— TESIS DOCTORAL —

EXPLICACIÓN EN REDES BAYESIANAS CAUSALES. APLICACIONES MÉDICAS.

Carmen Lacave Rodero

Licenciada en CC. Matemáticas
por la Universidad Complutense de Madrid

Departamento de Inteligencia Artificial
Universidad Nacional de Educación a Distancia

Madrid, 2002

— TESIS DOCTORAL —

Departamento de Inteligencia Artificial
Universidad Nacional de Educación a Distancia

Explicación en redes bayesianas causales. Aplicaciones médicas

Por: Carmen Lacave Rodero
Licenciada en CC. Matemáticas

Director de la Tesis: Dr. D. Francisco Javier Díez Vegas

Madrid, 5 de Diciembre de 2002

Felix qui potuit rerum cognoscere causas.
Virgilio, ***Geórgicas ii.490***

A Paco
A mi hijo, Daniel
A mis padres, Elías y Carmen

Agradecimientos

Son numerosas las personas que, directa o indirectamente, han contribuido a la realización de este trabajo y a las que quiero agradecerles su ayuda. En primer lugar, el profesor Francisco Javier Díez Vegas, director de esta tesis, a quien deseo expresar mi más sincero agradecimiento por haberme animado tanto desde mis comienzos, también en los momentos más difíciles, ofreciéndome siempre su apoyo y sus consejos, incluso fuera de horario, y contando conmigo para participar en distintos proyectos de investigación de los que he obtenido, aparte de la posibilidad de conocer a mucha gente relacionada con este campo, la mayoría del apoyo económico para realizar este trabajo. Además, le estoy muy agradecida por transmitirme su pasión por la investigación durante todo el tiempo que ha dedicado a atenderme, lo que ha sido en parte posible gracias a la amabilidad de Jesús Boticario y José Luis Fernández Vindel por dejarme compartir su despacho, e incluso su tiempo, durante nuestras reuniones de trabajo.

Agradezco a los integrantes del Proyecto Elvira todo lo que he aprendido en las intensas reuniones de trabajo (aunque llenas también de buenos ratos) que hemos mantenido desde el origen del primer proyecto, pero fundamentalmente a Andrés Cano, Luis Miguel de Campos, José Antonio Gámez, Manuel Gómez Olmedo, Pedro Larrañaga, Serafín Moral y Antonio Salmerón; sin olvidar a Roberto Atienza, por su experiencia con JAVA y la gran ayuda que nos prestó en los comienzos con la programación de la interfaz, pues la mayor parte del trabajo relacionado con su implementación se debe hoy a él.

Especialmente agradecida estoy al profesor D. Enrique Súcar que, además de proporcionarme ideas y colaboración permanente, me ha ayudado en la última etapa de trabajo con la revisión de esta memoria lo que ha servido, no sólo para detectar erratas, sino para aportar diversas sugerencias que hoy se ven reflejadas aquí.

Asimismo, esta tesis se ha enriquecido de los comentarios y sugerencias realizadas por el profesor Marek Druzdzel en varios trabajos relacionados con algunas de las aportaciones plasmadas en ella. Asimismo, la colaboración con la profesora Agniezska Oniśko ha sido muy valiosa, sobre todo por las aportaciones surgidas a raíz de su interacción con Elvira.

En este sentido, he intentado aprovechar también los comentarios y sugerencias de Linda van der Gaag, Silja Renooij y Adnan Darwiche, especialmente los relacionados con las posibilidades de trabajo futuro.

Quiero también manifestar mi gratitud a la Sección Departamental de Ciudad Real del Departamento de Informática, así como a la Escuela Superior de Informática (ESI) y a la Universidad de Castilla-La Mancha (UCLM), por haber puesto a mi disposición los medios materiales para poder llevar a cabo con éxito este proyecto. También he de valorar el respaldo de los miembros del Departamento de Informática y de la ESI, y en particular de Luis Jiménez, Camelia Muñoz y Pascual Iranzo, porque gracias a su colaboración he obtenido distintas ayudas económicas de la UCLM que me han permitido cubrir diversos gastos, sobre todo los de los numerosos viajes a Madrid. Por otra parte, no quiero olvidar el apoyo incondicional de Francisca Perea, Sebastián Reyes, Fernando Terán y Urbano Viñuela todos estos años.

Gracias además a mis amigos, especialmente a Diego, con quien empezó a tomar forma la idea de la aplicación médica y sin cuya ayuda e ilusión esta tesis no sería la que es hoy; a mi amiga Ester, cuya experiencia en estos "viajes" me ha servido de gran estímulo; a Juan, por su ayuda desinteresada, por los "empujones" que me ha dado en los peores momentos y por las experiencias compartidas; y a Mercedes y a Concha por su cariño, su ánimo constante y, por supuesto, por sus comidas.

Por último, mi familia ha constituido el principal pilar sobre el que se ha apoyado mi esfuerzo. Por eso, doy las gracias a mis padres, a mi hermano Elías y a mi hermana Yolanda por su constante aliento, cariño y comprensión; a Goyo, por estar ahí; a Lucía y a Carlos por los maravillosos momentos que me estáis haciendo pasar desde hace casi dos años; a Guillermo, por ser quien es, lo que compartimos y lo que significa para mí; a mi hijo Daniel, porque sin su sonrisa, su inquietud, su alegría,... no habría tenido fuerzas para robarle horas al sueño; y, especialmente, a Paco, por llevar más de catorce años sin esperar que le dé las gracias. Gracias.

Índice

Lista de Tablas

Lista de Figuras

Parte I

FUNDAMENTOS

Capítulo 1

Introducción

1.1. Motivación del proyecto

Los sistemas expertos surgieron en los años 70 como programas capaces de imitar el razonamiento seguido por expertos humanos e, incluso, con la intención de sustituirlos cuando fuera necesario. Sin embargo, en las décadas siguientes, durante los 80 y los 90, el principal objetivo de la inteligencia artificial pasó de ser el de *imitar* la inteligencia natural al de *colaborar* con los humanos, expertos o no, de forma sinérgica. De hecho, Clancey [25] resalta que la revista *Knowledge Acquisition* decía en su nota para los autores: "La cuestión clave no es la inteligencia *artificial*, sino cómo mejorar la inteligencia *natural* con la ayuda de los sistemas basados en conocimiento."[1] Unas de las principales aptitudes de los expertos humanos es su capacidad para comunicar su conocimiento y explicar el razonamiento llevado a cabo para obtener un resultado. Estas capacidades son especialmente importantes en el caso de los sistemas expertos, como se demostró en un experimento realizado durante el proyecto MYCIN, mediante el que se puso de manifiesto que los médicos se mostraban muy reacios a aceptar los consejos de la máquina si no entendían cómo se habían obtenido [177].

En este contexto, la colaboración entre los humanos y las máquinas requiere un entendimiento mutuo: los programas deberían tener en cuenta los procesos cognitivos de los usuarios humanos y, a la vez, deben hacer que éstos entiendan su proceso de razonamiento. Sin embargo, sorprende que la cantidad de esfuerzo dedicado a la investigación en este campo ha sido hasta ahora, relativamente pequeño, comparado con otras áreas de la inteligencia artificial: tan sólo existen trabajos aislados, raramente utilizados en aplicaciones del mundo real. De hecho, la capacidad de explicación de la mayoría de los sistemas expertos es más pobre que la que se desarrolló para MYCIN, que suele considerarse como

[1] "The key issue is not *artificial* intelligence, but how to extend *natural* intelligence through knowledge-based systems."

uno de los primeros sistemas expertos.

Por otra parte, en los años 80 hubo otro avance importante en el campo de la inteligencia artificial, debido a la necesidad de tratar la incertidumbre en casi todas las aplicaciones reales de los sistemas expertos, especialmente en medicina. Se demostró que el modelo de factores de certeza de MYCIN era inconsistente y podía llevar a obtener resultados erróneos. El modelo subjetivo bayesiano de PROSPECTOR [60, 61] no era mejor, y la teoría de Dempster-Shafer [38, 160], aparte de carecer de una interpretación universalmente aceptada, era poco práctica para problemas reales —al menos hasta que aparecieron los modelos gráficos. Sólo los métodos difusos [192, 193] parecían útiles para construir sistemas expertos en dominios en los que había incertidumbre.

En la misma década, los avances en redes bayesianas [15, 89, 141] y en diagramas de influencia [87, 158] demostraron de forma teórica que era posible construir sistemas expertos probabilísticos sin tener que introducir suposiciones poco realistas de independencia. Ambos modelos son capaces de representar el conocimiento de un sistema experto, utilizando la **probabilidad** como medida de incertidumbre. Los diagramas de influencia se pueden considerar como una generalización de las redes bayesianas, pero se diferencian de éstas en que están orientados a la toma de decisiones, lo que se sale del ámbito de nuestra investigación, centrada en los sistemas de ayuda al diagnóstico. Por ello, a partir de ahora, y a lo largo de todo el trabajo, nos centraremos exclusivamente en las redes bayesianas. Más información sobre diagramas de influencia y lo relacionado con el análisis de decisiones y sus aplicaciones puede encontrarse en [88].

1.2. Redes bayesianas

Una red bayesiana consta de un grafo dirigido acíclico (GDA),[2] cuyos nodos representan variables aleatorias, junto con una distribución de probabilidad condicionada para cada nodo X_i dados sus padres,[3] $P(x_i|pa(x_i))$. La probabilidad condicionada de un nodo sin padres es su probabilidad a priori $P(x_i)$. La figura 1.1 muestra un ejemplo de red bayesiana para el diagnóstico del cáncer de mama.

[2]Un ciclo es un camino dirigido cerrado, $X_i \to X_j \to X_k \to \ldots \to X_i$. Una restricción de las redes bayesianas es que los grafos que las representan no pueden contener ciclos.

[3]Un nodo X_i es un padre de X_j si y sólo si existe un enlace dirigido $X_i \to X_j$. El conjunto de padres se representa mediante $pa(x_i)$.

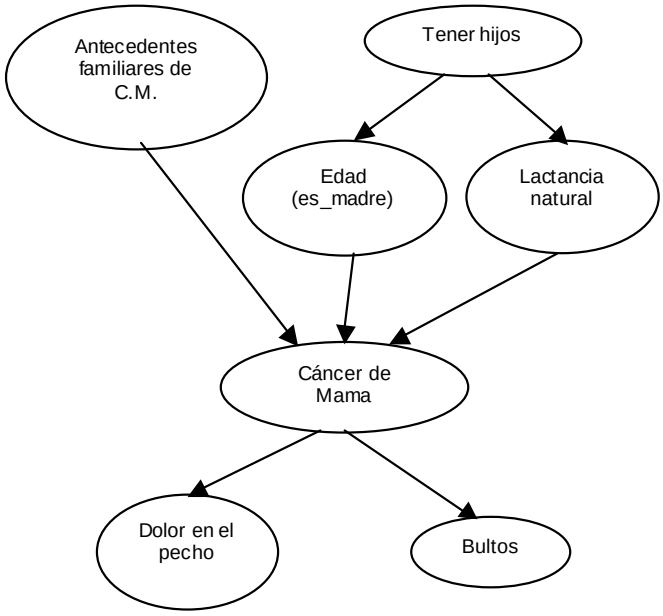

Figura 1.1: Ejemplo de grafo de la red bayesiana que representa los principales factores de riesgo, síntomas y signos del cáncer de mama.

Separación direccional

La probabilidad conjunta asociada a una red bayesiana satisface la propiedad de *separación direccional* [141], que es equivalente a la propiedad de Markov [89, 129], que afirma que un nodo es independiente de los nodos que no son descendientes suyos dados sus padres. Esta propiedad implica que un enlace $X \rightarrow Y$ en una red bayesiana representa una dependencia probabilística entre X e Y, lo que significa que la probabilidad de X afecta a la probabilidad de Y. La ausencia de enlace representa independencia probabilística, es decir, que la probabilidad de X no afecta a la probabilidad de Y y viceversa. Por ejemplo, la red de la figura 1.1, se basa en la hipótesis de que las variables Antecedentes familiares con C.M. y Tener hijos son independientes probabilísticamente, mientras que parece claro que el hecho de tener Antecedentes familiares con C.M. aumenta la probabilidad de padecer un Cáncer de mama. Por tanto, la propiedad de *separación direccional* nos permite deducir, por ejemplo, que la variable Dolor en el pecho es independiente de las variables Antecedentes familiares con C.M. y Bultos, conocido el valor de Cáncer de mama.

Como consecuencia de la propiedad de *separación direccional* se tiene que la probabilidad conjunta de una red bayesiana se puede obtener como el producto de las probabilidades de cada nodo condicionadas en sus padres, lo que se conoce como el **teorema de**

factorización de la probabilidad [141]:

$$P(x_1, ..., x_n) = \prod_i P(x_i|pa(x_i)) \tag{1.1}$$

1.2.1. Razonamiento

Un *hallazgo* es un dato que confirma con certeza el valor de una variable aleatoria; un hallazgo puede ser, por ejemplo, que la paciente tiene antecedentes familiares con cáncer de mama; otros hallazgos pueden ser que no tiene hijos, que tiene dolor en el pecho, etc. El conjunto de hallazgos se denomina *evidencia,* y se suele simbolizar por **e**.

Existen dos tipos de razonamiento que se puede llevar a cabo sobre una red bayesiana:

- **Propagación de cierta evidencia e,** con el objetivo de calcular la probabilidad a posteriori de las variables no observadas; por ejemplo, $P(x_i|\mathbf{e})$ o $P(x_i, x_j, x_k|\mathbf{e})$. Este proceso está basado, más o menos explícitamente, en la aplicación del teorema de Bayes; dependiendo de las características de la red, de las probabilidades asociadas, etc. existe la posibilidad de utilizar distintos algoritmos [11, 40, 86, 94, 103, 140, 145, 163] (véase [71] para una revisión más completa de los distintos algoritmos de propagación). Básicamente, el conjunto de métodos se puede clasificar en dos grupos:

  - **exactos,** que se basan en la idea de utilizar el paso de mensajes entre los nodos mediante *expresiones matemáticas exactas* para el cálculo de las probabilidades de los nodos.

  - **aproximados.** El problema de la propagación exacta es NP-completo [28], por lo que surge la necesidad de definir otros métodos que, a pesar de perder cierta exactitud en los resultados, sin embargo permiten obtener soluciones a aquellos problemas suficientemente complejos para los que no es posible obtener resultados en un tiempo razonable.

- **Inferencia abductiva,** que consiste básicamente en encontrar la configuración de valores más probable para las variables no observadas. Como la abducción se considera un tipo particular de explicación [141], se describe con mayor profundidad en la sección 2.4.1.

1.2.2. Causalidad

Desde un punto de vista matemático, una red bayesiana es simplemente un modelo para representar dependencias e independencias probabilísticas; en este caso, un enlace, considerado por sí mismo, no tiene ningún significado. Sin embargo, cuando se construye una red bayesiana como un modelo de un sistema del mundo real, un enlace $A \rightarrow B$ es *causal* si A es una causa de B, es decir, si existe un mecanismo mediante el cual el valor que tome A influye sobre el valor de B. Una red bayesiana se dice que es **causal** cuando todos sus enlaces son causales. Las razones para usar modelos causales en inteligencia artificial, especialmente en sistemas expertos probabilistas, son múltiples, entre las que destacamos las siguientes:

- Los seres humanos tendemos a interpretar los hechos en términos de relaciones causa-efecto [93, 143, 144]. Por tanto, los modelos causales son más fáciles de construir y de modificar [84, 144], puesto que tienden a ser más simples que los no causales [141]. En consecuencia, también son más fáciles de entender por los usuarios [50, 143, 174].

- Muchos dominios de aplicación reales están organizados en forma de jerarquías causales, como el caso de la medicina, cuyo conocimiento está estructurado en forma causa-efecto: desde la invasión de determinados organismos (virus, bacterias, etc) en el cuerpo humano, pasando por las anomalías físicas que produce, hasta sus complicaciones, síndromes, estados clínicos y los síntomas que genera [143].

- La identificación de relaciones causales *invariantes*[4] en un dominio permite predecir los efectos tanto de causas espontáneas (normalmente correspondientes a variables aleatorias) como de acciones (llamadas a veces manipulaciones o intervenciones) [142].

- Los modelos causales aportan mucha más información que los modelos meramente probabilísticos. Una distribución conjunta nos aporta información sobre las probabilidades de ciertos eventos y sobre cómo variarían éstas después de diferentes observaciones; un modelo causal además nos indica cómo variarían dichas probabilidades como resultado de intervenciones externas [143].

- Los conceptos de causalidad y probabilidad están íntimamente relacionados porque la causalidad normalmente implica un modelo de inter-dependencias probabilísti-

[4] "Invariante" significa que cuando un mecanismo está sujeto a cambios, los otros permanecen intactos.

cas. Recíprocamente, la observación de una correlación sugiere la existencia de una estructura causal subyacente [51].

- En la misma línea, la propiedades axiomáticas de las redes bayesianas (separación direccional y propiedad de Markov) corresponden a dependencias e independencias probabilísticas que aparecen en dominios causales [42].

- Existen modelos canónicos probabilísticos (puerta OR, puerta MAX, puerta AND, etc.) basados en la interpretación de los padres de un nodo como causas o condiciones para dicho nodo y sobre la asunción de independencia de interacciones causales. Estos modelos reducen el número de parámetros de la red, simplificando la adquisición del conocimiento e incluso produciendo cálculos más eficientes [43].

- Las redes bayesianas causales proporcionan determinados modelos de razonamiento cualitativo que pueden ser identificados con el objetivo de explicar los resultados de la inferencia [50, 52, 85]. Algunos de estos modelos son específicos de ciertos modelos canónicos; por ejemplo, *explicar y descartar*[5] es un fenómeno típico de la puerta OR.

- Aunque hasta hace relativamente poco la causalidad era un concepto un poco vago y abstracto, actualmente, la teoría de causalidad tiene una semántica bien definida y una lógica bien fundada [143]. Además, existen técnicas y métodos para obtener relaciones causales para representar conocimiento a partir de colecciones de datos "crudos"[73, 143, 168].

- Finalmente, el concepto de explicación está muy cerca de la noción de causa; de hecho, una de las modalidades de explicación científica consiste en encontrar las causas de los hechos observados [124].

1.2.3. Construcción

La construcción de una red bayesiana implica, básicamente, tres tareas [59]:

1. identificación de las variables y de sus valores;

2. identificación de las relaciones entre las variables, lo que completa la definición del grafo que representa el modelo; y

3. obtención de las probabilidades asociadas a cada nodo del grafo.

[5] En inglés se conoce como *explaining away*

Este proceso se puede realizar de tres formas distintas:

1. De modo *manual,* mediante la colaboración entre los ingenieros del conocimiento y los expertos humanos en el dominio que se desea representar. En este caso, los modelos que se suelen desarrollar corresponden a redes causales, pues los expertos humanos suelen tener el conocimiento estructurado en forma causal. Como la construcción manual de redes bayesianas constituye un tema básico en este trabajo de cara al diseño de la aplicación desarrollada, se analiza con más profundidad en la sección 1.3.

2. *Automáticamente,* mediante la aplicación de un conjunto de técnicas, conocidas como *algoritmos de aprendizaje,* que permiten extraer tanto los parámetros como la estructura de la red a partir de bases de datos. Actualmente existe una gran variedad de algoritmos para *aprender* la estructura de redes bayesianas, aunque la mayoría se pueden incluir dentro de una de las dos categorías siguientes [34]:

   - los que se basan en un procedimiento que busca en el espacio de posibles soluciones la mejor estructura, midiendo la calidad de cada red candidata mediante una función de evaluación. Cada uno de estos algoritmos se caracteriza por el tipo de función y por el procedimiento de búsqueda.
   - los basados en *detección de independencias,* que toman como entrada una lista de relaciones de independencia condicional y generan la red que mejor representan estas relaciones.
   - los híbridos, que emplean una combinación de ambas metodologías.

   Aunque existen métodos que tratan de construir redes causales [73, 168], en la mayoría de los casos las redes aprendidas no tienen por qué corresponder a modelos causales, lo que dificulta aún más su comprensión de cara al usuario, tanto del modelo como del razonamiento.

3. Combinando ambas posibilidades, lo que puede servir sobre todo para ayudar tanto al experto como al ingeniero a afianzar o corregir su conocimiento del dominio. Por ejemplo, puede ser interesante obtener el modelo de forma manual, con la ayuda de expertos humanos y aplicar alguno de los algoritmos de aprendizaje existentes para la obtención de las probabilidades. O al contrario, aprender primero la red a partir de una base de datos y posteriormente realizar su depuración, refinando tanto la estructura como los parámetros con la ayuda de expertos humanos, como en el

caso de Hepar II [135]. También puede resultar interesante de cara a depurar la estructura cualitativa dada por los expertos a partir de los datos [173].

1.3. Construcción manual de redes bayesianas

El principal inconveniente que surge al intentar construir una red bayesiana de modo manual es el de la poca literatura sobre el tema, aunque merece la pena citar algunos trabajos interesantes que pueden servir como guía en tan arduo proceso [45, 44, 90, 165, 173]. Esto se debe a que, como comentaremos en la sección 3.1, hasta ahora se ha prestado mucha atención a los algoritmos de inferencia y de construcción automática de redes, pero poca al campo de la construcción *manual* de modelos gráficos probabilísticos, cuando es una tarea nada trivial y sobre la que no existen criterios definidos: tan sólo las recomendaciones del sentido común y la experiencia de los colegas más cercanos. Afortunadamente, la tesis doctoral de la profesora Agniezska Oniśko [134] puede considerarse una referencia básica para construir redes bayesianas médicas, y su colaboración nos ha servido de una inestimable ayuda. Fruto de dicha colaboración surgió el trabajo [100].

En esta sección describimos los elementos para la construcción manual de redes bayesianas así como las distintas fases necesarias para ello.

1.3.1. Elementos y fases

Para la construcción manual de una red bayesiana o un diagrama de influencia es necesario recopilar todas las fuentes de información relacionadas con el dominio a modelar, es decir:

1. bibliografía sobre el dominio, que puede obtenerse de diversos medios como son libros, revistas, internet, etc.;

2. una herramienta de edición y procesamiento de redes bayesianas, con el fin de ayudar al ingeniero de conocimiento en la construcción y depuración de la red; y

3. la colaboración de al menos un experto humano, puesto que, desgraciadamente, la bibliografía sobre el dominio suele ser imprecisa, puede no estar actualizada o, lo que es más común, no contiene los datos que se necesitan. Por tanto, el único modo de conseguir la información ausente, errónea, etc, es contar con la ayuda de expertos en la materia.

Con estos elementos podemos abordar el proceso de construcción de la red que consta, básicamente, de tres partes:

- Definición del **grafo** que representa las variables del problema y las relaciones de dependencia e independencia entre ellas. Hay estudios [146] que aportan evidencia empírica sobre el hecho de que es más importante para el correcto funcionamiento del sistema la estructura de la red que la precisión de los datos numéricos. Sin embargo, es un proceso muy difícil y que, por diferentes razones que comentaremos más adelante, relacionadas sobre todo con la interacción con los expertos humanos, lleva mucho tiempo. Precisamente por ello, últimamente se empieza a dedicar atención a este proceso de *ingeniería del conocimiento* y se están intentando crear metodologías específicas que permitan llevar a cabo con éxito esta tarea [83, 102, 130], aunque todavía ninguna de ellas está lo suficientemente definida como para poder adoptarla como referencia.

- Identificación de los **modelos canónicos** [41, 43] a los que se pueden ajustar los grupos de variables de la red que los verifiquen. Como ya analizaremos en la sección 1.3.2, entre las grandes ventajas que aportan estos modelos, hay que resaltar la simplificación de la fase posterior, relacionada con la obtención de las probabilidades, pues se puede reducir considerablemente el número de parámetros a obtener, lo que es de gran ayuda especialmente cuando éstos se estiman de forma subjetiva.

- Obtención de los **datos cuantitativos**, es decir, de las probabilidades a priori de las variables que no tienen padres y de las probabilidades condicionadas para el resto. Esta fase es la más complicada del proceso de construcción manual de una red bayesiana, especialmente si se tienen que obtener de la interacción con un experto humano [58, 165] y, más aún, en medicina [45, Capítulos 3 y 11], [112, Capítulos 2 y 4] y [149, Capítulo 5]. De hecho, este problema se está empezando a abordar con cierta profundidad y existen algunas propuestas interesantes [30, 58, 149, 178] con la intención de facilitar tan ardua tarea. Sin embargo, no se han implementado aún en ninguna herramienta específica, por lo que utilizar cualquiera de ellas supone al ingeniero del conocimiento un esfuerzo adicional con el consiguiente consumo de tiempo, del que en muchos casos no se puede disponer. Únicamente las propuestas de [183] se han implementado en la nueva versión de GeNIe, pero al no estar todavía disponible, ha sido imposible su utilización.

No obstante, estas tres fases no son completamente independientes entre sí, ni una vez que estamos en una fase podemos dar por concluidas las anteriores: mediante un proceso

iterativo, cada una puede servir para obtener o modificar información de las otras. Por ello, como en cualquier proyecto de ingeniería, la idea consiste en utilizar el diseño descendente y los refinamientos sucesivos para obtener el grafo y las probabilidades que corresponda con una representación adecuada, lo que confirmará el experto humano. Esto implica que a lo largo del proceso de construcción se podrán ir obteniendo distintas versiones de la red.

Por tanto, el primer paso consiste en identificar las variables del modelo. Normalmente, en el primer intento no se obtendrá el conjunto completo de variables de la red, igual que ocurre cuando nos enfrentamos a cualquier problema de cierta complejidad. Lo importante es partir de un modelo inicial validado por los expertos humanos y que nos sirva de referencia para que, en etapas posteriores y mediante refinamientos sucesivos, podamos ir aproximándonos al diseño de la red definitivo. En esta fase es crucial transmitir correctamente el conocimiento representado en la red a los expertos, de forma precisa y comprensible para ellos porque es de quien depende la aceptación de la red.

Un problema puede ser, sin embargo, que el número de variables sea demasiado grande. En este caso, tanto la construcción de la red como su evaluación y aplicación pueden ser muy complicadas. Cuanto mayor sea el número de variables, en teoría, los resultados serán más precisos. Sin embargo, el proceso de construcción de la red puede ser demasiado complicado; tanto que sea imposible obtener la precisión teórica deseada. Por otra parte, cuanto menor sea el número de variables mayor riesgo hay de que los resultados obtenidos sean poco reales. Por tanto, hay que intentar buscar un equilibrio entre número de variables y eficiencia.

Una vez que el grafo está definido, el siguiente paso es el de la identificación de los modelos canónicos, explicados en la sección 1.3.2, con el objetivo de simplificar el proceso de la obtención de probabilidades. Además, este tipo de modelos resultan muy útiles de cara a la generación de explicaciones a los expertos.

Finalmente, durante la fase de obtención de probabilidades es posible que se detecten ciertas inconsistencias debido a que haya que añadir o eliminar variables y/o enlaces entre ellas, que los datos numéricos sean erróneos, etc. En ese caso, habría que volver a la primera fase otra vez para replantearse el modelo definido y, en caso de que haya modificaciones, refinar otra vez todo el proceso. Es obvio que para la depuración del modelo completo es fundamental aquí también la explicación del modelo representado.

El proceso completo lo podemos ver resumido en la figura 1.2.

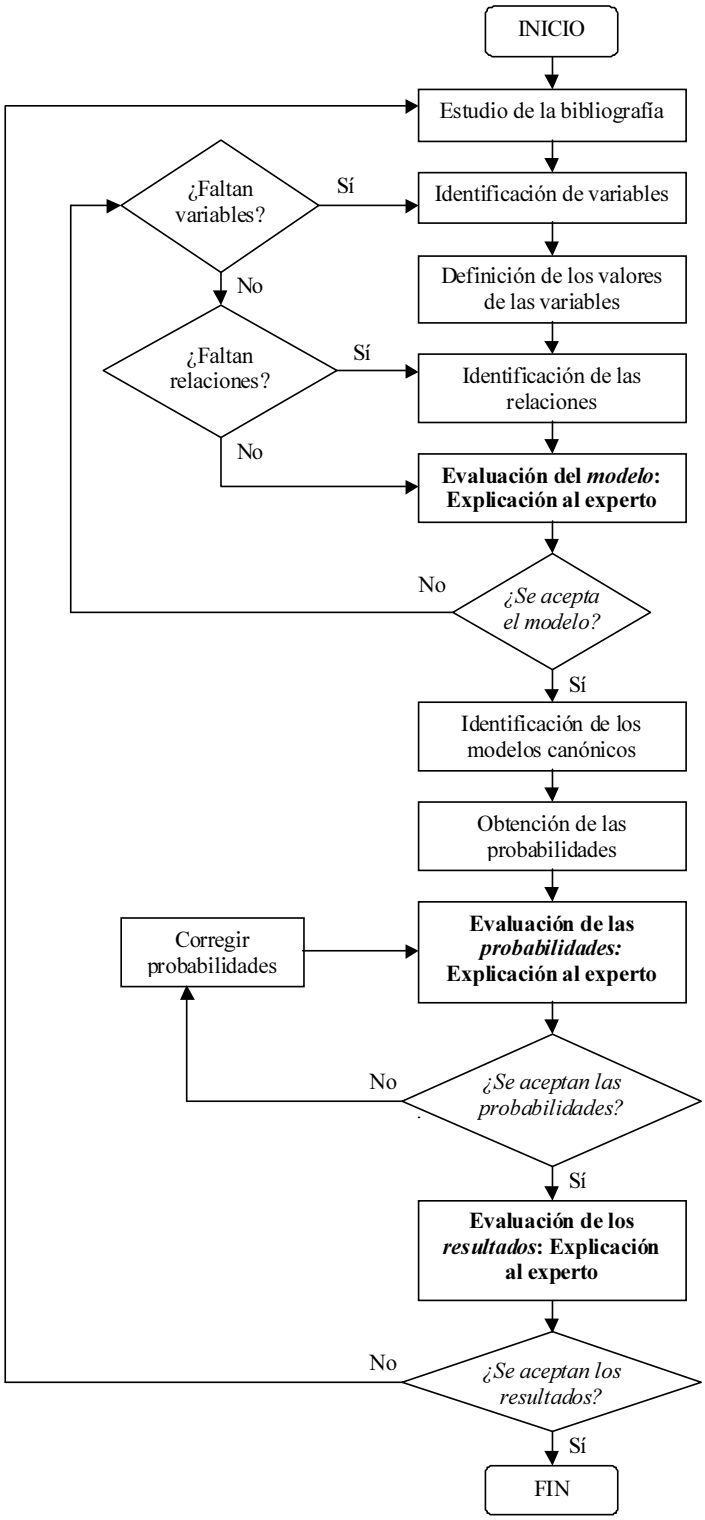

Figura 1.2: Algoritmo para la construcción manual de redes bayesianas

1.3.2. Modelos canónicos

Los principales problemas al construir redes bayesianas (en general, cualquier tipo de modelo gráfico probabilístico) están relacionados con la obtención de las probabilidades, especialmente en los casos en los que las tablas de probabilidad son considerablemente grandes, lo que puede ocurrir porque un nodo tenga muchos padres (más de tres) y/o porque los nodos tengan muchos estados posibles. Si la obtención de las probabilidades se hace de forma automática a partir de bases de datos puede ser que no existan registros con la información deseada; en el caso de que las probabilidades se asignen con la ayuda de expertos humanos, puede resultar muy complicado asignar la probabilidad para una configuración determinada de los padres por múltiples motivos: porque represente una situación poco frecuente, porque se desee que se cumplan unos requisitos específicos o porque el proceso sea demasiado tedioso. Por ejemplo, un caso bastante común es el de un nodo con 4 estados posibles, con 3 padres binarios y uno ternario; la tabla de probabilidad contendrá 96 valores. Si tenemos en cuenta que el tiempo del que se dispone para interaccionar con los expertos humanos suele estar bastante limitado, sería prácticamente imposible poder obtener valores reales en dichas situaciones.

Con la intención de solucionar estos problemas surgen los *modelos canónicos* [39, 43, 141], que permiten construir tablas de probabilidad con muchos valores a partir de un pequeño conjunto de parámetros, con el ahorro de espacio de almacenamiento y de tiempo de propagación que ello implica. Los modelos canónicos representan relaciones entre un nodo y sus padres y no son exclusivos de las redes bayesianas, sino que se usan también en diagramas de influencia y pueden aplicarse a las redes de Markov y a los *grafos en cadena*[6] [104]. Además, son aplicables a cualquier dominio que se desee modelar, no solo al campo médico. Las variables que forman parte de un modelo canónico son generalmente discretas, con un número finito de valores, y graduadas, es decir, representan una ordenación (creciente o decreciente) de sus valores. Otra propiedad de este tipo de modelos es que se les puede dar una interpretación causal, lo que permite reflejar de forma intuitiva el conocimiento sobre el mundo real. De hecho, en un estudio en el que participó la autora de esta tesis [99] se comprobó que la red obtenida al utilizar modelos canónicos producía mejores resultados que la que se había obtenido mediante el modelo general. Puesto que la única diferencia entre ambas redes eran las probabilidades y éstas habían sido obtenidas exclusivamente de expertos humanos, pudimos confirmar la idea de que, efectivamente, resulta mucho más sencillo para los expertos asignar probabilidades cuando se aplica alguno

[6] En inglés, *chain graphs*

de los modelos canónicos. Por otro lado, permiten generar explicaciones lingüísticas de una forma intuitiva, lo que implica que los modelos construidos serán mucho más fáciles de comprender por los usuarios y permitirá realizar sus evaluaciones de forma más eficaz.

Los modelos canónicos pueden ser [43]:

deterministas, caracterizados porque el valor del hijo es una función de los valores de los padres.

ruidosos, en los que se introducen un conjunto de variables auxiliares entre el nodo y sus padres de tal forma que la interacción entre el nodo y esas variables auxiliares corresponde a un modelo determinista.

residuales. En este tipo de modelos se realiza una simplificación del modelo real, representando únicamente un subconjunto de los posibles padres, antecesores y descendientes suyos.

Los principales modelos canónicos son la puerta OR y la puerta AND además de sus respectivas generalizaciones, la puerta MAX y la puerta MIN.

Puerta OR

La puerta OR, estudiada en profundidad por Pearl [141] está basada en la *interacción disyuntiva* [141] entre diversos nodos binarios del tipo ausente/presente, negativo/positivo, no/sí, etc. La idea es que ciertas causas binarias $X_1, ..., X_n$ pueden producir con cierta probabilidad un efecto binario Y, y la probabilidad de que éste ocurra no se ve alterada aun cuando un subconjunto de ellas se dan simultáneamente.

En el caso de la **puerta OR determinista**, la idea es que el efecto está presente cuando una de las causas al menos está presente, es decir,

$$P(+y|x) = \begin{cases} 1, & +x_1 \vee +x_2 \vee ... \vee +x_n \\ 0, & \text{en caso contrario} \end{cases} \tag{1.2}$$

En el caso de la **puerta OR ruidosa** (*noisy OR-gate*) supondremos además que asociada a cada una de las causas X_i existe un inhibidor I_i que puede bloquear su influencia. La probabilidad de que dicho inhibidor actúe es q_i, lo que significa que la probabilidad de que la causa X_i, actuando en ausencia de otras causas, llegue a producir el efecto Y es de $1 - q_i$. Además, el efecto Y sólo está ausente cuando todas las causas están ausentes

o, lo que es lo mismo, cuando para cada causa presente ha actuado el correspondiente inhibidor, es decir:

$$P(\neg y|x) = \prod_{i \in X_p} q_i \tag{1.3}$$

siendo $X_p \subseteq X$ el conjunto de las causas que están presentes.

Por último, la **puerta OR residual** (*leaky noisy OR*) es una generalización de la puerta OR ruidosa que permite simplificar los modelos reales al considerar la posibilidad de no tener que representar absolutamente todas las causas de posibles de un efecto, sino solamente las más relevantes para el experto. De este modo, el efecto puede estar presente incluso cuando ninguna de las causas explícitas está presente:

$$P(+y|\neg x_1 \wedge \neg x_2 \wedge ... \wedge \neg x_n) = c \tag{1.4}$$

lo que equivale a

$$P(\neg y|x) = \prod_{i \in X_p} (1-c)q_i \tag{1.5}$$

Puerta MAX

La puerta MAX causal [39, 84] es una generalización de la puerta OR en el sentido de que admite que las variables que forman parte del modelo no sean binarias, aunque forzosamente deben ser graduadas, es decir, que sus estados admitan una ordenación creciente dependiendo de los distintos grados de intensidad que representan. Este modelo se generaliza en [43].

La **puerta MAX determinista** se basa en considerar que el valor que toma Y es el estado que corresponde al máximo de los valores de Y producidos por las causas presentes si hubieran actuado independientemente.

La **puerta MAX ruidosa**, también conocida como **puerta OR graduada** [45] fue introducida por Henrion [84], aunque fue Díez quien la formalizó [39, 43]. Las condiciones que se deben cumplir para tener una puerta MAX ruidosa son:

$$P(Y = 0|\neg X_1 \wedge \neg X_2 \wedge ... \wedge \neg X_n) = 1 \tag{1.6}$$

$$P(Y \leq y|X) = \prod_{i \in X_p} P(Y \leq y|X_i = x_i) \tag{1.7}$$

siendo $X_p \subseteq X$ el conjunto que representa las causas que están presentes.

De forma análoga al caso de la puerta OR, la **puerta MAX residual** se basa en considerar la posibilidad de que el efecto esté presente incluso cuando ninguna de las causas lo está. La diferencia con la puerta OR residual está en que ahora necesitamos proporcionar un valor por cada uno de los estados del efecto Y.

Puertas AND y MIN

Los modelos AND y MIN han sido formalizados por Díez y Druzdzel [43]. Si las puertas OR y MAX se basaban en el máximo, las puertas AND y MIN se basan en el mínimo. Como el modelo AND se puede considerar un caso particular del modelo MIN, nos centraremos en éste último.

La **puerta MIN determinista** se basa en considerar que el valor que toma Y es el estado que corresponde al mínimo al que lo pueden hacer llegar las posibles causas presentes.

La **puerta MIN ruidosa**, también conocida como **puerta AND graduada** [45] fue formalizada por Díez [39]. Las condiciones que se deben cumplir para tener una puerta MIN ruidosa son:

$$P(Y = 0|\neg X_1 \wedge \neg X_2 \wedge ... \wedge \neg X_n) = 1 \tag{1.8}$$

$$P(Y \geq y|X) = \prod_{i \in X_p} P(Y \geq y|X_i = x_i) \tag{1.9}$$

siendo $X_p \subseteq X$ el conjunto que representa las causas que están presentes.

1.3.3. Interacción con los expertos

Una cuestión muy importante a la hora de construir redes bayesianas con la ayuda de expertos humanos es que hay que tener muy en cuenta las ventajas, pero sobre todo las dificultades inherentes al proceso de interacción con los expertos.

Entre las ventajas hay que destacar la gran experiencia que aportan basada en su experiencia, ya que muchas veces la información que se encuentra en la literatura no está actualizada, especialmente en medicina. Por otra parte, permiten al ingeniero de conocimiento resolver de forma más o menos inmediata aquéllas dudas que le pudieran surgir relacionadas con el conocimiento experto, ya que al ser tan especializado puede ser difícil de entender. Además, en muchas ocasiones puede resultar muy complicado obtener las fuentes de información necesarias para modelar el dominio.

No obstante, las dificultades son muy numerosas. La primera de ellas es que hay que encontrar a expertos que quieran colaborar en la tarea, lo que la mayoría de las veces no es

fácil por múltiples motivos: escasez de tiempo (sobre todo si no existen compensaciones económicas), escepticismo, incredulidad, etc. En el caso de que se trabaje únicamente con un experto, la información puede estar sesgada; sin embargo, si son varios los que colaboran los problemas pueden surgir al intentar buscar un acuerdo entre ellos. Por otro lado, la comunicación entre el ingeniero de conocimiento y los expertos es a veces muy difícil, pues la forma de expresarse de éstos suele ser muy distinta al modo en que lo hace el ingeniero de conocimiento, y viceversa, debido fundamentalmente a las posibles interpretaciones que admiten los discursos habituales en lenguaje natural. Esto exige que en las primeras sesiones de trabajo se dedique una parte del tiempo a definir e interpretar las formas de interacción. Finalmente, a la hora de asignar probabilidades, ya hemos indicado los trabajos que hay sobre los errores en sus estimaciones, pero nos gustaría incidir en las conclusiones que hemos sacado nosotros a la luz de nuestra interacción con expertos médicos. Hemos detectado dos grandes problemas: el primero es que no son capaces de pensar en la población en general, sino en la muestra constituida por el conjunto de sus pacientes; esto conlleva que las probabilidades asignadas de forma subjetiva sean mucho mayores que los reales. Otro problema es que al pedir que definan la frecuencia de un suceso, la estimación dependerá de la escala con la que trabajen, es decir, el valor será distinto si lo tienen que asignar entre 0 y 1, que si lo tienen que hacer entre 0 y 100 o entre 0 y 1000, por lo que no es sencillo.

1.4. Ventajas y desventajas

Entre las ventajas de las redes bayesianas hay que destacar el hecho de que tienen una clara interpretación, que son capaces de combinar de forma sencilla datos estadísticos y datos estimados de forma subjetiva y que se pueden justificar sobre bases teóricas sólidas. Por el contrario, su principal desventaja es que el método de razonamiento llevado a cabo se basa en una aproximación normativa, basado en una teoría matemática y, en consecuencia, la comprensión del proceso de inferencia es más difícil que en los métodos que tratan de imitar el razonamiento humano. Y es precisamente esta carencia la que ha constituido una de las grandes críticas que han sufrido los sistemas probabilísticos desde sus orígenes. El argumento esgrimido era la poca relación existente entre la forma de razonamiento humano y los métodos de razonamiento probabilístico. Sin embargo, estudios posteriores han demostrado que el problema no radica tanto en la ausencia de relación entre ambos tipos de razonamiento, sino en que la mayoría de la gente no hace interpretaciones correctas de las probabilidades [50].

Por otro lado, es precisamente este hecho uno de los que más dificulta el proceso de construcción manual de redes bayesianas, puesto que éste ha de hacerse casi exclusivamente con los datos que aportan los expertos en el dominio modelado. En este sentido, la definición del grafo puede ser más o menos sencilla; lo más complicado suele ser la asignación de valores a los parámetros.

Por eso, la aceptación de las redes bayesianas crecerá considerablemente si se dota a este tipo de sistemas de un método de explicación que, por un lado, ayude a los ingenieros de conocimiento y a los expertos humanos a *interaccionar* entre sí con el objetivo de crear un modelo del problema real; y, por otro, que permita *justificar* los resultados obtenidos y el proceso de razonamiento,

En consecuencia, la necesidad de métodos de explicación es incluso más importante en este tipo de modelos que en los sistemas expertos heurísticos.

1.5. Objetivos

A pesar de que, como veremos más adelante, ciertos autores han realizado grandes esfuerzos por dotar a estos modelos probabilísticos de cierta capacidad de explicación, no se ha encontrado todavía uno completamente satisfactorio, aunque existen algunos que proporcionan explicaciones útiles. Por eso, el principal objetivo inicial que nos planteamos fue el de la definición de un método para la **generación automática de explicaciones para redes bayesianas causales,** aunque adelantamos que algunas de las propuestas servirán también para otros tipos de modelos probabilísticos. Para conseguirlo, nos propusimos los siguientes objetivos previos:

1. analizar lo que se había hecho hasta el momento en el campo de la explicación, tanto en sistemas expertos heurísticos como en redes bayesianas. Es decir, intentar definir lo que se denomina en el argot científico el *estado del arte*;

2. extraer las características y propiedades básicas relacionadas con las capacidades de los métodos de explicación;

3. desarrollar un nuevo método de explicación para redes bayesianas que mejorase los existentes;

4. construir una red bayesiana sobre la que aplicar el método desarrollado, para así poder evaluar la eficacia del método de explicación y, por tanto, que sirviera a la vez de retroalimentación al objetivo anterior;

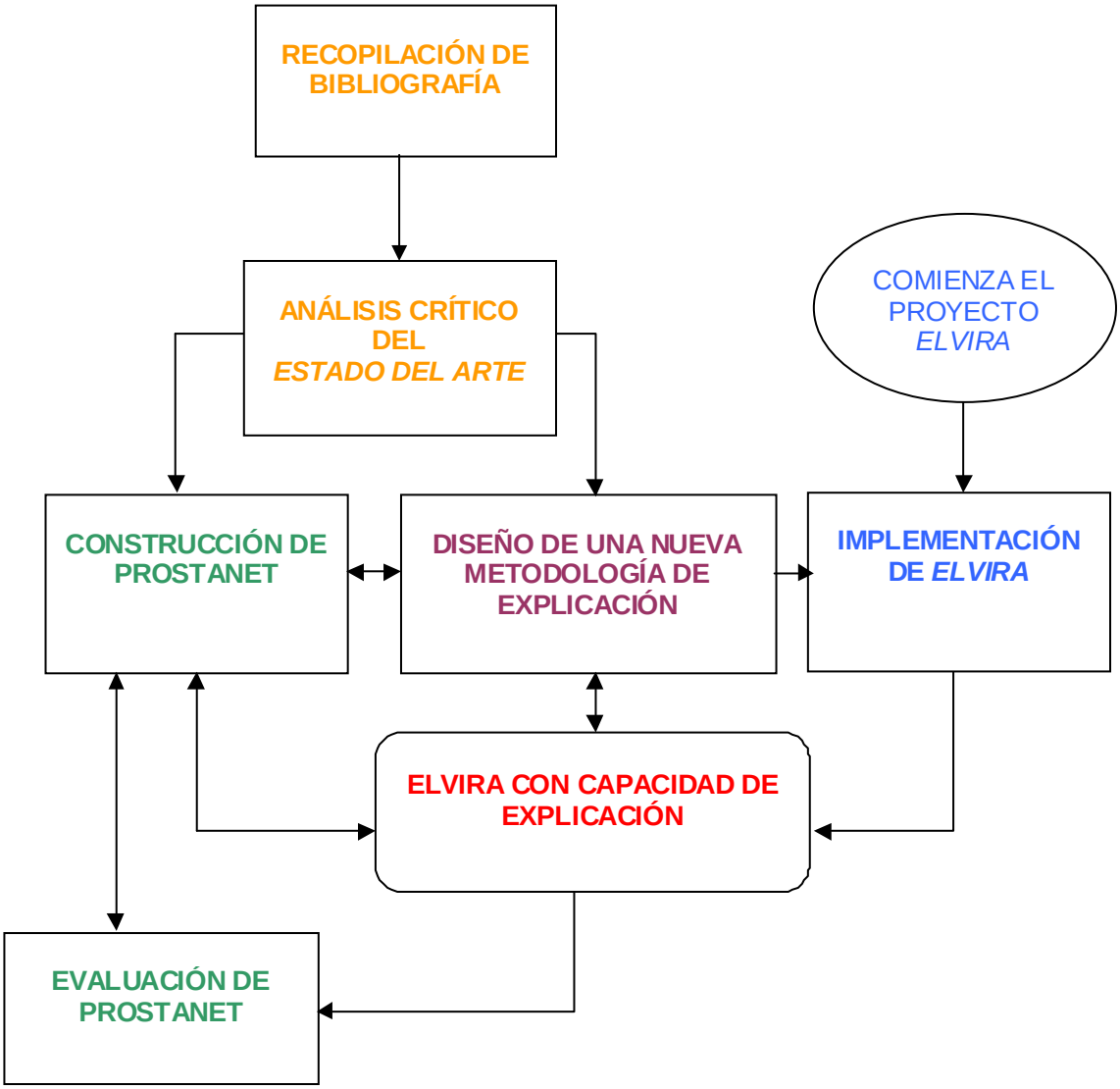

Figura 1.3: Fases en el desarrollo de la tesis doctoral

5. y, por último, implementarlo como parte de algún programa genérico para la edición y evaluación de redes bayesianas, integrándolo dentro de su interfaz gráfica.

1.6. Metodología

La metodología seguida para la consecución de los objetivos previamente citados ha constado de varias fases, tal y como podemos apreciar en la figura 1.3.

La primera de ellas consistió en recopilar la bibliografía relacionada con los métodos de explicación desarrollados hasta ahora, con el fin de analizar sus ventajas y sus carencias y poder extraer las características básicas de explicación. A partir de ahí, la

siguiente fase consistió en analizar el *estado del arte* en ese momento con el fin de intentar definir posteriormente una nueva metodología de explicación, haciendo hincapié como indicamos anteriormente, en la explicación gráfica. Por otra parte, en esa época comenzó a desarrollarse ELVIRA [26], un entorno para la creación y evaluación de modelos gráficos probabilísticos, como fruto de un proyecto de investigación entre varias universidades españolas financiado por la CICYT (TIC97-1135-C04). Por tanto, decidimos aprovechar la oportunidad que teníamos para dotar a ELVIRA de capacidad de explicación mediante la implementación de nuestra metodología, para lo que tuvimos que dedicar una gran cantidad de tiempo y esfuerzo a las tareas relacionadas con su implementación. Paralelamente a esta última, se inició la fase de construcción de la red bayesiana. Para ello, y gracias a la colaboración del urólogo el Dr. Diego Alberto Rodríguez Leal, y a su gran afición e interés por el mundo de la informática y de la inteligencia artificial, decidimos que el dominio de aplicación de nuestro método estaría centrada en urología. Sin embargo, al ser un campo tan amplio y teniendo en cuenta las necesidades del especialista, decidimos centrarnos en el diagnóstico del cáncer de próstata, que es el dominio de PROSTANET, una red bayesiana de ayuda al diagnóstico de dicha enfermedad.

Como podemos observar en la figura 1.3 las fases no se han desarrollado de forma secuencial, sino que algunas se han realimentado del trabajo de otras, como en el caso de la relacionada con la construcción de la red bayesiana y la del desarrollo de la nueva metodología, puesto que ésta nos ha servido como herramienta para la depuración de PROSTANET. A su vez, la construcción manual de esta red y la interacción con el experto nos ha servido como referencia para la evaluación del método de explicación desarrollado.

1.7. Organización de la tesis

La organización de esta memoria se ilustra en la figura 1.4 y está basada en una *jerarquía de niveles* [121, Capítulo 2].

La teoría de niveles fue introducida por David Marr [111] y Allen Newell [131]. En computación, los niveles propuestos por Marr son:

Teoría computacional, donde se indica el objetivo de la computación, su justificación y la teoría necesaria para implementarlo. Sería equivalente al nivel de conocimiento propuesto por Newell para la inteligencia artificial. En nuestro caso, corresponde con el estudio de la teoría de las redes bayesianas, haciendo especial hincapié en los métodos de explicación desarrollados hasta ahora para poder realizar la definición de uno nuevo.

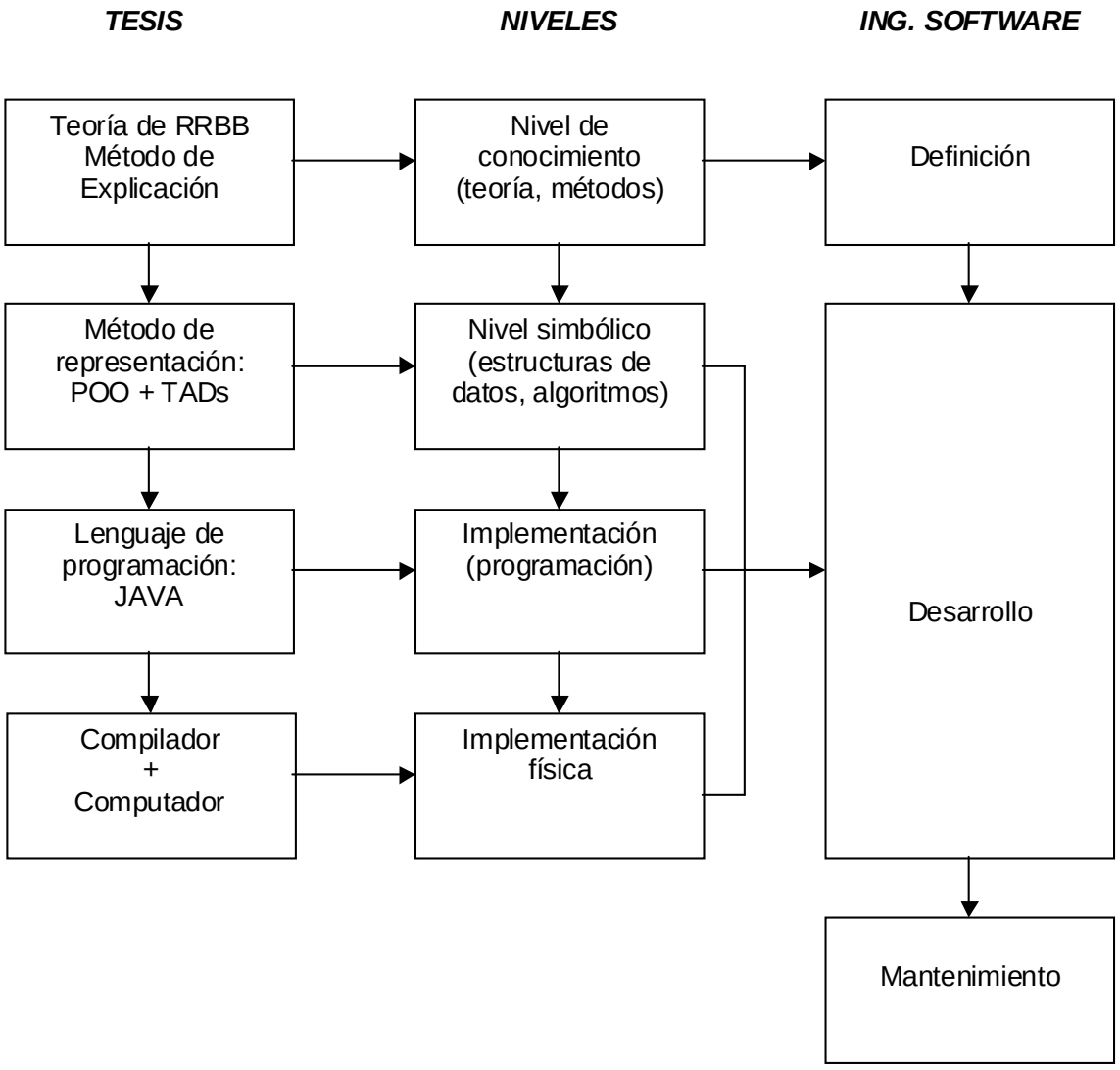

Figura 1.4: Organización de la tesis basada en una jerarquía de niveles

Algoritmo. En esta fase se elige una representación para los espacios de entrada y salida y el algoritmo que enlaza dichas representaciones. Este nivel coincide con el nivel simbólico de Newell. Para ello, es necesario establecer la metodología a seguir y nosotros la hemos basado en la programación orientada a objetos y los tipos abstractos de datos.

Implementación Finalmente, el último nivel es el de la elección de un lenguaje de programación y la construcción del programa y coincide con el nivel físico propuesto por Newell. Para implementar nuestro modelo, nosotros hemos elegido el lenguaje JAVA.

Podemos observar que esta jerarquía de niveles tiene también relación con las fases de

desarrollo de cualquier proyecto de Ingeniería del Software [147]:

Definición, que consta de las etapas correspondientes al análisis del sistema, planificación del proyecto y análisis de requisitos. Los objetivos de esta fase, como podemos observar, se corresponden con los del nivel de conocimiento descrito arriba.

Desarrollo. En esta fase, se lleva a cabo el diseño del programa, su codificación y las pruebas necesarias para comprobar el correcto funcionamiento del mismo. Como se puede ver, englobaría a los niveles simbólico y físico del apartado anterior.

Mantenimiento, para llevar a cabo las tareas de corrección, adaptación y mejora que pudieran surgir durante la implantación y funcionamiento del sistema. Esta última fase no tiene equivalente en el apartado anterior.

Con esta perspectiva, la tesis se estructura en siete capítulos, organizados así:

- El primero, como vemos, corresponde a una introducción dedicada a exponer la motivación que ha dado origen a este trabajo, así como a presentar los conocimientos básicos para un mejor entendimiento del resto de la tesis. Así pues, comenzamos definiendo los conceptos fundamentales sobre redes bayesianas para exponer posteriormente, puesto que constituye uno de nuestros objetivos, los aspectos básicos relacionados con su construcción manual: elementos necesarios y principales etapas; los modelos canónicos, que permiten simplificar el proceso de obtención de parámetros; y las ventajas y dificultades que supone la interacción con los expertos humanos.

- En el capítulo 2 realizamos un análisis de los métodos de explicación desarrollados hasta el momento, tanto para sistemas expertos heurísticos como para redes bayesianas. Aunque en algunos trabajos de investigación [50, 174] se proporcionan algunas revisiones del *estado del arte* en el campo de la explicación, sin embargo no existía hasta la fecha ningún tipo de referencia sobre la que realizar un estudio crítico. Por otra parte, a la luz de los trabajos analizados, tampoco existía unanimidad de criterios sobre lo que debe ser una explicación. Por tanto, nuestro primer objetivo, consistió en **definir** el término explicación para, a continuación, extraer una serie de **propiedades básicas** de los métodos de explicación. Estas propiedades se pueden clasificar atendiendo a tres conceptos, basándonos en la definición del término "explicación": el contenido de la explicación, la forma de presentarla al usuario y la posibilidad de adaptación a sus necesidades. Por tanto, la **revisión de**

los métodos de explicación la hemos hecho en base a estas propiedades, lo que nos ha permitido realizar un estudio crítico de cada uno de ellos. No obstante, el estudio lo hemos dividido en dos partes: la primera consiste en el análisis de los métodos de explicación desarrollados para sistemas expertos heurísticos, clasificándolos en dos grandes grupos según la capacidad de adaptación al usuario. La segunda parte es una presentación de los métodos de explicación para redes bayesianas, agrupados en tres grandes categorías según el objeto de la explicación: explicación de la evidencia, explicación del modelo representado y explicación del razonamiento llevado a cabo sobre la red.

- Después de este análisis, los capítulos 3, 4 y 5 presentan la nueva metodología que proponemos. Los capítulos 3 y 4 incluyen una descripción teórica del método, así como su justificación. Su implementación se desarrolla en el capítulo 5. Podríamos haber mezclado la descripción teórica con la implementación pero hemos preferido separar ambas partes, teniendo en cuenta la distinción de ambos conceptos que se realiza en distintos ámbitos de la informática: niveles de Newell [131], niveles de computación de D. Marr [111], fases de desarrollo de ingeniería del software [147], definición de tipos abstractos de datos [76], etc. Por otra parte, puesto que como veremos en el capítulo 2, la mayor parte de la investigación sobre explicación en redes bayesianas se ha centrado en explicar la evidencia introducida en la red, nosotros hemos preferido dedicarnos a la generación automática de explicaciones del modelo representado y del razonamiento.

- Así pues, el capítulo 3 comienza con la justificación de la necesidad de generar explicaciones automáticas sobre la información representada en una red bayesiana, de cara a la construcción manual de la misma y a posibles fines educativos. A continuación se presentan de forma detallada el modo de generar la **explicación del modelo** de una red bayesiana, dependiendo de si lo que se desea explicar es un nodo, un enlace o la red completa. Finalmente, se realiza una valoración del método propuesto, detallando sus características de acuerdo con las propiedades definidas en el capítulo 2, así como sus principales limitaciones.

- El capítulo 4 tiene la misma estructura que el anterior: comienza con una introducción en la que se justifica la necesidad y la utilidad de la generación automática de **explicación del razonamiento** de una red bayesiana, para describir después varias formas de explicación mediante la presentación gráfica de las probabilidades de los nodos, la explicación de casos de evidencia, la posibilidad de realizar análisis

de sensibilidad, y la visualización de los caminos de razonamiento desde la evidencia hasta una hipótesis determinada. El capítulo concluye con la valoración de las propuestas realizadas, también a la luz de las categorías definidas en el capítulo 2.

- El siguiente capítulo, el 5, expone la **implementación** de las propuestas de los capítulos 3 y 4 y las **herramientas utilizadas**. Ya hemos indicado que, aprovechando la puesta en marcha del proyecto ELVIRA, lo utilizamos para integrar en dicho entorno el método de explicación propuesto. Por eso, a continuación indicamos las herramientas de programación utilizadas para llevar a cabo el desarrollo de ELVIRA y su capacidad de explicación. Nuevamente, esta presentación la haremos teniendo en cuenta la jerarquía de niveles de Newell que distingue entre nivel de conocimiento, simbólico y físico. En el nivel de conocimiento detallamos los objetivos a alcanzar con la creación del nuevo entorno ELVIRA, así como la justificación del paradigma de la orientación a objetos como herramienta para la representación de la solución del problema. Las estructuras de datos y algoritmos principales los describiremos en el apartado relativo al nivel simbólico. En el nivel de implementación, justificamos la elección del lenguaje JAVA.

  Seguidamente describimos la interfaz gráfica de ELVIRA, tanto en modo edición como en modo inferencia y, por último, dedicamos la parte final del capítulo a presentar las capacidades de explicación de las que se ha dotado a ELVIRA, tanto del modelo como del razonamiento.

  Por último, el capítulo acaba con un análisis comparativo de las capacidades de ELVIRA frente a las que aportan las principales herramientas de edición y evaluación de redes bayesianas, tanto comerciales como educativas.

- No tendría sentido dedicar esfuerzos a definir un nuevo método de explicación sin demostrar las ventajas que éste aporta con respecto a los existentes. Por otro lado, puesto que una de las grandes deficiencias de las redes bayesianas es la falta de herramientas para la construcción manual de las mismas, pensamos que experimentando por nosotros mismos las carencias existentes podríamos aprovechar lo aprendido para incluirlo en nuestro método de explicación. Por eso, en el capítulo 6 presentamos la **aplicación desarrollada** para el diagnóstico del cáncer de próstata: la red PROSTANET. El capítulo comienza con una introducción a los conceptos médicos relacionados con la próstata y el diagnóstico del carcinoma de próstata, con el fin de familiarizar al lector con la terminología que va a aparecer a lo largo del tema y su significado. El resto del capítulo está dedicado a detallar el proceso de

construcción manual de la red Prostanet, según las fases descritas en la primera sección: construcción del grafo cualitativo, definición de los estados de las variables, identificación de los modelos canónicos y obtención de las probabilidades. Puesto que la red se ha construido mediante un proceso cíclico formado por distintas versiones, incluimos un apartado a su descripción, dependiendo de la fase de construcción de la misma.

