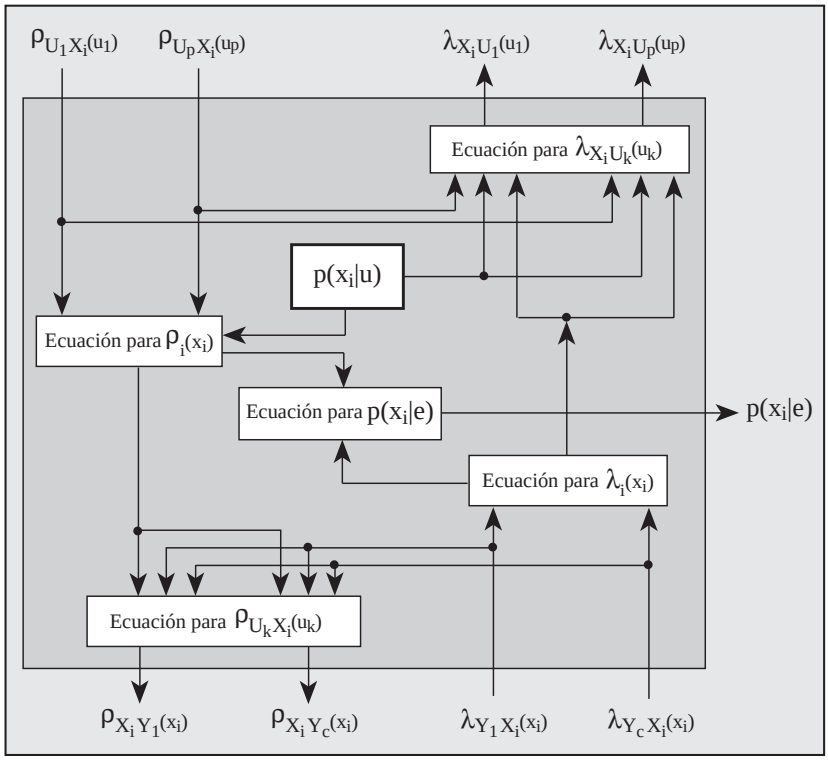


FIGURA 8.8. Cálculos realizados por el procesador de un nodo arbitrario X_i y mensajes que son recibidos y enviados a padres e hijos.

La complejidad del Algoritmo 8.1 es lineal en el número de nodos y aristas (el tamaño de la red). Por tanto, cuando una red Bayesiana tiene estructura de poliárbol, el proceso de propagación puede ser llevado a cabo de forma eficiente mediante este algoritmo. Como se verá más adelante, el problema de la propagación en redes Bayesianas con estructura arbitraria

es un problema NP-complejo[3] (ver Cooper (1990)). Esto significa que no existe ningún algoritmo que resuelva el problema en tiempo lineal para redes Bayesianas con cualquier topología. Las Secciones 8.5−8.7 presentan varios algoritmos de propagación exacta para este tipo de redes.

Peot y Shachter (1991) y Delcher y otros (1995) presentan algunas modificaciones del algoritmo de propagación en poliárboles que mejoran su eficiencia.

Ejemplo 8.2 Propagación en poliárboles (I). Considérese la red Bayesiana introducida en el Ejemplo 8.1 cuyo grafo dirigido acíclico se muestra en la Figura 8.2. La función de probabilidad de la red Bayesiana puede ser factorizada según este grafo como

$$p(a,b,c,d,e,f,g) = p(a)p(b)p(c|a)p(d|a,b)p(e)p(f|d)p(g|d,e).$$

Para simplificar los cálculos se supondrá que las variables son binarias. Las funciones de probabilidad condicionada que definen el modelo se muestran en la Tabla 8.1.

En primer lugar, se considera el caso de no disponer de evidencia. Por tanto, el algoritmo de propagación en poliárboles permitirá obtener las funciones de probabilidad iniciales $p(x_i)$, de todos los nodos del grafo. A continuación se considera la evidencia $D = 0$ y se utiliza el mismo algoritmo para actualizar las probabilidades, es decir, para obtener las funciones de probabilidad condicionada $p(x_i|D = 0)$ para todos los nodos $X_i \neq D$.

Para ilustrar las distintas etapas del algoritmo, se describen todos los pasos realizados, siguiendo el orden alfabético de los nodos en cada paso de iteración. En el caso de no disponer de evidencia, el Algoritmo 8.1 consta de los siguientes pasos:

Etapa de Iniciación:

- En este caso no se aplica el primer paso de la etapa de iniciación, pues no se tiene evidencia.

- Las funciones ρ de los nodos sin padres, A, B y E, se definen como

$$\rho_A(a) = p(a),\ \rho_B(b) = p(b),\ \rho_E(e) = p(e).$$

  Por tanto, a partir de los valores mostrados en la Tabla 8.1, se tiene

$$\begin{array}{lll} \rho_A(0) = 0.3, & \rho_B(0) = 0.6, & \rho_E(0) = 0.1, \\ \rho_A(1) = 0.7, & \rho_B(1) = 0.4, & \rho_E(1) = 0.9. \end{array}$$

[3]Para una introducción a la complejidad de algoritmos y a los problemas NP-complejo el lector puede consultar el libro de Garey y Johnson (1979).

a	$p(a)$
0	0.3
1	0.7

b	$p(b)$
0	0.6
1	0.4

e	$p(e)$
0	0.1
1	0.9

a	c	$p(c\|a)$
0	0	0.25
0	1	0.75
1	0	0.50
1	1	0.50

d	f	$p(f\|d)$
0	0	0.80
0	1	0.20
1	0	0.30
1	1	0.70

a	b	d	$p(d\|a,b)$
0	0	0	0.40
0	0	1	0.60
0	1	0	0.45
0	1	1	0.55
1	0	0	0.60
1	0	1	0.40
1	1	0	0.30
1	1	1	0.70

d	e	g	$p(g\|d,e)$
0	0	0	0.90
0	0	1	0.10
0	1	0	0.70
0	1	1	0.30
1	0	0	0.25
1	0	1	0.75
1	1	0	0.15
1	1	1	0.85

TABLA 8.1. Funciones de probabilidad condicionada de la red Bayesiana del Ejemplo 8.2.

- Las funciones λ de los nodos sin hijos, C, F y G, se definen como

$$\lambda_C(0) = 1.0, \quad \lambda_F(0) = 1.0, \quad \lambda_G(0) = 1.0,$$
$$\lambda_C(1) = 1.0, \quad \lambda_F(1) = 1.0, \quad \lambda_G(1) = 1.0.$$

La Figura 8.9 muestra las funciones ρ y λ calculadas en la etapa de iniciación. Los números indican el orden en el que se calculan las diferentes funciones.

Primer Paso de Iteración:

- *Nodo A:* Aplicando al nodo A las reglas dadas en la etapa de iteración del Algoritmo 8.1, se tiene

  (a) La función $\rho_A(a)$ ha sido calculada en la etapa de iniciación.

  (b) La función $\lambda_A(a)$ no puede ser calculada, pues A no ha recibido el mensaje λ de ninguno de sus dos hijos, C y D.

  (c) La función $\rho_A(a)$ ha sido calculada, pero A no puede enviar los mensajes $\rho_{AC}(a)$ y $\rho_{AD}(a)$ a sus hijos, ya que no ha recibido los mensajes λ de D y C, respectivamente.

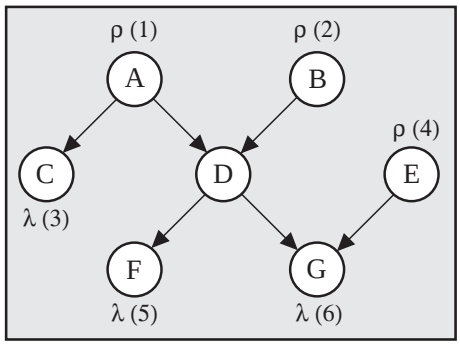


FIGURA 8.9. Etapa de iniciación del algoritmo de propagación en poliárboles.

(d) Dado que el nodo A no tiene padres, no tiene que enviar ningún mensaje λ.

Por tanto, no se realiza ningún cálculo asociado al nodo A en esta etapa de iteración.

- *Nodo B:* Dado que el nodo D es el único hijo del nodo B, entonces se puede calcular el mensaje ρ_{BD} utilizando (8.18) y enviárselo a D:

$$\rho_{BD}(b) = \rho_B(b) \prod_{y_j \setminus d} \lambda_{Y_j B}(b),$$

donde Y_j es el hijo j-ésimo del nodo B. Dado que B sólo tiene un hijo, esta ecuación se reduce a $\rho_{BD}(b) = \rho(b)$. Por tanto, se tiene $(\rho_{BD}(0), \rho_{BD}(1)) = (0.6, 0.4)$.

- *Nodo C:* Dado que A es el único padre del nodo C, se puede calcular y enviar el mensaje λ_{CA} utilizando (8.20). En este caso se tiene

$$\lambda_{CA}(a) = \sum_c \lambda_C(c) p(c|a),$$

que implica

$$\begin{aligned} \lambda_{CA}(0) &= \lambda_C(0)\, p(C=0|A=0) + \lambda_C(1)\, p(C=1|A=0) \\ &= 1 \times 0.25 + 1 \times 0.75 = 1.00, \end{aligned}$$

$$\begin{aligned} \lambda_{CA}(1) &= \lambda_C(0)\, p(C=0|A=1) + \lambda_C(1)\, p(C=1|A=1) \\ &= 1 \times 0.5 + 1 \times 0.5 = 1.00. \end{aligned}$$

- *Nodo D:* En esta etapa no se realiza ningún cálculo asociado al nodo D, pues éste no ha recibido ningún mensaje de sus padres e hijos.

- *Nodo E:* Como G es el único hijo del nodo E, se puede calcular y enviar el mensaje $\rho_{EG}(e)$ utilizando (8.18). Procediendo de la misma forma que para el nodo B, se obtiene $\rho_{EG}(e) = \rho(e)$. Por tanto, $(\rho_{EG}(0), \rho_{EG}(1)) = (0.1, 0.9)$.

- *Nodo F:* La situación del nodo F es similar a la del nodo C. En este caso se tiene
$$\lambda_{FD}(d) = \sum_f \lambda(f) p(f|d),$$
es decir, $(\lambda_{FD}(0), \lambda_{FD}(1)) = (1.0, 1.0)$.

- *Nodo G:* El nodo G tiene dos padres, D y E; además la función $\lambda_G(g)$ fue calculada en el paso de iniciación, y G ha recibido el mensaje ρ del nodo E. Por tanto, el nodo G puede calcular y enviar el mensaje λ a su otro padre, D. Utilizando (8.20) se tiene
$$\lambda_{GD}(d) = \sum_g \lambda_G(g) \sum_e p(g|d,e) \rho_{EG}(e),$$
que implica
$$\begin{aligned}
\lambda_{GD}(0) &= \lambda_G(0) \sum_e p(G=0|D=0,e)\, \rho_{EG}(e) \\
&\quad + \lambda_G(1) \sum_e p(G=1|D=0,e)\, \rho_{EG}(e) \\
&= 1.0 \times (0.9 \times 0.1 + 0.7 \times 0.9) \\
&\quad + 1.0 \times (0.1 \times 0.1 + 0.3 \times 0.9) = 1.0,
\end{aligned}$$
$$\begin{aligned}
\lambda_{GD}(1) &= \lambda_G(0) \sum_e p(G=0|D=1,e)\, \rho_{EG}(e) \\
&\quad + \lambda_G(1) \sum_e p(G=1|D=1,e)\, \rho_{EG}(e) \\
&= 1.0 \times (0.25 \times 0.1 + 0.15 \times 0.9) \\
&\quad + 1.0 \times (0.75 \times 0.1 + 0.85 \times 0.9) = 1.0.
\end{aligned}$$
Por tanto, se obtiene el mensaje $(\lambda_{GD}(0), \lambda_{GD}(1)) = (1.0, 1.0)$.

La Figura 8.10 muestra el orden en el que las funciones y mensajes ρ y λ han sido calculadas en el primer paso de iteración. Todas las funciones correspondientes a la etapa anterior se muestran con menor intensidad para distinguirlas de las nuevas funciones.

Segundo Paso de Iteración:

- *Nodo A:* El nodo A tiene dos hijos, C y D. La función $\rho_A(a)$ ya ha sido calculada y el nodo A ha recibido el mensaje λ del nodo

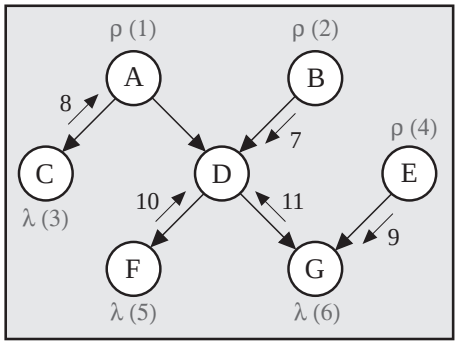


FIGURA 8.10. Primer paso de iteración del algoritmo de propagación en poliárboles.

C. Entonces, se puede calcular el mensaje ρ y enviárselo al nodo D utilizando (8.18):

$$\rho_{AD}(a) = \rho_A(a)\lambda_{CA}(a).$$

Por tanto, se tiene $(\rho_{AD}(0), \rho_{AD}(1)) = (0.3, 0.7)$.

- Los nodos B y C no han recibido los mensajes de los nodos D y A, respectivamente. Por tanto, no se puede efectuar ningún cálculo asociado a los nodos B y C en este paso de la iteración.

- *Nodo D:* El nodo D ha recibido los mensajes ρ de sus dos padres A y B. Por tanto, puede calcularse $\rho_D(d)$ utilizando (8.13):

$$\rho_D(d) = \sum_{a,b} p(d|a,b)\rho_{AD}(a)\rho_{BD}(b).$$

Por ejemplo, para $D = 0$ se tiene

$$\begin{aligned}\rho_D(0) &= p(D=0|A=0,B=0)\,\rho_{AD}(0)\rho_{BD}(0) \\ &\quad + p(D=0|A=0,B=1)\,\rho_{AD}(0)\rho_{BD}(1) \\ &\quad + p(D=0|A=1,B=0)\,\rho_{AD}(1)\rho_{BD}(0) \\ &\quad + p(D=0|A=1,B=1)\,\rho_{AD}(1)\rho_{BD}(1) \\ &= 0.4 \times 0.3 \times 0.6 + 0.45 \times 0.3 \times 0.4 \\ &\quad + 0.6 \times 0.7 \times 0.6 + 0.3 \times 0.7 \times 0.4 = 0.462.\end{aligned}$$

De forma similar, para $D = 1$ se tiene $\rho_D(1) = 0.538$. Además, el nodo D ha recibido el mensaje λ de sus dos hijos, F y G. Esto implica que $\lambda_D(d)$ puede ser calculado utilizando (8.16):

$$\lambda_D(d) = \lambda_{FD}(d)\lambda_{GD}(d),$$

obteniéndose $(\lambda_D(0), \lambda_D(1)) = (1.0, 1.0)$.

Como el nodo D ha recibido los mensajes de todos sus padres e hijos, entonces puede enviar todos los mensajes ρ y λ a sus padres e hijos. Por ejemplo, utilizando (8.18) se pueden calcular los mensajes $\rho_{DF}(d)$ y $\rho_{DG}(d)$ de la forma siguiente:

$$\rho_{DF}(d) = \rho_D(d)\lambda_{GD}(d),$$
$$\rho_{DG}(d) = \rho_D(d)\lambda_{FD}(d).$$

De forma similar, utilizando (8.20) se pueden calcular los mensajes $\lambda_{DA}(a)$ y $\lambda_{DB}(b)$ de la forma siguiente:

$$\lambda_{DA}(a) = \sum_d \lambda_D(d) \sum_b p(d|a,b)\rho_{BD}(b),$$
$$\lambda_{DB}(b) = \sum_d \lambda_D(d) \sum_a p(d|a,b)\rho_{AD}(a).$$

Los valores numéricos correspondientes a estos mensajes se muestran en la Figura 8.13.

- El nodo E no ha recibido el mensaje λ de su hijo G. Por tanto, no se puede realizar ningún cálculo con este nodo.

- *Nodo F:* El nodo F ha recibido el mensaje $\rho_{DF}(d)$ de su único padre, D. Por tanto, puede calcular la función $\rho_F(f)$:

$$\rho_F(f) = \sum_d p(f|d)\rho_{DF}(d),$$

obteniéndose $(\rho_F(0), \rho_F(1)) = (0.531, 0.469)$.

- *Nodo G:* El Nodo G ha recibido los dos mensajes ρ de sus dos padres, D y E. Por tanto, se puede calcular la función $\rho_G(g)$ utilizando (8.13):

$$\rho_G(g) = \sum_{d,e} p(g|d,e)\rho_{DG}(d)\rho_{EG}(e).$$

Por otra parte, la función $\lambda_G(g)$ también ha sido calculada. Por tanto, se puede calcular y enviar el mensaje λ al nodo E. Utilizando (8.20) se tiene:

$$\lambda_{GE}(e) = \sum_g \lambda_G(g) \sum_d p(g|d,e)\rho_{DG}(d).$$

La Figura 8.11 muestra el orden en que se calculan y envían las funciones y mensajes en el paso de iteración anterior.

Procediendo de la misma forma que en los pasos anteriores, en el último paso de iteración se calculan las funciones y mensajes siguientes: $\lambda_A(a)$,

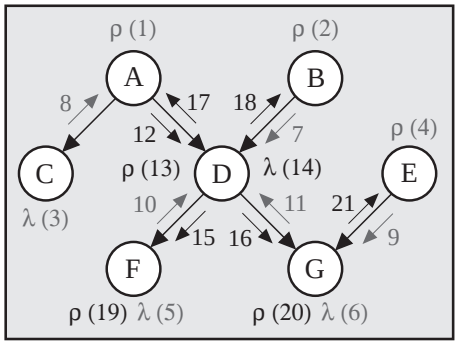


FIGURA 8.11. Segundo paso de iteración del algoritmo de propagación en poliárboles.

$\rho_{AC}(a)$, $\lambda_B(b)$, $\rho_C(c)$ y $\lambda_E(e)$. La Figura 8.12 ilustra este último paso de la etapa de iteración del algoritmo. En esta figura puede comprobarse que todos los mensajes han sido enviados y todas las funciones ρ y λ han sido calculadas. Por tanto, la etapa de iteración del algoritmo ha sido completada.

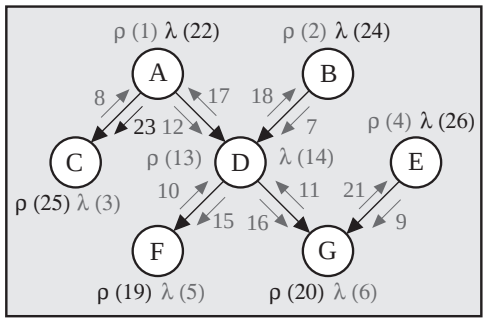


FIGURA 8.12. Último paso de iteración del algoritmo.

- En la Etapa 6 del algoritmo se calculan las funciones β. Dado que en este caso no se tiene evidencia, las funciones y mensajes λ son constantes para todos los nodos. Por tanto, en este caso $\beta_i(x_i) = \rho_i(x_i)$ para todos los nodos X_i.

- En la última etapa del algoritmo se obtienen las funciones de probabilidad marginal, $p(x_i)$, normalizando las correspondientes funciones $\beta(x_i)$. Sin embargo, en este caso las funciones β ya están normalizadas, por lo que no es necesario realizar ningún proceso de normalización. Este hecho se muestra en la Figura 8.13, que contiene los valores numéricos de todas las funciones y mensajes calculados en el proceso de propagación. ∎

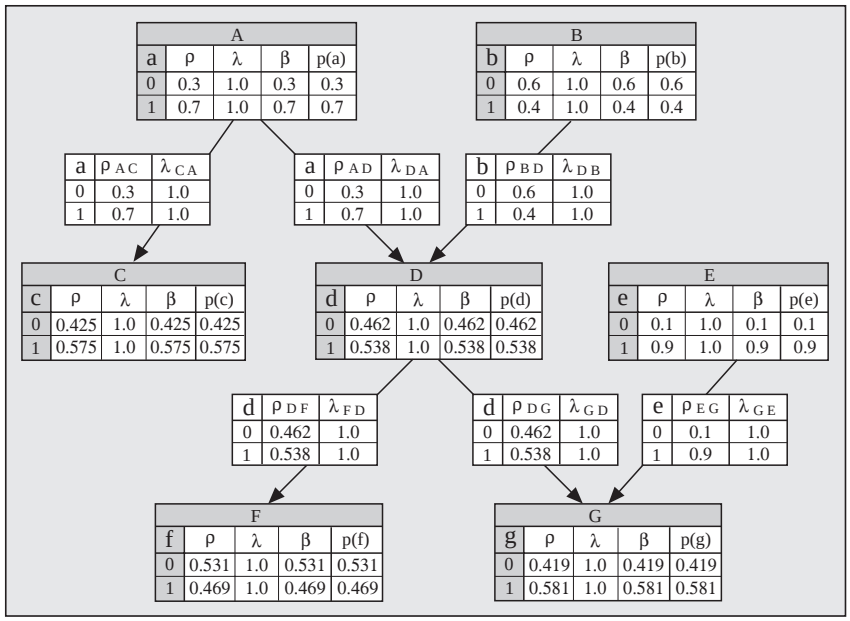


FIGURA 8.13. Valores numéricos de los mensajes y funciones calculados por el algoritmo de propagación en poliárboles cuando no se dispone de evidencia.

Obsérvese que cuando no se dispone de evidencia, todas las funciones y mensajes λ (ver ecuaciones (8.16) y (8.20)) toman el valor 1. Por tanto, en este caso no es necesario calcular estas funciones y mensajes. Sin embargo, cuando se dispone de evidencia, los mensajes λ propagan la información basada en esta evidencia de padres a hijos y, por tanto, son mensajes no triviales. Este hecho se ilustra en el ejemplo siguiente.

Ejemplo 8.3 Propagación en poliárboles (II). En este ejemplo se considera de nuevo la red Bayesiana utilizada en el Ejemplo 8.2, pero ahora se supone que se tiene la evidencia $D = 0$ y que se quieren actualizar las probabilidades iniciales obtenidas en el ejemplo anterior según esta evidencia. El Algoritmo 8.1 procede de la forma siguiente:

Etapa de Iniciación:

- En este caso se tiene la evidencia $D = 0$. Por tanto, se definen las funciones ρ y λ de este nodo de la forma siguiente:

$$\rho_D(0) = 1.0, \quad \lambda_D(0) = 1.0,$$
$$\rho_D(1) = 0.0, \quad \lambda_D(1) = 0.0.$$

- Las funciones ρ de los nodos sin padres se calculan directamente a partir de las funciones de probabilidad marginales contenidas en la

factorización del modelo probabilístico:

$$\begin{array}{lll} \rho_A(0) = 0.3, & \rho_B(0) = 0.6, & \rho_E(0) = 0.1, \\ \rho_A(1) = 0.7, & \rho_B(1) = 0.4, & \rho_E(1) = 0.9. \end{array}$$

- Se asignan valores constantes a las funciones λ de los nodos sin hijos, E, F y G:

$$\begin{array}{lll} \lambda_C(0) = 1.0, & \lambda_F(0) = 1.0, & \lambda_G(0) = 1.0, \\ \lambda_C(1) = 1.0, & \lambda_F(1) = 1.0, & \lambda_G(1) = 1.0. \end{array}$$

Primer Paso de Iteración:

- Los cálculos correspondientes al primer paso de la iteración coinciden con los realizados cuando no se tenía evidencia. Por tanto, en este paso se calculan los mensajes $\rho_{BD}(b)$, $\lambda_{CA}(a)$, $\rho_{EG}(e)$, $\lambda_{FD}(d)$ y $\lambda_{GD}(d)$ obteniéndose los mismos resultados mostrados en la Figura 8.13.

Segunda Etapa de Iteración:

- *Nodo A:* Se calcula el mensaje $\rho_{AD}(a)$ de la misma forma que cuando no se tenía evidencia, utilizando (8.18). En este caso se tiene el mensaje $(\rho_{AD}(0), \rho_{AD}(1)) = (0.3, 0.7)$.

- De la misma forma que cuando no se tenía evidencia, en este paso de la iteración no se puede realizar ninguna operación con los nodos B y C.

- *Nodo D:* Las funciones ρ y λ del nodo D ya han sido calculadas en la etapa de iniciación. El nodo D ha recibido todos los mensajes de sus padres e hijos. Por tanto, puede enviar los mensajes ρ y λ a todos sus hijos y padres, respectivamente. Aplicando (8.18) se calculan los mensajes $\rho_{DF}(d)$ y $\rho_{DG}(d)$ de la forma siguiente:

$$\begin{aligned} \rho_{DF}(d) &= \rho_D(d)\lambda_{GD}(d), \\ \rho_{DG}(d) &= \rho_D(d)\lambda_{FD}(d). \end{aligned}$$

  Por ejemplo, los valores numéricos asociados a $\rho_{DF}(d)$ son

$$\begin{aligned} \rho_{DF}(0) &= \rho_D(0)\lambda_{GD}(0) = 1.0 \times 1.0 = 1.0, \\ \rho_{DF}(1) &= \rho_D(1)\lambda_{GD}(1) = 0.0 \times 1.0 = 0.0. \end{aligned}$$

  De forma similar, para el mensaje $\rho_{DG}(d)$ se tiene $\rho_{DG}(0) = 1.0$ y $\rho_{DG}(1) = 0.0$. Aplicando (8.18) se pueden calcular los mensajes $\lambda_{DA}(a)$ y $\lambda_{DB}(b)$ de la forma siguiente:

$$\begin{aligned} \lambda_{DA}(a) &= \sum_d \lambda_D(d) \sum_b p(d|a,b)\rho_{BD}(b), \\ \lambda_{DB}(b) &= \sum_d \lambda_D(d) \sum_a p(d|a,b)\rho_{AD}(a). \end{aligned}$$

Por ejemplo, para el mensaje $\lambda_{DA}(a)$ se tienen los valores

$$\begin{aligned}\lambda_{DA}(0) &= \lambda_D(0)\sum_b p(D=0|A=0,b)\rho_{BD}(b)\\ &\quad + \lambda_D(1)\sum_b p(D=1|A=0,b)\rho_{BD}(b)\\ &= 1.0\times(0.4\times 0.6+0.45\times 0.4)\\ &\quad + 0.0\times(0.6\times 0.6+0.55\times 0.4)=0.42,\end{aligned}$$

$$\begin{aligned}\lambda_{DA}(0) &= \lambda_D(0)\sum_b p(D=0|A=1,b)\rho_{BD}(b)\\ &\quad +\lambda_D(1)\sum_b p(D=1|A=1,b)\rho_{BD}(b)\\ &= 1.0\times(0.6\times 0.6+0.3\times 0.4)\\ &\quad +0.0\times(0.4\times 0.6+0.7\times 0.4)=0.48.\end{aligned}$$

- De la misma forma que cuando no se tenía evidencia, no se puede realizar ningún cálculo asociado al nodo E. Sin embargo, se pueden calcular las funciones $\rho_F(f)$ y $\rho_G(g)$, asociadas a los nodos F y G, respectivamente, y el mensaje $\lambda_{GE}(e)$ que el nodo G envía al nodo E.

La Figura 8.14 muestra los valores numéricos correspondientes al resto de los mensajes.

La Figura 8.15(a) muestra las probabilidades iniciales de los nodos, cuando no se considera evidencia, y la Figura 8.15(b) muestra las probabilidades actualizadas cuando se considera la evidencia $D = 0$.[4] A partir de estas figuras, se puede ver que la evidencia no afecta al nodo E (la probabilidad marginal inicial coincide con la probabilidad condicionada actualizada). Sin embargo, la evidencia afecta de forma importante a algunos nodos como, por ejemplo, a los nodos F y G. La estructura de dependencia contenida en el grafo permite determinar qué variables serán afectadas por la evidencia, pero no la magnitud en que esta influencia modifica las probabilidades de los nodos. El Capítulo 10 introduce algunos algoritmos de propagación simbólica que permiten determinar el grado de influencia de la evidencia sobre cada nodo. ∎

[4]En la dirección WWW http://ccaix3.unican.es/~AIGroup se puede obtener la concha para redes Bayesianas *X-pert Nets* y los ficheros necesarios para resolver los ejemplos anteriores.

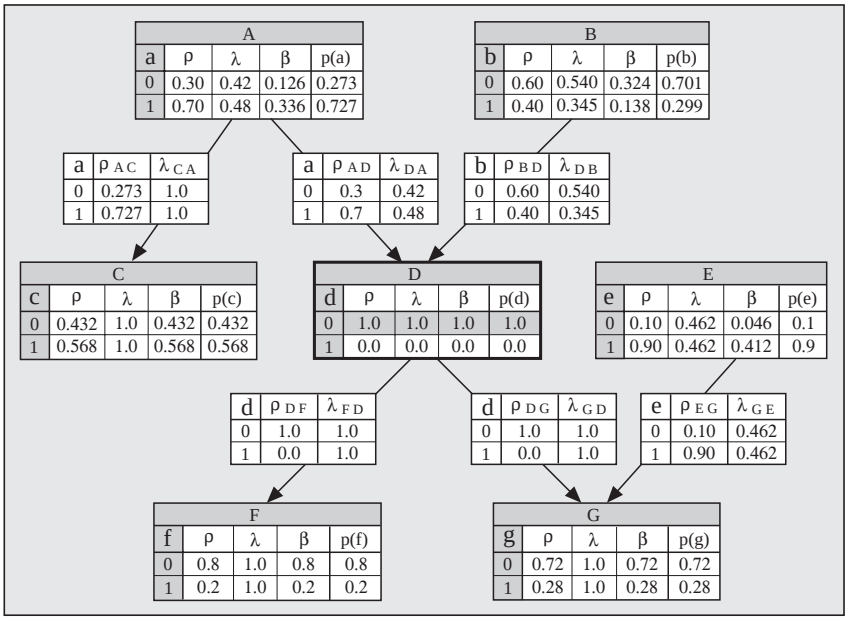


FIGURA 8.14. Valores numéricos de las funciones y mensajes calculados por el algoritmo de propagación en poliárboles considerando la evidencia $D = 0$.

8.4 Propagación en Redes Múltiplemente Conexas

El método de propagación en poliárboles descrito en la sección anterior es válido solamente para redes de estructura simple (poliárboles), en las cuales existe un único camino entre cada par de nodos. Por tanto, este tipo de redes carecen de generalidad y no son aplicables en numerosas situaciones prácticas. En estos casos es necesario trabajar con grafos múltiplemente conexos (grafos que contienen bucles) en los que pueden existir varios caminos entre dos nodos. Dos de los métodos de propagación más importantes para este tipo de redes son los denominados *método de condicionamiento* y *método de agrupamiento*. La idea fundamental del método de propagación por condicionamiento es cortar los múltiples caminos entre los nodos mediante la asignación de valores a un conjunto reducido de variables contenidas en los bucles (ver Pearl (1986a) y Suermondt y Cooper (1991b)). De esta forma se tendrá un poliárbol en el cual se podrá aplicar el algoritmo de propagación para poliárboles descrito en la sección anterior. Por otra parte, el método de agrupamiento construye representaciones auxiliares, de estructura más simple, uniendo conjuntos de nodos del grafo original (por ejemplo, un árbol de unión). De esta forma se puede obtener un grafo con estructura de poliárbol en el que pueden aplicarse las mismas ideas descritas en la sección anterior para propagar evidencia

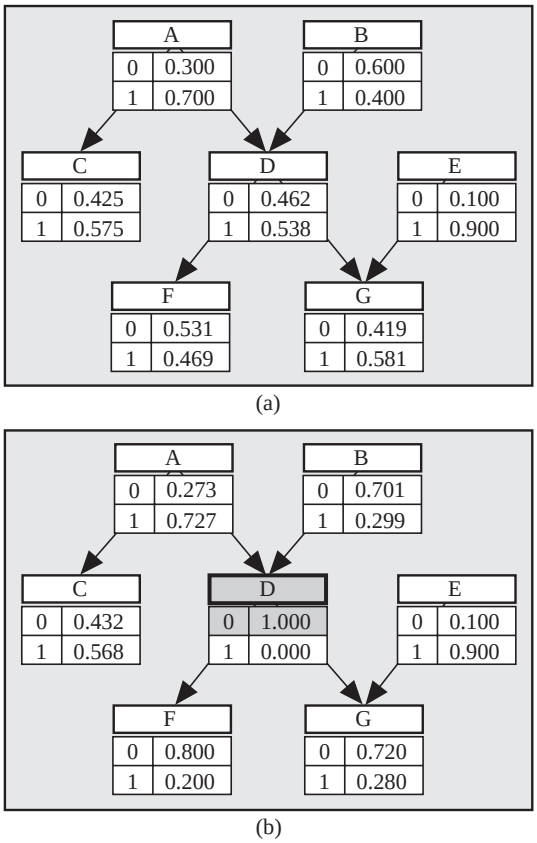


FIGURA 8.15. Probabilidades marginales (iniciales) de los nodos (a) y probabilidades condicionadas (actualizadas), dada la evidencia $D = 0$ (b).

(ver Lauritzen y Spiegelhalter (1988), Jensen, Olesen y Andersen (1990), y Shachter, Andersen y Szolovits (1994)).

A pesar de que la complejidad del algoritmo de propagación en poliárboles es lineal en el tamaño de la red, el problema de la propagación de evidencia en redes Bayesianas múltiplemente conexas es un problema NP-complejo (ver Cooper (1990)). En general, tanto el método de condicionamiento, como el de agrupamiento plantean este problema de complejidad. Sin embargo, las características particulares de estos métodos hacen que, en ocasiones, uno de ellos sea más eficiente que el otro en redes con estructura particular. Sin embargo, en general, ninguno de estos métodos es más eficiente que el otro, sino que son complementarios (Suermondt y Cooper (1991a)). Este hecho ha motivado la aparición de algunos algoritmos mixtos que combinan las ventajas de ambos métodos (ver, por ejemplo, Suermondt y Cooper (1991a) y Suermondt, Cooper y Heckerman (1991)).

La Sección 8.5 analiza el método de condicionamiento. Dado que este método utiliza el algoritmo de propagación en poliárboles, el algoritmo de condicionamiento es sólo válido para redes Bayesianas. Sin embargo, el algoritmo de agrupamiento (Sección 8.6) puede ser aplicado tanto a redes de Markov como a redes Bayesianas.

8.5 Método de Condicionamiento

En el caso de redes Bayesianas múltiplemente conexas ya no se cumple la propiedad de que un nodo cualquiera separa el grafo en dos partes inconexas. Por tanto, algunas de las propiedades de independencia aplicadas en el algoritmo de propagación en poliárboles no pueden ser aplicadas en esta situación. Por ejemplo, considérese el grafo múltiplemente conexo mostrado en la Figura 8.16, que ha sido creado añadiendo la arista $C \rightarrow F$ al poliárbol mostrado en la Figura 8.2. Esta arista produce un bucle en el grafo que contiene a los nodos A, C, D y F. En esta situación, ninguno de los nodos contenidos en el bucle separa el grafo en dos partes inconexas. Por ejemplo, se pueden asociar al nodo D los subgrafos $\{A, B, C\}$ y $\{E, F, G\}$, uno que contiene a sus padres y otro que contiene a los hijos, respectivamente. Sin embargo, estos subgrafos no están D-separados por el nodo D ya que la arista $C \rightarrow F$, contenida en el bucle, constituye una vía de comunicación alternativa entre los dos subgrafos.

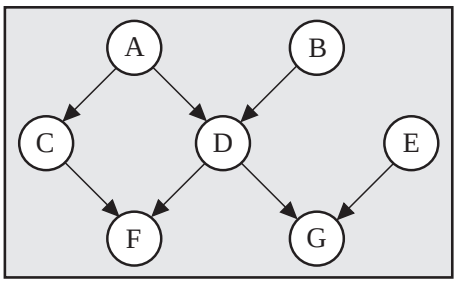


FIGURA 8.16. Grafo múltiplemente conexo.

La idea básica del algoritmo de condicionamiento es cortar estas vías alternativas de comunicación contenidas en los bucles asignando un valor arbitrario a un conjunto de nodos. Este conjunto de nodos se suele denominar *conjunto de corte* (en inglés, *cutset*). Por ejemplo, el nodo D no separa al grafo de la Figura 8.16 en dos partes inconexas, pero si se considera el conjunto de corte formado por el nodo C, entonces, el conjunto $\{C, D\}$ separa a $\{A, B\}$ de $\{E, F, G\}$, los subgrafos que contienen a los padres e hijos de D, respectivamente. Por tanto, se puede *cortar* el bucle

contenido en el grafo considerando el nodo C como un nodo evidencial, es decir, asignándole un valor arbitrario.

Esta idea de cortar los bucles para obtener un grafo de estructura más simple puede ser llevada a la práctica utilizando el método denominado *absorición de evidencia* (ver Shachter (1988, 1990a)). Este método muestra que la evidencia puede ser absorbida por el grafo cambiando su topología. De forma más precisa, si X_i es un nodo evidencial, se pueden eliminar del grafo todas las aristas de la forma $X_i \rightarrow X_j$ sustituyendo la función de probabilidad condicionada del nodo X_j, $p(x_j|\pi_j)$, por una función definida sobre un conjunto más reducido de variables:

$$p_1(x_j|\pi_j \setminus x_i) = p(x_j|\pi_j \setminus x_i, X_i = e_i).$$

Esta operación deja inalterado el modelo probabilístico, mientras que simplifica la topología del grafo al eliminar un conjunto de aristas. Obsérvese que el conjunto $\Pi_j \setminus X_i$ es el nuevo conjunto de padres del nodo X_j en el grafo modificado. Por ejemplo, si se asigna un valor arbitrario, $C = c$, al nodo C, es decir, si se convierte C en un nodo evidencial en el grafo de la Figura 8.16, entonces se puede absorber esta evidencia eliminando del grafo la arista $C \rightarrow F$, obteniendo así un nuevo grafo con estructura de poliárbol (ver Figura 8.17). Para mantener inalterada la función de probabilidad condicionada del conjunto de variables no evidenciales, $p(y|C = c)$, se reemplaza la función de probabilidad $p(f|c, d)$ por $p_1(f|d) = p(f|C = c, d)$, lo cual elimina la dependencia del nodo F respecto de la evidencia C.

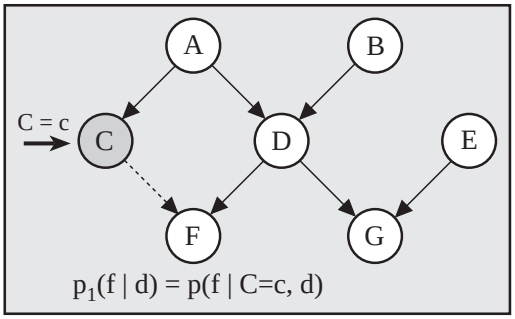


FIGURA 8.17. Absorción de la evidencia $C = c$ mediante la arista $C \rightarrow F$.

Por tanto, utilizando el método de absorción de evidencia se puede reducir un grafo múltiplemente conexo a un poliárbol, asignando un valor arbitrario a los nodos de un conjunto de corte $C = \{C_1, \ldots, C_m\}$. Entonces, se puede aplicar el algoritmo de propagación en poliárboles introducido en la Sección 8.3 para calcular las probabilidades $p(x_i|e, c_1, \ldots, c_m)$ para cada una de las realizaciones posibles $(c_1, \ldots, c_m)$ del conjunto de corte, dada la evidencia $E = e$. La función de probabilidad condicionada de cada nodo puede obtenerse combinando las distintas probabilidades obtenidas para

las distintas realizaciones del conjunto de corte:

$$p(x_i|e) = \sum_{c_1,\ldots,c_m} p(x_i|e, c_1, \ldots, c_m)p(c_1, \ldots, c_m|e). \tag{8.21}$$

La complejidad de este algoritmo radica en el hecho de que el número de realizaciones posibles de un conjunto de nodos crece de forma exponencial con el número de nodos. Por tanto, es conveniente obtener el conjunto de corte, de un grafo dado, con el mínimo número de nodos.

El problema de encontrar un conjunto de corte minimal es también un problema NP-complejo (Suermondt y Cooper (1990)), pero existen varios métodos heurísticos alternativos para obtener conjuntos de corte que, si bien no son minimales, se obtienen con un costo computacional razonable (ver, por ejemplo, Stillman (1991) y Suermondt y Cooper (1990)). Alguno de estos métodos proporciona una cota para el tamaño del conjunto de corte resultante. Por ejemplo, Becker y Geiger (1994) presentan un algoritmo para obtener un conjunto de corte que contenga menos del doble de las variables contenidas en un conjunto de corte minimal.

Obsérvese que los pesos $p(c_1, \ldots, c_m|e)$ dados en (8.21) no pueden ser calculados directamente en el poliárbol, pues en este caso no se está condicionando respecto del conjunto de corte. Pearl (1986a), Peot y Shachter (1991), y Suermondt y Cooper (1991b) introducen varios algoritmos para calcular estos pesos. Por ejemplo, el algoritmo de Suermondt y Cooper (1991b) descompone estos pesos como

$$p(c_1, \ldots, c_m|e) = \frac{p(e|c_1, \ldots, c_m)p(c_1, \ldots, c_m)}{p(e)},$$

que, al ser sustituidos en (8.21), dan

$$p(x_i|e) \propto \sum_{c_1,\ldots,c_m} p(x_i|e, c_1, \ldots, c_m)p(e|c_1, \ldots, c_m)p(c_1, \ldots, c_m). \tag{8.22}$$

Por tanto, las funciones de probabilidad de los nodos pueden ser calculadas a través de tres funciones distintas para cada combinación de valores de las variables contenidas en C. Como ya se ha mencionado anteriormente, la función $p(x_i|c, e)$ puede ser calculada utilizando el algoritmo de propagación en poliárboles. De forma similar, $p(e|c)$ puede ser calculada utilizando el mismo algoritmo, pero considerando la evidencia c. Por último, la función de probabilidad marginal del conjunto de corte, $p(c)$, puede calcularse asignando valores, secuencialmente, a los nodos de este conjunto, de tal forma que sólo sea necesaria una parte del grafo que tenga estructura de poliárbol para calcular la probabilidad marginal del subconjunto de nodos de corte asignados. Una descripción completa del proceso se tiene en Suermondt y Cooper (1991b).

El ejemplo siguiente ilustra el algoritmo de condicionamiento.

Ejemplo 8.4 Algoritmo de condicionamiento. Considérese el grafo múltiplemente conexo mostrado en la Figura 8.18, que implica la siguiente factorización de la función de probabilidad conjunta de las seis variables:

$$p(a,b,c,d,e,f) = p(a)p(b|a)p(c|a)p(d|b)p(e|b,c)p(f|c). \tag{8.23}$$

Los valores numéricos asociados al conjunto de funciones de probabilidad condicionada que componen esta factorización se muestran en la Tabla 8.2.

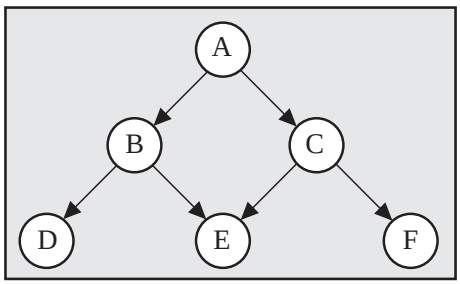


FIGURA 8.18. Grafo múltiplemente conexo.

Una forma de cortar el bucle $A-B-E-C-A$ es considerar el conjunto de corte formado por la variable A. La Figura 8.19 muestra dos opciones distintas para absorber la evidencia $A = a$, transformando el grafo original en un poliárbol. Por tanto, se puede aplicar el algoritmo de propagación en poliárboles al grafo resultante. Obsérvese que en ambos casos sólo se absorbe una de las dos aristas posibles para no transformar el grafo en inconexo.

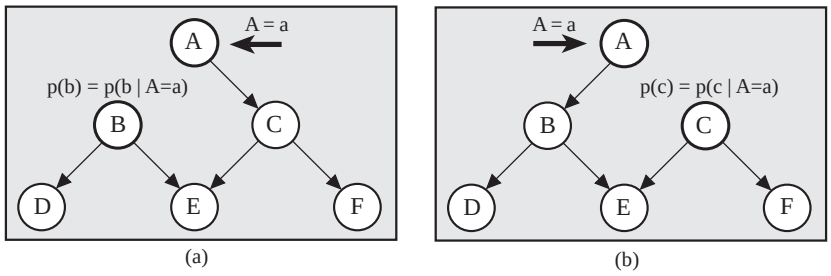


FIGURA 8.19. Dos opciones distintas de absorber la evidencia $A = a$.

En este ejemplo se considera la situación mostrada en la Figura 8.19(a). Para cada realización posible del conjunto de corte $\{A\}$, la nueva función de probabilidad asociada al poliárbol resultante se obtiene reemplazando $p(b|a)$ por $p_1(b) = p(b|A = a)$ en (8.23). En este caso se tiene

$$p(a,b,c,d,e,f|A=a) \propto p(a)p_1(b)p(c|a)p(d|b)p(e|b,c)p(f|c). \tag{8.24}$$

a	$p(a)$
0	0.3
1	0.7

a	b	$p(b\|a)$
0	0	0.4
0	1	0.6
1	0	0.1
1	1	0.9

a	c	$p(c\|a)$
0	0	0.2
0	1	0.8
1	0	0.5
1	1	0.5

b	d	$p(d\|b)$
0	0	0.3
0	1	0.7
1	0	0.2
1	1	0.8

c	f	$p(f\|c)$
0	0	0.1
0	1	0.9
1	0	0.4
1	1	0.6

b	c	e	$p(e\|b,c)$
0	0	0	0.4
0	0	1	0.6
0	1	0	0.5
0	1	1	0.5
1	0	0	0.7
1	0	1	0.3
1	1	0	0.2
1	1	1	0.8

TABLA 8.2. Valores numéricos de las funciones de probabilidad condicionada que forman la factorización (8.23).

El algoritmo de condicionamiento se ilustra en dos situaciones distintas. Primero se analiza el caso en el que no se dispone de evidencia, después se considera la evidencia $\{C = 1, D = 1\}$.

Para el caso en el que no se tiene evidencia, la ecuación (8.21) se reduce a

$$p(x_i) = \sum_{c_1,\ldots,c_m} p(x_i|c_1,\ldots,c_m)p(c_1,\ldots,c_m).$$

Por tanto, se tiene

$$p(x_i) = \sum_{a} p(x_i|a)p(a),$$

para todos los nodos X_i que no están contenidos en el conjunto de corte. Obsérvese que $p(a)$ es la función de probabilidad marginal del nodo A, que

puede ser calculada directamente de la Tabla 8.2:

$$(p(A=0), p(A=1)) = (0.3, 0.7).$$

Por tanto, sólo es necesario calcular $p(x_i|a)$ para los dos valores posibles de A. Estas probabilidades pueden obtenerse aplicando el algoritmo de poliárboles al grafo de la Figura 8.19(a) con las funciones de probabilidad condicionada dadas en (8.24). Las Figuras 8.20(a) y (b) muestran las probabilidades asociadas a las realizaciones $A = 0$ y $A = 1$, respectivamente. Utilizando los valores numéricos mostrados en estas figuras, se puede calcular la función de probabilidad marginal de los nodos de la red Bayesiana.

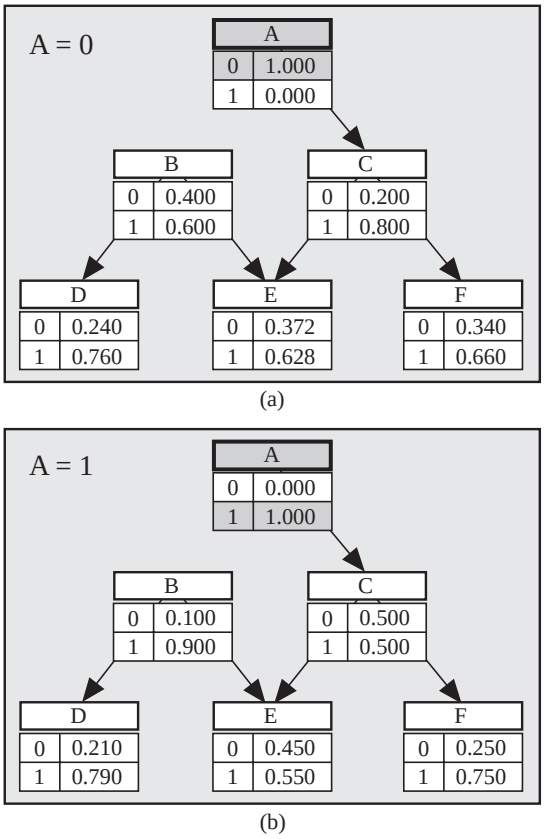


FIGURA 8.20. Probabilidades obtenidas aplicando el algoritmo de propagación en poliárboles para los dos valores posibles de A.

Por ejemplo, para el nodo B se tiene

$$p(B=0) \quad = \quad \sum_{a=0}^{1} p(B=0|a)p(a),$$

$$= \quad 0.4 \times 0.3 + 0.1 \times 0.7 = 0.19,$$

$$\begin{aligned} p(B=1) &= \sum_{a=0}^{1} p(B=1|a)p(a), \\ &= 0.6 \times 0.3 + 0.9 \times 0.7 = 0.81. \end{aligned}$$

Las probabilidades del resto de los nodos pueden ser calculadas de forma similar. La Figura 8.21 muestra los valores numéricos correspondientes a las funciones de probabilidad marginal de todos los nodos de la red.

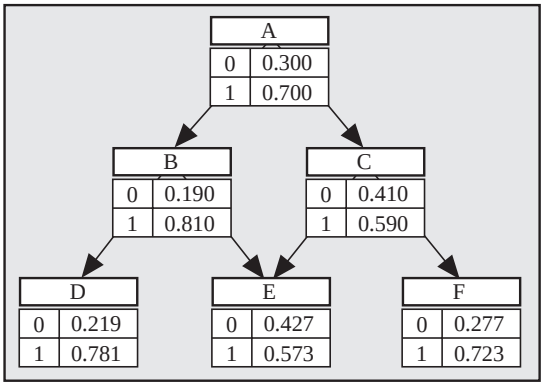


FIGURA 8.21. Funciones de probabilidad marginal de los nodos.

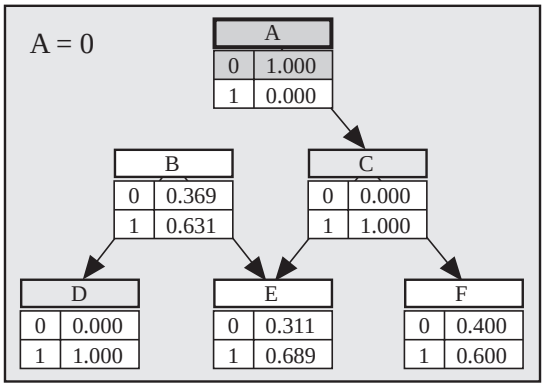
Supóngase ahora que se tiene la evidencia $\{C = 1, D = 1\}$. Entonces, aplicando (8.22) resulta

$$p(x_i|C=1, D=1) \propto \sum_{a=0}^{1} p(x_i|a, C=1, D=1)p(C=1, D=1|a)p(a). \tag{8.25}$$

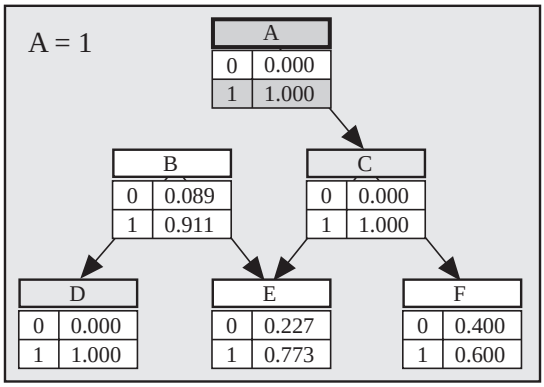
De la misma forma que en el caso anterior, $p(a)$ se obtiene directamente de la Tabla 8.2. Por tanto, sólo es necesario calcular $p(C = 1, D = 1|a)$ y $p(x_i|a, C = 1, D = 1)$ para los dos valores posibles del nodo A. Por otra parte, las probabilidades $p(x_i|a, C = 1, D = 1)$ pueden ser obtenidas aplicando el algoritmo de propagación en poliárboles considerando la evidencia $\{A = a, C = 1, D = 1\}$. La Figura 8.22 muestra los valores numéricos asociados a los dos valores posibles de A.

Por otra parte, la función de probabilidad $p(C = 1, D = 1|a)$ no puede obtenerse directamente aplicando el algoritmo de propagación en poliárboles pues no es una función de un sólo nodo. Sin embargo, esta función de probabilidad puede descomponerse aplicando la regla de la cadena:

$$p(C=1, D=1|a) \quad = \quad \frac{p(C=1, D=1, a)}{p(a)}$$



(a)



(b)

FIGURA 8.22. Probabilidades obtenidas aplicando el algoritmo de propagación en poliárboles considerando la evidencia $\{A = a, C = 1, D = 1\}$ para los dos valores posibles de la variable A.

$$\begin{aligned} &= \frac{p(C = 1|D = 1, a)p(D = 1|a)p(a)}{p(a)} \\ &= p(C = 1|D = 1, a)p(D = 1|a). \end{aligned}$$

Las probabilidades $p(D = 1|a)$ están contenidas en la Tabla 8.20 y $p(C = 1|D = 1, a)$ puede ser calculada de forma simultánea con $p(x_i|a, C = 1, D = 1)$ considerando secuencialmente los distintos elementos que componen la evidencia. Los valores numéricos de estas probabilidades son

$$(p(C = 1|D = 1, A = 0), p(C = 1|D = 1, A = 1)) = (0.8, 0.5).$$

Las probabilidades $p(C = 1, D = 1|a)$ pueden obtenerse a partir de los valores anteriores

$$\begin{aligned} p(C = 1, D = 1|A = 0) &= 0.8 \times 0.760 = 0.608, \\ p(C = 1, D = 1|A = 1) &= 0.5 \times 0.790 = 0.395. \end{aligned}$$

Por tanto, se pueden calcular las funciones de probabilidad condicionada de los nodos reemplazando los valores obtenidos en (8.25). Por ejemplo, para el nodo B se tiene

$$\begin{aligned} & p(B=0|C=1,D=1) \\ & \quad \propto \sum_{a=0}^{1} p(B=0|a,C=1,D=1)p(C=1,D=1|a)p(a), \\ & \quad = 0.369 \times 0.608 \times 0.3 + 0.089 \times 0.395 \times 0.7 = 0.092, \end{aligned}$$

$$\begin{aligned} & p(B=1|C=1,D=1) \\ & \quad \propto \sum_{a=0}^{1} p(B=1|a,C=1,D=1)p(C=1,D=1|a)p(a), \\ & \quad = 0.631 \times 0.608 \times 0.3 + 0.911 \times 0.395 \times 0.7 = 0.367. \end{aligned}$$

Finalmente, normalizando las funciones anteriores (dividiendo por $0.092 + 0.367 = 0.459$) se obtiene:

$$p(B=0|C=1,D=1) \quad = \quad 0.092/0.459 = 0.200,$$

$$p(B=1|C=1,D=1) \quad = \quad 0.367/0.459 = 0.800.$$

La Figura 8.23 muestra los valores numéricos resultantes para las funciones de probabilidad condicionada de los nodos de la red. ■

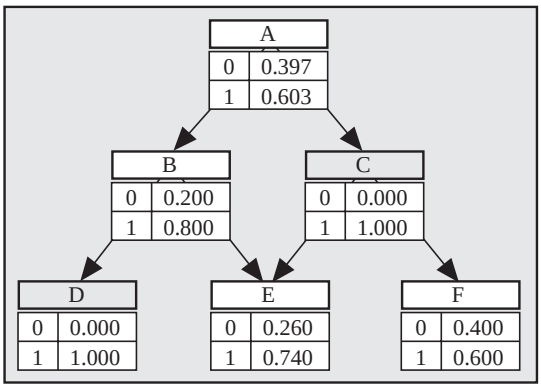


FIGURA 8.23. Probabilidades condicionadas de los nodos, dada la evidencia $\{C=1, D=1\}$.

Anteriormente se vio que la complejidad de este algoritmo reside en la aplicación múltiple del algoritmo de propagación en poliárboles para las distintas realizaciones del conjunto de corte. Sin embargo, estos procesos de propagación pueden involucrar cálculos repetitivos y redundantes. Para evitar este problema se han presentado algunas modificaciones de este

método como, por ejemplo, el *algoritmo de condicionamiento local* (Díez (1996)) y el *algoritmo de condicionamiento dinámico* (Darwiche (1995)), que aprovechan la estructura local del grafo para evitar cálculos redundantes. Esto conlleva una mejora notable de la complejidad del algoritmo y un importante ahorro en el tiempo de computación.

8.6 Métodos de Agrupamiento

El algoritmo de propagación en poliárboles y el algoritmo de condicionamiento introducidos en las secciones anteriores aprovechan la estructura particular de los grafos dirigidos para propagar la evidencia. Por tanto, estos algoritmos son sólo aplicables a redes Bayesianas. En esta sección se presenta un método de propagación distinto, el método de agrupamiento que, a partir de las estructuras locales contenidas en el grafo, produce representaciones alternativas para propagar la evidencia. Por tanto, estos métodos no dependen del tipo de grafo y son aplicables tanto a redes de Markov, como a redes Bayesianas.

El método de agrupamiento, inicialmente desarrollado por Lauritzen y Spiegelhalter (1988), se basa en la construcción de subconjuntos de nodos (aglomerados) que capturen las estructuras locales del modelo probabilístico asociado al grafo. De esta forma, el proceso de propagación de evidencia puede ser realizado calculando probabilidades locales (que dependen de un número reducido de variables), evitando así calcular probabilidades globales (que dependen de todas las variables). Como ya se vio en el Capítulo 4, los conglomerados de un grafo son los subconjuntos que representan sus estructuras locales. Por tanto, en primer lugar, el algoritmo de agrupamiento calcula los conglomerados del grafo; a continuación obtiene las funciones de probabilidad condicionada de cada conglomerado calculando de forma iterativa varias funciones de probabilidad locales. Por último, se obtiene la función de probabilidad condicionada de cada nodo marginalizando la función de probabilidad de cualquier conglomerado en el que esté contenido. En esta sección se presentan dos versiones de este algoritmo, una para redes de Markov y otra para redes Bayesianas.

8.6.1 Métodos de Agrupamiento en Redes de Markov

En el Capítulo 6 se analizaron dos representaciones alternativas del modelo probabilístico de una red de Markov. La representación básica de estos modelos es la *representación potencial*, dada por un conjunto de funciones positivas $\Psi = \{\psi(c_1), \ldots, \psi(c_m)\}$ definidas en los conglomerados del grafo $C = \{C_1, \ldots, C_m\}$ que permiten factorizar la función de probabilidad del

modelo como

$$p(x) = \frac{1}{k} \prod_{i=1}^{m} \psi_i(c_i), \tag{8.26}$$

donde $k = \sum_x \prod_{i=1}^{m} \psi_i(c_i)$ es una constante de normalización.

Si la red de Markov es descomponible, es decir, si el grafo asociado es triangulado, entonces, puede obtenerse una representación alternativa del modelo probabilístico considerando una *cadena de conglomerados* $(C_1, \ldots, C_m)$ que cumpla la propiedad de intersección dinámica (ver Sección 4.5 para más detalles). Esta cadena de conglomerados proporciona una factorización de la función de probabilidad como producto de funciones de probabilidad condicionada:

$$p(x) = \prod_{i=1}^{m} p(r_i|s_i), \tag{8.27}$$

donde $S_i = C_i \cap (C_1, \ldots, C_{i-1})$ es el conjunto *separador* del conglomerado C_i y $R_i = C_i \setminus S_i$ es el conjunto *residual*. La propiedad de intersección dinámica garantiza que S_i está contenido en alguno de los conglomerados anteriores, $C_1, \ldots, C_{i-1}$. Los conglomerados que contienen al separador S_i se denominan los *vecinos* de C_i y se denotan por B_i. Como se verá más adelante, la representación de la función de probabilidad dada por una cadena de conglomerados proporciona un algoritmo sencillo, utilizado por el método de agrupamiento, para calcular las funciones de probabilidad de los conglomerados y, por tanto, las funciones de probabilidad de los nodos.

Si una red de Markov no es descomponible, entonces el proceso de triangulación permite obtener un grafo no dirigido triangulado auxiliar cuyos conglomerados (estructuras locales) contendrán a los conglomerados del grafo original. Por tanto, la representación potencial de este nuevo grafo podrá ser obtenida a partir de la representación potencial del grafo original, lo que permitirá realizar el proceso de propagación en esta nueva red descomponible. Por tanto, se supondrá, sin pérdida de generalidad, que se tiene una red de Markov descomponible con la representación potencial (C, Ψ).

El algoritmo de agrupamiento se basa en los siguientes pasos:

1. Obtener una factorización de la función de probabilidad en la forma dada en (8.27).
2. Calcular las funciones de probabilidad de los conglomerados a partir de las funciones de probabilidad contenidas en (8.27).
3. Calcular las probabilidades de los nodos.

Etapa 1: En esta etapa es necesario calcular una cadena de conglomerados que cumpla la propiedad de intersección dinámica. Para ello se puede utilizar el Algoritmo 4.3. Entonces, las funciones de probabilidad condicionada $p(r_i|s_i)$ asociadas al conglomerado C_i se pueden obtener de forma iterativa

marginalizando la función de probabilidad $p(x)$, primero sobre C_m, después sobre C_{m-1}, y así sucesivamente. Estas funciones de probabilidad marginal se obtienen a partir de la representación potencial de la forma siguiente. Para el último conglomerado de la cadena, C_m, se tiene

$$\begin{aligned} p(c_1, \ldots, c_{m-1}) &= \sum_{c_m \setminus \{c_1, \ldots, c_{m-1}\}} p(x) \\ &= \sum_{r_m} k^{-1} \prod_{i=1}^{m} \psi_i(c_i) \\ &= k^{-1} \prod_{i=1}^{m-1} \psi_i(c_i) \sum_{r_m} \psi_m(c_m). \end{aligned} \tag{8.28}$$

Por tanto, marginalizar $p(c_1, \ldots, c_m)$ sobre C_m es básicamente lo mismo que marginalizar la función potencial asociada $\psi_m(c_m)$. Una vez que se ha obtenido $p(c_1, \ldots, c_{m-1})$, se puede aplicar de nuevo la misma idea para obtener $p(c_1, \ldots, c_{m-2})$, y así sucesivamente, hasta obtener $p(c_1)$. Obsérvese que el término $\sum_{r_m} \psi_m(c_m)$ en (8.28) depende solamente de las variables contenidas en S_m. Por tanto, se puede incluir este término en la función potencial de cualquier conglomerado C_j que contenga a S_m, es decir, en cualquier vecino del conglomerado C_m. De esta forma se tiene la nueva función potencial

$$\psi_j^*(c_j) = \psi_j(c_j) \sum_{r_m} \psi_m(c_m). \tag{8.29}$$

Para el resto de los conglomerados $\psi_k^*(c_k) = \psi_k(c_k)$, $k \neq j$. Por tanto, considerando estas nuevas funciones potenciales y (8.28), se tiene

$$p(c_1, \ldots, c_{m-1}) = k^{-1} \prod_{i=1}^{m-1} \psi_i^*(c_i). \tag{8.30}$$

Obsérvese que $\{\psi_1^*(c_1), \ldots, \psi_{m-1}^*(c_{m-1})\}$ es una representación potencial de la función de probabilidad $p(c_1, \ldots, c_{m-1})$.

Después de calcular la función $p(c_1, \ldots, c_{m-1})$, se puede utilizar de nuevo la ecuación (8.28) para calcular la función $p(c_1, \ldots, c_{m-2})$. Este proceso iterativo de marginalización de la función de probabilidad permite obtener las funciones de probabilidad condicionada que forman la factorización (8.27). Procediendo de nuevo de forma iterativa comenzando con el último término de la factorización, $p(r_m|s_m)$, y teniendo en cuenta que S_m separa R_m de $\{C_1, C_2, \ldots, C_{m-1}\}$, se tiene

$$\begin{aligned} p(r_m|s_m) &= p(r_m|c_1, c_2, \ldots, c_{m-1}) \\ &= \frac{p(c_1, c_2, \ldots, c_{m-1}, r_m)}{p(c_1, c_2, \ldots, c_{m-1})}. \end{aligned}$$

Aplicando la igualdad $\{C_1, \ldots, C_{m-1}, R_m\} = \{C_1, \ldots, C_{m-1}, C_m\} = X$, y utilizando la ecuación (8.28), se tiene

$$\begin{aligned} p(r_m|s_m) &= \frac{k^{-1} \prod_{i=1}^{m} \psi_i(c_i)}{k^{-1} \prod_{i=1}^{m-1} \psi_i(c_i) \sum_{r_q} \psi_m(c_m)} \\ &= \frac{\psi_m(C_m)}{\sum_{r_m} \psi_m(c_m)}. \end{aligned} \tag{8.31}$$

De esta forma se obtiene la función de probabilidad condicionada $p(r_m|s_m)$. Considerando ahora la función de probabilidad $p(c_1, \ldots, c_{m-1})$ dada en (8.30), y aplicando (8.28), se tiene de nuevo la función de probabilidad $p(c_1, \ldots, c_{m-2})$ marginalizando la función anterior en C_{m-1}. Entonces, aplicando (8.31) se obtiene $p(r_{m-1}|s_{m-1})$. Este mismo proceso puede repetirse hasta obtener el primer término de la factorización, $p(r_1|s_1)$.

Etapa 2: Una vez que se ha obtenido la representación del modelo probabilístico dada por la cadena de conglomerados, se pueden calcular las funciones de probabilidad de los conglomerados a partir de las funciones de probabilidad de los separadores correspondientes. En primer lugar, dado que $S_1 = \phi$, la función de probabilidad del primer conglomerado se obtiene directamente de $p(r_1|s_1) = p(c_1)$. El conjunto separador del conglomerado C_2 está contenido en C_1 (dado que $(C_1, \ldots, C_m)$ cumple la propiedad de intersección dinámica). Por tanto, $p(s_2)$ puede obtenerse marginalizando la función $p(c_1)$. Entonces,

$$p(c_2) = p(r_2, s_2) = p(r_2|s_2)p(s_2).$$

Una vez que $p(c_1)$ y $p(c_2)$ han sido calculados, se puede aplicar el mismo proceso de forma iterativa hasta obtener las funciones de probabilidad del resto de los conglomerados.

Etapa 3: Finalmente, una vez que han sido calculadas las funciones de probabilidad de los conglomerados, se puede obtener la función de probabilidad marginal de cualquier nodo X_i marginalizando la función de probabilidad de algún conglomerado que lo contenga. Si C_j contiene al nodo X_i, entonces se puede obtener la función de probabilidad del nodo mediante

$$p(x_i) = \sum_{c_j \setminus x_i} p(c_j). \tag{8.32}$$

Si el nodo X_i está contenido en más de un conglomerado, entonces es conveniente elegir aquél que tenga menor tamaño; de esta forma, el número de operaciones realizadas en el proceso de marginalización será mínimo.

Obsérvese que el tamaño de un conglomerado es el producto del número de valores posibles (la cardinalidad) de cada uno de sus nodos.

En el análisis anterior se ha considerado el caso de no disponer de evidencia. Sin embargo, si se conoce la evidencia $E = e$, donde E es un conjunto de variables, se puede aplicar el mismo método considerando las modificaciones siguientes. La función de probabilidad condicionada $p(x \setminus e|e) = p(x \setminus e, e)/p(e)$ es proporcional a $p(x \setminus e, e)$, que puede ser obtenida modificando las funciones potenciales originales, sustituyendo las variables contenidas en E por los valores observados de estas variables e. Este proceso se denomina *absorción de evidencia* y se puede llevar a cabo de dos formas alternativas:

1. Manteniendo el mismo conjunto de nodos X y conglomerados C. En este caso, sólo es necesario modificar las funciones potenciales que contengan nodos evidenciales de la forma siguiente. Para cada conglomerado C_i que contenga algún nodo evidencial, se define $\psi_i^*(c_i)$ como

$$\psi_i^*(c_i) = \begin{cases} 0, & \text{si algún valor de } c_i \text{ no es consistente con } e, \\ \psi_i(c_i), & \text{en otro caso.} \end{cases} \tag{8.33}$$

   Para el resto de los conglomerados no es necesario ningún cambio. Por tanto, se tiene

$$p(x|e) \propto \prod_{i=1}^{m} \psi_i^*(c_i).$$

2. Eliminar de X los nodos evidenciales. Este proceso también implica modificar el conjunto de conglomerados y la representación potencial. La nueva representación potencial, (C^*, Ψ^*), está definida en X^*, donde $X^* = X \setminus E$, C^* es el nuevo conjunto de conglomerados y Ψ^* son los nuevos potenciales, que contienen la evidencia, y que han sido obtenidos de la forma siguiente: Para cada conglomerado C_i contenido en C tal que $C_i \cap E \neq \phi$, se incluye el conjunto $C_i \setminus E$ en C^* y se define

$$\psi_i^*(c_i^*) = \psi_i(c_i \setminus e,\ E = e). \tag{8.34}$$

   Para el resto de los conglomerados que no contienen nodos evidenciales, no es necesario realizar ninguna modificación en las representaciones potenciales correspondientes. Con ello, se tiene

$$p(x^*|e) \propto \prod_{i=1}^{m} \psi_i^*(c_i).$$

Por tanto, en ambos casos, se puede aplicar el método anterior para obtener la función de probabilidad condicionada de los nodos, dada la evidencia $E = e$. En el primer caso se continúa con la misma estructura, pero se

utilizan más recursos de los necesarios. En el segundo caso, no se utilizan más recursos de los necesarios, pero se necesita modificar la estructura. Por tanto, se requiere un consenso entre ambas opciones con objeto de elegir la más adecuada en cada caso.

Algoritmo 8.2 Algoritmo de agrupamiento para redes de Markov descomponibles.

- **Datos:** Una red de Markov descomponible (C, Ψ) sobre un conjunto de variables X y una evidencia $E = e$.
- **Resultados:** Las funciones de probabilidad condicionada $p(x_i|e)$ para cada nodo $X_i \notin E$.

Etapa de Iniciación:

1. Absorber la evidencia $E = e$ en las funciones potenciales Ψ utilizando (8.33) ó (8.34).
2. Utilizar el Algoritmo 4.3 para obtener una cadena de conglomerados $(C_1, \ldots, C_m)$ que cumplan la propiedad de intersección dinámica.
3. Para cada conglomerado C_i, elegir como vecino cualquier otro conglomerado C_j, con $j < i$, tal que $S_i \subset C_j$.

Etapa de Iteración:

4. Para $i = m$ hasta 1 hacer
   (a) Calcular $m_i(s_i) = \sum_{r_i} \psi_i(c_i)$.
   (b) Asignar $p(r_i|s_i) = \psi_i(c_i)/m_i(s_i)$.
   (c) Reemplazar la función potencial $\psi_j(c_j)$ del conglomerado, C_j, vecino de C_i por $\psi_j(c_j) \leftarrow \psi_j(c_j)m_i(s_i)$.
5. Asignar $p(c_1) = p(r_1|s_1) = p(r_1)$.
6. Para $i = 2$ hasta m hacer
   (a) Calcular $p(s_i)$ marginalizando la función de probabilidad $p(c_j)$ del conglomerado, C_j, vecino de C_i.
   (b) Asignar $p(c_i) = p(r_i|s_i)p(s_i)$.
7. Para $i = 1$ hasta n hacer
   (a) Elegir el conglomerado de menor tamaño C_j que contenga al nodo X_i.
   (b) Asignar $p(x_i|e) \propto \sum_{c_j \setminus x_i} p(c_j)$.

(c) Normalizar los valores obtenidos. ∎

Obsérvese que este algoritmo puede ser utilizado para calcular no sólo las funciones de probabilidad condicionada de los nodos, sino también la función de probabilidad condicionada de cualquier subconjunto de nodos que esté contenido en algún conglomerado del grafo. Xu (1995) presenta una adaptación de este método para calcular la función de probabilidad de cualquier subconjunto de nodos, incluso si este conjunto no está contenido en ningún conglomerado. Este método modifica la representación potencial añadiendo a la cadena de conglomerados el conjunto cuya probabilidad se desea obtener. De esta forma, la función de probabilidad de este conjunto se obtiene conjuntamente con las probabilidades de los conglomerados en el proceso de propagación (una descripción completa del método se muestra en Xu (1995)).

Ejemplo 8.5 Algoritmo de Agrupamiento en Redes de Markov Descomponibles. Considérese el grafo no dirigido triangulado mostrado en la Figura 8.24(a). Este grafo define una red de Markov, cuyos conglomerados son

$$C_1 = \{A, B, C\}, \;\; C_2 = \{B, C, E\}, \;\; C_3 = \{B, D\} \; y \; C_4 = \{C, F\}.$$

La Figura 8.24(b) muestra este conjunto de conglomerados, que implican la siguiente representación potencial de la red de Markov:

$$p(a, b, c, d, e, f) = \psi_1(a, b, c)\psi_2(b, c, e)\psi_3(b, d)\psi_4(c, f). \tag{8.35}$$

Los valores numéricos correspondientes a estas funciones se muestran en la Tabla 8.3.

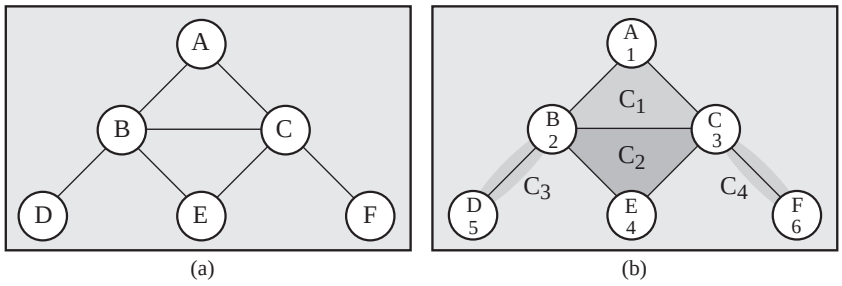


FIGURA 8.24. Un grafo triangulado no dirigido (a) y sus conglomerados (b).

Supóngase que no se dispone de evidencia y que se quieren obtener las funciones de probabilidad marginal de los nodos. En la primera etapa del algoritmo de agrupamiento es necesario calcular una cadena de conglomerados, es decir, es necesario ordenar los conglomerados de forma que cumplan la propiedad de intersección dinámica. Para ello, se puede aplicar el Algoritmo 4.3 (ver Capítulo 4). Este algoritmo calcula una numeración perfecta

a	b	c	$\psi_1(a,b,c)$
0	0	0	0.048
0	0	1	0.192
0	1	0	0.072
0	1	1	0.288
1	0	0	0.070
1	0	1	0.070
1	1	0	0.630
1	1	1	0.630

b	c	e	$\psi_2(b,c,e)$
0	0	0	0.08
0	0	1	0.12
0	1	0	0.10
0	1	1	0.10
1	0	0	0.14
1	0	1	0.06
1	1	0	0.04
1	1	1	0.16

b	d	$\psi_3(b,d)$
0	0	0.06
0	1	0.14
1	0	0.04
1	1	0.16

c	f	$\psi_4(c,f)$
0	0	0.02
0	1	0.18
1	0	0.08
1	1	0.12

TABLA 8.3. Valores numéricos de las funciones potenciales en (8.35).

Conglomerado	Separador S_i	Residuo R_i	Vecinos B_i
$C_1 = \{A, B, C\}$	ϕ	A, B, C	$-$
$C_2 = \{B, C, E\}$	B, C	E	C_1
$C_3 = \{B, D\}$	B	D	$\{C_1, C_2\}$
$C_4 = \{C, F\}$	C	F	$\{C_1, C_2\}$

TABLA 8.4. Separadores, residuos y vecinos de los conglomerados.

de los nodos y, después, ordena los conglomerados según el mayor número perfecto contenido en cada uno de ellos. Una numeración perfecta de los nodos del grafo de la Figura 8.24(a), obtenida aplicando el algoritmo de búsqueda de máxima cardinalidad (Algoritmo 4.1), se muestra junto a los nombres de los nodos en la Figura 8.24(b). Obsérvese que la ordenación natural de los conglomerados (C_1, C_2, C_3, C_4) cumple la propiedad de intersección dinámica. La Tabla 8.4 muestra los conjuntos separadores, los residuales y los de los vecinos asociados a cada uno de los conglomerados. A partir de esta tabla se deduce que el conjunto de funciones de probabilidad condicionada que forman la factorización (8.27) es

$$\begin{aligned} p(a,b,c,d,e,f) &= p(r_1|s_1)p(r_2|s_2)p(r_3|s_3)p(r_4|s_4) \\ &= p(a,b,c)p(e|b,c)p(d|b)p(f|c). \end{aligned} \tag{8.36}$$

Ahora se está en disposición de comenzar la etapa iterativa del Algoritmo 8.2. El Paso 4 va tomando los conglomerados sucesivamente, empezando por el último. En este caso $m = 4$, entonces se tiene

c	f	$p(f\|c)$
0	0	0.1
0	1	0.9
1	0	0.4
1	1	0.6

b	d	$p(d\|b)$
0	0	0.3
0	1	0.7
1	0	0.2
1	1	0.8

b	c	e	$p(e\|b,c)$
0	0	0	0.4
0	0	1	0.6
0	1	0	0.5
0	1	1	0.5
1	0	0	0.7
1	0	1	0.3
1	1	0	0.2
1	1	1	0.8

a	b	c	$p(a,b,c)$
0	0	0	0.024
0	0	1	0.096
0	1	0	0.036
0	1	1	0.144
1	0	0	0.035
1	0	1	0.035
1	1	0	0.315
1	1	1	0.315

TABLA 8.5. Funciones de probabilidad condicionada de los residuos, dados los separadores, $p(r_i|s_i)$.

- Para $i = 4$, la función de probabilidad correspondiente en la factorización (8.36) es $p(f|c)$. En primer lugar se calcula el término

$$m_4(s_4) = \sum_{r_4} \psi_4(c_4) = \sum_{f} \psi_4(c, f)$$

obteniéndose $(m_4(C = 0), m_4(C = 1)) = (0.2, 0.2)$. A continuación, utilizando (8.31), se calcula la función de probabilidad condicionada aplicando

$$p(f|c) = \frac{\psi_4(c, f)}{m_4(c)},$$

es decir,

$$\begin{aligned} p(F = 0|C = 0) &= 0.02/0.2 = 0.1, \\ p(F = 0|C = 1) &= 0.18/0.2 = 0.9, \\ p(F = 1|C = 0) &= 0.08/0.2 = 0.4, \\ p(F = 1|C = 1) &= 0.12/0.2 = 0.6. \end{aligned}$$

Esta función de probabilidad se muestra en la Tabla 8.5. Finalmente, se elige un conglomerado vecino de C_4, por ejemplo C_2, y se multiplica la función potencial $\psi_2(b, c, e)$ por $m_4(s_4)$, con lo que resulta la función $\psi_2^*(b, c, e)$, cuyos valores numéricos se muestran en la Tabla 8.6.

b	c	e	$\psi_2^*(b,c,e)$
0	0	0	0.016
0	0	1	0.024
0	1	0	0.020
0	1	1	0.020
1	0	0	0.028
1	0	1	0.012
1	1	0	0.008
1	1	1	0.032

a	b	c	$\psi_1^*(a,b,c)$
0	0	0	0.0096
0	0	1	0.0384
0	1	0	0.0144
0	1	1	0.0576
1	0	0	0.0140
1	0	1	0.0140
1	1	0	0.1260
1	1	1	0.1260

TABLA 8.6. Nuevas funciones potenciales de los conglomerados C_2 y C_1.

- Para $i = 3$, la función de probabilidad condicionada correspondiente en (8.36) es $p(d|b)$. En primer lugar se calcula el término

$$m_3(s_3) = \sum_{r_3} \psi_3(c_3) = \sum_{d} \psi_3(b,d),$$

obteniéndose $(m_3(B=0), m_3(B=1)) = (0.2, 0.2)$. A continuación se calcula la función de probabilidad

$$p(d|b) = \frac{\psi_3(b,d)}{m_3(b)},$$

cuyos valores numéricos son

$$\begin{aligned} p(D=0|B=0) &= 0.06/0.2 = 0.3, \\ p(D=0|B=1) &= 0.14/0.2 = 0.7, \\ p(D=1|B=0) &= 0.04/0.2 = 0.2, \\ p(D=1|B=1) &= 0.16/0.2 = 0.8. \end{aligned}$$

Esta función de probabilidad también se muestra en la Tabla 8.5.

Finalmente, se elige un conglomerado vecino de C_3, por ejemplo C_1, y se multiplica la función potencial $\psi_1(a,b,c)$ de C_1 por $m_3(s_3)$, con lo que se tiene la nueva función $\psi_1^*(a,b,c)$ cuyos valores numéricos se muestran en la Tabla 8.6.

- Para $i = 2$, la función de probabilidad condicionada correspondiente en (8.36) es $p(e|b,c)$. Procediendo de forma similar se calcula $m_2(b,c) = \sum_e \psi_2^*(b,c,e) = 0.04$, para todos los valores de b y c, y la función de probabilidad

$$p(e|b,c) = \frac{\psi_2^*(b,c,e)}{m_2(b,c)},$$

cuyos valores numéricos se muestran en la Tabla 8.5.

a	b	c	$\psi_1^*(a,b,c)$
0	0	0	0.000384
0	0	1	0.001536
0	1	0	0.000576
0	1	1	0.002304
1	0	0	0.000560
1	0	1	0.000560
1	1	0	0.005040
1	1	1	0.005040

TABLA 8.7. Función potencial del conglomerado C_1 en el último paso de la etapa de la marginalización.

Finalmente, dado que C_1 es el único conglomerado vecino de C_2, se multiplica por $m_2(s_2)$ la función potencial $\psi_1^*(a,b,c)$ de C_1 de la Tabla 8.6. En este caso se obtiene la nueva función $\psi_1^*(a,b,c)$ de la Tabla 8.7. La función de probabilidad del último conglomerado se obtiene directamente de la función potencial: $p(c_1) \propto \psi_1^*(a,b,c)$.

- Para $i = 1$, la función de probabilidad correspondiente en (8.36) es $p(a,b,c)$. Dado que $S_1 = \phi$, $R_1 = \{A,B,C\}$ y $B_1 = \phi$, se tiene que

$$m_1(\phi) = \sum_{a,b,c} \psi_1^*(a,b,c) = 0.016$$

y

$$p(r_1|s_1) = p(a,b,c) = \psi_1^*(a,b,c)/0.016.$$

Esta función se muestra en la Tabla 8.5. Por tanto, finaliza el Paso 4 de la etapa de iteración.

- En el quinto paso de la etapa de iteración, se tiene $p(c_1) = p(r_1) = p(a,b,c)$, que se puede obtener directamente de la Tabla 8.8. Seguidamente se comienza el Paso 6 considerando $i = 2$.

- Para $i = 2$, se tiene $S_2 = \{B,C\}$. Para calcular $p(b,c)$, se marginaliza la función de probabilidad $p(a,b,c)$ sobre A, obteniéndose

$$p(b,c) = \sum_a p(a,b,c),$$

que implica

$$\begin{aligned} p(B=0,C=0) &= 0.059, \\ p(B=0,C=1) &= 0.131, \\ p(B=1,C=0) &= 0.351, \\ p(B=1,C=1) &= 0.459. \end{aligned}$$

a	b	c	$p(a,b,c)$
0	0	0	0.024
0	0	1	0.096
0	1	0	0.036
0	1	1	0.144
1	0	0	0.035
1	0	1	0.035
1	1	0	0.315
1	1	1	0.315

b	c	e	$p(b,c,e)$
0	0	0	0.0236
0	0	1	0.0354
0	1	0	0.0655
0	1	1	0.0655
1	0	0	0.2457
1	0	1	0.1053
1	1	0	0.0918
1	1	1	0.3672

b	d	$p(b,d)$
0	0	0.057
0	1	0.133
1	0	0.162
1	1	0.648

c	f	$p(c,f)$
0	0	0.041
0	1	0.369
1	0	0.236
1	1	0.354

TABLA 8.8. Valores numéricos de la funciones de probabilidad de los conglomerados.

Por tanto,

$$p(b,c,e) = p(e|b,c)p(b,c).$$

Los valores numéricos asociados se muestran en la Tabla 8.8.

- Para $i = 3$, se tiene $S_3 = \{B\}$. En este caso $p(b)$ puede obtenerse marginalizando $p(a,b,c)$ sobre A y C, o $p(b,c,e)$ sobre C y E. En ambos casos se obtienen los valores numéricos $p(B = 0) = 0.19$, $p(B = 1) = 0.81$. Entonces, las probabilidades del conglomerado C_3 resultan

  $$p(b,d) = p(d|b)p(b),$$

  cuyos valores numéricos se muestran en la Tabla 8.8.

- Por último, para $i = 4$, $S_4 = \{C\}$. Para calcular $p(c)$, se marginaliza, o bien $p(a,b,c)$ sobre A y B, o bien $p(b,c,e)$ sobre B y E. En este caso se obtiene $p(C = 0) = 0.41$, $p(C = 1) = 0.59$. Entonces, la función de probabilidad del conglomerado C_4 se obtiene como

  $$p(c,f) = p(f|c)p(c).$$

  Los valores numéricos asociados se muestran en la Tabla 8.8.

En la etapa final del algoritmo de agrupamiento se calculan las probabilidades marginales de los nodos a partir de las funciones de probabilidad mostradas en la Tabla 8.8. El nodo A está únicamente contenido en el conglomerado C_1; por tanto, se calcula $p(a)$ marginalizando la función $p(a,b,c)$

a	b	c	$\psi_1^*(a,b,c)$
0	0	0	0.000
0	0	1	0.192
0	1	0	0.000
0	1	1	0.288
1	0	0	0.000
1	0	1	0.070
1	1	0	0.000
1	1	1	0.630

b	d	$\psi_3^*(b,d)$
0	0	0.00
0	1	0.14
1	0	0.00
1	1	0.16

TABLA 8.9. Absorción de la evidencia $\{C = 1, D = 1\}$ en las funciones potenciales de (8.35).

sobre B y C. El nodo B está contenido en tres conglomerados distintos, siendo C_3 el de menor tamaño. Por tanto, para obtener $p(b)$ se marginaliza $p(b,d)$ sobre D. El nodo C también está contenido en tres conglomerados, el menor de los cuales es C_4. Por tanto, para obtener $p(c)$ se marginaliza $p(c,f)$ sobre F. Por otra parte, el nodo D está contenido únicamente en el conglomerado C_3; luego, se marginaliza $p(b,d)$ sobre B. El nodo E está contenido en C_2, luego será necesario marginalizar $p(b,c,e)$ sobre B y C para obtener $p(e)$. Finalmente, el nodo F está contenido únicamente en C_4, por tanto, $p(f)$ se obtendrá marginalizando $p(c,f)$ sobre C. Todas las funciones de probabilidad marginales de los nodos han sido obtenidas previamente en el Ejemplo 8.4, aplicando el método de condicionamiento. La Figura 8.21 muestra las probabilidades obtenidas aplicando cualquiera de los métodos de propagación (condicionamiento o agrupamiento), que corresponden al caso en el que no se dispone de evidencia.

