UNIVERSIDAD DE CANTABRIA

TESIS DOCTORAL

MODELOS PROBABILÍSTICOS PARA UTILIZACIÓN EN SISTEMAS EXPERTOS

Presentada por: ELENA ÁLVAREZ SÁIZ
Dirigida por: ENRIQUE CASTILLO RON

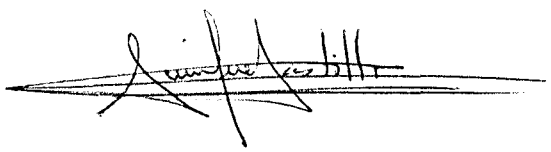
Santander, Diciembre 1989

Don Enrique Castillo Ron, catedrático de Matemática Aplicada de la Escuela Técnica Superior de Ingenieros de Caminos,

CERTIFICA,

Que el presente trabajo sobre "Modelos probabilísticos para utilización en Sistemas Expertos", ha sido efectuado bajo su dirección por Dña Elena Álvarez Sáiz en el Departamento de Matemática Aplicada y Ciencias de la Computación, y una vez revisado autoriza su presentación.

Y para que conste y a los efectos oportunos expido el presente certificado en Santander, a uno de diciembre de mil novecientos ochenta y nueve



Fdo: Enrique Castillo Ron

ÍNDICE

ÍNDICE

1 ESTADO DEL CONOCIMIENTO

ESTADO DEL CONOCIMIENTO Y CONTRIBUCIONES ORIGINALES

1.-ESTADO DEL CONOCIMIENTO Y CONTRIBUCIONES ORIGINALES

1.1.-¿QUÉ ES LA INTELIGENCIA ARTIFICIAL?

Antes de intentar definir la Inteligencia Artificial (Barr et al. (1981), Simons (1985), Gaines (1986)) es preciso hacer ciertas consideraciones sobre la inteligencia natural para comprender perfectamente las dificultades que encierra este propósito. A pesar de que el término "inteligencia" es utilizado con frecuencia en nuestro lenguaje habitual no existe una definición aceptada universalmente. La razón es que es un concepto con muchas dimensiones que va cambiando a medida que se producen nuevos conocimientos biológicos y tecnológicos. A lo largo los últimos años se han producido fuertes debates filosóficos al intentar definir la inteligencia. En la Tabla 1.1 se muestran algunas de las definiciones más utilizadas según un estudio de Pyle (1979).

La Inteligencia Artificial se centra en aspectos reducidos de las capacidades mentales: percepción, comprensión del lenguaje, sintetización de voz, manipulación de información, etc. A pesar de que los sistemas informáticos no poseen todas las características de la inteligencia natural pueden considerarse inteligentes en el sentido de que son capaces de enfrentarse a los problemas tomando decisiones óptimas mediante procesos de inferencia.

Por otra parte, llegar a medir la "inteligencia" de una máquina resulta difícil, ya que no puede pensarse en la actualidad en dar criterios objetivos para cuantificar el grado de inteligencia cuando este concepto resulta complejo y difuso. No obstante, se han propuesto algunas pruebas para ello. En este sentido Alan Turing, en 1950, propuso un test para determinar si una máquina era o no inteligente. Este método requiere de dos personas A y B y de la máquina C objeto de análisis. A, B y C se encuentran separadas en tres habitaciones diferentes, cada una de ellas con una terminal mediante la cual pueden intercomunicarse. Una de las personas A realiza el papel de interrogador y recibe las respuestas de B y C. Si A no puede determinar a través de las respuestas cual es la persona y cual la máquina entonces la máquina ha pasado satisfactoriamente el Test de Turing y puede considerarse inteligente.

La Inteligencia Artificial y los sistemas expertos han irrumpido vertiginosamente en el mundo actual. Puede decirse no sólo que hoy están de moda, sino que constituyen uno de los temas de más actualidad y que más recursos humanos y materiales está consumiendo. Antes de que los sistemas

expertos fueran realidad, ya existían éstos en la mente de muchos seres humanos; en efecto, las películas de ciencia ficción se encargaban de hacernos ver que las máquinas podían no sólo sustituir al hombre en muchas de sus funciones inteligentes sino también hacerlo mucho mejor que él, e incluso mucho más barato. La figura más representativa de este fenómeno es "el robot", al que tradicionalmente se le ha dado una forma humana para hacer resaltar aún más esta sustitución. Nos encontramos, pues, ante un reto de imprevisibles consecuencias, ya que son muchos los cambios que se avecinan y que trastocarán muchos de los paradigmas que rigen el mundo de hoy.

Investigador	Definición
Binot	Capacidad de discernimiento, comprensión y razonamiento
Heim	Capacidad para captar la esencia de las cosas y actuar en consecuencia
Spearman	Capacidad de relacionar y correlacionar
Terman	Capacidad para desarrollar conceptos y comprensión de su significado
Vernan	Capacidad de pensar
Wechster	Capacidad de actuar según su propósito, pensar racionalmente y relacionarse eficazmente con su entorno

Tabla 1.1.- Definiciones de inteligencia

Antes de centrarnos en los sistemas expertos es conveniente analizar el desarrollo de la llamada Inteligencia Artificial. Aunque su origen se remonta a mediados de este siglo, la idea de fabricar modelos computacionales capaces de realizar facultades propias de los seres humanos era ya ambicionada desde hace muchos años. Son muchas las obras de literatura a lo largo de la historia que predicen nuevas tecnologías. Las obras de ficción han "inventado", quizás desde un punto de vista muy imaginativo, máquinas electrónicas a imagen del hombre. En un principio consideraron a los robots (palabra derivada de la checa "robota" que significa trabajador) como máquinas para ayudar al hombre en aquellos trabajos más duros, pero no se tardó mucho en dotarles de sensibilidad y emotividad, pudiendo actuar incluso con inteligencia. En el mundo real, sin embargo, ciertas limitaciones producidas por el avance de la tecnología no han hecho posible la creación de máquinas con las posibilidades que les ha dado la fantasía.

El uso de computadores para estudiar actividades mentales ha sido posible sólo después de la invención del computador en los años 40. En 1937, Alan Turing

presentó el concepto de "máquina universal de Turing" en una ponencia sobre "números computables" que permitió la descripción de todos los ordenadores que aparecieron en los años posteriores.

Anteriormente se habían creado objetos de forma mecánica (juguetes, autómatas...), muy elaborados, y que daban la apariencia de tener vida. Hay documentos que muestran que ya en el siglo VIII en China se contaba con una especie de monje mecánico que pedía limosna extendiendo las manos. En Suiza, Pierre y Henr-Louis Jaquet Droz, en el siglo XVIII, inventaron un autómata que podía escribir, dibujar y tocar instrumentos musicales. En el siglo XIX aparecieron máquinas que podían hablar pudiendo interrogar y responder preguntas. Sin embargo, estos objetos mecánicos simulaban seres humanos, dando la impresión de que poseían vida e inteligencia.

Una de las primeras personas que contribuyó al desarrollo de la Inteligencia Artificial fue Turing (1963). Tanto él como H. Simon (1969), éste último de la Universidad de Cambridge y posteriormente premio Nobel, expusieron diferentes argumentos a favor de la programación de un ordenador con posibilidad de realizar comportamientos inteligentes. En una ponencia en el año 1950, Turing rebatió hasta nueve argumentos en contra de las máquinas inteligentes respondiendo a cada una de ellas.

Cuando finalizó la segunda guerra mundial científicos ingleses y americanos, en grupos separados, trabajaron sobre la creación de máquinas electrónicas que fueran capaces, mediante el suministro de instrucciones, de realizar operaciones numéricas complejas.

Un grupo reducido de científicos, preocupados por emular los razonamientos y comportamientos humanos, continuó investigando sobre el procesamiento de información simbólica. Ya que se pretendía trabajar con conocimientos e ideas, debían surgir ordenadores capaces de manipular símbolos que no fueran numéricos.

En 1956, John McCarthy, profesor auxiliar de matemáticas del Dartmounth College (Estados Unidos), organizó una conferencia sobre informática teórica. Se cree que fue entonces la primera vez que se utilizó el término "Inteligencia Artificial" para referirse a esta rama independiente de la Informática.

En esta conferencia estuvieron cuatro de los hombres que más influyeron en el desarrollo de la Inteligencia Artificial en los años posteriores: McCarthy, Marvin Minsky, Allen Newell y Herbert Simon. Newell y Simon (1962) presentaron un trabajo sobre demostración de teoremas (Logic Theorist). Junto con J.C. Show desarrollaron el Information Processing Language (IPL), lenguaje que permitía el

proceso de conceptos por ordenador. Marvin Minsky y McCarthy fueron cofundadores del Grupo de Inteligencia Artificial del Instituto Tecnológico de Massachussetts (MIT), además de ser el último creador de un lenguaje de procesamiento de listas: Lisp (Mc Carthy (1962)).

Aunque fueron pocas las personas interesadas en aquella conferencia quedó patente la importancia de las investigaciones llevadas a cabo en el campo de la Inteligencia Artificial y se vislumbraban los avances que tendrían lugar en la década siguiente.

En 1959 en el grupo de Inteligencia Artificial del MIT, Mc Carthy desarrolló el Advice Taker considerado por algunos autores el primer ejemplo de ingeniería del conocimiento.

En 1981 tuvo lugar, en la Universidad de British Columbia, Vancouver (Canadá), la séptima Conferencia Internacional de Inteligencia Artificial (IJCAI) donde se trataron diferentes temas: sistemas expertos, lenguaje natural, visión artificial, métodos de búsqueda, etc., además de celebrar el veinticinco aniversario de la célebre conferencia del Dartmounth College, que se considera la fecha del nacimiento de la Inteligencia Artificial.

Desde hace aproximadamente quince años se han creado grupos para obtener resultados prácticos de programas de Inteligencia Artificial que pudieran aplicarse socialmente permitiendo su utilización a los diferentes tipos de profesionales. En un principio el éxito no fue el esperado pues se carecía de ordenadores que permitieran desarrollar programas de esta naturaleza. Entre 1975 y 1980, con la llegada de las técnicas de miniaturización electrónica, surgieron ordenadores bastante más potentes, que dieron lugar a una generación de ordenadores más rápidos y con unas prestaciones mucho más amplias que las anteriores.

Son muchas las áreas de las que se ocupa la Inteligencia Artificial. Algunas de ellas son (ver figura 1.1):

(a) demostración de teoremas
(b) juegos inteligentes
(c) proceso de lenguaje natural
(d) robótica
(e) visión artificial
(f) sistemas expertos

En los años cincuenta y sesenta fueron desarrollados con gran ímpetu los dos primeros temas mencionados, quedando actualmente relegados a un segundo plano.

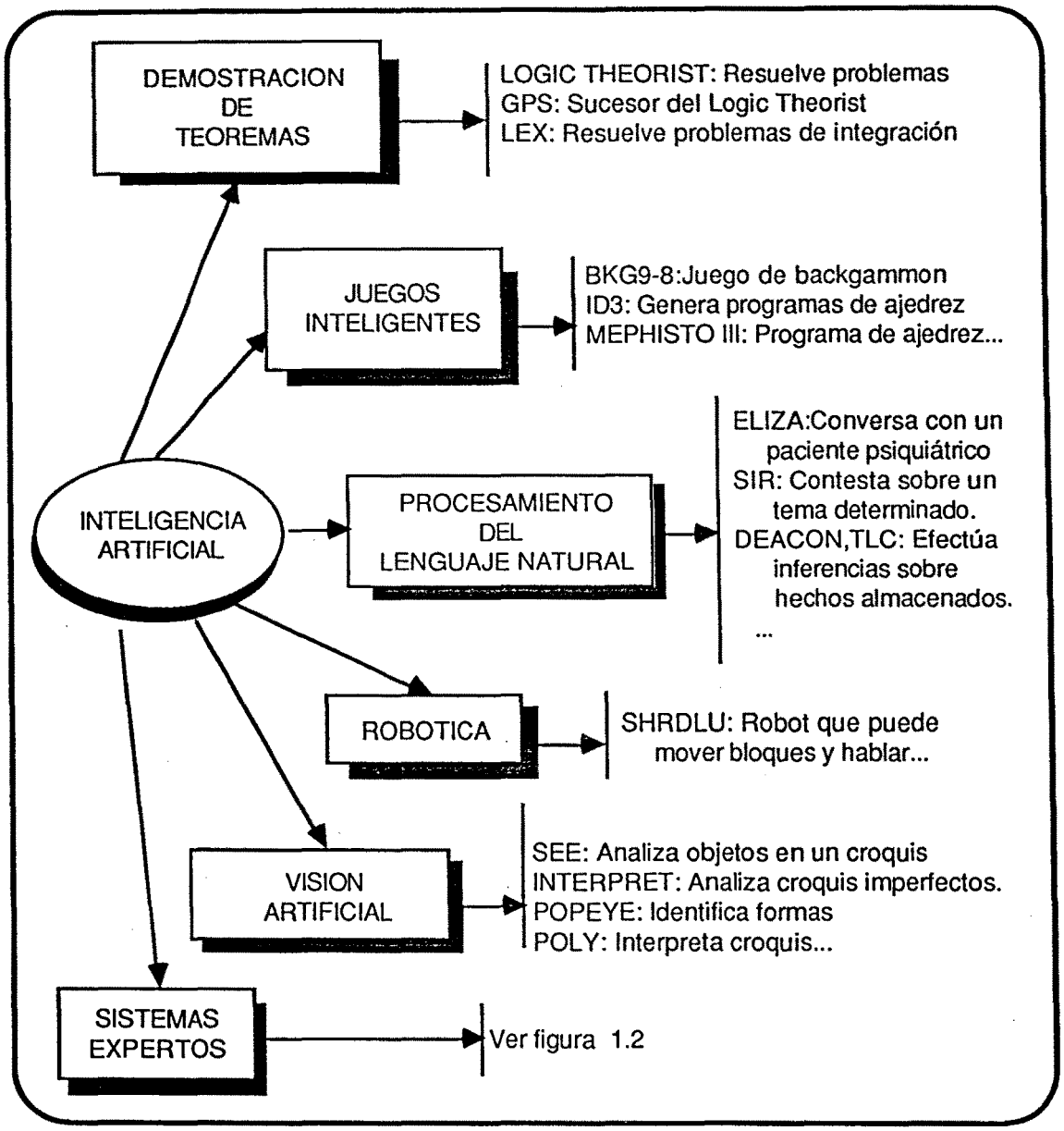


Figura 1.1.- Algunas ramas de la Inteligencia Artificial

En 1956 Herbert Gelenter construyó un programa capaz de resolver teoremas de Geometría plana. También en ese mismo año se desarrolló el programa Logic Theorist, cuyo objetivo era la demostración de teoremas propuestos en la obra "Principia Mathematica" de Russell y Whitehead. Logró demostrar 38 de los 52 teoremas que aparecen en el primer capítulo de esta obra.

El sistema GPS (General Problem Solver), desarrollado en 1957 por Newell, Shaw y Simon, es considerado el sucesor del Logic Theorist. GPS es un programa en LISP que trata de transformar un estado inicial en una situación final valiéndose de operadores que, mediante aplicaciones sucesivas, hacen alcanzar el objetivo deseado. Los operadores no son siempre aplicables sino que deben cumplir ciertas condiciones para que puedan actuar sobre un estado. GPS trabaja con una especie de encadenamiento hacia atrás buscando operadores que "aproximen" el estado inicial al final.

Los programas de partidas de juegos han alcanzado gran desarrollo. Hoy en día no resulta sorprendente que los ordenadores alcancen e incluso superen los niveles de muchos maestros en juegos como el ajedrez o el backgammon. Queda mejorado el programa de Arthur Samuel sobre juego de damas que le permitía, con la experiencia obtenida de partidas anteriores, "aprender" nuevas técnicas que mejoraran su juego. Por ejemplo, el programa Chess Champion Mark V, ha encontrado tres soluciones válidas al programa de ajedrez propuesto por Zagorutko pensándose antes que sólo tenía una, y el campeón de ajedrez Anatoly Karpov ha empatado con el programa de ajedrez Mephisto III.

Un grupo de personas trabajan para conseguir ordenadores que puedan leer, hablar y entender el lenguaje de las personas. Este tipo de programas es denominado comúnmente "proceso de lenguaje natural". La importancia de este logro es evidente, aparte de un manejo del ordenador más fácil por parte de personas no expertas, la aplicación a personas disminuidas físicamente es clara.

La traducción automática fue uno de los primeros campos de investigación dentro de esta área. En un principio se creyó que el ordenador sería capaz de comunicarse en un determinado idioma simplemente si poseía un diccionario bilingüe donde buscar las palabras que necesitara. Sin embargo, se necesitaban labores más complicadas para resolver el problema de la flexibilidad de comunicación entre el hombre y la máquina. Posteriormente se pensó que se solucionarían muchos problemas si el ordenador comprendía los textos que trataba, la principal dificultad radica en una buena modelización de la comunicación.

Son muchos los trabajos de investigación que se están llevando a cabo en este campo, por ejemplo, en Europa se trabaja en la Universidad de Hamburgo en el sistema Hamans y en la Universidad de Cambridge en una unidad de acceso a bases de datos, mientras en Japón desarrollan su proyecto de ordenadores de la quinta generación.

Conseguir de los ordenadores que sean capaces de interpretar información visual es uno de los objetivos de las investigaciones actuales en el desarrollo de la Inteligencia Artificial.

Algunos programas de visión existentes son: SEE (desarrollado en 1965, analiza en un croquis la cantidad de objetos tridimensionales que aparecen), INTERPRET (identifica en una fotografía los objetos y determina sus posiciones), VISIONS (proporciona visión artificial), POPEYE (identifica formas), POLY (interpreta croquis), etc.

Relacionado con el tema de comprensión del lenguaje y el de visión artificial, otro grupo de investigación trabaja en el desarrollo de la Robótica. El objetivo buscado es que los ordenadores interaccionen con el mundo exterior desarrollándoles capacidades sensoriales. Casi todos los robots están programados para trabajar en entornos predeterminados, pero las investigaciones se están centrando fundamentalmente en ampliar su autonomía haciéndoles capaces de responder a estímulos y situaciones adversas con mayor grado de movilidad, proporcionándoles sensores más desarrollados, etc.

Otro grupo de investigación está trabajando en el desarrollo de programas que ofrezcan soluciones a problemas que requerirían un experto humano conocedor del tema para su resolución, esto es los sistemas expertos, concepto introducido por Feigenbaum en 1977.

Existen diferentes etapas en la aparición y diseño de éstos sistemas expertos (ver figura 1.2):

1.- *Etapa de iniciación*, entre 1.965 y 1.970, en la que se desarrollan los primeros sistemas expertos: DENDRAL Y MACSIMA (Buchanan et al. (1969,70), Lindsay et al. (1980)).

Es en este periodo cuando Edward Feigenbaum desarrolló el programa DENDRAL (que siguió a la formulación del algoritmo DENDRAL de 1964), utilizado para resolver problemas de estructura química molecular a partir de análisis de espectrometría de masas. En sus primeros años DENDRAL utilizó el método algorítmico que apenas le diferenciaba de un programa estándar. Sin embargo, las necesidades de los químicos de introducir nueva información en el sistema obligó a los autores a combinar la anterior representación del conocimiento con una basada en reglas. Aunque el sistema DENDRAL ha sido utilizado con éxito por expertos químicos se sabe que esta limitado a estudiar estructuras químicas acíclicas y que sus conocimientos son inferiores a los de un experto humano.

MACSYMA es considerado también uno de los primeros sistemas expertos. En sus orígenes está el programa SAINT desarrollado en el MIT cuya función era resolver problemas elementales de integración simbólica. A partir de 1969 ampliaciones sucesivas dieron lugar a SIN y posteriormente a MACSYMA. MACSYMA es un sistema que trabaja con información simbólica tanto de entrada como de salida de datos y es un programa potente (su base de conocimiento le permite efectuar hasta 600 operaciones matemáticas diferentes).

Dentro del campo médico, en esta época, SHRINK crea un sistema cuya aplicación es la ayuda al diagnóstico en Psiquiatría.

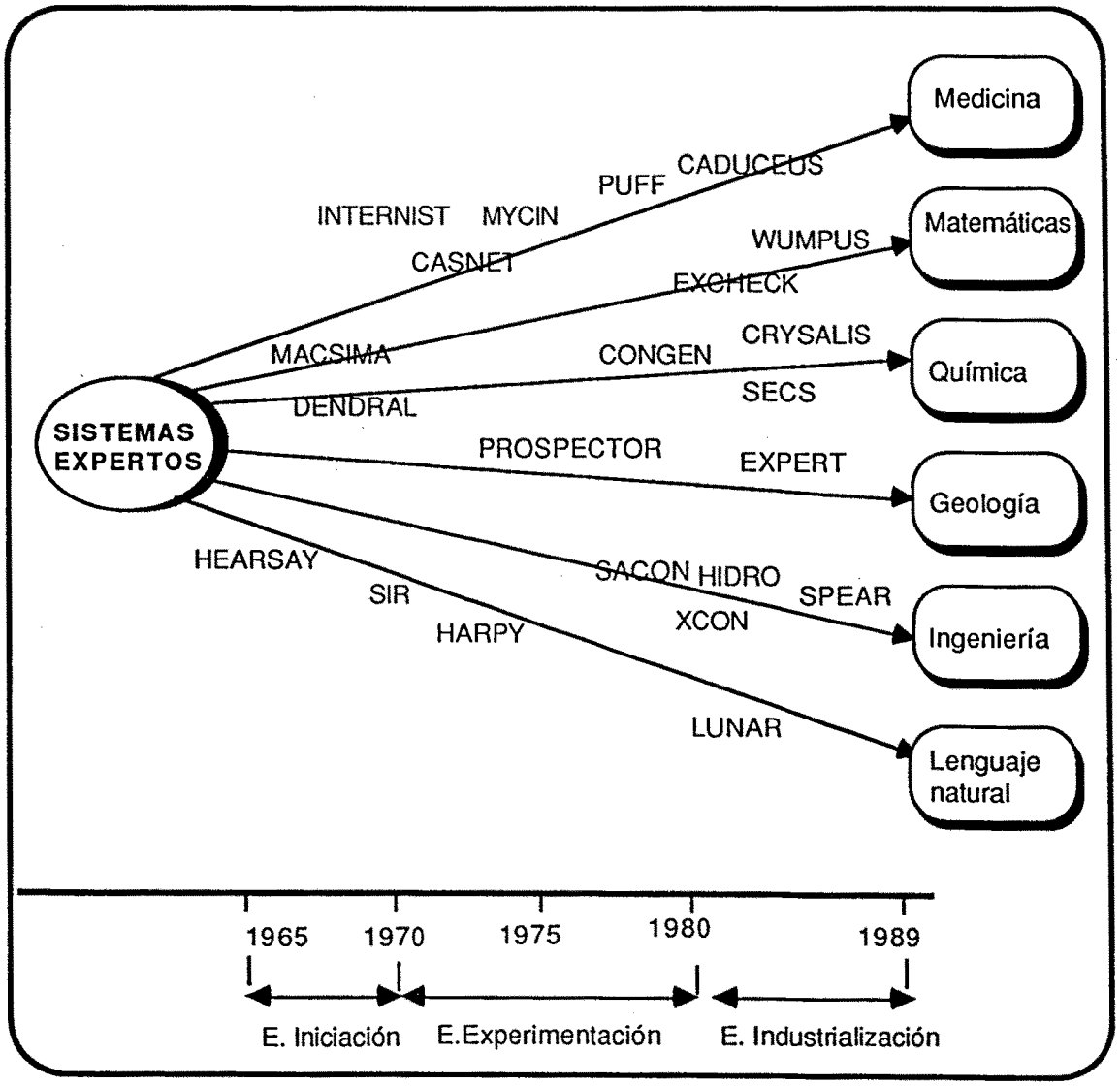


Figura 1.2.- Algunos sistemas expertos

2.-*Etapa de experimentación y desarrollo*, entre 1.970 y 1.980, en la que aparecen los sistemas expertos más conocidos.

En la Universidad de Stanford se desarrolló el programa MYCIN (Shortliffe et al. (1973,75), Shortliffe y Buchanan (1975), Adams (1976), Buchanan y Shortliffe (1984)) para consulta y diagnóstico de meningitis. Es considerado el primero de una nueva generación de programas que razonan, explican su razonamiento y que concluyen de forma análoga a como lo haría un experto humano. Anteriormente a su aparición la Inteligencia Artificial soportó duras críticas considerándola un área que sólo se ocupaba de resolver problemas de juegos y posiblemente su desarrollo no hubiera obtenido el nivel actual si el proyecto MYCIN no hubiese alcanzado tanto éxito.

MYCIN fue el primer sistema que utilizó el razonamiento impreciso, permitiendo asignar a los objetos unos valores acompañados de grados de certidumbre comprendidos entre 0 y 1.

En MYCIN aparecen dos partes claramente diferenciadas: la base del conocimiento y el mecanismo de inferencia, en contraste con DENDRAL y SAINT cuyo motor de inferencia era especialmente diseñado para trabajar con masa espectrométrica y álgebra respectivamente. La separación de estas dos partes permitió considerar el motor de inferencia aisladamente de la base de conocimiento. El resultado de efectuar esta operación recibe el nombre de "concha" o "sistema vacío". Así MYCIN dió lugar a EMYCIN (Essential MYCIN) con el que se construyó PUFF (Alello y Nii (1981), Aikins et al. (1983)), utilizado para el estudio de la función pulmonar, SACON (Bennet et al. (1978)) para ayuda en ingeniería de estructuras, GUIDON (Clancey (1983)) para resolver problemas relacionados con el diagnóstico de distintas patologías, etc.

MYCIN estimuló el desarrollo de PROSPECTOR (Duda et al. (1978b,79), Gaschning (1979)) por la compañía SRI, empleado para evaluar prospecciones geológicas con el fin de hallar yacimientos minerales. Unos años más tarde, en 1982, PROSPECTOR llegó a la conclusión, mediante el análisis de datos correspondientes a una región de Washington, de que en la zona existían yacimientos de molibdeno. A pesar de la incredulidad de los expertos, que afirmaban lo contrario, éstos no tuvieron más opción que rendirse ante la evidencia cuando realizaron la perforación. De PROSPECTOR se derivó la concha KAS (Knowledge Adquisition System) y en él se basó el sistema HYDRO, para la estimación de los parámetros de comportamiento de cuencas hidrográficas a partir de sus caracterísicas geológicas y morfológicas.

De esta época es también el HERSAY II (Reddy et al. (1973), Lesser et al. (1974,75), Erman et al. (1980)), desarrollado por la Universidad Carnegie Mellon, y que tenía por objeto la identificación de la palabra hablada. Del HERSAY II, deriva

el HASP. De esta época también destaca el MOLGEN, descrito por Stefick (1979), para ayudar a los biólogos en el diseño de experimentos de genética molecular.

Hacia mediados de los años 70 se desarrollaron programas que no tenían las limitaciones ya comentadas del sistema DENDRAL; así surgieron CONGEN (1976) y META-DENDRAL.

El programa CASNET (Causal Associational Network) fue escrito por Kulikowsky y Weiss (1971) en la Universidad de Rutgers para el diagnóstico del glaucoma. Es considerado un sistema potente debido a que posee una gran base de conocimiento. De CASNET (Kulikowsky (1982)) derivó la concha EXPERT con la que ya se han construido programas experimentales de consulta.

En la Universidad de Pittsburgh se creó el sistema de medicina interna INTERNIST (Myers et al. (1982), Miller et al. (1982)) que puede diagnosticar hasta 500 enfermedades distintas decidiendo cual de ellas tiene el paciente una vez que se le han suministrado los síntomas, las pruebas realizadas, etc.

La aparición de los sistemas citados anteriormente debe situarse en los Estados Unidos. La situación en Europa era muy distinta pues no se supo ver con suficiente claridad el futuro de la Inteligencia Artificial. En 1973 se produjo el informe de Sir James Lighthill en Gran Bretaña, en el que se expuso que la Inteligencia Artificial era demasiado costosa para los pocos frutos obtenidos hasta el momento, decidiendo no prestar más ayuda económica a los proyectos en los que se estaba trabajando. Las conclusiones de este informe estaban influidas por los fracasos obtenidos en algunos trabajos que intentaron resolver problemas demasiado amplios y generales. Evidentemente este informe frenó el desarrollo de Europa, produciéndose el retraso que actualmente existe respecto a paises como Estados Unidos y Japón que supieron intuir la influencia de la Inteligencia Artificial en años posteriores.

La mayor aportación de Europa a la Inteligencia Artificial y en particular a los sistemas expertos fue el desarrollo del lenguaje PROLOG en los primeros años de la década de los 70. Como ya se dijo anteriormente la demostración automática de teoremas fue uno de los primeros campos en los que se trabajó, demostrando Robinson su algoritmo en 1965. Este algoritmo sirvió de base para que en el año 1972 A. Colmerauer y P. Roussel, de la Universidad de Marsella, implementaran esta versión simplificada del cálculo de predicados en un sistema deductivo de preguntas y respuestas. Originalmente se llamó Q-SYSTEM para posteriormente llamarlo PROLOG (PROgraming language for LOGic). Más tarde Robert Kowalski en la Universidad de Edimburgo hizo notar que el Q-SYSTEM podría considerarse como un lenguaje para programar en lógica de predicados. El primer compilador PROLOG fue debido a David Warren de la Universidad de Edimburgo.

3.-*Etapa de industrialización,* a partir de 1980, en la que numerosas empresas de alta tecnología investigan en Inteligencia Artificial, desarrollando y comercializando sistemas expertos.

Como los sistemas tenían que resolver cada vez problemas más complejos, las estrategias de encadenamiento hacia adelante y hacia atrás debían ampliarse y mejorarse. Surgen los programas STRIPS diseñado por SRI para utilizar en robots móviles y una versión suya mejorada ABSTRIPS donde se trabaja ya con varios niveles de abstracción. Estos trabajos sirvieron de base para los sistemas MOLGEN, utilizado para síntesis y análisis del DNA, y R.1 (Mc Dermott (1980)) (actualmente llamado XCON) desarrollado por Digital Equipment Corporation conjuntamente con la Universidad Carnegie Mellon cuya función es configurar sistemas con PDP y VAX.

En esta etapa hay que señalar el desarrollo de la serie de sistemas de inferencia OPS (Forgy y Mc Dermott (1975)) y del DRILLING ADVISOR (Elf Aquitaine), utilizado para diagnosticar problemas en la perforación de pozos petrolíferos e interpretación de sondeos.

En octubre de 1981 se pusieron las bases para el desarrollo de los ordenadores de la quinta generación en una conferencia internacional en Tokio. Este proyecto japonés adoptó el PROLOG como el lenguaje fundamental que debía soportar el hardware que se debía construir.

Actualmente, la investigación no sólo se reduce a instituciones y universidades como Stanford, MIT y Carnegie Mellon, en Estados Unidos, y las universidades de Edinburgo y Sussex en el Reino Unido, sino que numerosas empresas poseen laboratorios de investigación: IBM, Tektronix, Fujitsu Digital Equipment Corporation, Hewlett Packard, etc., donde se diseñan y mejoran sistemas de alta tecnología.

1.2.-DEFINICION DE SISTEMA EXPERTO

Definir un sistema experto no resulta fácil, pues es un concepto en continua evolución, al que se le van añadiendo nuevas posibilidades y prestaciones como consecuencia de las investigaciones llevadas en este campo. Hace ya bastantes años, Edward Feigenbaum, de la Universidad de Stanford definió, en el Congreso Mundial de Inteligencia Artificial (Davis y Buchanan (1977)), un sistema experto como:

> *"un programa de computador inteligente que usa conocimiento y procedimientos de inferencia para resolver problemas que son lo*

suficientemente difíciles como para requerir la intervención de un experto humano para su resolución".

Hoy, con los avances conseguidos, resultaría más correcto definir un sistema experto como:

"un sistema informático que simula el proceso de aprendizaje, de memorización, de razonamiento, de comunicación y de acción de un experto humano en una determinada rama de la ciencia, suministrando, de esta forma, un consultor que puede sustituirle con unas ciertas garantías de éxito".

Estas características le permiten almacenar datos y conocimiento, sacar conclusiones lógicas, tomar decisiones, aprender de la experiencia y los datos existentes modificando su conocimiento ampliandolo o corrigiendo errores si los hubiera, comunicarse con otros expertos humanos o sistemas expertos, explicar el porqué de las decisiones tomadas y realizar acciones como consecuencia de todo lo anterior. En definitiva, se trata pues de dotarlos de procedimientos y técnicas heurísticas similares a las utilizadas por los expertos humanos.

Los sistemas expertos suponen un abandono en las técnicas tradicionales de programación a pesar de que muchos sistemas expertos actuales se reducen a un programa de computador convencional. A modo de ejemplo se utilizará el de las bases de datos que es, en cierto sentido, muy similar. No cabe duda de que ante un problema de bases de datos puede hacerse una programación que, partiendo de cero, lo resuelva. Sin embargo, la ventaja de las bases de datos es que suministran, de entrada, soluciones a muchos de los problemas que van a presentarse, como son: el acceso a ficheros de todo tipo, el control de restricciones, la elaboración de informes, etc., sin necesidad de que se descubran mediterráneos que hace mucho tiempo fueron descubiertos y de que se corran riesgos de errores innecesarios. De la misma forma, los sistemas expertos juegan ante un problema de Inteligencia Artificial el mismo papel que las bases de datos ante un problema de tratamiento de la información. En efecto, los mecanismos de almacenamiento del conocimiento, del razonamiento, del aprendizaje, de la explicación o de la comunicación, mediante interfases de usuario de calidad, son casi siempre los mismos y no es necesario, por tanto, partir de cero en cada aplicación práctica. Por ello, los sistemas expertos, al suministrar todos estos problemas resueltos, facilitan notablemente el desarrollo de aplicaciones, si bien, al igual que las bases de datos, no son insustituibles.

Los objetivos que se pretenden alcanzar con la construcción de un sistemas expertos son principalmente:

(a) mejorar la calidad del conocimiento de los expertos humanos. El contacto del experto humano con el ingeniero del conocimiento exige un esfuerzo al primero para ordenar, clarificar y dar rigor a las distintas reglas que de una forma o otra, explícita o implícitamente todos los expertos utilizan.
(b) conseguir la supervivencia del conocimiento y que no muera con la muerte física del experto humano, ya que una vez adquirido el conocimiento del experto humano éste se hace accesible y permanece para siempre.
(c) multiplicar el número de expertos y, por tanto, hacer más accesible el conocimiento existente, pues los sistemas expertos son fácilmente reproducibles y permiten analizar problemas en número ilimitado.
(d) disminuir el coste del conocimiento que se puede conseguir al resolverse infinidad de casos con una única inversión de desarrollo del sistema.

Según Addis (1982) la principal cualidad de los sistemas expertos es su capacidad para la optimización en la búsqueda de información, residiendo en ello su analogía con el experto humano. Sin embargo, los papeles que pueden realizar son muy variados y corresponden fundamentalmente a los que juegan normalmente los expertos humanos. Entre ellos merecen destacarse:

(a) como suministradores de información
(b) resolviendo problemas
(c) explicando

En muchos casos los expertos humanos no resuelven problemas directamente sino que se limitan a informar y orientar al cliente sobre la forma de abordarlos, las técnicas más utilizadas para su resolución, la experiencia personal que ellos tienen sobre casos análogos al planteado, etc. En otros casos, son ellos los que resuelven completamente el problema desde su inicio, bien sólos o acompañados por el cliente. Finalmente, a veces se limitan a explicar las causas de determinados resultados, comportamientos o problemas surgidos que son objeto de consulta al especialista.

La consulta a los sistemas expertos se hace por las mismas razones que a los expertos humanos, si bien hay algunas más, añadidas, que son exclusivas de los sistemas expertos. Entre las virtudes más importantes de los expertos humanos destacan (Hart (1986)):

(a) efectividad
(b) eficiencia
(c) sentido de sus limitaciones

La efectividad se refiere a que un problema dado a un experto humano tiene una alta probabilidad de ser resuelto con éxito. La eficiencia se refiere a que los problemas pueden ser resueltos con rapidez (la rapidez que da la experiencia y el conocimiento de los problemas en profundidad). Por último, el sentido de sus limitaciones hace referencia a que el experto humano tiene un conocimiento de sus posibilidades y sabe cuando se le presenta un problema que no es capaz de resolver, comunicándolo al cliente para que éste recurra a otros o sea consciente de los riesgos en que va a incurrir. Por ello, es importante que un sistema experto también controle sus limitaciones.