El final del capítulo se dedica a evaluar la red, describiendo la metodología seguida y los resultados obtenidos.

- El último capítulo de la tesis pretende resumir el trabajo completo, presentando las principales conclusiones del mismo, para lo que se enumeran las aportaciones realizadas y las cuestiones que han quedado abiertas para futuras líneas de investigación.

Capítulo 2

Métodos de explicación en sistemas expertos

En este capítulo, comenzamos definiendo el significado del término explicación (sección 2.1), para pasar posteriormente a estudiar y determinar las características básicas que distinguen a los distintos métodos de explicación (sección 2.2). Precisamente estas características son las que nos servirán de guía a lo largo de todo nuestro trabajo para describir, a continuación, el *estado del arte* de los métodos de explicación desarrollados hasta ahora. Éste se llevará a cabo en dos partes: en la primera, analizamos los diseñados para sistemas expertos heurísticos (sección 2.3) y en la segunda parte, analizamos los métodos diseñados específicamente para redes bayesianas (sección 2.4). Para una mejor comprensión del análisis de cada uno de los métodos, tendremos en cuenta las propiedades resumidas en la tabla 2.1 (sección 2.2).

2.1. Definición de explicación

El término explicación tiene múltiples acepciones, dependiendo del momento histórico y del contexto en el que se aplique [124]. Desde los orígenes de la filosofía y de la ciencia, numerosos estudiosos de ambas disciplinas han intentado determinar su significado, vinculándolo con la comprensión y la descripción, pero sin fijar un sentido único. Así, por ejemplo, para algunos autores, la explicación consiste en la búsqueda de las causas de un hecho; otros, distinguen entre explicación y comprensión; algunos científicos opinan, de manera diferente, que explicar consiste en deducir leyes y condiciones que describan un proceso. En el ámbito de la lingüística aparece también esta multiplicidad de significados. Por ejemplo, entre las acepciones que proporciona el diccionario de la Real Academia Española están "Declaración o exposición de cualquier materia, doctrina o texto con palabras claras o ejemplos, para que se haga más perceptible | ... | Manifestación o revelación

de la causa o motivo de alguna cosa". Según el diccionario de María Moliner, explicar consiste en "hablar sobre una cosa para hacerla comprender o conocer a otros". En el caso de los sistemas expertos, ocurre lo mismo que en otras disciplinas: no existe unanimidad de criterios sobre qué es la explicación y, no menos importante, sobre cómo ofrecerla al usuario. De hecho, podemos encontrar, desde los trabajos que consideran una explicación del modo más simple, como una mera descripción de datos, hasta los más sofisticados, en los que constituye un diálogo "inteligente" en lenguaje natural con el usuario del sistema. A pesar de ello, sí existe coincidencia en el objetivo que se persigue al dotar a los sistemas expertos de una herramienta de explicación, que es el de dar respuestas a preguntas del tipo "¿por qué el sistema ha obtenido este resultado?" o "¿cómo ha sido el proceso que ha llevado a esta conclusión?", con el fin de ayudar al usuario a comprender las conclusiones que le presenta el sistema. En el caso concreto de las redes bayesianas, se traduce en ayudar al usuario a entender la estructura y los parámetros que contiene, especialmente en el caso de redes aprendidas, o a interpretar las probabilidades a posteriori que el sistema le proporciona.

Por todo ello, podemos concluir que explicar consiste en **exponer algo** de tal forma que sea

- **comprensible** para el receptor de la explicación, lo que implica que mejore su conocimiento sobre el objeto de la explicación, y
- **satisfactorio**, en el sentido de que cubra las expectativas del usuario.

2.2. Características básicas

Pese a tanta diversidad de significados, el hecho de coincidir en un mismo objetivo permite aislar ciertas características básicas, comunes a todos los métodos de explicación y que intentaremos exponer en las secciones 2.3 y 2.4.

Para intentar hacer esta exposición lo más clara posible, nos basaremos en el resumen publicado por Wick [188] del seminario sobre explicación organizado por la AAAI en 1988, en el que se muestran las distintas líneas de investigación en el área de explicación en sistemas expertos, así como las cuestiones básicas que se plantean a la hora de construir este tipo de herramientas, que son: el contenido de la explicación, si hay que considerar la explicación como un proceso que hay que tener en cuenta durante la adquisición del conocimiento, los distintos tipos de usuario a los que va dirigida, los tipos de preguntas a realizar por éstos, y el modo de presentar la explicación.

Todas estas ideas, más las que han surgido durante el análisis de las herramientas de explicación desarrolladas hasta ahora, las podemos ver resumidas en la tabla 2.1.

El contenido de esta tabla lo analizamos con cierta profundidad en las secciones siguientes, y nos servirá como base para identificar las propiedades básicas de los métodos de explicación durante su estudio. La idea subyacente consiste en la agrupación de las características de la explicación en tres categorías:

- **contenido: qué** explicar;
- **comunicación**: **cómo** interactúa el sistema con el usuario;
- **adaptación**: **a quién** va dirigida la explicación.

Observamos que estas categorías corresponden con la definición de explicación dada en la sección 2.1: exponer algo (*contenido*) de tal forma (*comunicación*) que sea comprensible para el receptor de la explicación, lo que implica que mejore su conocimiento sobre el objeto de la explicación y resulte satisfactorio, en el sentido de que cubra las expectativas del usuario (*adaptación*).

2.2.1. Contenido

Uno de los aspectos más importantes de le explicación está relacionado con el contenido de la misma, es decir, *qué* es lo que debe incluir para que sea comprensible por el usuario. Esta es una cuestión no trivial porque depende de varios aspectos como son el objeto de la explicación, el propósito que se persigue y el nivel de detalle que requiere el usuario, así como si se va a basar en modelos causales o no.

Objeto de la explicación

En el caso de redes bayesianas, hay tres aspectos que pueden (y deben) ser explicados: la base de conocimiento, el proceso de razonamiento seguido para obtener ciertos resultados y la evidencia introducida. Por eso, los métodos de explicación los podemos clasificar en tres grupos, dependiendo del objeto de la explicación:

- Explicación de la **evidencia**: Consiste en determinar qué valores de las variables no observadas justifican la evidencia disponible. Este proceso se suele conocer como *abducción*, y está basado en la suposición (normalmente implícita) de que existe un modelo causal —véase la sección 1.2.2. En este contexto, una *explicación* es una configuración de las variables no observadas, y el objetivo del proceso de inferencia

Contenido	Objeto	Evidencia Modelo Razonamiento
	Propósito	Descripción Comprensión
	Nivel	Micro Macro
	Causalidad	Causal No Causal
Comunicación	Interacción Usuario-Sistema	Menú Preguntas predefinidas Diálogo en lenguaje natural
	Presentación	Texto Gráficos Multimedia
	Expresión de la probabilidad	Textual Numérica Ambas
Adaptación	Conocimiento del usuario sobre el dominio	Sin modelo Escala Modelo dinámico
	Conocimiento del usuario sobre el razonamiento	Sin modelo Escala Modelo dinámico
	Nivel de detalle	Fijo Manual Automático

Tabla 2.1: Propiedades de los métodos de explicación.

es obtener la explicación más probable (MPE)[1] o las k explicaciones más probables. En general, las variables que toman el valor "presente" o "positivo" en la MPE son consideradas como las causas que *explican* la evidencia. El objetivo de este tipo de explicación es básicamente ofrecer un diagnóstico para un conjunto de anomalías observadas. Por ejemplo, en sistemas expertos médicos, una explicación consistiría en determinar la enfermedad o enfermedades que explican la evidencia: síntomas, signos, resultados de pruebas, etc.

- Explicación del **modelo**: Consiste en mostrar (verbalmente, gráficamente o mediante marcos [120]) la información contenida en la base de conocimiento. Este tipo de explicación se conoce también como *estática* [85]. El objetivo es tanto permitir al experto humano examinar el contenido de la base de conocimiento durante la fase de *construcción* del sistema experto, como ofrecer al usuario novel algún tipo de conocimiento sobre el dominio modelado con *propósitos educativos*.

- Explicación del **razonamiento**: Este tipo de explicación se conoce a veces como *explicación dinámica* [85] y puede proporcionar tres tipos de justificación:

  - Los **resultados obtenidos** por el sistema y el **proceso de razonamiento** que los ha producido. Durante la *evaluación* del sistema experto, algunos de los casos de prueba son resueltos apropiadamente por el sistema y otros no. En el primer caso, los evaluadores (ingenieros del conocimiento y expertos humanos) pueden comprobar que todos los pasos en el camino a la solución fueron correctos (un argumento incorrecto podría llevar a una solución correcta por casualidad). En el caso de que el sistema experto cometa un error, la explicación del proceso de razonamiento es una herramienta inestimable para detectar y corregir las partes erróneas de información de la base de conocimiento. Durante el *funcionamiento* del sistema experto, la capacidad de explicación es muy útil para convencer al usuario de la corrección de los resultados. En el caso de los *sistemas tutoriales*, la capacidad de explicación es esencial para mejorar las capacidades de los estudiantes.

  - Los **resultados no obtenidos** por el sistema pero esperados por el usuario. Puede ser también interesante, especialmente para sistemas tutoriales, explicar por qué el sistema no produce una determinada conclusión esperada por el

[1] Del inglés **M**ost **P**robable **E**xplanation.

usuario; en particular, qué hallazgos son los que contradicen dicha conclusión y cuáles serían necesarios para obtenerla.

- **Razonamiento hipotético**, es decir, qué resultados habría producido el sistema si una o más variables hubieran tomado valores distintos a los observados.

Propósito

Existen dos objetivos diferentes que un método de explicación debe tratar de alcanzar, lo que conduce a dos tipos de explicación:

- **Descripción** de los estados que representan el modelo y de los valores resultantes que proporciona el sistema. En este caso, la explicación consistirá en presentar un conjunto de variables o un conjunto de reglas y, posiblemente, ciertos valores numéricos asociados para mostrar la *base de conocimiento* subyacente (en el caso de la explicación del modelo), o en proporcionar detalles adicionales sobre las *conclusiones*, o en mostrar *resultados intermedios* (en el caso de la explicación del razonamiento). Por ejemplo, en el caso concreto de las redes bayesianas, la descripción de un estado vendría dada por una colección de variables, los valores que toman y las probabilidades asociadas.

- **Comprensión**. En este caso, a diferencia del anterior, la explicación intenta hacer comprender al usuario las implicaciones del modelo o las conclusiones del sistema o las relaciones entre ellos. Normalmente, dicha explicación se presentará mediante un discurso, expresado en lenguaje natural, en modo gráfico, etc., en el que se indiquen, por ejemplo, las relaciones entre los datos, de qué manera influye cada uno, bien individualmente o junto con otros, sobre la conclusión; y, sobre todo, se deben identificar y justificar las causas que producen los resultados obtenidos.

Niveles de Explicación

Sember y Zukerman [156] propusieron otra clasificación de los métodos de explicación para redes bayesianas, aunque es útil para cualquier tipo de sistema experto:

- **Nivel Micro**. Una explicación a nivel micro, en el caso de las redes bayesianas, consiste en una justificación detallada de las variaciones que se producen en un nodo particular como consecuencia de las variaciones de sus vecinos. En el caso de un sistema experto basado en reglas, el nivel micro consistiría en el análisis de las variables de una determinada regla o de las reglas que contienen cierta variable.

- **Nivel Macro**. Este nivel de explicación analiza las principales líneas de razonamiento que llevan de la evidencia a una conclusión determinada. Una explicación macro para una red bayesiana vendrá dada por una serie de caminos en la red, mientras que para un sistema experto tradicional consistiría en una o varias cadenas de reglas.

Causalidad

Como ya hemos analizado en la sección 1.2.2, algunos de los métodos de explicación tanto para sistemas expertos heurísticos como para redes bayesianas están diseñados especialmente para modelos causales, lo que permite ofrecer una interpretación causal, tanto del modelo como del razonamiento. Sin embargo, existen otros métodos generales, en el sentido de que no requieren una interpretación causal del modelo que representan.

2.2.2. Comunicación

Un aspecto crucial de la explicación es la forma en que se ofrece al usuario. Ésta depende, básicamente, de la forma en la que se presenta y de los métodos de interacción entre el usuario y el sistema. En el caso de redes bayesianas merece especial interés el modo en que se expresan las probabilidades.

Interacción Usuario-Sistema

En algunos sistemas los usuarios pueden solicitar una explicación seleccionando algunas variables u opciones de un *menú* determinado. En otros, los usuarios pueden realizar *preguntas*, normalmente predefinidas por el sistema; por ejemplo, en MYCIN las preguntas disponibles eran "cómo [ha obtenido el sistema una conclusión]" y "por qué [el sistema solicita esta información]". Finalmente, hay otros sistemas que intentan aproximarse a la posibilidad de ofrecer un *diálogo* en lenguaje natural "recordando" (analizando y construyendo un modelo de) la conversación, con el objetivo de "comprender", en el contexto adecuado, cada pregunta y respuesta.

Por otro lado, en algunos sistemas el usuario puede solicitar una explicación después de que el programa haya presentado sus conclusiones, mientras que otros permiten al usuario interrumpir el proceso de inferencia para solicitar la explicación.

Presentación de la Explicación

Podemos distinguir los siguientes modos de presentar la explicación al usuario:

- **Verbalmente**, utilizando texto y números.
- **Gráficamente**, de dos formas:
  - utilizando diagramas de barras, gráficos, etc., que representen la relación entre valores o variables, la evolución de las probabilidades asociadas durante la propagación de la evidencia, etc.
  - utilizando el propio grafo de la red bayesiana, de tal forma que los cambios en las distribuciones de probabilidad se indican añadiendo texto, números o signos, o coloreando los nodos y/o los enlaces.
- **Multimedia**, integrando los anteriores en un entorno de hipertexto y multimedia que combine las explicaciones interactivas con imágenes, vídeo y sonido.

Expresión de la probabilidad

Las formas de expresar la probabilidad pueden ser *numérica* o *cuantitativa*, como 0'98 o 72'6 %, y *lingüística* o *cualitativa*, como "raramente", "muy probable" o "casi seguro". Del mismo modo, al comparar dos probabilidades es posible hacerlo con expresiones cuantitativas, como "la razón entre las probabilidades de A y B es 10^{-3}", o expresiones cualitativas, tales como "B es mucho más probable que A". Existe una cantidad significativa de trabajos de investigación sobre la asignación de expresiones lingüísticas de probabilidad a valores numéricos y viceversa —véase [49, 50, 182] para una revisión de ellos, y también la sección 2.4.3 de este trabajo. Sin embargo, también existen artículos relacionados con las posibles malinterpretaciones que se pueden generar al intentar expresar verbalmente las probabilidades, especialmente en medicina [126], por lo que algunos autores afirman que los médicos deberían utilizar sólo los valores numéricos para expresar probabilidades [8].

2.2.3. Adaptación al usuario

Explicar siempre implica exponer algo a *alguien*. Por tanto, una de las características claves de una explicación efectiva es la capacidad de adaptarse a las necesidades específicas de cada usuario y a sus expectativas, lo que depende básicamente del conocimiento que éste tiene. Sin embargo, éste es un problema complejo, pues implica conocer en cada momento cuál es el nivel de conocimiento del usuario e ir actualizándolo a la vez que éste progresa, no sólo durante el funcionamiento del sistema, sino incluso después, lo que supondría que

el nivel de conocimiento que tiene el usuario cada vez que maneja el sistema puede haber variado. Existen tres cuestiones que hay que considerar de modo independiente:

Conocimiento del usuario sobre el dominio

Algunos métodos de explicación no tienen en cuenta la variabilidad del conocimiento del dominio entre distintos usuarios. Por tanto, las explicaciones se generan únicamente para un usuario hipotético con un conocimiento determinado; por ejemplo hay métodos de explicación diseñados para principiantes mientras que otros presuponen que el usuario es un experto.

En otros muchos casos, el modelo de usuario se reduce a utilizar una **escala** con dos o tres categorías: "principiante, avanzado, experto". En principio, un sistema experto debería tener la capacidad de clasificar a los usuarios automáticamente, pero en la práctica el usuario es quien se tiene que autoclasificar según una escala dada, al comienzo de la interacción con el sistema.

Ambos tipos de sistemas coinciden en la utilización de un **modelo estático de usuario**, es decir, suponen que el conocimiento del usuario no se modifica conforme interactúa con el sistema. En estos sistemas el modelo de usuario está implícito, es decir, el diseñador del sistema lo tienen en mente al construirlo pero no lo representa explícitamente ni en el modelo ni en el motor de inferencia ni en los módulos de explicación.

En contraposición, otros métodos de explicación son capaces de tener en cuenta un **modelo dinámico** de usuario, que representa explícitamente su conocimiento y los cambios que éste experimenta al ir aprendiendo nuevos conceptos o relaciones durante su interacción son el sistema, lo que permite generar explicaciones adaptadas al conocimiento que tiene el usuario en cada momento.

Conocimiento del usuario sobre el método de razonamiento

De forma análoga, debería ser posible tener en cuenta el conocimiento del usuario sobre el método de razonamiento utilizado por el sistema experto. Por ejemplo, la explicación generada por un sistema experto bayesiano para un usuario que esté familiarizado con los conceptos de prevalencia, razón a priori/posteriori y razón de verosimilitud debería ser muy diferente de la explicación generada para un usuario que nunca haya oído hablar de ellos. Como en el caso del conocimiento sobre el dominio, las posibilidades varían entre suponer que el conocimiento del usuario sobre el método de razonamiento no varía hasta el modelo dinámico que representaría explícitamente su conocimiento en cada momento. Sin

embargo, como veremos más adelante, ninguno de los métodos de explicación desarrollados hasta la fecha ha modelado el conocimiento del usuario sobre el método del razonamiento.

Nivel de detalle

El nivel de detalle de la explicación está muy relacionado con el conocimiento del usuario pero puede establecerse de forma independiente. Por ejemplo, es posible para un sistema experto que no tenga en cuenta el conocimiento del usuario sobre el dominio, asignar un factor de importancia a cada elemento (cada regla o cada variable, por ejemplo) y un umbral de importancia, de tal forma que el método de explicación sólo muestre aquellos elementos cuyo factor de importancia supere dicho umbral. Disminuyendo este umbral el usuario puede incrementar el nivel de detalle y viceversa. De este modo es posible ofrecer diferentes niveles de detalle sin tener un modelo de usuario. Obviamente, en lugar de tener un factor de importancia fijo para cada elemento, sería deseable poder ajustar dinámicamente estos factores en función del conocimiento del usuario sobre el dominio, la evidencia disponible y la interacción con el sistema. Análogamente, los aspectos de la explicación relacionados con el método de razonamiento deberían ser ofrecidos en diferentes niveles de detalle.

2.3. Métodos de explicación en sistemas expertos heurísticos

Tras haber analizado en la sección anterior las características de los métodos de explicación, vamos a revisar ahora los trabajos realizados en este campo, tanto para sistemas expertos heurísticos (en esta sección) como para sistemas expertos probabilísticos (en la sección siguiente).

El auge que se viene produciendo durante esta última década de los entornos dedicados a la intercomunicación persona-computador, sobre todo en los sistemas aplicados al aprendizaje, se deja notar también a la hora de diseñar herramientas de explicación para sistemas expertos. Por esta razón, analizaremos dichos métodos de explicación atendiendo a su clasificación en dos grandes grupos:

1. **Métodos no intercomunicativos**, que son aquéllos en los que la explicación se genera de modo cerrado, sin adaptación a los diferentes tipos de usuarios que interaccionan con el sistema. Por tanto, la principal característica de este tipo de

métodos es que no existe posibilidad de que el sistema actualice, durante su funcionamiento, el conocimiento que a priori posee del usuario. Esto implica que tanto el **modelo de usuario** como el **nivel de detalle** de la explicación generada son **estáticos**. Otra característica propia de este tipo de métodos es que la **interacción usuario-sistema** normalmente se lleva a cabo a través de **preguntas u opciones predefinidas**.

2. **Métodos intercomunicativos**, que generan la explicación por medio de entornos en los que existe comunicación entre el sistema y el usuario y en los que, en consecuencia, tanto el usuario como el propio sistema, actualizan el conocimiento mutuo durante el proceso de explicación. Esto implica, fundamentalmente, la **adaptación** del sistema en cada momento a las necesidades y expectativas del usuario. Por esta razón a este tipo de sistemas se les denomina también *adaptativos*. Se caracterizan porque normalmente el **modelo de usuario** es **dinámico** y el **nivel de detalle** de la explicación se va actualizando **automáticamente**. Además, al contrario que en los métodos no intercomunicativos, una característica más de éstos es que la **interacción usuario-sistema** normalmente se lleva a cabo a través de **diálogo en lenguaje natural**.

Podemos adelantar que la mayoría de los métodos de explicación desarrollados hasta ahora se engloban dentro del primer tipo, los no intercomunicativos.

2.3.1. Métodos no intercomunicativos

Como se ha indicado, este tipo de métodos se caracterizan porque la presentación de la explicación correspondiente se genera de modo estático, sin permitir al usuario "conversar" con el sistema. Esto quiere decir que, por ejemplo, tanto los tipos de preguntas que el usuario puede realizar, como el nivel de detalle de la explicación están fijados previamente. Igualmente, y lo que es más significativo, el sistema no puede modificar, durante su funcionamiento, ni sus suposiciones sobre el conocimiento que posee el usuario, ni en general, sobre el contexto en el que se genera y presenta la explicación.

Como hay numerosos trabajos realizados en este campo, el estudio lo haremos atendiendo a una clasificación general de los tipos de sistemas expertos. Dentro de cada grupo, presentaremos los trabajos ordenados según su aparición cronológica.

Texto enlatado

El sistema de Bleich [6], diseñado para el diagnóstico de alteraciones relacionadas con ácidos y electrolitos, está considerado por algunos autores como uno de los primeros sistemas expertos. En él se utilizaba un mecanismo de explicación bastante rudimentario, conocido como técnica del "texto enlatado" (*canned text*) que consistía en almacenar, junto con el código del sistema, las respuestas de las preguntas que se le podían plantear. El método consistía en mostrar, dada cierta evidencia, las respuestas correspondientes y los valores de determinadas variables. El problema que conlleva la generación de este tipo de explicación es hay que prever todas las preguntas posibles, con la dificultad que esto supone. El objetivo de este tipo de explicación es la **comprensión del razonamiento** a **nivel macro** y las explicaciones se presentan de modo **textual**.

MYCIN

Uno de los grandes avances se produjo con el desarrollo del sistema médico MYCIN [9], para el diagnóstico y tratamiento de enfermedades infecciosas, y que fue pionero, entre otras cosas, en el campo de la generación de explicaciones. El funcionamiento del sistema consistía en la realización de preguntas al usuario con el fin de intentar recoger la información disponible. El método de explicación se basaba en traducir el conjunto de reglas utilizadas durante el proceso de razonamiento. Para ello, cada vez que el sistema generaba una pregunta, el usuario tenía la posibilidad de examinar el razonamiento que se había llevado a cabo hasta ese momento mediante dos tipos de preguntas:

- "¿Por qué [solicita el sistema esta información]?". En este caso, el sistema mostraba tanto los objetivos conseguidos hasta el momento relativos a las reglas que se habían ejecutado, como los que estaba intentando conseguir relativos a las reglas cuyo antecedente estaba tratando de confirmar. Además, se mostraban también las reglas que relacionaban estos objetivos, traducidas al inglés del lenguaje LISP.

- "¿Cómo [ha obtenido el sistema cierto objetivo]?". Para responder a este tipo de preguntas, el sistema mostraba la regla correspondiente, igualmente traducida al inglés.

De esta forma, si el usuario solicitaba sucesivamente los *porqués* y/o *cómos*, MYCIN era capaz de obtener la cadena de reglas que el sistema había encadenado o estaba intentando encadenar para obtener los resultados. El objetivo de este método es la **comprensión** tanto del **modelo** como del **razonamiento**. Las explicaciones se generan a **nivel micro**,

pues en cada momento se centra en una regla determinada, y la presentación de las mismas se realiza mediante **texto**; los valores de los grados de certeza asociados a cada regla se presentaban tanto de forma **numérica** como **lingüística**.

Sin embargo, hay que destacar dos inconvenientes de este método: la primera es que el mostrar el encadenamiento de reglas no implica que un usuario entienda el razonamiento seguido; la otra surge cuando hay un número elevado de reglas o de antecedentes de algunas de ellas. En este caso, la obtención de una línea de razonamiento coherente entre las mismas puede resultar bastante complicada.

NEOMYCIN

Para intentar solventar estas deficiencias se desarrolló NEOMYCIN [24, 80], un proyecto que incluía conocimiento **causal** explícito en forma de meta-reglas y una organización jerárquica de las hipótesis. Con ello se conseguía explicar no sólo los pasos que estaba siguiendo sino también su *estrategia de razonamiento*. El objetivo, por tanto, era también la **comprensión** tanto del **modelo** como del **razonamiento**. Las explicaciones se generan a **nivel macro** y la presentación de las mismas se realizaba mediante **texto**.

"MYCIN causal"

Uno de los primeros métodos que incluían la posibilidad de comunicación entre el usuario y el sistema fue el desarrollado por Wallis y Shortliffe [181], a partir del trabajo desarrollado en MYCIN. Estos autores diseñaron un método de explicación **causal**, utilizando una red de reglas para la representación de la base de conocimiento a partir de la diseñada para MYCIN. Esta representación causal servía tanto para la adquisición de nuevo conocimiento como para hacer de guía en la generación de explicaciones "a medida", dependiendo de las distintas necesidades y características de los distintos usuarios, ya que era capaz de tener en cuenta tanto el **modelo de usuario** como el **nivel de detalle** de la explicación. Para ello, asignaban una *medida de complejidad* y un *factor de importancia* tanto a las reglas como a los conceptos relacionados con las conclusiones del sistema. Además, definieron dos parámetros, la *experiencia* del usuario, para representar su nivel de conocimiento, y el *nivel de detalle* en el que el usuario deseaba obtener la explicación, aunque éste siempre tenía la opción de solicitar explicaciones con más nivel de detalle. Combinando estos parámetros se podían seleccionar de forma dinámica las reglas y los conceptos en función de las necesidades del usuario. El objetivo de este método es la **comprensión** tanto del **modelo** como del **razonamiento**. La explicación, a **nivel**

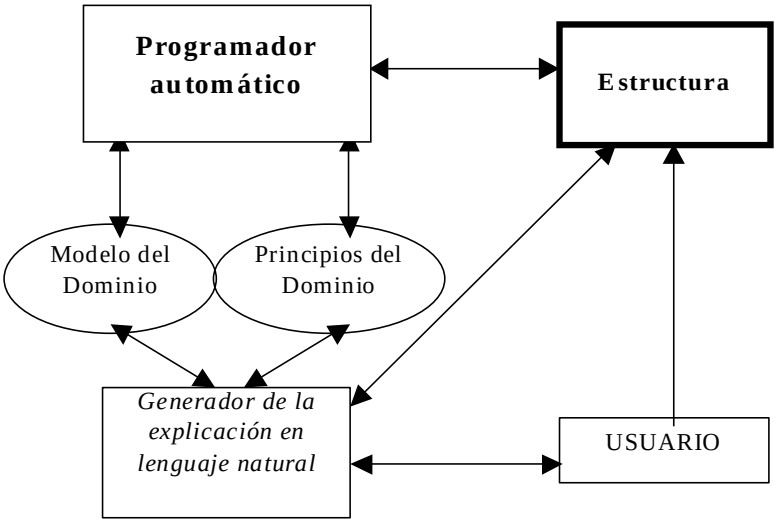

Figura 2.1: Elementos del sistema XPLAIN.

macro, la presentan mediante **texto** expresado en lenguaje natural y las probabilidades se expresan sólo de forma **lingüística**.

XPLAIN

Posteriormente Swartout [176], con su sistema XPLAIN, desarrolla un método consistente en la utilización de un programador automático para crear el sistema experto y para determinar el funcionamiento del mismo, lo que incluye también el mecanismo para la generación de explicaciones. El sistema XPLAIN se utilizó como sistema tutor en la administración de fármacos para la estabilización del ritmo cardíaco y constaba de varias partes, tal y como se muestra en la figura 2.1:

- El programador automático, encargado de generar el sistema.
- El modelo del dominio, que representa los hechos del conocimiento experto mediante una red **causal** (el *qué*).
- Los principios del dominio, que contiene los métodos y heurísticas del dominio (el *cómo*).
- La estructura para la generación de explicaciones.
- Un generador de explicaciones en lenguaje natural.

La novedad de este trabajo estriba en la utilización del programador automático para generar, durante el proceso de construcción del sistema experto, la información necesaria

para producir posteriormente las justificaciones de los resultados. Dicha información se representará en una estructura capaz de almacenar los mecanismos de razonamiento seguidos por el programador durante la construcción del sistema, en concreto, un árbol. Como el sistema se va generando mediante un proceso de refinamiento de objetivos generales a específicos, cada nodo del árbol corresponderá a cada uno de los objetivos; por tanto, un sub-árbol corresponde a una fase de refinamiento del objetivo que está representado en el nodo del nivel superior, es decir, en su nodo padre. Las hojas del árbol constituyen las operaciones básicas que el programa debe producir. Esta estructura servirá como entrada al módulo que se encarga de generar la explicación en lenguaje natural. Dicho generador de explicaciones consta de dos partes: una de bajo nivel, llamada *generador de frases*, que es la que construye las frases directamente de la base de conocimiento de XPLAIN; otra, llamada *generador de respuestas*, de alto nivel, que determina qué es lo que se debe decir; su misión es la de seleccionar aquellas partes de la representación del conocimiento que el generador de frases debe traducir a lenguaje natural. Además, otra tarea del generador de respuestas es la de seleccionar el nivel de detalle de la explicación, para lo que necesita conocer en cada momento cuál es el estado de la ejecución del programa, qué es lo que se ha dicho ya y qué es lo que el usuario necesita saber. El objetivo de este método es tanto la **comprensión del modelo como del razonamiento**. Las explicaciones se generan a **nivel macro**, presentándolas mediante **texto**. Como novedad frente a los métodos previamente descritos se encuentra la posibilidad de seleccionar el **nivel de detalle**.

ABEL

Este sistema, desarrollado por Patil et al. [138], utiliza también redes causales para representar el conocimiento médico, aunque distingue varios tipos de relación: causa-efecto, estar-asociado-a (no implica necesariamente causalidad) y agrupación. Las explicaciones generadas en este caso se basan en las propuestas por Swartout [176] (descrito previamente) y tienen como objetivo la **comprensión** tanto del **modelo** como del **razonamiento**. Se ofrecen a **nivel macro** y se presentan en **lenguaje natural** con cinco **niveles de detalle**, correspondientes a los distintos tipos de usuario para los que está pensado el sistema.

Toma de decisiones mediante razonamiento simbólico

En esta misma línea se enmarca el trabajo de Langlotz, Shortliffe y Fagan [101]. En él, y basándose en los principios de la teoría de la decisión, proponen una técnica de razonamiento simbólico para generar explicaciones **cualitativas** de los resultados del análisis de

las posibles decisiones. Al igual que los analistas generan explicaciones basadas en árboles de decisión, la idea es utilizar técnicas de representación del conocimiento para establecer una correspondencia entre los nodos del árbol de decisión y las situaciones médicas que representan. Dicho mecanismo de explicación está formado, fundamentalmente, por dos componentes:

- Representación del conocimiento mediante marcos. Cada nodo del árbol se representa mediante un marco, lo que permite definir sus características propias y las relaciones con otros nodos.

- Razonamiento matemático cualitativo, consistente a su vez en los cuatro pasos siguientes: cálculo de la utilidad esperada de cada decisión, comparación de las ramas para determinar qué conceptos son candidatos para aparecer en la explicación, selección de los candidatos que aparecerán en la explicación y de qué modo deben ser presentados y, por último, expresión en lenguaje natural utilizando las técnicas de traducción de sistemas de conocimiento basados en reglas.

La capacidad de decidir el modo más efectivo de organizar y presentar los conceptos se lleva a cabo por medio del análisis de una librería de plantillas y la selección de una de ellas. Las plantillas se representan mediante expresiones simbólicas que indican cómo se debe estructurar una explicación y de qué elementos debe constar. La traducción se realiza gracias a la asociación de un fragmento de texto en lenguaje natural con cada plantilla y con cada sentencia lógica que se utiliza para expresar situaciones o diferencias. En la tabla 2.2 se muestran algunos ejemplos de expresiones, junto con el texto asociado.

Expresión simbólica	*Texto asociado*
$P(X)$	"La probabilidad de X"
NEFROTOXICIDAD $\mid X$	"nefrotoxicidad debido a X"
$X > Y$	"X es mayor que Y"

Tabla 2.2: Ejemplos de expresiones simbólicas y textos asociados.

El objetivo de este método es la **descripción de la evidencia**. La explicación, que es **nivel macro**, se presenta mediante **texto** expresado en lenguaje natural. Como ya hemos indicado, las probabilidades se expresan sólo de forma **lingüística**.

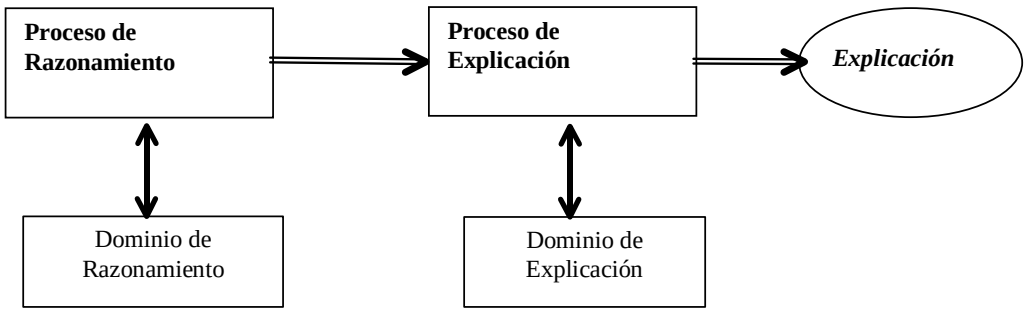

Figura 2.2: Proceso de explicación propuesto por Wick et al. [189].

Explicación reconstructiva

En la línea anterior encontramos las aportaciones de Wick y colaboradores [187, 189], quienes proponen una metodología de *explicación reconstructiva*, mediante la que se presentan los procesos de razonamiento y de explicación como tareas distintas e independientes, tanto en forma como en contenido. El argumento que utilizan es que cuando un humano proporciona una explicación de cómo ha resuelto un problema, normalmente no realiza una descripción de todos los pasos que le han llevado hasta la solución, sino que hay detalles que se omiten e, incluso, pueden aparecer otros que no han intervenido en su razonamiento. Por esta razón, consideran que las técnicas de generación de explicación deben tomar como entrada el proceso de razonamiento, posiblemente incompleto, que se haya llevado a cabo en el sistema para la solución de un determinado problema, y basarse en él para producir la explicación de las conclusiones. Para ello, se incluye un nuevo dominio que contiene todo el conocimiento necesario para la generación de dicha explicación (véase la figura 2.2).

Todas estas ideas las han plasmado Wick y Thompson en el prototipo REX [189], diseñado para proporcionar explicaciones de cómo un sistema experto obtiene los resultados que presenta. Dicho sistema está basado en el siguiente modelo para generar explicaciones reconstructivas: el sistema experto produce un conjunto de "pistas" del proceso de razonamiento que ha seguido; éstas pasan por un filtro que, basándose en las restricciones del problema, se encarga de determinar qué partes del razonamiento se pasarán al sistema generador de la explicación. A continuación, dicha información, filtrada y ajustada al problema concreto que se desee justificar, junto con las restricciones que debe verificar la solución, se proporcionan al módulo encargado de generar la explicación. Este módulo utiliza la información almacenada en el dominio de la explicación, para obtener la línea de la explicación que justifique las pistas de razonamiento proporcionadas; finalmente, la

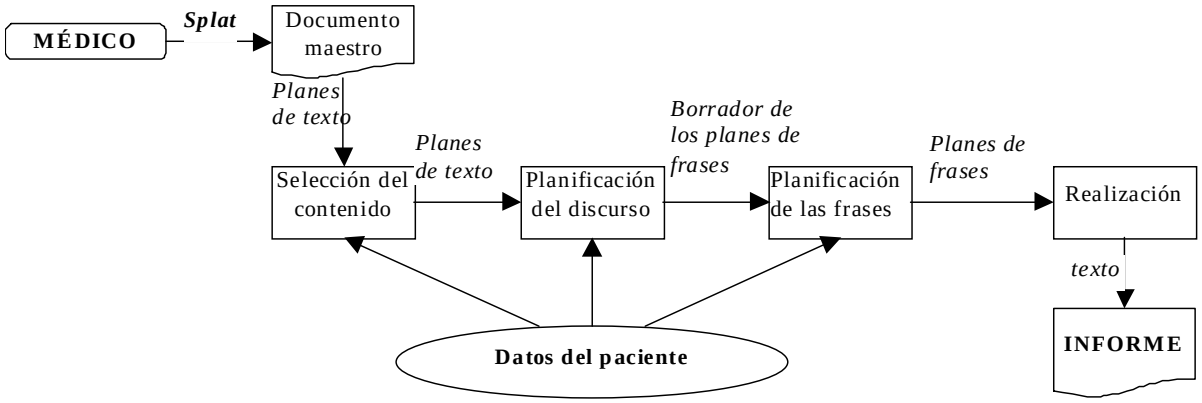

Figura 2.3: Proceso de generación de explicaciones en el sistema HEALTHDOC [110].

pasa a otro módulo encargado de traducirlo a lenguaje natural y de presentarlo al usuario. El objetivo de este método es la **comprensión del razonamiento**, para lo que se generan explicaciones a nivel **macro**, expresadas mediante **texto**.

HEALTHDOC

Di Marco y colaboradores [110] han desarrollado también un sistema para la generación de lenguaje natural en el ámbito de la medicina, llamado HEALTHDOC, cuyo objetivo es emitir informes médicos para los pacientes y materiales sobre educación para la salud, **adaptados** a las necesidades concretas de cada uno de ellos. El sistema posee las siguientes características:

- Los informes y materiales se presentan en forma de documentos escritos, por lo que no se permite la interacción entre el sistema y el paciente. Estos documentos se pueden enviar directamente a la impresora o ser enviados a un procesador de textos para que el médico lo edite y haga sobre él las modificaciones que estime oportunas.
- El proceso de creación de informes en lenguaje natural está basado en técnicas de *planificación de texto* [20, 122], tal y como indica la figura 2.3.
- La forma de adaptar el texto a las necesidades del paciente está determinada por el estado clínico del paciente y por sus características personales y culturales. Para ello, existe una base de datos con toda la información de los pacientes.
- Es posible generar el texto en varios idiomas.

La producción de cada informe concreto se realiza a partir de un documento maestro, específico del tema a tratar en dicho informe. Este documento contendrá toda la infor-

mación que el sistema tenga que incluir en él, incluidas las ilustraciones y las anotaciones sobre aquellos datos que sean considerados relevantes por el médico. En concreto, el documento maestro se representa mediante un conjunto de estructuras y anotaciones en SPL (*Sentence-Plan Language*), que se utiliza como paso intermedio para la generación de lenguaje natural. Además, cada frase se marca con relaciones de coherencia y referencia a otras frases. El documento maestro lo edita el médico en cuestión, mediante una herramienta proporcionada por el propio sistema, llamada *Splat*. El proceso de generación de un informe en lenguaje natural consta de varios pasos que, básicamente, son: a partir del documento maestro, se realiza una selección de su contenido, que es la que determinará los planes de texto que permitirán generar el consiguiente discurso en lenguaje natural. Estos planes, que son grupos de proposiciones, pasan a continuación a un módulo encargado de realizar la planificación del discurso, teniendo en cuenta las relaciones retóricas necesarias para la producción del texto. De esta fase se obtiene un conjunto de proposiciones estructuradas, que pasan al proceso encargado de realizar la planificación de las sentencias o frases; en concreto, se encarga de controlar el orden de las mismas, las relaciones entre ellas, el léxico, etc. La salida de este proceso consiste en una secuencia de especificaciones de frases, que son las que se pasan al módulo encargado de convertirlas a lenguaje natural. Como se puede apreciar, durante todo el proceso de planificación se tienen en cuenta los datos del paciente para así realizar la planificación en cada momento atendiendo a sus ca-rac-terísticas específicas. El objetivo de este método es la **comprensión del modelo**, para lo que se generan explicaciones a nivel **macro**, expresadas mediante **texto**. Como hemos indicado previamente, existe la posibilidad de **adaptar** cada explicación a las necesidades de los distintos usuarios.

OPADE

Por último, De Carolis y colaboradores [14] realizan un trabajo aplicado a la prescripción de fármacos en medicina, implementado en el sistema OPADE, en el que además de hacer un estudio previo de lo que debe contener una explicación, y teniendo en cuenta las **necesidades del usuario**, proponen un método para generar explicaciones basado en técnicas de *planificación de texto*. Así, clasifican a los usuarios en: *directos* e *indirectos*. El usuario directo es aquél que interacciona con el sistema, mientras que el indirecto es aquél que simplemente recibe un informe (posiblemente escrito) con la explicación de los resultados obtenidos. Por ejemplo, en el caso de la medicina, el usuario directo sería el médico y el indirecto el paciente, que es quien recibirá el informe con la explicación del diagnóstico médico. El método se propone exclusivamente para la generación de ex-

plicaciones al usuario indirecto, eso sí, teniendo en cuenta las características del usuario directo. Es decir, tratan de generar un informe tal y como lo haría el usuario directo que es quien lo "firmará". Para ello, se introduce como entrada la historia médica del paciente y el diagnóstico emitido y, a partir de aquí, se genera el texto de la explicación en **lenguaje natural** de dicha situación. Este proceso se lleva a cabo, fundamentalmente, gracias a las cuatro componentes de las que consta el sistema, que son las siguientes:

- *Modelo del usuario*, que contiene las características tanto del usuario directo como del indirecto. Del primero representa dos propiedades: su propensión a hablar y sus preferencias sobre los estilos de los mensajes. Del segundo, su grado de conocimiento y sus intereses sobre la información a recibir. Ambos modelos están diseñados de acuerdo con los resultados de algunos estudios psicológicos previos sobre las explicaciones que proporcionan los médicos y las necesidades de los pacientes a los que va dirigida dicha explicación.

- *Planificador de texto*, que es el que se encarga de definir la estructura del texto, en función de los tipos concretos de usuario, tanto directo como indirecto, según la clasificación citada anteriormente. También se encarga de determinar los conceptos que aparecerán en el texto de la explicación, su orden de presentación y las relaciones retóricas entre las partes del texto. Los criterios para la generación del texto se describen por medio de una librería de operadores de planificación, que son los que establecen cómo conseguir un objetivo cuando se verifican ciertas condiciones y qué efecto produce sobre el usuario. Un operador de planificación es una tupla de la forma (Encabezado, Restricciones, Intenciones, Precondiciones, Efectos, Descomposición, Relación_Retórica). La acción que realizan dichos operadores consiste en descomponer el objetivo en dos o más subobjetivos, o en una o más acciones que se incluirán en el texto del discurso. Gracias a estas descomposiciones se va obteniendo un plan de discurso en forma de árbol.

- *Módulo de resolución de conflictos*, que se encarga de encontrar un equilibrio entre el usuario directo y el indirecto cuando hay discrepancias.

- *Generador de texto*, basado en redes de transición aumentadas (ATN) anidadas en varios niveles. Dicho generador toma como entrada el plan de texto obtenido, en forma de árbol, y produce el texto asociado del modo siguiente: la red del primer nivel explora el árbol desde la raíz, examina el tipo de nodo y salta a los nodos de niveles inferiores; las redes de niveles intermedios generan términos lingüísticos de acuerdo

con el tipo de usuario (directo e indirecto) y las inferiores generan fragmentos de sentencias, de acuerdo con la acción asociada al nodo hoja. Estas frases se obtienen de la base de conocimiento.

Este proceso se puede ver resumido en la figura 2.4.

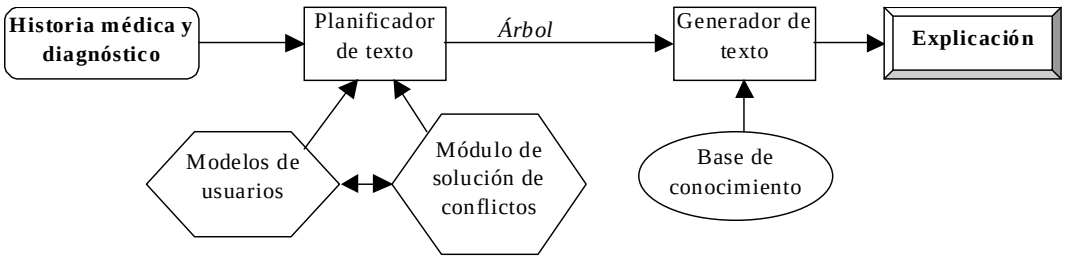

Figura 2.4: Proceso de explicación del sistema OPADE.

Por tanto, el objetivo de este método es la **comprensión del modelo a nivel macro** mediante la expresión **textual** de las explicaciones generadas. Como hemos indicado, es sólo el usuario directo el que interacciona con el sistema para producir el documento maestro, base de la explicación generada, adaptando su **nivel de detalle** según los distintos tipos de usuarios indirectos.

2.3.2. Métodos intercomunicativos

Este tipo de métodos surge frente a las limitaciones de las herramientas clásicas de explicación a la hora de realizar explicaciones complejas, debido sobre todo a la falta de comunicación bidireccional entre el sistema y el usuario. Uno de los objetivos que se persiguen al explicar algo es satisfacer las necesidades del usuario de cara a la comprensión del funcionamiento del sistema. Esto implica que, en determinadas situaciones, el usuario necesitará interrumpir el proceso de explicación para formular las preguntas que considere para completar su conocimiento o, por el contrario, para indicarle al sistema que no hace falta que profundice en un determinado tema que ya conoce.

Por tanto, la principal meta de estos métodos es llevar a cabo dicha comunicación, de tal modo que el sistema sea capaz de adaptarse a las nuevas necesidades y/o expectativas del usuario, a medida que varían durante el proceso de generación de la explicación. Por ejemplo, es posible que un usuario, al cabo de un cierto tiempo, pase progresivamente de inexperto a experto, por lo que el sistema debe ser capaz de actualizar dicho nivel de conocimiento.

Para conseguir la adaptación mutua entre el usuario y el sistema y que ambos puedan actualizar su conocimiento recíproco de forma dinámica, se debe permitir que el usuario solicite aclaraciones concretas sobre los datos de la explicación en los que necesite profundizar, realizando interrupciones cuando sea necesario; y a su vez se debe permitir al sistema revisar el nivel del usuario.

La idea sobre la que se fundamentan este tipo de herramientas es la de considerar la tarea de la generación de explicaciones como un problema independiente del tipo de sistema experto en sí pero, sin embargo, íntimamente relacionado con el problema de la transmisión de información de cierta complejidad de la forma más efectiva posible.

Explicación reactiva

Una de las investigadoras que más ha aportado al tema es J. Moore. En los trabajos realizados tanto con Swartout [123] como con otros autores [13, 122], plantea la explicación como una forma de conversación con el usuario del sistema y propone una técnica de explicación reactiva. Esto implica que el sistema tiene que ser "consciente" de todo lo que se haya dicho ya y reaccionar a cualquier pregunta del usuario de forma inteligente, identificando cuáles son sus objetivos. Para ello, estos autores proponen un método que se basa en planificar el texto que contendrá la explicación y en ir *memorizando* la explicación que se ha presentado ya, para basarse en ella cuando sea necesario. Las técnicas de planificación del texto permiten seleccionar y estructurar dinámicamente la información que se va a presentar en cada caso.

El proceso, en líneas generales, tiene lugar como sigue: cuando el usuario realiza una pregunta (o introduce un dato, quizás solicitado por el sistema), el analizador de preguntas genera un objetivo, que representa la abstracción de la respuesta que se desea producir en el discurso de la explicación. Este objetivo se pasa al planificador de texto, que busca entre las estrategias de explicación almacenadas en el sistema, normalmente en forma de reglas, aquéllas que verifican el objetivo deseado. La selección de una concreta entre todas ellas se lleva a cabo mediante técnicas de búsqueda heurística, utilizando además el conocimiento del sistema sobre el modelo del usuario, el diálogo mantenido hasta ese momento, etc. Una vez que se tiene el plan del texto de la explicación, representado en forma de árbol, se almacena en la historia del diálogo y se pasa a la interfaz gramatical para que sea transformado, mediante un traductor de lenguaje natural, en el texto de la explicación. Después de presentar dicho texto al usuario, el sistema espera a que éste le responda a dicha explicación para determinar si la ha entendido, si desea concretar algo más, etc. El sistema procesa esta información de modo similar al caso anterior, y se

repite el proceso hasta que el usuario queda satisfecho con la explicación presentada por el sistema.

En resumen, el propósito de este método es la **comprensión** tanto del **modelo** como del **razonamiento**. La explicación se genera a **nivel macro**, mediante un **diálogo** entre el usuario y el sistema expresado textualmente en **lenguaje natural**. Además es posible tener un **modelo dinámico de usuario**, en el que se ajusta en cada momento de forma **automática** el nivel de detalle de la explicación.

Planificación de diálogos

En una línea de investigación similar están los estudios de Cawsey [17, 18, 19, 20, 21], relacionados con la generación de explicaciones interactivas. El método que ella propone se basa, además de en la planificación del texto de las explicaciones, en la planificación del diálogo con el usuario. La explicación se genera mediante la combinación de las reglas que permiten planificar el texto con las de la planificación del diálogo. Dicho método lo han implantado en el sistema EDGE, desarrollado para la generación de tutoriales sobre circuitos electrónicos. Entre las capacidades que ofrece este sistema están: permitir al usuario que interrumpa el diálogo en cualquier momento y obligar al sistema, durante la generación de la explicación, a actualizar tanto el modelo del usuario como el contenido del resto de la explicación. En este sistema, el proceso de planificación del diálogo se lleva a cabo de modo incremental. Los objetivos de la explicación se ponen en una agenda. De ella, se selecciona el objetivo de mayor prioridad. A continuación, se selecciona una regla de planificación, la cual se usa para encontrar nuevos subobjetivos del objetivo inicial, que se colocan a su vez en la agenda. Así, se va creando una estructura jerárquica que servirá para representar el contexto en el que se debe desarrollar la explicación.

Por otra parte, las interrupciones del usuario pueden tener como resultado la inclusión de nuevos objetivos en la agenda y, además, pueden proporcionar información al sistema sobre el conocimiento del usuario para ser actualizado, lo que influirá en el plan del resto de la explicación. Además, después de cualquier intercambio con el usuario, el contexto del discurso se debe actualizar, para que la explicación se genere de acuerdo con las nuevas suposiciones realizadas.

Por tanto, el objetivo de este método es la **comprensión** tanto del **modelo** como del **razonamiento**. La explicación se genera a **nivel macro**, mediante un **diálogo** entre el usuario y el sistema expresado textualmente en **lenguaje natural**. Además es posible tener un **modelo dinámico de usuario**, en el que se ajusta en cada momento de forma **automática** el nivel de detalle de la explicación.

Sistema *P.rex* También se basa en la planificación de diálogos el sistema *P.rex* [68, 69], un sistema para la explicación de demostraciones de teoremas que es capaz de adaptarse a las distintas expectativas de los diferentes usuarios, al que se le permite además interaccionar con el sistema en cualquier momento. El objetivo es la **comprensión** tanto del **modelo** (teoremas), como del **razonamiento** (demostraciones). La explicación se genera a **nivel macro y micro**, mediante un diálogo con el usuario expresado en **lenguaje natural**, en el que se le va preguntando si ha entendido los pasos que ha dado hasta el momento el sistema. De este modo, se va adaptando la explicación al nivel de conocimiento del usuario, que se elige inicialmente de una base de datos para representar los distintos **modelos de usuario**. Un ejemplo de las explicaciones generadas las podemos ver en la figura 2.5.

Theorem 1 $((P \wedge Q) \supset (Q \wedge P))$.

Proof:

Let $(P \wedge Q)$. That implies that $(Q \wedge P)$ by $\wedge Comm$.
This step is too difficult.
$(Q \wedge P)$ since $(P \wedge Q)$ by $\wedge Comm$.
Do you understand this step? No.
Do you understand the premises? Yes.
Then Q by $\wedge E_r$. P by $\wedge E_l$. That implies that $(Q \wedge P)$ by $\wedge I$.
Do you understand this step? Yes.
Then $((P \wedge Q) \supset (Q \wedge P))$ by $\supset I$. ■

Figura 2.5: Ejemplo de explicación del sistema *P.rex* sacado de [68].

2.4. Métodos de explicación en redes bayesianas

En esta sección hacemos una revisión de los métodos desarrollados para generar explicaciones en redes bayesianas. Dicha revisión la haremos mediante la clasificación de los métodos en tres grupos, atendiendo al objeto de la explicación (véase la sección 2.2.1): explicación de la evidencia, explicación del modelo y explicación del razonamiento.

2.4.1. Explicación de la evidencia: abducción

En este caso, los métodos desarrollados se centran en la explicación de la evidencia. Como ya comentamos en la sección 2.2.1, en este contexto una explicación $\mathbf{w}$ es una asignación de valores a todas las variables de un subconjunto $\mathbf{W}$ de las variables de la red. Ya que los valores de las variables observadas se suelen conocer con certeza, sólo las

variables no observadas son objeto de interés en los métodos de abducción. Por tanto, el objetivo de la abducción es encontrar la explicación más probable (EMP), es decir, la configuración $\mathbf{w}$ que tiene la mayor probabilidad a posteriori $P(\mathbf{w}|\mathbf{e})$, donde $\mathbf{e}$ es la evidencia disponible. Algunos métodos son capaces de encontrar las k explicaciones más probables. Cuando $\mathbf{W}$ incluye todas las variables no observadas, el proceso se conoce como *abducción total*; en otro caso, se denomina *abducción parcial*.

Los métodos de abducción desarrollados hasta la fecha se han limitado a buscar las EMPs, sin intentar justificar por qué son más probables que otras; es decir, el propósito de estos métodos es la **descripción**, no la comprensión. La distinción entre los **niveles** micro y macro no tiene sentido en este caso porque no hay ningún análisis del proceso de razonamiento. La interpretación de una configuración (asignación de valores) como una explicación presupone implícitamente que existe un modelo **causal**, por lo que las variables que toman el valor "presente" en la EMP son las que *explican* las anomalías observadas. Sin embargo, ya que estos métodos son puramente matemáticos, pueden ser aplicados a cualquier red, independientemente de si es causal o no.

Con respecto a la comunicación usuario-sistema, la investigación sobre abducción en redes bayesianas normalmente se limita a encontrar las EMPs, como hemos dicho, sin prestar atención a la interacción y adaptación al usuario. Por esta razón, la mayoría de los criterios para analizar los métodos de explicación (sección. 2.2) no se aplican a la abducción. Como mucho, podemos mencionar que las probabilidades de cada explicación ofrecida se muestran **numéricamente**; no se ha hecho ningún esfuerzo por expresar las probabilidades de forma cualitativa.

A continuación exponemos en orden cronológico los trabajos más relevantes sobre abducción. Hay que señalar que todos los métodos están diseñados solamente para variables discretas.

Propagación π-λ de Pearl

El primer trabajo sobre explicación en redes bayesianas [141, cap. 5] fue un método de abducción total basado en la propiedad de que, para cada valor x de una variable $X \in \mathbf{W}$, existe la mejor explicación para el resto de variables $\mathbf{W} \setminus X$ (propiedad similar al principio de optimalidad de la programación dinámica). Por tanto, el valor de cada variable X en la EMP se puede obtener encontrando la mejor explicación para $\mathbf{W} \setminus X$ y eligiendo el mejor valor de X. Como las computaciones son locales para cada nodo, el proceso se puede diseñar como un intercambio de mensajes π-λ entre los nodos. El algoritmo tiene complejidad lineal para el caso de poliárboles y exponencial para redes con bucles. Sin

embargo, la búsqueda de las EMP puede ser muy complicada incluso para poliárboles, pues hay que tener en cuenta todas las variables, incluso aquéllas que no tienen ningún interés para la hipótesis. Una limitación de este método es que sólo puede encontrar las dos explicaciones más probables (k=2) [129]. En consecuencia, este método es sólo adecuado cuando las dos EMP son mucho más probables que el resto de las explicaciones.

No obstante, hay autores [23] que consideran básico el método del paso de mensajes para la generación de explicaciones, visto éste como una solución reconstructiva, con las posibilidades que ofrece de cara a incorporar además las características y expectativas del usuario.

Sistemas de restricciones lineales

Santos [154] propuso un método de abducción total que transforma una red bayesiana en un sistema de restricciones lineales equivalente. Un *sistema de restricciones lineales* $L(W)$ es una tupla de la forma (Γ, I, ψ) donde Γ es un conjunto de variables, I es un conjunto finito de desigualdades definido sobre las variables de Γ, y ψ es una función de $\Gamma \times \{\text{verdadero, falso}\}$ en $\mathbb{R}$. Las restricciones garantizan que cada variable tome un solo valor y que la probabilidad de una configuración de un conjunto de variables se calcule mediante el correspondiente conjunto de probabilidades condicionadas. La construcción del sistema se realiza en tiempo lineal respecto al número de variables de la red bayesiana. La búsqueda de todas las soluciones implica la resolución de una secuencia de sistemas de restricciones en el que cada uno se obtiene del anterior. Uno de los problemas de este método es que, como el de Pearl, asigna valores a todas las variables, incluidas aquéllas que no interesan para obtener el valor de la hipótesis.

Irrelevancia en abducción parcial

Shimony [164] y Suermondt [174] han estudiado el problema de la obtención de explicaciones que contuvieran sólo variables relevantes para la hipótesis. En el caso de la abducción parcial, Shimony propuso tres definiciones de *irrelevancia*:

- *Independencia estadística*: una variable X no debería formar parte de una explicación $\mathbf{w}$ si no afecta a la probabilidad de la evidencia: $P(\mathbf{e}|\mathbf{w}, x) = P(\mathbf{e}|\mathbf{w})$; se dice entonces que X es irrelevante;

- *independencia* δ, es menos restrictiva que la anterior e indica que un hecho es irrelevante si los hechos dados son independientes con una tolerancia de δ:

$$|P(\mathbf{e}|\mathbf{w}, x) - P(\mathbf{e}|\mathbf{w})| \leq \delta,$$

- y *cuasi-independencia*, que se basa en que el valor de una variable no interesa si no afecta "lo suficiente" a la probabilidad de la explicación.

La idea que subyace en el algoritmo para encontrar las explicaciones relevantes es la de que los antecesores no observados de aquellos nodos V con un valor conocido deben permanecer como no observados si no afectan a V, es decir, si no modifican suficientemente su probabilidad (y, por tanto, no se pueden utilizar para explicar V). Además, considera que todos los nodos que no son antecesores de V no se deben utilizar para explicarlo, pues no son posibles causas suyas. Observamos otra vez la suposición de un modelo **causal**.

Grafos de funciones booleanas valoradas

Charniak y Shimony [22] probaron que el problema de la abducción en redes bayesianas es equivalente a encontrar la asignación de mínimo coste para las variables de un determinado *grafo dirigido acíclico valorado por una función lógica (WBFDAG)* [2] que se obtiene mediante la transformación de la red bayesiana junto con la evidencia $\mathbf{e}$. El método para la construcción del consiguiente grafo valorado consiste, fundamentalmente, en crear nuevos nodos, similares a los de la red original pero con una función de coste asociada que define, para cada uno de ellos, un valor que depende de los valores de sus padres y que mide el coste de asignar un valor determinado a dicho nodo. El WBFDAG se resuelve mediante un algoritmo de búsqueda en calidad (*best-first*) que, aplicado sucesivamente, produce la enumeración de las asignaciones en orden decreciente según su coste, es decir, las explicaciones en orden decreciente de probabilidad. Una desventaja de este método es su complejidad tanto computacional como conceptual.

Abducción parcial aproximada

Gámez [71] ha realizado un estudio detallado sobre la abducción, tanto total como parcial, y sobre los algoritmos desarrollados hasta ahora para hacer inferencia abductiva en redes bayesianas con el objetivo de encontrar las explicaciones más probables para la evidencia observada. Entre sus contribuciones más relevantes podemos citar:

- Adaptación de los algoritmos de agrupamiento (propagación en árboles de cliques) a la abducción parcial.

- Nuevos algoritmos aproximados basados en *algoritmos genéticos* y *enfriamiento estocástico* (simulated annealing) [32] que se pueden utilizar en aquellos casos en los

[2] Del inglés **W**eighted **B**oolean **F**unction **D**irected **A**cyclic **G**raph.

que los métodos exactos son ineficientes.

- Definición de algunos criterios para simplificar las explicaciones [33]:

  Valor usual: la simplificación se basa en mostrar al usuario sólo aquellos valores que no son los *usuales*. Por ejemplo, en medicina, el valor usual de la variable Meningitis es ausente.

  Independencia, mediante el análisis del grafo que representa la red bayesiana. Ha desarrollado un algoritmo para detectar algunos tipos de independencias.

  Relevancia, lo que implica que la información que se omite es "casi" irrelevante (en el sentido de independencia estadística) para los hechos observados.

- Demostración de que aunque las simplificaciones basadas en normalidad son muy rápidas, los criterios de relevancia e independencia producen mejores resultados en aquellos casos en los que no es fácil decidir cuál es el valor *usual* de una variable, lo que ocurre cuando los estados de las variables tienen probabilidades parecidas.

2.4.2. Explicación del modelo

En la sección 2.2.1 mencionamos las ventajas que la explicación del modelo, conocida también como *explicación estática*, ofrece para construcción de la red bayesiana y para propósitos educativos. Ahora estudiaremos los métodos desarrollados en esta línea, clasificados según la forma de presentar las explicaciones.

Presentación gráfica de la red

La forma más directa e intuitiva de mostrar la información contenida en una red bayesiana es mostrar el grafo que la representa: cada nodo se muestra como un círculo o como un óvalo que contiene el nombre (o una breve descripción) de la variable asociada, y los enlaces se dibujan como flechas. Algunas de las primeras herramientas *software* para la edición y visualización de redes bayesianas fueron Analytica [3], IDEAL [169], DAVID [157], HUGIN [2] y PATHFINDER [82].