Supóngase ahora que se tiene la evidencia $\{C = 1, D = 1\}$. Esta evidencia puede absorberse en la representación potencial mostrada en la Tabla 8.3 utilizando, por ejemplo, la opción descrita en (8.33). En este caso, no es necesario modificar la estructura topológica, sino solamente las funciones potenciales asociadas a los nodos que contienen evidencia. Por ejemplo, se puede absorber la evidencia $C = 1$ en la función potencial correspondiente al conglomerado C_1 y la evidencia $D = 1$ en la función potencial asociada al conglomerado C_3. En consecuencia, sólo es necesario modificar las funciones $\psi_1(a,b,c)$ y $\psi_3(b,d)$, de la forma indicada en la Tabla 8.9.

Procediendo de la misma forma que en el caso anterior (sin evidencia), se pueden calcular las funciones de probabilidad condicionada de los nodos, dada la evidencia $\{C = 1, D = 1\}$. Estas mismas probabilidades también han sido obtenidas anteriormente con el algoritmo de condicionamiento (ver Figura 8.23). ∎

El algoritmo de agrupamiento anterior supone que la red de Markov es descomponible. Esta propiedad es necesaria para garantizar la existencia de

una cadena de conglomerados que permita factorizar la función de probabilidad conjunta de la forma (8.27). Sin embargo, como ya se indicó anteriormente, esta condición no es restrictiva, pues si la red no es descomponible, entonces el proceso de propagación puede ser realizado en una red descomponible auxiliar obtenida triangulando la red original. Dado que el proceso de triangulación añade nuevas aristas, todos los conglomerados del grafo original estarán contenidos en los conglomerados del nuevo grafo. Por tanto, las funciones potenciales de la nueva red podrán ser definidas agrupando las funciones potenciales de la red original en los nuevos conglomerados. De esta forma el modelo probabilístico asociado a ambas redes será el mismo y, por tanto, la propagación podrá ser realizada de forma equivalente en cualquiera de las redes de Markov. El ejemplo siguiente ilustra este hecho.

Ejemplo 8.6 Red de Markov no descomponible. Considérese el grafo no dirigido dado en la Figura 8.25. Este grafo no es triangulado, pues contiene el bucle de longitud cuatro $A - B - E - C - A$ que no posee ninguna cuerda. Por tanto, la red de Markov asociada a este grafo será no descomponible. Los conglomerados del grafo son: $C_1 = \{A, B\}$, $C_2 = \{A, C\}$, $C_3 = \{B, E\}$, $C_4 = \{C, E\}$, $C_5 = \{B, D\}$ y $C_6 = \{C, F\}$. Por tanto, una representación potencial de esta red de Markov viene dada por

$$p(a, b, c, d, e, f) = \psi_1(a, b)\psi_2(a, c)\psi_3(b, e)\psi_4(c, e)\psi_5(b, d)\psi_6(c, f). \quad (8.37)$$

El algoritmo de agrupamiento no puede ser aplicado en esta situación. Sin embargo, si se triangula el grafo añadiendo la arista $B - C$, se obtiene el grafo utilizado en el Ejemplo 8.5 que se muestra en la Figura 8.24. Los conglomerados asociados a este grafo son $C_1^* = \{A, B, C\}$, $C_2^* = \{B, C, E\}$, $C_3^* = \{B, D\}$ y $C_4^* = \{C, F\}$. Por tanto, se puede obtener una representación potencial para el nuevo grafo utilizando (8.37) de la forma siguiente:

$$\begin{aligned} \psi_1^*(a, b, c) &= \psi_1(a, b)\psi_2(a, c), \\ \psi_2^*(b, c, e) &= \psi_3(b, e)\psi_4(c, e), \\ \psi_3^*(b, d) &= \psi_5(b, d), \\ \psi_4^*(c, f) &= \psi_6(c, f). \end{aligned}$$

El grafo mostrado en la Figura 8.24(a) y la nueva representación potencial

$$p(a, b, c, d, e, f) = \psi_1^*(a, b, c)\psi_2^*(b, c, e)\psi_3^*(b, d)\psi_4^*(c, f), \quad (8.38)$$

proporcionan una red de Markov descomponible en la cual puede realizarse la propagación de la evidencia aplicando el algoritmo de agrupamiento descrito anteriormente. ∎

8.6.2 Algoritmo de Agrupamiento en Redes Bayesianas

En la sección anterior se presentó el método de agrupamiento para propagar evidencia en redes de Markov. En esta sección se presenta una adaptación

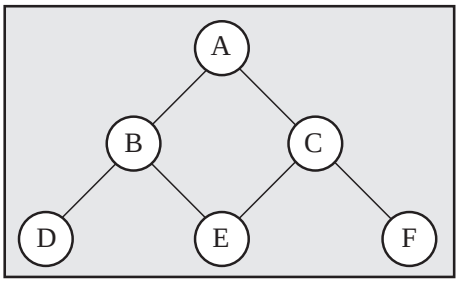


FIGURA 8.25. Grafo no dirigido no triangulado.

del método para propagar evidencia en redes Bayesianas. En la sección 6.4.4 se vio que dada una red Bayesiana (D, P), definida en un conjunto de variables $\{X_1, \ldots, X_n\}$, la función de probabilidad asociada podía ser factorizada en la forma

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i), \tag{8.39}$$

donde Π_i es el conjunto de padres del nodo X_i en D.

En este caso se puede transformar el grafo dirigido en un grafo no dirigido triangulado para poder aplicar el método de agrupamiento. En el Capítulo 4 se vieron algunos tipos de grafos no dirigidos asociados a un grafo dirigido. Considerando el grafo no dirigido obtenido triangulando el grafo moralizado del grafo dirigido original, se tiene que cada familia del grafo dirigido estará contenida en algún conglomerado del grafo triangulado. Esta propiedad permite definir una representación potencial para la red de Markov descomponible a partir de la representación de la función de probabilidad de la red Bayesiana dada en (8.39). Por tanto, el problema de propagar evidencia en redes Bayesianas puede ser resuelto aplicando el Algoritmo 8.2. Este proceso se presenta en el algoritmo siguiente.

Algoritmo 8.3 Algoritmo de agrupamiento en redes Bayesianas.

- **Datos:** Una red Bayesiana (D, P) definida en un conjunto de variables X y una evidencia $E = e$.
- **Resultados:** Las funciones de probabilidad condicionada $p(x_i|e)$ de cada nodo $X_i \notin E$.

1. Moralizar y triangular el grafo D, obteniendo un grafo no dirigido G.
2. Obtener el conjunto C de los conglomerados de G.
3. Asignar cada nodo X_i contenido en X a un único conglomerado que contenga a su familia. Sea A_i el conjunto de nodos asociados al conglomerado C_i.

4. Para cada conglomerado $C_i \in C$ definir $\psi_i(c_i) = \prod_{x_i \in A_i} p(x_i|\pi_i)$. Si $A_i = \phi$, definir $\psi_i(c_i) = 1$.

5. Aplicar el Algoritmo 8.2 a la red de Markov (C, Ψ) y a la evidencia $E = e$ para obtener las probabilidades de los nodos. ∎

El ejemplo siguiente ilustra este algoritmo.

Ejemplo 8.7 Algoritmo de agrupamiento en redes Bayesianas. Considérese el grafo dirigido mostrado en la Figura 8.26. En el Ejemplo 8.4 se ha visto que este grafo implica la siguiente factorización de la función de probabilidad:

$$p(a, b, c, d, e, f) = p(a)p(b|a)p(c|a)p(d|b)p(e|b, c)p(f|c), \tag{8.40}$$

con los valores numéricos asociados mostrados en la Tabla 8.2. En este ejemplo se aplica el Algoritmo 8.3 para obtener una red de Markov equivalente para propagar evidencia utilizando el algoritmo de agrupamiento.

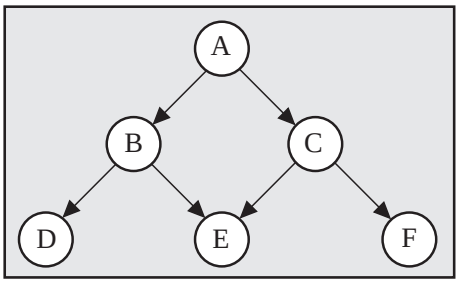


FIGURA 8.26. Grafo dirigido acíclico múltiplemente conexo.

Para moralizar el grafo dado en la Figura 8.26 es necesario añadir la arista $B-C$, ya que los nodos B y C tienen un hijo común, E. El grafo resultante es triangulado (ver Figura 8.24(a)), por lo que no es necesaria ninguna transformación adicional. Los conglomerados del grafo son $C_1 = \{A, B, C\}$, $C_2 = \{B, C, E\}$, $C_3 = \{B, D\}$ y $C_4 = \{C, F\}$. Cada uno de los nodos ha de ser asignado a un conglomerado que contenga a su familia. Por ejemplo, los nodos A, B y C se asignan al conglomerado C_1, el nodo E al conglomerado C_2, y los nodos D y F a los conglomerados C_3 y C_4, respectivamente. Por tanto, se obtiene la siguiente representación potencial de la red de Markov dada en la Figura 8.24(a):

$$\begin{aligned} \psi_1(a, b, c) &= p(a)p(b|a)p(c|a), \\ \psi_2(b, c, e) &= p(e|b, c), \\ \psi_3(b, d) &= p(d|b), \\ \psi_4(c, f) &= p(f|c). \end{aligned} \tag{8.41}$$

a	b	c	$\psi_1(a,b,c)$
0	0	0	0.024
0	0	1	0.096
0	1	0	0.036
0	1	1	0.144
1	0	0	0.035
1	0	1	0.035
1	1	0	0.315
1	1	1	0.315

b	c	e	$\psi_2(b,c,e)$
0	0	0	0.40
0	0	1	0.60
0	1	0	0.50
0	1	1	0.50
1	0	0	0.70
1	0	1	0.30
1	1	0	0.20
1	1	1	0.80

b	c	$\psi_3(b,d)$
0	0	0.30
0	1	0.70
1	0	0.20
1	1	0.80

c	f	$\psi_4(c,f)$
0	0	0.10
0	1	0.90
1	0	0.40
1	1	0.60

TABLA 8.10. Valores numéricos de las funciones potenciales de la representación (8.42).

Los valores numéricos correspondientes se muestran en la Tabla 8.10.

El grafo mostrado en al Figura 8.24(a) y la nueva representación potencial,

$$p(a,b,c,d,e,f) = \psi_1(a,b,c)\psi_2(b,c,e)\psi_3(b,d)\psi_4(c,f), \tag{8.42}$$

proporcionan una red de Markov descomponible que puede utilizarse para propagar evidencia. Obsérvese que los valores de cada función potencial son proporcionales a los valores correspondientes de la Tabla 8.3, utilizada en el Ejemplo 8.5. Por tanto, ambas representaciones potenciales definen la misma red de Markov. Así, al aplicar el algoritmo de agrupamiento en este ejemplo se obtienen las mismas probabilidades marginales que en el Ejemplo 8.5. ∎

8.7 Propagación en Árboles de Conglomerados

El algoritmo de agrupamiento introducido en la Sección 8.6 agrupa conjuntos de nodos con cierta estructura local creando una cadena de conglomerados para propagar evidencia. Algunas modificaciones de este método utilizan una representación gráfica de la cadena de conglomerados (por ejemplo, un árbol de unión) para propagar la evidencia de forma más eficiente. El método de los *universos de conocimiento* desarrollado por Jensen, Olesen y Andersen (1990) (ver también Jensen, Lauritzen y Olesen (1990))

transforma el grafo múltiplemente conexo en un árbol de conglomerados asociado al grafo original. Las dos operaciones básicas de *distribuir evidencia* y *agrupar evidencia*, se encargan de propagar la evidencia en el árbol de conglomerados de forma eficiente. Otra modificación del algoritmo de agrupamiento, introducida por Shachter, Andersen y Szolovits (1994), proporciona un marco general para entender las equivalencias entre las transformaciones y operaciones realizadas por los distintos algoritmos de propagación (para obtener más detalles, consultar Shachter, Andersen y Szolovits (1994)).

Esta sección presenta una modificación del algoritmo de agrupamiento que se basa en el envío de mensajes en un árbol de unión construido a partir de una cadena de conglomerados. De igual forma que el algoritmo de agrupamiento, este método de *propagación en árboles de unión* es válido para redes de Markov y redes Bayesianas.

En la Sección 8.6 se analizaron las representaciones potenciales de las funciones de probabilidad de las redes de Markov y las redes Bayesianas. Para redes de Markov la función de probabilidad de los nodos puede escribirse directamente como el producto de las funciones potenciales definidas en los conglomerados. Si el grafo no es triangulado, siempre puede hallarse un grafo triangulado equivalente cuya representación potencial venga dada a partir del grafo no triangulado original. En el caso de redes Bayesianas, la representación potencial se construye a través de grafo no dirigido obtenido moralizando y triangulando el grafo original. Esta representación potencial se obtiene asignando la función de probabilidad condicionada, $p(x_i|\pi_i)$, de cada nodo X_i a la función potencial de un conglomerado que contenga a la familia del nodo.

Por tanto, para describir el algoritmo de propagación en árboles de unión se supondrá que se tiene un grafo no dirigido triangulado con conglomerados $\{C_1, \ldots, C_m\}$ y una representación potencial $\{\psi_1(c_1), \ldots, \psi_m(c_m)\}$. Este algoritmo utiliza un árbol de unión del grafo para propagar la evidencia. El ejemplo siguiente ilustra el proceso de construcción de este árbol (ver Sección 4.6).

Ejemplo 8.8 Construyendo un árbol de unión. Considérese el grafo triangulado no dirigido mostrado en la Figura 8.24(a). En el Ejemplo 8.5 se obtuvo la cadena de conglomerados

$$C_1 = \{A, B, C\}, \;\; C_2 = \{B, C, E\}, \;\; C_3 = \{B, D\}, \;\; C_4 = \{C, F\},$$

que cumple la propiedad de intersección dinámica. Esta cadena define la estructura de los conjuntos separadores y determina los vecinos posibles de cada conglomerado, es decir, los conglomerados previos de la cadena que contienen al conjunto separador. Los conjuntos separadores y los vecinos de cada conglomerado se muestran en la Tabla 8.4. El Algoritmo 4.4 muestra que una cadena de conglomerados puede ser representada gráficamente mediante un árbol, en el que cada conglomerado está unido a alguno de

sus vecinos. En este ejemplo se tienen cuatro posibilidades[5], como se muestra en la Figura 8.27. Obsérvese que, por ejemplo, el árbol de unión de la Figura 8.27(a) ha sido construido asignando al conglomerado C_2 su único vecino posible C_1, y eligiendo el conglomerado C_2 como vecino para C_3 y C_4.

Por tanto, dado el grafo dirigido de la Figura 8.26, se puede aplicar el Algoritmo 4.5 (que incluye las operaciones de moralización y triangulación del grafo dirigido) para obtener un árbol de familias, es decir, un árbol de unión en el que la familia de cada nodo está contenida en algún conglomerado. Esto garantiza la existencia de una representación potencial del grafo no dirigido resultante definida por las funciones de probabilidad condicionada asociadas al grafo dirigido. ■

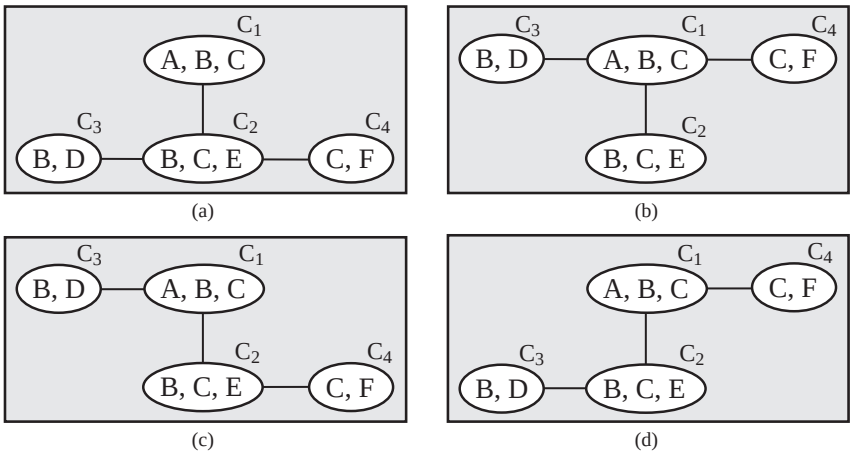


FIGURA 8.27. Cuatro árboles de unión posibles para el grafo no dirigido de la Figura 8.24(a).

La estructura de poliárbol dada por un árbol de unión proporciona un modelo gráfico adecuado para realizar las operaciones del algoritmo de agrupamiento de forma intuitiva. Por ejemplo, los separadores correspondientes a la cadena de conglomerados pueden representarse gráficamente mediante la intersección de conglomerados vecinos en el árbol de unión. Si los conglomerados C_i y C_j son vecinos en el árbol de unión, se define su conjunto separador S_{ij}, o S_{ji}, como $S_{ij} = C_i \cap C_j$. Por ejemplo, dado el árbol de unión de la Figura 8.27(a), los tres separadores posibles son $S_{12} = C_1 \cap C_2 = \{B, C\}$, $S_{23} = \{B\}$ y $S_{24} = \{C\}$, como se muestra en la

[5]Se tienen cuatro árboles distintos ya que C_1 no tiene vecinos, C_2 tiene un sólo vecino, y C_3 y C_4 tienen dos vecinos cada uno. Por tanto existen dos opciones para elegir el vecino de C_3 para cada uno de los vecinos posibles de C_4.

Figura 8.28. Obsérvese que estos conjuntos coinciden con los separadores obtenidos a partir de la cadena de conglomerados en el Ejemplo 8.5 (ver Tabla 8.4).

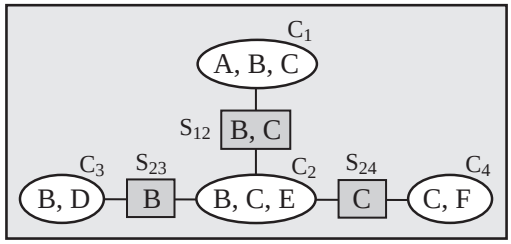


FIGURA 8.28. Conjuntos separadores del árbol de unión.

Suponiendo que la evidencia disponible, $E = e$, ya ha sido absorbida en las funciones potenciales, se necesita realizar las siguientes operaciones para calcular la probabilidad condicionada $p(x_i|e)$ de un nodo X_i:

- Calcular los mensajes entre conglomerados vecinos del árbol de unión.
- Calcular la función de probabilidad de cada conglomerado utilizando estos mensajes.
- Marginalizar la función de probabilidad de un conglomerado que contenga al nodo X_i sobre el resto de las variables.

El algoritmo de propagación en árboles de conglomerados realiza básicamente las mismas operaciones que el algoritmo de agrupamiento. Los cálculos locales son realizados ahora en forma de mensajes y propagados por el árbol mediante el envío de mensajes entre conglomerados vecinos, de forma análoga a la realizada en el algoritmo de propagación en poliárboles. Para ilustrar este proceso, supóngase que un conglomerado arbitrario C_i tiene q conglomerados vecinos $\{B_1, \ldots, B_q\}$. Sean C_{ij} y X_{ij} el conjunto de conglomerados y el conjunto de nodos contenidos en el subárbol correspondiente al conglomerado C_i cuando se elimina la arista $C_i - B_j$, respectivamente. La Figura 8.29 muestra el conjunto C_{ij} asociado a un conglomerado, C_i, y un vecino, B_j, arbitrarios. Obsérvese que el conjunto X_{ij} es la unión de los conglomerados contenidos en C_{ij}. Esta figura muestra también que los conjuntos C_{ij} y C_{ji} son complementarios, es decir, $X = C_{ij} \cup C_{ji} = X_{ij} \cup X_{ji}$.

Para calcular la función de probabilidad de un conjunto separador S_{ij}, primero se descompone el conjunto $X \setminus S_{ij}$ de la forma siguiente

$$X \setminus S_{ij} = (X_{ij} \cup X_{ji}) \setminus S_{ij} = (X_{ij} \setminus S_{ij}) \cup (X_{ji} \setminus S_{ij}) = R_{ij} \cup R_{ji},$$

donde $R_{ij} = X_{ij} \setminus S_{ij}$ es el conjunto de nodos contenidos en el subárbol asociado a C_i, pero no en el árbol asociado a B_j, cuando se elimina la

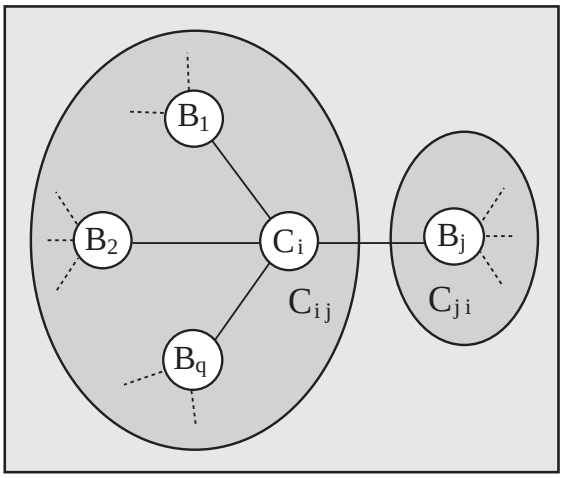


FIGURA 8.29. Descomposición en subconjuntos disjuntos.

arista $C_i - B_j$ del árbol de unión. Dado que todo nodo que esté contenido en dos conglomerados distintos tiene que estar contenido también en todos los conglomerados del camino que los une, los únicos nodos comunes de X_{ij} y X_{ji} deben de estar contenidos en su conjunto separador, S_{ij}. Por tanto, R_{ij} y R_{ji} son subconjuntos disjuntos. Este hecho puede ser utilizado para calcular la función de probabilidad del conjunto separador S_{ij} mediante

$$\begin{aligned} p(s_{ij}) &= \sum_{x \setminus s_{ij}} \prod_{k=1}^{m} \psi_k(c_k) \\ &= \sum_{r_{ij} \cup r_{ji}} \prod_{k=1}^{m} \psi_k(c_k) \\ &= \left(\sum_{r_{ij}} \prod_{c_k \in C_{ij}} \psi_k(c_k) \right) \left(\sum_{r_{ji}} \prod_{c_k \in C_{ji}} \psi_k(c_k) \right) \\ &= M_{ij}(s_{ij})\, M_{ji}(s_{ij}), \end{aligned}$$

donde

$$M_{ij}(s_{ij}) = \sum_{r_{ij}} \prod_{c_k \in C_{ij}} \psi_k(c_k) \tag{8.43}$$

es el mensaje que el conglomerado C_i envía a su vecino B_j y

$$M_{ji}(s_{ij}) = \sum_{r_{ji}} \prod_{c_k \in C_{ji}} \psi_k(c_k) \tag{8.44}$$

es el mensaje que el conglomerado B_j envía a C_i.

Obsérvese que la función de probabilidad del conjunto separador S_{ij} es el producto de esos dos mensajes, y que toda la información necesaria para calcular cada uno de estos mensajes está contenida en uno de los dos subgrafos separados por la arista $C_i - B_j$ del árbol de unión. Por tanto, estos mensajes propagan la información procedente de una de las ramas del grafo por la otra rama. Otra propiedad interesante, que hace posible la implementación paralela de este algoritmo, es que los mensajes pueden ser calculados de forma independiente.

Para calcular la función de probabilidad de un conglomerado C_i, se descompone el conjunto $X \setminus C_i$ según sea la información procedente de los conglomerados vecinos:

$$X \setminus C_i = \left(\bigcup_{k=1}^{q} X_{ki} \right) \setminus C_i = \bigcup_{k=1}^{q} (X_{ki} \setminus C_i) = \bigcup_{k=1}^{q} R_{ki},$$

donde la última igualdad resulta de la equivalencia, $X_{ki} \setminus C_i = R_{ki}$, que se deriva de la propiedad de los árboles de unión (cada nodo de X_{ki} que está contenido en C_i está también contenido en S_{ki}), es decir, $X_{ki} \setminus C_i = X_{ki} \setminus S_{ki} = R_{ki}$. Esta descomposición se ilustra en la Figura 8.30.

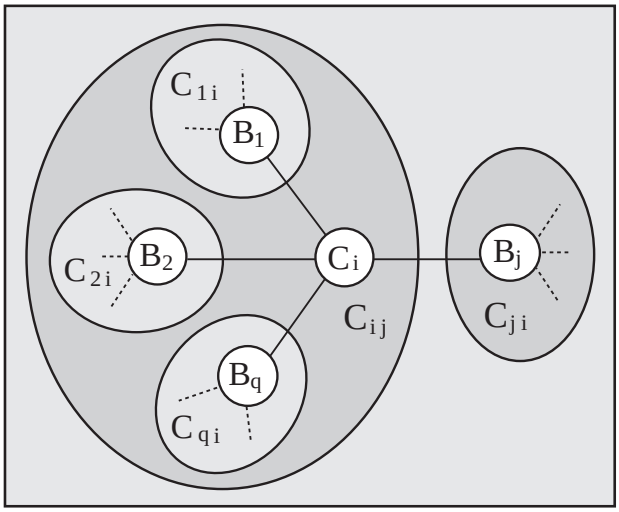


FIGURA 8.30. Descomposición en subconjuntos disjuntos.

Por tanto la función de probabilidad del conglomerado C_i puede escribirse como

$$\begin{aligned} p(c_i) &= \sum_{x \setminus c_i} \prod_{j=1}^{m} \psi_j(c_j) \\ &= \psi_i(c_i) \sum_{x \setminus c_i} \prod_{j \neq i} \psi_j(c_j) \end{aligned}$$

$$
\begin{aligned}
&= \psi_i(c_i) \sum_{r_{1i} \cup \ldots \cup r_{qi}} \prod_{j \neq i} \psi_j(c_j) \\
&= \psi_i(c_i) \left(\sum_{r_{1i}} \prod_{c_k \in C_{1i}} \psi_k(c_k) \right) \ldots \left(\sum_{r_{qi}} \prod_{c_k \in C_{qi}} \psi_k(c_k) \right) \\
&= \psi_i(c_i) \prod_{j=1}^{q} M_{ki}(s_{ij}), \qquad (8.45)
\end{aligned}
$$

donde $M_{ji}(s_{ij})$ es el mensaje enviado del conglomerado B_j al conglomerado C_i, de la forma definida en (8.44).

Una vez que todos los mensajes han sido calculados y enviados, puede obtenerse la función de probabilidad de los conglomerados utilizando (8.45). Obsérvese que para calcular el mensaje $M_{ij}(s_{ij})$ por medio de (8.43) es necesario utilizar un sumatorio sobre todas las variables del conjunto $X_{ij} \setminus S_{ij}$. Este cálculo puede ser simplificado considerando la descomposición

$$
X_{ij} \setminus S_{ij} = (C_i \setminus S_{ij}) \cup \left(\bigcup_{k \neq j} X_{ki} \setminus S_{ki} \right).
$$

Por tanto, se obtiene la relación siguiente entre los mensajes:

$$
\begin{aligned}
M_{ij}(s_{ij}) &= \sum_{x_{ij} \setminus s_{ij}} \prod_{c_s \in C_{ij}} \psi_s(c_s) \\
&= \sum_{c_i \setminus s_{ij}} \sum_{(x_{ki} \setminus s_{ki}),\, k \neq j} \prod_{c_s \in C_{ij}} \psi_s(c_s) \\
&= \sum_{c_i \setminus s_{ij}} \psi_i(c_i) \prod_{k \neq j} \sum_{x_{ki} \setminus s_{ki}} \prod_{c_s \in C_{ki}} \psi_s(c_s) \\
&= \sum_{c_i \setminus s_{ij}} \psi_i(c_i) \prod_{k \neq j} M_{ki}(s_{ki}). \qquad (8.46)
\end{aligned}
$$

Finalmente, la función de probabilidad condicionada de cualquier nodo X_i puede ser obtenida marginalizando la función de probabilidad de cualquier conglomerado que contenga al nodo X_i. Si el nodo está contenido en más de un conglomerado, entonces es conveniente elegir aquel que tenga menor tamaño para minimizar el número de operaciones.

A partir de la discusión anterior se deduce lo siguiente:

- La ecuación (8.45) muestra que la función de probabilidad $p(c_i)$ del conglomerado C_i puede ser calculada tan pronto como C_i haya recibido los mensajes de todos sus vecinos.
- La ecuación (8.46) muestra que el mensaje $M_{ij}(s_{ij})$, que el conglomerado C_i envía a su vecino B_j, puede ser calculado tan pronto como

C_i haya recibido los mensajes $M_{ki}(s_{ki})$ procedentes del resto de sus vecinos.

La discusión anterior sugiere el algoritmo siguiente.

Algoritmo 8.4 Propagación en redes de Markov utilizando un árbol de unión.

- **Datos:** Una red de Markov (C, Ψ) definida sobre un conjunto de variables X y una evidencia $E = e$.
- **Resultados:** La función de probabilidad condicionada $p(x_i|e)$ para cada nodo $X_i \notin E$.

Etapa de Iniciación:

1. Absorber la evidencia $E = e$ en las funciones potenciales Ψ utilizando (8.33) ó (8.34).
2. Utilizar el Algoritmo 4.4 para obtener un árbol de unión de la red de Markov.

Etapa Iterativa:

3. Para $i = 1, \ldots, m$ hacer: Para cada vecino B_j del conglomerado C_i, si C_i ha recibido los mensajes del resto de sus vecinos, calcular el mensaje $M_{ij}(s_{ij})$ y enviárselo a B_j, donde
$$M_{ij}(s_{ij}) = \sum_{c_i \setminus s_{ij}} \psi_i(c_i) \prod_{k \neq j} M_{ki}(s_{ki}). \tag{8.47}$$
4. Repetir el Paso 3 hasta que no se obtenga ningún nuevo mensaje.
5. Calcular la función de probabilidad de cada conglomerado C_i utilizando
$$p(c_i) = \psi_i(c_i) \prod_k M_{ki}(s_{ik}). \tag{8.48}$$
6. Para cada nodo de la red, X_i, calcular $p(x_i|e)$ utilizando
$$p(x_i|e) = \sum_{c_k \setminus x_i} p(c_k), \tag{8.49}$$
donde C_k es el conglomerado de menor tamaño que contiene a X_i. ∎

Obsérvese que en el Paso 3 del algoritmo pueden darse tres situaciones diferentes para cada conglomerado C_i:

- **Caso 1:** C_i ha recibido los mensajes de todos sus vecinos. En ese caso C_i puede calcular y enviar los mensajes a todos sus vecinos.

- **Caso 2:** C_i ha recibido los mensajes de todos sus vecinos, excepto de B_j. En ese caso, C_i sólo puede calcular y enviar su mensaje al conglomerado B_j.
- **Caso 3:** C_i no ha recibido los mensajes de dos o más vecinos. En ese caso, no se puede calcular ningún mensaje.

Este algoritmo puede ser modificado para propagar evidencia en redes Bayesianas de la forma siguiente.

Algoritmo 8.5 Propagación en redes Bayesianas utilizando un árbol de unión.

- **Datos:** Una red Bayesiana (D, P) sobre un conjunto de variables X y una evidencia $E = e$.
- **Resultados:** La función de probabilidad condicionada $p(x_i|e)$ para cada nodo $X_i \notin E$.

1. Utilizar el Algoritmo 4.5 para obtener un árbol de familias del grafo D. Sea C el conjunto de conglomerados resultante.
2. Asignar cada nodo X_i a un sólo conglomerado que contenga a su familia. Sea A_i el conjunto de nodos asignados al conglomerado C_i.
3. Para cada conglomerado C_i definir $\psi_i(c_i) = \prod_{x_i \in A_i} p(x_i|\pi_i)$. Si $A_i = \phi$, entonces definir $\psi_i(c_i) = 1$.
4. Aplicar el Algoritmo 8.4 a la red de Markov (C, Ψ) y a la evidencia $E = e$ para obtener las funciones de probabilidad condicionada de los nodos. ∎

La estructura de envío de mensajes de los Algoritmos 8.4 y 8.5 hace que estos sean apropiados para la implementación paralela. En este caso, puede asignarse un procesador a cada conglomerado del árbol de unión. El procesador asignado a un conglomerado C_i necesita conocer la información siguiente para calcular la función de probabilidad condicionada $p(c_i|e)$ (utilizando (8.48)) y para enviar los mensajes a sus vecinos:

- Una lista que contenga a los vecinos del conglomerado C_i. Esta información es independiente de la evidencia E.
- La función potencial asociada $\psi(c_i)$. Esta función puede depender de la evidencia, debido al proceso de absorción de evidencia.
- El mensaje $M_{ji}(s_{ij})$, recibido de cada uno de sus vecinos B_j. Este mensaje se calcula por el procesador asociado al conglomerado B_j, utilizando (8.47).

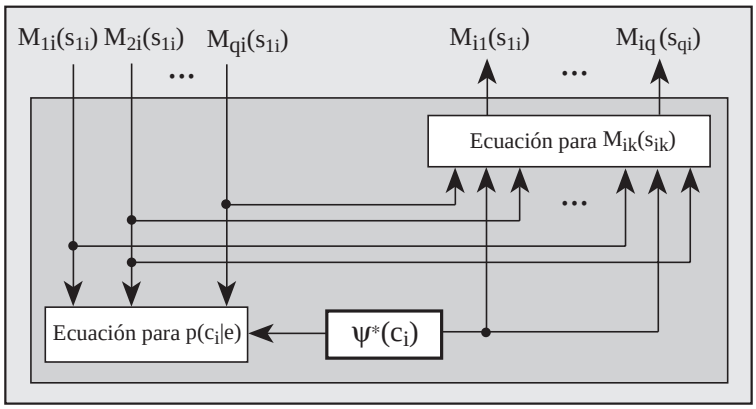


FIGURA 8.31. Cálculos realizados por el procesador del conglomerado C_i y los mensajes recibidos y enviados a sus vecinos.

En la Figura 8.31 se muestran de forma esquemática los cálculos realizados por el procesador de un nodo arbitrario C_i así como los mensajes enviados y recibidos.

Ejemplo 8.9 Propagación utilizando un árbol de unión. Considérese la red Bayesiana del Ejemplo 8.7, definida por el grafo múltiplemente conexo de la Figura 8.26 y por las funciones de probabilidad condicionada dadas en (8.40), cuyos valores numéricos se muestran en la Tabla 8.2. En el Ejemplo 8.7 se obtuvo un grafo moralizado y triangulado correspondiente a este grafo dirigido (ver Figura 8.24(a)) y el conjunto de conglomerados correspondiente (ver Figura 8.24(b)): $C_1 = \{A, B, C\}$, $C_2 = \{B, C, E\}$, $C_3 = \{B, D\}$ y $C_4 = \{C, F\}$. Las funciones potenciales se definen por medio de las funciones de probabilidad condicionada de la forma mostrada en (8.41), con los valores numéricos dados en la Tabla 8.10. En el Ejemplo 8.8 se obtuvieron cuatro árboles de unión distintos asociados a este grafo (ver Figura 8.27). En este ejemplo se trata de obtener las probabilidades de los nodos aplicando el Algoritmo 8.4 al árbol de unión mostrado en la Figura 8.27(a).

El algoritmo procede enviando mensajes entre conglomerados vecinos en el árbol de unión. La Figura 8.32 muestra el orden en el que se calculan y envían estos mensajes. Las flechas indican los mensajes y los números denotan el orden en que son calculados.

Se tienen los mensajes siguientes:

- El conglomerado C_1 tiene un único vecino, C_2; por tanto, se puede calcular el mensaje $M_{12}(s_{12})$ y enviárselo al conglomerado C_2. Utilizando (8.47), se tiene

$$M_{12}(s_{12}) = \sum_{c_1 \setminus s_{12}} \psi_1(c_1).$$

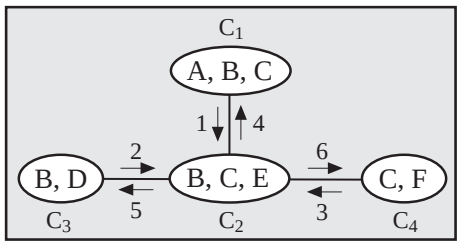


FIGURA 8.32. Orden en el que se calculan y envían los mensajes.

Por tanto,

$$M_{12}(b,c) = \sum_{a} \psi_1(a,b,c),$$

de lo cual se obtiene que

$$\begin{aligned} M_{12}(0,0) &= \psi_1(0,0,0) + \psi_1(1,0,0) = 0.024 + 0.035 = 0.059, \\ M_{12}(0,1) &= \psi_1(0,0,1) + \psi_1(1,0,1) = 0.096 + 0.035 = 0.131, \\ M_{12}(1,0) &= \psi_1(0,1,0) + \psi_1(1,1,0) = 0.036 + 0.315 = 0.351, \\ M_{12}(1,1) &= \psi_1(0,1,1) + \psi_1(1,1,1) = 0.144 + 0.315 = 0.459. \end{aligned}$$

- El conglomerado C_2 tiene tres vecinos, C_1, C_3 y C_4, pero ha recibido sólo el mensaje de C_1. Por tanto, todavía no se puede realizar ningún cálculo con este conglomerado (Caso 3).

- Dado que C_3 tiene un único vecino, C_2, se calcula y envía el mensaje $M_{32}(s_{23})$ utilizando (8.47):

$$M_{32}(s_{23}) = \sum_{c_3 \setminus s_{32}} \psi_3(c_3),$$

es decir,

$$M_{32}(b) = \sum_{d} \psi_3(b,d),$$

de lo cual se obtiene $(M_{32}(B=0), M_{32}(B=1)) = (1.0, 1.0)$.

- El conglomerado C_4 está en la misma situación que C_3. Por tanto, se calcula el mensaje $M_{42}(s_{24})$ y se envía al conglomerado C_2. Se tiene

$$M_{42}(s_{24}) = \sum_{c_4 \setminus s_{42}} \psi_4(c_4),$$

es decir

$$M_{42}(c) = \sum_{f} \psi_4(c,f),$$

de lo cual se obtiene $(M_{42}(C=0), M_{42}(C=1)) = (1.0, 1.0)$. Dado que se han obtenido varios mensajes en esta iteración, se vuelve a repetir el Paso 3.

- El conglomerado C_1 ya ha enviado el mensaje a su vecino C_2. El conglomerado C_2 ha recibido los mensajes de todos sus vecinos. Por tanto, puede calcular y enviar los mensajes $M_{21}(b,c)$, $M_{23}(b)$ y $M_{24}(c)$. Por ejemplo,

$$M_{23}(s_{23}) = \sum_{c_2 \setminus s_{23}} \psi_2(c_2) M_{12}(s_{12}) M_{42}(s_{24}),$$

o

$$M_{23}(b) = \sum_{c,e} \psi_2(b,c,e) M_{12}(b,c) M_{42}(c),$$

de lo cual se tiene

$$\begin{aligned} M_{23}(0) &= \sum_{c,e} \psi_2(0,c,e) M_{12}(0,c) M_{42}(c) \\ &= 0.4 \times 0.059 \times 1 + 0.6 \times 0.059 \times 1 \\ &\quad +0.5 \times 0.131 \times 1 + 0.5 \times 0.131 \times 1 = 0.19, \end{aligned}$$

$$\begin{aligned} M_{23}(1) &= \sum_{c,e} \psi_2(1,c,e) M_{12}(1,c) M_{42}(c) \\ &= 0.7 \times 0.351 \times 1 + 0.3 \times 0.351 \times 1 + \\ &\quad +0.2 \times 0.459 \times 1 + 0.8 \times 0.459 \times 1 = 0.81. \end{aligned}$$

Los valores numéricos correspondientes a todos los mensajes se muestran en la Figura 8.33.

- En el Paso 5 del algoritmo se calculan las funciones de probabilidad de los conglomerados utilizando (8.48). Se tiene

$$\begin{aligned} &p(c_1) = p(a,b,c) = \psi_1(a,b,c) M_{21}(b,c), \\ &p(c_2) = p(b,c,e) = \psi_2(b,c,e) M_{12}(b,c) M_{32}(b) M_{42}(c), \\ &p(c_3) = p(b,d) = \psi_3(b,d) M_{23}(b), \\ &p(c_4) = p(c,f) = \psi_4(c,f) M_{24}(c). \end{aligned}$$

Los valores numéricos de estas funciones son los mismos que han sido obtenidos con el algoritmo de agrupamiento (ver Tabla 8.8).

- En el Paso 6 se calculan las funciones de probabilidad condicionada de los nodos, de la misma forma que en el Ejemplo 8.5. Esto completa los cálculos en el caso de no disponer de evidencia.

Supóngase ahora que se tiene la evidencia $e = \{C = 1, D = 1\}$. Entonces, se puede absorber esta evidencia en la representación potencial utilizando (8.33) de la forma indicada en el Ejemplo 8.5 (ver Tabla 8.9), donde la evidencia $C = 1$ es absorbida en el conglomerado C_1 y la evidencia $D = 1$ es absorbida en el conglomerado C_3. El lector puede aplicar el Algoritmo 8.4 en este caso y obtener los mensajes mostrados en la Figura 8.34. ∎

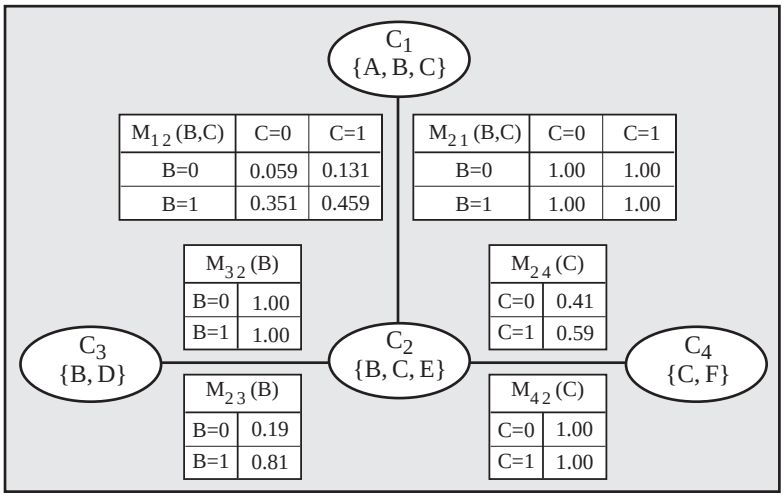


FIGURA 8.33. Valores numéricos de los mensajes calculados por el Algoritmo 8.4 cuando no se dispone de evidencia.

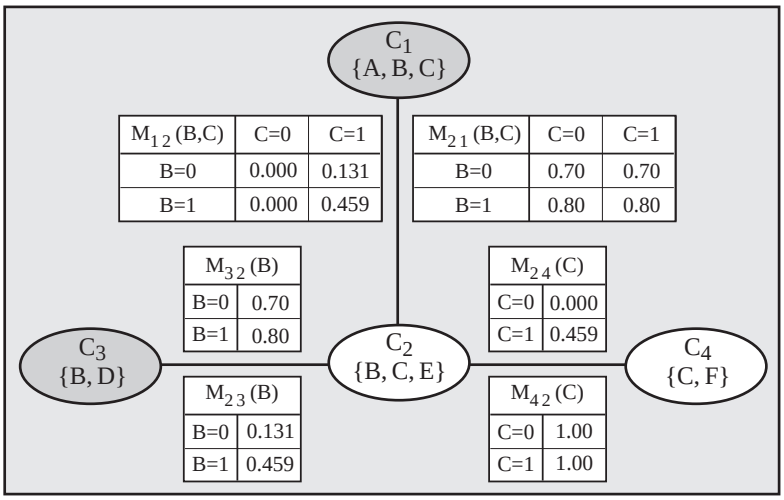


FIGURA 8.34. Valores numéricos de los mensajes calculados por el Algoritmo 8.4 cuando se dispone de la evidencia $e = \{C = 1, D = 1\}$. Los óvalos sombreados indican los conglomerados en los que se ha absorbido la evidencia.

8.8 Propagación Orientada a un Objetivo

En las secciones anteriores se han introducido varios algoritmos para la propagación de evidencia en redes de Markov y redes Bayesianas. El objetivo de estos algoritmos es obtener las funciones de probabilidad condicionada de todas las variables del modelo a partir de una evidencia observada,

$E = e$. Sin embargo, en algunos casos no se está interesado en todas las variables del modelo, sino que hay un pequeño conjunto de ellas por los que se tiene un interés particular. En estos casos el objetivo es obtener las funciones de probabilidad condicionada de esas variables, dada la evidencia. En esta situación, algunas de las variables del modelo pueden ser irrelevantes para el cálculo de estas funciones. Por tanto, se pueden evitar operaciones innecesarias hallando el conjunto de variables irrelevantes para la tarea concreta que se quiera realizar. Estas variables pueden ser eliminadas del grafo, permitiendo obtener un modelo equivalente más sencillo, que sólo contenga los nodos relevantes.

En esta sección se ilustra esta idea considerando el caso de las redes Bayesianas. En una red Bayesiana (D, P), definida en un conjunto de variables $X = \{X_1, \ldots, X_n\}$, cada nodo X_i tiene asociado un conjunto de parámetros correspondientes a la función de probabilidad condicionada $p(x_i|\pi_i)$. Supóngase que sólo se está interesado en un conjunto particular de variables $Y \subset X$, y que se desean calcular las funciones de probabilidad condicionada $p(x_i|e)$, para $X_i \in Y$, dada la evidencia $E = e$. Las variables contenidas en el conjunto Y son llamadas *variables objetivo*.

Los parámetros irrelevantes están asociados a los nodos cuyas funciones de probabilidad pueden se modificadas sin que ello afecte al valor de las probabilidades condicionadas de los nodos objetivo que se desean calcular. Estos parámetros pueden ser obtenidos utilizando la estructura de independencia contenida en el modelo probabilístico, como se muestra en Shachter (1988, 1990b) y Geiger, Verma, y Pearl (1990a, 1990b). Estos últimos autores proponen un algoritmo simple para hallar el conjunto de parámetros relevantes en tiempo polinomial en el número de variables. En este algoritmo, los parámetros asociados a la función de probabilidad condicionada $p(x_i|\pi_i)$ del nodo X_i son representados en el grafo mediante un nodo auxiliar Θ_i, asociado al nodo X_i. El conjunto de nodos relevantes puede ser obtenido utilizando el criterio de D-separación en el grafo resultante (ver Sección 5.2.2) de la forma indicada en el algoritmo siguiente.

Algoritmo 8.6 Identificación de los nodos relevantes.

- **Datos:** Una red Bayesiana (D, P) y dos conjuntos de nodos: el conjunto objetivo Y y el conjunto de evidencia E.
- **Resultados:** El conjunto de nodos relevantes R necesarios para calcular $p(y|e)$.

1. Construir un nuevo grafo dirigido, D', añadiendo un nodo auxiliar, Θ_i, y una arista $\Theta_i \rightarrow X_i$ a D para cada nodo X_i.
2. Identificar el conjunto Θ de nodos auxiliares que no están D-separados de Y por E en D' .
3. Asignar a R los nodos X_i cuyos nodos auxiliares Θ_i están contenidos en Θ. ∎

El nodo auxiliar Θ_i representa los parámetros asociados al nodo X_i. La segunda etapa del Algoritmo 8.6 puede ser desarrollada de forma eficiente utilizando un algoritmo introducido por Geiger, Verma, y Pearl (1990a, 1990b). Por tanto, si se considera el grafo reducido resultante de eliminar los nodos irrelevantes obtenidos mediante este algoritmo, se puede simplificar de forma importante el número de parámetros que han de ser considerados en el cálculo de las probabilidades. Este hecho se ilustra en el ejemplo siguiente.

Ejemplo 8.10 Propagación orientada a un objetivo. Considérese la red Bayesiana definida por el grafo dirigido acíclico mostrado en la Figura 8.35 y el modelo probabilístico asociado:

$$p(a, b, c, d, e, f) = p(a)p(b|a)p(c|a)p(d|b)p(e|b, c)p(f|c). \tag{8.50}$$

Supóngase que sólo se está interesado en las probabilidades del nodo C, después de conocer la evidencia $D = 0$. El nodo C está destacado en la Figura 8.35 para indiciar que es el nodo objetivo.

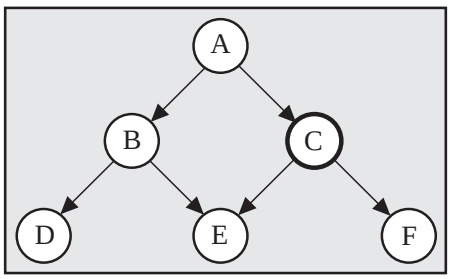


FIGURA 8.35. Grafo dirigido utilizado en un ejemplo de propagación orientada al nodo C (nodo objetivo).

Obsérvese que, en este caso, puede utilizarse alguno de los algoritmos de propagación descritos en secciones anteriores para calcular $p(x_i|e)$ para todos los nodos no evidenciales, incluyendo el nodo C. Sin embargo, basándose en lo comentado anteriormente, este proceso realizaría numerosas operaciones innecesarias. De forma alternativa, se puede reducir, en primer lugar, el grafo original, considerando únicamente el conjunto de nodos relevantes y, a continuación, realizar la propagación en el grafo reducido. Este grafo reducido puede obtenerse aplicando el Algoritmo 8.6 de la forma siguiente:

- En primer lugar, se construye un nuevo grafo añadiendo un nodo auxiliar para cada nodo del grafo. La Figura 8.36 muestra el nuevo grafo obtenido añadiendo los nodos $\{\Theta_A, \ldots, \Theta_F\}$ al grafo de la Figura 8.35.
- Utilizando el criterio de D-separación se pueden obtener los nodos auxiliares que no están D-separados del nodo objetivo C por el nodo

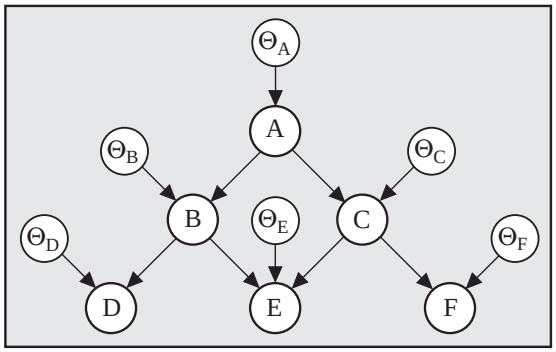


FIGURA 8.36. Grafo ampliado con un nodo auxiliar Θ_i y una arista $\Theta_i \rightarrow X_i$, para cada nodo X_i.

evidencial D en el grafo de la Figura 8.36. Por ejemplo, el nodo auxiliar Θ_A no está D-separado de C por D. Esto puede comprobarse construyendo el grafo ancestral moralizado asociado a los nodos C, D y Θ_A, como se muestra en la Figura 8.37 (ver Definición 5.4). En esta figura puede comprobarse que Θ_A no está D-separado de C por D, ya que existe un camino entre los nodos Θ_A y C que no incluye al nodo D. En este mismo grafo también puede observarse que los nodos auxiliares Θ_B, Θ_C y Θ_D no están D-separados de C por D. Sin embargo, los nodos Θ_E y Θ_F están D-separados de C por D (para ello pueden construirse los grafos moralizados ancestrales en cada caso.) Por tanto, el conjunto de parámetros relevantes para el cálculo de $p(c|d)$ es $\Theta = \{\Theta_A, \Theta_B, \Theta_C, \Theta_D\}$.

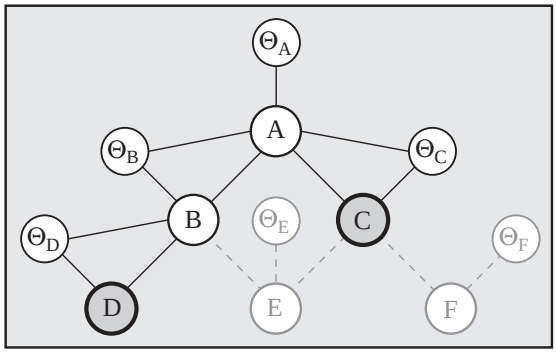


FIGURA 8.37. Grafo moralizado ancestral utilizado para decidir si los nodos auxiliares Θ_A, Θ_B, Θ_C y Θ_D están D-separados de C por D.

- El conjunto de nodos relevantes es $R = \{A, B, C, D\}$.

Una vez que se ha obtenido el conjunto de nodos relevantes, se puede simplificar el grafo eliminando los nodos irrelevantes. El grafo resultante se muestra en la Figura 8.38. Obsérvese que la función de probabilidad asociada al grafo reducido está definida por la factorización

$$p(a, b, c, d) = p(a)p(b|a)p(c|a)p(d|b). \tag{8.51}$$

Obsérvese también que mientras el grafo original es múltiplemente conexo, el grafo reducido es un poliárbol y, por tanto, la evidencia puede ser propagada de forma eficiente. Por tanto, la identificación de los parámetros relevantes permite simplificar tanto el tamaño, como la topología del grafo.

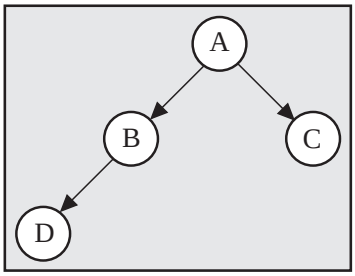


FIGURA 8.38. Grafo reducido para calcular $p(c|d)$.

Las Figuras 8.39(a) y (b) muestran las probabilidades resultantes de la propagación de la evidencia $D = 0$ en el grafo original (utilizando el algoritmo de agrupamiento, o el de propagación en árboles de conglomerados) y en el grafo reducido (utilizando el algoritmo para poliárboles), respectivamente. Como era de esperar, los dos métodos proporcionan las mismas probabilidades, $(p(C = 0|D = 0), p(C = 1|D = 0)) = (0.402, 0.598)$. ■

El ejemplo anterior ilustra el ahorro computacional que puede obtenerse aplicando el Algoritmo 8.6. Baker y Boult (1991) introducen una mejora de este algoritmo basada en el concepto de *redes computacionalmente equivalentes*, que formaliza la idea descrita en esta sección. Esta idea permite desarrollar un algoritmo para obtener el mínimo grafo computacionalmente equivalente a un grafo original para realizar una cierta tarea.

Por tanto, se pueden establecer las siguientes conclusiones:

- Dada la evidencia $E = e$, la función de probabilidad condicionada de cualquier nodo X_i de una red Bayesiana depende únicamente de los ascendientes del conjunto $\{X_i\} \cup E$. Como se ha visto en el Ejemplo 8.10, la función de probabilidad del nodo C dado $D = 0$ depende de $\{A, B, C, D\}$, que es el conjunto ancestral del conjunto de los nodos C y D.

- En el caso en que no exista evidencia, la función de probabilidad marginal de cualquier nodo X_i en una red Bayesiana depende sólo

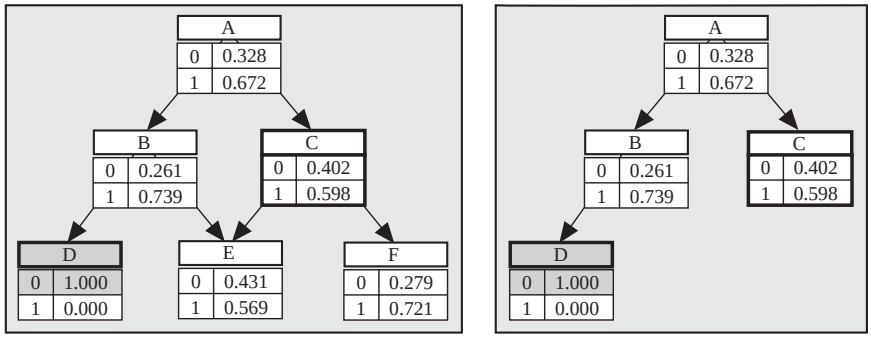


FIGURA 8.39. Probabilidades obtenidas de la propagación de la evidencia $D = 0$ en los grafos de las Figuras 8.35 y 8.38.

de su conjunto ancestral. Por ejemplo, la función de probabilidad marginal del nodo C en el Ejemplo 8.10 depende sólo de los nodos A y C.

- La función de probabilidad marginal de un nodo raíz X_i (un nodo sin padres) se puede obtener directamente de la correspondiente función de probabilidad condicional contenida en la factorización del modelo probabilístico $p(x_i|\pi_i) = p(x_i)$. Por ejemplo, la probabilidad del nodo A en el Ejemplo 8.10 puede obtenerse directamente de la Tabla 8.2: $(p(A = 0), p(A = 1)) = (0.3, 0.7)$.

8.9 Propagación Exacta en Redes Bayesianas Gausianas

En las secciones anteriores se han presentado varios algoritmos para la propagación de evidencia en modelos probabilísticos discretos multinomiales, que han tenido una amplia difusión sobre todo en aplicaciones prácticas. Sin embargo, en la literatura también han sido descritos varios modelos para tratar el caso de variables continuas. Algunos ejemplos son los modelos normales (ver Kenley (1986) y Shachter y Kenley (1989)), los modelos de Dirichlet (Geiger y Heckerman (1995) y Castillo, Hadi y Solares (1997)), e incluso modelos modelos mixtos de variables discretas y continuas (ver, por ejemplo, Lauritzen y Wermuth (1989), Olesen (1993), y Castillo, Gutiérrez y Hadi (1995b)).

Algunos de los algoritmos de propagación para estos modelos son adaptaciones de los algoritmos de propagación de evidencia en modelos discretos introducidos en las secciones anteriores. Por ejemplo, el algoritmo de propagación en redes Bayesianas normales introducido por Normand y Tritch-

ler (1992) utiliza las mismas ideas que el algoritmo de propagación en poliárboles descrito en la Sección 8.3. Por otra parte, Lauritzen (1992) presenta una modificación del algoritmo de propagación en árboles de unión (Sección 8.7) para propagar evidencia en modelos mixtos de variables discretas y continuas.

En la Sección 6.4.4 se analizaron las redes Bayesianas normales basándose en dos formas de representación diferentes: utilizando las matriz de covarianzas del modelo probabilístico normal, y utilizando una factorización de la función de la función de probabilidad como producto de funciones de probabilidad condicionada normales. Algunos algoritmos utilizan la segunda forma de representación para propagar evidencia (ver Xu y Pearl (1989), y Chang y Fung (1991)). En esta sección se presenta un algoritmo conceptualmente distinto que utiliza la representación del modelo dada por la matriz de covarianzas. Este algoritmo ilustra los conceptos básicos propios de la propagación en modelos normales. Una implementación iterativa permite actualizar las probabilidades en tiempo lineal, considerando un sólo nodo evidencial en cada paso. El teorema siguiente muestra el resultado principal que caracteriza las funciones de probabilidad condicionada obtenidas de un modelo normal (ver, por ejemplo, Anderson (1984)).

Teorema 8.1 Distribución normal multivariada. *Sean Y y Z dos conjuntos de variables aleatorias con función de distribución normal multivariada cuyo vector de medias y matriz de covarianzas son*

$$\mu = \begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix} \qquad \text{y} \qquad \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix},$$

donde μ_Y y Σ_{YY} son el vector de medias y la matriz de covarianzas de Y, μ_Z y Σ_{ZZ} son el vector de medias y la matriz de covarianzas de Z, y Σ_{YZ} es la matriz de covarianzas de Y y Z. Entonces, la función de probabilidad condicionada de Y dado $Z = z$ es una función normal multivariada con vector de medias $\mu_{Y|Z=z}$ y matriz de covarianzas $\Sigma_{Y|Z=z}$, donde

$$\mu_{Y|Z=z} = \mu_Y + \Sigma_{YZ}\Sigma_{ZZ}^{-1}(z - \mu_Z), \tag{8.52}$$

$$\Sigma_{Y|Z=z} = \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}. \tag{8.53}$$

Obsérvese que la media condicionada $\mu_{Y|Z=z}$ depende del valor z, pero no así la varianza condicionada $\Sigma_{Y|Z=z}$.

El Teorema 8.1 sugiere un procedimiento para obtener la media y la varianza de cualquier subconjunto de variables $Y \subset X$, dado un conjunto de nodos evidenciales $E \subset X$ que toman los valores $E = e$. Reemplazando Z por E en (8.52) y (8.53), se obtiene directamente el vector de medias y la matriz de covarianzas de la función de probabilidad de Y. Si se considera el conjunto $Y = X \setminus E$, entonces el procedimiento anterior permite calcular la función de probabilidad condicionada de los nodos no evidenciales, dado $E = e$. Obsérvese que esta función de probabilidad conjunta contiene toda

$Y \leftarrow X$
$\mu \leftarrow E[X]$
$\Sigma \leftarrow Var[X]$
Con $i \leftarrow 1$ hasta el número de elementos en E, **hacer**:
    $z \leftarrow i$-ésimo elemento de e
    $Y \leftarrow Y \setminus Z$
    $\mu \leftarrow \mu_y + \Sigma_{yz}\Sigma_z^{-1}(z - \mu_z)$
    $\Sigma \leftarrow \Sigma_y - \Sigma_{yz}\Sigma_z^{-1}\Sigma_{yz}^T$
$f(y|E = e) \sim N(\mu, \Sigma)$

FIGURA 8.40. Pseudocódigo del algoritmo iterativo para actualizar la función de probabilidad de los nodos sin evidencia Y, dada la evidencia $E = e$.

la información sobre las variables no evidenciales y sus relaciones, y no sólo la información usual de una sóla variable aislada obtenida por otros métodos. Los restantes métodos de propagación en redes normales utilizan la misma idea, pero realizan cálculos locales aprovechando la estructura de independencia dada por la factorización de la función de probabilidad como producto de funciones condicionadas.

Con objeto de simplificar los cálculos en el algoritmo anterior, es más conveniente utilizar un método iterativo, actualizando de uno en uno los nodos evidenciales contenidos en E. De esta forma, no es necesario invertir ninguna matriz ya que ésta se convierte en un escalar (un número) al considerar un sólo nodo evidencial. Además, μ_y y Σ_{yz} son vectores, y Σ_z es también un escalar. Entonces, el número de operaciones necesarias para actualizar la función de probabilidad de los nodos no evidenciales es lineal en el número de variables contenidas en X. Por tanto, este algoritmo proporciona un método sencillo y eficiente para propagar evidencia en modelos probabilísticos normales.

Dada la sencillez de este algoritmo iterativo, es muy fácil de implementar en el motor de inferencia de un sistema experto. La Figura 8.40 muestra el pseudocódigo correspondiente, que proporciona la función de probabilidad condicionada de los nodos no evidenciales Y, dada la evidencia $E = e$. El ejemplo siguiente ilustra el funcionamiento de este algoritmo.

Ejemplo 8.11 Propagación en redes Bayesianas normales. Considérese la red Bayesiana normal definida por el grafo mostrado en la Figura 8.41. Esta red Bayesiana fue introducida en el Ejemplo 6.18. La función de probabilidad de las variables puede ser factorizada en la forma siguiente:

$$f(a, b, c, d) = f(a)f(b)f(c|a)f(d|a, b),$$

donde

$$
\begin{aligned}
&f(a) \sim N\left(\mu_A, v_A\right), \\
&f(b) \sim N\left(\mu_B, v_B\right), \\
&f(c) \sim N\left(\mu_C + \beta_{CA}(a - \mu_A), v_C\right), \\
&f(d) \sim N\left(\mu_D + \beta_{DA}(a - \mu_A) + \beta_{DB}(b - \mu_B), v_D\right).
\end{aligned}
$$

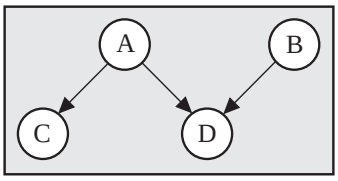


FIGURA 8.41. Ejemplo de un grafo dirigido acíclico utilizado para definir una red Bayesiana normal.

Por ejemplo, cuando todas las medias son cero, todas las varianzas son uno, y los parámetros de regresión son $\beta_{CA} = 1$, $\beta_{DA} = 0.2$ y $\beta_{DB} = 0.8$, se tiene

$$
\begin{aligned}
&f(a) \sim N\left(0, 1\right), \\
&f(b) \sim N\left(0, 1\right), \\
&f(c) \sim N\left(a, 1\right), \\
&f(d) \sim N\left(0.2a + 0.8b, 0\right).
\end{aligned}
$$

Este conjunto proporciona una de las dos representaciones alternativas de una red Bayesiana normal. La segunda representación utiliza la matriz de covarianzas del modelo. Esta matriz se puede obtener a partir de las funciones de probabilidad condicionada anteriores aplicando (6.40) (ver el Ejemplo 6.18)

$$
\Sigma = \begin{pmatrix} 1.0 & 0.0 & 1.0 & 0.20 \\ 0.0 & 1.0 & 0.0 & 0.80 \\ 1.0 & 0.0 & 2.0 & 0.20 \\ 0.2 & 0.8 & 0.2 & 1.68 \end{pmatrix}. \tag{8.54}
$$

Por tanto, se puede aplicar el algoritmo mostrado en la Figura 8.40 a la matriz de covarianzas (8.54) y el vector de medias $\mu = (0, 0, 0, 0)$ para propagar evidencia en la red Bayesiana normal. Supóngase que se tiene la evidencia $\{A = 1, B = 3, C = 2\}$. La Figura 8.42 muestra un programa de *Mathematica* que implementa este algoritmo iterativo para calcular el nuevo vector de medias y la nueva matriz de covarianzas en los distintos pasos del proceso.

En el primer paso de la iteración, se considera el primer nodo evidencial $A = 1$. En este caso, se obtiene el vector de medias y la matriz de

```
(* Definición de la función de probabilidad *)
M={0,0,0,0};
V={{1.0, 0.0, 1.0, 0.2},
   {0.0, 1.0, 0.0, 0.8},
   {1.0, 0.0, 2.0, 0.2},
   {0.2, 0.8, 0.2, 1.68}};
(* Nodos y evidencia *)
X={A,B,C,D};
Ev={A,B,C};
ev={1,3,2};
(* Actualización de M y V *)
NewM=Transpose[List[M]];
NewV=V;
For[k=1, k<=Length[Ev], k++,
(* Posición del elemento i-ésimo de E[[k]] en X *)
    i=Position[X,Ev[[k]]][[1,1]];
    My=Delete[NewM,i];
    Mz=NewM[[i,1]];
    Vy=Transpose[Delete[Transpose[Delete[NewV,i]],i]];
    Vz=NewV[[i,i]];
    Vyz=Transpose[List[Delete[NewV[[i]],i]]];
    NewM=My+(1/Vz)*(ev[[k]]-Mz)*Vyz;
    NewV=Vy-(1/Vz)*Vyz.Transpose[Vyz];
(* Eliminar el elemento i-ésimo *)
    X=Delete[X,i];
(* Imprimir los resultados *)
    Print["Iteracion = ",k];
    Print["Nodos restantes = ",X];
    Print["M = ",Together[NewM]];
    Print["V = ",Together[NewV]];
    Print["--------------------"];
]
```

FIGURA 8.42. Un programa de *Mathematica* para la propagación exacta de evidencia en una red Bayesiana normal.

covarianzas para el resto de los nodos, $Y = \{B, C, D\}$:

$$\mu_{Y|A=1} = \begin{pmatrix} 0.0 \\ 1.0 \\ 0.2 \end{pmatrix}; \; \Sigma_{Y|A=1} = \begin{pmatrix} 1.0 & 0.0 & 0.80 \\ 0.0 & 1.0 & 0.00 \\ 0.8 & 0.0 & 1.64 \end{pmatrix}.$$

En el segundo paso del algoritmo se añade la evidencia $B = 3$; en este caso, se obtiene el vector de medias y la matriz de covarianzas para los nodos

$Y = \{C, D\}$:

$$\mu_{Y|A=1,B=3} = \begin{pmatrix} 1.0 \\ 2.6 \end{pmatrix};\ \Sigma_{Y|A=1,B=3} = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}.$$

A partir de la estructura de esta matriz de covarianzas se puede observar que las dos variables restantes son independientes dadas las variables evidenciales A y B (este hecho también puede comprobarse en el grafo de la Figura 8.41 utilizando el criterio de D-separación). Por tanto, el último paso de la iteración no modifica ninguno de los valores anteriores. Finalmente, considerando la evidencia $C = 2$ se tiene la media y varianza para el nodo D dados por $\mu_{D|A=1,B=3,C=2} = 2.6$, $\sigma_{D|A=1,B=3,C=2} = 1.0$. ■

Ejercicios

8.1 Considérese la red Bayesiana definida por el grafo dirigido de la Figura 8.43 y las funciones de probabilidad condicionada dadas en la Tabla 8.11.

(a) Aplicar el algoritmo de poliárboles 8.1 para obtener las funciones de probabilidad marginales de los nodos.

(b) Hallar los mensajes y el orden en el que han de ser calculados y enviados en el proceso de propagación.

(c) Supóngase que se tiene la evidencia $\{D = 0, C = 1\}$. Repetir los cálculos anteriores para obtener las funciones de probabilidad condicionada de los nodos, dada la evidencia.

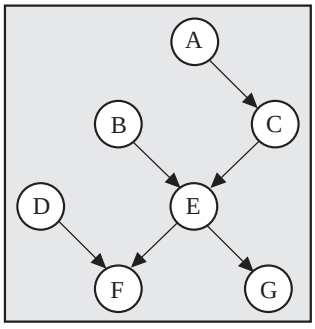


FIGURA 8.43. Grafo dirigido.

8.2 Repetir el ejercicio anterior utilizando el poliárbol dado en la Figura 8.44 y las funciones de probabilidad condicionada dadas en la Tabla 8.12.

a	$p(a)$
0	0.3
1	0.7

b	$p(b)$
0	0.6
1	0.4

e	$p(d)$
0	0.7
1	0.3

a	c	$p(c\|a)$
0	0	0.15
0	1	0.85
1	0	0.25
1	1	0.75

e	g	$p(g\|e)$
0	0	0.10
0	1	0.90
1	0	0.30
1	1	0.70

b	c	e	$p(e\|b,c)$
0	0	0	0.40
0	0	1	0.60
0	1	0	0.45
0	1	1	0.55
1	0	0	0.60
1	0	1	0.40
1	1	0	0.30
1	1	1	0.70

d	e	f	$p(f\|d,e)$
0	0	0	0.25
0	0	1	0.75
0	1	0	0.60
0	1	1	0.40
1	0	0	0.10
1	0	1	0.90
1	1	0	0.20
1	1	1	0.80

TABLA 8.11. Funciones de probabilidad condicionada para la red Bayesiana asociada al grafo de la Figura 8.43.

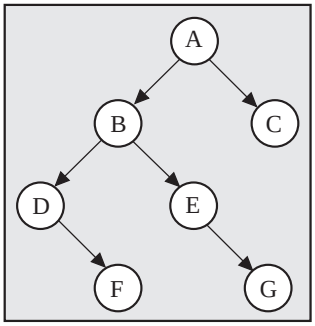


FIGURA 8.44. Un poliárbol.

8.3 Repetir los cálculos del Ejemplo 8.4 utilizando el grafo de la Figura 8.19(b).

8.4 Dado el grafo múltiplemente conexo de la Figura 8.45 y las funciones de probabilidad condicionada correspondientes mostradas en la Tabla 8.11 (reemplazando las funciones $p(b)$ y $p(d)$ por las funciones $p(b|a)$ y $p(d|b)$ dadas la Tabla 8.13),

a	$p(a)$
0	0.3
1	0.7

a	b	$p(b\|a)$
0	0	0.20
0	1	0.80
1	0	0.40
1	1	0.60

a	c	$p(c\|a)$
0	0	0.15
0	1	0.85
1	0	0.25
1	1	0.75

b	d	$p(d\|b)$
0	0	0.30
0	1	0.70
1	0	0.65
1	1	0.35

b	e	$p(e\|b)$
0	0	0.25
0	1	0.75
1	0	0.45
1	1	0.55

d	f	$p(f\|d)$
0	0	0.90
0	1	0.10
1	0	0.25
1	1	0.75

e	g	$p(g\|e)$
0	0	0.10
0	1	0.90
1	0	0.30
1	1	0.70

TABLA 8.12. Funciones de probabilidad condicionada para la red Bayesiana asociada al poliárbol de la Figura 8.44.

(a) Obtener un *conjunto de corte* que transforme este grafo en el poliárbol de la Figura 8.43 y utilizar el algoritmo de condicionamiento para obtener las funciones de probabilidad marginal (probabilidades iniciales) de los nodos, y las probabilidades condicionadas cuando se considera la evidencia $\{D = 0, C = 1\}$.

(b) ¿Es el conjunto obtenido un conjunto de corte minimal? Calcular todos los conjuntos de corte posibles para este grafo.

(c) Aplicar el algoritmo de condicionamiento con alguno de los conjuntos de corte minimales obtenidos.

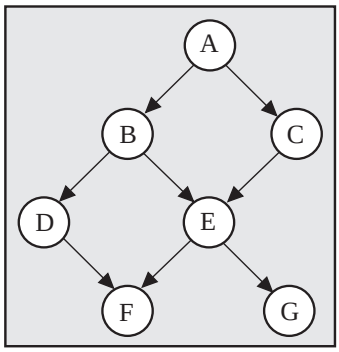


FIGURA 8.45. Un grafo dirigido múltiplemente conexo.

a	b	$p(b\|a)$
0	0	0.25
0	1	0.75
1	0	0.45
1	1	0.55

b	d	$p(d\|b)$
0	0	0.20
0	1	0.80
1	0	0.70
1	1	0.30

TABLA 8.13. Funciones de probabilidad condicionada para los nodos B y D asociadas al grafo de la Figura 8.45.

8.5 Considérese el grafo triangulado de la Figura 8.46.

(a) Obtener una cadena de conglomerados utilizando el Algoritmo 4.3.

(b) Encontrar los conjuntos separadores y residuales, así como los conjuntos de conglomerados vecinos.

(c) Construir un árbol de unión. ¿Es un árbol de familias para el grafo de la Figura 8.45?

(d) Calcular las funciones potenciales correspondientes a partir de las funciones de probabilidad condicionada asociadas a la red Bayesiana del ejercicio anterior.

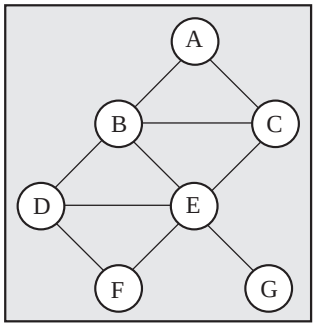


FIGURA 8.46. Un grafo dirigido múltiplemente conexo.

8.6 Considérese la cadena de conglomerados y las funciones potenciales del ejercicio anterior.

(a) Calcular las funciones de probabilidad iniciales de los nodos utilizando el Algoritmo 8.2 (algoritmo de agrupamiento).

(b) Calcular las funciones de probabilidad iniciales de los nodos utilizando el Algoritmo 8.4 y el árbol de unión obtenido en el ejercicio anterior.

(c) Considérese la evidencia $\{D = 0, C = 1\}$. ¿Cuántas posibilidades existen para absorber esta evidencia en la representación potencial?

(d) Dada la evidencia anterior, calcular las funciones de probabilidad condicionada de los nodos utilizando el Algoritmo 8.2.

(e) Dada la evidencia anterior, calcular las funciones de probabilidad condicionada de los nodos utilizando el Algoritmo 8.4.

8.7 Considérese el grafo mostrado en la Figura 8.47(a). Calcular el conjunto de corte y el árbol de unión más eficientes para este grafo. ¿Qué método de propagación es más conveniente en este caso?

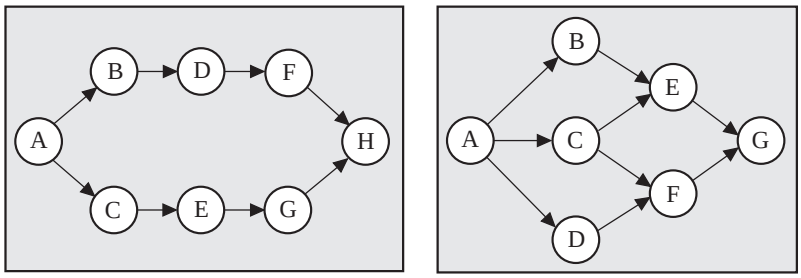


FIGURA 8.47. Dos grafos múltiplemente conexos.

8.8 Repetir el ejercicio anterior considerando el grafo de la Figura 8.47(b).

8.9 Utilizar el Algoritmo 8.4 para calcular los mensajes mostrados en la Figura 8.34.

8.10 Considérese el grafo mostrado en la Figura 8.36.

(a) Construir el grafo ancestral moralizado que muestra que D D-separa Θ_E de C.

(b) Construir el grafo ancestral moralizado que muestra que D D-separa Θ_F de C.

8.11 Dado el grafo dirigido acíclico de la Figura 8.43, y suponiendo que E es el nodo objetivo,

(a) Aplicar el Algoritmo 8.6 para reducir el grafo a aquel que únicamente contiene al conjunto relevante de nodos.

(b) Obtener los valores numéricos para las funciones de probabilidad condicionada a partir de los valores dados en la Tabla 8.11.

(c) Utilizar un método adecuado para calcular las funciones de probabilidad iniciales del nodo E.

8.12 Repetir el ejercicio anterior utilizando el grafo de la Figura 8.45 y las funciones de probabilidad condicionada dadas en la Tabla 8.12.

8.13 Supuesto que $(X_1, X_2, X_3, X_4, X_5)$ es un vector de variables aleatorias normales con vector de medias y matriz de covarianzas dados por

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} \qquad \text{y} \qquad \Sigma = \begin{pmatrix} 1 & 0.3 & 0 & 0.4 & 0 \\ 0.3 & 1 & 0 & 0.2 & 0 \\ 0 & 0 & 1 & 0 & 0.1 \\ 0.4 & 0.2 & 0 & 1 & 0 \\ 0 & 0 & 0.1 & 0 & 1 \end{pmatrix},$$

aplicar el algoritmo de propagación exacta en redes Gausianas (ver Figura 8.40) para calcular las probabilidades iniciales de los nodos, y las funciones de probabilidad condicionada, dada la evidencia $X_2 = 1$ y $X_4 = 2$.

Capítulo 9
Métodos de Propagación Aproximada

9.1 Introducción

En el Capítulo 8, se han presentado varios algoritmos para la propagación exacta de la evidencia en redes de Markov y Bayesianas. Sin embargo, estos métodos presentan algunos problemas. Por una parte, algunos de ellos no son aplicables a todos los tipos de estructuras de red. Por ejemplo, el algoritmo para poliárboles (Sección 8.3) se aplica sólo a las redes con estructura simple de poliárbol. Por otra parte, los métodos de validez general, que sirven para propagar evidencia en cualquier red Bayesiana o de Markov, se hacen cada vez más ineficientes, en ciertos tipos de estructuras, al crecer el número de nodos y su complejidad. Por ejemplo, los algoritmos basados en el condicionamiento (Sección 8.5) experimentan una explosión combinatoria cuando se tratan grandes conjuntos de corte, y los métodos de agrupamiento (Sección 8.6) requieren construir un árbol de unión, que puede ser una tarea cara en términos computacionales. Estos métodos también sufren esa explosión combinatoria cuando se tratan redes con grandes conglomerados. Esto no es sorprendente puesto que se ha demostrado que la tarea de la propagación exacta es NP-compleja (véase Cooper (1990)). Por ello, desde un punto de vista práctico, los métodos de propagación exacta pueden ser muy restrictivos e incluso ineficientes en situaciones en las que el tipo de estructura de la red requiere un gran número de cálculos y una gran capacidad de memoria y potencia computacional.

En este capítulo se introducen algunos métodos aproximados para propagar la evidencia, que son aplicables a todo tipo de estructuras de red. Se

recuerda al lector que por *algoritmos de propagación aproximada* nos referimos a algoritmos que calculan las probabilidades condicionales de los nodos de forma aproximada. La idea básica de estos métodos consiste en generar una muestra de tamaño N, a partir de la función de probabilidad conjunta de las variables, y luego utilizar la muestra generada para calcular valores aproximados de las probabilidades de ciertos sucesos dada la evidencia. Las probabilidades se aproximan mediante los cocientes de la frecuencia de aparición de los sucesos en la muestra y el tamaño de la muestra.

Los métodos de propagación aproximada pueden clasificarse en dos tipos: métodos de simulación estocástica y métodos de búsqueda determinista. Los métodos del primer grupo generan la muestra a partir de la función de probabilidad conjunta usando algunos mecanismos aleatorios, mientras que los de búsqueda determinista generan la muestra de forma sistemática.

La Sección 9.2 introduce los fundamentos intuitivos y teóricos de los métodos de simulación. La Sección 9.3 presenta una metodología general para estos métodos. Las Secciones 9.4 a 9.8 presentan cinco métodos de simulación estocástica: muestreo de aceptación-rechazo, muestreo uniforme, verosimilitud pesante, muestreo hacia adelante y hacia atrás, y el muestreo de Markov. Las Secciones 9.9 y 9.10 introducen el muestreo sistemático y el método de búsqueda de la máxima probabilidad, que pertenecen al grupo de los métodos de búsqueda determinista. Finalmente, en la Sección 9.11 se analiza la complejidad de la propagación aproximada.

9.2 Base Intuitiva de los Métodos de Simulación

En esta sección se ilustra un esquema general de simulación mediante un sencillo ejemplo. Considérese una urna que contiene seis bolas numeradas $\{1, \ldots, 6\}$. Supóngase que se quiere realizar el siguiente experimento. Se selecciona una bola al azar de la urna, se apunta su número, se devuelve a la urna, y se mezclan las bolas antes de proceder a extraer la bola siguiente. Este esquema de muestreo se denomina *muestreo con reemplazamiento*. Cada selección de una bola se llama una *extracción* o un *experimento*. En este caso cada extración tiene seis posibles resultados, $\{1, \ldots, 6\}$.

Sea X_i el resultado (el número de la bola) de la extracción i-ésima. Puesto que el muestreo se hace con reemplazamiento, las extracciones son independientes (el resultado de una extracción no influye en el resultado de las demás). Claramente, X_i es una variable *uniforme* con función de probabilidad $p(X_i = x_i) = 1/6$, para $x_i = 1, \ldots, 6$ y $i = 1, \ldots, N$, donde N es el número de extracciones (el tamaño de la muestra).

Como en los capítulos anteriores, por simplicidad, se representa $p(X_i = x_i)$ por $p(x_i)$. En este caso, la función de probabilidad conjunta, $p(x)$, de $X = \{X_1, \ldots, X_n\}$, es el producto de las probabilidades individuales, es

decir,

$$p(x) = \prod_{i=1}^{n} p(x_i). \tag{9.1}$$

Utilizando esta función de probabilidad conjunta, se pueden calcular las probabilidades exactas de ciertos sucesos tales como $p(X_1 = 1, \ldots, X_n = 1)$, $p(\textit{número de pares} = \textit{número de impares})$, etc. Estos cálculos son fáciles en este caso puesto que la distribución es uniforme (hay exactamente una bola para cada uno de los números $\{1, \ldots, 6\}$), las extracciones son idénticas (se usa la misma urna), y el resultado de cada extracción es independiente de los resultados de los demás (muestreamos con reemplazamiento). Los cálculos de las probabilidades exactas son complicados y costosos cuando la distribución no es uniforme (por ejemplo, se tiene distinto número de bolas de diferentes tipos), las extracciones no son idénticas (por ejemplo, se realiza un muestreo con diferentes números de bolas), y/o extracciones que no son independientes (por ejemplo, muestreo sin reemplazamiento).

En estas situaciones complicadas, se pueden calcular las probabilidades de ciertos sucesos de forma aproximada mediante técnicas de simulación. Se puede, por ejemplo, repetir un experimento N veces. Se obtiene lo que se llama una *muestra* de tamaño N. Entonces, la probabilidad de un suceso puede aproximarse por el cociente entre el número de veces que ocurre dicho suceso y el número total de simulaciones N. Claramente, cuanto mayor sea el tamaño de la muestra, más aproximada será la aproximación. Los métodos que se presentan en este capítulo difieren principalmente en la forma de generar la muestra a partir de la función de probabilidad conjunta de las variables.

Una forma equivalente de obtener una muestra de tamaño N con reemplazamiento de la Urna 1 es mediante el lanzamiento de un dado N veces. Sea X el número de bolas elegidas al azar de la Urna 1 e Y_i el número observado al lanzar un dado. Entonces Y tiene la misma función de probabilidad que X. Sea $p(x)$ la función de probabilidad de X y $h(y)$ la función de probabilidad de Y; entonces $p(x) = h(x)$, tal como se muestra en la Figura 9.1. Por ello, extraer N bolas con reemplazamiento de la Urna 1 puede simularse mediante el lanzamiento de un dado N veces.

Es útil distinguir en este momento entre la función de probabilidad $p(x)$ (generada por la Urna 1) y la función de probabilidad $h(y)$ (generada por el dado). Nos referiremos a la distribución de X, que es la distribución de la que se quiere obtener la muestra, como la *distribución de la población.* la distribución de Y se llama la *distribución simulada* puesto que es la que se utiliza para generar (simular) la muestra. En este caso, la distribución simulada es la misma que la de la población. Pero se verá más tarde que la distribución simulada puede ser diferente de la distribución de la población.

La razón para usar un dado (una distribución simulada) para simular la extracción de bolas de la Urna 1 (una distribución de la población) es

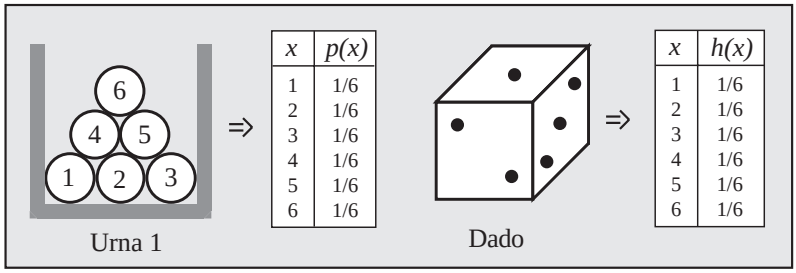


FIGURA 9.1. Simulando la extracción de bolas con reemplazamiento de la Urna 1 mediante un dado.

que es más fácil lanzar el dado que extraer la bola de una urna, devolverla y mezclar las bolas antes de la extracción siguiente. En otras palabras, si no es fácil obtener muestras de la distribución de la población se debe elegir otra distribución que resulte más sencilla para la simulación. Pero, ¿se puede hacer ésto siempre? Por ejemplo, se puede utilizar un dado para simular la extracción de bolas de urnas con diferentes números de bolas? La respuesta, afortunadamente, es positiva. Por ejemplo, supóngase que la urna contiene sólo cinco bolas numeradas $\{1, \ldots, 5\}$ tal como muestra la Figura 9.2 (Urna 2). Sea X el número de bolas con el número i sacadas al azar con reemplazamiento de la Urna 2. Entonces X es una variable aleatoria cuya función de probabilidad, $p(x)$, se muestra en la Figura 9.2 (Urna 2). En este caso, la distribución simulada (el dado) no es la misma que la distribución de la población (Urna 2), es decir, $p(x) \neq h(x)$ (las columnas etiquetadas $s(x)$ se explicarán en breve). A pesar del hecho de que la Urna 2 y el dado no tienen la misma distribución, se puede todavía utilizar el dado para simular la extracción de bolas de la Urna 2, pero se tiene que corregir por el hecho de que las distribuciones de la población y la simulada no coinciden.