1.3.-COMPONENTES DE UN SISTEMA EXPERTO

En la figura 1.3 se muestra un diagrama en el que se han incluido las partes o elementos más importantes de un sistema experto y las relaciones entre ambos. En un principio se creyó que no sería posible un sistema experto sin la existencia y colaboración de un experto humano. Sin embargo, hoy se sabe, como luego se verá, que son posibles sistemas expertos sin ni siquiera la existencia del experto humano. De todas formas la mayoría de los sistemas expertos existentes en la actualidad nacen de la colaboración de *expertos humanos* e *ingenieros del conocimiento* (Shortliffe et al (1979), Mulsant y Servan-Schreiber (1984)). El primero aporta el conocimiento en el área de interés y el segundo colabora poniendo ese conocimiento en forma tal que el sistema sea capaz de asimilarlo. En este trabajo conjunto está la clave del éxito de muchos sistemas y suele requerir enormes esfuerzos para conseguir los resultados apetecidos. No puede pensarse que se trata de una etapa fácil, ya que experto humano e ingeniero hablan lenguajes distintos y la comunicación no es en absoluto trivial. Como resultado del proceso, el experto humano resultará aún más experto en su especialidad, ya que se verá forzado a ordenar, estructurar y fundamentar su conocimiento al verse invadido por multitud de preguntas, formuladas por el ingeniero, para dar a ese conocimiento la estructura deseada. Este no se contentará con respuestas a sus preguntas, sino que exigirá razones, explicando los porqués de las mismas, y controlará la coherencia del conocimiento en su conjunto, no permitiendo contradicciones en el mismo.

A este mismo objetivo contribuirá el *subsistema de control de la coherencia* que, de una forma más ordenada y, por supuesto, más fiable, hará ese control y avisará de las imperfecciones que detecte. Éste es un elemento bastante reciente que no existe en muchos sistemas actuales, pero que, forzosamente, tendrá que estar presente en todos los sistemas del futuro. Así, es bastante frecuente encontrar sistemas con reglas y hechos que se contradicen y que, por tanto, impiden su correcto funcionamiento. También es frecuente, en los sistemas que

incluyen mecanismos o sistemas de propagación de incertidumbre, llegar a ciertas conclusiones absurdas como, por ejemplo, situaciones con probabilidades menores que cero o mayores que la unidad. Este sistema deberá asesorar al experto humano y al ingeniero del conocimiento antes de suministrar nueva información y deberá controlar la misma e informar, en su caso, de las incoherencias, no permitiendo en modo alguno que éstas pasen a la base del conocimiento. Supóngase que se pregunta, sucesivamente, a un médico sobre las probabilidades de aparición de los diferentes síntomas, aislados y en bloques, entre los pacientes de una determinada enfermedad. Es muy fácil que tras suministrar varias probabilidades se violen los axiomas de la probabilidad. Por ello, al preguntar al experto, y antes de que nos responda, puede requerirse del sistema que suministre unos límites superior e inferior de las probabilidades demandadas, de forma que no se produzcan esas violaciones.

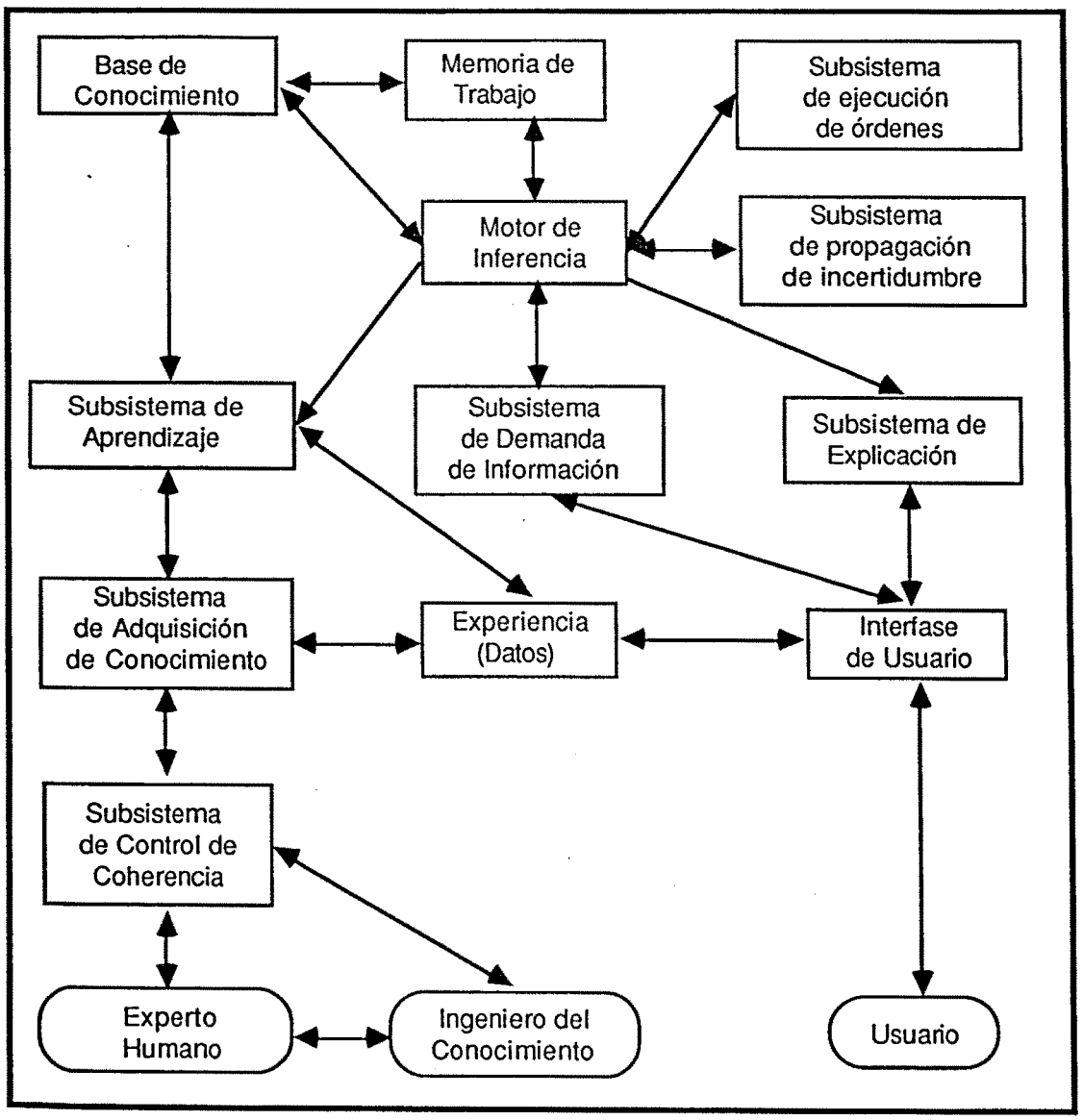


Figura 1.3.- Componentes de un sistema experto

El *subsistema de adquisición del conocimiento* es el encargado de recibir los elementos de conocimiento que proceden del tandem experto-ingeniero, comprobar que son elementos nuevos, es decir, todavía no incluidos en la base de conocimiento y, en su caso, transmitirlos a dicha base en forma por ella inteligible.

El conocimiento almacenado puede ser de tipo abstracto y concreto. El adjetivo abstracto se refiere al de validez general (reglas, espacios probabilísticos, etc.) y el concreto, al de validez particular. Así, en el caso de diagnóstico médico el conocimiento abstracto lo constituyen los síntomas de todas las diferentes enfermedades, sus nombres, los tratamientos, etc. y el conocimiento concreto está constituido por los síntomas particulares de cada paciente en estudio. Mientras que el abstracto es permanente y forma parte esencial del sistema, el concreto es efímero, es decir, se destruye y no forma parte del sistema propiamente dicho.

La *base de conocimiento* es el elemento que almacena el conocimiento abstracto, ya se verá más adelante en qué posibles formas, y lo pone a disposición del motor de inferencia para su posterior tratamiento. La elección de la forma de almacenamiento es fundamental de cara al rendimiento del sistema. Por ello, su elección y diseño requiere un estudio cuidadoso.

La *memoria de trabajo* es la que almacena el conocimiento concreto y todos los procedimientos de los diferentes sistemas y subsistemas. Su carácter es de tipo transitorio, es decir, cambiante.

El *motor de inferencia* es el corazón de todo sistema experto. Su función principal consiste en aplicar el conocimiento abstracto al conocimiento concreto para sacar conclusiones. Así, el diagnóstico de un paciente dado consiste en analizar sus síntomas (conocimiento concreto) y detectar la enfermedad que posee mediante un análisis de las diferentes sintomatologías (conocimiento abstracto). Durante este proceso puede ocurrir que el conocimiento concreto inicial sea muy limitado y que no puedan sacarse conclusiones fiables, por lo que el sistema experto debe disponer de un *subsistema de demanda de información* que complete ese conocimiento para proceder de nuevo a su reelaboración y repetir el ciclo hasta llegar a conclusiones válidas.

En muchos casos esta demanda de nueva información se hace preguntando al usuario, por lo que debe existir una *interfase de usuario* que la haga posible. Es muy recomendable cuidar con mimo esta interfase pues el reconocimiento de la bondad del sistema experto corre a cargo de los *usuarios*, que son los que, a la postre, determinan si el sistema les resulta útil o no. En esta interfase deben jugar papeles preponderantes los nuevos elementos de hardware (ratones, pantallas gráficas, pantallas en color, etc.) y software (menús, cuadros de diálogo, gráficos, etc.) introducidos hace ya varios años. Hay que darse cuenta de

que muchas veces los usuarios finales se dejan llevar más por la impresión que causan estos elementos que por la verdadera calidad de los sistemas, que es lo verdaderamente importante.

Una vez sacadas las conclusiones pertinentes, el sistema experto puede realizar ciertas acciones mediante el *subsistema de ejecución de órdenes*. Por ejemplo, un sistema experto para controlar el funcionamiento de una red de ferrocarriles puede dar lugar a la apertura o cierre de ciertas señales, al cambio de agujas, al arranque o la parada de los diferentes trenes, la comunicación de cierta información a los viajeros, etc., un sistema para control de una central nuclear puede producir la apertura de ciertas válvulas, la activación de ciertas alarmas, el movimiento de barras, etc.

Una vez producidas las conclusiones, el usuario puede interesarse también por las razones que han conducido a las mismas. Por ello, debe existir un *subsistema de explicación*, que, tras el análisis de los procesos seguidos por el motor de inferencia, comunique al usuario, en forma ordenada e inteligible, los hechos determinantes que diferencian las distintas alternativas de decisión. Es muy recomendable disponer de un mecanismo de elección que permita elegir el nivel del contenido de la explicación, ya que puede ocurrir que en unos casos con una explicación sucinta el usuario quede satisfecho, mientras que en otros se requiera una información exhaustiva. Para un ejemplo sencillo de subsistema de explicación ver Naylor (1986).

Es bastante frecuente que algunos de los hechos (conocimiento concreto) no sean conocidos con absoluta certeza. Piénsese, por ejemplo, en un enfermo que no está seguro de sus afirmaciones. También puede ocurrir que el conocimiento abstracto no esté definido en forma determinista o con total certeza, sino que haya duda, componentes aleatorias o información de tipo difusa. En este caso se necesita una base de conocimiento especial que permita almacenar este tipo de información y es inevitable la existencia de un mecanismo capaz de realizar la propagación de la incertidumbre al encadenar los distintos elementos de conocimiento. Este elemento se conoce con el nombre de *subsistema de propagación de la incertidumbre*. Su complejidad es tal que es probablemente el elemento más débil de los sistemas expertos que funcionan en la actualidad. Esto hay que entenderlo en el sentido de que la propagación de la incertidumbre es normalmente incorrecta y que, por tanto, las conclusiones que resultan son o pueden ser poco fiables.

Uno de los elementos que más recientemente se han incorporado a los sistemas expertos y que los hacen más sugestivos, es el *subsistema de aprendizaje*. No debe sorprendernos el oir que los sistemas expertos sean capaces de aprender. Hoy existen demostraciones palpables de que esto es ya

una realidad. Se diferenciará entre aprendizaje estructural y aprendizaje paramétrico. El *aprendizaje estructural* se refiere a los aspectos relacionados con la estructura del conocimiento (reglas, espacios probabilísticos, etc.). Así, el descubrir que un síntoma es relevante para una enfermedad o la incorporación de una nueva regla a la base de conocimiento constituye aprendizaje de tipo estructural. El *aprendizaje paramétrico* se refiere a los parámetros de la base de conocimiento. Así, la determinación de las probabilidades o las frecuencias de aparición de las diferentes enfermedades o de los síntomas de una enfermedad constituye aprendizaje paramétrico. El primer tipo de aprendizaje es muy superior al segundo, aunque ciertos artificios pueden permitir que el segundo suplante al primero, al menos parcialmente. Esto ocurre, por ejemplo, cuando se incluyen, ante la duda, como síntomas de una enfermedad algunos con una probabilidad nula y luego se permite su modificación. De esta forma los síntomas son excluidos inicialmente (al darles probabilidad nula) y luego, si la evidencia posterior pone de manifiesto que esa hipótesis no es correcta, una simple modificación de los parámetros (probabilidades) permite incluirlos. Este método suele ser muy útil en algunos casos.

Finalmente, se incluye *la base de datos y de experiencias* que se encarga de almacenar los datos o la experiencia existente sobre el tema que se trata. Al principio se decía que era posible la existencia de un sistema experto sin la necesidad de la existencia de un experto humano. Al hablar así se pretendía decir que se puede partir de datos o de experiencias recogidas por no expertos, en el sentido usual de la palabra, y desarrollar un sistema experto con las mismas garantías de calidad que pueden esperarse de la contribución de un experto humano durante el proceso de generación del sistema.

En este proceso juega un papel preponderante el *subsistema de aprendizaje*, que debe encargarse de realizar los aprendizajes estructural y paramétrico, bien conjuntamente o bien en fases escalonadas.

Los elementos anteriores sirven para definir claramente las funciones de un sistema experto, que es la mejor forma de entender lo que realmente hay detrás de estos sistemas y conocer sus posibilidades. Entre otras, a modo de resumen, podrían destacarse las siguientes:

(a) adquirir conocimiento
(b) almacenar conocimiento
(c) razonar e inferir
(d) demandar nueva información
(e) aprender
(f) propagar incertidumbre
(g) asistir al experto a dar información coherente

(h) explicar las conclusiones
(i) realizar ciertas acciones como consecuencia del razonamiento
(j) controlar la coherencia del conocimiento del sistema

A estas funciones podrán añadirse otras más, que resultarán de los avances que se produzcan en el futuro en el campo de los sistemas expertos. Recuérdese el carácter cambiante de la Informática y en particular de los sistemas expertos que evolucionan vertiginosamente, siendo impredecibles los límites a los que puede llegarse.

1.4.-ESTRATEGIAS PARA REPRESENTAR EL CONOCIMIENTO

1.4.1.-Introducción

Para abordar los problemas que resuelven los sistemas expertos se necesitan almacenar y manipular gran cantidad de conocimiento. El primer proceso necesario para ello es representar el conocimiento que encierra una base de conocimiento de manera que pueda ser manipulado por el ordenador. Esta representación puede hacerse de muy diversas maneras (Hendrix (1976,77), Harmon (1985), Alty (1986)). En este apartado se darán las cinco que se consideran más importantes y de más frecuente utilización, analizándolas brevemente. Estas son:

(a) redes semánticas
(b) ternas objeto-atributo-valor
(c) expresiones lógicas
(d) reglas
(e) marcos (frames)

La elección del sistema de representación depende del tipo de problema que pretenda resolver el sistema experto, por lo que no puede darse un sistema que sea el óptimo en todos los casos. Sin embargo, esta elección tiene mucha importancia, pues la eficiencia del sistema depende del método elegido.

1.4.2.-Redes semánticas

Es el sistema de representación más general de todos y también el más antiguo (Woods (1975), Brachman (1976), Duda et al. (1978a), Hendrix (1979)). Como su nombre indica, una red semántica es una colección de objetos, también llamados nodos, unidos mediante arcos o enlaces formando una red.

Normalmente los elementos de ambos grupos llevan asociada una etiqueta o palabra, que en el caso de los nodos es un nombre (común o propio) y en el de los enlaces es un verbo. Los nodos representan objetos y los enlaces, relaciones entre esos objetos.

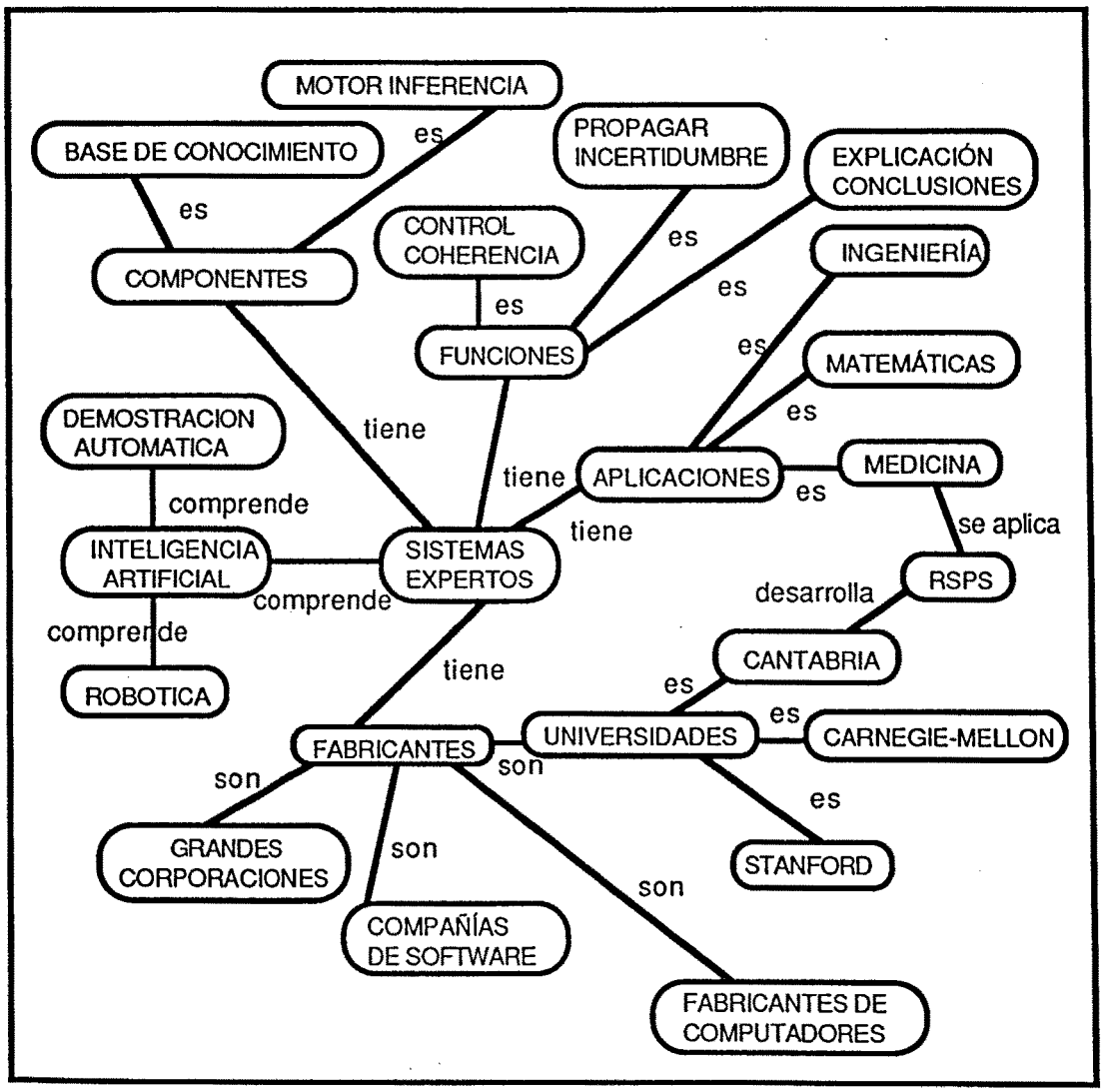


Figura 1.4.- Ejemplo de red semántica

La figura 1.4 muestra un ejemplo de red semántica que trata de representar el conocimiento sobre algunos aspectos de la Inteligencia Artificial. Los nodos se han representado mediante rectángulos de vértices redondeados y los enlaces mediante segmentos. Nótese como los nodos llevan etiquetas de nombres y los enlaces etiquetas de verbos.

Partiendo de esta red semántica, resulta fácil responder a multitud de preguntas como, por ejemplo, ¿es la base del conocimiento una componente de los sistemas expertos? o ¿es RSPS un sistema experto que se aplica en

Medicina?. Nótese la gran cantidad de información almacenada en una red semántica.

1.4.3.-Ternas objeto-atributo-valor

Este tipo de estrategia de almacenamiento de la información (Harmon (1985), Alty y Coombs (1986)) es un caso especial de red semántica. En este caso existen tres nodos: objetos, atributos (propiedades asociadas con los objetos) y valores (estado del atributo), y se suprimen los enlaces, que resultan triviales.

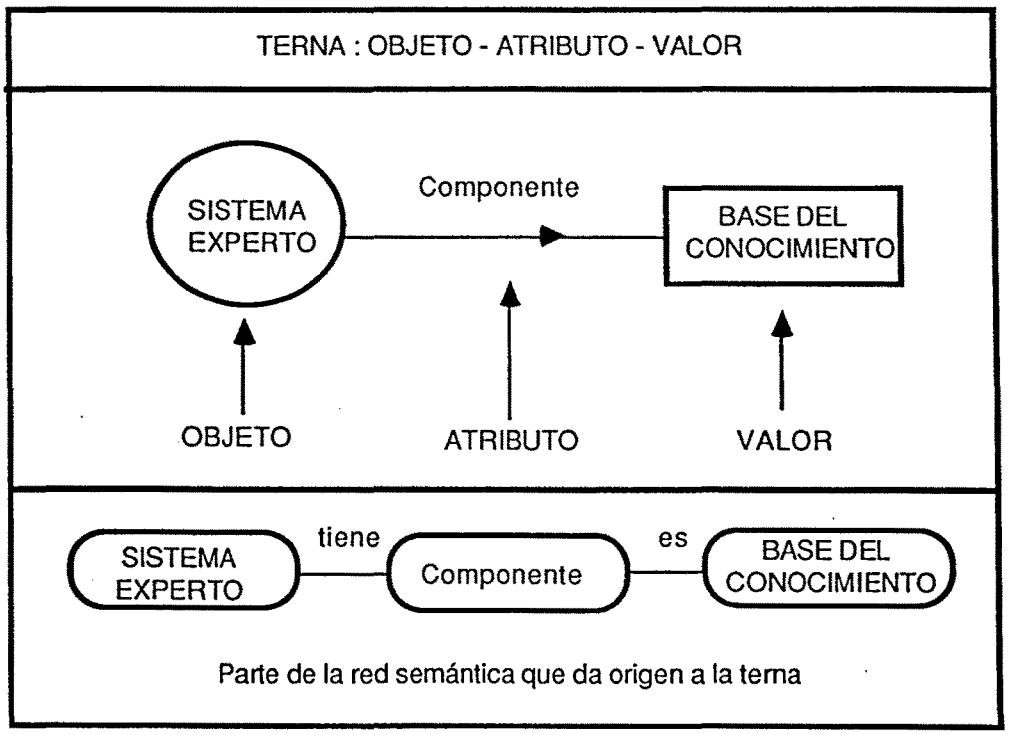


Figura 1.5.- Ejemplo de terna Objeto-Atributo-Valor

Un ejemplo se muestra en la figura 1.5, en la que en la parte superior se representa el objeto con un círculo, el atributo con una flecha y el valor con un rectángulo. En la parte inferior de la figura se muestra la parte de la red semántica de la que procede la terna. Nótese que han desaparecido los enlaces "tiene" entre "sistema experto" y "componente" y "es" entre "componente" y "base del conocimiento".

A veces se añade a las ternas un factor de certeza que sirve para medir la confianza que se tiene de que la terna sea correcta. En algunos casos se omite el atributo y se trabaja sólo con parejas objeto-valor. Este esquema es el utilizado por MYCIN.

1.4.4.-Expresiones lógicas

Otra forma de almacenar el conocimiento es la basada en las expresiones lógicas (Poper (1959), Carnap (1962), Hempel (1965), Davis (1977), etc.). La Lógica es una disciplina para analizar la validez de los razonamientos generados por la mente humana por dar métodos para determinar si unas conclusiones pueden deducirse correctamente a partir de unos hechos. Dentro de la Lógica, la noción de argumento verdadero juega un papel importante. Un argumento es verdadero si y solo si al ser todos sus supuestos verdaderos, entonces sus conclusiones también lo son. Para realizar una prueba dentro de la Lógica se comparan los hechos o supuestos con una serie de patrones abstractos de argumentos y se ve si alguno se ajusta con los nuestros. Tales patrones se denominan *modelos* y están constituidos por secuencias abstractas de hechos y por reglas que previamente se han demostrado que son válidas de una forma matemática ó formal. La lógica proporciona, pues, un formalismo bien definido para representar el conocimiento. Hay varias notaciones o sistemas de lógicas, pero las más comunes son: *la lógica proposicional y el cálculo de predicados.*

A	B	$\bar{A}$	$A \wedge B$	$A \vee B$	$A \rightarrow B$	$A \leftrightarrow B$
f	f	v	f	f	v	v
f	v	v	f	v	v	f
v	f	f	f	v	f	f
v	v	f	v	v	v	v

Tabla 1.2. Propagación de la veracidad

La lógica proposicional es un sistema de lógica común en el que las proposiciones son expresiones que pueden ser *verdaderas o falsas.* La llamada lógica proposicional se ocupa de las expresiones compuestas. Las expresiones que están unidas por las conectivas "y", "o", "implica" y "equivalente" (que representan la conjunción, disyunción, condicional y bicondicional, respectivamente) se denominan *expresiones compuestas.* Las reglas para propagar la veracidad de las expresiones dependiendo de las conectivas se resume en la Tabla 1.2.

En lógica de predicados la unidad fundamental es el objeto. Un predicado es la formalización de una propiedad o una relación. Es frecuente denotar a los predicados por el nombre de la propiedad o de la relación seguido a continuación de sus argumentos. En la figura 1.6 se muestran algunos ejemplos de predicados.

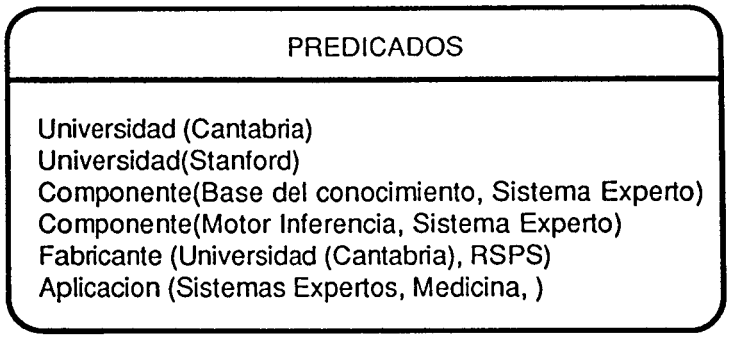


Figura 1.6.- Ejemplo de predicados

1.4.5.-Reglas

Otra forma de almacenar el conocimiento es mediante reglas (Davis et al. (1977), Davis (1977), Hii y Feigenbaum (1978), Buchanan y Mitchel (1978), Clancey (1979), Davis (1980), Buchanan y Duda (1983), etc.). Las reglas se usan para representar relaciones y constan de dos partes: "la premisa" y "la conclusión". A su vez, la premisa consta del condicional "Si" y de una expresión lógica (constituida por una o varias ternas objeto-atributo-valor unidas por los operadores lógicos "y", "o" y "no"). La conclusión consta del adverbio "entonces" y de una expresión lógica.

Si tras evaluar la expresión lógica de la premisa de una regla ésta resulta cierta, entonces se hace que la expresión lógica de la conclusión sea cierta.

La definición anterior es la más general de regla, sin embargo, algunos sistemas imponen ciertas limitaciones a la regla general. Así, en algunos las expresiones lógicas de la premisa sólo pueden utilizar el operador lógico "y", y en otros, la conclusión sólo admite una expresión y, por tanto, sin operadores lógicos. En algunos casos, las ternas objeto-atributo-valor de las reglas son sustituidas por parejas objeto-valor.

A veces, las reglas no son siempre ciertas o válidas y se les asocia, al igual que a las ternas, un factor de certeza.

En la figura 1.7 se muestra un ejemplo de reglas para determinar si la estructura algebraica de un conjunto A con dos operaciones internas, que llamaremos suma y producto, es un anillo.

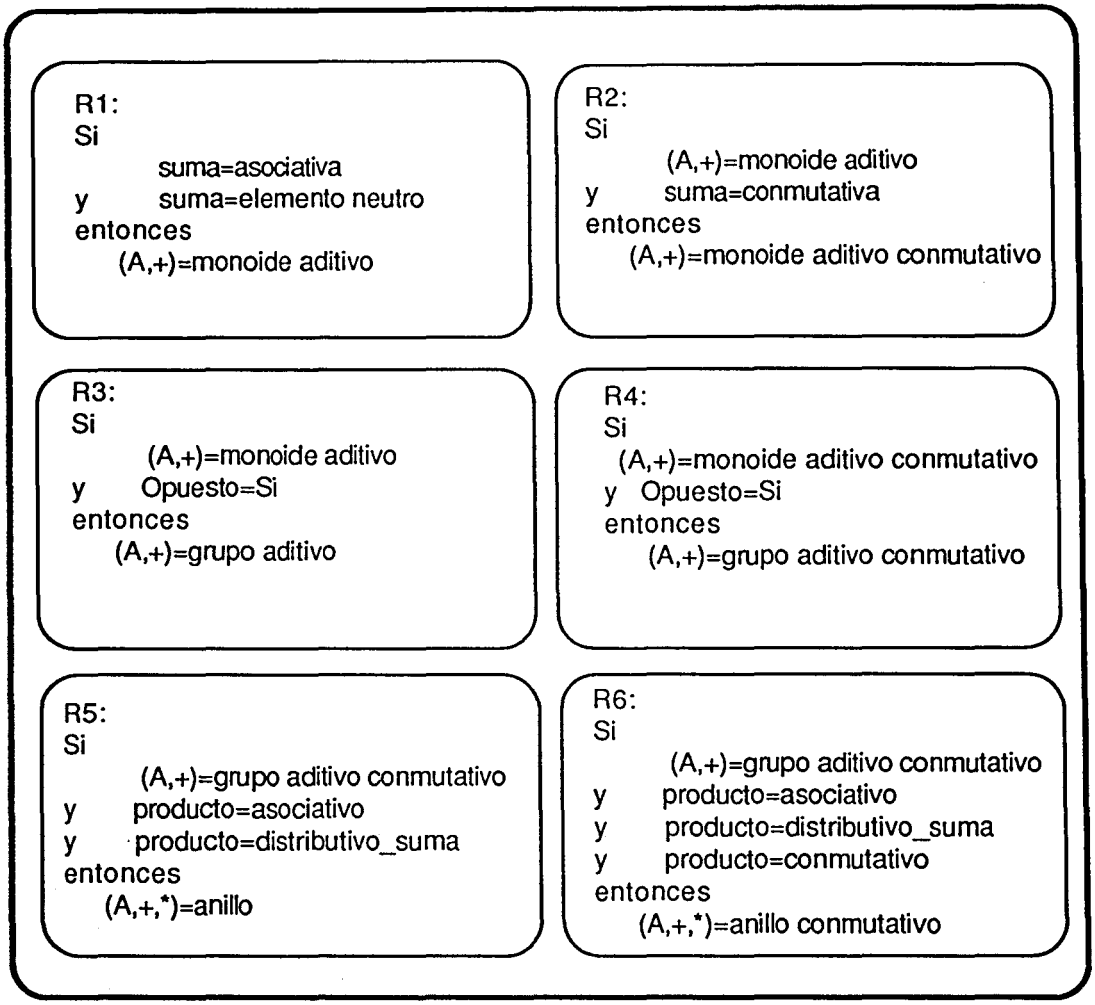


Figura 1.7.- Ejemplo de reglas

1.4.5.-Marcos

Los marcos (Minski (1975), Winograd (1975), Charniak (1977), Rosemberg (1977), Stefick (1979), etc.) suministran otra de las formas de representar objetos y relaciones y también pueden ser considerados como casos particulares de redes semánticas. Un marco suministra toda la información existente sobre un objeto. Esta información puede ser de tipo declarativa o descriptiva o de tipo indirecta, es decir, puede dar directamente las características del objeto o dar procedimientos o reglas para determinarlas. Esta posibilidad indirecta de trabajo es la que ha hecho a esta forma de almacenamiento alcanzar la mayor popularidad.

La figura 1.8 muestra un ejemplo de marco en el que el objeto es la el sistema expero MYCIN. Nótese como a los atributos "desarrollado por", "software",

"aplicación" y "concha" se les asocia valores directos y al atributo "reglas", cómo conseguirlos.

MYCIN	
ATRIBUTOS	VALORES
DESARROLLADO	UNIVERSIDAD DE STANFORD
SOFTWARE	LISP
APLICACIÓN	MEDICINA
CONCHA	EMYCIN
REGLAS	VER LISTA DE REGLAS

Figura 1.8.- Ejemplo de marco

1.5.-TIPOS DE SISTEMAS EXPERTOS

Entre los diferentes tipos de sistemas expertos destacan dos de ellos: los *basados en reglas* y los *basados en la probabilidad*. En lo que sigue se tratará de comentar las bases de los elementos más importantes que intervienen en un sistema experto y las diferencias entre ambos, aunque sólo sea a modo de introducción, pues los detalles concretos de cada uno de ellos serán objeto de los apartados 1.5.1 y 1.5.2.

En la Tabla 1.3 se muestran algunos de estos elementos y en qué estructura matemática (lógica o probabilística) se basan. En los sistemas basados en reglas la base de conocimiento está constituida por el conjunto de reglas y en los sistemas basados en la probabilidad, por el espacio probabilístico, en el que se incluyen los sucesos y sus dependencias. En ambos casos el conocimiento concreto se identifica con los hechos. El motor de inferencia consiste en el encadenamiento de reglas (hacia adelante o hacia atrás) o bien en la determinación o evaluación de las probabilidades condicionales. La explicación se basa en las reglas activas o en los valores relativos de las probabilidades condicionales. Finalmente, la adquisición de conocimiento y el aprendizaje consisten en la incorporación de nuevas reglas a la base, en la variación de los factores de certeza, en el cambio de la estructura probabilística o en la variación de las probabilidades. Detalles de todo esto se darán en los apartados que siguen.

La Tabla 1.4 muestra una comparación de las ventajas e inconvenientes de ambos. En primer lugar se dirá que el motor de inferencia de los métodos

probabilísticos es muy rápido, ya que en general todas las implicaciones están presentes en el modelo y sólo se trata de determinar con qué probabilidad se da una determinada implicación. En Medicina, por ejemplo, se trata de determinar con qué probabilidad puede darse una enfermedad cuando se tienen unos ciertos síntomas (probabilidad condicionada). La comparación de esta probabilidad para diferentes enfermedades permite decidir cuál de ellas es más verosímil. Por el contrario, el motor de inferencia en los métodos basados en reglas necesita encadenar las reglas, lo cual es muy costoso y puede producir problemas graves de memoria si se utilizan, como es habitual, los métodos recursivos. A veces se producen también problemas de tiempo de ejecución, que dan una muy mala imagen al usuario.

ELEMENTOS	MODELO PROBABILISTICO	MODELO BASADO EN REGLAS
BASE DE CONOCIMIENTO	Abstracto :Estructura probabilística (Sucesos dependientes) Concreto : Hechos	Abstracto : Reglas Concreto : Hechos
MOTOR DE INFERENCIA	Evaluación de probabilidades condicionales (Teorema de Bayes)	Encadenamientos hacia atrás y hacia adelante
SUBSISTEMA DE EXPLICACION	Basado en probabilidades condicionales	Basado en reglas activas
ADQUISICION DE CONOCIMIENTO	Espacio probabilístico Parámetros	Reglas Factores de certeza
SUBSISTEMA DE APRENDIZAJE	Cambio en la estructura del espacio probabilístico Cambio en los parámetros	Nuevas reglas Cambio en los factores de certeza

Tabla 1.3.- Comparación de elementos en los dos sistemas

Conclusiones análogas a las del apartado anterior son también aplicables a la propagación de incertidumbre, en cuyo caso hay que añadir además la imprecisión que se logra al utilizar fórmulas arbitrarias tales como la hipótesis de independencia en los métodos probabilísticos, las fórmulas clásicas de propagación de los factores de certeza, los métodos utilizados en la teoría de la evidencia, los métodos posibilísticos, etc.

En cuanto a la explicación puede decirse que es más fácil en los métodos basados en reglas, ya que se sabe perfectamente qué reglas están activas (han sido disparadas), mientras que en los métodos probabilísticos todas las reglas

están activas en un cierto grado y la explicación debe basarse en la medida de actividad, que no es otra cosa que la probabilidad.