Sin embargo, cuando el tamaño de la red es muy grande, resulta muy difícil visualizar el grafo. Por esta razón, Analytica y otros paquetes, como GeNIe [53], ofrecen la posibilidad de definir submodelos. Cuando un modelo está contraído, se representa por un tipo especial de nodo; cuando se expande, muestra todos los nodos y enlaces que contiene.

[3] `http://www.lumina.com`

En cualquier caso, este tipo de explicación constituye solamente una **descripción** del modelo y no pretende mejorar la comprensión del usuario. La interacción con el usuario se realiza a través de **menús**, que dan acceso a las propiedades de los nodos, enlaces y tablas de probabilidad condicionada, en las que los parámetros del modelo aparecen expresados de forma **numérica**. En las herramientas que hemos examinado, no existe la posibilidad de **adaptación** al usuario, excepto en opciones menores de formato o de edición.

MUNIN

El uso de redes causales en otros sistemas médicos (véase la sección 2.3) junto con los estudios que justifican que el conocimiento médico se puede estructurar de forma causal, motivaron la representación del conocimiento del sistema experto MUNIN [103] mediante una red causal. Este sistema fue diseñado mediante HUGIN en la universidad de Aalborg (Dinamarca) para el diagnóstico de problemas relacionados con los músculos y los nervios. Aunque los principales esfuerzos en este caso se dedicaron a proponer algoritmos "eficientes" de propagación de la evidencia en este tipo de redes, sin embargo sólo se proponen ciertas capacidades de explicación que no se implementan en el sistema, puesto que la única facilidad de explicación con la que contaba MUNIN consistía en explotar las capacidades intrínsecas que proporciona el grafo que representa la red causal, ya que permite visualizar las probabilidades de cada nodo e interpretar el conocimiento de forma más o menos intuitiva. La presentación de las explicaciones se realizaba de forma **gráfica**, expresando las probabilidades también de forma **numérica**.

Descripción verbal de la red

Druzdzel y Henrion [50, 85] propusieron un método para traducir la información cualitativa y cuantitativa de una red bayesiana a expresiones lingüísticas mediante plantillas que indican la probabilidad a priori de un nodo, como: "Estar constipado es muy poco probable (p=0.08)", y permite comparar probabilidades: "Estar constipado es algo menos probable que tener un gato (0.08/0.10)".

En este método, las **relaciones causales** juegan un papel importante. En particular, existen patrones para describir la puerta OR [141, 43], un modelo que supone la independencia de interacciones causales: "Estar constipado es una causa muy frecuente (0.9) de estornudar. Tener alergia es una causa muy frecuente (0.9) de estornudar. Estar constipado no afecta al hecho de tener alergia, y viceversa. Existen otras causas poco probables (p=0.1) de estornudar". Las relaciones de dependencia o independencia condicional pueden ser expresadas por frases como la siguiente: "Conocido el valor de estornudar, estar constipado y

tener alergia son independientes".

El objetivo de este tipo de explicación está entre la **descripción** y la **comprensión**. Las explicaciones se ofrecen al **nivel micro**. La **interacción** usuario-sistema no está descrita en las referencias. La presentación de la explicación es **textual** y contiene tanto expresiones **lingüísticas** como **numéricas** de la probabilidad.

Otros sistemas, como B2 (véase sección 2.4.3) y Elvira (capítulo 5) también ofrecen explicaciones verbales del modelo.

Navegación mediante menús

Díez [45], en su sistema experto DIAVAL, desarrolló un método para explicar tanto el modelo como el razonamiento. Dicho método distingue entre varios tipos de enlaces y nodos, correspondientes a los diferentes tipos de influencia **causal**: por ejemplo, en la puerta OR los padres de un nodo se consideran como *causas* en el sentido estricto, mientras que en el modelo general son considerados como *factores* que influyen sobre la variable hija.

La interacción usuario-sistema está basada en un sistema de ventanas y **menús**, que ofrecen diferentes opciones:

- Para cada *nodo*, el usuario puede ver su definición, prevalencia (probabilidiad a priori), probabilidad a posteriori, lista de causas o factores, tabla de probabilidad condicional (para nodos con padres) y lista de efectos (hijos).

- Para cada *parámetro*, que es un nodo con una medida numérica asociada (generalmente de ecocardiografía), el usuario puede ver el valor medido, intervalos, significado patofisiológico y fórmula (para aquellos parámetros calculados a partir de otros).

- Para los *enlaces* que forman parte de una puerta OR, es decir, los enlaces causales en el sentido estricto, las opciones son: causa, efecto, sensibilidad y especificidad (sólo para variables binarias), eficiencia (probabilidad de que la causa produzca el efecto) y un texto que explica el mecanismo causal asociado.

Mediante la visita de los padres (causas) y los hijos (efectos) de los nodos, el usuario puede navegar a través de la red. La explicación se sitúa dentro del **nivel micro**, porque se centra en cada momento en un nodo o en un enlace. Las explicaciones se muestran mediante **texto** dentro de diferentes ventanas y la probabilidad se expresa sólo **numéricamente**.

MEDICUS

Por último, cabe destacar las aportaciones del sistema MEDICUS, diseñado por Folckers y colaboradores [70, 161] y desarrollado como entorno para la construcción de modelos explicativos y de diagnóstico en los que el conocimiento es complejo e incierto, como en el caso de la medicina. La principal novedad de este proyecto es que se trata de un entorno en el que el usuario, independientemente de su **nivel de conocimientos** sobre probabilidad, construye una red bayesiana que modela el dominio correspondiente. Esto es posible gracias a las explicaciones, a **nivel micro** que genera un editor lingüístico del modelo, diseñado junto con un editor gráfico de la red. Tanto las explicaciones del modelo como del diagnóstico ofrecido se ofrecen de forma **cuantitativa** y **cualitativa** y se presentan tanto de forma **textual** como **gráfica**.

La falacia del jurado

La representación gráfica del modelo representado por una red bayesiana causal se ha utilizado para explicar dentro del ámbito jurídico, lo que sus autores [67] denominan *la falacia del observación del jurado*. La idea que subyace es que los miembros del jurado pierden parte de su credibilidad inicial sobre un veredicto de no culpabilidad en un delito después de que se les proporcione evidencia de una condena previa del acusado por un delito similar. Para llevar a cabo su experimento diseñaron una red bayesiana causal para representar el conocimiento implícito en sus razonamientos, y utilizaron HUGIN como herramienta para la presentación gráfica del modelo.

2.4.3. Explicación del razonamiento

En esta sección describimos en orden cronológico los métodos de explicación del razonamiento en redes bayesianas. Se caracterizan porque para generar las explicaciones se centran, en cada momento, en una sola variable, normalmente llamada *hipótesis*. También se presenta lo más relevante sobre explicación relacionado con la teoría de Dempster-Shafer por tener en común con las redes bayesianas ciertos aspectos probabilísticos, el tener un grafo y el realizar una propagación puramente numérica sobre él.

NESTOR

NESTOR, desarrollado por Greg Cooper [27], es una de las primeras redes bayesianas. Incluía una facilidad de explicación que ofrecía dos posibilidades:

- *comparar* cómo dos diagnósticos determinados justifican la evidencia, mostrando cómo cada hallazgo afectaba a la probabilidad relativa de los dos diagnósticos seleccionados por el usuario; y

- *criticar* una hipótesis con respecto al resto de posibles diagnósticos, mediante la visualización de la probabilidad relativa de una hipótesis con respecto a todas las demás.

En ambos casos, se generaba una explicación **verbal** que permitía describir **cualitativamente** los caminos entre la evidencia y una hipótesis dada. Básicamente, consistía en traducir al inglés los enlaces causales de cada camino. El método de explicación se propuso únicamente para redes **causales**. Además de las explicaciones verbales, también se presentaban de modo **gráfico**. Las probabilidades se expresaban **numéricamente**. La interacción con el usuario es a base de **preguntas predefinidas**, no tiene ningún modelo de usuario ni capacidad de adaptación.

GLADYS

También el sistema GLADYS, para el diagnóstico de la dispepsia y desarrollado por Spiegelhalter y Knill-Jones [167], contaba con capacidad de explicación basada en asignar a cada hallazgo un *peso de evidencia*:

$$\log\left(\frac{P(v_i|D_1,e)}{P(v_i|D_2,e)}\right) \tag{2.1}$$

Los pesos de evidencia habían sido propuestos por Good [74] y fueron utilizados también en otros sistemas [82, 108]. La explicación consiste en mostrar las probabilidades de las posibles enfermedades. Además, para las enfermedades plausibles, se realiza un "balance de evidencia", consistente en mostrar los hallazgos que contribuyen a padecer dicha enfermedad y los que no, junto con la "cantidad" de influencia que aporta cada uno. Así, el usuario puede identificar también los conflictos en los hallazgos y contrastar los síntomas y signos del paciente en cuestión. El objetivo de este método de explicación es, fundamentalmente, la **comprensión** del **razonamiento**. Las explicaciones se presentan de forma **textual** y los datos probabilísticos se presentan de forma **numérica**. No existe ningún tipo de interacción con el usuario ni tampoco hay posibilidad de ofrecer distintos niveles de detalle.

Clasificación bayesiana

En ocasiones, en los sistemas en los que además de emitir los posibles diagnósticos, tienen que ofrecer una decisión, no basta con dar una explicación del diagnóstico obtenido, sino que hay que considerar otras propuestas alternativas que, aunque no corresponden a la justificación de la conclusión, son igualmente útiles. Por ejemplo en medicina, además de emitir un diagnóstico hay que decidir el tratamiento a seguir. Entre los métodos de explicación desarrollados para este tipo de sistemas, destacaremos el de Reggia y Perricone [148], quienes proponen un método para la justificación de anomalías representadas mediante modelos basados en clasificación bayesiana.

El proceso consiste en realizar una ordenación de todas las posibles enfermedades que el sistema puede diagnosticar y después ofrecer la justificación de la posición relativa de una enfermedad concreta en la clasificación obtenida. Dicha justificación se realiza teniendo en cuenta una medida asociada a cada hallazgo relacionado con la enfermedad en cuestión. Esta medida se obtiene a partir de la comparación de las probabilidades a posteriori de dicho síntoma, suponiendo que se ha producido cada enfermedad.

El método propuesto es **interactivo** y se inicia pidiendo al usuario que introduzca la evidencia, para después mostrarle las cinco enfermedades más probables, junto con la probabilidad de cada una. Si el usuario lo desea, puede pedir la justificación de una de dichas enfermedades en relación con las demás; en ese caso, se muestran las probabilidades de cada una y las causas que han determinado la obtención de dicha probabilidad. Para la explicación de por qué una enfermedad D está en una posición determinada en relación con las demás enfermedades hay que detectar tanto las causas que hacen que D tenga una probabilidad menor que las que hay por encima de ella como las que hacen que D tenga una probabilidad mayor que las que hay por debajo de ella en la clasificación.

El objetivo de este método es la **descripción** del modelo y del razonamiento. La explicación se presenta a **nivel micro** mediante **texto** expresado en lenguaje natural, y para ello almacenan, junto con cada nodo y arco del grafo, el texto que representa el estado o el argumento para el cambio de estado, respectivamente. La interacción con el usuario se basa en **menús** en los que el usuario introduce las respuestas a las preguntas del sistema. Las probabilidades se expresan sólo de forma **numérica**. Los autores han aplicado su método al caso de los ataques neurológicos debidos a la ausencia de flujo sanguíneo o de hemorragias cerebrales. Sin embargo, una de las grandes limitaciones del método es que si el número de anomalías y síntomas es grande, las explicaciones generadas pueden llegar a ser muy confusas por la gran cantidad de datos presentados. Otra limitación del método estriba en que la hipótesis que tienen que verificar los modelos para poder aplicar este

algoritmo es demasiado restrictiva: todos los síntomas deben tener causas definidas.

Teoría de Dempster-Shafer

Entre los estudios sobre explicación relacionados con la teoría de Dempster-Shafer, destaca el de Strat y Lowrance [170, 171]. La idea principal que subyace en el método que proponen para la generación de las explicaciones es la del *análisis de sensibilidad*, idea que han utilizado también otros autores en otros tipos de sistemas [46, 47, 67, 79, 108, 174]. Este tipo de análisis consiste en realizar distintas variaciones en la evidencia para analizar la modificación de los resultados en las posibles soluciones. El proceso concreto que desarrollan Strat y Lowrance consta de varios pasos:

1. Análisis del efecto de la evidencia sobre la hipótesis a determinar. Este proceso genera un grafo en el que cada nodo define un estado, que representa una opinión, y cada arco indica el mecanismo por el que se cambia de un estado a otro.

2. Representación del flujo de información, mediante el grafo, desde los hallazgos hasta la conclusión, lo cual permitirá generar la explicación. Para ellos una explicación de las conclusiones consiste en un subconjunto de hallazgos que forman la evidencia. Para obtenerlos, se realiza un análisis de sensibilidad que consiste en eliminar uno o más hallazgos de la evidencia y actualizar el valor de la credibilidad de la hipótesis mediante diversas medidas de especificidad; éstas serán las que determinen la pertenencia de uno o más hallazgos a la explicación. Ya que en estos trabajos se contemplan distintos tipos de preguntas que puede realizar el usuario, las medidas de especificidad serán también distintas dependiendo de la clase de pregunta.

El objetivo de este método es la **descripción de la evidencia y del razonamiento**. La explicación la presentan a **nivel micro** mediante **texto** expresado en lenguaje natural y la interacción con el usuario se basa en **preguntas predefinidas**. Las probabilidades se expresan sólo de forma **numérica**.

MUNIN

Otras propuestas de este sistema experto para explicar el razonamiento consistían en mostrar los flujos de la evidencia, identificar los hallazgos que influyen sobre la hipótesis y en generar texto basado en ambos resultados. La explicación, generada de forma **textual** y **gráfica** no permitía ningún tipo de interacción con el usuario y las probabilidades se expresaban tanto **numérica** como **gráficamente**.

PATHFINDER

Otro de los primeros sistemas que incluyó también una herramienta de explicación parecida a la de GLADYS fue PATHFINDER [82]. El método consiste básicamente en comparar dos enfermedades, o dos grupos de enfermedades, mostrando cómo cada valor v_i de una variable determinada V afecta a su distribución de probabilidad. Para ello, para cada valor v_i de la variable seleccionada el sistema dibuja una barra proporcional a la razón de probabilidades de una enfermedad D_1 respecto a la otra D_2 dada la evidencia e. La medida utilizada es también el *peso de evidencia* [74].

Además, otra facilidad de explicación proporcionada por PATHFINDER es la de mostrar cómo la evidencia afecta a todas las posibles enfermedades. Con este objetivo, se muestra al usuario una lista ordenada de todas las posibles enfermedades junto con sus probabilidades asociadas, lo que define el orden de la lista.

El principal objetivo de este método es la **descripción** del razonamiento y puede ser aplicado tanto a redes **causales** como **no causales**. El primer tipo de explicación descrito se presenta de modo **gráfico** y el segundo en forma de lista en la que las probabilidades se expresan mediante **números**. La **interacción** con el usuario se lleva a cabo por medio de menús, ventanas y cajas de diálogo. No existe ningún tipo de **adaptación** al usuario, por lo que éste deberá tener cierto conocimiento sobre teoría de la probabilidad.

Explicación de las variaciones locales

En un poliárbol (una red sin bucles), la probabilidad a posteriori de una variable B se puede calcular mediante [141]

$$Bel(b) = \alpha\lambda(b)\pi(b) \tag{2.2}$$

donde $\lambda(b)$ representa la verosimilitud de cada valor b dados los efectos (hijos) de B observados y $\pi(b)$ es el soporte causal, es decir, la probabilidad de cada valor b dada la evidencia relativa a los antecesores de B en el grafo. Sember y Zukerman [156] desarrollaron un método de explicación para justificar el valor de $Bel(b)$ en términos de $\pi(b)$ y $\lambda(b)$. Las expectativas del usuario quedan caracterizadas por los cambios producidos en $\pi(b)$, ya que representa la información relativa a sus causas. Si $\pi(b)$ aumenta, disminuye o no varía, el usuario esperará que $Bel(b)$ aumente, disminuya o no varíe, respectivamente. Por tanto, el algoritmo para la generación de una explicación identifica una expectativa y, si ésta se alcanza, se genera la explicación; si no, se genera la explicación basada en identificar qué valores de $\pi(b)$ y $\lambda(b)$ han causado la desviación. Cuando la expectativa se

alcanza, el método genera la siguiente explicación: "La creencia en B ha aumentado debido al incremento del impacto de la evidencia correspondiente a las causas/efectos de B".

El objetivo de este método es la **comprensión**; se clasifica dentro del **nivel micro** (en cada momento sólo se analiza una variable, llamada *hipótesis focal*) y supone una red bayesiana **causal**. Las explicaciones se presentan en forma de **texto** y las variaciones de la probabilidad se expresan **lingüísticamente** ("ha aumentado/ha disminuido"). Sember y Zukerman no describen la **interacción** usuario-sistema ni la posibilidad de **adaptación**. Las explicaciones, por tanto, requieren que el usuario esté familiarizado con el método de razonamiento.

El método es sencillo e intuitivo en el caso de variables binarias, pero resulta complicado en el caso de variables multivaluadas. Sin embargo, la mayor limitación del método es que sólo sirve para poliárboles y, por tanto, raramente puede ser aplicado a problemas del mundo real en los que los modelos casi siempre tienen bucles.

Descripción de las variaciones de la probabilidad

Aunque este método no es específico de redes bayesianas, sin embargo lo incluímos aquí por la relación que tiene con la expresión lingüística de las probabilidades. La investigación experimental ha demostrado que los seres humanos entienden mejor las expresiones lingüísticas de la probabilidad que los datos numéricos [93]. Por esta razón, Elsaesser [63] propuso un método para generar explicaciones lingüísticas consistente en rellenar una plantilla, basada en los *modelos inductivos* de Polya, que son un conjunto de reglas heurísticas para describir los cambios en las probabilidades.

Por ejemplo, una de dichas reglas, compatible con la fórmula de Bayes, es: "Si $A \rightarrow B$ y B es verdadero, entonces la existencia de A es más creíble." La plantilla se rellena con términos expresados en lenguaje natural que representan los datos probabilísticos y sus variaciones. Las expresiones que denotan probabilidades varían entre los que reflejan probabilidades muy altas, como *casi seguro* (para valores entre 0.99 y 0.91) o *muy probable* —para el intervalo [0.90, 0.82]— y aquéllos que designan valores pequeños, como *improbable* (para el rango de 0.18 a 0.09) o *muy improbable* (para el rango de 0.08 a 0.01). Existen también expresiones para definir los cambios de la probabilidad p_1 a la probabilidad p_2; por ejemplo, cuando $p_2/p_1 \geq 5$, la expresión asociada es "muchísimo más probable"; cuando $2{,}5 < p_2/p_1 \leq 5$, la expresión asociada es "mucho más probable", etc.

Posteriormente, un experimento psicológico desarrollado por Elsaesser y Henrion [62] demostró que las expresiones lingüísticas para representar las variaciones entre las probabilidades (p_1, p_2) dependen más de la diferencia $p_2 - p_1$ que de la razón p_2/p_1 o de la

proporción de las razones de probabilidad $(p_1/(1-p_1))/(p_2/(1-p_2))$.

El objetivo de este método de explicación es la **descripción verbal** de los resultados del proceso de razonamiento, más que la comprensión de dicho proceso. Las explicaciones se corresponderían con el **nivel micro** y **no** existe ninguna suposición sobre **causalidad**. Este método está diseñado para usuarios que no estén familiarizados con la teoría de la probabilidad, aunque no existe ningún modelo explícito de usuario.

INSITE

El objetivo del método INSITE de Suermondt [174, 175] es identificar los hallazgos que influyen en la probabilidad a posteriori de una hipótesis dada, además de los caminos por los que fluye la evidencia. La influencia de la evidencia $\mathbf{e}$ sobre una variable determinada D se mide mediante una función de coste. Después de analizar diferentes funciones de coste, Suermondt concluye que la que mejor se adapta a sus objetivos es la *entropía cruzada*,

$$H(P(D|\mathbf{e}); P(D)) = \sum_i \left[p(d_i|\mathbf{e}) \log \left(\frac{p(d_i|\mathbf{e})}{p(d_i)} \right) \right] \tag{2.3}$$

donde d_i representa los posibles valores de D. La influencia de los hallazgos, tanto individualmente como en conjunto, se determina mediante un *análisis de sensibilidad*, consistente en calcular el coste omisión de tales hallazgos.

El método de Suermondt también busca, como hemos indicado, las *cadenas de razonamiento* más relevantes, es decir, los caminos desde la evidencia $\mathbf{e}$ a la variable de interés D que están relacionados computacionalmente con D dada $\mathbf{e}$. Analiza la *fuerza*[4] de cada una de ellas y determina si está en conflicto con el resultado de la inferencia. Las cadenas se muestran gráficamente al usuario, sombreando las que están en conflicto con el resultado y resaltando las consistentes con él. Además, si el usuario lo desea, éstas son descritas verbalmente.

El objetivo de INSITE es la **comprensión** del razonamiento. Puede ofrecer explicaciones tanto a **nivel micro** (midiendo la influencia entre un nodo y sus vecinos) como a **nivel macro** (identificando los caminos de razonamiento más relevantes). INSITE puede ser aplicado tanto a redes **causales** como **no causales**. La interacción con el usuario se lleva a cabo a través de menús, ventanas, cuadros de diálogo, botones y mediante la selección de nodos o enlaces en la representación gráfica de la red. Las explicaciones se presentan como una combinación de **gráficos y texto**. En este último caso, las probabilidades se expresan por números, junto con expresiones lingüísticas. Este método, sin

[4]La fuerza de una cadena puede considerarse como la importancia de dicha cadena en la obtención del resultado de la inferencia.

embargo, no tiene en cuenta ningún modelo de usuario y el nivel de detalle está fijo, aunque puede "adaptarse" a distintos usuarios ya que éstos tienen la posibilidad de modificar los ficheros de *ayuda* incluyendo sus experiencias para el futuro.

Uno de los aspectos más interesantes de INSITE es que es independiente del algoritmo de propagación de la evidencia. Por el contrario, su mayor inconveniente es la complejidad computacional, que crece con el número de nodos y arcos de la red.

Escenarios

Según Druzdzel y Henrion [50, 54, 85], un *escenario* es una asignación de valores para aquellas variables relevantes para una determinada conclusión, ordenadas de modo que formen una historia coherente —causal, si es posible— compatible con la evidencia. El uso de escenarios se basa en estudios psicológicos [144] que muestran que los seres humanos tienden a interpretar y explicar los procesos después de sopesar las historias más "creíbles" que incluyan la hipótesis a demostrar o su hipótesis contraria. (Una *hipótesis* en este contexto es la asignación de un valor a una variable discreta.) Por ejemplo, en el caso de un niño de 5 años con exantema y fiebre alta, un escenario podría ser: Edad *5*, Fiebre *alta*, Exantema *presente*, porque el exantema se produce con cierta frecuencia en los niños cuando padecen fiebres altas.

Aunque un escenario puede contener todos los nodos de la red, es más razonable incluir sólo aquellos nodos *relevantes* para una determinada tarea [57, 106, 107, 172]. Si existe cierta hipótesis focal H seleccionada por el usuario, los nodos relevantes son aquéllos que influyen en la probabilidad a posteriori de H dada la evidencia observada $\mathbf{e}$. En otro caso, los nodos relevantes son aquéllos cuyas probabilidades dependen de $\mathbf{e}$. Después de seleccionar los nodos relevantes, los escenarios se generan mediante alguno de los métodos de abducción parcial —véase la sección 2.4.1. La explicación consiste en mostrar la evidencia, los escenarios más probables compatibles con la hipótesis, aquéllos incompatibles con ella y una comparación entre las probabilidades de los escenarios más probables.

El objetivo de la explicación basada en escenarios es la **comprensión** del proceso de razonamiento, aunque el algoritmo de propagación no esté basado en escenarios. El nivel de explicación es **macro** (porque cada escenario contiene varias variables). El método está diseñado principalmente para redes bayesianas **causales**. Las explicaciones se presentan mediante **textos** que describen cada escenario en lenguaje natural. Las probabilidades de los escenarios se expresan **numéricamente**. Druzdzel y Henrion no describen la **interacción** usuario-sistema ni las posibilidades de **adaptación**. En principio, estas

explicaciones no requieren que el usuario esté familiarizado con el razonamiento probabilístico, aunque el conocimiento de los métodos de propagación utilizados ayudaría a entender las explicaciones.

Razonamiento cualitativo

Druzdzel y Henrion [50, 56, 55, 52, 85] también propusieron otro método de explicación basado en la transformación de una red bayesiana causal en una *red probabilística cualitativa* (RPC)[5] [184, 185]. La motivación de este método de explicación es que la gente normalmente razona y explica su razonamiento en términos de relaciones cualitativas. En una RPC la relación entre dos nodos adyacentes se marca como positiva (+), negativa (−), nula (0) o desconocida (?), basándose en el siguiente criterio: Se dice que A influye positivamente en C si

$$\forall b, \forall c, a_i > a_j,\ P(C \geq c|a_i, b) \geq P(C \geq c|a_j, b) \wedge \exists c,\ P(C \geq c|a_i, b) > P(C \geq c|a_j, b)$$

La idea que subyace es que al tomar la variable A valores más altos, aumenta la probabilidad de que la variable C también tome valores mayores. Las definiciones de influencia negativa, nula o desconocida son similares. Existen también relaciones que involucran a más de dos nodos, como las sinergias positivas o negativas. La principal ventaja de las RPC es que simplifican la construcción de modelos, porque no requieren la obtención de parámetros numéricos; en consecuencia, su principal desventaja es la ausencia de precisión en los resultados, especialmente porque la combinación de influencias "positivas" y "negativas" lleva a relaciones "desconocidas".

El algoritmo de propagación cualitativa propuesto por Druzdzel y Henrion consiste en el intercambio de mensajes entre nodos vecinos. Los nodos de evidencia se marcan con "+" o con "−". Los nodos no observados inicialmente se marcan con "0". El signo del mensaje (+, −, 0 o ?) desde X_i a X_j está determinado por el producto del signo de X_i y el signo del enlace.

El objetivo de la explicación es determinar el impacto cualitativo que cada hallazgo f ha producido sobre una determinada variable de interés V y encontrar los caminos activos desde f hasta V. Existen tres tipos de explicaciones elementales, correspondientes a los tres tipos básicos de inferencia cualitativa:

a) inferencia predictiva, que va desde las causas a los efectos, es decir, en la dirección de los enlaces; la explicación para este tipo de inferencia relativa a un enlace $A \rightarrow B$ es del tipo"A puede causar B".

[5]Del inglés **Q**ualitative **P**robabilistic **N**etwork (QPN).

b) inferencia abductiva, que va desde los efectos a las causas, es decir, en el sentido opuesto al de los enlaces; la explicación puede ser "B es evidencia para A".

c) inferencia intercausal, que determina el impacto cualitativo de la evidencia para una variable A sobre otra variable B cuando ambas son antecesoras de una tercera variable C, sobre la que existe evidencia independiente (por ejemplo, C ha sido observada); la explicación generada en este caso es del tipo: "A y B pueden causar C; como A explica C, no existe (casi) evidencia para B".

Si hay varios hallazgos y varias variables de interés, el proceso se repetirá varias veces. Análogamente, si existen varios caminos activos entre un hallazgo f y la variable de interés V, la explicación se muestra de forma secuencial, ya que el razonamiento humano y el lenguaje natural son esencialmente secuenciales.

El objetivo de este método es la **comprensión** del proceso de razonamiento. El nivel es **macro**. El método sólo sirve para redes **causales**. Las explicaciones se muestran mediante **texto**, que incluye expresiones **lingüísticas** para las probabilidades. Como en el caso de los escenarios, Druzdzel y Henrion no describen la **interacción** usuario-sistema ni las posibilidades de **adaptación**. En principio, estas explicaciones cualitativas no requieren que el usuario tenga conocimientos sobre los algoritmos de propagación cualitativa, aunque conocerlos le podría ayudar a comprender mejor las explicaciones.

En esta misma línea debemos mencionar los trabajos de Renooij y colaboradores cuyo objetivo es evitar en lo posible los resultados ambiguos que se suelen obtener al hacer inferencia en una red cualitativa. Por una parte, en [150] han diseñado un formalismo para la resolución de ambigüedades sin hacer uso de la información numérica. La idea consiste en utilizar *redes cualitativas aumentadas*, que son redes cualitativas pero en las que se distingue influencias fuertes y débiles, en contraste con las redes de Wellman [184, 185] que no hacen esta distinción. Para ello, a cada influencia se le asocia una fuerza relativa. El algoritmo para la propagación de los signos en este tipo de redes es una generalización del existente para redes cualitativas, aunque la principal diferencia es que tiene en cuenta que las influencias fuertes dominan sobre las débiles.

Además, han diseñado otro algoritmo, descrito en [151], para calcular resultados informativos para una hipótesis, en los casos en los que su signo sea ambiguo después de hacer inferencia. El algoritmo se basa en centrarse en la subred formada por los caminos entre la evidencia y el nodo de interés y determinar así el signo del *nodo pivote*[6], que es

[6]El nodo pivote es un nodo que separa la parte de la red que contiene los intercambios del resto de la

el que define el signo de la hipótesis. La ambigüedad en el pivote se resuelve en términos de fuerzas relativas de influencia entre ellos, además de los signos de los nodos que determinan su solución.

Basándose también en las redes cualitativas de Wellman, Parsons [136, Capítulo 7] aporta otro enfoque similar al anterior, aunque con ciertas diferencias, al definir las *redes de certeza cualitativas* (RCCs).[7] La principal diferencia con las anteriores radica en que en el contexto de las RCCs, lo que se analiza son las influencias que ejerce cada estado de la variable A sobre cada estado de otra variable B, a diferencia de las RPCs donde se comparan distribuciones de probabilidad. Hay que hacer notar que en el caso de que las variables sean binarias, ambos paradigmas coinciden. Sin embargo, no nos extenderemos más porque no se han diseñado métodos de explicación específicos para este tipo de redes.

DIAVAL

La mayoría de las herramientas para redes bayesianas limitan el proceso de inferencia a mostrar las probabilidades de los estados de cada variable. Esto es claramente insuficiente para los sistemas expertos prácticos, especialmente cuando la red contiene un gran número de nodos. Por esta razón, el sistema experto DIAVAL [45] implementó un método para seleccionar los diagnósticos más probables y relevantes. El concepto de relevancia en DIAVAL se define como una medida de la importancia desde el punto de vista médico: por ejemplo, una enfermedad como la estenosis mitral es más "relevante" que una variable patofisiológica intermedia como la hipertensión de la aurícula izquierda; durante la construcción de la red bayesiana, a cada nodo V se le asignan dos factores de forma subjetiva: la relevancia positiva (negativa) se aplica cuando el valor más probable es $+v$ ($\neg v$). El sistema sólo muestra los diagnósticos cuya probabilidad a posteriori y relevancia superan tanto el *umbral de certeza* como el *umbral de relevancia*, respectivamente.

El usuario puede seleccionar con el ratón cada diagnóstico y abrir un menú con diferentes opciones, incluyendo la posibilidad de visitar los padres y los hijos de cada nodo, como se describió en la sección 2.4.2. De este modo, el usuario puede investigar en qué vecinos del nodo ha aumentado o disminuido su probabilidad. Además, si un nodo Y es el hijo de una puerta OR, la probabilidad de que cada padre X_i haya producido Y se puede calcular como $P(+x_i|\mathbf{e}) \times c_i$, donde c_i es el parámetro de la puerta OR asociado al enlace $X_i \rightarrow Y$, y representa la probabilidad de que X_i, estando presente, produzca Y. El cálculo

red

[7] Del inglés **Q**ualitative **C**ertainty **N**etwork (QCNs).

para la puerta MAX es similar.

El objetivo de este método de explicación es fundamentalmente **describir** los resultados de la inferencia y ofrece una ayuda básica para la **comprensión** del razonamiento. El nivel de explicación es **micro**. El método supone que el modelo es **causal** e incluye patrones específicos de explicación para las puertas OR y MAX. La **interacción** usuario-sistema está implementada mediante ventanas y menús. Las probabilidades se expresan **numéricamente**. No existe ningún **modelo de usuario** y la única capacidad de adaptación es que el usuario puede controlar el **nivel de detalle** en la selección de diagnósticos modificando los umbrales de relevancia y de certeza.

BANTER

Basándose en el método INSITE, Haddaway, Jacobson y Kahn [77, 78, 79] han desarrollado BANTER, una herramienta para la ayuda a la decisión y para la educación, especialmente en medicina, aunque funciona con cualquier red cuyos nodos se puedan clasificar en hipótesis, observaciones y pruebas. Dada cierta evidencia, BANTER puede ofrecer la probabilidad de una hipótesis o seleccionar la prueba que en mayor medida confirma o descarta un posible diagnóstico. Cuando se usa como herramienta educativa, BANTER genera de forma aleatoria una serie de escenarios y el usuario tiene la posibilidad de realizar un diagnóstico o de seleccionar una prueba para confirmarlo o descartarlo.

BANTER puede generar también explicaciones verbales (utilizando una modificación del método de Suermondt), identificando los hallazgos más influyentes y seleccionando los caminos más fuertes[8] y con menor número de nodos.

Ambos métodos comparten el propósito de la **comprensión** del proceso de razonamiento y ofrecen explicaciones a **nivel micro**. Difieren en que BANTER está diseñado para redes bayesianas **causales** y sólo genera explicaciones **verbales**, mientras que INSITE puede mostrar también explicaciones gráficas. En BANTER, las expresiones de la probabilidad son sólo **cuantitativas**. Ninguno de ellos ofrece la posibilidad de **adaptación al usuario**. Como BANTER está especialmente concebido para usuarios noveles e inexpertos, no se requiere ningún conocimiento sobre redes bayesianas, aunque sí del dominio de aplicación y unas nociones básicas sobre probabilidad.

[8] Dado un camino C, se define $Fuerza(C) = \text{mín}\,|(P(n) - P(n|e)|, \forall n \in C$

B2

Sin embargo, existen dos grandes inconvenientes de BANTER cuando se desea utilizar como herramienta educativa: que no es capaz de realizar razonamiento hipotético y que ofrece demasiada información que a veces no es relevante para la conclusión. Como la forma de responder a las preguntas del usuario está muy alejada de la forma de expresarse de los humanos, las explicaciones no son fáciles de entender. Con la intención de solventar estas deficiencias, McRoy y colaboradores [114, 115, 116] han desarrollado el sistema B2 como una extensión de BANTER, que tiene la capacidad de generar explicaciones **verbales** en lenguaje natural, de manera más consistente que como lo hacía BANTER, y ofrece además explicaciones **gráficas**.

En el sistema B2 se ha incluido la capacidad de **explicación estática** del dominio médico representado, así como un modelo del discurso, representado mediante una red semántica proposicional. Esto permite al sistema "razonar" sobre el contenido y la estructura de la interacción usuario-sistema, lo que evita producir información irrelevante. Por tanto, podemos decir, en resumen, que el objetivo básico de este sistema es el de la **comprensión** tanto del **modelo** como del **razonamiento**, además de permitir realizar **razonamiento hipotético**. Al igual que en BANTER, las expresiones de la probabilidad son sólo **cuantitativas** y tampoco ofrece la posibilidad de adaptación al usuario. La **interacción** entre el usuario y el sistema se lleva a cabo a través de ventanas y de preguntas en **lenguaje natural**.

Peso gráfico de la evidencia

Madigan, Mosurski y Almond [108] proponen un **método gráfico** que consiste en mostrar cómo la evidencia se propaga a través de una red bayesiana causal. El impacto de la evidencia sobre un nodo binario H se mide calculando el *peso de la evidencia* de Good [74],

$$W(H:\mathbf{e}) = \log\left(\frac{P(\mathbf{e}|+h)}{P(\mathbf{e}|\neg h)}\right) \tag{2.4}$$

y mostrando en el propio grafo de la red cómo varían de color y de anchura los nodos y los enlaces, etc. El sistema proporciona explicaciones para dos tipos de preguntas:

- *¿Cuál es la importancia relativa de cada hallazgo f sobre la variable de interés H?* Ésta se responde visualizando $W(H:f)$ en un *balance gráfico* en el que se muestran todos los hallazgos junto con su importancia sobre la hipótesis. Hay que señalar que dicha importancia depende del orden en el que el usuario asigna los valores.

- *¿Por qué un nodo concreto tiene tanta influencia sobre la variable de interés?* La respuesta consiste en mostrar los caminos relevantes, de forma similar a las *cadenas de razonamiento* de Suermondt.

El objetivo de este método de explicación es la **comprensión** del proceso de razonamiento. El nivel es **macro**. Aunque Madigan et al. hablan de redes bayesianas causales, su método funcionaría también para redes **no causales**. La interacción usuario-sistema está basada en **menús**, las explicaciones se muestran **gráficamente** y las probabilidades se codifican de forma **cualitativa** por medio de colores y grosores. No se ha desarrollado ninguna capacidad de adaptación para este método.

Otros trabajos

En la actualidad diferentes grupos de investigación trabajan en el desarrollo de nuevos métodos de explicación que intentan mejorar algunos de los problemas que existen en los que hemos estudiado, especialmente aquellos relacionados con la complejidad computacional (exponencial en la mayoría de los casos). Entre los proyectos que se están llevando a cabo, merece la pena destacar los siguientes:

- Los últimos trabajos de Zukerman y colaboradores [113, 195, 196] se centran en el uso de redes bayesianas para la generación de *argumentos* en lenguaje natural. En sus últimos trabajos proponen la utilización de redes bayesianas como base del sistema NAG (*Nice Argument Generator*) para el análisis y generación de razonamientos que sean "suficientemente buenos" para el usuario, es decir, que sean convincentes. Así, dada una proposición objetivo introducida por el usuario, y el contexto en el que se realiza dicha proposición, junto con el grado de credibilidad que se desee alcanzar, el sistema genera los argumentos que la justifican; además, analiza aquéllos razonamientos proporcionados por el usuario y prepara otros para refutarlos cuando sea necesario. Ambas bases de conocimiento, el modelo del usuario y el modelo normativo, se representan por medio de redes bayesianas. Los nuevos argumentos que se elaboran, se van añadiendo al modelo normativo. Ya que dichas redes pueden ser demasiado complejas, se utiliza un mecanismo para, en ambos modelos, centrar la atención en las subredes relevantes sobre las que se generará el razonamiento. La intersección de las estructuras de dichas subredes forma el grafo de razonamiento. Para analizar dicho grafo se realiza la propagación sobre las redes bayesianas que representan el modelo del usuario y el modelo normativo, utilizando la información probabilística asociada a cada modelo.

- En la misma línea se engloba el trabajo de Biris [5] en el que se utilizan redes bayesianas para generar explicaciones con el objetivo de modelar un sistema de forma automática. La motivación del trabajo se basa en la utilización del *modelado composicional (MC)* [64, 75, 105, 128] como técnica mediante la que generar explicaciones informativas en el campo de la formación industrial, dependiendo de los distintos niveles de conocimiento del usuario y de las diferentes complejidades del dominio a modelar. Esta técnica se basa en componer distintos fragmentos de modelo, cada uno de los cuales describe sólo algunos aspectos del comportamiento de ciertos elementos. Por ello, el MC permite variar el detalle de la representación del modelo completo combinando el detalle de los distintos fragmentos utilizados como "bloques de construcción". Las redes bayesianas se utilizan para seleccionar los fragmentos de modelo adecuado en cada momento. Para ello, se diseña una red cuyos nodos representan la selección de los distintos fragmentos del modelo y los enlaces representan las relaciones entre ellos, dependiendo de la descripción global del sistema.

- En un campo distinto están los trabajos de Dittmer y Jensen [46], quienes utilizan herramientas de análisis de sensibilidad para generar explicaciones en redes bayesianas y que han aplicado a la red BOBLO para determinar si el pedigrí asignado al ganado es correcto. Utilizan el algoritmo *HUGIN* para detectar conflictos entre los datos, y el análisis de sensibilidad para determinar cómo los cambios en la evidencia afectan a la conclusión.

- También se han utilizado técnicas de análisis de sensibilidad para explicar gráficamente el razonamiento de la red causal que representa *la falacia del jurado* [67] (véase la sección 2.4.2), con el objetivo de demostrar que, en muchas ocasiones, es incorrecta.

- Por último, el proyecto ARPI, desarrollado en el Laboratorio de Palo Alto del Centro Científico Internacional de Rockwell, tiene como uno de sus objetivos la evaluación y explicación de planes representados por medio de redes causales. Como parte del mismo, se intentan crear nuevos algoritmos para explicación en redes causales [47]. Sin embargo, hemos de lamentar no haber podido encontrar ninguna otra referencia bibliográfica en relación con este proyecto ni con sus trabajos relacionados.

Teoría de la decisión

Finalmente, aunque no estudiaremos los estudios relacionados con la teoría de la decisión, Druzdzel, en el capítulo 9 de su tesis [50], realiza un análisis amplio sobre este tema y podría servirnos de referencia si quisiéramos ampliar el problema de la explicación en redes bayesianas al caso de diagramas de influencia, donde sí aparecen nodos decisión y utilidad.

Parte II

NUEVO MÉTODO DE EXPLICACIÓN

Capítulo 3

Explicación del modelo

En este capítulo proponemos un método para explicar el modelo representado por cualquier red bayesiana, aunque describimos una serie de capacidades específicas para redes bayesianas causales médicas. La explicación del modelo la hemos dividido en tres apartados: explicación de nodos (sección 3.2.1), explicación de enlaces (sección 3.2.2) y explicación global de la red (sección 3.2.3). Finalizamos realizando una valoración del método de explicación para lo que presentamos sus principales propiedades y limitaciones (sección 3.3).

3.1. Introducción

En cualquier sistema experto, y en concreto en redes bayesianas, tan importante como la corrección y eficiencia del método de razonamiento, es que la red refleje exactamente el conocimiento experto que se desea modelar, ya que si éste no se representa correctamente no se puede esperar ningún resultado fiable del sistema.

Sin embargo, aunque se ha investigado bastante sobre los algoritmos de propagación y de aprendizaje de la evidencia dentro de la red, no ocurre lo mismo con el proceso de construcción de redes bayesianas. De hecho, al no existir una metodología para llevar a cabo esta tarea, está considerada más como un arte que como una técnica, lo que hace que la construcción de la red sea un proceso difícil y laborioso.

Por otra parte, una vez finalizada la construcción del sistema experto, el usuario debería tener la posibilidad de obtener algún tipo de explicación del conocimiento que está representado en dicho sistema.

Como ya comentamos en la sección 2.2.1, la explicación del modelo consiste en mostrar la información contenida en la base de conocimientos, con el objetivo de *ayudar* al ingeniero de conocimiento que está modelando la red, y al experto humano a detectar posibles errores durante la fase de *construcción* del sistema experto y a depurarlos. Así mismo,

serviría para proporcionar al usuario, sobre todo al que no es experto en el dominio representado, un modo de obtener nuevo conocimiento sobre el dominio, con *propósitos educativos*.

Como podemos deducir del análisis de los métodos desarrollados hasta ahora (sección 2.4), la mayoría de ellos ofrecen como explicación del modelo la mera visualización del grafo que representa la red. Esto puede ser útil para la detección de relaciones de dependencia y/o independencia, pero sólo en los casos en los que se dan las siguientes condiciones:

- *Que el número de enlaces y nodos de la red no sea muy grande.* Normalmente, este no es el caso en los dominios que modelan situaciones reales, y concretamente en el campo de la medicina, por lo que la visualización del grafo se ha de hacer por partes, cuando las herramientas lo permiten, lo que no ocurre siempre. Esto puede provocar confusión en el usuario al no poder tener una visión generalizada del modelo o incluso al no poder centrarse en el fragmento de red en la que esté interesado.

- *Que el usuario esté familiarizado con las redes bayesianas.* En este punto podemos pensar que quien va a utilizar una herramienta para el procesamiento de redes bayesianas, debe tener conocimiento del paradigma utilizado. Sin embargo, pensemos en el ingeniero de conocimiento que está diseñando un modelo con la ayuda de un experto que no tiene ninguna idea de lo que representa el grafo (dependencias, relaciones, etc). En ese caso, la responsabilidad de la correcta construcción del modelo depende casi exclusivamente del ingeniero de conocimiento y de su capacidad de explicación para transmitirle al experto toda la información que hay representada en la red. Por tanto, la validez del modelo y la aceptación del sistema por los usuarios dependerán de algo tan subjetivo y concreto como es la capacidad que tenga el ingeniero para llevar a cabo dicha explicación.

Por otro lado, simplemente con la visualización del grafo, la única información que posee el usuario (sea experto o no) está relacionada con las influencias entre las variables, pero no tiene ningún tipo de información sobre si son positivas o negativas, ni sobre la cantidad de influencia que transmite una variable a otra, etc. Por estos motivos, se hace necesario proporcionar un método de explicación que sirva, por un lado, para convencer tanto al usuario como a los expertos de los que se extrae el conocimiento, de que la información con la que trabajará el sistema es correcta, y por otro, para transmitir al usuario el significado de dicha información.

3.2. Descripción

El método consiste, a grandes rasgos, en mostrar la información que representa cada nodo y cada enlace individualmente, aunque el usuario también tiene la posibilidad de solicitar la explicación de varios nodos y enlaces a la vez e, incluso, de la red completa. Para ello, el usuario simplemente debe indicar el objeto sobre el que desea que el sistema le genere la explicación (nodo, enlace o red completa). En las subsecciones siguientes analizamos más detenidamente cada una de estas opciones.

3.2.1. Explicación de nodos

La explicación de un nodo consiste en mostrar toda la información asociada al mismo de forma comprensible al usuario. Atendiendo a la clasificación hecha en la sección 2.2.2, relativa a la presentación al usuario, detallamos las formas de presentar la explicación de un nodo determinado, tanto de forma textual como gráfica. Finalmente, concluimos esta sección ofreciendo diversas posibilidades de **simplificación** de explicaciones gráficas de los nodos, de cara a una mejor comprensión por parte del usuario.

Presentación textual

Este tipo de presentación, a nivel micro, está orientada a aquellos usuarios que deseen un cierto **nivel de detalle** a la hora de obtener la explicación, por lo que se les supone algún tipo de conocimiento sobre probabilidades. En este caso, el método propone la presentación al usuario de la siguiente información relativa a un nodo, seleccionado por el usuario:

- *Nombre* que identifica y diferencia el nodo dentro de la red.
- *Estados* que puede tener el nodo en cuestión, junto con sus probabilidades asociadas.
- *Definición*. Es un texto que proporciona al usuario la posibilidad de profundizar en el significado y en el papel que desempeña el nodo en la red. Este dato puede modificarse, por lo que se puede utilizar para generar explicaciones de la red completa con distinto nivel de detalle.
- *Padres* del nodo, identificados por su nombre. Así, aunque el número elevado de padres pudiera generar cierta confusión en la representación gráfica de la red, el usuario tendría la posibilidad de conocer con certeza cuáles son los antecesores inmediatos del nodo en cuestión.

- *Hijos* del nodo, por razones análogas a las comentadas anteriormente.
- *Razones de probabilidad a priori* entre los estados del nodo, lo que permite al usuario comprobar si las probabilidades asignadas durante la construcción de la red, bien mediante tablas de probabilidad o árboles, funciones, etc., reflejan las diferencias cuantitativas de la "importancia" de cada uno de los estados. Por ejemplo, supongamos que un nodo X tiene tres estados x_1, x_2 y x_3. Al asignar probabilidades a dichos estados, el usuario puede otorgar: 0.1, 0.6 y 0.3, respectivamente, sin pensar en que el estado x_2 es 6 veces más probable que el estado x_1 y el doble que x_3. Al mostrarle las razones de probabilidad, puede detectar ciertas inconsistencias entre los valores asignados y proceder a su corrección. La experiencia nos ha enseñado que, a la hora de obtener valores probabilísticos, los médicos son bastante dados a ofrecerlos sin pensar detenidamente en las relaciones y diferencias entre los valores. Por eso, en nuestro trabajo proponemos que los estados del nodo se muestren ordenados de mayor a menor según sus probabilidades. Así el usuario puede localizar fácilmente cuál es el estado más probable y cuántas veces lo es más que el resto, lo que le permitirá hacer las modificaciones que considere necesarias hasta que las razones de probabilidad se ajusten a su estimación subjetiva.
- *Razones de probabilidad a posteriori.* Por los mismos motivos citados en el apartado anterior, se muestran estos valores con la intención de que el usuario pueda comprobar cómo se ha modificado la probabilidad de cada estado al propagar la evidencia. Como además la probabilidad a posteriori de un estado determinado es proporcional a la probabilidad a priori de dicho estado $P(x|e) = \alpha P(x)P(e|x)$ le puede dar una idea al usuario de por qué la probabilidad a posteriori de un estado ha variado más o menos con relación a otros.
- *Otras propiedades del nodo*, como son el *factor de importancia* y la *función* que desempeña el nodo en la red, así como el *valor "normal"* del nodo en cuestión y que detallamos más adelante.

Presentación gráfica

Al ser ésta una de las formas más accesibles a cualquier tipo de usuario, se plantea la necesidad de incluirla como modo de presentación básica de una explicación, orientada especialmente a aquellos usuarios que no tienen ningún tipo de conocimiento sobre teoría de la probabilidad, algoritmos de inferencia probabilística, etc. Por esta razón, como características primordiales han de destacar su *sencillez* y *simplicidad* a la hora de

mostrar la información asociada a cada nodo, independientemente del número de nodos que contenga la red. Además, debe ser comprensible para cualquier tipo de usuario. Por estas razones, proponemos la visualización gráfica de la información básica asociada a cada nodo, a saber:

- el nombre del nodo, que permite distinguirlo entre los demás;
- los nombres de los estados, para identificar el conjunto de valores que puede tomar la variable que representa, destacando gráficamente el más probable;
- las probabilidades asociadas a cada estado, representadas de forma gráfica mediante una barra de longitud proporcional a su valor.

La principal ventaja que tiene este tipo de explicación es que el usuario puede conocer simultáneamente los valores más probables de cada nodo. Esto le permitirá detectar y corregir posibles errores antes de comenzar a utilizar el sistema.

Simplificación de la explicación gráfica

Es obvio que mostrar esta información de modo gráfico puede crear mucha confusión cuando la red tiene un gran número de nodos. Por otro lado, incluso en aquellos casos en los que la red no sea muy grande puede ser que el usuario no desee visualizar gráficamente la información asociada a todos los nodos de la red, sino que prefiera centrarse en uno o varios determinados. Por estos motivos, el método planteado proporciona la posibilidad de mostrar la explicación gráfica solamente de aquellos nodos seleccionados por el usuario.

La selección puede realizarse explícitamente, indicando exactamente el nodo o los nodos sobre los que desea obtener este tipo de explicación gráfica; o implícitamente, es decir, de modo automático, de acuerdo con los siguientes criterios:

Función. La *función* de un nodo se asigna durante la construcción de la red y representa el papel que ejerce dicho nodo en la red. Al estar definiendo un método de explicación para redes bayesianas causales, todos los antecedentes de un nodo podrían ser considerados como causas y los descendientes como efectos. Sin embargo, es obvio que esta es una clasificación demasiado genérica, pues el conjunto de funciones se puede ampliar, sobre todo dependiendo del dominio representado. Sin embargo, después de analizar distintos tipos de redes, hemos encontrado propiedades comunes en la mayoría de ellas. Así, en aplicaciones médicas un nodo puede englobarse dentro de una de las siguientes categorías:

- Enfermedad/anomalía. Normalmente, las redes bayesianas se utilizan en sistemas expertos de diagnóstico, cuyo principal objetivo es el de mostrar la probabilidad de padecer una enfermedad o de sufrir una anomalía.
- Factor de riesgo, utilizado para representar las variables que favorecen una enfermedad, una anomalía, etc. Por ejemplo, la edad es un factor de riesgo para la hipertensión.
- Síntoma. Se utiliza para representar un indicio de algo que está sucediendo. En medicina es lo que el paciente experimenta y constituye un factor revelador de una enfermedad; por ejemplo, la fiebre es un síntoma de ciertas enfermedades.
- Signo de la exploración. En medicina, la diferencia entre signo y síntoma es que los signos son factores medidos de forma objetiva por el médico, mientras que los síntomas suelen ser relatados por los pacientes de forma subjetiva. Por ejemplo, una erupción cutánea es un signo; el dolor es un síntoma.
- Prueba de laboratorio, como dato que ayuda a descartar o confirmar la presencia de una enfermedad o anomalía; por ejemplo, un análisis de sangre, un electrocardiograma, una radiografía.
- Auxiliar, para aquellos nodos que se definen con la intención de simplificar la representación de la red. Por ejemplo, en la red ASIA [82], el nodo Tuberculosis_o_Cancer sería un nodo auxiliar que permite simplificar la red.
- Definido por el usuario.
- Otro, para el resto.

Podría pensarse que esta clasificación no es exclusiva porque un mismo nodo puede desempeñar distintas funciones dependiendo de la relación en la que se le considere.

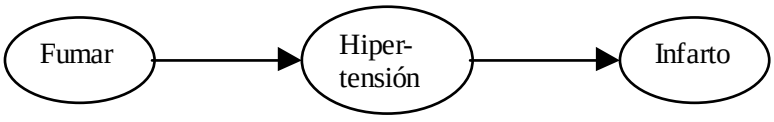

Figura 3.1: Ejemplo de distintas funciones de un nodo en varias relaciones

Por ejemplo, en la figura 3.1 observamos que si consideramos la variable *hipertensión* con arreglo a la relación *Tabaco-Hipertensión*, desempeña el papel de *Enfermedad*. Sin embargo, de acuerdo a la relación *Hipertensión-Infarto*, se consideraría como *Factor de riesgo*. La cuestión radica en que no estamos analizando el papel del nodo

en cuestión sino el tipo de relación en la que está implicado y que se define por los enlaces que la integran. Por tanto, distinguiremos la función que desempeña el nodo respecto a la red entera, del papel asignado al enlace, del que hablaremos en la sección 3.2.2.

El papel de cada nodo dentro de la red se asigna, opcionalmente, durante la fase de construcción de la misma, por lo que se requiere para ello que el usuario tenga cierto conocimiento del dominio representado. En caso contrario, por defecto se asigna el papel *Otro* a cada nodo, pero en este caso no se podrá sacar partido de ciertas facilidades de explicación, sobre todo de cara a la generación de explicaciones de la red completa, o a nivel macro, etc.

Factor de importancia. Se define como el valor, asignado en una escala de 0 a 10, que representa el interés del usuario en dicho nodo, no solo a la hora de obtener explicaciones sino también a la hora de analizar los resultados obtenidos por el sistema. El usuario es quien asigna subjetivamente este valor de acuerdo con el conocimiento que tenga del dominio representado. En caso contrario, se asigna un valor dependiendo de la función que desempeña el nodo en la red. Los factores por defecto serían: Enfermedad/Anomalía $\equiv 9$; Síntoma y Signo $\equiv 8$; Factor de riesgo $\equiv 7$; Otro $\equiv 5$ y Auxiliar $\equiv 3$. La justificación para la asignación de estos valores radica en el hecho de que normalmente los sistemas expertos se diseñan con la intención de ayudar al usuario a realizar un diagnóstico. Por lo tanto, los nodos que lo representan serían los que más importancia tendrían, a priori, en la red. A continuación le siguen todos aquellos nodos que representan las causas o los hallazgos que favorecen a dicho diagnóstico. Los factores de riesgo se consideran menos importantes que los anteriores. Por último, se supone que los nodos auxiliares juegan el papel menos importante para la emisión de un diagnóstico. Además, como veremos más adelante, este factor de importancia sirve también para generar explicaciones de la red completa dependiendo del nivel de detalle requerido por el usuario.

Valor normal. Después de analizar los tipos de nodos de una red bayesiana médica, podemos clasificarlos en ordinales y no ordinales, entendiendo por nodo *ordinal* aquél que admite una ordenación de sus estados, por ejemplo, el nodo Fiebre puede tomar los valores ausente, leve, moderado y severo. En este tipo de nodos, normalmente el primer estado es el que representa el estado normal del nodo. La utilidad de distinguir un estado de ese modo es que, como ya han apuntado otros autores [71], permite realizar simplificaciones de la red, en los casos en los que ésta contiene

un gran número de nodos. Por tanto, el hecho de mostrar solamente la información relacionada con las variables que toman un valor distinto del habitual, puede simplificar enormemente la explicación, tanto del modelo como del razonamiento.

No obstante, el valor normal puede ser otro distinto del primero; por ejemplo, la variable altura de una persona, puede tomar los valores bajo, medio y alto, siendo medio el valor normal. Por tanto, en el método propuesto existe la posibilidad de que el usuario defina explícitamente cuál es el estado normal de cada nodo, tal y como proponen Madigan et al. [108].

Además, hay que tener en cuenta que existen nodos no ordinales, como por ejemplo el sexo. Para estos casos, para poder sacar partido a la facilidad de explicación, se propone la ordenación arbitraria de los estados del nodo, lo que permitirá al usuario analizar cómo ciertos valores del nodo afectan a determinadas variables. Por ejemplo, en el caso del sexo, se puede observar si cierta enfermedad es más frecuente en los hombres que en las mujeres.

Las ventajas que ofrece la simplificación con arreglo a estos criterios son:

- el usuario puede solicitar al sistema que le muestre solamente la información gráfica relativa a ciertos nodos; por ejemplo, los que representan síntomas;
- puede solicitar que le muestre únicamente los nodos cuyos valores más probables sean distintos de los usuales;
- también puede solicitar mostrar sólo aquellos nodos cuyo factor de importancia es superior a un cierto valor;
- combinando sendos factores se pueden mostrar los nodos que desempeñan un determinado papel y cuyo factor de importancia es superior a uno dado, según las necesidades del usuario.

3.2.2. Explicación de enlaces

De forma similar al caso de los nodos, para los enlaces se distinguen también dos clases de explicación:

Gráfica

Este tipo de explicación consiste en representar cada enlace dependiendo de la influencia que transmite cada nodo a sus hijos, de forma similar a como propuso Wellman para

redes cualitativas [184, 185].

Definición 3.1 *Suponiendo A y B padres del nodo C, la influencia del enlace del nodo A al nodo C es **positiva** (negativa) si:*

$$\forall i, \forall j \ a_i < a_j \Rightarrow \begin{cases} \forall k, Acum(c_k|a_i, b) \leq Acum(c_k|a_j, b) \\ y \\ \exists l | Acum(c_l|a_i, b) < Acum(c_l|a_j, b) \end{cases} \tag{3.1}$$

donde $Acum(c_k)$ representa la probabilidad de que el nodo C tome un valor mayor o igual que c_k:

$$Acum(c_k) = P(C \geq c_k) = \sum_{i \geq k} p(c_i) \tag{3.2}$$

*La influencia es **nula** cuando*

$$\forall i, \forall j, \ a_i < a_j \Rightarrow \forall k, Acum(c_k|a_i, b) = Acum(c_k|a_j, b)$$

lo cual equivale a

$$\forall i, \forall j, \ a_i < a_j \Rightarrow \forall k, P(c_k|a_i, b) = P(c_k|a_j, b)$$

*La influencia es **desconocida** si no es ni positiva ni negativa ni nula, lo que sólo puede ocurrir cuando la variable tiene más de dos valores.*

La interpretación intuitiva de una influencia positiva (negativa) entre los nodos A y C es que a mayores valores de A la probabilidad de que C tome mayores valores aumenta (disminuye).

Corolario 3.1 *Sean A, B y C variables binarias, con valores $+a$ y $\neg a$, $+b$ y $\neg b$, y $+c$ y $\neg c$ respectivamente. Si A y B son condicionalmente independientes, la definición de influencia positiva entre A y C equivale a una correlación positiva entre dichas variables.*

Demostración

$$\forall b, P(+c|\neg a, b) \leq P(+c| + a, b). \tag{3.3}$$

En este caso se tiene que

$$\forall b, \frac{P(+c| + a, b)}{P(+c|\neg a, b)} \geq 1,$$

Por otro lado,

$$\begin{aligned}
P(+c) &= \sum_{a,b} P(+c|a,b)P(a,b) = \\
&= P(+c|+a,+b)P(+a,+b) + P(+c|\neg a,+b)P(\neg a,+b) + \\
&\quad +P(+c|+a,\neg b)P(+a,\neg b) + P(+c|\neg a,\neg b)P(\neg a,\neg b) \leq \\
&\leq P(+c|+a,+b)P(+a,+b) + P(+c|+a,+b)P(\neg a,+b) + \\
&\quad +P(+c|+a,\neg b)P(+a,\neg b) + P(+c|+a,\neg b)P(\neg a,\neg b) = \\
&= P(+c|+a,+b)[P(+a,+b)P(\neg a,+b)] + P(+c|+a,\neg b)[P(+a,\neg b)P(\neg a,\neg b)] = \\
&= P(+c|+a,+b)[P(+b)] + P(+c|+a,\neg b)[P(\neg b)]
\end{aligned}$$

Dado que

$$P(+b|+a) = P(+b)$$

y

$$P(\neg b|+a) = P(\neg b)$$

se tiene que

$$P(+c|+a,+b)[P(+b)] + P(+c|+a,\neg b)[P(\neg b)] = P(+c,+a)$$

Por tanto,

$$P(+c) \leq P(+c|+a)$$

Entonces,

$$P(+c,+a) = P(+c|+a)P(+a) \geq P(+c)P(+a),$$

que corresponde a la definición de correlación positiva entre A y C. □

El criterio definido para establecer las comparaciones entre las distribuciones debe cumplir la propiedad de *dominación estocástica de primer orden* [119]. Esta propiedad le confiere un carácter robusto frente a las distintas escalas ordinales sobre las que se pueda medir una variable.