Una forma de tener en cuenta esta diferencia es la siguiente: cuando en el dado sale un 6, se ignora la tirada y se repite de nuevo hasta que salga un valor menor que 6, en cuyo caso se hace y igual al número que salga y se toma y como valor generado de la población $p(x)$. Este ejemplo es en realidad un caso especial del método conocido como *método de aceptación-rechazo*. Los fundamentos teóricos se presentan en el teorema siguiente, que se debe a Von Neumann (1951) (véase también Rubinstein (1981), Devroye (1986), y Ripley (1987)).

Teorema 9.1 El método de aceptación-rechazo. *Sea X una variable aleatoria con función de probabilidad $p(x)$. Supóngase que $p(x)$ puede ser expresada como*

$$p(x) = c\, g(x)\, h(x), \tag{9.2}$$

donde $c \geq 1, 0 \leq g(x) \leq 1$ y $h(x)$ es una función de probabilidad. Sea U una variable aleatoria uniforme $U(0,1)$ y sea Y una variable aleatoria con

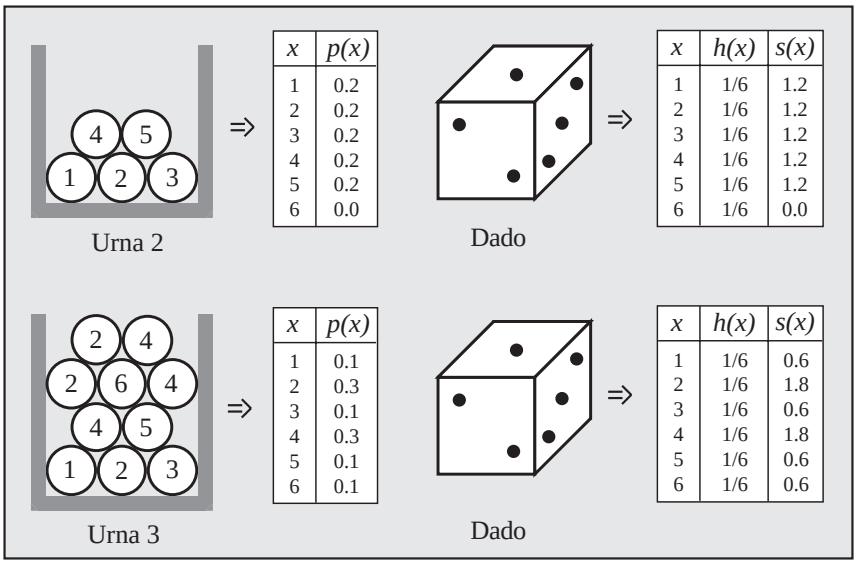


FIGURA 9.2. Una ilustración de un esquema general de simulación.

función de probabilidad $h(y)$ independiente de U. Entonces, la función de probabilidad condicional de Y dado que $u \leq g(y)$ coincide con la función de probabilidad de X. Por otra parte, la probabilidad de aceptar la muestra (eficiencia) es $1/c$.

Por ejemplo, en el caso de la Urna 2 que se muestra en la Figura 9.2, se puede escribir $p(x) = cg(x)h(x)$, donde $p(x)$ y $h(x)$ se muestran en la Figura 9.2, $c = 6/5$ y

$$g(x) = \begin{cases} 0, & \text{si } x = 6, \\ 1, & \text{en otro caso.} \end{cases} \tag{9.3}$$

Por ello, utilizando el teorema anterior, se puede obtener una muestra de $p(x)$ (Urna 2) usando $h(x)$ (el dado) y comprobando la condición $u \leq g(x)$ para todo valor x que se simule de $h(x)$, donde u es un número obtenido de la distribución uniforme $U(0, 1)$. Por tanto, en este caso, el suceso $x = 6$ siempre se rechaza, ya que $g(6) = 0$, y los restantes sucesos se aceptan siempre.

Antes de ilustrar el método de aceptación-rechazo mediante un ejemplo, se describe cómo se simula una muestra procedente de una determinada población con función de probabilidad $h(x)$.

Ejemplo 9.1 Simulando muestras de una población dada. Para generar una muestra procedente de una determinada función de proba-

bilidad $h(x)$, se calcula en primer lugar la función de distribución,

$$H(x) = p(X \leq x) = \int_{-\infty}^{x} h(x)dx.$$

Seguidamente, se genera una sucesión de números aleatorios $\{u_1, \ldots, u_N\}$ de una distribución uniforme $U(0,1)$ y se obtienen los valores correspondientes $\{x_1, \ldots, x_N\}$ resolviendo la ecuación $H(x_i) = u_i$, $i = 1 \ldots, N$, que da $x_i = H^{-1}(u_i)$, donde $H^{-1}(u_i)$ es la inversa de la función de distribución evaluada en el punto u_i. Por ejemplo, la Figura 9.3 muestra la función de distribución $H(x)$ y los dos valores x_1 y x_2 correspondientes a los números u_1 y u_2 procedentes de la población uniforme $U(0,1)$. ∎

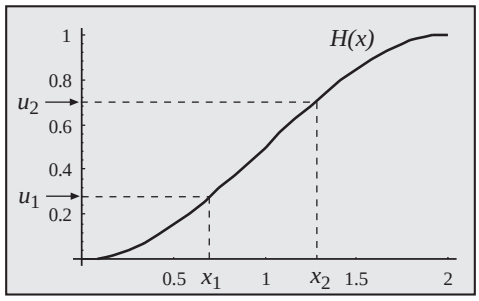


FIGURA 9.3. Generando muestras de una población con función de densidad $h(x)$ utilizando la función de distribución $H(x)$.

Ejemplo 9.2 El método de aceptación-rechazo. Supóngase que $p(x) = 3x(2-x)/4$, $0 \leq x \leq 2$, y que se desea obtener una muestra de tamaño N de un población con función de densidad $p(x)$. En este caso, $p(x)$ puede factorizarse en la forma (9.2) tomando $h(x) = 1/2$, $0 \leq x \leq 2$, $g(x) = x(2-x)$ y $c = 3/2$. Estas funciones se muestran en la Figura 9.4(a). La función de distribución correspondiente a $h(x)$ es

$$H(x) = \int_{0}^{x} h(x)dx = \int_{0}^{x} (1/2)dx = x/2.$$

Nótese que esta función es más fácil de simular que la de la población con función de distribución $P(x)$. Las funciones $P(x)$, $H(x)$ y $g(x)$ se muestran en la Figura 9.4(b).

Supóngase que se genera un número aleatorio y de $h(y)$ y un número aleatorio u de $U(0,1)$. Se calcula $g(y)$, y si $u \leq g(y)$, se acepta y como número procedente de $p(x)$. En otro caso, se rechazan ambos u y y, y se repite el proceso de nuevo. Por ejemplo, si $y = 1.5$, entonces $g(1.5) = 0.75$,

lo que significa que la probabilidad de aceptar $y = 1.5$ como un valor aleatorio generado con $p(x)$ es 0.75. Por otra parte, si $y = 1$, $g(1) = 1$, lo que significa que la probabilidad de aceptar $y = 1$ como número aleatorio generado con $p(x)$ es 1. Esto no debe sorprender, puesto que, como se puede ver en la Figura 9.4(b), cuando $x = 1$, $P(x) = H(x)$. También puede verse que $g(x)$ alcanza el máximo para $x = 1$, ya que, a medida que y se aleja de 1, la probabilidad de aceptar el valor simulado decrece. La probabilidad de aceptar un número aleatorio generado con $h(y)$ es igual a $1/c = 2/3$. ∎

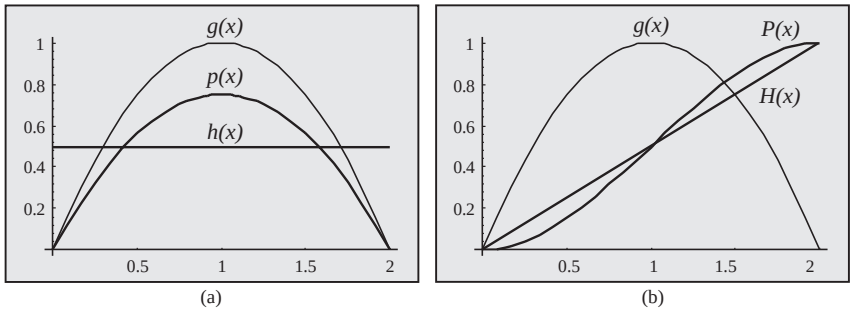


FIGURA 9.4. Una ilustración del método de aceptación-rechazo.

Según el Teorema 9.1, la probabilidad de aceptar es $1/c$. Por ello, la probabilidad de aceptar es alta cuando c está cercano a 1, y esto sucede normalmente cuando $h(x)$ es parecida a $p(x)$. Por ello, por una parte, se quiere que $h(x)$ esté tan próxima a $p(x)$ como sea posible de forma que la probabilidad de aceptar sea alta. Por otra parte, se quiere que $h(x)$ sea tan fácil de simular como sea posible.

El Teorema 9.1 sugiere el algoritmo siguiente para generar una muestra aleatoria de tamaño N de $p(x)$ pero utilizando $h(x)$.

Algoritmo 9.1 El Método de rechazo.

- **Datos:** Función de probabilidad de la población $p(x)$, función de probabilidad simulada $h(x)$, y el tamaño de muestra N.
- **Resultados:** Una muestra $\{x_1, \ldots, x_N\}$ procedente de $p(x)$.

1. Hacer $i = 1$.
2. Generar un valor aleatorio u de la distribución $U(0, 1)$.
3. Generar un valor aleatorio y de $h(y)$.
4. Si $u \leq g(y)$, hacer $x_i = y$. En otro caso ir a la Etapa 2.
5. Si $i < N$, aumentar i en una unidad e ir a la Etapa 2. En otro caso devolver $\{x_1, \ldots, x_N\}$. ∎

Cuando c es grande, la eficiencia del algoritmo anterior es baja y el porcentaje de valores rechazados es alto. Consecuentemente, hay que generar un número muy alto de valores aleatorios de $h(y)$ para obtener una muestra válida pequeña de $p(x)$. Sin embargo, el algoritmo de aceptación-rechazo, se puede hacer mucho más eficiente con la modificación siguiente. Escribir $p(x)$ en la forma

$$p(x) = \frac{p(x)}{h(x)} h(x) = s(x)h(x), \tag{9.4}$$

donde

$$s(x) = \frac{p(x)}{h(x)}, \tag{9.5}$$

es una *función de peso*. Por ello, el peso del suceso x es el cociente entre la probabilidad real, $p(x)$, y la probabilidad simulada, $h(x)$.

De (9.4) y (9.5), se ve que $s(x) = c\, g(x)$, es decir, el peso es proporcional a $g(x)$. Por tanto, en vez de rechazar un número x que se ha generado de $h(x)$, se le asigna una probabilidad proporcional a $s(x)$ o $g(x)$. Al final de las simulaciones se normalizan los pesos (dividiendo cada peso por la suma de todos ellos) y se utilizan los pesos normalizados para estimar la probabilidad de cualquier suceso de interés. Esto conduce a un aumento considerable de la eficiencia del proceso.

Por ejemplo, los pesos correspondientes a nuestro ejemplo del dado se muestran en la Figura 9.2. Nótese que en el caso de la Urna 2 el peso asociado al suceso $x = 6$ es cero, y el peso asociado al resto de sucesos es el mismo. Por ello, en este caso, el uso de los pesos produce el mismo resultado que la aplicación del método de aceptación-rechazo. Sin embargo, la situación de la Urna 3 es más complicada, y el método de los pesos es más eficiente que el método de aceptación-rechazo.

9.3 Metodología General para los Métodos de Simulación

La discusión anterior sugiere un marco general para los métodos de simulación. Sea $X = \{X_1, \ldots, X_n\}$ un conjunto de variables con función de probabilidad conjunta $p(x)$. Supóngase que el subconjunto E de las variables de X toman valores conocidos e. Las variables de E se llaman entonces variables *evidenciales* o se dice que constituyen la *evidencia* y a las restantes variables se les llama variables *no evidenciales*. Nuestro objetivo consiste en calcular la función de probabilidad de cada variable no evidencial dada la evidencia. En general, dado un conjunto de variables $Y \subset X$, se desea calcular la probabilidad condicional de y dada e, es decir, los valores de las variables evidenciales. Esta probabilidad condicional puede

ser escrita como

$$p(y|e) = \frac{p_e(y)}{p(e)} \propto p_e(y), \tag{9.6}$$

donde $p(e)$ es la probabilidad de e y

$$p_e(y) = \begin{cases} p(y \cup e), & \text{si } y \text{ es consistente con } e, \\ 0, & \text{en otro caso.} \end{cases} \tag{9.7}$$

Nótese que si $Y \cap E = \phi$, entonces $p_e(y) = p(y, e)$. Se puede ver de (9.6) que para calcular $p(y|e)$, sólo se necesita calcular y normalizar $p_e(y)$. Nótese también que cuando $E = \phi$ (no hay evidencia disponible), $p(y|e) = p(y)$ es simplemente la probabilidad marginal de Y.

Tal como se ha mencionado previamente, el cálculo exacto de $p(x_i|e)$ puede ser muy costoso computacionalmente y a menudo imposible. Por ello, es necesario recurrir a los métodos aproximados para obtener una estimación de $p(x_i|e)$. En primer lugar se genera una muestra de tamaño N de $p(x)$ pero utilizando una función de densidad diferente $h(x)$. Seguidamente, se calculan y se normalizan los pesos. Entonces, la probabilidad $p(x_i|e)$ puede aproximarse mediante la suma de todos los pesos de las realizaciones que son consistentes con los dos sucesos x_i y e.

Tras obtener una muestra de tamaño N, y disponer de las realizaciones, $x^j = \{x_1^j, \ldots, x_n^j\}$, $j = 1, \ldots, N$, la probabilidad condicional de cualquier subconjunto $Y \subset X$ dada la evidencia $E = e$ se estima por la suma normalizada de los pesos de todas las realizaciones en las que ocurre y, es decir

$$p(y) \approx \frac{\sum\limits_{y \in x^j} s(x^j)}{\sum\limits_{j=1}^{N} s(x^j)}. \tag{9.8}$$

El procedimiento anterior se implementa en el algoritmo que sigue:

Algoritmo 9.2 Metodología general de simulación.

- **Datos:** *Las funciones de probabilidad real $p(x)$ y la de simulación $h(x)$, el tamaño de la muestra N, y un subconjunto $Y \subset X$.*
- **Resultados:** *Una aproximación de $p(y)$ para todo valor posible y de Y.*

1. *Para $j = 1$ a N*
   - *Generar $x^j = (x_1^j, \ldots, x_n^j)$ usando $h(x)$.*
   - *Calcular $s(x^j) = \dfrac{p(x^j)}{h(x^j)}$.*
2. *Para todo valor posible y de Y, aproximar $p(y)$ utilizando (9.8).* ∎

La calidad de la aproximación obtenida usando el Algoritmo 9.2 depende de los factores siguientes:

- La función de probabilidad $h(x)$ elegida para obtener la muestra.
- El método usado para generar las realizaciones a partir de $h(x)$.
- El tamaño de muestra N.

La selección de la distribución de simulación influye decididamente en la calidad de la aproximación y es difícil pronunciarse sobre cuál o cuales son las más convenientes. Sin embargo, puede decirse que esquemas de muestreo en los que los pesos sean similares conducen a buenos resultados, mientras que los esquemas con pesos muy diferentes suelen conducir a malos resultados.

El ejemplo que sigue ilustra este marco general de simulación.

Ejemplo 9.3 Modelo de Red Bayesiana. Supóngase que el grafo dirigido de la Figura 9.5 se da como modelo gráfico para definir la estructura de independencia de la función de probabilidad conjunta de seis variables binarias $X = \{X_1, \ldots, X_6\}$ que toman valores en el conjunto $\{0, 1\}$. Este grafo define una red Bayesiana cuya función de probabilidad conjunta puede ser factorizada en la forma

$$p(x) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_2, x_3)p(x_6|x_3). \tag{9.9}$$

Los valores numéricos necesarios para especificar el conjunto de probabilidades condicionales requeridas para poder conocer la función de probabilidad conjunta, $p(x)$, se dan en la Tabla 9.1.

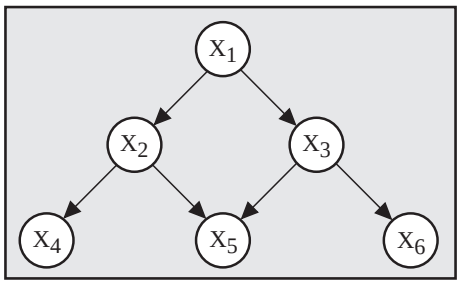


FIGURA 9.5. Ejemplo de grafo dirigido acíclico.

Se genera una muestra de N observaciones a partir de esta distribución. Cada observación $x^j = \{x_1^j, \ldots, x_6^j\}$ consiste en un valor observado para cada una de las seis variables en X. Una observación x^j se denomina una *realización*. Puesto que las seis variables son binarias, hay $2^6 = 64$ posibles realizaciones. Con el fin de ilustrar, supóngase que se seleccionan

x_1	$p(x_1)$
0	0.3
1	0.7

x_1	x_2	$p(x_2\|x_1)$
0	0	0.4
0	1	0.6
1	0	0.1
1	1	0.9

x_1	x_3	$p(x_3\|x_1)$
0	0	0.2
0	1	0.8
1	0	0.5
1	1	0.5

x_2	x_4	$p(x_4\|x_2)$
0	0	0.3
0	1	0.7
1	0	0.2
1	1	0.8

x_3	x_6	$p(x_6\|x_3)$
0	0	0.1
0	1	0.9
1	0	0.4
1	1	0.6

x_2	x_3	x_5	$p(x_5\|x_2, x_3)$
0	0	0	0.4
0	0	1	0.6
0	1	0	0.5
0	1	1	0.5
1	0	0	0.7
1	0	1	0.3
1	1	0	0.2
1	1	1	0.8

TABLA 9.1. Probabilidades condicionales requeridas para definir la función de probabilidad conjunta correspondiente a la red Bayesiana dada en la Figura 9.5.

cinco de las 64 realizaciones al azar y con reemplazamiento.[1] Esto implica que $h(x^j) = 1/64$; $j = 1, \ldots, 5$. Las realizaciones seleccionadas y sus correspondientes $p(x)$, $h(x)$ y $s(x)$ se dan en la Tabla 9.2.

Por tanto, basándose en las realizaciones de la Tabla 9.2, se aproxima la probabilidad de las variables sin más que normalizar los pesos dividiendo por su suma. La probabilidad deseada se obtiene sumando los pesos normalizados de todas las realizaciones en las que se da el suceso de interés.

[1] Hay que mencionar aquí que la aproximación basada en sólo cinco realizaciones será, casi con seguridad, poco aproximada. Por otra parte, tal como se verá más tarde en este capítulo, la elección de realizaciones uniformemente al azar no es necesariamente la elección óptima como distribución para simular.

Realización x^j	$p(x^j)$	$h(x^j)$	$s(x^j)$
$x^1 = (0, 1, 1, 1, 0, 0)$	0.0092	1/64	0.5898
$x^2 = (1, 1, 0, 1, 1, 0)$	0.0076	1/64	0.4838
$x^3 = (0, 0, 1, 0, 0, 1)$	0.0086	1/64	0.5529
$x^4 = (1, 0, 0, 1, 1, 0)$	0.0015	1/64	0.0941
$x^5 = (1, 0, 0, 0, 1, 1)$	0.0057	1/64	0.3629

TABLA 9.2. Cinco realizaciones obtenidas aleatoriamente de las 64 realizaciones posibles de las seis variables binarias.

Por ejemplo,

$$p(X_1 = 0) \approx \frac{s(x^1) + s(x^3)}{\sum_{i=1}^{5} s(x^i)} = \frac{0.5898 + 0.5529}{2.0835} = 0.5485$$

y

$$p(X_2 = 0) \approx \frac{s(x^3) + s(x^4) + s(x^5)}{\sum_{i=1}^{5} s(x^i)} = \frac{0.5529 + 0.0941 + 0.3629}{2.0835} = 0.4847.$$

Estos valores se obtienen porque $X_1 = 0$ aparece en las realizaciones x^1 y x^3, mientras que $X_2 = 0$ aparece en las realizaciones x^3, x^4 y x^5. Las probabilidades condicionales para otras variables pueden calcularse de forma análoga.

Las probabilidades de las que hemos hablado hasta ahora son las marginales de los nodos, es decir, en el caso de que no haya evidencia disponible. Si se dispone de alguna evidencia, se procede de forma análoga. Con el fin de ilustrar cómo, supóngase que $E = \{X_3, X_4\}$ son los nodos evidenciales y que los valores que toman son $e = \{X_3 = 1, X_4 = 1\}$. Puesto que los valores de los nodos X_3 y X_4 ya se han fijado, se tienen sólo $2^4 = 16$ realizaciones posibles. De nuevo, supóngase que se seleccionan al azar cinco de las 16 realizaciones con reemplazamiento. Esto implica que $h(x^j) = 1/16$; $j = 1, \ldots, 5$. Las realizaciones seleccionadas y sus correspondientes $p(x)$, $h(x)$ y $s(x)$ se dan en la Tabla 9.3.

Basándose en las realizaciones de la Tabla 9.3, se aproximan las probabilidades condicionales de las variables no evidenciales dada la evidencia utilizando los pesos normalizados. Por ejemplo,

$$p(X_1 = 0 | X_3 = 1, X_4 = 1) \approx \frac{s(x^1) + s(x^4)}{\sum_{i=1}^{5} s(x^i)} = \frac{0.2212 + 0.5898}{1.3296} = 0.6099$$

Realización x^j	$p(x^j)$	$h(x^j)$	$s(x^j)$
$x^1 = (0, 1, 1, 1, 0, 1)$	0.0138	1/16	0.2212
$x^2 = (1, 0, 1, 1, 0, 0)$	0.0049	1/16	0.0784
$x^3 = (1, 0, 1, 1, 1, 1)$	0.0073	1/16	0.1176
$x^4 = (0, 1, 1, 1, 1, 0)$	0.0369	1/16	0.5898
$x^5 = (1, 1, 1, 1, 0, 0)$	0.0202	1/16	0.3226

TABLA 9.3. Cinco realizaciones obtenidas al azar entre las 16 possibles de las seis variables binarias cuando los nodos evidenciales toman los valores $X_3 = 1$ y $X_4 = 1$.

y

$$p(X_2 = 0|X_3 = 1, X_4 = 1) \approx \frac{s(x^2) + s(x^3)}{\sum_{i=1}^{5} s(x^i)} = \frac{0.0784 + 0.1176}{1.3296} = 0.1474.$$

Esto resulta porque $X_1 = 0$ aparece en las realizaciones x^1 y x^4, mientras que $X_2 = 0$ aparece en las realizaciones x^2 y x^3. Las probabilidades condicionales de las restantes variables pueden calcularse de forma similar.

Merece la pena hacer notar que las realizaciones obtenidas pueden ser usadas para aproximar no sólo la distribución de probabilidad univariada sino también las multivariadas. Esto puede hacerse siempre que se almacenen las frecuencias de aparición de cada uno de los posibles valores de las variables discretas y todos los valores simulados de las continuas. Por ejemplo, utilizando las realizaciones de la Tabla 9.2, se puede calcular

$$\begin{aligned}
p(X_5 = 0, X_6 = 0) &\approx \frac{s(x^1)}{\sum_{i=1}^{5} s(x^i)} = \frac{0.5898}{2.0835} = 0.2831, \\
p(X_5 = 0, X_6 = 1) &\approx \frac{s(x^3)}{\sum_{i=1}^{5} s(x^i)} = \frac{0.5529}{2.0835} = 0.2654, \\
p(X_5 = 1, X_6 = 0) &\approx \frac{s(x^2) + s(x^4)}{\sum_{i=1}^{5} s(x^i)} = \frac{0.4838 + 0.0941}{2.0835} = 0.2773, \\
p(X_5 = 1, X_6 = 1) &\approx \frac{s(x^5)}{\sum_{i=1}^{5} s(x^i)} = \frac{0.3629}{2.0835} = 0.1742.
\end{aligned}$$

Nótese, sin embargo que el tamaño de muestra necesario para calcular las probabilidades con cierta precisión aumenta a medida que aumenta la dimensión del espacio muestral. Por ello, el tamaño de muestra requerido para

calcular las probabilidades multivariadas resulta necesariamente mayor que el tamaño de muestra requerido para calcular probabilidades univariadas con la misma precisión. ∎

Especialmente interesante es el caso en que ambas, la distribución de la población real y la de la simulada pueden factorizarse en la forma

$$p(x) = \prod_{i=1}^{n} p(x_i|s_i) \tag{9.10}$$

y

$$h(x) = \prod_{i=1}^{n} h(x_i|s_i), \tag{9.11}$$

donde $S_i \subset X$ es un subconjunto de variables y $h(x_i)$ es la distribución simulada del nodo X_i. En esta situación, el proceso de simulación puede simplificarse sin más que simular secuencialmente los nodos como, por ejemplo, en el caso del ejemplo del lanzamiento de un dado. Entonces, (9.10) y (9.11) permiten calcular el peso de una determinada realización $x = (x_1, \ldots, x_n)$ mediante el producto de los pesos de las variables

$$s(x) = \frac{p(x)}{h(x)} = \prod_{i=1}^{n} \frac{p(x_i|s_i)}{h(x_i|s_i)} = \prod_{i=1}^{n} s(x_i|s_i). \tag{9.12}$$

Nótese que todos los modelos de redes Bayesianas, todos los modelos de Markov descomponibles, y otros varios modelos probabilísticos pueden expresarse en la forma (9.10) (véase el Capítulo 6). Por razones de simplicidad, se consideran los modelos de redes Bayesianas para ilustrar las diferentes metodologías.

Cuando se sabe que un conjunto de nodos evidencial E toma los valores $E = e$, se puede utilizar también el Algoritmo 9.2 para calcular la función de probabilidad condicionada, $p(y|e)$, pero teniendo en cuenta la evidencia, es decir, que la distribución de la población es en este caso

$$p_e(x) \propto \prod_{i=1}^{n} p_e(x_i|\pi_i), \tag{9.13}$$

donde

$$p_e(x_i|\pi_i) = \begin{cases} p(x_i|\pi_i), & \text{si } x_i \text{ y } \pi_i \text{ son consistentes con } e, \\ 0, & \text{en otro caso,} \end{cases} \tag{9.14}$$

es decir, $p_e(x_i|\pi_i) = 0$ si X_i o alguno de sus padres son inconsistentes con la evidencia, en otro caso $p_e(x_i|\pi_i) = p(x_i|\pi_i)$. Por ello, la función de probabilidad condicionada $p(y|e)$ puede ser aproximada usando el Algoritmo 9.2 con la nueva función de probabilidad conjunta $p_e(x)$ sin normalizar.

De la discusión anterior se ve que todo método de simulación consta de tres componentes:

1. Una distribución para simular, $h(x)$, utilizada para generar la muestra.
2. Un método para obtener las realizaciones de $h(x)$.
3. Una fórmula para calcular los pesos.

Muchos de los métodos de simulación existentes son variantes del método anterior. Normalmente difieren entre sí en una o varias de las dos últimas componentes. En este capítulo se discuten los métodos siguientes:

- El método de aceptación-rechazo.
- El método del muestreo uniforme.
- El método de la función de verosimilitud pesante.
- EL método de muestreo hacia adelante y hacia atrás.
- El método del muestreo de Markov.
- El método del muestreo sistemático.
- El método de la búsqueda de la probabilidad máxima.

Dada una distribución de probabilidad $p(x)$, cada uno de estos métodos genera una muestra de tamaño N a partir de $p(x)$. Éstos difieren sólo en cómo se genera la muestra y en la elección de la función para cálculo de los pesos. De estos métodos, los cinco primeros pertenecen a la clase de los métodos de simulación estocástica, y los dos últimos son métodos de búsqueda determinista.

Por razones de simplicidad, se ilustran los diferentes algoritmos en el caso de variables aleatorias discretas, si bien, el mismo esquema es aplicable a variables continuas o al caso mixto de variables discretas y continuas (Castillo, Gutiérrez y Hadi (1995b)).

9.4 El Método de Aceptación-Rechazo

Henrion (1988) sugiere un método de simulación que genera las variables, una a una, con un muestreo hacia adelante, es decir, se muestrea una variable sólo cuando ya han sido muestreados todos sus padres. Según este método, se simulan todas las variables, incluyendo las evidenciales, en caso de que las haya. La distribución con la que se simula X_i es su función de probabilidad condicionada asociada (9.10), es decir

$$h(x_i|\pi_i) = p(x_i|\pi_i),\ i \in \{1, \ldots, n\}. \tag{9.15}$$

Por ello, las variables deben ordenarse de tal forma que los padres de una variable deben preceder a ésta en la simulación. Una ordenación de los

```
Iniciar
    Ordenar los nodos ancestralmente
Ciclo Principal
    para j ← 1 a N hacer
        para i ← 1 a n hacer
            x_i ← generar un valor a partir de p(x_i|π_i)
            si X_i ∈ E y x_i ≠ e_i entonces, repetir el ciclo i
```

FIGURA 9.6. Pseudocódigo para el método del muestreo lógico o de aceptación-rechazo.

nodos que satisface tal propiedad se llama una *ordenación ancestral*. Esta estrategia de simulación se llama muestreo *hacia adelante* porque va de padres a hijos. Una vez simulados los padres de X_i y asignados sus valores, se simula un valor de X_i usando la distribución de simulación $h(x_i|\pi_i)$, que en este caso es $p(x_i|\pi_i)$. Por tanto, los pesos se obtienen mediante

$$s(x) = \frac{p_e(x)}{h(x)} = \frac{\prod\limits_{X_i \notin E} p_e(x_i|\pi_i) \prod\limits_{X_i \in E} p_e(x_i|\pi_i)}{\prod\limits_{X_i \notin E} p(x_i|\pi_i) \prod\limits_{X_i \in E} p(x_i|\pi_i)}. \tag{9.16}$$

De (9.14) y (9.16), se deduce que

$$s(x) = \begin{cases} 1, & \textit{si } x_i = e_i, \textit{ para todo } X_i \in E, \\ 0, & \textit{en otro caso.} \end{cases} \tag{9.17}$$

Nótese que si $x_i \neq e_i$ para algún $X_i \in E$, entonces el peso es cero; por tanto, tan pronto como el valor simulado para los nodos evidenciales no coincida con el valor observado, se rechaza la muestra (peso cero). Por ello, este método es una variante del método de aceptación-rechazo (véase el Teorema 9.1). Sin embargo, Henrion (1988) denominó a este método *muestreo lógico*. El pseudocódigo para este algoritmo se da en la Figura 9.6.

El proceso de simulación procede secuencialmente, variable a variable. Consecuentemente, no es posible tener en cuenta la evidencia de que se dispone hasta que las variables correspondientes han sido muestreadas. Una vez obtenidos los valores simulados de las variables evidenciales, si dichos valores coinciden con la evidencia, la muestra se da por válida; en otro caso, se rechaza. Por tanto, las probabilidades condicionales se aproximan calculando el cociente entre los casos que están en concordancia con la evidencia y el número de casos totales.

En algunos casos, este método conduce a un porcentaje de rechazo muy alto y puede requerir un altísimo número de simulaciones, especialmente en los casos en los que las probabilidades de la evidencia son pequeñas, es

decir, en redes con probabilidades extremas. El ejemplo siguiente ilustra el método.

Ejemplo 9.4 Muestreo por el método de aceptación-rechazo. Considérese la red con seis nodos de la Figura 9.5 y las correspondientes funciones de probabilidad condicionada de la Tabla 9.1. La función de probabilidad conjunta de las seis variables puede factorizarse como en (9.9). Dada la evidencia $X_3 = 1, X_4 = 1$ (véase la Figura 9.7), se desea calcular la función de probabilidad "a posteriori" de cada una de las otras cuatro variables aplicando el método de aceptación-rechazo.

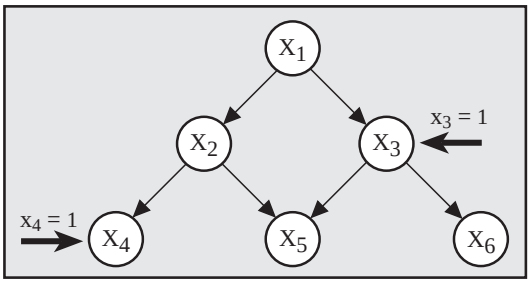


FIGURA 9.7. Añadiendo evidencia a la red de la Figura 9.5.

En primer lugar debe elegirse una numeración ancestral de los nodos. La Figura 9.8 muestra la estructura ancestral asociada a este ejemplo. En consecuencia, las variables tienen que ser numeradas de arriba a abajo, eligiendo los números de las variables en la misma fila arbitrariamente. Por ejemplo, se elige la ordenación $(X_1, X_2, X_3, X_4, X_5, X_6)$.

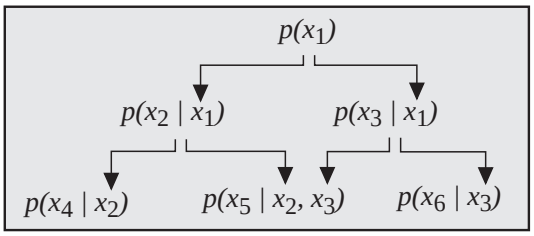


FIGURA 9.8. Estructura ancestral de la distribución de probabilidad conjunta asociada a la red de la Figura 9.7.

Entonces, las variables se muestrean a partir de las distribuciones de probabilidad condicionada de la Tabla 9.1 como sigue:

1. Muestreo de la variable X_1: Basándose en $p(x_1)$, se consulta un generador de números aleatorios que de un cero con probabilidad 0.3 y un uno con probabilidad 0.7. Con objeto de ilustrar el método, supóngase que el valor obtenido para X_1 es $x_1^1 = 1$. Seguidamente se

utiliza este valor x_1^1 para calcular las probabilidades de los restantes nodos de esta extracción.

2. Muestreo de la variable X_2: dada $X_1 = 1$, de la Tabla 9.1 las probabilidades de que X_2 tome los valores 0 y 1 son $p(X_2 = 0|x_1^1) = p(X_2 = 0|X_1 = 1) = 0.1$ y $p(X_2 = 1|X_1 = 1) = 0.9$, respectivamente. Por ello, se utiliza un generador de números aleatorios con estas características para obtener el valor simulado de X_2. Supóngase que dicho valor para X_2 resulta ser $x_2^1 = 0$.

3. Muestreo de la variable X_3: dado $X_1 = 1$, de la Tabla 9.1 X_3 toma los valores 0 y 1 con igual probabilidad. Se genera un valor de X_3 con esta distribución de probabilidad. Si el valor resultante es 0, entonces se rechaza esta extracción porque no coincide con la evidencia $X_3 = 1$ y se comienza con la Etapa 1 de nuevo. En otro caso, X_3 toma el valor 1 y la simulación continúa. Por tanto, se siguen simulando valores para X_1, X_2 y X_3 hasta que X_3 resulta $x_3^3 = 1$.

4. Muestreo de las variables X_4, X_5, X_6: la situación es similar a la de las etapas previas. Si el valor simulado del nodo X_4 no coincide con la evidencia $X_4 = 1$, se rechaza la muestra completa y se comienza de nuevo; en otro caso, se simulan los nodos X_5 y X_6. Supóngase que se obtiene $x_5^1 = 1$ y $x_6^1 = 0$.

La extracción concluye con la primera realización

$$x^1 = (x_1^1, x_2^1, x_3^1, x_4^1, x_5^1, x_6^1) = (1, 0, 1, 1, 1, 0).$$

Se repite el proceso hasta que se obtienen N realizaciones. Entonces, la distribución de probabilidad de cualquier variable se puede estimar por el porcentaje de muestras en las que ocurre el suceso de interés, como se muestra en el Ejemplo 9.3. ∎

El método de muestreo de aceptación-rechazo puede escribirse con la notación del Teorema 9.1 como

$$p_e(x) = c\, g(x)\, h(x) = c\, g(x) \prod_{i=1}^{n} p(x_i|\pi_i),$$

donde $p_e(x)$ está dado en (9.13), y

$$g(x_i) = \begin{cases} 1, & \text{si } x_i = e_i, \textit{ para todo } X_i \in E, \\ 0, & \text{en otro caso.} \end{cases}$$

Esto significa que la condición de aceptación dada por el Teorema 9.1 se cumplirá siempre si la realización es consistente con la evidencia y fallará, si no lo es.

Una desventaja clara del método del muestreo de aceptación-rechazo es que las evidencias que son conocidas no pueden ser tenidas en cuenta hasta que se conozcan los valores de las variables implicadas en ellas. Como respuesta a este problema, se han desarrollado otros métodos de simulación que no tienen este problema. Estos se presentan en las secciones siguientes.

9.5 Método del Muestreo Uniforme

La distribución de la población para la variable X_i está dada por (9.13), donde la función $p_e(x_i|\pi_i)$ es $p(x_i|\pi_i)$ con las variables evidenciales sustituídas por sus valores observados o cero si los valores x_i o π_i son inconsistentes con la evidencia. En este método la distribución con la que se simula la variable X_i, $h(x_i)$, es uniforme, es decir,

$$h(x_i) = \begin{cases} \dfrac{1}{card(X_i)}, & si\ X_i \notin E, \\ 1, & si\ X_i \in E\ y\ x_i = e_i, \\ 0, & en\ otro\ caso, \end{cases} \tag{9.18}$$

donde $card(X_i)$ denota la cardinalidad (número de posibles valores) de X_i. Esto garantiza que las variables evidenciales siempre tomarán sus valores observados y que todas las realizaciones serán compatibles con la evidencia (no se rechazan realizaciones). Por tanto, sólo es necesario simular las variables no evidenciales.

Tan pronto como se genera una realización $x = \{x_1, \ldots, x_n\}$, el peso correspondiente puede calcularse mediante

$$\begin{aligned} s(x) &= \frac{p_e(x)}{h(x)} \\ &= \frac{p_e(x)}{\prod_{X_i \notin E} \frac{1}{card(X_i)} \prod_{X_i \in E} 1} \propto p_e(x) = p(x), \end{aligned} \tag{9.19}$$

donde no es necesario considerar el factor $\prod_{X_i \notin E} card(X_i)$ puesto que es constante para todas las realizaciones. La última igualdad en (9.19) se cumple porque a las variables evidenciales se les asignan sus correspondientes valores en cada realización. La Figura 9.9 muestra el pseudocódigo para el método del muestreo uniforme.

Este método puede ser aplicado a un conjunto de nodos en cualquier orden, puesto que $h(x_i)$ no depende del valor de ninguna otra variable. Este método es simple, pero desgraciadamente, sus resultados son poco satisfactorios en casos en los que las realizaciones están lejos de estar distribuídas uniformemente. Si es éste el caso, se pueden generar muchas muestras poco representativas, lo que conduce a valores poco fiables.

Iniciar
    Para $X_i \in E$ **hacer**
        $x_i = e_i$
Ciclo Principal
    Para $j \leftarrow 1$ **a** N **hacer**
        Para $X_i \notin E$ **hacer**
            $x_i \leftarrow$ generar un valor con $h(x_i)$ en (9.18)
        $s_j = \prod_{i=1}^{n} p(x_i|\pi_i)$
Normalizar los pesos

FIGURA 9.9. Pseudocódigo para el método del muestreo uniforme.

Ejemplo 9.5 Método del muestreo uniforme. Considérese de nuevo la situación del Ejemplo 9.4, en el que se tenían seis variables, dos de las cuales eran evidenciales: $e = \{X_3 = 1, X_4 = 1\}$. Para obtener una realización $x = (x_1, \ldots, x_6)$ de X, en primer lugar se asignan a las variables evidenciales sus correspondientes valores observados $x_3^1 = 1$ y $x_4^1 = 1$. Seguidamente, se selecciona una ordenación arbitraria para los nodos no evidenciales, por ejemplo, (X_1, X_2, X_5, X_6) y se generan valores para cada una de estas variables al azar con idénticas probabilidades (0.5 y 0.5), puesto que en este caso la cardinalidad de los nodos es 2. Supóngase que en la primera extracción se obtienen los valores $x_1^1 = 0, x_2^1 = 1, x_5^1 = 0$ y $x_6^1 = 1$. Entonces, se calculan los pesos asociados a la realización

$$x^1 = (x_1^1, x_2^1, x_3^1, x_4^1, x_5^1, x_6^1) = (0, 1, 1, 1, 0, 1),$$

mediante (9.9), (9.19) y las distribuciones de probabilidad condicionadas de la Tabla 9.1:

$$\begin{aligned} s(x^1) &= p(x_1^1)p(x_2^1|x_1^1)p(x_3^1|x_1^1)p(x_4^1|x_2^1)p(x_5^1|x_2^1, x_3^1)p(x_6^1|x_3^1) \\ &= 0.3 \times 0.6 \times 0.8 \times 0.8 \times 0.2 \times 0.6 = 0.0138. \end{aligned}$$

El proceso se repite hasta que se obtiene el número deseado de realizaciones. ■

9.6 El Método de la Función de Verosimilitud Pesante

El método de la función de verosimilitud pesante fue desarrollado independientemente por Fung y Chang (1990) y Shachter y Peot (1990). Trata de resolver los problemas de alto rechazo (como en el método de aceptación-rechazo) y de los pesos descompensados de las realizaciones (como en el

```
Iniciar
    Ordenar los nodos ancestralmente
    para X_i ∈ E hacer
        x_i = e_i
Ciclo principal
    para j ← 1 a N hacer
        para X_i ∉ E hacer
            x_i ← generar un valor de h(x_i) en (9.20)
        s_j = ∏_{X_i∈E} p(e_i|π_i)
Normalizar los pesos
```

FIGURA 9.10. Pseudocódigo para el método de la función de verosimilitud pesante.

método del muestreo uniforme). La distribución de la población es (9.13) con (9.14) y la distribución que se utiliza en la simulación

$$h(x_i) = \begin{cases} p_e(x_i|\pi_i), & \text{si } X_i \notin E, \\ 1, & \text{si } X_i \in E \text{ y } x_i = e_i, \\ 0, & \text{en otro caso.} \end{cases} \tag{9.20}$$

Puesto que $h(x_i)$ depende de π_i, el orden de los nodos debe ser tal que los padres se muestreen antes que sus hijos (muestreo hacia adelante). Por ello, se ordenan los nodos según una *ordenación ancestral*.

El peso asociado a una realización $x = (x_1, \ldots, x_n)$ resulta

$$\begin{aligned} s(x) &= \frac{p_e(x)}{h(x)} \\ &= \prod_{X_i \notin E} \frac{p_e(x_i|\pi_i)}{p_e(x_i|\pi_i)} \prod_{X_i \in E} \frac{p_e(x_i|\pi_i)}{1} \\ &= \prod_{X_i \in E} p_e(x_i|\pi_i) = \prod_{X_i \in E} p(e_i|\pi_i). \end{aligned} \tag{9.21}$$

La última igualdad se verifica debido a que cuando $X_i \in E$, entonces $x_i = e_i$ (todas las realizaciones son consistentes con la evidencia). El pseudocódigo para el método de la función de verosimilitud pesante se muestra en la Figura 9.10. Este método da lugar a realizaciones representativas mientras que la probabilidad de la evidencia observada no esté muy próxima a cero.

Ejemplo 9.6 Método de la función de verosimilitud pesante. Considérese de nuevo la red Bayesiana del Ejemplo 9.5 con dos nodos evidenciales: $e = \{X_3 = 1, X_4 = 1\}$. Se ordenan los nodos no evidenciales con un criterio ancestral, por ejemplo, $X = (X_1, X_2, X_5, X_6)$. Entonces,

se comienza por asignar a los nodos evidenciales sus correspondientes valores observados $x_3^1 = 1$ y $x_4^1 = 1$. Siguiendo la ordenación de los nodos no evidenciales, se simula X_1 usando la función $p(x_1)$ dada en la Tabla 9.1, es decir, se utiliza un generador de números aleatorios que devuelva 0 con probabilidad 0.3 y 1 con probabilidad 0.7. Supóngase que el valor simulado para X_1 es $x_1^1 = 0$. Seguidamente se simula X_2 usando $p(x_2|X_1 = x_1^1) = p(x_2|X_1 = 0)$, es decir, se selecciona un cero con probabilidad 0.4 y un uno con probabilidad 0.6. Supóngase que se obtiene $x_2^1 = 1$. Seguidamente se simula X_5 usando $p(x_5|x_2^1, x_3^1) = p(x_5|X_2 = 1, X_3 = 1)$, que asigna una probabilidad 0.2 a cero y 0.8 a uno. Entonces se simula X_6 usando $p(x_6|x_3^1) = p(x_6|x_3 = 1)$, que asigna 0.4 a cero y 0.6 a uno. Supóngase que se obtiene $x_5^1 = 0$ y $x_6^1 = 1$. Entonces, aplicando (9.21), el peso de la muestra obtenida $x^1 = (x_1^1, x_2^1, x_3^1, x_4^1, x_5^1, x_6^1) = (0, 1, 1, 1, 0, 1)$ resulta

$$s(x^1) = p(x_3^1|x_1^1)p(x_4^1|x_2^1) = 0.8 \times 0.8 = 0.64.$$

Por ello, se obtiene la primera realización, $x^1 = (0, 1, 1, 1, 0, 1)$, con el peso asociado $s(x^1) = 0.64$. Se repite el proceso hasta que se obtiene el número deseado de realizaciones. ■

Nótese que cuando las probabilidades de las realizaciones no son uniformes, el método de la función de verosimilitud pesante exhibe un mejor comportamiento en la simulación que los métodos de aceptación-rechazo y que el muestreo uniforme.

El método de la función de verosimilitud pesante es simple pero eficiente y potente para propagar incertidumbre, especialmente en los casos en los que no existen métodos de propagación aproximada, por ejemplo, redes con variables discretas y continuas (véase Castillo, Gutiérrez y Hadi (1995b)). Por estas razones, se utiliza este método en las aplicaciones de la vida real analizadas en el Capítulo 12.

9.7 El Muestreo Hacia Adelante y Hacia Atrás

Los métodos de aceptación-rechazo y de la verosimilitud pesante pueden ser considerados como métodos de muestreo hacia adelante, en el sentido de que las variables se muestrean sólo tras haber muestreado sus padres respectivos. De esta forma, se obtiene $p(x_i|\pi_i)$ de las tablas de distribuciones condionales (véase, por ejemplo, la Tabla 9.1). Un método que no corresponde a esta estructura ha sido presentado recientemente por Fung y Del Favero (1994). Este método no requiere muestrear los padres antes de los hijos. El método utiliza $p_e(x)$ en (9.13) con (9.14) como función de distribución simulada, pero considera un método diferente para obtener las realizaciones a partir de esta distribución. Fung y Del Favero (1994) se refieren a este método como *muestreo hacia atrás*, pero en realidad combina

los muestreos hacia adelante y hacia atrás. El método que se utiliza para muestrear cada nodo depende de la topología de la red. Para nodos sin evidencia procedente de sus descendientes, se utiliza el muestreo hacia adelante. Para nodos que tienen algún descendiente evidencial el muestreo se realiza, una vez que se ha asignado un valor al nodo, utilizando la función de probabilidad condicionada correspondiente para obtener un valor de los padres del nodo (muestreo hacia atrás).

La primera etapa en el método de simulación hacia atrás consiste en ordenar los nodos, para obtener lo que se denomina una *ordenación válida*. Para ser considerada válida, una ordenación debe satisfacer:

1. Un nodo debe ser simulado o ser un nodo evidencial antes de ser utilizado para muestrear sus padres hacia atrás,
2. Los predecesores de un nodo deben ser simulados antes de que dicho nodo sea utilizado para muestrear hacia adelante, y
3. Cada nodo de la red debe ser o un nodo perteneciente a la ordenación o un ascendiente directo de un nodo que se utilice para muestrear hacia atrás.

Las condiciones 1 y 2 se imponen para garantizar que son posibles los muestreos hacia adelante y hacia atrás. La condición 3 garantiza que a todos los nodos se les asigna un valor.

En el muestreo hacia atrás, se simulan los padres de un nodo X_i siempre que ese nodo X_i haya sido ya simulado. El nodo X_i puede haber tomado un valor, bien porque sea un nodo evidencial o porque haya sido muestreado con anterioridad. Los valores de π_i se generan según la función asociada al nodo X_i, es decir, con probabilidad

$$h(\pi_i^*) = \frac{p(x_i|\pi_i)}{\alpha_i}, \tag{9.22}$$

donde π_i^* es el conjunto de los padres de X_i, con valores desconocidos y

$$\alpha_i = \sum_{x_j \in \pi_i^*} p(x_j|\pi_j) \tag{9.23}$$

es una constante de normalización que garantiza que la suma de las probabilidades asociadas a todos los valores posibles sea la unidad. Puesto que no todos los nodos pueden muestrearse hacia atrás, por ejemplo, porque no hay evidencia todavía, el muestreo hacia atrás debe ser combinado con otro método de muestreo hacia adelante, tal como el de la función de verosimilitud pesante. Entonces, los restantes nodos se muestrean hacia atrás por este método. El peso resulta entonces

$$s(x) \quad = \quad \frac{p_e(x)}{h(x)}$$

Iniciar
  Ordenación válida de los nodos
  Crear la lista (B) de nodos muestreados hacia adelante y la lista (F) de nodos muestreados hacia atrás
  para $X_i \in E$ **hacer**
    $x_i = e_i$
Ciclo Principal
  para $j \leftarrow 1$ **a** N **hacer**
    para $i \leftarrow 1$ **a** n **hacer**
      si $X_i \in F$ **entonces**
        $x_i \leftarrow$ generar un valor de $p(x_i|\pi_i)$
      si $X_i \in B$ **entonces**
        $\pi_i^* \leftarrow$ generar un valor a partir de $p(x_i|\pi_i)/\alpha_i$
    $s_j = \prod_{X_i \notin B \cup F} p(x_i|\pi_i) \prod_{X_i \in B} \alpha_i$
Normalizar los pesos

FIGURA 9.11. Pseudocódigo para el método de muestreo hacia adelante y hacia atrás.

$$= \frac{p_e(x_i|\pi_i)}{\prod_{X_i \in B} \frac{p_e(x_i|\pi_i)}{\alpha_i} \prod_{X_i \in F} p_e(x_i|\pi_i)}$$

$$= \prod_{X_i \notin B \cup F} p(x_i|\pi_i) \prod_{X_i \in B} \alpha_i, \tag{9.24}$$

donde B y F son los nodos que se muestrean hacia adelante y hacia atrás, respectivamente.

El pseudocódigo para el método del muestreo hacia adelante y hacia atrás se muestra en la Figura 9.11. A continuación se da un ejemplo ilustrativo.

Ejemplo 9.7 Muestreo hacia adelante y hacia atrás. Considérese la red Bayesiana de la Figura 9.5, con las correspondientes funciones de probabilidad condicionada de la Tabla 9.1, y la evidencia $X_3 = 1, X_4 = 1$. En primer lugar se necesita elegir una ordenación válida de los nodos que satisfaga las tres condiciones anteriores. Hay varias posibilidades, por ejemplo, $\{X_4, X_2, X_5, X_6\}$, $\{X_4, X_3, X_5, X_6\}$ y $\{X_6, X_4, X_5, X_3\}$ son ordenaciones válidas. Supóngase que se elige la ordenación válida $\{X_4, X_2, X_5, X_6\}$, donde X_4 y X_2 se muestrean hacia atrás y X_5 y X_6 se muestrean hacia adelante. Se procede como sigue:

1. En primer lugar, se asigna a las variables evidenciales los correspondientes valores de la evidencia. Por ello, se tiene $x_3^1 = x_4^1 = 1$.

2. Se muestrea hacia atrás la variable X_2 usando el nodo X_4: Para ello se usa $p(x_4^1|x_2)$, es decir, $p(X_4 = x_4^1|X_2 = 0) = 0.7$ y $p(X_4 = x_4^1|X_2 = 1) = 0.8$, lo que conduce a $\alpha_4 = 1.5$. Entonces se utiliza un generador de números aleatorios con $h(X_2 = 0) = 0.7/1.5$ y $h(X_2 = 1) = 0.8/1.5$. Supóngase que se obtiene $x_2^1 = 1$.

3. Se muestrea hacia atrás la variable X_1 usando el nodo X_2: Se utiliza $p(x_2^1|x_1)$, es decir, $p(X_2 = x_2^1|X_1 = 0) = 0.6$ y $p(X_2 = x_2^1|X_1 = 1) = 0.9$, lo que conduce a $\alpha_1 = 0.6 + 0.9 = 1.5$. Entonces se utiliza un generador de números aleatorios con $h(X_1 = 0) = 0.6/1.5$ y $h(X_1 = 1) = 0.9/1.5$. Supóngase que se obtiene $x_1^1 = 0$.

4. Se muestrea hacia adelante la variable X_5: Se usa $p(x_5|x_2^1, x_3^1)$, es decir, $p(X_5 = 0|X_1 = x_2^1, X_3 = x_3^1) = 0.2$ y $p(X_5 = 1|X_1 = x_2^1, X_3 = x_3^1) = 0.8$. Entonces se utiliza un generador de números aleatorios con estas probabilidades y se obtiene, por ejemplo, $x_5^1 = 0$.

5. Se muestrea hacia adelante la variable X_6: Se usa $p(x_6|x_3^1)$, es decir, $p(X_6 = 0|X_3 = x_3^1) = 0.4$ y $p(X_6 = 1|X_3 = x_3^1) = 0.6$ y se obtiene, por ejemplo, $x_6^1 = 0$.

Por ello, se obtiene la realización $x^1 = (0, 1, 1, 1, 0, 0)$. Entonces se calcula el peso usando (9.24):

$$s(x^1) = p(x_1^1)p(x_3^1|x_1^1)\alpha_4\alpha_2 = 0.3 \times 0.8 \times 1.5 \times 1.5 = 0.54.$$

Se puede repetir el mismo proceso hasta que se alcance el número deseado de realizaciones. ∎

9.8 Método de Muestreo de Markov

El método de muestreo de Markov se debe a Pearl (1987b) y consiste en asignar a los nodos evidenciales su correspondiente evidencia y en simular estocásticamente en la red resultante. Inicialmente, se genera una realización aleatoriamente, bien eligiendo al azar entre todas las posibles realizaciones, o bien aplicando alguno de los métodos previos. Seguidamente, se simulan las variables no evidenciales, una a una, siguiendo un orden arbitrario de las variables. Se genera un valor aleatorio para la variable seleccionada utilizando su función de probabilidad condicionada a todas las demás (se entiende que éstas toman los valores últimamente simulados). Esta función de probabilidad puede obtenerse de la forma indicada en el teorema siguiente (Pearl (1987b)).

Teorema 9.2 Función de probabilidad de una variable condicionada a todas las demás *La función de probabilidad de una variable X_i,*

Iniciar
  para $X_i \in E$ **hacer**
    $x_i = e_i$
  para $X_i \notin E$ **hacer**
    $x_i \leftarrow$ valor generado con la ley uniforme
Ciclo Principal
  para $j \leftarrow 1$ **a** N **hacer**
    para $X_i \notin E$ **hacer**
      para cada valor x_i de X_i **hacer**
        $q(x_i) \leftarrow p(x_i|\pi_i) \prod_{X_j \in C_i} p(x_j|\pi_j)$
      Normalizar $q(x_i)$
      $x_i \leftarrow$ generar a partir de $q(x_i)$ normalizada

FIGURA 9.12. Pseudocódigo para el método del muestreo de Markov.

condicionada a todas las demás, es la dada por

$$h(x_i) = p(x_i|x \setminus x_i) \propto p(x_i|\pi_i) \prod_{X_j \in C_i} p(x_j|\pi_j), \tag{9.25}$$

donde C_i es el conjunto de hijos de X_i y $X \setminus X_i$ denota todas las variables de X que no están en X_i.

Una vez que todas las variables no evidenciales han sido muestreadas, se obtiene una realización. A continuación se utilizan los valores de las variables obtenidos en esta realización para generar la siguiente realización. El pseudocódigo para este método se muestra en la Figura 9.12. Nótese que los únicos nodos que intervienen en la función asociada al nodo X_i ,son el propio X_i y el conjunto formado por los padres de X_i, los hijos de X_i y los padres de los hijos de X_i, excepto el mismo X_i. Nótese también que para las variables a las que no se les ha asignado todavía un valor se utilizan los valores de las simulaciones previas. Por ello, este método no requiere una ordenación de los nodos.

Puesto que se simula con las probabilidades reales, el peso asociado a este método es constante e igual a 1.

Tras generar el número deseado de realizaciones, la función de probabilidad condicional de cualquier nodo dada la evidencia se estima como se muestra en la Sección 9.3:

- Por la proporción de las extracciones en las que ocurre un elemento dado, o
- Por la media de la probabilidad condicional del suceso en todas las extracciones.

Este método produce muestras representativas, siempre que la función de probabilidad conjunta no contenga valores extremos.

Ejemplo 9.8 Método de muestreo de Markov. Considérese de nuevo la red Bayesiana usada en los ejemplos anteriores. Como antes, la evidencia obervada es $X_3 = 1$ y $X_4 = 1$. Según (9.25), las distribuciones que se utilizan en la simulación para las variables no evidenciales son

$$\begin{array}{l} h(x_1) = p(x_1|x \setminus x_1) \propto p(x_1)p(x_2|x_1)p(x_3|x_1), \\ h(x_2) = p(x_2|x \setminus x_2) \propto p(x_2|x_1)p(x_4|x_2)p(x_5|x_2, x_3), \\ h(x_5) = p(x_5|x \setminus x_5) \propto p(x_5|x_2, x_3), \\ h(x_6) = p(x_6|x \setminus x_6) \propto p(x_6|x_3). \end{array} \tag{9.26}$$

El método del muestreo de Markov comienza asignando a las variables evidenciales sus correspondientes valores de la evidencia. Por ello, $x_3^j = 1$ y $x_4^j = 1$, para todo $j = 1, \ldots, N$. Seguidamente, el método continúa como sigue:

1. Se asigna a cada una de las variables no evidenciales un valor arbitrario inicial. Supóngase que se obtiene $X_1 = 0, X_2 = 1, X_5 = 0, X_6 = 1$. Por ello, la realización inicial resulta $x^0 = (0, 1, 1, 1, 0, 1)$.

2. Se elige una ordenación arbitraria para seleccionar los nodos no evidenciales, por ejemplo $\{X_1, X_2, X_5, X_6\}$. Para cada variable de esta lista, se genera un valor aleatorio a partir de la correspondiente distribución de probabilidad conjunta en (9.26) dados los valores que toman las restantes variables, como sigue:

   Variable X_1: Los valores que toman las restantes variables están dados por x^0. Por ello, usando la Tabla 9.1, se calcula

   $$\begin{array}{rcl} p(X_1 = 0|x \setminus x_1) & \propto & p(X_1 = 0)p(X_2 = 1|X_1 = 0)p(X_3 = 1|X_1 = 0) \\ & = & 0.3 \times 0.6 \times 0.8 = 0.144, \end{array}$$

   $$\begin{array}{rcl} p(X_1 = 1|x \setminus x_1) & \propto & p(X_1 = 1)p(X_2 = 1|X_1 = 1)p(X_3 = 1|X_1 = 1) \\ & = & 0.7 \times 0.9 \times 0.5 = 0.315. \end{array}$$

   Normalizando las probabilidades dividiendo por su suma, $0.144 + 0.315 = 0.459$, se obtiene $p(X_1 = 0|x \setminus x_1) = 0.144/0.459 = 0.314$ y $p(X_1 = 1|x \setminus x_1) = 0.315/0.459 = 0.686$. Por tanto, se genera un valor para X_1 usando un generador de números aleatorios que devuelve 0 con probabilidad 0.314 y 1 con probabilidad 0.686. Supóngase que el valor obtenido es 0.

   Variable X_2: Utilizando el estado actual de las variables, las expresiones (9.26) y la Tabla 9.1 se obtiene

   $$\begin{array}{rcl} p(X_2 = 0|x \setminus x_2) & \propto & 0.4 \times 0.3 \times 0.5 = 0.06, \\ p(X_2 = 1|x \setminus x_2) & \propto & 0.6 \times 0.2 \times 0.2 = 0.024. \end{array}$$

Normalizando las probabilidades anteriores, se obtiene $p(X_2 = 0|x \setminus x_2) = 0.06/0.084 = 0.714$ y $p(X_2 = 1|x \setminus x_2) = 0.024/0.084 = 0.286$. Por ello, se genera un valor para X_2 a partir de esta distribución. Supóngase que el valor obtenido es 1.

Las variables X_5 y X_6 se simulan de forma similar. Supóngase que se obtiene $X_5 = 0$ y $X_6 = 1$. Por ello, tras la primera extracción, se obtiene la primera realización de la muestra $x^1 = (0, 1, 1, 1, 0, 1)$.

3. Se repite la Etapa 2 para N extracciones.

Ahora, la función de probabilidad condicional de un suceso puede aproximarse por el porcentaje de muestras en las que el suceso es cierto, o mediante la media de su función de probabilidad condicionada en todas las extracciones. Por ejemplo, si se quiere conocer la probabilidad del suceso $X_2 = 1$, se puede elegir una de las dos alternativas siguientes:

1. Dividir el número de muestras en las que $X_2 = 1$ por el número total de extracciones.

2. Calcular la media de las $p(X_2 = 1|x \setminus x_2)$ usadas en todas las extracciones. ■

La convergencia del método de muestreo de Markov está garantizada, bajo ciertas condiciones, por un teorema debido a Feller (1968), referente a la existencia de una distribución límite para cadenas de Markov. En cada extracción, las realizaciones cambian del estado i al estado j, y el cambio está gobernado por la probabilidad de transición de la variable simulada. En esta situación, cuando las probabilidades de transición son estríctamente positivas, la probabilidad de que el sistema se encuentre en un estado dado se aproxima a un límite estacionario. El caso de que algunas de las probabilidades de transición sean nulas corresponde a una cadena de Markov reducible y limita la aplicabilidad de los esquemas de simulación estocástica, como éste.

El método de muestreo de Markov evita claramente el problema de rechazo, pero desgraciadamente, tiene sus propios problemas. El método tendrá problemas de convergencia cuando la red contenga variables con probabilidades extremas. Estos problemas son debidos al hecho de que los ciclos sucesivos en los esquemas de simulación de Markov no son independientes y la simulación puede bloquearse en ciertos estados o conjuntos de estados.

9.9 Método del Muestreo Sistemático

Recientemente, Bouckaert (1994) y Bouckaert, Castillo y Gutiérrez (1996) introdujeron un nuevo método para generar una muestra de realizaciones

de forma sistemática. Al contrario que los algoritmos introducidos en la sección anterior, que son de naturaleza estocástica, este método procede de forma determinista.

9.9.1 La Idea Básica

Para ilustrar la idea básica de este método, se necesita la definición siguiente.

Definición 9.1 Ordenación de realizaciones. *Sea* $X = \{X_1, \ldots, X_n\}$ *un conjunto de variables discretas y sean* $\{0, \ldots, r_i\}$ *los valores posibles de* X_i*. Sea* $x^j = (x_1^j, \ldots, x_n^j)$*,* $j = 1, \ldots, m$*, el conjunto de todas las realizaciones posibles de* X*. Supóngase que las realizaciones se dan en un orden que satisface*

$$x^j < x^{j+1} \Leftrightarrow \exists\, k \text{ tal que } \forall\, i < k,\ x_i^j = x_i^{j+1} \text{ y } x_k^j < x_k^{j+1}. \tag{9.27}$$

Entonces, se dice que $(x_1^j, \ldots, x_n^j)$ precede *a* $(x_1^{j+1}, \ldots, x_n^{j+1})$*. A toda ordenación que satisface (9.27) se la conoce como ordenación de realizaciones.*

Ejemplo 9.9 Ordenación de realizaciones. Considérese la red Bayesiana de la Figura 9.13. Supóngase que X_1 y X_2 son binarias y X_3 es ternaria. La primera columna de la Tabla 9.4 contiene todas las realizaciones posibles en el orden que satisface (9.27) de arriba a abajo. Por ejemplo, para $j = 1$, se tiene $x^1 < x^2$ puesto que $x_1^1 = x_1^2$, $x_2^1 = x_2^2$ y $x_3^1 < x_3^2$. El lector puede verificar que $x^j < x^{j+1}$, para $j = 2, \ldots, 12$, en el sentido de la ordenación de realizaciones en (9.27). ∎

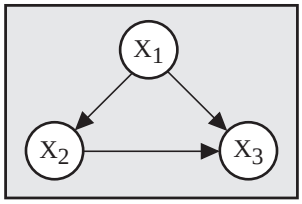


FIGURA 9.13. Una red Bayesiana con tres variables.

Utilizando la ordenación de las realizaciones, el intervalo unidad $[0, 1]$ puede dividirse en subintervalos asociados a las diferentes realizaciones. Por ello, se puede asociar cada realización, x^j, a un intervalo $I_j = [l_j, u_j) \subset [0, 1)$, cuya cota inferior la dan las probabilidades acumuladas de todas las realizaciones previas:

$$l_j = \sum_{x^i < x^j} p_e(x^i) \geq 0,$$

Realización (x_1, x_2, x_3)	Probabilidad $p(x_1, x_2, x_3)$	Probabilidad Acumulada	Intervalo $[l_i, u_i)$
(0,0,0)	$0.4 \times 0.6 \times 0.3 = 0.072$	0.072	[0.000, 0.072)
(0,0,1)	$0.4 \times 0.6 \times 0.3 = 0.072$	0.144	[0.072, 0.144)
(0,0,2)	$0.4 \times 0.6 \times 0.4 = 0.096$	0.240	[0.144, 0.240)
(0,1,0)	$0.4 \times 0.4 \times 0.3 = 0.048$	0.288	[0.240, 0.288)
(0,1,1)	$0.4 \times 0.4 \times 0.3 = 0.048$	0.336	[0.288, 0.336)
(0,1,2)	$0.4 \times 0.4 \times 0.4 = 0.064$	0.400	[0.336, 0.400)
(1,0,0)	$0.6 \times 0.4 \times 0.3 = 0.072$	0.472	[0.400, 0.472)
(1,0,1)	$0.6 \times 0.4 \times 0.4 = 0.096$	0.568	[0.472, 0.568)
(1,0,2)	$0.6 \times 0.4 \times 0.3 = 0.072$	0.640	[0.568, 0.640)
(1,1,0)	$0.6 \times 0.6 \times 0.3 = 0.108$	0.748	[0.640, 0.748)
(1,1,1)	$0.6 \times 0.6 \times 0.4 = 0.144$	0.892	[0.748, 0.892)
(1,1,2)	$0.6 \times 0.6 \times 0.3 = 0.108$	1.000	[0.892, 1.000)

TABLA 9.4. Realizaciones ordenadas y probabilidades absolutas y acumuladas con sus cotas inferior y superior.

y cuya cota superior es

$$u_j = l_j + p_e(x^j) \leq 1,$$

donde

$$p_e(x^j) = \begin{cases} p(x^j), & \text{si } x^j \text{ es consistente con } e, \\ 0, & \text{en otro caso.} \end{cases}$$

Nótese que en el caso de no evidencia, la función $p_e(x)$ es simplemente $p(x)$.

El método de muestreo sistemático genera las realizaciones sistemáticamente, eligiendo una sucesión de valores igualmente espaciados en el intervalo unidad $(0, 1)$ y buscando las realizaciones que les corresponden, es decir, las realizaciones cuyos intervalos asociados contienen los valores dados. Para generar una muestra de tamaño N se toman los valores

$$f_j = (j - 0.5)/N, \; j = 1, 2, \ldots, N, \tag{9.28}$$

de donde resulta que $0 < f_j < 1$, $j = 1, \ldots, N$. La realización cuyo intervalo contiene el valor f_j es elegida como la j-ésima realización de la muestra. Nótese que debido al carácter determinista del procedimiento, no se utilizan números aleatorios.

Ejemplo 9.10 Intervalos asociados a las realizaciones. Considérese de nuevo la red Bayesiana de la Figura 9.13. La función de probabilidad conjunta de las tres variables puede factorizarse como (véase la Sección 6.4.4)

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2). \tag{9.29}$$

x_1	$p(x_1)$
0	0.4
1	0.6

x_1	x_2	$p(x_2\|x_1)$
0	0	0.6
0	1	0.4
1	0	0.4
1	1	0.6

x_1	x_2	x_3	$p(x_3\|x_1, x_2)$
0	0	0	0.3
0	0	1	0.3
0	0	2	0.4
0	1	0	0.3
0	1	1	0.3
0	1	2	0.4
1	0	0	0.3
1	0	1	0.4
1	0	2	0.3
1	1	0	0.3
1	1	1	0.4
1	1	2	0.3

TABLA 9.5. Funciones de probabilidad condicional requeridas para especificar la función de probabilidad conjunta de los nodos de la red Bayesiana de la Figura 9.13.

Las funciones de probabilidad condicional necesarias para especificar la función de probabilidad conjunta en (9.29) se dan en la Tabla 9.5. La segunda columna de la Tabla 9.4 muestra la probabilidad de cada realización, la tercera columna da las probabilidades acumuladas, y la última columna muestra los intervalos correspondientes. La Figura 9.14 también ilustra las realizaciones, las probabilidades acumuladas, y sus intervalos asociados.

Veamos cómo se genera una muestra de tamaño $N = 4$ usando el método de muestreo sistemático. La sucesión $f_j = (j - 0.5)/4$ es (0.125, 0.375, 0.625 y 0.875). De la Tabla 9.4 ó la Figura 9.14, se ve que la muestra generada consta de las realizaciones siguientes: $\{(001), (012), (102), (111)\}$. La primera realización (001) es generada porque $f_1 = 0.125$ está contenido en el intervalo correspondiente a esta realización (el segundo intervalo de la Tabla 9.4), la segunda realización (012) se genera porque el intervalo correspondiente a esta realización (el sexto intervalo de la Tabla 9.4) contiene al valor $f_2 = 0.375$, y así sucesivamente. ■

Nótese que para generar una muestra usando el método del muestreo sistemático, se enumeran todas las posibles realizaciones de las variables como en la Figura 9.14. Esto se hace sólo con el propósito de ilustrar el método puesto que éste no requiere la generación de todas ellas (véase la Sección 9.9.2). Esto es afortunado, puesto que el número de realizaciones diferentes crece exponencialmente con el número de variables de tal forma que incluso con sólo diez variables binarias hay 1024 realizaciones diferentes. Por ello, varios de los valores f_j pueden caer en el intervalo correspondiente a la realización dada. Si la probabilidad de una realización x es $p_e(x)$, el número de valores diferentes de f_j asociados a esta realización en una

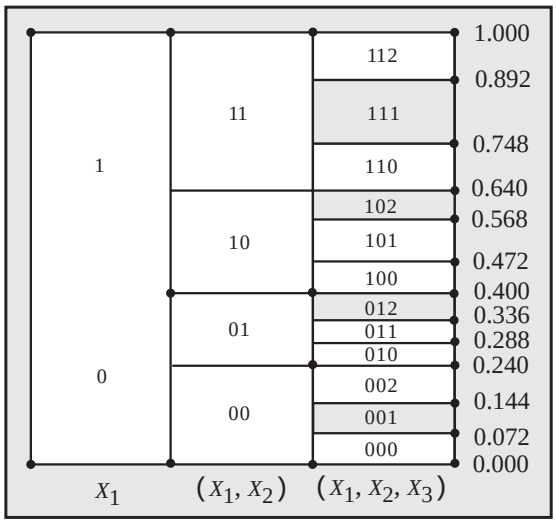


FIGURA 9.14. Realizaciones ordenadas y probabilidades acumuladas asociadas.

muestra de tamaño N es igual a $Np_e(x)$ si $p_e(x) \geq 1/N$. Esto significa que la realización correspondiente aparecerá $Np_e(x)$ veces en la muestra. Por ello, se tiene $h(x) = p_e(x)$ cuando $N \to \infty$. En otro caso, es decir, si $p_e(x) < 1/N$, se tendrá en la muestra o una realización, con probabilidad $Np_e(x)$, o ninguna realización, con probabilidad $1 - N\,p_e(x)$, que de nuevo implica que $h(x) = p_e(x)$, ya que

$$h(x) = \frac{1 \times Np_e(x) + 0 \times (1 - Np_e(x))}{N} = p_e(x).$$

Es claro que cuando N crece, la frecuencia de una realización dada converge a la frecuencia exacta, y, por tanto, los valores aproximados obtenidos con este método convergen a los valores exactos.

Puesto que el método genera las realizaciones sistemáticamente (de forma similar al muestreo aleatorio sistemático), este método se denomina *muestreo sistemático*. Bouckaert (1994) se refiere a este método como *muestreo estratificado*, puesto que todos los posibles conjuntos de realizaciones están divididos en subconjuntos mútuamente exhaustivos y exclusivos, o estratos,[2] y se considera un muestreo aleatorio en cada estrato. El muestreo estratificado es una técnica estadística bien conocida que conduce a muestras más representativas que las obtenidas mediante el muestreo aleatorio simple.

[2]Estos estratos se obtienen ordenando las realizaciones y usando los valores f_j que dividen el intervalo $(0, 1)$ en subintervalos iguales y disjuntos, cada uno de los cuales contiene un cierto número de realizaciones.

9.9.2 Algunos Aspectos de la Implementación

El método del muestreo sistemático es conceptuelmente muy simple, pero su implementacioón es complicada. En general, no se pueden calcular las probabilidades de todas las realizaciones debido al alto esfuerzo computacional requerido (para n variables binarias, hay 2^n realizaciones posibles). Debido al hecho de que este método se suele utilizar cuando los métodos exactos no pueden ser utilizados, se supone que el número de realizaciones es mucho mayor que el tamaño de la muestra N. Esto implica que muchas de las realizaciones (la mayoría de ellas con pequeñas probabilidades) no aparecerán en la muestra simulada. Un método que sea eficiente computacionalmente debe ser capaz de saltarse estas realizaciones, evitando cálculos innecesarios. El método procede de una forma sistemática, yendo de la primera realización $(0, \ldots, 0)$ a la última realización $(r_1, \ldots, r_n)$ y teniendo en cuenta la ordenación de éstas y el carácter determinista de la sucesión seleccionada para determinar la sucesión de valores f_j. La principal ventaja de este procedimiento es que para obtener una nueva realización sólo se necesita actualizar los valores de las últimas k variables. Esto permite un procedimiento rápido que salta muchas realizaciones de golpe. Sin embargo, es necesario determinar qué variables deben ser actualizadas.

Con este objetivo, supóngase que las variables $X_1, \ldots, X_n$ se dan en un orden ancestral y que se tiene la realización, $(x_1^{j-1}, \ldots, x_n^{j-1})$, correspondiente al valor f_{j-1} de la sucesión. Entonces, se define una cota inferior $l(i)$ y una cota superior $u(i) \geq l(i)$ para cada variable X_i, que indica los valores de la probabilidad para los que cada variable dará lugar a los dos próximos cambios de valor. Por ejemplo, obsérvese en la Figura 9.15 que en la Etapa 4, que corresponde a un valor de f_j en el intervalo $(0.240, 0.288)$, la realización asociada es $(0, 1, 0)$. El cambio siguiente de la variable X_3 tiene lugar en el valor 0.288 (realización $(0, 1, 1)$) y el siguiente ocurrirá en el valor 0.336 (realización $(0, 1, 2)$). Por ello, para pasar a la realización siguiente, se cambia $l(3)$ de 0.240 a 0.288 y $u(3)$ de 0.288 a 0.336. Nótese que las funciones límites para las variables X_1 y X_2 no cambian. La Figura 9.15 muestra los valores de las funciones límites $l(i)$ y $u(i)$ para las tres variables cuando el valor f_j está en cada uno de los intervalos sombreados.

Una vez que se ha calculado la realización asociada a f_{j-1} y las funciones límites de los nodos han sido actualizadas, el problema clave para obtener la realización asociada al próximo valor, f_j, es la determinación del número k de la variable a partir de la cual no es necesario proceder a su actualización. Los límites inferior y superior definidos arriba se usan para identificar esta variable de forma eficiente.

De la discusión anterior se observa que el algoritmo del muestreo sistemático puede resumirse en las etapas siguientes:

1. Iniciar las funciones límites.

2. Generar la sucesión de valores f_j.

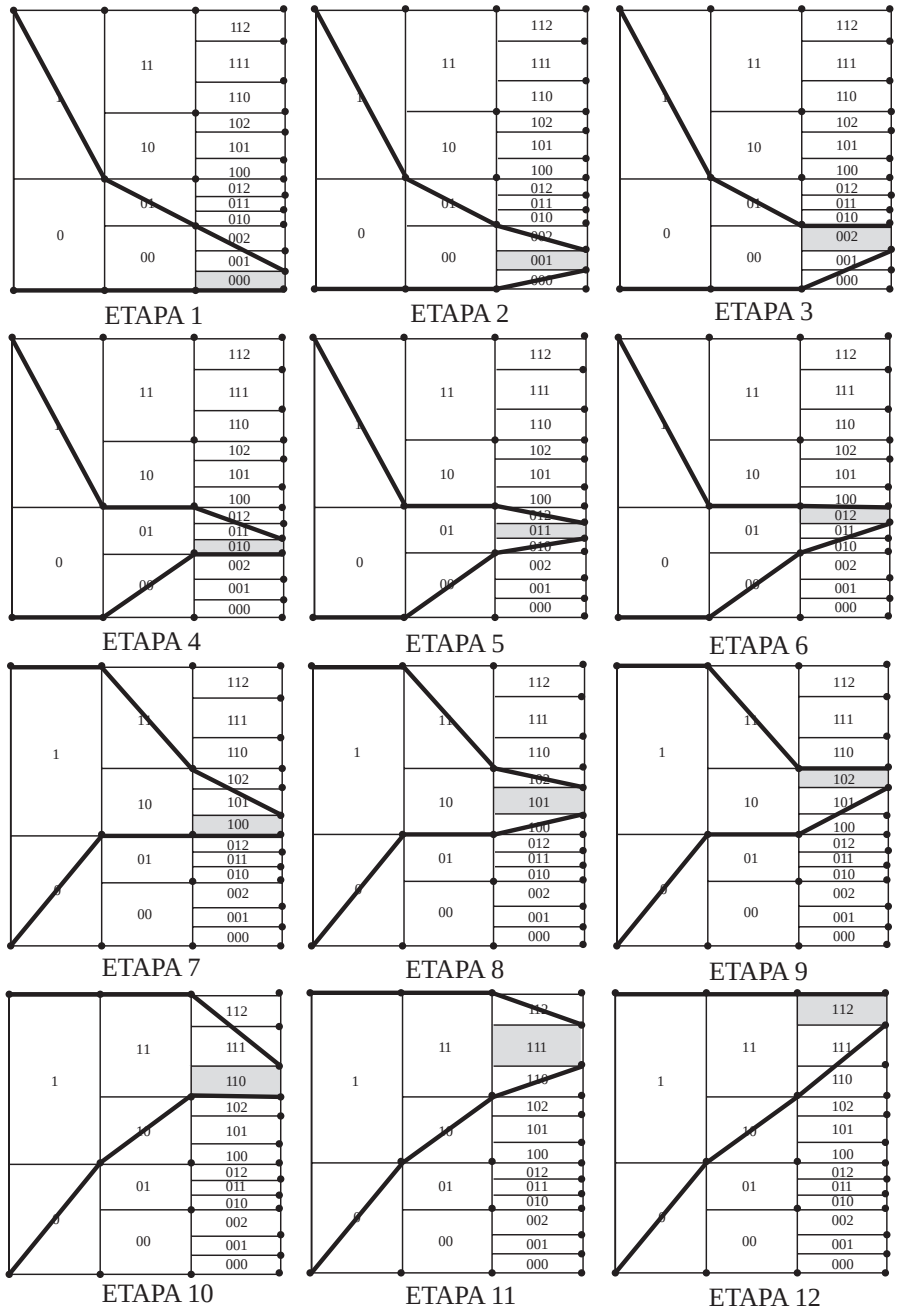


FIGURA 9.15. Una ilustración de los límites $l(j)$ y $u(j)$.

```
Iniciar
    Ordenar los nodos ancestralmente
    Iniciar las funciones límites (usar el código de la Figura 9.17)
Ciclo Principal
    Generar la sucesión de valores f_j usando (9.28)
    para j ← 1 a N hacer
        x^j ← determinar la realización asociada a f_j
              (usar el código de la Figura 9.18)
        s(x^j) ← ∏_{X_i ∈ E} p(e_i|π_i)
Normalizar los pesos
```

FIGURA 9.16. Pseudocódigo para la obtención de una muestra de tamaño N, mediante el algoritmo de muestreo sistemático.

3. Determinar qué realizaciones corresponden a estos valores.

4. Calcular y normalizar los pesos.

La Figura 9.16 da el pseudocódigo correspondiente. Nótese que en el caso de que no haya evidencia disponible no hacen falta pesos. Los códigos correspondientes a las dos etapas principales, iniciación de las funciones límites y determinación de las realizaciones asociadas a los valores f_j, se dan en las Figuras 9.17 y 9.18. En estas Figuras $p(X_i = k|\pi_i^j)$ es la probabilidad de que el nodo X_i tome el valor k dado que sus padres toman valores en x^j.

En el proceso de iniciación, se elige la primera realización, $(0, \ldots, 0)$, para iniciar los valores de las funciones límites $l()$ y $u()$. La Etapa 1 de la Figura 9.15 muestra los valores límites correspondientes a esta realización. Estos límites se calculan como sigue:

- Para el nodo X_1 se tiene $x_1^0 = 0$, por lo que las cotas inferior y superior para la primera variable son $l(1) = 0.0$ y $u(1) = p(X_1 = 0) = 0.4$.
- Para X_2, se tiene $x_2^0 = 0$, y por tanto $l(2) = 0.0$ y $u(2) = u(1) \times p(X_2 = 0|x_1^0) = 0.4 \times 0.6 = 0.24$.
- Finalmente, se tiene $l(3) = 0.0$ y $u(3) = u(2) \times p(X_3 = 0|x_1^0, x_2^0) = 0.24 \times 0.3 = 0.072$.

Una vez que se han iniciado las funciones límite, se usan los valores obtenidos para calcular la primera realización, asociada al valor f_1. Para ello se utiliza una función para la búsqueda binaria, $Binsearch(f_1, l, u)$, que localiza en $\log_2 n$ operaciones el índice mayor i tal que $l(i) \leq f_1 \leq u(i)$. Por tanto, para obtener la primera realización los valores $(x_1^0, \ldots, x_i^0)$ son los mismos y sólo es necesario actualizar los valores para el resto de los nodos, $(x_{i+1}, \ldots, x_n)$. Entonces, los nuevos valores para las funciones límites se

```
l(0) ← 0;  u(0) ← 1
para i ← 1 a n hacer
    l(i) ← 0
    si X_i ∈ E entonces
        x_i^0 ← e_i
        u(i) ← u(i − 1)
    en otro caso
        x_i^0 ← 0
        u(i) ← u(i − 1) × p(X_i = 0|π_i^0)
```

FIGURA 9.17. Pseudocódigo para la etapa de iniciación de las funciones límites en el algoritmo del muestreo sistemático.

```
i ← Binsearch (f_j, l, u)
mientras i <= n hacer
    si X_i ∈ E entonces
        l(i) ← l(i − 1)
        u(i) ← u(i − 1)
        x_i^j ← e_i
    en caso contrario
        k ← 0
        l(i) ← l(i − 1)
        u(i) ← l(i) + (u(i − 1) − l(i − 1)) × p(X_i = k|π_i^j)
        mientras f_j > u(i) hacer
            k ← k + 1
            l(i) ← u(i)
            u(i) ← l(i) + (u(i − 1) − l(i − 1)) × p(X_i = k|π_i^j)
        x_i^j ← k
    i ← i + 1
devolver (x_1^{j−1}, ... x_{i−1}^{j−1}, x_i^j, ... x_n^j)
```

FIGURA 9.18. Pseudocódigo para la determinación de las realizaciones asociadas a f_j en el algoritmo de muestreo sistemático.

utilizan para obtener la realización asociada a f_2, y así sucesivamente hasta obtener las N realizaciones que forman la muestra.

La Figura 9.18 muestra el pseudocódigo de este algoritmo optimizado para reducir el número de multiplicaciones, que son operaciones costosas. Esto es especialmente útil para el caso de variables con altas cardinalidades.

Recientemente se ha introducido una modificación a este algoritmo de muestreo sistemático por Bouckaert, Castillo y Gutiérrez (1996). Puesto que se trabaja con una sucesión determinista, una vez que se genera una realización asociada a un determinado valor f_j, se puede conocer cuántos valores de la sucesión dan lugar a la misma realización usando la fórmula (véase la Figura 9.19)

$$\delta = \left\lfloor \frac{u(n) - f_j}{N} \right\rfloor + 1, \tag{9.30}$$

donde $\lfloor . \rfloor$ es la parte entera y n es el número de variables. Entonces se incrementa δ unidades el contador j de la sucesión f_j, en vez de una unidad. De esta forma, se ahorra el trabajo de buscar la misma realización varias veces cuando los valores de f_j corresponden a la misma realización. Con esta técnica el tiempo de simulación se reduce notablemente.

La Figura 9.20 da el pseudocódigo para esta modificación del método del muestreo sistemático. Las únicas diferencias entre los códigos de las Figuras 9.16 y 9.20 consisten en que en el último código el contador j se incrementa en δ unidades en vez de en una sola unidad.

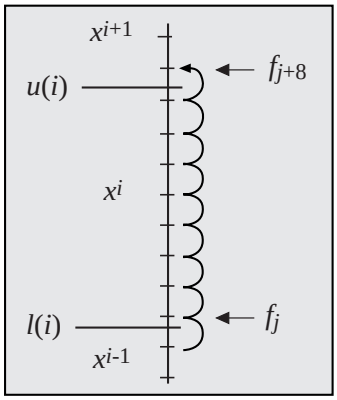


FIGURA 9.19. Una ilustración de cómo se saltan los mismos valores de la sucesión f_j cuando corresponden a la misma realización.