	PROBABILISTICO	BASADO EN REGLAS
VENTAJAS	- Motor de inferencia rápido - Aprendizaje paramétrico fácil - Propagación de incertidumbre fácil	- Explicación fácil - Sólo implicaciones deseadas
DEFECTOS	- Elevado número de parámetros - Implicaciones superfluas	- Motor de inferencia lento - Dificultad de propagación de incertidumbre

Tabla 1.4.- Ventajas e inconvenientes de los dos tipos de sistemas

Un inconveniente de los métodos probabilísticos, es el elevado número de parámetros que resultan, lo que lleva asociada la inclusión de reglas superfluas. Por el contrario, el aprendizaje paramétrico de estos modelos es muy sencillo y fácil de implementar.

Aún cuando entre los sistemas expertos existentes predominan los basados en reglas, no es porque sean superiores a los basados en la probabilidad. El principal defecto de los primeros es la propagación de la incertidumbre, que deja mucho que desear. Este problema no es conocido por muchos de los usuarios, lo cual puede ser grave a efectos prácticos, ya que las conclusiones a las que se llegan pueden no ser correctas. Por ello, debe cuidarse especialmente este problema.

La clasificación anterior de sistemas expertos se basa exclusivamente en la forma de almacenar el conocimiento; sin embargo, existen muchas clasificaciones más completas y ricas que esa. A continuación se analizará una de ellas basada en un trabajo de Reichgel y Harmelen (1985). En esta clasificación se atiende a diferentes aspectos de los sistemas. La Tabla 1.5 muestra estos aspectos.

El primero de ellos se refiere a la naturaleza de la tarea a realizar por el sistema experto. En este sentido cabe distinguir entre los cuatro tipos de la Tabla 1.5. En el diagnóstico o clasificación se parte de un espacio de soluciones totalmente conocido y se trata de clasificar o diagnosticar en ese espacio en función de una serie de datos. En este grupo se incluyen todos los sistemas de diagnóstico médico o de averías de máquinas, por ejemplo. En la monitorización, que también parte de un espacio de soluciones conocido, el sistema experto observa iterativamente el comportamiento de un cierto sistema y analiza su comportamiento para detectar posibles fallos en el mismo. Aún cuando tiene

bastante en común con el diagnóstico, la diferencia fundamental está en que ahora lo que ocurre en una iteración puede depender de lo observado en la anterior, cosa que no ocurría en el caso del diagnóstico, para el que valores idénticos de los datos son siempre valorados por el sistema de la misma manera. En el diseño se trata de construir una solución a un problema a partir de ciertas restricciones que deben ser satisfechas. La diferencia fundamental con los dos anteriores está en que aquí el espacio de soluciones es totalmente desconocido y hay que generarlo basándose en las restricciones. Finalmente, en la predicción o simulación se suministran al sistema experto ciertos cambios de un sistema y se le pide hacer una predicción de su comportamiento. En este caso es imposible enumerar inicialmente el espacio de todas las posibles soluciones, que se van generando automáticamente por el sistema experto.

Aspecto	Variantes	Subvariantes
Por las estructuras de la tarea a realizar	Diagnóstico o Clasificación Monitorización Diseño Predicción	
Por el papel del sistema en la interacción con el usuario	Como ayuda o apoyo Dictatorial Crítica	
Por la limitación del tiempo	Tiempo limitado (tiempo real) Conocimiento causal	
Por la naturaleza del conocimiento almacenado en la base de conocimiento	Conocimiento de experiencias Conocimiento causal	
Por la naturaleza temporal del conocimiento	Estática Dinámica →	Predecible Impredecible Adicional Diferente
Por la certeza de la información	Perfecta Imperfecta →	Incompleta Terminología incierta Datos inciertos Conocimiento incierto

Tabla 1.5.- Aspectos para clasificar los sistemas expertos

Con respecto al papel del sistema en la interacción con el usuario hay que distinguir entre los sistemas que funcionan ayudando al usuario, pero

manteniendo éste la decisión última, los que toman las decisiones directamente sin consultar con el usuario, o a lo más informándole de ellas, y los que simplemente evaluan decisiones tomadas por el usuario y las someten a crítica.

En cuanto a la limitación de tiempo disponible para tomar decisiones cabe distinguir entre los sistemas que disponen de tiempo limitado, entre los que se encuentran todos los que funcionan en tiempo real, y los que, en teoría, disponen de tiempo ilimitado. El someter a un sistema a limitaciones de tiempo impone importantes restricciones a los algorítmos de búsqueda de soluciones por lo que éstos deben estar preparados para trabajar bajo esas limitaciones de tiempo.

El conocimiento puede ser de varios tipos, entre los que destacan dos de ellos: el basado en experiencias y el de tipo causa-efecto. En el primero, el experto humano o la base de datos es simplemente portadora de hechos acaecidos, en los que está contenido el conocimiento, pero se desconocen claramente las causas de los efectos observados. Por ello, en muchos casos la información es difusa o aleatoria y se requieren mecanismos de tratamiento y propagación capaces de trabajar con este tipo especial de información. Por el contrario, si el conocimiento es de tipo causa-efecto (nótese que éste es mucho más profundo que el anterior) es posible utilizar la lógica clásica y el análisis es mucho más sencillo.

Puede ocurrir que la base del conocimiento no se altere durante una sesión de utilización del sistema experto, en cuyo caso se habla de conocimiento estático. Si, por el contrario, éste cambia durante dicho proceso se dice que es dinámico. Entre los diferentes tipos de cambios o alteraciones que pueden producirse cabe destacar los predecibles, los impredecibles, el aumento progresivo de información y la alteración de la misma.

Por último, es importante señalar las diferentes variantes según el grado de certeza de la información. No siempre ésta es perfecta, ya que se pueden dar casos de información incompleta (no se conoce toda la información requerida por el sistema experto para tomar decisiones), de terminología imprecisa (diferentes términos se utilizan con el mismo sentido o el mismo término se utiliza con dos sentidos diferentes), de datos inciertos (no confirmados o conocidos con precisión) y de conocimiento incierto (reglas que no siempre son válidas).

De todo lo anterior se deduce, como consecuencia, que puede darse una clasificación muy completa considerando todos los aspectos mencionados y sus posibles variantes.

1.5.1.-Sistemas expertos basados en reglas

En este apartado se analizará uno de los tipos más importantes de sistemas expertos: los "basados en reglas", dedicando especial atención a la base de conocimiento y al motor de inferencia.

1.5.1.1.-La base del conocimiento

El centro de este tipo de sistemas lo constituye el conjunto de reglas de la base de conocimiento, que forman lo que se ha llamado conocimiento abstracto (ver ejemplo en figura 1.9). Cuando las premisas de algunas reglas coinciden, en su totalidad o en parte, con las conclusiones de otras, se produce lo que se llama un encadenamiento de reglas.

La figura 1.10 muestra las reglas del ejemplo anterior encadenadas. Nótese que en dos reglas encadenadas (por ejemplo, la 1 y la 4) la conclusión de una coincide (en todo o en parte) con la premisa de la otra .

1.5.1.2.-Motor de inferencia

Las reglas se aplican sobre la base de hechos para obtener nuevos hechos. Tanto los hechos iniciales como los que resultan de la aplicación de las reglas forman el conocimiento concreto, que reside en la memoria de trabajo. Se llamarán conclusiones simples a aquellas que resultan de la aplicación de sólo una regla y conclusiones compuestas a las que resultan del encadenamiento de varias reglas.

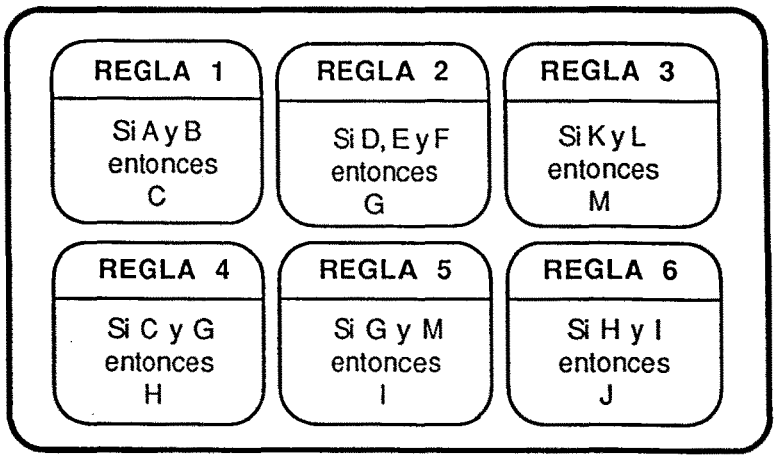


Figura 1.9.- Ejemplo de reglas

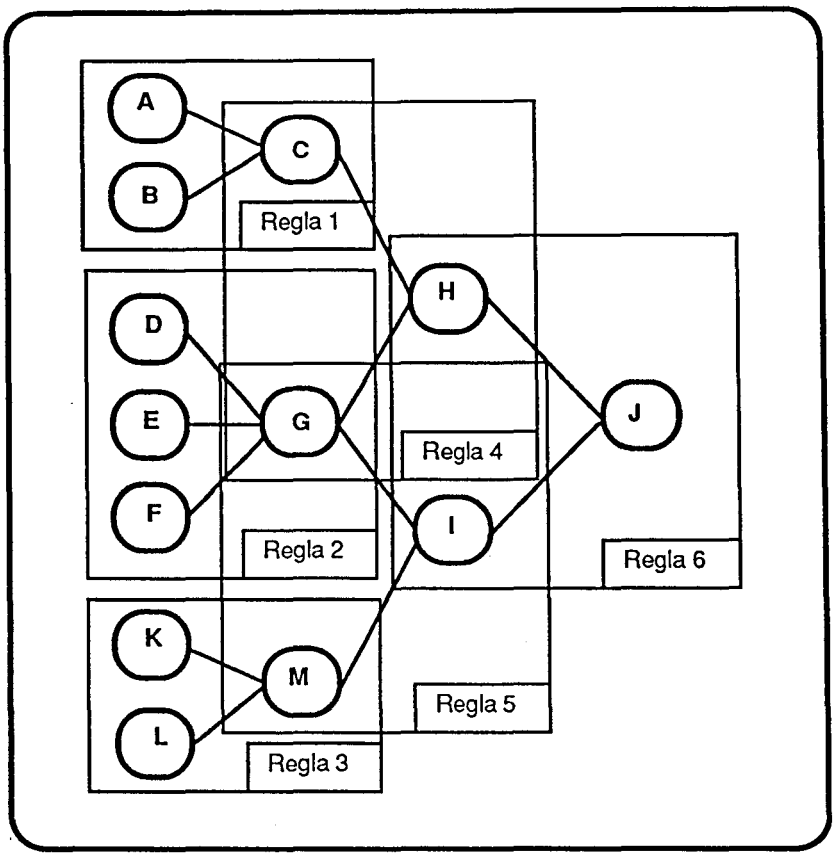


Figura 1.10.- Reglas encadenadas

Para la obtención de conclusiones en los sistemas expertos se utilizan diferentes tipos de estrategias de inferencia y control del razonamiento (Friedman (1971), Waterman (1978), Davis (1979), Winograd (1980), Garvey et al. (1981), etc.).

Para las conclusiones simples existen dos tipos de estrategias:

(a) modus ponens
(b) modus tollens

El "modus ponens", que es la estrategia más común, afirma que si se tiene la regla:

"Si A es cierto entonces B es cierto"

y se sabe que "A es cierto", entonces puede afirmarse que "B es cierto" (ver figura 1.11). Esta estrategia, que nos resulta trivial por su familiaridad, es la base de una gran parte de los sistemas expertos basados en reglas.

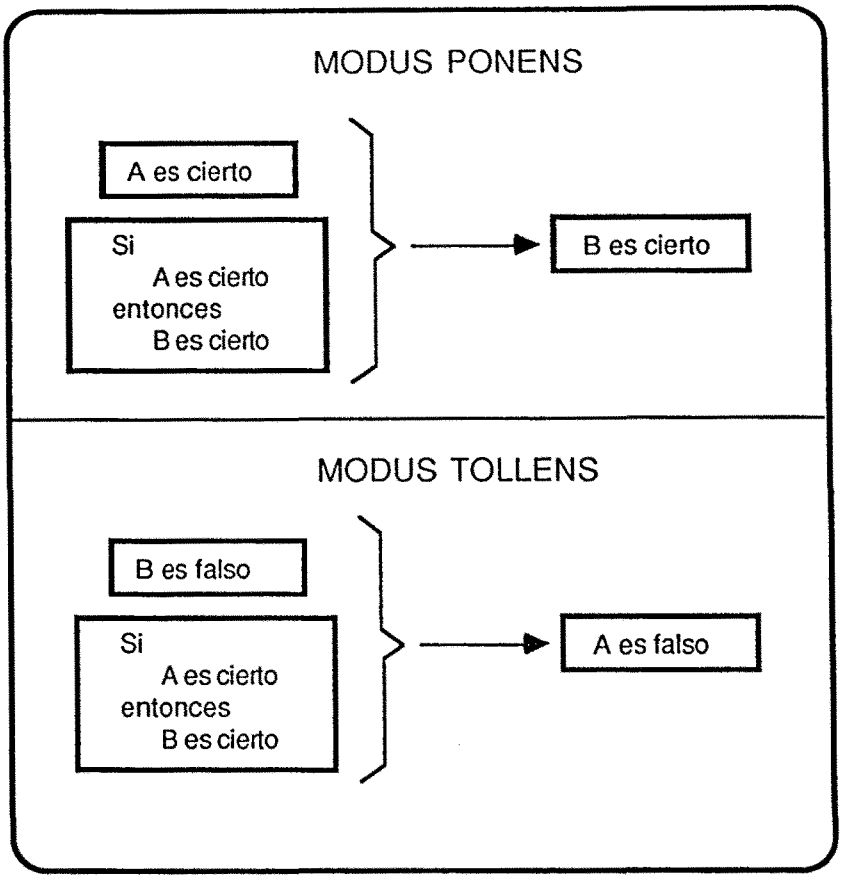


Figura 1.11.- Ilustración del "Modus Ponens" y "Modus Tollens"

El "modus tollens", que es una estrategia bastante menos corriente que el "modus ponens", pero que tiene la misma potencia que la anterior, afirma que si se tiene la regla anterior y se sabe que "B es falso", entonces puede afirmarse que "A es falso" (ver figura 1.11). Es curioso y sorprendente comprobar que muchos sistemas expertos no incluyen esta estrategia de inferencia, a pesar de que resulta muy sencilla y práctica su aplicación.

Para la obtención de conclusiones compuestas a partir de dos reglas encadenadas se utiliza el mecanismo llamado "resolución", que consta de los siguientes pasos:

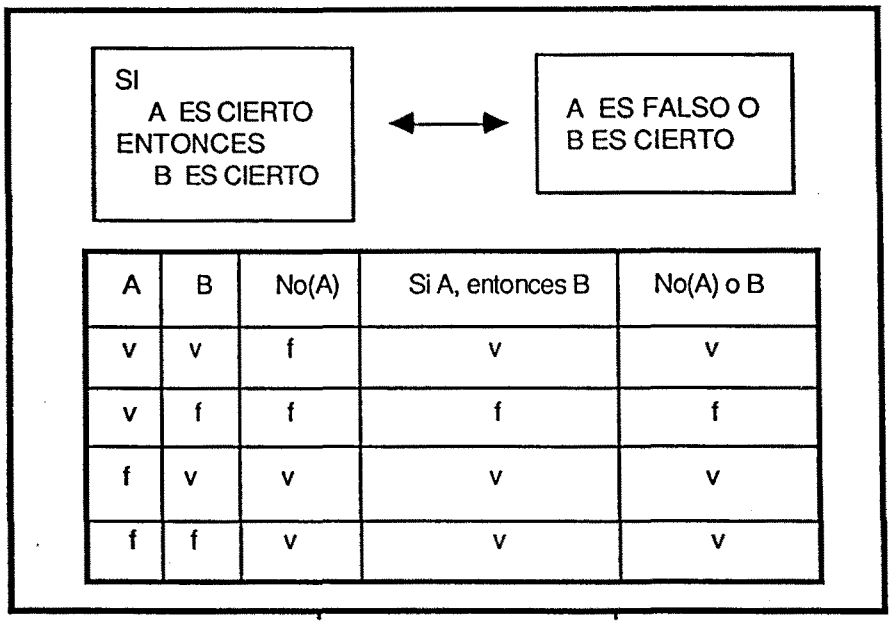
(a) se sustituyen las reglas por expresiones lógicas equivalentes
(b) se combinan éstas entre sí para dar una nueva expresión lógica
y (c) se combina ésta con la evidencia de los hechos.

Esto se hace teniendo en cuenta que:

(a) la regla: "Si A es cierto entonces B es cierto" es equivalente a la expresión lógica: "A es falso o B es cierto". Una demostración de esta

equivalencia se muestra en la tabla de verdad de la figura 1.12 (nótese la coincidencia en las dos últimas columnas).

(b) las expresiones lógicas "A es falso o B es cierto" y "B es falso o C es cierto" implican la expresión "A es falso o C es cierto". Esto se muestra en la tabla de verdad de la figura 1.13.



A	B	No(A)	Si A, entonces B	No(A) o B
v	v	f	v	v
v	f	f	f	f
f	v	v	v	v
f	f	v	v	v

Figura 1.12.- Ilustración del mecanismo de resolución

Conviene señalar que no siempre tienen por qué resultar conclusiones del análisis de una regla o del encadenamiento de dos o más reglas, ya que puede no conocerse la verdad o falsedad de las premisas. El sistema experto o, más precisamente, su motor de inferencia ante una situación como ésta, podrá optar por:

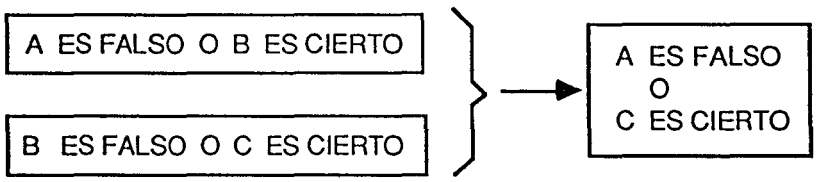
(a) abandonar la regla por no poder concluir nada, o

(b) preguntar, a través del subsistema de demanda de información, sobre la verdad o falsedad de las premisas y continuar el proceso hasta producirse una conclusión.

Otra forma de combinar varias reglas encadenadas para sacar conclusiones compuestas es mediante las técnicas de razonamiento incierto, que se describen brevemente a continuación.

En Lógica clásica, que es la que se ha utilizado hasta ahora ("modus ponens" y "modus tollens"), siempre que la premisa de una regla es cierta, la

conclusión es cierta (ver figura 1.14); así, dada la regla: "Si A es cierto entonces B es cierto" se puede decir que A implica B con probabilidad 1. Sin embargo, este modelo clásico tiene muchas limitaciones, ya que existen multitud de casos prácticos en los que esto no es así. Por ejemplo, la presencia de ciertos síntomas no siempre garantiza la existencia de una enfermedad, aún cuando estos síntomas sean un indicio muy fuerte de que así ocurra. Por ello, resulta mucho más práctico e interesante generalizar la lógica clásica por una de tipo incierto. De esta forma, la regla anterior, vista en forma generalizada, se interpreta diciendo que "A implica B con una probabilidad P(B/A)" (probabilidad de B condicionado por A o probabilidad de B supuesto que A es cierto) (ver segundo caso en figura 1.14). Además, incluso el caso clásico de no implicación, "A no implica B" puede tratarse como caso extremo de implicación: "A implica B con probabilidad nula" (ver tercer caso en figura 1.14).



A	B	C	No(A)	No(A) ó B	No(B) ó C	(No(A) ó B) y (No(b) ó C)	No(A) ó C
v	v	v	f	v	v	v	v
v	v	f	f	v	f	f	f
v	f	v	f	f	v	f	v
v	f	f	f	f	v	f	f
f	v	v	v	v	v	v	v
f	v	f	v	v	f	f	v
f	f	v	v	v	v	v	v
f	f	f	v	v	v	v	v

Figura 1.13.- Ilustración del mecanismo de resolución

Uno de los problemas que se presentan con este tipo de Lógica es la propagación de incertidumbre. Nótese que ahora toda afirmación (hecho) debe ir acompañada de una cierta medida de incertidumbre (probabilidad, factor de certeza, etc.) y que al combinar varios hechos inciertos hay que dotar a las conclusiones de su correspondiente medida de incertidumbre (Grayson (1960)). El problema radica en que fijadas las medidas de incertidumbre de las premisas, la incertidumbre de la conclusión normalmente no sólo no es única sino que puede tomar un infinidad de valores. Si a esto se añade la posibilidad de reglas inciertas, la cosa se complica aún más. Desgraciadamente los sistemas expertos actuales han cuidado muy poco el rigor en los métodos de propagación de incertidumbre y es éste uno de los puntos más débiles de los mismos. De todas formas este tema no es objeto de este apartado y se tratará con mucho más detalle en el apartado 1.6.3.

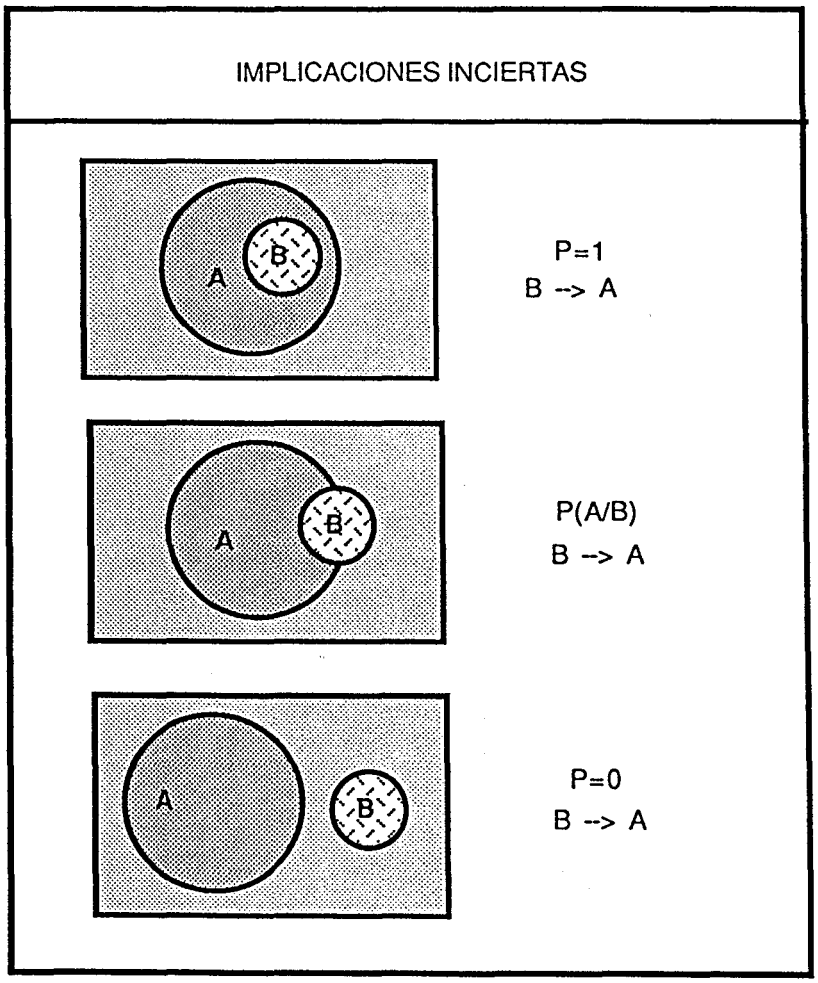


Figura 1.14.- Implicaciones inciertas

La elección de los grupos de reglas encadenadas para obtención de reglas compuestas puede hacerse siguiendo varias estrategias de lo que se llama control

del razonamiento. De estas estrategias depende la total eficacia del sistema. Normalmente, éste se realiza mediante:

(a) un sistema con un procedimiento para decidir por donde empezar. Nótese que los datos y las reglas están en una base de datos estática, y que debe existir un mecanismo que les ponga a funcionar. Si se desea conocer el valor de un cierto objeto, deberán localizarse reglas que hagan referencia a él.

(b) un mecanismo para que el sistema pueda decidir cómo continuar cuando se le presenten varias alternativas al mismo tiempo. A veces, el proceso de razonamiento llega a un punto en el que varias reglas están dispuestas para *"disparar"* su conclusión. El sistema se ve obligado a elegir qué regla debe examinar a continuación.

Entre las estrategias más comunes destacan:

- encadenamiento hacia adelante
- encadenamiento hacia atrás

Las figuras 1.15 y 1.16 ilustran estos dos tipos de encadenamiento, donde los números indican el orden en que se van evaluando las diferentes componentes (premisas o conclusiones) de las reglas. Si durante este proceso una regla concluye, se dice que esta regla "se ha disparado" o que esta regla "está activa". Las reglas disparadas o activas forman la base del mecanismo de explicación, que corre a cargo del subsistema de explicación, ya que toda conclusión nace de las reglas activas.

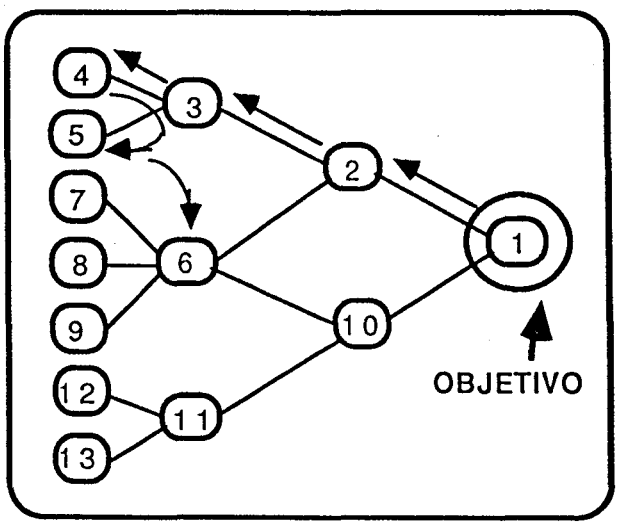


Figura 1.15.- Encadenamiento hacia atrás

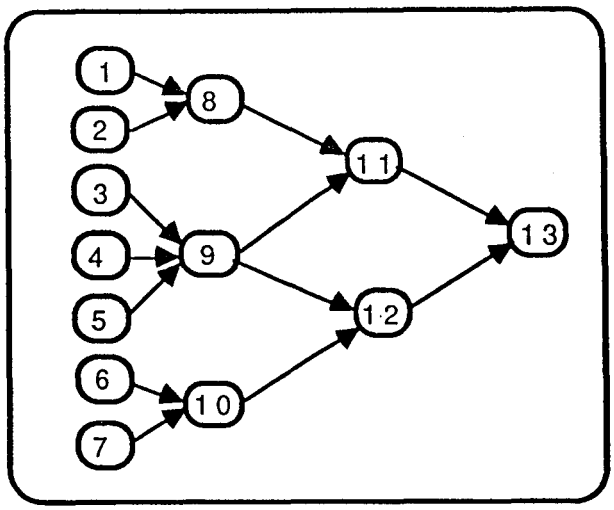


Figura 1.16.- Encadenamiento hacia adelante

Los sistemas de encadenamiento hacia adelante diferencian claramente la memoria de trabajo del banco de datos. La memoria de trabajo contiene los datos iniciales más los que surgen durante una consulta. Las premisas de las reglas que están en la base de datos se confrontan con los contenidos de la memoria de trabajo y los resultados que puedan derivarse de éstas pasan también a la memoria incrementando el conocimiento.

Conviene señalar que los mecanismos de encadenamiento hacia adelante se utilizan en problemas orientados a los datos o de diagnóstico, en los que se tienen unos hechos y se quiere saber cuales son las conclusiones que pueden derivarse de ellos. Por el contrario, los mecanismos de encadenamiento hacia atrás se utilizan en problemas orientados al objetivo, en los que se tienen unos objetivos y se quiere saber qué hechos son necesarios para conseguirlos.

Además de distinguir entre los encadenamientos hacia adelante y hacia atrás es importante distinguir entre la *búsqueda vertical* (depth first) y la *búsqueda horizontal* (breadth first) de un árbol de reglas encadenadas.

La forma de actuar de un sistema de encadenamiento hacia atrás verticalmente es buscar primero los detalles. Una búsqueda horizontal barre todas las premisas de una regla antes de enterarse de ningún detalle. La búsqueda horizontal será más eficiente si una regla funciona y aparece rápidamente el valor del objetivo.

Otro de los aspectos diferenciales del motor de inferencia es su capacidad para realizar razonamiento no monótono (Mac Dermot y Doyle (1980)). Se dice que un motor de inferencia es capaz de realizar un razonamiento monótono

cuando durante la consulta los hechos toman un único valor y las reglas no cambian. Si por el contrario, se permite que las reglas o los hechos cambien durante la sesión de consulta, se dice que el razonamiento es de tipo no monótono. Nótese que el cambio de uno de los valores de un atributo puede implicar el cambio de otros muchos, derivados de la aplicación de las reglas, y que llevar este control es muy complejo. Por ello, no todos los sistemas expertos están en condiciones de realizar un razonamiento no monótono.

Especial dificultad entraña la revisión de las conclusiones obtenidas en base a las reglas que en un determinado momento cambian. Por ello, los sistemas expertos con esta capacidad suelen ser muy caros.

1.5.2.-Sistemas expertos de tipo probabilístico

En este capítulo se analizará otro de los tipos más importantes de sistemas expertos: los "basados en la probabilidad". De ellos se destacará especialmente la estructura del espacio probabilístico (base de conocimiento) y el motor de inferencia, que se basa en las probabilidades condicionales (Carnap (1950), Parzen (1960), Luce y Suppers (1965), Harré (1970), Smart (1972), Duda et al. (1976), Domenech (1977), Castillo (1978), Castillo et al. (1987), Lindley (1987), etc.).

1.5.2.1.-La base del conocimiento

El centro de este tipo de sistemas lo constituye el espacio probabilístico, que forma lo que se ha llamado conocimiento abstracto o de aplicación general.

1.5.2.1.1.-Síntomas binarios y con múltiples opciones

El problema que se planteará es el de diagnosticar, en una población Ω, un conjunto de problemas suponiendo que los datos son síntomas binarios de un conjunto dado. Con objeto de ser más concreto, supóngase que se tiene una población, representada en la figura 1.17 por los puntos interiores al rectángulo que limita su contorno. Supóngase también que cada elemento de esa población tienen uno y sólo uno de los problemas que se muestran, y ninguno, uno o varios de los síntomas $\{S_j; i = 1, 2, .., m\}$ que aparecen en la misma figura. Elegido un caso, que vendrá representado por un punto del rectángulo, resulta muy fácil decir el problema y los síntomas que tiene, ya que basta observar a qué conjuntos de problemas o síntomas pertenece. Así, el individuo X de la figura tiene el problema E_1 con los síntomas S_1, S_2 y S_3. Si fuera posible conocer la frecuencia (número de casos) de todos los conjuntos (huecos) en que queda dividida la población

(rectángulo) por las intersecciones de los conjuntos que definen los problemas y los síntomas, se tendría la máxima información posible para definir los problemas en función de los síntomas y se estaría en unas condiciones óptimas para hacer el diagnóstico.

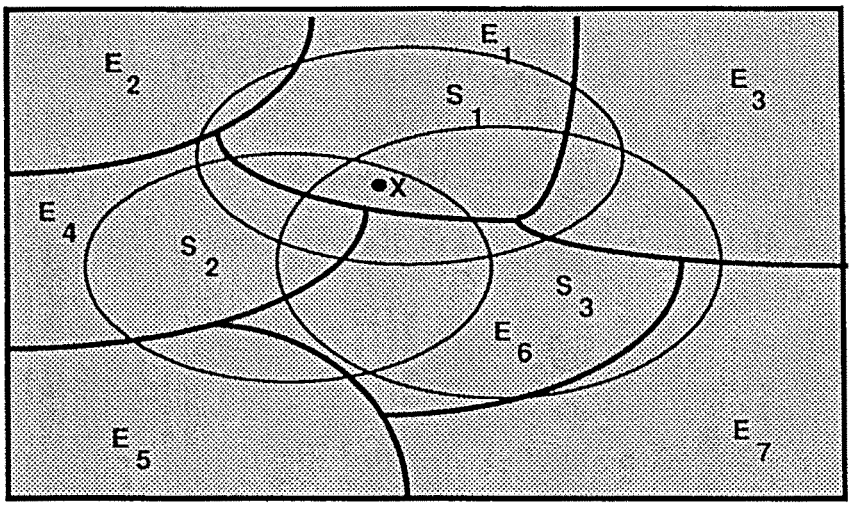


Figura 1.17.- Ilustración de un problema de diagnóstico

Sin embargo, eso, que es la situación ideal, sólo es posible cuando se trata con pocos síntomas. En efecto, en el caso de 100 problemas y 200 síntomas (caso no muy complicado por otra parte), el número de huecos resultante y, por tanto, el número de frecuencias a almacenar (parámetros), es 10^{62} que es tan grande que ningún ordenador de los actuales sería capaz de almacenar. A este modelo, que es el más general posible, es decir, el que tiene total libertad en sus parámetros, se le llama "modelo de dependencia general".

Debido a las dificultades anteriores, es necesario simplificar el modelo y suponer que hay algunas relaciones entre las diferentes frecuencias o lo que es igual restringir su libertad a la hora de fijar sus parámetros. En el capítulo 2 se darán distintas simplificaciones que darán lugar a otros tipos de modelos.

Hasta ahora se ha tratado con síntomas binarios, es decir, con síntomas con sólo dos opciones. Sin embargo, en muchos casos la solución de dos opciones no es satisfactoria y hay que acudir a los síntomas con varias opciones o niveles, con lo que el problema se complica.

Los síntomas múltiples se definen mediante particiones del conjunto población, es decir, que para cada uno de estos síntomas hay que dar el conjunto de individuos que poseen cada nivel del síntoma. Ahora la base de conocimiento almacenará las frecuencias de cada uno de los conjuntos intersección de cada

problema con cada uno de los subconjuntos que definen las particiones de los síntomas.

1.5.2.1.2.-El problema de la información incompleta

Otro de los problemas que se presentan en el diagnóstico es el de la información incompleta. Para presentar este problema se pondrá un ejemplo con síntomas binarios. Supóngase que se tiene un problema E en el que los síntomas relevantes para dicho problema son $\{S_j; i = 1, 2, .., m\}$. Los casos con este problema pueden ser clasificados en algún elemento del conjunto $\mathfrak{D}$:

$$\mathfrak{D} = \left\{ E \cap S_1^{\delta_1} \cap .. \cap S_m^{\delta_m} / \delta_i \in \{0, 1\}\ i = 1, .., m \right\}$$

donde

$$S_i^{\delta_i} = \left\{ \begin{matrix} S_i & \text{si } \delta_i = 1 \\ \overline{S_i} & \text{si } \delta_i = 0 \end{matrix} \right\}$$

El experto tendrá una información completa sobre los síntomas relevantes, si conoce la presencia o ausencia de todos los síntomas, o una información incompleta, si desconoce alguno de ellos.

Nótese que en el caso de m síntomas binarios serían necesarios 2^m parámetros (frecuencias) y que en el caso de información incompleta sería necesario conocer las frecuencias de los elementos del conjunto

$$\mathfrak{C} = \left\{ E \cap S_1^{\delta_1} \cap .. \cap S_m^{\delta_m} / \delta_i \in \{0, 1, 2\}\ i = 1, .., m \right\}$$

donde

$$S_i^{\delta_i} = \left\{ \begin{matrix} \overline{S_i} & \text{si } \delta_i = 0 \\ S_i & \text{si } \delta_i = 1 \\ \Omega & \text{si } \delta_i = 2 \end{matrix} \right\}$$

que hacen un total de 3^m parámetros.

1.5.2.2.-Motor de inferencia

El motor de inferencia probabilístico por excelencia es el que se basa en las probabilidades condicionadas y consiste en actualizar las probabilidades de que el caso en estudio tenga los diferentes problemas a medida que van conociéndose los síntomas que presenta o se van incorporando nuevos datos.

Antes de definir una probabilidad hace falta una estructura previa que se conoce con el nombre de espacio probabilístico. Éste a su vez consta de dos elementos: un conjunto Ω (un espacio muestral) y una clase, $\mathcal{A}$, de subconjuntos de Ω con estructura de álgebra.

Definición 1.1.- *(Álgebra).-* Sea Ω un conjunto. Una clase $\mathcal{A}=\{A_i \subseteq \Omega / i \in I\}$ de subconjuntos de Ω, donde I es un conjunto de índices, se dice que tiene estructura de álgebra si verifica las condiciones siguientes:

A1.– $\forall A_i \in \mathcal{A} \Rightarrow \overline{A}_i \in \mathcal{A}$

A2.– $\forall A_i, A_j \in \mathcal{A} \Rightarrow A_i \cup A_j \in \mathcal{A}$

donde $\overline{A}_i$ representa el conjunto complementario de A_i.