Teorema 3.1 *La propiedad de dominación estocástica de primer orden para las funciones de densidad, en el caso de variables continuas, y de distribuciones de probabilidad, para variables discretas, es equivalente a la propiedad de monotonía de las razones de verosimilitud.*

Demostración Véase [119].
□

Corolario 3.2 *A influye positivamente sobre B cualquiera que sea la configuración de padres X si y sólo si*

$$\forall x, \forall a_i < a_j, \forall b_k < b_l, \ \frac{P(a_j|b_l, x)}{P(a_j|b_k, x)} \geq \frac{P(a_i|b_l, x)}{P(a_i|b_k, x)} \tag{3.4}$$

Teorema 3.2 *El tipo de influencia que transmite una variable A a su hijo B es la misma que la que transmite B a A.*

Demostración La haremos para el caso de las influencias positivas, puesto que las de las demás serían similares.

Caso binario. (Véase [136])) Supongamos que A puede tomar los valores $\neg a$ y $+a$ y B los valores $\neg b$ y $+b$, suponiendo en ambos casos una ordenación de los estados de forma que el negativo "es menor" que el positivo.
En este caso, el problema se reduce a demostrar que

$$P(+b|\neg a) < P(+b|+a) \Rightarrow P(+a|\neg b) < P(+a|+b).$$

Aplicando el teorema de Bayes se tiene que

$$P(+a|\neg b) = \frac{P(\neg b|+a)P(+a)}{P(\neg b)} \tag{3.5}$$

y

$$P(+a|+b) = \frac{P(+b|+a)P(+a)}{P(+b)} \tag{3.6}$$

Dividiendo (3.5) entre (3.6) se tiene

$$\frac{P(+a|\neg b)}{P(+a|+b)} = \frac{\dfrac{P(\neg b|+a)P(+a)}{P(\neg b)}}{\dfrac{P(+b|+a)P(+a)}{P(+b)}} \tag{3.7}$$

Como

$$P(b) = \sum_a P(b|a),$$

la expresión 3.7 es equivalente a

$$\frac{\frac{P(\neg b|+a)}{P(\neg b|\neg a)P(\neg a)+P(\neg b|+a)P(+a)}}{\frac{P(+b|+a)}{P(+b|\neg a)P(\neg a)+P(+b|+a)P(+a)}} = \frac{\frac{P(+b|\neg a)P(\neg a)+P(+b|+a)P(+a)}{P(+b|+a)}}{\frac{P(\neg b|\neg a)P(\neg a)+P(\neg b|+a)P(+a)}{P(\neg b|+a)}}$$

Simplificando tanto el numerador como el denominador tenemos

$$\frac{P(+a)+P(\neg a)\frac{P(+b|\neg a)}{P(+b|+a)}}{P(+a)+P(\neg a)\frac{P(\neg b|\neg a)}{P(\neg b|+a)}} \tag{3.8}$$

Como la influencia de A a B es positiva, eso significa que

$$P(+b|\neg a) \leq P(+b|+a) \tag{3.9}$$

Como

$$P(\neg b|\neg a) = 1 - P(+b|\neg a) \text{ y } P(\neg b|+a) = 1 - P(+b|+a),$$

podemos deducir que

$$P(\neg b|\neg a) \geq P(\neg b|+a) \tag{3.10}$$

Entonces, (3.8) es equivalente a

$$\frac{P(+a)+P(\neg a)X}{P(+a)+P(\neg a)Y} \tag{3.11}$$

Como $X \leq 1$ (por (3.9)) e $Y \geq 1$ (por (3.10)), podemos deducir entonces que

$$\frac{P(+a|\neg b)}{P(+a|+b)} \leq 1 \tag{3.12}$$

lo que implica que

$$P(+a|\neg b) \leq P(+a|+b) \tag{3.13}$$

c.q.d.

Caso no binario. Realizada por Druzdzel [50], se basa en aplicar el teorema de Bayes a la propiedad 3.2.

Supongamos $a_i < a_j$ y $b_k < b_l$

$$\frac{P(a_j|b_l,x)}{P(a_j|b_k,x)} = \frac{\frac{P(b_l|a_j,x)P(a_j)}{P(b_l)}}{\frac{P(b_k|a_j,x)P(a_j)}{P(b_k)}} = \frac{\frac{P(b_l|a_j,x)}{P(b_l)}}{\frac{P(b_k|a_j,x)}{P(b_k)}} \tag{3.14}$$

$$\frac{P(a_i|b_l,x)}{P(a_i|b_k,x)} = \frac{\dfrac{P(b_l|a_i,x)P(a_i)}{P(b_l)}}{\dfrac{P(b_k|a_i,x)P(a_i)}{P(b_k)}} = \frac{\dfrac{P(b_l|a_i,x)}{P(b_l)}}{\dfrac{P(b_k|a_i,x)}{P(b_k)}} \tag{3.15}$$

Según el corolario 3.2, se tiene que

$$\frac{P(a_j|b_l,x)}{P(a_j|b_k,x)} \geq \frac{P(a_i|b_l,x)}{P(a_i|b_k,x)}$$

lo que equivale a:

$$\frac{\dfrac{P(b_l|a_j,x)}{P(b_l)}}{\dfrac{P(b_k|a_j,x)}{P(b_k)}} \geq \frac{\dfrac{P(b_l|a_i,x)}{P(b_l)}}{\dfrac{P(b_k|a_i,x)}{P(b_k)}} \tag{3.16}$$

Como $P(b_k) > 0$ y $P(b_l) > 0$, la expresión anterior se puede simplificar, obteniendo

$$\frac{P(b_l|a_j,x)}{P(b_k|a_j,x)} \geq \frac{P(b_l|a_i,x)}{P(b_k|a_i,x)} \tag{3.17}$$

que es equivalente, por la propiedad 3.2, a la definición de influencia positiva de B hacia A. □

Otra forma equivalente de definir los distintos tipos de influencia se basa en la diferencia de distribuciones para el caso de variables discretas.

Definición 3.2 *Sea* $\Delta_k = Acum(c_k|v) - Acum(c_k|v')$. *Definimos la* diferencia de distribuciones *de un nodo de la siguiente forma:*

$$Dist(C|v) - Dist(C|v') = \begin{cases} \max\limits_k \Delta_k & si\ \forall k, \Delta_k \geq 0\ \ y\ \exists l/\Delta_l > 0 \\ \min\limits_k \Delta_k & si\ \forall k, \Delta_k \leq 0\ \ y\ \exists l/\Delta_l < 0 \\ 0 & si\ \forall k, \Delta_k = 0 \end{cases}, \tag{3.18}$$

Para variables binarias, con valores $+c$ y $\neg c$, por ejemplo, se tiene que

$$Dist(C|v) - Dist(C|v') = P(+c|v) - P(+c|v')$$

A partir de la diferencia de distribuciones, se puede establecer un orden parcial entre las mismas, de la forma:

$$Dist(C|v) > Dist(C|v') \Leftrightarrow Dist(C|v) - Dist(C|v') > 0 \tag{3.19}$$

Por tanto, la definición 3.1 se puede reescribir así:

$$\forall b, \forall a_i, \forall a_j, a_i < a_j \Rightarrow Dist(C|a_i, b) < Dist(C|a_j, b) \tag{3.20}$$

La ecuación anterior refleja de modo más intuitivo el significado de una influencia positiva: a mayores valores del nodo padre, la probabilidad de que el hijo tome valores mayores aumenta. Si disminuye, sería una influencia negativa.

Un ejemplo en el que se reflejan influencias positivas y negativas vendría dado por la red bayesiana de la figura 3.2, en la que los tres nodos son ordinales.

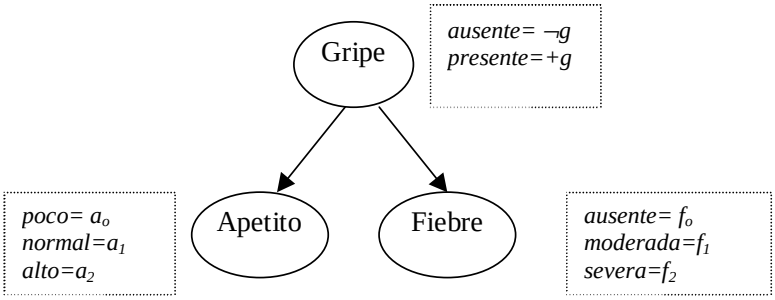

Figura 3.2: Ejemplo de red bayesiana con influencias positivas y negativas

Las probabilidades asociadas a los nodos de la red bayesiana del ejemplo son las siguientes:

$P(g)$
$P(\neg g) = 0{,}8$
$P(+g) = 0{,}2$

$P(f_i\|g)$	$\neg g$	$+g$
f_2	0.2	0.3
f_1	0.3	0.6
f_0	0.5	0.1

$P(a_i\|g)$	$\neg g$	$+g$
a_2	0.3	0.1
a_1	0.5	0.3
a_0	0.2	0.6

Según la representación de la red, el conocimiento que tenemos del dominio nos hace pensar que la influencia Gripe-Fiebre es positiva, es decir, que si una persona tiene gripe la probabilidad de que tenga fiebre será mayor que cuando no tiene gripe. La misma idea es la que nos hace pensar que la influencia Gripe-Apetito es negativa. Veremos si efectivamente se verifican ambas propiedades según las probabilidades asignadas.

Para la influencia Gripe-Fiebre:

$$Dist(Fiebre|\neg g) = \begin{pmatrix} 0{,}2 \\ 0{,}5 \\ 1 \end{pmatrix} < \begin{pmatrix} 0{,}3 \\ 0{,}9 \\ 1 \end{pmatrix} = Dist(Fiebre| + g)$$

Como $\neg g < +g$, podemos deducir que la influencia es positiva. En el caso de la influencia Gripe-Apetito, vemos que la influencia es negativa, puesto que:

$$Dist(Apetito|\neg g) = \begin{pmatrix} 0{,}3 \\ 0{,}8 \\ 1 \end{pmatrix} > \begin{pmatrix} 0{,}1 \\ 0{,}4 \\ 1 \end{pmatrix} = Dist(Apetito| + g)$$

Este tipo de explicación ha demostrado ser de mucha utilidad tanto para construir la red como al hacer inferencia. En el primer caso, porque ayuda a ver la información cuantitativa de forma cualitativa, lo que suele ser preferible por el usuario, puesto que es más fácil de entender [49, 63, 62, 182]. Esto permite analizar y corregir las probabilidades condicionadas asignadas a cada nodo, ya que la gente normalmente comete muchos errores al hacer estimaciones de probabilidades [51, 93]. Además, puede ayudar a estudiar la interacción entre variables.

Factor de influencia. Análogamente al caso de los nodos, para los enlaces también existe un modo de medir la influencia que se transmite a través de él. Para ello, se define el *factor de influencia del enlace entre los nodos A y B*, que representamos mediante $FI(AB)$ como el valor que representa la influencia que se transmite a través del enlace desde un nodo hacia su hijo y se calcula automáticamente dependiendo únicamente de las tablas de probabilidad del mismo. Para ello, sea

$$\Delta_k = \max_i Acum(b_k|a_i) - Acum(b_k|a_0) \tag{3.21}$$

Definición 3.3 *El factor de influencia del enlace cuyo origen es el nodo A y su destino es el nodo B se define como*

$$FI(AB) = \max_k \Delta_k \tag{3.22}$$

Lo que se intenta reflejar con este factor es la importancia explicativa de un nodo con respecto a su hijo, midiendo cuánto hace cambiar la probabilidad de todos los estados de un nodo cuando su padre pasa del estado "normal" a otro estado "no normal", cualquiera que sea la configuración de valores de sus padres.

En el caso en que ambos nodos sean binarios, con estados, por ejemplo, $+a$ y $\neg a$; $+b$ y $\neg b$, respectivamente, donde los valores *normales* son $\neg a$ y $\neg b$, la fórmula anterior es equivalente a:

$$FI(AB) = P(+b| + a) - P(+b|\neg a)$$

La ventaja que tiene esta medida frente a otras, como por ejemplo la entropía cruzada, es que es independiente de la población, lo que implica que no se ve alterada por la

probabilidad a priori del origen o el destino del enlace, lo que en ocasiones puede conducir a sesgos en la obtención de la medida de la influencia que un nodo (o varios) ejerce sobre otro.

Teorema 3.3 *$FI(AB) > (<)0$ si y sólo si existe una correlación positiva (negativa) entre las variables A y B. Del mismo modo, FI(AB)=0 si y sólo si no existe correlación.*

Demostración

Es trivial por la definición de correlación (véase la Eq. 3.3). Para la demostración de las correlaciones negativa y nula sería similar; simplemente bastaría con cambiar la desigualdad. □

Este teorema permite interpretar de modo intuitivo el significado del signo del factor de influencia, puesto que refleja la idea de que si la influencia es positiva, lógicamente debe existir una correlación positiva entre ambos nodos.

La utilidad de esta herramienta es la posibilidad de que el usuario identifique fácilmente cuáles son los enlaces que más influencia ejercen sobre un nodo que tiene varios padres, lo que le puede ayudar a predecir ciertos valores. Este tipo de facilidad es muy interesante de cara a sistemas orientados a la educación del usuario. Por otra parte, resulta muy provechoso también para complementar las explicaciones verbales de los modelos canónicos [43].

Verbal

En este caso, después de que el usuario seleccione el enlace del que quiere obtener la explicación, se le muestra la siguiente información:

- *tipo de relación;* en este caso, se pueden clasificar los distintos tipos de relaciones, tal y como proponen Díez [45] y Druzdzel [50]. Así, tendríamos ES_UN_TIPO_DE, PRODUCE, FAVORECE, CAUSA, SE_DIAGNOSTICA_CON, etc.;
- *origen,* mediante el nombre del nodo del que parte el enlace;
- *destino,* reflejado por el nombre del nodo al que llega el enlace;
- *razones de verosimilitud* para cada estado del nodo destino, que sirven para conocer qué estado del nodo padre explica mejor cada uno de los estados del nodo hijo. Además, se presentan ordenadamente de mayor a menor, junto con las proporciones correspondientes, lo que permite al usuario localizar fácilmente cuál es el estado del

nodo padre que mejor explica cada uno de los estados del nodo hijo y en qué proporción con respecto al resto.

- *Acceso a las propiedades del enlace,* lo que permite conocer otras características del mismo, como por ejemplo, si es dirigido o no, etc.

3.2.3. Explicación de la red

Al igual que para los nodos y los enlaces, la red se puede explicar de forma textual y de forma gráfica.

Verbal

En este caso, la explicación de la red está orientada a mostrar los nodos más importantes de cara al diagnóstico, que serían los que desempeñan el papel de "Enfermedad/Anomalía". Por eso, para cada uno de ellos se genera un texto con todos los datos asociados: otras posibles enfermedades o anomalías que lo provocan, factores de riesgo que contribuyen, síntomas y signos y pruebas que ayudan a confirmar o descartar la presencia de la enfermedad o anomalía correspondiente. Finalmente, se indica también el papel auxiliar que juegan el resto de los nodos.

Para ello, se utiliza la siguiente plantilla:

La red [nombre_red] está diseñada con el objetivo de realizar un diagnóstico sobre la/s siguiente/s variables:

[Enfermedad/anomalía_1]: Las enfermedades / anomalías que lo provocan son: [lista de enfermedades / anomalías]. Los factores de riesgo que contribuyen son: [lista de factores]. Se identifica por los siguientes síntomas: [lista de síntomas] y signos:[lista de signos]. Existen las siguientes pruebas para su diagnóstico [lista de pruebas]. Además, se han representado los siguientes datos auxiliares: [lista de nodos auxiliares y otros].

Esta plantilla se repite para todos los nodos cuyo papel es el de *Enfermedad/Anomalía.*

Gráfica

La explicación gráfica de la red consistirá en mostrar la información gráfica asociada a cada nodo y a cada enlace, según lo comentado en las secciones anteriores. No obstante, puesto que puede ser que la red sea muy grande, se proporcionará al usuario la posibilidad de centrarse únicamente en la porción de red que desee, lo que implica mostrar sólo la explicación de los enlaces o sólo la de los nodos. Incluso, se permitirá que el usuario

pueda seleccionar aquellos nodos en los que esté interesado en obtener su explicación gráfica, lo que puede hacerse por distintos criterios: importancia, función que desempeña, una combinación de ambos o ninguno de ellos, seleccionándolos directamente el propio usuario.

3.3. Valoración del método propuesto

A la luz de la descripción del método desarrollado y teniendo en cuenta las propiedades de la tabla 2.1, podemos destacar como principales aportaciones las siguientes:

En cuanto al *contenido* de la explicación, éstas se ofrecen tanto a **nivel micro** como a **nivel macro**, para redes bayesianas **causales**. No obstante, algunas de las facilidades que proporciona el método diseñado también pueden servir para redes **no causales**, por lo que incidiremos en ello durante su descripción. Las explicaciones a nivel macro ofrecen una visión global del modelo, lo que puede servir para confirmar a grandes rasgos si la representación del dominio es la correcta y para ayudar al usuario a profundizar en las relaciones definidas. Por el contrario, las explicaciones a nivel micro permiten profundizar con cierto nivel de detalle en el dominio representado, facilitando así la detección y corrección de los posibles errores que se hubieran podido producir durante la fase de construcción de la red. El método se ha diseñado con el propósito tanto de **describir** los distintos estados que caracterizan el modelo, indicado para aquellos usuarios con experiencia en redes bayesianas, como de hacer **comprender** al usuario las relaciones y propiedades representadas, pensado fundamentalmente para los usuarios sin conocimientos sobre teoría de la probabilidad o de redes bayesianas.

Con respecto a la *comunicación* entre el usuario y el sistema, la interacción entre ambos se lleva a cabo mediante la utilización de **menús y ventanas**, ya que en la actualidad constituye la forma más común de comunicación entre cualquier usuario, sobre todo los no expertos en informática, y el ordenador. La principal ventaja que ofrece este tipo de interacción es que es adecuada tanto para los usuarios expertos como para los no expertos.

Finalmente, en cuanto a la *adaptación* al usuario, el método desarrollado no tiene en cuenta ningún **modelo**. Lo que sí está contemplado es la capacidad de adaptación del **nivel de detalle**, dependiendo de los requisitos y necesidades de cada usuario, mediante la modificación de los valores de su *factor de importancia* y de la *función* que desempeña el nodo en la red. Por último, en cuanto a la forma de presentar las explicaciones el método propuesto es capaz de ofrecerlas no solo de modo **gráfico** sino también de modo **textual**, por las razones comentadas en la sección 3.1. Las probabilidades se expresan de

forma **numérica**, ya que es la única forma de no caer en interpretaciones subjetivas de las mismas, aunque se complementan con expresiones **lingüísticas** que ayudan al usuario a obtener información cualitativa de los datos presentados.

Sin embargo, también somos conscientes de las limitaciones de nuestro método. Las principales están relacionadas con la imposibilidad de adaptar las explicaciones al nivel de conocimiento del usuario, y con la "rigidez" de los textos generados al solicitar la explicación verbal de la red completa. Esto se debe a que al no poder modificar el nivel de detalle de las mismas, las explicaciones contendrán absolutamente toda la información que el usuario haya definido para cada dato, lo que puede generar mucha confusión al usuario en redes grandes. En la sección 7.2 proponemos distintas líneas de trabajo para el futuro con el objetivo de corregir estas carencias.

Capítulo 4

Explicación del razonamiento

En este capítulo presentamos los mecanismos teóricos propuestos para explicar el razonamiento seguido en una red bayesiana al propagar determinada evidencia. El método consta de varias facilidades que el usuario podrá utilizar dependiendo de sus necesidades y preferencias. Así, después de justificar la necesidad de ofrecer este tipo de explicación (sección 4.1), exponemos cada una de las herramientas de las que consta: la visualización gráfica de los resultados obtenidos al hacer inferencia (sección 4.2.1) y la posibilidad de gestionar simultáneamente distintos casos de evidencia (sección 4.2.2). Además, el usuario podrá solicitar al sistema la explicación de un caso de evidencia determinado (secciones 4.2.3 y 4.2.4). Por último, analizamos las principales propiedades del método desarrollado para generar explicaciones del razonamiento, así como sus limitaciones (sección 4.3).

4.1. Introducción

Hasta ahora hemos detallado las formas de explicación **estática** de una red bayesiana, que consiste en ofrecer distintos tipos de explicación sobre el modelo representado. Sin embargo, el "caballo de batalla" de la explicación de redes bayesianas radica en convencer al usuario de los resultados que el sistema produce cuando se realiza la propagación de determinada evidencia sobre la red, es decir, la explicación **dinámica**, puesto que hay determinados tipos de usuarios, como por ejemplo los médicos, que son reacios a aceptar los consejos de un ordenador si no conocen el proceso por el que han sido obtenidos [177]. Puesto que los resultados que ofrecen los sistemas basados en redes bayesianas se obtienen aplicando algún algoritmo de inferencia (véase sección 1.2.1), una posible solución podría ser mostrar de alguna forma (gráfica o textual) la traza de la ejecución de dicho algoritmo, teniendo en cuenta la evidencia y las variables de interés para cada caso concreto. Por tanto, en cada caso habría que saber cuál es el algoritmo que se ha utilizado, lo que en numerosas ocasiones no es factible, pues forma parte de la representación interna del

sistema.

No obstante, este tipo de explicación únicamente sería válida para aquellos usuarios especializados en este campo y además, sólo sería más o menos intuitivo en los casos en los que la red tuviera una cantidad de nodos y enlaces lo suficientemente pequeña como para que fuera sencillo visualizar los algoritmos de propagación, de forma similar a como se explican en los libros en los que se proponen. En otro caso, al usuario no le queda más remedio que confiar de modo más o menos ciego en el correcto funcionamiento del algoritmo aplicado.

Por otra parte, aun en el mejor de los casos en que el usuario conozca y entienda perfectamente el algoritmo de propagación utilizado y que la red tenga una cantidad de nodos y enlaces suficientemente "pequeño", para la mente humana no es sencillo realizar los cálculos de las probabilidades, con lo que es fácil que los resultados calculados mentalmente de forma "heurística" no coincidan con los ofrecidos por el sistema.

Con todo, aunque fuera lo suficientemente sencilla como para poder realizar cálculos mentalmente, podría darse la situación de que el sistema ofreciera un resultado y el usuario esperase otro, debido fundamentalmente a las siguientes causas:

- el usuario no es capaz de analizar el efecto conjunto de los hallazgos que forman la evidencia, bien porque son numerosos o bien porque existen conflictos entre los hallazgos [174];
- el modelo no está representado correctamente o los algoritmos no están bien implementados;
- las probabilidades no se han asignado correctamente, sobre todo en los casos en los que se tiene que hacer de forma subjetiva.

Por otra parte, en el caso de la justificación de los resultados, es interesante que el usuario pueda conocer también cuál es la influencia que ejerce cada hallazgo sobre la variable de interés; o qué variables se han visto afectadas por un hallazgo o conjunto de hallazgos, qué variables han aumentado su probabilidad, cuáles han disminuido, etc.

Por todas estas razones es conveniente dotar a los sistemas basados en redes bayesianas de algún tipo de explicación del proceso de razonamiento seguido en la red, con el objetivo de aportar conocimiento al usuario sobre el modo de obtener los resultados proporcionados y de su significado.

Las principales ventajas que proporciona la explicación del razonamiento son:

- ayuda a detectar posibles conflictos entre los hallazgos que forman la evidencia;

- se dispone de una herramienta para detectar los errores cometidos en la representación del modelo;

- se tiene un método eficaz para convencer al usuario de los resultados obtenidos e, incluso, de por qué son distintos a los que él esperaba;

- es posible mejorar el conocimiento del usuario, tanto del modelo como del razonamiento;

- además, ayuda al usuario a entender los resultados que se hubieran obtenido en caso de que los hallazgos hubieran sido distintos

4.2. Descripción

Después de analizar las características citadas en la sección anterior se puede deducir que para poder ofrecer todas las opciones que en él se proponen se necesitan varias herramientas que se encarguen de realizar distintas tareas de forma independiente del resto. Por ejemplo, no se puede ofrecer a la vez una explicación a nivel micro y a nivel macro; tampoco se puede generar la explicación de los resultados a la vez que la explicación del razonamiento hipotético; etc. Así pues, en las siguientes secciones se detallan cada una de las herramientas que conforman el método propuesto para la explicación del razonamiento.

4.2.1. Presentación gráfica de las probabilidades

Ya hemos indicado previamente que para el usuario la forma más intuitiva de entender los resultados es visualizarlos gráficamente. Por eso, proponemos mostrar gráficamente las probabilidades de los nodos de la red después de una propagación. En concreto, de cada nodo se mostrarán las probabilidades de todos sus posibles estados, resaltando de algún modo el más probable, con el objetivo de que el usuario pueda identificarlo fácilmente y analizar la diferencia con respecto al resto de los estados. Puesto que la red puede tener una gran cantidad de nodos, la visualización de las probabilidades se produce sólo en los nodos seleccionados por el usuario, bien por la función que desempeñan, bien por tener un factor de importancia mayor que un umbral dado o bien porque el estado más probable sea igual o no al valor "normal".

4.2.2. Casos de evidencia

Después de analizar tanto los métodos de explicación desarrollados hasta la fecha como las aplicaciones (comerciales o gratuitas) para editar y procesar redes bayesianas, nos dimos cuenta de que una de las grandes carencias comunes a todos ellos era la imposibilidad de analizar los resultados de distintas propagaciones, puesto que cuando se visualizan gráficamente los datos de una propagación, los de la anterior se pierden. Esto impide comparar los resultados con otros ya calculados, salvo que el usuario sea capaz de retener los datos que le interesen. Pero en la práctica, resulta muy complicado memorizar más de 3 ó 4 valores para uno o dos nodos. Por eso, hemos intentado solventar esta deficiencia y una de las principales aportaciones del método propuesto radica en la posibilidad de gestionar simultáneamente los resultados de diferentes propagaciones. Esto es posible gracias a la definición y manipulación de distintos *casos de evidencia.*

Definición 4.1 *Dada una red bayesiana B y cierta evidencia e se define un* Caso de Evidencia (CE) *como un par (E, P) donde E está determinado por el conjunto de hallazgos de la evidencia $\boldsymbol{e}$ y P es el conjunto de probabilidades asociadas a los nodos no observados.*

$$CE(B|\boldsymbol{e}) = (\{h \in \boldsymbol{E}\}, \{P(y|\boldsymbol{e})|y \in nodos(B) \backslash \boldsymbol{E}\}) \tag{4.1}$$

Existe un caso especial, llamado *caso a priori*, que corresponde a la ausencia de evidencia y cuyas probabilidades asociadas son las probabilidades a priori de cada nodo de la red:

$$CE_0(B) = (\emptyset, \{P(y)|y \in nodos(B)\}) \tag{4.2}$$

Por tanto, un caso de evidencia no se define sólo por la evidencia introducida sino que tiene también en cuenta el estado de la red después de la propagación de la misma.

La creación y manipulación de casos de evidencia tiene como principal objetivo facilitar la generación de explicaciones gráficas asociadas a la propagación de dicha evidencia sobre una red bayesiana.

Explicación gráfica

Una de las principales aportaciones del método que proponemos consiste en la visualización gráfica de cada caso de evidencia, tanto de forma individual como en relación con otros. Puesto que el *caso a priori* se corresponde con el estado inicial de la red, la explicación de dicho caso se realizará con el método diseñado que el visto para la explicación estática (sección 3.2.3). Para los otros casos distintos del *caso a priori*, en los

que la evidencia no es vacía, proponemos la posibilidad de mostrar de modo gráfico, y *simultáneamente*, los resultados de tantas propagaciones como el usuario desee. De esta forma, se pueden conocer y analizar los cambios producidos en las probabilidades de los nodos dependiendo de la evidencia introducida.

Entre las principales utilidades de esta herramienta merece la pena destacar:

- se puede analizar de un modo sencillo los resultados asociados a distintos casos;
- permite realizar análisis de sensibilidad entre los distintos hallazgos, pues se puede asociar un caso distinto por cada hallazgo o por cada conjunto de hallazgos, y se tiene así una forma sencilla y rápida de analizar el efecto de cada uno de ellos sobre la o las variables que sean de interés para el usuario;
- sirve también como herramienta para realizar razonamiento hipotético, mediante la generación y propagación de un nuevo caso para luego comparar los resultados con el resto de casos. Así el usuario puede predecir el efecto que produce observar cada uno de los posibles valores de una variable determinada.
- Además, dada cierta evidencia, es posible calcular el valor diagnóstico de cada uno de sus componentes mediante la asignación de diferentes casos para cada hallazgo o conjunto de hallazgos, tanto individualmente como de modo incremental (primero introduciendo un hallazgo, luego otro, etc.). Esta posibilidad resulta muy útil para los sistemas médicos, pues sirve tanto para asociar un caso por paciente, como para analizar el efecto que producen determinados hallazgos frente a otros;
- por último, proporciona un mecanismo que ayuda al usuario a detectar posibles conflictos entre los datos.

4.2.3. Análisis de sensibilidad

El análisis de sensibilidad (AS) consiste en estudiar cómo la variación en los resultados de un modelo (matemático o computacional) puede depender, cualitativa o cuantitativamente, de las variaciones sufridas por distintas fuentes de entrada [4]. Su principal objetivo, por tanto, consiste en determinar cómo dependen los modelos de la información representada en ellos, de su estructura y de las suposiciones hechas al construirlo. Puesto que la representación o definición del modelo lleva asociada normalmente cierta incertidumbre, el AS está muy relacionado con su análisis. De hecho, el AS surgió como método para procesar la incertidumbre asociada a las entradas y a los parámetros del

modelo, aunque a lo largo del tiempo estas ideas se han generalizado hasta incorporarla también a la estructura del modelo.

En general, el AS se utiliza para aumentar la confianza en el modelo y en sus resultados pues ayuda a comprender cómo responden éstos a los cambios en los datos de entrada. Por estos motivos, aunque en el contexto de las redes bayesianas se ha utilizado para facilitar el proceso de obtención de las probabilidades [16, 30], una de sus principales aplicaciones es su utilización como herramienta para la generación de explicaciones [46, 47, 79, 103, 108, 174], siguiendo las mismas ideas utilizadas para explicar el razonamiento en otros tipos de sistemas expertos [167, 170, 171].

Como indica Jensen [90, cap. 6], una parte de la explicación es el **análisis de la sensibilidad a la evidencia**, que consiste en analizar el efecto de los hallazgos que forman la evidencia sobre una o varias variables de interés con el objetivo de:

- identificar los hallazgos que afectan (positiva o negativamente) o no a la probabilidad de una hipótesis determinada, o
- conocer qué evidencia sirve para discriminar una hipótesis de otra.

La idea desarrollada en nuestro método consiste en mostrar al usuario los cambios que producen los hallazgos asociados a un caso concreto, tanto individualmente como en conjunto, sobre una variable determinada o sobre la red entera.

Objeto del análisis

Existen numerosos métodos para realizar análisis de sensibilidad, cada uno con ciertas ventajas e inconvenientes. La elección de uno determinado depende de numerosos factores: las propiedades del modelo, el número de factores involucrados en el análisis, la complejidad computacional y, naturalmente, el objeto del análisis. En el caso de las redes bayesianas y la explicación del razonamiento dicho objeto puede ser una variable particular o la red completa. Por tanto, dependiendo de cada caso, el método contempla distintas opciones:

Análisis sobre la red entera. Cuando el usuario desee conocer cómo afectan determinados hallazgos al conjunto de la red, simplemente debe crear un nuevo caso que contenga dichos hallazgos. Al propagar dicho caso, se obtiene la representación gráfica de los nuevos valores de probabilidad de los nodos seleccionados por el usuario. A pesar de que la visualización gráfica de las probabilidades es una herramienta muy sencilla e intuitiva para cualquier tipo de usuario, independientemente de su nivel de

conocimiento sobre probabilidades o sobre el dominio representado, puede llegar a ser algo confusa en los casos en los que la red tiene muchos nodos, cuando los nodos tienen muchos estados o cuando se han hecho varias propagaciones de evidencia. Por tanto, se hace necesario definir una herramienta de explicación mediante la cual el usuario tenga la posibilidad de identificar fácilmente los nodos cuya probabilidad ha variado al hacer la propagación. Para ello, se propone la posibilidad de representar gráficamente los nodos en función del cambio sufrido en su probabilidad al propagar cierta evidencia e, dependiendo de si ha aumentado, disminuido o se ha mantenido igual. El criterio elegido para realizar la comparación se basa nuevamente, como ya ocurriera para la explicación del modelo (véase sección 3.2.2), en considerar que todos los nodos tienen un valor "normal", que generalmente coincide con el valor "ausente", "negativo", etc. de cada nodo. Como medida para determinar el cambio de probabilidad se propone la distribución acumulada, de acuerdo con la siguiente definición (compárese con la definición 3.2):

Definición 4.2 *Sean dos distribuciones de probabilidad, P1 y P2, de una variable ordinal V con estados $v_1, ..., v_s$. Considerando*

$$\forall i \in \{1,2\}, \forall j \in \{1..s\}, Acum_i(v_k = \sum_{j>k} P_i(v_j)$$

definimos un orden entre P1 y P2 del siguiente modo:

- *P1 es* ***mayor (menor)*** *que P2 si*

$$\begin{gathered} Acum_1(v_k) \geq (\leq) Acum_2(v_k) \forall k \in \{1..s\} \\ y \\ \exists l \in \{1..s\} / Acum_1(v_l) > (<) Acum_2(v_l). \end{gathered} \tag{4.3}$$

- *P1 es* ***igual*** *que P2 si*

$$Acum_1(v_k) = Acum_2(v_k) \forall k \in \{1..s\}. \tag{4.4}$$

- *Cuando no se cumpla ninguna de las situaciones anteriores diremos que el orden entre las distribuciones es* ***desconocida***.

Lo que intentamos reflejar es que si las probabilidades de todos los estados coinciden, ambas distribuciones son iguales. En caso contrario, comparamos los valores de las distribuciones acumuladas.

Definición 4.3 *Diremos que la evidencia e del caso C* tiene influencia sobre la variable *V si la diferencia entre las distribuciones de probabilidad de V, antes y después de propagar la evidencia* $\boldsymbol{e}$*, es mayor que* θ*, es decir, si*

$$\forall k \in \{1..s\}, |Acum(v_k|\boldsymbol{e}) - Acum(v_k)| > \theta \tag{4.5}$$

El valor de θ lo asignará el usuario de acuerdo a sus necesidades, aunque por defecto se definirá con el valor 10^{-2}.

Definición 4.4 *Si la evidencia e tiene influencia sobre una variable V, diremos que ésta es* ***positiva (negativa)*** *si*

$$\begin{gathered} Acum(v_k|\boldsymbol{e}) \geq (\leq) Acum(v_k) \forall k \in \{1..s\} \\ y \\ \exists l \in \{1..s\} / Acum(v_l) > (<) Acum(v_l|\boldsymbol{e}) \end{gathered} \tag{4.6}$$

De forma similar, si

$$\begin{gathered} \exists k \in \{1..s\} / Acum(v_k) < Acum(v_k|\boldsymbol{e}) \\ y \\ \exists l \in \{1..s\} / Acum(v_l) > Acum(v_l|\boldsymbol{e}) \end{gathered} \tag{4.7}$$

entonces se dice que la influencia es ***desconocida.***

En el caso de variables binarias, con valores $\neg v$ y $+v$ la ecuación 4.5 será equivalente a comprobar si $P(+v|e)$ es mayor, menor o igual que $P(+v)$. Así, si $P(+v|e) > P(+v)$, diremos que la probabilidad de V ha **aumentado.** Habrá **disminuido** si $P(+v|e) < P(+v)$ y será **igual** si $P(+v|e) = P(+v)$. En el caso de variables binarias la influencia no puede ser desconocida.

Lo que esta fórmula pretende reflejar es que si ha habido un cambio "notable" en las probabilidades de la variable, se analizará si los valores de dicha variable han aumentado, disminuido o se han mantenido (casi) iguales.

Las definiciones están inspiradas en los trabajos de Wellman sobre redes cualitativas [184]. Sin embargo, el hecho de que nuestras redes contengan probabilidades numéricas y que la propagación de la evidencia se haga mediante algoritmos cuantitativos permite determinar el signo de los cambios en la probabilidad en muchos casos en los que los algoritmos de Wellman obtendrían el signo "?".

Definición 4.5 *Si la evidencia e tiene influencia no desconocida sobre una variable V, definimos la* magnitud de influencia que ejerce la evidencia e sobre V *como*

$$MI_e(V) = \max_k |Acum(v_k|\boldsymbol{e}) - Acum(v_k)| \tag{4.8}$$

En el caso de que V sea una variable binaria con estados $\neg v$ y $+v$, la ecuación 4.8 será equivalente a calcular $|P(+v|e) - P(+v)|$. Por tanto, la idea subyacente en esta definición es tener una forma de analizar cuánto hace variar la evidencia e la probabilidad de que la variable V alcance su estado "anormal".

Esta medida la utilizaremos para generar explicaciones gráficas del razonamiento de tal forma que los nodos se representen de distinta forma dependiendo del impacto que la evidencia ejerce sobre ellos.

Análisis sobre una variable. En caso de que el usuario desee un análisis más detallado del efecto de cierta evidencia sobre una variable determinada, se le ofrece al usuario la posibilidad de analizar cómo ha variado la probabilidad de los estados de una variable determinada, también llamada *variable de interés* o *hipótesis*, con respecto a las probabilidades asociadas a un caso de evidencia CE_j determinado. Por defecto, j=0, lo que significa que el caso que se toma como referencia es el *caso a priori*. No obstante, nuestra propuesta permite además seleccionar el caso de evidencia, de entre los creados, con respecto al cual el usuario desea realizar este análisis. Para llevarlo a cabo, deberá crear un nuevo caso con la evidencia CE_n y seleccionar la variable V a analizar. Hecho esto, para cada estado v_k de la hipótesis, se mostrará su probabilidad acumulada en el caso de evidencia CE_j, la probabilidad acumulada de cada uno de ellos después de propagar el caso de evidencia CE_n y el valor que refleja el cambio de probabilidad sufrido en cada uno de los estados, mediante la razón de dichas probabilidades:

$$CP(v_k|\mathbf{e}) = \lg\left(\frac{Acum(v_k|\mathbf{e})}{Acum(v_k)}\right) \tag{4.9}$$

El conjunto de funciones candidatas a medir la sensibilidad de un nodo a distintas variaciones puede ser muy amplio [174], pero la razón por la que se ha utilizado ésta es que se entiende fácilmente por cualquier tipo de usuario y es muy intuitiva, por lo que se explica por sí misma. Hay que tener en cuenta que el método no tiene en cuenta distintos niveles de usuario, por lo que las medidas utilizadas deben ser lo suficientemente intuitivas para todo tipo de usuario.

Análisis de los hallazgos

Otra posibilidad que incluye el método propuesto consiste en determinar qué hallazgos influyen positiva o negativamente sobre una variable determinada v. Se puede hacer de dos formas:

Automáticamente. Para ello, basándonos en las ideas de Suermondt [174], el sistema analiza para cada hallazgo h la diferencia entre los valores de la distribución de una variable cuando se tiene en cuenta toda la evidencia propagada y cuando no se considera el hallazgo en cuestión, según la ecuación 4.3 que define la diferencia entre distribuciones.

Definición 4.6 *Un hallazgo h influye* ***positivamente (negativamente)*** *sobre la variable V si $Acum(V|e) < (>) Acum(V|e\backslash h)$. En el caso de que $Acum(V|e) = Acum(V|e\backslash h)$, diremos que el hallazgo h* ***no influye*** *sobre V. Si no se puede establecer un orden entre las distribuciones, entonces la influencia del hallazgo es* ***desconocida****.*

Con esta definición pretendemos analizar si el hallazgo hace aumentar (disminuir) la probabilidad de los estados "anormales" de la variable V. Se podría haber definido la influencia de los hallazgos de forma individual, analizando únicamente la diferencia de distribuciones $Dist(V) - Dist(V|h)$. Sin embargo, en ese caso perderíamos información porque no podríamos tener en cuenta la acción sinérgica del hallazgo seleccionado con el resto de los que forman parte de la evidencia. Por tanto, para analizar el impacto del hallazgo h se estudian las diferencias entre las distribuciones de las variables cuando h forma parte de la evidencia y cuando se elimina. Con el objetivo de mostrar toda esta información de modo comprensible para el usuario, se propone la clasificación de los hallazgos según el tipo y la cantidad de influencia que transmitan a la variable V, de acuerdo con la siguiente definición:

Definición 4.7 *Dado la evidencia e, la* ***magnitud de influencia*** *que ejerce un hallazgo h sobre una variable V se define como*

$$MI_h(V) = |Dist(V|\boldsymbol{e}) - Dist(V|\boldsymbol{e}\backslash h)| \tag{4.10}$$

La idea es identificar cuál es el cambio máximo que el hallazgo h puede llegar a producir sobre la probabilidad de la variable V, puesto que según la ecuación 3.18, es equivalente a

$$MI_h(V) = \max_k |Acum(v_k|\mathbf{e}) - Acum(v_k|\mathbf{e}\backslash h)| \tag{4.11}$$

De modo manual, en los casos en los que la evidencia e no contiene demasiados hallazgos y, además, la red bayesiana B es lo suficientemente "pequeña" como para que la explicación gráfica de los casos de evidencia sea cómoda. Para ello, el usuario debe crear un caso $CE(B|e)$ con la evidencia y otro caso $CE(B|e\backslash h)$ por cada hallazgo h, que contenga toda la evidencia e exceptuando a h. De este modo, si el usuario expande los nodos que corresponden a las variables en las que está interesado, puede observar y analizar los cambios que se producen en las probabilidades de los estados de cada variable, y así deducir la influencia que ejerce cada hallazgo sobre cada una de ellas. La ventaja de esta opción es que permite analizar gráficamente los cambios que produce sobre cada variable de interés la introducción de cada hallazgo por separado.

4.2.4. Caminos de razonamiento

Según indica Suermondt en su tesis doctoral [174], para explicar a un experto en el dominio representado los resultados obtenidos bastará con identificar la evidencia "relevante" para una hipótesis. En la misma línea, Johnson [92] afirma que cuando los médicos se encuentran con una serie de hallazgos clínicos, no razonan explícitamente sobre los caminos que les llevan a obtener la probabilidad de la hipótesis, sino que directamente actualizan su creencia en las probabilidades de varios diagnósticos de un modo que no son capaces de expresar, aunque parece que sí aplican un proceso de razonamiento paso a paso cuando tienen que resolver problemas que son difíciles o con los que no están familiarizados. Por otra parte, según Dreyfus [48], solamente los usuarios noveles razonan explícitamente según las reglas mediante las que ciertas características afectan a las soluciones. Esta teoría ha sido confirmada en parte por estudios empíricos [65, 66]. Por tanto, para los usuarios expertos en el dominio, puede bastar con mostrar la influencia de cada hallazgo sobre la variable de interés; sin embargo, para los usuarios no expertos, esto no es suficiente, puesto que en muchas ocasiones la relación entre los hallazgos y la hipótesis no son nada intuitivos. Por ejemplo, puede darse el caso de que un hallazgo no habitual afecte a la hipótesis a través de alguna variable intermedia que no forme parte del modelo usual observado por el usuario. En ese caso, merecerá la pena mostrar los caminos entre los hallazgos y la evidencia. Según Druzdzel [50, pág. 153], en redes cualitativas sólo los nodos que forman parte de estos caminos son computacionalmente relevantes.

Por tanto, otra de las aportaciones de nuestro método consiste en mostrar de modo gráfico únicamente los nodos y los enlaces que forman parte de todos los caminos entre los hallazgos que forman la evidencia de un caso de evidencia determinado y la variable

seleccionada por el usuario. En el caso de los nodos, la idea consiste en representar, además, la variación sufrida en sus distribuciones de probabilidad con respecto a otro caso de evidencia, según la definición 4.5, por lo que se propone representarlos gráficamente dependiendo de dicha variación. De los enlaces se tendrá en cuenta el sentido y la magnitud de influencia que transmiten. Esto resulta muy útil cuando se utilizan las redes bayesianas con fines educativos, puesto que a la luz de toda la información recibida el usuario puede hacerse una idea del tipo de razonamiento llevado a cabo en la red.

4.3. Valoración del método propuesto

Las principales aportaciones de nuestra propuesta, de acuerdo con las propiedades reflejadas en la tabla 2.1, son las siguientes:

Con respecto al *contenido* de la explicación, se pretende ofrecer explicaciones del razonamiento tanto a **nivel micro** como a **nivel macro**. La razón por la que se decide ofrecer explicaciones a ambos niveles es que las explicaciones a nivel micro son útiles en aquellos casos en los que un nodo no tiene demasiados vecinos (padres e hijos) porque, en caso contrario, puede ser demasiado difícil para el usuario intentar analizar todas las influencias asociadas a un mismo nodo. No obstante, hay que tenerlas en cuenta ya que ofrecen un método de análisis detallado de las variaciones producidas en un nodo determinado. Sin embargo, en los casos en los que existen nodos en la red que tienen demasiados vecinos o en los que el usuario no necesita tanto detalle en la explicación, lo que interesa entonces es simplemente analizar a nivel macro cuáles han sido los principales caminos que se han seguido para transmitir influencia desde la evidencia hasta la hipótesis.

Al contrario que ocurre con la explicación del modelo, la causalidad no es un factor determinante en el método propuesto. En cuanto al objetivo, se pretende que el método sirva tanto para **describir** el razonamiento como para hacer **comprender** al usuario:

- cómo se han obtenido los resultados proporcionados por el sistema,
- por qué no se han obtenido los resultados que el usuario esperaba, y
- qué hubiera pasado si los hallazgos hubieran sido otros.

La *comunicación* entre el usuario y el sistema se llevará a cabo mediante **menús y ventanas**, puesto que es el modo usual, y además sirve tanto para los usuarios expertos como para los principiantes. Las explicaciones generadas se presentarán tanto de modo **gráfico** como de modo **textual**. Las probabilidades se expresarán de forma **numérica**

aunque se podrán complementar con expresiones **lingüísticas** que ayuden al usuario a obtener información cualitativa de los datos presentados.

En cuanto a la *adaptación* al usuario, no se tiene en cuenta ningún **modelo**. Al usuario se le presupone un conocimiento básico de teoría de probabilidades. Puesto que como ya hemos comentado en la introducción los algoritmos de propagación no son únicos y, además, requieren un conocimiento específico por parte del usuario sobre redes bayesianas, probabilidad, teoría de grafos, etc., el principal objetivo que nos planteamos es que la explicación del razonamiento sea **independiente del algoritmo de propagación**. Con esto pretendemos conseguir que la explicación sirva para distintos tipos de usuario, independientemente de su nivel de conocimiento sobre los métodos de razonamiento. Lo que sí está contemplado es la capacidad de modificar el **nivel de detalle**, dependiendo de los requisitos y necesidades de cada usuario, mediante la modificación del *factor de importancia* y de la *función* que desempeña cada nodo.

En cuanto a las limitaciones del método descrito una de las más importantes está relacionada, al igual que ocurre con la explicación del modelo (véase la sección 3.3), con la imposibilidad de adaptar la explicación del razonamiento a distintos tipos de usuario. Otro gran inconveniente es el la elevada complejidad computacional de los algoritmos de generación de explicaciones gráficas, ya que en la mayoría de los casos, crece exponencialmente con el número de nodos de la red.

Capítulo 5

Implementación en ELVIRA

En este capítulo vamos a introducir los principales aspectos de ELVIRA [26], un entorno para la construcción y el procesamiento de modelos gráficos probabilísticos (sección 5.1). Para ello, después de presentar sus principales objetivos, justificamos el paradigma de la Programación Orientada a Objetos (POO) basada en Tipos Abstractos de Datos (TAD) como metodología a utilizar para el desarrollo del programa. Esta sección, junto con los cuatro capítulos anteriores finalizarían el estudio del *nivel de conocimiento* de nuestro trabajo, tal y como expusimos en la sección 1.7 (véase la figura 1.4). A partir de este momento, nos dedicamos a describir el *nivel simbólico*, correspondiente a las estructuras de datos y algoritmos desarrollados en ELVIRA (sección 5.1.4), para después describir su entorno gráfico (sección 5.2), en el que se ha incluido el método de explicación que hemos descrito en los capítulos anteriores, y cuya implementación exponemos posteriormente (sección 5.3). Acabamos este capítulo realizando una comparación de ELVIRA con otras aplicaciones existentes en lo relativo a edición y evaluación de redes bayesianas (sección 5.4).

5.1. ELVIRA

El programa ELVIRA[1] surge como resultado de un proyecto de investigación (TIC-1997-1135-C04) desarrollado entre los años 1997 a 2000 por varias universidades españolas, con el objeto de crear un entorno de desarrollo para modelos gráficos probabilísticos cuyo código fuente estuviera disponible para que los investigadores del proyecto pudieran trabajar y experimentar nuevos métodos de propagación, aprendizaje y explicación. Como meta adicional, intentamos cubrir las deficiencias que el resto de herramientas poseen.

[1]Se puede obtener más información de todo lo relacionado con ELVIRA en la dirección de Internet `http://www.leo.ugr.es/~elvira`, a través de la que cualquier persona interesada puede descargar la última versión que exista del programa en ese momento.

Actualmente, se está trabajando en su mejora y depuración, como parte del proyecto concedido para el trienio 2001-2004 (TIC-2001-2973-C05), que se plantea como continuación del anterior. En él se pretenden desarrollar diversas aplicaciones (médicas, agrícolas, etc.) utilizando ELVIRA como herramienta.

5.1.1. Objetivos

Los principales objetivos que se plantearon inicialmente fueron los siguientes:

- debería ser independiente de la plataforma sobre la que se ejecutara;
- serviría para la edición y procesamiento de redes bayesianas y de diagramas de influencia;
- las tareas a realizar serían: edición, inferencia, abducción, aprendizaje y fusión;
- cada una de esas tareas se podría realizar con la ayuda de una interfaz gráfica o mediante órdenes, invocando los distintos métodos desde el propio sistema operativo;
- además, el usuario podría seleccionar el algoritmo que deseara utilizar para realizar cada tarea específica;
- la interfaz gráfica estaría basada en un sistema de ventanas y menús, confiriéndole un aspecto amigable y sencillo;
- las redes creadas se podrían almacenar y recuperar de dispositivos externos en distintos formatos, incluso a través de Internet;
- el usuario podría obtener distintos tipos de explicación, a nivel micro y a nivel macro, tanto del modelo como del razonamiento.

Teniendo esto en cuenta, la solución que se planteó al problema se ilustra en la figura 5.1.

La idea es proporcionar al usuario la posibilidad de realizar las principales tareas de procesamiento de redes bayesianas desde línea de comandos, para interaccionar directamente con ELVIRA invocando a los distintos métodos del conjunto de clases básicas, lo que permite analizar la complejidad de los métodos implementados de inferencia, aprendizaje, etc., sin tener en cuenta el consumo de tiempo relacionado con la presentación de los datos a través de una interfaz gráfica de usuario (IGU). Por el contrario, las ventajas que proporcionan este tipo de interfaces para la edición de redes, introducción de

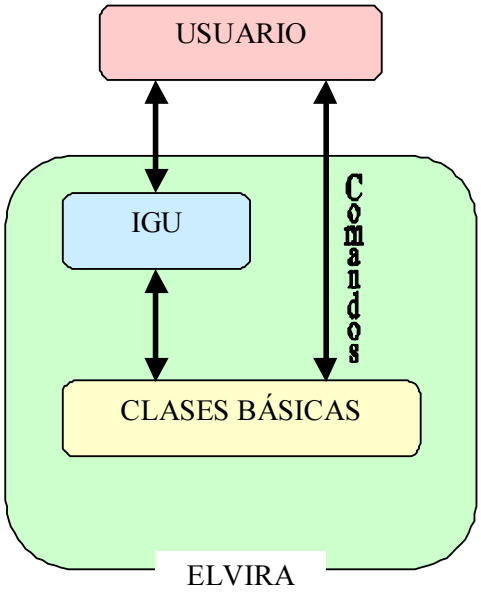

Figura 5.1: Componentes de ELVIRA

evidencia y el análisis de los resultados, especialmente si además cuentan con distintas opciones de explicación, nos hizo pensar en diseñar una IGU. A la aplicación integrada por el conjunto de clases de ELVIRA junto con la IGU se le denomina BayElvira 1.0. De momento, la mayoría de las funcionalidades proporcionadas por el conjunto de clases básicas de ELVIRA también se obtienen a través de su IGU, aunque muchas de ellas todavía no están integradas. Por otra parte, las facilidades que ofrece la IGU relacionadas con la edición y explicación de redes, no se pueden obtener cuando se interacciona con ELVIRA a través de la línea de órdenes del sistema operativo.

5.1.2. Metodología: Programación Orientada a Objetos

Un proyecto tan ambicioso como ELVIRA requería la elección de una buena metodología de programación, puesto que además iba a ser creado por distintos grupos de trabajo, repartidos físicamente por distintos puntos de la geografía española. Por ello, y dado el carácter abstracto de este nivel, se eligió la POO como paradigma de programación sobre el que se basara la creación de ELVIRA y se eligió JAVA como lenguaje de programación.

Paradigma de la Orientación a Objetos

El paradigma de la POO surge a finales de los años 60, como intento de aportar alguna solución a la denominada *crisis del software* [147], aunque empieza a adquirir relevancia a finales de los 80. Sin embargo, es durante la década de los 90 cuando alcanza su máxi-

mo apogeo, siendo actualmente el paradigma de moda, debido en gran medida al auge de lenguajes de POO como C++ y, especialmente, Java. Debido a la gran cantidad de aplicaciones en distintos ámbitos (lenguajes de programación, bases de datos, inteligencia artificial, sistemas operativos, interfaces de usuario, hardware ...), muchas universidades, tanto españolas como extranjeras, han incluido en sus planes de estudio como materia básica el estudio de la POO [1, 180].

Calidad del software. Teniendo en cuenta sus orígenes, el principal objetivo de la POO es el de producir *software* de calidad, cuyas principales propiedades se pueden clasificar en externas e internas [118]. Entre las primeras destacaremos:

- corrección, realizando con exactitud sus tareas, tal y como se definen en las especificaciones;
- robustez, es decir, la capacidad de responder apropiadamente ante situaciones excepcionales;
- extensibilidad, adaptándose con facilidad a los cambios de especificación;
- reutilización, lo que permite construir diferentes aplicaciones con los mismos elementos;
- compatibilidad, lo que permite combinar unos elementos *software* con otros;
- eficiencia, consumiendo la menor cantidad de recursos;
- portabilidad, para poder transferir los productos *software* a diferentes entornos y plataformas;
- facilidad de uso, para que personas con diferentes formaciones y aptitudes pueden aprender a usar los programas y aplicarlos a la resolución de problemas;
- facilidad de mantenimiento, pues se estima en un 70 % el coste de la dedicación al mantenimiento del *software*.

Fundamentos de la POO. Para construir *software* que cumpla todas estas propiedades se deben asegurar otros tipos de factores internos, que dependen, sobre todo, del paradigma de programación utilizado. En el caso de la POO, éstos se pueden resumir en cuatro conceptos clave:

Abstracción. Es el mecanismo que utiliza la mente humana para la comprensión de fenómenos o situaciones que involucren una gran cantidad de detalles. La abstracción permite estudiar fenómenos complejos siguiendo un método jerárquico, es decir, por sucesivos niveles de detalle, puesto que se basa en **destacar** los detalles relevantes del objeto en estudio, y **omitir** los detalles irrelevantes del objeto en ese nivel de abstracción, aunque podrían ser relevantes al descender de nivel.

Modularidad. Es el atributo que permite a un programa ser intelectualmente manejable, por lo que este concepto se lleva teniendo en cuenta desde los años 50. Supongamos que la solución de un problema P consiste en combinar las soluciones de los subproblemas $P1$ y $P2$. La idea que subyace bajo el concepto de modularidad es que $E(P) > E(P1) + E(P2)$, siendo $E(X)$ el esfuerzo que hay que realizar para resolver el problema X. Un módulo es una unidad independiente, y generalmente pequeña, que realiza completamente una (y sólo una) tarea y se comunica con otros elementos del sistema a través de parámetros. Para conseguir una modularidad efectiva, los módulos deben tener el máximo de independencia entre sí. No existe una técnica concreta para ello, pero hay que intentar asegurar tres características básicas:

- comprensibilidad, e.d., los módulos deben tener un significado propio para ser entendibles por sí mismos;
- continuidad modular, lo que significa que un pequeño cambio en las especificaciones implique cambios en la menor cantidad de módulos posible
- protección, para que una situación excepcional en un módulo sólo le afecte a él.

Ocultamiento de la información. Para conseguir las propiedades citadas, los módulos se han de caracterizar por decisiones de diseño que oculten los detalles internos de unos a los otros, de forma que la información contenida en cada módulo sea inaccesible a otros que no necesiten tal información. Precisamente, el ocultamiento de la información es lo que hace que el sistema mantenga su integridad durante la prueba y el mantenimiento del mismo, ya que evita la propagación de errores.

Reutilización de componentes de *software*, con el objetivo de conseguir distintos beneficios como son la disminución del esfuerzo de mantenimiento, fiabilidad, eficiencia, consistencia, etc. La forma apropiada de reutilización es alguna forma de módulo abstracto que encapsule algún tipo de funcionalidad mediante una interfaz bien definida. Para ello, existen dos técnicas básicas: la *sobrecarga*, que permite utilizar

un mismo nombre para más de una operación, y la *genericidad*, que proporciona la posibilidad de definir módulos parametrizados.

Elementos de la POO. La tecnología de la POO se basa en construir cada módulo sobre la base de algún tipo de **objeto**, entendido éste como la instancia de una **clase**, que es la que representa al conjunto de elementos que tienen el mismo comportamiento y características. Una clase, por tanto, contiene el estado de un objeto, definido mediante mediante *atributos*, que representan las características del objeto, y *métodos*, que representan las operaciones.

A pesar de que no existe una metodología precisa basada en OO para la construcción de *software*, sí que se puede obtener una buena definición de clases basando su diseño en los Tipos Abstractos de Datos. De hecho, una clase se puede considerar un TAD dotado de una implementación, posiblemente parcial [118].

Tipos Abstractos de Datos

El concepto de tipo abstracto de datos y la demostración de su utilidad para el desarrollo de programas fue propuesto por John Guttag [76].

Definición 5.1 *Un Tipo Abstracto de Datos (TAD) es una colección de valores y de operaciones que se definen mediante una especificación que es independiente de cualquier representación [139].*

Básicamente, constituye un concepto matemático que se define con la intención de especificar las propiedades lógicas de un tipo de datos o de un objeto.

La definición de un TAD implica la definición de dos partes diferenciadas:

A. Interfaz de usuario [2] Permite definir la forma de uso del TAD por los programadores que deseen utilizarlo. Con el objetivo de evitar las ambigüedades que se pudieran generar al expresarla en lenguaje natural, la interfaz de usuario vendrá determinada por la especificación algebraica del TAD, que consta de:

1. *Especificación sintáctica,* mediante la que se establecen el conjunto de operaciones básicas para la creación y manipulación del TAD, así como su perfil sintáctico. Las operaciones pueden ser de tres tipos:

   Generadoras. Mediante éstas, ha de ser posible generar cualquier valor del TAD, pues ahora no existe otro modo de hacerlo.

[2] "usuario" es el programador que "usa" el TAD

Modificadoras. Permiten modificar los valores del TAD definido, de tal forma que se eviten los errores sintácticos, semánticos o de ejecución, que se podrían producir si el programador tuviera la posibilidad de manipular directamente los valores del TAD a través de su representación en memoria;

De acceso. Permiten conocer los valores definidos por los términos que representan valores del TAD.

2. *Especificación semántica,* por la que se pretende dotar de significado a las operaciones definidas en la especificación sintáctica.

B. La representación en memoria, que queda oculta al usuario del TAD, de modo que, una vez establecida la interfaz, el programador del TAD es libre para escoger la representación que más le convenga. Ésta dependerá de las operaciones definidas sobre él y de la eficiencia que se desee (en tiempo y/o espacio). Para realizar un tratamiento homogéneo, la definición de la representación de las operaciones debería realizarse en un ámbito de declaración inaccesible al resto del programa. Los demás programadores (usuarios) sólo conocerían el nombre del tipo y la especificación sintáctica y semántica de las operaciones. Si el programador que ideó el TAD decide cambiar su representación por otra, no debería afectar al resto del programa. Así, el tratamiento dado por el lenguaje a los tipos definidos por el programador sería equivalente al que da a sus propios tipos, con lo que se logran tres objetivos: privacidad de la representación, protección y facilidad de uso.

Entre las ventajas de la programación basada en TADs [139] destacaremos las facilidades que proporciona para las siguientes tareas:

- La programación modular, concepto básico de la POO, puesto que la definición de un TAD es un candidato idóneo para el diseño de un módulo o clase, al reunir los requisitos y propiedades exigidas a los módulos: pocas conexiones, independencia y tamaño adecuado. Además, las modificaciones y mejoras del programa no deben afectar al módulo y viceversa.

- De cara a la utilización de la técnica de los refinamientos sucesivos al realizar el diseño del programa, evita definir demasiado temprano la forma en la que se van a representar los datos.

- En cuanto a la programación a gran escala, permite repartir el trabajo, lo que es indispensable cuando se trata de desarrollar proyectos de cierta envergadura.

- Por otro lado, proporciona una herramienta para la especificación de tipos u objetos en los que algunas de sus características están indefinidas, además de permitir la reutilización de *software*.

- Finalmente, constituyen una ayuda inestimable de cara a la verificación formal de programas.

Así pues, la programación con TADs permite describir los objetos en los que se basa la arquitectura del programa a diseñar, de tal forma que el proceso de construcción de *software* OO se puede considerar como la construcción de colecciones estructuradas de implementaciones de TADs. La estructuración se consigue gracias a dos relaciones entre las distintas clases: clientelismo y herencia. La herencia es una forma de reutilización de software que permite crear nuevas clases a partir de otras ya existentes. Básicamente representa la relación *ES_UN* entre objetos. Una clase es cliente de otra cuando forma parte de sus atributos.

Por último, hemos de concluir este apartado indicando que, desde nuestra experiencia en la programación de ELVIRA y en la docencia universitaria de las asignaturas relacionadas con programación y estructuras de datos, podemos afirmar que una aplicación informática de las características de ELVIRA, tanto por su envergadura como por las circunstancias en que se ha generado el código, no se habría podido desarrollar con éxito sin el enfoque de la OO y de los TAD.