La eficiencia del método modificado de muestreo sistemático aumenta en las situaciones en las que hay un gran número de valores f_j que caen en el intervalo asociado a una realización. En redes probabilísticas con probabilidades extremas, la mayor parte de las realizaciones tienen probabilidades muy bajas, y sólo unas pocas realizaciones tienen probabilidades altas. En esta situación, la modificación anterior proporciona una mejora sustancial de la eficiencia del método. Por ejemplo, considérese la red Bayesiana de la Figura 9.13 y supóngase que se toman los nuevos valores de las probabilidades condicionales dados en la Tabla 9.6. Los valores de esta tabla son más extremos que los de la Tabla 9.5.

Iniciar
    Ordenar los nodos en orden ancestral
    Iniciar las funciones límites
Ciclo principal
    Generar la sucesión de valores f_j usando (9.28)
    para $j \leftarrow 1$ **a** N **hacer**
        $x^j \leftarrow$ generar la realización asociada a f_j
        $\delta \leftarrow \lfloor \frac{u(n) - f_j}{N} \rfloor$
        $s(x^j) \leftarrow \delta$
        $i \leftarrow i + \delta$
Normalizar los pesos

FIGURA 9.20. Marco general para el algoritmo modificado de muestreo sistemático.

x_1	$p(x_1)$
0	0.9
1	0.1

x_1	x_2	$p(x_2\|x_1)$
0	0	0.2
0	1	0.8
1	0	0.1
1	1	0.9

X_1	X_2	X_3	$p(X_3\|X_1, X_2)$
0	0	0	0.2
0	0	1	0.2
0	0	2	0.6
0	1	0	0.8
0	1	1	0.1
0	1	2	0.1
1	0	0	0.3
1	0	1	0.4
1	0	2	0.3
1	1	0	0.7
1	1	1	0.1
1	1	2	0.2

TABLA 9.6. Nuevos valores de las funciones de probabilidad condicionales requeridos para especificar la función de probabilidad conjunta de los nodos de la red Bayesiana de la Figura 9.13.

La nueva estructura de las realizaciones se muestra en la Figura 9.21. En esta Figura, se puede ver que la probabilidad de la realización $(0, 1, 0)$ es mayor que la mitad de la probabilidad total. Por ello, el algoritmo modificado ahorrará más de $N/2$ iteraciones para obtener una muestra de tamaño N.

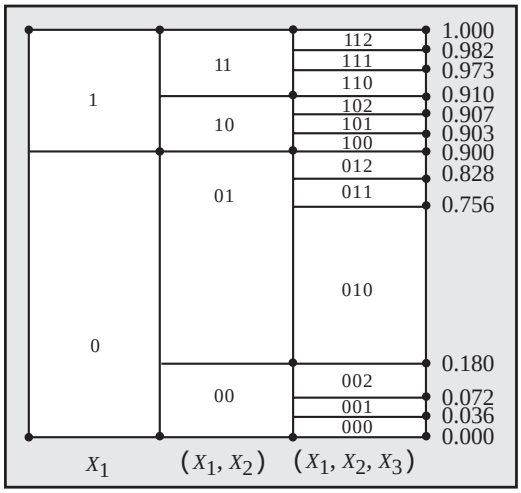


FIGURA 9.21. Realizaciones y probabilidades acumuladas.

9.9.3 Comparación de los Algoritmos Estocásticos

Bouckaert, Castillo y Gutiérrez (1996) comparan los cuatro métodos de propagación aproximada presentados en las secciones anteriores: función de verosimilitud pesante, muestreo de Markov, muestreo sistemático, y muestreo sistemático modificado. Ellos prueban los métodos en diez redes Bayesianas diferentes y usan dos medidas de rendimiento: el tiempo medio de ejecución del algoritmo y el error medio de la aproximación, concluyendo lo siguiente:

1. En general, el muestreo sistemático modificado es el mejor de los cuatro métodos.

2. En términos de error, los dos métodos de muestreo sistemático tienen igual calidad. Esto se debe a que la muestra producida por ambos algoritmos es la misma. En términos de tiempo de computación, para pequeños valores de N, los métodos de muestreo sistemático y el modificado dan rendimientos similares cuando las probabilidades de la red Bayesiana se eligen en el intervalo unidad. Cuando las probabilidades son extremas (elegidas, por ejemplo, en el intervalo compuesto $[0, 0.1] \cup [0.9, 1]$), el método de muestreo sistemático modificado se comporta mejor que el del método original. La razón de este comportamiento es que las redes con probabilidades extremas dan lugar a estratos grandes, en los que el salto conduce a notables ahorros. Estos estratos grandes son menos frecuentes en redes con probabilidades no extremas, para las que saltar no ayuda mucho.

3. El método de la función de verosimilitud pesante resulta crecientemente ineficiente con redes que tienen probabilidades extremas (cer-

canas a cero). Sin embargo, la eficiencia del método de muestreo sistemático crece en este tipo de redes. Por ello, este algoritmo elimina el problema de las probabilidades extremas asociadas a los métodos estocásticos.

9.10 Método de Búsqueda de la Máxima Probabilidad

En la Sección 9.9 se ha discutido un método determinista que genera las realizaciones de forma sistemática. En esta sección se describe un *algoritmo determinista de búsqueda*, que llamaremos *método de búsqueda de la máxima probabilidad*. Los métodos de búsqueda de la máxima probabilidad calculan las N realizaciones de la muestra mediante la creación de un árbol cuyas ramas se asocian a realizaciones parciales de las variables. En cada etapa, el proceso elige una de las ramas del árbol asociada a una realización parcial $(x_1^i, \ldots, x_m^i)$. Si la correspondiente realización es completa, es decir, si $m = n$, se corta la rama del árbol y se incluye dicha realización en la muestra. En otro caso, se aumenta el árbol con tantas ramas como valores posibles pueda tomar la siguiente variable, x_{m+1}. Por ello, la rama original, $(x_1^i, \ldots, x_m^i)$, se reemplaza por las ramas $(x_1^i, \ldots, x_m^i, x_{m+1})$ para todos los valores posibles de X_m.

Se han propuesto varios métodos en la literatura (véase, por ejemplo, Pearl (1987a), Henrion (1991), Poole (1993a, 1993b) y Srinivas y Nayak (1996)). La principal diferencia entre ellos es el criterio de elección de las ramas en cada etapa. Por ejemplo, el algoritmo de búsqueda de la máxima probabilidad (Poole (1993a)) usa el criterio consistente en maximizar la probabilidad para elegir las ramas en cada etapa. El algoritmo procede como sigue.

Dada una ordenación ancestral de las variables $X = \{X_1, \ldots, X_n\}$, se comienza con un árbol con tantas ramas como número de valores posibles tome la variable X_1. Se calcula la probabilidad de cada una de esas ramas, y se elige la rama con mayor probabilidad. Seguidamente, se aumenta esta rama con un subárbol que contiene tantas ramas como posibles valores tome X_2. Seguidamente, se calculan las probabilidades de cada una de estas ramas multiplicando las probabilidades (marginales y/o condicionadas) de sus nodos. Se elige la rama (del árbol ampliado) con máxima probabilidad. El árbol se expande con las ramas de X_3, y así sucesivamente hasta que se alcanza la última variable X_n. Nos referiremos a toda rama con n nodos como *rama completa*. La primera realización está formada por los valores de los nodos asociados a la rama completa con máxima probabilidad. Para proseguir con la búsqueda, se corta esta rama del árbol.

La siguiente realización se obtiene buscando la rama con máxima probabilidad de entre las restantes ramas del árbol. Si la rama elegida es

completa, la segunda realización constará de los valores de las variables asociadas a la misma. Si, por el contrario, la rama elegida es incompleta, el árbol se aumenta con un subárbol correspondiente a la siguiente variable en la ordenación ancestral. Este proceso se continúa hasta que se generan las realizaciones suficientes para garantizar la calidad de la estimación. Sin embargo, este método se explica mucho mejor mediante un ejemplo.

Ejemplo 9.11 Método de búsqueda de la máxima probabilidad. Considérese la red Bayesiana de la Figura 9.13 y las probabilidades condicionales de la Tabla 9.6. Las variables $X = (X_1, X_2, X_3)$ tienen cardinalidades $(2, 2, 3)$, respectivamente. La función de probabilidad conjunta de las tres variables puede escribirse en la forma

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2).$$

Las variables ya están dadas en orden ancestral. Puesto que X_1 toma dos valores posibles, se comienza con un árbol que tiene dos ramas, correspondientes a $X_1 = 0$ y $X_1 = 1$. De la Tabla 9.6, las probabilidades de las dos ramas son $p(X_1 = 0) = 0.9$ y $p(X_1 = 1) = 0.1$. Esto se ilustra en la Figura 9.22(a). Por tanto, se tienen las ramas iniciales

$$\begin{aligned} R_1^1 &= (X_1 = 0), \quad p(R_1^1) = 0.9, \\ R_2^1 &= (X_1 = 1), \quad p(R_2^1) = 0.1. \end{aligned}$$

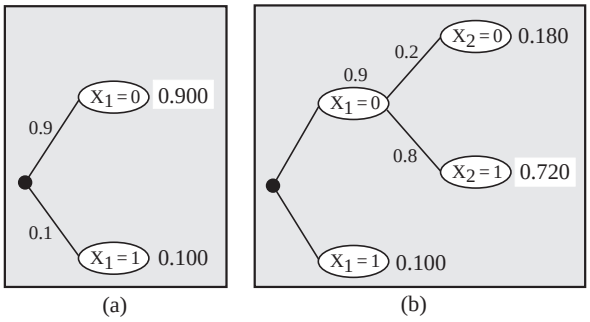


FIGURA 9.22. Dos primeras etapas del algoritmo de búsqueda de la máxima probabilidad.

A la primera rama, R_1^1 le corresponde la probabilidad máxima. Como la rama no es completa, se considera la variable siguiente, X_2, que toma dos posibles valores. En la Etapa 2, se aumenta el árbol mediante un subárbol con dos ramas, $X_2 = 0$ y $X_2 = 1$, como se muestra en la Figura 9.22(b). Las probabilidades correspondientes, obtenidas de la Tabla 9.6, son $p(X_2 = 0|X_1 = 0) = 0.2$ y $p(X_2 = 1|X_1 = 0) = 0.8$. Ahora el árbol contiene tres

ramas:

$$\begin{array}{ll} R_1^2 = (X_1 = 0, X_2 = 0), & p(R_1^2) = 0.9 \times 0.2 = 0.180, \\ R_2^2 = (X_1 = 0, X_2 = 1), & p(R_2^2) = 0.9 \times 0.8 = 0.720, \\ R_3^2 = (X_1 = 1), & p(R_3^2) = 0.100. \end{array}$$

En la Etapa 3, se selecciona la rama con probabilidad máxima. En este caso, la rama R_2^2. Se aumenta el árbol con las tres ramas asociadas a los tres valores posibles del nodo X_3, como se muestra en la Figura 9.23(a). Las correspondientes probabilidades, obtenidas de la Tabla 9.6, son $p(X_3 = 0|X_1 = 0, X_2 = 1) = 0.8$, $p(X_3 = 1|X_1 = 0, X_2 = 1) = 0.1$ y $p(X_3 = 2|X_1 = 0, X_2 = 1) = 0.1$. El nuevo árbol contiene tres ramas:

$$\begin{array}{ll} R_1^3 = (X_1 = 0, X_2 = 0), & p(R_1^3) = 0.180, \\ R_2^3 = (X_1 = 0, X_2 = 1, X_3 = 0), & p(R_2^3) = 0.720 \times 0.8 = 0.576, \\ R_3^3 = (X_1 = 0, X_2 = 1, X_3 = 1), & p(R_3^3) = 0.720 \times 0.1 = 0.072, \\ R_4^3 = (X_1 = 0, X_2 = 1, X_3 = 2), & p(R_4^3) = 0.720 \times 0.1 = 0.072, \\ R_5^3 = (X_1 = 1), & p(R_5^3) = 0.100. \end{array}$$

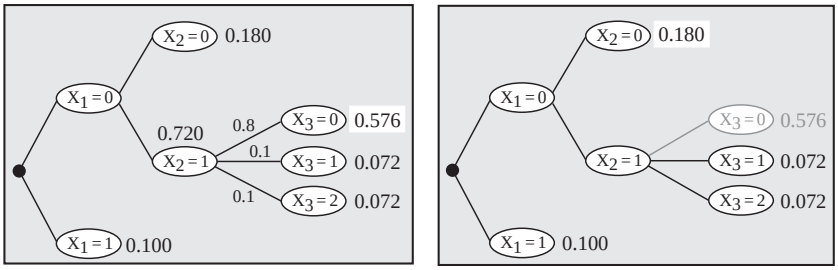


FIGURA 9.23. Método de búsqueda de la máxima probabilidad: Etapa 3 (a) y Etapa 4 (b).

La rama con máxima probabilidad es $R_2^3 = (X_1 = 0, X_2 = 1, X_3 = 0)$. Puesto que esta rama es completa, se ha conseguido la primera realización: $x^1 = (010)$, y se elimina dicha rama del árbol (véase la Figura 9.23(b)).

Ahora, de las restantes cuatro ramas, R_1^3 tiene la máxima probabilidad. Por ser incompleta, se aumenta el árbol en las tres nuevas ramas para X_3. Esto se muestra en la Figura 9.24(a). Las correspondientes probabilidades son $p(X_3 = 0|X_1 = 0, X_2 = 0) = 0.2$, $p(X_3 = 1|X_1 = 0, X_2 = 0) = 0.2$ y $p(X_3 = 3|X_1 = 0, X_2 = 0) = 0.6$. El árbol contiene ahora seis ramas:

$$\begin{array}{ll} R_1^4 = (X_1 = 0, X_2 = 0, X_3 = 0), & p(R_1^4) = 0.180 \times 0.2 = 0.036, \\ R_2^4 = (X_1 = 0, X_2 = 0, X_3 = 1), & p(R_2^4) = 0.180 \times 0.2 = 0.036, \\ R_3^4 = (X_1 = 0, X_2 = 0, X_3 = 2), & p(R_3^4) = 0.180 \times 0.6 = 0.108, \\ R_4^4 = (X_1 = 0, X_2 = 1, X_3 = 1), & p(R_4^4) = 0.072, \\ R_5^4 = (X_1 = 0, X_2 = 1, X_3 = 2), & p(R_5^4) = 0.072, \\ R_6^4 = (X_1 = 1), & p(R_6^4) = 0.100. \end{array}$$

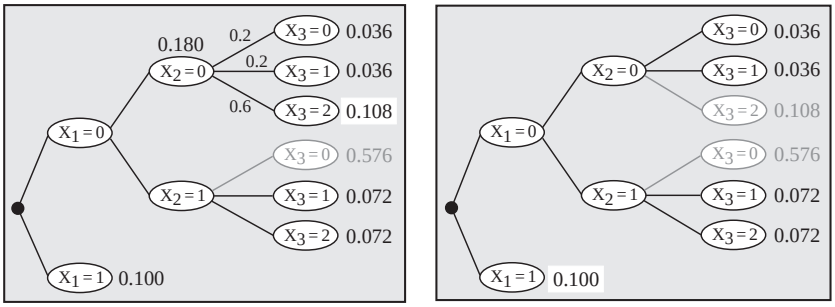


FIGURA 9.24. Método de búsqueda de la máxima probabilidad: Etapa 5 (a) y Etapa 6 (b).

La rama R_3^4 tiene la máxima probabilidad y es completa. Por tanto, $x^2 = (002)$. El proceso continúa hasta que se genera el número deseado de realizaciones o se acumula una probabilidad suficiente. Si se desea calcular la probabilidad exacta, en vez de la aproximada, se continúa el proceso hasta generar todas las realizaciones. En este caso, el árbol completo contiene $2 \times 2 \times 3 = 12$ posibles ramas (realizaciones). Este árbol se muestra en la Figura 9.25. Sin embargo, en la mayor parte, si no en todos, los casos prácticos, no se necesitan valores exactos, y se puede interrumpir el proceso una vez que se obtienen N realizaciones. Para ilustrar este hecho, la Tabla 9.7 muestra las probabilidades aproximadas de cada nodo basándose en las primeras j realizaciones, para $j = 1, \ldots, 12$. La última fila contiene las probabilidades exactas puesto que se basan en las 12 realizaciones posibles. Puede verse que la aproximación mejora notablemente a medida que el número de realizaciones aumenta.

El proceso puede ser visto como un proceso en el que la probabilidad total 1 se distribuye secuencialmente entre las realizaciones o grupos de realizaciones. Por ejemplo, en la Etapa 1 se asigna una probabilidad 0.9 al grupo de realizaciones de la forma $(0, \ldots)$, y una probabilidad 0.1 al grupo de realizaciones de la forma $(1, \ldots)$, y en la Etapa 2 se distribuye una probabilidad 0.9 entre los grupos de realizaciones de las formas $(0, 0, \ldots)$ y $(0, 1, \ldots)$, asignándoles 0.18 y 0.72, respectivamente. ■

Una ventaja importante de este método es que, en el caso de no evidencia, permite calcular las cotas inferior y superior de las probabilidades marginales de todas las variables, en cada una de las etapas. Si se tiene alguna evidencia, estas cotas pueden calcularse nada más incluir los nodos evidenciales en todas las ramas del árbol, es decir, una vez que el árbol ha sido completado hasta la última variable evidencial, puesto que sólo entonces puede conocerse la constante de normalización. Estas cotas pueden utilizarse para interrumpir el proceso de muestreo cuando las cotas están dentro de una cierta tolerancia. Esto da lugar a un gran ahorro de cálculos.

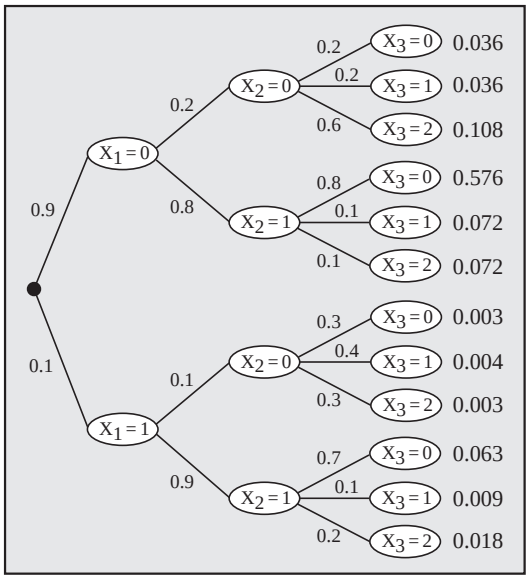


FIGURA 9.25. Árbol de búsqueda mostrando las probabilidades asociadas a todas las realizaciones.

De hecho, el método es óptimo, en el sentido de que no hay ordenación mejor para las realizaciones, es decir, las realizaciones con grandes probabilidades asociadas son generadas en primer lugar.

Durante el proceso de simulación, las cotas de las probabilidades se calculan como sigue. Se comienza con las cotas obvias $(0, 1)$ y en cada etapa se alteran sólo las cotas de las variables a las que se han añadido nuevas ramas. Se procede de la forma siguiente:

- **Cotas inferiores:** Se añaden a las cotas inferiores previas de la variable en curso todas las probabilidades asignadas a dicha variable en las nuevas ramas.

- **Cotas superiores:** Se calcula la anchura del intervalo sumando todas las probabilidades no asignadas, es decir, la suma de las probabilidades en los nodos terminales que no corresponden a la variable. Seguidamente, se añade esta anchura a la cota inferior de la variable para obtener la cota superior. Nótese que esta anchura es idéntica para todos los valores posibles de la variable, puesto que la probabilidad todavía por asignar puede asignarse posteriormente a cualquiera de sus valores posibles. Esto se ilustra en el ejemplo siguiente.

Ejemplo 9.12 Calculando cotas de probabilidad. Considérese la red Bayesiana del Ejemplo 9.11. Inicialmente, las cotas inferior y superior son 0 y 1. En la Etapa 1, se asigna 0.9 a las realizaciones de la forma $(0, x_2, x_3)$ y 0.1 a las de la forma $(1, x_2, x_3)$. Por ello, el mínimo valor de $p(X_1 = 0)$ es 0.9

		$p(x_1)$		$p(x_2)$		$p(x_3)$		
x^j	$p(x^j)$	0	1	0	1	0	1	2
$(0,1,0)$	0.576	1.000	0.000	0.000	1.000	1.000	0.000	0.000
$(0,0,2)$	0.108	1.000	0.000	0.158	0.842	0.842	0.000	0.158
$(0,1,2)$	0.072	1.000	0.000	0.143	0.857	0.762	0.000	0.238
$(0,1,1)$	0.072	1.000	0.000	0.130	0.870	0.696	0.087	0.217
$(1,1,0)$	0.063	0.929	0.071	0.121	0.879	0.717	0.081	0.202
$(0,0,1)$	0.036	0.932	0.068	0.155	0.845	0.689	0.117	0.194
$(0,0,0)$	0.036	0.935	0.065	0.187	0.813	0.701	0.112	0.187
$(1,1,2)$	0.018	0.917	0.083	0.183	0.817	0.688	0.110	0.202
$(1,1,1)$	0.009	0.909	0.091	0.182	0.818	0.682	0.118	0.200
$(1,0,1)$	0.004	0.905	0.095	0.185	0.815	0.679	0.122	0.199
$(1,0,2)$	0.003	0.903	0.097	0.188	0.812	0.677	0.121	0.202
$(1,0,0)$	0.003	0.900	0.100	0.190	0.810	0.678	0.121	0.201

TABLA 9.7. Realizaciones generadas por el método de búsqueda de la máxima probabilidad y las probabilidades aproximadas de cada nodo basadas en las j primeras realizaciones.

(cota inferior) y el máximo es el mismo valor, puesto que la probabilidad restante, 0.1 ha sido ya asignada a las realizaciones con $X_1 = 1$, y ya no queda más probabilidad disponible. Similarmente, las cotas inferior y superior para $p(X_1 = 1)$ son iguales a 0.1. Debido a la amplitud nula de este intervalo, las cotas coinciden con el valor exacto y ya no serán modificadas en las etapas posteriores.

En la Etapa 2, se modifican las cotas de la variable X_2 puesto que esta variable se ha utilizado en el proceso de ramificación. Se asignan probabilidades 0.18 y 0.72 (un total de 0.90) a los grupos de realizaciones de la forma $(0,0,x_3)$ y $(0,1,x_3)$, respectivamente, que son las cotas inferiores de $p(X_2 = 0)$ y $p(X_2 = 1)$, respectivamente. La probabilidad restante, 0.1, queda por asignar. Por ello, podría ser asignada a cualquiera de ellas. En consecuencia, se pueden obtener cotas superiores añadiendo 0.1 a las cotas inferiores. El mismo método se aplica para obtener las cotas en las restantes etapas.

Las Tablas 9.8 y 9.9 muestran las probabilidades marginales exactas de los tres nodos X_1, X_2 y X_3 y las cotas inferiores y superiores correspondientes a las Etapas 1−8. ■

De la discusión anterior es claro que usando el método de la búsqueda de la máxima probabilidad se puede

- Controlar el error de las probabilidades marginales de los nodos.
- Tratar con diferentes errores para las diferentes variables.

		Etapa 1		Etapa 2		Etapas 3 y 4	
Marginal	Exacto	Inf.	Sup.	Inf.	Sup.	Inf.	Sup.
$X_1 = 0$	0.900	0.9	0.9	–	–	–	–
$X_1 = 1$	0.100	0.1	0.1	–	–	–	–
$X_2 = 0$	0.190	0.0	1.0	0.180	0.280	0.180	0.280
$X_2 = 1$	0.810	0.0	1.0	0.720	0.820	0.720	0.820
$X_3 = 0$	0.678	0.0	1.0	0.0	1.0	0.576	0.856
$X_3 = 1$	0.121	0.0	1.0	0.0	1.0	0.072	0.352
$X_3 = 2$	0.201	0.0	1.0	0.0	1.0	0.072	0.352

TABLA 9.8. Valores exactos y cotas inferiores y superiores de las probabilidades marginales (Etapas 1 a 4).

		Etapas 5 y 6		Etapa 7		Etapa 8	
Marginal	Exacto	Inf.	Sup.	Inf.	Sup.	Inf.	Sup.
$X_1 = 0$	0.900	–	–	–	–	–	–
$X_1 = 1$	0.100	–	–	–	–	–	–
$X_2 = 0$	0.190	0.180	0.280	0.190	0.190	–	–
$X_2 = 1$	0.810	0.720	0.820	0.810	0.810	–	–
$X_3 = 0$	0.678	0.612	0.712	0.612	0.712	0.675	0.685
$X_3 = 1$	0.121	0.108	0.208	0.108	0.208	0.117	0.127
$X_3 = 2$	0.201	0.180	0.280	0.180	0.280	0.198	0.208

TABLA 9.9. Valores exactos y cotas inferiores y superiores de las probabilidades marginales (Etapas 5 a 8).

- Utilizar los errores como criterio de parada.

9.10.1 Tratando con la Evidencia

Si se dan una serie de nodos evidenciales, E, es necesario modificar el proceso de la forma siguiente. Se cambia la probabilidad condicional $p(x_i|\pi_i)$ asociada a los nodos evidenciales $X_i \in E$ asignando a dichos nodos la evidencia correspondiente. Entonces, se consideran las funciones de probabilidad condicionada $p_e(x_i|\pi_i)$, definidas en (9.14), como

$$p_e(x_i|\pi_i) = \begin{cases} p(x_i|\pi_i), & \text{si } x_i \text{ y } \pi_i \text{ son consistentes con } e, \\ 0, & \text{en otro caso.} \end{cases}$$

Por tanto, se añade al árbol solamente la rama correspondiente al valor evidencial e_i. Esto se ilustra a continuación, mediante un ejemplo.

Ejemplo 9.13 Tratando con la evidencia. Considérese la red Bayesiana del Ejemplo 9.11 y supóngase que es conocida la evidencia $X_2 = 0$. Antes se creaba un árbol de búsqueda, ahora es necesario cambiar la función de probabilidad condicionada de X_2 a

$$p(X_2 = 0|X_1 = 0) = 0.2, \qquad p(X_2 = 0|X_1 = 1) = 0.1,$$
$$p(X_2 = 1|X_1 = 0) = 0.0, \qquad p(X_2 = 1|X_1 = 1) = 0.0.$$

Las restantes funciones de probabilidad condicionadas permanecen inalteradas. El correspondiente árbol de búsqueda se muestra en la Figura 9.26. Nótese que la asignación de probabilidades anterior fuerza al nodo evidencial X_2 se le ha forzado a tomar el valor evidencial 0.

Finalmente, se normalizan las probabilidades. ■

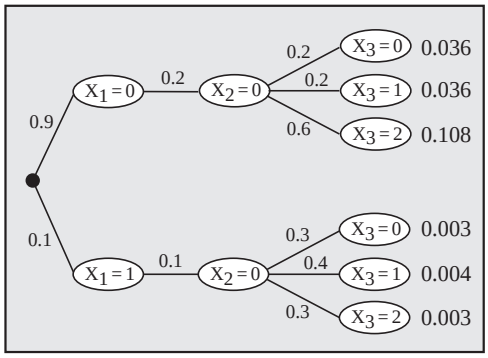


FIGURA 9.26. Incorporando la evidencia al método de la máxima probabilidad.

9.10.2 Aspectos de Implementación

El pseudocódigo para el método de búsqueda de la máxima probabilidad se da en la Figura 9.27. Por razones de simplicidad, en el algoritmo se consideran variables binarias. Se utiliza una cola de prioridades (una estructura de datos en la que se definen las operaciones "insertar" y "obtener el máximo", siendo posible realizar ambas en un máximo de log(tamaño de la cola) operaciones) para almacenar la estructura del árbol de búsqueda y para obtener la rama con máxima probabilidad. Para toda *rama* del árbol, utilizando la cola se almacena la realización parcial asociada, *rama.rea*, y su correspondiente probabilidad, *rama.prob*.

Nótese que para calcular las cotas, se utiliza la suma de los logaritmos de las probabilidades en vez del producto de las probabilidades. Esto disminuye la inestabilidad numérica que ocurre generalmente cuando se multiplican muchos números pequeños.

```
Iniciar
    cola ← {(X1 = 0, p(X1 = 0)), (X1 = 1, p(X1 = 1))}
    ProbTotal ← 0

Ciclo Principal
    mientras ProbTotal < δ hacer
        rama ← obtener el máximo elemento de cola.
        si (|rama.rea| = n) entonces
            Añadir rama.rea a la muestra
            Añadir rama.prob a ProbTotal
            Eliminar rama de cola
        si (|rama.rea| < n) entonces
            expandir rama en cola
```

FIGURA 9.27. Pseudocódigo para el método de búsqueda de la máxima probabilidad.

9.10.3 Resultados Experimentales

Para evaluar el comportamiento del algoritmo de búsqueda de la máxima probabilidad se han realizado algunos experimentos. Se ha generado aleatoriamente una red Bayesiana con diez variables binarias. En el primer experimento, las probabilidades condicionales han sido generadas con números aleatorios del intervalo unidad y en el segundo experimento los valores se han obtenido de $[0, 0.1] \cup [0.9, 1]$. Los experimentos se han realizado utilizando los algoritmos con un mínimo de probabilidad acumulada δ de $\{0.8, 0.9, 0.95, 0.975, 0.98125, 0.99\}$. Los rendimientos se han evaluado por (a) el número de realizaciones completas generadas, (b) el máximo tamaño de la cola, y (c) el tiempo empleado en la simulación. Los resultados que se muestran son las medias de diez simulaciones realizadas en idénticas condiciones.

La Figura 9.28 muestra los resultados para el caso en el que las tablas de probabilidad condicional se han seleccionado del intervalo unidad. La Figura 9.29 muestra los resultados para el intervalo $[0, 0.1] \cup [0.9, 1]$. Como cabía esperar, el tiempo de ejecución aumenta cuando la probabilidad acumulada aumenta. Cuando se comparan las Figuras 9.28 y 9.29, se comprueba que todas las calidades son del orden de diez veces peores para las distribuciones cuyas probabilidades proceden del intervalo unidad. Por tanto, el método de búsqueda de la máxima probabilidad da mejores resultados en los casos en que hay probabilidades extremas. Esto se debe al mayor tamaño de los mayores intervalos que aparecen cuando las distribuciones contienen probabilidades extremas.

En la Figura 9.29 es sorprendente comprobar que el tamaño máximo de la cola es constante para todos los valores de δ. Aparentemente, tras

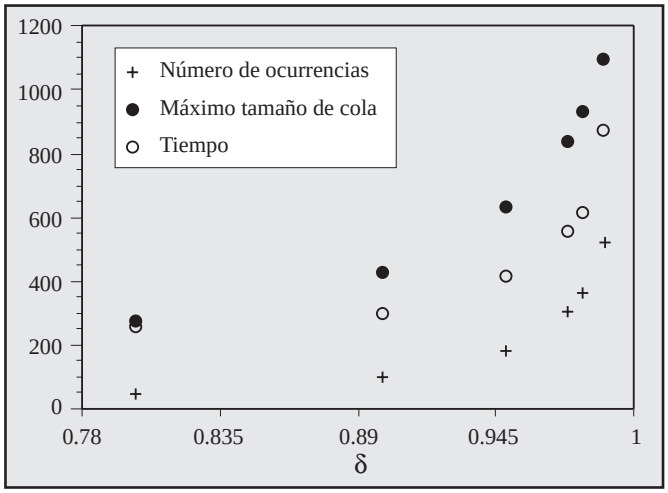


FIGURA 9.28. Diagrama de dispersión de tres medidas de rendimiento en función de δ para el caso en que las probabilidades se han elegido del intervalo $[0, 1]$.

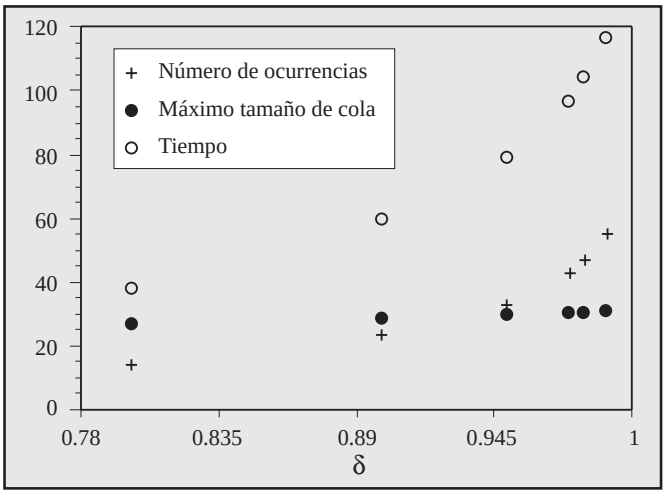


FIGURA 9.29. Diagrama de dispersión de tres medidas de rendimiento en función de δ para el caso en que las probabilidades se han elegido de $[0, 0.1] \cup [0.9, 1]$.

el almacenamiento de un cierto número de realizaciones, la expansión de la cola está equilibrada con el número de realizaciones que se eliminan de ésta. Esto es interesante, puesto que significa que no es necesario almacenar todas las realizaciones posibles en la cola al mismo tiempo.

9.11 Análisis de Complejidad

Cooper (1990) muestra que la complejidad de la inferencia exacta en redes Bayesianas es un problema NP-complejo. Recientemente, Dagum y Luby (1993) han demostrado que la complejidad de la propagación aproximada es también NP-complejo en la precisión de la aproximación. Sin embargo, si la precisión se mantiene fija, entonces, el tiempo de ejecución se mantiene lineal en el número de variables. El hecho de que el caso más desfavorable sea NP-complejo no significa que esto sea cierto para todos los casos que se encuentren en la práctica. La ventaja principal de los métodos de simulación frente a los métodos exactos es que, para un nivel de precisión dado, sus complejidades son lineales en el número de nodos independientemente del grado de conectividad del grafo.

Los problemas de complejidad se deben al hecho de que los errores absolutos en p_i pueden ser una función exponencial del número de nodos. Esto ocurre para todas las probabilidades pequeñas p_i, tales como las que se utilizan en problemas de fiabilidad. Sin embargo, se puede fijar un error absoluto tal que para la estimación anterior tal precisión es innecesaria. Por ejemplo, en una central nuclear la probabilidad de algunos sucesos críticos debe ser estimada con un error absoluto de 10^{-7} ó 10^{-8}, pero este error es fijo y no depende del número de nodos. Además, en muchas aplicaciones prácticas, los intervalos de confianza simultáneos sólo son necesarios para unas pocas celdas.

Ejercicios

9.1 Considérese la función de probabilidad $p(x) = 3x(2-x)/4$ dada en el Ejemplo 9.2. Obtener una muestra de tamaño 20 usando

(a) La función de distribución asociada $p(x) = (3x^2 - x^3)/4$ (véase la Figura 9.3).

(b) El método de aceptación-rechazo con $h(x) = 1/2$ y $g(x) = x(2-x)$.

Comparar las muestras obtenidas.

9.2 Considérese la función de probabilidad poblacional $p(x) = x^2$ y la de simulación $h(x) = x$. Calcular una función $g(x)$ y la constante asociada c que satisface (9.2) y repetir los mismos cálculos mostrados en el ejercicio anterior.

9.3 Determinar todas las ordenaciones ancestrales de la red Bayesiana de la Figura 9.5 (usar la Figura 9.8). ¿Cuál es la ordenación más eficiente para rechazar las realizaciones que no satisfacen la evidencia?

x_1	x_3	$p(x_3\|x_1)$
0	0	0.01
0	1	0.99
1	0	0.01
1	1	0.99

x_2	x_4	$p(x_4\|x_2)$
0	0	0.01
0	1	0.99
1	0	0.01
1	1	0.99

TABLA 9.10. Nuevas distribuciones de probabilidad condicionadas de los nodos X_3 y X_4 de la red Bayesiana dada en la Figura 9.5.

9.4 Considérese el muestreo de aceptación-rechazo dado en el Ejemplo 9.4:

- Calcular la función de probabilidad "a posteriori" de los cuatro nodos no evidenciales de la Figura 9.5 usando muestras de tamaño 10, 20 y 100, respectivamente.
- Utilizar cualquiera de los métodos de propagación exacta del Capítulo 8 para calcular los valores exactos de estas probabilidades.
- Analizar la convergencia de las probabilidades aproximadas para los tres tamaños de muestra.
- ¿Cuál es el porcentaje de realizaciones rechazadas en este ejemplo?

9.5 Repetir el ejercicio anterior considerando las nuevas probabilidades condicionales de los nodos X_3 y X_4 dados en la Tabla 9.10. ¿Cuál es el porcentaje de realizaciones rechazadas?

9.6 Repetir los dos ejercicios previos usando el método de la función de verosimilitud pesante. Comparar los resultados obtenidos con los que resultan de aplicar el método de aceptación-rechazo.

9.7 Repetir las etapas del algoritmo del muestreo hacia adelante y hacia atrás mostrado en el Ejemplo 9.7, considerando las ordenaciones alternativas (X_4, X_3, X_5, X_6) y (X_6, X_4, X_5, X_3), donde los nodos X_2, X_3 y X_4 se muestrean hacia atrás y los nodos X_5 y X_6 se muestrean hacia adelante.

9.8 Generar dos muestras sistemáticas, una de tamaño cinco y otra de tamaño diez, de la red de la Figura 9.13 y las funciones de probabilidad condicionada de la Tabla 9.5.

9.9 Considérese la red Bayesiana dada en la Figura 9.13 con los valores de las probabilidades condicionadas de la Tabla 9.6. Utilizar el muestreo sistemático para obtener una muestra de tamaño 10

(véase la Figura 9.21). Repítase el mismo proceso usando el método modificado. Compárese el rendimiento de ambos métodos.

9.10 Considérese la red Bayesiana con seis nodos del Ejemplo 9.3 y utilícese el método de búsqueda de la máxima probabilidad, etapa a etapa, calculando las correspondientes cotas inferior y superior para las probabilidades de las variables. ¿Son las cotas de los errores dependientes del orden de las variables?

Capítulo 10
Propagación Simbólica de Evidencia

10.1 Introducción

En los capítulos 8 y 9 se han introducido varios métodos para la propagación exacta y aproximada de evidencia en los modelos de redes probabilísticas. Estos métodos requieren que la función de probabilidad conjunta del modelo se especifique numéricamente, es decir, que se asignen valores numéricos fijos a todos los parámetros. Sin embargo, la especificación numérica de estos parámetros puede no ser posible. También puede suceder que los especialistas sólo conozcan intervalos de valores para algunos/todos los parámetros en vez de sus valores exactos. En tales casos, los métodos numéricos de propagación deben reeemplazarse por métodos simbólicos, que son capaces de tratar los parámetros mismos, sin necesidad de asignarles valores.

Los métodos de propagación simbólica conducen a soluciones que se expresan como funciones de los parámetros. Por ello, las respuestas a cuestiones generales pueden darse en forma simbólica en función de los parámetros, y las respuestas a preguntas específicas pueden obtenerse sin más que sustituir los valores de los parámetros en la solución simbólica, sin necesidad de rehacer la propagación. Por otra parte, la propagación simbólica permite estudiar con escaso esfuerzo computacional la sensibilidad de los resultados a cambios en los valores de los parámetros. La propagación simbólica es especialmente útil en los casos siguientes:

1. Cuando no se dispone de la especificación numérica del modelo probabilístico.

2. Cuando los especialistas sólo son capaces de especificar intervalos de los parámetros en vez de valores concretos. En este caso, los métodos de propagación simbólica pueden utilizarse para obtener cotas inferiores y superiores de las probabilidades para todos los valores posibles de los parámetros en los intervalos dados.

3. Cuando se requiere un análisis de sensibilidad. Una de las cuestiones que surge normalmente en este contexto es, ¿cómo son de sensibles los resultados a cambios en los parámetros y a los valores evidenciales?

Los algoritmos de propagación simbólica han sido introducidos recientemente en la literatura. Por ejemplo, Castillo, Gutiérrez y Hadi (1995c, 1995d) realizan la propagación simbólica adaptando algunos de los métodos numéricos de propagación descritos en el Capítulo 8 a este tipo de propagación. Estos métodos realizan los cálculos simbólicos necesarios utilizando paquetes de cálculo con posibilidades simbólicas (tales como *Mathematica* y *Maple*).

Otro método con capacidades de cálculo simbólico es el algoritmo de inferencia probabilística simbólica (SPI) (Shachter, D'Ambrosio y DelFabero (1990) y Li y D'Ambrosio (1994)). Este método es orientado a un objetivo y analiza sólo los cálculos que se requieren para responder a la pregunta en estudio. Con este método, los resultados se obtienen posponiendo la evaluación de las expresiones y manteniéndolas en forma simbólica.

Sin embargo, los dos métodos anteriores tienen el mismo problema: necesitan utilizar programas especiales, o un esfuerzo computacional extra, para poder tratar las expresiones simbólicas. Por otra parte, el cálculo y la simplificación de expresiones simbólicas es una tarea computacionalmente cara, y a veces progresivamente ineficiente cuando se trata con grandes redes o grandes conjuntos de parámetros simbólicos. Recientemente, Castillo, Gutiérrez y Hadi (1996c) han introducido una solución a la propagación simbólica que utiliza ventajosamente la estructura simbólica polinomial de las probabilidades de los nodos (véase la Sección 7.5.1) para evitar los cálculos simbólicos. La principal idea del método consiste en obtener las expresiones simbólicas mediante un algoritmo numérico que calcula los coeficientes de los polinomios correspondientes. Entonces, todos los cálculos se hacen de forma numérica, evitando la manipulación de expresiones simbólicas costosas.

En la Sección 10.2 se introduce la notación y el marco de trabajo de los métodos simbólicos. La Sección 10.3 discute la generación automática de código simbólico. La estructura algebraica de las probabilidades se analiza en la Sección 10.4. La Sección 10.5 muestra cómo se pueden utilizar los métodos de propagación simbólica de evidencia para obtener las expresiones simbólicas de las probabilidades de ocurrencia de los nodos. La Sección 10.6 presenta una mejora del método anterior para el caso de tareas orientadas a un objetivo. La Sección 10.7 trata del problema de la evidencia aleatoria simbólica. La Sección 10.8 muestra cómo hacer un estudio de sensibilidad

mediante los métodos simbólicos. Finalmente, la Sección 10.9 analiza el problema de la propagación simbólica de la evidencia en redes Bayesianas normales.

10.2 Notación y Conceptos Preliminares

En el Capítulo 6 se ha visto que la función de probabilidad conjunta asociada a las redes probabilísticas de Markov descomponibles y Bayesianas puede darse mediante una factorización como producto de probabilidades condicionales

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i). \tag{10.1}$$

En el caso de redes Bayesianas, los conjuntos condicionantes son los padres del nodo, Π_i, $i = 1, \ldots, n$. En el caso de redes de Markov descomponibles, estos conjuntos se obtienen aplicando la regla de la cadena a la factorización obtenida a partir de la cadena de conglomerados (véase el Capítulo 6). Por tanto, aunque algunos de los métodos introducidos en este capítulo pueden ser fácilmente extendidos para tratar una representación potencial de la función de probabilidad conjunta, por simplicidad, pero sin pérdida de generalidad, se utiliza el conjunto de probabilidades condicionales en (10.1) como representación paramétrica básica de la función de probabilidad conjunta.

Sea $X = \{X_1, \ldots, X_n\}$ un conjunto de n variables discretas, cada una de las cuales puede tomar valores en el conjunto $\{0, 1, \ldots, r_i\}$, y sea $B = (D, P)$ una red Bayesiana definida sobre X, donde el grafo dirigido acíclico D determina la estructura del conjunto de probabilidades condicionales, y $P = \{p(x_1|\pi_1), \ldots, p(x_n|\pi_n)\}$ es el conjunto de probabilidades condicionales que se necesitan para especificar la función de probabilidad conjunta.

Algunas de las probabilidades condicionales en (10.1) pueden darse en forma numérica y otras en forma simbólica, es decir, $p(x_i|\pi_i)$ pueden ser familias paramétricas o probabilidades totalmente especificadas numéricamente.

Definición 10.1 Nodo Simbólico. *Cuando $p(x_i|\pi_i)$ es una familia paramétrica simbólica (es decir, depende de al menos un parámetro en forma simbólica), el nodo X_i se denomina un nodo simbólico, y se utiliza Θ_i para denotar sus correspondientes parámetros simbólicos.*

Como se ha visto en la Sección 7.5.1, cuando $p(x_i|\pi_i)$ es una familia paramétrica, es decir, cuando X_i es un nodo simbólico, una elección conveniente de los parámetros es la siguiente

$$\theta_{ij\pi} = p(X_i = j|\Pi_i = \pi),\ j \in \{0, \ldots, r_i\}, \tag{10.2}$$

donde π es cualquier posible realización de los padres, Π_i, de X_i. Por ello, el primer subíndice de $\theta_{ij\pi}$ se refiere al número del nodo, el segundo subíndice se refiere al estado del nodo, y los restantes subíndices se refieren a las realizaciones de sus padres. Puesto que $\sum_{j=0}^{r_i} \theta_{ij\pi} = 1$, para todo i y π, no todos los parámetros son libres, es decir, uno cualquiera de ellos puede ser escrito como la unidad menos la suma del resto. Por ejemplo, el primer parámetro puede escribirse como

$$\theta_{i0\pi} = 1 - \sum_{j=1}^{r_i} \theta_{ij\pi}. \tag{10.3}$$

Para simplificar la notación en los casos en los que la variable X_i no tiene padres, se utiliza θ_{ij} para denotar $p_i(X_i = j),\ j \in \{0, \ldots, r_i\}$. Se ilustra esta notación usando el ejemplo siguiente.

Ejemplo 10.1 Nodos simbólicos. Considérese una red Bayesiana discreta consistente en las variables $X = \{X_1, \ldots, X_8\}$ cuyo correspondiente grafo dirigido acíclico se muestra en la Figura 10.1. La estructura del grafo implica que la probabilidad conjunta del conjunto de nodos puede escribirse en la forma (10.1), como

$$p(x) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3)p(x_5|x_3)p(x_6|x_4)p(x_7|x_4)p(x_8|x_5). \tag{10.4}$$

Por simplicidad, y sin pérdida de generalidad, supóngase que todos los nodos representan variables binarias con valores en el conjunto $\{0, 1\}$. Esto y la estructura de la distribución de probabilidad en (10.4) implica que la función de probabilidad conjunta de las ocho variables depende de 34 parámetros $\Theta = \{\theta_{ij\pi}\}$. Nótese, sin embargo, que solamente 17 de ellos son libres (puesto que las probabilidades en cada una de las probabilidades condicionales deben sumar la unidad). Estos 17 parámetros se dan en la Tabla 10.1.

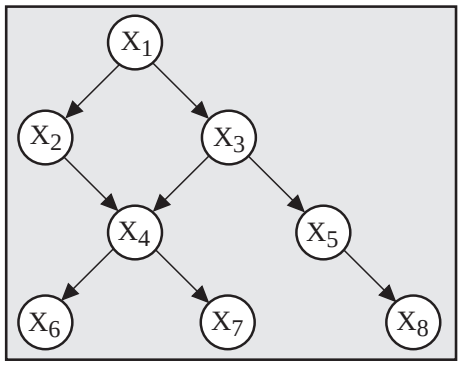


FIGURA 10.1. Un grafo dirigido acíclico.

X_i	Π_i	Parámetros libres
X_1	ϕ	$\theta_{10} = p(X_1 = 0) = 0.2$
X_2	X_1	$\theta_{200} = p(X_2 = 0\|X_1 = 0) = 0.3$ $\theta_{201} = p(X_2 = 0\|X_1 = 1) = 0.5$
X_3	X_1	$\theta_{300} = p(X_3 = 0\|X_1 = 0)$ $\theta_{301} = p(X_3 = 0\|X_1 = 1) = 0.5$
X_4	X_2, X_3	$\theta_{4000} = p(X_4 = 0\|X_2 = 0, X_3 = 0) = 0.1$ $\theta_{4001} = p(X_4 = 0\|X_2 = 0, X_3 = 1) = 0.8$ $\theta_{4010} = p(X_4 = 0\|X_2 = 1, X_3 = 0) = 0.3$ $\theta_{4011} = p(X_4 = 0\|X_2 = 1, X_3 = 1) = 0.4$
X_5	X_3	$\theta_{500} = p(X_5 = 0\|X_3 = 0) = 0.3$ $\theta_{501} = p(X_5 = 0\|X_3 = 1) = 0.1$
X_6	X_4	$\theta_{600} = p(X_6 = 0\|X_4 = 0)$ $\theta_{601} = p(X_6 = 0\|X_4 = 1) = 0.9$
X_7	X_4	$\theta_{700} = p(X_7 = 0\|X_4 = 0) = 0.3$ $\theta_{701} = p(X_7 = 0\|X_4 = 1) = 0.6$
X_8	X_5	$\theta_{800} = p(X_8 = 0\|X_5 = 0) = 0.2$ $\theta_{801} = p(X_8 = 0\|X_5 = 1) = 0.4$

TABLA 10.1. El conjunto de parámetros libres asociados a las distribuciones condicionales en (10.4).

En este ejemplo, sólo los nodos X_3 y X_6 son nodos simbólicos puesto que sus correspondientes funciones de probabilidad condicionada contienen al menos un parámetro simbólico. Se tienen los conjuntos de parámetros $\Theta_3 = \{\theta_{300}, \theta_{310}\}$ y $\Theta_6 = \{\theta_{600}, \theta_{610}\}$. Nótese que estos conjuntos incluyen todos los parámetros simbólicos, no sólo los parámetros libres. Por ello, el conjunto de parámetros simbólicos asociados a la red Bayesiana es $\Theta = \{\Theta_3, \Theta_6\}$. ∎

10.3 Generación Automática de Código Simbólico

El tratamiento con parámetros simbólicos es idéntico al tratamiento con valores numéricos, con la única diferencia de que las operaciones requeridas deben realizarse con un programa capaz de manipular símbolos en vez de números. Los cálculos simbólicos, sin embargo, son mucho más lentos que los numéricos y requieren más memoria. Se ha visto en los Ejemplos 7.13 y 7.14 que la propagación simbólica puede realizarse directamente usando las expresiones para las probabilidades marginales y condicionales dadas por

(3.4). Sin embargo, este método de resolver el problema es muy costoso computacionalmente, y resulta ineficiente incluso con números reducidos de variables.

Una alternativa a este método consiste en adaptar algunos de los algoritmos de propagación numérica introducidos en el Capítulo 8 para la computación simbólica. Castillo, Gutiérrez y Hadi (1995d) muestran que la adaptación simbólica de estos métodos requiere sólo pequeñas modificaciones. Por ejemplo, el algoritmo de propagación por agrupamiento (Algoritmo 8.4) puede adaptarse fácilmente a la propagación simbólica utilizando una herramienta informática simbólica, tal como *Mathematica*. En esta sección se examinan las capacidades simbólicas de este algoritmo de propagación. Seguidamente se ilustra su aplicación mediante un ejemplo.

Del análisis de los algoritmos de agrupamiento introducidos en el Capítulo 8, en particular, el algoritmo para propagación en modelos de redes probabilísticas usando árboles de unión (Algoritmos 8.4 y 8.5), se puede concluir lo siguiente:

1. El número de argumentos de las funciones que intervienen (probabilidades condicionales, funciones potenciales, y funciones de probabilidad de los conglomerados) depende de la topología de red. Además, las funciones potenciales pueden construirse de varias formas distintas a partir de las funciones de probabilidad condicionada que aparecen en (10.1).

2. Dados dos conglomerados vecinos C_i y C_j con conjunto separador S_{ij}, el mensaje $M_{ij}(s_{ij})$ que el conglomerado C_i envía a C_j es

$$M_{ij}(s_{ij}) = \sum_{c_i \setminus s_{ij}} \psi_i(c_i) \prod_{k \neq j} M_{ki}(s_{ki}), \tag{10.5}$$

   una función que depende de los parámetros simbólicos contenidos en C_i y todos sus conglomerados vecinos excepto C_j. Por ello, el máximo número de parámetros simbólicos que intervienen en este mensaje está acotado por el máximo número de vecinos de un conglomerado.

3. Los mensajes entre conglomerados deben ser ordenados de tal forma que no se utilice ningún mensaje antes de que haya sido definido. Una vez que todos los mensajes han sido definidos, la función de probabilidad conjunta de los conglomerados ya puede ser calculada usando la expresión

$$p(c_i) = \psi_i(c_i) \prod_{k} M_{ki}(s_{ik}). \tag{10.6}$$

   Esta expresión depende de los parámetros simbólicos contenidos en C_i y en todos sus conglomerados vecinos C_k.

4. La función de probabilidad marginal de un nodo puede ser calculada de varias formas mediante la marginalización de la función de probabilidad conjunta de cualquier conglomerado que contenga al nodo. Sin embargo, cuanto más pequeño sea el tamaño del conglomerado seleccionado, menor será el número de cálculos requeridos.

Por ello, la escritura del código simbólico correspondiente a un árbol de conglomerados requiere analizar su topología para construir un árbol de unión para la propagación simbólica. Puede construirse fácilmente un generador con este propósito, sin más que basarse en el Algoritmo 8.4. En este algoritmo, el código simbólico se escribe en el orden indicado por el algoritmo, de tal forma que las definiciones de funciones, mensajes y probabilidades coincide con el orden requerido por el algoritmo. Tal como se muestra en el ejemplo siguiente, los códigos de las Figuras 10.3 y 10.4 han sido generados por un programa escrito en lenguaje C++.[1]

Ejemplo 10.2 Código simbólico. En este ejemplo se ilustra una implementación simbólica del Algoritmo 8.4 usando *Mathematica* para resolver el problema de la propagación de evidencia. Considérese de nuevo el grafo dirigido de la Figura 10.1 cuya función de probabilidad conjunta asociada es (10.4). La Figura 10.2 muestra el correspondiente árbol de unión con los conglomerados del grafo moralizado y triangulado (véase la Sección 4.5).

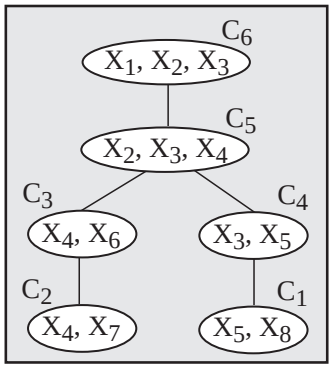


FIGURA 10.2. Un Árbol de unión para el grafo moralizado y triangulado correspondiente al grafo dirigido acíclico de la Figura 10.1.

Las Figuras 10.3 y 10.4 muestran el código en *Mathematica* para la propagación simbólica. Un código similar puede obtenerse utilizando *Maple*, *Axiom*, o cualquier otro paquete de cálculo simbólico. Estos códigos simbólicos

[1]El programa *X-pert simbólico* para la generación automática de código simbólico para *Mathematica* y *Maple* pueden obtenerse en la dirección de World Wide Web http://ccaix3.unican.es/~AIGroup.

han sido generados automáticamente por el generador de código anteriormente mencionado. Este generador lee, en primer lugar, las estructuras gráficas y probabilísticas de la red Bayesiana, en términos de parámetros numéricos y simbólicos. Seguidamente, genera el código simbólico correspondiente, que permite la propagación.

Inicialmente, se definen las probabilidades requeridas para especificar $p(x)$ en (10.4) utilizando a y b para denotar los parámetros simbólicos θ_{300} y θ_{600}, contenidos en las probabilidades condicionales dadas en la Tabla 10.1.

Por claridad de exposición, se seleccionan algunos parámetros numéricos y otros simbólicos. Sin embargo, el tratamiento con todos los parámetros simbólicos es idéntico y no afecta al resto del código. Bajo el encabezamiento "Tablas de Probabilidad" se asignan los valores numéricos y simbólicos a los parámetros, y se da una lista de comandos para construir las probabilidades condicionales de cada nodo. Los nombres de las funciones que generan estas tablas son $P1, P2, \ldots, P8$. Nótese que el generador automático de código simbólico tiene en cuenta los diferentes números de argumentos que intervienen en las diferentes tablas de probabilidad. Nótese también que sólo se necesitan los parámetros libres para la definición de la función de probabilidad conjunta. La etapa siguiente del algoritmo es la definición de las funciones potenciales ψ_i, una por conglomerado. Los conglomerados correspondientes al grafo de la Figura 10.1 se muestran en la Figura 10.2. Entonces, las funciones $\{F1, F2, F3, F4, F5, F6\}$ se definen asignando cada una de la tablas de probabilidad condicional a un conglomerado que incluya su correspondiente familia de nodos. Seguidamente, se definen los rangos de las variables, $\{(inf[i], sup[i]),\ i = 1, 2, \ldots, 8\}$, que son $\{0, 1\}$, puesto que se trata de variables binarias.

En el código de la Figura 10.4, los mensajes que se envían entre conglomerados vecinos se definen usando (10.5). Estos mensajes se denotan por

$$\{M14, M23, M34, M45, M56, M65, M53, M54, M32, M41\},$$

donde los índices se refieren a los conglomerados que envían y reciben los mensajes, respectivamente. El orden de estos comandos está impuesto parcialmente por la topología de red, puesto que algunos mensajes dependen de los mensajes anteriores. Sin embargo, el generador de código simbólico tiene esto en cuenta a la hora de decidir este orden.

Seguidamente, utilizando las funciones potenciales y los mensajes se calcula la función de probabilidad conjunta de todos los conglomerados $\{Q1, Q2, Q3, Q4, Q5, Q6\}$, usando (10.6). Los comandos necesarios de *Mathematica* se muestran en la Figura 10.4 bajo el encabezamiento "FPCs de los conglomerados."

Seguidamente, se calculan las funciones de probabilidad marginales de cada nodo sin más que marginalizar uno de los conglomerados que contenga

```
(* Tablas de Probabilidad *)

T={0.2,0.8};
n=1; Do[P1[i1]=T[[n]];n++,{i1,0,1}];
T={0.3,0.5,0.7,0.5};
n=1; Do[P2[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];
T={a,0.5,1-a,0.5};
n=1; Do[P3[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];
T={0.1,0.8,0.3,0.4,0.9,0.2,0.7,0.6};
n=1; Do[P4[i1,i2,i3]=T[[n]];n++,{i1,0,1},{i2,0,1},{i3,0,1}];
T={0.3,0.1,0.7,0.9};
n=1; Do[P5[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];
T={b,0.9,1-b,0.1};
n=1; Do[P6[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];
T={0.3,0.6,0.7,0.4};
n=1; Do[P7[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];
T={0.2,0.4,0.8,0.6};
n=1; Do[P8[i1,i2]=T[[n]];n++,{i1,0,1},{i2,0,1}];

(* Funciones Potenciales *)

F1[X8_,X5_]:=P8[X8,X5];
F2[X7_,X4_]:=P7[X7,X4];
F3[X6_,X4_]:=P6[X6,X4];
F4[X5_,X3_]:=P5[X5,X3];
F5[X4_,X2_,X3_]:=P4[X4,X2,X3];
F6[X3_,X1_,X2_]:=P1[X1]*P2[X2,X1]*P3[X3,X1];

(* Iniciar Rangos *)

Do[inf[i]=0; sup[i]=1, {i,1,8}];
```

FIGURA 10.3. Un programa en *Mathematica* para la propagación simbólica de la evidencia (Parte 1).

al nodo correspondiente. Nótese que para un cierto nodo hay tantas posibilidades como conglomerados lo contienen. La elección del conglomerado con el mínimo tamaño conduce a un esfuerzo computacional mínimo. El generador de código selecciona el óptimo para cada nodo.

Finalmente, se normalizan las probabilidades dividiendo por la suma correspondiente y se imprimen los resultados. Las probabilidades marginales (antes de conocer la evidencia) de los nodos obtenidas con este programa se muestran en la Tabla 10.2. Nótese que se obtienen polinomios en los parámetros y que los exponentes de estos parámetros son siempre la unidad.

En lo que sigue se considera la evidencia $\{X_2 = 1, X_5 = 1\}$. En este caso, la propagación de evidencia puede ser realizada usando el mismo

```
(* Definición de Mensajes *)

M14[X5_]:=Sum[F1[X8,X5],{X8,inf[8],sup[8]}];
M23[X4_]:=Sum[F2[X7,X4],{X7,inf[7],sup[7]}];
M35[X4_]:=Sum[F3[X6,X4]*M23[X4],{X6,inf[6],sup[6]}];
M45[X3_]:=Sum[F4[X5,X3]*M14[X5],{X5,inf[5],sup[5]}];
M56[X2_,X3_]:=Sum[F5[X4,X2,X3]*M35[X4]*M45[X3],{X4,inf[4],sup[4]}];
M65[X2_,X3_]:=Sum[F6[X3,X1,X2],{X1,inf[1],sup[1]}];
M53[X4_]:=Sum[F5[X4,X2,X3]*M45[X3]*M65[X2,X3],{X2,inf[2],sup[2]},
{X3,inf[3],sup[3]}];
M54[X3_]:=Sum[F5[X4,X2,X3]*M35[X4]*M65[X2,X3],{X4,inf[4],sup[4]},
{X2,inf[2],sup[2]}];
M32[X4_]:=Sum[F3[X6,X4]*M53[X4],{X6,inf[6],sup[6]}];
M41[X5_]:=Sum[F4[X5,X3]*M54[X3],{X3,inf[3],sup[3]}];

(* FPCs de los Conglomerados *)

Q1[X8_,X5_]:=F1[X8,X5]*M41[X5];
Q2[X7_,X4_]:=F2[X7,X4]*M32[X4];
Q3[X6_,X4_]:=F3[X6,X4]*M23[X4]*M53[X4];
Q4[X5_,X3_]:=F4[X5,X3]*M14[X5]*M54[X3];
Q5[X4_,X2_,X3_]:=F5[X4,X2,X3]*M35[X4]*M45[X3]*M65[X2,X3];
Q6[X3_,X1_,X2_]:=F6[X3,X1,X2]*M56[X2,X3];

(* Probabilidades Marginales de los Nodos *)

P[1,X1_]:=Sum[Q6[X3,X1,X2],{X3,inf[3],sup[3]},{X2,inf[2],sup[2]}];
P[2,X2_]:=Sum[Q6[X3,X1,X2],{X3,inf[3],sup[3]},{X1,inf[1],sup[1]}];
P[3,X3_]:=Sum[Q4[X5,X3],{X5,inf[5],sup[5]}];
P[4,X4_]:=Sum[Q3[X6,X4],{X6,inf[6],sup[6]}];
P[5,X5_]:=Sum[Q4[X5,X3],{X3,inf[3],sup[3]}];
P[6,X6_]:=Sum[Q3[X6,X4],{X4,inf[4],sup[4]}];
P[7,X7_]:=Sum[Q2[X7,X4],{X4,inf[4],sup[4]}];
P[8,X8_]:=Sum[Q1[X8,X5],{X5,inf[5],sup[5]}];

(* Normalización e Impresión *)

Do[c=Chop[Simplify[Sum[P[i,t],{t,inf[i],sup[i]}]]];
      Do[R[i,t]:=Simplify[Chop[P[i,t]]/c];
        Print["p(Node",i,"=",t,")=",R[i,t]],
      {t,inf[i],sup[i]}],
{i,1,8}]
```

FIGURA 10.4. Un programa en *Mathematica* para la propagación simbólica de la evidencia (Parte 2).

X_i	$p(x_i = 0)$
X_1	0.2
X_2	0.46
X_3	$0.4 + 0.2\,\theta_{300}$
X_4	$0.424 - 0.056\,\theta_{300}$
X_5	$0.18 + 0.04\,\theta_{300}$
X_6	$0.5184 + 0.0504\,\theta_{300} + 0.424\,\theta_{600} - 0.056\,\theta_{300}\,\theta_{600}$
X_7	$0.4728 + 0.0168\,\theta_{300}$
X_8	$0.364 - 0.008\,\theta_{300}$

TABLA 10.2. Probabilidades iniciales de los nodos.

código, pero igualando los rangos de las variables X_2 y X_5 a $(1, 1)$, es decir, $inf[2] = inf[5] = 1$ y repitiendo los cálculos. La Tabla 10.3 da las nuevas probabilidades de los nodos dada esta evidencia. Nótese que se obtienen funciones racionales, es decir, cocientes de funciones polinómicas en los parámetros con exponentes unidad.

Las Tablas 10.2 y 10.3 pueden utilizarse para responder a todas las preguntas referentes a las probabilidades iniciales o las condicionadas tras conocer la evidencia, simplemente sustituyendo los valores concretos que tomen los parámetros.

En la Figura 10.5 se muestran las probabilidades condicionales de los nodos, dados $X_2 = 1$ y $X_5 = 1$ para los valores de los parámetros $\theta_{300} = 0.4, \theta_{600} = 0.8$, que pueden ser obtenidas de la información de la Tabla 10.3 por simple sustitución. ∎

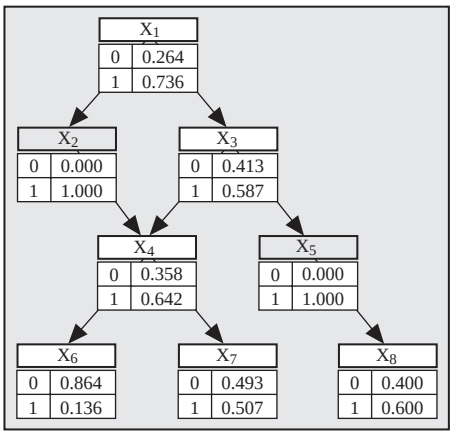


FIGURA 10.5. Probabilidades condicionales de los nodos dada la evidencia $\{X_2 = 1, X_5 = 1\}$.

X_i	$p(x_i = 0 \mid X_2 = 1, X_5 = 1)$
X_1	$\dfrac{0.126 - 0.028\,\theta_{300}}{0.446 - 0.028\,\theta_{300}}$
X_2	0
X_3	$\dfrac{0.14 + 0.098\,\theta_{300}}{0.446 - 0.028\,\theta_{300}}$
X_4	$\dfrac{0.1644 - 0.021\,\theta_{300}}{0.446 - 0.028\,\theta_{300}}$
X_5	0
X_6	$\dfrac{0.25344 - 0.0063\,\theta_{300} + 0.1644\,\theta_{600} - 0.021\,\theta_{300}\theta_{600}}{0.446 - 0.028\,\theta_{300}}$
X_7	$\dfrac{0.21828 - 0.0105\,\theta_{300}}{0.446 - 0.028\,\theta_{300}}$
X_8	$\dfrac{0.178 - 0.011\,\theta_{300}}{0.446 - 0.028\,\theta_{300}} = 0.4$

TABLA 10.3. Probabilidades condicionales de los nodos dada la evidencia $\{X_2 = 1, X_5 = 1\}$.

10.4 Estructura Algebraica de las Probabilidades

Las probabilidades marginales y condicionales poseen una estructura algebraica interesante. Los métodos simbólicos que se presentan en las secciones siguientes utilizan ventajosamente esta estructura. En la Sección 7.5.1 se ha analizado la estructura algebraica de las probabilidades marginales y condicionadas de los nodos en los modelos de redes probabilísticas. Los resultados siguientes proceden de Castillo, Gutiérrez, y Hadi (1995c).

Teorema 10.1 *La probabilidad marginal de una realización dada,* $(x_1, \ldots, x_n)$*, de los nodos de una red Bayesiana es un polinomio en los parámetros simbólicos de grado menor o igual que el número de nodos simbólicos. Sin embargo, es un polinomio de primer grado en cada parámetro.*

Demostración: Según la expresión (10.1) la probabilidad de una realización $(x_1, \ldots, x_n)$ es

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|\pi_i) = \prod_{i=1}^{n} \theta_{ix_i\pi_i}.$$

Nótese que todos los parámetros que aparecen en el producto anterior se asocian a diferentes variables, y algunos de ellos pueden ser especificados numéricamente. Por ello, $p(x_1, \ldots, x_n)$ es un monomio de grado menor que o igual al número de nodos simbólicos. ∎

Corolario 10.1 *La probabilidad marginal de cualquier conjunto de nodos $Y \subset X$ es un polinomio en los parámetros de grado menor que o igual al número de nodos simbólicos. Sin embargo, es un polinomio de primer grado en cada parámetro.*

Por ejemplo, como puede verse en la Tabla 10.2, las probabilidades marginales de todos los nodos de la red Bayesiana son polinomios de primer grado en cada uno de los parámetros simbólicos θ_{300} y θ_{600}.

Corolario 10.2 *La probabilidad condicional de cualquier conjunto de nodos Y dada la evidencia $E = e$ es una función racional (un cociente de dos funciones polinomiales) de los parámetros. Por otra parte, el polinomio denominador es el mismo para todos los nodos.*

Por ejemplo, la Tabla 10.3 muestra que las probabilidades condicionales de los nodos dada la evidencia $\{X_2 = 1, X_5 = 1\}$ es el cociente de dos polinomios, y que el polinomio denominador es el mismo para todos ellos.

Esta estructura polinomial de las probabilidades en los modelos de redes probabilísticas da lugar a un método muy eficiente para la computación simbólica. Este método se presenta en la sección siguiente.

10.5 Propagación Simbólica Mediante Métodos Numéricos

10.5.1 Estructura Polinomial

Supóngase que se trata con un conjunto de nodos simbólicos $\{X_{i_1}, \ldots, X_{i_s}\} \subset X$. Sea $\Theta = \{\Theta_1, \ldots, \Theta_s\}$ el conjunto de parámetros simbólicos asociados, donde Θ_k se refiere a los parámetros simbólicos asociados al nodo simbólico X_{i_k}, con $k = 1, \ldots, s$.

Los Corolarios 10.1 y 10.2 garantizan que las probabilidades condicionales de un nodo típico X_i, dada la evidencia $E = e$, $p(X_i = j|E = e), j = 0, \ldots, r_i$, es o un polinomio o el cociente de dos polinomios de los parámetros simbólicos. Además, el Teorema 10.1 garantiza que cada monomio que forma parte de estos polinomios contiene a lo sumo un parámetro de Θ_k, para cada $k = 1, \ldots, s$. Por tanto, se construye el conjunto de los monomios posibles, M, tomando el producto cartesiano de los conjuntos de parámetros simbólicos libres correspondientes a los diferentes nodos simbólicos, incluyendo un valor numérico 1 para tener en cuenta los posibles valores

numéricos asignados a algunos de los parámetros numéricos del nodo simbólico. Entonces, se tiene

$$M = \{1, \Theta_1\} \times \ldots \times \{1, \Theta_s\}, \tag{10.7}$$

donde Θ_i representa, en este caso, el conjunto de parámetros simbólicos asociados al nodo X_i. Por ello, la forma general de estos polinomios es

$$\sum_{m_r \in M} c_r m_r, \tag{10.8}$$

donde c_r es el coeficiente numérico asociado al monomio $m_r \in M$.

Por ejemplo, la estructura polinomial inicial de las probabilidades de la red Bayesiana dada en el Ejemplo 10.1 resulta

$$M = \{1, \theta_{300}\} \times \{1, \theta_{600}\}.$$

En consecuencia, los polinomios asociados a las probabilidades marginales y condicionales de los nodos son de la forma

$$c_1 + c_2\,\theta_{300} + c_3\,\theta_{600} + c_4\,\theta_{300}\theta_{600}. \tag{10.9}$$

Una representación alternativa de este polinomio puede darse considerando no sólo los parámetros simbólicos libres, sino todos los parámetros simbólicos asociados a los nodos simbólicos. En el ejemplo anterior, el producto cartesiano de todos los parámetros simbólicos asociados a los diferentes nodos simbólicos es

$$M = \{\theta_{300}, \theta_{310}\} \times \{\theta_{600}, \theta_{610}\},$$

que define la siguiente estructura polinomial

$$c_1\,\theta_{300}\theta_{600} + c_2\,\theta_{310}\theta_{600} + c_3\,\theta_{300}\theta_{610} + c_4\,\theta_{310}\theta_{610}. \tag{10.10}$$

Las representaciones en (10.9) y (10.10) son equivalentes. Esto se debe al hecho de que los parámetros de cada nodo deben sumar la unidad. Por ello, comenzando con (10.10), se tiene

$$\begin{aligned} c_1\theta_{300}\theta_{600} + c_2\theta_{310}\theta_{600} + c_3\theta_{300}\theta_{610} + c_4\theta_{310}\theta_{610} &= c_1\theta_{300}\theta_{600} \\ &+ c_2(1-\theta_{300})\theta_{600} + c_3\theta_{300}(1-\theta_{600}) + c_4(1-\theta_{300})(1-\theta_{600}) \\ &= c_4 + \ (c_3 - c_4)\theta_{300} + (c_2 - c_4)\theta_{600} + (c_4 + c_1 - c_2 - c_3)\theta_{300}\theta_{600}, \end{aligned}$$

que es de la misma forma que (10.9). Se verá más tarde que, desde un punto de vista computacional, la representación dada en (10.10) es más conveniente que la dada en (10.9). Por ello, para representar los polinomios en (10.8), se utilizará el conjunto producto

$$M = \Theta_1 \times \ldots \times \Theta_s, \tag{10.11}$$

donde Θ_i representa el conjunto de todos los parámetros simbólicos asociados al nodo simbólico X_{i_s}.

En esta sección se desarrolla un método, que se denomina el *método de las componentes canónicas*, para calcular los coeficientes c_r asociados a una representación de los polinomios. El método asigna en primer lugar valores a los parámetros simbólicos Θ y calcula las probabilidades numéricas resultantes. Una vez que los coeficientes ya son conocidos, los polinomios y, por lo tanto, las probabilidades $p(x_i|e)$ se obtienen fácilmente.

Se mostrará que hay ciertas analogías entre el método de las componentes canónicas y el de los algoritmos de propagación basados en el condicionamiento presentados en la Sección 8.5. Ambos métodos realizan varias propagaciones por métodos numéricos asociadas a diferentes realizaciones de ciertos parámetros de la red para obtener la solución buscada.

10.5.2 Propagación Simbólica Mediante Métodos Numéricos

Sea M el conjunto de los monomios necesarios para calcular $p(X_i = j|E = e)$ para $j = 0, \ldots, r_i$. Sea m el número de monomios en M. De (10.8), el polinomio necesario para calcular $p(X_i = j|E = e)$ es de la forma

$$p(X_i = j|E = e) \propto \sum_{m_k \in M} c_k^{ij} m_k = p_{ij}(\Theta),\ j = 0, \ldots, r_i. \tag{10.12}$$

El término $p_{ij}(\Theta)$ representa las probabilidades sin normalizar $p(X_i = j|E = e)$. Por ello, $p_{ij}(\Theta)$ puede escribirse como una combinación lineal de los monomios en M. Nuestro objetivo ahora consiste en calcular los coeficientes c_k^{ij}.

Si a los parámetros Θ se les asignan valores numéricos, por ejemplo, θ, entonces $p_{ij}(\theta)$ puede obtenerse reemplazando Θ por θ y usando cualquier método de propagación numérica para calcular $p(X_i = j|E = e, \Theta = \theta)$. Similarmente, el monomio m_k toma un valor numérico, que es el producto de los parámetros que aparecen en m_k. Por ello, se tiene

$$p(X_i = j|E = e, \Theta = \theta) \propto \sum_{m_k \in M} c_k^{ij} m_k = p_{ij}(\theta). \tag{10.13}$$

Nótese que en (10.13) todos los monomios m_k y las probabilidades sin normalizar $p_{ij}(\theta)$ son números conocidos, y las únicas incógnitas son los coeficientes c_k^{ij}, $k = 1, \ldots, m$. Para calcular estos coeficientes, se necesita construir un conjunto de m ecuaciones linealmente independientes, cada una de la forma (10.13). Estas ecuaciones pueden obtenerse utilizando m realizaciones diferentes de Θ. Denotemos estos valores por $C = \{\theta_1, \ldots, \theta_m\}$. Sea $\mathbf{T}_{ij}$ la matriz no singular de dimensión $m \times m$ cuyo rk-ésimo elemento es el valor del monomio m_k obtenido reemplazando Θ por θ_r, la r-ésima realización de Θ. Nos referiremos a la matriz $\mathbf{T}_{ij}$ como la *matriz canónica*

asociada al conjunto de componentes canónicas C. Sea

$$\mathbf{c}_{ij} = \begin{pmatrix} c_1^{ij} \\ \vdots \\ c_m^{ij} \end{pmatrix} \quad \text{y} \quad \mathbf{p}_{ij} = \begin{pmatrix} p_{ij}(\theta_1) \\ \vdots \\ p_{ij}(\theta_m) \end{pmatrix}. \tag{10.14}$$

De (10.13) las m ecuaciones linealmente independientes pueden ser escritas como

$$\mathbf{T}_{ij}\ \mathbf{c}_{ij}\ =\ \mathbf{p}_{ij}. \tag{10.15}$$

Puesto que $\mathbf{T}_{ij}$ no es singular, existe $\mathbf{T}_{ij}^{-1}$ y la solución de (10.14) resulta

$$\mathbf{c}_{ij}\ =\ \mathbf{T}_{ij}^{-1}\ \mathbf{p}_{ij}. \tag{10.16}$$

Los valores de los coeficientes en $\mathbf{c}_{ij}$ pueden ser sustituídos en (10.12) y las probabilidades sin normalizar $p_{ij}(\Theta)$ pueden expresarse como una función de Θ.

Por tanto, las ecuaciones (10.12)−(10.16) suministran un algoritmo eficiente para la propagación simbólica, que no requiere ninguna computación simbólica. Nos referiremos a este algoritmo como *algoritmo o método de las componentes canónicas*. Este algoritmo se resume como sigue:

Algoritmo 10.1 Componentes canónicas.

- **Datos:** Una red Bayesiana (D, P), donde P está definida con parámetros numéricos y simbólicos, y la evidencia $E = e$.
- **Resultados:** Las probabilidades simbólicas $p(X_i = j|E = e)$, $i = 1, \ldots, n$.

1. Construir m conjuntos de realizaciones de Θ: $\theta_1, \ldots, \theta_m$, que den lugar a m ecuaciones linealmente independientes en $\mathbf{c}_{ij}$, tras la sustitución de los valores de θ_i en (10.13).
2. Calcular la matriz $\mathbf{T}_{ij}$ no singular de dimensión $m \times m$ cuyo rk-ésimo elemento es el valor del monomio m_k obtenido reemplazando Θ por θ_r, la r-ésima realización de Θ.
3. Calcular el vector de probabilidades $\mathbf{p}_{ij}$ en (10.14) usando un método estándar de propagación numérica.
4. Resolver el sistema lineal de ecuaciones (10.15) para obtener los coeficientes deseados, $\mathbf{c}_{ij}$.
5. Sustituir los valores obtenidos de $\mathbf{c}_{ij}$ en (10.12) y normalizar para obtener la expresión simbólica de las probabilidades $p(X_i = j|E = e)$.

∎

Nótese que la Etapa 3 del Algoritmo 10.1 requiere el uso de un método de propagación numérica para propagar la incertidumbre tantas veces como número de posibles combinaciones de los parámetros simbólicos haya. Esto significa que el número de propagaciones numéricas se incrementa exponencialmente con el número de parámetros simbólicos. Este problema también aparece en otros algoritmos de propagación. Por ejemplo, el algoritmo basado en el condicionamiento también presenta este problema con respecto al número de nodos en el conjunto de corte (véase el Capítulo 8). Por tanto, el papel de los nodos simbólicos en el método de las componentes canónicas es similar al papel de los nodos condicionantes en los algoritmos de condicionamiento. Sin embargo, allí se dan valores a las evidencias y aquí, a los parámetros.

El Algoritmo 10.1 requiere calcular y resolver un sistema lineal de ecuaciones (véase, por ejemplo, Press et al. (1992)). En lo que sigue, se muestra que imponiendo ciertas condiciones a los parámetros simbólicos, es siempre posible encontrar un conjunto de componentes canónicas cuya matriz asociada $\mathbf{T}_{ij}$ coincida con la identidad. Por ello, las expresiones simbólicas asociadas a las probabilidades pueden obtenerse directamente, sin necesidad de resolver sistemas de ecuaciones como (10.15) o invertir matrices como $\mathbf{T}_{ij}$ en (10.16).

Considérese un nodo simbólico X_i con sus parámetros asociados $\theta_{ij\pi}$. Algunos de estos parámetros pueden especificarse numéricamente, y algunos pueden darse en forma simbólica. Supóngase que un subconjunto de parámetros $\theta_{ij\pi}$ se da en forma simbólica para una cierta realización π de Π_i, y que los parámetros para todas las restantes realizaciones son numéricos. Por ejemplo, la red Bayesiana del Ejemplo 10.1 satisface esta hipótesis puesto que los parámetros simbólicos correspondientes a cada nodo simbólico están asociados a la misma realización del conjunto de padres (véase la Tabla 10.1). Por ejemplo, X_1 es el único padre del nodo simbólico X_3, y ambos parámetros simbólicos θ_{300} y θ_{310} están asociados a la misma realización, $X_1 = 0$, de X_1.

En esta situación, las componentes canónicas resultantes de considerar valores extremos para los parámetros simbólicos producen un matriz canónica $\mathbf{T}_{ij}$ igual a la matriz identidad. El teorema siguiente prueba este hecho.

Teorema 10.2 *Dado un conjunto de nodos simbólicos* $\{X_{i_1}, \ldots, X_{i_s}\}$ *con parámetros simbólicos asociados* $\Theta = \{\Theta_1, \ldots, \Theta_s\}$, *donde* $\Theta_k = \{\theta_{i_k j \pi_k}, j = 0, \ldots, r_{i_k}\}$, *y* π_k *es una realización dada de* Π_{i_k}, *entonces la matriz canónica asociada al conjunto de componentes canónicas* C *definida por el producto cartesiano* $C = C_1 \times \ldots \times C_n$, *donde*

$$C_k = \begin{pmatrix} \{\theta_{i_k 0 \pi_k} = 1, \theta_{i_k 1 \pi_k} = 0, \ldots, \theta_{i_k r_{i_k} \pi_k} = 0\} \\ \vdots \\ \{\theta_{i_k 0 \pi_k} = 0, \theta_{i_k 1 \pi_k} = 0, \ldots, \theta_{i_k r_{i_k} \pi_k} = 1\} \end{pmatrix},$$

es la matriz identidad de dimensión $m \times m$.

Demostración: De (10.11), el conjunto de monomios m_k está dado por

$$M = \Theta_1 \times \ldots \times \Theta_s.$$

En este caso, utilizando la hipótesis $\Theta_k = \{\theta_{i_k j \pi_k},\ j = 0, \ldots, r_{i_k}\}$, se tiene

$$M = \{\{\theta_{i_1 0 \pi_1}, \ldots, \theta_{i_1 r_{i_1} \pi_1}\} \times \ldots \times \{\theta_{i_s 0 \pi_s}, \ldots, \theta_{i_1 r_{i_s} \pi_s}\}\}.$$

Por tanto, cualquier realización de los parámetros simbólicos en C elimina todos los monomios de M salvo uno. Sea $\theta \in C$ una componente canónica típica. Entonces, todos los parámetros en Θ_k salvo uno son cero, $\theta_{i_k j_k \pi_k}$, para $k = 1, \ldots, m$. Por ello, sustituyendo estos valores numéricos en los monomios de M, el monomio $\theta_{i_1 j_1 \pi_1} \ldots \theta_{i_s j_s \pi_s}$ toma el valor 1, y el resto de monomios se anulan. Por tanto, toda fila de la matriz $\mathbf{T}_{ij}$ contiene un elemento con el valor 1 y todos los demás son 0. Finalmente, el proceso de la construcción de M y C garantiza que la matriz $\mathbf{T}_{ij}$ es la matriz identidad de orden $m \times m$. ∎

Del Teorema 10.2 se deduce que la solución $\mathbf{c}_{ij}$ del sistema lineal de ecuaciones (10.15) resulta

$$\mathbf{c}_{ij} \ = \ \mathbf{p}_{ij}. \tag{10.17}$$

Por tanto, en esta situación, el Algoritmo 10.1 puede simplificarse como sigue:

Algoritmo 10.2 Componentes canónicas modificado.

- **Datos:** Una red Bayesiana (D, P), donde P se define usando parámetros simbólicos y numéricos que satisfacen la condición del Teorema 10.2, y la evidencia $E = e$.
- **Resultados:** Las probabilidades $p(X_i = j|E = e)$, $i = 1, \ldots, n$ en forma simbólica.

1. Construir m conjuntos de realizaciones de Θ en la forma canónica indicada en el Teorema 10.2, $C = \{\theta_1, \ldots, \theta_m\}$.
2. Calcular el vector de probabilidades $\mathbf{p}_{ij}$ en (10.14) usando cualquier método de propagación numérica.
3. Sea $\mathbf{c}_{ij} = \mathbf{p}_{ij}$, sustituir $\mathbf{c}_{ij}$ en (10.12), y normalizar para obtener la expresión simbólica de las probabilidades $p(X_i = j|E = e)$. ∎

Nótese que con la hipótesis del Teorema 10.2, las componentes canónicas nos permiten realizar la propagación simbólica de una forma fácil y eficiente. Si no se satisface la hipótesis del Teorema 10.2, entonces la matriz canónica resultante asociada al conjunto de parámetros C puede ser

diferente de la matriz identidad, y será necesario resolver el sistema de ecuaciones (10.15) para obtener los coeficientes de los polinomios.

Seguidamente se ilustra este algoritmo mediante un ejemplo.