Las propiedades más inmediatas que se derivan de esta definición son:

PA1.– Si $\mathcal{A} \neq \phi$ entonces Ω y $\phi \in \mathcal{A}$

PA2.– $\forall A_i, A_j \in \mathcal{A} \Rightarrow A_i \cap A_j \in \mathcal{A}$

PA3.– $\forall A_i, A_j \in \mathcal{A} \Rightarrow A_i - A_j \in \mathcal{A}$

Nótese que puede obtenerse una definición equivalente a la de álgebra sustituyendo en la definición el axioma A2 por la propiedad PA2.

Definición 1.2.-*(σ-álgebra).*-Un álgebra, $\mathcal{A}$, se dice que es un σ-álgebra si

$$\forall A_i \in \mathcal{A}, i \in J \subseteq I, J \text{ numerable} \Rightarrow \bigcup_{i \in J} A_i \in \mathcal{A}$$

Además de las propiedades de las álgebras, las σ-álgebra tienen las siguientes propiedades

PA4.– $\forall A_i \in \mathcal{A}, i \in J \subseteq I, J \text{ numerable} \Rightarrow \bigcap_{i \in J} A_i \in \mathcal{A}$

PA5.- Toda σ-álgebra es cerrado respecto al paso al límite.

Es fácil demostrar que la intersección de un número cualquiera de álgebras (σ-álgebra) es un álgebra (σ-álgebra).

Definición 1.3.-*(Álgebra engendrada por una clase de conjuntos).*- Dado un conjunto Ω y una clase $\mathcal{C}$ no vacía de subconjuntos de Ω, se llama álgebra (σ-álgebra) generada por $\mathcal{C}$ al mínimo álgebra (σ-álgebra) que contiene a $\mathcal{C}$.

Uno de los métodos para obtención de esta mínima álgebra es el siguiente.

Proposición 1.1.- *(Mínima álgebra que contiene a una clase de conjuntos)*.-Sea $\mathcal{C}$ una clase cualquiera de sucesos de Ω y $\mathcal{C}_1$ la formada por intersecciones finitas de ϕ, Ω, y A ó $\bar{A}$ con $A \subset \Omega$. La mínima álgebra que contiene a $\mathcal{C}$ es la constituida por uniones finitas de elementos de $\mathcal{C}_1$ dos a dos disjuntos.

Definición 1.4.-*(σ-álgebra de Borel)*.- Considerese $\Omega=\mathfrak{R}$ y la clase $\mathcal{C}$ constituida por los intervalos (a, b], donde a, b pertenecen a $\mathfrak{R}$. La mínima σ-álgebra generada por esta clase se conoce con el nombre de σ-álgebra de Borel de $\mathfrak{R}$.

Definición 1.5.- *(Espacio probabilístico)*.- Un espacio probabilístico es un par $(\Omega,\mathcal{A})$ donde Ω es un espacio muestral y $\mathcal{A}$ es un álgebra o un σ-álgebra de sucesos de Ω.

Definición 1.6.- *(Probabilidad)*.-Sea $(\Omega,\mathcal{A})$ un espacio probabilístico. Una función que asocia a cada subconjunto de $\mathcal{A}$ un número real, comprendido entre 0 y 1, de tal forma que se cumplan los siguientes axiomas:

P1. – $P(\Omega) = 1$

P2. – $P(\bigcup_{j \in J} A_j) = \sum_{j \in J} P(A_j)$ si $A_k \cap A_j = \phi$ para todo $k \neq j$, (J numerable)

se dice que es una probabilidad.

Ello quiere decir que la probabilidad del conjunto universal Ω siempre es la unidad y que la probabilidad de cualquier conjunto unión de conjuntos disjuntos es la suma de sus probabilidades. A la terna $(\Omega,\mathcal{A}, P)$ se le llama espacio de probabilidad.

De estas propiedades se deducen otras más como, por ejemplo, las siguientes:

PP1. – $P(\bar{A}_i) = 1 - P(A_i) \quad \forall A_i \in \mathcal{A}$

PP2. – $P(\phi) = 0$

PP3. – si $A_i \subset A_j \Rightarrow P(A_i) \leq P(A_j) \quad \forall A_i, A_j \in \mathcal{A}$

PP4. – $P(A_i) \leq 1 \quad \forall A_i \in \mathcal{A}$

PP5. – $P(A_i \cup A_j) \leq P(A_i) + P(A_j) \quad \forall A_i, A_j \in \mathcal{A}$

PP6. – $P(\bigcup_{j \in J} A_j) \leq \sum_{j \in J} P(A_j)$ si J es numerable

PP7. – $P(A_i \cap A_j) \leq \min[P(A_i), P(A_j)] \leq \max[P(A_i), P(A_j)] \leq P(A_i \cup A_j)$

1.5.2.2.1.-Probabilidades condicionadas

Un tipo especial de probabilidad, es la probabilidad condicionada. El objeto de introducir el concepto de probabilidad condicionada es el de estudiar cómo se va modificando las probabilidades iniciales de los sucesos al conocer nueva información. Es decir, se pretende conocer cual es la probabilidad de una hipótesis "h" conocida una evidencia "e".

Definición 1.7.- *(Espacio probabilístico condicionado por un suceso).*-Sea $(\Omega,\mathcal{A})$ un espacio probabilístico y $e \in \mathcal{A}$ con $P(e)>0$. Se llama espacio probabilístico condicionado por e, al espacio probabilístico $(e, \mathcal{A}_e)$ donde

$$\mathcal{A}_e = \{B \in \mathcal{A} / \exists\, C \in \mathcal{A} \text{ tal que } B = C \cap e\}$$

Nótese que la clase $\mathcal{A}_e$ es una σ-álgebra en e y que, por tanto, $(e, \mathcal{A}_e)$ es un espacio probabilístico.

Definición 1.8.- *(Probabilidad condicionada por un suceso).*- Sea $(\Omega,\mathcal{A},P)$ un espacio de probabilidad y $(e, \mathcal{A}_e)$ el espacio probabilístico condicionado por el suceso $e \in \mathcal{A}$ con $P(e)>0$. Entonces se llama probabilidad condicionada por e a la función P_e

$$P_e(B) = \frac{P(B)}{P(e)} = \frac{P(e \cap C)}{P(e)} \quad \forall\, h \in \mathcal{A}_e \;\; (B = e \cap C)$$

Con frecuencia se utiliza la notación $P(h/e) \equiv P_e(h)$. Es fácil demostrar que P_e verifica los axiomas P1, P2 y que, por tanto, define una probabilidad sobre el espacio probabilístico $(\Omega,\mathcal{A})$ de la forma

$$(1.1) \qquad P(h/e) \equiv P_e(h) = \frac{P(h \cap e)}{P(e)} \quad \forall\, h \in \mathcal{A}$$

que recibe también el nombre de probabilidad condicionada por e.

Definición 1.9.- *(Independencia de sucesos).*- Dados dos sucesos e y h se dice que h es independiente de e si

$$(1.2) \qquad P(h/e) = P(h) \quad \text{es decir}, P(h \cap e) = P(h)\, P(e)$$

Por definición, se cumple que si "h" es independiente de "e" entonces "e" es independiente de "h".

1.5.2.2.3.-Teorema de Bayes

Para el cálculo de estas probabilidades condicionadas, además de utilizar directamente su definición, puede utilizarse el teorema de Bayes (Warner et al. (1961,64), Armitage (1971), Szolovits (1978), Castillo (1978), Edwards and Davis (1984), Charniak y Mac Dermot (1985), Fred et al. (1987), etc.) que se da en este apartado.

Teorema 1.1.- *(De la probabilidad total)*.-Sea $\{h_i\}_{i \in I}$ (I numerable) una clase de sucesos incompatibles dos a dos, y tal que $\bigcup_{i \in I} h_i = \Omega$. Entonces

$$P(e) = \sum_{i \in I} P(e / h_i)\, P(h_i) \qquad (1.3)$$

Demostración.-

Por el axioma (P2) se tiene

$$P(e) = P(e \cap \Omega) = P(e \cap (\bigcup_{i \in I} h_i)) = P(\bigcup_{i \in I} (e \cap h_i)) =$$
$$= \sum_{i \in I} P(e \cap h_i) = \sum_{i \in I} P(e / h_i)\, P(h_i)$$

Teorema 1.2.- *(De Bayes)*.- En las mismas condiciones que el teorema anterior se tiene que

$$P(h_i / e) = \frac{P(e / h_i)\, P(h_i)}{\sum_{i \in I} P(e / h_i)\, P(h_i)} \qquad (1.4)$$

Demostración.-

En efecto, ya que

$$P(h_i \cap e) = P(e / h_i)\, P(h_i) = P(h_i / e)\, P(e)$$

se tiene despejando y aplicando el teorema anterior

$$P(h_i / e) = \frac{P(e / h_i)\, P(h_i)}{P(e)} = \frac{P(e / h_i)\, P(h_i)}{\sum_{i \in I} P(e / h_i)\, P(h_i)}$$

$P(h_i)$ recibe el nombre de probabilidad "a priori" (antes de conocerse ninguna información) de que se dé el suceso h, $P(h_i/e)$ es la probabilidad "a posteriori" (después del conocimiento del suceso e) de que se dé el suceso h_i, $P(e/h_i)$ son las "verosimilitudes"

La importancia del teorema de Bayes está, precisamente, en que permite determinar las probabilidades "a posteriori" en función de las probabilidades "a priori" y de las "verosimilitudes".

Por tanto, las probabilidades "a priori" son las probabilidades de que el caso en estudio tenga un determinado problema antes de iniciar la experimentación o toma de datos, las probabilidades "a posteriori" son las probabilidades de que pertenezca a un determinado tipo después de conocer sobre él una cierta información "e", y las "verosimilitudes" son las probabilidades de que se obtenga esa información "e" en el supuesto de que el caso pertenezca realmente a cada uno de los grupos h_i.

1.6.-EL PROBLEMA DE LA MEDIDA DE LA INCERTIDUMBRE

1.6.1.-Introducción

Anteriormente, al hablar de implicaciones inciertas, se introdujo el concepto de lógica incierta, que generaliza a la lógica clásica, y se vió como toda afirmación debe ir acompañada de una medida de incertidumbre, que expresa la confianza que se tiene de que esa afirmación sea cierta. En la Lógica clásica toda afirmación es siempre segura; sin embargo, cuando se trabaja con implicaciones inciertas, las afirmaciones hay que entenderlas como posibles en vez de como seguras. Las ventajas que se derivan de esta generalización de la Lógica clásica son claras y fueron ya descritas en el apartado 1.5.1.2. En este apartado se analizan algunas medidas de incertidumbre y se discute el problema de su propagación.

La medida de incertidumbre más antigua e intuitiva es la probabilidad. Prácticamente todo el mundo está familiarizado con este concepto y con su equivalencia frecuentista, pues no en vano se utiliza mucho en el lenguaje vulgar. En el campo de los sistemas expertos, además de esta medida, se utilizan otras más, como los factores de certeza, las medidas de la teoría de la evidencia, de Dempster y Shafer, y las funciones de posibilidad de la Lógica difusa. Sin

embargo, algunas de ellas son poco intuitivas y de manipulación matemática compleja. Por ello, cabe preguntarse por las razones que han llevado a su aparición en este campo. De hecho existe una gran controversia al respecto y pueden encontrarse grandes defensores de la probabilidad, como Adams (1976), Cheeseman (1985), Lindley (1987), etc.), y también detractores, como Shafer (1982), Zadeh (1983), etc.). Como ejemplo, Lindley dice textualmente:

> *"la única descripción satisfactoria de la incertidumbre es la probabilidad. Al decir esto se hace referencia a que toda afirmación incierta debe estar en la forma de una probabilidad; que varias incertidumbres deben ser combinadas usando las reglas de la probabilidad; y que el cálculo de probabilidades es adecuado para manejar todas las situaciones con incertidumbre. En particular, las descripciones alternativas de incertidumbre son innecesarias".*

En este trabajo se defenderán los métodos que se basan en la probabilidad, si bien se pondrá de manifiesto los problemas que presentan. Estos problemas, no solo no han sido resueltos por los demás métodos sino que los han complicado aún más.

Los modelos de tipo probabilístico han sido criticados, bien por el elevado número de parámetros o bien por las dificultades de estimarlos a partir de los datos. De hecho, hay un acuerdo generalizado en que el modelo general de dependencia no es práctico en la gran mayoría de los casos. Otra de las fuentes de crítica proviene del hecho de que los modelos de independencia son demasiado simples para reproducir muchos casos prácticos. Sin embargo, entre estos dos extremos (dependencia general e independencia) existen muchas posibilidades, algunas de ellas se indicarán en el próximo capítulo.

No debe olvidarse, sin embargo, que muchos de los métodos alternativos a la probabilidad necesitan una infraestructura probabilística. Este es el caso de los factores de certeza y de las medidas de la teoría de la evidencia. Estos métodos podrían ser defendidos en aras a utilizar conceptos más sencillos o intuitivos, o de facilitar la propagación de la incertidumbre. Sin embargo, éste no es el caso. En realidad no hay problema alguno en definir nuevos conceptos para medir la incertidumbre sino en olvidarse que detrás de ellos están unos axiomas que, por coherencia, deben satisfacerse. Lindley (1987) afirma:

> *"bajo ningún concepto los axiomas de la probabilidad deben ser violados si directa o indirectamente se utiliza una probabilidad".*

Por ello, hay que clarificar qué métodos son coherentes y cuáles no lo son. Ni siquiera el hecho de que un experto humano utilice conocimiento no coherente

justifica un modelo incoherente. El objetivo del sistema experto es el de adquirir el conocimiento humano pero no, los errores. En este sentido el sistema experto mejoraría al experto humano.

El problema principal de coherencia surge al propagar la incertidumbre mediante fórmulas sin base axiomática (Klahr y Waterman (1986)). De hecho, muchas de las fórmulas de propagación no son mejores que la, tan criticada, hipótesis de independencia. Por ello, no se entienden las razones de su proliferación en el ámbito de los sistemas expertos.

1.6.2.-Medidas de incertidumbre

En este apartado se dan la definición y propiedades (Frost (1986)) de las principales medidas de incertidumbre a parte de la probabilidad, analizada en el apartado 1.5.2.2, esto es, los factores de certeza, las de la teoría de la evidencia y las de la teoría difusa, dejando para el apartado siguiente el problema de su propagación.

1.6.2.1.-Teoría matemática de la evidencia

Esta teoría fue desarrollada inicialmente por Dempster en los años 60 y posteriormente por Shafer como alternativa a los modelos probabilísticos y a los factores de certeza tan fuertemente criticados por el excesivo número de parámetros y la arbitrariedad a la hora de propagar la incertidumbre en uno y otro caso respectivamente.

El objetivo es definir unas medidas que asignen a cada $A \in \mathcal{P}(U)$ un valor que refleje el grado de certidumbre de que un determinado elemento de U pertenezca a A, desconociéndose a priori si esto es cierto o falso. Serán las medidas inciertas (Bratnagal y Kanal(1986), Klir (1989)).

Definición 1.10.- *(Medida incierta).-* Sea $g: \mathcal{P}(U) \rightarrow [0,1]$. Si g verifica

(I1) $g(\phi) = 0$ y $g(U) = 1$

(I2) para cada $A, B \in \mathcal{P}(U)$ si $A \subseteq B$ entonces $g(A) \leq g(B)$

(I3) para cada $(A_i \in \mathcal{P}(U) / i \in \mathcal{N})$ con $A_1 \subseteq A_2 \subseteq \ldots$ *ó* $A_1 \supseteq A_2 \supseteq \ldots$ se tiene

$$\lim_{i \to \infty} g(A_i) = g\left(\lim_{i \to \infty} A_i\right)$$

entonces se dice que es una medida incierta.

Esta definición puede restringirse aún más considerando como conjunto de partida de la función g un conjunto con estructura de álgebra de Borel o σ-álgebra.

Definición 1.11.- *(Medida de credibilidad).-* Sea Cr: $\mathcal{P}(U) \rightarrow [0,1]$ una medida incierta. Se dice que es una medida de credibilidad si cumple

(CR1) $$Cr(A_1 \cup A_2 \cup ... \cup A_n) \geq \sum_i Cr(A_i) - \sum_{i<j} Cr(A_i \cap A_j) +$$

$$+ \sum_{i<j<K} Cr(A_i \cap A_j \cap A_k) + ... + (-1)^{n+1} Cr(A_1 \cap A_2 \cap ... \cap A_n)$$

Nótese que la medida de credibilidad es monótona. En efecto, sea $A \subseteq B$. Si llamamos C=B-A entonces

$$Cr(A \cup C) = Cr(B) \geq Cr(A) + Cr(C) - Cr(A \cap C) \Rightarrow$$
$$\Rightarrow Cr(B) \geq Cr(A) + Cr(C) \Rightarrow Cr(B) \geq Cr(A)$$

Por otra parte si tomamos $A_1 = A$ y $A_2 = \overline{A}$ entonces

$$Cr(A) + Cr(\overline{A}) \leq 1$$

Definición 1.12.- *(Medida de plausibilidad).-* Sea Pl: $\mathcal{P}(U) \rightarrow [0,1]$ una medida incierta. Se dice que es una medida de plausibilidad si cumple

(Pl1) $$Pl(A_1 \cap A_2 \cap ... \cap A_n) \leq \sum_i Pl(A_i) - \sum_{i<j} Pl(A_i \cup A_j) +$$

$$+ \sum_{i<j<K} Pl(A_i \cup A_j \cup A_k) + ... + (-1)^{n+1} Pl(A_1 \cup A_2 \cup ... \cup A_n)$$

Nótese que si tomamos $A_1 = A$ y $A_2 = \overline{A}$ entonces

(1.5) $$Pl(A) + Pl(\overline{A}) \geq 1$$

Alternativamente una medida de plausibilidad Pl puede definirse como una función de $\mathcal{P}(U)$ en [0,1] que está relacionada con la medida de credibilidad Cr mediante

(1.6) $$Pl(A) = 1 - Cr(\overline{A}) \quad \forall A \in \mathcal{P}(U)$$

En la teoría de Dempster-Shafer se parte de un conjunto U que se llama *marco de discernimiento* formado por hipótesis mutuamente exclusivas y exhaustivas. Una unidad de evidencia, es decir, la que resulta de una cierta información disponible, se define como una generalización de la probabilidad.

Definición 1.13.-*(Asignación de probabilidad básica)*- Se llama asignación de probabilidad básica, m, a una aplicación del conjunto $\mathcal{P}(U)$ formado por todos los subconjuntos posibles de U en el intervalo [0,1], verificando las dos condiciones siguientes:

(AP1) $\sum_{A \in \mathcal{P}(U)} m(A) = 1$

(AP2) $m(\phi) = 0$

La condición (AP2) se añade para representar que este conjunto corresponde a una hipótesis falsa, sin embargo, esto no es cierto si todas las hipótesis en U son exhaustivas. Esta condición ha sido cuestionada en muchas ocasiones, pero a pesar de ello, se supondrá en lo que sigue que $m(\Phi)=0$. Las asignaciones de probabilidad básicas que verifican esta propiedad se llaman *normales*.

Nótese que m no es una medida incierta.

A cada conjunto $A \in \mathcal{P}(U)$ le asociamos $m(A) \in [0,1]$ que representa la evidencia que se tiene sobre la ocurrencia del conjunto A, o de otra forma, la evidencia de que se presente un suceso cuya descripción está incluida en A, sin tener en cuenta cómo está repartido esta medida entre los subconjuntos de A.

Shafer probó que dos medidas inciertas P^* y P_* se pueden expresar de forma única a partir de una asignación de probabilidad m mediante

$$P^*(A) = \sum_{B \subseteq A} m(B) \qquad \text{y} \qquad P_*(A) = \sum_{B \cap A \neq \phi} m(B)$$

si y solamente si son respectivamente superaditivas y subaditivas, es decir, si y solamente si P^* es una medida de credibilidad y P_* es una medida de incredibilidad. Por esta razón se puede definir la credibilidad y la plausibilidad mediante

(1.7) $$Cr(A) = \sum_{B \subseteq A} m(B)$$

(1.8) $$Pl(A) = \sum_{B \cap A \neq \phi} m(B)$$

Esta definición permite interpretar la credibilidad como una cota inferior de la probabilidad de A, ya que sólo se suman las probabilidades de los sucesos en los que la ocurrencia de A es segura. Nótese que los conjuntos asociados a los casos de duda (superconjuntos de A) no se incluyen en la suma anterior. Análogamente, la plausibilidad puede interpretarse como una cota superior de la probabilidad de A, ya que se suman tanto las probabilidades de los casos en los que A es dudoso como las de los casos en que A es seguro. Nótese que las dos medidas acotan la probabilidad, es decir, la probabilidad de $A \in \mathcal{P}(U)$ está contenido en el intervalo [Cr(A),Pl(A)]. Por esta razón, Dempster llamaba a la credibilidad y a la plausibilidad "lower probability" y "upper probability" respectivamente.

Proposición 1.2.- *(Propiedades de la credibilidad y plausibilidad)*- Entre las propiedades más importantes de estas dos medidas destacan

(1) $Pl(\phi) = Cr(\phi) = 0$
(2) $Pl(A) \geq Cr(A)$
(3) $Cr(U) = Pl(U) = 1$

Demostración.-
(1).
$Pl(\phi) = Cr(\phi) = m(\phi)$
(2).
Si $A = \phi$ por (1) se tiene probado (2). En el caso de que $A \neq \phi$ si $B \subseteq A$ entonces $B \cap A \neq \phi$ y, por tanto, $Cr(A) \leq Pl(A)$.
(3).
Como $\forall A \in \mathcal{P}(U)$, $A \subseteq U$ entonces $Cr(U) = Pl(U) = \sum_{A \in \mathcal{P}(U)} m(A) = 1$

Definición 1.14.- *(Elemento focal)*- Se llama elemento focal a cada $A \in \mathcal{P}(U)$ con $m(A)>0$. El conjunto de elementos focales lo representaremos por $\mathcal{F}$.

Definición 1.15.- *(Cuerpo de evidencia)*- Al par $(\mathcal{F},m)$ se le llama cuerpo de evidencia.

Proposición 1.3.-*(Propiedad de la credibilidad y plausibilidad)*- Sea $\mathcal{F}=\{A \in \mathcal{P}(U) / m(A)>0\}$ el conjunto de elementos focales para una unidad de evidencia m. Supongamos que está ordenado por inclusión, es decir, $A_1 \subseteq A_2 \subseteq ... \subseteq A_n$ con $A_i \in \mathcal{F}$, entonces se verifica

$$\forall\ A,B \in \mathcal{P}(U) \qquad Cr(A \cap B) = \min(Cr(A), Cr(B))$$
$$Pl(A \cup B) = \max(Pl(A), Pl(B))$$

Demostración.-

En efecto, sea $\mathcal{F}=\{A_1, A_2,...,A_n\}$ con $A_1 \subseteq A_2 \subseteq ... \subseteq A_n$ y A y B subconjuntos arbitrarios de U. Si se llama

$$i_1 = \max\{i / A_i \subseteq A\}, \ i_2 = \max\{i / A_i \subseteq B\}$$

entonces

$$A_i \subseteq A \cap B \ \text{ sii } \ i \leq \min(i_1, i_2).$$

Por tanto,

$$Cr(A \cap B) = \sum_{i=1}^{\min(i_1, i_2)} m(A_i) = \min\left[\sum_{i=1}^{i_1} m(A_i), \sum_{i=1}^{i_2} m(A_i) \right] =$$

$$= \min(Cr(A), Cr(B))$$

Por otra parte,

$$Pl(A \cup B) = 1 - Cr(\overline{A \cup B}) = 1 - Cr(\bar{A} \cap \bar{B}) = 1 - \min(Cr(\bar{A}), Cr(\bar{B})) =$$
$$\max(1 - Cr(\bar{A}), 1 - Cr(\bar{B})) = \max(Pl(A), Pl(B))$$

Nótese que si los elementos focales son disjuntos entonces la función de credibilidad y plausibilidad coinciden:

$$\forall\ A \in \mathcal{P}(U) \qquad Cr(A) = \sum_{B \subseteq A} m(B) = m(A) = \sum_{B \cap A \neq \phi} m(B) = Pl(A)$$

Definición 1.16.- *(Centro de $\mathcal{F}$)*- La unión de todos los elementos focales recibe el nombre de centro de $\mathcal{F}$ y se denotará por $\mathcal{C}$.

Proposición 1.4.- Si $\mathcal{C}$ es el centro de $\mathcal{F}$ para una unidad de evidencia sobre U entonces se cumple que para cada $A \in \mathcal{P}(U)$,

$$Cr(A) = 1 \Leftrightarrow \mathcal{C} \subseteq A$$

Dempster y Shafer proponen utilizar el intervalo [Cr(A),Pl(A)] para medir la incertidumbre de A. En los siguientes ejemplos y en la Tabla 1.6 se muestran algunas situaciones en las que este intervalo nos sirve para medir la ignorancia y la incertidumbre de los subconjuntos posibles de un marco de discernimiento U.

(a) Ignorancia.
Una situación de ignorancia puede representarse, por ejemplo, como

$$\mathcal{F} = \{U\} \text{ con } m(U) = 1 \text{ entonces } \forall A \in \mathcal{P}(U) \text{ es } Cr(A) = 0 \text{ y } Pl(A) = 1$$

(b) Certidumbre.
Una situación de certidumbre puede representarse, por ejemplo, como

$$\mathcal{F} = \{A\} \text{ con } m(A) = 1 \text{ entonces } Cr(A) = Pl(A) = 1$$

Si además A está formado por un único elemento entonces el caso anterior indica ausencia de ignorancia, pero, sin embargo, si A tiene subconjuntos propios existiría ignorancia al no poder distinguir entre los elementos de A.

(c) Incertidumbre.
Una situación de incertidumbre puede darse, por ejemplo, cuando los elementos focales son disjuntos dos a dos.

$$\mathcal{F} = \{A, B\} \text{ con } m(A) = m(B) = 0.5 \text{ entonces } \begin{Bmatrix} Cr(A) = Pl(A) = 0.5 \\ Cr(B) = Pl(B) = 0.5 \end{Bmatrix}$$

Nótese que puede darse el caso de ausencia total de ignorancia (Cr(A)=Pl(A)) con incertidumbre alta.

CASO	HECHO EN EL QUE SE REFLEJA	EJEMPLO [Cr(A),Pl(A)]
IGNORANCIA	Cr(A)<<Pl(A)	[0,1]
INFORMACION MAXIMA	Cr(A)=Pl(A)	[0.6,0.6]
CERTIDUMBRE	Cr(A) y Pl(A) cercanas a 1	[0.99,1]
INCERTIDUMBRE	Cr(A) y Pl(A) cercanas a 0.5	[0.49,0.50]

Tabla 1.6- Ilustración de la relación entre el intervalo [Cr(A),Pl(A)], ignorancia e incertidumbre

La teoría de medidas inciertas surge para generalizar la teoría de probabilidad. A pesar de las similudes y semejanzas vamos a ver como estas teorías se desarrollan paralelamente y que llegado un determinado punto se produce la divergencia entre ambas.

Definición 1.17.- *(Medida de credibilidad bayesiana).-* Sea Bel: $\mathcal{P}(U) \rightarrow [0,1]$ una medida de credibilidad. Se dice que es bayesiana si

$$\forall A, B \in \mathcal{P}(U) \text{ con } A \cap B = \phi \text{ entonces } Bel(A \cup B) = Bel(A) + Bel(B)$$

Teorema 1.3.- Una medida de credibilidad definida en $\mathcal{P}(U)$ con U un conjunto finito es una medida de credibilidad bayesiana sii m viene dada por

$$m(\{x\}) = Bel(\{x\}) \quad y \quad m(A) = 0 \quad con\ A \in \{B \in \mathcal{P}(U) / card(B) \neq 1\}$$

Demostración.-

* Implicación directa.

Por definición de m, $m(\Phi)=0$. Tomamos $A \in \mathcal{P}(U)$, $A=\{a_1,a_2,...,a_n\}$ entonces

$$Bel(A) = Bel(\{a_1\}) + Bel(\{a_2,..,a_n\}) = .. = Bel(\{a_2\}) + .. + Bel(\{a_n\})$$

como $Bel(\{a_i\}) = m(\{a_i\})$ se tiene que

$$Bel(A) = \sum_{i=1}^{m} m(\{a_i\})$$

y, por tanto,

$$m(A) = 0 \qquad \forall\ A \in \{B \in \mathcal{P}(U) / card(B) \neq 1\}$$

* Implicación recíproca.

Se considera $A,B \in \mathcal{P}(U)$ con $A \cap B = \phi$

$$Bel(A) + Bel(B) = \sum_{x \in A} m(\{x\}) + \sum_{x \in B} m(\{x\}) = \sum_{x \in A \cup B} m(\{x\}) = Bel(A \cup B)$$

entonces se ha probado que Bel es una medida de credibilidad bayesiana.

La medida de credibilidad bayesiana no es otra que la probabilidad que surge, por tanto, de imponer una condición adicional a las medidas de credibilidad.

1.6.2.2.-Lógica difusa

En teoría de la probabilidad, el conocimiento de P(A) y de P(B) no permite conocer los valores de $P(A \cap B)$ y $P(A \cup B)$. Esto quiere decir que dados P(A) y P(B) existen infinitos valores de $P(A \cap B)$ y $P(A \cup B)$ que satisfacen los axiomas de la probabilidad. Sin embargo, estos dos valores, $P(A \cap B)$ y $P(A \cup B)$, no son independientes, ya que están ligados por la propiedad PD5. Por tanto, existe sólo un grado de libertad, ya que fijado uno de ellos puede calcularse el otro mediante la propiedad PD5.

En teoría difusa se pretende definir una medida incierta g de manera que dados dos conjuntos A y B se cumpla que $g(A \cup B)$ y $g(A \cap B)$ sean funciones de g(A) y g(B), es decir,

$$g(A \cap B) = h\,(g(A), g(B))$$
$$g(A \cup B) = f\,(g(A), g(B))$$

Nótese que si g es una medida incierta se debe de cumplir por (I2) que

(1.9) $\quad h(g(A), g(B)) \leq \min(g(A), g(B))$

(1.10) $\quad f(g(A), g(B)) \geq \max(g(A), g(B))$

A las medidas que verifican la igualdad en (1.9) y (1.10) se llamarán medidas de necesidad y posibilidad respectivamente.

Definición 1.18.- *(Medidas de necesidad y posibilidad).-* Sean $\eta, \pi : \mathcal{P}(U) \rightarrow [0,1]$ dos medidas difusas. Se dice que

(N1) η es una medida de necesidad si $\eta(A \cap B) = \min(\eta(A), \eta(B))$

(P1) π es una medida de posibilidad si $\pi(A \cup B) = \max(\pi(A), \pi(B))$

Nótese que una medida de posibilidad puede definirse a partir de una de necesidad mediante la siguiente expresión

(1.11) $\quad \eta(A) = 1 - \pi(\overline{A}) \quad \forall\, A \in \mathcal{P}(U)$

La proposición 1.3 pone de manifiesto que Pl es una medida de posibilidad y Cr es una medida de necesidad cuando los elementos focales están ordenados por inclusión.

Proposición 1.5.- *(Propiedades).-* Sea A un elemento de $\mathcal{P}(U)$. Se cumple:

(1) $\quad \eta(A) > 0 \Rightarrow \pi(A) = 1$

(2) $\quad \pi(A) < 1 \Rightarrow \eta(A) = 0$

Demostración.-

(1).

Por definición de medida de necesidad

$$0 = \eta(\phi) = \eta(A \cap \overline{A}) = \min(\eta(A), \eta(\overline{A}))$$

y, como por hipótesis, $\eta(A) > 0$,

$$\eta(A) > 0 \Rightarrow \eta(\overline{A}) = 0 \Rightarrow \pi(A) = 1$$

(2).

Se deduce trivialmente de (1).

En la figura 1.18 aparecen representados gráficamente los diferentes tipos de medidas inciertas.

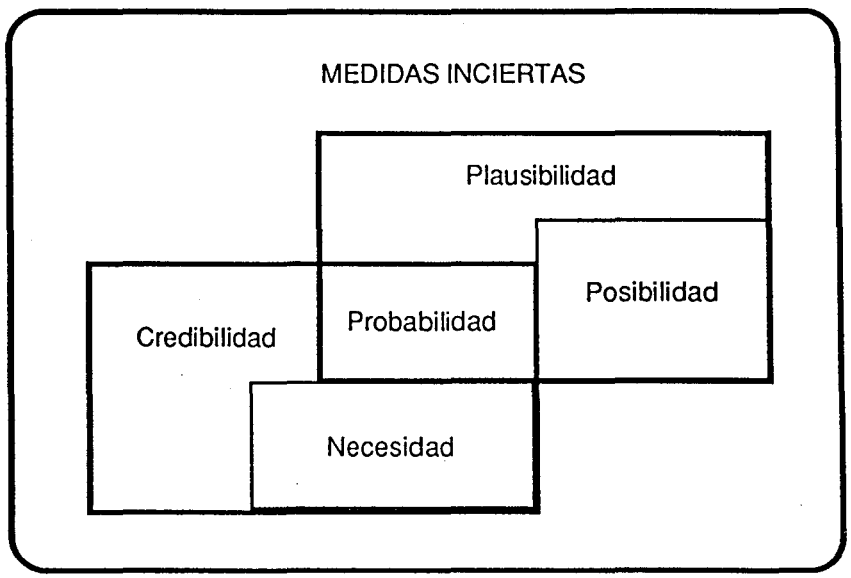


Figura 1.18.-Tipos de medidas inciertas

Nótese que la única medida de probabilidad que lo es también de necesidad y posibilidad es aquella que tiene un único elemento focal que es simple. En efecto, como P es una medida de necesidad y posibilidad por (I1) se tiene

(1.12) $$0 = P(\phi) = P(A \cap \overline{A}) = \min(P(A), P(\overline{A}))$$
$$1 = P(U) = P(A \cup \overline{A}) = \max(P(A), P(\overline{A}))$$

que permite deducir

(1.13) $$P(A) \in \{0, 1\} \quad \forall A \in \mathcal{P}(U)$$

Por otra parte, si P es una probabilidad por el Teorema 1. 3

(1.14) $$\mathcal{F} \subseteq \{ \{x\} / x \in U \} \quad P(\{x\}) = m(\{x\}) \quad \forall \{x\} \in \mathcal{F}$$
$$P(A) = 0 \quad \forall A \text{ con } \text{card}(A) \neq 1$$

y por (1.13)

(1.15) $$P(\{x\}) = m(\{x\}) = 1 \quad \forall \{x\} \in \mathcal{F}$$

de lo que se sigue que $\mathcal{F}$ únicamente está formado por un subconjunto de U con un elemento sobre el cual P vale la unidad.

Aunque se han introducido las medidas de posibilidad y necesidad a partir de las medidas de credibilidad podría haberse optado por definirlas a partir de la teoría de los conjuntos difusos como en sus orígenes lo hizo Zadeh. Brevemente se indicará el proceso aunque para ello se deberá estudiar previamente la base sobre la que se apoya esta teoría, esto es, los conjuntos difusos o borrosos.

1.6.2.2.1.-Conjuntos difusos

La Lógica clásica define la pertenencia de los distintos elementos a un conjunto haciéndoles corresponder el valor 1 si pertenecen y cero si no. Zadeh (1965), sin embargo, define una función, que les asocia valores en el intervalo [0,1] indicando con ello el grado de pertenencia a dicho conjunto. El valor uno va asociado a los elementos que con toda seguridad pertenecen al conjunto y cero a los que no, mientras que los valores intermedios se asocian a elementos de pertenencia dudosa. Esta idea implica un cambio de perspectiva frente a la idea clásica de pertenencia y se hace muy costosa inicialmente. No obstante, produce un enriquecimiento notable ya que la primera es un caso particular de la segunda y, en consecuencia, cualquier problema planteado en forma clásica puede también ser resuelto en forma difusa. Esto ha motivado un desarrollo notable de la lógica difusa en los últimos años.

Dado un conjunto universal U, según la teoría clásica de conjuntos, existe asociado a cada subconjunto $A \in \mathcal{P}(U)$ una función que nos refleja si un elemento de U lo es o no de A. Esta función llamada característica viene dada por

$$\mu_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases}$$

Sin embargo, en la práctica existen a menudo conjuntos imprecisos en los que la pertenencia o no de un elemento no resulta fácil de determinar. A este tipo de conjuntos se los conoce con el nombre de conjuntos difusos. Un ejemplo de conjunto difuso puede ser el que representa a las personas muy jovenes de una determinada población, ya que no existe ningún criterio objetivo para decidir si una persona en concreto pertenece a dicho conjunto.