Lenguaje de implementación: JAVA

El lenguaje JAVA fue creado por James Goslings, de la empresa norteamericana Sun Microsystems, a raíz de un proyecto de investigación corporativo interno relacionado con el impacto del papel de los microprocesadores en los dispositivos electrónicos de consumo inteligente. Uno de los objetivos de dicho proyecto fue el de crear un nuevo lenguaje de programación basándose en los dos más utilizados en todo el mundo, C y C++, intentando mejorar sus limitaciones y haciéndolo portable al máximo para poder utilizarlo en la implementación de aplicaciones basadas en Internet y la World Wide Web (WWW). El nombre asignado originalmente a dicho lenguaje fue Oak, por el roble que crecía fuera de la ventana del despacho de su creador. Sin embargo, al estar ya registrado, se decidió cambiarle el nombre por el de JAVA[3] cuando un grupo de empleados de Sun visitó una cafetería local. Se presentó por primera vez en el año 1995 en una conferencia sobre Internet y la WWW, con el objetivo de dotar de "contenido dinámico" a las páginas de

[3]En inglés americano coloquial se denomina *java* al café.

la WWW, integrando en ellas audio, video y otros elementos multimedia. Sin embargo, a partir de entonces su fama ha ido en aumento hasta ser considerado hoy día como uno de los lenguajes de programación más utilizados en todo el mundo [37]. Por otra parte, las empresas cada vez demandan más expertos en el lenguaje JAVA, puesto que sus características lo hacen especialmente apropiado para la implementación de aplicaciones relacionadas con la planificación estratégica de los sistemas de información, cada vez más en auge debido a la nueva era de globalización de la información. Entre las principales aplicaciones de Java podemos destacar la implementación de sistemas operativos, sistemas de comunicaciones, sistemas de bases de datos, etc.

Las ventajas más reseñables del lenguaje JAVA son [127, 179]:

- es un lenguaje de propósito general *Orientado a Objetos*, lo que permite desarrollar grandes aplicaciones utilizando el paradigma de la POO. Pero, en este sentido, se diferencia de otros lenguajes "puristas", en los que todo es POO, en que también permite programar mediante el paradigma de la programación imperativa.

- Es *portable*, por lo que los programas se compilan una vez y se pueden "ejecutar" en cualquier sistema que tenga una "máquina virtual" de Java, ya sea linux, Windows, unix, Macintosh, etc., lo cual constituye una de las características más importantes de cara al desarrollo de una herramienta como ELVIRA, que pretende ser utilizada en diferentes plataformas;

- Actualmente, es capaz de manejar correctamente las condiciones anormales gracias, fundamentalmente, al tratamiento de las excepciones y a la gestión de la memoria. Precisamente, ésta última tarea constituye una de las principales causas de los errores de los programadores (y de sus dolores de cabeza intentando encontrarlos...). Además, al ser un lenguaje muy estricto en relación a los tipos y a las declaraciones de datos, muchos de los errores típicos se pueden detectar durante la compilación.

- Además, es un lenguaje diseñado especialmente para crear programas *interactivos* de cara a su implantación en las páginas de la WWW. Por eso, soporta la concurrencia entre distintos procesos, así como la capacidad de ejecutarse en los navegadores y en los servidores *web*.

- El conjunto de clases incorporadas al lenguaje Java, contribuyen a aumentar su *potencia* y a hacer que sea lo suficientemente *rico* para la resolución de cualquier tipo de problema.

- Finalmente, el hecho de que el entorno básico de desarrollo sea *gratuito*, contribuye a su difusión e implantación.

A pesar de todas las ventajas citadas, no debemos olvidar sus mayores limitaciones:

- Rendimiento. El hecho de que sea un lenguaje "interpretado"[4] hace que no se elija para implementar soluciones de problemas en los que sea necesario realizar gran cantidad de cálculos. No obstante, en las evaluaciones que se han hecho con ELVIRA podemos decir que los resultados han sido bastante satisfactorios, incluso en redes con un elevado número de nodos.

- Robustez. En la última versión, la 1.4, parece ser que ha disminuido el número de errores de ejecución que existían en las versiones anteriores. No obstante, sigue teniendo algunos fallos.

- Compatibilidad entre versiones. El proyecto ELVIRA comenzó a implementarse con la versión 1.2; posteriormente, cuando surgió la 1.3, se producían algunos errores al comenzar la ejecución de ELVIRA en Windows. En la última versión vuelve a funcionar, aunque con algunas diferencias de presentación en las componentes gráficas.

5.1.3. Arquitectura

Puesto que otra de las aportaciones de est memoria consiste en el diseño e implementación de la interfaz gráfica de usuario de ELVIRA, en esta sección pretendemos profundizar en los elementos *software* de los que consta ELVIRA, así como en sus relaciones, lo que esperamos que ayude a comprender cómo se ha llevado a cabo el proceso de desarrollo.

Básicamente, el diseño de ELVIRA se ha organizado en torno a cuatro módulos principales, cada uno de ellos encargado de resolver un conjunto específico de tareas, tal y como se ilustra en la figura 5.2:

1. **Representación de datos.** En este módulo, que es sobre el que se fundamentan todos los demás, se definen las estructuras de datos para representar en memoria (interna o externa) toda la información asociada a las redes bayesianas y los diagramas de influencia, además de las clases auxiliares.

[4] Realmente JAVA no es un lenguaje interpretado, puesto que el intérprete no se ejecuta sobre el código fuente sino sobre código intermedio expresado en *bytecode* resultante de la compilación del código fuente.

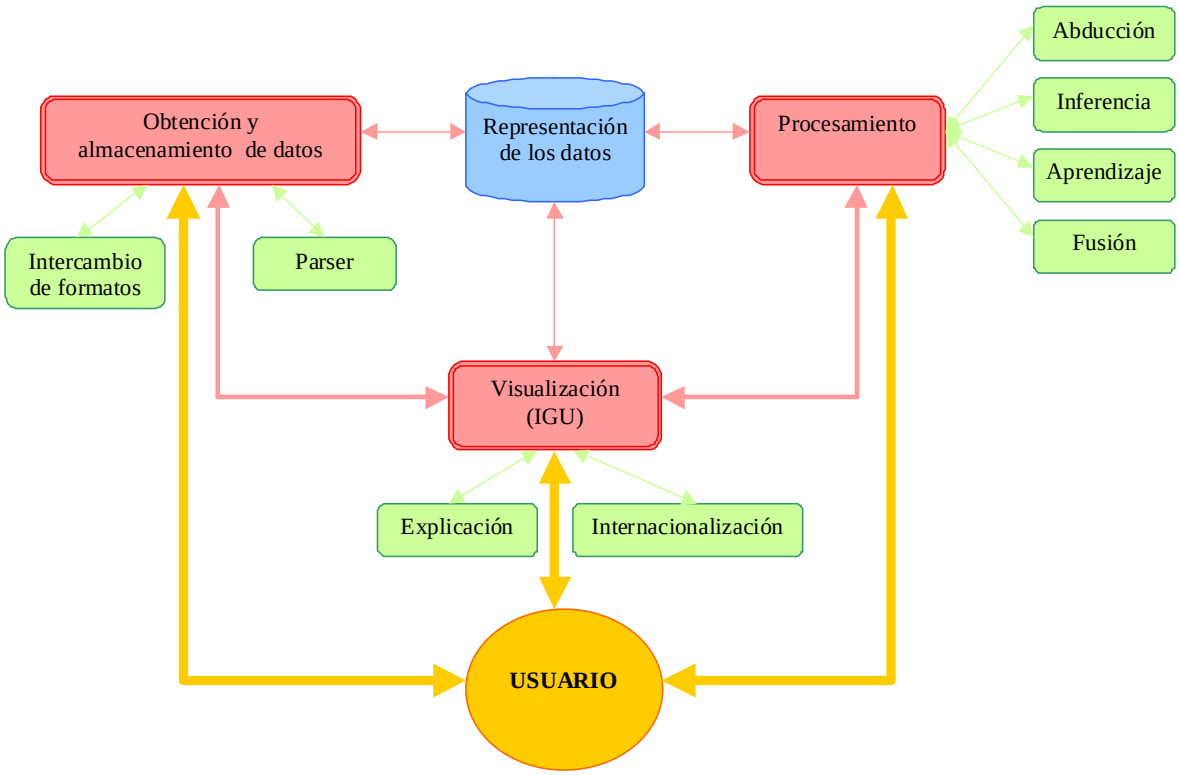

Figura 5.2: Esquema de los distintos elementos de ELVIRA y sus relaciones.

2. **Obtención y almacenamiento de datos,** que es el que se encarga de cargar y/o almacenar una red o un diagrama de influencia, los casos de evidencia, las bases de datos para los algoritmos de aprendizaje, etc. desde/en algún dispositivo de almacenamiento (interno o externo). Todo ello es posible gracias a uno de los submódulos que lo integran, el *analizador sintáctico*, que es el que se encarga de hacer la traducción de/a los correspondientes archivos de texto. Existe otro submódulo de *intercambio de formatos*, que permite cargar redes y almacenarlas desde y en distintos formatos de redes.

3. **Procesamiento,** que se encarga de las tareas de procesamiento y evaluación de modelos probabilísticos. Para ello, se ha organizado en un conjunto de submódulos más pequeños con el objetivo de realizar las tareas de *inferencia, abducción, aprendizaje* y *fusión.*

4. **Visualización.** Este es el módulo encargado de implementar la IGU de ELVIRA, por lo que necesita interaccionar constantemente con el resto de los módulos, con el fin de permitir al usuario realizar las tareas relacionadas con ellos. Además, utiliza las funcionalidades proporcionadas por los submódulos de *explicación,* que es el que

genera los distintos tipos de explicación comentados en los capítulos anteriores, y de *internacionalización*, lo que permite ofrecer la información y los mensajes de texto de la IGU en distintos idiomas.

Hay que indicar también que el usuario puede interaccionar directamente con los módulos de obtención y almacenamiento de datos, inferencia y visualización, pero nunca con el de representación de los datos, manteniendo así la coherencia del principio de encapsulamiento y ocultación de la información comentado en las secciones anteriores.

Las ventajas más relevantes que proporciona este diseño están relacionadas con las posibilidades que ofrece para realizar el desarrollo del programa entre distintos grupos de programadores. Además, está preparado para modificar y dotar de nuevas funcionalidades a ELVIRA, lo que constituye uno de los principales objetivos de cualquier proyecto de ingeniería del software.

En la actualidad, la implementación del programa ELVIRA cuenta de unas 118.000 líneas de código JAVA, organizadas en 15 paquetes que contienen unas 250 clases, de las cuales unas 50, equivalentes a unas 20.000 líneas de código han sido programadas por la autora de esta memoria.

5.1.4. Estructuras de datos y algoritmos en ELVIRA

En esta sección se presentan las estructuras de datos y los algoritmos sobre los que se basan los elementos descritos en la sección anterior. Como la implementación de los algoritmos depende de las estructuras de datos definidas, es importante realizar una definición precisa de las mismas, lo que se consigue gracias al enfoque de los TADs. Como ejemplo describiremos la estructuras de datos más elemental: el grafo.

Definición del TAD Grafo

Una red bayesiana y un diagrama de influencia se definen a partir del concepto de grafo, al que se le añaden ciertas propiedades y restricciones. Por tanto, la primera tarea consistió en la definición del TAD grafo y su correspondiente implementación en una clase, de la que heredasen tanto las clases para redes bayesianas como para diagramas de influencia.

Un grafo se define como un conjunto de vértices y un conjunto de aristas entre los vértices del mismo. En el caso de los modelos gráficos probabilísticos, los vértices se denominan *nodos* y las aristas, *enlaces*. La definición precisa de los grafos vendrá determinada

por su especificación algebraica, una parte de la cual se incluye en el apéndice A a modo de ejemplo. Para la representación en memoria del TAD grafo, existen distintas posibilidades: con memoria estática, mediante matrices de adyacencias, o con memoria dinámica, mediante listas de adyacencias. La primera posibilidad es muy eficiente en cuanto a la complejidad de los algoritmos, y muchos de ellos son más fáciles de implementar; sin embargo, conlleva los problemas de exceso o defecto de memoria que requiere la estimación previa del tamaño de la matriz de adyacencias. Por tanto, elegimos la segunda opción por las ventajas que conlleva: el acceso directo a los adyacentes de cada vértice, que constituyen la base de los algoritmos de procesamiento de grafos. Por otra parte, aunque las redes bayesianas y los diagramas de influencia son grafos dirigidos, sin embargo diferentes operaciones de manipulación de los mismos lo hacen sobre la versión "no dirigida". Incluso, existen casos en los que interesa considerar grafos mixtos, es decir, una parte es dirigido y otra es no dirigido. Por ello, se propone la siguiente definición:

clase Grafo{

Variables de instancia

Privadas

tipo:(dirigido, no dirigido, mixto);

Protegidas

nodos:Lista (de Vértice)
enlaces:Lista (de Arco)

}

Definición del TAD Vértice

Normalmente, las operaciones de manipulación de los vértices consisten simplemente en obtener los adyacentes y los incidentes. En grafos no dirigidos, los adyacentes se denominan hermanos; en grafos dirigidos, pueden llamarse padres o hijos, dependiendo de si son el origen o el destino de un arco, respectivamente, por lo que interesará distinguir los padres, hijos y/o hermanos de un nodo. Pero además, puesto que se utilizarán para representar los nodos de los distintos modelos probabilísticos, interesará conocer el tipo de nodo en el modelo (aleatorio, decisión o utilidad), así como el tipo de variable que representa(continua, discreta, etc.) En este caso, la novedad reside en su representación en memoria, puesto que tradicionalmente la forma de hacerlo es asociando a cada vértice

la lista de sus adyacentes. La ventaja de esta representación es que el acceso a los descendientes de un vértice es muy sencillo (hablando en términos de complejidad), pero ocurre lo contrario con el acceso a sus ascendientes. Para evitarlo, merecería la pena tener acceso también a los padres de un nodo. Sin embargo, en lugar de hacer referencia a los vértices, y con el objeto de mantener la coherencia de la representación completa, hemos decidido hacer que los vértices hagan referencia a los enlaces que contienen la información sobre sus padres y sus hijos. Por otro lado, como en grafos no dirigidos todos los nodos son hermanos, y pensando también en la posibilidad de incluir en un futuro la posibilidad de trabajar en ELVIRA con redes de Markov y grafos en cadena, donde también aparece el concepto de "hermano", decidimos representarlos explícitamente de forma análoga a los "padres" e "hijos". Por tanto, la definición de la clase Vértice que se propone es:

clase Vértice{

Variables de instancia

Privadas

nombre: Cadena
título: Cadena
tipodeNodo:(aleatorio, decisión, utilidad)
tipodeVariable: (continua, de estados finitos, discreta
infinita, mixta)

Protegidas

hijos: Lista (de Arco)
padres: Lista (de Arco)
hermanos: Lista (de Arco)

}

Definición del TAD Arco

La información que contienen los arcos básicamente consiste en los vértices origen y destino que lo definen, también llamados cola y cabeza, respectivamente. Sin embargo, con el fin de optimizar el rendimiento del sistema, e intentar conseguir el menor número de inconsistencias, conviene saber en cada momento si el arco es dirigido o no. La representación en memoria es bastante sencilla, puesto que sólo es necesario conocer el origen y el destino del arco.

clase Arco{

Variables de instancia

Privadas

cabeza: Vértice

cola: Vértice

}

Como resumen del diseño de clases presentado, ilustramos en la figura 5.4 cómo quedaría representado internamente, según lo muestra la figura 5.3.

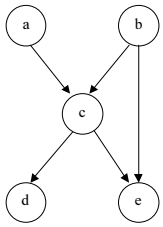

Figura 5.3: Ejemplo de grafo

El resto de las clases de ELVIRA se definen mediante la extensión, agregación y uso de estas clases básicas.

ALGORITMOS

Respecto a los algoritmos implementados, los citaremos según su funcionalidad:

Para edición En este caso, los más relevantes son los relacionados con la creación de redes o diagramas de influencia. Inicialmente, se parte de una red o diagrama vacío sobre los que se van insertando nodos y enlaces.

Para inferencia ELVIRA cuenta con un amplio repertorio de métodos de inferencia. Entre los métodos exactos implementados actualmente están: el de Hugin [91], el método de Shenoy-Shafer [162], la propagación perezosa [109] y la eliminación de variables [194]. Los métodos aproximados con los que cuenta ELVIRA son: propagación barata (penniless propagation) [11], propagación perezosa barata (lazy-penniless propagation) [12], muestreo por importancia y muestreo sistemático [86]. Además, se han incluido versiones aproximadas del de Hugin y el de eliminación de

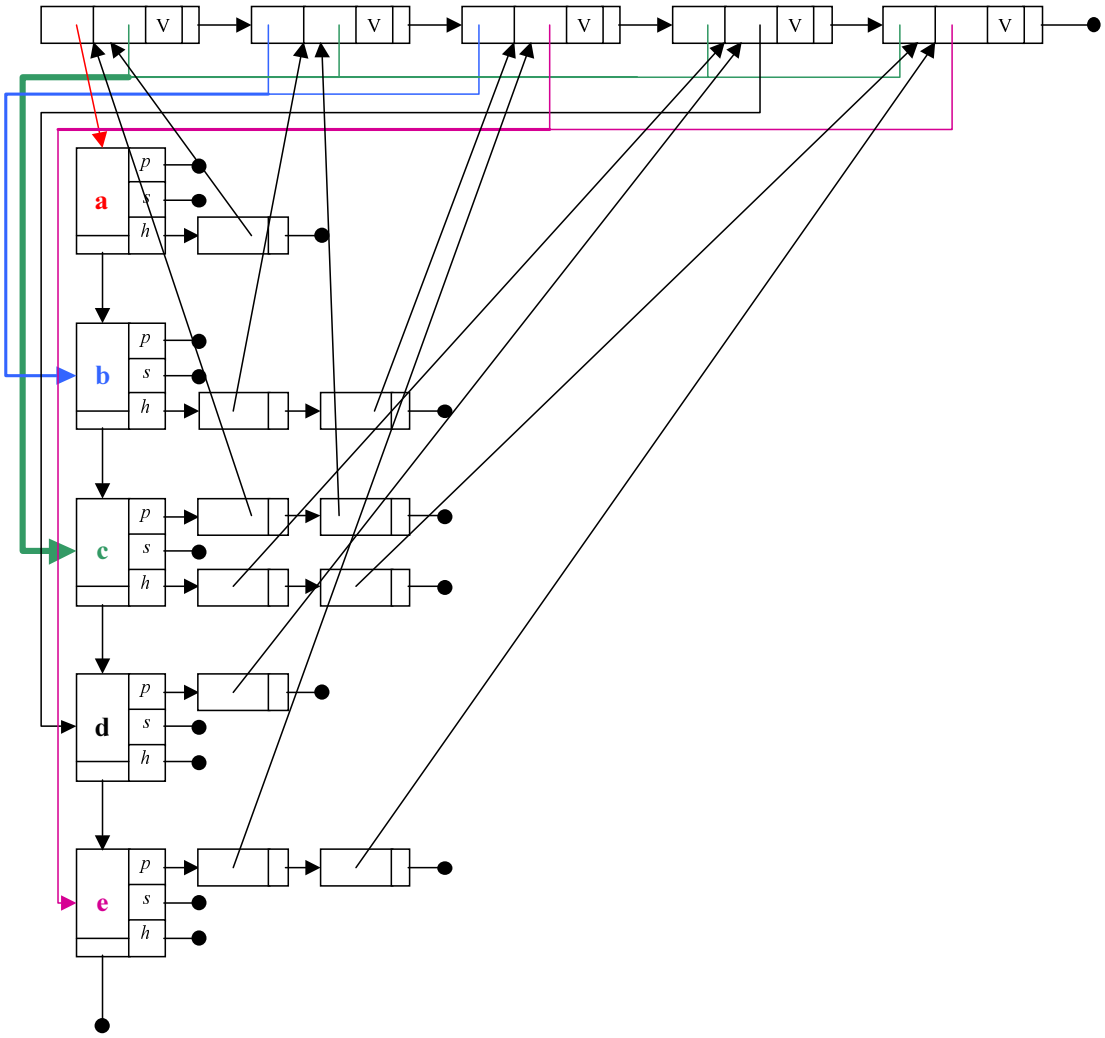

Figura 5.4: Representación en memoria del grafo de la figura 5.3

variables basados en el uso de árboles de probabilidad [10] para representar los potenciales. En ELVIRA, el algoritmo de Shenoy-Shafer es un caso particular del de propagación barata, en el que el nivel de error se establece a 0 y el número de etapas se limita a dos. Con respecto al de muestreo por importancia se han implementado tres versiones: basadas en tablas de probabilidad [86], en árboles de probabilidad [152] y en variables antitéticas [153].

También está implementado (aunque no integrado en la IGU) el método de propagación de Monte Carlo con cadenas de Markov para variables continuas cuyas distribuciones de probabilidad vienen dadas por una mezcla de potenciales truncados [125]. Este modelo permite trabajar simultáneamente con variables continuas y discretas en la misma red.

Para explicación En este caso distinguimos dos tipos de métodos: los relacionados con la abducción y los que sirven para la explicación del modelo y del razonamiento. Para el primer caso, se ha implementado el algoritmo de Nilsson [132] para abducción total; para abducción parcial, primero se obtiene un árbol de uniones válido, que solo contiene las variables del conjunto explicación, y después se aplica el algoritmo de Nilsson [132].

En cuanto a los algoritmos de explicación del modelo y del razonamiento, se han implementado los métodos descritos en los capítulos 3 y 4.

Para aprendizaje Respecto a la capacidad de aprendizaje con la que se ha dotado a ELVIRA, hay que destacar la implementación de los siguientes algoritmos para la obtención de la estructura de las redes bayesianas: dentro de los métodos basados en detección de independencias, se ha incluido el algoritmo PC [168], que es el más conocido y utilizado. Además, se han incluido métodos basados en métricas y técnicas de búsqueda. De momento, las métricas implementadas son: K2 [29], BIC [155] and BDeu [81].

Además, teniendo en cuenta la métrica y el algoritmo de búsqueda, los usuarios de ELVIRA pueden seleccionar actualmente cualquiera de los siguientes algoritmos: Búsqueda Local (LS) [81], el algoritmo K2 y una versión distribuida del método de búsqueda en la vecindad (Variable Neighbourhood Search) [35]. Todos estos algoritmos tienen en común el mismo espacio de búsqueda, que es el de los grafos dirigidos acíclicos. Una opción alternativa, que toma como espacio de búsqueda las ordenaciones de las variables, está implementada mediante el algoritmo K2SN [36].

Con respecto al aprendizaje de parámetros, se han proporcionado los dos estimadores siguientes: el estadístico de máxima verosimilitud (MLE) y el bayesiano por aproximación de Laplace [29].

De momento, los archivos que contienen las bases de datos utilizadas como entradas de los algoritmos deben tener un formato determinado. Además, ELVIRA no es capaz todavía de tratar los valores *perdidos*, es decir, aquéllos que no están recogidos en la base de datos, ni las *variables latentes*, que son aquellas que no han sido observadas pero deben incluirse en el modelo aprendido para reflejar las dependencias probabilísticas que se observan en los datos.

5.2. Interfaz gráfica con capacidad de explicación

La interfaz gráfica de ELVIRA, como ya hemos dicho antes, está basada en un sistema de menús y ventanas, con el objetivo de hacer lo más cómodo y sencillo posible al usuario su relación con el entorno. De hecho, el usuario puede seleccionar el **idioma** en el que desea trabajar. De momento sólo están incluidos el inglés y el castellano, pero en un futuro próximo se prevee aumentar la lista de idiomas. Además, existen tres modos distintos de trabajo, cada uno con su correspondiente interfaz gráfica, dependiendo del tipo de tarea que el usuario desee realizar: edición, inferencia y aprendizaje. El paso de uno a otro se realiza simplemente mediante la selección con el ratón. No obstante, puesto que el usuario puede tener abiertas varias redes o diagramas a la vez, el modo de trabajo no tiene por qué ser el mismo para todas, lo que implica que con una de las redes abiertas se puede estar en modo edición y con otra en modo inferencia, por ejemplo. Así pues, la interfaz de ELVIRA la presentaremos dependiendo del modo de trabajo, aunque no expondremos el caso del aprendizaje porque no hemos sido nosotros los responsables de su desarrollo.

5.2.1. Interfaz para edición

En modo edición, el usuario puede crear redes bayesianas y diagramas de influencia de un modo sencillo e intuitivo. En la figura 5.5 podemos ver la ventana para la edición de modelos gráficos probabilísticos. Las principales diferencias con otras herramientas (véase sección 5.4) aparecen en la edición de nodos y de enlaces.

En el caso de la edición de **nodos**, cada vez que se crea un nodo aparece automáticamente una ventana de edición mediante la que se ofrece al usuario la posibilidad de incluir todos los datos relacionados con él, como son el nombre, los estados, la tabla de probabi-

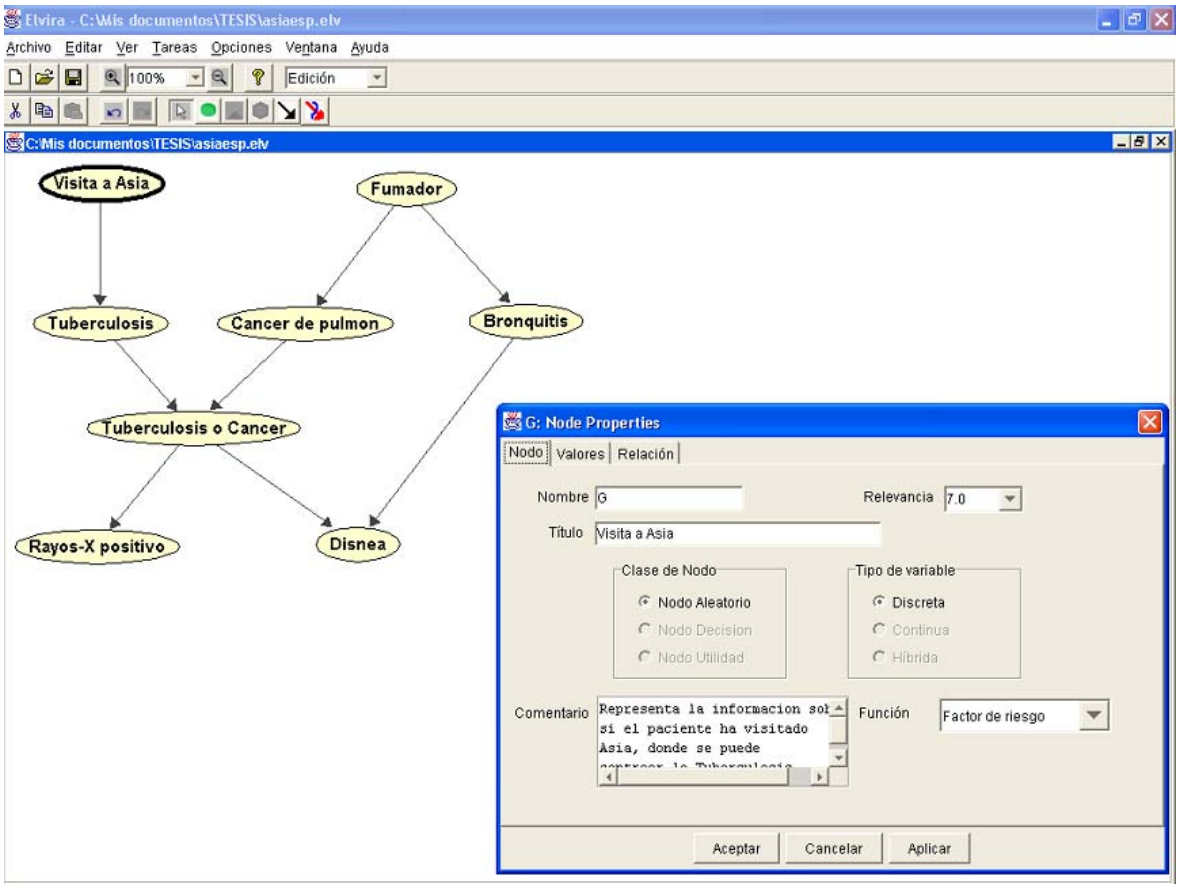

Figura 5.5: Interfaz para edición en ELVIRA para la red ASIA con una muestra de las ventanas de edición de nodos

lidad, etc., y otra serie de propiedades que se utilizarán posteriormente para la generación de explicaciones y que son: la *Definición* de la variable que representa, su *factor de importancia* y el *rol* que desempeña en la red. El factor de importancia es un valor, comprendido entre 0 y 10 (por defecto tiene asignado el valor 7), que el usuario puede modificar en cualquier momento. Para la definición del rol que desempeña el nodo, se ofrece una lista desplegable con las opciones más comunes (véase sección 3.2.1). Por defecto, no se asigna ninguno. Además, puesto que es muy complicado realizar una clasificación exhaustiva de todas las posibilidades, el usuario puede definir otras que considere necesarias para su modelo. Para ello, únicamente debe seleccionar el valor "Definir" e introducir el nuevo nombre en el área de texto reservada para ello.

Otras propiedades de ELVIRA útiles para la edición son el *"zoom"* y la posibilidad de *deshacer* y *rehacer* los últimos cambios realizados.

La edición de **enlaces** es similar. Una ventana permite definir si el enlace es dirigido o no, el nombre (por si interesa identificarlo para que aparezca luego en las explicaciones

verbales), el origen y el destino, y el tipo de relación que representa entre ambos, tal y como se indicó en la sección 3.2.2.

5.2.2. Interfaz para inferencia

Básicamente, en modo inferencia las operaciones que se pueden hacer son las de introducción de evidencia y su propagación. Sin embargo, la diferencia con otras herramientas radica en la posibilidad de seleccionar el método de propagación, la forma de visualizar sus resultados y el modo en que se puede introducir la evidencia. Estas propiedades las describimos a continuación.

Selección de opciones de inferencia

Desde el modo edición, el usuario puede pasar a modo inferencia, habiendo seleccionado previamente:

- El *objetivo* de la propagación. Con la ayuda de una ventana como la de la figura 5.6 se puede elegir si se desea obtener las probabilidades a posteriori de todas las variables no observadas o bien realizar una tarea de abducción total o parcial, también conocida como la búsqueda de la explicación más probable (EMP). En caso de que se elija esta última posibilidad, aparece una ventana nueva mediante la que se pueden seleccionar las variables que se desea incluir en la explicación. En el caso de la abducción, el usuario puede seleccionar también el número de explicaciones más probables que desea obtener.

- El *método* de propagación utilizado. Por defecto, se utiliza el algoritmo exacto de HUGIN aunque ya hemos descrito en la sección 5.1.4 el repertorio de métodos implementados.

- La *forma* en que desea que se realice la inferencia: automática o manual, lo que se hace también a través de la ventana anterior. En el primer caso, cada vez que se pasa de modo edición a modo inferencia, la red se compila [5] automáticamente. Si ya se está en modo inferencia, siempre que se introduce nueva evidencia en la red, se realiza la propagación de dicha evidencia. En el caso de que se seleccione la opción

[5] La compilación de la red consiste en hacer una propagación inicial sin evidencia teniendo en cuenta el objetivo previamente seleccionado. Por defecto, es el cálculo de las probabilidades a posteriori de todas las variables de la red.

manual, tanto la compilación de la red como la propagación de la evidencia se hace cuando el usuario lo solicita para lo que existe un icono expecífico para ello.

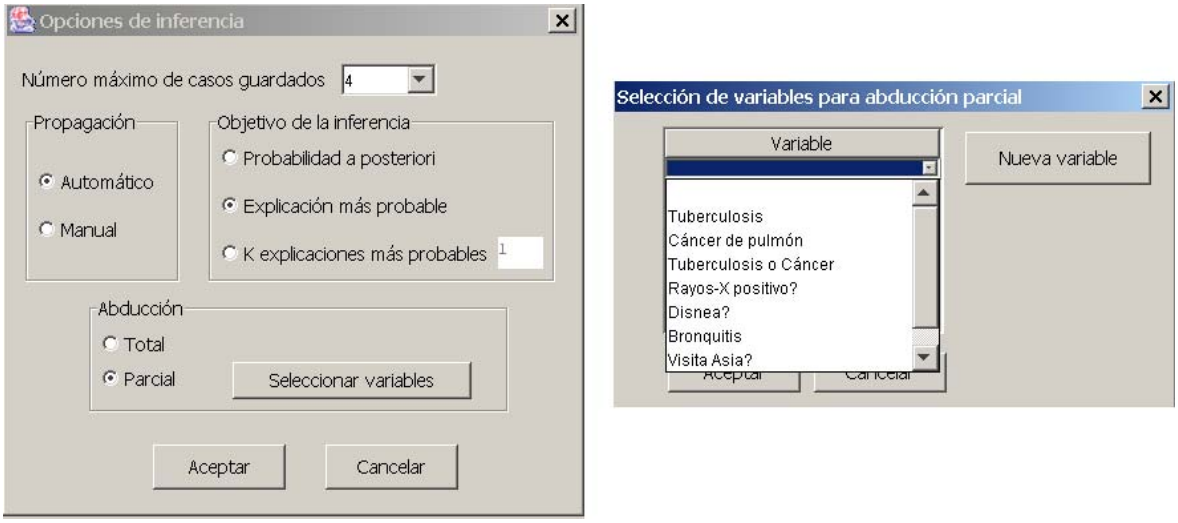

Figura 5.6: Ventanas para seleccionar las opciones de inferencia en ELVIRA

- Ciertas cuestiones relacionadas con la forma de obtener la *explicación del razonamiento* como, por ejemplo, el número de casos de evidencia guardados en memoria. Estas opciones las describiremos más adelante.

A pesar de que todas estas opciones deben estar definidas antes de pasar a modo inferencia, una vez en dicho modo se pueden volver a modificar en cualquier momento.

Expansión selectiva de nodos

Por otra parte, siempre que se pasa de modo edición a modo inferencia, independientemente de si la red se compila o no, los nodos cuyo factor de importancia es mayor o igual que el *umbral de expansión* se muestran gráficamente en **modo expandido**; los demás aparecerán en **modo contraído**. En este caso, la representación gráfica de los nodos es la misma que en modo edición. La figura 5.7 representa la red ASIA y se puede observar que hay cuatro nodos expandidos (Tuberculosis, Cáncer de pulmón, Bronquitis y Disnea) y el resto están contraídos. La información que presentan los nodos expandidos es la siguiente: el nombre del nodo y un área de datos, de fondo blanco, que contiene los identificadores de todos los estados del nodo, destacando con un recuadro el más probable; una barra al lado de cada estado de longitud proporcional a su probabilidad, junto con el valor numérico que ésta representa. Además, podemos ver que el valor del umbral de expansión es 5. Por defecto, el umbral de expansión está definido con el valor 7, pero es

modificable por el usuario por medio de una lista desplegable de la barra de herramientas del modo inferencia.

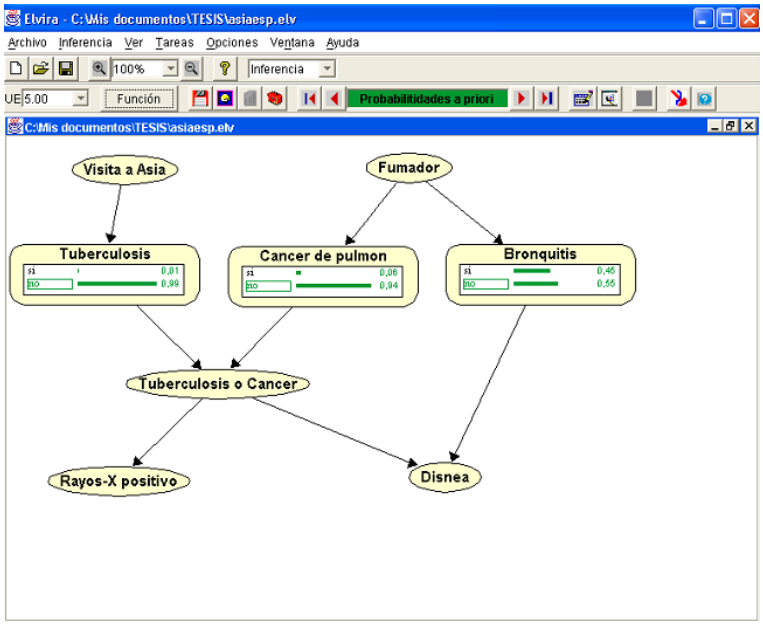

Figura 5.7: Ejemplo de la red ASIA con los nodos cuyo factor de importancia es mayor o igual que 8 expandidos

El usuario también puede modificar el criterio de expansión según el *rol* de cada nodo. Para ello, simplemente debe presionar el botón "Función" de la barra de herramientas para que le aparezca un cuadro en el que puede señalar el rol o roles deseados. De esta forma, combinando el umbral de expansión y la función que desempeña cada nodo se tiene una herramienta muy potente para generar explicaciones gráficas del modelo representado en la red. En la figura 5.7 se han expandido los nodos cuyo factor de importancia es mayor que 5 y cuyo rol es el de signo o enfermedad. Además, el usuario tiene también la posibilidad de expandir o contraer los nodos previamente seleccionados por el usuario mediante un icono específico para ello en la barra de herramientas.
Si al pasar a modo inferencia la red no está compilada, el área de datos de cada nodo expandido solamente contendrá el nombre de cada estado.

En todos los casos, el *tamaño* de cada nodo expandido es proporcional al número de estados que contenga, por lo que no todos los nodos expandidos tienen por qué tener el mismo tamaño, a diferencia de lo que ocurre en otras herramientas (sección 5.4).

Otra importante aplicación de esta herramienta es la posibilidad de mostrar gráficamente también los resultados de los procesos de *abducción*. En este caso, como una explicación está definida por la asignación de valores a ciertas variables, el método con-

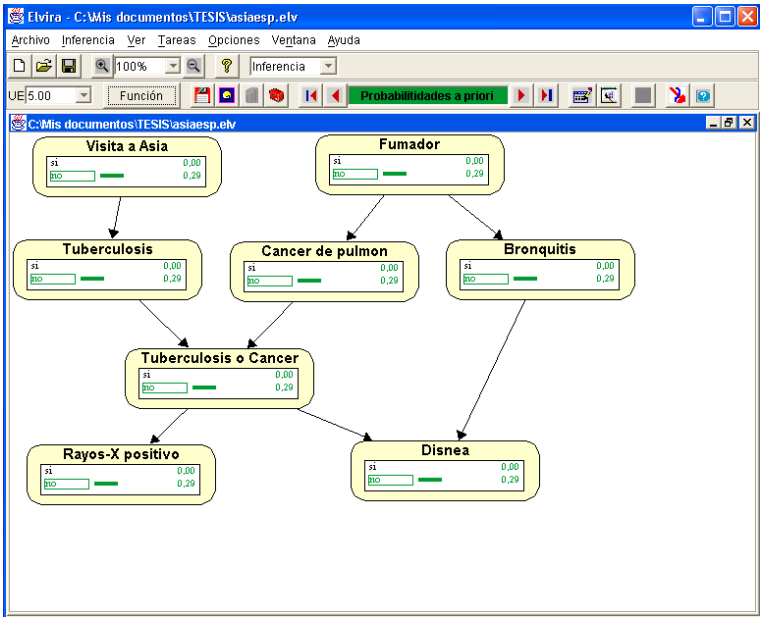

Figura 5.8: Ejemplo de abducción total sobre la red ASIA. Observamos que la explicación más probable correspondería a la configuración C={Visita a Asia=no, Fumador=no, Tuberculosis=no, Cáncer de pulmón=no, Bronchitis=no, Tuberculosis o Cáncer=no, Disnea=no, Rayos-X positivo=no}, siendo además su probabilidad P(C)=0.29

siste en mostrar en los nodos expandidos la representación gráfica mediante diagramas de barras únicamente para los estados de cada nodo que forman parte de la explicación. Para ello, se asocia a dicho estado la probabilidad de la explicación; el resto de los estados tendrían una probabilidad 0. De esta forma, al igual que en el caso anterior, se pueden identificar fácilmente las variables que definen la explicación junto con sus valores asociados, tal y como ilustra la figura 5.8.

Introducción de evidencia

Por último, la forma en que se *introduce la evidencia* es bastante sencilla. Se puede hacer de modo *dinámico*, durante el funcionamiento de ELVIRA, o de modo *estático*, cargando los datos desde un archivo. En el primer caso, si el nodo sobre el que queremos definir un hallazgo dado está expandido, el usuario únicamente debe pinchar con el ratón en la franja correspondiente al estado observado. En ese caso, la probabilidad de dicho estado se actualiza a 1 y la del resto a 0, y el color del fondo del nodo cambia de color, pasando de amarillo a gris, con el objetivo de visualizar gráficamente de un vistazo cuáles son los nodos cuyos valores se conocen con certeza. En la figura 5.9 podemos ver un ejemplo de nodo expandido antes y después de ser observado.

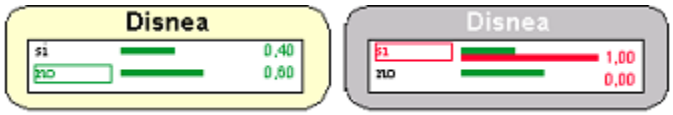

Figura 5.9: Ejemplo del nodo Disnea antes de ser observado y después de introducir como evidencia el estado yes

Para eliminar un hallazgo, simplemente basta con pinchar dos veces en su zona del área de datos. Si el nodo está contraído, al pulsar dos veces sobre él aparece un nuevo cuadro de diálogo en el que se muestran dos listas desplegables: una para seleccionar el nombre del nodo y la otra para seleccionar el estado que se desea introducir. Esta última contiene un valor más, denominado "borrar hallazgo", que sirve para eliminar dicho la evidencia que se tiene de dicho nodo cuando está contraído. Igualmente, después de introducir evidencia sobre un nodo no expandido su fondo pasa de amarillo a gris.

A pesar de que la idea de "hacer clic" sobre el nodo que se desea observar es muy cómodo e intuitivo, sin embargo, si la red tiene muchos nodos es posible que no sea fácil identificar las variables deseadas, bien porque no sea posible verlas todas a la vez en la pantalla, o bien porque se solapen unas con otras. Para este tipo de situaciones se ha implementado un **Editor de Casos de Evidencia**, detallado más adelante en la sección 5.3.2, que permite introducir la evidencia de una forma cómoda.

5.3. Explicación en ELVIRA

En esta sección presentamos la forma en que se han implementado en ELVIRA los mecanismos de explicación desarrollados en este trabajo, según los objetivos de cada uno de ellos: explicación del modelo (capítulo 3) y explicación del razonamiento (capítulo 4). Todo lo que se ha descrito en dichos capítulos está implementado en ELVIRA.

5.3.1. Explicación del modelo

Para la explicación del modelo, hemos tenido en cuenta la explicación por separado de los distintos elementos que lo integran: nodos, enlaces y la red completa. Casi todas estas opciones se ofrecen exclusivamente en *modo inferencia*, puesto que en modo edición todos los elementos son susceptibles de sufrir modificaciones, por lo que la explicación que se ha generado en un momento puede dejar de tener sentido inmediatamente después. Por eso, la explicación solo se ofrece cuando se ha salido del modo edición. Por otra parte, otra de

las principales ventajas del método y, por tanto de ELVIRA, es que las explicaciones sólo se generan a petición del usuario, por lo que no se muestra ningún tipo de información adicional que no haya sido requerida por él en un momento determinado.

Explicación de nodos. La explicación *gráfica* de un nodo determinado consiste en mostrarlo en modo expandido, según hemos comentado en la sección anterior, "haciendo clic" sobre el icono correspondiente. Para la explicación *verbal*, después de que el usuario haya seleccionado el nodo sobre el que desea obtener la explicación, aparece una ventana como la de la figura 5.10 que, como se puede observar, ofrece información de todas sus propiedades, como ya se comentó en la sección 3.2.1: nombre, estados, definición, causas, efectos y razones de probabilidad a priori y a posteriori. En el campo *definición* aparece la información incluida como comentario durante la edición del nodo, pero hay que indicar que, a diferencia de otras herramientas, este campo es editable. Esto permite que la explicación generada se pueda ir adaptando a distintas necesidades de los usuarios sin tener que editar el nodo en cuestión. Además, para mostrar las razones de probabilidad se incluye un pequeño texto, escrito en lenguaje natural, que indica si son mayores, menores o iguales a 1, además de incluir el valor de dicha razón. El texto es del tipo:"La razón de probabilidad es MAYOR/MENOR/IGUAL A 1, lo que implica que el estado s_1 (s_0 en el caso MENOR) es más probable que el estado s_0 (s_1 en el caso MENOR)" En el caso de que la variable no sea binaria, se muestran las razones de probabilidad para cada par de estados, ordenadas de mayor a menor, junto con el texto explicativo.

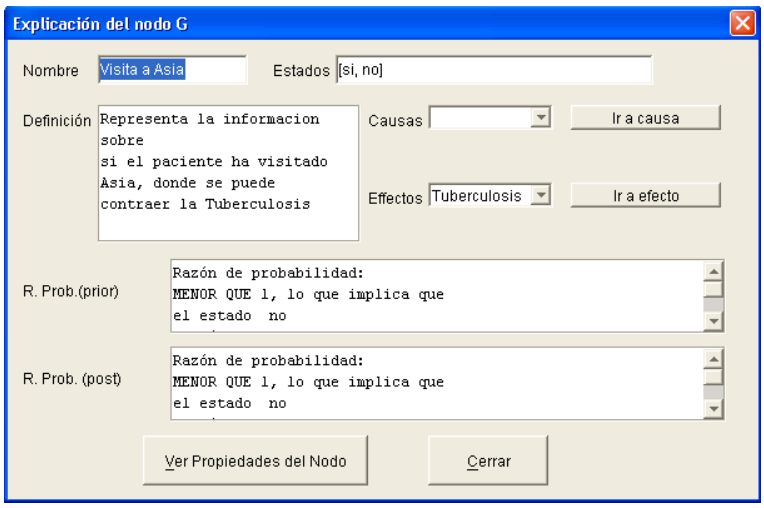

Figura 5.10: Ejemplo de explicación textual en ELVIRA para el nodo Visita a Asia de la red ASIA

Como facilidad suplementaria hay que destacar la capacidad de navegación por los nodos de la red, como ya propusiera Díez en su sistema DIAVAL [45]. Para ello, dentro de la ventana existen dos listas desplegables, correspondientes a las causas (padres) y a los efectos (hijos) del nodo, junto con dos botones al lado que permiten obtener la explicación del padre o del hijo seleccionado. Así, el usuario puede irse desplazando por aquellos caminos del grafo que considere necesarios con el fin de analizar los distintos datos que éste contiene. Para conocer el resto de las propiedades del nodo existe un botón que permite abrir una ventana igual a la de edición del nodo, con la salvedad de que ahora ningún campo es editable.

Por otra parte, la explicación gráfica de los nodos consiste en mostrarlos expandidos, con la información que ya hemos indicado previamente.

Explicación de enlaces. Cuando el usuario solicita la explicación verbal del enlace seleccionado, aparece una ventana como la de la figura 5.11, que contiene los nombres de los nodos origen y destino, el tipo de relación entre ambos nodos y las razones de verosimilitud para cada estado del nodo, junto con un pequeño texto explicativo sobre su significado, del tipo: "Para el estado d_i de [Destino] la R.V. de los estados de [Origen] es MAYOR/MENOR/IGUAL A 1, lo que implica que el estado o_1 (o_0 en el caso MENOR) lo explica MEJOR/IGUAL que el estado o_0 (o_1 en el caso MENOR); concretamente RV() veces mejor." Para ver otras propiedades del enlace existe un

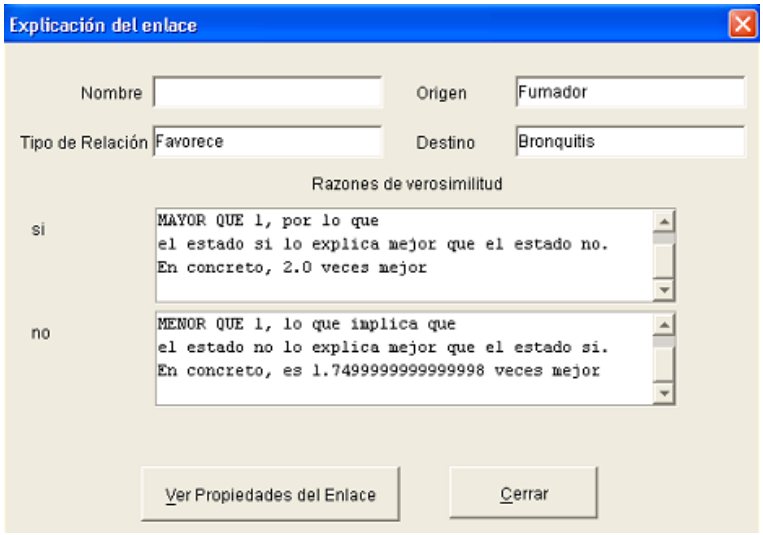

Figura 5.11: Ejemplo de explicación textual en ELVIRA para el enlace de Fumador a Bronquitis de la red ASIA

botón que abre una ventana igual que la de edición pero cuyos valores no se pueden

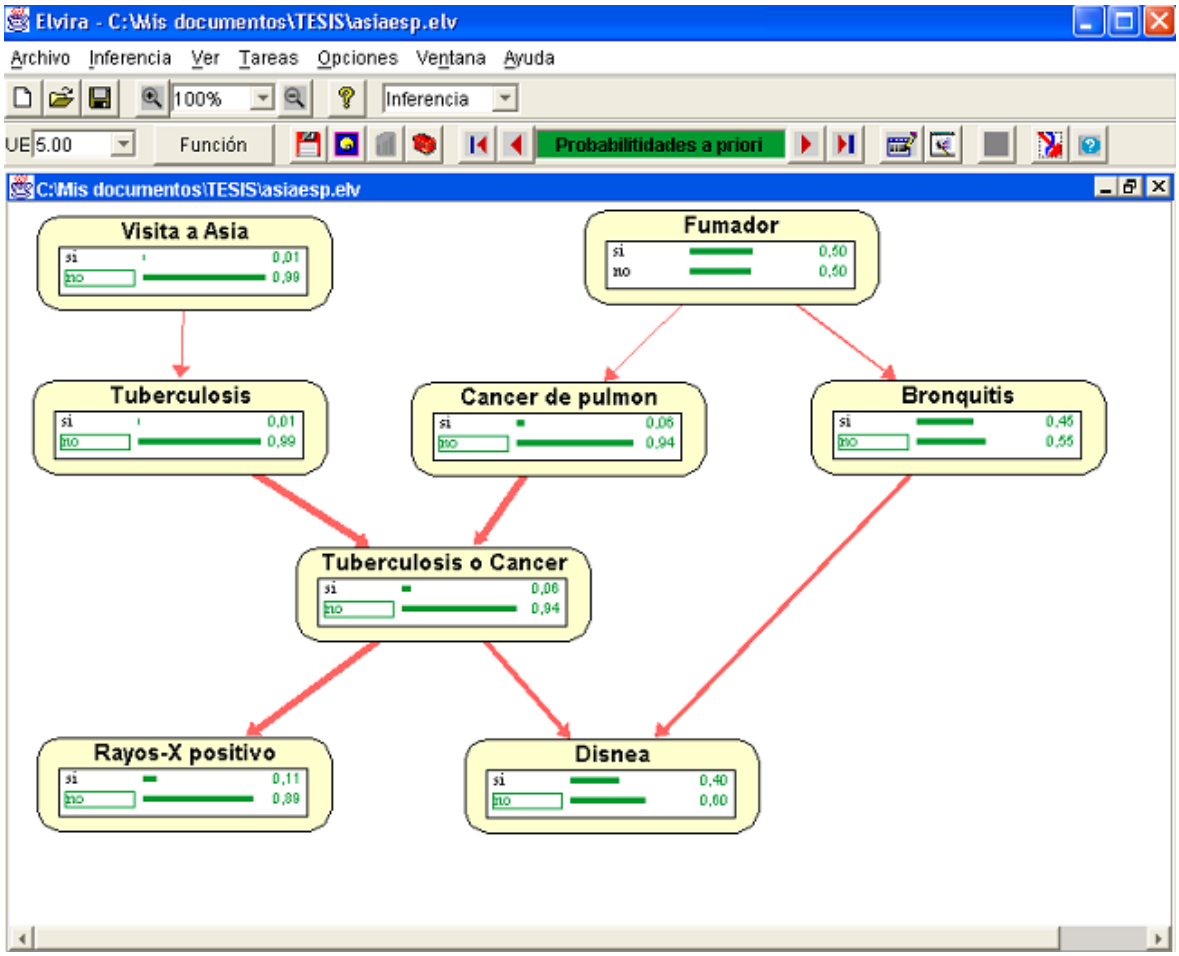

Figura 5.12: Ejemplo de explicación gráfica de los enlaces de la red ASIA

modificar.

Para la explicación gráfica, los enlaces se colorean, tanto en modo edición como en modo inferencia, teniendo en cuenta la influencia que transmiten y su cantidad, según comentamos en la sección 3.2.2. Para representar el tipo de influencia (definición 3.1), los enlaces cuyo impacto es positivo se colorean en rojo, los negativos en azul, los nulos en negro y los de influencia indefinida en violeta. Como consecuencia, en redes causales la mayoría de los enlaces están representados en rojo. Por otra parte, los enlaces se dibujan con distinto grosor, siendo éste proporcional a la magnitud de influencia que transmiten (definición 3.3). En la figura 5.12 vemos que todos los enlaces de la red representan influencias positivas, lo que es de esperar según el significado intuitivo de cada uno de ellos. Por otro lado, cada uno tiene un grosor distinto, lo que aporta información sobre la magnitud de la influencia de cada enlace. Por ejemplo, el nodo Fumador ejerce menor influencia sobre Cáncer de

pulmón que sobre Bronquitis, lo que permite deducir que si un paciente es fumador hará que la probabilidad de padecer Bronquitis varíe más que la de padecer Cáncer de pulmón. Por otra parte, vemos también que ambas influencias son menores que la que ejerce, por ejemplo, Cáncer de pulmón sobre Tuberculosis o Cáncer. Eso permite deducir que si observamos Fumador=sí, la probabilidad de que sus hijos estén presentes no es demasiado diferente de la original. Sin embargo, si observamos Cáncer de pulmón=sí, la probabilidad de que Tuberculosis o Cáncer esté presente es 1 (por ser una puerta OR determinista).

Explicación de la red. Para la explicación *verbal* aparece una nueva ventana de edición que contiene, por defecto, el conocimiento del dominio representado por la red. La idea consiste en utilizar la información de los nodos y de los enlaces y construir un texto coherente y cohesionado que refleje el significado de la red, según la plantilla descrita en la sección 3.2.3. Este texto puede ser modificado por el usuario y seleccionado para cortarlo y pegarlo en cualquier editor de texto, y así poder darle otro formato, guardarlo, imprimirlo, etc. La explicación *gráfica* de la red se puede obtener solicitando la explicación gráfica del conjunto de nodos seleccionado por el usuario (véase la sección 5.2.2), o de todos los enlaces, o de ambas cosas a la vez.

5.3.2. Explicación del razonamiento

Finalmente, en este apartado presentamos las facilidades con las que cuenta ELVIRA para explicar los resultados obtenidos y el razonamiento seguido por el sistema al realizar la propagación de la evidencia, que se basan en la creación y gestión de distintos casos de evidencia y en su correspondiente explicación.

Gestión de casos de evidencia

En cada momento, puede haber varios casos de evidencia (véase definición 4.1) almacenados en ELVIRA en una lista, que se vacía cada vez que se pasa a modo inferencia; uno de los casos almacenados es el que constituye el *caso actual*. Cuando se compila la red por primera vez se genera el *caso a priori* (CE_0), que se inserta como primer elemento de dicha lista. Cada vez que el usuario introduce nueva evidencia, los hallazgos correspondientes se añaden al caso actual, a menos que decida generar un caso nuevo o modificar uno existente. Todos los casos son modificables salvo el primero, por lo que cuando la red se compila por primera vez se genera un segundo caso de evidencia CE_1, que se incluye también en la lista y que es el que constituye el caso actual. El objetivo es que el usuario

no se tenga que preocupar de estar generando nuevos casos la primera vez que desee introducir un hallazgo en la red. El caso CE_1 está vacío, es decir, es exactamente igual al CE_0, puesto que cada vez que se genera un nuevo caso C_i, éste contiene la misma evidencia que el caso que le precede, es decir, $C_i = C_{i-1}$. Además, el nuevo caso creado pasa a ser el *caso actual*, sobre el que se pueden introducir o eliminar hallazgos. El número máximo de casos a almacenar lo determina el usuario junto con el resto de opciones de inferencia (véase la figura 5.6).

Las posibilidades que ofrece ELVIRA relacionadas con la gestión de casos son las siguientes:

- **Visualizar Casos**. En los nodos expandidos, se incluye una barra por cada uno de los casos creados, aunque solo se muestra la probabilidad numérica del caso actual (véase la figura 5.14). El color de las barras y de los rectángulos son específicos de cada caso. Puesto que el color de los nodos que representan los hallazgos que definen la evidencia de cada caso es distinto al del resto, el usuario puede identificarlos fácilmente. La barra de explicación (véase la figura 5.13) contiene botones para

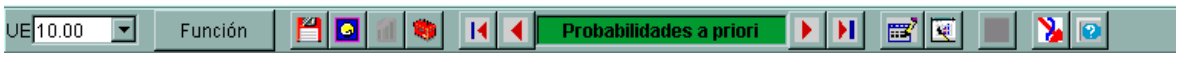

Figura 5.13: Barra de explicación en ELVIRA

generar nuevos casos, expandir o contraer nodos y guardarlos en disco. Además, existe un campo de texto que contiene el nombre y el color del caso actual, junto con botones que permiten "navegar" a lo largo de la lista de casos de evidencia. En la Figura 5.14 podemos ver que se han creado cuatro casos de evidencia: el primero es el caso a priori, que no contiene evidencia; en el segundo, el único hallazgo es Visita a Asia; el tercero está formado sólo por el hallazgo No fumador; y el cuarto, formado por ambos hallazgos, que corresponde en la figura al caso actual.

- **Editor de Casos**. En ELVIRA un hallazgo se puede introducir "haciendo clic" sobre el nodo en cuestión o abriendo el *Editor de Casos*, como se muestra en la figura 5.15. Para ello, se incluyen tres botones:

  - *Nuevo hallazgo* permite introducir un nuevo hallazgo. En ese caso, se crea una nueva fila de la tabla con dos columnas: una para introducir el nombre de la variable observada, mediante una lista desplegable que contiene los nombres de todos los nodos de la red, y otra para introducir su valor, de forma similar, mediante otra lista desplegable.

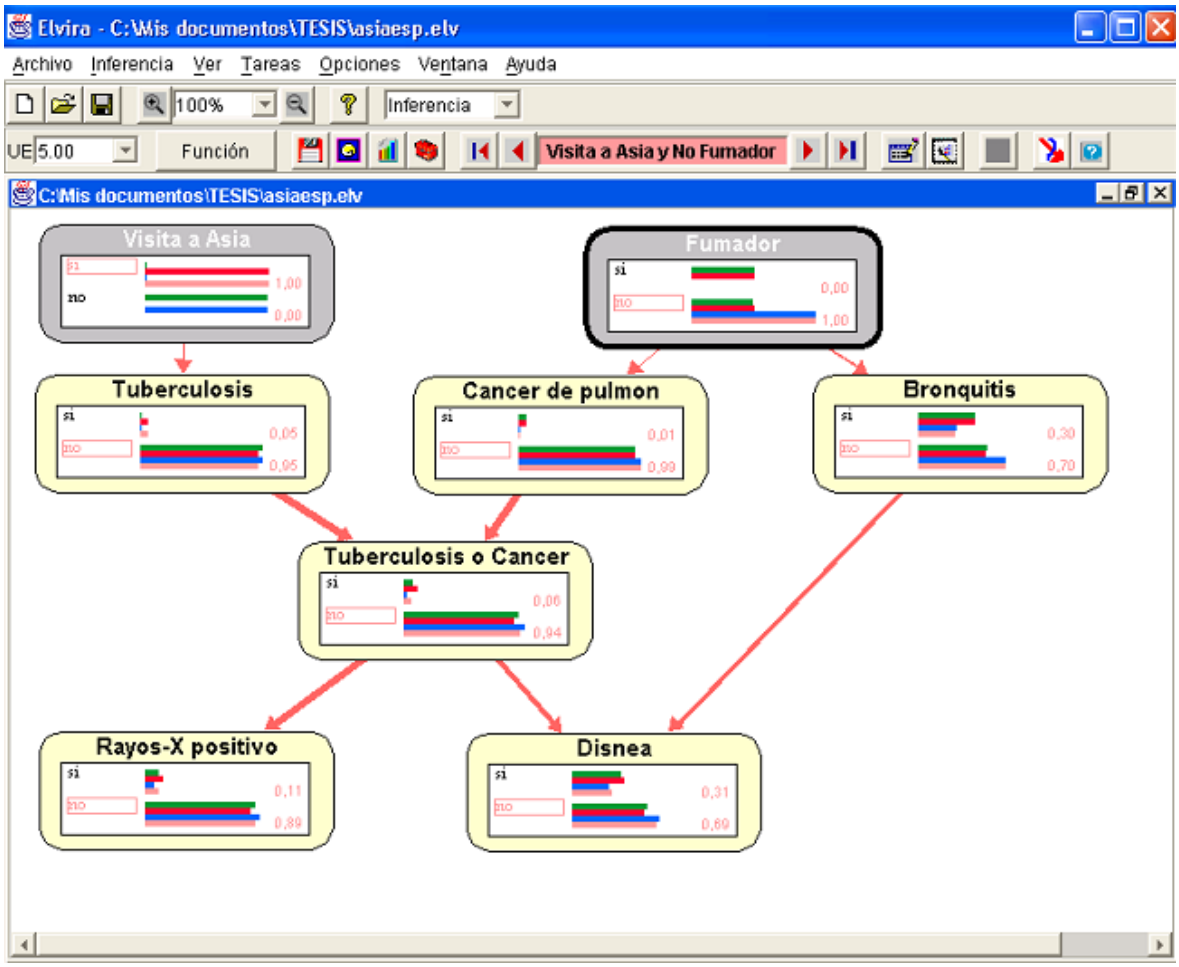

Figura 5.14: Ejemplo de creación de cuatro casos de evidencia en ELVIRA para la red ASIA

- *Borrar hallazgo*, para borrar el hallazgo correspondiente a la fila seleccionada por el usuario.
- *Propagar*, para propagar la evidencia introducida hasta el momento (en el caso de que la propagación se vaya a hacer de modo manual).