Ejemplo 10.3 Componentes canónicas. Considérese la red de la Figura 10.1 y la evidencia $e = \{X_2 = 1, X_5 = 1\}$. Se desea analizar la influencia de los parámetros simbólicos en las probabilidades condicionales de los restantes nodos. En este ejemplo el conjunto de nodos simbólicos es $\{X_3, X_6\}$ y el conjunto de parámetros es $\Theta = \{\Theta_3, \Theta_6\} = \{\{\theta_{300}, \theta_{310}\}, \{\theta_{600}, \theta_{610}\}\}$ (véase Tabla 10.1). Entonces, el conjunto de posibles monomios resulta

$$\begin{aligned} M &= \Theta_3 \times \Theta_6 \\ &= \{\theta_{300}\theta_{600}, \theta_{300}\theta_{610}, \theta_{310}\theta_{600}, \theta_{310}\theta_{610}\} \\ &= \{m_1, m_2, m_3, m_4\}. \end{aligned}$$

De (10.12), la probabilidad condicional sin normalizar $p(X_i = j|e)$ es una función polinomial de la forma

$$p(X_i = j|e) \propto \sum_{k=1}^{4} c_k^{ij}\ m_k = p_{ij}(\Theta). \tag{10.18}$$

Por ello, nuestro objetivo es obtener los coeficientes $\{c_k^{ij}\ ;\ k = 1, \ldots, 4\}$ para cada nodo X_i y cada posible valor j. Con este objetivo, se consideran las componentes canónicas asociadas al conjunto Θ de parámetros simbólicos. En este caso, dado que se está tratando con variables binarias, y que se cumplen las condiciones del Teorema 10.2, hay sólo dos posibles combinaciones canónicas de los parámetros en Θ_i, $\{1, 0\}$ y $\{0, 1\}$. Por ello, se tiene el conjunto siguiente de componentes canónicas:

$$\begin{aligned} C &= \{\{1, 0\}, \{0, 1\}\} \times \{\{1, 0\}, \{0, 1\}\} \\ &= \{\{1, 0; 1, 0\}, \{1, 0; 0, 1\}, \{0, 1; 1, 0\}, \{0, 1; 0, 1\}\} \\ &= \{c_1, c_2, c_3, c_4\}. \end{aligned}$$

Entonces, dando a los parámetros simbólicos los valores que les corresponden según sus componentes canónicas, todos los monomios que aparecen en (10.18) toman valores 0 ó 1. De esta forma, se obtiene una expresión que depende sólo de los coeficientes c^{ij}:

$$\begin{aligned} p_{ij}(\Theta = c_1) &= c_1^{ij}, \qquad p_{ij}(\Theta = c_2) = c_2^{ij}, \\ p_{ij}(\Theta = c_3) &= c_3^{ij}, \qquad p_{ij}(\Theta = c_4) = c_4^{ij}. \end{aligned}$$

Por ello, en este caso, la matriz $\mathbf{T}_{ij}$ es la matriz identidad puesto que todos los parámetros simbólicos de los nodos simbólicos están asociados a

la misma realización del conjunto de padres. Entonces, se tiene

$$\begin{pmatrix} c_1^{ij} \\ c_2^{ij} \\ c_3^{ij} \\ c_4^{ij} \end{pmatrix} = \begin{pmatrix} p_{ij}(c_1) \\ p_{ij}(c_2) \\ p_{ij}(c_3) \\ p_{ij}(c_4) \end{pmatrix}. \tag{10.19}$$

Es interesante hacer notar aquí que el conjunto factible (el conjunto generado por todos los valores de los parámetros) para las probabilidades de cualquier conjunto de nodos es la envolvente convexa generada por las probabilidades canónicas.

En la Figura 10.6 se muestran las probabilidades condicionales sin normalizar $p_{ij}(c_k)$ de todos los nodos, dada la evidencia $e = \{X_2 = 1, X_5 = 1\}$, asociadas a las cuatro posibles componentes canónicas. Usando estos valores se pueden obtener todas las funciones racionales de la Tabla 10.3. Por ejemplo, de la Figura 10.6 se obtienen los siguientes valores para el nodo X_6:

$$\begin{pmatrix} c_1^{60} \\ c_2^{60} \\ c_3^{60} \\ c_4^{60} \end{pmatrix} = \begin{pmatrix} 0.390 \\ 0.247 \\ 0.418 \\ 0.254 \end{pmatrix} \tag{10.20}$$

y

$$\begin{pmatrix} c_1^{61} \\ c_2^{61} \\ c_3^{61} \\ c_4^{61} \end{pmatrix} = \begin{pmatrix} 0.028 \\ 0.171 \\ 0.029 \\ 0.193 \end{pmatrix}, \tag{10.21}$$

es decir, los coeficientes de los polinomios del numerador para $X_6 = 0$ y $X_6 = 1$, respectivamente. Entonces, sustituyendo estos valores en (10.18) se obtiene

$$\begin{aligned} p(X_6 = 0|e) \quad &\propto \quad 0.390\,\theta_{300}\theta_{600} + 0.247\,\theta_{300}\theta_{610} \\ &\quad + 0.418\,\theta_{310}\theta_{600} + 0.254\,\theta_{310}\theta_{610}, \end{aligned}$$

$$\begin{aligned} p(X_6 = 1|e) \quad &\propto \quad 0.028\,\theta_{300}\theta_{600} + 0.171\,\theta_{300}\theta_{610} \\ &\quad + 0.029\,\theta_{310}\theta_{600} + 0.193\,\theta_{310}\theta_{610}. \end{aligned}$$

Sumando ambos polinomios, se obtiene el polinomio del denominador que representa la constante de normalización, es decir,

$$\begin{pmatrix} c_1^{60} \\ c_2^{60} \\ c_3^{60} \\ c_4^{60} \end{pmatrix} + \begin{pmatrix} c_1^{61} \\ c_2^{61} \\ c_3^{61} \\ c_4^{61} \end{pmatrix} = \begin{pmatrix} 0.418 \\ 0.418 \\ 0.447 \\ 0.447 \end{pmatrix}.$$

Por ello, se tiene

$$\begin{aligned} p(X_6 = 0|e) &= (0.390\theta_{300}\theta_{600} + 0.247\theta_{300}\theta_{610} \\ &\quad +0.418\theta_{310}\theta_{600} + 0.254\theta_{310}\theta_{610})/d, \\ p(X_6 = 1|e) &= (0.028\theta_{300}\theta_{600} + 0.171\theta_{300}\theta_{610} \\ &\quad +0.029\theta_{310}\theta_{600} + 0.193\theta_{310}\theta_{610})/d, \end{aligned}$$

donde $d = 0.418\,\theta_{300}\theta_{600} + 0.418\,\theta_{300}\theta_{610} + 0.447\,\theta_{310}\theta_{600} + 0.447\,\theta_{310}\theta_{610}$.

Finalmente, eliminando los parámetros dependientes, θ_{310} y θ_{610}, se obtiene la expresión de la Tabla 10.3. Nótese que la única operación simbólica de este proceso consiste en simplificar la expresión final para eliminar los parámetros dependientes. Sin embargo, ésta es una operación opcional, y en algunos casos, es más conveniente mantener la expresión con todos los parámetros y simplificar los resultados numéricos tras sustituir valores concretos en los parámetros. ■

Debe mencionarse aquí que aunque se están utilizando sólo métodos de propagación exacta para calcular $\mathbf{p}_{ij}$, la metodología sigue siendo válida si se utilizan los métodos aproximados descritos en el Capítulo 9. En este caso, se obtienen soluciones simbólicas aproximadas para las probabilidades.

10.5.3 Computación Eficiente de las Componentes Canónicas

El método de propagación simbólica propuesto anteriormente requiere varias aplicaciones de un método numérico, aproximado o exacto, para calcular numéricamente las probabilidades $\mathbf{p}_{ij}(c_k)$ asociadas a cada una de las componentes canónicas C_k. Por tanto, el número de propagaciones aumenta exponencialmente con el número de parámetros simbólicos. Sin embargo, se pueden ahorrar muchos cálculos cuando se propaga la incertidumbre en los casos canónicos, puesto que algunos de los mensajes (véase el Capítulo 8) son comunes a varias componentes canónicas. En esta sección se ilustra este hecho en dos de los algoritmos que utilizan este paso de mensajes: el algoritmo para poliárboles (véase la Sección 8.3) y el algoritmo de agrupamiento (Sección 8.6).

La Figura 10.7 ilustra el proceso de paso de mensajes correspondiente a un nodo típico X_i en un poliárbol cuando se aplica el Algoritmo 8.1. En esta figura, Θ^k representa el conjunto de parámetros contenidos en la componente conexa asociada al nodo X_k cuando se elimina la arista $X_k - X_i$, mientras que Θ_k es el conjunto de parámetros asociados al nodo X_k. Nótese que el mensaje que va del nodo X_k al nodo X_i depende sólo de esos parámetros, mientras que el mensaje que va de X_i a X_k depende de los restantes parámetros de la red Bayesiana. Nótese también que si Θ^k no contiene ningún parámetro simbólico, entonces todos los mensajes que van de esta región del grafo a X_i necesitan ser calculados una sóla vez, puesto que tienen el mismo valor para todas las componentes canónicas.

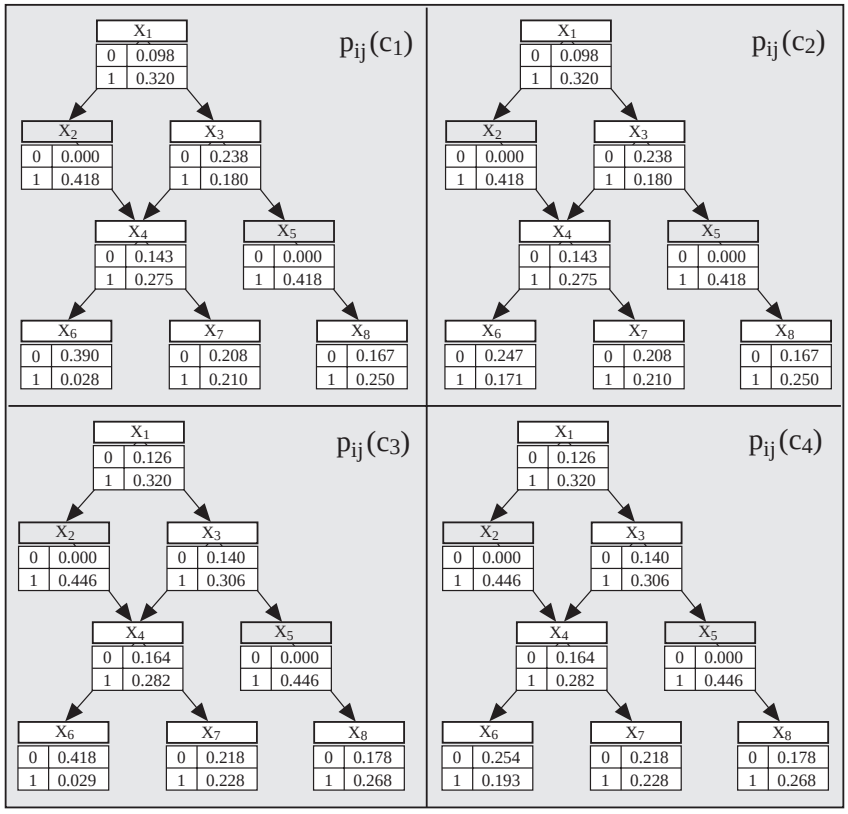


FIGURA 10.6. Los cuatro casos canónicos elementales. La primera columna de la tabla de cada nodo X_i representa el estado de X_i y la segunda columna es la probabilidad condicionada sin normalizar de X_i dada la evidencia $e = \{X_2 = 1, X_5 = 1\}$. Los rectángulos sombreados indican los nodos evidenciales.

Los mismos ahorros computacionales, en el envío de mensajes, pueden obtenerse cuando se aplican los algoritmos de agrupamiento. En este caso, la situación es la misma, excepto que ahora se trata de un árbol de aglomerados (conjuntos de nodos) en vez de un árbol de nodos. Por ejemplo, considérese el grafo múltiplemente conexo de la Figura 10.1 con las probabilidades numéricas y simbólicas de la Tabla 10.1. La Figura 10.8 muestra todos los mensajes necesarios para propagar evidencia en una familia de árboles asociados a la red Bayesiana de la Figura 10.1, usando el método de agrupamiento. En la Figura 10.8, se indican con flechas los mensajes. Se puede distinguir entre dos tipos diferentes de mensajes:

1. Mensajes sin índice. Estos mensajes son comunes para todas las componentes canónicas. Por tanto, sólo hay que calcularlos una vez.

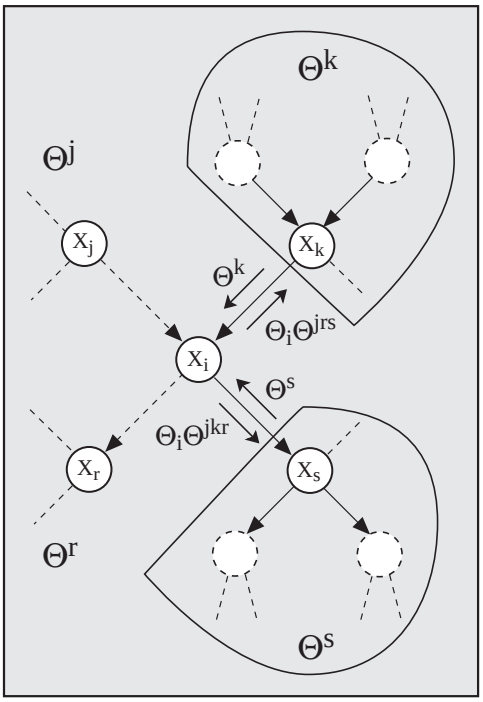


FIGURA 10.7. Dependencia paramétrica de los mensajes en un nodo general de un poliárbol.

2. Mensajes con uno o más índices tales como Θ_3, Θ_6, ó $\Theta_3\Theta_6$. Estos mensajes dependen de los parámetros, y por tanto deben calcularse tantos mensajes como parámetros tengan asociados.

Por ello, en este ejemplo se puede construir la función racional asociada a cada nodo haciendo sólo la mitad de los cálculos.

Una mejora adicional del método de las componentes canónicas y que reduce el número de monomios identificando aquellos cuyos coeficientes son diferentes de cero (consistentes con la evidencia) se ha sugerido por Castillo, Gutiérrez, y Hadi (1996a). Este método se ilustra en la sección siguiente considerando el caso de propagación orientada a un objetivo.

10.6 Propagación Simbólica Orientada a un Objetivo

En la Sección 8.8 se introdujo un algoritmo para calcular el conjunto de nodos relevantes necesarios para una cierta tarea. En el caso de la propagación simbólica, esas reducciones son de una importancia muy significativa, ya

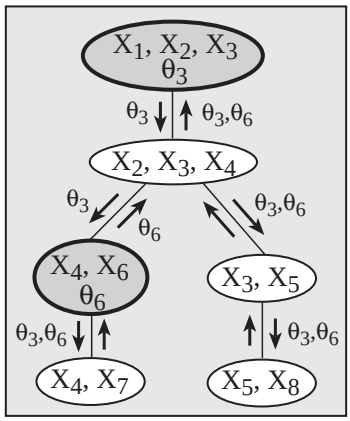


FIGURA 10.8. Árbol de unión y mensajes afectados por los parámetros relevantes. Los conglomerados cuyas funciones potenciales incluyen parámetros simbólicos Θ_3 ó Θ_6 se muestran con borde grueso.

que la complejidad de la propagación simbólica viene dada por el número total de parámetros simbólicos.

Supóngase que se está interesado en un determinado nodo X_i (nodo objetivo), y que se quiere obtener la función de probabilidad condicional $p(X_i = j|E = e)$, donde E es un conjunto de nodos evidenciales que toman los valores $E = e$. En esta sección se muestran las etapas a seguir para la propagación simbólica orientada a un objetivo. El algoritmo consta de cuatro partes principales (véase Castillo, Gutiérrez y Hadi (1996a)):

- **Parte I: Identificar todos los nodos relevantes.**
  Como se ha visto en la Sección 8.8, la función de probabilidad condicionada $p(X_i = j|E = e)$ no depende necesariamente de los parámetros asociados a todos los nodos. Por ello, se puede identificar el conjunto de nodos que son relevantes para el cálculo de $p(X_i = j|E = e)$ usando el Algoritmo 8.6. Una vez que se ha identificado el conjunto de nodos relevantes, se identifica el correspondiente conjunto de parámetros relevantes Θ y los restantes nodos se eliminan del grafo. Los cálculos pueden hacerse ahora en el grafo reducido considerando sólo el conjunto de nodos relevantes sin que cambien los resultados.

- **Parte II: Identificar los parámetros suficientes.**
  El conjunto de parámetros Θ puede reducirse aún más mediante la identificación y eliminación del conjunto de parámetros que están en contradicción con la evidencia. Estos parámetros se eliminan aplicando las dos reglas siguientes:

  - **Regla 1:** Eliminar los parámetros $\theta_{jk\pi}$ si $e_j \neq k$ para todo $X_j \in E$.

  - **Regla 2:** Eliminar los parámetros $\theta_{jk\pi}$ si los valores de los padres π son incompatibles con la evidencia.

- **Parte III: Identificar los monomios factibles.**
  Una vez que los subconjuntos mínimos de parámetros suficientes han sido identificados, los parámetros se combinan en monomios tomando el producto cartesiano de los subconjuntos suficientes minimales de parámetros y se eliminan los conjuntos de todas las combinaciones no factibles de los parámetros aplicando la regla siguiente:

  - **Regla 3:** Los parámetros asociados a realizaciones condicionantes contradictorias no pueden aparecer en el mismo monomio.

- **Parte IV: Calcular los coeficientes de todos los monomios.**
  En esta parte se calculan los coeficientes aplicando el método de las componentes canónicas descrito en la Sección 10.5.2.

Para ilustrar este algoritmo se usa la red Bayesiana introducida en el Ejemplo 10.1.

Ejemplo 10.4 Considérese de nuevo la red de la Figura 10.1. Por razones ilustrativas, la Tabla 10.4 da los valores numéricos y simbólicos de los parámetros asociados a la función de probabilidad condicional en (10.4). Por ello, los nodos simbólicos son X_1, X_3, X_5, X_6, X_7 y X_8. Supóngase que el nodo objetivo es X_7 y que la evidencia es $e = \{X_1 = 1\}$. Entonces, se desea calcular las probabilidades condicionales $p(x_7|X_1 = 1)$. Se procede como sigue:

Parte I:

- **Etapa 1:** Siguiendo las etapas del Algoritmo 8.6, se necesita añadir al grafo inicial de la Figura 10.1 los nodos auxiliares $V_1, \ldots, V_8$. Esto conduce al nuevo grafo de la Figura 10.9.

- **Etapa 2:** El conjunto V de nodos auxiliares D-separados de X_7 (el nodo objetivo) por X_1 (el nodo evidencial) resulta ser $V = \{V_5, V_6, V_8\}$. El nuevo grafo resultante de eliminar los correspondientes nodos X_5, X_6 y X_8 se muestra en la Figura 10.10. Por ello, el conjunto de todos los parámetros asociados a los nodos simbólicos auxiliares que no están incluídos en V es

$$\Theta = \{\{\theta_{300}, \theta_{301}, \theta_{310}, \theta_{311}\}; \{\theta_{700}, \theta_{701}, \theta_{710}, \theta_{711}\}\}.$$

  Este es el conjunto de parámetros relevantes. Nótese que los parámetros del nodo X_1 (θ_{10} y θ_{11}) no están incluídos en Θ puesto que X_1 es un nodo evidencial. Nótese también que en esta etapa se ha reducido el número de parámetros simbólicos de 22 a 8 (o el número de parámetros libres de 11 a 4).

X_i	Π_i	Parámetros libres
X_1	ϕ	$\theta_{10} = p(X_1 = 0)$
X_2	$\{X_1\}$	$\theta_{200} = p(X_2 = 0\|X_1 = 0) = 0.2$ $\theta_{201} = p(X_2 = 0\|X_1 = 1) = 0.5$
X_3	$\{X_1\}$	$\theta_{300} = p(X_3 = 0\|X_1 = 0)$ $\theta_{301} = p(X_3 = 0\|X_1 = 1)$
X_4	$\{X_2, X_3\}$	$\theta_{4000} = p(X_4 = 0\|X_2 = 0, X_3 = 0) = 0.1$ $\theta_{4001} = p(X_4 = 0\|X_2 = 0, X_3 = 1) = 0.2$ $\theta_{4010} = p(X_4 = 0\|X_2 = 1, X_3 = 0) = 0.3$ $\theta_{4011} = p(X_4 = 0\|X_2 = 1, X_3 = 1) = 0.4$
X_5	$\{X_3\}$	$\theta_{500} = p(X_5 = 0\|X_3 = 0)$ $\theta_{501} = p(X_5 = 0\|X_3 = 1)$
X_6	$\{X_4\}$	$\theta_{600} = p(X_6 = 0\|X_4 = 0)$ $\theta_{601} = p(X_6 = 0\|X_4 = 1)$
X_7	$\{X_4\}$	$\theta_{700} = p(X_7 = 0\|X_4 = 0)$ $\theta_{701} = p(X_7 = 0\|X_4 = 1)$
X_8	$\{X_5\}$	$\theta_{800} = p(X_8 = 0\|X_5 = 0)$ $\theta_{801} = p(X_8 = 0\|X_5 = 1)$

TABLA 10.4. Un conjunto de parámetros libres asociados a la función de probabilidad condicionada en (10.4).

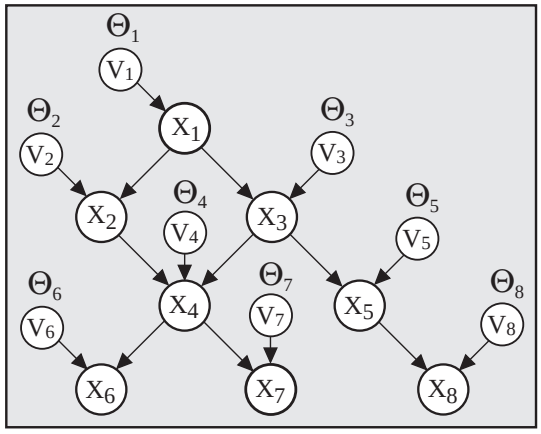


FIGURA 10.9. El grafo aumentado obtenido añadiendo un nodo auxiliar V_i a todo nodo X_i.

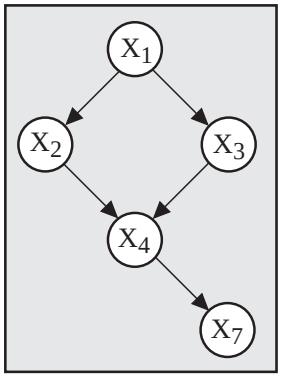


FIGURA 10.10. El grafo que contiene el conjunto de nodos cuyos nodos auxiliares no están D-separados de X_7 por X_1.

M_0	M_1
$\theta_{301}\theta_{700}$	$\theta_{301}\theta_{710}$
$\theta_{301}\theta_{701}$	$\theta_{301}\theta_{711}$
$\theta_{311}\theta_{700}$	$\theta_{311}\theta_{710}$
$\theta_{311}\theta_{701}$	$\theta_{311}\theta_{711}$

TABLA 10.5. Monomios necesarios para determinar las probabilidades que se indican.

Parte II:

- **Etapa 3:** El conjunto Θ no contiene parámetros asociados al nodo evidencial X_1. Por ello, no es posible una reducción aplicando la Regla 1.
- **Etapa 4:** Puesto que θ_{300} y θ_{310} son inconsistentes con la evidencia (puesto que indican que $X_1 = 0$), estos parámetros se pueden eliminar de Θ, obteniendo el mínimo conjunto de parámetros suficientes:

$$\Theta = \{\{\theta_{301}, \theta_{311}\}; \{\theta_{700}, \theta_{701}, \theta_{710}, \theta_{711}\}\}.$$

Parte III:

- **Etapa 5:** El conjunto inicial de monomios candidatos consiste en el producto casrtesiano de los subconjuntos minimales suficientes, es decir, $M = \{\theta_{301}, \theta_{311}\} \times \{\theta_{700}, \theta_{701}, \theta_{710}, \theta_{711}\}$. Por ello, los monomios candidatos se muestran en la Tabla 10.5.
- **Etapa 6:** Los padres de los nodos X_3 y X_7 no tienen elementos comunes; por tanto todos los monomios que se muestran en la Tabla 10.5 son monomios factibles.

Parte IV:

- **Etapa 7:** Los conjuntos de monomios M_0 y M_1 necesarios para calcular $p(X_7 = 0|X_1 = 1)$ y $p(X_7 = 1|X_1 = 1)$, respectivamente, se muestran en la Tabla 10.5.

- **Etapa 8:** Para $j = 0$ se tiene la ecuación polinomial siguiente:

$$\begin{aligned} p_0(\Theta) &= c_{01}m_{01} + c_{02}m_{02} + c_{03}m_{03} + c_{04}m_{04} \\ &= c_{01}\theta_{301}\theta_{700} + c_{02}\theta_{301}\theta_{701} \\ &\quad + c_{03}\theta_{311}\theta_{700} + c_{04}\theta_{311}\theta_{701}. \end{aligned} \tag{10.22}$$

  Por ello, eligiendo las componentes canónicas

$$\begin{aligned} \theta_1 = (1,0,1,0,1,0), \quad \theta_2 = (1,0,0,1,1,0), \\ \theta_3 = (0,1,1,0,1,0), \quad \theta_4 = (0,1,0,1,1,0) \end{aligned}$$

  para el conjunto $\Theta = \{\theta_{301}, \theta_{311}, \theta_{700}, \theta_{701}, \theta_{710}, \theta_{711}\}$ de parámetros suficientes, se obtiene el sistema de ecuaciones siguiente:

$$c_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} p_0(\theta_1) \\ p_0(\theta_2) \\ p_0(\theta_3) \\ p_0(\theta_4) \end{pmatrix} = \begin{pmatrix} 0.15 \\ 0.85 \\ 0.35 \\ 0.65 \end{pmatrix}. \tag{10.23}$$

  Similarmente, para $j = 1$ se obtiene

$$c_1 = \begin{pmatrix} p_1(\theta_1) \\ p_1(\theta_2) \\ p_1(\theta_3) \\ p_1(\theta_4) \end{pmatrix} = \begin{pmatrix} 0.15 \\ 0.85 \\ 0.35 \\ 0.65 \end{pmatrix}. \tag{10.24}$$

  La Tabla 10.6 muestra los resultados de calcular las probabilidades numéricas necesarias para obtener las expresiones anteriores. Combinando (10.22) y (10.23) se obtienen las expresiones polinomiales finales

$$\begin{aligned} p(X_7 = 0|X_1 = 1) \quad \propto \quad & 0.15\theta_{301}\theta_{700} + 0.85\theta_{301}\theta_{701} \\ & + 0.35\theta_{311}\theta_{700} + 0.65\theta_{311}\theta_{701}. \end{aligned} \tag{10.25}$$

  Similarmente, para $X_7 = 1$ se obtiene

$$\begin{aligned} p(X_7 = 1|X_1 = 1) \quad \propto \quad & 0.15\theta_{301}\theta_{710} + 0.85\theta_{301}\theta_{711} \\ & + 0.35\theta_{311}\theta_{710} + 0.65\theta_{311}\theta_{711}. \end{aligned} \tag{10.26}$$

- **Etapa 9:** Sumando las probabilidades sin normalizar en (10.25) y (10.26) se obtiene la constante de normalización. Puesto que $\theta_{i0\pi} + \theta_{i1\pi} = 1$, para todo i, la constante de normalización resulta ser 1.

$X_7 = 0$			
$(\theta_{301}, \theta_{311}, \theta_{700}, \theta_{701}, \theta_{710}, \theta_{711})$	$p_0(\theta)$	Monomios	Coeficientes
(1,0,1,0,1,0)	0.15	$\theta_{301}\theta_{700}$	$c_{01} = 0.15$
(1,0,0,1,1,0)	0.85	$\theta_{301}\theta_{701}$	$c_{02} = 0.85$
(0,1,1,0,1,0)	0.35	$\theta_{311}\theta_{700}$	$c_{03} = 0.35$
(0,1,0,1,1,0)	0.65	$\theta_{311}\theta_{701}$	$c_{04} = 0.65$
$X_7 = 1$			
$(\theta_{301}, \theta_{311}, \theta_{700}, \theta_{701}, \theta_{710}, \theta_{711})$	$p_1(\theta)$	Monomios	Coeficientes
(1,0,1,0,1,0)	0.15	$\theta_{301}\theta_{710}$	$c_{11} = 0.15$
(1,0,0,1,1,0)	0.85	$\theta_{301}\theta_{711}$	$c_{12} = 0.85$
(0,1,1,0,1,0)	0.35	$\theta_{301}\theta_{710}$	$c_{13} = 0.35$
(0,1,0,1,1,0)	0.65	$\theta_{311}\theta_{711}$	$c_{14} = 0.65$

TABLA 10.6. Coeficientes de los monomios y sus correspondientes valores $p_j(\theta)$.

- **Etapa 10:** Las expresiones en (10.25) y (10.26) pueden simplificarse reemplazando θ_{311}, θ_{710} y θ_{711} por $(1-\theta_{301})$, $(1-\theta_{700})$, y $(1-\theta_{701})$, respectivamente. Finalmente, se obtiene

$$\begin{aligned} p(X_7 = 0|X_1 = 1) &\propto 0.15\theta_{301}\theta_{700} + 0.85\theta_{301}\theta_{701} \\ &\quad + (1-\theta_{301})(0.35\theta_{700} + 0.65\theta_{701}) \\ &= 0.35\theta_{700} - 0.2\theta_{301}\theta_{700} \\ &\quad + 0.65\theta_{701} + 0.2\theta_{301}\theta_{701} \end{aligned}$$

y

$$\begin{aligned} p(X_7 = 1|X_1 = 1) &\propto 0.15\theta_{301}(1-\theta_{700}) + 0.85\theta_{301}(1-\theta_{701}) \\ &\quad + 0.35(1-\theta_{301})(1-\theta_{700}) \\ &\quad + 0.65(1-\theta_{301})(1-\theta_{701}) \\ &= 1 + 0.2\theta_{301}\theta_{700} - 0.35\theta_{700} \\ &\quad - 0.65\theta_{701} - 0.2\theta_{301}\theta_{701}. \end{aligned}$$

■

10.7 Tratamiento Simbólico de la Evidencia Aleatoria

En las secciones amteriores, se ha supuesto que la evidencia disponible es determinista, es decir, se sabe que el conjunto evidencial E toma los valores

e. En algunas situaciones la evidencia disponible puede ser estocástica, es decir, puede incorporar algún grado de incertidumbre. Por ejemplo, se puede decir que $E = e$ con probabilidad $q(e)$, donde $\sum_e q(e) = 1$. Por ello, cuando $q(e) = 1$, se tiene una evidencia determinista. En esta sección se trata de la evidencia estocástica.

Supóngase que se tiene una evidencia estocástica $E = e$ con probabilidad $q(e)$. Se desea calcular la probabilidad condicionada $p(x_i|(e\,con\,prob\,q(e)))$, para todos los nodos no evidenciales X_i. Esta probabilidad condicionada es

$$p(x_i|(E = e\,con\,prob\,q(e))) = \sum_e q(e)p(x_i|e). \tag{10.27}$$

Por ello, la probabilidad condicionada en (10.27) puede calcularse aplicando el método de las componentes canónicas una vez para cada combinación de valores de las variables evidenciales.

Es importante hacer notar que $p(x_i|(e\,con\,prob\,q(e)))$ es también una función racional de dos polinomios puesto que se trata de una combinación lineal convexa de $p(x_i|e)$, que es una función racional (véase el Corolario 10.2). Sin embargo, en este caso los parámetros pueden aparecer con exponentes mayores que la unidad, lo que implica polinomios de grado mayor que uno. Este resultado se muestra en el teorema que sigue (Castillo, Gutiérrez y Hadi (1995c)).

Teorema 10.3 *La probabilidad condicionada de cualquier nodo no evidencial dada una evidencia estocástica es el cociente de dos funciones polinomiales, en las que el grado de los polinomios es como mucho igual a la suma de las cardinalidades de los nodos evidenciales.*

Demostración: Los polinomios del denominador de la función racional son en general diferentes para diferentes combinaciones del conjunto de evidencias e. Por ello, el denominador común es el producto de diferentes polinomios. El número de estos productos, y por tanto el grado del polinomio, no puede exceder de la suma de las cardinalidades de los nodos evidenciales. ∎

Ejemplo 10.5 Considérese la red Bayesiana del Ejemplo 10.1 y supóngase que se conoce una evidencia determinista $X_2 = 1$ y una evidencia estocástica $X_5 = x_5$ con probabilidad $q(x_5)$, donde $q(0) = p$ y $q(1) = 1 - p$. Con esta información, se desea calcular $p(X_6 = 0|(X_2 = 1, X_5 = x_5\,con\,prob\,q(x_5)))$. En este caso, se usa el método de las componentes canónicas dos veces (una vez con la evidencia $(X_2 = 1, X_5 = 0)$ y una vez con la evidencia $(X_2 = 1, X_5 = 1)$) y se obtiene

$$\begin{array}{rcl} p(X_6 = 0|X_2 = 1, X_5 = 0) & = & s_0, \\ p(X_6 = 0|X_2 = 1, X_5 = 1) & = & s_1, \end{array} \tag{10.28}$$

donde

$$\begin{aligned} s_0 &= \frac{0.056 + 0.019\,\theta_{300} + 0.032\,\theta_{600} + 0.007\,\theta_{300}\,\theta_{600}}{0.094 + 0.028\,\theta_{300}}, \\ s_1 &= \frac{0.253 - 0.006\,\theta_{300} + 0.164\,\theta_{600} - 0.021\,\theta_{300}\,\theta_{600}}{0.446 - 0.028\,\theta_{300}}. \end{aligned}$$

Sustituyendo s_0, s_1 y $q(x_5)$ en (10.27), se obtiene

$$p(x_i|(X_2 = 1, X_5 = 0\,con\,q(x_5))) = ps_0 + (1-p)s_1 = a/b,$$

donde

$$\begin{aligned} a &= -30.334 - 1.522\,p - 8.316\,\theta_{300} - 0.492\,p\,\theta_{300} + 0.214\,\theta_{300}^2 \\ &\quad + 0.464\,p\,\theta_{300}^2 - 19.663\,\theta_{600} + 1.459\,p\,\theta_{600} - 3.339\,\theta_{300}\,\theta_{600} \\ &\quad + 0.5\,p\,\theta_{300}\,\theta_{600} + 0.75\,\theta_{300}^2\,\theta_{600} - 0.5\,p\,\theta_{300}^2\,\theta_{600}, \\ b &= -53.474 - 12.571\,\theta_{300} + \theta_{300}^2. \end{aligned}$$

Por ello, la probabilidad del suceso $X_6 = 0$ en este caso es la combinación lineal convexa de s_0 y s_1 que es una función racional en la que intervienen dos polinomios a y b. Nótese que los polinomios son de segundo grado en θ_{300}, como cabe esperar por el Teorema 10.3. ∎

Nótese también que en los casos en que las combinaciones lineales convexas son expresiones complicadas, pueden mantenerse éstas en forma expandida, y el proceso de simplificación puede realizarse una vez que hayan sido sustituídos los valores numéricos concretos.

10.8 Análisis de Sensibilidad

Por análisis de sensibilidad se entiende el estudio de los efectos que producen los cambios de los valores de alguno de los parámetros en las probabilidades condicionales de los nodos no evidenciales dada la evidencia. Una forma de realizar el análisis de sensibilidad consiste en cambiar los valores del parámetro y a continuación determinar los efectos de estos cambios en las probabilidades condicionales sin más que repetir los cálculos usando un método de propagación apropiado. Claramente, este método de resolver el problema, por la fuerza bruta, es computacionalmente costoso.

Otras formas de realizar este análisis de sensibilidad pueden encontrarse en la literatura. Por ejemplo, Breese y Fertig (1991) y Tessem (1992) propagan intervalos de probabilidades en vez de valores concretos, y Laskey (1995) mide el impacto de pequeños cambios en un parámetro sobre la probabilidad de interés usando las derivadas parciales de $p(x_i|e)$ con respecto a los parámetros.

En esta sección se muestra que el análisis de sensibilidad puede hacerse también usando métodos de propagación simbólica con muy poco esfuerzo computacional adicional (véase Castillo, Gutiérrez y Hadi (1996d)). Se muestra que el método de las componentes canónicas permite obtener las cotas inferiores y superiores de las probabilidades, obtenidas de los resultados de la propagación simbólica. Estas cotas suministran información muy válida para realizar el análisis de sensibilidad de una red Bayesiana.

Las expresiones simbólicas de las probabilidades obtenidas por el Algoritmo 10.1 pueden ser usadas para obtener cotas inferiores y superiores de las probabilidades marginales. Para calcular estos límites se necesita previamente el siguiente resultado, debido a Martos (1964):

Teorema 10.4 Cotas de las probabilidades. *Si la función fraccional lineal del vector* $\mathbf{u}$*,*

$$f(\mathbf{u}) = \frac{\mathbf{c}\,\mathbf{u} - c_0}{\mathbf{d}\,\mathbf{u} - d_0}, \tag{10.29}$$

donde $\mathbf{c}$ *y* $\mathbf{d}$ *son coeficientes vectoriales y* c_0 *y* d_0 *son constantes reales, está definida en el poliedro convexo* $A\mathbf{u} \leq \mathbf{a_0}, \mathbf{u} \geq 0$*, donde* A *es una matriz constante y* $\mathbf{a_0}$ *es un vector constante y el denominador en (10.29) no se anula en el poliedro anterior, entonces el máximo de* $f(\mathbf{u})$ *tiene lugar en uno de los vértices del poliedro.*

En nuestro caso, $\mathbf{u}$ es el conjunto de parámetros simbólicos y $f(\mathbf{u})$ es el conjunto de expresiones simbólicas asociadas a las probabilidades $p(x_i|e)$. En este caso, el poliedro convexo se define por $\mathbf{u} \leq 1, \mathbf{u} \geq 0$, es decir, A es la matriz identidad. Entonces, usando el Teorema 10.4, la cotas inferiores y superiores de las expresiones simbólicas asociadas a las probabilidades se alcanzan en los vértices de este poliedro, es decir, para alguna de las componentes canónicas asociadas al conjunto de parámetros simbólicos.

Por ello, las cotas inferiores y superiores del cociente de los polinomios en las probabilidades $p(x_i|e)$ corresponden al mínimo y máximo, respectivamente, de los valores numéricos obtenidos para estas probabilidades en todas las componentes canónicas asociadas a los parámetros contenidos en Θ, es decir, para todos los posibles combinaciones de valores extremos de los parámetros (los vértices del conjunto de parámetros).

Ejemplo 10.6 Cotas de probabilidad. Calculemos las cotas inferiores y superiores asociadas a todas las probabilidades de las variables de la red Bayesiana del Ejemplo 10.4, en primer lugar para el caso en el que no se dispone de evidencia, y luego para el caso de evidencia $X_1 = 0$. Se supone que el experto humano da, como información inicial, cotas inferiores y superiores de todos los parámetros, tal como se muestra en la Tabla 10.7.

Con objeto de realizar una comparación, se calculan las probabilidades para dos modelos:

- El modelo especificado por los once parámetros simbólicos libres de la Tabla 10.4.

Parámetro	Inferior	Superior
θ_{10}	0.7	0.9
θ_{300}	0.3	0.5
θ_{301}	0.1	0.4
θ_{500}	0.0	0.2
θ_{501}	0.8	1.0
θ_{600}	0.3	0.7
θ_{601}	0.4	0.5
θ_{700}	0.2	0.3
θ_{701}	0.5	0.7
θ_{800}	0.0	0.3
θ_{801}	0.4	0.6

TABLA 10.7. Cotas inferiores y superiores de los parámetros suministradas por el experto humano.

- El modelo reducido obtenido del modelo anterior reemplazando los parámetros de las variables X_3 y X_6 por valores numéricos fijos, es decir,

$$\theta_{300} = 0.4, \quad \theta_{301} = 0.2, \quad \theta_{600} = 0.5, \quad \theta_{601} = 0.4.$$

  Este modelo tiene siete parámetros simbólicos libres.

Para cada uno de los casos anteriores, se calculan las cotas para el caso sin evidencia y el caso con evidencia $X_1 = 0$. Las Tablas 10.8 y 10.9 muestran las cotas inferiores y superiores para los cuatro casos diferentes. A partir de su análisis, pueden hacerse los siguientes comentarios:

1. Comparando los modelos de siete y once parámetros, se ve que el rango (la diferencia entre las cotas superiores e inferiores) es no decreciente en el número de parámetros simbólicos. Por ejemplo, los rangos para el caso de siete parámetros no son mayores que los del caso de once parámetros (en realidad, con la excepción de X_1 y X_2, los rangos del caso de siete parámetros son menores que los correspondientes rangos en el caso de once parámetros). Esto es normal, ya que un número menor de parámetros simbólicos significa menor incertidumbre.

2. Comparando el caso sin evidencia (Tabla 10.8) con el de evidencia (Tabla 10.9), se ve que los rangos del primero son generalmente mayores que los del segundo. De nuevo, éste es un resultado esperado, pues más evidencia implica menos incertidumbre. ■

	Caso 1: Once Parámetros			Caso 2: Siete Parámetros		
Nodo	Inf.	Sup.	Rango	Inf.	Sup.	Rango
X_1	0.7000	0.9000	0.2000	0.7000	0.9000	0.2000
X_2	0.2300	0.2900	0.0600	0.2300	0.2900	0.0600
X_3	0.2400	0.4900	0.2500	0.3400	0.3800	0.0400
X_4	0.2770	0.3230	0.0460	0.3010	0.3030	0.0020
X_5	0.4080	0.8080	0.4000	0.4960	0.7280	0.2320
X_6	0.3677	0.5646	0.1969	0.4301	0.4303	0.0002
X_7	0.4031	0.5892	0.1861	0.4091	0.5796	0.1705
X_8	0.0768	0.4776	0.4008	0.1088	0.4512	0.3424

TABLA 10.8. Cotas inferiores y superiores para las probabilidades marginales (iniciales) $p(X_i = 0)$ (caso sin evidencia).

	Caso 1: Once Parámetros			Caso 2: Siete Parámetros		
Nodo	Inf.	Sup.	Rango	Inf.	Sup.	Rango
X_1	1.0000	1.0000	0.0000	1.0000	1.0000	0.0000
X_2	0.2000	0.2000	0.0000	0.2000	0.2000	0.0000
X_3	0.3000	0.5000	0.2000	0.4000	0.4000	0.0000
X_4	0.2800	0.3200	0.0400	0.3000	0.3000	0.0000
X_5	0.4000	0.7600	0.3600	0.4800	0.6800	0.2000
X_6	0.3680	0.5640	0.1960	0.4300	0.4300	0.0000
X_7	0.4040	0.5880	0.1840	0.4100	0.5800	0.1700
X_8	0.0960	0.4800	0.3840	0.1280	0.4560	0.3280

TABLA 10.9. Cotas inferiores y superiores de las probabilidades condicionales $p(X_i = 0|X_1 = 0)$.

10.9 Propagación Simbólica en Redes Bayesianas Normales

En la Sección 8.9 se han presentado varios métodos para la propagación exacta en redes Bayesianas normales. Algunos de estos métodos se han extendido a la computación simbólica (véase, por ejemplo, Chang y Fung (1991) y Lauritzen (1992)). En esta sección se ilustra la propagación simbólica en redes Bayesianas nomales usando el método, conceptualmente más sencillo, de la Sección 8.9. Cuando se trata con cálculos simbólicos, todas las operaciones requeridas deben realizarse con un programa que disponga de capacidades de manipulación simbólica de expresiones. La Figura 10.11 muestra el código del programa realizado con *Mathematica* para implementar el método dado en la Sección 8.9. El código calcula la media y la varianza de todos los nodos dada la evidencia en una lista.

```
(* Vector de medias y matriz de covarianzas *)
mean={m,0,0,1,0};
var={{a,1,1,2,1},{1,2,1,-1,1},{1,1,b,-1,c},
     {2,-1,-1,4,-2},{1,1,c,-2,6}};
(* Orden de conocimiento de la evidencia *)
evidencia={3,5}

(* Probabilidades Marginales *)
For[k=0,k<=Length[evidencia],k++,
 For[i=1,i<=Length[mean],i++,
  If[MemberQ[Take[evidencia,k],i],
   cmean=x[i];
   cvar=0,
   meany=mean[[i]];
   meanz=Tabla[{mean[[evidencia[[j]]]]},{j,1,k}];
   vary=var[[i,i]];
   If[k==0,
    cmean=Together[meany];
    cvar=Together[vary],
    varz=Tabla[Tabla[
         var[[evidencia[[t]],evidencia[[j]]]],
       {t,1,k}],{j,1,k}];
    covaryz=Tabla[{var[[evidencia[[t]]]][[i]]},{t,1,k}];
    zaux=Tabla[{x[evidencia[[t]]]},{t,1,k}];
    aux=Inverse[varz];
    cmean=meany+Transpose[covaryz].aux.(zaux-meanz);
    cvar=vary-Transpose[covaryz].aux.covaryz;
   ]
  ];
  Print["Nodos evidenciales ",k," Nodo ",i];
  Print["Media ",Together[cmean],"Var ",Together[cvar]];
 ]
]
```

FIGURA 10.11. Un programa en *Mathematica* para propagar evidencia en una red Bayesiana normal.

Ejemplo 10.7 Considérese el conjunto de variables $X = \{X_1, \ldots, X_5\}$ con vector de medias y matriz de covarianzas

$$\mu = \begin{pmatrix} m \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \quad \text{y} \quad \Sigma = \begin{pmatrix} a & 1 & 1 & 2 & 1 \\ 1 & 2 & 1 & -1 & 1 \\ 1 & 1 & b & -1 & c \\ 2 & -1 & -1 & 4 & -2 \\ 1 & 1 & c & -2 & 6 \end{pmatrix}. \tag{10.30}$$

Nótese que la media de X_1, la varianza de X_1 y X_3, y la covarianza de X_3 y X_5 están especificadas en forma simbólica.

Se usa el código de *Mathematica* de la Figura 10.11 para calcular las medias y las varianzas condicionales de todos los nodos. La primera parte del código define el vector de medias y la matriz de covarianzas de la red Bayesiana. La Tabla 10.10 muestra las probabilidades marginales iniciales de los nodos (no evidencia) y las probabilidades condicionales de los nodos dadas cada una de las evidencias $\{X_3 = x_3\}$ y $\{X_3 = x_3, X_5 = x_5\}$. Un examen de los resultados de la Tabla 10.10 muestra que las medias y varianzas condicionadas son expresiones racionales, es decir, cocientes de polinomios en los parámetros. Nótese, por ejemplo, que para el caso de evidencia $\{X_3 = x_3, X_5 = x_5\}$, los polinomios son de primer grado en m, a, b, x_3 y x_5, es decir, en los parámetros que representan la media y la varianza y en las variables evidenciales, y de segundo grado en c, el parámetro de covarianza. Nótese también que el denominador de las funciones racionales es común para la media condicionada y la varianza condicionada. ∎

El hecho de que las medias y las varianzas de las distribuciones condicionales de los nodos sean funciones racionales se demuestra en el teorema que sigue (véase Castillo, Gutiérrez, Hadi y Solares (1997)).

Teorema 10.5 *Considérese una red Bayesiana normal definida en un conjunto de variables $X = \{X_1, \ldots, X_n\}$ con vector de media μ y matriz de covarianzas Σ. Particionemos X, μ y Σ en la forma $X = \{Y, Z\}$,*

$$\mu = \begin{pmatrix} \mu_Y \\ \mu_Z \end{pmatrix} \qquad \text{y} \qquad \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix},$$

donde μ_Y y Σ_{YY} son el vector de medias y la matriz de covarianzas de Y, μ_Z y Σ_{ZZ} son el vector de medias y la matriz de covarianzas de Z, y Σ_{YZ} es la matriz de covarianzas de Y y Z. Supóngase que Z es el conjunto de nodos evidenciales. Entonces la distribución condicional de cualquier variable $X_i \in Y$ dada Z, es normal, cuya media y varianza son cocientes de funciones polinomiales en las variables evidenciales y los correspondientes parámetros en μ y Σ. Los polinomios son como mucho de grado uno en las variables condicionantes, en los parámetros de medias y varianzas, y son de grado dos en los parámetros de covarianzas. Finalmente, el polinomio del denominador es el mismo para todos los nodos.

Demostración: Del Teorema 8.1 se tiene que

$$\mu_{Y|Z=z} = \mu_Y + \Sigma_{YZ}\Sigma_{ZZ}^{-1}(z - \mu_Z). \tag{10.31}$$

Nótese que $\Sigma_{YZ}\Sigma_{ZZ}^{-1}(z - \mu_Z)$ es una función racional puesto que puede ser escrita como cociente de los polinomios $\Sigma_{YZ}adj(\Sigma_{ZZ})(z - \mu_Z)$ y $det(\Sigma_{ZZ})$,

No Evidencia		
X_i	Media	Varianza
X_1	m	a
X_2	0	2
X_3	0	b
X_4	1	4
X_5	0	6
Evidencia $X_3 = x_3$		
X_i	Media	Varianza
X_1	$(bm + x_3)/b$	$(ab - 1)/b$
X_2	$(x_3)/b$	$(2b - 1)/b$
X_3	x_3	0
X_4	$(b - x_3)/b$	$(4b - 1)/b$
X_5	$(cx_3)/b$	$(6b - c^2)/b$
Evidencia $X_3 = x_3$ y $X_5 = x_5$		
X_i	Media	Varianza
X_1	$\frac{6bm - c^2m + (6-c)x_3 + (b-c)x_5}{6b - c^2}$	$\frac{6ab + 2c - ac^2 - b - 6}{6b - c^2}$
X_2	$\frac{(6-c)x_3 + (b-c)x_5}{6b - c^2}$	$\frac{11b + 2c - 2c^2 - 6}{6b - c^2}$
X_3	x_3	0
X_4	$\frac{6b - c^2 + (2c-6)x_3 + (c-2b)x_5}{6b - c^2}$	$\frac{20b + 4c - 4c^2 - 6}{6b - c^2}$
X_5	x_5	0

TABLA 10.10. Medias y varianzas de los nodos, inicialmente y después de conocer la evidencia.

donde $adj(\Sigma_{ZZ})$ es la matriz adjunta de Σ_{ZZ} y $det(\Sigma_{ZZ})$ es el determinante de Σ_{ZZ}. Por tanto, la esperanza condicional $\mu_{Y|Z=z}$ en (10.31) es μ_Y más una función racional, lo que implica que $\mu_{Y|Z=z}$ es una función racional con polinomio denominador $det(\Sigma_{ZZ})$. Nótese también que cada

parámetro aparece en uno sólo de los tres factores anteriores, lo que implica la linealidad en cada parámetro.

Similarmente, del Teorema 8.1 la varianza condicional es

$$\Sigma_{Y|Z=z} = \Sigma_{YY} - \Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}, \tag{10.32}$$

que es Σ_{YY} menos una función racional $\Sigma_{YZ}\Sigma_{ZZ}^{-1}\Sigma_{ZY}$. Esto implica que $\Sigma_{Y|Z=z}$ es una función racional con polinomio denominador $|\Sigma_{ZZ}|$. Nótese también que todos los parámetros excepto los de Σ_{YZ} aparecen en uno sólo de los factores, lo que implica de nuevo linealidad en estos parámetros. Por el contrario, los parámetros en Σ_{YZ} aparecen en dos factores, y por tanto pueden generar términos de segundo grado en los correspondientes polinomios.

Finalmente, el polinomio denominador puede ser de segundo grado en los parámetros de covarianza, debido a la simetría de la matriz de covarianzas. ∎

Nótese que por ser el polinomio denominador idéntico para todos los nodos, para su implementación es más conveniente calcular y almacenar los polinomios numeradores y el denominador separadamente, y no, como funciones racionales.

Ejercicios

10.1 Considérese el grafo dirigido de la Figura 10.12 y supóngase que las tres variables se dan en forma simbólica. Escribir los parámetros libres utilizando la notación en (10.2) en cada uno de los casos siguientes:

(a) Todas las variables son binarias.

(b) X_1 y X_2 son binarias pero X_3 es ternaria.

(c) X_1 y X_3 son ternarias pero X_2 es binaria.

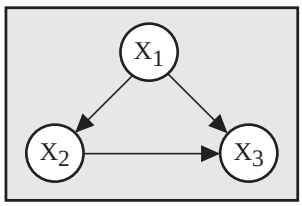


FIGURA 10.12. Un grafo dirigido con tres variables.

10.2 Considérese el grafo dirigido de la Figura 9.5 y las funciones de probabilidad condicionada de la Tabla 9.1. Escribir los parámetros libres usando la notación en (10.2) en cada uno de los casos siguientes:

$p(X_1 = 0) = 0.3$	$p(X_2 = 0) = \theta_{20}$
$P(X_3 = 0\|X_1 = 0, X_2 = 0) = 0.1$ $P(X_3 = 0\|X_1 = 0, X_2 = 1) = 0.3$ $P(X_3 = 0\|X_1 = 1, X_2 = 0) = 0.4$ $P(X_3 = 0\|X_1 = 1, X_2 = 1) = 0.5$	$P(X_6 = 0\|X_3 = 0, X_4 = 0) = 0.1$ $P(X_6 = 0\|X_3 = 0, X_4 = 1) = 0.4$ $P(X_6 = 0\|X_3 = 1, X_4 = 0) = 0.3$ $P(X_6 = 0\|X_3 = 1, X_4 = 1) = 0.2$
$p(X_4 = 0\|X_2 = 0) = 0.3$ $p(X_4 = 0\|X_2 = 1) = \theta_{401}$	$p(X_5 = 0\|X_3 = 0) = \theta_{500}$ $p(X_5 = 0\|X_3 = 1) = 0.8$

TABLA 10.11. Conjunto de parámetros de la red Bayesiana de la Figura 10.13

(a) Las funciones de probabilidad condicionales de X_3 y X_4 se reemplazan por parámetros simbólicos.

(b) Las funciones de probabilidad condicionales de X_1 y X_5 se reemplazan por parámetros simbólicos.

10.3 Considérese la red Bayesiana de la Figura 10.13 y las correspondientes funciones de probabilidad condicional numéricas y simbólicas dadas en la Tabla 10.11. Escribir un programa en *Mathematica* (usando las Figuras 10.3 y 10.4 como referencia de partida) para realizar la propagación simbólica de incertidumbre en cada uno de los casos siguientes:

(a) No se dispone de evidencia.

(b) Se da la evidencia $\{X_2 = 1, X_5 = 1\}$.

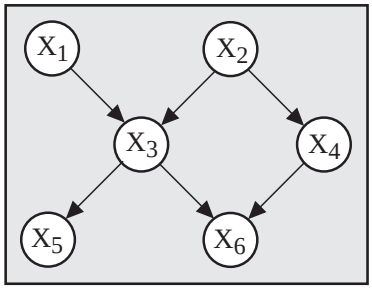


FIGURA 10.13. Un grafo múltiplemente conexo.

10.4 Considérese la red Bayesiana de la Figura 10.13 y la Tabla 10.11. Escribir la forma general de la estructura polinomial de las probabilidades marginales iniciales (sin evidencia) de los nodos, como en (10.8).

10.5 Considérese la red Bayesiana de la Figura 10.13 y la Tabla 10.11. Usar el método de las componentes canónicas para calcular $p(X_6 = 0|E = e)$ en cada uno de los casos siguientes:

(a) No hay evidencia disponible ($E = \phi$).

(b) La evidencia es $X_2 = 0$.

10.6 Considérese la red Bayesiana del Ejemplo 10.1. Usar el método de las componentes canónicas para calcular las probabilidades marginales de X_6 en cada uno de los casos siguientes:

(a) Las funciones de probabilidad condicionales de X_3 y X_4 se reemplazan por parámetros simbólicos.

(b) Las funciones de probabilidad condicionales de X_1 y X_2 se reemplazan por parámetros simbólicos.

10.7 Considérese la situación del Ejemplo 10.6. Calcular las cotas inferiores y superiores de las probabilidades de los nodos en los siguientes modelos reducidos:

(a) El modelo que resulta del modelo de siete parámetros reemplazando los parámetros de X_8 por $\theta_{300} = 0.2$ y $\theta_{801} = 0.4$.

(b) El modelo resultante del anterior reemplazando los parámetros de la variable X_7 por $\theta_{700} = 0.1$ y $\theta_{701} = 0.6$.

10.8 Repetir los cálculos del Ejemplo 10.7 para cada uno de los siguientes casos:

(a) Los nodos evidenciales son $X_1 = 0$ y $X_2 = 0$.

(b) Los nodos evidenciales son $X_1 = 0$ y $X_5 = 0$.

10.9 Dada la red Bayesiana de la Figura 9.5, supóngase que los nodos X_4 y X_8 son nodos simbólicos. Identificar los mensajes que están afectados por ambos parámetros, los que están afectados por sólo uno de ellos, y los no afectados.

Capítulo 11
Aprendizaje en Redes Bayesianas

11.1 Introducción

En los capítulos previos se ha supuesto que tanto la estructura de dependencia del modelo como las distribuciones condicionales de probabilidad asociadas se dan por los expertos humanos. En muchas situaciones prácticas, esta información puede no estar disponible. Además, diferentes expertos pueden dar diferentes valores, a veces contradictorios, debido al carácter subjetivo del proceso. En estas situaciones, la estructura de dependencia y las distribuciones condicionales de probabilidad asociadas pueden estimarse a partir de un conjunto de datos. Esto se conoce como *aprendizaje*.

Justo antes de enviar este libro a impresión, nos llamó la atención un libro interesante editado por Fisher y Lenz (1996). Los artículos contenidos en este libro tratan fundamentalmente del problema del aprendizaje, pero cubren otras áreas relacionadas con la inteligencia artificial y la estadística.

El aprendizaje puede estudiarse en los casos de las redes de Markov y en los de redes Bayesianas. Debido a limitaciones de espacio en este capítulo nos concentramos sólo en el aprendizaje de redes Bayesianas. Para el estudio del aprendizaje en las redes de Markov, se recomienda al lector consulte a Dawid y Lauritzen (1993), Madigan y Raftery (1994), y las referencias en ellos.

Tal como se ha visto en el Capítulo 6, una red Bayesiana es un modelo

$$B(\theta) = (D, P(\theta))$$

definido sobre el conjunto de n variables $X = \{X_1, \ldots, X_n\}$ y consiste en un grafo dirigido acíclico D y un conjunto $P(\theta)$ de n distribuciones condicionadas cada una de la forma

$$p(x_i|\pi_i; \theta_i); \quad i = 1, \ldots, n,$$

donde π_i es una realización del conjunto de padres del nodo X_i, θ_i es un conjunto de parámetros desconocidos de las distribuciones condicionales asociadas al nodo X_i, y $p(x_i|\pi_i; \theta_i)$ es una familia paramétrica que depende de θ_i. Por ello, $\theta = \{\theta_1, \ldots, \theta_n\}$ es el conjunto de parámetros de la distribución de probabilidad conjunta definida sobre X, y $\theta \in \Theta = \{\Theta_1, \ldots, \Theta_n\}$. Cuando no sea necesaria la referencia explícita a θ, se escribirá $B = (D, P)$ en vez de $B(\theta) = (D, P(\theta))$. La función de probabilidad conjunta sobre X se define como el producto de las n funciones de probabilidad condicionadas en $P(\theta)$. Esta factorización de la función de probabilidad conjunta está definida por la estructura topológica de D (véase la Sección 6.4.4).

Cuando D y/o $P(\theta)$ son desconocidas, pueden estimarse o aprenderse usando información "a priori" y/o un conjunto de datos S. La información previa refleja la opinión del experto humano sobre si una estructura gráfica o probabilística es más razonable que otras para representar las relaciones existentes entre las variables. Esta información se da a veces en la forma de probabilidades "a priori" en el conjunto de posibles estructuras gráficas y/o parámetros de la estructura probabilística.

El conjunto de datos S consta de N observaciones, cada una de las cuales consta de n valores, un valor para cada una de las variables $\{X_1, \ldots, X_n\}$. El conjunto de datos S puede contener valores desconocidos. Cuando S no contiene ningún valor desconocido, dicho conjunto se dice que es *un conjunto completo de datos*; en otro caso, el conjunto se dice que es de *datos incompletos*.

Cuando se trata del problema del aprendizaje de redes Bayesianas, es importante hacer notar que diferentes grafos dirigidos pueden representar las mismas estructuras de independencia y/o las mismas distribuciones conjuntas para el conjunto de sus variables. En otras palabras, cuando se resuelve el problema del aprendizaje en el conjunto de todas las redes Bayesianas, uno puede obtener muchas, aparentemente diferentes, soluciones, pero algunas de estas soluciones, de hecho, representan las mismas estructuras de independencia y/o las mismas distribuciones conjuntas. Desde el punto de vista del aprendizaje, los grafos dirigidos con las mismas estructuras de independencias y/o las mismas distribuciones conjuntas son equivalentes. Consecuentemente, si las aristas no tienen interpretaciones causales, se pueden obtener considerables ahorros en el cálculo, considerando sólo una estructura gráfica para cada conjunto de estructuras equivalentes. Puede ser interesante que el lector revise el material de la Sección 6.5 sobre redes Bayesianas equivalentes.

Por ello, usando la noción de independencia, o equivalencia de distribuciones, se puede dividir el conjunto de todos los grafos dirigidos posibles

sobre el conjunto de n variables o nodos en clases de equivalencia tal como se discute en la Sección 6.5.

La importancia práctica del concepto de redes independientemente equivalentes es que es necesario seleccionar un único grafo dirigido a partir del conjunto de sus equivalentes (clase de equivalencia) puesto que todos ellos conducen a la misma solución. Por ello, en este capítulo se toma la clase D de todos los grafos dirigidos acíclicos que conducen a la misma estructura de independencia, es decir, la clase de todos los grafos independientemente equivalentes. Esta clase puede, naturalmente, ser representada por cualquiera de ellos. Se ha elegido esta definición de D para evitar problemas con las medidas de calidad. Sin embargo, en lo que sigue, D puede ser cualquiera de estos grafos, siendo el resultado independiente de cuál se seleccione.

Se diferenciará en este capítulo entre dos tipos conceptualmente diferentes de aprendizaje:

1. **Aprendizaje Estructural:** Este se refiere al aprendizaje de la estructura (dependencia) gráfica de D, es decir, determinar las aristas a incluir en D.

2. **Aprendizaje Paramétrico:** Esto se refiere a aprender la estructura paramétrica (probabilidades) P. En el lenguaje de los estadísticos, el aprendizaje paramétrico se llama *estimación*.

Todo método de aprendizaje consta de dos elementos:

1. Una *medida de calidad*, que se usa para decidir cuál es la mejor de un conjunto de redes Bayesianas candidatas. Esto es una medida global de calidad, ya que mide la calidad de la red Bayesiana en su totalidad, es decir, ambas, la calidad de la estructura gráfica y la calidad de los parámetros estimados. En las Secciones 11.2−11.8 se discuten varias medidas de calidad.

2. Un *algoritmo de búsqueda*, que se usa para seleccionar un pequeño subconjunto de redes Bayesianas de alta calidad, del que se selecciona el mejor. Nótese que el número de todas las posibles redes, incluso para un pequeño número de variables, es extremadamente grande, y es virtualmente imposible determinar la calidad de todas las redes posibles. Los métodos de búsqueda se presentan en la Sección 11.9.

Por ello, aprender ambas, la estructura gráfica y la estructura probabilística de las redes Bayesianas, implica las siguientes etapas:

1. Elegir una medida de calidad y un algoritmo de búsqueda.

2. Utilizar el algoritmo de búsqueda para seleccionar un subconjunto de redes Bayesianas con calidades altas. Esto requiere estimar los parámetros de las redes seleccionadas usando métodos de estimación,

y evaluando las calidades de todas las redes Bayesianas en el conjunto elegido.

3. Seleccionar la estructura de la red con mayor calidad del conjunto anterior.

En este capítulo se discuten estas etapas. Las Secciones 11.2−11.8 introducen y discuten algunas medidas de calidad. Este material requiere algunos resultados de la estadística Bayesiana. Un resumen de ésta se da en el apéndice a este capítulo. En la Sección 11.9 se dan dos algoritmos de búsqueda. Finalmente, el problema de datos incompletos se trata en la Sección 11.10.

11.2 Midiendo la Calidad de una Red Bayesiana

Una medida de calidad, $Q(B|S,\xi)$, es un criterio mediante el cual se puede ordenar el conjunto de todas las redes Bayesianas posibles por su calidad, donde B es una red Bayesiana, ξ la información "a priori", y S un conjunto de datos. Por ello, dada la información "a priori" ξ y/o un conjnunto de datos S, nuestro objetivo consiste en obtener una red Bayesiana de alta calidad. Una medida de calidad debe satisfacer algunas propiedades deseables. Por ejemplo, debe asignarse la misma calidad a las redes que conduzcan a la misma estructura de independencia. A continuación se define esta importante propiedad.

Definición 11.1 Equivalencia en peso. *Dado un conjunto de datos S, una medida de calidad $Q(B|S,\xi)$ se dice que es equivalente en peso si asigna el mismo valor a todo par de redes Bayesianas equivalentes B_1 y B_2, es decir, si $Q(B_1|S,\xi) = Q(B_2|S,\xi)$.*

Otras propiedades de las medidas de calidad son:

- A las redes recomendadas por los expertos se les debe asignar calidades más altas que a las rechazadas por ellos.
- Las representaciones perfectas deben recibir calidades mayores que las imperfectas.
- Las I-representaciones mínimas deben recibir calidades mayores que las no mínimas.
- Las redes con reducido número de parámetros a igualdad del resto de propiedades deben recibir calidades mayores que las de elevado número de parámetros.
- A las redes que confirmen la información contenida en los datos debe asignársele una calidad mayor que a aquellas que contradigan a éstos.

Para ampliar conocimientos sobre éstas y otras propiedades se remite al lector a consultar el trabajo de Bouckaert (1995).

Las medidas de calidad dependen de la incertidumbre de la información disponible. Dos posibles situaciones son:

1. Una situación en la que las estructuras probabilísticas y gráficas están ambas sometidas a incertidumbre. En este caso, se dispone de la información "a priori" ξ y el conjunto de datos S, y el objetivo consiste en encontrar la mejor red Bayesiana $B(\theta) = (D, P(\theta))$ usando algún criterio de calidad. Nótese que ξ contiene información "a priori" referente a ambas estructuras, la gráfica y la paramétrica. Dados ξ y S, la calidad de una red Bayesiana $B(\theta)$ depende de la calidad de sus subcomponentes, D y $P(\theta)$. Se usa

$$Q(B(\theta)|S,\xi) \text{ o } Q(D, P(\theta)|S,\xi) \tag{11.1}$$

   para denotar la medida de calidad de la red Bayesiana en su totalidad y para indicar que la medida depende de S y ξ. Sin embargo, en algunos casos se puede estar interesado sólo en el aprendizaje estructural. En tales casos, se puede obtener una medida de la calidad de la estructura gráfica maximizando la calidad de sus redes Bayesianas $Q(B(\theta)|S,\xi)$ con respecto a θ, es decir,

$$Q(D|S,\xi) = Q(D, P(\hat{\theta})|S,\xi), \tag{11.2}$$

   donde $\hat{\theta}$ es el valor de θ que maximiza $Q(D, P(\theta)|S,\xi)$. Alternativamente, se puede usar cualquier otra estimación de θ, tal como la *estimación de máxima verosimilitud*, una *estimación Bayesiana*, etc.

2. Una situación en la que la estructura gráfica D es conocida y sólo la estructura probabilística está sometida a incertidumbre. En este caso, se está interesado sólo en el aprendizaje paramétrico, y el objetivo consiste en encontrar la mejor estructura probabilística $P(\theta)$, utilizando algún criterio de calidad. Dados S, D y ξ, la calidad de $P(\theta)$ depende de la calidad de los parámetros estimados. Se usa

$$Q(P(\theta)|D,S,\xi) \tag{11.3}$$

   para denotar la medida de calidad de la estructura probabilística de la red Bayesiana y para enfatizar que está condicionada a D, S, y ξ. Nótese que ξ sólo contiene información "a priori" sobre la estructura paramétrica ya que se conoce con certeza la estructura gráfica.

Algunas medidas de calidad se definen como la suma de tres términos o componentes:

$$Q = f(\text{información "a priori"}) + g(\text{datos disponibles}) + h(\text{complejidad}),$$

donde $f(.), g(.)$ y $h(.)$ son funciones conocidas. El significado de estos términos se explica a continuación:

1. La *información "a priori":* La función f(información "a priori") asigna una probabilidad alta a las redes que han sido indicadas como altamente probables por la información "a priori" y una probabilidad baja a las que han sido indicadas como poco probables. Cuanto mayor sea la contribución de este término a la medida de calidad, mayor será el peso del concocimiento "a priori" frente al aportado por los datos. Este término contribuye decisivamente a la calidad de la red en estudio cuando no existen datos disponibles o son muy reducidos, pero es despreciable cuando los datos disponibles son abundantes. Una elección típica para este término es $\log p(B)$, donde $p(B) = p(D, \theta)$ es la probabilidad "a priori" asignada a la red B, donde θ se usa en vez de P para mostrar la dependencia explícita de P del parámetro θ. Si no hay conocimiento "a priori" disponible, este término se sustituye por cero, lo que es equivalente a suponer que $p(B)$ es una distribución uniforme.

2. Los *datos disponibles*: La función g(datos disponibles) es un término de bondad de ajuste que mide lo bien o mal que una red Bayesiana reproduce los datos S. Da una alta calidad a las redes que están de acuerdo con los datos y una baja calidad a las que los contradicen. La contribución de este término aumenta cuando se añaden aristas a la red. En tal caso se tienen más parámetros o grados de libertad y, normalmente, se puede obtener un mejor ajuste a los datos. Algunas elecciones típicas para este término son las siguientes:

   (a) El logaritmo de la verosimilitud de los datos: $\log p(S|D, \theta)$.

   (b) El logaritmo de la probabilidad "a posteriori" de θ dada la estructura D y los datos S: $\log p(\theta|S, D)$.

3. La *complejidad*: La función h(complejidad) penaliza las redes con estructura compleja (por ejemplo, redes con un gran número de aristas y/o un número alto de parámetros). Por ello, la función $h(.)$ conduce a una calidad alta para las redes simples con un número reducido de aristas y parámetros, y a una baja calidad para las redes con muchas aristas y/o parámetros.

Para medir la complejidad de una red Bayesiana es importante conocer su dimensión.

Definición 11.2 Dimensión de una red Bayesiana. *Sea X un conjunto de variables y $B = (D, P)$ una red Bayesiana definida sobre X. La dimensión de esta red Bayesiana, $Dim(B)$, se define como el número de parámetros necesarios para especificar su función de probabilidad conjunta asociada.*

Chickering (1995a) muestra que las redes Bayesianas independientemente equivalentes tienen la misma dimensión.

En la literatura existente se han propuesto varias medidas de calidad para redes Bayesianas. Estas se han clasificado en los tipos siguientes:

- Medidas de calidad Bayesianas.
- Medidas de mínima longitud de descripción.
- Medidas de información.

Estos tipos de medidas de calidad se discuten en las secciones siguientes.

11.3 Medidas de Calidad Bayesianas

Las medidas de calidad Bayesianas se basan en la filosofía de la estadística Bayesiana (teorema de Bayes y las familias conjugadas en particular). En el apéndice de este capítulo se dan los conceptos necesarios para los lectores que no estén familiarizados con los conceptos de la teoría estadística Bayesiana (por ejemplo, distribuciones "a priori", distribuciones "a posteriori", distribuciones conjugadas, etc.). En este apéndice se incluyen también algunas familias conjugadas interesantes de distribuciones de probabilidad. Para una presentación más completa de la estadística Bayesiana se recomienda al lector que consulte alguno de los libros clásicos de la estadística Bayesiana, por ejemplo, DeGroot (1970), Press (1992), o Bernardo y Smith (1994).

En la teoría estadística Bayesiana, se supone inicialmente que la distribución "a priori" $p(B) = p(D, \theta)$ la dan los expertos. Esta distribución refleja la opinión de los expertos sobre la frecuencia relativa de ocurrencia de diferentes redes Bayesianas $B = (D, P(\theta))$. Para mejorar el conocimiento, se obtienen unos datos S y, mediante el teorema de Bayes, la distribución "a posteriori" $p(B, \theta|S)$ como sigue:

$$\begin{aligned} p(D, \theta|S) &= \frac{p(D, \theta, S)}{p(S)} = \frac{p(D, \theta, S)}{\sum\limits_{D,\theta} p(D, \theta, S)} \\ &= \frac{p(D, \theta)p(S|D, \theta)}{\sum\limits_{D,\theta} p(D, \theta)p(S|D, \theta)} \\ &= \frac{p(S, D, \theta)}{\sum\limits_{D,\theta} p(D, \theta)p(S|D, \theta)}. \end{aligned} \tag{11.4}$$

La idea básica de las medidas de calidad Bayesianas consiste en asignar a toda red una calidad que es una función de su probabilidad "a posteriori" (véase Heckerman (1995)). Por ejemplo, una medida de calidad puede definirse como una función que asigna a toda red Bayesiana valores proporcionales a su probabilidad "a posteriori" $p(B, \theta|S)$.

Nótese que el denominador en (11.4) es una constante de normalización, por lo que puede escribirse

$$p(D, \theta|S) \propto p(S, D, \theta) = p(D)p(\theta|D)p(S|D, \theta), \tag{11.5}$$

donde el símbolo $\propto$ significa "proporcional a". La igualdad en (11.5) se obtiene aplicando la regla de la cadena (véase la Sección 5.5). Usando esta regla puede también escribirse

$$p(S, D, \theta) = p(D)p(\theta|S, D)p(S|D) \propto p(D)p(\theta|S, D). \tag{11.6}$$

Esto indica que se puede utilizar $p(S, D, \theta)$ o $p(D)p(\theta|S, D)$ en vez de $p(D, \theta|S)$ para medir la calidad de la red. Los significados de los factores que aparecen en (11.5) y (11.6) son los siguientes:

- El factor $p(D)$ es la probabilidad "a priori" de la estructura gráfica, que es dada por los expertos humanos.
- El factor $p(\theta|D)$ es la probabilidad "a priori" de los parámetros de la red dada su estructura gráfica D. Esta distribución "a priori" también es dada por los expertos humanos.
- El factor $p(S|D, \theta)$ es la verosimilitud de los datos dada la red Bayesiana $B = (D, P(\theta))$. Este factor se calcula usando $P(\theta)$.
- El factor $p(\theta|S, D)$ es la distribución "a posteriori" de los parámetros de la red dada su estructura gráfica D y los datos S. Este factor se calcula basándose en los datos y en la distribución "a priori".

Por ello, para calcular las medidas de calidad de redes Bayesianas es necesario proceder mediante las siguientes etapas:

1. Asignar una distribución "a priori" $p(D)$ a la estructura gráfica.
2. Asignar una distribución "a priori" $p(\theta|D)$ a los parámetros.
3. Calcular la distribución "a posteriori" $p(\theta|S, D)$ usando la metodología Bayesiana usual (véase el apéndice de este capítulo).
4. Obtener una estimación $\hat{\theta}$ de θ.
5. Calcular la medida de calidad deseada, $Q(B(\theta), S)$.

La distribución "a priori", $p(D)$, de la Etapa 1 la suminstran los expertos humanos basándose en su conocimiento previo de las áreas de su especialidad. Ellos asignan una probabilidad a cada estructura de la red D. La especificación de la distribución "a priori", $p(\theta|D)$, de la Etapa 2, para todas las posibles estructuras de red es muy complicada por la enorme cantidad de información requerida. Una forma de evitar este problema consiste

en comenzar con una distribución de probabilidad, $p(\theta|D_c)$, para una red Bayesiana completa $B_c = (D_c, P_c)$. A partir de esta distribución, se puede calcular $p(\theta|D)$ para cualquier estructura de red D. Geiger y Heckerman (1995) hacen dos importantes hipótesis que permiten calcular $p(\theta|D)$ para cualquier estructura gráfica D especificando sólo la distribución "a priori" de los parámetros para la red completa B_c. Estas dos hipótesis son:

1. **Independencia parámetrica:** Supóngase que para toda red Bayesiana $B = (D, P)$, se tiene

$$p(\theta|D) = \prod_{i=1}^{n} p(\theta_i|D). \tag{11.7}$$

A veces, se supone también que θ_i puede ser particionada en la forma $(\theta_{i1}, \ldots, \theta_{is_i})$ con $\theta_{ik} \in \Theta_{ik}$ y

$$p(\theta_i|D) = \prod_{k=1}^{s_i} p(\theta_{ik}|D). \tag{11.8}$$

La hipótesis (11.7) dice que los parámetros asociados a cada variable (o nodo) de D son independientes. Esto se conoce como la hipótesis de *independencia de parámetros global*. La hipótesis (11.8) dice que los parámetros asociados a cada variable pueden particionarse en s_i conjuntos de parámetros independientes. Esto se conoce como la hipótesis de *independencia de parámetros local*. Si se verifican ambas, la independencia de parámetros local y la global, se dice simplemente que se tiene *independencia de parámetros*. Esta hipótesis significa que los parámetros que definen diferentes distribuciones de probabilidad condicional son mútuamente independientes. Nótese que estas distribuciones de probabilidad condicionales pueden ser dadas independientemente, sin más restricciones que las impuestas por los axiomas de la probabilidad. Como se verá, en algunos casos, tales como el modelo multinomial, con una distribución "a priori" de tipo Dirichlet y el normal con una distribución "a priori" de tipo normal-Wishart (véase el apéndice de este capítulo), la hipótesis de independencia de parámetros se satisface automáticamente; por tanto, no es necesario hacerla explícitamente.

2. **Modularidad paramétrica:** Si dos redes Bayesianas D_1 y D_2 tienen los mismos padres para un nodo dado, X_i, entonces

$$p(\theta_i|D_1) = p(\theta_i|D_2). \tag{11.9}$$

Esta hipótesis significa que la incertidumbre de los parámetros de diferentes módulos, o familias,[1] es independiente de cómo se ensamblan estos módulos para formar la red.

[1] La familia de un nodo consta del nodo y de sus padres.

Las hipótesis de independencia y modularidad paramétrica permiten obtener la distribución "a priori" $p(\theta|D)$ para toda estructura de red D, dada una simple distribución "a priori" $p(\theta|D_c)$ para una estructura de red D_c que sea completa. Supóngase que nos dan tal distribución "a priori" $p(\theta|D_c)$ para la estructura completa D_c y que se desea obtener la distribución "a priori" $p(\theta|D)$ para la estructura D. Esto implica la necesidad de calcular $p(\theta_i|D)$ para todos los nodos $X_i; i = 1, \ldots, n$. Estos cálculos se hacen utilizando el algoritmo siguiente:

Algoritmo 11.1 Calcula la distribución "a priori" de los parámetros de una red dada, a partir de la distribución "a priori" de una red completa.

- **Datos:** *Una red Bayesiana completa $B_c = (D_c, P_c)$, una distribución "a priori" $p(\theta|D_c)$, y otra red Bayesiana $B = (D, P)$.*
- **Resultados:** *La distribución "a priori" $p(\theta|D)$ correspondiente a D.*

1. *Sea $i = 1$, es decir, selecciónese el primer nodo.*
2. *Buscar una estructura completa D_c^i que tenga el nodo X_i y los mismos padres Π_i, utilizando una ordenación de los nodos en la que los nodos de Π_i sean anteriores a X_i.*
3. *Determinar la distribución "a priori" $p(\theta|D_c^i)$ asociada a esta estructura a partir de la distribución "a priori" $p(\theta|D_c)$ mediante un cambio de variable. Nótese que la hipótesis de independencia paramétrica garantiza la obtención de $p(\theta_i|D_c^i)$.*
4. *Usar la hipótesis de modularidad para obtener $p(\theta_i|D) = p(\theta_i|D_c^i)$ y $p(\theta_{ik}|D) = p(\theta_{ik}|D_c^i), k = 1, \ldots, s_i$.*
5. *Si $i < n$, hacer $i = i + 1$ e ir a la Etapa 2. En otro caso, continuar.*
6. *Usar la independencia paramétrica para obtener la función de probabilidad conjunta "a priori"*

$$p(\theta|D) = \prod_{i=1}^{n} p(\theta_i|D) = \prod_{i=1}^{n} \prod_{k=1}^{s_i} p(\theta_{ik}|D). \tag{11.10}$$

■

Aunque las hipótesis de independencia y modularidad paramétricas son útiles para obtener la distribución de cualquier red a partir de la distribución "a priori" de la red completa, una desventaja clara resultante de este método es que la especificación de la distribución "a priori" de la red completa puede resultar una tarea muy difícil, especialmente cuando el número de variables es grande.

En las Secciones 11.4 y 11.5 se desarrollan medidas de calidad Bayesianas para las redes multinomiales (como ejemplo de redes discretas) y las redes multinormales (como ejemplo de las continuas), respectivamente.

11.4 Medidas Bayesianas para Redes Multinomiales

En esta sección se presentan las medidas de calidad Bayesianas para redes multinomiales, que son unas de las redes de variables discretas más importantes. Sea $X = (X_1, \ldots, X_n)$ una variable aleatoria, es decir, un conjunto de n variables discretas, en el que cada variable X_i puede tomar uno de r_i valores diferentes $0, 1, \ldots, r_i - 1$. Por ello, se supone que $p(S|D, \theta)$ en (11.5) es una función de verosimilitud multinomial con parámetros desconocidos θ. Sea S un conjunto de datos (por ejemplo, una muestra) de N casos, es decir, cada caso consiste en un valor para cada variable, que representa un muestra aleatoria $(x_{1j}, \ldots, x_{nj}), j = 1, \ldots, N$, procedente de una distribución multinomial. Sea D una estructura de red definida sobre X y $N_{x_1,\ldots,x_n}$ el número de casos de S tales que $X_1 = x_i, \ldots, X_n = x_n$. Para toda variable X_i, sea

$$s_i = \prod_{X_j \in \Pi_i} r_j \tag{11.11}$$

el número de todas las realizaciones posibles de los padres de X_i.

11.4.1 Hipótesis y Resultados

Para obtener una medida de calidad Bayesiana para una red Bayesiana, se hacen las siguientes hipótesis:

- **Hipótesis 1:** El conjunto de datos S es *completo*, es decir, no hay datos incompletos. El caso en el que se tengan *datos incompletos* se trata en la Sección 11.10.
- **Hipótesis 2:** Modularidad paramétrica, tal como se ha descrito en la Sección 11.3.
- **Hipótesis 3:** Para una red completa, D_c, la distribución "a priori" de los parámetros $\theta_{x_1,\ldots,x_n}$, donde $\theta_{x_1,\ldots,x_n} = p(X_1 = x_1, \ldots, X_n = x_n)$, es una distribución de Dirichlet

$$p(\theta_{x_1,\ldots,x_n}|D_c) \propto \prod_{x_1,\ldots,x_n} \theta_{x_1,\ldots,x_n}^{\eta_{x_1,\ldots,x_n}-1}, \tag{11.12}$$

  donde $\eta_{x_1,\ldots,x_n}$ son sus parámetros (hiperparámetros).

Una razón para usar la distribución de Dirichlet es que es la distribución conjugada de la multinomial, es decir, la distribución conjunta "a posteriori" $p(\theta_{x_1,\ldots,x_n}|S, D_c)$, dados los datos y la distribución "a priori", es también Dirichlet

$$p(\theta_{x_1,\ldots,x_n}|S, D_c) \propto \prod_{x_1,\ldots,x_n} \theta_{x_1,\ldots,x_n}^{\eta_{x_1,\ldots,x_n}+N_{x_1,\ldots,x_n}-1}, \tag{11.13}$$

donde $N_{x_1,\ldots,x_n}$ es el número de casos de S con $X_i = x_i; i = 1, \ldots, n$.

Como se ha indicado anteriormente, una forma de especificar una distribución "a priori" de los parámetros θ es mediante una distribución "a priori" de una red completa (D_c, P_c). A partir de esta distribución "a priori" de la red completa se pueden calcular las distribuciones "a priori" de los valores de los hiperparámetros $\eta_{x_1,\ldots,x_n}$ mediante

$$\eta_{x_1,\ldots,x_n} = \eta \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}, D_c). \tag{11.14}$$

El hiperparámetro η mide la importancia (peso) de la distribución "a priori" y puede interpretarse como un tamaño de muestra equivalente.

Debido al hecho de que la función de probabilidad conjunta puede factorizarse como un producto de probabilidades condicionales, es más conveniente elegir los nuevos parámetros

$$\theta_{x_i|x_1,\ldots,x_{i-1}},\ \forall\, x_1, \ldots, x_{i-1},\ x_i = 0, 1, \ldots, r_i - 1,\ i = 1, \ldots, n, \tag{11.15}$$

donde

$$\theta_{x_1,\ldots,x_n} = \prod_{i=1}^{n} \theta_{x_i|x_1,\ldots,x_{i-1}},\ \forall\, (x_1, \ldots, x_n) \neq (r_1, \ldots, r_n), \tag{11.16}$$

y

$$\theta_{r_i|x_1,\ldots,x_{i-1}} = 1 - \sum_{x_i=0}^{r_i-1} \theta_{x_i|x_1,\ldots,x_{i-1}}.$$

Por razones de simplicidad, se usará también θ_{ijk} para denotar $\theta_{x_i|x_1,\ldots,x_{i-1}}$. Por ello, θ_{ijk} se refiere a la probabilidad de que $X_i = j$ dada la k-ésima realización del conjunto de padres Π_i. Nos referiremos a $\theta_{x_1,\ldots,x_n}$ como los parámetros iniciales y a θ_{ijk} como los parámetros condicionales. La distribución de probabilidad "a priori" de los parámetros condicionales es también Dirichlet. Esto se prueba en el teorema que sigue (Geiger y Heckerman (1995)).

Teorema 11.1 Distribución de los parámetros condicionales. *Si la función de densidad conjunta de los parámetros iniciales de una red completa es Dirichlet de la forma (11.12), la función de densidad de los parámetros condicionales de cada nodo de (11.15) es independiente y conjuntamente distribuida como una distribución de Dirichlet con función de densidad*

$$p(\theta_{ijk}) \propto \prod_{j=0}^{r_i} \theta_{ijk}^{\eta_{ijk}-1};\ k = 1, \ldots, s_i;\ i = 1, \ldots, n, \tag{11.17}$$

donde

$$\eta_{ijk} = \eta_{x_i|x_1\ldots x_{i-1}} = \sum_{x_{i+1},\ldots,x_n} \eta_{x_1,\ldots,x_n}, \tag{11.18}$$

y r_i es el número de valores distintos de X_i.

El Teorema 11.1 muestra que la independencia paramétrica, que no se verifica para los parámetros iniciales, se cumple para los nuevos parámetros. Esta independencia resulta como una consecuencia de la hipótesis de una distribución de Dirichlet para los parámetros iniciales y no es una hipótesis adicional.

El Teorema 11.1, combinado con el carácter conjugado de la distribución de Dirichlet para el caso de la verosimilitud multinomial, muestra que tanto las distribuciones "'a priori" como las distribuciones "a posteriori" para los parámetros iniciales y condicionales son también de Dirichlet. Por tanto, la distribución "a posteriori" de todos los parámetros $\theta = (\theta_1, \ldots, \theta_n)$ es una distribución de Dirichlet con

$$p(\theta|S, D_c) \propto \prod_{i=1}^{n} \prod_{j=0}^{r_i} \prod_{k=1}^{s_i} \theta_{ijk}^{\eta_{ijk}+N_{ijk}-1}, \tag{11.19}$$

donde $N_{ik} = \sum_{j=0}^{r_i} N_{ijk}$, y N_{ijk} es el número de casos de S en los que $X_i = j$ y el conjunto de padres de X_i toma valores asociados a la realización k-ésima de Π_i.