Definición 1.19.- *(Conjunto difuso).-* Sea U un conjunto universal. Un subconjunto difuso $\tilde{A}$ de U es el conjunto de pares

$$\{(x, \mu_A(x)) / x \in U\} \quad \text{con} \quad \mu_A : U \to [0, 1]$$

donde el valor $\mu_A(x)$ representa el grado de pertenencia de x al concepto representado por $\tilde{A}$.

Esta definición puede generalizarse considerando la función $\mu_A : U \to L$ con L cualquier conjunto parcialmente ordenado. Normalmente, se considera que L es un retículo y, por ello, estos conjuntos reciben el nombre de L-conjuntos donde L es una abreviatura de retículo en inglés (lattice).

Se define el conjunto difuso vacío como aquel que $\mu_\phi(x) = 0$ para cada x en U. Se denotará por $\tilde{\mathcal{P}}(U)$ al conjunto de todos los subconjuntos difusos de U.

Definición 1.20.- *(Soporte de un conjunto difuso)* .- Sea $\tilde{A}$ un conjunto difuso de U. El soporte de $\tilde{A}$ es el conjunto

$$\sup \tilde{A} = \{x \in U / \mu_A(x) > 0\}$$

Una vez formalizado la noción de conjunto difuso se necesita definir ciertas operaciones que permitan su manejo. Originalmente la teoría de conjuntos difusos definía la complementación, unión e intersección de la siguiente forma

$$\begin{aligned} &\mu_{\overline{A}}(x) = 1 - \mu_A(x) \\ &\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x)) \\ &\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x)) \end{aligned} \qquad (1.16)$$

que incluían como caso particular a los conjuntos clásicos. A estas operaciones se las conoce con el nombre de operaciones estandar de conjuntos difusos.

Definición 1.21.- *(Complementario de un conjunto difuso).*- Sea $\tilde{A}$ un conjunto difuso de U. Un complementario de $\tilde{A}$ viene dado por la funcion c:[0,1] → [0,1] cumpliendo

(CD1) c(0)=1, c(1)=0

(CD2) para cada a, b ∈ [0, 1] con a < b se cumple c(a) ≥ c(b)

siendo $c(\mu_{\tilde{A}}(x))$ el grado de pertenencia de x en el concepto negación del representado por $\tilde{A}$.

Nótese que existen muchas funciones que verifican estas dos condiciones para c, por este motivo cuando se hace referencia al complementario de $\tilde{A} \in \tilde{\mathcal{P}}(U)$ es necesario especificar referido a qué función c se trata. De este modo se denotará por $C(\tilde{A})$ al complementario según

$$C: \tilde{\mathcal{P}}(U) \to \tilde{\mathcal{P}}(U) \quad \text{con} \quad c(\mu_{\tilde{A}}(x)) = \mu_{C(\tilde{A})}(x) \quad \forall x \in U$$

Obsérvese que para cada conjunto $\tilde{A} \in \tilde{\mathcal{P}}(U)$ existe, por tanto, una clase de funciones complementarias que puede restringirse si se añaden más propiedades adicionales como por ejemplo

(CD3) c es una función continua

(CD4) c es involutiva, es decir, c(c(a))=a para cada $a \in [0,1]$.

Ejemplo 1.1.- Se considera la siguiente clase de funciones, que recibe el nombre de clase de Sugeno (1977), definida mediante

$$c_\lambda(a) = \frac{1-a}{1+\lambda a} \quad \lambda \in (-1, \infty)$$

Se trata de una clase de complementarios difusos involutivos. En efecto,

(CD1)

$$c_\lambda(0) = \frac{1-0}{1+0} = 1 \qquad c_\lambda(1) = \frac{1-1}{1+\lambda} = 0$$

(CD2)

$$\frac{1-a}{1+\lambda a} \geq \frac{1-b}{1+\lambda b} \Leftrightarrow 1+\lambda b - a - \lambda ab \geq 1+\lambda a - b - \lambda ab \Leftrightarrow$$

$$\Leftrightarrow \lambda(b-a) \geq -(b-a) \Leftrightarrow \lambda \geq -1$$

(CD4)

$$c_\lambda(c_\lambda(a)) = \frac{1-c_\lambda(a)}{1+c_\lambda(a)} = \frac{\lambda a - a}{1+\lambda} = a$$

Si se particulariza para $\lambda=1$ y $U=\{1,2,3,4,5,6,7\}$ y se toma el conjunto difuso (figura 1.19)

$$\tilde{\tilde{A}} = \{(1, 0.3), (3, 0.5), (5, 0.7), (7, 0.3)\}$$

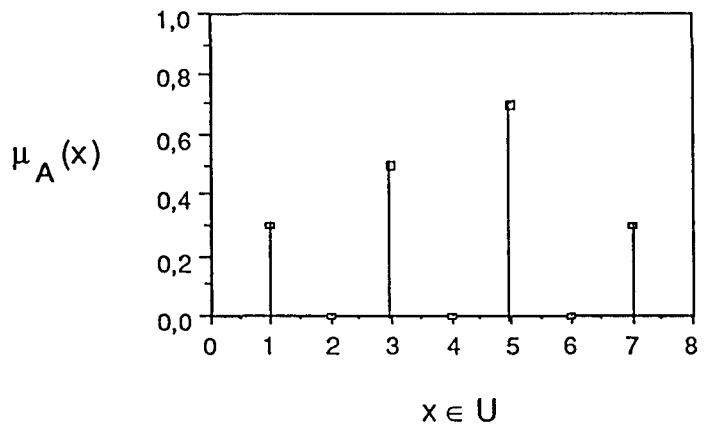


Figura 1.19- Representación del conjunto difuso $\tilde{\tilde{A}}$

entonces el complementario estándar y el asociado para la función c_1 son

$$\overline{\tilde{\tilde{A}}} = \{(1, 0.7), (2, 1), (3, 0.5), (4, 1), (5, 0.3), (6, 1), (7, 0.7)\}$$

$$C_1(\tilde{\tilde{A}}) = \{(1, 7/3), (2, 1), (3, 1/3), (4, 1), (5, 3/17), (6, 1), (7, 7/13)\}$$

que vienen representados en la figura 1.20.

¤

Análogamente se pueden definir la unión e intersección difusa, siendo también válidas las observaciones comentadas anteriormente para la complementación.

Definición 1.22.- *(Unión difusa).-* Una unión de dos conjuntos difusos $\tilde{\tilde{A}}$ y $\tilde{\tilde{B}}$ viene dada por una función u: [0,1] x[0,1]→ [0,1] cumpliendo

(UD1) u(0,0)=0, u(0,1)=u(1,0)=u(1,1)=1

(UD2) u(a, b)=u(b, a) para cada a, b pertenecientes a [0,1]

(UD3) Si a≤a´ y b≤b´ se cumple u(a, b)≤u(a´, b´) siendo a, a´, b, b´ ∈ [0, 1]

(UD4) Para cada a,b,c en [0,1] se tiene u(u(a,b),c)=u(a,u(b,c))

siendo $c(\mu_A(x), \mu_B(x))$ el grado de pertenencia de x en el concepto unión de los representados por $\tilde{\tilde{A}}$ y $\tilde{\tilde{B}}$.

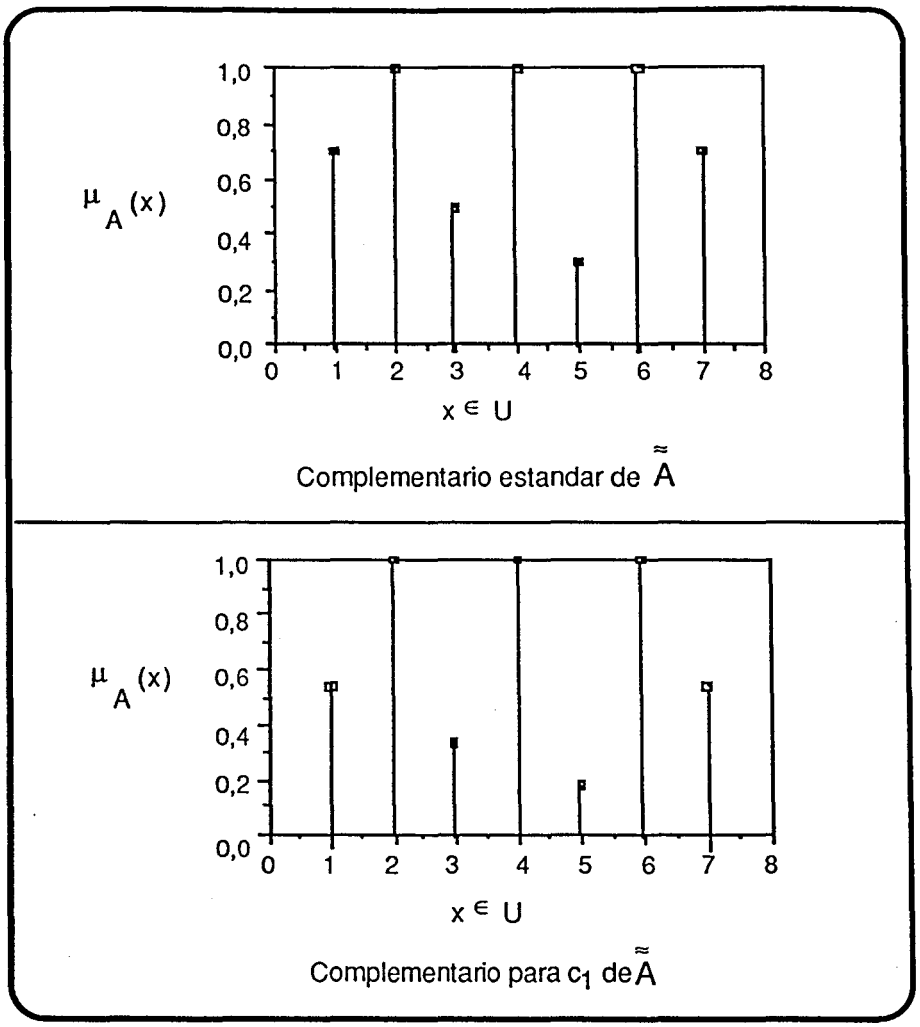


Figura 1.20.- Representación del complementario estandar y $C_1(\tilde{A})$

Es frecuente en la literatura sobre teoría de conjuntos difusos llamar a las intersecciones y uniones difusas "normas triangulares" (t-normas) y conormas triangulares (t-conormas) respectivamente (Schweizer y Sklar (1983)). Estos conceptos fueron introducidos por Menger (1942).

Ejemplo 1.2.- La clase de uniones de Yager (1980,82) es una clase de uniones difusas formada por las funciones

$$u_w(a,b) = \min\left[1,\ (a^w + b^w)^{1/w}\right] \quad \text{con } w \in (0,\infty)$$

que verifica que cuando w tiende a infinito se obtiene la unión estandar de conjuntos difusos.

¤

Teorema 1.4.- Se verifica

$$\lim_{w \to \infty} \min\left[1,\ (a^w + b^w)^{1/w}\right] = \max(a, b)$$

Definición 1.23.- *(Intersección difusa).*- Una intersección de dos conjuntos difusos $\tilde{\tilde{A}}$ y $\tilde{\tilde{B}}$ viene dada por una función i: [0,1] x[0,1]→ [0,1] cumpliendo

(ID1) i(1,1)=1, u(0,1)=u(1,0)=u(0,0)=0

(ID2) i(a, b)=i(b, a) para cada a, b pertenecientes a [0,1]

(ID3) Si a≤a´ y b≤b´ se cumple i(a, b)≤i(a´, b´) siendo a, a´, b, b´ ∈ [0, 1]

(ID4) Para cada a,b,c en [0,1] se tiene i(i(a,b),c)=i(a,i(b,c))

siendo $c(\mu_A(x), \mu_B(x))$ el grado de pertenencia de x en el concepto intersección de los representados por $\tilde{\tilde{A}}$ y $\tilde{\tilde{B}}$.

Una de las clases de funciones que satisfacen las propiedades anteriores es la definida por

$$i_w(a, b) = 1 - \min(1, ((1-a)^w + (1-b)^w)^{1/w}) \quad \text{con } w \in (0, \infty)$$

que recibe el nombre de clase de intersecciones de Yaper. Se puede demostrar fácilmente que

$$i_\infty(a, b) = \min(a, b)$$

El siguiente teorema nos dice que la unión e intersección estandar de conjuntos difusos representa respectivamente la cota inferior y superior de todas las posibles funciones unión e intersección difusa.

Teorema 1.5.- Se verifica

$$\max(a, b) \le u(a, b) \le u_{max}(a, b)$$
$$\min(a, b) \le i(a, b) \le u_{min}(a, b)$$

donde

$$u_{max}(a, b) = \begin{cases} a & \text{si } b = 0 \\ b & \text{si } a = 0 \\ 1 & \text{en otro caso} \end{cases} \qquad u_{min}(a, b) = \begin{cases} a & \text{si } b = 1 \\ b & \text{si } a = 1 \\ 0 & \text{en otro caso} \end{cases}$$

A continuación se definirá una relación binaria entre dos conjuntos difusos $\tilde{\tilde{A}}$ y $\tilde{\tilde{B}}$, de U y V respectivamente, como un conjunto difuso de UxV.

Definición 1.24.- *(Relación binaria).-* Sean $\tilde{\tilde{A}}$ y $\tilde{\tilde{B}}$ dos conjuntos difusos de U y V respectivamente. Se define la relación binaria R(U,V) como el conjunto difuso

$$\{(u, v, \mu_R(u, v)) / u \in U, v \in V\}$$

El valor $\mu_R(u, v)$ debe interpretarse como una medida del grado de relación entre u y v.

Si R(U,V) y Q(V,W) son dos relaciones binarias difusas la composición es otra composición binaria difusa S(U,W) que se denota por

(1. 17) $$S(U, W) = R(U, V) \circ Q(V, W)$$

La composición más utilizada es la max-min que viene dada por

(1. 18) $$\mu_{R \circ Q}(u, w) = \max_{v \in V} \min \{\mu_R(u, v), \mu_Q(v, w)\} \quad \forall u \in U, w \in W$$

1.6.2.2.2.-Medidas difusas

Los conjuntos difusos representan de alguna forma el grado de incertidumbre que se tiene sobre la pertenencia de un elemento de un conjunto universal U en el concepto que define el conjunto difuso. Esto nos va permitir definir a partir de ello las medidas de posibilidad y necesidad.

Supongamos ahora $\tilde{\tilde{A}}$ un conjunto difuso de U con función característica μ_A. Entonces se puede considerar una función de distribución de posibilidad dada por

(1. 19) $$r : U \to [0, 1]$$
$$x \to r(x) = \mu_A(x)$$

que define una función de posibilidad cuyo significado es el grado de posibilidad de que los elementos de un subconjunto de U pertenezcan al concepto representado por el conjunto difuso $\tilde{\tilde{A}}$, es decir,

(1. 20) $$\forall B \subseteq U \quad \pi(B) = \max_{x \in B} (r(x))$$

La necesidad vendrá dada por

(1.21) $$\forall B \subseteq U \quad \eta(B) = \inf_{x \notin B} (1 - r(x))$$

Si r está definido en UxW es una distribución de posibilidad conjunta entonces las distribuciones de posibilidad marginales serán las proyecciones y vendrán dadas mediante

(1.22) $$r_U(x) = \max_{y \in W} (r(x, y)) \qquad \forall x \in U$$

(1.23) $$r_W(y) = \max_{x \in U} (r(x, y)) \qquad \forall y \in W$$

En efecto, nótese que para cada $x \in U$ la distribución marginal está definida para los $\{(x, y) / y \in W\}$ en los que la distribución conjunta está definida. Por tanto,

(1.24) $$\pi_U(\{x\}) = \pi(\{(x, y) / y \in W\})$$

donde π_U y π representan las medidas correspondientes a r_U y r respectivamente y, por (1.20)

$$r_U(x) = \max_{y \in W} (r(x, y))$$

Definición 1.25.- *(Interactivos).-* Se dice que U y W son interactivos en el sentido de posibilidad sii

$$r(x, y) = \min\{r_U(x), r_W(y)\}$$

De (1.22) y (1.23) se deduce que

(1.25) $$r_U(x) \geq r(x, y) \qquad \forall y \in W$$
$$r_W(y) \geq r(x, y) \qquad \forall x \in U$$

es decir,

(1.26) $$r(x, y) \leq \min\{r_U(x), r_W(y)\}$$

1.6.2.3.-Factores de certeza

En este apartado se introduce otro mecanismo para medir el razonamiento impreciso: el factor de certeza (Barr y Feingenbaum (1981), Buchanan y Shortliffe (1984), Cuena (1987), etc). Fue desarrollado por Shortliffe y Buchanan en MYCIN, sistema experto para el diagnóstico de enfermedades infecciosas, para medir la confianza que merece una hipótesis h dada una cierta evidencia e. El factor de

certeza combina en un sólo número el grado de credibilidad y de incredibilidad en una conclusión supuesta una evidencia. Es por ello que antes de dar su definición se necesite definir previamente otras dos medidas.

Se empezará en primer lugar analizando brevemente su significado físico. Se denotará por MC(h,e), y se llamará medida de credibilidad, a aquella que representa el decremento relativo de incredibilidad de una hipótesis h debida a una evidencia o información e. Es decir, si [1-P(h)] es la incredibilidad en la hipótesis h antes de conocer la evidencia e, [1-P(h/e)] es la incredibilidad en la hipótesis h después de conocer la evidencia e, entonces MC(h,e) viene dada por la expresión:

$$(1.27) \qquad MC(h,e) = \frac{[1-P(h)]-[1-P(h/e)]}{1-P(h)} = \frac{P(h/e)-P(h)}{1-P(h)}$$

Nótese que el término [1-P(h)] del denominador lo convierte en un valor relativo en vez de un valor absoluto.

Se denotará por MI(h,e), y se llamará medida de incredibilidad, a la medida que representa el decremento relativo de credibilidad de una hipótesis h debida a una evidencia o información e, es decir,

$$(1.28) \qquad MI(h,e) = \frac{P(h)-P(h/e)}{P(h)}$$

Nótese que, al igual que el caso anterior, P(h) es la credibilidad en la hipótesis h antes de conocer la evidencia e, que P(h/e) es la credibilidad en la hipótesis, h, después de conocer la evidencia, e, y que el término P(h) del denominador lo convierte en un valor relativo en vez de uno absoluto.

Por otro lado, nótese también que sólo se pueden dar uno de los tres casos siguientes:

(a) $P(h/e)<P(h)$. La evidencia e disminuye la credibilidad en h, mientras aumenta su incredibilidad en la verdad de h.
(b) $P(h/e)>P(h)$. La evidencia e aumenta la credibilidad en h, mientras disminuye su incredibilidad en la verdad de h.
(c) $P(h/e)=P(h)$. La evidencia e no influye en la credibilidad o incredibilidad de h.

Los autores de estos conceptos reflejaron los tres casos en términos de las medidas de credibilidad y de incredibilidad de la siguiente manera:

(a) Si $P(h/e)<P(h) \Rightarrow MC(h,e)=0$, $MI(h,e)>0$.
(b) Si $P(h/e)>P(h) \Rightarrow MC(h,e)>0$, $MI(h,e)=0$.
(c) Si $P(h/e)=P(h) \Rightarrow MC(h,e)=0$, $MI(h,e)=0$.

Por todo ello, modificaron esta idea intuitiva limitándola a valores positivos, es decir, dándole a estas medidas el valor cero cuando al aplicar (1.27) y (1.28) se obtiene un valor negativo. La definición formal en términos de probabilidades a priori y condicionales es la siguiente.

Definición 1.26.- *(Medida de credibilidad y de incredibilidad)* :
(MC1) Se define la medida de credibilidad de una hipótesis h debida a una evidencia e como

$$MC(h,e)=\begin{cases} 1 & \text{si } P(h)=1 \\ \dfrac{\max[P(h/e),P(h)]-P(h)}{1-P(h)} & \text{en otro caso} \end{cases}$$

(MI1) Se define la medida de incredibilidad de una hipótesis h debida a una evidencia e como

$$MI(h,e)=\begin{cases} 1 & \text{si } P(h)=0 \\ \dfrac{P(h)-\min[P(h/e),P(h)]}{P(h)} & \text{en otro caso} \end{cases}$$

(CF1) El factor de certeza se define mediante la expresión

$$CF(h,e)=MC(h,e)-MI(h,e)$$

Según estas definiciones, las medidas MC y MI toma valores entre cero y uno, es decir, que $0 \le MC(h,e), MI(h,e) \le 1$, y, por tanto, $-1 \le CF(h,e) \le 1$.

Nótese que cuando P(h) es próximo a cero entonces el factor de certeza es aproximadamente la probabilidad condicionada P(h/e)

$$CF(h,e) \approx \frac{P(h/e)-P(h)}{1-P(h)} - 0 \approx P(h/e)$$

Las propiedades más importantes de estas medidas se destacan en la Proposición 1. 6.

Proposición 1.6.- *(Propiedades de las medidas de credibilidad e incredibilidad)*- Se verifican las siguientes propiedades:

(PC1) si h es cierta supuesto e entonces MC(h,e)=1, MI(h,e)=0 y CF(h,e)=1

(PC2) si h es falsa supuesto e, entonces MC(h,e)=0 MI(h,e)=1 y CF(h,e)=-1

(PC3) $MC(h,e) = MI(\bar{h},e)$

(PC4) $CF(h,e) + CF(\bar{h},e) = 0$

Demostración.-

(PC1) y (PC2).

Si h es cierta o falsa supuesto e se traduce en el valor de P(h/e) que tomará los valores 1 ó 0 respectivamente. Por lo tanto, teniendo en cuenta esta observación, las propiedades (1) y (2) son evidentes.

(PC3).

Si $P(h/e) \leq P(h) \Rightarrow P(\bar{h}) \leq P(\bar{h}/e)$ y en este caso $MC(h,e)=MI(\bar{h},e)=0$.

Si P(h/e)<P(h) entonces

$$MC(h,e) = \frac{P(h/e) - P(h)}{1 - P(h)} = \frac{P(\bar{h}) - P(\bar{h}/e)}{P(\bar{h})} = MI(\bar{h},e)$$

(PC4).

Por (PC3) es evidente.

Las medidas MC, MI y CF no son medidas de probabilidad. No es de extrañar, por tanto, que, dada una evidencia, los factores de certeza correspondientes a una hipótesis y su complementaria no sumen la unidad sino que sean opuestos.

Teorema 1.6.- *(Control de la consistencia)*- Sean $h_1, h_2, ..., h_k$, k hipótesis mutuamente excluyentes. Supongamos también que e es una evidencia a favor de todas las hipótesis, $P(h_i/e)>P(h_i)$ para i=1,...,k.

(1) En estas condiciones se verifica:

$$-k \leq \sum_{i=1}^{k} CF(h_i, e) \leq 1$$

(2) Si se supone además que para cualquier evidencia, e, una de las hipótesis es siempre cierta, es decir, para cada e, $\sum_{i=1}^{k} P(h_i/e) = 1$, entonces

$$\sum_{i=1}^{k} CF(h_i, e) = 1 \Leftrightarrow k = 1 \text{ y } P(h_1/e) = 1$$

Demostración.-

(1).

Se aplicará inducción sobre el número de hipótesis.

* Para k=1 el resultado es trivial.

* Se supone k>1 y que $-k+1 \le \sum_{i=1}^{k-1} CF(h_i, e) \le 1$.

Para la acotación superior,

$$\sum_{i=1}^{k} CF(h_i, e) = \sum_{i=1}^{k} \frac{P(h_i / e) - P(h_i)}{1 - P(h_i)} =$$

$$= \frac{\sum_{i=1}^{k} \left([P(h_i / e) - P(h_i)] \prod_{j=1,\, j \neq i}^{k} (1 - P(h_i)) \right)}{\prod_{i=1}^{k} (1 - P(h_i))} \le \frac{\sum_{i=1}^{k} (P(h_i / e) - P(h_i))}{\prod_{i=1}^{k} (1 - P(h_i))}$$

y aplicando que

$$\prod_{i=1}^{k} (1 - P(h_i)) \ge 1 - \sum_{i=1}^{k} P(h_i)$$

se tiene

$$\frac{\sum_{i=1}^{k} (P(h_i / e) - P(h_i))}{\prod_{i=1}^{k} (1 - P(h_i))} \le \frac{\sum_{i=1}^{k} P(h_i / e) - \sum_{i=1}^{k} P(h_i)}{1 - \sum_{i=1}^{k} P(h_i)} \le 1$$

Para la acotación inferior,

$$-k \le -K + 1 + CF(h_k / e) \le \sum_{i=1}^{k} CF(h_i / e)$$

(2).

* Implicación directa.

Supongamos que k>1. Como

$$\sum_{i=1}^{k} CF(h_i, e) = 1 \quad y \quad CF(h_i, e) > 0 \quad \text{para } i = 1, .., k$$

existe $j \in \{1,...,k\}$ de forma que si se llama $I_j = \{1,...,k\} - \{j\}$ se cumple

$$\sum_{i \in I_j} CF(h_i, e) < 1$$

Por otro lado,

$$\sum_{i \in I_j} \left\{ (P(h_i/e) - P(h_i)) \prod_{r \in I_j - \{i\}} (1 - (P(h_r))) \right\} \leq \sum_{i \in I_j} (P(h_i/e) - P(h_i)) \leq$$

$$\leq \sum_{i \in I_j} P(h_i/e) - \sum_{i \in I_j} P(h_i) = P(h_j) - P(h_j/e) \leq 0$$

que contradice el hecho de que e sea una evidencia a favor de las hipótesis. Por lo tanto, k=1 y $CF(h_1,e)=1$. Es decir, k=1 y $P(h_1/e)=1$.

* La implicación recíproca se obtiene trivialmente sustituyendo en (CF1).

1.6.3.-Propagación de la incertidumbre

Hasta este apartado, los nuevos conceptos utilizados para definir la incertidumbre no han planteado problemas importantes, salvo su más o menos complicada definición o su falta de significado intuitivo. Sin embargo, como ahora se verá, es a la hora de propagar la incertidumbre cuando surgen problemas serios.

Cuando se trata de evidencias simples, el cálculo de las medidas de incertidumbre no presenta problemas, pero ¿qué pasa cuando se trata de combinar varias evidencias simples para dar lugar a una compuesta?, es decir, ¿Cómo mejora la información por el hecho de conocer nuevos datos? En este apartado se analizará este problema y se plantearán los problemas existente.

El problema de la propagación de incertidumbre en el caso de la probabilidad se reduce al cálculo de las probabilidades condicionadas por todas las evidencias o unidades de información y fue ya tratado en el apartado 1.5.2.2. En los siguientes apartados se estudiarán la propagación para el resto de medidas de incertidumbre: factores de certeza, medidas de credibilidad e incredibilidad y medidas de necesidad y posibilidad. Antes de ello se dará un ejemplo que pondrá de manifiesto la invalidez de la hipótesis de independencia para ciertos casos y sobre su utilización indiscriminada, además de los graves errores que pueden cometerse si se prescinde de la estructura probabilística para simplificarla o para componer las evidencias.

Ejemplo 1.3.- Supongamos que un experto sospecha de la presencia del problema E y que según una serie de datos que posee, determina que la probabilidad a priori para E es 0.8. Como el decidirse por E supone cometer un 0.2 de error, decide obtener más información. El problema E se describe a partir de tres síntomas que se denotan por S_1, S_2 y S_3. En la figura 1.21a se representa la

base del conocimiento de la que dispone el experto en relación a las frecuencias de casos que presentan el problema y las combinaciones de síntomas.

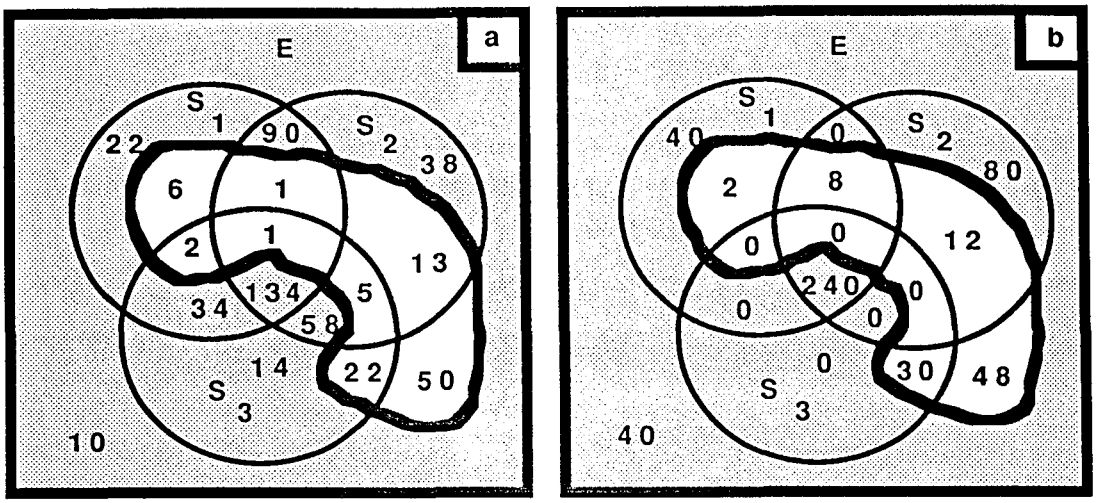


Figura 1.21.- Dos soluciones diferentes con las mismas probabilidades a priori y verosimilitudes

En términos de probabilidades la información de la figura 1.21 puede escribirse

$P(E) = 0.80$	$P(\text{no } E) = 0.20$	$P(S_1 / E) = 0.70$	$P(S_1 / \text{no } E)=0.10$
$P(S_2/ E)=0.80$	$P(S_2/ \text{no } E)=0.20$	$P(S_3/ E)=0.60$	$P(S_3 / \text{no } E)=0.30$

Conviene destacar que esta información no es suficiente para definir completamente la probabilidad. Es decir, existen muchas distribuciones de probabilidad que tienen como valores de las probabilidades a priori y de verosimilitudes las anteriores. En la figura 1.21 se dan dos de ellas.

Supongamos ahora que el experto recibe nueva información en el siguiente orden

1.- Datos iniciales
2.- Presencia de S_1.
3.- Presencia de S_2.
4.- Presencia de S_3.

	$P(E)$	$P(E / S_1)$	$P(E / S_1 \cap S_2)$	$P(E / S_1 \cap S_2 \cap \overline{S_3})$	
				real	independencia
caso a	0.80	0.966	0.994	0.989	0.989
caso b	0.80	0.966	0.968	0.000	0.989

Tabla 1.7.- Actualización de las probabilidades

En la Tabla 1.7 se muestran las probabilidades actualizadas de E despues de conocer la información anterior en el orden indicado en los dos casos de la figura 1.21. La probabilidad $P(E/S_1 \cap S_2 \cap \overline{S_3})$ se ha obtenido por los dos métodos siguientes

- método exacto basado en las frecuencias de la figura.
- utilizando hipótesis de independencia.

Nótese que para el caso (b) el valor real para $P(E/S_1 \cap S_2 \cap \overline{S_3})$ es cero y que suponiendo hipótesis de independencia es 0.989.

¤

1.6.3.1.-Teoría de la evidencia

Dadas dos asignaciones básicas de probabilidad basadas en dos observaciones distintas pero con el mismo marco de discernimiento, Dempster propone unas reglas que las combinan dando lugar a una nueva medida de evidencia que auna los efectos de ambas.

Definición 1.27.- *(Composición de evidencias)-* Dadas m_1 y m_2 dos asignaciones de probabilidad básicas se define su composición

$$(1.29) \qquad m_1 \oplus m_2(A) = \begin{cases} \dfrac{1}{1-K} \displaystyle\sum_{B \cap C = A} m_1(B)\, m_2(C) & \text{si } A \neq \phi \\ 0 & \text{si } A = \phi \end{cases}$$

donde

$$(1.30) \qquad K = \sum_{B \cap C = \phi} m_1(B)\, m_2(C)$$

Proposición 1.7.- En las condiciones anteriores se cumple:

(1) La composición de dos asignaciones de probabilidad básica es también una asignación de probabilidad básica.

(2) La composición de evidencias es conmutativa

$$m_1 \oplus m_2 = m_2 \oplus m_1$$

(3) La composición de evidencias es asociativa

$$m_1 \oplus (m_2 \oplus m_3) = (m_1 \oplus m_2) \oplus m_3$$

La composición de credibilidades y plausibilidades vendrá dada por la fórmula (1.7) y (1.8) donde la asignación de probabilidad básica asociada es $m_1 \oplus m_2$.

Ejemplo 1.4.- Supongamos que se tienen las evidencias m_1, m_2 y m_3 de la Tabla 1.8 que se han obtenido de la figura 1.21b. cuando se conoce la información de presencia de S_1, presencia de S_2 y ausencia de S_3 respectivamente.

$\mathcal{P}(U)$	m_1	m_2	m_3	m_{12}	m_{123}	Cr_{123}	Pl_{123}
ϕ	0.000	0.000	0.000	0.000	0.000	0.000	0.000
E	0.966	0.941	0.695	0.998	0.999	0.999	0.999
$\overline{E}$	0.058	0.058	0.695	0.022	0.001	0.001	0.001
$\{E, \overline{E}\}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Tabla 1.8.- Unidades de evidencia y funciones de credibilidad y plausibilidad asociadas a la composición

Aplicando la fórmula de propagación de evidencias a m_1, m_2 y m_3 se obtienen los resultados de la Tabla 1.8. Nótese que la credibilidad y plausibilidad para E conocidos $S_1 \cap S_2 \cap \overline{S_3}$ es 0.0 que es un resultado muy malo. Esto cuestiona seriamente las fórmulas de propagación.

¤

Son muchas las críticas que se han venido produciendo en los últimos años en contra de la teoría de la evidencia, centradas casi todas ellas en la fórmula arbitraria con la que se propaga la incertidumbre, dando lugar en muchos casos a resultados erroneos. Algunos argumentos en contra de esta teoría se refieren a las limitaciones que supone que la composición se realice sobre el mismo cuerpo de discernimiento y que no permita combinar evidencias sobre marcos de discernimiento distintos pero con restricciones y relaciones entre ellos.

Por otro lado, se restringe a marcos de discernimientos U formados por hipótesis mutuamente exclusivas y exhaustivas, no contemplando aquellos problemas reales en los que esta condición no se verifique.

Desde el punto de vista computacional, este método crece exponencialmente con el tamaño del problema aunque en la actualidad se han

desarrollado algunas teorías para reducir su computación a tiempo polinomial (Barnett (1981)).

1.6.3.2.-Lógica difusa

Vamos a ver ahora como se realiza la inferencia y consiguientemente la propagación de incertidumbre en el caso de que la base del conocimiento contenga un conjunto de reglas inciertas, es decir, reglas de la forma

$$\text{Si } x \text{ es } \tilde{\tilde{A}} \text{ entonces } y \text{ es } \tilde{\tilde{B}}$$

con $\tilde{\tilde{A}}$ y $\tilde{\tilde{B}}$ conjuntos difusos de U y V y con funciones de posibilidad μ_A y μ_B respectivamente.

En lógica clásica una regla se dice que es correcta si cuando las premisas son ciertas lo son también las conclusiones. La generalización para la lógica difusa se traduce en que una regla es correcta cuando la función de posibilidad de las premisas es menor o igual que la de las conclusiones. Es decir,

$$(1.31) \qquad \forall v \in V \qquad \mu_B(v) \geq \mu_{A^\circ R}(u,v) = \mu_{A \to B}(u,v)$$

donde R es una relación binaria difusa representando $\mu_R(u,v)$ el grado en que u influye en v, pudiendo considerarse como una función de posibilidad condicional sobre UxV. El hecho de considerar una desigualdad en (1.31) en lugar de una igualdad se debe a que si B está contenido en B´entonces se debe también de deducir B´.

Si la composición es la max-min (1.31) pude escribirse como

$$(1.32) \qquad \forall v \in V \qquad \mu_B(v) \geq \max_{u \in U} (\min (\mu_A(u), \mu_R(u,v)))$$

Nótese la analogía de (1.32) con (1.3).