La principal utilidad de este editor es introducir o eliminar evidencia cuando el número de variables es tan grande que es imposible o difícil mostrar la red completa en la pantalla.

- **Monitor de casos**. Además de poder crear tantos casos de evidencia como desee el usuario, también puede controlar cuáles de ellos quiere que se le muestren, así como su nombre y su color, mediante el monitor de casos (véase la figura 5.16). Los dos primeros botones sirven, respectivamente, para crear nuevos casos de evidencia y

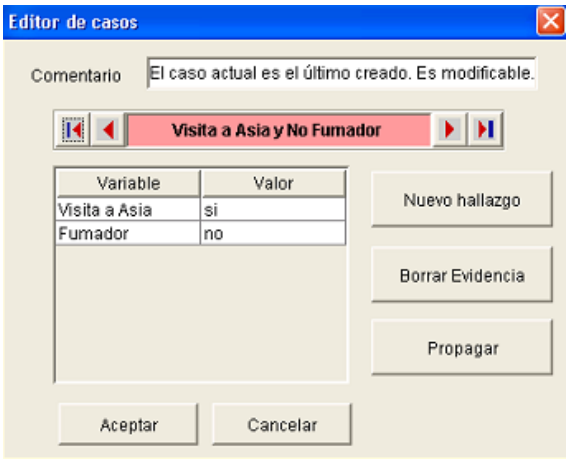

Figura 5.15: Editor de Casos

Figura 5.16: Monitor de casos correspondiente a los casos creados en la figura 5.14

eliminarlos de la memoria. El tercero permite editar casos de evidencia mediante el editor de casos. Cada caso almacenado se corresponde con una fila de la tabla: si la primera columna no está marcada, dicho caso de evidencia no se mostrará en los nodos expandidos; la segunda columna es un campo de texto editable que sirve para modificar su nombre; y la tercera columna permite al usuario cambiar el color asignado a dicho caso. Además, el botón *Explicar* sirve para solicitar la explicación del caso seleccionado, lo que describimos a continuación.

Explicación de casos de evidencia

Desde el monitor de casos se puede solicitar la explicación del caso seleccionado por el usuario tal y como propusimos en los capítulos 3 y 4; en la barra de explicación existe también un botón para explicar el caso actual. En ambos casos, aparece una nueva ventana como la de la figura 5.17, que ofrece la siguiente información:

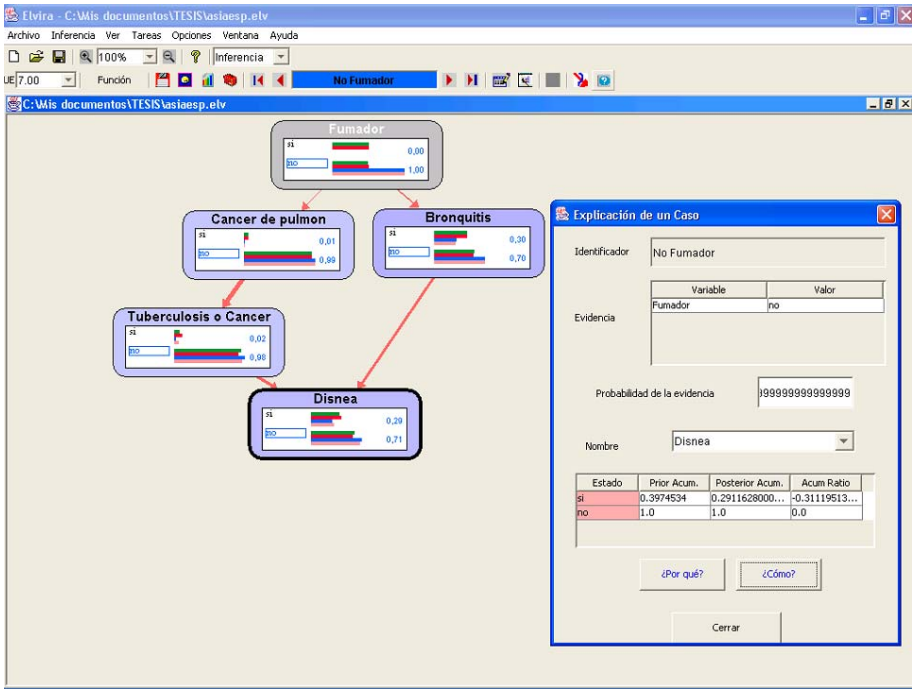

Figura 5.17: Explicación gráfica del caso **No fumador**

- El *nombre* del caso y la *evidencia* asociada.
- La *probabilidad de la evidencia,* porque, especialmente en medicina, ayuda al usuario a estimar la frecuencia de aparición de un caso determinado. Así, si la probabilidad de un conjunto de hallazgos es muy baja, al usuario no debe extrañarle que el sistema ofrezca como resultado un diagnóstico "raro", puesto que la probabilidad de una hipótesis h dada le evidencia e es inversamente proporcional a la probabilidad de la evidencia.
- Un panel que muestra los resultados del *análisis de sensibilidad a la evidencia* [90, 174] realizados sobre una hipótesis determinada seleccionada previamente por el usuario, mediante una lista desplegable que ofrece los nombres de todas las variables de la red. Dichos resultados se presentan en una tabla que contiene el nombre de

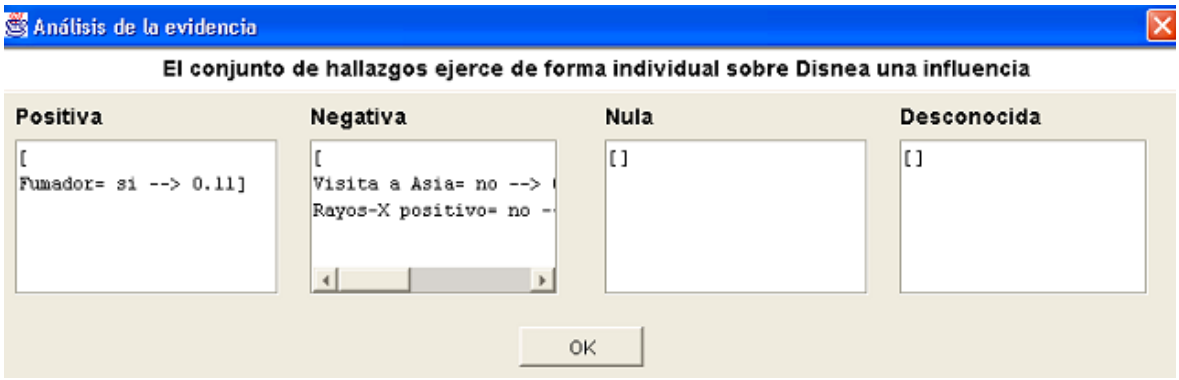

Figura 5.18: Clasificación de los hallazgos del caso C=Visita a Asia=no, Fumador=sí, Rayos-X positivo=no y variable de interés Disnea.

cada uno de sus estados junto con su probabilidad a priori, su probabilidad después de propagar la evidencia del caso, y la razón logarítmica de ambos valores. Por ejemplo, en la figura 5.17 la hipótesis seleccionada es Disnea.

- El botón *Por qué* permite ver en una nueva ventana la clasificación de los hallazgos dependiendo del tipo y la magnitud de la influencia que ejercen sobre una variable determinada, tal y como se ilustra en la figura 5.18 (definiciones 4.6 y 4.7).

- El botón *Cómo* muestra al usuario la cadena o cadenas de razonamiento seguidas en la red desde la evidencia hasta la hipótesis (sección 4.2.4). Cuando se presiona este botón, **sólo** se muestran al usuario los caminos desde los hallazgos de la evidencia hasta la hipótesis. Los nodos de estos caminos se colorean dependiendo del tipo de influencia recibida. En ELVIRA, como en casos anteriores, las influencias positivas se colorean en rojo, las negativas en azul y las desconocidas, en violeta. Los nodos con influencia nula no varían el color, por lo que se muestran en amarillo, el mismo que tienen inicialmente en ELVIRA.

  En la figura 5.17 podemos observar los caminos de razonamiento entre Fumador, que constituye el único hallazgo del caso de evidencia seleccionado, y Disnea, que es la variable de interés seleccionada. Como vemos, todos los nodos se han coloreado con el color azul, lo que es lógico puesto que como todas las influencias son negativas, al tener como evidencia el valor Fumador=no, esto hace que las probabilidades de que sus hijos tomen el valor "sí" disminuyen. De forma similar, esta disminución se va transmitiendo de cada nodo a sus hijos y deducimos que, entonces, la probabilidad de padecer Disnea ha disminuido con respecto al caso a priori.

Hay que destacar también que los nodos se colorean con distinta intensidad, dependiendo de la variación que ha sufrido la probabilidad de alcanzar un valor "anormal", según la definición 4.5 (véase sección 4.2.3). Por eso, en la figura 5.17 observamos que el color de la variable Bronquitis es más intenso que el de la variable Cáncer de pulmón, puesto que la probabilidad de ésta ha variado menos que la de Bronquitis.

Opciones de explicación

De modo similar a como ocurría con la inferencia, existe también la posibilidad de seleccionar distintas opciones de explicación mediante una ventana como la de la figura 5.19, en la que se le ofrecen al usuario las siguientes posibilidades de elección:

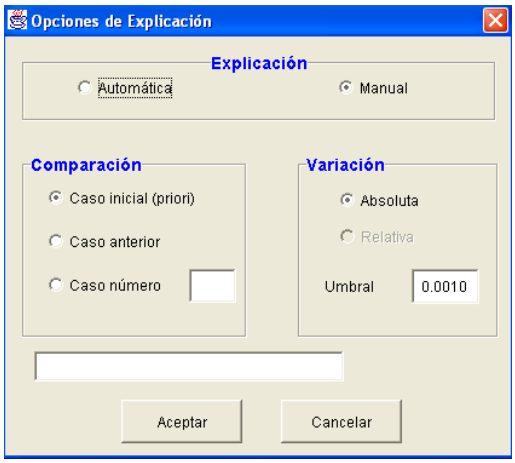

Figura 5.19: Ventana para seleccionar las distintas opciones relacionadas con la explicación de casos de evidencia

- *Caso de evidencia* sobre el que realizar el análisis de sensibilidad. Por defecto, las comparaciones se realizan siempre sobre el caso a priori, salvo que el usuario seleccione el caso previo, con lo que el análisis se realizará tomando como referencia el caso anterior al actual; o bien que indique la posición del caso con respecto al cual realizar las comparaciones. De este modo, al poder seleccionar el caso con respecto al cual hacer el análisis, el usuario dispone de una herramienta gráfica para poder estudiar las variaciones de las probabilidades de los nodos.

- *Umbral de variación.* Además de seleccionar el caso sobre el que realizar el análisis, el usuario puede seleccionar la cantidad mínima necesaria para considerar las variaciones en las probabilidades de los nodos. Esta variación puede medirse en términos

absolutos o relativos y, en cada caso, es el usuario quien establece el umbral mínimo de variación. Por defecto, dichas variaciones se miden en valores absolutos y el umbral establecido es de 10^{-3}.

- *Modo de explicación.* La explicación cualitativa proporcionada por ELVIRA para una variable determinada se puede obtener también para todos los nodos de la red. Para ello, el usuario simplemente debe seleccionar en esta ventana la opción de explicación automática. En este caso, todos los nodos no observados de la red se colorean dependiendo de las influencias emitidas por la evidencia, tomando como referencia el caso de evidencia seleccionado.

5.4. Comparación con otras aplicaciones

En esta sección presentamos las características de las aplicaciones (comerciales o no) que existen en la actualidad para la edición y manipulación de redes bayesianas cada una de ellas, haciendo hincapié en las diferencias con ELVIRA.

Aunque existen numerosos programas para la edición y procesamiento de redes bayesianas[6], nos centraremos en el análisis de GeNIe [53][7], HUGIN [8], JavaBayes [9], Netica [10], MSBN [11] y XBaies [12], por las razones siguientes:

- **HUGIN** fue una de las primeras herramientas que se creó para la edición y evaluación de diagramas de influencia y redes bayesianas. Además, HUGIN es en la actualidad, junto con **Netica**, la aplicación más usada para el desarrollo y procesamiento de redes bayesianas.

- **GeNIe** y **JavaBayes** son aplicaciones no comerciales desarrolladas por grupos de investigación con los que nuestro equipo trabaja o ha trabajado en alguna ocasión. Concretamente, las primeras versiones de ELVIRA estuvieron inspiradas en la interfaz y en los algoritmos implementados en JavaBayes.

[6] Se pueden encontrar la mayoría de ellas en `http://www.ia.uned.es/~fjdiez/bayes/#software` y en `http://www.ai.mit.edu/~murphyk/Bayes/bnsoft.html`.

[7] GeNIe: a Graphical Network Interface. `http://www2.sis.pitt.edu/genie`

[8] Hugin Expert A/S. Hugin Lite 6.1. `http://www.hugin.com`

[9] F.G. Cozman. JavaBayes version 0.346. `http://www.cs.cmu.edu/~fgcozman/home.html`

[10] Netica 1.12. `http://www.norsys.com`

[11] Microsoft Research. MSBNx, version 1.4.2. `http://www.msbn.com`

[12] R.G. Cowell. Xbaies, version 2.0. `http://www.xbaies.com`

- **XBaies** fue la herramienta de libre distribución que utilizamos para la creación de nuestra primera red bayesiana, como fruto de un trabajo para un curso de doctorado.
- **MSBN** es la herramienta comercial diseñada por Microsoft para plataformas Windows. Puesto que la mayoría de los usuarios de PC tienen instalado el SO Windows, la hemos elegido también por la repercusión que puede tener.

El estudio de las distintas herramientas lo realizaremos analizando las tres tareas que más relevancia tienen para nuestro trabajo: edición, inferencia y explicación.

Edición

Las facilidades para edición proporcionadas por todas las herramientas analizadas, salvo XBaies, consisten en permitir al usuario construir y evaluar redes de **modo gráfico**. Para construir una red, el usuario simplemente debe ir añadiendo nodos y enlaces mediante el uso de menús o "haciendo clic" con el ratón en el lugar donde desee situarlos. En este sentido, ELVIRA es similar a todas las demás herramientas. Además, ELVIRA, XBaies, Netica y Hugin permiten hacer *zoom* del modelo representado (GeNIe lo incorporará en la versión 2.0); sin embargo, la posibilidad de *deshacer* y *rehacer* los últimos cambios realizados, sólo la proporcionan ELVIRA y Netica. En GeNIe, Netica, HUGIN, MSBN y XBaies se pueden seleccionar las características gráficas del modelo, como el color de los nodos, etc. Otras diferencias significativas entre unas herramientas y otras radican en que:

- GeNIe y Netica ofrecen la posibilidad de definir submodelos, lo que resulta muy útil de cara a la edición gráfica de redes en los que el número de nodos es grande, pues se pueden construir mediante refinamientos sucesivos.
- A la hora de introducir las probabilidades en MSBN, además de permitir crear las tablas de probabilidad como el resto de herramientas, el usuario puede refinar los valores ayudándose de un diagrama de barras que se puede ajustar manualmente.
- La interfaz gráfica más simple es JavaBayes como se puede apreciar en la figura 5.20.
- ELVIRA, GeNIe y MSBN permiten construir tablas de probabilidades para modelos canónicos: en el caso de MSBN solo para la puerta OR residual; en GeNIe, para la puerta OR, MAX y para la puerta AND. Hay que indicar que la nueva versión de GeNIe, GeNIe 2.0, que se espera que aparezca a principios del año 2003, entre otras

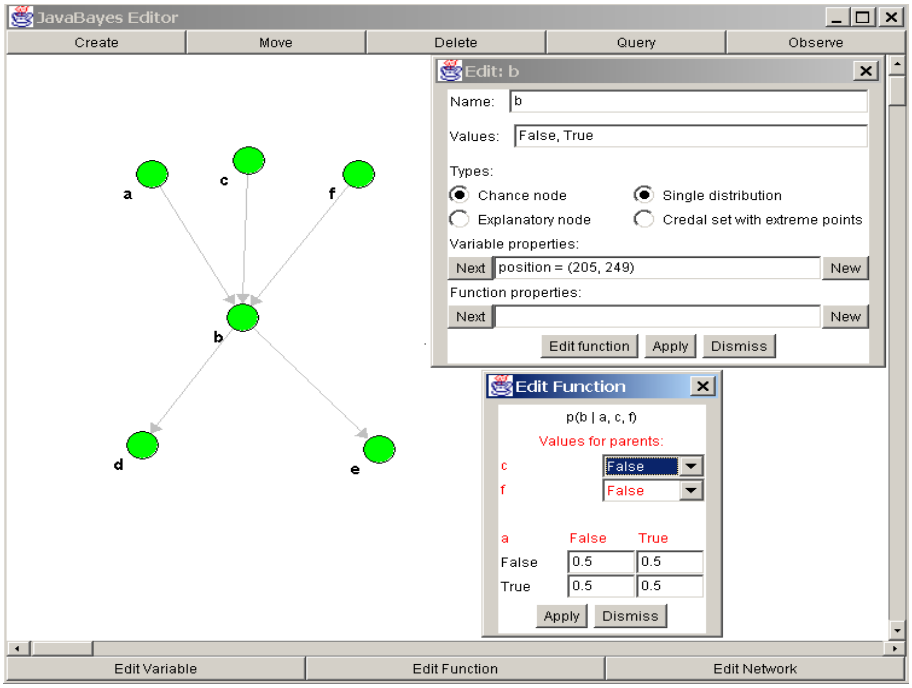

Figura 5.20: Interfaz gráfica de JavaBayes.

muchas mejoras con respecto a la edición anterior, proporciona novedosas formas de obtención de probabilidades, especialmente útiles para la asignación de datos en forma cualitativa o en los casos en los que los nodos tienen una gran cantidad de padres [183].

- Netica y HUGIN son las únicas herramientas que permiten trabajar con variables continuas. ELVIRA lo ofrece pero no a través de la interfaz gráfica; está previsto integrarlas próximamente.

En XBaies el proceso de construcción de una red bayesiana consiste en realizar una serie de pasos secuenciales mediante los que primero se define el dominio, posteriormente se crean los nodos, después las relaciones, etc. Y todo ello mediante la ayuda de ventanas en las que hay que rellenar una serie de campos en un orden determinado, que sólo se indica en el manual de instrucciones. Como, además, dicho orden no es demasiado intuitivo, si por casualidad no se sigue provoca que las acciones realizadas no tengan ningún efecto, con la pérdida de tiempo que conlleva, o en caso contrario, los posibles errores que se pueden producir. Como ejemplo, en la figura 5.21 se muestra la ventana para la creación de variables.

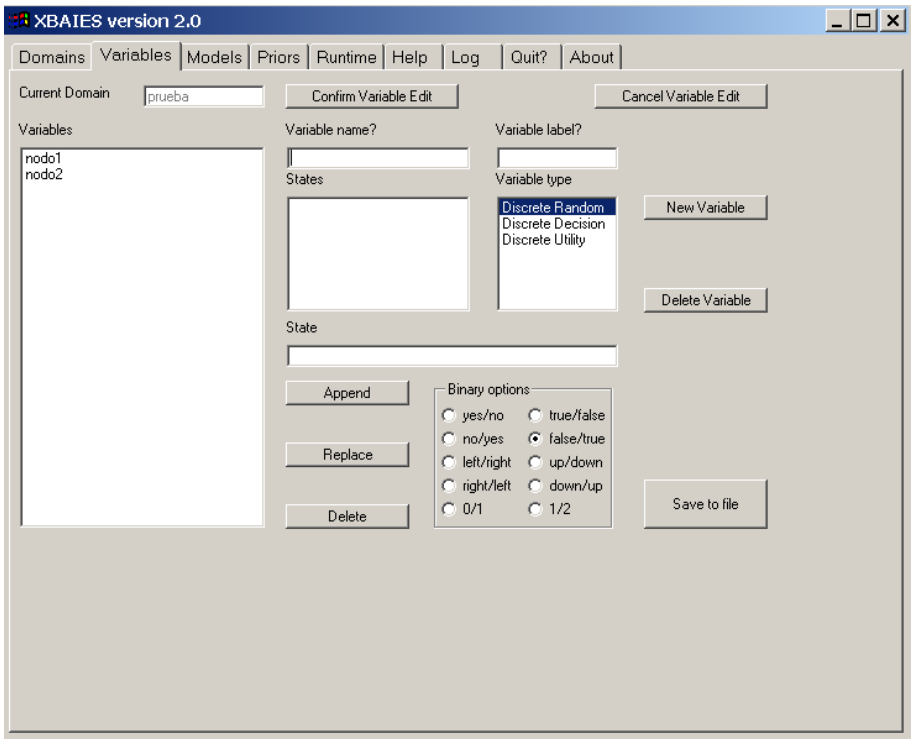

Figura 5.21: Ventana para la creación de variables en XBaies 2.0.

Inferencia

En el apartado relativo a la propagación de la evidencia y obtención de resultados, realizaremos el análisis de acuerdo con las principales cuestiones para el usuario relacionados con la inferencia, a saber:

- **Introducción de evidencia.** En todos los casos, salvo en JavaBayes, la evidencia se introduce después de compilar la red. Las herramientas que permiten editar gráficamente la red (todas las analizadas excepto XBaies), facilitan también la introducción de nuevos hallazgos "haciendo clic" sobre el nodo que se desea observar, destacando de algún modo el nodo o el hallazgo observado. También tienen en común la posibilidad de visualizar los nodos de la red en una lista e introducir la evidencia a partir de ella.

- **Tareas.** Todas las herramientas permiten, como mínimo, *propagar la evidencia* introducida para obtener las probabilidades a posteriori de las variables no observadas. En GeNIe existe también la posibilidad de realizar *razonamiento basado en relevancia*, que consiste en identificar las variables no observadas relacionadas con la variable o variables de interés seleccionadas por el usuario y calcular sus probabilidades a posteriori (véase la figura 5.22); si no hay ninguna, se obtienen las probabilidades a

posteriori de todas las variables no observadas. Los algoritmos para la identificación de las variables relevantes a una dada se describen en [57] y [106]. En ELVIRA,

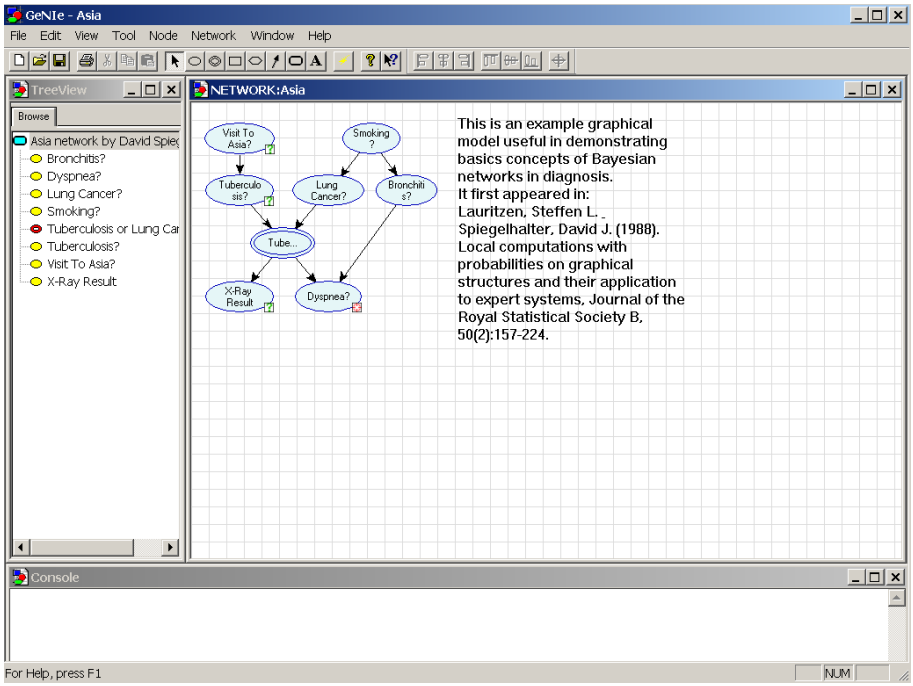

Figura 5.22: Ejemplo de razonamiento basado en relevancia en GeNIe para la red ASIA [103].

HUGIN, Netica, JavaBayes y XBaies es posible realizar *abducción* total y/o parcial, mediante la que se muestran, tanto de forma gráfica como textual, las configuraciones más probables junto con sus probabilidades. Además, en ELVIRA, GeNIe, Netica y HUGIN se puede realizar *análisis de sensibilidad.* En el caso de GeNIe hay que añadir un nodo de decisión relacionado mediante un arco con el nodo que representa la variable sobre la que se desea hacer el análisis. Por otra parte, una de las principales utilidades de XBaies es la posibilidad de mostrar los resultados correspondientes a otras tareas, como es el árbol de eliminación, las fases de propagación en el árbol de uniones, etc.

- **Algoritmos de propagación** En ELVIRA y GeNIe, el usuario tiene la posibilidad de elegir el algoritmo que desea utilizar para realizar la inferencia, aunque por defecto, en ELVIRA se utiliza el mismo que en HUGIN [91]y en GeNIe, el de agrupamiento de Lauritzen y Spiegelhalter [103]. En JavaBayes existen dos métodos implementados: eliminación de variables y árbol de eliminación de grupos [31]. En los demás, sólo se ha implementado un algoritmo y el usuario no tiene posibilidad de

elección: XBaies y MSBN utilizan el de Nilsson [132]; Netica, el de paso de mensajes en árboles de grupos (join tree), [89, 129, 166].

- **Visualización de resultados.** JavaBayes y XBaies ofrecen los resultados de la propagación en forma de *texto* en una ventana distinta a la que contiene la representación gráfica de la red. En GeNIe hay que seleccionar el nodo del que se desea conocer su probabilidad y editar sus propiedades para ver los resultados. Sin embargo, en ELVIRA, HUGIN, MSBN y Netica existe la posibilidad de visualizar *gráficamente* los resultados de los estados de cada nodo sobre el propio grafo de la red mediante diagramas de barras. En Netica todos los nodos tienen el mismo tamaño; en HUGIN, sin embargo, aparece una barra de desplazamiento para moverse por los estados por lo que en nodos con muchos estados (más de 6) no es posible verlos todos simultáneamente. En el caso de MSBN, al evaluar la red se genera una ventana nueva en la que se muestran las probabilidades de los nodos seleccionados por el usuario, tanto con sus valores numéricos como mediante diagramas de barras.

- **Otras facilidades.** En MSBN, los modelos creados se pueden clasificar en tres grupos: diagnóstico, solución de problemas u otro, definido por el usuario. En los dos primeros casos, dada cierta evidencia, el programa hace sugerencias sobre ciertos hallazgos no observados que ayudarían a realizar un diagnóstico determinado. Para ello, se muestra una lista ordenada de los hallazgos junto con el *valor de la información* que aporta cada nodo para realizar un diagnóstico claro. En Netica existe la posibilidad de introducir hallazgos *negativos* que son aquéllos de los que sabemos con certeza que **no** tienen un valor determinado; por ejemplo, un hallazgo negativo para la variable altura, con estados bajo, normal y alto, podría ser: altura=no alto. Además, permite guardar en archivos distintos casos de evidencia para poder recuperarlos después, al igual que en ELVIRA. En MSBN, GeNie y JavaBayes existe la posibilidad de guardar la red en distintos formatos, aunque no existe ninguno común a los tres. Sin embargo, es curioso que en GeNie se pueden guardar las redes en formato de HUGIN. En todas ellas se han implementado opciones para el aprendizaje de redes bayesianas.

Explicación

Aunque las capacidades de explicación que incluyen las herramientas estudiadas son bastante rudimentarias, para su presentación tendremos en cuenta las propiedades descritas en la tabla 2.1:

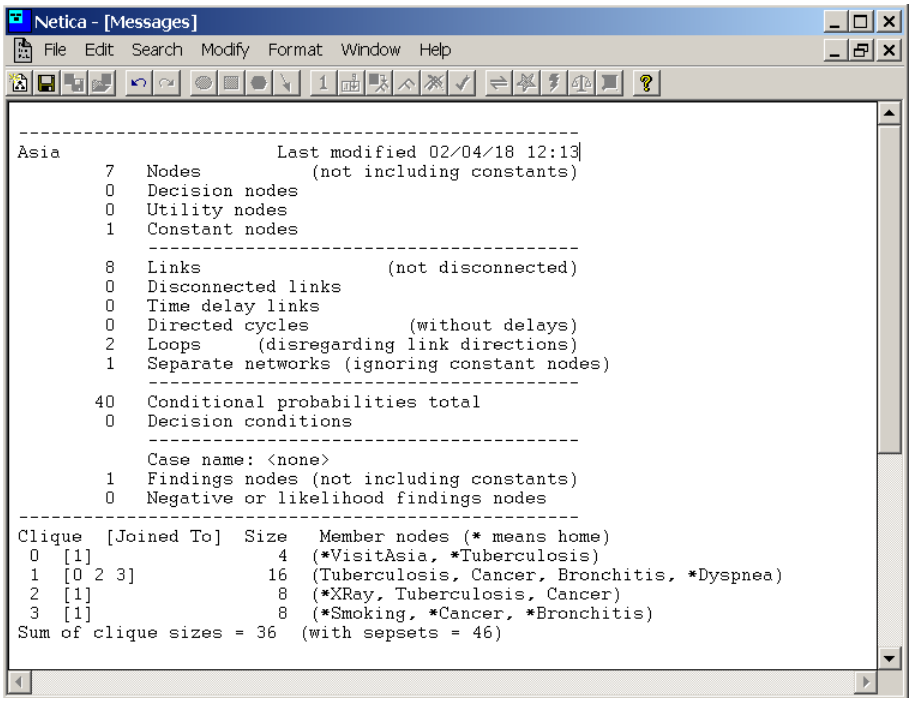

Figura 5.23: Ejemplo de explicación verbal en Netica de la red ASIA.

- **Contenido.** La característica común a todas las herramientas es que las facilidades de explicación se limitan simplemente a mostrar las probabilidades obtenidas al realizar ciertas tareas, por lo que el propósito común a todas es el de la *descripción* de los resultados. Como ya hemos indicado en el apartado anterior, en ELVIRA, HUGIN, Netica, JavaBayes y XBaies se ofrecen explicaciones de la *evidencia*, con el objeto de describir el estado del sistema que mejor explica los hallazgos introducidos.

  Ninguna herramienta, salvo ELVIRA, ofrece explicación textual del *modelo*. Tan solo en GeNIe, HUGIN y Netica se puede almacenar una descripción del nodo durante el proceso de edición del mismo. Netica ofrece al usuario la posibilidad de generar una descripción *verbal* del modelo representado en la red y guardarla con ella. Lo que el programa genera automáticamente es un informe, que el usuario puede editar, sobre determinadas propiedades de la red, tanto estáticas (número de nodos, enlaces, etc.) como dinámicas (árbol de grupos, probabilidad de la evidencia, etc.). Un ejemplo de la explicación generada por Netica para la red ASIA lo podemos ver en la figura 5.23. Además, sólo ELVIRA proporciona la explicación gráfica de los enlaces.

  En cuanto a la explicación *macro* del *razonamiento*, en ELVIRA, GeNIe, Netica y

Hugin se puede realizar el análisis de sensibilidad sobre una variable respecto a la evidencia introducida, con el propósito de describir cómo los hallazgos afectan a la probabilidad de cierta variable. Sin embargo, sólo ELVIRA es capaz de representar gráficamente los resultados del análisis de sensibilidad, además de mostrar los caminos de razonamiento y de clasificar los hallazgos en función del tipo de impacto que ejercen sobre una variable. En XBaies se puede visualizar a grandes rasgos el proceso de propagación de la inferencia, puesto que el usuario puede ver cómo se moraliza el grafo, el orden de eliminación de cada variable, el árbol de eliminación, así como los resultados de muchas otras operaciones relacionadas con la propagación y el aprendizaje. El único problema, insistimos, es que la interfaz no es muy intuitiva ni fácil de utilizar.

- **Comunicación.** La interacción usuario-sistema se lleva a cabo exclusivamente mediante *menús* y la presentación de las explicaciones se realiza de forma *gráfica*, aunque normalmente se limita a representar las probabilidades en forma de diagramas de barras, bien sobre el propio grafo de la red, como en HUGIN (véase la figura 5.24) y Netica, o bien en otra ventana distinta, como en MSBN (véase la figura 5.25).

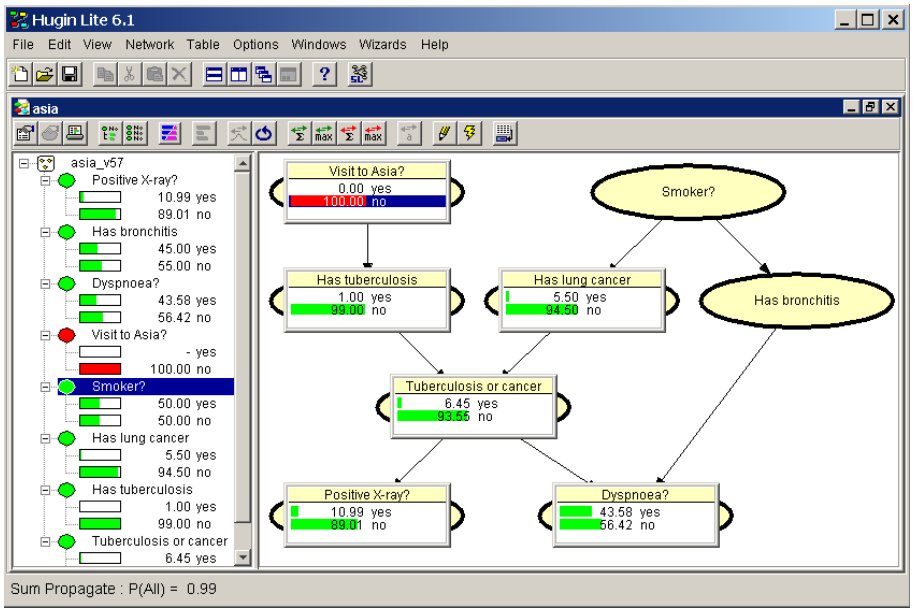

Figura 5.24: Ejemplo de visualización en HUGIN de la red ASIA

En XBaies se pueden visualizar de forma gráfica ciertas propiedades cualitativas de la red, como la independencia de unos nodos respecto a otros, los antecedentes de

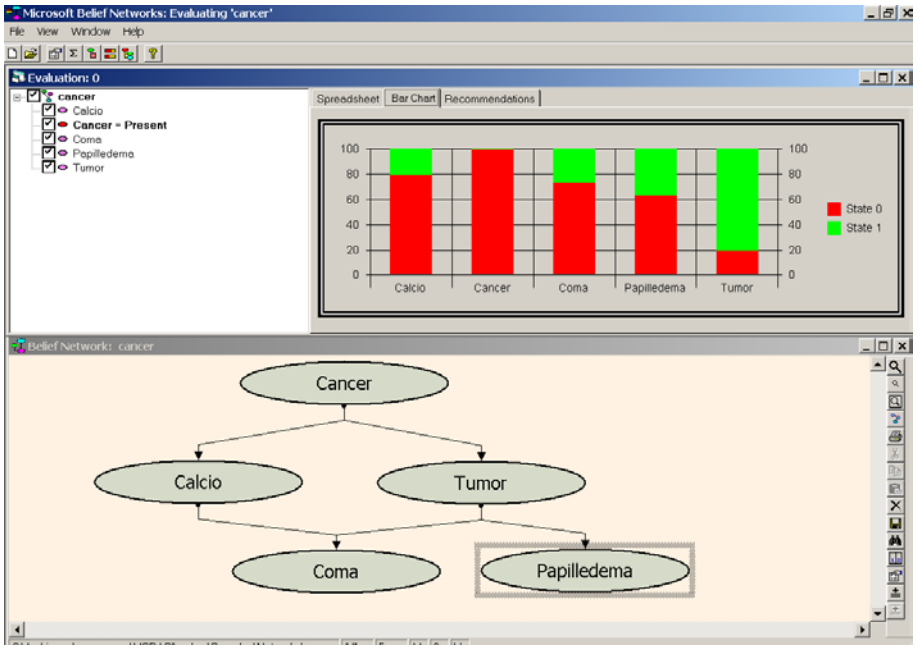

Figura 5.25: Ejemplo de explicación gráfica en MSBN de la red CANCER [129].

un nodo, etc., lo que ayuda a comprender el algoritmo de propagación. En todas estas herramientas, en las que los resultados se presentan gráficamente, también es posible visualizarlos de forma *textual*, que es la forma en la que se muestran en ELVIRA, GeNie y en JavaBayes. En todos los casos, las probabilidades se expresan sólo de forma *numérica*.

- **Adaptación.** En ninguna de las herramientas analizadas existe la posibilidad de adaptar las explicaciones a distintos modelos de usuario, ni según el conocimiento que tenga sobre el dominio representado ni sobre los métodos de razonamiento. Además, el nivel de detalle de las explicaciones es *fijo*, orientado a un usuario que tenga experiencia en el campo de las redes bayesianas o los diagramas de influencia y los algoritmos de propagación.

A la luz de este análisis podemos concluir que, a pesar de que las herramientas actuales ofrecen grandes facilidades para el procesamiento de redes bayesianas, sin embargo ninguna de ellas es lo suficientemente completa, aunque ELVIRA constituye una mejora con respecto a las demás en los aspectos relacionados con la explicación.

Por el contrario, aún posee ciertas limitaciones, que está previsto mejorar en el futuro, entre las que destacan su poca robustez, la imposibilidad de representar submodelos y la lentitud de los algoritmos, además de no ofrecer ningún tipo de ayuda al usuario, ni en

línea ni de otro tipo. Tan sólo cuando se realizan operaciones en la interfaz que no están permitidas, como por ejemplo la inserción de un arco que genere un ciclo, o se introduce evidencia imposible, etc. ELVIRA informa del error o de la imposibilidad de realizar dicha operación.

Parte III

APLICACIÓN

Capítulo 6

Aplicación: Diagnóstico del cáncer de próstata

Al disponer en ELVIRA de un método de explicación, decidimos ponerlo a prueba mediante su aplicación a la urología. El cáncer de próstata es el tumor maligno más común entre los hombres mayores de 50 años de edad y la segunda causa de muerte por cáncer (el cáncer de pulmón es la primera). La probabilidad de recuperación depende de la etapa del cáncer (si se encuentra localizado exactamente en la próstata o se ha diseminado a otras partes del cuerpo) y de la salud del paciente en general. Por eso, es importante diagnosticarla en una fase temprana. Sin embargo, los síntomas del cáncer de próstata son muy parecidos a los de la hiperplasia prostática benigna (HPB) o de otros problemas de la próstata, por lo que es fácil confundirlos. Por ello, resulta útil disponer de una herramienta que ayude al médico, especialmente de medicina general, a realizar un diagnóstico diferencial entre las posibles enfermedades basado en sus probabilidades.

En este capítulo, hacemos una introducción de los principales conceptos médicos relacionados con el cáncer de próstata y su diagnóstico (secciones 6.1 y 6.2.1), presentamos el proceso de construcción de la red bayesiana PROSTANET para el diagnóstico del cáncer de próstata (sección 6.3) y su correspondiente evaluación (sección 6.4).

6.1. Conceptos médicos

Comenzaremos en esta sección haciendo un repaso de la anatomía, fisiología y patología de la próstata para que el lector no experto en el tema pueda familiarizarse con la terminología que va a utilizarse a lo largo del capítulo. No obstante, no pretende ser éste un tratado médico sobre la próstata sino un resumen de las principales características de cada apartado. Para profundizar sobre cualquiera de los temas que aquí aparecen recomendamos las siguientes referencias [3, 7, 117, 190].

6.1.1. Anatomía y fisiología de la próstata

La próstata es una de las glándulas sexuales masculinas que tradicionalmente se ha descrito como un cono invertido. Está situada en la pelvis por detrás del pubis, por delante del recto y debajo de la vejiga urinaria (ver figura 6.1[1]).

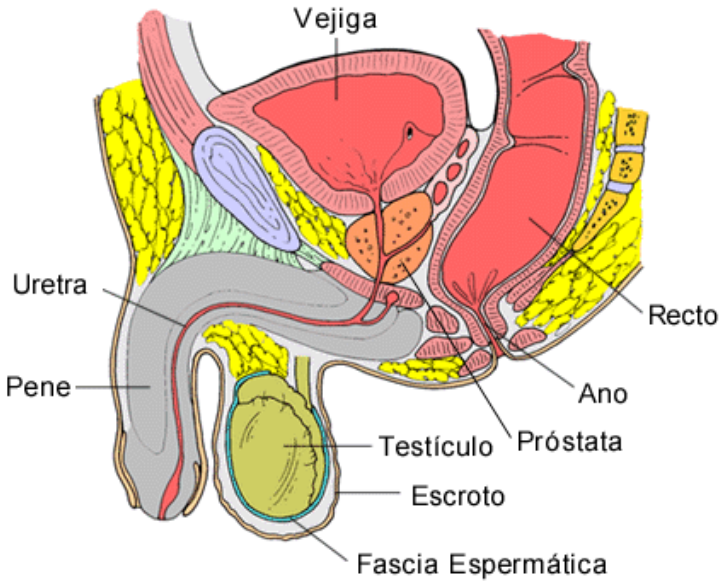

Figura 6.1: Anatomía de la glándula prostática

Tiene aproximadamente el tamaño de una nuez y rodea parte de la uretra, el tubo que transporta la orina de la vejiga al exterior del cuerpo. Pesa alrededor de 20 g. y se encuentra envuelta en una cápsula de tejido fibroconectivo. Es parcialmente muscular y parcialmente glandular, con conductos que se abren a la porción prostática de la uretra. Consta de tres lóbulos: un lóbulo central con un lóbulo a cada lado. En la actualidad no se conocen todas las funciones de la glándula prostática. Sin embargo, se sabe que la glándula prostática juega un papel importante tanto en la función sexual como en la urinaria.

La glándula prostática secreta un líquido lechoso alcalino que contiene ácido cítrico, calcio, fosfatasa ácida y fibrolisina, entre otros, y que forma parte del líquido seminal. La estimulación alfaadrenérgica genera la secreción prostática hacia la porción ampular del conducto deferente, por el que circulan los espermatozoides generados en los testículos, y a la uretra posterior. Esta misma estimulación alfaadrenérgica es responsable de cerrar el cuello vesical durante la eyaculación. El papel de estas secreciones en la fertilización

[1]Tomado de `http://www.uuhsc.utah.edu/healthinfo/spanish/Prostate/index.htm`

se desconoce con exactitud. Sin embargo se ha visto que el zinc tiene una función bacteriostática, estabiliza el núcleo del espermatozoide y mejora la capacidad de fertilización. Es común que la glándula prostática se agrande a medida que el hombre envejece y la mayor parte de los hombres en algún momento de su vida tiene algún tipo de problema prostático.

Los andrógenos son los principales reguladores del crecimiento y actividad de la próstata. La castración antes de la pubertad provoca un desarrollo anormal de la glándula. La castración después de la pubertad produce una involución glandular y atrofia. Los andrógenos adrenales participan en el crecimiento prostático y se ha demostrado experimentalmente que estimular con hormona adrenocorticotrópica a animales castrados puede inducir el crecimiento prostático.

6.1.2. Patología de la próstata

Entre las alteraciones más comunes de la próstata, y que pueden afectar a hombres de todas las edades, se incluyen las siguientes:

Prostatitis: inflamación de la glándula prostática que puede ir acompañada de malestar, dolor, micción frecuente y, algunas veces, de fiebre.

Hiperplasia prostática benigna (HPB): también llamada hipertrofia prostática benigna, define la condición de una próstata agrandada y representa el problema prostático no canceroso más común. Aproximadamente a los 40 años, en la glándula prostática se inicia un crecimiento singular que constituye lo que conocemos con el nombre de hiperplasia. Este crecimiento puede obstruir el flujo de orina o interferir con la función sexual. Una glándula prostática agrandada podría necesitar tratamiento con medicamentos o cirugía para aliviar los síntomas. Esta condición prostática benigna común puede causar muchos de los mismos síntomas que el cáncer de próstata.

Incontinencia urinaria: pérdida del control de la vejiga.

Cáncer de próstata, descrito con detalle en la siguiente sección.

6.1.3. Carcinoma de la próstata

Este tipo de tumor es raro antes de los 50 años y aumenta su incidencia a partir de esa edad, de forma que el 50 % de los hombres de más de 80 años tienen cáncer de próstata.

En la mayoría de los hombres afectados, el cáncer es de crecimiento lento y muchos de ellos no mueren debido al cáncer sino por otras causas. Aunque las verdaderas causas del cáncer de próstata se desconocen, sí se pueden enumerar una serie de factores de riesgo. Hay que insistir en que, por ser factores de riesgo, conocerlos puede ayudar a tomar las decisiones apropiadas con el fin de evitar la enfermedad. En el caso del cáncer de próstata se incluyen los siguientes:

Edad. La edad es un factor de riesgo para el cáncer de próstata, sobre todo en los hombres de 50 años de edad o mayores. Más del 80 % de todos los cánceres de próstata se les diagnostican a hombres mayores de 65 años de edad. En cuanto a la incidencia por grupos de edad, va del 15 % en hombres de edades comprendidas entre 50 y 60 años, aumenta hasta el 30 % en la séptima década de la vida y llega hasta el 50 % de los mayores de 80.

Dieta. Los datos epidemiológicos sugieren que la dieta de los países occidentales industrializados puede ser uno de los factores que más contribuyen en el desarrollo del cáncer de próstata [190]. Los estudios han demostrado que los hombres que consumen dietas de alto contenido en grasas tienen más probabilidades de desarrollar cáncer de próstata. La cantidad de fibra en la dieta puede influir en los niveles circulantes de testosterona y estradiol, los cuales, a su vez, pueden disminuir el progreso del cáncer de próstata. Además de disminuir la ingestión de grasas, otra gran diferencia entre las dietas de Asia y de Estados Unidos es el consumo de soja, con un promedio de 35 gramos diarios por persona. La soja contiene isoflavones, y varios estudios han demostrado que éstos inhiben el crecimiento del cáncer de próstata. Se ha demostrado que la vitamina E, un antioxidante, combinada con el selenio, inhibe el crecimiento de tumores en animales en el laboratorio. Además, se ha demostrado que los carotenoides que contienen licopenos inhiben el crecimiento de las células cancerosas prostáticas humanas en cultivos de tejidos. La fuente principal de licopenos es el tomate.

Raza. El cáncer de próstata es casi dos veces más frecuente entre los hombres afroamericanos que entre los hombres americanos caucásicos (de raza blanca). Los hombres japoneses y chinos residentes en sus países de origen tienen los índices más bajos de cáncer de próstata. Es interesante que cuando los hombres japoneses y chinos emigran a Estados Unidos, los índices de riesgo y mortalidad del cáncer de próstata aumentan. En Japón, la incidencia del cáncer de próstata ha aumentado desde que se han adoptado dietas y estilos de vida occidentales. Por tanto, parece ser que la

raza sólo influye a través de la dieta.

Obesidad. La obesidad no solamente contribuye a la diabetes y al colesterol alto, sino que también se ha asociado con algunos cánceres comunes, incluyendo los tumores relacionados con las hormonas, como los cánceres de próstata, de mama y de ovario.

Historia familiar de cáncer de próstata Si el padre o un hermano tienen cáncer de próstata, el riesgo de desarrollar la enfermedad se duplica. El riesgo es aún más alto para los hombres que tienen varios familiares afectados, particularmente si los familiares eran jóvenes cuando se les diagnosticó la enfermedad. Los geneticistas dividen a las familias en tres grupos según el número de hombres con cáncer de próstata y las edades a las que se les diagnosticó la enfermedad:

Esporádico: familia en la que se le ha diagnosticado cáncer de próstata a un hombre a la edad típica;

Familiar: familia en la que el cáncer de próstata afecta a más de una persona, pero sin patrón definitivo de herencia y que por lo general empieza en personas de edad avanzada.

Hereditario: familia en la que hay un grupo de tres o más familiares afectados en cualquier núcleo familiar (padres y hermanos), una familia con cáncer de próstata en tres generaciones, ya sea de parte del padre o de la madre, o un grupo de dos familiares afectados a edad temprana (55 años o menos). Del 5 al 10 por ciento de los casos de cáncer de próstata se consideran hereditarios.

Sin embargo, los urólogos suelen eludir este tipo de información. A ellos lo único que les interesa saber, generalmente, es si hay antecedentes familiares con cáncer de próstata pero no el número.

Factores genéticos. En el centro de cada célula del cuerpo humano se encuentra nuestro material genético: los cromosomas. Normalmente, las células contienen 46 cromosomas, es decir, 23 pares. Los cromosomas contienen los genes, que son los que determinan nuestras características, tales como el color de los ojos y el grupo sanguíneo, y también controlan las funciones reguladoras importantes del cuerpo, como el índice de crecimiento celular. Algunos genes, cuando se alteran o mutan, establecen un riesgo mayor para el crecimiento incontrolado de células, el cual a su vez puede llevar al desarrollo de un tumor. Estos genes tienen varios nombres, pero en conjunto se les llama "genes susceptibles al cáncer". Aproximadamente el 9 % de todos los cánceres de próstata y el 45 % de los casos en hombres menores de 55

años de edad pueden atribuirse al gen susceptible al cáncer, que se hereda como característica dominante (de padres a hijos).

6.2. Diagnóstico del cáncer de próstata

De acuerdo con los factores de riesgo que hemos comentado en la sección 6.1.3, además de los hallazgos clínicos, el médico deberá conocer la historia del paciente para poder determinar la probabilidad de que los síntomas presentados correspondan a un cáncer de próstata o a otra enfermedad relacionada.

Historia clínica

Por ello, deberá conocer ciertos datos del paciente como su edad, país de origen y residencia, actividad sexual, hábitos alimenticios y si tiene antecedentes familiares de cáncer de próstata. A continuación se le realizan las preguntas correspondientes al protocolo de puntuación internacional de síntomas de próstata (IPSS[2]), con el objetivo de determinar posibles patologías relacionadas con la prostáta como la HPB, la prostatitis crónica y la congestión prostática. Consta de 7 preguntas relacionadas con el flujo y cantidad de orina, la frecuencia, etc. Cada respuesta se valora entre 0 y 5, por lo que el resultado del test tendrá un valor entre 0 y 35. Por ejemplo, una de las preguntas es: "Durante el pasado mes, ¿cuántas veces ha tenido la sensación de no vaciar su vejiga completamente después de la micción?". Las posibles respuestas son: Ninguna (0), Menos de una vez cada 5 veces (1), Menos de la mitad de las veces (2), Alrededor de la mitad de las veces (3), Más de la mitad de las veces (4), Casi siempre(5). Este protocolo permite clasificar los síntomas de la próstata en leves (de 0 a 7), moderados (8-19) y graves (20-35).

Por último, otros datos interesantes para el urólogo son conocer si ha sido operado (sondaje), por la posibilidad de infecciones urinarias, si tiene disuria (dolor al orinar) o síndrome constitucional, que agrupa el siguiente conjunto de síntomas: astenia, anorexia, pérdida de peso y febrícula o fiebre no filiada.

6.2.1. Hallazgos médicos

Aunque la única prueba que sirve para efectuar un diagnóstico preciso del cáncer de próstata es la biopsia de la próstata, existen una serie de síntomas, signos y pruebas de

[2]Del inglés, International Prostatic Symptom Score.

laboratorio que, junto con la historia clínica del paciente, permiten sospechar o descartar la presencia del tumor. Son los que describimos a continuación.

Síntomas

Aunque, por lo general, el cáncer de próstata no tiene señales o síntomas específicos en las primeras etapas, los más comunes son:

- Iniciales: Los que primero se presentan, en un 75 % de los casos, son la obstrucción del flujo de la orina y infección. Estos síntomas son semejantes a los de la hipertrofia prostática benigna.
- Vesicales: Pérdida de fuerza y de calibre al orinar, goteo terminal, orinar con frecuencia (principalmente durante la noche), dificultad para orinar o contener la orina, incapacidad para orinar, dolor o sensación de ardor al orinar, sangre en la orina o en el semen y dificultad para conseguir la erección.
- Debidos a la metástasis: El 5 % de los enfermos presenta sus primeros síntomas debidos a metástasis, que ocasiona: dolor persistente en la espalda, la cadera o la pelvis, presencia de una masa en el cuadrante abdominal superior derecho (afección del hígado), presencia de una masa supraclavicular (metástasis en el ganglio linfático centinela), anemia, pérdida de peso, hematuria tardía (cuando se invaden la vejiga y la uretra), etc.

Signos

El principal signo para la detección del cáncer de próstata es el resultado de la prueba del **tacto rectal**, en la cual el médico palpa la próstata a través del recto con el fin de encontrar áreas duras o abultadas.

Datos de laboratorio

Otro de los principales hallazgos que ayudan a realizar el diagnóstico del cáncer de próstata cuando no hay síntomas es el resultado de una prueba de sangre que se utiliza para detectar una sustancia producida por la próstata, llamada **antígeno prostático específico (PSA)**[3]. Sin embargo, la sensibilidad y la especificidad de esta prueba no son del 100 %, por lo que la mayoría de los hombres con niveles moderadamente elevados de PSA no tienen cáncer de la próstata y, al contrario, algunos hombres con cáncer de

[3]Se le suele denominar PSA por las siglas de su nombre en inglés: Prostatic Specific Antigen

próstata tienen niveles normales de PSA. La primera situación se puede dar, entre otras causas, por haber realizado la prueba del tacto rectal previamente a los análisis del PSA, puesto que dicha prueba aumenta los niveles en sangre del PSA. El hecho de que haya algunos hombres con cáncer de la próstata con niveles normales de PSA puede deberse a estar siendo medicados con Finasteride, un fármaco utilizado para tratamientos capilares y para la HPB, lo que disminuye los niveles de PSA.

En etapas tardías, la anemia puede ser intensa, cuando la médula ósea es reemplazada por tumor, a lo cual contribuyen también las hemorragias y las infecciones. El examen de orina puede mostrar o no infección, pudiendo existir hematíes.

Si los resultados del tacto rectal o del PSA son anormales, se pueden repetir los exámenes o solicitar otras pruebas, que incluyen:

- Ecografía transrectal, que permite observar condiciones anormales, como el agrandamiento de las glándulas, los nódulos, la penetración del tumor a través de la cápsula de la glándula o la invasión de las vesículas seminales. También puede orientar durante las biopsias con aguja de la glándula prostática.

- Tomografía axial computerizada (TAC). Es un procedimiento de diagnóstico por imagen que utiliza una combinación de rayos X y tecnología computerizada para obtener imágenes de cortes transversales (que suelen llamarse "rebanadas") del cuerpo, tanto horizontales como verticales. Una tomografía computarizada muestra imágenes detalladas de cualquier parte del cuerpo, incluyendo los huesos, los músculos, la grasa y los órganos. La tomografía computarizada muestra más detalles que los rayos X regulares.

- Resonancia magnética, que utiliza una combinación de imanes grandes, radiofrecuencias y una computadora para producir imágenes detalladas de los órganos y estructuras dentro del cuerpo.

- Gammagrafía ósea, que es un método nuclear de creación de imágenes que ayuda a mostrar si el cáncer se ha diseminado desde la glándula prostática a los huesos. El procedimiento comprende la inyección de material radioactivo que ayuda a localizar las células óseas enfermas por todo el cuerpo y sugiere la posible metástasis del cáncer.

- Biopsia de la próstata o de los nódulos linfáticos, o ambos. Es una prueba en la que se extraen del cuerpo muestras de tejido (con una aguja o durante la cirugía) para

examinarlas con un microscopio con el fin de determinar si existen células cancerosas o anormales. **El diagnóstico de cáncer se confirma únicamente por biopsia**.

6.2.2. Grado y etapas del cáncer de próstata

Una vez diagnosticada la presencia del carcinoma de la próstata es importante conocer la agresividad del tumor y si se ha diseminado; de ser así, conviene conocer las áreas del cuerpo que están afectadas, porque las decisiones con respecto al tratamiento dependerán de estos resultados. Para ello, se suelen evaluar dos características del cáncer de próstata: el grado de Gleason y la etapa.

Grado de Gleason. Representa el grado de diferenciación tumoral y describe las semejanzas entre el tumor y el tejido normal. Basándose en la apariencia microscópica de un tumor, los patólogos (los médicos que identifican las enfermedades a través del estudio de tejidos bajo el microscopio) pueden describir el grado del cáncer como bajo, medio o alto. De acuerdo con el Instituto Nacional del Cáncer de EE.UU., una forma de clasificar el cáncer de próstata por grados es el *sistema de Gleason*, que asigna números del 2 al 10. Los tumores de grados altos crecen más rápidamente que los tumores de grados bajos y es probable que se diseminen con mayor rapidez a otras partes del cuerpo. Los grados menores de 4 indican que las células cancerosas se parecen a las células normales y es probable que el cáncer sea menos agresivo. Los grados del 8 al 10 indican que es más probable que las células cancerosas crezcan muy agresivamente.

Etapa. La clasificación por etapas del cáncer de la próstata pretende determinar el sitio y la localización de la enfermedad. Hay varios sistemas diferentes que describen las etapas del cáncer. Los diferentes sistemas pueden utilizar letras del alfabeto, números romanos u ordinales, o cualquier combinación de ellos, para designar las etapas apropiadas del cáncer.

- *Etapa I (A):* En esta etapa, el cáncer no se detecta en la exploración rectal y no suele causar síntomas. Se encuentra sólo en la próstata y se suele detectar por casualidad al realizar cirugía por otras razones, como por ejemplo, por una HPB. Las células cancerosas se pueden encontrar en una o varias áreas de la próstata.
- *Etapa II (B):* Las células cancerosas se encuentran sólo en la próstata, pero los niveles de PSA suelen ser elevados. El tumor puede detectarse por medio de

una biopsia o mediante un examen rectal.

- *Etapa III (C):* Hay extensión extracapsular del tumor. Las vesículas seminales pueden estar afectadas por el cáncer.
- *Etapa IV (D):* Se ha diseminado, por metástasis, a otros órganos o tejidos. Normalmente la metástasis afecta a los ganglios linfáticos o a los huesos, hígado y pulmones.

Sistema de clasificación TNM. Este sistema de clasificación, desarrollado por el American Joint Committee on Cancer, contempla el tamaño del tumor (T), la extensión a los ganglios linfáticos (N) y la extensión o diseminación a otras partes del cuerpo (M). Para conocer la etapa de la enfermedad se utilizan varios exámenes de sangre y de imágenes.

T. En cuanto al tamaño del tumor, podemos distinguir las siguientes fases:

TX. El tumor no se puede valorar.

T0. No hay evidencia del tumor.

T1. Corresponde a la etapa A. El tumor se localiza sólo en la próstata y no causa síntomas. Es todavía demasiado pequeño para palparlo en la exploración rectal o visualizarlo en un escáner. Normalmente, se diagnostica casualmente, cuando se realiza cirugía por otras razones, como por ejemplo, la HPB, aunque los valores del PSA pueden estar ya elevados. Esta etapa puede subdividirse en las etapas a,b y c:

T1a. (Etapa A1) El tumor se encuentra en un 5 % de tejido de la próstata.

T1b. (Etapa A2) El tumor se encuentra casualmente en más de 5 % de tejido de la próstata.

T1c. (Etapa B0) El tumor se ha detectado mediante una biopsia al encontrar valores elevados de PSA.

T2. Corresponde a la etapa B. En esta etapa, el tumor está localizado sólo dentro de la próstata; aunque los pacientes normalmente no experimentan síntomas, es lo suficientemente grande para ser palpado en la exploración rectal. Esta etapa puede subdividirse en las etapas a,b y c:

T2a. (Etapa A1) El tumor ocupa menos de medio lóbulo de la próstata. Normalmente se detecta en el examen rectal.

T2b. (Etapa B1) El tumor ocupa más de medio lóbulo de la próstata y suele ser detectable en el examen rectal.

T2c. (Etapa B2) El tumor ocupa ambos lóbulos y siempre se toca en el examen rectal.

T3. Corresponde a la etapa C. En esta etapa, el tumor se ha extendido a los tejidos circundantes, por lo que las vesículas seminales pueden estar afectadas. Esta etapa puede subdividirse en las etapas *a*, *b* y *c*:

T3a. (Etapa C1) El tumor se ha extendido fuera de la cápsula de la próstata sólo por un lado. Se suele denominar extensión extracapsular unilateral.

T3b. (Etapa C2) El tumor se ha extendido fuera de la cápsula de la próstata por ambos lados. Se suele denominar extensión extracapsular bilateral.

T3c. (Etapa C3) El tumor se ha extendido a una o ambas vesículas seminales.

T4. No corresponde con ninguna etapa A, B, C o D. En esta etapa, el tumor está todavía en la región pélvica pero se puede haber extendido a otras áreas. Esta etapa puede subdividirse en las etapas *a* y *b*:

T4a. El tumor está extendido a alguna de las siguienes áreas: el cuello de la vejiga, el esfínter externo y/o el recto.

T4b. El tumor puede haber afectado a los músculos que ayudan a la erección y/o a las paredes pélvicas.

N. En cuanto a la metástasis a los ganglios linfáticos, podemos distinguir las siguientes fases:

N0. No hay metástasis a los ganglios linfáticos.

N1. (Etapa D1) Las células se encuentran en uno o más ganglios linfáticos del área pélvica y su tamaño es nemor de 2 cm.

N2. (Etapa D1) El cáncer se describe como N2 si las células se encuentran en uno o más ganglios linfáticos del área pélvica y su tamaño es mayor de 2 cm y menor de 5 cm.

N3. El cáncer se describe como N3 si el tamaño de los ganglios es mayor de 5 cm.

M. Por último, en cuanto a la metástasis a otros órganos, se tiene:

M0. No existe metástasis.

M1. (Etapa D2) La metástasis puede afectar a la médula espinal, por lo que los síntomas más comunes son dolor de huesos, pérdida de peso y cansancio.