Según (11.5), para calcular $p(S, D, \theta)$ se necesita:

1. La distribución "a priori" $p(D)$, que está dada por expertos humanos.
2. La verosimilitud de los datos S, que es
$$p(S|D, \theta) \propto \prod_{i=1}^{n} \prod_{j=0}^{r_i} \prod_{k=1}^{s_i} \theta_{ijk}^{N_{ijk}}. \tag{11.20}$$
3. La distribución "a priori", $p(\theta|D)$, de los parámetros θ que se calcula utilizando las hipótesis de independencia y el Algoritmo 11.1. Por ello, según (11.10), se obtiene
$$p(\theta|D) = \prod_{i=1}^{n} \prod_{j=0}^{r_i} \prod_{k=1}^{s_i} \theta_{ijk}^{\eta_{ijk}-1}. \tag{11.21}$$
4. La distribución "a posteriori" por tanto resulta
$$p(\theta|S, D) \propto \prod_{i=1}^{n} \prod_{j=0}^{r_i} \prod_{k=1}^{s_i} \theta_{ijk}^{N_{ijk}+\eta_{ijk}-1}. \tag{11.22}$$

Los resultados anteriores se utilizan para obtener tres medidas diferentes:

- La medida de Geiger y Heckerman.
- La medida de Cooper-Herskovits.
- La medida Bayesiana usual.

11.4.2 La Medida de Geiger y Heckerman

Geiger y Heckerman (1995) proponen la siguiente medida de calidad Bayesiana

$$\begin{aligned} Q_{GH}(D,S) &= \log p(D) + \log \int p(S|D,\theta)p(\theta|D)d\theta \\ &= \log p(D) + \sum_{i=1}^{n}\left[\sum_{k=1}^{s_i}\left[\log\frac{\Gamma(\eta_{ik})}{\Gamma(\eta_{ik}+N_{ik})}\right.\right. \\ &\quad \left.\left. + \sum_{j=0}^{r_i}\log\frac{\Gamma(\eta_{ijk}+N_{ijk})}{\Gamma(\eta_{ijk})}\right]\right], \end{aligned} \tag{11.23}$$

donde $\Gamma(.)$ es la función *gamma*. Nótese que (11.23) no incluye un término explícito de penalización para la complejidad de la red.

11.4.3 La Medida de Cooper-Herskovits

Cooper y Herskovits (1992) proponen la medida de calidad siguiente:

$$\begin{aligned} Q_{CH}(D,S) &= \log p(D) + \sum_{i=1}^{n}\left[\sum_{k=1}^{s_i}\left[\log\frac{\Gamma(r_i)}{\Gamma(N_{ik}+r_i)}\right.\right. \\ &\quad \left.\left. + \sum_{j=0}^{r_i}\log\Gamma(N_{ijk}+1)\right]\right]. \end{aligned} \tag{11.24}$$

Esta medida tampoco incluye un término de penalización explícita de la complejidad de la red.

11.4.4 La Medida de Calidad Bayesiana Usual

Como alternativa a (11.23) y (11.24), se introduce aquí una medida de calidad que se obtiene evaluando la verosimilitud en el punto modal "a posteriori", y considerando un término de penalización explícito. La moda posterior, que es el máximo de (11.22) con respecto a θ, es

$$\hat{\theta}_{ijk} = \frac{\eta_{ijk}+N_{ijk}-1}{\eta_{ik}+N_{ik}-r_i}. \tag{11.25}$$

Asintóticamente, (11.25) tiende a

$$\hat{\theta}_{ijk} = \frac{N_{ijk}}{N_{ik}}. \tag{11.26}$$

Por ello, sustituyendo (11.25) en la verosimilitud, se obtiene la medida Bayesiana estándar o usual

$$\begin{aligned} Q_{SB}(D,S) &= \log p(D) \\ &\quad + \sum_{i=1}^{n}\sum_{j=0}^{r_i}\sum_{k=1}^{s_i}(N_{ijk}+\eta_{ijk}-1)\log\left(\frac{\eta_{ijk}+N_{ijk}-1}{\eta_{ik}+N_{ik}-r_i}\right) \\ &\quad -\frac{1}{2}Dim(B)\log N, \end{aligned} \tag{11.27}$$

donde $Dim(B)$ es la dimensión (el número de parámetros necesarios para especificar completemente la función de probabilidad conjunta de X) de la red Bayesiana B y N es el tamaño de la muestra. El último término en (11.27) es un término que penaliza las redes con mayor número de parámetros.

Alternativamente, se puede utilizar la estimación asintótica (11.26) en vez de la (11.25) y obtener

$$\begin{aligned} Q_{SB}(D,S) &= \log p(D) + \sum_{i=1}^{n}\sum_{j=0}^{r_i}\sum_{k=1}^{s_i} N_{ijk}\log\frac{N_{ijk}}{N_{ik}} \\ &\quad -\frac{1}{2}Dim(B)\log N. \end{aligned} \tag{11.28}$$

11.4.5 Ejemplos y Códigos de Ordenador

Para ilustrar las medidas de calidad anteriores mediante un ejemplo y suministrar una evaluación limitada, se necesita la definición siguiente.

Definición 11.3 Muestra perfecta. *Supóngase que se tiene un conjunto discreto de variables $(X_1, X_2, \ldots, X_n)$ con una función de probabilidad conjunta $p(x_1, \ldots, x_n)$. Una muestra S de $p(x_1, \ldots, x_n)$ se dice que es perfecta si y sólo si toda posible combinación de las variables, $(x_1, \ldots, x_n)$, aparece en S con una frecuencia proporcional a su probabilidad $p(x_1, \ldots, x_n)$.*

En lo que sigue se usarán muestras perfectas para evaluar el comportamiento de las diferentes medidas de calidad.

Ejemplo 11.1 Muestra perfecta. Considérese la red Bayesiana de la Figura 11.1, donde cada una de las variables (X_1, X_2, X_3) es binaria, y toma valores 0 ó 1. La red implica que la función de probabilidad conjunta puede factorizarse en la forma

$$p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3|x_1). \tag{11.29}$$

Supóngase también que las distribuciones de probabilidad condicionales necesarias para especificar $p(x_1, x_2, x_3)$ se dan en la Tabla 11.1. Las ocho posibles combinaciones de las tres variables, junto con la probabilidad de

x_1	$p(x_1)$
0	0.4
1	0.6

x_2	$p(x_2)$
0	0.2
1	0.8

x_1	x_3	$p(x_3\|x_1)$
0	0	0.1
0	1	0.9
1	0	0.6
1	1	0.4

TABLA 11.1. Distribuciones de probabilidad condicionales requeridas para especificar la función de probabilidad conjunta de los nodos de la red Bayesiana de la Figura 11.1.

$\{x_1, x_2, x_3\}$	$p(x_1, x_2, x_3)$	Frecuencia
$\{0, 0, 0\}$	0.008	8
$\{0, 0, 1\}$	0.072	72
$\{0, 1, 0\}$	0.032	32
$\{0, 1, 1\}$	0.288	288
$\{1, 0, 0\}$	0.072	72
$\{1, 0, 1\}$	0.048	48
$\{1, 1, 0\}$	0.288	288
$\{1, 1, 1\}$	0.192	192
Total	1.000	1000

TABLA 11.2. Muestra perfecta asociada a la red del Ejemplo 11.1.

cada una, se muestran en las dos primeras columnas de la Tabla 11.2. Una muestra perfecta de tamaño 1000 obtenida de la función de probabilidad anterior, consiste en todas las combinaciones mostradas en la Tabla 11.2. Son las frecuencias de la última columna de la misma. Por ello, la frecuencia de cada combinación de variables es proporcional a su probabilidad, $p(x_1, x_2, x_3)$. ∎

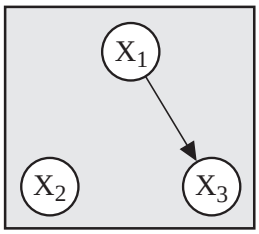


FIGURA 11.1. Un ejemplo de una red Bayesiana.

Ejemplo 11.2 Medidas de calidad para una red multinomial. Dadas las tres variables binarias del Ejemplo 11.1, la Figura 11.2 muestra las 11 estructuras de red posibles que conducen a diferentes clases de equivalencia.

Supóngase que se tiene la muestra perfecta S de la Tabla 11.2. Recuérdese que esta muestra se obtiene de la función de probabilidad conjunta dada por (11.29) y la Tabla 11.1. Esto significa que el modelo verdadero que genera S es la red Bayesiana número 4 de la Figura 11.2.

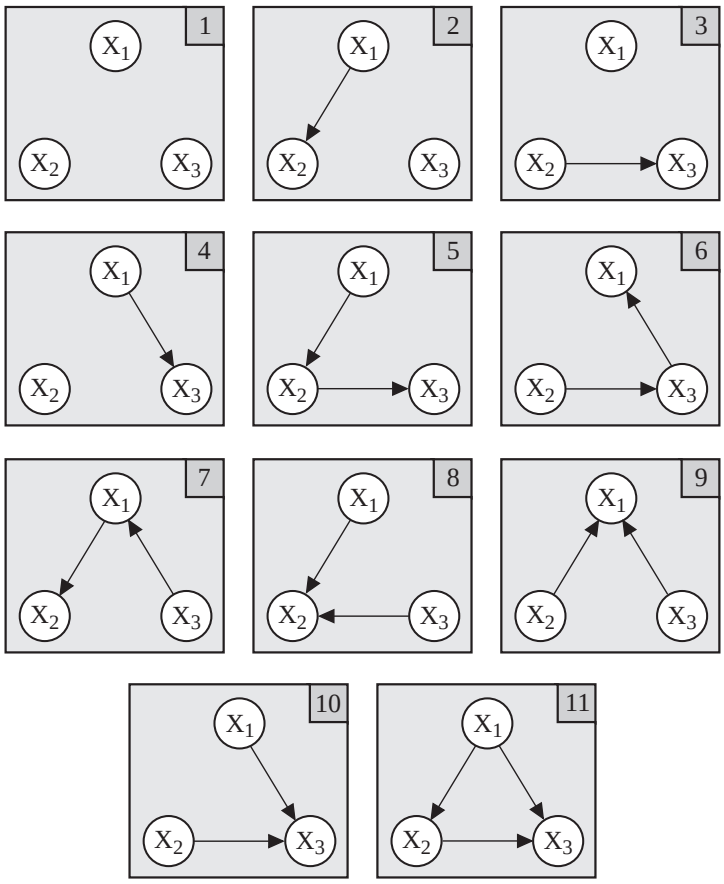


FIGURA 11.2. Todas las clases de redes Bayesianas diferentes, es decir, clases que conducen a diferentes estructuras de independencia.

La Figura 11.3 da un *programa en Mathematica* para calcular las medidas de calidad en (11.27). Pequeñas modificaciones de este programa permiten calcular las restantes medidas en (11.23), (11.24) y (11.28).

Ahora se utilizan muestras perfectas de tamaños 1000, 5000 y 10000 para calcular las medidas de calidad para cada una de las 11 estructuras de la Figura 11.2. Se utiliza una distribución de probabilidad "a priori" que asigna la misma frecuencia a cada celda (se ha seleccionado un valor 20, lo que implica $\eta = 160$).

```
TI={{{0,0,0},{20}},{{0,0,1},{20}},{{0,1,0},{20}},{{0,1,1},{20}},
{{1,0,0},{20}},{{1,0,1},{20}},{{1,1,0},{20}},{{1,1,1},{20}}};
TD={{{0,0,0},{8}},{{0,0,1},{72}},{{0,1,0},{32}},{{0,1,1},{288}},
{{1,0,0},{72}},{{1,0,1},{48}},{{1,1,0},{288}},{{1,1,1},{192}}};

SampleSize=Sum[TotalData[[i,2,1]],{i,1,Length[TotalData]}];
Netw[1]={{1},{2},{3}}; Netw[2]={{1},{2,1},{3}};
Netw[3]={{1},{2},{3,2}}; Netw[4]={{1},{2},{3,1}};
Netw[5]={{1},{2,1},{3,2}}; Netw[6]={{1,3},{2},{3,2}};
Netw[7]={{1,3},{2,1},{3}}; Netw[8]={{1},{2,1,3},{3}};
Netw[9]={{1,2,3},{2},{3}}; Netw[10]={{1},{2},{3,1,2}};
Netw[11]={{1},{2,1},{3,2,1}};

Card={2,2,2};
For[Net=1,Net<=11,Net++,Nnodes=Length[Netw[Net]];
For[i=0,i<Nnodes,i++,For[k=0,k<Card[[i+1]],k++,
Nsons[i]=Length[Netw[Net][[i+1]]]-1;
For[jj=0;fact=1;q1=0,jj<Nsons[i],jj++,
fact1=Card[[Netw[Net][[i+1]][[jj+2]]]];
q1+=fact*(fact1-1);fact*=fact1];q[i]=q1;
For[j=0,j<=q1,j++,N2[i][j][k]=0;N21[i][j][k]=0]]];
KK=Sum[(Card[[i+1]]-1)*(q[i]+1),{i,0,Nnodes-1}];
For[Ndat=1,Ndat<=Length[TD],Ndat++,
Data=TD[[Ndat]];DatInit=TI[[Ndat]];
For[i=0,i<Nnodes,i++,k=Data[[1]][[i+1]];
kInit=DatInit[[1]][[i+1]];
For[jj=0;j=0;jInit=0;fact=1,jj<Nsons[i],jj++,
fact1=Card[[Netw[Net][[i+1]][[jj+1]]]];
j+=fact*Data[[1]][[Netw[Net][[i+1]][[jj+2]]]];
jInit+=fact*DatInit[[1]][[Netw[Net][[i+1]][[jj+2]]]];fact*=fact1];
N2[i][j][k]+=Data[[2]][[1]];N21[i][jInit][kInit]+=DatInit[[2]][[1]]];
For[i=0,i<Nnodes,i++,
For[j=0,j<=q[i],j++,For[k=0;N3[i][j]=0;N31[i][j]=0,k<Card[[i+1]],
k++,N3[i][j]+=N2[i][j][k];N31[i][j]+=N21[i][j][k] ]]];
PNetw[Net]=N[Exp[-KK*Log[SampleSize]/2
+Sum[(N2[i][j][k]+N21[i][j][k]-1)*Log[(N2[i][j][k]+N21[i][j][k]-1)/
(N3[i][j])+N31[i][j][k]-Card[i+1]],
{i,0,Nnodes-1},{j,0,q[i]},{k,0,Card[[i+1]]-1}]]];
Print["Net=",Net," p=",PNetwork[Net]]
For[S=0;jj=1,jj<=11,jj++,S+=PNetw[jj]]
For[L={};jj=1,jj<=11,jj++,L=Append[L,{PNetw[jj]/S,jj}]];Sort[L]
```

FIGURA 11.3. Un programa en *Mathematica* para calcular la medida de calidad Bayesiana estándar en (11.27) usando la muestra perfecta de la Tabla 11.2.

La Tabla 11.3 muestra las distribuciones de probabilidad "a posteriori" de las 11 estructuras de red para diferentes tamaños de muestra para la medida de calidad de Geiger y Heckerman en (11.23). Para una muestra perfecta de tamaño 1000, esta medida no da la mayor probabilidad "a posteriori" a la red número 4, que es la que generó la muestra. Sin embargo, para muestras de tamaños 5000 y 10000, la red número 4 ya recibe la máxima probabilidad. Una de las razones para que ocurra esto es debida al uso de una distribución uniforme de probabilidad "a priori". Otra razón es que el peso de la distribución "a priori" domina la medida. Por ejemplo, cuando esta distribución se fija en 1 en vez de 20 (es decir $\eta = 8$), la red número 4 es identificada como la red con la mayor probabilidad "a posteriori", tal como se muestra en la Tabla 11.4.

Las Tablas 11.5, 11.6 y 11.7 muestran las probabilidades "a posteriori" asociadas a las 11 diferentes estructuras de red para diferentes tamaños de muestra y para las tres medidas de calidad en (11.24), (11.27) y (11.28), respectivamente. Como puede verse, las tres medidas identifican correctamente a la red 4 como la más probable para todos los tamaños de muestra. Sin embargo, la medida Bayesiana estándar en (11.28), da la mayor probabilidad "a posteriori" para esta red. Esto se debe al hecho de que η no influye en este caso asintótico. Nótese también que cuando aumenta el tamaño de la muestra, la probabilidad "a posteriori" para la medida asintótica tiende a la correspondiente a la medida Bayesiana estándar, como cabe esperar.

Los resultados de las Tablas 11.3−11.7 pueden resumirse como sigue:

1. A la red 4 (que es la estructura usada para generar S) le corresponde la calidad más alta, tal como cabría esperar. A las restantes redes les corresponden valores mucho menores de calidad.

2. Las calidades asignadas a la red número 4 aumentan con el tamaño de muestra, tal como es lógico. Esto significa que la probabilidad de que la medida de calidad seleccione la red correcta aumenta con el tamaño de muestra.

3. La medida de calidad Bayesiana estándar en (11.28) da mejores resultados que las otras tres medidas. Sin embargo, el comportamiento de una medida puede depender de otros factores no considerados en el análisis anterior, tales como la complejidad de la red y/o la función de probabilidad conjunta que se haya utilizado para generar los datos. Por tanto, son necesarias más comparaciones y una investigación profunda de las propiedades de las diferentes medidas de calidad para determinar cuál de ellas es más ventajosa. ■

Tamaño de muestra					
1000		5000		10000	
Red	Prob.	Red	Prob.	Red	Prob.
10	0.310	4	0.505	4	0.671
9	0.310	10	0.117	6	0.109
11	0.160	9	0.117	7	0.109
4	0.108	7	0.117	10	0.051
6	0.056	6	0.117	9	0.051
7	0.056	11	0.027	11	0.008
8	0.000	1	0.000	1	0.000
1	0.000	2	0.000	3	0.000
3	0.000	3	0.000	2	0.000
2	0.000	8	0.000	5	0.000
5	0.000	5	0.000	8	0.000

TABLA 11.3. La medida de Geiger y Heckerman en (11.23) con $\eta = 160$: Probabilidades "a posteriori" de las 11 estructuras de red de la Figura 11.2 para diferentes tamaños de muestra.

	Tamaño de muestra		
Red	1000	5000	10000
4	0.808	0.910	0.936
6	0.083	0.042	0.031
7	0.083	0.042	0.031
10	0.012	0.003	0.001
9	0.012	0.003	0.001
11	0.001	0.000	0.000
1	0.000	0.000	0.000
3	0.000	0.000	0.000
2	0.000	0.000	0.000
5	0.000	0.000	0.000
8	0.000	0.000	0.000

TABLA 11.4. La medida de Geiger y Heckerman en (11.23) con $\eta = 8$: Probabilidades "a posteriori" de las 11 estructuras de red de la Figura 11.2 para diferentes tamaños de muestra.

11.5 Medidas Bayesianas para Redes Multinormales

En esta Sección se consideran las medidas de calidad para redes Bayesianas normales, como un caso particular de redes Bayesianas continuas.

	Tamaño de muestra		
Red	1000	5000	10000
4	0.844	0.928	0.949
6	0.082	0.040	0.029
7	0.054	0.027	0.019
10	0.010	0.002	0.001
9	0.010	0.002	0.001
11	0.001	0.000	0.000
1	0.000	0.000	0.000
3	0.000	0.000	0.000
2	0.000	0.000	0.000
5	0.000	0.000	0.000
8	0.000	0.000	0.000

TABLA 11.5. La medida de Cooper y Herskovits en (11.24): Probabilidades "a posteriori" de las 11 estructuras de red de la Figura 11.2 para tres muestras perfectas.

Tamaño de muestra					
1000		5000		10000	
Red	Prob.	Red	Prob.	Red	Prob.
4	0.488	4	0.831	4	0.891
10	0.153	7	0.075	7	0.056
9	0.153	6	0.051	6	0.038
7	0.099	10	0.020	10	0.007
6	0.075	9	0.020	9	0.007
11	0.031	11	0.002	11	0.000
1	0.000	1	0.000	1	0.000
2	0.000	2	0.000	3	0.000
3	0.000	3	0.000	2	0.000
8	0.000	8	0.000	5	0.000
5	0.000	5	0.000	8	0.000

TABLA 11.6. La medida de calidad Bayesiana estándar en (11.28): Probabilidades "a posteriori" de las 11 estructuras de red de la Figura 11.2 para diferentes tamaños de muestras.

11.5.1 Hipótesis y Resultados

Para obtener las medidas de calidad de redes Bayesianas multinormales, se hacen las siguientes hipótesis:

1. **Hipótesis 1:** El conjunto de datos S es completo. El caso de datos incompletos se aborda en la Sección 11.10.

	Tamaño de muestra		
Red	1000	5000	10000
4	0.938	0.972	0.980
6	0.030	0.014	0.010
7	0.030	0.014	0.010
10	0.001	0.000	0.000
9	0.001	0.000	0.000
11	0.000	0.000	0.000
1	0.000	0.000	0.000
3	0.000	0.000	0.000
2	0.000	0.000	0.000
5	0.000	0.000	0.000
8	0.000	0.000	0.000

TABLA 11.7. La medida de calidad Bayesiana asintótica en (11.27): Probabilidades "a posteriori" de las 11 estructuras de red de la Figura 11.2 para diferentes tamaños de muestras.

2. **Hipótesis 2:** Modularidad paramétrica, tal como se ha descrito en la Sección 11.3.

3. **Hipótesis 3:** El conjunto de datos S consiste en una muestra aleatoria de una distribución normal n-variada $N(\mu, \Sigma)$ con vector de medias desconocido μ y matriz de covarianzas Σ. Por ello, $p(S|D, \theta)$ en (11.5) se supone que es una verosimilitud normal multivariada. En vez de la matriz de covarianzas Σ, es a veces conveniente trabajar con la matriz de precisión $W = \Sigma^{-1}$, que es la inversa de Σ. Por ello, aquí $\theta = \{\mu, W\}$.

4. **Hipótesis 4:** La distribución "a priori" de los parámetros $\theta = \{\mu, W\}$,

$$p(\theta|D_c) = p(\mu, W|D_c),$$

donde D_c es una red completa, es una distribución normal-Wishart (véase el apéndice de este capítulo), es decir, la distribución condicional de μ dado W es $N(\mu_0, \nu W)$, donde ν es una constante y μ_0 es un vector de constantes, y la distribución marginal de W es una distribución de Wishart con $\alpha > n - 1$ grados de libertad y matriz de precisión W_0.

Una razón para utilizar la distribución normal-Wishart es que es la conjugada natural de la distribución normal multivariada, es decir,

- La función de densidad conjunta "a posteriori", $p(\mu, W|D_c, S)$, de μ y W dados los datos es también una distribución normal-Wishart,

Hiperparámetros "a priori"	Hiperparámetros "a posteriori"
μ_0	$\frac{\nu\mu_0+N\bar{x}}{\nu+N}$
ν	$\nu+N$
α	$\alpha+N$
W_0	$W_N = W_0 + S + \frac{\nu N}{\nu+N}(\mu_0-\bar{x})(\mu_0-\bar{x})^T$

TABLA 11.8. Hiperparámetros "a priori" y "a posteriori" para el caso normal multivariado.

- La distribución condicional de μ dado W es $N(\mu_N, \left((\nu+N)W\right)^{-1})$ y la distribución marginal de W es una distribución de Wishart con $\alpha+N$ grados de libertad y matriz de precisión

$$W_N = W_0 + S + \frac{\nu N}{\nu+N}(\mu_0-\bar{x})(\mu_0-\bar{x})^T, \tag{11.30}$$

donde

$$\bar{x} = N^{-1}\sum_{i=1}^{N} x_i \tag{11.31}$$

y

$$S = \sum_{i=1}^{N}(x_i-\bar{x})(x_i-\bar{x})^T \tag{11.32}$$

son la media y la matriz de covarianzas muestrales, respectivamente.

La obtención de estos resultados se da en el apéndice a este capítulo (Ejemplo 11.9). La Tabla 11.8 da los valores de los hiperparámetros "a priori" y "a posteriori" y muestra cómo pueden ser actualizados a partir de los valores muestrales.

Los estimadores de Bayes son

$$\hat{\mu}_N = \frac{\nu\mu_0+N\bar{x}}{\nu+N} \quad \text{y} \quad \hat{W}_N = (\alpha+N)W_N^{-1},$$

que son asintóticamente equivalentes a los estimadores $\bar{x}$ y NS^{-1}, respectivamente.

La distribución "a priori" se especifica, de forma similar al caso multinomial, dando una red normal completa. A partir de esto se puede calcular μ_0, W_0, y los dos pesos ν y α que miden la importancia relativa de la información "a priori" frente a la contenida en los datos.

Puesto que se utilizan probabilidades condicionales, y por conveniencia, en vez de usar los parámetros μ (media) y W (matriz de precisión), se usan los nuevos parámetros m_i, $\beta_i = \{\beta_{ij}, j = 1, \ldots, i-1\}$, y v_i, donde los β_{ij} son los coeficientes de regresión cuando la variable X_i se expresa en función de las variables $X_1, \ldots, X_{i-1}$, m_i es la ordenada en el origen de esta regresión, y v_i es la inversa de la varianza residual, es decir,

$$p(x_i|x_1, \ldots, x_{i-1}) = N(\mu_i, 1/v_i),$$

donde

$$\mu_i = m_i + \sum_{j=1}^{i-1} \beta_{ij}\mu_j. \tag{11.33}$$

Esto implica que se puede interpretar una distribución normal multivariada como una red Bayesiana, en la que no hay arista entre X_j e X_i si $\beta_{ij} = 0, i < j$.

Por ello, se está interesado en conocer las distribuciones de probabilidad "a priori" y "a posteriori" de los nuevos parámetros, como funciones de las correspondientes distribuciones de probabilidad de los parámetros viejos. Heckerman y Geiger (1995) dan el teorema siguiente:

Teorema 11.2 Independencia global de parámetros normales. *Si los parámetros viejos (μ, W) tienen una distribución normal-Wishart, entonces los nuevos parámetros satisfacen*

$$p(m, v, D) = \prod_{i=1}^{n} p(m_i, v_i, \beta_i). \tag{11.34}$$

Por ello, la independencia global de parámetros normales se satisface automáticamente y se puede usar la distribución de probabilidad "a priori" de una red Bayesiana completa para obtener la distribución "a priori" de cualquier otra red (véase la Sección 11.3). La hipótesis de modularidad se transforma en

$$p(m_i, v_i\beta_i|D_1) = p(m_i, v_i\beta_i|D_2). \tag{11.35}$$

Una vez que se conoce la distribución "a priori" para una red Bayesiana dada, se puede calcular la distribución "a posteriori" y definir medidas de calidad basadas en ellas. Aquí se dispone de dos medidas: la medida de Geiger y Heckerman y la medida Bayesiana estándar.

11.5.2 La Medida de Geiger y Heckerman

Geiger y Heckerman (1994) proponen una medida Bayesiana de calidad similar a la (11.23), es decir,

$$Q_{GH}(D, S) = \log p(D) + \log \int p(S|D, \theta)p(\theta|D)d\theta, \tag{11.36}$$

donde $\theta = (\mu, W)$. Es bien sabido (véase DeGroot (1970)) que la distribución Bayesiana o mixta,

$$p(x|D_c) = \int p(x|\mu, W, D_c)p(\mu, W|D_c)d\mu dW$$

es una distribución t n-dimensional con $\gamma = \alpha - n + 1$ grados de libertad, vector de localización μ_0, y matriz de precisión $W_0' = (\nu\gamma/(\nu+1))W_0^{-1}$. Esto da una medida de calidad

$$\begin{aligned} Q_{GH}(D,S) \quad \propto \quad & \log p(D) \\ & + \log\left[(\pi)^{-nN/2}\left(\frac{\nu}{\nu+N}\right)^{n/2}\right] \\ & + \log\left[\frac{c(n,\alpha)d(N,\alpha)}{c(n,\alpha+N)}\right], \end{aligned} \tag{11.37}$$

donde

$$c(n,\alpha) = \left[\prod_{i=1}^{n}\Gamma\left(\frac{\alpha-i+1}{2}\right)\right]^{-1}$$

y

$$d(N,\alpha) = [det(W_0)]^{\alpha/2}[det(W_N)]^{-(\alpha+N)/2}.$$

Aquí $det(A)$ denota el determinante de la matriz A. Como en la medida de calidad (11.23), esta medida de calidad no incluye un término explícito de penalización por la complejidad de la red.

11.5.3 La Medida Bayesiana Estándar

Una medida de calidad para redes Bayesianas multinormales, similar a la de (11.27) para las redes Bayesianas multinomiales, puede obtenerse fácilmente sin más que evaluar la verosimilitud en la moda posterior de θ y considerar el término de penalización. Esto da

$$\begin{aligned} Q_B(B(\theta),S) \quad = \quad & \log p(D) + \log p(S|D,\hat{\theta}) \\ & -\frac{1}{2}Dim(B)\log N, \end{aligned} \tag{11.38}$$

donde $\hat{\theta}$ es la moda posterior de θ, es decir,

$$\hat{\theta} = arg\max_{\theta} \log p(\theta|D,S). \tag{11.39}$$

11.5.4 Ejemplos de código de ordenador

Seguidamente se ilustran las medidas de calidad para redes Bayesianas normales mediante un ejemplo. Se dan también los códigos de ordenador correspondientes en *Mathematica*.

Ejemplo 11.3 Medidas de calidad para redes Bayesianas normales. Considérese una variable tridimensional (X_1, X_2, X_3). Supóngase que la distribución "a priori" de los parámetros es una distribución normal-Wishart con

$$\nu = \alpha = 6;\ \hat{\mu}_0 = (0.1, -0.3, 0.2), v = (1, 1, 1), \beta_2 = (0);\ \beta_3 = (1, 1),$$

donde β_2 y β_3 se refieren a los coeficientes de regresión β_{ij} en (11.33) y

$$W_0 = \begin{pmatrix} 12/7 & 0 & 12/7 \\ 0 & 12/7 & 12/7 \\ 12/7 & 12/7 & 36/7 \end{pmatrix}.$$

Usando la muestra dada en la Tabla 11.9, y según (11.30), (11.32) y (11.31), se obtiene

$$\bar{x} = \begin{pmatrix} 0.510 \\ -0.362 \\ -0.785 \end{pmatrix}, \qquad S_N = \begin{pmatrix} 11.337 & 11.439 & 6.806 \\ 11.439 & 34.014 & 25.717 \\ 6.806 & 25.717 & 31.667 \end{pmatrix},$$

$$\mu_N = \begin{pmatrix} 0.415 \\ -0.348 \\ -0.558 \end{pmatrix}, \quad W_N = \begin{pmatrix} 13.825 & 11.322 & 6.659 \\ 11.322 & 35.746 & 27.713 \\ 6.659 & 27.713 & 41.288 \end{pmatrix}.$$

La Figura 11.4 lista un programa en *Mathematica* para calcular la medida de Geiger y Heckerman para redes normales. La Tabla 11.10 muestra las 11 estructuras de redes Bayesianas diferentes y sus calidades asociadas normalizadas. Nótese que a la estrucrura de red número 5 le corresponde la máxima calidad con un valor 0.763. El resultado no es sorprendente puesto que la muestra fue generada a partir de esta estructura con parámetros $\mu_0 = (0.5, 0.2, -0.5)$, $v = (1, 1, 1)$, $\beta_2 = (1)$, y $\beta_3 = (0, 1)$. Entonces, los estimadores son

$$\hat{\mu} = \mu_N = \begin{pmatrix} 0.415 \\ -0.348 \\ -0.558 \end{pmatrix},$$

$$\hat{W} = (\alpha + N)W_N^{-1} = \begin{pmatrix} 2.597 & -1.038 & 0.278 \\ -1.038 & 1.931 & -1.129 \\ 0.278 & -1.129 & 1.343 \end{pmatrix},$$

y las estimaciones asintóticas resultan

$$\hat{\mu} = \bar{x} = \begin{pmatrix} 0.510 \\ -0.362 \\ -0.785 \end{pmatrix},$$

$$\hat{W} = NS^{-1} = \begin{pmatrix} 2.773 & -1.249 & 0.418 \\ -1.249 & 2.086 & -1.426 \\ 0.418 & -1.425 & 1.699 \end{pmatrix},$$

Caso	X_1	X_2	X_3
1	−0.78	−1.55	0.11
2	0.18	−3.04	−2.35
3	1.87	1.04	0.48
4	−0.42	0.27	−0.68
5	1.23	1.52	0.31
6	0.51	−0.22	−0.60
7	0.44	−0.18	0.13
8	0.57	−1.82	−2.76
9	0.64	0.47	0.74
10	1.05	0.15	0.20
11	0.43	2.13	0.63
12	0.16	−0.94	−1.96
13	1.64	1.25	1.03
14	−0.52	−2.18	−2.31
15	−0.37	−1.30	−0.70
16	1.35	0.87	0.23
17	1.44	−0.83	−1.61
18	−0.55	−1.33	−1.67
19	0.79	−0.62	−2.00
20	0.53	−0.93	−2.92

TABLA 11.9. Un conjunto de datos generado al azar a partir de la estructura número 5 de la Figura 11.3, suponiendo normalidad.

que están cercanas pero no mucho, indicando que la información "a priori" tiene un cierto peso, debido al hecho de que el tamaño de muestra $N = 20$ es pequeño. ∎

11.6 Medidas de Mínimo Requerimiento Descriptivo

Las medidas de calidad descritas en las secciones anteriores requieren especificar la información "a priori" de ambas, la estructura gráfica y la probabilística. Esta información puede no ser accesible. Las medidas de mínimo requerimiento descriptivo son una forma alternativa de medir la calidad de una estructura de red. Este concepto viene de la *teoría de la codificación*, en la que una cadena de caracteres se codifica en el mínimo número de bits. La idea básica consiste en comprimir una cadena dividiéndola en subcadenas; las cadenas más frecuentes se codifican en mensajes cortos y las menos frecuentes en mensajes más largos, conduciendo a una longitud media lo más pequeña posible. La codificación consta de dos partes: la descripción

```
Needs["Statistics`ContinuousDistributions`"];
dist=NormalDistribution[0,1];
x=Table[Random[dist],{i,1,3},{j,1,2000}];
cc={{1,0,0},{0,1,0},{1,0,1}};
X=Transpose[cc.x];
mu0={0.0,0.0,0.0};v={1,1,2};b2={0};b3={0,0};
nu=6;alp=6;
c[n_,v_]:=N[Product[Gamma[(v+1-i)/2],{i,1,n}]),1000];
L=Length[X];n=Length[X[[1]]];
XL={};;For[i=1,i<=n,i++,M=0;For[j=1,j<=L,j++,M+=X[[j]][[i]]];
AppendTo[XL,{M/L}]];
muL=(nu*mu0+L*XL)/(nu+L);
SL=Sum[(X[[i]]-XL).Transpose[(X[[i]]-XL)],{i,1,L}];
W={{1/v[[1]]+b3[[1]]^ 2/v[[3]],b3[[1]]*b3[[2]]/v[[3]],-b3[[1]]/v[[3]]},
{b3[[1]]*b3[[2]]/v[[3]],1/v[[2]]+b3[[2]]^ 2/v[[3]],-b3[[2]]/v[[3]]},
{-b3[[1]]/v[[3]],-b3[[2]]/v[[3]],1/v[[3]]}};
W1=Inverse[W]; T0=(nu*(alp-n-1))/(nu+1)*W1;
TL=T0+SL+(nu*L)/(nu+L)*(mu0-XL).Transpose[(mu0-XL)];
P[n_,L_,T0_,TL_]:=N[(Pi)^ (-n*L/2)*(nu/(nu+L))^ (n/2)*
Det[T0]^ (alp/2)*Det[TL]^ (-(alp+L)/2)*c[n,alp]/c[n,alp+L],1000];
T012={{T0[[1]][[1]],T0[[1]][[2]]},{T0[[2]][[1]],T0[[2]][[2]]}};
TL12={{TL[[1]][[1]],TL[[1]][[2]]},{TL[[2]][[1]],TL[[2]][[2]]}};
T023={{T0[[2]][[2]],T0[[2]][[3]]},{T0[[3]][[2]],T0[[3]][[3]]}};
TL23={{TL[[2]][[2]],TL[[2]][[3]]},{TL[[3]][[2]],TL[[3]][[3]]}};
T013={{T0[[1]][[1]],T0[[1]][[3]]},{T0[[3]][[1]],T0[[3]][[3]]}};
TL13={{TL[[1]][[1]],TL[[1]][[3]]},{TL[[3]][[1]],TL[[3]][[3]]}};
T01={{T0[[1]][[1]]}};TL1={{TL[[1]][[1]]}};
T02={{T0[[2]][[2]]}};TL2={{TL[[2]][[2]]}};
T03={{T0[[3]][[3]]}};TL3={{TL[[3]][[3]]}};
P12=P[2,L,T012,TL12];P13=P[2,L,T013,TL13];
P23=P[2,L,T023,TL23];P1=P[1,L,T01,TL1];
P2=P[1,L,T02,TL2];P3=P[1,L,T03,TL3];P123=P[3,L,T0,TL];
PNetw[1]=P1*P2*P3;PNetw[2]=P12*P3;PNetw[3]=P23*P1;
PNetw[4]=P13*P2;PNetw[5]=P12*P23/P2;PNetw[6]=P13*P23/P3;
PNetw[7]=P12*P13/P1;PNetw[8]=P123*P1*P3/P13;
PNetw[9]=P123*P2*P3/P23;
PNetw[10]=P123*P1*P2/P12;PNetw[11]=P123;
SS=Sum[PNetw[i],{i,1,11}];
Sort[Table[PNetw[i]/SS,{i,1,11}]]
```

FIGURA 11.4. Un programa en *Mathematica* para simular datos de una red Bayesiana normal y calcular la medida de calidad de Geiger y Heckerman (11.37).

Red	Probabilidad
5	0.763
8	0.136
11	0.049
3	0.036
6	0.013
10	0.002
2	0.000
7	0.000
9	0.000
1	0.000
4	0.000

TABLA 11.10. La medida de Geiger y Heckerman en (11.37): Probabilidades "a posteriori" correspondientes a las 11 estructuras de la Figura 11.2.

de la transformación utilizada en la compresión y la cadena comprimida. La longitud de descripción de una cadena de símbolos es la suma de las longitudes de estas dos partes. El principio de mínimo requerimiento selecciona la transformación que conduce a una mínima longitud de descripción.

En el caso de las redes Bayesianas, la longitud de descripción incluye:

1. La longitud requerida para almacenar la estructura de la red Bayesiana. Puesto que el número máximo de aristas en una red Bayesiana con n nodos es $n(n-1)/2$, y se puede almacenar un 1 si existe la arista y un 0 en otro caso,[2] entonces el máximo número de aristas (longitud) requerido es $n(n-1)/2$. Nótese que este número no depende de la estructura en particular. Por ello, no necesita ser incluída en la medida de calidad.

2. La longitud requerida para almacenar los parámetros θ. Es bien sabido que la longitud media necesaria para almacenar un número en el rango 0 a N es $\frac{1}{2}\log N$ (véase Rissanen (1983, 1986)). Por ello, para almacenar los parámetros de la función de probabilidad conjunta, se necesita una longitud de $\frac{1}{2}Dim(B)\log N$, donde N es el tamaño de la muestra y

$$Dim(B) = \sum_{i=0}^{n}(r_i - 1)\prod_{X_j \in \Pi_i} r_j = \sum_{i=1}^{n}(r_i - 1)s_i \qquad (11.40)$$

es el número de parámetros libres (grados de libertad) asociados a la función de probabilidad conjunta. Nótese que s_i es el número de

[2]Esta puede no ser la forma óptima de almacenar la estructura. Se ha elegido por razones de simplicidad.

parámetros libres (grados de libertad) asociados a la distribución de probabilidad condicional para X_i, $p(x_i|\pi_i)$.

3. La longitud de descripción del conjunto de datos S comprimidos usando la distribución asociada a (D, P), que en el caso de las redes Bayesianas multinomiales resulta ser

$$\sum_{i=1}^{n}\sum_{j=0}^{r_i}\sum_{k=1}^{s_i} N_{ijk}\log\frac{N_{ijk}}{N_{ik}},$$

que es $-N$ veces la entropía

$$H(S,B) = \sum_{i=1}^{n}\sum_{j=0}^{r_i}\sum_{k=1}^{s_i} -\frac{N_{ijk}}{N}\log\frac{N_{ijk}}{N_{ik}}. \tag{11.41}$$

Bouckaert (1995) añade un término adicional para incorporar la información "a priori" y propone la siguiente medida de mínima longitud de descripción

$$\begin{aligned} Q_{MLD}(B,S) &= \log p(D) + \sum_{i=1}^{n}\sum_{j=0}^{r_i}\sum_{k=1}^{s_i} N_{ijk}\log\frac{N_{ijk}}{N_{ik}} \\ &\quad -\frac{1}{2}Dim(B)\log N. \end{aligned} \tag{11.42}$$

Nótese que las medidas de calidad (11.28) y (11.42) son las mismas, y que las medidas de calidad (11.27) y (11.42) son asintóticamente equivalentes.

11.7 Medidas de Información

Otra forma de medir la calidad de una red es mediante las medidas de información. La idea básica consiste en seleccionar la estructura de red que mejor se ajusta a los datos, penalizada por el número de parámetros necesarios para especificar su correspondiente función de probabilidad conjunta. Esto conduce a la medida de información

$$Q_I(B,S) = \log p(D) + \sum_{i=1}^{n}\sum_{j=0}^{r_i}\sum_{k=1}^{s_i} N_{ijk}\log\frac{N_{ijk}}{N_{ik}} - Dim(B)f(N), \tag{11.43}$$

donde $f(N)$ es una función de penalización no negativa.

En la literatura existente se han propuesto muchas funciones de penalización, tales como *el criterio de máxima verosimilitud de información* ($f(N) = 0$), el *criterio de información de Akaike* (Akaike (1974)), ($f(N) = 1$) y el *criterio de información de Schwarz* (Schwarz (1978)), ($f(N) = \log(N)/2$). Nótese que la medida de mínima longitud de descripción (11.42) es un caso particular de esta medida.

11.8 Análisis Posterior de las Medidas de Calidad

Las medidas de calidad presentadas en las secciones anteriores son capaces de distinguir I-representaciones y representaciones perfectas. Bouckaert (1995) estudia estas medidas de calidad para redes Bayesianas discretas y da los teoremas siguientes:

Teorema 11.3 Medidas de calidad e I-representaciones. *Sea X un conjunto de variables y $\alpha(X)$ una ordenación total de X. Supóngase que la distribución "a priori" definida sobre todas las estructuras sobre X es positiva. Sea P una función de probabilidad conjunta sobre X tal que D es una I-representación mínima de P que satisface $\alpha(X)$ y que ninguna otra estructura de red que satisface $\alpha(X)$ es una I-representación mínima de P. Sea S una muestra de tamaño N obtenida de P. Sea Q la medida Bayesiana de Cooper-Herskovits, la medida de mínima longitud de descripción, o la medida de información con función de penalización no nula f, donde $lim_{N\to\infty} f(N) = \infty$ y $lim_{N\to\infty} f(N)/N = 0$. Entonces, para cualquier estructura de red D' sobre X que satisfaga la ordenación $\alpha(X)$, se tiene*

$$\lim_{N\to\infty} [N(Q(D', S) - Q(D, S))] = -\infty$$

si y sólo si D' no es una I-representación mínima de P.

Este teorema demuestra que asintóticamente, a las I-representaciones se les asignan mejores calidades que a las que no lo son.

Teorema 11.4 Medidas de calidad y representaciones perfectas. *Sea X un conjunto de variables. Supónganse que la distribución "a priori" sobre todas las estructuras de red sobre X es positiva. Sea P una distribución positiva sobre X tal que exista una representación perfecta para P. Sea B tal representación perfecta para P. Ahora, sea S una muestra de tamaño N de P. Sea Q la medida Bayesiana, la de mínima longitud de descripción, o una medida de información con función de penalización, f, no nula donde $lim_{N\to\infty} f(N) = \infty$ y $lim_{N\to\infty} f(N)/N = 0$. Entonces, para cualquier estructura de red D' sobre X se tiene*

$$\lim_{N\to\infty} [N(Q(D, S) - Q(D', S))] = -\infty$$

si y sólo si D' no es una representación perfecta de P.

Este teorema muestra que asintóticamente a las representaciones perfectas se les asigna una calidad mejor que a las que no lo son.

Finalmente, el teorema siguiente muestra que existen grafos dirigidos óptimos con un número reducido de aristas.

Teorema 11.5 *Sea X un conjunto de variables. Supóngase una distribución "a priori" uniforme definida sobre el conjunto de todas las estructuras*

posibles en X. Sea S una muestra de tamaño N de X. Sea B una estructura de red sobre X con un conjunto de padres que contiene más de $\log(N/f(N)+1)$ variables. Entonces, existe una estructura de red D' tal que $Q_f(D',S) > Q_f(D,S)$ con menos aristas, donde Q_f es una medida de calidad de información con función de penalización no nula f.

11.9 Algoritmos de Búsqueda de Redes Bayesianas

En las secciones anteriores se han discutido varias medidas de calidad de redes Bayesianas. Estas medidas son usadas por los algoritmos de búsqueda para encontrar redes Bayesianas de alta calidad. El número de posibles estructuras de red puede ser tan grande que es prácticamente imposible evaluar cada una de ellas. Esta sección presenta dos algoritmos de búsqueda que intentan buscar la red Bayesiana con la mayor calidad dada una cierta información "a priori" y un conjunto de datos.

Se hace notar que hay métodos de búsqueda que trabajan en el espacio de todas las redes Bayesianas y otros que lo hacen en el de las clases de equivalencia de las estructuras de red. Para ampliar detalles se remite al lector a los trabajos de Spirtes y Meek (1995), y Chickering (1995a).

11.9.1 El Algoritmo K2

Cooper y Herskovits (1992) proponen un algoritmo para encontrar redes Bayesianas de alta calidad. Ellos le denominan el algoritmo $K2$. Este algoritmo comienza con la red más simple posible, es decir, una red sin aristas, y supone que los nodos están ordenados. Para cada variable X_i, el algoritmo añade a su conjunto de padres Π_i el nodo con número menor que X_i que conduce a un máximo incremento en calidad correspondiente a la medida de calidad elegida para el proceso de búsqueda. El proceso se repite hasta que, o bien no se incrementa la calidad, o se llega a una red completa.

El pseudocódigo para este algoritmo se da en la Figura 11.5, donde $q_i(\Pi_i)$ es la contribución de la variable X_i con conjunto de padres Π_i a la calidad de la red. Por ejemplo, utilizando la medida de calidad de Geiger y Heckerman (11.23), se tiene

$$q_i(\Pi_i) = \sum_{k=1}^{s_i} \log \frac{\Gamma(\eta_{ik})}{\Gamma(\eta_{ik}+N_{ik})} + \sum_{j=0}^{r_i} \log \frac{\Gamma(\eta_{ijk}+N_{ijk})}{\Gamma(\eta_{ijk})}, \tag{11.44}$$

mientras que usando la de mínima longitud de descripción de Bouckaert (1995) (11.42), resulta

$$q_i(\Pi_i) = \sum_{j=0}^{r_i} \sum_{k=1}^{s_i} N_{ijk} \log \frac{N_{ijk}}{N_{ik}} - \frac{1}{2} s_i (r_i - 1) \log N. \tag{11.45}$$

```
Etapa de Iniciación:
    Ordenar las variables
    para i ← 1 a n hacer
        Π_i ← φ
Etapa Iterativa:
    para i ← 1 a n hacer
    repetir
        seleccionar el nodo Y ∈ {X_1, ..., X_{i-1}} \ Π_i que
            maximiza g = q_i(Π_i ∪ {Y})
        δ ← g − q_i(Π_i)
        si δ > 0 entonces
            Π_i ← Π_i ∪ {Y}
    hasta que δ ≤ 0 ó Π_i = {X_1, ..., X_{i-1}}
```

FIGURA 11.5. Pseudocódigo para el algoritmo $K2$.

```
Etapa de Iniciación:
    para i ← 1 a n hacer
        Π_i ← φ
    para i ← 1 a n y j ← 1 a n hacer
        si i ≠ j entonces
            A[i, j] ← m_i(X_j) − m_i(φ)
        en otro caso
            A[i, j] ← −∞  {no permitir X_i → X_i}
Etapa Iterativa:
    repetir
        seleccionar los i, j, que maximizan A[i, j]
        si A[i, j] > 0 entonces
            Π_i ← Π_i ∪ {X_j}
            para X_a ∈ Ascen_i, X_b ∈ Desc_i hacer
                A[a, b] ← −∞  {no permitir ciclos}
            para k ← 1 a n hacer
                si A[i, k] > −∞ entonces
                    A[i, k] ← m_i(Π_i ∪ {X_k}) − m_i(Π_i)
    hasta que A[i, j] ≤ 0 o A[i, j] = −∞, ∀ i, j
```

FIGURA 11.6. Pseudocódigo para el algoritmo B.

Un problema del algoritmo $K2$ es que requiere una ordenación previa de los nodos. El algoritmo que sigue no requiere tal ordenación.

11.9.2 Algoritmo B

Este algoritmo fué propuesto por Buntine (1991). El algoritmo comienza con conjuntos de padres vacíos como el algoritmo $K2$. En cada etapa se añade una nueva arista que no de lugar a ciclos y que maximice el incremento de calidad. El proceso se repite hasta que no se consigue incrementar más la calidad o se obtiene una red completa.

La Figura 11.6 muestra el pseudocódigo para este algoritmo, donde $Ascen_i$ es el conjunto de ascendientes del nodo X_i y $Desc_i$ es el conjunto de descendientes de X_i.

11.10 El Caso de Datos Incompletos

En las secciones anteriores se ha supuesto que el conjunto de datos disponible S es completo, es decir, cada unidad de información contiene datos para todas las variables. En esta sección se presentan algunos métodos para tratar el caso de datos incompletos:

- El muestreo de Gibbs.
- El algoritmo EM.

11.10.1 El Muestreo de Gibbs

El muestreo de Gibbs (véase, por ejemplo, Gelfand y Smith (1990), Casella y George (1992), y Gilks y Wild (1992) para una descripción detallada del mismo) se basa en simular secuencialmente las distribuciones univariadas de las variables dadas todas las demás, tal como se muestra en el algoritmo que sigue.

Algoritmo 11.2 El muestreo de Gibbs estándar.

- **Datos:** Una función $f(x_1, \ldots, x_n)$, un conjunto de probabilidades condicionales $p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n), i = 1, \ldots, n$, y dos enteros m_1 y m_2, que son los números de iteraciones de las fases de iniciación y estimación, respectivamente.
- **Resultados:** El valor esperado de $f(x_1, \ldots, x_n)$ respecto a la probabilidad conjunta correspondiente a $p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$.

1. Iniciación: Hacer $mediatheta = 0$ y θ^1 igual a su estimación "a priori".
2. Para $j = 1$ a $j = m_1$ hacer:
   (a) Usar el muestreo estándar de Gibbs para simular los datos incompletos usando la probabilidad condicional $p(x_k|x^j_{(k)}, \theta^j)$.

(b) Usar el método para datos completos para obtener la nueva estimación $\theta^{j+1} = \hat{\theta}^j$ de θ.

3. Para $j = m_1$ a $j = m_1 + m_2$ hacer:

   (a) Usar el muestreo estándar de Gibbs para simular los datos incompletos usando la probabilidad condicional $p(x_k|x^j_{(k)}, \theta^j)$.

   (b) Usar el método para datos completos para obtener la nueva estimación $\theta^{j+1} = \hat{\theta}^j$ de θ.

   (c) $mediatheta = mediatheta + \theta^{j+1}$.

4. Devolver $mediatheta/m_2$. ∎

11.10.2 El Algoritmo EM para Datos Incompletos

El algoritmo EM consta de dos etapas. La primera es la etapa de cálculo de los valores esperados, en la que se calcula la esperanza de los datos incompletos o funciones de ellos. La segunda etapa es una etapa de maximización, en la que se maximiza una cierta función. Las dos etapas se iteran hasta conseguir la convergencia del proceso, que está garantizado bajo ciertas condiciones de regularidad (véase Dempster, Laird y Rubin (1977)).

Seguidamente se ilustra el método usando redes Bayesianas multinomiales. Para ello, se inicia θ a algún valor θ^1, y para la iteración k-ésima se tiene

1. **Etapa de cálculo de esperanzas:** Calcular la esperanza de N_{ijk} usando la expresión

$$E[N_{ijk}|\theta^k] = \sum_{\ell=1}^{N} \sum_{y_\ell} p(X_i = k, (\pi_i^j|S_\ell)|\theta^k, D), \tag{11.46}$$

   donde $\pi_i^j|S_\ell$, lo que significa que en π_i^j se han asignado a los datos completos sus correspondientes valores muestrales en la realización j-ésima de Π_i, y y_ℓ es el conjunto de valores desconocidos (incompletos) de S_ℓ. Los términos en (11.46) se calculan como sigue:

   - Si X_i y sus padres son conocidas en S_ℓ, $p(X_i = k, (\pi_i^j|S_\ell)|\theta, D)$ se hace cero o uno dependiendo de si los valores muestrales coinciden con los de dicho término.
   - En otro caso, se usa un algoritmo de inferencia de redes Bayesianas para evaluar dicho término.

2. **Etapa de maximización:** Se obtienen los valores de θ^k que maximizan

$$p(\theta|S, D), \tag{11.47}$$

lo que conduce a

$$\theta_{ijk}^{k} = \frac{E[N_{ijk}|\theta^{k-1}] + \eta_{ijk}}{E[N_{ik}|\theta^{k-1}] + \eta_{ik}}. \tag{11.48}$$

Por ello, se elige inicialmente algún valor de θ^1, y se repite este proceso hasta su convergencia para obtener las estimaciones finales de θ. Nótese que, en vez de los estimadores estándar de la media posterior para los parámetros θ, se usa su moda posterior.

Apéndice al Capítulo 11: Estadística Bayesiana

El material que se presenta en este capítulo requiere un conocimiento previo de la teoría de la estadística Bayesiana. En este apéndice se introducen los conceptos más importantes de esta teoría. Para un tratamiento más completo del tema se remite al lector a otros trabajos como los de DeGroot (1970), Press (1992), o Bernardo y Smith (1994).

Supóngase que una variable aleatoria X pertenece a alguna familia paramétrica $f(x|\theta)$ que depende de un conjunto de parámetros θ. En la estadística Bayesiana se supone que θ es una variable aleatoria con función de densidad $g(\theta;\eta)$, donde η son parámetros desconocidos. La función de densidad $g(\theta;\eta)$ se llama *función de probabilidad "a priori"* del parámetro θ. Para mejorar el conocimiento sobre θ, se obtiene una muestra (los datos x) a partir de $f(x|\theta)$ mediante un proceso de muestreo. Basándose en la distribución "a priori" $g(\theta;\eta)$ y en la muestra x, se obtiene la *distribución de probabilidad "a posteriori"*, $p(\theta|x,\eta)$, de θ. Bajo las hipótesis anteriores, la variable aleatoria X es una variable aleatoria mixta y puede usarse el teorema de la probabilidad total para obtener su distribución

$$u(x;\eta) = \int f(x|\theta)g(\theta;\eta)d\theta. \tag{11.49}$$

Una etapa clave en la estadística Bayesiana es la selección de la distribución "a priori" $g(\theta;\eta)$. Esta distribución debe seleccionarse sin el conocimiento de los datos; sin embargo, la información contenida en los datos previos puede incorporarse en la distribución "a priori". Un problema de la estadística Bayesiana es que a menos que la distribución "a priori" sea seleccionada cuidadosamente, la distribución "a posteriori" resultante puede no pertenecer a ninguna familia conocida y el tratamiento matemático consiguiente complicarse notablemente. Para resolver este problema, una posibilidad consiste en elegir la distribución "a priori" entre la clase de *distribuciones conjugadas*. Otra posibilidad consiste en elegir *distribuciones posteriores convenientes*. Ambas alternativas se describen a continuación.

11.11.1 Distribuciones Conjugadas

Definición 11.4 Distribuciones conjugadas. *Si las distribuciones "a priori" y "a posteriori" pertenecen a la misma familia de distribuciones para un mismo proceso muestral, se dice que son* conjugadas naturales *con respecto al mismo.*

Por ello, la propiedad más importante de las familias conjugadas es que son *cerradas* con respecto al muestreo, es decir, si se usa una muestra para actualizar la distribución "a priori" en una familia conjugada, la distribución posterior resultante también pertenece a la misma familia. La idea de utilizar familias de distribuciones paramétricas conjugadas con respecto a un proceso muestral fue propuesta por Raiffa y Schlaifer (1961).

Los beneficios derivados de la conjugación incluyen la facilidad de tratamiento matemático. Se obtienen fórmulas cerradas para muchos de los estadísticos usados en inferencia Bayesiana. De hecho, se pueden obtener fórmulas muy simples que dan los parámetros de la distribución posterior en función de la distribución "a priori" y de los datos. Consecuentemente, la obtención de la distribución posterior se limita a evaluar estas fórmulas en vez de utilizar la fórmula de Bayes. Cuando se usa una familia de distribuciones conjugada, la distribución posterior resulta

$$g(\theta, h(\eta; x)) \propto g(\theta, \eta) f(x, \theta),$$

donde $f(x, \theta)$ es la función de verosimilitud, $g(\theta; \eta)$ es una familia de distribuciones, $h(\eta; x)$ da los parámetros de la posterior en función de los parámetros de la distribución "a priori" y de la muestra x, y los parámetros η y θ son cantidades vectoriales. A los parámetros η se les llama *hiperparámetros* para distinguirlos de los parámetros θ.

Las distribuciones utilizadas con más frecuencia pertenecen a una clase de distribuciones conocida como la *familia exponencial*. La función de densidad de una variable aleatoria multidimensional X que pertenezca a la familia exponencial puede escribirse en la forma

$$f(x; \theta) = \exp \left[\sum_{j=0}^{m} g_j(\theta) T_j(x) \right], \tag{11.50}$$

donde $T_j(x)$ es una función de x y $g_j(\theta)$ es una función de los m parámetros θ; ver, por ejemplo, Bickel y Doksum (1977) y Brown (1986). Ejemplos típicos de distribuciones de probabilidad que pertenecen a la familia exponencial son la normal, la binomial, la de Poisson, y la exponencial. En esta sección se presentan ejemplos de familias conjugadas que pertenecen a la familia exponencial.

Se comienza con el teorema siguiente, debido a Arnold, Castillo y Sarabia (1993), que da las familias exponenciales más generales que son conjugadas a una familia exponencial dada.

Teorema 11.6 Distribuciones "a priori" conjugadas para verosimilitudes exponenciales. *Supóngase que la verosimilitud puede expresarse en la forma correspondiente a la familia exponencial,*

$$f(x;\theta) = \exp\left[Na(\theta) + \sum_{j=0}^{m} g_j(\theta)T_j(x)\right],\ m < t, \tag{11.51}$$

donde por convenio $g_0(\theta) = 1$, $\{T_j(x); j = 0, 1, \ldots, m\}$ *es un conjunto de funciones linealmente independientes,* N *es el tamaño de muestra, y* $a(\theta)$ *es una función de* θ*. Entonces, la familia exponencial* t*-paramétrica más general de distribuciones "a priori" para* $\theta = (\theta_1, \ldots, \theta_m)$ *que es conjugada con respecto a (11.51) es*

$$q(\theta;\eta) = \exp\left[\nu(\eta) + u(\theta) + \sum_{j=1}^{m} \eta_i g_i(\theta) + \eta_{m+1}a(\theta) + \sum_{j=m+2}^{t} \eta_i s_i(\theta)\right], \tag{11.52}$$

donde $s_{m+2}(\theta), \ldots, s_t(\theta)$*;* $\nu(\eta)$*, y* $u(\theta)$ *son funciones arbitrarias de* θ*; y* $\eta_1, \ldots, \eta_m$ *son los hiperparámetros. El vector de hiperparámetros de la distribución posterior es*

$$(\eta_1 + T_1(x), \ldots, \eta_m + T_m(x), \eta_{m+1} + N, \eta_{m+2}, \ldots, \eta_t), \tag{11.53}$$

donde los hiperparámetros $\eta_{m+1}, \ldots, \eta_t$ *no dependen de los datos de la muestra.*

Nótese que (11.53) da los hiperparámetros posteriores en función de los hiperparámetros "a priori" y de los valores muestrales. En lo que sigue se dan importantes ejemplos de aplicaciones de este teorema.

Ejemplo 11.4 Conjugada de un proceso muestral de Bernouilli. Supóngase que X es una variable aleatoria de Bernoulli que toma los valores $X = 1$ con probabilidad θ y $X = 0$ con probabilidad $1 - \theta$. Supóngase también que se obtiene una muestra aleatoria simple $x = (x_1, \ldots, x_N)$ de tamaño N. Sea $r = \sum_{i=1}^{N} x_i$, es decir, el número de unos en la muestra. Entonces, r es una variable aleatoria binomial $B(N, \theta)$, y la función de verosimilitud resulta

$$f(x;\theta) = \exp\left\{\log\left(\begin{array}{c} N \\ r \end{array}\right) + r\log\left(\frac{\theta}{1-\theta}\right) + N\log(1-\theta)\right\},$$

donde

$$\left(\begin{array}{c} N \\ r \end{array}\right) = \frac{N!}{r!(N-r)!}.$$

Identificando términos con (11.51) se obtiene

$$\begin{aligned} a(\theta) &= \log(1-\theta), \\ T_0(x) &= \log \left(\begin{array}{c} N \\ r \end{array} \right), \\ T_1(x) &= r, \\ g_1(\theta) &= \log \left(\tfrac{\theta}{1-\theta} \right), \end{aligned}$$

y sustituyendo en (11.52) conduce a

$$q(\theta, \eta) = \exp \left\{ \nu(\eta) + u(\theta) + \sum_{i=3}^{t} \eta_i s_i(\theta) + \eta_1 \log \left(\tfrac{\theta}{1-\theta} \right) + \eta_2 \log(1-\theta) \right\},$$

que incluye, pero no se limita a, la distribución *Beta*. Esto muestra que la distribución Beta es conjugada natural del proceso de muestreo de Bernoulli anterior. Se deduce de (11.53) que los hiperparámetros posteriores en este caso son

$$(\eta_1 + r, \eta_2 + N),$$

lo que muestra que una distribución "a priori" $Beta(\eta_1, \eta_2)$ conduce a una distribución "a posteriori" $Beta(\eta_1 + r, \eta_2 + N)$. ∎

Ejemplo 11.5 Conjugada de un proceso muestral de Poisson. Considérese un proceso de *Poisson* con intensidad θ sucesos por unidad de tiempo. Se cuenta el número de sucesos X que ocurren en un periodo de duración N unidades temporales. Entonces X es una variable aleatoria de Poisson con media $N\theta$. La función de verosimilitud en este caso es

$$f(x; \theta) = \exp \left\{ -N\theta + \log \left(\frac{N^x}{x!} \right) + x \log \theta \right\}.$$

Identificando términos con (11.51) se obtiene

$$\begin{aligned} a(\theta) &= -\theta, \\ T_0(x) &= \log \left(\tfrac{N^x}{x!} \right), \\ T_1(x) &= x, \\ g_1(\theta) &= \log \theta, \end{aligned}$$

y sustituyendo en (11.52) se llega a

$$q(\theta, \eta) = \exp \left\{ \nu(\eta) + u(\theta) + \sum_{i=3}^{t} \eta_i s_i(\theta) + \eta_1 \log \theta - \eta_2 \theta \right\},$$

que incluye pero no se limita a, la distribución *gamma*. Esto muestra que la distribución gamma es la conjugada natural del proceso de Poisson.

La fórmula para los hiperparámetros (11.53) muestra que a una distribución de probabilidad "a priori" $Gamma(\eta_1, \eta_2)$ le corresponde una distribución de probabilidad "a posteriori" $Gamma(\eta_1 + x, \eta_2 + N)$. ∎

Ejemplo 11.6 Conjugada de un proceso de muestreo normal. Supóngase que X es una variable aleatoria normal univariada $N(\theta, \sigma^2)$ con media desconocida θ y varianza conocida σ^2. Se obtiene una muestra aleatoria simple $(x_1, \ldots, x_N)$ a partir de $N(\theta, \sigma^2)$. Entonces, la función de verosimilitud resulta

$$f(x;\theta) = \exp\left\{ -\log\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{N} \frac{x_i^2}{2\sigma^2} - N\frac{\theta^2}{2\sigma^2} + \theta \sum_{i=1}^{N} \frac{x_i}{\sigma^2} \right\}.$$

Identificando términos con (11.51) se obtiene

$$\begin{aligned} a(\theta) &= -\tfrac{\theta^2}{2\sigma^2}, \\ T_0(x) &= -\log\left(\sigma\sqrt{2\pi}\right) - \sum_{i=1}^{N} \tfrac{x_i^2}{2\sigma^2}, \\ T_1(x) &= \sum_{i=1}^{N} \tfrac{x_i}{\sigma^2}, \\ g_1(\theta) &= \theta, \end{aligned}$$

y sustituyendo en (11.52) se llega a

$$q(\theta;\eta) = \exp\left\{ \nu(\eta) + u(\theta) + \sum_{i=3}^{t} \eta_i s_i(\theta) + \eta_1\theta - \eta_2 \tfrac{\theta^2}{2\sigma^2} \right\},$$

que incluye, pero no se limita a, la distribución normal $N(\eta_1\sigma^2/\eta_2, \sigma^2/\eta_2)$. Esto muestra que la distribución normal es la conjugada natural del proceso normal de muestreo anterior.

De la expresión (11.53) se deduce fácilmente que a una distribución "a priori" normal $N(\eta_1\sigma^2/\eta_2, \sigma^2/\eta_2)$ le corresponde una distribución "a posteriori" normal $N((\eta_1\sigma^2 + N\bar{x})/(\eta_2 + N), \sigma^2/(\eta_2 + N))$. ∎

Ejemplo 11.7 Conjugada de un proceso de muestreo multinomial (I). Sea $(x_1, \ldots, x_k)$ una muestra aleatoria simple procedente de una distribución *multinomial* $M(N; p_1, \ldots, p_{k-1})$, donde $\sum_{i=1}^{k} x_i = N$. Entonces, la función de verosimilitud es

$$f(x;\theta) = \exp\left\{ \log\left(\tfrac{N!}{x_1!\ldots x_k!}\right) + \sum_{i=1}^{k-1} x_i \log\theta_i + \left(N - \sum_{i=1}^{k-1} x_i\right) \log\left(1 - \sum_{i=1}^{k-1} \theta_i\right) \right\}.$$

Identificando términos con (11.51) se obtiene

$$\begin{aligned} a(\theta) &= \log(1 - \theta_1 - \ldots - \theta_{k-1}), \\ T_0(x) &= \log\left(\tfrac{N!}{x_1!\ldots x_k!}\right), \\ T_j(x) &= x_j;\ j = 1, \ldots k-1, \\ g_j(\theta) &= \log\left(\tfrac{\theta_j}{1-\theta_1-\ldots-\theta_{k-1}}\right);\ j = 1, \ldots, k-1, \end{aligned}$$

y sustituyendo en (11.52) conduce a

$$q(\theta;\eta) = \exp\left\{\nu(\eta) + u(\theta) + \sum_{i=k+1}^{t} \eta_i s_i(\theta)\right\} \times \exp\left\{\sum_{i=1}^{k-1} \eta_i \log\left(\frac{\theta_i}{1-\theta_1-\ldots-\theta_{k-1}}\right) + \eta_k \log(1-\theta_1-\ldots-\theta_{k-1})\right\},$$

que incluye, pero no se limita a, la distribución de *Dirichlet*. Esto muestra que la distribución de Dirichlet es la conjugada natural del proceso de muestreo multinomial.

El resultado que se muestra en (11.53) prueba que a una distribución "a priori" $Dirichlet(k;\eta_1,\ldots,\eta_k)$ le corresponde una distribución "a posteriori" $Dirichlet(k;\eta_1+x_1,\ldots,\eta_{k-1}+x_{k-1},\eta_k+N)$. ∎

Ejemplo 11.8 Conjugada de un proceso de muestreo multinomial (II). Considérese de nuevo el proceso de muestreo multinomial del Ejemplo 11.7, pero con una nueva reparametrización. Los nuevos parámetros son

$$\pi_i = \log\left(\frac{\theta_i}{1-\theta_1-\ldots-\theta_{k-1}}\right); i = 1,\ldots,k-1.$$

Con esta reparametrización, la función de verosimilitud resulta

$$\begin{aligned} f(x;\pi) &= \exp\{-N\log(1+\exp(\pi_1)+\ldots+\exp(\pi_{k-1})) + \log(N!)\} \\ &\quad \times \exp\left\{-\log\left(x_1!\ldots x_{k-1}!\left(N-\sum_{i=1}^{k-1} x_i\right)!\right) + \sum_{i=1}^{k-1}\pi_i x_i\right\}. \end{aligned}$$

Identificando términos con (11.51) se obtiene

$$\begin{aligned} &a(\pi) = -\log(1+\exp(\pi_1)+\ldots+\exp(\pi_{k-1})), \\ &T_0(x) = +\log(N!) - \log\left[x_1!\ldots x_{k-1}!\left(N-\sum_{i=1}^{k-1} x_i\right)!\right], \\ &T_j(x) = x_j;\ j = 1,\ldots,k-1, \\ &g_j(\pi) = \pi_j;\ j = 1,\ldots,k-1, \end{aligned}$$

y sustituyendo en (11.52) conduce a

$$\begin{aligned} q(\pi;\eta) &= \exp\left\{\nu(\eta) + u(\pi) + \sum_{i=k+1}^{t} \eta_i s_i(\pi)\right\} \\ &\quad \times \exp\left\{\sum_{i=1}^{k-1} \eta_i\pi_i - \eta_k \log(1+\exp(\pi_1)+\ldots+\exp(\pi_{k-1}))\right\}. \end{aligned}$$

Por tanto, a una distribución "a priori" $q(\pi;\eta_1,\ldots,\eta_k)$ le corresponde una distribución "a posteriori" $q(\pi;\eta_1+x_1,\ldots,\eta_{k-1}+x_{k-1},\eta_k+N)$ (véase (11.53)). ∎

Ejemplo 11.9 Conjugada de un proceso de muestreo multinormal. Sea $(x_1, \ldots, x_N)$ una muestra aleatoria simple procedente de una distribución multinormal $N(\mu, W^{-1})$, donde los parámetros μ y W, son el vector de medias y la matriz de precisión (la inversa de la matriz de covarianzas), respectivamente. Nótese que x_i es un vector. Entonces, la función de verosimilitud es

$$\begin{aligned} & f(x;\mu,W) = \\ & \exp\left\{\log\left[\sqrt{det(W)}\right] - k\log(2\pi)/2 - \tfrac{1}{2}\left[\sum_{r=1}^{N}\sum_{i,j=1}^{k} w_{ij}z_{ir}z_{jr}\right]\right\} \\ & \qquad = \exp\left\{\log\left[\sqrt{det(W)}\right] - k\log(2\pi)/2 - \tfrac{N}{2}\sum_{i,j=1}^{k} w_{ij}\mu_i\mu_j\right\} \\ & \qquad \times \exp\left\{\sum_{i=1}^{k}\left(\sum_{j=1}^{k} w_{ij}\mu_j\right)\sum_{r=1}^{N} x_{ir} - \tfrac{1}{2}\sum_{i,j=1}^{k} w_{ij}\sum_{r=1}^{N} x_{ir}x_{jr}\right\}. \end{aligned}$$

donde $z_{ir} = x_{ir} - \mu_i$. Identificando términos con (11.51) y utilizando índices dobles se obtiene

$$\begin{aligned} a(\theta) &= -\tfrac{1}{2}\sum_{i,j=1}^{k} w_{ij}\mu_i\mu_j, \\ W_0(x) &= -k\log(2\pi)/2 + \log\left(\sqrt{det(W)}\right), \\ W_i(x) &= \sum_{r=1}^{N} x_{ir};\ i = 1, \ldots k, \\ W_{ij}(x) &= -\tfrac{1}{2}\sum_{r=1}^{N} x_{ir}x_{jr};\ i,j = 1, \ldots k, \\ g_i(\mu, W) &= \sum_{j=1}^{k} w_{ij}\mu_j;\ i = 1, \ldots k, \\ g_{ij}(\mu, W) &= w_{ij};\ i,j = 1, \ldots k, \end{aligned}$$

y sustituyendo en (11.52) conduce a

$$\begin{aligned} & q(\mu, W; \eta) = \exp\left\{\nu(\eta) + u(\mu, W) + \sum_{i=k+k^2+2}^{t} \eta_i s_i(\mu, W)\right\} \\ & \times \exp\left\{\sum_{i=1}^{k}\eta_i\left(\sum_{j=1}^{k} w_{ij}\mu_j\right) + \sum_{i,j=1}^{k}\eta_{ij}w_{ij} + \eta_{k+k^2+1}\left[-\tfrac{1}{2}\sum_{i,j=1}^{k} w_{ij}\mu_i\mu_j\right]\right\}, \end{aligned}$$

que incluye, pero no se limita a, la distribución *normal-Wishart*.

Por tanto, con una reparametrización, si la distribución "a priori" es una normal-Wishart$(\nu, \mu_0, \alpha, W_0)$ con función de densidad

$$\begin{aligned} & g(\mu, W; \nu, \mu_0, \alpha, W_0) \propto \\ & det(W)^{(\alpha-k)} \exp\left\{-\tfrac{1}{2}\left[tr(W_0W) + \sum_{i,j=1}^{k} \nu w_{ij}(\mu_i - \mu_{0i})(\mu_j - \mu_{0j})\right]\right\}, \end{aligned}$$

donde $tr(A)$ es la traza de la matriz A (la suma de los elementos de la diagonal principal de A), entonces la distribución "a posteriori" es una normal-Wishart$(\nu + N, \mu_N, \alpha + N, W_N)$, donde

$$\begin{aligned}
&\bar{x} = \tfrac{1}{N} \sum_{i=1}^{N} x_i, \\
&\mu_N = \tfrac{\nu\mu_0 + N\bar{x}}{\nu + N}, \\
&S = \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T, \\
&W_N = W_0 + S + \tfrac{\nu N}{\nu + N}(\mu_0 - \bar{x})(\mu_0 - \bar{x})^T.
\end{aligned}$$

Por tanto, la distribución normal-Wishart es la conjugada natural del proceso de muestreo normal multivariado. ∎

Ejemplo 11.10 Conjugada de un proceso muestral de Dirichlet. Considérese una muestra aleatoria simple $(x_1, \ldots, x_N)$ procedente de una distribución de Dirichlet con parámetros $\theta_1, \ldots, \theta_{k+1}$. Nótese que x_i es un vector k-dimensional. La función de densidad de la distribución de Dirichlet es

$$f(x_1, \ldots, x_k; \theta_1, \ldots, \theta_{k+1}) = \frac{\Gamma(\sum_{j=1}^{k+1} \theta_j)}{\prod_{j=1}^{k+1} \Gamma(\theta_j)} \left(\prod_{j=1}^{k} x_j^{\theta_j - 1} \right) \left(1 - \sum_{j=1}^{k} x_j \right)^{\theta_{k+1} - 1},$$

donde $\Gamma(x)$ es la función gamma. Nótese que si x es un entero, se tiene $\Gamma(x) = (x-1)!$. Los parámetros $\theta_i, i = 1, \ldots, k$ son proporcionales a los valores medios de $X_1, \ldots, X - k$ y tales que las varianzas decrecen con su suma. Entonces la verosimilitud de la muestra es

$$L = \prod_{i=1}^{N} f(x_i; \theta) = \prod_{i=1}^{N} \left[\frac{\Gamma\left(\sum_{j=1}^{k+1} \theta_j\right)}{\prod_{j=1}^{k+1} \Gamma(\theta_j)} \left(\prod_{j=1}^{k} x_{ij}^{\theta_j - 1} \right) \left(1 - \sum_{j=1}^{k} x_{ij} \right)^{\theta_{k+1} - 1} \right],$$

que puede escribirse en la forma de (11.51) como

$$\begin{aligned}
L \;=\; & \exp\left[N \log\left[\Gamma\left(\sum_{j=1}^{k+1} \theta_j \right) - \sum_{j=1}^{k+1} \log \Gamma(\theta_j) \right] \right. \\
& \left. + \sum_{j=1}^{k} (\theta_j - 1) \sum_{i=1}^{N} \log x_{ij} + (\theta_{k+1} - 1) \sum_{i=1}^{N} \log\left(1 - \sum_{j=1}^{k} x_{ij} \right) \right].
\end{aligned} \tag{11.54}$$

Entonces, se tiene

$$
\begin{aligned}
& m = k+1, \\
& a(\theta) = \log\Gamma\left(\sum_{j=1}^{k+1}\theta_j\right) - \sum_{j=1}^{k+1}\log\Gamma(\theta_j), \\
& g_j(\theta) = \theta_j, \;\; j = 1,\ldots,k, \\
& g_{k+1}(\theta) = \theta_{k+1}, \\
& T_0(x) = -\sum_{j=1}^{k}\sum_{i=1}^{N}\log x_{ij} - \sum_{i=1}^{N}\log\left(1-\sum_{j=1}^{k}x_{ij}\right), \\
& T_j(x) = \sum_{j=1}^{N}\log x_{ij}, \;\; j = 1,\ldots,k, \\
& T_{k+1}(x) = \sum_{i=1}^{N}\log\left(1-\sum_{j=1}^{k}x_{ij}\right).
\end{aligned}
$$

Se deduce de (11.52) que la distribución de probabilidad "a priori" de θ, $q(\theta;\eta)$, es proporcional a

$$
\exp\left(\nu(\eta) + u(\theta) + \sum_{i=1}^{k+1}\eta_i\theta_i + \eta_{k+2}\log\frac{\Gamma\left(\sum_{j=1}^{k+1}\theta_j\right)}{\prod_{j=1}^{k+1}\Gamma(\theta_j)} + \sum_{i=k+3}^{t}\eta_i s_i(\theta)\right),
$$

donde $\nu_(\eta)$, $u(\theta)$ y $s_i(\theta)$; $i = k+3,\ldots,t$, son funciones arbitrarias. Los hiperparámetros posteriores resultan

$$
(\eta_1 + T_1(x),\ldots,\eta_{k+1} + T_{k+1}(x), \eta_{k+2} + N, \eta_{k+3},\ldots,\eta_t).
$$

Debido al hecho de que los hiperparámetros η_i, $i = k+3,\ldots,t$, son parámetros estáticos (no se alteran por la información), se puede tomar

$$
q(\theta;\eta) \propto \exp\left(\sum_{i=1}^{k+1}\eta_i\theta_i + \eta_{k+2}\log\Gamma\left(\sum_{j=1}^{k+1}\theta_j\right) - \eta_{k+2}\sum_{j=1}^{k+1}\log\left(\Gamma(\theta_j)\right)\right). \tag{11.55}
$$

Puesto que la media de la distribución conjugada en (11.55) es difícil de obtener analíticamente, se puede usar la moda de (11.55) para estimar los parámetros de Dirichlet. La moda puede ser obtenida maximizando (11.55) con respecto a θ. ∎

11.11.2 Distribuciones "a posteriori" Convenientes

Cuando se selecciona una familia de distribuciones de probabilidad "a priori" para combinar con la verosimilitud, la principal consideración es que la

distribución posterior resultante pertenezca a una familia conocida y fácil de tratar. Por ello, no es necesario que ambas, la distribución "a priori" y la distribución "a posteriori", pertenezcan a la misma familia. Es suficiente con que las distribuciones posteriores pertenezcan a una familia conocida, como la exponencial. A estas distribuciones las llamamos distribuciones "a posteriori" *convenientes*.

El teorema siguiente debido a Arnold, Castillo y Sarabia (1994) identifica la familia de distribuciones "a priori" que conducen a familias convenientes.

Teorema 11.7 Distribuciones posteriores convenientes *Considérese una muestra aleatoria simple de N observaciones posiblemente vectoriales $x_1, \ldots, x_N$ procedentes de una familia exponencial m-paramétrica de la forma*

$$f(x;\theta) = \exp\left[a(\theta) + \sum_{j=0}^{m} \theta_j T_j(x)\right], \tag{11.56}$$

donde $\theta_0 = 1$. Entonces, la clase de distribuciones "a priori" más general sobre Θ que conduce a una distribución de probabilidad para θ que pertenece a la familia t-paramétrica exponencial de la forma

$$f(\theta,\eta) = \exp\left[\nu(\eta) + \sum_{s=0}^{t} \eta_s g_s(\theta)\right],\ t \leq m,$$

donde $\eta_0 = 1$, es de la forma

$$f(\theta;c) = \exp\left[c_{00} - a(\theta) + g_0(\theta) + \sum_{i=1}^{t} c_{i0} g_i(\theta)\right], \tag{11.57}$$

y

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \cdots \\ \theta_m \end{bmatrix} = \begin{bmatrix} c_{01} & c_{11} & \ldots & c_{t1} \\ c_{02} & c_{12} & \ldots & c_{t2} \\ \ldots & \ldots & \ldots & \ldots \\ c_{0m} & c_{1m} & \ldots & c_{tm} \end{bmatrix} \begin{bmatrix} g_0(\theta) \\ g_1(\theta) \\ g_2(\theta) \\ \cdots \\ g_t(\theta) \end{bmatrix}. \tag{11.58}$$

Los hiperparámetros resultan

$$\begin{bmatrix} \eta_1(x) \\ \eta_2(x) \\ \cdots \\ \eta_t(x) \end{bmatrix} = \begin{bmatrix} c_{10} & c_{11} & \ldots & c_{1m} \\ c_{20} & c_{21} & \ldots & c_{2m} \\ \ldots & \ldots & \ldots & \ldots \\ c_{t0} & c_{t1} & \ldots & c_{tm} \end{bmatrix} \begin{bmatrix} T_0(x) \\ T_1(x) \\ T_2(x) \\ \cdots \\ T_m(x) \end{bmatrix}. \tag{11.59}$$

Los coeficientes $\{c_{ij}; i = 0, \ldots t; j = 0, \ldots, m\}$ deben elegirse para que la función (11.57) sea integrable.

Además estos autores han encontrado que este problema tiene solución sólo en la familia exponencial. Sin embargo, nótese que (11.58) impone severas restricciones sobre las distribuciones posteriores candidatas.

Ejercicios

11.1 Para cada uno de los 11 grafos dirigidos de la Figura 11.2

(a) Escribir la función de probabilidad conjunta en forma factorizada tal como la sugiere el grafo.

(b) Escribir los parámetros correspondientes usando la notación $\theta_{ij\pi} = p(X_i = j|\Pi_i = \pi)$.

(c) Encontrar la dimensión de cada una de las redes Bayesianas implicadas por el grafo y sus correspondientes funciones de probabilidad condicionales.

11.2 Considérese la red Bayesiana definida por el grafo dirigido completo de la Figura 11.2 (grafo número 11) y su función de probabilidad conjunta

$$p(x_1, x_2, x_3) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2). \tag{11.60}$$

Denótese este grafo por D_{11}. Supóngase que todas las variables son binarias.

(a) Mostrar que la función de probabilidad conjunta en (11.60) depende de siete parámetros libres.

(b) Dar un ejemplo numérico de distribución de probabilidad condicional en (11.60) asignando a cada uno de los parámetros un valor a elegir.

(c) Sea θ_k el k-ésimo parámetro. Supóngase que se especifica una distribución de probabilidad "a priori" uniforme $p(\theta_1, \ldots, \theta_7|D_c)$, es decir, $p(\theta_1, \ldots, \theta_7|D_c) = 1; 0 \leq \theta_k \leq 1, k = 1, \ldots, 7$. Supóngase independencia paramétrica y modularidad. Usar el Algoritmo 11.1 para calcular la distribución de probabilidad "a priori" para la red Bayesiana (D_5, P_5), donde D_5 es el grafo número 5 de la Figura 11.2 y P_5 es el conjunto de distribuciones condicionales sugeridas por D_5.

(d) Bajo las condiciones anteriores, usar el Algoritmo 11.1 para calcular la distribución de probabilidad para la red Bayesiana (D_8, P_8), donde D_8 es el grafo número 8 de la Figura 11.2 y P_8 es el conjunto de distribuciones condicionales sugeridas por D_8.

11.3 Repetir el ejercicio previo cuando X_1 y X_2 son binarias pero X_3 es ternaria.

11.4 Considérese la red Bayesiana definida por el grafo número 10 de la Figura 11.2 y las distribuciones de probabilidad condicionales de la Tabla 11.11. Nótese que una de las variables es ternaria y las otras dos son binarias.

x_1	$p(x_1)$
0	0.7
1	0.3

x_2	$p(x_2)$
0	0.6
1	0.4

x_1	x_2	x_3	$p(x_3\|x_1, x_2)$
0	0	0	0.1
0	0	1	0.6
0	0	2	0.3
0	1	0	0.4
0	1	1	0.2
0	1	2	0.4
1	0	0	0.7
1	0	1	0.1
1	0	2	0.2
1	1	0	0.1
1	1	1	0.1
1	1	2	0.8

TABLA 11.11. Un conjunto de distribuciones de probabilidad condicionales.

(a) ¿Cuál es la dimensión de esta red Bayesiana?

(b) Calcular la frecuencia de cada posible realización de las tres variables en una muestra perfecta de tamaño 1000.

11.5 Supóngase una distribución de probabilidad que asigna la misma probabilidad a cada uno de los parámetros libres de una red Bayesiana. Usar la muestra perfecta de tamaño 1000 del ejercicio anterior para calcular las siguientes medidas de calidad de cada una de las 11 estructuras de la Figura 11.2 y comparar las medidas mediante un análisis de los resultados obtenidos:

(a) La medida de Geiger y Heckerman (11.23) con $\eta = 160$.

(b) La medida de Geiger y Heckerman (11.23) con $\eta = 8$.

(c) La medida de Cooper y Herskovits (11.24).

(d) La medida Bayesiana estándar (11.27).

(e) La medida Bayesiana estándar asintótica (11.28).

(f) La medida de mínima longitud de descripción (11.42).

(g) La medida de información (11.43) para cualquier función de penalización $f(N)$ a elegir.

11.6 Los datos de la Tabla 11.12 han sido generados a partir de una distribución normal trivariada. Con las hipótesis de la Sección 11.5.1 y

Caso	X_1	X_2	X_3	Caso	X_1	X_2	X_3
1	4.3	10.3	16.0	16	4.8	10.3	16.3
2	6.5	10.8	16.3	17	4.5	6.2	10.5
3	4.3	8.6	13.1	18	4.8	14.2	17.8
4	3.0	10.5	14.6	19	4.3	7.6	12.5
5	4.9	8.9	14.4	20	5.4	10.5	15.1
6	5.5	8.3	14.0	21	3.3	11.4	15.1
7	5.2	12.6	17.4	22	6.0	9.8	16.2
8	2.9	9.8	13.5	23	5.9	8.6	15.7
9	6.2	9.3	17.2	24	4.4	6.5	9.5
10	5.7	10.1	16.4	25	3.4	7.2	11.5
11	4.6	12.1	16.6	26	4.2	7.6	12.1
12	4.4	13.8	16.9	27	5.2	11.4	15.4
13	5.7	8.8	13.2	28	5.2	11.4	17.0
14	5.6	9.2	16.3	29	2.2	9.4	10.9
15	2.8	8.4	11.4	30	4.6	9.0	11.3

TABLA 11.12. Una muestra procedente de una distribución normal trivariada.

la distribución de probabilidad "a priori" de los parámetros dados en el Ejemplo 11.3, calcular las medidas de Geiger y Heckerman en (11.37) para cada una de las 11 estructuras de la Figura 11.2.