1.6.3.2.1-Modus Ponens

Supongamos que se quiere realizar un encadenamieno hacia adelante y se tiene

$$\text{Si } x \text{ es } \tilde{\tilde{A}} \text{ entonces } y \text{ es } \tilde{\tilde{B}}$$

y el hecho

$$x \text{ es } \tilde{\tilde{A}}'$$

con $\tilde{A}$ y $\tilde{A}'$ conjuntos difusos en U y $\tilde{B}'$ de V. Entonces el modus ponens en lógica difusa permite obtener un conjunto difuso $\tilde{B}'$ con la siguiente función de posibilidad

(1.33) $$\forall v \in V \quad \mu_{B'}(v) = \mu_{A' \circ R}(u, v)$$

Análogamente si tomamos la composición max-min (1.33) se transforma en

(1.34) $$\forall v \in V \quad \mu_{B'}(v) = \max_{u \in U} (\min (\mu_{A'}(u), \mu_R(u, v)))$$

1.6.3.2.2-Modus Tollens

El modus tollens en lógica difusa permite obtener un conjunto difuso $\tilde{A}'$ con la siguiente función de posibilidad

(1.35) $$\forall u \in U \quad \mu_{A'}(u) = \mu_{B' \circ R'}(v, u)$$

siendo

(1.36) $$\mu_{R'}(v, u) = \mu_R(\bar{u}, \bar{v})$$

con el fin de que se mantenga la equivalencia entre las reglas

Si x es $\tilde{A}$ entonces y es $\tilde{B}$

Si y es no $\tilde{B}$ entonces y es no $\tilde{A}$

Análogamente si tomamos la composición max-min (1.35) se transforma en

(1.37) $$\forall u \in U \quad \mu_{B'}(u) = \max_{v \in V} (\min (\mu_{B'}(u), \mu_{R'}(v, u)))$$

Ejemplo 1.5.- Supongamos el ejemplo de la figura 1.21. Consideramos los conjuntos

$$U = \{S_1, S_2, S_3\} \quad y \quad V = \{E, \bar{E}\}$$

y los conjuntos difusos

$$\tilde{A} = \{(S_1, 0.58), (S_2, 0.68), (S_3, 0.48)\} \qquad \tilde{B} = \{(E, 0.80), (\bar{E}, 0.20)\}$$

$$\tilde{A}' = \{(S_1, 0.42), (S_2, 0.32), (S_3, 0.54)\}$$

Entonces se define la relación R que nos da el grado de implicación como

$$\mu_R(u,v) = \mu_{(A \cap B) \cup \bar{A}}(u,v) = \max\,(\min\,(\mu_A(u), \mu_B(v)), 1 - (\mu_A(u))$$

que matricialmente se puede escribir

$$R = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \begin{array}{c} \begin{array}{cc} E & \bar{E} \end{array} \\ \begin{pmatrix} 0.58 & 0.42 \\ 0.68 & 0.32 \\ 0.54 & 0.54 \end{pmatrix} \end{array}$$

Supongamos ahora que tenemos un caso cuyo grado de presencia de los síntomas viene dado como el conjunto difuso

$$\tilde{A}' = \{(S_1, 0.95), (S_1, 0.99), (S_3, 0.01)\}$$

El problema que presenta este caso se calcula mediante la fórmula (1.33). Si además consideramos la composición la max-min se tiene el siguiente resultado

$$(E \;\; \bar{E}) = (0.95 \quad 0.99 \quad 0.01) \begin{pmatrix} 0.58 & 0.42 \\ 0.68 & 0.32 \\ 0.54 & 0.54 \end{pmatrix} = (0.68 \quad 0.42)$$

donde el producto de estas matrices debe realizarse sustituyendo en la fórmula general del producto la suma de dos números por el máximo y el producto por el mínimo, obteniéndose igual resultado que en (1.34). Es decir, se concluye el conjunto

$$\tilde{B}' = \{(E, 0.68), (\bar{E}, 0.42)\}$$

¤

1.6.3.3.-Factores de certeza

La propagación de las medidas de credibilidad, MC, e incredibilidad, MI, y de los factores de certeza, CF, suele hacerse en los tres casos representados en la Tabla 1.9 mediante las fórmulas (Heckerman (1986))

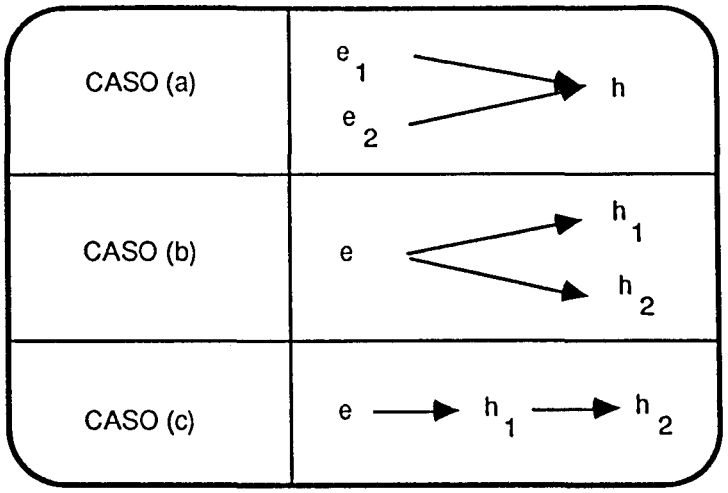


Tabla 1.9.- Representación de los casos de combinación de los factores de certeza

(a) Combinación de evidencias sobre una misma hipótesis, h,

$$(1.38)\quad MC(h, e_1 \cap e_2) = \begin{cases} 0 & \text{si } MI(h, e_1 \cap e_2) = 1 \\ MC(h, e_1) + MC(h, e_2)[1 - MC(h, e_1)] & \text{en otro caso} \end{cases}$$

$$(1.39)\quad MI(h, e_1 \cap e_2) = \begin{cases} 0 & \text{si } MC(h, e_1 \cap e_2) = 1 \\ MI(h, e_1) + MI(h, e_2)[1 - MI(h, e_1)] & \text{en otro caso} \end{cases}$$

Nótese que

$$MC(h, e_1 \cap e_2) = 1 \Leftrightarrow MC(h, e_1) = 1 \text{ ó } MC(h, e_2) = 1$$
$$MI(h, e_1 \cap e_2) = 1 \Leftrightarrow MI(h, e_1) = 1 \text{ ó } MI(h, e_2) = 1$$

y, por tanto, si llamamos $x=CF(h,e_1)$ e $y=CF(h,e_2)$ entonces

$$(1.39)\quad CF(h, e_1 \cap e_2) = \begin{cases} x + y - xy & \text{si } x, y \geq 0 \\ \dfrac{x + y}{1 - \min(|x|, |y|)} & \text{si } x, y \text{ tienen dist int o signo} \\ x + y + xy & \text{si } x, y < 0 \end{cases}$$

(b) Combinación de hipótesis con una misma evidencia, e.

(1. 40) $$MC(h_1 \cap h_2, e) = \min(MC(h_1, e), MC(h_2, e))$$

(1. 41) $$MI(h_1 \cap h_2, e) = \max(MI(h_1, e), MI(h_2, e))$$

(1. 42) $$MC(h_1 \cup h_2, e) = \max(MC(h_1, e), MC(h_2, e))$$

(1. 43) $$MI(h_1 \cup h_2, e) = \min(MI(h_1, e), MI(h_2, e))$$

(c) En los sistemas expertos basados en reglas es frecuente que en la búsqueda de un objetivo, el objeto de una premisa de una regla aparezca en la conclusión de otra regla distinta, y, por ello, el motor de inferencia encadene unas reglas con otras. Esto pone de manifiesto la necesidad de encadenar también los factores de certeza asociados.

Si se supone que e implica h_1 y que h_1 implica h_2 entonces Shortliffe y Buchanan proponen las siguientes fórmulas para calcular $MC(h_2,e)$ y $MI(h_2,e)$

(1. 44) $$MC(h_2, e) = MC(h_2, h_1)\ \max(0, CF(h_1, e))$$
$$MI(h_2, e) = MI(h_2, h_1)\ \max(0, CF(h_1, e))$$

y, por tanto,

(1. 45) $$CF(h_2, e) = \begin{cases} CF(h_1, e) * CF(h_2, h_1) & \text{si } CF(h_1, e) \geq 0 \\ -CF(h_1, e) * CF(h_2, \overline{h_1}) & \text{si } CF(h_1, e) < 0 \end{cases}$$

Ejemplo 1.6.- Con objeto de ilustrar algunos de los problemas asociados con las fórmulas de propagación se muestra en la Tabla 1.10. los valores exactos según las definiciones de MC, MI, y CF y utilizando las fórmulas de propagación de los dos casos del Ejemplo 1.3. Nótese la diferencia entre el valor exacto y el valor de certeza calculado para en el caso b. Este sorprendente resultado prueba que estas fórmulas de propagación no son satisfactorias en este caso, y nos previene sobre la utilización incontrolada de las expresiones anteriores.

	$CF(E; S_1)$	$CF(E; S_1, S_2)$	$CF(E; S_1, S_2, \text{no } S_3)$	$CF(E; S_1, S_2, \text{no } S_3)$ Fórmulas propagación
caso a	0.873	0.970	0.945	0.989
caso b	0.828	0.839	-1.000	0.972

Tabla 1.10.- Actualización de los factores de certeza

¤

1.7.-MODELO DE REDES CAUSALES

En este apartado se describe una versión modificada del modelo de Lauritzen y Spiegelhalter (1988), que es uno de los métodos basados en la idea de redes causales. La idea de Lauritzen y Spiegelhalter consiste en utilizar una estructura probabilística tal que la propagación de probabilidades sea exacta, rápida y no cause problemas de excesivo número de parámetros. Por ello, suponen que el conocimiento puede ser representado mediante lo que ellos llaman un "diagrama de influencias". El diagrama de influencias no es otra cosa que un grafo dirigido, es decir, un conjunto, V, de nodos y un conjunto de aristas orientadas entre pares de nodos. Una arista orientada entre los nodos "A" y "B" puede representarse mediante la notación $A \rightarrow B$ y entonces se dice que el nodo "A" es un padre y que el nodo "B" es un hijo.

En el caso de que las aristas del grafo anterior sean no orientadas (no importa el orden de los dos nodos que definen la arista) se tiene un grafo no dirigido. Un conjunto de nodos $\mathbf{C}$ se dice que es "*completo*" si existen aristas entre todos sus pares de nodos y se dice que forma un "*aglomerado*" si es maximal, es decir, si no puede ampliarse a otro conjunto completo. Los aglomerados serán en lo sucesivo designados como $(C_1, C_2, \ldots, C_q)$ y el conjunto de ellos como Δ. Al conjunto de nodos extremos de aristas de uno dado "A" se le llama "*frontera de A*" y se le designa como Fr(A).

Una numeración de nodos se dice "*perfecta*" si todo conjunto de nodos $Fr(i) \cap \{1,2,\ldots,i-1\}$ es completo. A todo grafo que admite una numeración perfecta se le llama "triangulado". Lauritzen y Spiegelhalter transforman el diagrama de influencias en un grafo no dirigido y triangulado, de forma que si el diagrama de influencias original no conduce a este tipo de grafo, se completa con nuevas aristas hasta conseguir que lo sea.

Los nodos del grafo representan los objetos, que pueden tomar un conjunto finito de valores. Como datos de partida se dan unas "tablas de probabilidades condicionadas". Estas tablas contienen las probabilidades de que cada nodo tome cada uno de sus posibles valores para todas las combinaciones posibles de los valores de sus padres. Además se supone que el conjunto de padres, Π_A, de un nodo A tiene toda la información sobre él, es decir, que:

$$(1.46) \qquad P(A / B, U) = P(A / B) \qquad \forall A, U \subset V \; ; \; B \subset \Pi_A \quad U \cap \Pi_A = \phi$$

A esta hipótesis se la llama hipótesis de Markov o de "dependencia a través de los padres".

Esto implica que la función de probabilidad conjunta, P(V), de todos los nodos puede ponerse como el producto de las probabilidades condicionadas (ver Ejemplo 1.7), sin más que aplicar la conocida fórmula

(1.47) $P(A_1,A_2,...,A_n)=P(A_1)P(A_2/A_1)P(A_3/A_1\cap A_2)...P(A_n/A_1\cap A_2\cap...\cap A_{n-1})$

y tener en cuenta la hipótesis de Markov o de dependencia a través de los padres.

El objetivo perseguido por Lauritzen y Spiegelhalter (1988) no es otro que conseguir algoritmos rápidos para calcular probabilidades condicionadas y marginales a partir de la conjunta. Nótese que el método directo es muy lento y requiere mucho tiempo de ordenador.

A veces resulta más fácil utilizar una representación de la función de probabilidad conjunta, P(V), diferente de la indicada anteriormente como producto de probabilidades condicionadas. Esta nueva forma se consigue mediante producto de funciones de los nodos en los aglomerados (ver Ejemplo 1.7). En este caso, ni siquiera es necesario conocer la función de probabilidad conjunta y basta una función proporcional a ella, ya que, al tratarse de una probabilidad, siempre puede normalizarse. A estas funciones definidas en los aglomerados se les llama "potenciales de evidencia", que se designarán con la letra ψ. De esta forma se tiene:

(1.48) $$P(V)=\prod_{i=1}^{q}\psi(C_i)/Z$$

donde Z es una constante de normalización.

A partir de los potenciales de evidencia (Δ,ψ) resulta fácil obtener las funciones de probabilidad marginales de un conjunto de nodos dado $U\subset V$. A este proceso se le llama marginalización sobre $\bar{U}$. Para definir esta probabilidad conjunta (de los nodos de U) se utilizará una representación en potenciales de evidencia $(\bar{\Delta},\bar{\psi})$.

Sea $\Delta_1=\{A\in\Delta / A\cap\bar{U}=\phi\}$ y $\Delta_2=\Delta-\Delta_1$, entonces se tiene

(1.49) $$P(U)=\sum_U P(U,\bar{U})=\sum_U Z^{-1}\prod_{A\in\Delta}\psi(A)=Z^{-1}\prod_{A\in\Delta_1}\psi(A)\sum_U\prod_{A\in\Delta_2}\psi(A)$$

$$=Z^{-1}\phi(B)\prod_{A\in\Delta_1}\psi(A)$$

donde

$$B = \bigcup_{A \in \Delta_2} A - \overline{U} \tag{1.50}$$

$$\phi(B) = \sum_{\overline{U}} \prod_{A \in \Delta_2} \psi(A) \tag{1.51}$$

De (1.49) se deduce que el nuevo conjunto de aglomerados y los nuevos potenciales de evidencia para el nuevo conjunto de nodos U son:

$$\left.\begin{array}{l} \overline{\Delta} = \Delta_1 \cup B \\ \overline{\psi}(A) = \left\{\begin{array}{ll} \phi(B) & \text{si } A = B \\ \psi(A) & \text{en otro caso} \end{array}\right\} \end{array}\right\} \text{ si } B \not\subset A \; ; \forall A \in \Delta_1$$

(1.52)

$$\left.\begin{array}{l} \overline{\Delta} = \Delta_1 \\ \overline{\psi}(A) = \left\{\begin{array}{ll} \psi(A)\phi(B) & \text{si } A = A_1 \\ \psi(A) & \text{en otro caso} \end{array}\right\} \end{array}\right\} \text{ en otro caso}$$

donde A_1 es un elemento de Δ_1 tal que $B \subset A_1$, y además que la constante de normalización Z no varía en este proceso.

Con objeto de obtener de forma más sencilla las distribuciones de probabilidad marginales de los aglomerados y de los nodos, Lauritzen y Spiegelhalter (1988) utilizan una tercera representación de la función de probabilidad conjunta de todos los nodos mediante una "cadena de conjuntos" (aglomerados) que gozan de la propiedad de que todos los nodos de un aglomerado contenidos en los aglomerados anteriores pertenecen a un único aglomerado anterior. Esta propiedad permite obtener las probabilidades conjuntas de los aglomerados muy fácilmente. En efecto, la cadena anterior es tal que se tiene:

$$P(V) = \prod_{i=1}^{q} P(R_i / S_i) \tag{1.53}$$

donde

$$S_i = C_i \cap (C_1 \cup C_2 \cup \ldots \cup C_{i-1}) \tag{1.54}$$

(1.55) $$R_i = C_i - S_i$$

Los conjuntos S_i y R_i suelen llamarse separadores y residuos de los aglomerados, respectivamente.

Además, teniendo en cuenta que $V = \{ R_q\, C_1\, C_2 \ldots C_{q-1} \}$ y que R_q es independiente de $\{ C_1\, C_2 \ldots C_{q-1} \} - S_q$, para el último aglomerado se tiene:

(1.56) $$P(R_q / S_q) = P(R_q / C_1\, C_2 \ldots C_{q-1}) = P(V) / P(C_1\, C_2 \ldots C_{q-1}) =$$

$$= \frac{Z^{-1} \prod_{A \in \{C_1\, C_2 \ldots C_q\}} \psi(A)}{(Z^{-1} \prod_{A \in \{C_1\, C_2 \ldots C_{q-1}\}} \psi(A) \sum_{R_q} \prod_{A \in C_q} \psi(A)} =$$

$$= \psi(C_q) / \sum_{R_q} \psi(C_q)$$

por lo que mediante marginalización progresiva y la fórmula (1.56) pueden obtenerse todas las probabilidades $P(R_i/S_i)$ que aparecen en (1.53). En efecto, $P(R_q/S_q)$ se obtiene directamente de (1.56). A continuación se marginaliza sobre C_q, utilizando las fórmulas (1.52) y se calcula $P(R_{q-1}/S_{q-1})$, aplicando (1.56) otra vez. Luego se marginaliza sobre C_{q-1} y se calcula $P(R_{q-2}/S_{q-2})$, repitiéndose el proceso hasta obtener $P(R_1/S_1)$.

Si se quiere propagar una evidencia, es decir, si se conocen los valores de los nodos de un determinado conjunto T y se desea conocer cómo se altera la función de probabilidad conjunta P(V) por esa información, se utilizan los potenciales de evidencia, modificándolos de la forma siguiente: la nueva representación (Δ^*, ψ^*), definida en V^*, donde $V^* = V\text{-}T$, Δ^* es la nueva lista de aglomerados y ψ^* son los nuevos potenciales de evidencia, se obtiene como sigue: para cada conjunto A de la lista Δ tal que $A \cap T \neq \Phi$ se incluye A - T en Δ^* y se hace

(1.57) $$\psi^*_{A-T}(.) = \psi_A(.) \mid_T$$

donde $\psi_A(.) \mid_T$ indica la sustitución de los valores de los nodos de T en $\psi_A(.)$.

El principal uso que hacen Lauritzen y Spiegelhalter (1988) de las diferentes representaciones de la probabilidad es como sigue:

(a) tablas de probabilidades condicionadas: inicialmente, para facilitar la asignación del experto
(b) potenciales de evidencia: para propagar evidencia y obtener distribuciones marginales
(c) cadena de conjuntos: para obtener distribuciones marginales de aglomerados o nodos.

Lauritzen y Spiegelhalter (1988) dan también fórmulas sencillas para cambiar de representación, de forma que, dependiendo del objetivo perseguido, pueda pasarse de unas a otras.

Inicialmente, las tablas de probabilidades condicionadas se obtienen del experto o de la base de datos y los potenciales de evidencia iniciales se calculan a partir de ellas (ver Ejemplo 1.7).

Ejemplo 1.7.- Con objeto de ilustrar el método se aplica éste, a continuación, al caso del conjunto de reglas de la figura 1.9. En la figura 1.22a se muestra el diagrama de influencias correspondiente al conjunto de reglas encadenadas.

Como ejemplos de padres e hijos, los nodos A, B, D, E, F, K y L no tienen padres, los padres del nodo G son D, E y F y sus hijos son los nodos H e I.

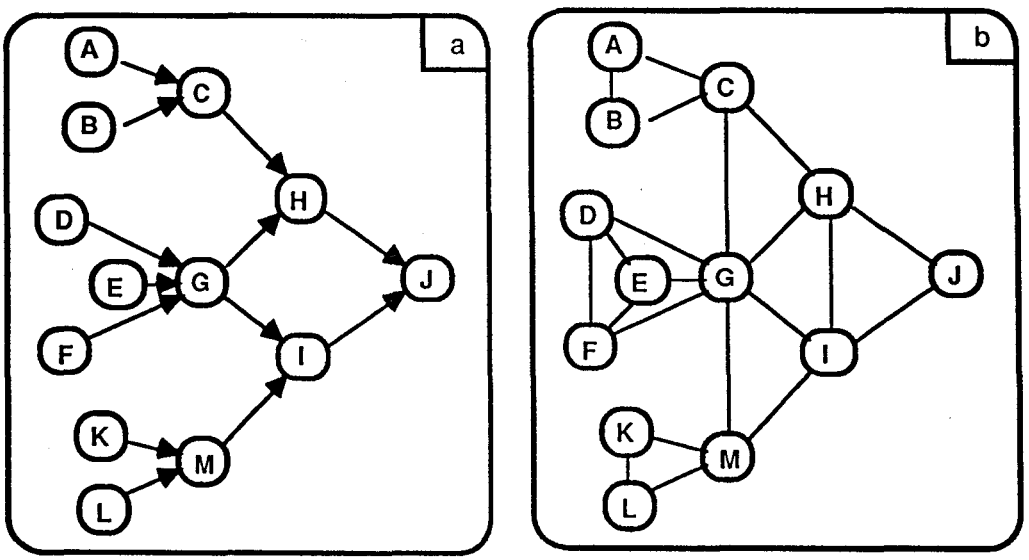


Figura 1.22.- Diagrama de influencias y grafo no dirigido y triangulado correspondiente

La figura 1.22b muestra el grafo no dirigido y triangulado que corresponde al diagrama de influencias anterior. Las aristas no orientadas (A,B), (D,E), (D,F), (E,F) y (K,L) se han incluido para tener en cuenta que los conjuntos {A,B,C}, {D,E,F,G} y {K,L,M} resultan de premisas y conclusiones de 3 reglas. Nótese que su omisión indicaría 7 reglas (con una sola premisa) en vez de las tres reglas asociadas a los conjuntos anteriores. La misma razón justifica las aristas (C,G), (G,M) y (H,I).

Los conjuntos de nodos y de aristas de este último grafo son, por tanto:

V={ A,B,C,D,E,F,G,H,I,J,K,L,M }
y
{ AC,AB,BC,CH,HJ,CG,DG,DE,DF,EF,EG,FG,GH,GI,GM,IJ,HI,KM,KL,LM,MI }

respectivamente.

Aunque los conjuntos {A,B,C}, {D,E,G} y {E,F,G} son completos, sólo el primero de ellos es un aglomerado, ya que los otros dos no son maximales. Nótese que pueden ampliarse al conjunto completo {D,E,F,G}. El conjunto frontera de G, que se denota por Fr(G) es {D,E,F,C,M,H,I}.

Si se supone que los nodos son binarios, es decir, que toman sólo dos valores cada uno (cierto y falso, sí o no, blanco o negro, etc.) resulta la tabla de probabilidades condicionadas siguiente:

(1.58)

P(A,B)
P(D,E,F)
P(K,L)
P(C/A,B)
P(G/D,E,F)
P(M/K,L)
P(H/C,G)
P(I/G,M)
P(J/H,I)

Con ello, la función de probabilidad conjunta de todos los nodos puede escribirse

(1.59) $$P(V) = P(A,B,C,D,E,F,G,H,I,J,K,L,M) =$$
$$= P(A,B)P(C/A,B)P(D,E,F)P(G/D,E,F)P(K,L)P(M/K,L)P(H/C,G)P(I/G,M)P(J/H,I)$$

Esta expresión se obtiene de aplicar (1.47) de la siguiente forma

$$P(V)=P(A,B)xP(C/A,B)xP(D,E,F/A,B,C)xP(G/A,B,C,D,E,F)x$$
$$xP(K,L/A,B,C,D,E,F,G)xP(M/A,B,C,D,E,F,G,K,L)x$$
$$xP(H/A,B,C,D,E,F,G,K,L,M)xP(I/A,B,C,D,E,F,G,K,L,M,H)x$$
$$xP(J/A,B,C,D,E,F,G,K,L,M,H,I)$$

y teniendo en cuenta la hipótesis de Markov o de dependencia a través de los padres, resulta (1.59).

Puesto que el conjunto de aglomerados resultante es

$\Delta = \{(A,B,C), (D,E,F,G), (K,L,M), (C,G,H), (G,I,M), (G,H,I), (H,I,J)\}$,

la función de probabilidad conjunta en función de los potenciales de evidencia es:

(1.60) $$P(V) = P(A,B,C,D,E,F,G,H,I,J,K,L,M) =$$
$$= \psi(A,B,C)\, \psi(D,E,F,G)\, \psi(K,L,M)\, \psi(C,G,H)\, \psi(G,I,M)\, \psi(G,H,I)\, \psi(H,I,J)$$

Inicialmente, entre las muchas posibilidades existentes, según se deduce de (1.59), puede elegirse

(1.61)
$$\psi(A,B,C) = P(A,B)\, P(C / A,B)$$
$$\psi(D,E,F,G) = P(D,E,F)\, P(G / D,E,F)$$
$$\psi(K,L,M) = P(K,L)\, P(M / K,L)$$
$$\psi(C,G,H) = P(H / C,G)$$
$$\psi(G,I,M) = P(I / G,M)$$
$$\psi(G,H,I) = 1$$
$$\psi(H,I,J) = P(J / H,I)$$

Una numeración perfecta de nodos es la de la figura 1.23. De ella se obtiene la Tabla 1.11 en la que se muestran los aglomerados, con sus separadores y residuos. Así resulta la representación en cadena de conjuntos siguiente:

(1.62) $$P(V) = P(A,B,C,D,E,F,G,H,I,J,K,L,M) =$$
$$= P(A,B,C)\, P(G,H / C)\, P(I / G,H)\, P(J / H,I)\, P(M / G,I)\, P(D,E,F / G)\, P(K,L / M)$$

Esta expresión se basa también en la fórmula (1.47), que en este caso conduce a

$$P(V) = P(A,B,C)\, P(G,H / A,B,C)\, P(I / A,B,C,G,H)\, P(J / A,B,C,G,H,I)\, x$$
$$xP(M/A,B,C,G,H,I,J)P(D,E,F/A,B,C,G,H,I,J,M)P(K,L/ A,B,C,G,H,I,J,M,D,E,F)$$

y teniendo en cuenta, una vez más, la hipótesis de Markov o de dependencia a través de los padres, resulta (1.62).

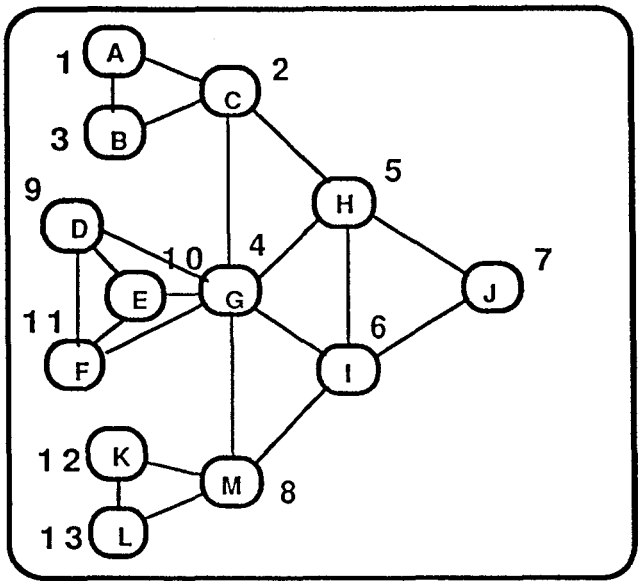


Figura 1.23.- Numeración perfecta de nodos

número i	aglomerado C_i	residuo R_i	separador S_i
1	ABC	ABC	F
2	CGH	GH	C
3	GHI	I	GH
4	HIJ	J	HI
5	GIM	M	GI
6	DEFG	DEF	G
7	KLM	KL	M

Tabla 1.11.- Descomposición en cadena

Como ejemplo, considérese la tabla de probabilidades condicionadas (sólo se dan las probabilidades condicionadas del valor "cierto", pues las del valor "falso" son los complementos a uno, ya que se trata de nodos binarios (dos valores posibles solamente)):

$P(a, b) = 0.4$	$P(d, e, f) = 0.2$	$P(c / a, b) = 0.3$	$P(g / d, e, f) = 0.5$
$P(a, \bar{b}) = 0.3$	$P(d, e, \bar{f}) = 0.2$	$P(c / a, \bar{b}) = 0.5$	$P(g / d, e, \bar{f}) = 0.8$
$P(\bar{a}, b) = 0.2$	$P(d, \bar{e}, f) = 0.1$	$P(c / \bar{a}, b) = 0.4$	$P(g / d, \bar{e}, f) = 0.9$
$P(\bar{a}, \bar{b}) = 0.1$	$P(d, \bar{e}, \bar{f}) = 0.1$	$P(c / \bar{a}, \bar{b}) = 0.8$	$P(g / \bar{d}, e, f) = 0.6$
$P(k, l) = 0.3$	$P(\bar{d}, e, f) = 0.1$	$P(m / k, l) = 0.4$	$P(g / d, \bar{e}, \bar{f}) = 0.2$
$P(k, \bar{l}) = 0.4$	$P(\bar{d}, e, \bar{f}) = 0.1$	$P(m / k, \bar{l}) = 0.8$	$P(g / \bar{d}, e, \bar{f}) = 0.1$
$P(\bar{k}, l) = 0.2$	$P(\bar{d}, \bar{e}, f) = 0.1$	$P(m / \bar{k}, l) = 0.6$	$P(g / \bar{d}, \bar{e}, f) = 0.3$
$P(\bar{k}, \bar{l}) = 0.1$	$P(\bar{d}, \bar{e}, \bar{f}) = 0.1$	$P(m / \bar{k}, \bar{l}) = 0.3$	$P(g / \bar{d}, \bar{e}, \bar{f}) = 0.4$

$P(h / c, g) = 0.2$	$P(h / \bar{c}, \bar{g}) = 0.7$	$P(i / \bar{g}, m) = 0.4$	$P(j / h, i) = 0.8$
$P(h / c, \bar{g}) = 0.1$	$P(i / g, m) = 0.3$	$P(i / \bar{g}, \bar{m}) = 0.8$	$P(j / \bar{h}, i) = 0.6$
$P(h / \bar{c}, g) = 0.8$	$P(i / g, \bar{m}) = 0.5$	$P(j / h, i) = 0.9$	$P(j / \bar{h}, \bar{i}) = 0.7$

donde "a" representa A = cierto y "$\bar{a}$" , A = falso y se utiliza análoga notación para los demás nodos.

A partir de las probabilidades anteriores pueden calcularse los potenciales de evidencia iniciales siguientes

$\psi(a, b, c) = 0.12$	$\psi(d, e, f, g) = 0.10$	$\psi(k, l, m) = 0.12$	$\psi(g, i, m) = 0.3$
$\psi(a, b, \bar{c}) = 0.28$	$\psi(d, e, f, \bar{g}) = 0.10$	$\psi(k, l, \bar{m}) = 0.18$	$\psi(g, i, \bar{m}) = 0.5$
$\psi(a, \bar{b}, c) = 0.15$	$\psi(d, e, \bar{f}, g) = 0.16$	$\psi(k, \bar{l}, m) = 0.32$	$\psi(g, \bar{i}, m) = 0.7$
$\psi(a, \bar{b}, \bar{c}) = 0.15$	$\psi(d, e, \bar{f}, \bar{g}) = 0.04$	$\psi(k, \bar{l}, \bar{m}) = 0.08$	$\psi(g, \bar{i}, \bar{m}) = 0.5$
$\psi(\bar{a}, b, c) = 0.08$	$\psi(d, \bar{e}, f, g) = 0.09$	$\psi(\bar{k}, l, m) = 0.12$	$\psi(\bar{g}, i, m) = 0.4$
$\psi(\bar{a}, b, \bar{c}) = 0.12$	$\psi(d, \bar{e}, f, \bar{g}) = 0.01$	$\psi(\bar{k}, l, \bar{m}) = 0.08$	$\psi(\bar{g}, i, \bar{m}) = 0.8$
$\psi(\bar{a}, \bar{b}, c) = 0.08$	$\psi(d, \bar{e}, \bar{f}, g) = 0.02$	$\psi(\bar{k}, \bar{l}, m) = 0.03$	$\psi(\bar{g}, \bar{i}, m) = 0.6$
$\psi(\bar{a}, \bar{b}, \bar{c}) = 0.02$	$\psi(d, \bar{e}, \bar{f}, \bar{g}) = 0.08$	$\psi(\bar{k}, \bar{l}, \bar{m}) = 0.07$	$\psi(\bar{g}, \bar{i}, \bar{m}) = 0.2$

$\psi(c, g, h) = 0.2$	$\psi(d, e, f, g) = 0.06$	$\psi(h, i, j) = 0.9$	$\psi(g, h, i) = 1.0$
$\psi(c, g, \bar{h}) = 0.8$	$\psi(\bar{d}, e, f, \bar{g}) = 0.04$	$\psi(h, i, \bar{j}) = 0.1$	$\psi(g, h, \bar{i}) = 1.0$
$\psi(c, \bar{g}, h) = 0.1$	$\psi(\bar{d}, e, \bar{f}, g) = 0.01$	$\psi(h, \bar{i}, j) = 0.8$	$\psi(g, \bar{h}, i) = 1.0$
$\psi(c, \bar{g}, \bar{h}) = 0.9$	$\psi(\bar{d}, e, \bar{f}, \bar{g}) = 0.09$	$\psi(h, \bar{i}, \bar{j}) = 0.2$	$\psi(g, \bar{h}, \bar{i}) = 1.0$
$\psi(\bar{c}, g, h) = 0.8$	$\psi(\bar{d}, \bar{e}, f, g) = 0.03$	$\psi(\bar{h}, i, j) = 0.6$	$\psi(\bar{g}, h, i) = 1.0$
$\psi(\bar{c}, g, \bar{h}) = 0.2$	$\psi(\bar{d}, \bar{e}, f, \bar{g}) = 0.07$	$\psi(\bar{h}, i, \bar{j}) = 0.4$	$\psi(\bar{g}, h, \bar{i}) = 1.0$
$\psi(\bar{c}, \bar{g}, h) = 0.7$	$\psi(\bar{d}, \bar{e}, \bar{f}, g) = 0.04$	$\psi(\bar{h}, \bar{i}, j) = 0.7$	$\psi(\bar{g}, \bar{h}, i) = 1.0$
$\psi(\bar{c}, \bar{g}, \bar{h}) = 0.3$	$\psi(\bar{d}, \bar{e}, \bar{f}, \bar{g}) = 0.06$	$\psi(\bar{h}, \bar{i}, \bar{j}) = 0.3$	$\psi(\bar{g}, \bar{h}, \bar{i}) = 1.0$

a partir de los cuales pueden obtenerse las probabilidades de los términos que figuran en (1.62), mediante el procedimiento basado en (1.56). Éstos resultan ser

$P(a, b, c) = 0.12$ $P(g, h / c) = 0.102$ $P(d, e, f / g) = 0.196$ $P(i / g, h) = 0.382$
$P(a, b, \bar{c}) = 0.28$ $P(g, h / \bar{c}) = 0.408$ $P(d, e, f / \bar{g}) = 0.204$ $P(i / g, \bar{h}) = 0.382$
$P(a, \bar{b}, c) = 0.15$ $P(g, \bar{h} / c) = 0.408$ $P(d, e, \bar{f} / g) = 0.314$ $P(i / \bar{g}, h) = 0.564$
$P(a, \bar{b}, \bar{c}) = 0.15$ $P(g, \bar{h} / \bar{c}) = 0.102$ $P(d, e, \bar{f} / \bar{g}) = 0.082$ $P(i / \bar{g}, \bar{h}) = 0.564$
$P(\bar{a}, b, c) = 0.08$ $P(\bar{g}, h / c) = 0.049$ $P(d, \bar{e}, f / g) = 0.176$ $P(i / g, h) = 0.618$
$P(\bar{a}, b, \bar{c}) = 0.12$ $P(\bar{g}, h / \bar{c}) = 0.343$ $P(d, \bar{e}, f / \bar{g}) = 0.020$ $P(i / g, \bar{h}) = 0.618$
$P(\bar{a}, \bar{b}, c) = 0.08$ $P(\bar{g}, \bar{h} / c) = 0.441$ $P(d, \bar{e}, \bar{f} / g) = 0.039$ $P(i / \bar{g}, h) = 0.436$
$P(\bar{a}, \bar{b}, \bar{c}) = 0.02$ $P(\bar{g}, \bar{h} / \bar{c}) = 0.147$ $P(d, \bar{e}, \bar{f} / \bar{g}) = 0.163$ $P(i / \bar{g}, \bar{h}) = 0.436$

$P(j / h, i) = 0.9$ $P(m / g, i) = 0.463$ $P(\bar{d}, e, f / g) = 0.117$ $P(k, l / m) = 0.203$
$P(j / h, \bar{i}) = 0.8$ $P(m / g, \bar{i}) = 0.668$ $P(\bar{d}, e, f / \bar{g}) = 0.082$ $P(k, l / \bar{m}) = 0.439$
$P(j / \bar{h}, i) = 0.6$ $P(m / \bar{g}, i) = 0.418$ $P(\bar{d}, e, \bar{f} / g) = 0.020$ $P(k, \bar{l} / m) = 0.542$
$P(j / \bar{h}, \bar{i}) = 0.7$ $P(m / \bar{g}, \bar{i}) = 0.812$ $P(\bar{d}, e, \bar{f} / \bar{g}) = 0.184$ $P(k, \bar{l} / \bar{m}) = 0.195$
$P(\bar{j} / h, i) = 0.1$ $P(\bar{m} / g, i) = 0.537$ $P(\bar{d}, \bar{e}, f / g) = 0.059$ $P(\bar{k}, l / m) = 0.203$
$P(\bar{j} / h, \bar{i}) = 0.2$ $P(\bar{m} / g, \bar{i}) = 0.332$ $P(\bar{d}, \bar{e}, f / \bar{g}) = 0.143$ $P(\bar{k}, l / \bar{m}) = 0.195$
$P(\bar{j} / \bar{h}, i) = 0.4$ $P(\bar{m} / \bar{g}, i) = 0.582$ $P(\bar{d}, \bar{e}, \bar{f} / g) = 0.078$ $P(\bar{k}, \bar{l} / m) = 0.050$
$P(\bar{j} / \bar{h}, \bar{i}) = 0.3$ $P(\bar{m} / \bar{g}, \bar{i}) = 0.188$ $P(\bar{d}, \bar{e}, \bar{f} / \bar{g}) = 0.122$ $P(\bar{k}, \bar{l} / \bar{m}) = 0.170$

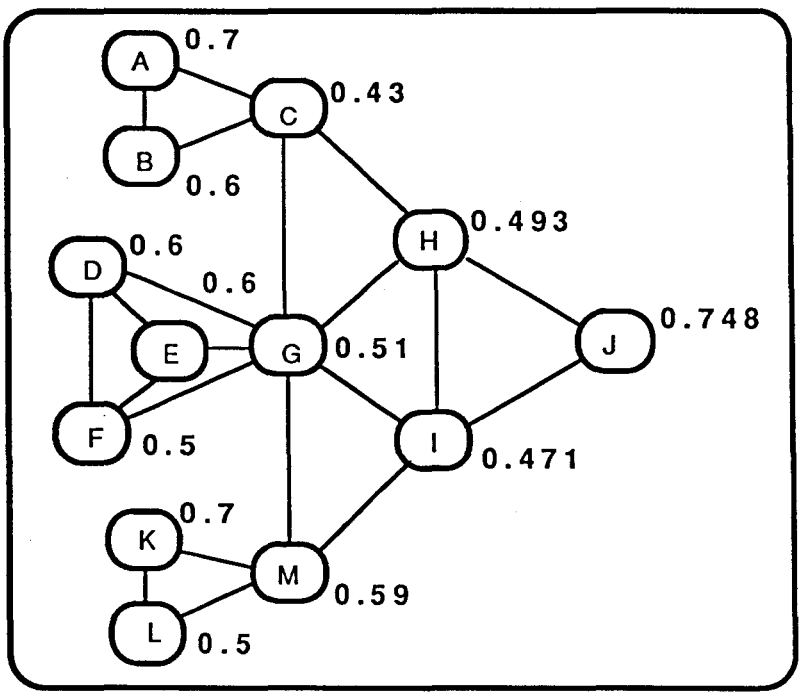


Figura 1.24.- Probabilidades iniciales de los nodos

Finalmente, las probabilidades marginales de los aglomerados o de los nodos se calculan a partir de los términos que figuran en (1.62). El primer factor del miembro de la derecha de (1.62) da la función de probabilidad marginal del aglomerado {A,B,C}, de la que, por marginalización (suma en el conjunto

adecuado) pueden obtenerse las marginales de los nodos A, B y C. Multiplicando luego la marginal, P(C), de C, por P(G,H/C) se obtiene la marginal del aglomerado {G,H,C} y de ella las marginales aisladas y la conjunta de G y H. Multiplicando ahora P(G,H) por P(I/G,H) se obtiene P(G,H,I), y así sucesivamente.