6.2.3. Tablas de Partin: Índice de supervivencia-extensión

Las tablas de Partin [137] representan un protocolo para determinar la extensión del cáncer de próstata, lo que servirá para definir el índice de supervivencia del enfermo. Se basan en un estudio estadístico realizado sobre 4.133 pacientes que no habían recibido radioterapia ni tratamiento hormonal alguno y mediante el que los autores hallaron una correlación estadística entre el grado del cáncer, su etapa y el valor del PSA. La idea subyacente es proporcionar la probabilidad de que el cáncer de próstata se encuentre:

1. confinado en la próstata;
2. que haya habido penetración capsular;
3. que las vesículas seminales estén invadidas;
4. que haya metástasis a los ganglios linfáticos.

Por tanto, para cada una de estas posibilidades existe una tabla con los índices de supervivencia correspondientes a cada caso, obtenidos a partir de los valores de los tres principales indicadores del tumor: la clasificación TNM, el grado de gleason y el valor del PSA. La figura 6.1 ilustra una parte de una de las tablas.

Gl.	T1a	T1b	T1c	T2a	T2b	T2c	T3a
PSA de 0 a 4							
2-4	90 (84-95)	80 (72-86)	89 (86-92)	81 (75-86)	72 (65-79)	77 (69-83)	
5	82 (73-90)	66 (57-73)	81 (76-84)	68 (63-72)	57 (52-62)	62 (55-69)	40 (26-53)
6	78 (66-88)	61 (52-69)	78 (74-81)	64 (59-68)	52 (46-57)	57 (51-64)	35 (22-48)
7		43 (34-53)	63 (55-68)	47 (41-52)	34 (29-39)	38 (32-45)	19 (11-29)
8-10		31 (20-43)	52 (41-62)	36 (27-45)	24 (17-32)	27 (18-36)	

Tabla 6.1: Probabilidades de padecer cáncer confinado en la próstata cuando los valores del PSA están comprendidos entre 0 y 4

6.3. Construcción de Prostanet con la facilidad de explicación de ELVIRA

Después de implementar el nuevo método de explicación en ELVIRA (descrito en el capítulo 5), decidimos escoger un dominio de aplicación para poner a prueba las facilidades

de explicación desarrolladas. En las secciones siguientes describimos cómo se ha llevado a cabo el proceso de construcción manual de la red PROSTANET, siguiendo el algoritmo descrito y cómo ELVIRA nos ha facilitado la tarea.

El experto que nos ha ayudado ha sido el Dr. Diego Alberto Rodríguez Leal, urólogo del Hospital de Alarcos, de Ciudad Real, con más de 20 años de experiencia clínica en distintos hospitales. Aunque la construcción de la red bayesiana suscitó en él un gran interés, hay que tener en cuenta que, como era de esperar, el urólogo no tenía ningún tipo de conocimiento sobre la teoría y significado de las redes bayesianas. Sin embargo, las capacidades de explicación que proporciona ELVIRA han constituído una potente herramienta que nos ha facilitado enormemente el proceso de construcción de la red, especialmente en lo relacionado con la transmisión de información al experto.

6.3.1. Construcción del grafo cualitativo

El proceso de identificación de variables en PROSTANET se llevó a cabo a partir del estudio de la bibliografía específica [3, 7, 117, 190].

En medicina las variables normalmente corresponden a posibles *causas* y *factores de riesgo* de una *enfermedad*, junto con los *síntomas*, *signos* y *pruebas de laboratorio* que permiten confirmar o descartar la enfermedad. Afortunadamente para los ingenieros de conocimiento, entre los que nos incluímos, todos estos datos suelen aparecer fácilmente identificados en la literatura médica.

En nuestro caso, como el objetivo era el diagnóstico del cáncer de próstata, partimos de la creación de este nodo y comenzamos a identificar los factores de riesgo, signos, síntomas y pruebas relacionados directamente con la enfermedad, obteniendo la primera versión de la red, tal y como aparece en la figura 6.2.

Observamos que es una primera aproximación muy sencilla donde prácticamente se distinguen tres niveles: el primero, correspondiente a los nodos que están por encima del nodo Cáncer de próstata, que representa los factores de riesgo; el siguiente sería el que representa la enfermedad en cuestión y, por último, el conjunto de nodos asociados a los signos, síntomas y pruebas.

Para la construcción de esta primera versión la principal dificultad que tuvimos fue la representación del sentido de los arcos. El problema es que, frecuentemente, los médicos suelen definirlos en el sentido opuesto, que es el sentido del razonamiento médico: de los efectos a las causas. Por ejemplo, en la primera entrevista, el urólogo definió las relaciones entre los nodos Cáncer de Prostata, Gleason, PSA y Exploración Rectal como se ilustra en la figura 6.3. En este caso, la razón por la que razonaba así era que estaba acostumbrado a

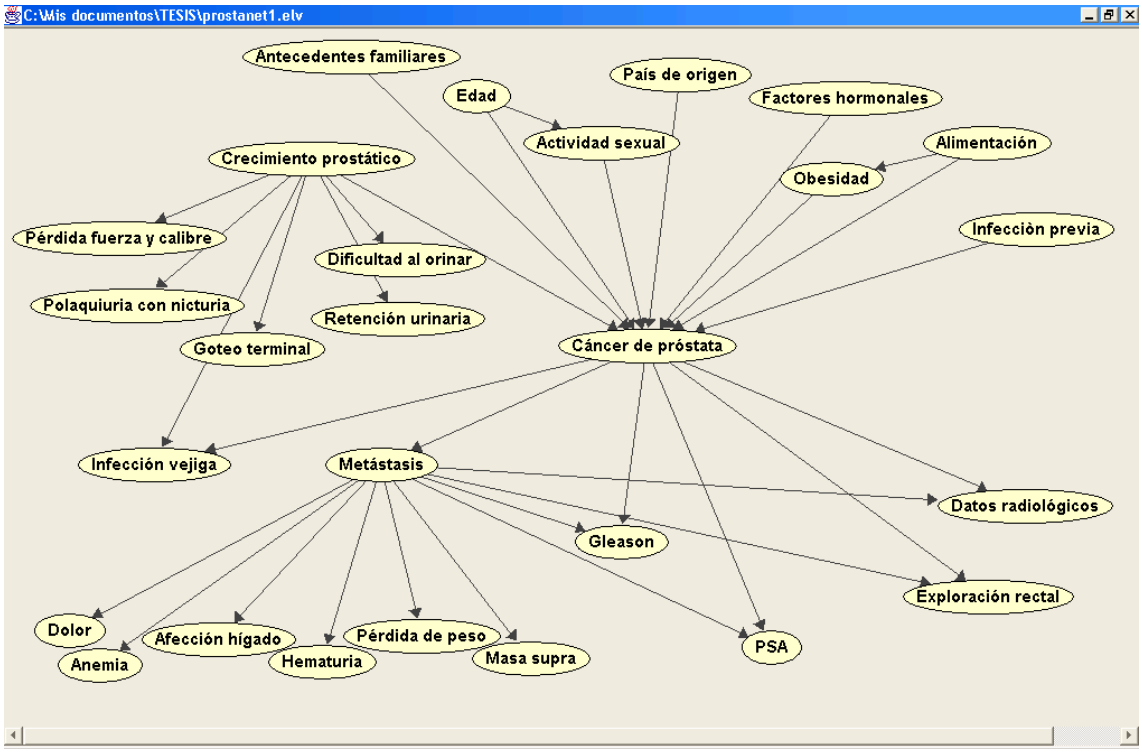

Figura 6.2: Primera versión de la red Prostanet

utilizar las Tablas de Partin que, precisamente lo que permiten es calcular la probabilidad de tener localizado el cáncer en un determinado lugar dados los valores del Gleason, del PSA y de la exploración rectal.

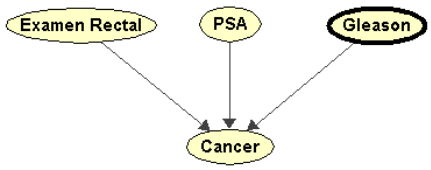

Figura 6.3: Dibujo de los arcos entre los nodos Cancer, Gleason, PSA y Exploración rectal por el urólogo

Hay que indicar, sin embargo, que el experto enseguida se familiarizó con la *interpretación causal* de la representación gráfica de la red que se estaba construyendo. Para ello, la explicación verbal generada por ELVIRA fue de gran ayuda, puesto que el experto tuvo la oportunidad de ver reflejado el conocimiento representado en Prostanet de modo similar a como suele aparecer en la literatura sobre el tema. Como ejemplo, ilustramos un fragmento del texto:

> *La enfermedad Cáncer de próstata tiene los siguientes FACTORES DE RIESGO: Edad, Actividad sexual, País de origen, Factores hormonales, Infección previa, Antecedentes familiares, Obesidad. Se manifiesta a través de los siguientes SÍNTOMAS: Dolor y SIGNOS: Metástasis, Anemia, Pérdida de peso, Afección hígado, Hematuria, Masa supra, Infección vejiga. Existen las siguientes PRUEBAS que ayudan a confirmarla o descartarla: Exploración rectal, Gleason, PSA, Datos radiológicos.*

A la luz de dicha explicación, al experto le resultó más sencillo detectar la ausencia de otras enfermedades que pueden cursar con síntomas y signos similares, o que son a su vez causas o factores de riesgo del cáncer de próstata. Por tanto, después de analizar la explicación verbal procedimos a identificar las variables correspondientes a otras enfermedades relacionadas con la próstata y cuya manifestación podría dar lugar a confusión con el cáncer, a saber: Hipertrofia Prostática Benigna (HPB), Prostatitis crónica, Congestión Prostática, Displasia, Cistitis e Infección del tracto urinario. La siguiente versión de la red (véase la figura 6.4) ya contenía 34 nodos, que representaban 8 enfermedades, 7 factores de riesgo, 7 pruebas de laboratorio, 2 signos de exploraciones por el médico y 8 a otros síntomas y signos.

Sin embargo, al volver a analizar la explicación verbal generada para esta segunda versión, el experto consideró que el modelo representado seguía siendo un poco simple porque había relaciones que no se habían tenido en cuenta. Por ejemplo, cuando los valores del PSA están entre 4 y 10, los especialistas suelen solicitar un nuevo análisis para la obtención de la razón entre el PSA libre [4] y el PSA total. Si ése es menor que 0.1 aumenta la probabilidad de padecer cáncer. Por ello, cambiamos el nombre de PSA por PSA total, para reflejar la cantidad total de PSA que circula en sangre, ligado o no; además, incluímos dos nuevos nodos binarios: el nodo binario PSA libre y el nodo binario PSAl/PSAt.

Otro caso curioso fue el relacionado con las infecciones urinarias. Según el experto, debíamos incluir un nuevo enlace desde Prostatitis Crónica a Infección urinaria pues ésta se puede producir también como consecuencia de una prostatitis. Sin embargo, dicho enlace provocaría un bucle en la red (ELVIRA avisa cuando se producen este tipo de situaciones). El problema surgía porque hasta entonces el experto no había considerado los distintos momentos en que se pueden producir las infecciones urinarias. Por tanto, se modificó el nombre del nodo ITUs por ITUS_uno, reflejando las infecciones previas a la prostatitis, y se añadió un nuevo nodo, ITUS_dos, representando las infecciones que se producen como

[4] El PSA libre representa únicamente la proporción de PSA que no va ligado a las proteínas en sangre.

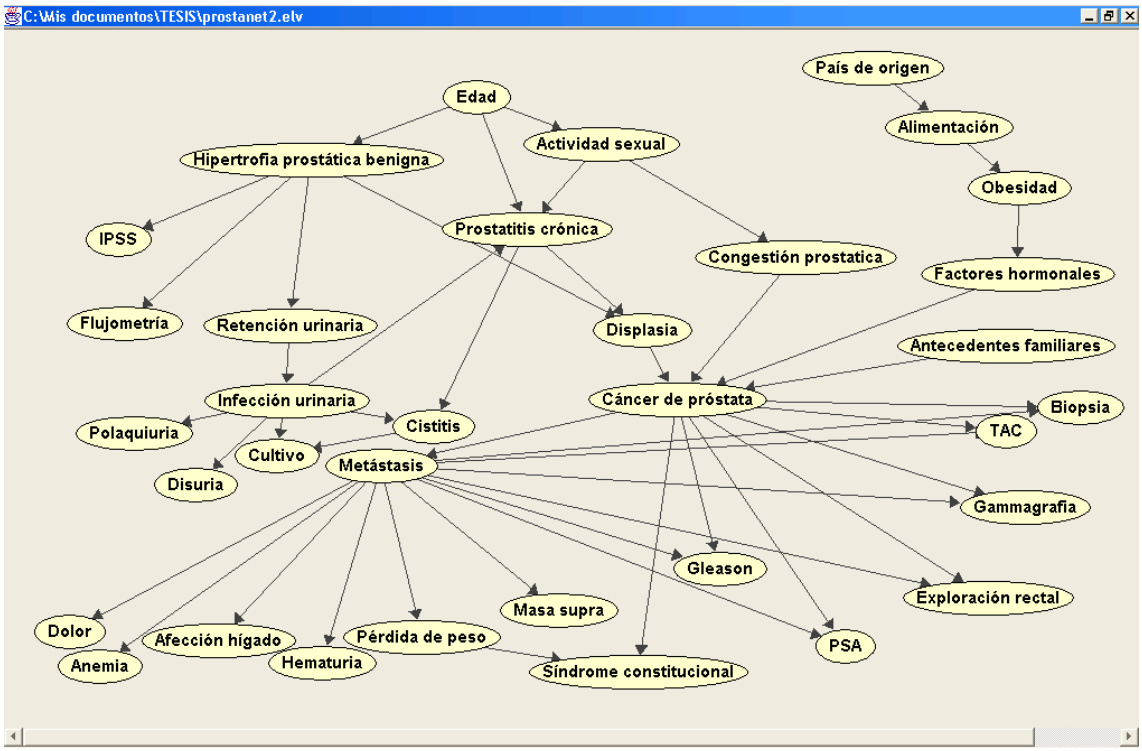

Figura 6.4: Segunda versión de la red Prostanet en la que se incluyen otras posibles patologías de la próstata que pueden cursar con signos y síntomas parecidos a los del cáncer.

consecuencia de una prostatitis.

Por tanto, después de un proceso de análisis en las posibles relaciones existentes entre las enfermedades representadas, durante el que se incluyeron algunas variables y enlaces nuevos, obtuvimos la tercera versión de la red, que se muestra en la figura 6.5

Después de estudiar la explicación verbal generada para esta versión, el experto consideró concluido el proceso de representación de la información cualitativa, aunque, como iremos viendo a lo largo de las siguientes secciones, más tarde esta versión sufrió varias modificaciones.

6.3.2. Discretización de variables

Aunque las variables de una red bayesiana pueden ser discretas, continuas o mixtas, nosotros hemos utilizado únicamente variables discretas. Las razones para ello son diversas, pero fundamentalmente se debe a la dificultad para representar distribuciones continuas (salvo casos particulares, como gaussianas) y a que todavía no hay algoritmos eficientes de propagación con dichas variables. Por tanto, a lo largo de este capítulo las

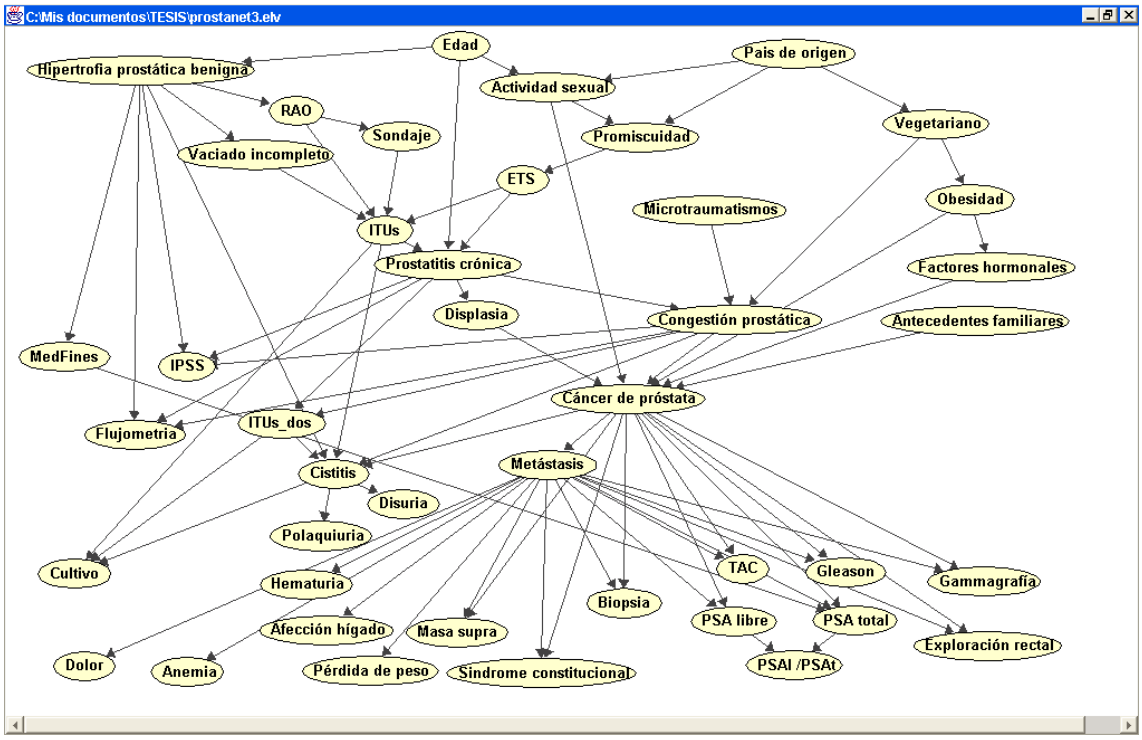

Figura 6.5: Tercera versión de la red PROSTANET.

referencias serán únicamente a variables discretas.

Durante el proceso de construcción del grafo cualitativo otro de los pasos importantes es el de la definición de los estados de cada variable que, normalmente, se realiza a la vez que se van creando las variables. En este proceso hay que intentar buscar el equilibrio respecto al número de estados que se elija para cada una de ellas, pues cuanto mayor es el número de estados, más se incrementa el número de probabilidades a obtener. En el caso de PROSTANET todas las variables, excepto el País de origen, son **ordinales**. Todas tienen un nombre significativo, salvo las tres siguientes que se han identificado mediante abreviaturas:

- ITUs representa las infecciones del tracto urinario;
- ITUs_dos para las infecciones urinarias debidas a gérmentes acantonados en la próstata y que se producen después de haber tenido una infección urinaria previa;
- RAO para la retención aguda de orina;
- ETS para las enfermedades de transmisión sexual.

Además, casi todas son **binarias** con valores del tipo ausente-presente, no-si, negativo-positivo, puesto que lo que interesa conocer de cada una de ellas es si tienen un valor

distinto del habitual.

Sin embargo, las siguientes variables se definieron como **no binarias** por las siguientes razones:

- Edad. Representa la edad del hombre y sus valores están ordenados de menor a mayor con la intención de que los enlaces causales que partan de esta variable representen influencias positivas, ya que el riesgo de desarrollar un cáncer de próstata aumenta con la edad. La edad a partir de la que comienza a aumentar el riesgo del cáncer es a los 50 años. Por tanto, el primer valor de la variable se definió como menor50, que representa el tener menos de 50 años. El problema era intentar definir el resto de valores, puesto que el número de ellos podía ser demasiado grande. Por tanto, el siguiente paso fue determinar cuál iba a ser el estado máximo a tener en cuenta. Como muchos hombres hoy viven más de 80 años y, a partir de esa edad, si hubiera un cáncer de próstata el crecimiento es más lento y, por otro lado, puede ser que no supusiera la causa del fallecimiento del paciente, consideramos que no merecía la pena hacer distinciones más allá de esta edad. Por eso, el último valor considerado es mayor80. Ahora bien, entre los 50 y los 80 años existían 30 valores posibles. Sin embargo, resulta del todo ineficiente definirlos todos, por lo que decidimos representarlos por franjas de edad de 10 años, que es como aparecen reflejadas en la literatura sobre el tema. Así pues, definimos tres valores más, 50-59, 60-69 y 70-79.

- Actividad Sexual. Este fue uno de los nodos más problemáticos pues no existe ningún tipo de dato fiable sobre él. Simplemente se sabe que cuanta mayor actividad sexual con eyaculación se tenga a lo largo de la vida, menor es la probabilidad de padecer una congestión prostática y, en consecuencia, menor es la probabilidad de desarrollar el cáncer. Por tanto, la definimos también como una variable ordinal con los valores siguientes: poca_o_nada, normal y mucha. Observamos que son valores "difusos" y que habría sido mejor intentar cuantificarlos. Sin embargo, no es fácil hacerlo porque no hay estudios que relacionen la frecuencia de relaciones sexuales con el riesgo de cáncer de próstata. Por eso, decidimos que no podía ser binaria porque un valor representaría la ausencia de relaciones y el otro la presencia y no podría considerarse la frecuencia, que se puede reflejar con un valor más. Por otra parte hay que indicar que no hemos considerado aisladamente el valor "nada" porque pensamos, especialmente el urólogo, que la diferencia entre no tener relaciones y tener muy pocas no influye significativamente en la probabilidad de padecer un

cáncer, al contrario de lo que parece que ocurre entre los valores normal y mucha.

- PSA. Este nodo representa los niveles en sangre del antígeno prostático específico, y tiene los siguientes valores: de 0 a 4, suele indicar la ausencia del cáncer; de 4 a 10, es un valor dudoso, por lo que se recomienda hacer otro tipo de pruebas; de 10 a 20, hay bastantes indicios de tener un cáncer de próstata; y mayor que 20, indica casi con certeza que el paciente tiene un cáncer de próstata.

- IPSS. Este nodo toma tres valores para representar el tipo de patología: leve, moderado, grave, que corresponden respectivamente a tres intervalos, que dependen de la puntuación obtenida: de0a7, de8a19 y mayor20.

- Gleason. Se mide con seis intervalos: cero, de 2 a 4, cinco, seis, siete y de 8 a 10. El valor 0 representa la ausencia de cáncer de próstata; el valor 2-4 indica que las células cancerosas se parecen a las células normales y, por tanto, el cáncer es menos agresivo; el valor entre 8 y 10 indica que el cáncer crece muy rápidamente. En un primer momento se pensó en asignar 11 valores a dicho nodo pero después de analizar el problema con el experto decidimos simplificar el número de valores y asignar sólo 3, cero, menorque6, mayorque6 puesto que los especialistas no tienen en cuenta todo ese rango de valores a la hora de emitir un diagnóstico.

- Exploración rectal, que sirve para representar el resultado de dicha prueba, mediante el que se mide la densidad de la próstata y si está entera o no. En las primeras versiones, esta variable se consideró binaria, con valores negativo-positivo. Sin embargo, durante el proceso de evaluación se consideró mucho más apropiado añadir un nuevo estado que representara estados intermedios, para reflejar mejor la realidad. Por tanto, finalmente los valores de dicha variable son normal, fibrosa, pétrea.

- En cuanto a la variable Cancer de Próstata, inicialmente se representó mediante una variable no binaria con 5 estados: ausente, confinado a la próstata, penetración capsular, invasión de vesículas seminales, e invasión de ganglios linfáticos. El objetivo era poder determinar no sólo la presencia del cáncer sino, además, su localización y, con ello, el índice de supervivencia del enfermo. Sin embargo, el experto nos indicó que los últimos cuatro estados no son exclusivos. Por otra parte, la definición de los cinco estados complicaba enormemente el proceso de obtención de las probabilidades. No obstante, dichos estados se podrían clasificar en dos grupos que son los correspondientes a la localización del cáncer en la cápsula prostática y la existencia de metástasis. Por tanto, decidimos definir el nodo Cáncer de próstata como nodo

binario, el nodo Metástasis que representa la simplificación de las cuatro posibles fases de localización del cáncer y un enlace de Cáncer de próstata a Metástasis.

No incluimos figura de la cuarta versión de la red Prostanet porque sólo se diferencia de la tercera en los valores de las variables.

6.3.3. Aplicación de modelos canónicos

El siguiente paso consistió en la identificación de los modelos canónicos (véase la sección 1.3.2) sobre la red de la figura 6.5. Todos los modelos identificados correspondieron a puertas OR y MAX. Hay que indicar que en todas las ocasiones fue el experto el que determinó la no existencia de sinergia entre las posibles causas de cada nodo, aunque desconocía si existen estudios científicos en los que basar su afirmación. Por otra parte, casi todos las relaciones correspondieron a modelos residuales, ya que se desconocen con exactitud cuáles son todas las causas de los nodos involucrados.

En general, la mayoría de los modelos canónicos definidos en Prostanet fueron detectados de forma inmediata y sin demasiados problemas; en concreto, los correspondientes a las siguientes relaciones:

- Flujometría - HPB, Prostatitis crónica, Congestión Prostática. En este caso se definió una puerta MAX residual puesto que la variable que representa la flujometría no es binaria; además, no parece existir sinergia significativa entre las distintas causas que pueden influir en los valores obtenidos en dicha prueba.

- IPSS - HPB, Prostatitis crónica, Congestión Prostática En este caso se definió también una puerta MAX residual por las mismas razones.

Por otro lado, nos encontramos diversos problemas en las siguientes situaciones:

- El nodo Cáncer de Próstata tenía 6 padres, cinco binarios y uno con tres estados, y no se ajustaba a ninguno de los modelos canónicos que permiten reducir el número de parámetros. Esto significaba tener que asignar 96 probabilidades. Pero el problema no sólo era el número de datos numéricos a estimar, sino la cantidad de situaciones distintas en las que había que pensar. Un ejemplo de pregunta que tendría que resolver el urólogo sería:

  ¿Cuál es la probabilidad de que haya cáncer de próstata teniendo en cuenta que la actividad sexual del paciente es mucha, que no tiene antecedentes familiares con cáncer de próstata, que es obeso, que tiene congestión prostática, que no hay displasia pero sí tiene alteraciones hormonales?

Evidentemente, responder a 96 preguntas de este tipo dando valores fiables es imposible. No obstante, después de analizar con el experto la situación observamos que los padres del nodo se podían separar en dos grupos: uno con las posibles enfermedades que podían degenerar en cáncer y el otro con los factores de riesgo que, en principio, no tienen nada que ver con ellas y que se consideran que actúan independientemente del resto. Para ello, definimos dos nuevos nodos auxiliares, Cancer_f_r y Cancer_o_c con el objetivo de aplicar la técnica conocida como "divorcio de los padres" [90, 133]. Así, conseguimos separar los factores de riesgo y las anomalías relacionadas con la aparición del cáncer de próstata en dos grupos: por un lado, el grupo relativo a los factores de riesgo y el otro relacionado con las causas físicas, lo que supuso una gran mejora de cara a la obtención de las probabilidades. Con esta nueva situación, el modelo asociado al nodo Cáncer de Próstata corresponde a una puerta OR determinista, cuyas probabilidades las refleja la siguiente tabla:

Cancer_f_r	presente	presente	ausente	ausente
Cancer_o_c	presente	ausente	presente	ausente
CP=presente	1	1	1	0
CP=ausente	0	0	0	1

Por otro lado, para la relación del nuevo nodo Cancer_f_r se podía identificar ahora una puerta MAX residual, puesto que podríamos considerar como hipótesis simplificadora que cada uno de los padres de dicho nodo actúan independientemente del resto a la hora de favorecer la aparición del cáncer. El hecho de que sea residual radica en que no se conocen con exactitud todas las causas que pueden producir el cáncer de próstata.

En consecuencia, el número de parámetros a estimar se redujo considerablemente puesto que para el nodo Cancer_o_c, al tener solo 2 padres binarios, se debían obtener únicamente 2 valores; para el nodo Cancer_f_r el número de valores que se debían obtener se redujo a 6. En total, 8 probabilidades asignadas frente a las 96 iniciales. Pero entre las ventajas conseguidas no sólo tenemos que hablar de la reducción del número de valores, que no es poco, sino de la fiabilidad de los mismos, puesto que ahora el experto no tenía que pensar en todos los posibles factores actuando conjuntamente, sino que podía centrarse en cada uno de ellos independientemente del resto.

- Al intentar identificar algún modelo para el nodo Cultivo cuyos padres eran ITUS, ITUS_dos y Cistitis, observamos que este nodo representa una prueba de laboratorio

que sirve para detectar una cistitis o infección de orina, independientemente de si ésta es primaria o secundaria. Por tanto, decidimos incluir un nuevo nodo auxiliar, ITUS, que sirviera para representar este hecho, de tal forma que los padres de Cultivo fueran Cistitis e Itus, siendo los padres de ITUS los nodos ITUS_uno e ITUS_dos, tal y como refleja la figura 6.6. Además, la relación entre estos últimos tres nodos correspondía al de la puerta OR determinista, de forma similar al caso del nodo Cáncer de Próstata.

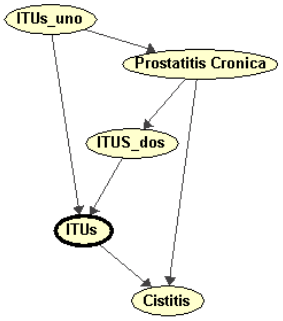

Figura 6.6: Definición del nodo ITUS.

- Como consecuencia de lo anterior, conseguimos reducir el número de padres del nodo Cistitis, que inicialmente tenía cinco (ITUS_uno, ITUS_dos, HPB, Cáncer de próstata, Congestión Prostática). Además, cada uno de sus padres actúa independientemente de los demás para producir una cistitis, por lo que se identificó otra puerta OR residual para definir la relación de este nodo. Con ello sólo se tuvieron que asignar 5 probabilidades.

- Otras relaciones para las que pudimos aplicar también una puerta OR residual fueron las definidas por ITUS_uno- Vaciado incompleto, ETS, RAO, Sondaje y por Prostatitis crónica - ITUS_uno, ETS.

Así obtuvimos la siguiente versión de la red, que podemos ver en la figura 6.7.

6.3.4. Obtención de las probabilidades

Como ya hemos comentado en la sección 1.3, ninguna de las metodologías propuestas para obtención de probabilidades se ha implementado en ninguna de las herramientas existentes para la construcción de redes bayesianas, salvo en prototipos que no están aún disponibles públicamente, por lo que la utilización de cualquiera de estos métodos supone

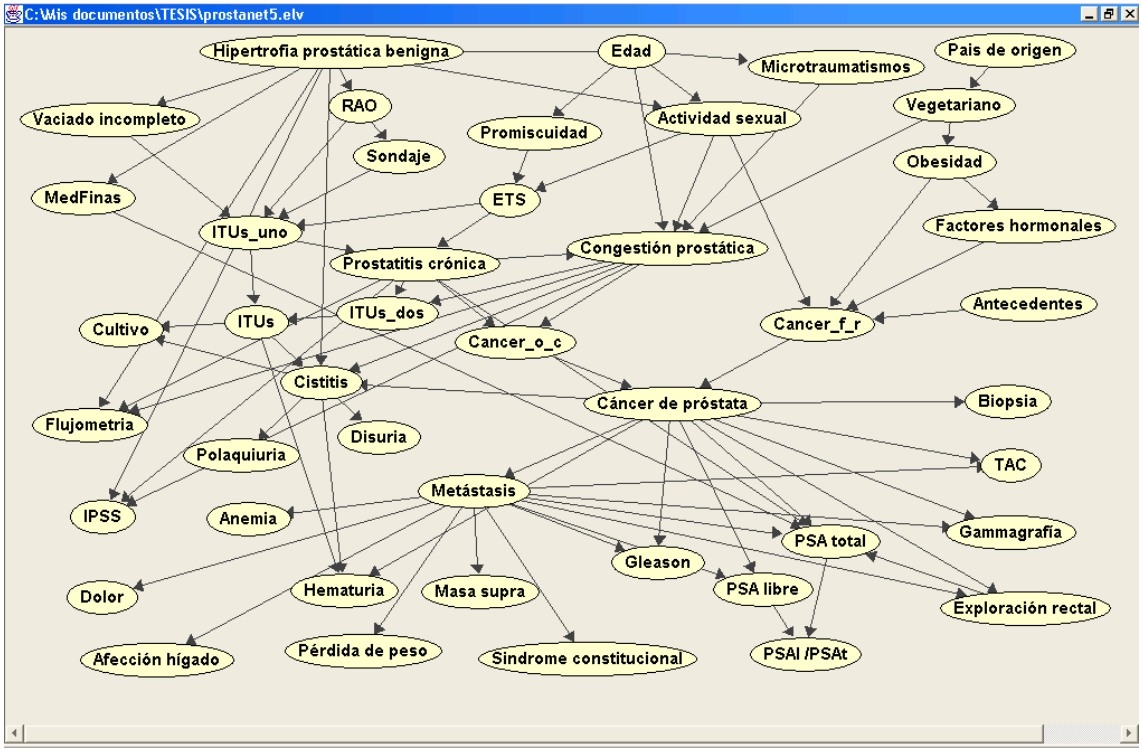

Figura 6.7: Quinta versión de la red PROSTANET después de aplicar los modelos canónicos.

hacer un gran esfuerzo previo, conocido en la literatura como "calibración del experto". En nuestro caso, sin embargo, el proceso se ha podido realizar de forma dinámica e interactiva, gracias a las posibilidades que ofrece ELVIRA, lo que ha permitido ir realizando a la vez un proceso de depuración de la red. A continuación describimos cómo las posibilidades de explicación implementadas en ELVIRA nos han ayudado a conseguir la versión definitiva de PROSTANET.

Representación de influencias, que ha sido una de las facilidades que más ayuda nos ha proporcionado para las siguientes tareas:

- *Edición de probabilidades.* Cuando las redes se editan con la interfaz gráfica de ELVIRA, durante el proceso de construcción del grafo cualitativo todos los enlaces están coloreados en negro —como podemos observar en las figuras 6.2, 6.4 y 6.5—, representando el hecho de que aún no se han asignado las probabilidades. Así, cuando se han asignado las probabilidades asociadas a un nodo, el color de los arcos incidentes en él cambian al color de la influencia transmitida por cada uno de sus padres. En consecuencia, los enlaces en negro nos sirven para conocer los nodos cuyas probabilidades todavía no se han asignado, lo que resulta muy útil en los casos en los que la red tiene un gran número de nodos.

- *Eliminación de enlaces* Por otra parte, un arco puede tener el color negro incluso después de asignar las probabilidades, cuando los valores que toma el padre no modifican la probabilidad del hijo. En este caso, el enlace puede (y debe) eliminarse porque no es necesario. Por ejemplo, en la tercera versión de Prostanet (figura 6.5) los padres de Gammagrafía eran dos: Cáncer de Próstata y Metástasis. Al asignar sus probabilidades, el enlace entre Cáncer de Próstata y Gammagrafía aparecía en negro. Al analizarlo con el experto, éste cayó en la cuenta de que la gammagrafía ósea es una prueba de diagnóstico exclusivo de la metástasis y no del cáncer de próstata, a pesar de que anteriormente había incluido ambos enlaces porque le costaba trabajo desligar mentalmente la metástasis del cáncer.

- *Análisis de influencias no positivas.* En redes bayesianas causales la mayoría de los enlaces representan influencias positivas, aunque en ocasiones es necesario reflejar influencias negativas o, incluso, indefinidas. Sin embargo, durante las primeras etapas de obtención de probabilidades a algunos nodos, el coloreado de enlaces en ELVIRA nos ayudó a constatar que las probabilidades eran incorrectas, pues las influencias no tenían el signo esperado. Por ejemplo, en Prostanet las siguientes influencias son negativas: de Edad a Actividad Sexual, puesto que normalmente de los 50 a los 80 años disminuye la actividad sexual; de Vegetariano a Congestión Prostática, puesto que la alimentación a base de soja disminuye el riesgo de padecer una congestión prostática; de Actividad Sexual a Congestión Prostática y a Cáncer_f_r, porque a mayor actividad sexual menor es el riesgo de padecer una congestión prostática y también un cáncer de próstata; de MedFinas a PSA, porque, como hemos dicho antes, la medicación con Finasteride reduce los valores del PSA. Sin embargo, podemos observar en la figura 6.7 que algunas influencias (por ejemplo, Vegetariano → Congestión Prostática o Actividad Sexual → Congestión Prostática) no se reflejaban correctamente.

  Por otro lado, la representación gráfica de las influencias nos permitió comprobar que, por ejemplo, aunque la relación Edad → Congestión Prostática está representada como indefinida (podría parecer errónea), es correcta puesto que, según el hombre va envejeciendo, la próstata se va agrandando, por lo que las probabilidades de padecer una congestión prostática aumentan; sin embargo, a partir de un momento determinado, cuando los hombres tienen ya 70 años o más, la próstata tiende a atrofiarse. En consecuencia, su tamaño se reduce

considerablemente y, por tanto, no puede congestionarse.

- *Modificación de la estructura de la red.* Durante el proceso de obtención de probabilidades observamos que en muchas ocasiones el conocimiento que tienen los expertos sobre la relación entre dos variables es simplemente la influencia que ejerce una sobre otra. Por ejemplo, se sabe que la medicación con Finasteride reduce en un 30 % los valores del PSA. Sin embargo, al intentar asignar las probabilidades, especialmente en los casos en los que el nodo destino del enlace tiene más de un padre (como en el caso del PSA), es muy complicado reflejar estas influencias, pues es fácil perderse entre la cantidad de valores y situaciones reflejados en una probabilidad. Por eso, si no tenemos una herramienta que refleje este tipo de influencias es muy difícil comprobar si los parámetros están más o menos bien definidos o no. Por otro lado, cuando los médicos tienen los resultados del cociente entre el PSA libre y el total, no les interesa conocer el valor absoluto del PSA libre. Por tanto, el valor del PSA libre es completamente irrelevante. Así pues, decidimos eliminar ese nodo con el objetivo de simplificar la red, ya que resultaba muy complicado discretizar dicha variable y, por tanto, asignar las probabilidades al nodo PSAl PSAt. Volviendo al ejemplo del PSA total, este nodo tenía inicialmente cinco padres binarios (ver figura 6.7): Exploración Rectal, MedFinas, Cáncer de Próstata, Prostatitis crónica y Metástasis. Como el nodo PSA tiene cuatro valores y no se ajusta a ninguno de los modelos canónicos, el experto tenía que asignar ¡96 probabilidades! Obviamente, a pesar de intentarlo en varias ocasiones, fue imposible asignar todos los valores de forma que se reflejaran correctamente las distintas influencias ejercidas entre los nodos. Es más, todas las influencias salían representadas como "desconocidas", en color violeta, cuando deberían haber sido positivas, salvo la que se ejerce desde MedFinas. Por tanto, la solución consistió en añadir un nodo auxiliar, PSA aux, con el objetivo de "divorciar" a sus padres, como se ilustra en la figura 6.8, y así poder simplificar el proceso de asignación de probabilidades. De este modo, para el nodo PSA aux solo habría que obtener 18 valores. Además, el nodo PSA se ajustaba a una puerta MAX graduada, por lo que el número de valores a estimar sería simplemente 9. En total, fue suficiente obtener 27 valores frente a los 96 iniciales.

- Siguiendo con el caso anterior del PSA, al asignar las probabilidades al nodo

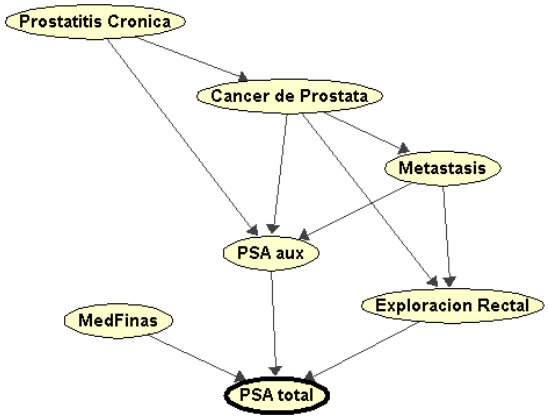

Figura 6.8: Representación del nodo PSA total después de añadir el nodo auxiliar PSA aux

PSA total nos dimos cuenta de que el resultado de la prueba de la Exploración Rectal no era lo que influía en el valor del PSA sino la realización de la prueba en sí. Por eso, después de eliminar el enlace entre Exploración Rectal y PSA total, pensamos en añadir un nodo nuevo que reflejara el hecho de que la prueba se había realizado. Sin embargo, el experto consideró que hoy en día todos los especialistas (incluso los médicos de atención primaria) conocen que la exploración rectal puede modificar los valores del PSA, por lo que no realizan la prueba hasta que el paciente no se ha hecho los análisis para la obtención del PSA. Por tanto, no se añadió ningún nodo nuevo.

6.3.5. Depuración de la red

A partir de aquí, las siguientes versiones corresponden a la fase de depuración de la red, que consistió en realizar las siguientes tareas:

Análisis del impacto de los hallazgos. La posibilidad de ver simultáneamente los efectos de varios casos de evidencia nos permitió analizar la influencia que determinados hallazgos tenían sobre una variable de interés. Por ejemplo, en las primeras etapas del proceso de obtención de las probabilidades detectamos que, según el modelo representado en la figura 6.11, y de acuerdo a las probabilidades asignadas por el urólogo, la edad no tenía mucho impacto sobre la variable Cancer de Prostata. En la figura 6.10 podemos apreciar que, prácticamente, la variación en las probabilidades es nula. Sin embargo, se sabe que la edad es uno de los factores de riesgo que más

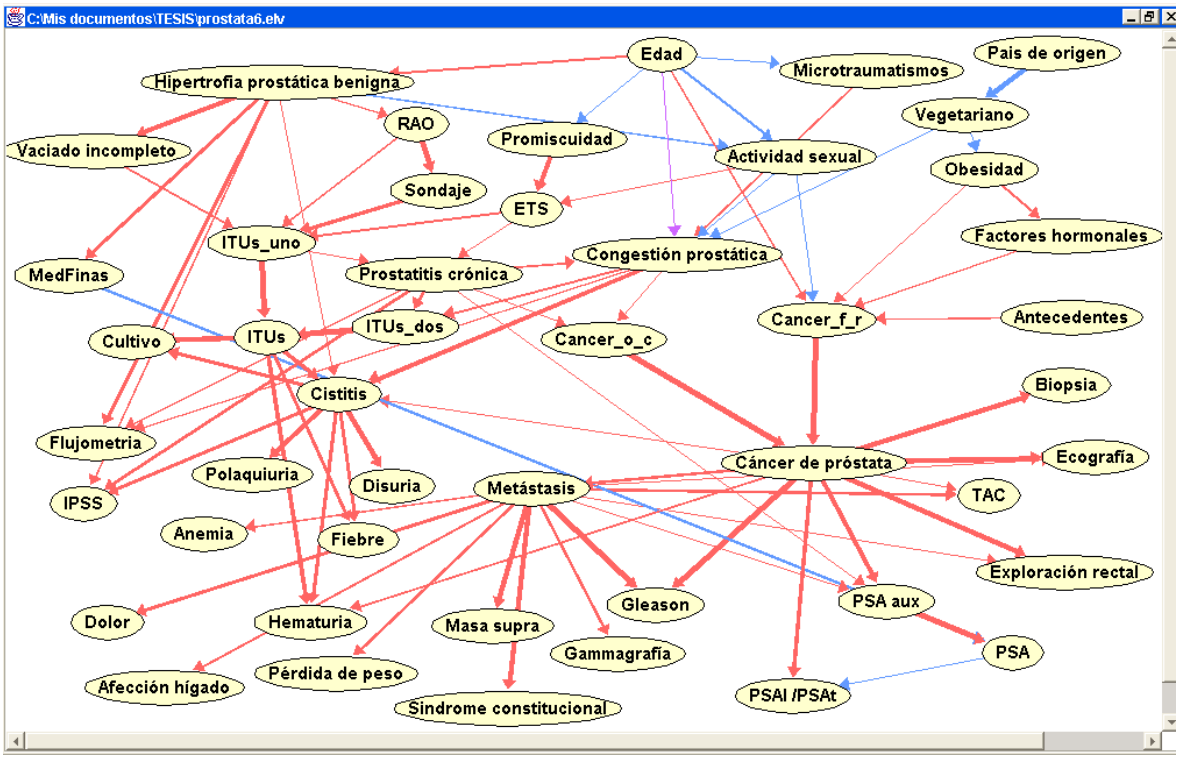

Figura 6.9: Sexta versión de la red PROSTANET

hacen aumentar la probabilidad de desarrollar un cáncer de próstata. Por tanto, eso nos hizo analizar las tablas de probabilidad del resto de los nodos y, al ver que eran aceptables, pudimos deducir que el modelo no era correcto. Por tanto, incluimos un nuevo enlace de Edad a Cancer_f_r. Un caso parecido nos ocurrió con las probabilidades del nodo PSA: al evaluar los casos de determinados pacientes vimos que al introducir los hallazgos correspondientes al PSA, la probabilidad de padecer un cáncer no aumentaba suficientemente, cuando se sabe que es uno de los factores más relevantes. Por ello, tuvimos que estudiar detenidamente las tablas de probabilidad para modificar sus valores.

Refinar probabilidades. Algo distinto ocurrió al analizar las probabilidades de tener cáncer después de introducir algunos otros hallazgos sobre los nodos Vegetariano y Antecedentes. En ese caso no es que el grafo fuera incorrecto, sino que lo que no estaba bien asignado era el conjunto de las probabilidades.

Explicación de los estados. La explicación verbal de los estados a través de las explicaciones de los enlaces y de los nodos de la red nos ha servido para refinar algunos valores de probabilidad. En el caso de la explicación de los nodos, porque los médicos suelen hablar en términos de que un valor es "dos veces más probable que otro",

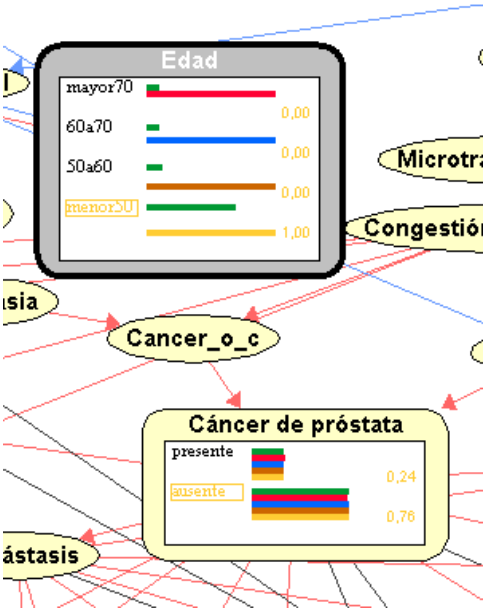

Figura 6.10: Variación de la probabilidad del nodo Cáncer de próstata dependiendo de las variaciones en el nodo Edad en la versión 6

por ejemplo, lo que resulta muy útil en el caso de variables com más de dos estados. Concretamente, a nosotros nos ha sido de especial utilidad en el caso del nodo PSA, al que tan complicado ha sido asignar correctamente sus probabilidades.

En cuanto a las razones de verosimilitud que se ofrecen en la explicación de los enlaces, permiten explicar qué estado de un nodo explica mejor un estado determinado de uno de sus descendientes. Esto resulta muy útil pues la forma de razonar de información que los médicos a veces se hace mediante razones de verosimilitud, lo que permite en muchas ocasiones poder refinar los datos asignados de forma subjetiva por el experto.

Expansión selectiva de nodos. Por último hay que destacar que en todo el proceso de obtención de probabilidades, nos ha sido de una gran ayuda la posibilidad de expandir únicamente los nodos seleccionados por el experto. Para ello, dependiendo de los casos interesaba utilizar únicamente el umbral de expansión o el rol desempeñado por el nodo o, en otras ocasiones, una combinación de ambos.

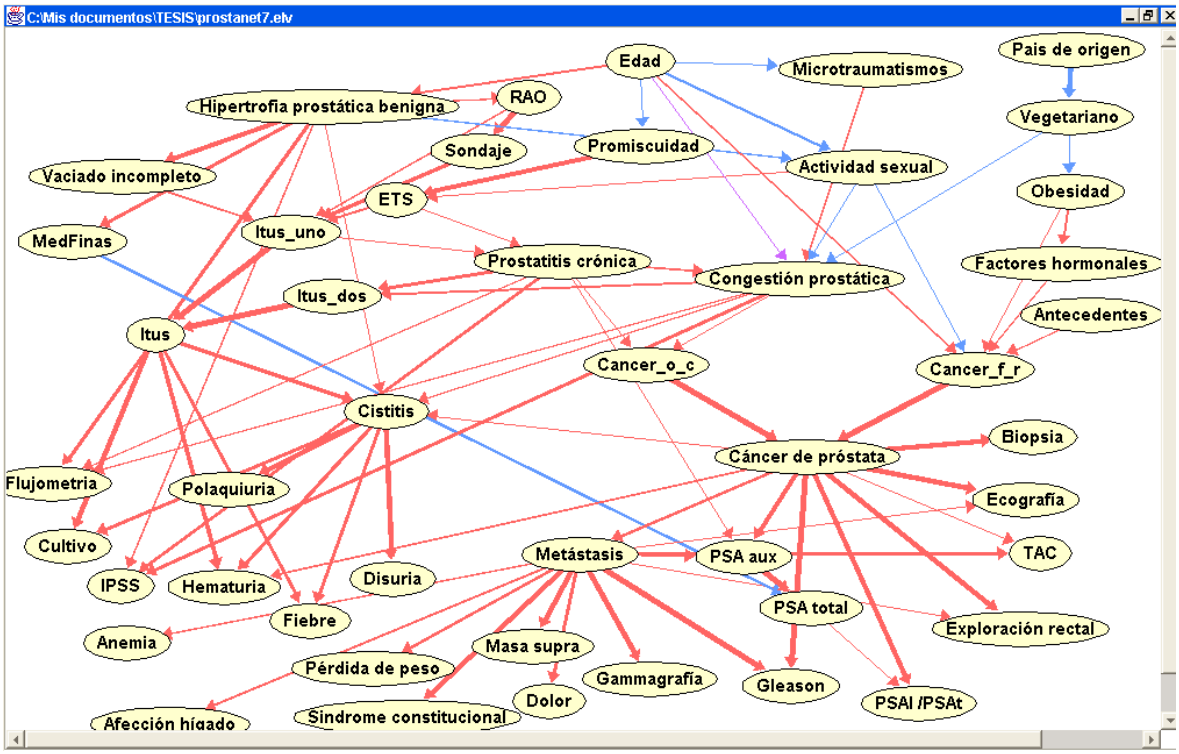

Figura 6.11: Última versión de la red PROSTANET

6.3.6. Versiones

En resumen, podemos observar que hemos tenido que trabajar sobre 6 versiones distintas de la red antes de conseguir la definitiva. En las filas de la tabla 6.2 se reflejan las principales características de cada una de ellas: en primer lugar, el identificador de la versión (*Versión*) y, a continuación, como curiosidad, la fecha en la que se creó (*Fecha*), ya que ayuda a corroborar la afirmación habitual de la gran cantidad de tiempo que lleva construir una red bayesiana con la ayuda de expertos humanos. El resto de las columnas corresponden, respectivamente, al número de nodos (*Nodos*) y de enlaces (*Enlaces*) de la red, a la cantidad de parámetros necesarios (*Parám.*), al número real de probabilidades a definir por el experto después de aplicar los modelos canónicos (*Real*) y el número de probabilidades que el experto asignó en dicha versión (*As. Exp.*), al número máximo de padres de los nodos de la red (*Padres*) y si en el modelo se habían detectado puertas OR y/o MAX (*OR/MAX*).

Si analizamos la primera columna, puede llamar la atención que desde la versión 4 hasta la 5 pasa más de un año. Sin embargo hemos de destacar que desde finales de marzo de 2001 hasta principios de enero de 2002 hubo un parón en el trabajo por motivos personales no achacables al experto. Por ello, esos 9 meses no se deben tener en cuenta a

Versión	Fecha	Nodos	Enlaces	Parám.	Real	As. Exp.	Padres	OR/MAX
1	14/12/00	26	30	1128	564	0	8	No
2	28/1/01	34	46	184	92	0	4	No
3	15/2/01	43	75	564	282	0	6	No
4	4/3/01	43	75	564	259	68	6	No
5	27/5/02	45	79	812	169	132	4	Sí
6	29/6/02	45	77	836	165	165	4	Sí
7	22/8/02	47	81	850	170	170	4	Sí

Tabla 6.2: Versiones de la red Prostanet

efectos de cómputos de tiempo.

Las versiones 1, 2 y 3 corresponden a las primeras fases de construcción de la red en que el modelo era sólo cualitativo (grafo de la red). Se caracterizan porque todas sus variables se consideraban binarias. En la columna "As. Exp." aparece un 0 lo que refleja el hecho de que ni siquiera se intentó la obtención de dichos parámetros. La versión 4 corresponde al resultado de discretizar las variables, siendo casi todas binarias salvo 3 ternarias y otras 3 con cuatro estados. A continuación aparecen las últimas versiones de la red que representan distintos refinamientos, todos asociados a puertas OR o MAX. Podemos observar que, a pesar de que el número de parámetros de la red es considerablemente mayor que en el resto, sin embargo, la cantidad de valores a obtener por el experto se reduce considerablemente. De hecho, si analizamos las fechas de creación de cada red, observamos que conseguir los 170 valores del experto nos ha llevado 8 meses, por lo que conseguir los aproximadamente 400 que hubiera tenido que calcular sin las puertas OR y MAX habría costado ¡casi 2 años!, suponiendo que el coste de obtener cada parámetro fuera el mismo, lo cual no es así. Creemos que no habría sido posible llevar a cabo con éxito dicha tarea.

A la luz de la tabla podemos concluir también que a partir de la tercera versión de la red el número de nodos y enlaces varía muy poco. En cualquier caso, la mayor parte de los esfuerzos se centraron en refinar el modelo de tal forma que se redujera al máximo el número de probabilidades a estimar por el experto pero reflejando fielmente el conocimiento médico. Ello se ha conseguido reduciendo el número máximo de padres de los nodos e identificando los modelos canónicos asociados a las relaciones entre las variables.

6.4. Evaluación de Prostanet

En la construcción de un sistema experto es esencial la evaluación. Díez [45], en su tesis doctoral, dedica un capítulo a este tema, destacando las tareas a realizar en todo procedimiento de evaluación:

- **Verificación,** para comprobar que satisface las especificaciones planteadas inicialmente. En el caso de redes bayesianas construidas con la ayuda de expertos humanos consistirá en comprobar que el dominio está modelado correctamente, tanto de forma cualitativa como cuantitativa.

- **Validación,** para determinar en qué medida el sistema se comporta de forma satisfactoria al resolver problemas reales. Esto se traduce en comprobar si los diagnósticos de los casos de pacientes que se introduzcan en la red son correctos. Cuando hay discrepancia entre el experto humano y el sistema, éste deberá intentar "convencer" al experto de que los resultados obtenidos son correctos.

- **Aceptación.** Para que un sistema experto sea utilizado en la práctica no basta con que sus resultados sean correctos: hace falta que tenga una interfaz fácil de manejar, que ahorre tiempo a los usuarios, que explique su razonamiento, etc. Todos estos factores influyen, junto con la calidad de los consejos que ofrece el sistema experto, en su aceptación.

Uno de los estudios más completos sobre evaluación de sistemas expertos es el aplicado a MYCIN [191]. Merece la pena citar sus resultados para tenerlos en cuenta a la hora de evaluar redes bayesianas médicas, porque en él se han basado otros autores para realizar la evaluación de sus trabajos [45, 134]. Las principales conclusiones de la evaluación de MYCIN son las siguientes:

- un nivel de aceptación de resultados por encima del 75 % es bastante bueno, pues el acuerdo entre los propios especialistas médicos no suele superar el 80 %;

- además, los expertos deben solucionar los mismos problemas que el sistema para que los evaluadores no sepan si el resultado lo ha producido uno u otro;

- es conveniente desarrollar cuestionarios breves con el fin de poder recoger los resultados lo más rápidamente posible.

6.4.1. Métodos

Para la evaluación de Prostanet hemos seguido varias técnicas, que han sido utilizadas en diferentes momentos del proceso de construcción de la red, ya que cada vez que se creaba una nueva versión de la red, ésta se evaluaba para comprobar que se adecuaba a las especificaciones médicas y, fundamentalmente, para comprobar que la nueva versión mejoraba en algún aspecto a las anteriores. Las herramientas utilizadas han sido, básicamente dos:

1. **Preguntas al experto,** tanto de forma oral como escrita. Es uno de los métodos más rápidos y sencillos, mediante el que hemos pretendido conseguir los siguientes objetivos:

   - *Detectar las dependencias/independencias* entre las variables de la red. Es bien sabido que una red bayesiana causal, aparte de representar relaciones causa-efecto, fundamentalmente tiene que reflejar relaciones de dependencia o independencia entre variables. Por tanto, hay que asegurarse de que dichas relaciones se representen correctamente en la red. Por ejemplo, para comprobar si el nodo ITUS separa condicionalmente los nodos ITUS_uno e ITUS_dos de Cistitis, le planteamos al experto preguntas del tipo: *Una vez que conocemos que el paciente tiene (o no) una infección urinaria (ITUS), ¿la probabilidad de que el cultivo dé positivo o negativo se ve afectada al conocer si la infección ha sido primaria (ITUS_uno) o secundaria (ITUS_dos)?.*
     Obviamente, este tipo de método de evaluación lo hemos realizado desde las primeras etapas de construcción de la red pues ha constituido la principal herramienta para detectar las relaciones entre los nodos.
   - *Comprobar la validez de los resultados* obtenidos por el sistema. Por ejemplo, si el valor que ofrecía la red para el estado presente de la variable Cáncer era de 0.12 preguntábamos al experto: *¿Confirma el dato de que 12 de cada 100 hombres padecen en la actualidad cáncer de próstata?* El hecho de poder interaccionar con el experto nos ha permitido poder reajustar los datos del modelo, especialmente las probabilidades, de forma inmediata, lo que ha reducido bastante el tiempo de construcción de la red.
   - Obtener un indicador de aceptación de la red.

2. **la capacidad de explicación de ELVIRA**, que nos ha permitido:

- *Comprobar la corrección de los resultados.* Esto lo hemos podido llevar a cabo de forma más o menos rápida y sencilla gracias a la posibilidad de seleccionar los nodos a expandir y de visualizar numérica y gráficamente las probabilidades en los nodos expandidos.
- *Analizar casos de pacientes.* Para ello, el experto nos proporcionó un conjunto de casos de pacientes ya diagnosticados por él. La evaluación consistió en introducir los hallazgos correspondientes a cada paciente en la red y comprobar si las probabilidades que ofrecía el sistema se correspondían con los datos reales. En este sentido, la posibilidad de gestionar el conjunto de casos de evidencia de forma simultánea nos permitió ir analizando cómo las probabilidades se iban modificando dependiendo de los nueva evidencia introducida, en el mismo orden en que el médico la recibe.
- *Realizar razonamiento hipotético.* En este caso, presentamos al experto varios hallazgos correspondientes a casos ficticios de pacientes y le pedíamos que nos proporcionara la probabilidad de padecer las 6 posibles enfermedades que hay representadas en PROSTANET.

6.4.2. Resultados

Para presentar los resultados de la evaluación, distinguiremos los relativos a cada uno de los apartados citados previamente.

Verificación

Como ya comentamos en la sección 6.3 del capítulo 6, las explicaciones (verbales y gráficas) generadas por ELVIRA nos han ayudado a confirmar que el modelo cualitativo está representado correctamente. Además, el experto se ha mostrado de acuerdo con las probabilidades a priori que se obtienen al compilar la red, lo que nos permite afirmar que los datos cuantitativos son adecuados.

Validación

Para la validación hemos utilizado las técnicas de análisis de sensibilidad sobre una variable, el análisis de casos de pacientes y el razonamiento hipotético para confirmar que la red resuelve de forma similar al experto los problemas propuestos.

El *análisis de sensibilidad* que hemos realizado ha consistido en analizar cómo la modificación de los valores de determinados hallazgos afectaban a la probabilidad de de-

terminadas variables. Para ello, se creaba un caso de evidencia por cada uno de los estados del hallazgo (o conjunto de hallazgos) sobre el que se realizaba el análisis. Por ejemplo, los datos estadísticos sobre la prevalencia del cáncer de próstata que manejan los médicos están basados, principalmente, en la edad. Por tanto, para ver si los resultados ofrecidos por la red eran los correctos creábamos un caso de evidencia por cada uno de los cuatro estados posibles del nodo Edad y centrábamos nuestra atención en la variable Cáncer de Próstata.

Respecto a los *casos de pacientes*, se han analizado un total de 20 casos de las historias clínicas del urólogo, obteniendo un diagnóstico similar al del experto en todos ellos, lo que permite confirmar que la red construida refleja adecuadamente el conocimiento médico. Como ejemplo, presentamos uno de los casos analizados:

Caso 1. Se trata del paciente Nº 14 (L.R.A), de 55 años de edad. El paciente tiene prostatitis crónica, además de disuria y poliaquiuria. El resultado de la prueba del tacto rectal detecta una consistencia fibrosa irregular. El valor del PSA es de 9.5 por lo que se solicita el del PSA libre, obteniendo un valor de 0.03. Se decide hacer una biopsia resultando positiva con un grado de gleason de 7. Por tanto, el diagnóstico final es de cáncer de próstata. En la tabla 6.3 mostramos los resultados que proporciona Prostanet cuando se ejecuta con ELVIRA. Por otro lado, la figura 6.12 muestra cómo ayuda ELVIRA al experto a comprobar que los resultados se ajustan a los diagnósticos realizados por ambos especialistas.

El *razonamiento hipotético* nos ha servido para comprobar que la red resuelve de forma similar los casos ficticios propuestos al experto. Para ello, al experto le hemos propuesto varios casos "inventados" y le pedíamos que asignara las probabilidades de padecer cada una de las posibles enfermedades de Prostanet. Nuevamente, no detectamos diferencias significativas entre los datos proporcionados por el experto y los que ofrecía la red.