11.7 Supóngase una distribución de probabilidad "a priori" que asigna probabilidades idénticas a cada uno de los parámetros libres de una red Bayesiana. Considérese la muestra perfecta de tamaño 1000 generada a partir de la función de probabilidad conjunta definida por las distribuciones de probabilidad condicionales de la Tabla 11.11. Encontrar la red Bayesiana con la máxima calidad usando los algoritmos de búsqueda siguientes:

(a) El algoritmo $K2$ usando la ordenación natural X_1, X_2, X_3.

(b) El algoritmo $K2$ usando la ordenación invertida X_3, X_2, X_1.

(c) El algoritmo B.

Capítulo 12
Ejemplos de Aplicación

12.1 Introducción

En este capítulo se aplica la metodología presentada en los capítulos anteriores a tres casos de la vida real:

- El problema del tanque de presión (Sección 12.2).
- El problema del sistema de distribución de energía (Sección 12.3).
- El problema de daño en vigas de hormigón armado (Secciones 12.4 y 12.5).

Con estos tres ejemplos se ilustra y refuerza el conocimiento de las etapas que deben seguirse cuando se modelan casos reales con los diferentes modelos probabilísticos que se han introducido en los capítulos anteriores.

Tal como cabe esperar, estas aplicaciones son más complicadas que los simples ejemplos que se han utilizado en este libro para ilustrar ciertos métodos. Por otra parte, muchas de las hipótesis que se hacen para simplificar las cosas no suelen verificarse en la práctica. Por ejemplo:

- Las variables pueden ser discretas (binarias, categóricas, etc.), continuas, o mixtas (algunas discretas y otras continuas).
- Las relaciones entre las variables pueden ser muy complicadas y el modelo puede incluir retroalimentación o ciclos. Como consecuencia, la especificación de los modelos probabilísticos puede ser difícil y dar problemas.

- La propagación de evidencia puede requerir mucho tiempo, debido al gran número de parámetros y a la complejidad de las estructuras de la red.

En los tres ejemplos que se presentan en este capítulo aparecen muchos de los problemas anteriores.

En algunas situaciones prácticas, las relaciones entre las variables del modelo son claras y permiten definir el modelo probabilístico correspondiente de una forma inmediata y fácil. En algunos casos, las redes de Markov y Bayesianas (véase el Capítulo 6) suministran metodologías apropiadas para definir un modelo probabilístico consistente, basado en una representación gráfica. En las Secciones 12.2 y 12.3 se utiliza esta metodología para definir un modelo de red probabilística para los problemas del *tanque de presión* y del *sistema de distribución de energía*. Sin embargo, en el caso general, las relaciones entre las variables de un problema dado pueden ser muy complejas, y por tanto, los modelos especificados gráficamente (véase el Capítulo 6) pueden ser inapropiados. En estas situaciones hace falta utilizar esquemas más generales para analizar las relaciones entre las variables que intervienen. Por ejemplo, en la Sección 12.4 se aplican los modelos especificados condicionalmente (véase la Sección 7.7) al problema del *daño en vigas de hormigón armado*. Se verá que esta metodología suministra modelos probabilísticos consistentes en este caso.

Todos los cálculos se han hecho utilizando los programas de ordenador escritos por los autores.[1]

12.2 El Sistema del Tanque de Presión

12.2.1 Definición del problema

La Figura 12.1 muestra un diagrama de un tanque de presión con sus elementos más importantes. Se trata de un tanque para almacenar un fluido a presión, que se introduce con la ayuda de una bomba activada por un motor eléctrico. Se sabe que el tanque no tiene problemas si la bomba funciona durante un periodo inferior a un minuto. Por tanto, se incorpora un mecanismo de seguridad, basado en un relé, F, que interrumpe la corriente tras funcionar 60 segundos. Además, un interruptor de presión, A, corta también la corriente si la presión en el tanque alcanza un cierto valor umbral, que se considera peligroso. El sistema incluye un interruptor, E, que inicia la operación del sistema; un relé, D, que suministra corriente tras la etapa de iniciación y la interrumpe tras la activación del relé F; y el C, que

[1] Estos programas pueden obtenerse en la dirección World Wide Web (WWW) http://ccaix3.unican.es/~AIGroup.

activa la operación del circuito eléctrico del motor. El objetivo del estudio consiste en conocer la probabilidad de fallo del tanque de presión.

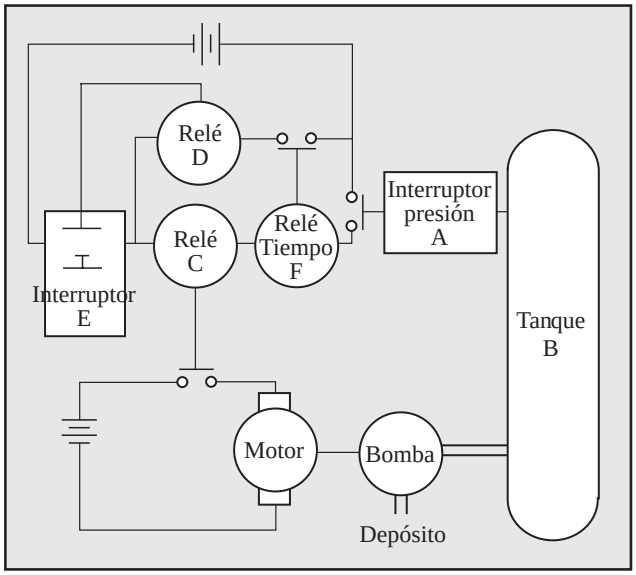


FIGURA 12.1. Un diagrama del sistema del tanque de presión.

12.2.2 Selección de Variables

Puesto que se está interesado en el análisis de todas las posibles causas de fallo del tanque B, se introduce una nueva variable K que denota este suceso. Se usará la notación $K = k$ para indicar el fallo del tanque, y $K = \bar{k}$ para el suceso complementario de no fallo. Similarmente, a, ..., f representan los fallos de las respectivas componentes A, ..., F y $\bar{a}$, ..., $\bar{f}$ representan los sucesos correspondientes al no fallo.

Basándose en la descripción previa del problema, se puede escribir la siguiente expresión lógica para el fallo del tanque:

$$\begin{aligned} k &= b \vee c \vee (a \wedge e) \vee (a \wedge d) \vee (a \wedge f) \\ &= b \vee c \vee (a \wedge (e \vee d \vee f)), \end{aligned} \tag{12.1}$$

donde los símbolos $\vee$ y $\wedge$ se usan para *o* e *y*, respectivamente. Esta expresión se obtiene combinando todas las posibilidades de fallo de las diferentes componentes que conducen al fallo del tanque. La expresión lógica alternativa para el correcto funcionamiento del tanque puede obtenerse mediante el complemento de ambos lados de la expresión (12.1). De esta forma

se obtiene

$$\begin{aligned}\bar{k} = \; & \overline{b \vee c \vee (a \wedge e) \vee (a \wedge d) \vee (a \wedge f)} \\ & = \bar{b} \wedge \bar{c} \wedge \overline{(a \wedge e)} \wedge \overline{(a \wedge d)} \wedge \overline{(a \wedge f)} \\ & = \bar{b} \wedge \bar{c} \wedge (\bar{a} \vee \bar{e}) \wedge (\bar{a} \vee \bar{d}) \wedge (\bar{a} \vee \bar{f}) \\ & = \left(\bar{b} \wedge \bar{c} \wedge \bar{a}\right) \vee \left(\bar{b} \wedge \bar{c} \wedge \bar{e} \wedge \bar{d} \wedge \bar{f}\right).\end{aligned} \tag{12.2}$$

Las ecuaciones (12.1) y (12.2) constituyen la base para obtener el conjunto de reglas (para un sistema experto basado en reglas) o la estructura de dependencia (de un sistema de red probabilística). Estas ecuaciones pueden expresarse de una forma mucho más intuitiva usando lo que se llama un *árbol de fallos*. Por ejemplo, la Figura 12.2 muestra el árbol de fallos correspondiente a la expresión (12.1). En este árbol, los fallos de los relés D y F se combinan para dar una causa de fallo intermedia, G; seguidamente G se combina con E para definir otra causa intermedia, H, y así sucesivamente. Este árbol incluye las variables iniciales $\{A, \ldots, F\}$ así como los fallos intermedios $\{G, \ldots, J\}$ que implican el fallo del tanque. Por tanto, el conjunto final de variables usadas en este ejemplo es $X = \{A, \ldots, K\}$.

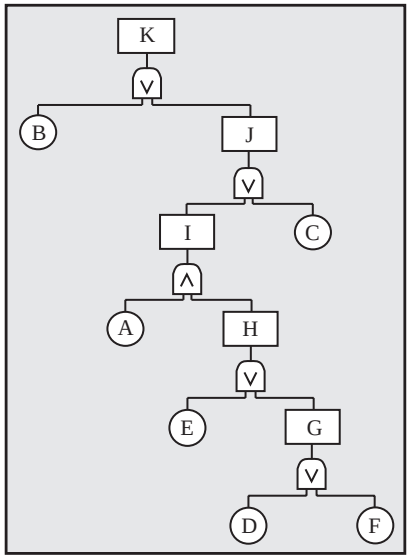


FIGURA 12.2. Árbol de fallos del sistema del tanque de presión.

12.2.3 Selección del Modelo

El ejemplo del tanque de presión definido anteriormente puede ser analizado desde un punto de vista determinista usando los sistemas expertos basados en reglas introducidos en el Capítulo 2. La Figura 12.3 muestra las reglas

que resultan del árbol de fallos de la Figura 12.2 así como el encadenamiento entre las premisas y las conclusiones de las diferentes reglas. La definición formal de la base de conocimiento (las reglas) se deja, como ejercicio, al lector.

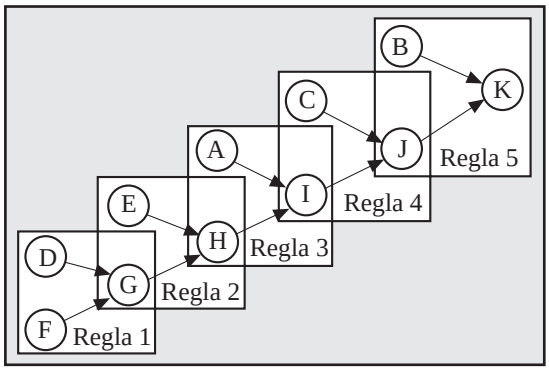


FIGURA 12.3. Reglas encadenadas para el sistema del tanque de presión.

Por ello, una vez que se dispone de cierta evidencia, los algoritmos para encadenamiento de reglas pueden ser utilizados para obtener conclusiones.

Las reglas establecen las condiciones bajo las cuales puede derivarse la verdad de los objetos conclusiones a partir de la verdad de los objetos premisas. Sin embargo, desde el punto de vista del paradigma aleatorio, los objetos pueden tener asociada una cierta medida de incertidumbre, tal como una probabilidad de ser ciertos. Además, las reglas pueden contener cierta información que permita obtener la incertidumbre de las conclusiones a partir de las incertidumbres de los objetos en las premisas de dichas reglas. De hecho, el problema clave consiste en obtener la incertidumbre asociada a algunos objetos (los objetos que figuran en las conclusiones de las reglas) cuando se conoce la de otros objetos (los objetos que figuran en la premisa de las reglas). Como se ha visto en el Capítulo 3, desde un punto de vista estadístico, se puede dar una interpretación mucho más amplia a estas reglas dando tablas de probabilidades condicionales, e incluso obteniendo fórmulas de agregación para combinar la incertidumbre de los objetos de las premisas con la incertidumbre propia de la regla, para obtener la incertidumbre de los objetos de las conclusiones. Por ejemplo, la incertidumbre de las reglas "Si A y B y C entonces D" y "Si A o B o C entonces D" puede medirse mediante la probabilidad condicional $p(d|a,b,c)$. Esto incluye las reglas deterministas con

$$p(c|a,b,c) = \begin{cases} 1, & si\ A = B = C = cierto, \\ 0, & en\ otro\ caso, \end{cases} \tag{12.3}$$

para la primera regla, y

$$p(d|a,b,c) = \begin{cases} 1, & si\ A\ o\ B\ o\ C\ es\ cierto, \\ 0, & en\ otro\ caso, \end{cases} \tag{12.4}$$

para la segunda.

El estudio de este ejemplo desde un punto de vista determinista se deja, como ejercicio, al lector[2].

La estructura dada por las reglas encadenadas puede utilizarse para definir la estructura gráfica de un sistema experto probabilístico. Por ejemplo, puesto que los fallos de las diferentes componentes del sistema son las causas de los fallos intermedios y, finalmente, del fallo del tanque, se puede obtener un grafo dirigido que reproduzca estas dependencias entre las variables que intervienen en el modelo (véase la Figura 12.4). Como se ha visto en el Capítulo 6, este grafo contiene la estructura de dependencia de la red Bayesiana. De este grafo se deduce que la función de probabilidad conjunta de todos los nodos puede escribirse en la forma

$$\begin{aligned} p(x) &= p(a)p(b)p(c)p(d)p(e)p(f)p(g|d,f)p(h|e,g) \\ &\quad p(i|a,h)p(j|c,i)p(k|b,j). \end{aligned} \tag{12.5}$$

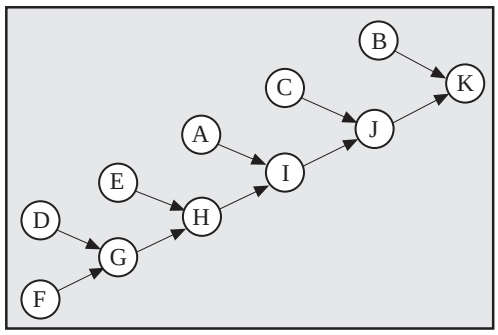


FIGURA 12.4. Grafo dirigido correspondiente al tanque de presión.

Por otra parte, las distribuciones de probabilidad condicionales asociadas a las causas intermedias en el árbol de fallos se definen utilizando (12.3) y (12.4) tal como se muestra en la Tabla 12.1, donde se dan sólo las probabilidades condicionales de los fallos, puesto que $p(no\ fallo) = 1 - p(fallo)$. Por otra parte, las probabilidades marginales asociadas a las componentes del sistema representan las probabilidades iniciales de fallo de cada una de

[2]Se invita y anima al lector a utilizar el programa *X-pert Rules* para el análisis de este problema. El programa puede obtenerse de la dirección WWW http://ccaix3.unican.es/~AIGroup.

D	F	$p(g\|D,F)$
d	f	1
d	$\bar{f}$	1
$\bar{d}$	f	1
$\bar{d}$	$\bar{f}$	0

E	G	$p(h\|E,G)$
e	g	1
e	$\bar{g}$	1
$\bar{e}$	g	1
$\bar{e}$	$\bar{g}$	0

A	H	$p(i\|A,H)$
d	h	1
a	$\bar{h}$	0
$\bar{a}$	h	0
$\bar{a}$	$\bar{h}$	0

C	I	$p(j\|C,I)$
e	i	1
c	$\bar{i}$	1
$\bar{c}$	i	1
$\bar{c}$	$\bar{i}$	0

B	J	$p(k\|B,J)$
b	j	1
b	$\bar{j}$	1
$\bar{b}$	j	1
$\bar{b}$	$\bar{j}$	0

TABLA 12.1. Probablidades condicionales de fallo de las variables intermedias en el sistema del tanque de presión.

sus componentes. Supóngase que las probabilidades son

$$\begin{array}{lll} p(a) = 0.002, & p(b) = 0.001, & p(c) = 0.003, \\ p(d) = 0.010, & p(e) = 0.001, & p(f) = 0.010. \end{array} \tag{12.6}$$

El grafo de la Figura 12.4, junto con las tablas de probabilidad que se muestran en (12.6) y en la Tabla 12.1, define una red Bayesiana que corresponde al ejemplo del tanque de presión. La correspondiente función de probabilidad conjunta se da en (12.5).

12.2.4 Propagación de Evidencia

El grafo de la Figura 12.4 es un poliárbol, lo que significa que se puede utilizar el algoritmo para poliárboles (Algoritmo 8.1) para la propagación de la evidencia. Supóngase, en primer lugar, que no hay evidencia disponible. En este caso, el Algoritmo 8.1 da las probabilidades marginales de los nodos, que se muestran en la Figura 12.5. Nótese que la probabilidad de fallo inicial del tanque es $p(k) = 0.004$.

Supóngase ahora que las componentes F y D fallan, es decir, se tiene la evidencia $F = f, D = d$. Se puede usar el Algoritmo 8.1 para propagar esta evidencia. Las nuevas probabilidades condicionales de los nodos $p(x_i|f,d)$ se muestran en la Figura 12.6. Nótese que los fallos de los relés F y D inducen el fallo de los nodos intermedios G y H, pero la probabilidad de fallo del tanque es todavía pequeña ($p(k) = 0.006$).

Para continuar la ilustración, supóngase que el interruptor de presión A también falla ($A = a$). Si se propaga la evidencia acumulada ($F = f, D = d, A = a$) usando el Algoritmo 8.1, se obtienen las nuevas probabilidades condicionales de los nodos que se muestran en la Figura 12.7. Ahora, puesto

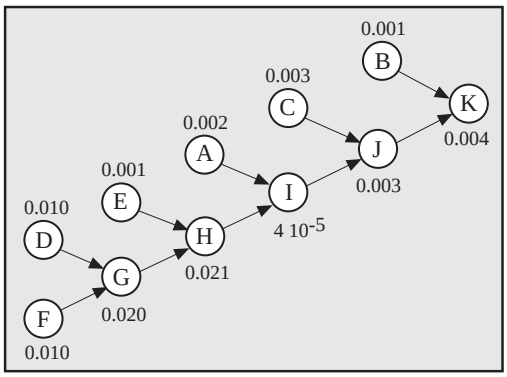


FIGURA 12.5. Las probabilidades marginales iniciales de los nodos (cuando no hay evidencia disponible) para el tanque de presión.

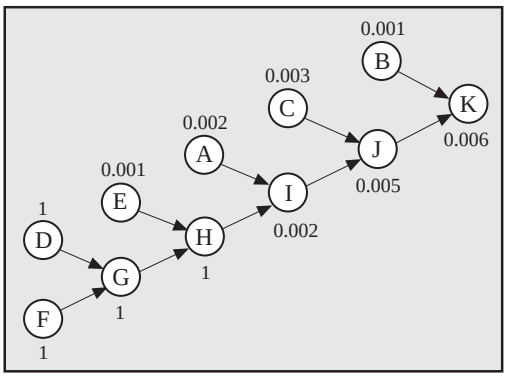


FIGURA 12.6. Probabilidades condicionales de los nodos dada la evidencia $F = f$ y $D = d$ para el sistema del tanque de presión.

que $p(k) = 1$, el fallo de estas componentes F, D y A, implican el fallo de todos los nodos intermedios y el fallo del tanque.

12.2.5 Considerando Causas Comunes de Fallo

Supóngase ahora que hay una causa común de fallo para los relés C, D y F. Por ejemplo, supóngase que estos relés han sido construidos en las mismas circunstancias. Por ello, una posibilidad consiste en dibujar nuevos enlaces entre ellos, para indicar las nuevas relaciones de dependencia en el modelo (véase Castillo y otros (1994)), tal como se muestra en la Figura 12.8, que se obtiene de la Figura 12.4 tras enlazar los nodos C, D y F. Nótese que los nodos se han reordenado para evitar que el nuevo grafo resulte confuso. Ahora, el grafo de la Figura 12.8 es un grafo múltiplemente conexo, y el algoritmo para poliárboles 8.1 ya no puede aplicarse. En este

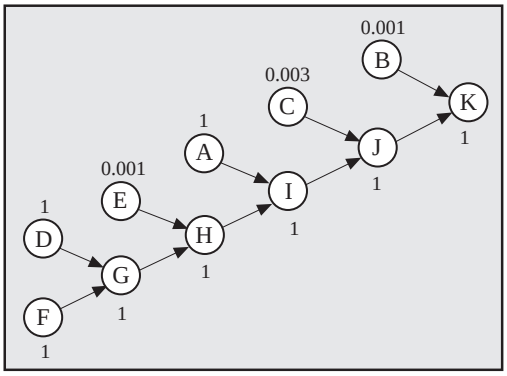


FIGURA 12.7. Probabilidades condicionales de los nodos dada la evidencia $F = f$, $D = d$ y $A = a$ para el sistema del tanque de presión.

caso, tiene que utilizarse un algoritmo de propagación más general tal como el de agrupamiento 8.5 para propagar la evidencia en un árbol de unión asociado al grafo.

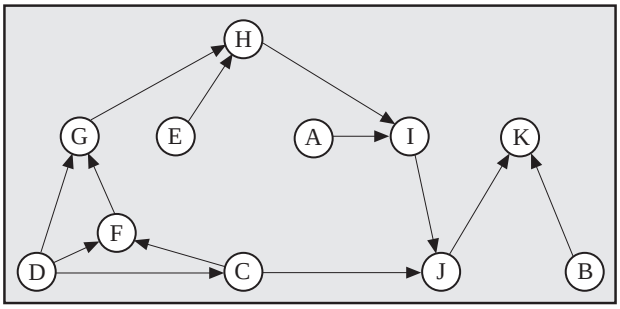


FIGURA 12.8. Grafo dirigido para el caso del tanque de presión cuando se consideran causas comunes de fallo.

Según el grafo de la Figura 12.8, la función de probabilidad conjunta de los nodos puede factorizarse en la forma

$$\begin{aligned} p(x) &= p(d)p(c|d)p(f|c,d)p(g|d,f)p(e) \\ &\quad p(h|e,g)p(a)p(i|h,a)p(j|c,i)p(b)p(k|b,j). \end{aligned} \tag{12.7}$$

Las correspondientes funciones de probabilidad condicionales se dan en la Tabla 12.2.

Para usar el Algoritmo de agrupamiento 8.5, se necesita en primer lugar moralizar y triangular el grafo de la Figura 12.8. El grafo no dirigido moralizado y triangulado resultante se muestra en la Figura 12.9.

D	$p(c\|D)$
d	0.750
$\bar{d}$	0.001

C	D	$p(f\|C,D)$
c	d	0.99
c	$\bar{d}$	0.75
$\bar{c}$	d	0.75
$\bar{c}$	$\bar{d}$	0.01

D	F	$p(g\|D,F)$
d	f	1
d	$\bar{f}$	1
$\bar{d}$	f	1
$\bar{d}$	$\bar{f}$	0

E	G	$p(h\|E,G)$
e	g	1
e	$\bar{g}$	1
$\bar{e}$	g	1
$\bar{e}$	$\bar{g}$	0

A	H	$p(i\|A,H)$
d	h	1
a	$\bar{h}$	0
$\bar{a}$	h	0
$\bar{a}$	$\bar{h}$	0

C	I	$p(j\|C,I)$
e	i	1
c	$\bar{i}$	1
$\bar{c}$	i	1
$\bar{c}$	$\bar{i}$	0

B	J	$p(k\|B,J)$
b	j	1
b	$\bar{j}$	1
$\bar{b}$	j	1
$\bar{b}$	$\bar{j}$	0

TABLA 12.2. Probabilidades de fallo para el tanque de presión cuando se consideran causas comunes de fallo.

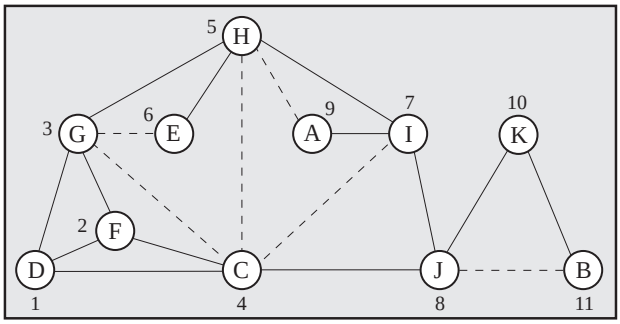


FIGURA 12.9. Un grafo moralizado y triangulado asociado al grafo dirigido de la Figura 12.8. Se muestra una numeración perfecta de los nodos.

Los conglomerados de este grafo son

$$C_1 = \{A, H, I\}, \quad C_2 = \{C, H, I\}, \quad C_3 = \{C, G, H\}, \quad C_4 = \{C, D, F, G\},$$
$$C_5 = \{B, J, K\}, \quad C_6 = \{C, I, J\}, \quad C_7 = \{E, G, H\}.$$

Por ello, (12.7) puede escribirse también usando la representación potencial

$$\begin{aligned} p(x) &= \psi(a,h,i)\psi(c,h,i)\psi(c,g,h) \\ &\quad \psi(c,d,f,g)\psi(b,j,k)\psi(c,i,j)\psi(e,g,h), \end{aligned} \tag{12.8}$$

donde

$$\psi(a,h,i) = p(a)p(i|h,a),$$
$$\psi(c,h,i) = 1,$$
$$\psi(c,g,h) = 1,$$
$$\psi(c,d,f,g) = p(d)p(c|d)p(f|c,d)p(g|d,f),$$
$$\psi(b,j,k) = p(b)p(k|j,b),$$
$$\psi(j,c,i) = p(j|c,i),$$
$$\psi(e,g,h) = p(e)p(h|e,g).$$

EL árbol de unión obtenido usando el Algoritmo 4.4 se muestra en la Figura 12.10.

Supóngase en primer lugar que no hay evidencia disponible. Aplicando el Algoritmo de agrupamiento 8.5 a este árbol de unión, se obtienen las probabilidades marginales iniciales de los nodos que se muestran en la Figura 12.11. La probabilidad inicial de fallo del tanque es $p(k) = 0.009$. Nótese que esta probabilidad cuando no se consideraban las causas comunes era $p(k) = 0.004$.

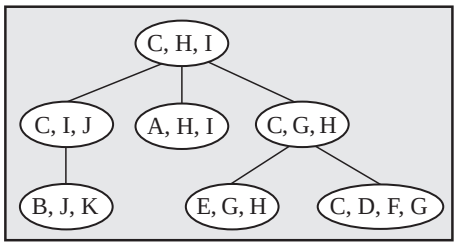


FIGURA 12.10. Un árbol de unión obtenido a partir del grafo no dirigido moralizado y triangulado en 12.9.

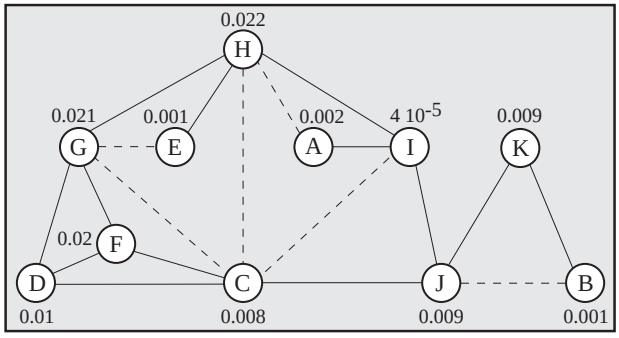


FIGURA 12.11. Las probabilidades marginales iniciales de los nodos (cuando no se dispone de evidencia) para el sistema del tanque de presión con causas comunes de fallo.

Ahora se considera la evidencia $F = f$ y $D = d$. Usando el Algoritmo 8.5 para propagar esta evidencia, se obtienen las probabilidades condicionales de la Figura 12.12. La probabilidad condicional actualizada de fallo es ahora $p(k) = 0.799$. Nótese que con esta misma evidencia la probabilidad de fallo del tanque en el caso de no considerar causas comunes de fallo era $p(k) = 0.006$. La razón que explica esta diferencia es que se ha considerado que el relé C tiene causas comunes de fallo con los relés F y D, por lo que, el fallo de aquel relé implica un aumento considerable de la probabilidad de fallo de éstos.

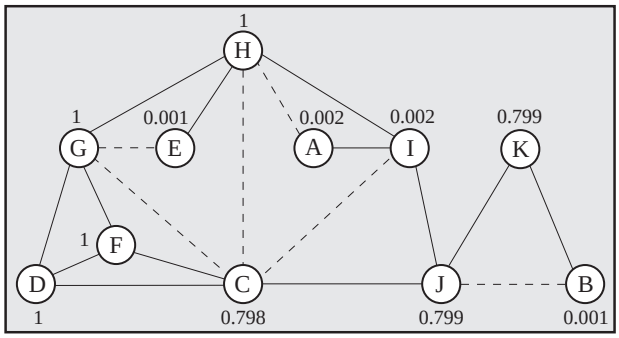


FIGURA 12.12. Probabilidades condicionales de los nodos dada la evidencia $F = f$ y $D = d$ para el tanque de presión con causas comunes de fallo.

Finalmente, cuando se considera la evidencia adicional $A = a$, se obtiene $p(k) = 1$, lo que indica que el tanque falla en este caso (véase la Figura 12.13).

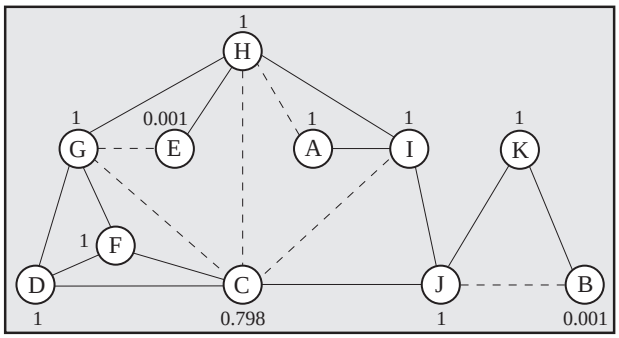


FIGURA 12.13. Probabilidades condicionales de los nodos dada la evidencia $F = f$, $D = d$ y $A = a$ para el tanque de presión con causas comunes de fallo.

X_i	x_i	$p(x_i)$
A	a	0.002
B	b	0.001
C	c	$0.0075 + 0.99\theta_C$
D	d	0.01
E	e	θ_E
F	f	$0.0192 + 0.7326\theta_C$
G	g	$0.0199 + 0.7326\theta_C$
H	h	$0.0199 + 0.7326\theta_C + 0.9801\theta_E - 0.7326\theta_C\theta_E$
I	i	$0.0001 + 0.0015\theta_C + 0.0019\theta_E - 0.0014\theta_C\theta_E$
J	j	$0.0075 + 0.9900\theta_C + 0.0019\theta_E - 0.0019\theta_C\theta_E$
K	k	$0.0085 + 0.9890\theta_C + 0.0019\theta_E - 0.0019\theta_C\theta_E$

TABLA 12.3. Probabilidades marginales iniciales de los nodos (sin evidencia) como una función de los parámetros θ_C y θ_E.

12.2.6 Propagación Simbólica de Evidencia

En esta sección se aplican los métodos de propagación simbólica de evidencia introducidos en el Capítulo 10 para realizar un análisis de sensibilidad, es decir, se desea estudiar el efecto de cambiar las probabilidades asociadas a algunos nodos en las probabilidades de otros nodos de la red. Como ejemplo, modifiquemos algunas de las probabilidades condicionales en (12.6) y la Tabla 12.2 incluyendo algunos parámetros simbólicos para los nodos C y E. Se reemplazan las probabilidades de los nodos C y E por

$$\begin{array}{ll} p(c|d) = 0.75, & p(c|\bar{d}) = \theta_C, \\ p(\bar{c}|d) = 0.25, & p(\bar{c}|\bar{d}) = 1 - \theta_C; \\ p(e) = \theta_E, & p(\bar{e}) = 1 - \theta_E, \end{array}$$

donde $0 < \theta_C < 1$ y $0 < \theta_E < 1$.

Para el caso sin evidencia, usando el método simbólico discutido en el Capítulo 10 (Algoritmo 10.1), se obtienen las probabilidades marginales de los nodos que se muestran en la Tabla 12.3. En esta tabla, se ve que las probabilidades marginales de los nodos C, F y G dependen de θ_C pero no de θ_E. También se puede ver que las probabilidades marginales de los nodos H a K dependen de ambas θ_C y θ_E. Sin embargo, las probabilidades marginales de los nodos J y K dependen mucho más de θ_C que de θ_E (los coeficientes de θ_C son mucho mayores que los de θ_E). También, las probabilidades marginales del nodo I dependen de θ_C y θ_E débilmente.

Los métodos simbólicos pueden usarse también para calcular las probabilidades condicionales de los nodos dada cualquier evidencia. Por ejemplo,

X_i	x_i	$p(x_i)$
A	a	0.002
B	b	0.001
C	c	$(0.007 + 0.743\theta_C)/(0.019 + 0.733\theta_C)$
D	d	$(0.009)/(0.019 + 0.733\theta_C)$
E	e	θ_E
F	f	1
G	g	1
H	h	1
I	i	0.002
J	j	$(0.007 + 0.742\theta_C)/(0.019 + 0.733\theta_C)$
K	k	$(0.007 + 0.742\theta_C)/(0.019 + 0.733\theta_C)$

TABLA 12.4. Probabilidades condicionales de los nodos dada la evidencia $F = f$ como funciones de los parámetros θ_C y θ_E.

la Tabla 12.4 da las probabilidades condicionales de los nodos dada la evidencia $F = f$.

También se ha visto en el Capítulo 10 cómo pueden usarse las expresiones simbólicas, tales como las de la Tabla 12.3, para obtener cotas para las probabilidades marginales y condicionales de los nodos. Para el caso sin evidencia, la Tabla 12.5 muestra las probabilidades marginales iniciales de los nodos y sus correspondientes cotas inferior y superior, que se obtienen cuando los parámetros simbólicos se fijan a sus valores extremos (los llamados *casos canónicos*):

$$\begin{array}{ll} p_{00} = (\theta_C = 0, \theta_E = 0), & p_{01} = (\theta_C = 0, \theta_E = 1), \\ p_{10} = (\theta_C = 1, \theta_E = 0), & p_{11} = (\theta_C = 1, \theta_E = 1). \end{array} \tag{12.9}$$

Nótese que el rango de la variable, es decir, la diferencia entre las cotas superior e inferior, puede utilizarse como un indicador para medir la sensibilidad de las probabilidades a cambios en los valores de los parámetros (un rango reducido significa que es poco sensible).

Una tabla similar puede obtenerse cuando alguna evidencia está disponible. Por ejemplo, las probabilidades condicionales de los nodos y las nuevas cotas inferior y superior se muestran en la Tabla 12.6 para los dos casos de evidencia: $F = f$ y $\{F = f, E = e\}$. Se puede concluir, por ejemplo, que dados $\{F = f, E = e\}$, las probabilidades condicionales de los restantes nodos no dependen de los parámetros θ_C y θ_E.

X_i	x_i	p_{00}	p_{01}	p_{10}	p_{11}	Inf.	Sup.	Rango
A	a	0.002	0.002	0.002	0.002	0.002	0.002	0.000
B	b	0.001	0.001	0.001	0.001	0.001	0.001	0.000
C	c	0.0075	0.007	0.997	0.997	0.007	0.997	0.990
D	d	0.010	0.010	0.010	0.010	0.010	0.010	0.000
E	e	0.000	1.000	0.000	1.000	0.000	1.000	1.000
F	f	0.019	0.019	0.752	0.752	0.019	0.752	0.733
G	g	0.020	0.020	0.752	0.752	0.020	0.753	0.733
H	h	0.020	1.000	0.752	1.000	0.020	1.000	0.980
I	i	10^{-4}	0.002	0.002	0.002	10^{-4}	0.002	0.002
J	j	0.008	0.009	0.997	0.997	0.008	0.998	0.990
K	k	0.008	0.010	0.997	0.997	0.008	0.997	0.989

TABLA 12.5. Probabilidades marginales iniciales de los nodos y sus correspondientes cotas inferior y superior para los casos canónicos en (12.9).

		$F = f$			$(F = f,\, E = e)$		
X_i	x_i	Inf.	Sup.	Rango	Inf.	Sup.	Rango
A	a	0.002	0.002	0.000	0.002	0.002	0.000
B	b	0.001	0.001	0.000	0.001	0.001	0.000
C	c	0.387	0.997	0.610	0.798	0.798	0.000
D	d	0.012	0.484	0.472	1.000	1.000	0.000
E	e	0.000	1.000	1.000	1.000	1.000	0.000
F	f	1.000	1.000	0.000	1.000	1.000	0.000
G	g	1.000	1.000	0.000	1.000	1.000	0.000
H	h	1.000	1.000	0.000	1.000	1.000	0.000
I	i	0.002	0.002	0.000	0.002	0.002	0.000
J	j	0.388	0.997	0.609	0.799	0.799	0.000
K	k	0.388	0.997	0.608	0.799	0.799	0.000

TABLA 12.6. Cotas inferiores y superiores de las probabilidades condicionales para los casos canónicos dados dos casos de evidencia, $F = f$ y $\{F = f, E = e\}$.

12.3 Sistema de Distribución de Energía

12.3.1 Definición del problema

La Figura 12.14 muestra un sistema de distribución con tres motores, 1, 2, y 3 y tres temporizadores, A, B y C, que están normalmente cerrados. Una

pulsación momentánea del pulsador F suministra energía de una batería a los relés G e I. A partir de ese instante G e I se cierran y permanecen activados eléctricamente. Para comprobar si los tres motores están operando propiamente, se envía una señal de prueba de 60 segundos a través de K. Una vez que K se ha cerrado, la energía de la batería 1 llega a los relés R y M. El cierre de R arranca el motor 1. El cierre de T envía energía de la batería 1 a S. El cierre de S arranca el motor 3.

Tras un intervalo de 60 segundos, K debe abrirse, interrumpiendo la operación de los tres motores. Si K dejase de cerrarse tras los 60 segundos, los tres temporizadores A, B y C se abrirían, dejando sin energía a G y por tanto parando el sistema. Supóngase que K se abre para dejar sin energía a G y el motor 1 para. B y C actúan de forma similar para parar el motor 2 ó el motor 3, por lo que M o S deberían dejar de estar cerrados. En lo que sigue se analiza sólo el efecto sobre el motor 2. El análisis de los motores 1 y 3 se dejan como ejercicio al lector.

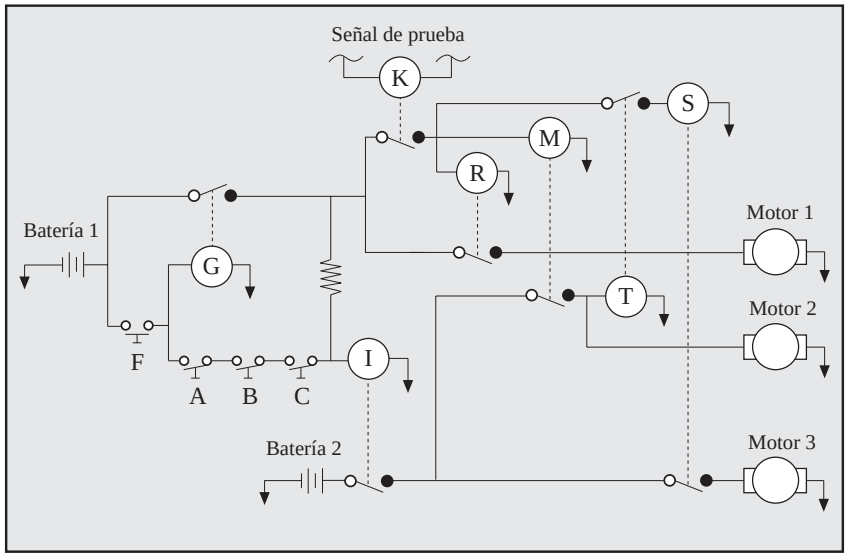


FIGURA 12.14. Un diagrama del sistema de distribución de energía.

12.3.2 Selección de Variables

Se está interesado en conocer el estado de operación del motor 2. Denotemos a esta variable aleatoria por Q. Por tanto, q significa fallo y $\bar{q}$ significa no fallo. Se desea calcular $p(q)$. La Figura 12.15 muestra el árbol de fallos y los conjuntos que conducen al fallo del sistema. Nótese que el fallo del motor 2 es igual a la expresión lógica

$$q = [(m \vee (k \wedge g) \vee k \wedge (a \wedge b \wedge c) \vee (k \wedge f)] \wedge (i \vee g \vee b \vee f), \qquad (12.10)$$

donde los símbolos $\vee$ y $\wedge$ se usan para o e y, respectivamente.

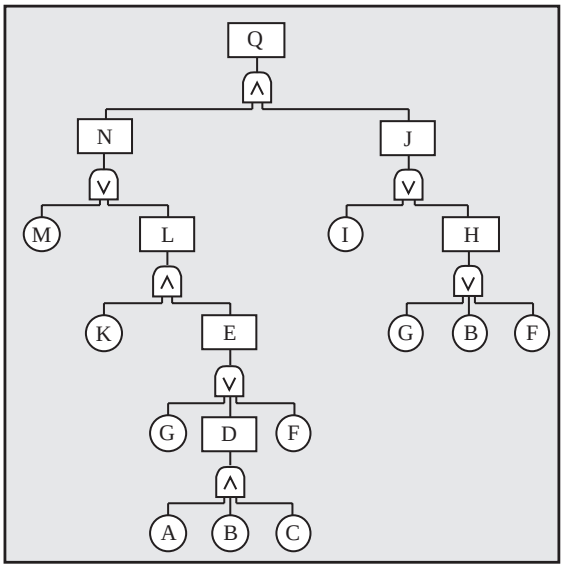


FIGURA 12.15. árbol de fallos para el motor 2.

La ecuación (12.10) puede utilizarse para obtener el conjunto de reglas de un sistema experto determinista. La Figura 12.16 muestra el conjunto de reglas obtenido de (12.10). El conjunto de variables usadas en este ejemplo es

$$X = \{A, B, C, D, E, F, G, H, I, J, K, L, M, N, Q\},$$

donde D, E, H, J, L y N son fallos intermedios.

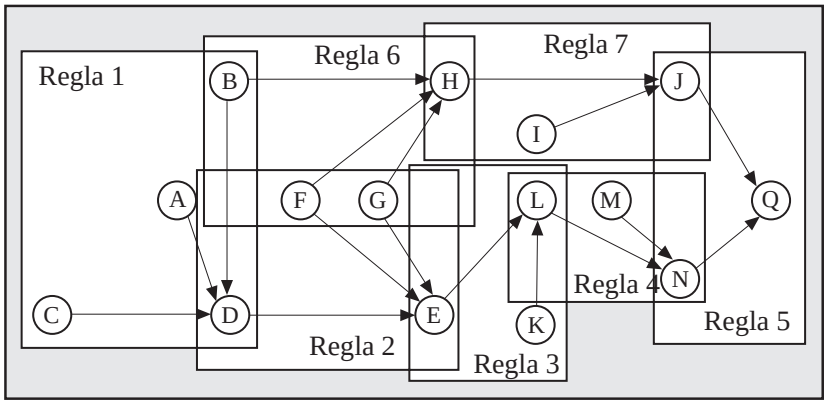


FIGURA 12.16. Reglas encadenadas para el motor 2.

El estudio de este ejemplo desde un punto de vista determinista se deja como ejercicio al lector.

Como en el ejemplo anterior, la estructura encadenada, dada por las reglas, nos permite definir un modelo muy potente para este ejemplo. Por ello, se utilizará el grafo dirigido de la Figura 12.17 como modelo gráfico para una red Bayesiana cuya función de probabilidad conjunta puede factorizarse en la forma

$$\begin{aligned} p(x) &= p(a)p(b)p(c)p(d|a,b,c)p(f)p(g)p(e|d,f,g)p(h|b,f,g) \\ &\quad p(i)p(j|h,i)p(k)p(l|e,k)p(m)p(n|l,m)p(q|j,n). \end{aligned}$$

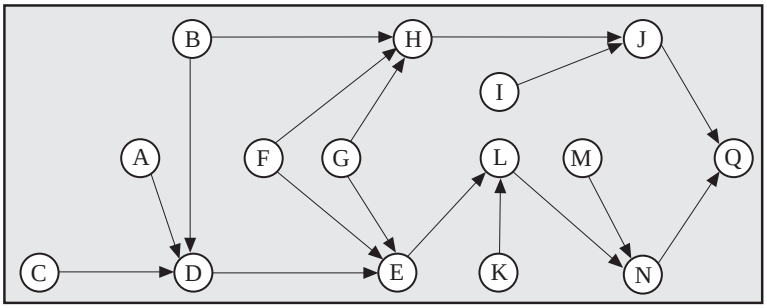


FIGURA 12.17. Grafo dirigido múltiplemente conexo para el sistema de distribución de energía (motor 2).

Las funciones de probabilidad condicionada necesarias para definir la función de probabilidad conjunta se dan en la Tabla 12.7 (se dan sólo las probabilidades de fallo puesto que $p(no\ fallo) = 1 - p(fallo)$). Las probabilidades marginales de los nodos terminales A, B, C, F, G, I, K y M son

$$\begin{array}{llll} p(a) = 0.010, & p(b) = 0.010, & p(c) = 0.010, & p(f) = 0.011 \\ p(g) = 0.011, & p(i) = 0.001, & p(k) = 0.002, & p(m) = 0.003. \end{array}$$

Para ilustrar mejor el procedimiento de trabajo, se usa un método exacto y otro aproximado para la propagación de evidencia en esta red Bayesiana.

12.3.3 Propagación Exacta de Evidencia

La Figura 12.18 muestra el grafo no dirigido moralizado y triangulado que corresponde al grafo dirigido de la Figura 12.17. En la Figura 12.18 se da una numeración perfecta de los nodos.

Los conglomerados, que pueden obtenerse del grafo de la Figura 12.18, son

$$\begin{array}{lll} C_1 = \{A,B,C,D\}, & C_2 = \{B,D,E,F,G\}, & C_3 = \{B,E,F,G,H\} \\ C_4 = \{E,H,L\}, & C_5 = \{H,I,J,L\}, & C_6 = \{E,K,L\}, \\ C_7 = \{J,L,M,N\}, & C_8 = \{J,N,Q\}, & \end{array}$$

I	H	$p(j\|I,H)$
i	h	1
i	$\bar{h}$	1
$\bar{i}$	h	1
$\bar{i}$	$\bar{h}$	0

E	K	$p(l\|E,K)$
e	k	1
e	$\bar{k}$	0
$\bar{e}$	k	0
$\bar{e}$	$\bar{k}$	0

L	M	$p(n\|L,M)$
l	m	1
l	$\bar{m}$	1
$\bar{l}$	m	1
$\bar{l}$	$\bar{m}$	0

J	N	$p(q\|J,N)$
j	n	1
j	$\bar{n}$	0
$\bar{j}$	n	0
$\bar{j}$	$\bar{n}$	0

A	B	C	$p(d\|A,B,C)$
a	b	c	1
a	b	$\bar{c}$	0
a	$\bar{b}$	c	0
a	$\bar{b}$	$\bar{c}$	0
$\bar{a}$	b	c	0
$\bar{a}$	b	$\bar{c}$	0
$\bar{a}$	$\bar{b}$	c	0
$\bar{a}$	$\bar{b}$	$\bar{c}$	0

D	F	G	$p(e\|D,F,G)$
d	f	g	1
d	f	$\bar{g}$	1
d	$\bar{f}$	g	1
d	$\bar{f}$	$\bar{g}$	1
$\bar{d}$	f	g	1
$\bar{d}$	f	$\bar{g}$	1
$\bar{d}$	$\bar{f}$	g	1
$\bar{d}$	$\bar{f}$	$\bar{g}$	0

B	F	G	$p(h\|B,F,G)$
b	f	g	1
b	f	$\bar{g}$	1
b	$\bar{f}$	g	1
b	$\bar{f}$	$\bar{g}$	1
$\bar{b}$	f	g	1
$\bar{b}$	f	$\bar{g}$	1
$\bar{b}$	$\bar{f}$	g	1
$\bar{b}$	$\bar{f}$	$\bar{g}$	0

TABLA 12.7. Probabilidades condicionales de fallo de las variables del sistema de distribución de energía (motor 2).

lo que implica que la función de probabilidad conjunta de los nodos puede escribirse como una función de las funciones potenciales como sigue:

$$\begin{aligned} p(x) \quad = \quad & \psi(a,b,c,d)\ \psi(b,d,e,f,g)\ \psi(b,e,f,g,h)\ \psi(e,h,l) \\ & \times \psi(h,i,j,l)\ \psi(e,k,l)\ \psi(j,l,m,n)\ \psi(j,n,q), \end{aligned} \tag{12.11}$$

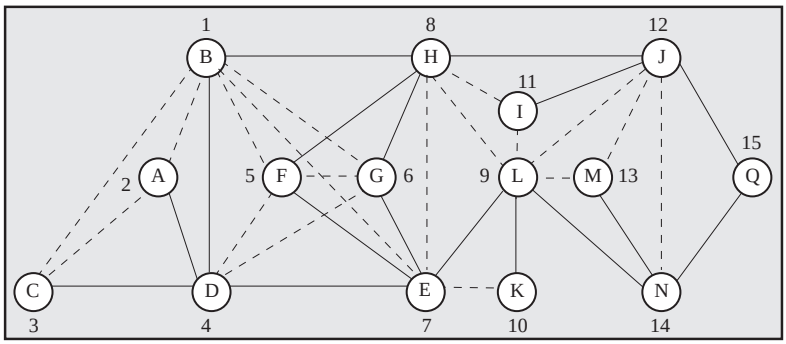


FIGURA 12.18. Grafo moralizado y triangulado asociado al grafo dirigido de la Figura 12.17. Se muestra una numeración perfecta de los nodos.

donde

$$\begin{aligned}
&\psi(a,b,c,d) = p(a)p(b)p(c)p(d|a,b,c),\\
&\psi(b,d,e,f,g) = p(f)p(g)p(e|d,f,g),\\
&\psi(b,e,f,g,h) = p(h|b,f,g),\\
&\psi(e,h,l) = 1,\\
&\psi(h,i,j,l) = p(i)p(j|i,h),\\
&\psi(e,k,l) = p(k)p(l|e,k),\\
&\psi(j,l,m,n) = p(m)p(n|l,m),\\
&\psi(j,n,q) = p(q|j,n).
\end{aligned} \tag{12.12}$$

El árbol de unión correspondiente se muestra en la Figura 12.19.

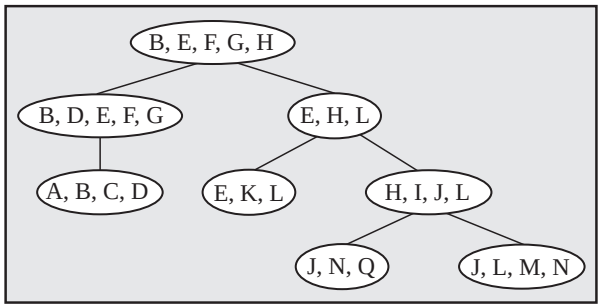


FIGURA 12.19. Un árbol de unión obtenido del grafo moralizado y triangulado de la Figura 12.18.

Se usa el Algoritmo de agrupamiento 8.5 para obtener las probabilidades marginales iniciales de los nodos cuando no hay evidencia disponible. Estas probabilidades se muestran en la Figura 12.20. Supóngase ahora que se tiene la evidencia $K = k$. Las probabilidades condicionales de los nodos dada esta evidencia se obtienen usando el Algoritmo 8.5 y se muestran en

la Figura 12.21. En este caso, la probabilidad de fallo aumenta pasando del valor inicial $p(q) = 0.0001$ al valor $p(q|K = k) = 0.022$.

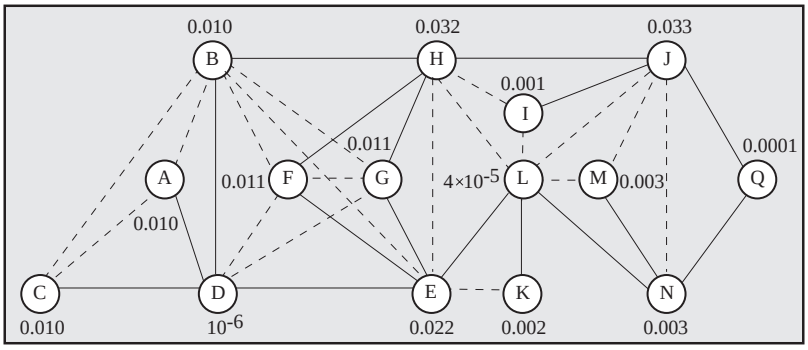


FIGURA 12.20. Probabilidades marginales de los nodos cuando no hay evidencia disponible para el sistema de distribución de energía (motor 2).

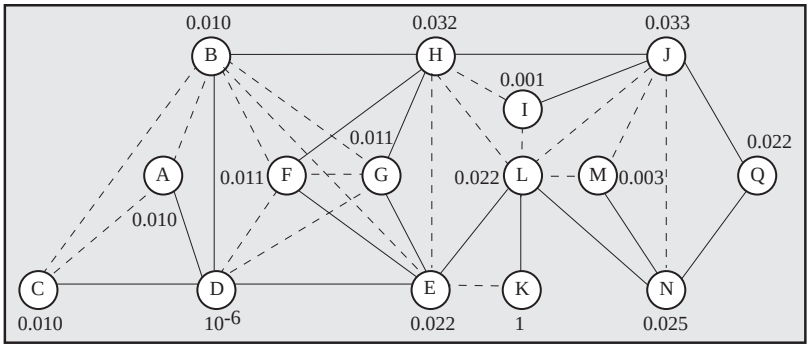


FIGURA 12.21. Probabilidades condicionales de los nodos dada la evidencia $K = k$, para el sistema de distribución de energía (motor 2).

Cuando se introduce la evidencia adicional $E = e$, entonces L y N también fallan. Consecuentemente, el sistema falla: $p(q|E = e, K = k) = 1$ (véase la Figura 12.22).

12.3.4 Propagación Aproximada de Evidencia

En el Capítulo 9 se han introducido varios algoritmos para propagar la evidencia de forma aproximada. Se ha visto que el método de la *verosimilitud pesante* es uno de los más eficientes dentro de los métodos estocásticos y que el *muestreo sistemático* y el de *búsqueda de la máxima probabilidad* son los más eficientes dentro de los de tipo determinista en el caso de redes con probabilidades extremas. En este caso, se tiene una red Bayesiana con

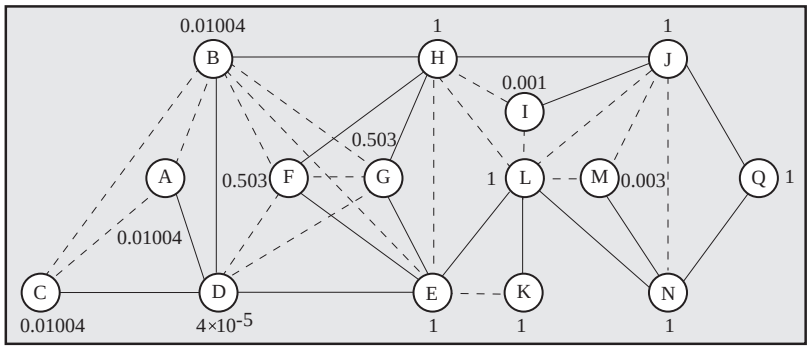


FIGURA 12.22. Probabilidades condicionales de los nodos dada la evidencia $E = e$ y $K = k$, para el sistema de distribución de energía (motor 2).

tablas de probabilidad que contienen valores extremos (ceros y unos), una situación en la que el método de la verosimilitud pesante se sabe que es ineficiente. Sin embargo, en lo que sigue se comparan los métodos anteriores en el caso de esta red Bayesiana (antes y después de conocer la evidencia $E = e, K = k$).

La Tabla 12.8 da el error de la aproximación,

$$\text{error} = |\text{exacta} - \text{aproximada}|,$$

para ambos métodos y para diferente número de réplicas. Claramente, el algoritmo de muestreo sistemático (véase la Sección 9.9) vence al de la verosimilitud pesante, conduciendo a errores mucho más pequeños para el mismo número de réplicas. La ineficiencia del algoritmo de la verosimilitud pesante es parcialmente debida a las probabilidades extremas. Puesto que la mayoría de las ocurrencias tienen asociada una probabilidad nula, el método más eficiente aquí es el de *búsqueda de la máxima probabilidad*. Por ejemplo, aún en el caso de que se considere un número de ocurrencias tan bajo como 10, el error obtenido (no mostrado) es menor que 3×10^{-6}.

12.4 Daño en Vigas de Hormigón Armado

En las Secciones 12.2 y 12.3 se han usado modelos de redes probabilísticas para definir funciones de probabilidad conjunta consistentes de una forma sencilla y directa para el caso de dos problemas de la vida real. El uso de redes Bayesianas fue sencillo en esos casos porque las relaciones de dependencia entre las variables no eran complicadas. Por ejemplo, no había retroalimentación o ciclos. En esta sección se presenta un problema que inicialmente no puede ser tratado mediante una red Bayesiana por la presencia de ciclos. En este caso se utilizan los modelos especificados condicionalmente de la Sección 7.7, que son más generales, para modelar las

Número de Simulaciones	Error			
	Sin Evidencia		$E = e, K = k$	
	Verosimilitud	Sistemático	Verosimilitud	Sistemático
100	0.00205	0.00023	0.19841	0.00650
1,000	0.00021	5.25×10^{-6}	0.04300	0.00292
2,000	6.26×10^{-5}	3.91×10^{-6}	0.01681	0.00109
10,000	1.49×10^{-5}	4.35×10^{-7}	0.00302	3.34×10^{-5}
20,000	9.36×10^{-6}	1.22×10^{-7}	0.00265	1.78×10^{-5}
50,000	5.79×10^{-6}	3.08×10^{-8}	0.00053	7.66×10^{-6}
100,000	1.26×10^{-6}	3.06×10^{-9}	0.00011	2.08×10^{-6}

TABLA 12.8. Rendimiento de los métodos de la verosimilitud pesante y de muestreo sistemático con diferentes números de réplicas.

relaciones entre variables. Los problemas de compatibilidad y eliminación de información redundante, que no surgen en los modelos de redes Bayesianas, son ahora importantes y tienen que ser abordados durante el proceso de modelización.

12.4.1 Definición del problema

En este caso, el objetivo consiste en determinar el daño de vigas de hormigón armado. En esta sección se ilustra este problema usando un modelo mixto con variables discretas y continuas. Alternativamente, en la Sección 12.5, se usan modelos de redes Bayesianas normales (Gausianas) en los que todas las variables son continuas. Este ejemplo, que está tomado de Liu y Li (1994) (véase también Castillo, Gutiérrez y Hadi (1995b)), ha sido modificado ligeramente por motivos ilustrativos. La primera parte de la formulación del modelo consta de dos etapas: selección de las variables e identificación de las dependencias.

12.4.2 Selección de las Variables

El proceso de la formulación del modelo comienza generalmente con la selección o especificación de un conjunto de variables de interés. Esta especificación corresponde a los expertos humanos en la especialidad (ingenieros civiles, en este caso). En nuestro ejemplo, la variable objetivo (el daño de una viga de hormigón armado) se denota por X_1. Un ingeniero civil identifica inicialmente 16 variables $(X_9, \ldots, X_{24})$ como las variables principales que influyen en el daño de una viga de hormigón armado. Además, el ingeniero identifica siete variables intermedias no observables $(X_2, \ldots, X_8)$ que definen estados parciales de la estructura. La Tabla 12.9 muestra la lista de variables y sus respectivas definiciones. La tabla también muestra

X_i	Tipo	Valores	Definición
X_1	Discreta	$\{0,1,2,3,4\}$	Daño de la viga
X_2	Discreta	$\{0,1,2\}$	Estado de agrietamiento
X_3	Discreta	$\{0,1,2\}$	Agrietamiento por cortante
X_4	Discreta	$\{0,1,2\}$	Corrosión del acero
X_5	Discreta	$\{0,1,2\}$	Agrietamiento por flexión
X_6	Discreta	$\{0,1,2\}$	Agrietamiento por retracción
X_7	Discreta	$\{0,1,2\}$	Peor grieta por flexión
X_8	Discreta	$\{0,1,2\}$	Estado de corrosión
X_9	Continua	$(0-10)$	Debilidad de la viga
X_{10}	Discreta	$\{0,1,2\}$	Flecha de la viga
X_{11}	Discreta	$\{0,1,2,3\}$	Posición de la peor grieta de cortante
X_{12}	Discreta	$\{0,1,2\}$	Tamaño de la peor grieta de cortante
X_{13}	Discreta	$\{0,1,2,3\}$	Posición de la peor grieta de flexión
X_{14}	Discreta	$\{0,1,2\}$	Tamaño de la peor grieta de flexión
X_{15}	Continua	$(0-10)$	Longitud de la peor grieta de flexión
X_{16}	Discreta	$\{0,1\}$	Recubrimiento
X_{17}	Continua	$(0-100)$	Edad de la estructura
X_{18}	Continua	$(0-100)$	Humedad
X_{19}	Discreta	$\{0,1,2\}$	PH del aire
X_{20}	Discreta	$\{0,1,2\}$	Contenido de cloro en el aire
X_{21}	Discreta	$\{0,1,2,3\}$	Número de grietas de cortante
X_{22}	Discreta	$\{0,1,2,3\}$	Número de grietas de flexión
X_{23}	Discreta	$\{0,1,2,3\}$	Retracción
X_{24}	Discreta	$\{0,1,2,3\}$	Corrosión

TABLA 12.9. Definiciones de las variables que intervienen en el problema de daño de vigas de hormigón armado.

el carácter continuo o discreto de cada variable. Las variables se miden usando una escala que está ligada directamente a la variable objetivo, es decir, cuanto mayor es el valor de la variable mayor es la posibilidad de daño. Sea $X = \{X_1, \ldots, X_{24}\}$ el conjunto de todas las variables.

12.4.3 Identificación de las Dependencias

La etapa siguiente en la formulación del modelo consiste en la identificación de la estructura de las dependencias entre las variables seleccionadas. Esta identificación corresponde también a un ingeniero civil y se hace normalmente identificando el menor conjunto de variables, $Vec(X_i)$, para cada variable X_i tales que

$$p(x_i|x \setminus x_i) = p(x_i|Vec(X_i)), \tag{12.13}$$

X_i	$Vec(X_i)$	Π_i
X_1	$\{X_9, X_{10}, X_2\}$	$\{X_9, X_{10}, X_2\}$
X_2	$\{X_3, X_6, X_5, X_4, X_1\}$	$\{X_3, X_6, X_5, X_4\}$
X_3	$\{X_{11}, X_{12}, X_{21}, X_8, X_2\}$	$\{X_{11}, X_{12}, X_{21}, X_8\}$
X_4	$\{X_{24}, X_8, X_5, X_2, X_{13}\}$	$\{X_{24}, X_8, X_5\}$
X_5	$\{X_{13}, X_{22}, X_7, X_2, X_4\}$	$\{X_{13}, X_{22}, X_7\}$
X_6	$\{X_{23}, X_8, X_2\}$	$\{X_{23}, X_8\}$
X_7	$\{X_{14}, X_{15}, X_{16}, X_{17}, X_8, X_5\}$	$\{X_{14}, X_{15}, X_{16}, X_{17}, X_8\}$
X_8	$\{X_{18}, X_{19}, X_{20}, X_7, X_4, X_6, X_3\}$	$\{X_{18}, X_{19}, X_{20}\}$
X_9	$\{X_1\}$	ϕ
X_{10}	$\{X_1\}$	ϕ
X_{11}	$\{X_3\}$	ϕ
X_{12}	$\{X_3\}$	ϕ
X_{13}	$\{X_5, X_4\}$	$\{X_4\}$
X_{14}	$\{X_7\}$	ϕ
X_{15}	$\{X_7\}$	ϕ
X_{16}	$\{X_7\}$	ϕ
X_{17}	$\{X_7\}$	ϕ
X_{18}	$\{X_8\}$	ϕ
X_{19}	$\{X_8\}$	ϕ
X_{20}	$\{X_8\}$	ϕ
X_{21}	$\{X_3\}$	ϕ
X_{22}	$\{X_5\}$	ϕ
X_{23}	$\{X_6\}$	ϕ
X_{24}	$\{X_4\}$	ϕ

TABLA 12.10. Variables y sus correspondientes vecinos, $Vec(X_i)$ y padres, Π_i, para el caso del daño de una viga de hormigón armado.

donde el conjunto $Vec(X_i)$ se llama el conjunto de *vecinos* de X_i. La ecuación (12.13) indica que la variable X_i es condicionalmente independiente del conjunto $R_i = X \setminus \{X_i, Vec(X_i)\}$ dado $Vec(X_i)$. Por ello, utilizando la notación de independencia condicional (véase la Sección 5.1), se puede escribir $I(X_i, R_i|Vec(X_i))$. Las variables y sus correspondientes vecinos se muestran en las dos primeras columnas de la Tabla 12.10. Se sigue que si $X_j \in Vec(X_i)$, entonces $X_i \in Vec(X_j)$.

Adicionalmente, pero opcionalmente, el ingeniero puede imponer ciertas relaciones de causa-efecto entre las variables, es decir, especificar qué variables entre las del conjunto $Vec(X_i)$ son causas directas de X_i y cuáles son los efectos directos de X_i. El conjunto de las causas directas de X_i se conoce como el conjunto de *padres* de X_i y se denota por Π_i.

En nuestro ejemplo, el ingeniero especifica las siguientes relaciones de causa-efecto, tal como se muestra en la Figura 12.23. La variable objetivo

X_1, depende fundamentalmente de tres factores: X_9, la debilidad de la viga, disponible en la forma de un factor de daño, X_{10}, la flecha de la viga, y X_2, su estado de agrietamiento. El estado de agrietamiento, X_2, depende de cuatro variables: X_3, el estado de agrietamiento por cortante, X_6, el agrietamiento por retracción, X_4, la corrosión del acero, y X_5, el estado de agrietamiento por flexión. El agrietamiento por retracción, X_6, depende de la retracción, X_{23} y el estado de corrosión, X_8. La corrosión del acero, X_4, está ligada a X_8, X_{24} y X_5. El estado de agrietamiento por cortante, X_3, depende de cuatro factores: X_{11}, la posición de la peor grieta de cortante, X_{12}, la anchura de la misma, X_{21}, el número de grietas de cortante, y X_8. En el estado de agrietamiento por flexión, X_5 influyen tres variables: X_{13}, la posición de la peor grieta de flexión, X_{22}, el número de grietas de flexión, y X_7, la peor grieta de flexión. La variable X_{13} depende de X_4. La variable X_7 es una función de cinco variables: X_{14}, la anchura de la peor grieta de flexión, X_{15}, la longitud de la peor grieta de flexión, X_{16}, el recubrimiento, X_{17}, la edad de la estructura, y X_8, el estado de corrosión. La variable X_8 está ligada a tres variables: X_{18}, la humedad, X_{19}, el PH del aire, y X_{20}, el contenido de cloro en el aire.

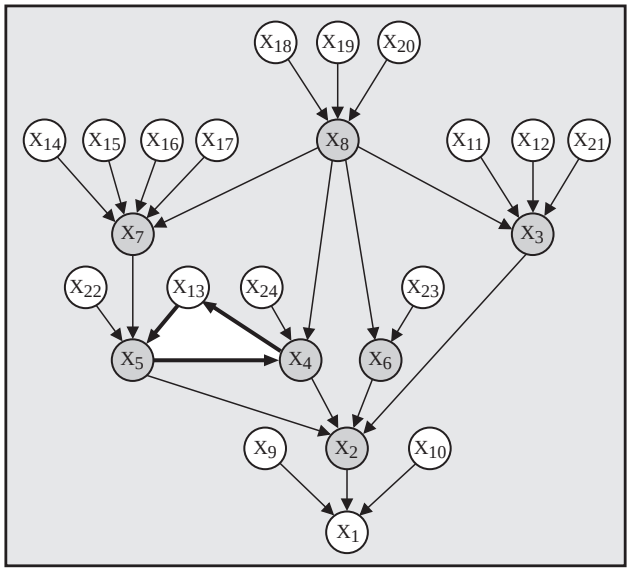


FIGURA 12.23. Grafo dirigido para representar el problema de determinación de daño en vigas de hormigón armado. Los nodos sombreados representan variables auxiliares (no observables) del modelo.

El conjunto, de padres Π_i de cada una de las variables de la Figura 12.23 se muestra en la tercera columna de la Tabla 12.10. Si no se diesen relaciones causa-efecto las relaciones se representarían mediante aristas no dirigidos (una línea que conecta dos nodos).

12.4.4 Especificación de Distribuciones Condicionales

Una vez que se ha especificado la estructura gráfica, el ingeniero suministra un conjunto de probabilidades condicionales sugeridas por el grafo. Para simplificar la asignación de probabilidades condicionales, el ingeniero supone que éstas pertenecen a familias paramétricas (por ejemplo, Binomial, Beta, etc.). El conjunto de probabilidades condicionales se da en la Tabla 12.11, donde las cuatro variables continuas se suponen de tipo $Beta(a, b)$ con los parámetros indicados y las variables discretas se suponen *Binomiales* $B(n, p)$. La razón para esta elección es que la distribución beta tiene rango finito y una gran variedad de formas dependiendo de la elección de los parámetros.

La variable X_1 puede tomar sólo cinco valores (estados): $0, 1, 2, 3, 4$. Con 0 se indica que la viga está libre de daño y con 4 que está seriamente dañada. Los valores intermedios, entre 0 y 4, son estados intermedios de daño. Las restantes variables se definen de forma similar usando una escala que está directamente ligada a la variable objetivo, es decir, cuanto mayores sean sus valores mayor es el daño.

Todas las variables discretas se supone que siguen una distribución binomial con parámetros N y p, con $N + 1$ estados posibles para cada variable. Sin embargo, estas distribuciones pueden reemplazarse por otras cualesquiera. El parámetro $0 \leq p \leq 1$ se especifica como sigue. Sean π_i los valores observados de los padres de un nodo dado X_i. La función $p_i(\pi_i)$, $i = 1, \ldots, 8$, de la Tabla 12.11 es una función que toma π_i como dato y produce una probabilidad asociada al nodo X_i, es decir, $p_i(\pi_i) = h(\pi_i)$. Por simplicidad considérese $\Pi_i = \{X_1, \ldots, X_m\}$. Entonces, algunos posibles ejemplos de $h(\pi_i)$ son

$$h(\pi_i) = \sum_{j=1}^{m} \frac{x_j/u_j}{m} \tag{12.14}$$

y

$$h(\pi_i) = 1 - \prod_{j=1}^{m} (1 - x_j/u_j), \tag{12.15}$$

donde u_j es una cota superior (por ejemplo, el valor máximo) de la variable aleatoria X_j. Las funciones $h(\pi_i)$ en (12.14) y (12.15) crecen con valores crecientes de Π_i. También satisfacen el axioma de la probabilidad $0 \leq h(\pi_i) \leq 1$. Debe señalarse aquí que estas funciones son sólo ejemplos, dados con la intención de ilustrar, y que pueden reemplazarse por otras funciones con las mismas propiedades.

La Tabla 12.12 da las funciones $h(\pi_i)$ utilizadas para calcular las probabilidades condicionales de la Tabla 12.11. Alternativamente, pudiera darse una tabla de distribuciones marginales o condicionales para cada variable discreta.

X_i	$p(x_i\|u_i)$	Familia
X_1	$p(x_1\|x_9, x_{10}, x_2)$	$B(4, p_1(x_9, x_{10}, x_2))$
X_2	$p(x_2\|x_3, x_6, x_4, x_5)$	$B(2, p_2(x_3, x_6, x_4, x_5))$
X_3	$p(x_3\|x_{11}, x_{12}, x_{21}, x_8)$	$B(2, p_3(x_{11}, x_{12}, x_{21}, x_8))$
X_4	$p(x_4\|x_{24}, x_8, x_5)$	$B(2, p_4(x_{24}, x_8, x_5))$
X_5	$p(x_5\|x_{13}, x_{22}, x_7)$	$B(2, p_5(x_{13}, x_{22}, x_7))$
X_6	$p(x_6\|x_{23}, x_8)$	$B(2, p_6(x_{23}, x_8))$
X_7	$p(x_7\|x_{14}, x_{15}, x_{16}, x_{17}, x_8)$	$B(2, p_7(x_{14}, x_{15}, x_{16}, x_{17}, x_8))$
X_8	$p(x_8\|x_{18}, x_{19}, x_{20})$	$B(2, p_8(x_{18}, x_{19}, x_{20}))$
X_9	$f(x_9)$	$10 * Beta(0.5, 8)$
X_{10}	$p(x_{10})$	$B(2, 0.1)$
X_{11}	$p(x_{11})$	$B(3, 0.2)$
X_{12}	$p(x_{12})$	$B(2, 0.1)$
X_{13}	$p(x_{13}\|x_4)$	$B(3, p_{13}(x_4))$
X_{14}	$p(x_{14})$	$B(2, 0.1)$
X_{15}	$f(x_{15})$	$10 * Beta(1, 4)$
X_{16}	$p(x_{16})$	$B(1, 0.1)$
X_{17}	$f(x_{17})$	$100 * Beta(2, 6)$
X_{18}	$f(x_{18})$	$100 * Beta(2, 6)$
X_{19}	$p(x_{19})$	$B(2, 0.2)$
X_{20}	$p(x_{20})$	$B(2, 0.2)$
X_{21}	$p(x_{21})$	$B(3, 0.2)$
X_{22}	$p(x_{22})$	$B(3, 0.2)$
X_{23}	$p(x_{23})$	$B(3, 0.1)$
X_{24}	$p(x_{24})$	$B(3, 0.1)$

TABLA 12.11. Probabilidades marginales y condicionadas correspondientes a la red de la Figura 12.23.

X_i	$p(\pi_i)$	$h(\pi_i)$
X_1	$p_1(x_9, x_{10}, x_2)$	(12.15)
X_2	$p_2(x_3, x_6, x_5, x_4)$	(12.14)
X_3	$p_3(x_{11}, x_{12}, x_{21}, x_8)$	(12.14)
X_4	$p_4(x_{24}, x_8, x_5, x_{13})$	(12.14)
X_5	$p_5(x_{13}, x_{22}, x_7)$	(12.14)
X_6	$p_6(x_{23}, x_8)$	(12.14)
X_7	$p_7(x_{14}, x_{15}, x_{16}, x_{17}, x_8)$	(12.14)
X_8	$p_8(x_{18}, x_{19}, x_{20})$	(12.14)

TABLA 12.12. Funciones de probabilidad requeridas para calcular las probabilidades condicionales de la Tabla 12.11.

12.4.5 Diagnóstico del modelo

Una vez que se ha dado el conjunto de probabilidades condicionales de la Tabla 12.11, comienza la tarea del experto estadístico. Antes de que la propagación de evidencia pueda comenzar, hay que comprobar que el conjunto de probabilidades condicionales cumple las condiciones de unicidad y compatibilidad puesto que el grafo de la Figura 12.23 contiene un ciclo.

Dado un conjunto de nodos $X = \{X_1, \ldots, X_n\}$ y un conjunto de probabilidades condicionales de los nodos de X, el experto estadístico determina si las probabilidades condicionales corresponden a una función de probabilidad conjunta de las variables de la red bien definida.

En algunos casos particulares (cuando el modelo no tiene ciclos), se puede fijar una ordenación de los nodos, por ejemplo $X_1, \ldots, X_n$, tal que

$$\Pi_i = Vec(X_i) \cap A_i, \tag{12.16}$$

donde $A_i = \{X_{i+1}, \ldots, X_n\}$, es decir, el conjunto de padres de un nodo es el conjunto de vecinos con número mayor que el nodo dado.[3] En este caso, el grafo asociado es un grafo dirigido acíclico , que define una red Bayesiana cuya función de probabilidad conjunta puede factorizarse en la forma

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{24} p(x_i|a_i) = \prod_{i=1}^{24} p(x_i|\pi_i). \tag{12.17}$$

Sin embargo, cuando los nodos no pueden ordenarse como en (12.16), la función de probabilidad conjunta no puede ser definida directamente mediante (12.17). Por ejemplo, en nuestro ejemplo del daño de estructuras de hormigón armado, las variables no pueden ser ordenadas como en (12.16) puesto que los nodos X_5, X_4 y X_{13} forman un ciclo (véase la Figura 12.23). En esta situación, el conjunto de probabilidades condicionales

$$P = \{p(x_i|\pi_i) : i = 1, \ldots, 24\}$$

dado por el ingeniero en la Tabla 12.11 tiene que ser comprobado para ver si verifica las condiciones de unicidad y consistencia.

Nótese que si no se hubieran dado relaciones causa-efecto, siempre hubiera podido encontrarse una ordenación satisfaciendo (12.16) puesto que no hay restricción en la elección de los padres, es decir, cualesquiera subconjuntos de $Vec(x_i)$ pueden servir como π_i. Por otra parte, si se dan algunas relaciones causa-efecto, la ordenación de los nodos deberá satisfacer (12.16), es decir, un hijo debe ser numerado con un número más pequeño que sus padres. Se deduce entonces que si las relaciones causa-efecto tienen ciclos, no existe ninguna ordenación que satisfaga (12.16). Sin embargo, como se

[3] Nótese que una definición equivalente puede darse considerando $\Pi_i = B_i \cap Vec(X_i)$, donde $B_i = \{X_1, \ldots, X_{i-1}\}$.

ha visto en la Sección 7.7, los ciclos siempre conducen a información redundante, y por tanto, pueden ser eliminados del modelo. Para obtener un modelo reducido, debe comprobarse si el conjunto dado de probabilidades condicionales satisface las condiciones de unicidad y consistencia con el fin de detectar los enlaces asociados a información redundante.

Dada una ordenación de las variables $\{X_1, \ldots, X_n\}$, se ha mostrado, en la Sección 7.7, que un conjunto de probabilidades condicionales que contiene una sucesión de la forma $p(x_i|a_i, u_i)$, $i = 1, \ldots, n$, donde $A_i = \{X_{i+1}, \ldots, X_n\}$ y $U_i \subset B_i = \{X_1, \ldots, X_{i-1}\}$, contiene suficiente información para definir como mucho una función de probabilidad conjunta. Las variables de la Tabla 12.11 se han ordenado de forma que no sea necesario reordenarlas para comprobar la unicidad. Por tanto, puede verse fácilmente que las dadas en la Tabla 12.11 satisfacen dicha condición, ya que se supone (12.13).

El teorema de unicidad asegura sólo la unicidad, pero no garantiza la existencia de una función de probabilidad conjunta para el conjunto X. En la Sección 7.7 se da un teorema mediante el cual uno puede determinar si un conjunto dado de probabilidades condicionales define una función de probabilidad conjunta para X (compatibilidad). En esta sección también se muestra que la información redundante está asociada a cierta estructura de las probabilidades condicionales. Más precisamente, las probabilidades condicionales de la forma $p(x_i|a_i, u_i)$, con $U_i \neq \phi$ pueden ser inconsistentes con las probabilidades condicionales previas, y, por tanto, contienen información redundante. Por tanto, cualquier enlace $Y_j \rightarrow X_i$, con $Y_j \in U_i$, puede ser eliminado sin restringir las funciones de probabilidad conjuntas asociadas a la estructura gráfica. Además, eliminando los ciclos se elimina la información redundante del modelo y el proceso de asignación de probabilidades puede hacerse sin restricciones en los valores numéricos que se asignan a los parámetros (aparte de los axiomas de la probabilidad que individualmente debe satisfacer cada una de las distribuciones condicionadas).

De la Tabla 12.11 se puede ver que todas las probabilidades condicionales salvo la $p(x_{13}|x_4)$ satisfacen la condición de compatibilidad. Por ello, el enlace $X_4 \rightarrow X_{13}$ puede ser eliminado (o invertido) del grafo. Esta información redundante se muestra gráficamente en el diagrama de la Figura 12.23 mediante el ciclo formado por los nodos X_5, X_4 y X_{13}. Este ciclo implica que en la mente del ingeniero, el estado de agrietamiento por flexión afecta a la corrosión del acero, ésta afecta a la posición de la peor grieta por flexión, y ésta última afecta al estado de agrietamiento por flexión.

Por ello, se puede eliminar este ciclo sin afectar a la asignación de la probabilidad conjunta. En consecuencia se ha invertido la dirección del enlace $X_4 \rightarrow X_{13}$, con lo que se obtiene otro grafo sin ciclos (Figura 12.24), lo que nos permite definir la función de probabilidad conjunta de los nodos sin restricciones en la selección de las probabilidades condicionales. Nótese que la inversión de este enlace exige cambiar las probabilidades condicionales

para X_4 y X_{13} en la Tabla 12.11. Las probabilidades condicionales de X_4 tienen que cambiarse de $p(x_4|x_{24}, x_8, x_5)$ a

$$p(x_4|x_{24}, x_8, x_5, x_{13}) = B(2, p_4(x_{24}, x_8, x_5, x_{13})). \tag{12.18}$$

Análogamente, las de X_{13} deben cambiar de $p(x_{13}|x_4)$ a

$$p(x_{13}) = B(3, 0.2). \tag{12.19}$$

Por ello, se llega a un conjunto de probabilidades condicionales en forma canónica estándar y la asignación de probabilidad no ofrece problemas de compatibilidad. En otras palabras, las probabilidades condicionales pueden elegirse libremente.

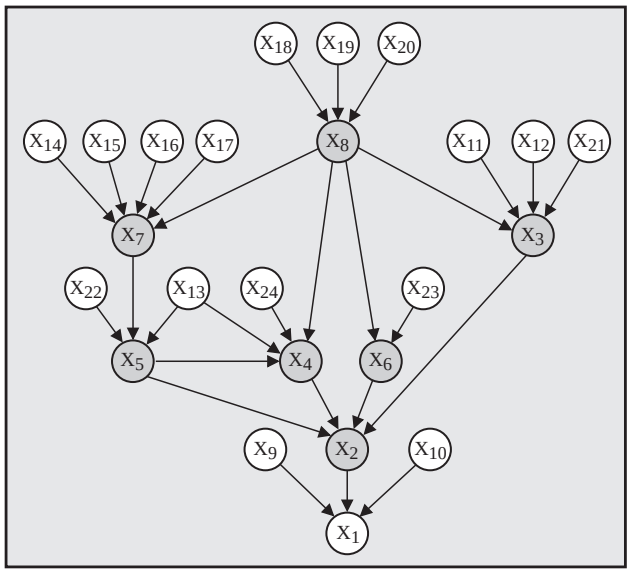


FIGURA 12.24. La red de la Figura 12.23 tras invertir el enlace de X_4 a X_{13}.

12.4.6 Propagación de Evidencia

En este ejemplo se trata con variables discretas y continuas en la misma red. Por ello, se necesita un método de propagación de evidencia para tratar este tipo de red. El caso de variables continuas complica las cosas porque las sumas deben reemplazarse por integrales y el número de posibles resultados se hace infinito. Los métodos de propagación exacta no pueden ser usados aquí porque son aplicables sólo cuando las variables son discretas o pertenecen a familias simples (tales como la normal, véase el Capítulo 8), y no existen métodos generales para redes mixtas de variables (para un caso especial véase Lauritzen y Wermouth (1989)).

Evidencia Disponible	$p(X_1 = x_1 \mid evidencia)$				
	$x_1 = 0$	$x_1 = 1$	$x_1 = 2$	$x_1 = 3$	$x_1 = 4$
Ninguna	0.3874	0.1995	0.1611	0.1235	0.1285
$X_9 = 0.01$	0.5747	0.0820	0.1313	0.1002	0.1118
$X_{10} = 0$	0.6903	0.0651	0.0984	0.0606	0.0856
$X_{11} = 3$	0.6154	0.0779	0.1099	0.0783	0.1185
$X_{12} = 2$	0.5434	0.0914	0.1300	0.0852	0.1500
$X_{13} = 3$	0.3554	0.1033	0.1591	0.1016	0.2806
$X_{14} = 2$	0.3285	0.1052	0.1588	0.1043	0.3032
$X_{15} = 9.99$	0.3081	0.1035	0.1535	0.1096	0.3253
$X_{16} = 1$	0.2902	0.1054	0.1546	0.1058	0.3440
$X_{17} = 99.9$	0.2595	0.1029	0.1588	0.1064	0.3724
$X_{18} = 99.9$	0.2074	0.1027	0.1513	0.1010	0.4376
$X_{19} = 2$	0.1521	0.0937	0.1396	0.0908	0.5238
$X_{20} = 2$	0.1020	0.0813	0.1232	0.0786	0.6149
$X_{21} = 3$	0.0773	0.0663	0.1062	0.0698	0.6804
$X_{22} = 3$	0.0325	0.0481	0.0717	0.0437	0.8040
$X_{23} = 3$	0.0000	0.0000	0.0000	0.0001	0.9999
$X_{24} = 3$	0.0000	0.0000	0.0001	0.0000	0.9999

TABLA 12.13. Distribución aproximada del daño, X_1, dadas las evidencias acumuladas de $x_9, \ldots, x_{24}$ tal como indica la tabla. Los resultados se basan en 10000 réplicas.

Sin embargo, se pueden utilizar los métodos de propagación aproximada del Capítulo 9. Por su eficacia computacional y generalidad, se elige el de la verosimilitud pesante. La propagación de evidencia se hace usando el conjunto de probabilidades marginales y condicionales de la Tabla 12.11 tal como ha sido modificada en (12.18) y (12.19). Para ilustrar la propagación de evidencia y para responder a ciertas preguntas del ingeniero, se supone que éste examina una viga de hormigón y obtiene los valores $x_9, \ldots, x_{24}$ correspondientes a las variables observables $X_9, \ldots, X_{24}$. Nótese que estos valores pueden medirse secuencialmente. En este caso, la inferencia puede hacerse también secuencialmente. La Tabla 12.13, muestra las probabilidades de daño X_1 de una viga dada para varios tipos de evidencia que van desde la evidencia nula al conocimiento de los valores que toman todas las variables $x_9, \ldots, x_{24}$. Los valores de la Tabla 12.13 se explican e interpretan a continuación.

Como ejemplo ilustrativo, supóngase que se desea determinar el daño (la variable objetivo X_1) en cada una de las situaciones siguientes:

- **No hay evidencia disponible.** La fila correspondiente a la evidencia acumulada "Ninguna" de la Tabla 12.13 da la probabilidad marginal inicial de cada uno de los estados de la variable objetivo

X_1. Por ejemplo, la probabilidad de que una viga seleccionada al azar no esté dañada ($X_1 = 0$) es 0.3874 y la probabilidad de que esté seriamente dañada ($X_1 = 4$) es 0.1285. Estas probabilidades pueden ser interpretadas como que el 39% de las vigas son seguras y el 13% están seriamente dañadas.

- **Evidencia de daño alto.** Supóngase que se tienen los datos de todas las variables observables que se dan en la Tabla 12.13, donde la evidencia se obtiene secuencialmente en el orden dado en la tabla. Las probabilidades en la fila i-ésima de la Tabla 12.13 se calculan usando $x_9, \ldots, x_i$, es decir, se basan en evidencias acumuladas. Excepto para las variables clave X_9 y X_{10}, los valores de las restantes variables alcanzan valores altos, lo que da lugar a altas probabilidades de daño. Como puede verse en la última fila de la tabla, cuando se consideran todas las evidencias, se obtiene $p(X_1 = 4) \simeq 1$, una indicación de que la viga está seriamente dañada.