De esta manera se han obtenido las probabilidades marginales de los nodos que se muestran en la figura 1.24.

Si ahora se conoce que "G = cierto" (G="g") resulta

$\Delta^* = \{(A,B,C), (D,E,F), (K,L,M), (C,H), (I,M), (H,I,J)\}$

$\psi^*(A\ B\ C) = \psi(A\ B\ C)$ $\qquad$ $\psi^*(D\ E\ F) = \psi(D\ E\ F\ g)$

$\psi^*(K\ L\ M) = \psi(K\ L\ M)$ $\qquad$ $\psi^*(C\ H) = \psi(C\ g\ H)$

$\psi^*(I\ M) = \psi(g\ I\ M)$ $\qquad$ $\psi^*(H\ I\ J) = \psi(H\ I\ J)\psi(g\ H\ I)$

Nótese que ahora aparece un aglomerado menos que antes. Por tanto resulta:

$\psi^*(a, b, c) = 0.12$	$\psi^*(d, e, f) = 0.10$	$\psi^*(k, l, m) = 0.12$	$\psi^*(h, i, j) = 0.9$
$\psi^*(a, b, \bar{c}) = 0.28$	$\psi^*(d, e, \bar{f}) = 0.16$	$\psi^*(k, l, \bar{m}) = 0.18$	$\psi^*(h, i, \bar{j}) = 0.1$
$\psi^*(a, \bar{b}, c) = 0.15$	$\psi^*(d, \bar{e}, f) = 0.09$	$\psi^*(k, \bar{l}, m) = 0.32$	$\psi^*(h, \bar{i}, j) = 0.8$
$\psi^*(a, \bar{b}, \bar{c}) = 0.15$	$\psi^*(d, \bar{e}, \bar{f}) = 0.02$	$\psi^*(k, \bar{l}, \bar{m}) = 0.08$	$\psi^*(h, \bar{i}, \bar{j}) = 0.2$
$\psi^*(\bar{a}, b, c) = 0.08$	$\psi^*(\bar{d}, e, f) = 0.06$	$\psi^*(\bar{k}, l, m) = 0.12$	$\psi^*(\bar{h}, i, j) = 0.6$
$\psi^*(\bar{a}, b, \bar{c}) = 0.12$	$\psi^*(\bar{d}, e, \bar{f}) = 0.01$	$\psi^*(\bar{k}, l, \bar{m}) = 0.08$	$\psi^*(\bar{h}, i, \bar{j}) = 0.4$
$\psi^*(\bar{a}, \bar{b}, c) = 0.08$	$\psi^*(\bar{d}, \bar{e}, f) = 0.03$	$\psi^*(\bar{k}, \bar{l}, m) = 0.03$	$\psi^*(\bar{h}, \bar{i}, j) = 0.7$
$\psi^*(\bar{a}, \bar{b}, \bar{c}) = 0.02$	$\psi^*(\bar{d}, \bar{e}, \bar{f}) = 0.04$	$\psi^*(\bar{k}, \bar{l}, \bar{m}) = 0.07$	$\psi^*(\bar{h}, \bar{i}, \bar{j}) = 0.3$

$\psi^*(i, m) = 0.3$	$\psi^*(\bar{i}, m) = 0.7$	$\psi^*(c, h) = 0.2$	$\psi^*(\bar{c}, h) = 0.8$
$\psi^*(i, \bar{m}) = 0.5$	$\psi^*(\bar{i}, \bar{m}) = 0.5$	$\psi^*(c, \bar{h}) = 0.8$	$\psi^*(\bar{c}, \bar{h}) = 0.2$

Siguiendo ahora un proceso análogo al anterior, se calculan las nuevas (después de conocido el valor de G) probabilidades marginales de los nodos que resultan ser las de la figura 1.25.

De esta manera, cuando llega nueva información pueden reactualizarse los potenciales de evidencia y, a partir de ellos, recalcular las probabilidades marginales deseadas. Este proceso de cálculo constituye precisamente la base del motor de inferencia de los sistemas expertos basados en redes causales, mientras que la estructura de la red (aglomerados, relaciones de paternidad, tablas, potenciales de evidencia, etc.) constituyen la base de conocimiento.

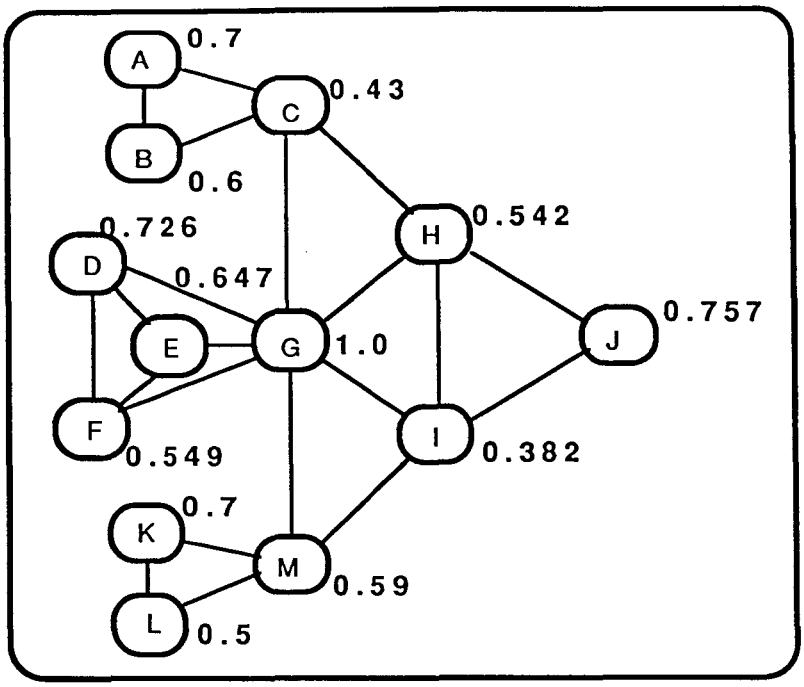


Figura 1.25.- Probabilidades actualizadas de los nodos conocido G

¤

1.8.-APRENDIZAJE

La posibilidad de corregir o aumentar el conocimiento almacenado en la base de conocimiento es una de las funciones más interesantes de un sistema experto (Barr y Feigenbaum (1981), Buchanan y Mitchel (1978)). La incorporación del subsistema de aprendizaje le da potencia y permite solventar en algunos casos el problema que supone el desconocimiento del experto de ciertos aspectos relacionados con la especialidad para la que se diseña el sistema experto.

Hoy en día no existe una teoría general sobre técnicas de aprendizaje. Por ello, en este apartado se darán algunas que tienen validez en casos particulares, centrándonos en aquellas aplicables al aprendizaje de reglas (Angluin (1978,1986), Dietterich y Michalski (1983), Valiant (1984), etc).

A la hora de clasificar las técnicas de aprendizaje se puede distinguir dos tipos fundamentalmente:

(a) mediante comunicación experto-máquina. En este caso el experto ayudado por el ordenador corrige y mejora las reglas almacenadas. Este es el procedimiento empleado por el sistema TEIRESIAS que se ha utilizado para elaborar la base de reglas de MYCIN.

(b) mediante un proceso automático. A partir de datos experimentales el ordenador genera nuevas reglas.

Básicamente la segunda técnica se reduce a un proceso que consta de las siguientes etapas:

(a) consideración de una estructura básica de aprendizaje

(b) formulación de nuevas afirmaciones mediante un mecanismo de generalización, que consiste en la obtención de afirmaciones más débiles que las básicas, en términos lógicos se trata de dado A obtener B de manera que $A \rightarrow B$.

(c) evaluación del nivel de generalización mediante un mecanismo llamado de especificación, que es recíproco de la generalización y consiste en dado A encontrar B de forma que $B \rightarrow A$. Según los resultados obtenidos en este paso se puede volver a una nueva formulación (vuelta a (b)).

1.8.1-Método de Michalski

Este método permite inferir reglas con conclusiones prefijadas, es decir, se pretende obtener reglas que concluyan un objeto concreto dado. Ha sido utilizado ya con éxito para construir las reglas de un sistema para diagnóstico de enfermedades de soja (Michalski y Chilansky, (1980)). El algoritmo propuesto se basa en las etapas anteriormente comentadas, y es el siguiente (Michalski (1983), Michalski et al (1982)).

Supongamos que se pretende aprender una regla cuya conclusión es A y que la base de aprendizaje básica está formada por afirmaciones positivas, que son aquellas que concluyen A y negativas las que concluyen no A. Se llama POS al conjunto de afirmaciones positivas y NEG al de afirmaciones negativas. Entonces los pasos a seguir son:

(a) se selecciona aleatoriamente una afirmación positiva O_i

(b) se procede a crear un árbol de generalización de O_i mediante un proceso que Michalski especifica y que entre sus propiedades cumple que ninguna expresión generalizada se satisface por ningún elemento de NEG

(c) se elige la generalización óptima B_i según un proceso de evaluación

(d) si todas las afirmaciones positivas verifican B_i entonces ir a (f)

(e) si no es así eliminar de POS todas las afirmaciones que verifican B_i y volver a (a)

(f) la fórmula $B_1 \vee B_2 \vee \ldots \vee B_n$ es satisfecha por todos los elementos de POS y no lo es por ninguno de NEG.

(g) simplificar la fórmula utilizando la base de aprendizaje o cálculo de predicados

1.8.2-Método de Quinlan

Este método propuesto por Quinlan (1979), desarrollado a partir del sistema CLS (Concept Learning System, Hunt (1966)), es al igual que el anterior un modelo para aprendizaje de conceptos.

La representación tanto de las premisas como de las conclusiones de las reglas es del tipo atributo-valor. Para cada atributo A_j se considera una serie de valores A_j^k $k=1,\ldots,h_j$. Quinlan propone para deducir el concepto, que se describe a partir de los atributos, crear un árbol de clasificación cuyos nodos terminales sean los valores asociados al concepto. Para construir este árbol se parte de un atributo, que se considera como raiz, del cual salen tantas ramas como valores posea A_i en el conjunto de experiencias. A partir de cada nodo A_j^k salen tantas ramas como valores tenga el siguiente atributo y así sucesivamente.

El objetivo del método de Quinlan es dar un criterio que permita construir un árbol mínimo. Para ello propone como selección del atributo a clasificar en cada vértice aquel que tiene mayor contenido de información respecto del conjunto de experiencias a clasificar en cada vértice. Para medir este contenido de información utiliza la entropía de la siguiente manera.

Supongamos que estamos en un nodo del árbol de clasificación de "n" tipos $C_1, C_2,\ldots,C_n$ que se definen en función de los atributos $A_1, A_2, \ldots, A_m$ tomando A_j los valores

$$A_j^1, A_j^2,\ldots, A_j^{h_j}$$

La entropía en el nodo es

$$H = -\sum_i P_i \log P_i \qquad (1.63)$$

siendo P_i la probabilidad de que se dé una experiencia en la clase o el tipo "i." Si se clasifica según A_j hay h_j valores asociados a este atributo en los cuales se puede calcular la entropía parcial HP_k con $k=1,...,h_j$

$$HP_k = -\sum_i P_i^k \log P_i^k \qquad (1.64)$$

donde P_i^k es la probabilidad de que se dé una experiencia del tipo "i" supuesto que el atributo es A_j^k. Si se estima la nueva entropía por el promedio

$$H(A_j) = \sum_{k=1}^{h_j} prob(A_j^k)\, HP_k \qquad (1.65)$$

y el incremento de la información aportado por la clasificación A_j se evalúa mediante

$$\Delta I_j = H - H(A_j) \qquad (1.66)$$

Quinlan propone para clasificar el árbol aquel atributo A_j con mayor ΔI_j. El método consiste en calcular $\Delta I_1, \Delta I_2, ..., \Delta I_m$ y elegir aquel con el mayor valor.

Suceso	Frecuencia
$E \cap S_1 \cap \overline{S_2} \cap \overline{S_3}$	325
$E \cap S_1 \cap S_2 \cap \overline{S_3}$	25
$E \cap S_1 \cap \overline{S_2} \cap S_3$	25
$E \cap \overline{S_1} \cap S_2 \cap S_3$	50
$E \cap S_1 \cap S_2 \cap S_3$	25
$\overline{E} \cap \overline{S_1} \cap \overline{S_2} \cap \overline{S_3}$	50

Tabla 1.12.- Datos de síntomas y problema

Ejemplo 1.8.- Para mostrar el método anterior supongamos que se tiene los datos obtenidos experimentalmente que aparecen en la Tabla 1.12 en relación a unos síntomas, S_1, S_2 y S_3, y un problema, E. Se pretende determinar cuales son las combinaciones de síntomas que nos determinan el problema.

Para decidir cual será el atributo, S_1, S_2 ó S_3, a tomar como raiz, se calcula la entropía y la información ganada por clasificar respecto de él. La entropía inicial es

$$H = -P(E)\log_2 P(E) - P(\bar{E})\log_2 P(\bar{E}) = -0.9\log_2 0.9 - 0.1\log_2 0.1 = 0.468$$

* Calculamos la entropía en el nodo S_1
  Valor positivo:

$$P(E/S_1) = 1 \qquad P(\bar{E}/S_1) = 0$$
$$HP_1 = -1\log_2 1 = 0$$

Valor negativo:

$$P(E/\overline{S_1}) = 0.5 \qquad P(\bar{E}/\overline{S_1}) = 0.5$$
$$HP_2 = -2 * 0.5\log_2 0.5 = 1$$

Entonces

$$HP(S_1) = P(S_1) * 0 + P(\overline{S_1}) * 1 = 0.2$$

$$\Delta I_1 = H - HP(S_1) = 0.468 - 0.2 = 0.268$$

* Calculamos la entropía en el nodo S_2
  Valor positivo:

$$P(E/S_2) = 1 \qquad P(\bar{E}/S_2) = 0$$
$$HP_1 = -1\log_2 1 = 0$$

Valor negativo:

$$P(E/\overline{S_2}) = 0.875 \qquad P(\bar{E}/\overline{S_2}) = 0.125$$
$$HP_2 = -0.875\log_2 0.875 - 0.125\log_2 0.125 = 0.543$$

Entonces

$$HP(S_2) = P(S_2)^* 0 + P(\overline{S_2})^* 0.543 = 0.434$$

$$\Delta I_2 = H - HP(S_2) = 0.468 - 0.434 = 0.034$$

* Calculamos la entropía en el nodo S_3

Valor positivo:

$$P(E/S_3) = 1 \qquad P(\bar{E}/S_1) = 0$$
$$HP_1 = -1 \log_2 1 = 0$$

Valor negativo:

$$P(E/\overline{S_3}) = 0.875 \qquad P(\overline{E}/\overline{S_3}) = 0.125$$
$$HP_2 = -0.875 \log_2 0.875 - 0.125 \log_2 0.125 = 0.543$$

Entonces

$$HP(S_3) = P(S_3)^* 0 + P(\overline{S_3})^* 0.543 = 0.434$$

$$\Delta I_3 = H - HP(S_3) = 0.468 - 0.434 = 0.034$$

Como ΔI_1 es el valor máximo tomamos S_1 como nodo raiz. Para cada una de las ramas, valor positivo, S_1, y negativo, $\overline{S_1}$, se vuelve a repetir el proceso. En este nodo la entropía inicial es

$$H = -P(E/\overline{S_1}) \log_2 P(E/\overline{S_1}) - P(\bar{E}/\overline{S_1}) \log_2 P(\bar{E}/\overline{S_1}) = -2^* 0.5 \log_2 0.5 = 1$$

* Calculamos la entropía en el nodo S_2

Valor positivo:

$$P(E/\overline{S_1} \cap S_2) = 1 \qquad P(\bar{E}/\overline{S_1} \cap S_2) = 0$$
$$HP_1 = -1 \log_2 1 = 0$$

Valor negativo:

$$P(E/\overline{S_1} \cap \overline{S_2}) = 0 \qquad P(\bar{E}/\overline{S_1} \cap \overline{S_2}) = 1$$
$$HP_2 = -1 \log_2 1 = 0$$

Entonces

$$HP(S_2) = P(S_2) * 0.0 + P(\overline{S_2}) * 0.0 = 0.0$$

$$\Delta I_2 = H - HP(S_2) = 1.0 - 0.0 = 1.0$$

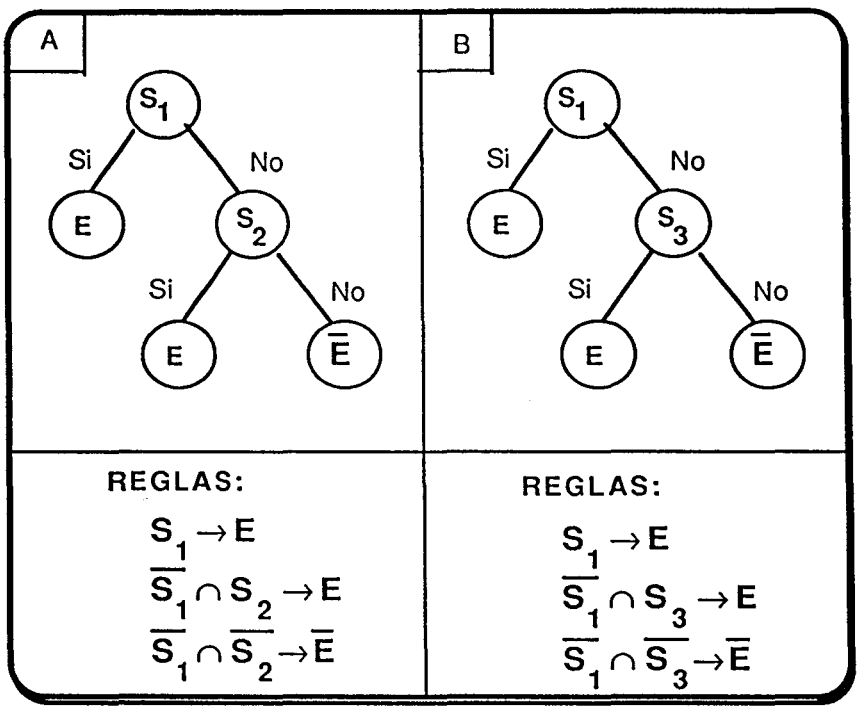


Figura 1.26.- Conjunto de reglas de la Tabla 1.12

* Calculamos la entropía en el nodo S_3
Valor positivo:

$$P(E / \overline{S_1} \cap S_3) = 1 \qquad P(\overline{E} / \overline{S_1} \cap S_2) = 0$$

$$HP_1 = -1 \log_2 1 = 0$$

Valor negativo:

$$P(E / \overline{S_1} \cap \overline{S_3}) = 0 \qquad P(\overline{E} / \overline{S_1} \cap \overline{S_3}) = 1$$

$$HP_2 = -1 \log_2 1 = 0$$

Entonces

$$HP(S_3) = P(S_3)^* 0.0 + P(\overline{S_3})^* 0.0 = 0.0$$

$$\Delta I_3 = H - HP(S_3) = 1.0 - 0.0 = 1.0$$

Entonces podemos tomar S_2 ó S_3 como el siguiente nodo pudiendo plantearse dos alternativas al final de la rama. En la figura 1.26 aparecen representadas los dos conjuntos de reglas posibles que pueden deducirse.

¤

1.8.3-Método de Valiant

Valiant (1984) propone un método para aprender conceptos a partir de una base de experiencias analizando algunas clases de conceptos aprendibles.

La principal contribución de este autor es que muestra que es posible diseñar máquinas que pueden aprender con las siguientes características:

(a) las clases de conceptos que pueden ser aprendidos mediante esta máquina pueden ser caracterizadas

(b) el proceso por el cual la máquina puede deducir reglas es polinomial.

Brevemente se analizarán los algoritmos dados por Valiant que permiten aprender conceptos. Para ello se considerarán t variables booleanas A_1, ..., A_t que sólo pueden tomar los valores 0 ó 1 indicando con ello que la variable es falsa si toma el valor cero y cierta en caso contrario.

Definición 1.28.- *(Vector).-* Un vector es una asignación de un valor del conjunto {0,1,*} a cada una de las variables booleanas A_1, ..., A_t donde "*" significa que la variable no está determinada.

Dado un vector v se denotará por v_i al valor asignado a la variable A_i en v. Así por ejemplo un vector es la asignación A_1=0, A_2=1 y A_3=0 siendo en este caso v_1=0, v_2=1 y v_3=0.

Definición 1.29.- *(Vector total).-* Un vector es total cuando todas las variables están determinadas, es decir, si

$$v_i \in \{0, 1\} \qquad \forall\, i \in \{1, 2, .., t\}$$

Definición 1.30 .- *(Función booleana).-* Una función booleana es una aplicación del conjunto de los 2^t vectores totales en el conjunto {0,1}.

Un concepto C puede considerarse como la extensión de una función booleana F extendiendo el dominio de F al conjunto de todos los vectores de la siguiente manera, para cada vector v

$$(1.67) \qquad F(v) = 1 \Leftrightarrow F(w) = 1 \quad \forall\, w \in \mathfrak{D}_v$$

siendo

$$(1.68) \qquad \mathfrak{D}_v = \{ w \text{ vector } / \forall\, i \in \{1,..,t\} \;\; w_i = v_i \;\; \text{si } v_i \in \{0, 1\} \}$$

Hay dos componentes en este proceso de aprendizaje. En primer lugar una base de experiencias y un procedimiento deductivo que consiste en un mecanismo por el cual los conceptos son definidos. Se supone también que existen dos subrutinas llamadas

(a) EJEMPLOS que da como salida un vector v de forma que F(v)=1. Se denota por D a la distribución de probabilidad sobre los vectores v en los que F vale 1.

(b) ORACLE que tiene como input un vector v y como salida un valor 0 ó 1 dependiendo si F(v)=0 ó 1.

Definición 1.31.- (*Programa aprendible*).- Sea X una clase de programas representando conceptos. Entonces X es aprendible mediante la base de experiencias si y solamente si existe un algoritmo A con las siguientes características

(a) para todos los programas f en X y todas las distribuciones D sobre los vectores v en los cuales f toma el valor 1 el algoritmo permite deducir un programa g en X verificando

(i) $\forall\, v \text{ vector } \; g(v) = 1 \Rightarrow f(v) = 1$

(ii) $$\sum_{\substack{v\,/\,f(v)=1 \\ g(v)\neq 1}} D(v) \leq h^{-1}$$

donde h es un parámetro variable.

(b) el algoritmo termina en tiempo polinomial con respecto al número de variables t y al parámetro h.

En un caso más general se puede necesitar considerar una distribución de probabilidad D_1 en el conjunto de vectores tales que f(v) es distinto de 1. En este caso la condición (i) de (a) puede sustituirse por

$$\text{(iii)} \quad \sum_{\substack{v\,/\,f(v)\neq 1 \\ g(v)=1}} D_1(v) \leq h_1^{-1}$$

Vamos a definir a continuación diferentes clases de conceptos que pueden ser aprendidos.

Definición 1.32.- *(Forma normal conjuntiva).-* Una forma normal conjuntiva es un producto $C_1\ C_2 \ldots\ C_r$ de cláusulas donde cada una de ellas es la suma de literales (A_i ó la negación de A_i). Para cada entero positivo K se puede definir una k-forma normal conjuntiva como una forma normal conjuntiva donde cada clausula es la suma de a lo sumo k literales.

Por ejemplo $(A_1 + \overline{A_2})(A_1 + A_3)$ es una forma normal conjuntiva.

Definición 1.33. *-(Implicación de conceptos).-* Sean F y G dos conceptos. Diremos que F implica a G, $F \Rightarrow G$, si para cada vector tal que F(v)=1 entonces G(v)=1.

Valiant demostró que cada k-forma normal conjuntiva es aprendible con llamadas a EJEMPLOS y no es requerida ninguna llamada a ORACLE.

Algoritmo.-Se comienza inicializando g como el producto de todas las posibles clausulas de k literales de $\{A_1, \overline{A_1}, A_2, \overline{A_2}, \ldots, A_t, \overline{A_t}\}$. El algoritmo mediante llamadas a EJEMPLOS produce afirmaciones positivas v de f. Para cada clausula C_i se analiza si v implica C_i si no es así se elimina de g.

Este algoritmo podemos escribirlo en pseudocódigo de la manera siguiente:

begin
  v:=EJEMPLOS
  Para cada C_i en g
    If v no implica C_i **then** eliminar C_i de g
end

Ejemplo 1.9.-Consideremos t=2 y k=2, y que el concepto que queremos aprender es A_1+A_2. En este caso el conjunto de afirmaciones positivas para este concepto es

$$E_p = \{A_1 \overline{A_2}, \overline{A_1} A_2, A_1 A_2\}$$

La inicialización de g como producto de todas las clausulas de dos literales resulta ser

$$g = (A_1)\ (\overline{A_1})(A_2)(\overline{A_2})\ (A_1 + A_2)\ (\overline{A_1} + A_2)\ (A_1 + \overline{A_2})\ (\overline{A_1} + \overline{A_2})$$

Supongamos que en la primera llamada a EJEMPLOS se tiene v:=A_1 entonces eliminando de g todas aquellas clausulas que no implican v queda

$$g = (A_1)\ (A_1 + A_2)\ (A_1 + \overline{A_2})$$

Si en una segunda llamada a EJEMPLOS se tiene v:=A_2 borrando de g las dos clausulas que no implica a A_2, esto es $(A_1 + \overline{A_2})$ y (A_1), nos queda g como el concepto que se pretendía aprender. Cualquier elemento que tomemos del conjunto de afirmaciones positivas no puede elimininar de g ninguna otra clausula. ¤

Definición 1.34.- *(Forma normal disyuntiva).-* Una forma normal disyuntiva es un producto $m_1 + m_2 + \ldots + m_r$ de monomios donde cada uno de ellos es un producto de literales.

Por ejemplo $A_1 \overline{A_2} + A_1 A_3 + A_2$ es una forma normal disyuntiva.

Valiant demostró que cada forma normal disyuntiva es aprendible con llamadas a EJEMPLOS y a ORACLE. El algoritmo es el siguiente.

Algoritmo.-Se inicializa el concepto g a cero. El algoritmo mediante llamadas a EJEMPLOS produce afirmaciones positivas v de f. Para cada variable A_i se analiza si esta determinada en v. Si esto es así se construye otro vector w con los mismos valores para las variables que los asignados por v salvo la i-ésima que se hace indeterminada, es decir,

$$w_j = v_j \quad \text{si} \quad j \neq i$$
$$w_i = *$$

Si ORACLE(w) nos devuelve el valor 1 se cambia v por w y se añade a g el monomio m que resulta de realizar el producto de aquellos literales q tales que v implica q. Este proceso se repite con la siguiente variable.

Este algoritmo podemos escribirlo en pseudocódigo de la manera siguiente:

```
begin
    v:=EJEMPLOS
    if v no implica g then
        begin
            for I:=1 to t do
                if v_i es 0 ó 1 then
                    begin
                        sea w igual a v pero con v_i:=*;
                        if ORACLE(w)=1 then v:=w;
                    end
            hacemos m igual al producto de todos los literales q
            tales que v implica q
            g:=g+m;
        end
end
```

Ejemplo 1.10.- Supongamos igual que en el Ejemplo 1.9 que el número de variables booleanas es 2, t=2, y que queremos aprender el concepto A_1+A_2. En este caso el conjunto de afirmaciones positivas para este concepto es

$$E_p = \{A_1\overline{A_2}, \overline{A_1}A_2, A_1A_2\}$$

Se comienza inicializando el concepto g a cero. En la Tabla 1.13 aparece el proceso seguido mediante este algoritmo para aprender el concepto. Despues de dos llamadas a EJEMPLOS se obtiene el concepto A_1+A_2.

EJEMPLOS	w		ORACLE(w)	g
$v = A_1\overline{A_2}$	i= 1	$\overline{A_2}$	0	$A_1\overline{A_2}$
	i= 2	A_1	1	$A_1\overline{A_2} + A_1$
$v = \overline{A_1}A_2$	i= 1	A_2	1	$A_1\overline{A_2} + A_1 + A_2$
	i= 2	Ω	0	$A_1\overline{A_2} + A_1 + A_2$

Tabla 1.13.- Aprendizaje de la forma normal disyuntiva A_1+A_2

¤

1.9.-CONTRIBUCIONES ORIGINALES DE LA TESIS

En este apartado se describen esquemáticamente, y sólo a modo de introducción, las aportaciones originales de esta tesis, que se desarrollan con detalle en los capítulos siguientes.

La Tabla 1.14 muestra un resumen de los problemas principales que se han abordado, las soluciones que aparecen en la literatura existente, las deficiencias que se han observado en éstas y, finalmente, la solución propuesta en la tesis.

A continuación se comenta brevemente esta tabla y se dan algunas de las motivaciones que han llevado al estudio de estos problemas.

Uno de los problemas más candentes en Sistemas Expertos es la discusión existente entre los defensores de los métodos probabilísticos y otros modelos alternativos, que han surgido ante los problemas detectados en éstos. El estudio de estos métodos alternativos, incluyendo los mecanismos de propagación, ha demostrado su falta de capacidad para reproducir la realidad en muchos casos prácticos, o su incoherencia interna y falta de rigor. Poner de manifiesto estos problemas es una contribución de esta tesis.

Entre las causas de crítica de los métodos basados en la probabilidad destacan la insuficiencia del método basado en la hipótesis de independencia y la imposibilidad de reflejar la dependencia total, debida al excesivo número de parámetros requerido. En la tesis se propone un método alternativo llamado método de dependencia de síntomas relevantes, que resuelve el problema.

Este modelo ha sido implementado en su totalidad en un ordenador Macintosh utilizando todas sus posibilidades de interfase con el usuario, como son el ratón, los menús, las ventanas, los cuadros de diálogo, etc. Además, han sido incorporados unos subsistemas de aprendizaje, de explicación y de acceso y modificación de la base de datos, lo cual constituye una aportación importante a la tesis.

Después de analizar los diferentes sistemas de representación de reglas (conjunción de formas normales, disyunción de formas normales, etc.) se consideró importante el problema de caracterizar los tipos de reglas (o sus conjuntos asociados) que correspondían a cada una de estas formas y el de analizar la representación mínima posible, lo cual se ha hecho en esta tesis. Además, en muchos casos resultaba necesario pasar de una representación a

otra. Por ello, se han desarrollado algoritmos que permiten el cambio de representación.

Uno de los problemas de los sistemas basados en reglas consiste en la lentitud del motor de inferencia, debido a que cada vez que actua se ve obligado a encadenar las reglas. La solución más típica a este problema es la compilación de las reglas previamente a su almacenamiento en la base de conocimiento. En la tesis se dan dos modelos de compilación, uno de implicación de reglas y otro de equivalencia según el paradigma de trabajo que se considere.

En muchos ambientes de la Inteligencia Artificial se habla de sistemas de reglas incoherentes. Aquí se muestra que con la concepción elegida no existen nunca reglas incoherentes, sino hechos que contradicen las reglas. La metodología que se describe permite detectar estas incoherencias.

Otra de las contribuciones de este trabajo consiste en poner de manifiesto la facilidad de que existan incoherencias en la base de conocimiento de los sistemas probabilísticos. Ello quiere decir que dicha base puede contener probabilidades que contradicen los axiomas de la probabilidad. Con objeto de evitar este problema se dan dos soluciones originales al mismo. En una de ellas, se fuerza a una aportación ordenada de probabilidades y se controla su coherencia. En la otra, se da un método, basado en las técnicas de programación lineal, para asistir al experto humano a dar su información en forma de probabilidades iniciales o condicionadas. En este último caso el orden de suministro de la información es completamente libre.

Con cierta frecuencia, resulta imposible forzar al experto humano a dar probabilidades en forma puntual, por lo que se ve uno obligado a aceptarla en forma de intervalos. Ello exige un tratamiento adecuado y totalmente diferente. La teoría de la evidencia sumistra una posible solución a este problema, pero tiene el gravísimo inconveniente de requerir muchos más parámetros que los métodos probabilísticos usuales. En este trabajo se da un método que permite, mediante la programación lineal, utilizar bases de conocimiento en forma de intervalos y que es un verdadero motor de inferencia que suministra también intervalos.

El contacto con la programación lineal para resolver los problemas anteriores ha motivado la necesidad de resolver problemas paramétricos asociados a la información por intervalos descrita en el párrafo anterior. Ello lleva a que los términos independientes de las desigualdades o igualdades sean parámetros. En la tesis se describe un algoritmo para resolver este tipo de problemas en forma simbólica, tal como lo hacen los conocidos paquetes REDUCE, MAPLE, MACSYMA, etc., en otros problemas matemáticos.

Durante el estudio bibliográfico de esta tesis pudo comprobarse que hay bastantes técnicas estadísticas, totalmente contrastadas en la práctica, que tienen grandes posibilidades de ser aplicadas en los problemas de los Sistemas Expertos, sin que hayan sido aplicadas hasta el momento. Una aportación de la tesis consiste en señalar algunas de estas técnicas y describir como pueden ser utilizadas en las bases de conocimiento, motor de inferencia, explicación y aprendizaje.