Aceptación

Inicialmente, elaboramos un cuestionario con el objetivo de determinar el grado de utilidad de la implantación de Prostanet, puesto que nuestra intención era probar la red con más médicos, tanto especialistas como médicos de atención primaria. Sin embargo, aunque la autora de esta memoria intentó contactar con otros especialistas en urología, no hubo nadie que quisiera prestar su colaboración. La principal razón esgrimida fue la falta de tiempo libre para la dedicación al proyecto, aunque otras causas no menos importantes

Hallazgo	HPB	ITUS	Cistitis	Prostat.	Cg. Prost.	Cáncer	Metástasis
$\emptyset$	0.35	**0.46**	**0.44**	0.37	0.21	0.12	0.06
Edad=55	0.4	0.47	0.46	0.36	0.29	0.14	0.07
Dis=sí	0.49	0.79	**0.94**	0.53	0.43	0.18	0.09
Prost=sí	0.46	0.93	0.98	**1**	0.53	0.22	0.11
Polaq=sí	0.46	0.94	**1**	**1**	0.53	0.22	0.11
TactoR=fibr	0.46	0.94	1	1	0.52	0.07	0.02
PSA=9.5	0.41	0.94	1	1	0.53	0.1	0.03
PSA libre=0.03	0.45	0.92	1	1	0.56	**0.65**	0.2
Biop=pos	0.45	0.92	1	1	0.56	**0.97**	0.3
Gleason=7	0.47	0.91	1	1	0.58	**1**	0.57

Tabla 6.3: Probabilidades de cada uno de los posibles diagnósticos del paciente del caso 1 según el orden en el que se van introduciendo los hallazgos en PROSTANET.

detectadas fueron el desconocimiento del método de representación y el escepticismo que generan los sistemas informáticos, sobre todo en la clase médica. Por tanto, no podemos aportar más datos que los proporcionados por el experto que ha colaborado con nosotros. En este sentido, la valoración general que él hace es bastante positiva puesto que, según su opinión, PROSTANET representa de forma adecuada el conocimiento médico y el índice de acierto de los resultados es bastante alto. Por otra parte, le ha ayudado a considerar otro tipo de relaciones (causa-efecto, en lugar de efecto-causa) a las que no estaba acostumbrado. Además, le gustaría poder trabajar con un programa que esté basado en PROSTANET, al menos en su clínica privada. Incluso piensa que podría ser útil su implantación en los centros de atención primaria para poder ayudar a los médicos de familia a refinar más sus creencias sobre la existencia o no de cáncer de próstata, lo que permitiría agilizar los trámites para la derivación del paciente al especialista.

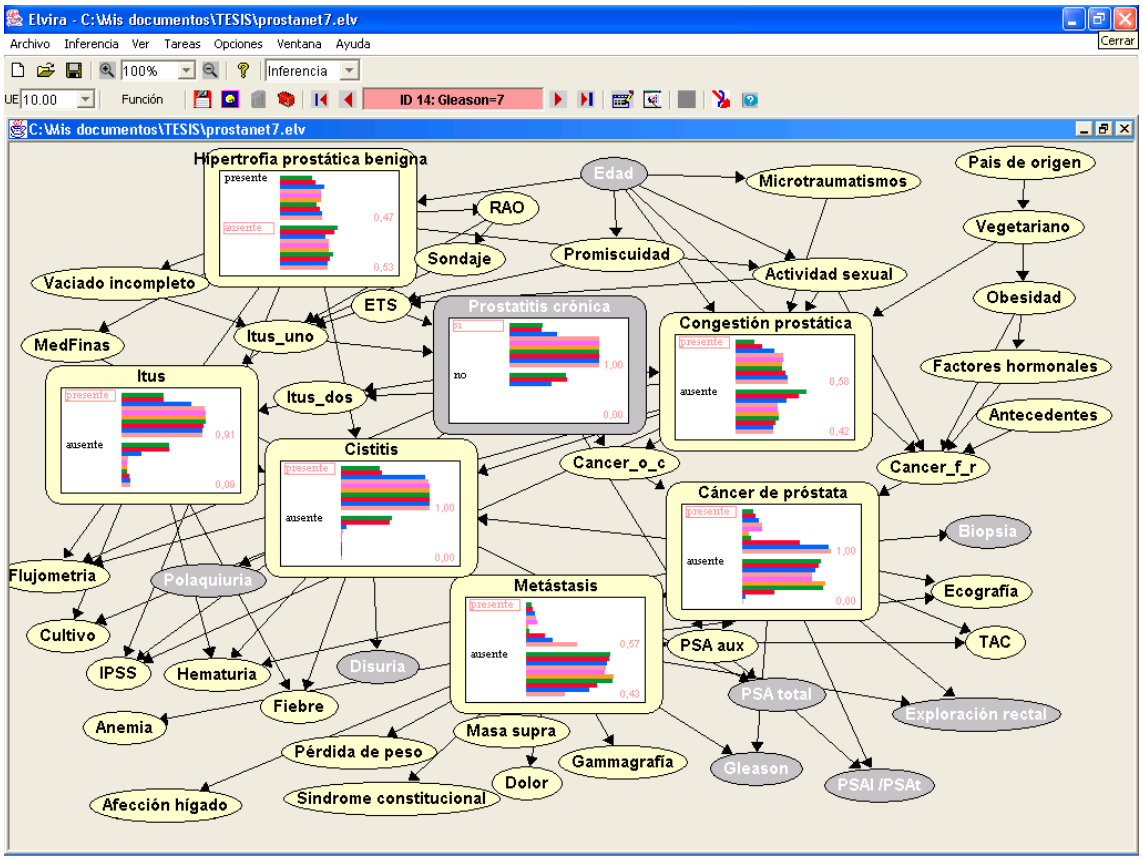

Figura 6.12: Análisis del caso 1 en ELVIRA.

Parte IV

CONCLUSIÓN

Capítulo 7
Conclusiones

Finalizamos esta memoria presentando las principales aportaciones que se pueden extraer del trabajo realizado (sección 7.1) y las posibles líneas de trabajo que quedan abiertas para continuar nuestra investigación en el futuro (sección 7.2).

7.1. Aportaciones

Al principio de nuestro trabajo nos planteábamos como principal objetivo el desarrollo de un método de explicación automática para redes bayesianas causales y desglosábamos este objetivo en una serie de tareas (sección 1.5). Ahora, presentamos las principales contribuciones de nuestro trabajo en función de este objetivo y estas tareas:

1. Hemos estudiado el **estado del arte** de la explicación en redes bayesianas (capítulo 2). Las principales conclusiones de este estudio son [96, 98]:

   - Puesto que no existe unanimidad en la interpretación del término *explicación* en el campo de los sistemas expertos, comenzamos ofreciendo nuestra propia definición: *"Explicar consiste en exponer algo de tal forma que sea comprensible para el receptor de la explicación, lo que implica que mejore su conocimiento sobre el objeto de la explicación, y satisfactorio, en el sentido de que cubra las expectativas del usuario"*(sección 2.1). De ella extrajimos una serie de características básicas inherentes a todo proceso de explicación en sistemas expertos, relacionadas con el contenido del mismo, la forma de comunicarlo y las posibilidades de adaptación al usuario (sección 2.2).
   - A la luz de estas propiedades, hemos realizado un estudio crítico de los métodos de explicación más relevantes desarrollados hasta la fecha para sistemas expertos, tanto para sistemas heurísticos (sección 2.3) como para redes bayesianas (sección 2.4).

- Una de las principales conclusiones de este estudio es que, lamentablemente hasta la fecha, no existe ningún método de explicación que satisfaga a los usuarios de redes bayesianas, puesto que los pocos que existen tienen múltiples carencias.

2. Con ese fin hemos desarrollado un **método de explicación** para redes bayesianas causales (capítulos 3 y 4), aunque muchas de sus capacidades son aplicables también a redes no causales. Teniendo en cuenta los resultados del estudio sobre los métodos de explicación tanto para sistemas expertos heurísticos como para redes bayesianas, hemos intentado aprovechar algunas de las ventajas de los mejores y cubrir las principales deficiencias encontradas. En este sentido, las características más significativas del método desarrollado son:

   - contempla la posibilidad de generar **explicaciones** tanto del **modelo** representado en la red bayesiana causal como del **razonamiento**, a nivel micro y a nivel macro. Además, estas explicaciones se pueden presentar de forma verbal y de modo gráfico.
   - La **explicación verbal** del modelo consiste en mostrar la información asociada a un nodo o a un enlace seleccionado por el usuario, o de la red completa.
   - Para ello, otra de las aportaciones consiste en **clasificar los nodos y los enlaces**, como ya propusiera Druzdzel [50], con el fin, entre otros, de generar textos coherentes y comprensibles para el usuario.
   - La **expansión de los nodos** permite presentar de forma numérica y gráfica las probabilidades asociadas a cada estado, además de destacar el estado más probable.
   - La **explicación gráfica de los enlaces** se basa en los trabajos de Wellman [184, 185, 186] sobre la noción del signo de la influencia en redes cualitativas, y consiste en representar el tipo de influencia que transmite cada nodo a sus hijos, dibujando los enlaces con distintos colores. La novedad de enfoque reside en la definición de un orden parcial entre las distribuciones acumuladas de probabilidad de dos nodos como forma de analizar la influencia entre ellos. La principal ventaja de nuestro método reside en que al realizar la propagación de la evidencia mediante algoritmos cuantitativos, no existen tantas ambigüedades como las que se producen en la propagación con algoritmos cualitativos de Wellman.

- Por otro lado, hemos propuesto una forma de representar la **magnitud de la influencia** que se transmite de un nodo a cada uno de sus hijos, basándonos exclusivamente en sus tablas de probabilidad. En la interfaz gráfica, la magnitud de la influencia se representa mediante el grosor de los enlaces.
- La **expansión selectiva** de los nodos dependiendo del *factor de relevancia* y de la *función* de cada uno de ellos permite simplificar las explicaciones gráficas de la red completa, además de facilitar al usuario el análisis de la información relativa a un nodo determinado.
- La posibilidad de gestionar distintos **casos de evidencia** de forma simultánea propicia la visualización de los resultados del análisis de sensibilidad de cada nodo respecto de la evidencia, y el razonamiento hipotético, ofreciendo un modo sencillo de estudiar los resultados obtenidos en ambos casos.
- Para ello, hemos definido nuevas formas de medir las variaciones en la probabilidad de los nodos, siempre y cuando éstos sean ordinales, haciendo uso de su *valor normal*.
- La representación gráfica de las **variaciones de probabilidad** en los nodos con respecto al caso seleccionado por el usuario, dependiendo del *umbral de variación*, constituye otra poderosa herramienta para analizar la sensibilidad de los nodos a determinados hallazgos.
- La visualización de los **caminos** por los que fluye la evidencia, junto con la representación gráfica del **cambio** producido en cada nodo y de las influencias asociadas a cada enlace, ayuda al usuario a comprender cómo se han obtenido los resultados proporcionados por el sistema.
- La **clasificación de los hallazgos** que forman la evidencia dependiendo del tipo y la cantidad de influencia que ejercen sobre una variable determinada permite obtener información sobre el porqué de los resultados obtenidos sobre dicha variable.

3. Hemos realizado un **análisis comparativo de las principales herramientas** de edición y evaluación de redes bayesianas (sección 5.4), basándonos en las posibilidades que ofrece cada una de ellas para los siguientes aspectos claves: la edición, la inferencia y la explicación.

4. Dicho estudio, y en particular lo referente a explicación, nos ha servido para incluir en ELVIRA distintas herramientas que corresponden a la **implementación del**

método de explicación (capítulo 5) desarrollado en este trabajo.

5. Con este objetivo, la autora de esta memoria ha colaborado activamente en la **programación de ELVIRA**, para lo que se ha utilizado el paradigma de la Programación Orientada a Objetos, apoyado por la definición de Tipos Abstractos de Datos como metodología de representación, y JAVA como lenguaje de programación.

6. Las principales propiedades de las facilidades de explicación con las que cuenta ELVIRA son:

   - el usuario tiene la posibilidad de solicitar diferentes tipos de explicación, que ELVIRA genera de forma automática, tanto verbalmente (en castellano y en inglés) como gráficamente, a niveles micro y macro, de cualquier red bayesiana que se edite en Elvira.
   - Las explicaciones únicamente se ofrecen cuando el usuario lo solicita, lo que significa que no se ofrece al usuario más información de la que desea en cada momento, con lo que se evita:
     - la posibilidad de "sobreinformación", con los problemas que conlleva de cara a la presentación de los resultados y a su evaluación por parte del usuario; y
     - la disminución del rendimiento del sistema, ya que el resto de las tareas se han implementado de forma independiente a las de explicación.
   - Además, ELVIRA controla las posibles inconsistencias que se pueden producir al intentar realizar acciones no permitidas como, por ejemplo, la inserción de un arco que define un ciclo en la red o la introducción de evidencia imposible. En estos casos, el programa proporciona un aviso al usuario del tipo de error cometido o de la imposibilidad de realizar la tarea seleccionada, ofreciéndole, en su caso, información sobre cómo corregirlo.

7. El método de explicación implementado en ELVIRA lo hemos puesto a prueba mediante la **construcción de la red Prostanet** (capítulo 6). Para llevar a cabo esta tarea, hemos refinado el proceso de construcción manual de redes bayesianas, añadiendo nuevas etapas relacionadas con la evaluación de la red mediante las explicaciones proporcionadas a los expertos humanos y haciendo hincapié en la identificación de modelos canónicos por las ventajas que aportan de cara a la obtención de probabilidades.

8. Hemos demostrado también cómo la herramienta de explicación desarrollada facilita la construcción manual de redes bayesianas, tanto en la definición del modelo como en la parte más complicada: la asignación de las probabilidades por el experto (sección 6.3). Probablemente, sin una herramienta de explicación como la que incluye Elvira la estimación satisfactoria de las probabilidades de la red habría sido mucho más compleja y los resultados mucho menos fiables.

9. Como prueba de esto es que PROSTANET ha conseguido un elevado índice de exactitud en todos los diagnósticos de los casos evaluados (sección 6.4).

10. Además, debido a nuestra participación en el proyecto *Diagnostic Systems Based on Graphical Decision-Theoretic Methods* (PST.CLG.976167), subvencionado por el Programa Científico de la OTAN, el método de explicación desarrollado se ha utilizado también para **construir y depurar Hepar II** [135]. Esta experiencia ha servido para confirmar nuevamente la utilidad de la capacidad de explicación de ELVIRA. Especialmente ventajoso han resultado las explicaciones gráficas de los nodos y de los enlaces, puesto que ha servido para facilitar tanto la construcción del grafo como la depuración de las probabilidades aprendidas de una base de datos. Además, la gestión simultánea de distintos casos de evidencia ha ayudado a justificar los resultados que aparentemente podrían no estar suficientemente claros [100].

11. Esto demuestra que los métodos desarrollados en este trabajo no son útiles solamente para el dominio particular en que hemos construido nuestra red, sino que se pueden aplicar también en otro dominio médico en el que no habíamos pensado inicialmente.

7.2. Trabajo futuro

A partir de ahora, nuestro trabajo de investigación queda abierto en diferentes líneas:

1. Las **explicaciones textuales** se pueden mejorar para adaptarlas a las nuevas posibilidades que ofrecen las tecnologías de la información, por lo que sería interesante poder generar automáticamente textos hipermedia con distintos niveles de detalle, incluso con otro tipo de datos como vídeo, voz, imágenes, etc. Para ello habría que refinar las distintas taxonomías para poder mejorar las clasificaciones de los nodos y los enlaces y para tener en cuenta otros dominios de aplicación, no sólo la medicina. Además, se podrían incluir técnicas de planificación de diálogos que permitieran interactuar de modo más directo con el usuario dependiendo de sus posibilidades.

En este sentido, otra posibilidad consiste en facilitar al usuario la explicación verbal únicamente de la porción de la red que le interese, lo que puede ser útil en los casos en los que la red tenga muchos nodos y enlaces, puesto que, en este caso, la explicación verbal de la red puede ser algo confusa por incluir demasiada información. Así, mediante un proceso de *refinamientos sucesivos*, el usuario puede centrarse en la parte que le interese.

2. Uno de los aspectos en los que no hemos llegado a profundizar en esta memoria es el de la **adaptación al usuario.** Por ello, una de las líneas de trabajo de cara al futuro consiste en la definición de distintos tipos de usuario, dependiendo del conocimiento que tengan tanto del dominio representado, como de la teoría subyacente a las redes bayesianas. En este sentido, sería deseable poder contar con un método de explicación capaz de adaptarse de forma dinámica al usuario.

3. Sería interesante mejorar la **representación gráfica** de los datos teniendo en cuenta distintas técnicas de visualización de datos, como por ejemplo el tamaño de los nodos dependiendo de su importancia, la intensidad de color dependiendo de la influencia recibida, etc.

4. Haciendo uso del **factor de importancia** de cada nodo, se podría comparar dicho valor con la relevancia computacional que tiene. Sería muy interesante poder analizar la importancia estadística respecto de la importancia médica asignada subjetivamente por los expertos. Por otra parte, la modificación de este factor de importancia de forma dinámica en función de la evidencia observada en cada caso se podría aprovechar para explicar las líneas de razonamiento, especialmente con fines educativos,

5. **Simplificación de la red**. A pesar de que el grafo que representa la red es un medio muy útil para generar explicaciones intuitivas, en los casos en los que la red tiene un gran número de nodos, esto puede ocasionar más confusión que claridad. Por otra parte, no todos los nodos de la red tienen por qué afectar a una variable: dependerá de la evidencia introducida, de la estructura de la red, etc. Por tanto, sería conveniente no incluir en la explicación gráfica aquellos nodos que no estén relacionados ni con la evidencia ni con la variable de interés, con el objetivo de no confundir al usuario ofreciéndole información irrelevante. Por otra parte, dado que los algoritmos de propagación son computacionalmente complejos (su coste crece exponencialmente con el tamaño de la red en el caso de los algoritmos exactos, y

exponencialmente con los valores de los parámetros en el caso de los algoritmos aproximados), el funcionamiento del sistema puede ganar precisión y eficiencia al omitir la información que no es necesaria y centrarse únicamente en la que es relevante para cada situación, analizado en [50, 72, 159, 172, 174]. Por todos estos motivos, otra de las herramientas proporcionadas por nuestro método para explicar el efecto de la evidencia sobre una hipótesis se basa en la simplificación gráfica de la red. Continuando en esta línea, la idea consistiría en mostrar sólo los nodos relevantes sobre los que influye la evidencia correspondiente al caso que se desea visualizar. Para ello, se pueden utilizar los criterios de separación direccional y de irrelevancia, entendida ésta como independencia estadística [71, 72, 164] y/o como independencia computacional [159].

Otra simplificación que sería conveniente tener en cuenta aquí es la propuesta por Druzdzel [50] y Suermondt [174] de eliminar los nodos *molestos*, es decir, los nodos que están computacionalmente relacionados con la hipótesis dada la evidencia pero que no forman parte de los caminos entre la evidencia y la hipótesis.

Por último, se puede hacer uso también del factor de importancia y de la función de cada nodo para hacer simplificaciones más generales, además de aplicar el criterio de "normalidad" [71], es decir, mostrar solamente aquellos nodos cuyo estado más probable sea distinto del estado normal.

6. También habrá que trabajar en la mejora de los algoritmos de explicación del razonamiento con el objetivo de reducir su complejidad ya que actualmente, en la mayoría de ellos, crece exponencialmente dependiendo del número de nodos de la red, lo que constituye una de sus mayores desventajas.

7. Debido a las dificultades que existen para desarrollar redes bayesianas con la ayuda de expertos humanos, sería deseable poder ampliar ELVIRA con una **interfaz** para creación de redes bayesianas de forma manual, de modo que los expertos no dependieran tanto de los ingenieros de conocimiento.

8. Gracias a la evaluación positiva de PROSTANET, y el interés del urólogo con el que hemos trabajado por los sistemas expertos, nos ha propuesto como tarea inminente la creación de una interfaz para poder trabajar en su clínica privada con PROSTANET (en el hospital público no disponen, hasta el momento, de los medios necesarios). Además, otro de sus deseos es que intentemos ampliarla para incluir otras posibles patologías relacionadas con la urología. Sin embargo, antes de eso nos gustaría

realizar un análisis comparativo de los resultados ofrecidos por Prostanet en contraposición con los datos que proporcionan las Tablas de Partin, que es el protocolo que utilizan los urólogos habitualmente para el diagnóstico del cáncer de próstata.

9. Por último, otro de nuestros objetivos es el de la aplicación de redes bayesianas causales a otros dominios, especialmente los relacionados con la educación y con el procesamiento del lenguaje natural.

Bibliografía

[1] ACM/IEEE. Computing Curricula 2001. Final Draft. Informe técnico, Computer Society of the Institute for Electrical and Electronic Engineers (IEEE) and the Association for Computing Machinery (ACM), 2001.

[2] S.K. Andersen, K.G. Olesen, F.V. Jensen, y F. Jensen. HUGIN: A shell for building Bayesian belief universes for expert systems. En *Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI–89)*, páginas 1080–1085, Detroit, MI, 1989.

[3] G. Andriola y W.J. Catalona. Using PSA to screen for prostatae cancer. *Urol Clin North America*, 20:647–651, 1993.

[4] Informatics Applied Statistics Group of the Institute for Systems and Safety. Sensitivity analysis. http://sensitivity-analysis.jrc.cec.eu.int, 2000.

[5] E. Biris. Automatic modelling using Bayesian networks for explanation generation. En *Proceedings of the 13th International Workshop on Qualitative Reasoning (QR'99)*, Pavia, Italy, 1999.

[6] H.L. Bleich. Computer-based consultation: Electrolyte and acid-base disorders. *American Journal of Medicine*, 53:285–291, 1972.

[7] C.C. Boring, T.S. Squieres, y T. Tong et al. Cancer statistics. *Cancer Journal Clinique*, 44:7–26, 1994.

[8] G. D. Bryant y G. R. Norman. Expressions of probability: Words and numbers. *New England Journal of Medicine*, 302:411, 1980.

[9] B.G. Buchanan y E.H. Shortliffe, editores. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA, 1984.

[10] A. Cano y S. Moral. Propagación exacta y aproximada con árboles de probabilidad. En *Actas de la VII Conferencia de la Asociación Española para la Inteligencia Artificial*, páginas 635–644, 1997.

[11] A. Cano, S. Moral, y A. Salmerón. Penniless propagation in join trees. *International Journal of Intelligent Systems*, 15:1027–1059, 2000.

[12] A. Cano, S. Moral, y A. Salmerón. Lazy evaluation in Penniless propagation over join trees. *Networks*, 39:175–185, 2002.

[13] G. Carenini, V. Mittal, y J. Moore. Generating patient-specific interactive natural language explanations. En *Proceedings of 18th Symposium on Computer Applications in Medical Care (SCAMC-94)*. McGraw-Hill, 1994.

[14] B. De Carolis, F. De Rosis, F. Grasso, A. Rossiello, D. Berry., y T. Gillie. Generating recipient-centered explanations about drug prescription. *Artificial Intelligence in Medicine*, 8:123–145, 1996.

[15] E. Castillo, J. M. Gutiérrez, y A. S. Hadi. *Expert Systems y Probabilistic Network Models*. Springer Verlag, Nueva York, 1997. Versión española: *Sistemas Expertos y Modelos de Redes Probabilísticas*, Academia de Ingeniería, Madrid, 1997.

[16] E. Castillo, J. M. Gutiérrez, y A. S. Hadi. Sensitivity analysis in discrete Bayesian networks. *IEE Transactions on Systems, Man and Cybernetics*, 27(4):412–423, 1997.

[17] A. Cawsey. Generating interactive explanations. En *Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91)*, páginas 86–91, Anaheim, California, 1991.

[18] A. Cawsey. User-modelling in interactive explanations. *Journal of User Modelling and User Adapted Interaction*, 3(3):221–247, 1993.

[19] A. Cawsey. Developing an explanation component for a knowledge-based system: Discussion. *Expert Systems with Applications*, 8(4):527–531, 1994. Special issue on explanation.

[20] A. Cawsey. *Explanation and Interaction*. MIT Press, Cambridge, MA, 1995.

[21] A. Cawsey, J. Galliers, B. Logan, S. Reece, y K. S. Jones. Revising beliefs and intentions: A unified framework for agent interaction. En A. Sloman et al., editor,

Proceedings of the Conference of the Society for Artificial Intelligence and Simulation of Behaviour, páginas 130–139, Amsterdam, 1993. IOS Press.

[22] E. Charniak y S.E. Shimony. Cost-based abduction and MAP explanation. *Artificial Intelligence*, 66(2):345–374, 1994.

[23] Z. Chen. Explanation as a solution reconstruction in causal networks and expert systems. *Kybernetes*, 24(8):41–46, 1995.

[24] W. J. Clancey. The epistemology of rule-based expert systems. A framework for explanation. *Artificial Intelligence*, 20:215–251, 1983.

[25] W. J. Clancey. Notes on "Heuristic classification". *Artificial Intelligence*, 59:191–196, 1993.

[26] Elvira Consortium. Elvira: An environment for probabilistic graphical models. En *Proceedings of the First International Workshop on Probabilistic Graphical Models*, páginas 222–230, Cuenca, España, 2002.

[27] G. Cooper. *NESTOR: A computer-based medical diagnostic aid that integrates causal and diagnostic knowledge*. Tesis doctoral, Stanford University, Stanford, CA, 1984.

[28] G. F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligente*, 42:393–405, 1990.

[29] G.F. Cooper y E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–348, 1992.

[30] V.M.H. Coupe, L.C. van der Gaag, y J.D.F. Habbema. Sensitivity analysis: an aid for probability elicitation. *Knowledge Engineering Review*, 15(3):1–18, 2000.

[31] F.G. Cozman. Generalizing variable elimination in Bayesian networks. En *Proceedings of the Workshop on Probabilistic Reasoning in Artificial Intelligence*. Atibaia, Brasil, 2000.

[32] L. M. de Campos, J.A. Gámez, y S. Moral. Partial abductive inference in Bayesian belief networks using a genetic algorithm. *Pattern Recognition Letters*, 20:1211–1217, 1999.

[33] L. M. de Campos, S. Moral, y J.A. Gámez. Simplifying explanations in Bayesian belief networks. Informe técnico DECSAI-99-02-01, Dept. Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, España, 1999.

[34] L.M. de Campos. Aprendizaje automático de modelos gráficos I: Métodos básicos. En J. A. Gámez y J. M. Puerta, editores, *Sistemas Expertos Probabilísticos*, páginas 113–140. Universidad de Castilla-La Mancha, Cuenca, 1998.

[35] L.M. de Campos y J.M. Puerta. Stochastic local and distributed search algorithms for learning belief networks. En *Proceedings of the III International Symposium on Adaptive Systems: Evolutionary Computation and Probabilistic Graphical Model*, páginas 109–115, 2001.

[36] L.M. de Campos y J.M. Puerta. Stochastic local search algorithms for learning belief networks: Searching in the space of orderings. En *Lecture Notes in Artificial Intelligence*, volumen 2143, páginas 228–239, 2001.

[37] H.M. Deitel y P.J. Deitel. *Cómo programar en JAVA*. Prentice Hall, México, 1ª edición, 1998.

[38] A. P. Dempster. A generalization of Bayesian inference. *Journal of the Royal Statistical Society, Series B*, 30:205–247, 1968.

[39] F. J. Díez. Parameter adjustment in Bayes networks. The generalized noisy OR–gate. En *Proceedings of the 9th Conference on Uncertainty in Artificial Intelligence (UAI'93)*, páginas 99–105, Washington D.C., 1993. Morgan Kaufmann, San Mateo, CA.

[40] F. J. Díez. Local conditioning in Bayesian networks. *Artificial Intelligence*, 87:1–20, 1996.

[41] F. J. Díez. Causality and probabilistic independence in graphical models. En *EURO XV - INFORMS XXXIV*, Barcelona, España, 1997.

[42] F. J. Díez. From causal graphs to Bayesian networks. En *Proceedings of the Workshop on Conditional Independence Structures and Graphical Models*, Toronto, Canadá, 1999.

[43] F. J. Díez y M. Druzdzel. Canonical probabilistic models for knowledge engineering. Informe técnico IA-02-01, Dpto. Inteligencia Artificial, UNED, Madrid, 2002.

[44] F. J. Díez, J. Mira, E. Iturralde, y S. Zubillaga. Diaval, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine*, 10:59–73, 1997.

[45] F.J. Díez. *Sistema Experto Bayesiano para Ecocardiografía.* Tesis doctoral, UNED, Dept. de Informática y Automática, 1994.

[46] S.L. Dittmer y F.V. Jensen. Tools for explanation in Bayesian networks with application to an agricultural problem. En *Proceedings of the First European Conference for Information Technology in Agriculture*, Copenague, Dinamarca, 1997.

[47] D. Draper. Plan explanation. Informe técnico, Rockwell Palo Alto Laboratory, Palo Alto, CA, 1997. URL: www.rpal.rockwell.com/∼draper/ARPI/expl.html.

[48] H.L. Dreyfus. *Mind Over Machine.* Free Press, Nueva York, 1986.

[49] M. Druzdzel. Verbal uncertainty expressions: Literature review. Informe técnico CMU-EPP-1990-03-02, Dept. Engineering and Public Policy, Carnegie Mellon University, 1989.

[50] M. Druzdzel. *Probabilistic reasoning in decision support systems: From computation to common sense.* Tesis doctoral, Dept. Engineering and Public Policy, Carnegie Mellon University, 1993.

[51] M. Druzdzel. Explanation in probabilistic systems: Is it feasible? Will it work? En *Proceedings of the Workshop on Intelligent Information Systems*, Deblin, Polonia, 1996.

[52] M. Druzdzel. Qualitative verbal explanations in Bayesian belief networks. *Artificial Intelligence and Simulation of Behaviour Quarterly*, 94:43–54, 1996. Special issue on Bayesian Belief networks.

[53] M. Druzdzel. GeNIe: A development environment for graphical decision-analytic models. En *Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association (AMIA-1999)*, pág. 1206. Washington, D.C., 1999.

[54] M. Druzdzel y M. Henrion. Using scenarios to explain probabilistic inference. En *Working Notes of the AAAI-90 Workshop on Explanation*, páginas 133–141, Boston, MA, 1990.

[55] M. Druzdzel y M. Henrion. Belief propagation in qualitative probabilistic networks. En N. Piera Carrete y M.G. Singh, editores, *Qualitative Reasoning and Decision Technologies*, páginas 451–460. CIMNE, Barcelona, España, 1993.

[56] M. Druzdzel y M. Henrion. Efficient reasoning in qualitative probabilistic networks. En Richard Fikes y Wendy Lehnert, editores, *Proceedings of the 11th National Conference on Artificial Intelligence*, páginas 548–553, Menlo Park, CA, 1993. AAAI Press.

[57] M. Druzdzel y H.J. Suermondt. Relevance in probabilistic models: Backyards in a small world. En *Working Notes of the AAAI-94 Workshop on Fall Symposium Series: Relevance*, páginas 60–63, Nueva Orleans, LA, 1994.

[58] M. Druzdzel y L. van der Gaag. Elicitation of probabilities for belief networks: combining qualitative and quantitative information. En *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, páginas 141–148, Montreal, Canadá, 1995.

[59] M. Druzdzel y L.C. van der Gaag. Building probabilistic networks:"Where do the numbers come from?". Guest Editors Introduction. *IEEE Transactions on Knowledge and Data Engineering*, 12:481–486, 2000.

[60] R. O. Duda, J. Gaschnig, y P. E. Hart. Model design in the PROSPECTOR consultant program for mineral exploration. En D. Michie, editor, *Expert Systems in the Microelectronic Age*, páginas 153–167. Edinburgh University Press, Edimburgo, UK, 1979.

[61] R.O. Duda, P.E. Hart, y N.Ñilsson. Subjective Bayesian methods for rule-based inference systems. *Proceedings of the 1976 National Computer Conference (AFIPS)*, 45:1075–1082, 1976.

[62] C. Elsaesser. Verbal expressions for probability updates. How much more probable is "much more probable"? En P.P. Bonisone, M. Henrion, L.N. Kanal, y J.F. Lemmer, editores, *Uncertainty in Artificial Intelligence*, volumen 6, páginas 387–400. Elsevier Science Publishers B.V. (North-Holland) Science Publishers, Amsterdam, 1990.

[63] C. Elsaesser y M. Henrion. Explanation of probabilistic inference. En M. Henrion, R.D. Shachter, L.N. Kanal, y J.F. Lemmer, editores, *Uncertainty in Artificial Intelligence*, volumen 5, páginas 319–328. Elsevier Science Publishers B.V. (North-Holland), 1989.

[64] B. Falkenheiner y K. Forbus. Compositional Modelling: finding the right model for the job. *Artificial Intelligence*, 51:95–143, 1991.

[65] P.J. Feltovitch. *Knowledge-Based Components of Expertise in Medical Diagnosis.* Tesis doctoral, Deparment of Educational Psychology, University of Minnesota, St. Paul, MN, 1981.

[66] P.J. Feltovitch, P.E. Johnson, J.H. Moller, y D.B. Swanson. LCS: The role and development of medical knowledge in diagnostic expertise. En W.J. Clancey y E.H. Shortliffe, editores, *Readings in Medical Artificial Intelligence: The First Decade*, páginas 275–319. Addison-Wesley, Reading, MA, 1984.

[67] N. Fenton y M. Neil. The "jury observation fallacy" and the use of Bayesian networks to present probabilistic legal arguments. *Mathematics Today (Bulletin of the IMA)*, 36(6):180–187, 2000.

[68] A. Fiedler. Dialog-driven adaptation of explanations of proofs. En *Proceedings of the Seventeenth International Joint Conference of Artificial Intelligence*, páginas 1295–1300, Seattle, Washington, 2001.

[69] A. Fiedler. P.rex: An interactive proof explainer. En *Proceedings of the First International Conference on Automated Reasoning*, Siena, Italia, 2001.

[70] J. Folckers, C. Möbus, O. Schröder, y H. Thole. An intelligent problem solving environment for designing explanation models and for diagnostic reasoning in probabilistic domains. En *Proceedings of the Third International Conference on Intelligent Tutoring Systems, (ITS-96)*, páginas 353–362, Montreal, Canadá, 1996.

[71] J.A. Gámez. *Inferencia Abductiva en Redes Causales.* Tesis doctoral, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, España, 1998.

[72] D. Geiger, T. Verma, y J. Pearl. Identifying independence in Bayesian networks. *Networks*, 20:507–534, 1990.

[73] C. Glymour y G.F. Cooper, editores. *Computation, Causation and Discovery.* MIT Press, Cambridge, MA, 1999.

[74] I.J. Good. Explicativity: A mathematical theory of explanation with statistical applications. *Proceedings of the Royal Statistical Society*, (354):303–330, 1977.

[75] T.R. Gruber y P. Gautier. Machine-generated explanations of engineering models: A compositional modelling approach. En *Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93)*, páginas 1502–1508, 1993.

[76] J. Guttag. Abstract data types and the development of data structures. *Communications of the ACM*, 20(6):396–404, 1977.

[77] P. Haddawy, J. Jacobson, y C. Kahn. An educational tool for high-level interaction with Bayesian networks. En *Proceedings of the 6th IEEE International Conference on Tools with Artificial Intelligence*, New Orleans, Louisiana, 1994.

[78] P. Haddawy, J. Jacobson, y C. Kahn. Generating explanations and tutorial problems from Bayesian networks. En *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, páginas 770–774, Washington, 1994.

[79] P. Haddawy, J. Jacobson, y C. Kahn. BANTER: A Bayesian network tutoring shell. *Artifical Intelligence in Medicine*, 10:177–200, 1997.

[80] D. W. Hasling, W. J. Clancey, y G. Rennels. Strategic explanations for a diagnostic consultation system. *International Journal of Man-Machine Studies*, 20:3–19, 1984.

[81] D. Heckerman, D. Geiger, y D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[82] D. E. Heckerman. *Probabilistic Similarity Networks*. MIT Press, Cambridge, MA, 1991.

[83] E.M. Helsper y L. van der Gaag. Building Bayesian networks through ontologies. En F. van Harmelen, editor, *Proceedings of the 15th European Conference on Artificial Intelligence (ECAI-2002)*, páginas 680–684, Lyon, Francia, 2002.

[84] M. Henrion. Some practical issues in constructing belief networks. En M. Henrion, R.D. Shachter, L.N. Kanal, y J.F. Lemmer, editores, *Uncertainty in Artificial Intelligence*, volumen 5, páginas 161–173. Elsevier Science Publishers B.V. (North-Holland), 1989.

[85] M. Henrion y M. Druzdzel. Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning. En *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, páginas 17–32, 1990.

[86] L.D. Hernández, S. Moral, y A. Salmerón. A Monte Carlo algorithm for probabilistic propagation in belief networks based on importance sampling and stratified simulation techniques. *International Journal of Approximate Reasoning*, 18:53–91, 1998.

[87] R.A. Howard y J.E. Matheson. Influence diagrams. En R.A. Howard y J.E. Matheson, editores, *Readings on the Principles and Applications of Decision Analysis*, páginas 719–762. Strategic Decisions Group, Menlo Park, CA, 1984.

[88] S. Ríos Insua, M.C. Bielza Lozoya, y A. Mateos Caballero. *Fundamentos de los Sistemas de Ayudas a la Decisión*. RA-MA, Madrid, España, 2002.

[89] F.V. Jensen. *An Introduction to Bayesian Networks*. UCL Press, 1996.

[90] F.V. Jensen. *Bayesian Networks and Decision Graphs*. Statistics for Engineering and Information Science. Springer, New York, USA, 2001.

[91] F.V. Jensen, S.L. Lauritzen, y K.G. Olesen. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly*, 4:269–282, 1990.

[92] P.E. Johnson. What kind of expert should a system be? *Journal of Medicine and Philosophy*, 8:77–97, 1983.

[93] D. Kahneman, P. Slovic, y A. Tversky, editores. *Judgement under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK, 1982.

[94] J.H. Kim y J. Pearl. A computational model for combined causal and diagnostic reasoning in inference system. En *Proceedings of the 8th International Joint Conference on Artificial Intelligence (IJCAI-83)*, páginas 190–193. Morgan-Kaufmann, 1983.

[95] C. Lacave, R. Atienza, y F. J. Díez. Graphical explanations in Bayesian networks. En R.W.Brause y E. Hanisch, editores, *Lecture Notes in Computer Science. Medical Data Analysis. First International Symposium, ISMDA 2000*, volumen 1933, páginas 122–129, Frankfurt, Germany, September 2000. Springer.

[96] C. Lacave y F. J. Díez. Estudio de los métodos de explicación en redes bayesianas. Informe técnico IA-98-01, Dpt. Inteligencia Artificial, UNED, Madrid, 1998.

[97] C. Lacave y F. J. Díez. A review of explanation methods for Bayesian networks. Informe técnico IA-00-01, Dpt. Inteligencia Artificial, UNED, Madrid, 2000.

[98] C. Lacave y F.J. Díez. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, 17, December 2002.

[99] C. Lacave, J. Giralt, y F. J. Díez. Red bayesiana para el diagnóstico de enfermedades infecciosas. Informe Técnico IA-98-02, Dept. Inteligencia Artificial, UNED, Madrid, 1998.

[100] C. Lacave, A. Oniśko, y F. J. Díez. Debugging medical Bayesian networks with Elvira's explanation capability. En *Working Notes of the Workshop on Bayesian Models in Medicine, Eighth European Conference on Artificial Intelligence in Medicine (AIME-2001)*, páginas 47–52, Cascais, Portugal, 2001.

[101] C.P. Langlotz, E.H. Shortliffe, y L.M. Fagan. A methodology for generating computer-based explanations of decision-theoretic advice. *Medical Decision Making*, 8(4):290–303, 1988.

[102] K.B. Laskey y S.M. Mahoney. Network engineering for agile belief network models. *IEEE Transactions on Knowledge and Data Engineering*, 12:487–498, 2000.

[103] S.L. Lauritzen y D.J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, 50(2):157–224, 1988.

[104] S.L. Lauritzen y N. Wermuth. Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, (17):31–54, 1989.

[105] A.Y. Levy, Y. Iwasaki, y R. Fikes. Automated model selection for simulation based on relevance reasoning. *Artificial Intelligence*, 96:351–394, 1997.

[106] Y. Lin y M. Druzdzel. Computational advantages of relevance reasoning in Bayesian belief networks. En *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, páginas 342–350. Morgan Kaufmann, 1997.

[107] K. Ludwig. Functionalism, causation and causal relevance. *Psyche. An Interdisciplinary Journal of Research and Consciousness*, 4(3), 1998. http://psyche.cs.monash.edu.au/v4/psyche-4-03-ludwig.html.

[108] D. Madigan, K. Mosurski, y R. Almond. Graphical explanations in belief networks. *Journal of Computational and Graphic Statistics*, 6(2):160–181, 1997. Special issue on Bayesian belief networks.

[109] A.L. Madsen y F.V. Jensen. Lazy propagation: a junction tree inference algorithm based on lazy evaluation. *Artificial Intelligence*, 113:203–245, 1999.

[110] C. Di Marco, G. Hirst, L. Wanner, y J. Wilkinson. Healthdoc: Customizing patient information and health education by medical condition and personal characteristics. En *First International Workshop on Artificial Intelligence in Patient Education*, Glasgow, UK, 1995.

[111] D. Marr. *Vision*. Freeman, Nueva York, 1982.

[112] P. Juez Martel y F.J. Díez Vegas. *Probabilidad y Estadística en Medicina. Aplicaciones en la Práctica Clínica y en la Gestión Sanitaria*. Ed. Díaz de Santos, Madrid, 1996.

[113] R. McConachy, K. Korb, y I. Zukerman. Deciding what not to say: An attentional-probabilistic approach to argument presentation. En *Proceedings of the Twentieth Annual Meeting of the Cognitive Science Society*, páginas 669–674, Madison, WI, 1998.

[114] S. McRoy, S. Haller, y S. Ali. B2: An interactive tool for explaining Bayesian reasoning in natural language. En *Conference Proceedings of Energy-Week 97*, volumen V, páginas 153–159, Houston, TX, 1997.

[115] S. McRoy, S. Haller, y S. Ali. Uniform knowledge representation for language processing in the B2 system. *Journal of Natural Language Engineering*, 3(2):123–145, 1997.

[116] S. McRoy, A. Liu-Perez, J. Helwig, y S. Haller. B2: A tutoring shell for Bayesian networks that supports natural language interaction. En *Working Notes of the AAAI-96 Spring Symposium on Artificial Intelligence*, Stanford, CA, 1996.

[117] E. Mendoza y A. Fernández. Hipertrofia benigna de próstata. *MEDIFAM*, 10(6):334–344, 2000.

[118] B. Meyer. *Construcción de Software Orientado a Objetos*. Prentice Hall, Madrid, 2ª edición, 1999.

[119] P.R. Milgrom. Good news and bad news: Representation theorems and applications. *Bell Journal of Economics*, 12:380–391, 1981.

[120] M.L. Minsky. A framework for representing knowledge. *Mind Design*, páginas 95–128, 1981.

[121] J. Mira, A.E. Delgado, J.G. Boticario, y F.J. Díez. *Aspectos Básicos de la Inteligencia Artificial*. Sanz y Torres, Madrid, 1995.

[122] J. Moore. *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context*. MIT Press, Cambridge, MA, 1994.

[123] J. Moore y W.R. Swartout. Pointing: a way toward explanation dialogue. En *Proceedings of the 8th Natural Conference on Artificial Intelligence*, páginas 457–464, Boston, MA, 1990.

[124] J. Ferrater Mora. *Diccionario de Filosofía Abreviado*. Editorial Sudamericana, Buenos Aires, 1970.

[125] S. Moral, R. Rumí, y A. Salmerón. Mixtures of truncated exponentials in hybrid Bayesian networks. En *Lecture Notes in Artificial Intelligence*, volumen 2143, páginas 156–167, 2001.

[126] M.A. Nakao y S. Axelrod. Numbers are better than words: Verbal specifications of frequency have no place in medicine. *American Journal of Medicine*, 74:1061–1065, 1983.

[127] P. Naughton. *Manual de JAVA*. Osborne/McGraw-Hill, Madrid, 1996.

[128] P.P. Nayak. Causal approximations. *Artificial Intelligence*, 70:277–334, 1994.

[129] R.E. Neapolitan. *Probabilistic Reasoning in Expert Systems: Theory and Algorithms*. Wiley-Interscience, Nueva York, 1990.

[130] M. Neil, N. Fenton, y L. Nielsen. Building large-scale Bayesian networks. *Knowledge Engineering Review*, 15(3):257–284, 2000.

[131] A. Newell. The knowledge level. *AI Magazine*, páginas 1–20, 1981.

[132] D. Nilsson. An efficient algorithm for finding the M most probable configurations in Bayesian networks. *Statistics and Computing*, 9:159–173, 1998.

[133] K.G. Olesen, U. Kjaerul, F. Jensen, F.V. Jensen, B. Falk, S. Andraessen, y S.K. Andersen. A MUNIN network for the median nerve. A case study in loops. *Artificial Intelligence*, 3:385–404, 1989.

[134] A. Oniśko. *Probabilistic Causal Models in Medicine: Application to Diagnosis of Liver Disorders*. Tesis doctoral, Institute of Computer Science, Białystok University of Technology, Białystok, Polonia, 2002.

[135] A. Oniśko, M. Druzdzel, y H. Wasyluk. Extension of the Hepar II model to multiple-disorder diagnosis. En M. Kłopotek, M. Michalewicz, y S.T. Wierzchoń, editores, *Intelligent Information Systems. Advances in Soft Computing Series*, páginas 303–313. Springer-Verlag, Heidelberg, Alemania, 2000.

[136] S. Parsons. *Qualitative Methods for Reasoning under Uncertainty*. MIT Press, Cambridge, MA, 2001.

[137] A.W. Partin, M.W. Kattan, y E.N. Subong et al. Combination of prostate-specific antigen, clinical stage, and Gleason score to predict pathologic stage of localized prostate cancer. A multi-institutional update. *JAMA*, 277:1445–1451, 1977.

[138] R. Patil, P. Szolovits, y W. Schwartz. Modeling knowledge of the patient in acid-base and electrolyte disorders, 1982.

[139] R. Peña. *Diseño de Programas. Formalismo y Abstracción*. Prentice Hall, Madrid, 2ª edición, 1998.

[140] J. Pearl. Fusion, propagation and structuring in belief netowrks. *Artificial Intelligence*, 29:241–288, 1986.

[141] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Mateo, CA, 1988.

[142] J. Pearl. Reasoning with cause and effect. Informe técnico, DeptD of Computer Science. University of California, Los Angeles, California, 1999. http://bayes.cs.ucla.edu/jphome.html.

[143] J. Pearl. *Causality. Models, Reasoning and Inference*. Cambridge University Press, Cambridge, UK, 2001.

[144] N. Pennington y R. Hastie. Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14(3):521–533, 1988.

[145] M.A. Peot y R.D. Shachter. Fusion and propagation with multiple observations in belief networks. *Artificial Intelligence*, 48:299–318, 1991.

[146] M. Pradham, M. Henrion, G. Provan, B. del Favero, y K. Huang. The sensitivity of belief networks to imprecise probabilities: An experimental investigation. *Artificial Intelligence*, 85(1-2):363–397, 1996.

[147] R.S. Pressman. *Ingeniería del Software. Un Enfoque Práctico.* McGraw-Hill, Madrid, 3ª edición, 1993.

[148] J.A. Reggia y B.T. Perricone. Answer justification in medical decision support systems based on Bayesian classification. *Computers in Biology and Medicine*, 15(4):161–167, 1985.

[149] S. Renooij. *Qualitative Approaches to Quantifying Probabilistic Networks.* Tesis doctoral, Institute for Information and Computing Sciences, Utrecht University, Holanda, 2001.

[150] S. Renooij y L. van der Gaag. Enhancing QPNs for trade-off resolution. En *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, páginas 559–566, Stocolmo, Suecia, 1999. Morgan Kaufmann.

[151] S. Renooij, L. van der Gaag, S. Parsons, y S. Green. Pivotal prunning of trade-offs in QPNs. En *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, páginas 515–522, Stanford, CA, 2000. Morgan Kaufmann.

[152] A. Salmerón, A. Cano, y S. Moral. Importance sampling in Bayesian networks using probability trees. *Computational Statistics and Data Analysis*, 34:387–413, 2000.

[153] A. Salmerón y S. Moral. Importance sampling in Bayesian networks using antithetic variables. *Lecture Notes in Artificial Intelligence*, 2143:168–179, 2001.

[154] E. Santos. On the generation of alternative explanations with implications for belief revision. En *Proceedings of the 7th Conference on Uncertainty in Artificial Intelligence*, páginas 339–347. Morgan Kaufmann, San Mateo, CA, 1991.

[155] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[156] P. Sember y I. Zukerman. Strategies for generating micro explanations for Bayesian belief networks. En P.P. Bonissone, M. Henrion, L.N. Kanal, y J. F. Lemmer, editores, *Uncertainty in Artificial Intelligence*, volumen 6, páginas 295–302. Elsevier Science Publishers B.V. (North-Holland), 1990.

[157] R.D. Shachter. DAVID: Influence diagram processing system for the Macintosh. En *Proceedings of the Second Workshop on Uncertainty in Artificial Intelligence*, páginas 243–248, Stanford, CA, 1986. Morgan Kaufmann.

[158] R.D. Shachter. Evaluating influence diagrams. *Operations Research*, 34(6):871–882, 1986.

[159] R.D. Shachter. An ordered examination of influence diagrams. *Networks*, 20:535–563, 1990.

[160] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

[161] O. Shcröder, C. Mobus, J. Folckers, y H. Thole. Supporting the construction of explanation models and diagnostic reasoning in probabilistic domains. Informe técnico, OFFIS Institute, Oldenburg, Alemania, 1996.

[162] P.P. Shenoy. Binary join trees for computing marginals in the Shenoy-Shafer architecture. *International Journal of Approximate Reasoning*, 17:239–263, 1997.

[163] P.P. Shenoy y G.R. Shafer. Axioms for probability and belief-function propagation. En P.P. Bonissone, M. Henrion, L.N. Kanal, y J. F. Lemmer, editores, *Uncertainty in Artificial Intelligence*, volumen 6, páginas 169–198. Elsevier Science Publishers B.V. (North-Holland), 1990.

[164] S.E. Shimony. *A Probabilistic Framework for Explanation*. Tesis doctoral, DeptD of Computer Science, Brown University, 1991. Informe técnico CS-91-57.

[165] D. J. Spiegelhalter, R. C. G. Frankling, y K. Bull. Assessment, criticism and improvement of imprecise subjective probabilities. En M. Henrion, R. D. Shachter, L.Ñ. Kanal, y J. F. Lemmer, editores, *Uncertainty in Artificial Intelligence 5*, páginas 285–294. Elsevier Science Publishers, Amsterdam, 1990.

[166] D.J. Spiegelhalter, A.P. Dawid, S.L. Lauritzen, y R.G. Cowell. Bayesian analysis in expert systems. *Statistical Science*, 8(3):219–283, 1993.

[167] D.J. Spiegelhalter y R.P. Knill-Jones. Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroentorology. *Journal of the Royal Statistical Society*, 147:35–77, 1984.

[168] P. Spirtes, C. Glymour, y R. Scheines. *Causation, Prediction and Search*. Springer-Verlag, Nueva York, 1993.

[169] S. Srinivas y J. Breese. IDEAL: A software package for analysis of influence diagrams. En *Proceedings of the 6th Workshop on Uncertainty in Artificial Intelligence*, páginas 212–219, Cambridge, MA, 1990.

[170] T.M. Strat. The generation of explanations within evidential reasoning systems. En *Proceedings of the 10th International Joint Conference on Artificial Intelligence*, páginas 1097–1104, Milán, Italia, 1987.

[171] T.M. Strat y J.D. Lowrance. Explaining evidential analysis. *International Journal of Approximate Reasoning*, 3(4):299–353, 1989.

[172] D. Subramanian, R. Greiner, y J. Pearl. The relevance of relevance. *Artificial Intelligence*, 97(1-2):1–5, 1997.

[173] L. E. Súcar, D. F. Gillies, y D. A. Gillies. Objective probabilities in expert systems. *Artificial Intelligence*, 61:187–208, 1993.

[174] H.J. Suermondt. *Explanation in Bayesian Belief Networks*. Tesis doctoral, Dept. Computer Science. Stanford University, Stanford, CA, 1992. STAN–CS–92–1417.

[175] H.J. Suermondt y G. Cooper. An evaluation of explanations of probabilistic inference. *Computers and Biomedical Research*, 26:242–254, 1993.

[176] W. Swartout. XPLAIN: A system for creating and explaining expert consulting programs. *Artificial Intelligence*, 21:285–325, 1983.

[177] R. L. Teach y E. H. Shortliffe. An analysis of physician's attitudes. En B. G. Buchanan y E. H. Shortliffe, editores, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, chapter 34, páginas 635–652. Addison-Wesley, Reading, MA, 1984.

[178] L.C. van der Gaag, S. Renooij, C.L.M. Witteman, B. Aleman, y B.G. Taal. How to elicit many probabilities. En *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, páginas 647–654, Stocolmo, Suecia, 1999.

[179] P. van der Linden. *Just JAVA 2*. JAVA series. The Sun Microsystem Press - Prentice Hall, CA, USA, 4ª edición, 1999.

[180] A. Vizcaino. Enseñando POO en el primer curso de programación. ¿Acierto o error? 2002.

[181] J.W. Wallis y E.H. Shortliffe. Customized explanations using causal knowledge. En B.G. Buchanan y E.H. Shortliffe, editores, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, chapter 20, páginas 371–388. Addison-Wesley, Reading, MA, 1984.

[182] T.S. Wallsten y D.V. Budescu. A review of human linguistic probability processing: General principles and empirical evidence. *Knowledge Engineering Review*, 10(1):43–62, 1995.

[183] H. Wang y M. J. Druzdzel. User interface tools for navigation in conditional probability tables and elicitation of probabilities in Bayesian networks. En *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, páginas 617–625, Stanford, CA, 2000. Morgan Kaufmann.

[184] M. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44:257–303, 1990.

[185] M. Wellman. Graphical inference in qualitative probabilistic networks. *Networks*, 20:687–701, 1990.

[186] M. Wellman y M. Henrion. Explaining "explaining away". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):287–292, 1993.

[187] M. Wick y J.R. Slagle. An explanation facility for today's expert systems. *IEEE Expert*, 4:24–36, 1989.

[188] M.R. Wick. The 1988 Workshop on Explanation. *Artifical Intelligence Magazine*, 10(3):22–26, 1989.

[189] M.R. Wick y W.B. Thompson. Reconstructive expert system explanation. *Artifical Intelligence*, 54:33–70, 1992.

[190] I. Yip, D. Heber, y W. Aronson. Nutrición y cáncer de próstata. En K.J. Dienta, editor, *Clínicas de Urología de Norteamérica*, volumen 2, páginas 429–435. McGraw-Hill Interamericana, 1999.

[191] V.L. Yu, L.M. Fagan, S.M. Wraith, W.J. Clancey, A.C. Scott, J.F. Hannigan, R.L. Blum, B.G. Buchanan, y S.N. Cohen. An evaluation of MYCIN's advice. En B. G. Buchanan y E. H. Shortliffe, editores, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*, chapter 31, páginas 589–596. Addison-Wesley, Reading, MA, 1984.

[192] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

[193] L.A. Zadeh. Fuzzy sets as a basis for the theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[194] N.L. Zhang y D. Poole. Exploiting causal independence in Bayesian network inference. *Journal of Artificial Intelligence Research*, 5:301–328, 1996.

[195] I. Zukerman, K. Korb, y R. McConachy. Perambulations on the way to an architecture for a nice argument generator. En *Notes of the ECAI-96 Workshop on Gaps and Bridges: New Directions in Planning and Natural Language Generation*, páginas 31–36, Budapest, Hungría, 1998.

[196] I. Zukerman, R. McConachy, y R. Korb. Bayesian reasoning in an abductive mechanism for argument generation and analysis. En *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, páginas 833–838, Madison, WI, 1998.

Apéndice A

Especificación algebraica del TAD *Grafo*

Especificación sintáctica, para la que suponemos los siguientes TADs auxiliares:
VÉRTICE (sección 5.1.4), ARCO (sección 5.1.4), CONJUNTO, LÓGICO

Operaciones generadoras:
GRAFOVACIO: → GRAFO
GRAFOVACIO(): Genera el grafo vacío
INSVERTICE: GRAFO x VÉRTICE → GRAFO
INSVERTICE(G,v): Añade el vértice v *a G*
INSARCO: GRAFO x ARCO → GRAFO
INSARCO(G,a): Añade el arco a *a G*

Operaciones de acceso:
ESGRAFOVACIO: GRAFO → LÓGICO
ESGRAFOVACIO(G) devuelve verdadero si G es vacío y falso en caso contrario
PERTV: VÉRTICE x GRAFO → LÓGICO
PERTV(v,G): Decide si el vértice v *pertenece al grafo G*
PERTAR: ARCO x GRAFO → LÓGICO
PERTAR(a,G): Decide si el arco a *pertenece al grafo G*
VERTICES: GRAFO → CONJUNTO
VERTICES(G): Devuelve el conjunto de vértices del grafo G
ARCOS: GRAFO → CONJUNTO
ARCOS(G): Devuelve el conjunto de arcos del grafo G
ADYACENTES: GRAFO x VÉRTICE → CONJUNTO
ADYACENTES(G,v): devuelve el conjunto de vértices adyacentes a v *en el grafo G*
INCIDENTES: GRAFO x VÉRTICE → CONJUNTO

INCIDENTES(G,v): devuelve el conjunto de vértices incidentes a v *en el grafo G*

Operaciones modificadoras:

BORRARV: GRAFO x VÉRTICE → GRAFO

BORRARV(G,v) elimina del grafo G el vértice v, si está y todos los arcos relacionados con él

BORRARA: GRAFO x ARCO → GRAFO

BORRARA(G,a) elimina del grafo G el arco a, si está

Especificación semántica $\forall G \in GRAFO,\ \forall v1, v2 \in VERTICE,\ \forall a1, a2 \in ARCO$

Axiomas de equivalencia, que definen la igualdad de los valores representados por términos sintácticamente distintos:

$$\text{INSVERTICE(INSVERTICE(G,v1),v2)} = \begin{cases} \text{INSVERTICE}(G, v1), & v1 = v2 \\ \text{INSVERTICE}(\text{INSVERTICE}(G, v2), v1), & c.c. \end{cases}$$

$$\text{INSARCO}(\text{INSARCO}(G, a1), a2) = \begin{cases} \text{INSARCO}(G, a1), & a1 = a2 \\ \text{INSARCO}(\text{INSARCO}(G, a2), a1), & c.c. \end{cases}$$

$$\text{INSARCO}(G, a1) = G, \neg\ \text{PERTV}(\text{ORIGEN}(a1), G)\ \wedge\ \neg\text{PERTV}(\text{DESTINO}(a1), G)$$

Axiomas de acceso

$$\text{ESVACIO}(\text{GRAFOVACIO}) = Verdadero$$
$$\text{ESVACIO}(\text{INSVERTICE}(G, v1)) = Falso$$
$$\text{ESVACIO}(\text{INSARCO}(G, a1)) = \text{ESVACIO}(G)$$

$$\text{PERTV}(v1, \text{GRAFOVACIO}) = Falso$$
$$\text{PERTV}(v1, \text{INSVERTICE}(G, v2)) = \begin{cases} Verdadero & v1 = v2, \\ \text{PERTV}(v1, G) & c.c. \end{cases}$$

$$\text{PERTAR}(a1, \text{GRAFOVACIO}) = Falso$$
$$\text{PERTAR}(a1, \text{INSVERTICE}(G, v3)) = \text{PERTAR}(a1, G)$$
$$\text{PERTAR}(a1, \text{INSARCO}(a2, G)) = \begin{cases} \text{PERTAR}(a2, G) & a1 = a2, \\ \text{PERTV}(a1, G) & c.c. \end{cases}$$

VERTICES (GRAFOVACIO) = CJTOVACIO
VERTICES(INSVERTICE($v1, G$)) = INSCJTO(*v1*,VERTICES(G))
VERTICES(INSARCO(*a1, G*)) = VERTICES(G)

ARCOS(GRAFOVACIO) = CJTOVACIO
ARCOS(INSVERTICE(*v1, G*)) = ARCOS(G)
ARCOS(INSARCO(*a1, G*))=INSCJTO(*a1*, ARCOS(G))

ADYACENTES (GRAFOVACIO, *v1*) = CJTOVACIO
ADYACENTES (INSVERTICE (*G, v1), v2*) = ADYACENTES (*G, v2*)
ADYACENTES (INSARCO (*G, a1), v1*) =

$$\begin{cases} \text{INSCJTO (}v2\text{,ADYACENTES (}G\text{, DESTINO(}a1\text{)))} & \text{ORIGEN(}a1\text{)=}v1\text{,} \\ \text{ADYACENTES (}G\text{, }v1\text{)} & c.c. \end{cases}$$

INCIDENTES (GRAFOVACIO, *v1*) = CJTOVACIO
INCIDENTES (INSVERTICE (*G, v1), v2*) = INCIDENTES (*G, v2*)
INCIDENTES (INSARCO (*G, a1), v1*) =

$$\begin{cases} \text{INSCJTO (}v2\text{,INCIDENTES (}G\text{, ORIGEN(}a1\text{)))} & \text{DESTINO(}a1\text{)=}v1\text{,} \\ \text{INCIDENTES (}G\text{, }v3\text{)} & c.c. \end{cases}$$

Axiomas de modificación

BORRARV (GRAFOVACIO, *v1*) = GRAFOVACIO
BORRARV (INSVERTICE (*G, v1), v2*) =

$$\begin{cases} G & v1 = v2, \\ \text{INSVERTICE (BORRARV (}G\text{, }v2\text{), }v1\text{)} & c.c. \end{cases}$$

BORRARV(INSARCO(*G,a1),v*)=

$$\begin{cases} \text{BORRARV(}G\text{,}v\text{)} & \text{ORIGEN(}a1\text{)=}v \ \vee \ \text{DESTINO(}a1\text{)=}v, \\ \text{INSARCO(BORRARV(}G\text{,}v\text{),}a1\text{)} & c.c. \end{cases}$$

BORRARA(GRAFOVACIO,*a1*)=GRAFOVACIO
BORRARA(INSVERTICE(*G,v),a1*)= INSVERTICE(BORRARA(*G,a1),v*)
BORRARA(INSARCO(*G,a1),a2*)=

$$\begin{cases} G & a1 = a2, \\ \text{INSARCO(BORRARA(}G\text{,}a2\text{),}a1\text{)} & c.c. \end{cases}$$