- **Evidencia de daño bajo.** Ahora, supóngase que se tienen los datos de las variables observables dados en la Tabla 12.14, donde los datos se miden secuencialmente en el orden dado en la tabla. En este caso todas las variables toman valores bajos, lo que indica que la viga está en buenas condiciones. Cuando se considera toda la evidencia, la probabilidad de ausencia daño es tan alta como 1.

- **Observando una evidencia clave.** Supóngase que únicamente se observa el valor de una variable clave X_9, la debilidad de la viga, y resulta ser $X_9 = 8$, una indicación de que la viga es muy débil. Propagando la evidencia $X_9 = 8$ se obtiene

$$\begin{array}{ll} p(X_1 = 0|X_9 = 8) = 0.0012, & p(X_1 = 1|X_9 = 8) = 0.0168, \\ p(X_1 = 2|X_9 = 8) = 0.0981, & p(X_1 = 3|X_9 = 8) = 0.3320, \\ p(X_1 = 4|X_9 = 8) = 0.5519. & \end{array}$$

  Nótese que tras observar la evidencia $X_9 = 8$, $p(X_1 = 4)$ ha aumentado de 0.1285 a 0.5519 y $p(X_1 = 0)$ ha disminuido de 0.3874 a 0.0012. La razón es que (12.15) se ha usado para evaluar la probabilidad condicional

$$p(X_1 = x_1|X_9 = x_9, X_{10} = x_{10}, X_2 = x_2).$$

  La función en (12.15) es similar a una puerta O, lo que significa que un valor muy alto de uno de los tres padres de X_1 es suficiente para que X_1 sea muy alto.

- **Observando evidencia parcial.** Finalmente, supóngase que la evidencia disponible consiste en un subconjunto de las variables observables, tal como se muestra en la Tabla 12.15. Las probabilidades se muestran en la Tabla 12.15 y pueden interpretarse de forma similar

Evidencia Disponible	$p(X_1 = x_1 \mid evidencia)$				
	$x_1 = 0$	$x_1 = 1$	$x_1 = 2$	$x_1 = 3$	$x_1 = 4$
Ninguna	0.3874	0.1995	0.1611	0.1235	0.1285
$X_9 = 0$	0.5774	0.0794	0.1315	0.1002	0.1115
$X_{10} = 0$	0.6928	0.0630	0.0984	0.0603	0.0855
$X_{11} = 0$	0.7128	0.0550	0.0872	0.0615	0.0835
$X_{12} = 0$	0.7215	0.0571	0.0883	0.0551	0.0780
$X_{13} = 0$	0.7809	0.0438	0.0685	0.0469	0.0599
$X_{14} = 0$	0.7817	0.0444	0.0686	0.0466	0.0587
$X_{15} = 0$	0.7927	0.0435	0.0680	0.0441	0.0517
$X_{16} = 0$	0.7941	0.0436	0.0672	0.0421	0.0530
$X_{17} = 0$	0.8030	0.0396	0.0630	0.0428	0.0516
$X_{18} = 0$	0.8447	0.0330	0.0525	0.0316	0.0382
$X_{19} = 0$	0.8800	0.0243	0.0434	0.0269	0.0254
$X_{20} = 0$	0.9079	0.0217	0.0320	0.0217	0.0167
$X_{21} = 0$	0.9288	0.0166	0.0274	0.0172	0.0100
$X_{22} = 0$	0.9623	0.0086	0.0125	0.0092	0.0074
$X_{23} = 0$	0.9857	0.0030	0.0049	0.0037	0.0027
$X_{24} = 0$	1.0000	0.0000	0.0000	0.0000	0.0000

TABLA 12.14. Probabilidades aproximadas del daño, X_1, dada la evidencia acumulada de $x_9, \ldots, x_{24}$ como se indica en la tabla. Los resultados se basan en 10000 réplicas.

Evidencia Disponible	$p(X_1 = x_1 \mid evidencia)$				
	$x_1 = 0$	$x_1 = 1$	$x_1 = 2$	$x_1 = 3$	$x_1 = 4$
Ninguna	0.3874	0.1995	0.1611	0.1235	0.1285
$X_{11} = 2$	0.3595	0.1928	0.1711	0.1268	0.1498
$X_{12} = 2$	0.3144	0.1868	0.1700	0.1427	0.1861
$X_{21} = 2$	0.2906	0.1748	0.1784	0.1473	0.2089
$X_{18} = 80$	0.2571	0.1613	0.1764	0.1571	0.2481
$X_{19} = 2$	0.2059	0.1434	0.1797	0.1549	0.3161
$X_{20} = 1$	0.1835	0.1431	0.1716	0.1575	0.3443

TABLA 12.15. Probabilidades aproximadas del daño de la viga, X_1, dada la evidencia acumulada de las variables indicadas en la tabla. Los resultados se basan en 10000 réplicas.

Puede verse de los ejemplos anteriores que cualquier pregunta hecha por los ingenieros puede ser contestada sin más que propagar la evidencia disponible mediante el método de la verosimilitud pesante. Nótese también que es posible realizar el proceso de inferencia de forma secuencial, a medida que se va conociendo la evidencia. Una ventaja de esto es que se puede

tomar una decisión con respecto al daño de la viga inmediatamente después de observar un subconjunto de variables. Por ejemplo, nada más observar un valor alto de X_9 ó X_{10}, la inspección podría interrumpirse y declarar la viga como seriamente dañada.

12.5 Daño en Vigas de Hormigón Armado: El Modelo Normal

12.5.1 Especificación del modelo

En esta sección se presenta una formulación alternativa al ejemplo de daño en vigas de hormigón armado introducido en la Sección 12.4. Aquí se supone que todas las variables son continuas y se distribuyen según una distribución normal.

Es importante notar que en la práctica diferentes especialistas pueden desarrollar diferentes estructuras de dependencia para el mismo problema. Por otra parte, el desarrollo de una red probabilística consistente y no redundante es una tarea dura, a menos que el problema pueda ser descrito mediante una red Bayesiana o Markoviana, que automáticamente conducen a consistencia. En la Sección 12.4 se ha estudiado este problema desde un punto de vista práctico, describiendo las etapas a seguir para generar un diagrama causa-efecto único y consistente. Se ha visto en la Figura 12.23 que la red inicial dada por el ingeniero contiene un ciclo $X_4 - X_{13} - X_5 - X_4$. El ciclo se ha eliminado mediante la inversión de la dirección del enlace $X_4 \rightarrow X_{13}$. Ahora se supone que la función de densidad conjunta de $X = \{X_1, X_2, \ldots, X_{24}\}$ es normal multivariada $N(\mu, \Sigma)$, donde μ es el vector de medias de dimensión 24, Σ es la matriz de covarianzas de dimensión 24×24, y las variables $X_1, \ldots, X_{24}$ se miden utilizando una escala continua que es consistente con la hipótesis de normalidad. Entonces, tal como se ha visto en el Capítulo 6, la función de densidad conjunta de X puede escribirse como

$$f(x_1, \ldots, x_{24}) = \prod_{i=1}^{24} f_i(x_i|\pi_i), \tag{12.20}$$

donde

$$f_i(x_i|\pi_i) \sim N\left(m_i + \sum_{j=1}^{i-1} \beta_{ij}(x_j - \mu_j); v_i\right), \tag{12.21}$$

m_i es la media condicional de X_i, v_i es la varianza condicional de X_i dados los valores de Π_i, y β_{ij} es el coeficiente de regresión asociado a X_i y X_j (véase la Sección 6.4.4). Nótese que si $X_j \notin \Pi_i$ entonces $\beta_{ij} = 0$.

Alternativamente, se puede definir la función de densidad conjunta dando el vector de medias y la matriz de covarianzas. En el Capítulo 6 se ha dado un método para pasar de una a otra forma de representación.

Por ello, se puede considerar el grafo de la Figura 12.24 como la estructura de una red Bayesiana normal. Entonces, la etapa siguiente consiste en la definición de la función de densidad conjunta usando (12.20). Supóngase que las medias iniciales de todas las variables son ceros, los coeficientes β_{ij} de (12.21) se definen como se indica en la Figura 12.25, y las varianzas condicionales están dadas por

$$v_i = \begin{cases} 10^{-4}, & \text{si } X_i \text{ es no observable,} \\ 1, & \text{en otro caso.} \end{cases}$$

Entonces la red Bayesiana normal está dada por (12.20). En lo que sigue se dan ejemplos que ilustran la propagación numérica y simbólica de evidencia.

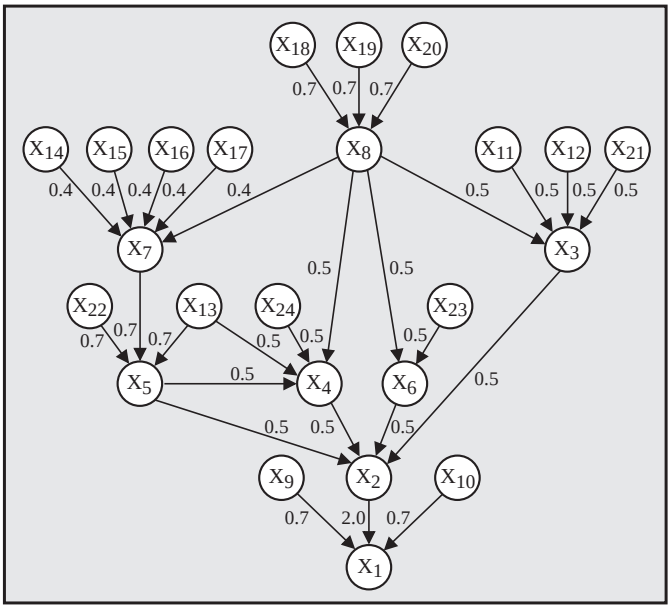


FIGURA 12.25. Grafo dirigido para evaluar el daño de una viga de hormigón armado. Los números cercanos a los enlaces son los coeficientes de regresión β_{ij} en (12.21) usados para definir la red Bayesiana.

12.5.2 Propagación Numérica de Evidencia

Para propagar evidencia en la red Bayesiana anterior, se usa el algoritmo incremental descrito en el Capítulo 8. Para ilustrar el proceso, se supone que el ingeniero examina una viga y obtiene secuencialmente los valores $\{x_9, x_{10}, \ldots, x_{24}\}$ correspondientes a las variables observables $X_9, \ldots, X_{24}$. Por simplicidad, supóngase que la evidencia es $e = \{X_9 = 1, \ldots, X_{24} = 1\}$, que indica que la viga está seriamente dañada.

De nuevo, se desea evaluar el daño (la variable objetivo, X_1). El vector de medias y la matriz de covarianzas condicionales de las variables restantes $Y = (X_1, \ldots, X_8)$ dado e, que se han obtenido usando el algoritmo incremental, son

$$E(y|e) = (1.8, 2.32, 1.4, 3.024, 3.412, 2.4, 5.118, 11.636),$$

$$Var(y|e) = \begin{pmatrix} 0.00010 & \ldots & 0.00006 & 0.00005 & 0.00009 & 0.00019 \\ 0.00004 & \ldots & 0.00006 & 0.00002 & 0.00009 & 0.00018 \\ 0.00005 & \ldots & 0.00003 & 0.00003 & 0.00010 & 0.00020 \\ 0.00003 & \ldots & 0.00009 & 0.00001 & 0.00014 & 0.00028 \\ 0.00006 & \ldots & 0.00018 & 0.00003 & 0.00017 & 0.00033 \\ 0.00005 & \ldots & 0.00003 & 0.00012 & 0.00010 & 0.00020 \\ 0.00010 & \ldots & 0.00017 & 0.00010 & 0.00035 & 0.00070 \\ 0.00013 & \ldots & 0.00033 & 0.00019 & 0.00072 & 1.00100 \end{pmatrix}.$$

Por ello, la distribución condicional de las variables en Y es normal con el vector de medias y la matriz de covarianzas anterior.

Nótese que en este caso, todos los elementos de la matriz de covarianzas excepto la varianza condicionada de X_1 son cercanos a cero, lo que indica que los valores medios son muy buenos estimadores de $E(X_2, \ldots, X_8)$ y razonables de $E(X_1)$.

Se puede considerar también la evidencia en forma secuencial. La Tabla 12.16 muestra la media y la varianza condicionales de X_1 suponiendo que la evidencia se obtiene secuencialmente en el orden indicado en la tabla. La evidencia oscila desde ausencia total de evidencia a un completo conocimiento de todas las variables $X_9, X_{10}, \ldots, X_{24}$. Por ejemplo, la media y la varianza inicial de X_1 son $E(X_1) = 0$ y $Var(X_1) = 12.861$, respectivamente; y la media y la varianza condicionales de X_1 dado $X_9 = 1$ son $E(X_1|X_9 = 1) = 0.70$ y $Var(X_1|X_9 = 1) = 12.371$. Nótese que tras observar la ecidencia clave $X_9 = 1$, la media de X_1 aumenta de 0 a 0.7 y la varianza decrece de 12.861 a 12.371. Como puede verse en la última fila de la tabla, cuando se consideran todas las evidencias, $E(X_1|X_9 = 1, \ldots, X_{24} = 1) = 12.212$ y $Var(X_1|X_9 = 1, \ldots, X_{24} = 1) = 1.001$, una indicación de que la viga está seriamente dañada. En la Figura 12.26 se muestran varias de las funciones de densidad de X_1 resultantes de añadir nuevas evidencias. La figura muestra el daño creciente de la viga en las diferentes etapas, tal como cabría esperar. Nótese que la media aumenta y la varianza disminuye, una indicación de que la incertidumbre decrece.

Puede verse de los ejemplos anteriores que cualquier pregunta hecha por el ingeniero puede ser contestada simplemente mediante la propagación de evidencia usando el algoritmo incremental.

12.5.3 Cálculo Simbólico

Supóngase ahora que se está interesado en analizar el efecto de la flecha de la viga, X_{10}, en la variable objetivo, X_1. Entonces, se considera X_{10}

Etapa	Evidencia Disponible	Daño	
		Media	Varianza
0	None	0.000	12.861
1	$X_9 = 1$	0.700	12.371
2	$X_{10} = 1$	1.400	11.881
3	$X_{11} = 1$	1.900	11.631
4	$X_{12} = 1$	2.400	11.381
5	$X_{13} = 1$	3.950	8.979
6	$X_{14} = 1$	4.370	8.802
7	$X_{15} = 1$	4.790	8.626
8	$X_{16} = 1$	5.210	8.449
9	$X_{17} = 1$	5.630	8.273
10	$X_{18} = 1$	6.974	6.467
11	$X_{19} = 1$	8.318	4.660
12	$X_{20} = 1$	9.662	2.854
13	$X_{21} = 1$	10.162	2.604
14	$X_{22} = 1$	11.212	1.501
15	$X_{23} = 1$	11.712	1.251
16	$X_{24} = 1$	12.212	1.001

TABLA 12.16. Medias y varianzas del daño, X_1, dada la evidencia acumulada de $x_9, x_{10}, \ldots, x_{24}$.

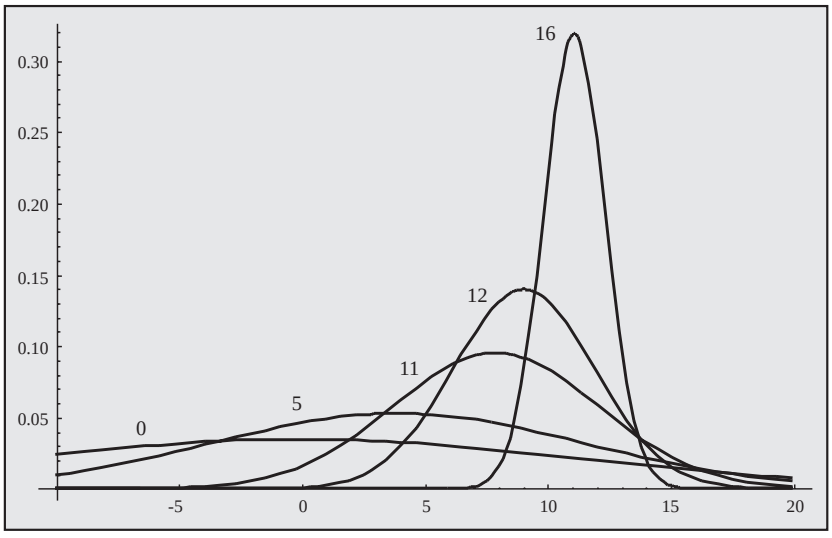


FIGURA 12.26. Distribuciones condicionadas del nodo X_1 correspondientes a la evidencia acumulada de la Tabla 12.16. El número de la etapa se muestra cerca de cada gráfica.

Evidencia Disponible	Daño	
	Media	Varianza
Ninguna	0	12.861
$X_9 = 1$	0.7	12.371
$X_{10} = 1$	$\frac{c - cm + 0.7v}{v}$	$\frac{-c^2 + 12.371v}{v}$
$X_{11} = x_{11}$	$\frac{c - cm + 0.7v + 0.5vx_{11}}{v}$	$\frac{-c^2 + 12.121v}{v}$
$X_{12} = 1$	$\frac{c - cm + 1.2v + 0.5vx_{11}}{v}$	$\frac{-c^2 + 11.871v}{v}$
$X_{13} = x_{13}$	$\frac{c - cm + 1.2v + 0.5vx_{11} + 1.55vx_{13}}{v}$	$\frac{-c^2 + 9.467v}{v}$
$X_{14} = 1$	$\frac{c - cm + 1.62v + 0.5vx_{11} + 1.55vx_{13}}{v}$	$\frac{-c^2 + 9.292v}{v}$

TABLA 12.17. Medias y varianzas condicionales de X_1, inicialmente y tras la evidencia acumulada.

como un nodo simbólico. Sea $E(X_{10}) = m, Var(X_{10}) = v, Cov(X_{10}, X_1) = Cov(X_1, X_{10}) = c$. Las medias y varianzas condicionales de todos los nodos pueden calcularse aplicando el algoritmo para propagación simbólica en redes Bayesianas normales, introducido en el Capítulo 10. Las medias y varianzas condicionales de X_1 dadas las evidencias secuenciales $X_9 = 1, X_{10} = 1, X_{11} = x_{11}, X_{12} = 1, X_{13} = x_{13}, X_{14} = 1$, se muestran en la Tabla 12.17. Nótese que algunas evidencias (X_{11}, X_{13}) se dan en forma simbólica.

Nótese que los valores de la Tabla 12.16 son un caso especial de los de la Tabla 12.17. Pueden ser obtenidos haciendo $m = 0$, $v = 1$ y $c = 0.7$ y considerando los valores evidenciales $X_{11} = 1, X_{13} = 1$. Por ello, las medias y varianzas de la Tabla 12.16 pueden en realidad obtenerse de la Tabla 12.17 sin más que reemplazar los parámetros por sus valores. Por ejemplo, para el caso de la evidencia $X_9 = 1, X_{10} = 1, X_{11} = x_{11}$, la media condicional de X_1 es $(c - cm + 0.7v + 0.5vx_{11})/v = 1.9$. Similarmente, la varianza condicional de X_1 es $(-c^2 + 12.121v)/v = 11.631$.

D	F	$p(g\|D,F)$
d	f	θ_1
d	$\bar{f}$	θ_2
$\bar{d}$	f	1
$\bar{d}$	$\bar{f}$	0

E	G	$p(h\|E,G)$
e	g	1
e	$\bar{g}$	1
$\bar{e}$	g	θ_3
$\bar{e}$	$\bar{g}$	θ_4

TABLA 12.18. Probabilidades condicionales modificadas de los fallos de las variables G y H en el sistema del tanque de presión con causas comunes de fallo.

Ejercicios

12.1 Definir la base de conocimiento para un sistema experto correspondiente al ejemplo del tanque de presión introducido en la Sección 12.2. Seguidamente:

(a) Usar el Algoritmo de encadenamiento 2.1 para sacar nuevas conclusiones cuando se conocen los hechos:
   i. $D = d$, $E = e$ y $F = f$.
   ii. $A = a$, $D = d$, $E = e$ y $F = f$.
   iii. $D = \bar{d}$, $E = e$ y $F = \bar{f}$.
   iv. $A = a$, $D = \bar{d}$, $E = e$ y $F = \bar{f}$.

(b) Usar el algoritmo de encadenamiento de reglas orientado a un objetivo para obtener las causas que pueden dar lugar al fallo del tanque.

12.2 Usar el Algoritmo de agrupamiento 8.5 para propagar evidencia en el ejemplo del tanque de presión dado en la Sección 12.2. Verificar los resultados de las Figuras 12.5, 12.6 y 12.7, que fueron obtenidos usando el (más eficiente) Algoritmo para poliárboles 8.1.

12.3 Reemplazar las probabilidades de los nodos G y H en la Tabla 12.2 por los dados en la Tabla 12.18. Usar los métodos simbólicos para calcular las probabilidades condicionales de los nodos dada la evidencia $F = f$.

12.4 Definir la base de conocimiento para un sistema experto basado en reglas correspondiente al ejemplo del sistema de distribución de energía dado en la Sección 12.3. Seguidamente:

(a) Usar el Algoritmo de encadenamiento 2.1 para sacar nuevas conclusiones cuando se conocen los nuevos hechos:
   i. $A = a$, $G = g$ y $M = m$.
   ii. $A = a$, $G = g$, $M = m$ y $F = f$.

I	H	$p(j\|I,H)$
i	h	θ_1
i	$\bar{h}$	1
$\bar{i}$	h	θ_2
$\bar{i}$	$\bar{h}$	0

E	K	$p(l\|E,K)$
e	k	θ_3
e	$\bar{k}$	θ_4
$\bar{e}$	k	0
$\bar{e}$	$\bar{k}$	0

TABLA 12.19. Probabilidades condicionales modificadas del fallo de las variables J y L en el sistema de distribución de energía (motor 2).

iii. $A = \bar{a}$, $G = \bar{g}$ y $M = m$.

iv. $A = \bar{a}$, $G = \bar{g}$, $M = m$ y $F = f$.

(b) Usar el algoritmo de encadenamiento de reglas orientado a un objetivo para obtener las causas que pueden dar lugar al fallo del motor 2.

12.5 Reemplazar las probabilidades de los nodos J y L de la Tabla 12.7 por los de la Tabla 12.19. Usar los métodos simbólicos para calcular las probabilidades condicionales de los nodos dada la evidencia $K = k$.

12.6 Considérese el sistema de distribución de energía de la Figura 12.14 y supóngase que se está interesado en conocer el estado de operación de los motores 1 y 3. Repetir las etapas descritas en la Sección 12.3 para el motor 1 ó el motor 3.

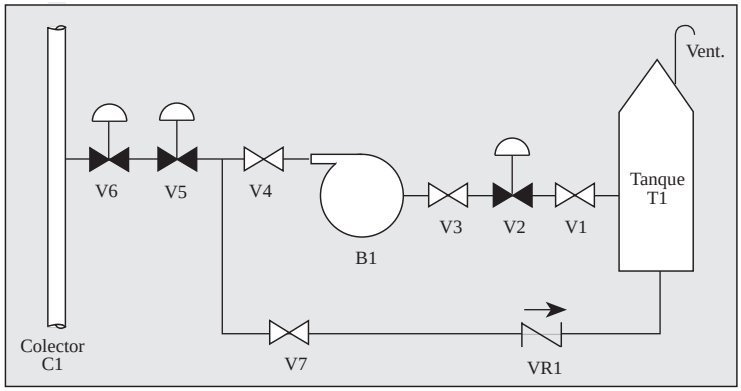


FIGURA 12.27. Diagrama simplificado de un sistema de suministro de agua en espera.

12.7 La Figura 12.27 muestra un diagrama simplificado de un sistema de suministro de agua en espera. El sistema suministra agua procedente del tanque $T1$ al colector $C1$. Para que funcione correctamente, se usa

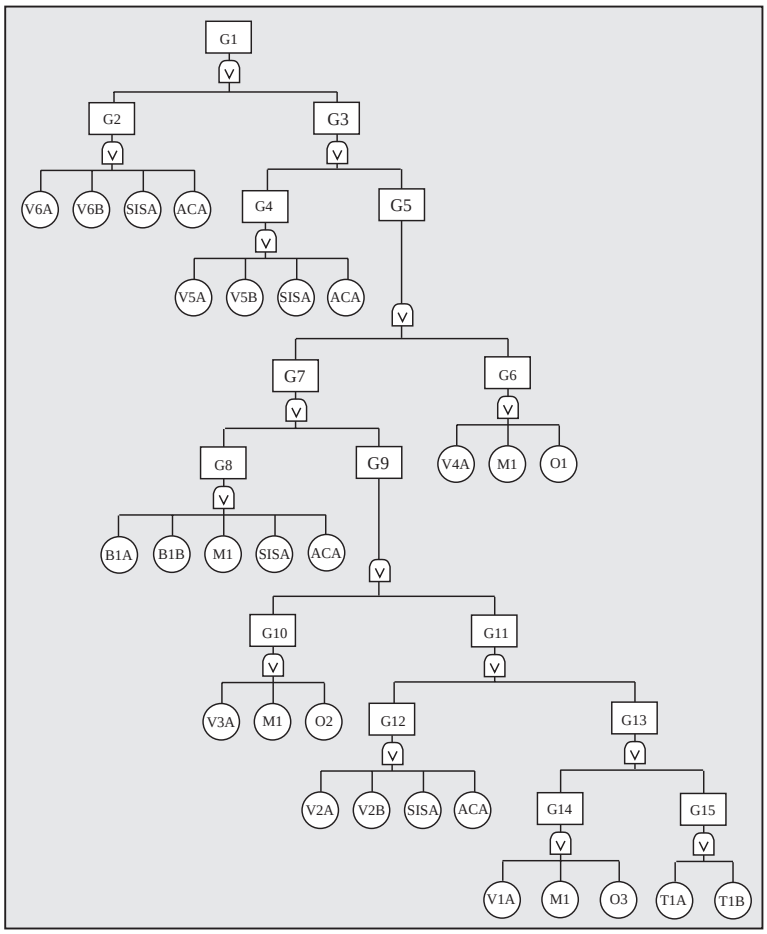


FIGURA 12.28. El árbol de fallos del sistema de suministro de agua.

la bomba $B1$ para abrir las válvulas motorizadas $V2$, $V5$ y $V6$. Nótese que todas las válvulas mostradas en la Figura 12.27 están en sus posiciones normales (sistema en espera). La bomba $B1$ se comprueba una vez al mes. Durante la prueba, se arranca la bomba y se hace funcionar durante diez minutos, permitiendo al agua fluir a través de la válvula manual $V7$ y la válvula de retención $VR1$, tras abrir la válvula motorizada $V2$, que devuelve el agua al tanque $T1$.

El sistema funciona correctamente incluso cuando la válvula $V7$ está abierta. El mantenimiento de la bomba $B1$ se hace dos veces por año. Durante el tiempo que dura el mantenimiento, las válvulas $V3$ y $V4$ están cerradas, y el sistema permanece no disponible durante el periodo de mantenimiento, que dura siete horas. Se está interesado en

Variable	Definición
ACA	Fallo del suministro eléctrico
$B1A$	Bomba $B1$ no arranca
$B1B$	Bomba $B1$ falla tras el arranque
$G1$	Colector no recibe agua
$G2$	Válvula $V6$ no se abre
$G3$	Válvula $V6$ no recibe agua
$G4$	Válvula $V5$ no se abre
$G5$	Válvula $V5$ no recibe agua
$G6$	Válvula $V4$ está cerrada
$G7$	Válvula $V4$ no recibe agua
$G8$	Bomba $B1$ falla
$G9$	Bomba $B1$ no recibe agua
$G10$	Válvula $V3$ está cerrada
$G11$	Válvula $V3$ no recibe agua
$G12$	Válvula $V2$ no se abre
$G13$	Válvula $V2$ no recibe agua
$G14$	Válvula $V1$ está cerrada
$G15$	Válvula $V1$ no recibe agua
$M1$	Elemento fuera de servicio por mantenimiento
$O1$	Operador olvida abrir la válvula $V4$ tras el mantenimiento
$O2$	Operador olvida abrir la válvula $V3$ tras el mantenimiento
$O3$	Operador olvida abrir la válvula $V1$ tras el mantenimiento
$SISA$	Fallo de señal lógica
$T1A$	Fallo del tanque
$T1B$	Fallo de la ventilación del tanque
$V1A$	Válvula $V1$ está bloqueada
$V2A$	Fallo mecánico en válvula $V2$
$V2B$	Válvula $V2$ está bloqueada
$V3A$	Válvula $V3$ está bloqueada
$V4A$	Válvula $V4$ está bloqueada
$V5A$	Fallo mecánico en válvula $V5$
$V5B$	Válvula $V5$ está bloqueada
$V6A$	Fallo mecánico en válvula $V6$
$V6B$	Válvula $V6$ está bloqueada

TABLA 12.20. Variables y sus correspondientes significados.

determinar la indisponibilidad del sistema. Las variables relevantes y sus definiciones se dan en la Tabla 12.20.

La Figura 12.28 muestra el diagrama de fallos correspondiente. Nótese que algunos nodos aparecen repetidos (por ejemplo, $M1$, $SISA$ y ACA) para mantener la estructura de árbol.

La Figura 12.29 muestra un grafo dirigido acíclico que evita la replicación de nodos del árbol de fallos y muestra la estructura de dependencia correspondiente. Se ha supuesto que hay una causa común de fallo para las válvulas $V2A$, $V5A$ y $V6A$. Calcular la fiabiliadd del colector $C1$, sin y con causas comunes de fallo.

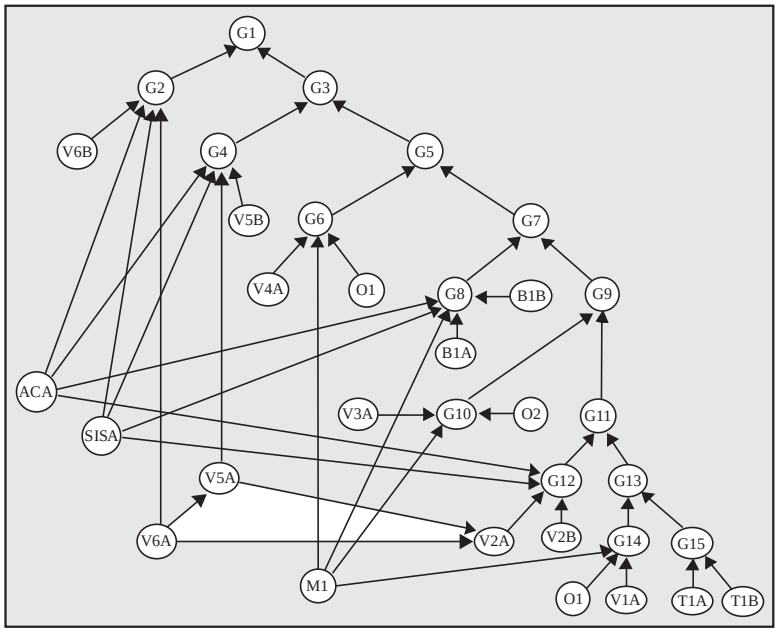


FIGURA 12.29. Un grafo dirigido para el sistema de suministro de agua suponiendo una causa común de fallo para las válvulas $V2A$, $V5A$ y $V6A$.

Notación

En este apéndice se presenta la notación utilizada en este libro. Se ha tratado de mantener la notación lo más consistente posible. Por ello, primero se presenta la notación común, utilizada en todos los capítulos y, a continuación, se describe la notación específica de cada capítulo.

Notación Común

$A \subseteq B$ ..... A es un subconjunto de B
$A \subset B$ ..... A es un subconjunto propio de B
A^T ..... Traspuesta de la matriz A
A^{-1} ..... Inversa de la matriz A
A_i ..... Variables posteriores a Y_i, $\{Y_{i+1}, \ldots, Y_n\}$
B_i ..... Variables anteriores a Y_i, $\{Y_1, \ldots, Y_{i-1}\}$
$card(A)$ ..... Número de elementos (cardinal) del conjunto A
D ..... Grafo dirigido
$det(A)$ ..... Determinante de la matriz A
$D(X, Y|Z)_G$ .... X es condicionalmente dependiente de Y dado Z en G
$D(X, Y|Z)_P$ .... X es condicionalmente dependiente de Y dado Z en P
G ..... Grafo (dirigido o no dirigido)
$I(X, Y|Z)_G$ ..... X es condicionalmente independiente de Y dado Z en G
$I(X, Y|Z)_P$ ..... X es condicionalmente independiente de Y dado Z en P
M ..... Lista inicial o modelo de dependencia
$M_{ij}(s_{ij})$ ..... Mensaje enviado del conglomerado C_i al C_j
n ..... Número de variables o nodos
$p(c_i)$ ..... Función de probabilidad del conglomerado C_i
$p(x_i|\pi_i)$ ..... Probabilidad condicional de $X_i = x_i$ dado $\Pi_i = \pi_i$

$p(X = x|Y = y)$ Probabilidad condicional de $X = x$ dado $Y = y$
$p(x_1, \ldots, x_n)$ .... Función de probabilidad conjunta
$tr(A)$ ..... Traza de la matriz A
X ..... Conjunto de variables o nodos
$X \setminus Y$ ..... X menos Y (diferencia de conjuntos)
$\{X_1, \ldots, X_n\}$ ... Conjunto de variables o nodos
X_i ..... Variable i-ésima del conjunto $\{X_1, \ldots, X_n\}$
$X_i - X_j$ ..... Arista no dirigida entre los nodos X_i y X_j
$X_i \notin S$ ..... X_i no está contenida en S
$X_i \rightarrow X_j$ ..... Arista dirigida de X_i a X_j
WWW ..... World Wide Web
$\bar{A}$ ..... Conjunto complementario de A
$\cap$ ..... Intersección de conjuntos
$\cup$ ..... Unión de conjuntos
ϕ ..... Conjunto vacío
Π_i ..... Conjunto de padres del nodo X_i
π_i ..... Realización de los padres, Π_i, de X_i
$\Psi_i(c_i)$ ..... Función potencial del conglomerado C_i
μ_X ..... Media de X
Σ_X ..... Matriz de covarianzas de X
σ_i^2 ..... Varianza de X_i
σ_{ij} ..... Covarianza de X_i y X_j
Σ_{XX} ..... Matriz de covarianzas de X
Σ_{XY} ..... Covarianza de X e Y
$\exists$ ..... Existe
$\prod_{i=1}^{n} x_i$ ..... $x_1 \times \ldots \times x_n$
$\sum_{i=1}^{n} x_i$ ..... $x_1 + \ldots + x_n$
$\propto$ ..... Proporcional a
$\wedge$ ..... y
$\vee$ ..... o

Capítulo 1

E_i ..... Enfermedad i-ésima
IA ..... Inteligencia Artificial
S_i ..... Síntoma i-ésimo

Capítulo 2

A, B, C ..... Objectos de un sistema experto basado en reglas
F ..... Falso
NIP ..... Número de identificación personal
T ..... Cierto

Capítulo 3

A_i ..... Subconjunto de un espacio muestral S
D ..... Síntoma *dolor*
E ..... Conjunto de enfermedades
e_i ..... Enfermedad i-ésima
$E(u)$ ..... Valor medio, o esperanza, de u
G ..... Adenocarcinoma gástrico
lim ..... Límite
m ..... Número de enfermedades
$\max_i(a_1, \ldots, a_k)$ Valor máximo de $\{a_1, \ldots, a_k\}$
$\min_i(a_1, \ldots, a_k)$ Valor mínimo de $\{a_1, \ldots, a_k\}$
MSD ..... Modelo de síntomas dependientes
MSD ..... Modelo de síntomas independientes
$MSRD$ ..... Modelo de síntomas relevantes dependientes
$MSRI$ ..... Modelo de síntomas relevantes independientes
n ..... Número de síntomas
P ..... Síntoma *pérdida de peso*
S ..... Espacio muestral
S_j ..... Síntoma j-ésimo
$u_i(x)$ ..... Función de utilidad
V ..... Síntoma *vómitos*

Capítulo 4

$A - D - H$ ..... Camino no dirigido entre A y H
$A \rightarrow D \rightarrow H$ .... Camino dirigido de A a H
$Ady(X_i)$ ..... Conjunto de adyacencia del nodo X_i
C ..... Conjunto de conglomerados
C_i ..... Conglomerado i-ésimo
$Frn(S)$ ..... Frontera del conjunto S
L ..... Conjunto de aristas
L_{ij} ..... Arista del nodo X_i al X_j
NA_k ..... k-ésimo nivel dirigido ascendente
ND_k ..... k-ésimo nivel dirigido descendente
$PA(X_i)$ ..... Profundidad ascendente del nodo X_i

$PD(X_i)$ ..... Profundidad descendente del nodo X_i
T ..... Matriz de alcanzabilidad
$Vec(X_i)$ ..... Vecinos del nodo X_i
$\alpha(i)$ ..... Nodo i-ésimo en el orden α

Capítulo 5

D ..... Donaciones
E ..... Situación económica
F ..... Felicidad
H ..... Salud
I ..... Ganancias por inversiones
L ..... Número de variables en U_i
m ..... Número de probabilidades condicionales $p_i(u_i|v_i)$
T ..... Situación laboral
UF ..... Unión fuerte
U_i, V_i ..... Subconjuntos disjuntos de X
X, Y, Z ..... Subconjuntos disjuntos de variables

Capítulo 6

$Frn(S)$ ..... Frontera del conjunto S
C_j ..... Conglomerado j-ésimo
L_{ij} ..... Arista entre los nodos X_i y X_j
R_i ..... Conjunto residual i-ésimo
S_i ..... Conjunto separador i-ésimo
v_A ..... Varianza de A
v_i ..... Varianza condicional de X_i dado $\Pi_i = \pi_i$
$W = \Sigma^{-1}$ ..... Inversa de la matriz de covarianzas Σ
X, Y, Z ..... Conjuntos disjuntos de variables aleatorias
β_{CA} ..... Coeficiente de A en la regresión de C sobre Π_C
β_{ij} ..... Coeficiente de X_j en la regresión de X_i sobre Π_i
μ ..... Media de una variable normal
μ_i ..... Media de X_i
μ_X ..... Media de X
Σ ..... Matriz de covarianzas de una variable normal
Σ_{XX} ..... Matriz de covarianzas de X
Σ_{XY} ..... Matriz de covarianzas de X e Y

Capítulo 7

$B = (D, P)$ ..... Red Bayesiana

(G^ℓ, P^ℓ) ..... Red probabilística ℓ-ésima
L_{ij} ..... Arista entre los nodos X_i y X_j
$p^\ell(x_i^\ell|s_i^\ell)$ ..... Probabilidades condicionadas del modelo ℓ-ésimo
R_i ..... Función de probabilidad condicionada $p(y_i|a_i)$
r_i^ℓ ..... Cardinalidad de Y_i^ℓ
S_i ..... Subconjunto de $B_i = \{Y_1, \ldots, Y_{i-1}\}$
U_i ..... Subconjunto de $A_i = \{Y_{i+1}, \ldots, Y_n\}$
X, Y, Z ..... Variables aleatorias
β_{ij} ..... Coeficiente de X_j en la regresión de X_i sobre el resto de las variables
ϵ_i ..... Variable aleatoria normal, $N(0, \sigma_i^2)$
$\theta_{ij\pi}$ ..... Parámetros $p(X_i = j|\Pi_i = \pi)$
Θ^ℓ ..... Conjunto $\theta_{ij\pi}$ para el modelo ℓ-ésimo

Capítulo 8

C_i ..... Conglomerado i-ésimo del grafo
E ..... Conjunto de nodos evidenciales
$E_i^+, E_{X_i}^+$ ..... Subconjunto de E en la componente de los padres de X_i
$E_i^-, E_{X_i}^-$ ..... Subconjunto de E en la componente de los hijos de X_i
$E_{U_j X_i}^+$ ..... Subconjunto de E en la parte U_j de la arista $U_j \to X_i$
$E_{X_i Y_j}^-$ ..... Subconjunto de E en la parte Y_j de la arista $X_i \to Y_j$
e_i ..... Valor del nodo evidencial E_i
M_{ij} ..... Mensaje que el conglomerado C_i envía a su vecino B_j
R_i ..... i-ésimo conjunto residual
S_i ..... i-ésimo conjunto separator
$S_{ij} = C_i \cap S_j$ ... Separator de los conglomerados C_i y C_j
v_i ..... Varianza condicional de X_i dado $\Pi_i = \pi_i$
$\beta_i(x_i)$ ..... $\lambda_i(x_i)\rho_i(x_i)$
$\lambda_i(x_i)$ ..... Probabilidad condicional $p(e_i^-|x_i)$
$\lambda_{Y_j X_i}(x_i)$ ..... Mensaje λ que Y_j envía a su padre X_i
μ_X ..... Media de X
$\psi_j(c_j)$ ..... Función potencial del conglomerado C_j
$\psi_j^*(c_j)$ ..... Nueva función potencial del conglomerado C_j
$\rho_i(x_i)$ ..... Función de probabilidad conjunta $p(x_i, e_i^+)$
$\rho_{U_j X_i}(u_j)$ ..... Mensaje ρ que envía el nodo U_j a su hijo X_i
Θ_i ..... Padre auxiliar del nodo X_i

Capítulo 9

B ..... Nodos utilizados en el muestreo hacia atrás
C_i ..... Conjunto de los hijos de X_i

E ..... Conjunto de nodos evidenciales
F ..... Nodos utilizados en el muestreo hacia adelante
I_J ..... Intervalo $[l_j, u_j) \subset [0, 1)$
l_j ..... Cota inferior, de valor $\sum_{x^i < x^j} p_e(x^i) \geq 0$
m ..... Número de realizaciones posibles del conjunto X
N ..... Número de simulaciones (tamaño de la muestra)
$p(e)$ ..... Probabilidad de e
$p_e(x_i|\pi_i)$ ..... ver (9.14)
$p_e(y)$ ..... ver (9.7)
R_i^j ..... Rama ij-ésima
r_i ..... Cardinalidad de X_i
$s(x)$ ..... Peso de la realización x, $p(x)/h(x)$
u ..... Valor aleatorio uniforme obtenido de $U(0, 1)$
$U(0, 1)$ ..... Distribución uniforme estándar
u_j ..... Cota superior, de valor $l_j + p_e(x^j) \leq 1$
x^j ..... Realización j-ésima, $\{x_1^j, \ldots x_n^j\}$
π_i^* ..... Conjunto de padres de X_i no simulados

Capítulo 10

Θ_i ..... Conjunto de parámetros simbólicos asociados al nodo simbólico X_i
c_r ..... Coeficientes numéricos asociados al monomio m_r
M ..... Conjunto de todos los monomios
E ..... Conjunto de nodos con evidencia
$q(e)$ ..... Probabilidad de la evidencia estocástica e
$adj(\Sigma_z)$ ..... Matriz adjunta de Σ_z

Capítulo 11

$Desc_i$ ..... Conjunto de descendientes de X_i
$Dim(B)$ ..... Dimensión del modelo de red Bayesiana B número de parámetros libres, o grados de libertad, requeridos para especificar la función de probabilidad conjunta de X)
D_c ..... Grafo dirigido completo
$f(N)$ ..... Función de penalización no negativa
$f(x, \theta)$ ..... Función de verosimilitud
$g(\theta, \eta)$ ..... η-familia de distribuciones "a priori" para θ
$g(\theta, \eta)$ ..... η-familia de distribuciones "a posteriori" para θ
$h(\eta; x)$ ..... Función que da los parámetros "a posteriori" en función de los parámetros "a priori" y la muestra x
iid ..... Independientes e idénticamente distribuidos

m_i ..... Ordenada en el origen de la regresión de X_i sobre Π_i
N ..... Tamaño de muestra
$N!$ ..... Factorial de N $(1 \times 2 \times \ldots \times N)$
$\begin{pmatrix} N \\ r \end{pmatrix}$ ..... $\frac{N!}{r!(N-r)!}$
N_{ik} ..... $\sum_{j=0}^{r_i} N_{ijk}$
N_{ijk} ..... Número de casos en la muestra S con $X_i = k$ y Π_i igual a la realización j-ésima de Π_i
$N_{x_1,\ldots,x_n}$ ..... Número de casos en la muestra S con $X_i = x_i$
$p(B) = p(D, \theta)$ .. Probabilidad "a priori" asignada al modelo de red B
$Pred_i$ ..... Conjunto de ascendientes del nodo X_i
$Q_B(B(\theta), S)$ .... Medida de calidad Bayesiana estándar $(B(\theta), S)$
$Q_f(D', S)$ ..... Medida de calidad (D', S) con función de penalización f
$Q_I(B, S)$ ..... Medida de Información
$q_i(\Pi_i)$ ..... Contribución de X_i con conjunto de padres Π_i a la calidad de la red
$Q_{CH}(D, S)$ ..... Medida de calidad de Cooper y Herskovits
$Q_{GH}(D, S)$ ..... Medida de calidad de Geiger y Heckerman
$Q_{MDL}(B, S)$ .... Medida de calidad de mínima descripción
$Q_{SB}(D, S)$ ..... Medida de calidad estándar
r_i ..... Cardinalidad de X_i
S ..... Muestra
$W\Sigma^{-1}$ ..... Matriz de precisión (inversa de la matriz de covarianzas)
$X_{(k)}$ ..... Todas las variables salvo X_k
$\bar{x}$ ..... Media aritmética de $x_1, \ldots, x_N$
β_{ij} ..... Coeficientes de regresión de la variable X_i respecto a las variables $X_1, \ldots, X_{i-1}$
η..... Tamaño de muestra equivalente de la distribución "a priori".
$\eta_{x_1,\ldots,x_n}$ ..... Parámetros de la distribución de Dirichlet
η_{ik} ..... $\sum_{j=0}^{r_i} \eta_{ijk}$
η_{ijk} ..... Parámetros "a priori" correspondientes a N_{ijk}
$\Gamma(.)$ ..... Función Gamma
$\hat{\theta}$ ..... Moda posterior de θ
μ ..... Vector de medias
Ω ..... Conjunto universal

Capítulo 12

μ ..... Vector de medias
Σ ..... Matriz de covarianzas
θ_C ..... Parámetro asociado al nodo C
$Vec(X_i)$ ..... Vecinos del nodo X_i

Referencias

Abell, M. and Braselton, J. P. (1994), *Maple V by Example*. Academic Press, New York.

Akaike, H. (1974), A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Allen, J. (1995), *Natural Language Understanding, 2nd edition*. Addison-Wesley, Reading, MA.

Allen, J., Hendler, J., and Tate, A., editors (1990), *Readings in Planning*. Morgan Kaufmann Publishers, San Mateo, CA.

Almond, R. G. (1995), *Graphical Belief Modeling*. Chapman and Hall, New York.

Almulla, M. (1995), *Analysis of the Use of Semantic Trees in Automated Theorem Proving*. Ph.D. Thesis, McGill University.

Andersen, S. K., Olesen, K. G., Jensen, F. V., and Jensen, F. (1989), HUGIN: A Shell for Building Bayesian Belief Universes for Expert Systems. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI-89)*. Morgan Kaufmann Publishers, San Mateo, CA, 1080–1085.

Anderson, T. W. (1984), *An Introduction to Multivariate Statistical Analysis, 2nd edition*. John Wiley and Sons, New York.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1992), *Conditionally Specified Distributions*. Springer-Verlag, New York.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1993), Conjugate Exponential Family Priors for Exponential Family Likelihoods. *Statistics*, 25:71–77.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1994), Priors with Convenient Posteriors. Technical Report, TR94-2, Department of Applied Mathematics, Cantabria University, Santander, Spain.

Arnold, B. C., Castillo, E., and Sarabia, J. M. (1996), Specification of Distributions by Combinations of Marginal and Conditional Distributions. *Statistics and Probability Letters*, 26:153–157.

Bachman, R. J., Levesque, H. J., and Reiter, R., editors (1991), *Artificial Intelligence, Special Volume on Knowledge Representation*, Volume 49.

Baker, M. and Boult, T. E. (1991), Pruning Bayesian Networks for Efficient Computation. In Bonissone, P. P., Henrion, M., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 6*. North Holland, Amsterdam, 225–232.

Balas, E. and Yu, C. S. (1986), Finding a Maximum Clique in an Arbitrary Graph. *SIAM Journal on Computing*, 15(4):1054–1068.

Barr, A. and Feigenbaum, E. A. (1981), *The Handbook of Artificial Intelligence, Volume I*. William Kaufman, Los Altos, CA.

Barr, A. and Feigenbaum, E. A. (1982), *The Handbook of Artificial Intelligence, Volume II*. William Kaufman, Los Altos, CA.

Becker, A. and Geiger, D. (1994), Approximation Algorithms for the Loop Cutset Problem. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 60–68.

Beeri, C., Fagin, R., Maier, D., and Yannakis, M. (1993), On the Desirability of Acyclic Database Schemes. *Journal of the ACM*, 30:479–513.

Bench-Capon, T. J. M. (1990), *Knowledge Representation: An Approach to Artificial Intelligence*. Academic Press, San Diego.

Berge, C. (1973), *Graphs and Hypergraphs*. North-Holland, Amsterdam.

Bernardo, J. and Smith, A. (1994), *Bayesian Theory*. John Wiley and Sons, New York.

Bickel, P. J. and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, Oakland, CA.

Biggs, N. L. (1989), *Discrete Mathematics, 2nd edition*. Oxford University Press, New York.

Billingsley, P. (1995), *Probability and Measure*. John Wiley and Sons, New York.

Bouckaert, R. R. (1994), A Stratified Simulation Scheme for Inference in Bayesian Belief Networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 110–117.

Bouckaert, R. R. (1995), *Bayesian Belief Networks: From Construction to Inference*. Ph.D. Thesis, Department of Computer Science, Utrecht University, Netherlands.

Bouckaert, R. R., Castillo, E., and Gutiérrez, J. M. (1996), A Modified Simulation Scheme for Inference in Bayesian Networks. *International Journal of Approximate Reasoning*, 14:55–80.

Bond, A. H. and Gasser, L., editors (1988), *Readings in Distributed Reasoning*. Morgan Kaufmann Publishers, San Mateo, CA.

Bondy, J. A. and Murty, U. S. R. (1976), *Graph Theory with Applications*. North Holland, New York.

Breese, J. S. and Fertig, K. W. (1991), Decision Making with Interval Influence Diagrams. In Bonissone, P. P., Henrion, M., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 6*. North Holland, Amsterdam, 467–478.

Brown, L. D. (1986), *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, Hayward, CA.

Brown, J. R. and Cunningham, S. (1989), *Programming the User Interface: Principles and Examples*. John Wiley and Sons, New York.

Buchanan, B. G. and Shortliffe, E. H. (1984), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA.

Buckley, J. J., Siler, W., and Tucker, D. (1986), A Fuzzy Expert System. *Fuzzy Sets and Systems*, 20:1–16.

Bundy, A. (1983), *The Computer Modelling of Mathematical Reasoning*. Academic Press, New York.

Buntine, W. (1991), Theory Refinement on Bayesian Networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Mateo, CA, 52–60.

Campos, L. M. D. and Moral, S. (1995), Independence Concepts for Convex Sets of Probabilities. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 108–115.

Cano, J., Delgado, M., and Moral, S. (1993), An Axiomatic Framework for Propagating Uncertainty in Directed Acyclic Networks. *International Journal of Approximate Reasoning*, 8:253–280.

Casella, G. and George, E. I. (1992), Explaining the Gibbs Sampler. *The American Statistician*, 46(3):167–174.

Castillo, E. and Alvarez, E. (1990), Uncertainty Methods in Expert Systems. *Microcomputers in Civil Engineering*, 5:43–58.

Castillo, E. and Alvarez, E. (1991), *Expert Systems: Uncertainty and Learning*. Computational Mechanics Publications and Elsevier Applied Science, London, U.K.

Castillo, E., Bouckaert, R., Sarabia, J. M., and Solares, C. (1995), Error Estimation in Approximate Bayesian Belief Network Inference. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 55–62.

Castillo, E., Cobo, A., Gutiérrez, J. M., Iglesias, A., and Sagástegui, H. (1994), Causal Network Models in Expert Systems. *Microcomputers in Civil Engineering, Special Issue on Uncertainty in Expert Systems*, 9:55–60.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1995a), An Introduction to Expert Systems for Medical Diagnoses. *Biocybernetics and Biomedical Engineering*, 15:63–84.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1995b), Modelling Probabilistic Networks of Discrete and Continuous Variables. Technical Report, TR95-11, Department of Social Statistics, Cornell University, Ithaca, NY.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1995c), Parametric Structure of Probabilities in Bayesian Networks. *Lecture Notes in Artificial Intelligence*, 946:89–98.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1995d), Symbolic Propagation in Discrete and Continuous Bayesian Networks. In Keranen, V. and Mitic, P., editors, *Mathematics with Vision: Proceedings of the First International Mathematica Symposium*. Computational Mechanics Publications, Southampton, U.K., 77–84.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1996a), Goal Oriented Symbolic Propagation in Bayesian Networks. In *Proceedings of the Thirteenth National Conference on AI (AAAI-96)* AAAI Press/MIT Press, Menlo Park, CA, 1263–1268.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1996b), Multiply Factorized Bayesian Network Models. Technical Report, TR96-10, Department of Social Statistics, Cornell University, Ithaca, NY.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1996c), A New Method for Efficient Symbolic Propagation in Discrete Bayesian Networks. *Networks*, 28:31–43.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1996d), Sensitivity Analysis in Discrete Bayesian Networks. *IEEE Transactions on Systems, Man and Cybernetics*, 26. In press.

Castillo, E., Gutiérrez, J. M., and Hadi, A. S. (1996e), Constructing Probabilistic Models Using Conditional Probability Distributions. Technical Report, TR96-11, Department of Social Statistics, Cornell University, Ithaca, NY.

Castillo, E., Gutiérrez, J. M., Hadi, A. S, and Solares, C. (1997), Symbolic Propagation and Sensitivity Analysis in Gaussian Bayesian Networks with Application to Damage Assessment. *Artificial Intelligence in Engineering*, 11:173-181.

Castillo, E., Hadi, A. S., and Solares, C. (1997), Learning and Updating of Uncertainty in Dirichlet Models. *Machine Learning*, 26:43–56. In press.

Castillo, E., Iglesias, A., Gutiérrez, J. M., Alvarez, E., and Cobo, A. (1993), *Mathematica*. Editorial Paraninfo, Madrid.

Castillo, E., Mora, E. and Alvarez, E. (1994), Log-Linear Models in Expert Systems. *Microcomputers in Civil Engineering, Special Issue on Uncertainty in Expert Systems*, 9:347–357.

Castillo, E., Solares, C., and Gómez, P. (1996a), Tail Sensitivity Analysis in Bayesian Networks. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 133–140.

Castillo, E., Solares, C., and Gómez, P. (1996b), Estimating Extreme Probabilities Using Tail Simulated Data. *International Journal of Approximate Reasoning*. In press.

Chandrasekaran, B. (1988), On Evaluating AI Systems for Medical Diagnosis. *AI Magazine*, 4:34–37.

Chang, K.-C. and Fung, R. (1991), Symbolic Probabilistic Inference with Continuous Variables. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Mateo, CA, 77–85.

Char, B., Geddes, K., Gonnet, G., Leong, B., Monagan, M., and Watt, S. (1991), *Maple V Language Reference Manual*. Springer-Verlag.

Charniak, E. and McDermott, D. (1985), *Introduction to Artificial Intelligence*. Addison-Wesley, Reading, MA.

Chavez, R. and Cooper, G. (1990), A Randomized Approximation Algorithm for Probabilistic Inference on Bayesian Belief Networks. *SIAM Journal on Computing*, 20:661–685.

Cheeseman, P. (1985), In Defense of Probability. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence (IJCAI–85)*. Morgan Kaufmann Publishers, San Mateo, CA, 1002–1009.

Chickering, D. M. (1995a), Search Operators for Learning Equivalence Classes of Bayesian Network Structures. Technical Report, R231, UCLA Cognitive Systems Laboratory, Los Angeles.

Chickering, D. M. (1995b), A Transformational Characterization of Equivalent Bayesian Network Structures. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 87–98.

Clancey, W. J. (1983), The Epistemology of Rule-Based Expert Systems. A Framework for Explanation. *Artificial Intelligence*, 20:215–251.

Cohen, P. R. and Feigenbaum, E. A. (1982), *The Handbook of Artificial Intelligence, Volume III*. William Kaufman, Los Altos, CA.

Cooper, G. F. (1984), *Nestor: A Computer-Based Medical Diagnostic Aid that Integrates Causal and Probabilistic Knowledge*. Ph.D. Thesis, Department of Computer Science, Stanford University.

Cooper, G. F. (1990), The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks. *Artificial Intelligence*, 42:393–405.

Cooper, G. F. and Herskovits, E. (1992), A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9:309–347.

Cormen, T. H., Leiserson, C. E., and Rivest, R. L. (1990), *Introduction to Algorithms*. MIT Press, Boston, MA.

Dagum, P. and Luby, M. (1993), Approximating Probabilistic Inference in Bayesian Belief Networks is NP-hard. *Artificial Intelligence*, 60:141–153.

D'Ambrosio, B. (1994), SPI in Large BN2O Networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 128–135.

Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980), Markov Fields and Log-linear Models for Contingency Tables. *Annals of Statistics*, 8:522–539.

Darwiche, A. (1995), Conditioning Algorithms for Exact and Approximate Inference in Causal Networks. In *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 99–107.

Davis, R. and Buchanan, B. G. (1977), Meta Level Knowledge. Overview and Applications. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence (IJCAI-77)*. Morgan Kaufmann Publishers, San Mateo, CA, 920–927.

Dawid, A. (1979), Conditional Independence in Statistical Theory. *Journal of the Royal Statistical Society, Series B*, 41:1–33.

Dawid, A. (1980), Conditional Independence for Statistical Operations. *Annals of Statistics*, 8:598–617.

Dawid, A. and Lauritzen, S. L. (1993), Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models. *Annals of Statistics*, 21:1272–1317.

DeGroot, M. H. (1970), *Optimal Statistical Decisions*. Mc Graw Hill, New York.

DeGroot, M. H. (1987), *Probability and Statistics*. Addison-Wesley, Reading, MA.

Delcher, A. L., Grove, A., Kasif, S., and Pearl, J. (1995), Logarithmic-Time Updates and Queries in Probabilistic Networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 116–124.

Dempster, A., Laird, N., and Rubin, D. (1977), Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society, B*, 39:1–38.

Devroye, L. (1986), *Non-Uniform Random Variate Generations*. Springer Verlag, New York.

Díez, F. J. (1994), *Sistema Experto Bayesiano para Ecocardiografía*. Ph.D. Thesis, Departamento de Informática y Automática, U.N.E.D., Madrid.

Díez, F. J. (1996), Local Conditioning in Bayesian Networks. *Artificial Intelligence*. In press.

Díez, F. J. and Mira, J. (1994), Distributed Inference in Bayesian Networks. *Cybernetics and Systems*, 25:39–61.

Dirac, G. A. (1961), On Rigid Circuit Graphs. *Abh. Math. Sem. Univ. Hamburg*, 25:71–76.

Duda, R. O., Gaschnig, J. G., and Hart, P. E. (1980), Model Design in the Prospector Consultant System for Mineral Exploration. In Michie, D., editor, *Expert Systems in the Microelectronic Age*. Edinburgh University Press, Edinburgh, 153–167.

Duda, R. O., Hart, P. E., and Nilsson, N. (1976), Subjective Bayesian Methods for Rule-Based Inference Systems. *Proceedings of the 1976 National Computer Conference (AFIPS)*, 45:1075–1082.

Durkin, J. (1994), *Expert Systems: Design and Development*. Maxwell Macmillan, New York.

Durrett, R. (1991), *Probability: Theory and Examples*. Wadsworth, Pacific Grove, CA.

Feller, W. (1968), *An Introduction to Probability Theory and Applications*. John Wiley and Sons, New York.

Fischler, M. A. and Firschein, O., editors (1987), *Readings in Computer Vision*. Morgan Kaufmann Publishers, San Mateo, CA.

Fisher, D. and Lenz, H. J., editors (1996), *Learning from Data: Artificial Intelligence and Statistics V*. Springer Verlag, New York.

Freeman, J. A. and Skapura, D. M. (1991), *Neural Networks: Algorithms, Applications, and Programming Techniques*. Addison-Wesley, Reading, MA.

Frydenberg, M. (1990), The Chain Graph Markov Property. *Scandinavian Journal of Statistics*, 17:333–353.

Fulkerson, D. R. and Gross, O. A. (1965), Incidence Matrices and Interval Graphs. *Pacific J. Math.*, 15:835–855.

Fung, R. and Chang, K. (1990), Weighing and Integrating Evidence for Stochastic Simulation in Bayesian Networks. In Henrion, M., Shachter, R. D., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 5*. North Holland, Amsterdam, 209–219.

Fung, R. and Favero, B. D. (1994), Backward Simulation in Bayesian Networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 227–234.

García, O. N. and Chien, Y.-T., editors (1991), *Knowledge-Based Systems: Fundamentals and Tools*. IEEE Computer Society Press, Los Alamitos, CA.

Garey, M. R. and Johnson, D. S. (1979), *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman and Company, New York.

Gavril, T. (1972), Algorithms for Minimum Coloring, Maximum Clique, Minimum Covering by Cliques, and Maximum Indpendent Set of a Chordal Graph. *SIAM Journal of Computing*, 1:180–187.

Gavril, T. (1974), An Algorithm for Testing Chordality of Graphs. *Inform. Process. Lett.*, 3:110–112.

Geiger, D. (1987), Towards the Formalization of Informational Dependencies (M.S. Thesis). Technical Report, R-102, UCLA Cognitive Systems Laboratory, Los Angeles.

Geiger, D. (1990), *Graphoids: A Qualitative Framework for Probabilisitc Inference*. Ph.D. Thesis, Department of Computer Science, University of California.

Geiger, D. and Heckerman, D. (1994), Learning Gaussian Networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 235–243.

Geiger, D. and Heckerman, D. (1995), A Characterization of the Dirichlet Distribution with Application to Learning Bayesian Networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 196–207.

Geiger, D. and Pearl, J. (1990), On the Logic of Causal Models. In Shachter, R. D., Levitt, T. S., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 4*. North Holland, Amsterdam, 3–14.

Geiger, D., Verma, T., and Pearl, J. (1990a), D-separation: From Theorems to Algorithms. In Henrion, M., Shachter, R. D., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 5*. North Holland, Amsterdam, 139–148.

Geiger, D., Verma, T., and Pearl, J. (1990b), Identifying Independence in Bayesian Networks. *Networks*, 20:507–534.

Gelfand, A. E. and Smith, A. F. (1990), Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85:398–409.

Gelman, A. and Speed, T. P. (1993), Characterizing a Joint Probability Distribution by Conditionals. *Journal of the Royal Statistical Society, Series B*, 55:185–188.

Gibbons, A. (1985), *Algorithmic Graph Theory*. Cambridge University Press, Cambridge.

Gilks, W. R. and Wild, P. (1992), Adaptive Rejection Sampling for the Gibbs Sampling. *Journal of the Royal Statistical Society, Series C*, 41:337–348.

Ginsberg, M. L. (1993), *Essentials of Artificial Intelligence*. Morgan Kaufmann, San Mateo, CA.

Golumbic, M. C. (1980), *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, New York.

Gutiérrez, J. M. (1994), *Sistemas Expertos, Grafos y Redes Bayesianas*. Ph.D. Thesis, Departamento de Matemática Aplicada y Ciencias de la Computación, Universidad de Cantabria, Spain.

Gutiérrez, J. M. and Solares, C. (1995), Some Graph Algorithms for Causal Network Expert Systems. In Keranen, V. and Mitic, P., editors, *Mathematics with Vision: Proceedings of the First International Mathematica Symposium*. Computational Mechanics Publications, Southampton, U.K., 183–190

Hadi, A. S. (1996), *Matrix Algebra as a Tool*. Duxbury Press, Belmont, CA.

Harary, F., editor (1969), *Graph Theory*. Addison-Wesley, Reading, MA.

Hayes-Roth, F. (1985), Rule-Based Systems. *Comunications of the ACM*, 28:921–932.

Hayes-Roth, F., Waterman, D. A., and Lenat, D. B., editors (1983), *Building Expert Systems*. Addison-Wesley, Reading, MA.

Heckerman, D. (1990a), *Probabilistic Similarity Networks*. Ph.D. Thesis, Program in Medical Information Sciences, Stanford University.

Heckerman, D. (1990b), Probabilistic Similarity Networks. *Networks*, 20:607–636.

Heckerman, D. (1995), A Tutorial on Learning With Bayesian Networks. Technical Report, Msr TR-95-06, Microsoft Research, Redmond, WA.

Heckerman, D. and Geiger, D. (1995), Learning Bayesian Networks: A Unification for Discrete and Gaussian Domains. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 274–284.

Heckerman, D., Geiger, D., and Chickering, D. M. (1994), Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 293–301.

Henrion, M. (1988), Propagation of Uncertainty by Logic Sampling in Bayes' Networks. In Lemmer, J. F. and Kanal, L. N., editors, *Uncertainty in Artificial Intelligence 2*. North Holland, Amsterdam, 149–164.

Henrion, M. (1991), Search-Based Methods to Bound Diagnostic Probabilities in Very Large Belief Nets. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Mateo, CA, 142–150.

Hernández, L. D. (1995), *Diseño y Validación de Nuevos Algoritmos para el Tratamiento de Grafos de Dependencias*. Ph.D. Thesis, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada. Spain.

Hogg, R. V. (1993), *Probability and Statistical Inference*. Maxwell Macmillan International, New York.

Horvitz, E., Breese, J., and Henrion, M. (1988), Decision Theory in Expert Systems and Artificial Intelligence. *International Journal of Approximate Reasoning*, 2:247–302.

Horvitz, E., Heckerman, D., and Langlotz, C. (1986), A Framework for Comparing Alternative Formalisms for Plausible Reasoning. In *Fifth National Conference on Artificial Intelligence (AAAI-86)*. AAAI Press/MIT Press, Menlo Park, CA, 210–214.

Isham, B. (1981), An Introduction to Spatial Point Processes and Markov Random Fields. *International Statistical Review*, 49:21–43.

Jackson, P. (1990), *Introduction to Expert Systems, 2nd edition*. Addison-Wesley, Reading, MA.

Jensen, F. V. (1988), *Junction Trees and Decomposable Hypergraphs*. JUDEX Research Report, Aalborg, Denmark.

Jensen, F. V. (1996), *An Introduction to Bayesian Networks*. Springer-Verlag, New York.

Jensen, F., Lauritzen, S., and Olesen, K. (1990), Bayesian Updating in Causal Probabilistic Networks by Local Computations. *Computational Statistics Quarterly*, 4:269–282.

Jensen, F. V., Olesen, K. G., and Andersen, S. K. (1990), An Algebra of Bayesian Belief Universes for Knowledge-Based Systems. *Networks*, 20:637–660.

Johnson, L. and Keravnou, E. T. (1988), *Expert System Architectures*. Kogan Page Limited, London, U.K.

Johnson, R. A. and Wichern, D. W. (1988), *Applied Multivariate Analysis, 2nd edition*. Prentice Hall, Englewood Cliffs, NJ.

Jones, J. L. and Flynn, A. M. (1993), *Mobile Robots: Inspiration to Implementation*. A. K. Peters, Wellesley, MA.

Jordan, M. and Jacobs, R. (1993), Supervised Learning and Divide-And-Conquer: A Statistical Approach. In *Machine Learning: Proceedings of the Tenth International Conference*. Morgan Kaufmann Publishers, San Mateo, CA, 159–166.

Jubete, F. and Castillo, E. (1994), Linear Programming and Expert Systems. *Microcomputers in Civil Engineering, Special Issue on Uncertainty in Expert Systems*, 9:335–345.

Kenley, C. R. (1986), *Influence Diagram Models with Continuous Variables*. Ph.D. Thesis, Stanford, Stanford University.

Kim, J. H. (1983), *CONVINCE: A Conversation Inference Consolidation Engine*. Ph.D. Thesis, Department of Computer Science, University of California.

Kim, J. H. and Pearl, J. (1983), A Computational Model for Combined Causal and Diagnostic Reasoning in Inference Systems. In *Proceedings of the 8th*

International Joint Conference on Artificial Intelligence (IJCAI-83). Morgan Kaufmann Publishers, San Mateo, CA, 190–193.

Larrañaga, P. (1995), *Aprendizaje Estructural y Descomposición de Redes Bayesianas Via Algoritmos Genéticos*. Ph.D. Thesis, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad del País Vasco. Spain.

Larrañaga, P., Kuijpers, C., Murga, R., and Yurramendi, Y. (1996), Searching for the Best Ordering in the Structure Learning of Bayesian Networks. *IEEE Transactions on Systems, Man and Cybernetics*, 26. In press.

Laskey, K. B. (1995), Sensitivity Analysis for Probability Assessments in Bayesian Networks. *IEEE Transactions on Systems, Man and Cybernetics*, 25:901–909.

Lauritzen, S. L. (1974), Sufficiency, Prediction and Extreme Models. *Scandinavian Journal of Statistics*, 1:128–134.

Lauritzen, S. L. (1982), *Lectures on Contingency Tables, 2nd edition*. Aalborg University Press, Aalborg, Denmark.

Lauritzen, S. L. (1992), Propagation of Probabilities, Means, and Variances in Mixed Graphical Association Models. *Journal of the American Statistical Association*, 87:1098–1108.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H. G. (1990), Independence Properties of Directed Markov Fields. *Networks*, 20:491–505.

Lauritzen, S. L., Speed, T. P., and Vijayan, K. (1984), Decomposable Graphs and Hypergraphs. *Journal of the Australian Mathematical Society, Series A*, 36:12–29.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988), Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society, Series B*, 50:157–224.

Lauritzen, S. L. and Wermuth, N. (1989), Graphical Models for Association Between Variables, Some of Which are Qualitative and Some Quantitative. *Annals of Statistics*, 17:31–54.

Levy, D. N. L., editor (1988), *Computer Games*. Springer Verlag, New York.

Li, Z. and D'Ambrosio, B. (1994), Efficient Inference in Bayes Nets as a Combinatorial Optimization Problem. *International Journal of Approximate Reasoning*, 11:55–81.

Lindley, D. V. (1987), The Probability Approach to the Treatment of Uncertainty in Artificial Intelligence. *Statistical Science*, 2:17–24.

Lisboa, P. G. L., editor (1992), *Neural Networks: Current Applications*. Chapman and Hall, New York.

Liu, C. (1985), *Elements of Discrete Mathematics*. McGraw-Hill, New York.

Liu, X. and Li, Z. (1994), A Reasoning Method in Damage Assessment of Buildings. *Microcomputers in Civil Engineering, Special Issue on Uncertainty in Expert Systems*, 9:329–334.

Luger, G. F. and Stubblefield, W. A. (1989), *Artificial Intelligence and the Design of Expert Systems*. Benjamin/Cummings, Redwood City, CA.

Madigan, D. and Raftery, A. (1994), Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *Journal of the American Statistical Association*, 89:1535–1546.

Martos, B. (1964), Hyperbolic Programming. *Naval Research Logistic Quarterly*, 11:135–156.

McHugh, J. A. (1990), *Algorithmic Graph Theory*. Prentice Hall, Englewood Cliffs, NJ.

McKeown, K. R. (1985), *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, New York.

McKerrow, P. (1991), *Introduction to Robotics*. Addison-Wesley, Reading, MA.

Moore, J. D. and Swartout, W. R. (1990), Pointing: A Way Toward Explanation Dialogue. In *Proceedings of the 8th National Conference on AI (AAAI-90)*. AAAI Press/MIT Press, Menlo Park, CA, 457–464.

Naylor, C. (1983), *Build Your Own Expert System*. Sigma Press, Wilmslow, U.K.

Neapolitan, R. E. (1990), *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. Wiley-Interscience, New York.

Newborn, M. (1994), *The Great Theorem Prover, Version 2.0*. NewBorn Software, Westmount, Quebec.

Newell, A., Shaw, J. C., and Simon, H. A. (1963), Chess-Playing Programs and the Problem of Complexity. In Feigenbaum, E. and Feldman, J., editors, *Computers and Thought*, McGraw-Hill, New York.

Niemann, H. (1990), *Pattern Analysis and Understanding, 3rd edition*. Series in Information Sciences. Springer-Verlag, Berlin.

Normand, S.-L. and Tritchler, D. (1992), Parameter Updating in Bayes Network. *Journal of the American Statistical Association*, 87:1109–1115.

Norton, S. W. (1988), An Explanation Mechanism for Bayesian Inference Systems. In Lemmer, J. F. and Kanal, L. N., editors, *Uncertainty in Artificial Intelligence 2*. North Holland, Amsterdam, 165–174.

O'Keefe, R. M., Balci, O., and Smith, E. P. (1987), Validating Expert System Performance. *IEEE Expert*, 2:81–90.

Olesen, K. G. (1993), Causal Probabilistic Networks with Both Discrete and Continuous Variables. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3:275–279.

Patrick, E. A. and Fattu, J. M. (1984), *Artificial Intelligence with Statistical Pattern Recognition*. Prentice-Hall, Englewood Cliffs, N.J.

Paz, A. (1987), A Full Characterization of Pseudographoids in Terms of Families of Undirected Graphs. Technical Report, R-95, UCLA Cognitive Systems Laboratory, Los Angeles.

Paz, A. and Schulhoff, R. (1988), Closure Algorithms and Decision Problems for Graphoids Generated by Two Undirected Graphs. Technical Report, R-118, UCLA Cognitive Systems Laboratory, Los Angeles.

Pearl, J. (1984), *Heuristics*. Addison–Wesley, Reading, MA.

Pearl, J. (1986a), A Constraint-Propagation Approach to Probabilistic Reasoning. In Kanal, L. N. and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence*. North Holland, Amsterdam, 357–369.

Pearl, J. (1986b), Fusion, Propagation and Structuring in Belief Networks. *Artificial Intelligence*, 29:241–288.

Pearl, J. (1987a), Distributed Revision of Compatible Beliefs. *Artificial Intelligence*, 33:173–215.

Pearl, J. (1987b), Evidential Reasoning Using Stochastic Simulation of Causal Models. *Artificial Intelligence*, 32:245–257.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA.

Pearl, J., Geiger, D., and Verma, T. (1989), The Logic of Influence Diagrams. In Oliver, R. M. and Smith, J. D., editors, *Influence Diagrams, Belief Networks and Decision Analysis*. John Wiley and Sons, New York, 67–87.

Pearl, J. and Paz, A. (1987), Graphoids: A Graph-Based Logic for Reasoning about Relevance Relations. In Boulay, B. D., Hogg, D., and Steels, L., editors, *Advances in Artificial Intelligence-II*. North Holland, Amsterdam, 357–363.

Pearl, J. and Verma, T. (1987), The Logic of Representing Dependencies by Directed Graphs. In *Proceedings of the National Conference on AI (AAAI-87)*. AAAI Press/MIT Press, Menlo Park, CA, 374–379.

Pedersen, K. (1989), *Expert Systems Programming: Practical Techniques for Rule-Based Expert Systems*. John Wiley and Sons, New York.

Peot, M. A. and Shachter, R. D. (1991), Fusion and Propagation With Multiple Observations in Belief Networks. *Artificial Intelligence*, 48:299–318.

Poole, D. (1993a), Average-Case Analysis of a Search Algorithm for Estimating Prior and Posterior Probabilities in Bayesian Networks with Extreme Probabilities. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*. Morgan Kaufmann Publishers, San Mateo, CA, 606–612.

Poole, D. (1993b), The Use of Conflicts in Searching Bayesian Networks. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Mateo, CA, 359–367.

Preece, A. D. (1990), Towards a Methodology for Evaluating Expert Systems. *Expert Systems*, 7:215–293.

Preparata, F. P. and Shamos, M. I. (1985), *Computational Geometry*. Springer-Verlag, New York.

Press, S. J. (1992), *Bayesian Statistics: Principles, Models, and Applications*. John Wiley and Sons, New York.

Press, W. H., Teulosky, S. A., Vetterling, W. T., and Flannery, B. P. (1992), *Numerical Recipies, 2nd edition*. Cambridge University Press, Cambridge.

Quinlan, J., editor (1987), *Applications of Expert Systems, Volume 1*. Addison-Wesley, Reading, MA.

Quinlan, J., editor (1989), *Applications of Expert Systems, Volume 2*. Addison-Wesley, Reading, MA.

Rabiner, L. and Juang, B. H. (1993), *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, N.J.

Raiffa, H. and Schlaifer, R. (1961), *Applied Statistical Decision Theory*. Division of Research, Harvard Business School, Boston.

Rencher, A. C. (1995), *Methods of Multivariate Analysis*. John Wiley and Sons, New York.

Ripley, B. D. (1987), *Stochastic Simulation*. John Wiley and Sons, New York.

Rissanen, J. (1983), A Universal Data Compression System. *IEEE Transactions on Information Theory*, IT-29:656–664.

Rissanen, J. (1986), Stochastic Complexity and Modeling. *Annals of Statistics*, 14:1080–1100.

Rich, E. and Knight, K. (1991), *Artificial Intelligence, 2nd edition*. McGraw-Hill, New York.

Rose, D. J., Tarjan, R. E., and Leuker, G. S. (1976), Algorithmic Aspects of Vertex Elimination on Graphs. *SIAM Journal of Computing*, 5:266–283.

Ross, K. A. and Wright, C. R. (1988), *Discrete Mathematics*. Prentice Hall, Englewood Cliffs, N.J.

Rubinstein, R. Y. (1981), *Simulation and the Monte Carlo Method*. John Wiley and Sons, New York.

Russell, S. J. and Norvig, P. (1995), *Artificial Intelligence: A Modern Approach*. Prentice Hall, Englewood Cliffs, N.J.

Schank, R. C. and Abelson, R. P. (1977), *Scripts, Plans, Goals and Understanding. An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum Associates, Hillsdale, N.J.

Schwarz, G. (1978), Estimation the Dimension of a Model. *Annals of Statistics*, 17(2):461–464.

Shachter, R. D. (1986), Evaluating Influence Diagrams. *Operations Research*, 34:871–882.

Shachter, R. D. (1988), Probabilistic Inference and Influence Diagrams. *Operations Research*, 36:589–605.

Shachter, R. D. (1990a), Evidence Absorption and Propagation Through Evidence Reversals. In Henrion, M., Shachter, R. D., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 5*. North Holland, Amsterdam, 173–190.

Shachter, R. D. (1990b), An Ordered Examination of Influence Diagrams. *Networks*, 20:535–563.

Shachter, R. D., Andersen, S. K., and Szolovits, P. (1994), Global Conditioning for Probabilistic Inference in Belief Networks. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 514–522.

Shachter, R., D'Ambrosio, B., and DelFavero, B. (1990), Symbolic Probabilistic Inference in Belief Networks. In *Proceedings of the 8th National Conference on AI (AAAI-90)*. AAAI Press/MIT Press, Menlo Park, CA, 126–131.

Shachter, R. and Kenley, C. (1989), Gaussian Influence Diagrams. *Management Science*, 35(5):527–550.

Shachter, R. and Peot, M. (1990), Simulation Approaches to General Probabilistic Inference on Belief Networks. In Henrion, M., Shachter, R. D., Kanal, L. N., and Lemmer, J. F., *Uncertainty in Artificial Intelligence 5*. North Holland, Amsterdam, 221–231.

Shafer, G. (1976), *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ.

Shapiro, L. G. and Rosenfeld, A. (1992), *Computer Vision and Image Processing*. Academic Press, Boston, MA.

Shapiro, S. C., editor (1987), *Encyclopedia of Artificial Intelligence*, John Wiley and Sons, New York.

Shneiderman, B. (1987), *Designing the Human Interface*. Addison-Wesley, Reading, MA.

Shwe, M. and Cooper, G. (1991), An Empirical Analysis of Likelihood-Weighting Simulation on a Large Multiply Connected Medical Belief Network. *Computers and Biomedical Research*, 24:453–475.

Simons, G. L. (1985), *Introducing Artificial Intelligence*. John Wiley and Sons, New York.

Sing-Tze, B. (1984), *Pattern Recognition*. Marcel Dekker, New York.

Skiena, S. S. (1990), *Implementing Discrete Mathematics: Combinatorics and Graph Theory with Mathematica*. Addison-Wesley, Reading, MA.

Smith, C. A. B. (1961), Consistency in Statistical Inference and Decision. *Journal of the Royal Statistical Society, Series B*, 23:1–37.

Spirtes, P. and Meek, C. (1995), Learning Bayesian Netwoks with Discrete Variables from Data. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, 294–300.

Srinivas, S. and Nayak, P. (1996), Efficient Enumeration of Instantiations in Bayesian Networks. In *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 500–508.

Stevens, L. (1984), *Artificial Intelligence. The Search for the Perfect Machine*. Hayden Book Company, Hasbrouck Heights, N.J.

Stillman, J. (1991), On Heuristics for Finding Loop Cutsets in Multiply Connected Belief Networks. In Bonissone, P. P., Henrion, M., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 6*. North Holland, Amsterdam, 233–343.

Strat, T. M. and Lowrance, J. D. (1989), Explaining Evidential Analysis. *International Journal of Approximate Reasoning*, 3:299–353.

Studený, M. (1989), Multiinformation and the Problem of Characterization of Conditional Independence Relations. *Problems of Control and Information Theory*, 18:3–16.

Studený, M. (1992), Conditional Independence Relations Have No Finite Complete Characterization. In Kubík, S. and Vísek, J. A., editors, *Information Theory, Statistical Decision Functions and Random Processes: Transactions of 11th Prague Conference B*. Kluwer, Dordrecht, 377–396.

Studený, M. (1994), Semigraphoids are Two-Antecedental Approximations of Stochastic Conditional Independence Models. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers, San Francisco, CA, 546–552.

Suermondt, H. J. (1992), *Explanation in Bayesian Belief Networks*. Ph.D. Thesis, Department of Computer Science, Stanford University.

Suermondt, H. J. and Cooper, G. F. (1990), Probabilistic Inference in Multiply Connected Belief Networks Using Loop Cutsets. *International Journal of Approximate Reasoning*, 4:283–306.

Suermondt, H. J. and Cooper, G. F. (1991a), A Combination of Exact Algorithms for Inference on Bayesian Belief Networks. *International Journal of Approximate Reasoning*, 5:521–542.

Suermondt, H. J. and Cooper, G. F. (1991b), Initialization for the Method of Conditioning in Bayesian Belief Networks. *Artificial Intelligence*, 50:83–94.

Suermondt, H., Cooper, G., and Heckerman, D. (1991), A Combination of Cutset Conditioning with Clique-Tree Propagation in the Pathfinder System. In Bonissone, P. P., Henrion, M., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 6*. North-Holland, Amsterdam, 245–253.

Tamassia, R. and Tollis, I. G., editors (1995), *Graph Drawing (Proceedings of GD'94)*, Lecture Notes in Computer Science. Springer-Verlag, New York.

Tarjan, R. E. (1983), *Data Structures and Network Algorithms*. SIAM (Society for Industrial and Applied Mathematics), Philadelphia, PA.

Tarjan, R. E. and Yannakakis, M. (1984), Simple Linear-Time Algorithms to Test Chordality of Graphs, Test Acyclity of Hypergraphs and Selectively Reduce Acyclic Hypergraphs. *SIAM Journal of Computing*, 13:566–579.

Tessem, B. (1992), Interval Probability Propagation. *International Journal of Approximate Reasoning*, 7:95–120.

Ur, S. and Paz, A. (1994), The Representation Power of Probabilistic Knowledge by Undirected Graphs and Directed Acyclic Graphs: A Comparasion. *International Journal of General Systems*, 22:219–231.

Verma, T. S. (1987), Some Mathematical Properties of Dependency Models. Technical Report, R-103, UCLA Cognitive Systems Laboratory, Los Angeles.

Verma, T. and Pearl, J. (1990), Causal Networks: Semantics and Expressiveness. In Shachter, R. D., Levitt, T. S., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 4*. North Holland, Amsterdam, 69–76.

Verma, T. and Pearl, J. (1991), Equivalence and Synthesis of Causal Models. In Bonissone, P. P., Henrion, M., Kanal, L. N., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 6*. North Holland, Amsterdam, 255–268.

Von Neumann, J. (1951), Various Techniques Used in Connection with Random Numbers. *U.S. Nat. Bur. Stand. Appl. Math. Ser.*, 12:36–38.

Waterman, D. A. (1985), *A Guide to Expert Systems*. Addison-Wesley, Reading, MA.

Weiss, S. M. and Kulikowski, C. A. (1984), *A Practical Guide to Designing Expert Systems*. Rowman and Allanheld, Totowa, N.J.

Wermuth, N. and Lauritzen, S. L. (1983), Graphical and Recursive Models for Contingency Tables. *Biometrika*, 70:537–552.

Whittaker, J. (1990), *Graphical Models in Applied Mathematical Multivariate Statistics*. John Wiley and Sons, New York.

Wick, M. R. and Thompson, W. B. (1992), Reconstructive Expert System Explanation. *Artificial Intelligence*, 54:33–70.

Winston, P. H. (1992), *Artificial Intelligence, 3rd edition*. Addison-Wesley, Reading, MA.

Wolfram, S. (1991), *Mathematica: A System for Doing Mathematics by Computer*. Addison-Wesley, Reading, MA.

Wos, L., Overbeek, R., Lusk, E., and Boyle, J. (1984), *Automated Reasoning. Introduction and Applications*. Prentice-Hall, Englewood Cliffs, N.J.

Xu, H. (1995), Computing Marginals for Arbitrary Subsets from Marginal Representation in Markov Trees. *Artificial Intelligence*, 74:177–189.

Xu, L. and Pearl, J. (1989), Structuring Causal Tree Models with Continuous Variables. In Kanal, L. N., Levitt, T. S., and Lemmer, J. F., editors, *Uncertainty in Artificial Intelligence 3*. North Holland, Amsterdam, 209–219.

Yager, R. R., Ovchinnikov, S., Yong, R. M., and Nguyen, H. T., editors (1987), *Fuzzy Sets and Applications: Selected Papers by L. A. Zadeh*. John Wiley and Sons, New York.

Yannakakis, M. (1981), Computing the Minimal Fill-in is NP-Complete. *SIAM Journal of Algebraic Discrete Methods*, 2:77–79.

Yu, Q., Almulla, M., and Newborn, M. (1996), Heuristics for a Semantic Tree Theorem Prover. In *Proceedings of the Fourth International Symposium on Artificial Intelligence and Mathematics (AI/MATH-96)*, Fort Lauderdale, Florida, 162–165.

Zadeh, L. A. (1983), The Role of Fuzzy Logic in the Management of Uncertainty in Expert Systems. *Fuzzy Sets and Systems*, 11:199–227.

Índice