Otra de las carencias observada en los sistemas expertos comerciales es la ausencia de métodos de aprendizaje, en especial cuando la información disponible es parcial. Se aporta un método de aprendizaje válido para los modelos de independencia y de dependencia de síntomas relevantes que mantiene la coherencia de la base de conocimiento en todo momento.

Uno de los modelos más interesantes para el tratamiento de la incertidumbre mediante métodos probabilísticos es el de redes causales, propuesto por Lauritzen y Spiegelhalter (1988). Sin embargo, estos autores mencionan que es necesario dotarle de métodos de aprendizaje. En la tesis se da un método, basado en el método de la máxima verosimilitud, que permite éste a partir de información total. Además se demuestra que el aprendizaje puede tener lugar localmente por ser equivalente, para este método, al aprendizaje global.

Uno de los temas de más actualidad en la Inteligencia Artificial es el del aprendizaje de conceptos. Valiant, Pitt, Dietterich y Michalski, etc. han estudiado este interesante problema desde el punto de vista determinista. La tesis generaliza este problema a los conceptos probabilísticos y caracteriza, mediante técnicas estadísticas, los tipos de conceptos que pueden ser aprendidos mediante algoritmos que terminan en tiempo polinómico. Con algunos ejemplos se pone de manifiesto el notable incremento de recursos necesarios para incluir la aleatoriedad.

Problema	Soluciones aportadas	Deficiencias observadas	Solución propuesta
Métodos alternativos a los probabilísticos	Teoría de la evidencia Factores de certeza Lógica difusa	- Mecanismos de propagación inadecuados - Reproducción incorrecta de realidad	- Poner de manifiesto sus peligros
Modelos Probabilísticos en Sistemas Expertos	- Modelos de independencia - Modelos de dependencia	- No reproduce la realidad - Excesivo número de parámetros	- Modelo de dependencia de síntomas relevantes
Concha de tipo probabilístico	Muchas	- Utilización hipótesis independencia - Ausencia de aprendizaje	- RSPS con aprendizaje y modelo de dependencia con síntomas relevantes
Representación de reglas	Varias (CNF, DNF, etc.)	- Falta de optimización - Ausencia de métodos de cambio de representación	- Cota mínima de la representación mediante conjuntos pseudobásicos - Algoritmos de cambio de representación
Compilación de reglas	Varias	- Lentitud del motor de inferencia	- Modelo de equivalencia de reglas - Modelo de implicación de reglas
Coherencia de la base del conocimiento en sistemas basados en reglas	Varias	- Incoherencia de la base de hechos y de reglas	- Modelo basado en la contrastación de hechos y reglas compiladas

Problema	Soluciones aportadas	Deficiencias observadas	Solución propuesta
Coherencia de la base del conocimiento en sistemas de tipo probabilístico	No explicitas	- Incoherencia de la base de conocimiento - Falta de ayuda al experto humano	- Modelo basado en p. lineal - Modelo de aportación ordenada de información
Tratamiento de la incertidumbre mediante intervalos	Teoría de la evidencia	- Muy elevado número de parámetros - Problema en la propagación de la incertidumbre	- Motor de inferencia basado en programación lineal - Método de acotación de la probabilidad
Programación lineal paramétrica	Tratamiento a nivel analítico	- Necesidad de estudiar casos aisladamente	- Tratamiento simbólico
Modelos estadísticos en Sistemas Expertos	Varias	- Ausencia de modelos contrastados - Ausencia de métodos para tratamiento de síntomas continuos	- Utilización de modelos de la Estadística a los Sistemas Expertos
Aprendizaje en Sistemas Probabilísticos	- Métodos basados en el experto humano	- Ausencia de coherencia - Necesidad de información total - Falta de automatismo	- Modelo de aprendizaje con información parcial y coherente
Aprendizaje en modelos redes causales	No propuestas	- Ausencia de un modelo de aprendizaje	- Modelo basado en el método de la máxima verosimilitud
Aprendizaje de conceptos	Angluin (1978,1986), Valiant (1984, 1988), Pitt and Valiant (1988), Dietterich and Michalski (1983)	Sólo incluye conceptos deterministas	- Extensión a los conceptos probabilísticos - Caracterización de conceptos aprendibles

2 HIPOTESIS DE TRABAJO MATERIAL Y METODO DE ESTUDIO

MODELOS PROBABILÍSTICOS

2.-MODELOS PROBABILÍSTICOS

2.1.- INTRODUCCIÓN

Uno de los tipos de sistemas expertos más importantes son los basados en probabilidad. El nucleo central de este tipo de sistemas lo constituye el espacio probabilístico y el procedimiento capaz de utilizar esta representación para producir respuestas (motor de inferencia) se apoya en las probabilidades condicionadas.

Se estudiarán en este capítulo diferentes modelos que nos permitirán realizar un diagnóstico de un conjunto de problemas $\{E_i ; i = 1, 2,..., n\}$, no necesariamente disjuntos, en una población dada Ω, suponiendo que se dispone de unos datos en forma de síntomas $\{S_j ; i = 1, 2,.., m\}$, donde un síntoma, S_j, es una partición $S_j = \{B_{j1}, B_{j2},.., B_{jr_j}\}$ de Ω (ver figura 2.1).

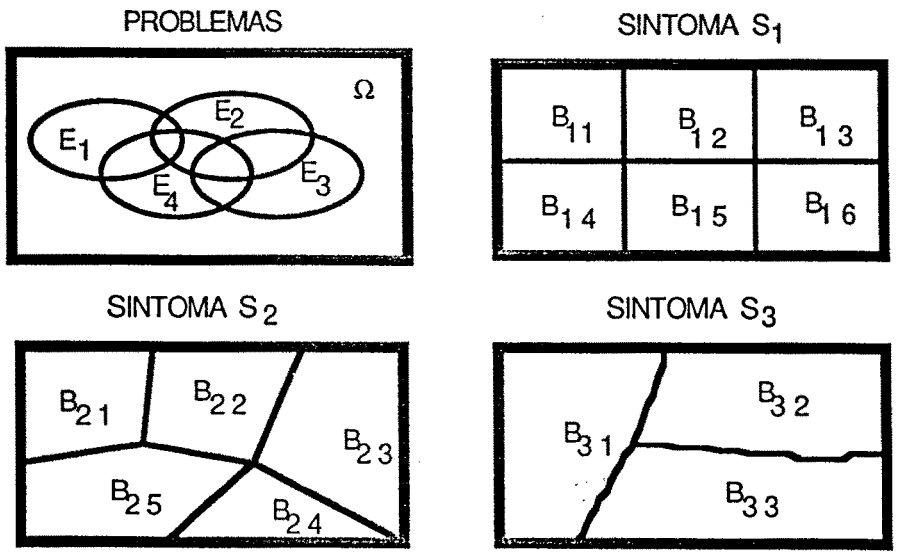


Figura 2.1.- Representación gráfica de los problemas y los síntomas.

Elegido un caso, que vendrá representado por unos síntomas $A_1 \cap A_2 \cap .. \cap A_m$, donde $A_j \in \mathcal{P}(S_j) - \{\phi\}$ y $\mathcal{P}(S_j)$ es el conjunto de partes de

S_j, el objetivo del motor de inferencia es obtener las probabilidades de cada uno de los problemas supuesto conocidos estos síntomas:

$$(2.1) \qquad P(E_i / A_1 \cap A_2 \cap .. A_m) = P(E_i \cap A_1 \cap A_2 \cap .. A_m) / P(A_1 \cap A_2 \cap .. A_m) \qquad i = 1, 2, .., n$$

Nótese que si A_j es Ω significa falta de información sobre el síntoma S_j.

En la fórmula (2.1), el papel de la probabilidad $P(A_1 \cap A_2 \cap ... \cap A_m)$ es actuar como una constante de normalización, y una decisión basada en el máximo valor de $P(E_i/A_1 \cap A_2 \cap ... \cap A_m)$ coincide con la basada en el valor máximo de $P(E_i \cap A_1 \cap A_2 \cap ... \cap A_m)$. Por ello, a veces, en vez de utilizar el criterio anterior, se utiliza uno equivalente, que consiste en calcular los cocientes de verosimilitud

$$(2.2) \qquad V_i = \frac{P(E_i \cap A_1 \cap A_2 \cap .. \cap A_m)}{\underset{i}{\text{Max}}\, P(E_i \cap A_1 \cap A_2 \cap .. \cap A_m)} \quad ; \quad i = 1, 2, ..., n$$

y decidirse por el problema E_i que da el valor unidad.

En lo que sigue, se llamará $\mathbf{C}$ a la clase de conjuntos de la forma $(A_1 \cap A_2 \cap ... \cap A_m)$. Esta clase incluye conjuntos con un nivel de información que va desde la ausencia total de la misma ($A_j = \Omega$, para j=1,2,...,m) al conocimiento perfecto (para todo j, $A_j = B_{jk}$,con $k \in \{1, ..., r_j\}$).

Si todos los A_j son diferentes, habrá $\prod_{j=1}^{m} (2^{r_j} - 1)$ conjuntos distintos en la clase $\mathbf{C}$, pero todos ellos pueden ser descritos mediante uniones de conjuntos de la subclase $\mathbf{D}$ de $\mathbf{C}$ de conjuntos tales que para cada j existe $k \in \{1, ..., r_j\}$ de forma que $A_j = B_{jk}$. La clase $\mathbf{D}$ incluye sólo conjuntos con información sobre todos los síntomas. Nótese que $\mathbf{D}$ es una partición de Ω y que su cardinal es $\prod_{j=1}^{m} r_j$. Se supone que el diagnóstico se hará sobre conjuntos de $\mathbf{D}$.

Con objeto de presentar los siguientes modelos se analizarán:

(a) el número de parámetros independientes existentes en cada uno de los modelos
(b) cuáles son (se ha seleccionado una de las posibles alternativas)
(c) cuáles son las restricciones para dichos parámetros
(d) cómo son actualizados cuando se dispone de nueva información.

2.2.- MODELO GENERAL DE DEPENDENCIA (GD)

En el modelo general de dependencia se permite cualquier tipo de dependencia en el conjunto de síntomas $S_1,S_2,...,S_m$. Esto implica que se necesita conocer para cada problema los valores $P(E_i \cap D)$ donde $D \in \mathfrak{D}$, por tanto, se debe disponer de $\prod_{j=1}^{m} r_j$ grados de libertad o parámetros. En la figura 2.2 aparece representada la población por los puntos interiores al rectángulo que limita su contorno. Los parámetros quedan reflejados gráficamente por el área de todos los conjuntos que definen los problemas y los síntomas al intersecarse. Con objeto de actualizar las probabilidades cuando se conoce un nuevo caso se necesita además el parámetro N (tamaño de la población). Por tanto, el número total de ellos es

$$P_{GD} = n \prod_{j=1}^{m} r_j + 1 \qquad (2.3)$$

La expresión (2.3) muestra que el modelo de dependencia general no es factible en muchos casos ya que su número es demasiado elevado. Si se supone, por ejemplo, que el número de problemas es 200 y que se consideran 300 síntomas binarios, el número que resulta es $P_{GD} \approx 4x10^{92}$ siendo tan elevado que no es posible almacenarlo en ningún ordenador de los actuales.

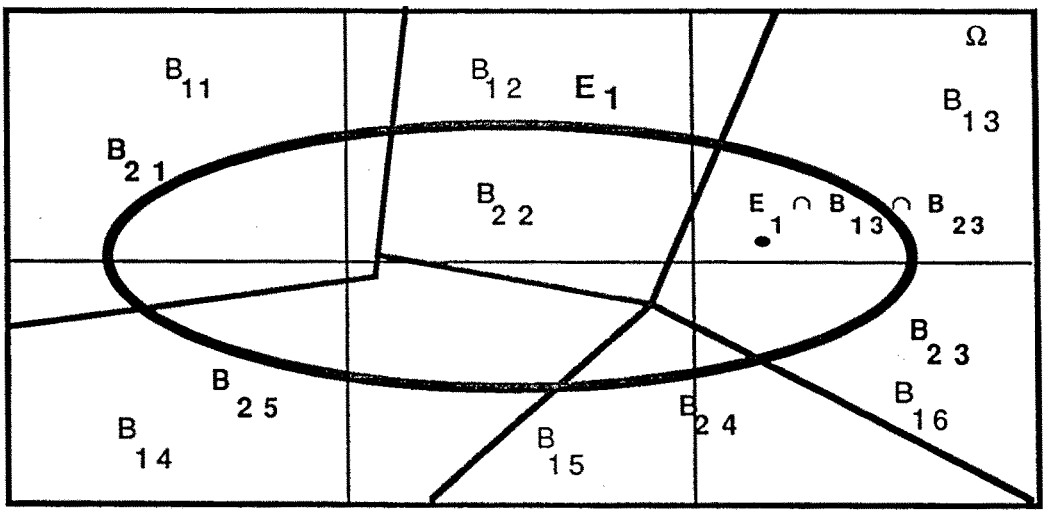


Figura 2.2.- Ilustración del diagnóstico de un problema.

Los parámetros anteriores, por ser probabilidades, deben estar sujetos a las siguientes restricciones

$$(2.4)\qquad \left.\begin{array}{l} \sum_{D\in \mathfrak{D}} P(E_i \cap D) \le 1 \\ \\ P(E_i \cap D) \ge 0 \text{ , para todo } D \in \mathfrak{D} \\ \\ N > 0 \end{array}\right\} \quad i = 1, 2, .., n$$

La actualización de parámetros, cuando un nuevo dato con síntomas $A_1^+ \cap A_2^+ \cap .. \cap A_m^+ \in \mathfrak{C}$ y problemas conocidos ($\mathfrak{P}$) puede hacerse mediante las fórmulas

$$(2.5)\qquad P^*(E_i \cap A) = \begin{cases} P(E_i \cap A)[N + 1/P(E_i \cap A_1^+ \cap A_2^+ \cap .. \cap A_m^+)] / (N+1) \\ \qquad\qquad\qquad\qquad\qquad\qquad \text{si } i \in \mathfrak{P} \quad \text{y } A \in \mathfrak{D}^+ \\ P(E_i \cap A)N/(N+1) \qquad \text{en otro caso} \end{cases}$$

$$N^* = N + 1$$

donde $\mathfrak{P}$={j / el dato tiene el problema j }, los asteriscos en P y N indican los valores actualizados y $\mathfrak{D}^+ = \left\{ D \in \mathfrak{D} / D \subseteq A_1^+ \cap A_2^+ \cap .. \cap A_m^+ \right\}$.

Nótese que

$$(2.6)\qquad P(E_i \cap A_1^+ \cap A_2^+ \cap .. \cap A_m^+) = \sum_{D \in \mathfrak{D}^+} P(E_i \cap D)$$

Los parámetros actualizados según (2.5) satisfacen también las restricciones iniciales dadas en (2.4).En efecto, por (2.5) y (2.6) se tiene que si i$\in \mathfrak{P}$

$$(2.7)\qquad \sum_{D\in \mathfrak{D}} P^*(E_i \cap D) = \sum_{D\in \mathfrak{D}^+} P^*(E_i \cap D) + \sum_{D\notin \mathfrak{D}^+} P^*(E_i \cap D) =$$

$$= \frac{1}{N+1}\left(\sum_{D\in \mathfrak{D}^+} P(E_i \cap D)\left[N + 1/P(E_i \cap A_1^+ \cap A_2^+ \cap .. \cap A_m^+)\right] \right) +$$

$$+\frac{1}{N+1}\left(\sum_{D\in \mathfrak{D}^-} P(E_i\cap D)\,N\right)=\frac{1}{N+1}\left(1+\sum_{D\in \mathfrak{D}} P(E_i\cap D)\,N\right)$$

En el caso de que $i\notin \mathfrak{P}$

$$(2.8)\qquad \sum_{D\in \mathfrak{D}} P^*(E_i\cap D) = \frac{1}{N+1}\left(\sum_{D\in \mathfrak{D}} P(E_i\cap D)\,N\right)$$

Por (2.7) y (2.8)

$$\text{si } \sum_{D\in \mathfrak{D}} P(E_i\cap D)\le 1 \quad \text{entonces} \quad \sum_{D\in \mathfrak{D}} P^*(E_i\cap D)\le 1$$

2.3.- MODELO DE DEPENDENCIA DE SÍNTOMAS RELEVANTES (DR)

El modelo anterior, a pesar de ser el más general posible, presenta, sin embargo, el inconveniente de requerir excesivo número de parámetros que lo hace inviable en la práctica cuando el número de síntomas es elevado. Con objeto de reducir este número de parámetros se permite la dependencia entre los llamados síntomas relevantes para cada problema, es decir, se seleccionan para cada uno de los problemas E_i un subconjunto de síntomas

$\{S_j;\ j\in Q_i=\{k_1^i,k_2^i,\ldots,k_{l_i}^i\}\}$ para los cuales todas las dependencias están permitidas (los que son relevantes para la identificación de ese problema). Se supone además que el resto de los síntomas son independientes entre sí y con los anteriores. En otras palabras se supone

$$(2.9)\qquad P(E_i\cap A_1\cap A_2\cap \ldots A_m)=P(E_i)\,P(A_1\cap A_2\cap \ldots\cap A_m/E_i)=$$
$$=P(E_i\cap(\underset{j\in Q_i}{\cap} A_i))\prod_{j\notin Q_i} r_{ij}$$

donde

$$(2.10)\qquad r_{ij}=P(A_j/E_i)=P(B_{jk}/E_i)=p_{ijk}\quad \text{si } A_j=B_{jk}$$

y $p_{ijK}\in[0,1]$.

En lo que sigue se denota mediante $\mathfrak{D}_i$ a la subclase de $\mathfrak{C}$, tal que $A_j\in S_j$ si $j\in Q_i$ y $A_j=\Omega$ si $j\notin Q_i$. Nótese que esta clase es una partición de Ω de cardinal

$$\prod_{j=1}^{l_i} r_{k_j}^{i}$$

y que sólo incluye conjuntos con información completa sobre todos los síntomas relevantes del problema E_i. Los parámetros del modelo son los de la forma

$$P(E_i \cap D) \text{ con } D \in \mathfrak{D}_i,\ p_{ijk}\ (i = 1, 2, \ldots, n;\ k = 1, 2, \ldots, r_j;\ j \notin Q_i) \text{ y } N,$$

que hacen un total de

$$P_{DR} = \sum_{i=1}^{n} \left[\prod_{j=1}^{l_i} r_{k_j}^{i} + \sum_{j \notin Q_i} r_j \right] + 1 \qquad (2.11)$$

Estos deben satisfacer las siguientes restricciones,

$$\left.\begin{array}{l} \displaystyle\sum_{D \in \mathfrak{D}_i} P(E_i \cap D) \leq 1 \\ P(E_i \cap D) \geq 0 \text{ , para todo } D \in \mathfrak{D}_i \\ \left.\begin{array}{l} \displaystyle\sum_{k=1}^{r_j} p_{ijk} = 1 \\ p_{ijk} \geq 0 \end{array}\right\} \ k = 1, 2, \ldots, r_j \quad j \notin Q_i \end{array}\right\} \ i = 1, 2, \ldots, n \qquad (2.12)$$

$$N > 0$$

La actualización de estos parámetros puede hacerse mediante las fórmulas

$$P^*(E_i \cap A) = \begin{cases} P(E_i \cap A)[N + 1 / P(E_i \cap (\underset{j \in Q_i}{\cap} A_j^+))] / (N + 1) & \\ & \text{si } i \in \mathbf{P} \text{ y } A \in \mathbf{D}_i^+ \\ P(E_i \cap A) N / (N + 1) & \text{en otro caso} \end{cases} \qquad (2.13)$$

$$p^*_{ijk} = \begin{cases} [p_{ijk}(NP(E_i) + P(E_i) / P(E_i \cap A_j^+)] / (NP(E_i) + 1) & \text{si } i \in \mathbf{P} \text{ y } A_j \subset A_j^+ \\ p_{ijk} NP(E_i) / (NP(E_i) + 1) & \text{si } i \in \mathbf{P} \text{ y } A_j \not\subset A_j^+ \\ p_{ijk} & \text{en otro caso} \end{cases}$$

$$N^* = N + 1$$

donde

$$\mathfrak{D}_i^+ = \left\{ D \in \mathfrak{D}_i / D \subseteq \bigcap_{j \in Q_i} A_j^+ \right\}$$

y $P(E_i)$ se puede expresar en función de los parámetros como

$$P(E_i) = \sum_{D \in \mathfrak{D}_i} P(E_i \cap D)$$

Nótese que

(2.14) $$P\left(E_i \cap \left(\bigcap_{j \in Q_i} A_j^+ \right)\right) = \sum_{D \in \mathfrak{D}_i^+} P(E_i \cap D)$$

Los parámetros actualizados según (2.13) verifican también las restricciones iniciales dadas en (2.12). En efecto, con un razonamiento análogo a (2.7) y (2.8) se obtiene que la primera restricción de (2.12) se cumple para los parámetros actualizados. Por otro lado, si $i \in \mathfrak{P}$ y $j \notin Q_i$ llamando

$$I_{1j} = \left\{ k \in \{1,..,r_j\} / B_{jk} \subseteq A_j^+ \right\}, I_{2j} = \{1,..,r_j\} \setminus I_{1j}$$

por (2.13) y (2.14) se tiene que

(2.15) $$\sum_{k=1}^{r_j} p_{ijk}^* = \sum_{k \in I_{1j}} p_{ijk}^* + \sum_{k \in I_{2j}} p_{ijk}^* =$$

$$= \frac{1}{NP(E_i)+1} \left[\sum_{k \in I_{1j}} p_{ijk} \left(NP(E_i) + P(E_i) / P(E_i \cap A_j^+) \right) + \sum_{k \in I_{2j}} p_{ijk} NP(E_i) \right] =$$

$$= \frac{1}{NP(E_i)+1} \left[NP(E_i) + \sum_{k \in I_{1j}} P(E_i \cap B_{jk}) / P(E_i \cap A_j^+) \right] = 1$$

Este modelo puede ser simplificado suponiendo modelos logarítmico lineales para $P(E_i \cap D)$ $(D \in D_i, i = 1, 2,.., n)$, es decir,

(2. 16) $$\log(P(E_i \cap (\underset{j \in Q_i}{\cap} A_j)) = u_i + \sum_{k \in K_i} v_{ik}$$

donde u_i y v_{ik} $(k \in \mathcal{K}_i ; i = 1, 2, .., n)$ son constantes y los cardinales, m_i, de $\mathcal{K}_i$ dependen del modelo logarítmico-lineal seleccionado.

Ahora, las expresiones (2.9) y (2.11) pueden escribirse

$$\log (P (E_i \cap A \cap A_2 \cap ... \cap A_m)) = \log [P (E_i \cap (\underset{j \in Q_i}{\cap} A_j))] +$$

(2. 17) $$+ \sum_{j \in Q_i} \log r_{ij} = u_i + \sum_{k \in K_i} v_{ik} + \sum_{j \in Q_i} \log r_{ij}$$

(2. 18) $$P_{DR} = 1 + \sum_{i=1}^{n} (1 + m_i + \sum_{j \in Q_i} r_j)$$

y las dos primeras restricciones de (2.12) se reducen a

(2. 19) $$u_i \leq - \log \left(\sum_{D \in \mathcal{D}_i} \exp \left(\sum_{k \in K_i} v_{ik} \right) \right)$$

En efecto, por (2.16)

(2. 20) $$P(E_i \cap (\underset{j \in Q_i}{\cap} A_j)) = \exp (u_i + \sum_{K \in \mathcal{K}_i} v_{ik})$$

y, por tanto,

$$\sum_{D \in \mathcal{D}_i} P(E_i \cap D) = \sum_{D \in \mathcal{D}_i} \exp (u_i + \sum_{K \in \mathcal{K}_i} v_{ik}) \leq 1 \quad \Leftrightarrow$$

$$\sum_{D \in \mathcal{D}_i} \exp \left(\sum_{K \in \mathcal{K}_i} v_{ik} \right) \leq \exp (- u_i) \quad \Leftrightarrow$$

$$u_i \leq - \log \left(\sum_{D \in \mathcal{D}_i} \exp \left(\sum_{K \in \mathcal{K}_i} v_{ik} \right) \right)$$

2.4.- MODELO DE INDEPENDENCIA (GI)

En este modelo se supone independencia entre cualquier conjunto de síntomas para un problema dado, E_i, y se tiene entonces

$$(2.21) \qquad P(E_i \cap A_1 \cap A_2 \cap .. A_m) = P(E_i) P(A_1 \cap A_2 \cap ... A_m / E_i) =$$

$$= P(E_i) P(A_1 / E_i) P(A_2 / E_i) ... P(A_m / E_i)$$

en esta expresión cada factor $P(A_j/E_i)$ modifica la probabilidad inicial $P(E_i)$ según se conoce el nuevo síntoma A_j, reflejando la contribución de cada síntoma al valor de $P(E_i \cap A_1 \cap A_2 \cap ... \cap A_m)$.

En este modelo, se tienen los siguientes parámetros:

$P(E_i)$ (i=1,2,...n), $P(B_{jk} / E_i)$ (i = 1, 2,.., n; k = 1, 2,.., r_j ; j = 1, 2,.., m) y N.

El número de parámetros es

$$(2.22) \qquad P_{GI} = n \left(\sum_{i=1}^{m} r_i + 1 \right) + 1$$

Las restricciones que deben verificar son

$$(2.23) \qquad \left.\begin{array}{l} 0 \le P(E_i) \le 1 \\ 0 \le P(B_{jk} / E_i) \quad k = 1, 2, .., r_j \quad j = 1, 2, .., m \\ \sum_{k=1}^{r_j} P(B_{jk} / E_i) = 1 \end{array}\right\} \quad i = 1, 2, .., n$$

$$N > 0$$

y su actualización puede hacerse mediante

$$P^*(E_i) = \begin{cases} [P(E_i)N + 1]/(N+1) & si\ i \in \mathcal{P} \\ [P(E_i)N]/(N+1) & si\ i \notin \mathcal{P} \end{cases}$$

(2.24)
$$P^*(B_{jk}/E_i) = \begin{cases} P(B_{jk}/E_i)\left[P(E_i)N + 1/P(A_j^+/E_i)\right]/(NP(E_i)+1) & si\ i \in \mathcal{P}\ y\ B_{jk} \subset A_j^+ \\ P(B_{jk}/E_i)P(E_i)N/(NP(E_i)+1) & si\ i \in \mathcal{P}\ y\ B_{jk} \not\subset A_j^+ \\ P(B_{jk}/E_i) & en\ otro\ caso \end{cases}$$

$$N^* = N + 1$$

En el cuadro 2.1 se resumen los parámetros de cada uno de los modelos anteriores y el número de parámetros.

MODELO	PARAMETROS	NUMERO DE PARAMETROS
Dependencia general	$P(E_i \cap A_1 \cap \ldots \cap A_m)$ N	$P_{GD} = n\prod_{j=1}^{m} r_j + 1$
Dependencia síntomas relevantes	$P(E_i \cap (\bigcap_{j \in Q_i} A_j))$ $P(A_j/E_i)\ \ j \in Q_i$ N	$P_{DR} = \sum_{i=1}^{n}\left[\prod_{j=1}^{l_i} r_{k_j^i} + \sum_{j \notin Q_i} r_j\right] + 1$
Independencia general	$P(E_i)$ $P(A_j/E_i)$ N	$P_{GI} = n\left(\sum_{i=1}^{m} r_i + 1\right) + 1$

Cuadro 1.- Parámetros de los diferentes modelos probabilísticos

2.5.- CONCHA RSPS

En este apartado se describe la concha RSPS (Relevant Symptoms Probabilistic Shell), que ha sido desarrollada como parte de esta tesis (Castillo et al. (1987,1988)). En el apartado 7.1 se incluye el listado completo del programa. Aunque en lo que sigue se hablará de enfermedades y de síntomas y se mostrará un ejemplo de Medicina para hacer más fácilmente comprensible la explicación, la concha es válida para todo problema de diagnosis que pueda definirse mediante síntomas y opciones (enfermedades).

La concha es de tipo probabilístico y se basa en el modelo de síntomas relevantes descrito en el apartado 2.3. Esta hipótesis reduce considerablemente el número de parámetros sin afectar de forma significativa a las conclusiones que pueden derivarse en el diagnóstico ya que la hipótesis de independencia afecta sólo a los síntomas no relevantes.

La implementación se ha hecho en un ordenador Macintosh y se ha aplicado para la ayuda al diagnóstico de enfermedades comunes y no comunes por médicos, estudiantes de Medicina y enfermeras. Se ha ensayado con un ejemplo de 90 enfermedades y 143 síntomas binarios, lo que supone un total de 32022 parámetros. La implementación del modelo se ha hecho en Lightspeed Pascal y se han incluido ventanas, menús e instrucciones autoexplicativas, en español e inglés, con objeto de conseguir una interfase de usuario sencilla y agradable. La figura 2.3 muestra los menús, que por sí solos ya dan una idea de las posibilidades del sistema experto en cuanto a información sobre enfermedades y síntomas, diagnóstico, explicación, simulación de pacientes, acceso a la base de datos e idioma.

Seleccionando la opción "ayuda" del menú "información", el usuario entra en una sesión de ayuda (figura 2.4), durante la cual se inactivan los menús y se convierten en autoexplicativos (ver ejemplo en figura 2.5). A partir de esa selección pueden consultarse todos los menús y descubrir las posibilidades de la concha hasta que, siguiendo las instrucciones indicadas en la parte inferior de la figura 2.5, se selecciona la opción "ayuda" del menú "informar" para concluir dicha fase. Esta fase de ayuda tiene una cierta importancia ya que el usuario puede en cualquier momento acceder a ella y consultar las dudas que se le presenten sin necesidad de recurrir a un manual que resulta bastante engorroso. También es posible hacer la consulta en inglés sin más que seleccionar dicho idioma en el menú del mismo nombre.

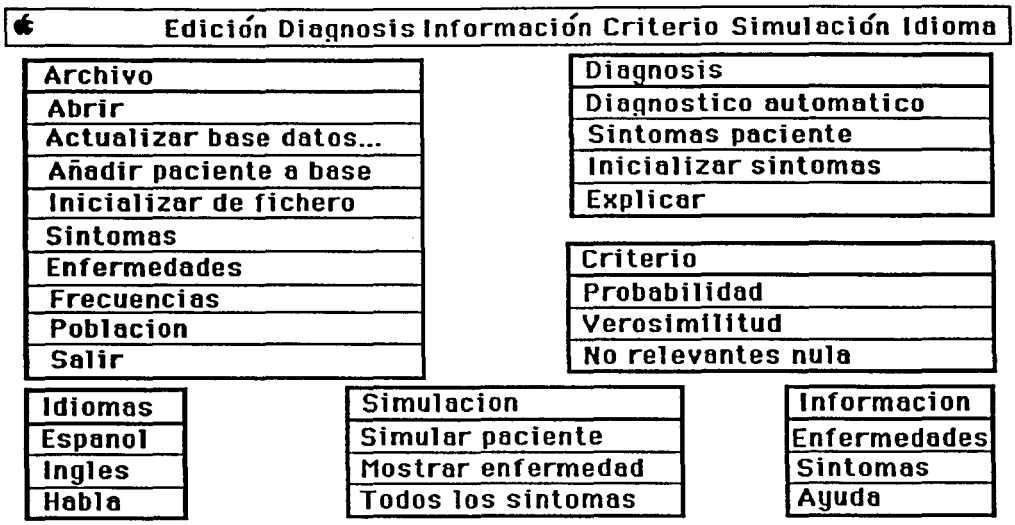


Figura 2.3.- Detalle de menús utilizados por la concha

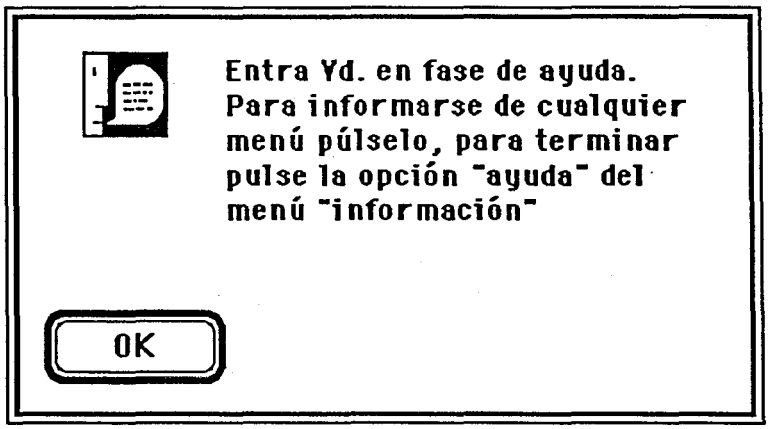


Figura 2.4.- Entrada en la fase de ayuda

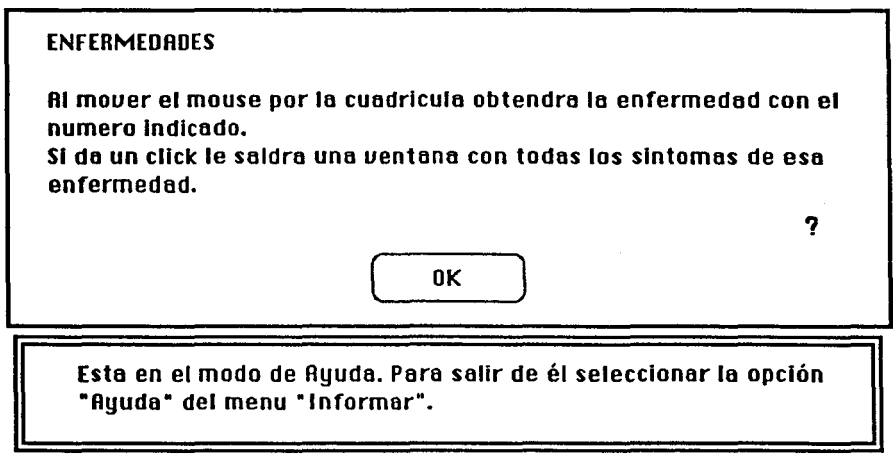


Figura 2.5.- Ejemplo de autoexplicación

Otro de los menús ("información") permite, mediante el desplazamiento del ratón por la red de las iniciales de las enfermedades, ordenadas alfabéticamente, el acceso a la información de las enfermedades y de los números de sus síntomas relevantes (ver figura 2.6) y, pulsando el ratón, a los nombres de estos síntomas (figura 2.7). De esta forma se tiene un acceso rápido a los síntomas relevantes de las enfermedades. Esta opción tiene mucha importancia para el caso de usuarios en fase de aprendizaje, ya que de una forma muy sencilla y rápida pueden consultar los principales síntomas de las diferentes enfermedades y realizar las comparaciones pertinentes. Además dichos síntomas le orientan para continuar con el diagnóstico, pues son los que tiene que conocer del paciente para poder tomar una decisión.

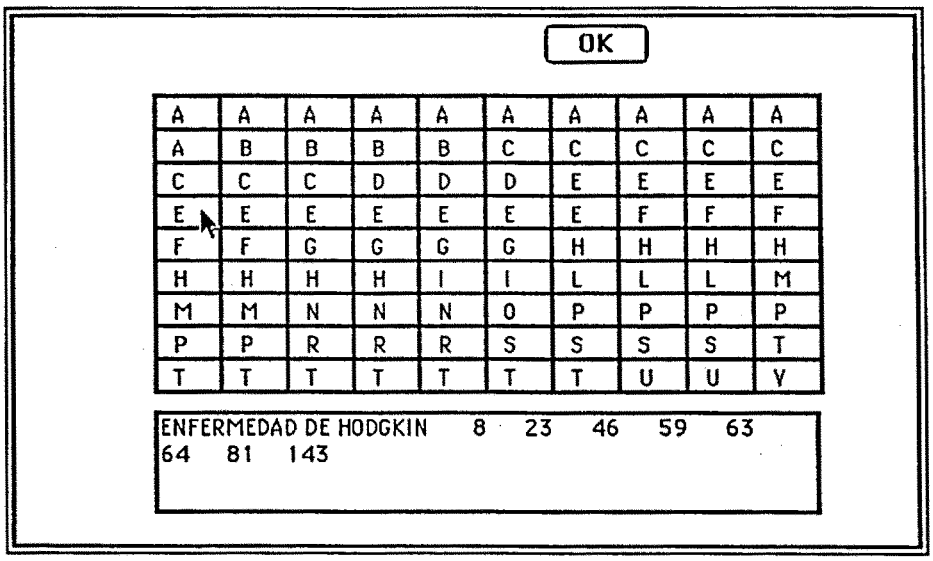


Figura 2.6.- Acceso a enfermedades y sus síntomas

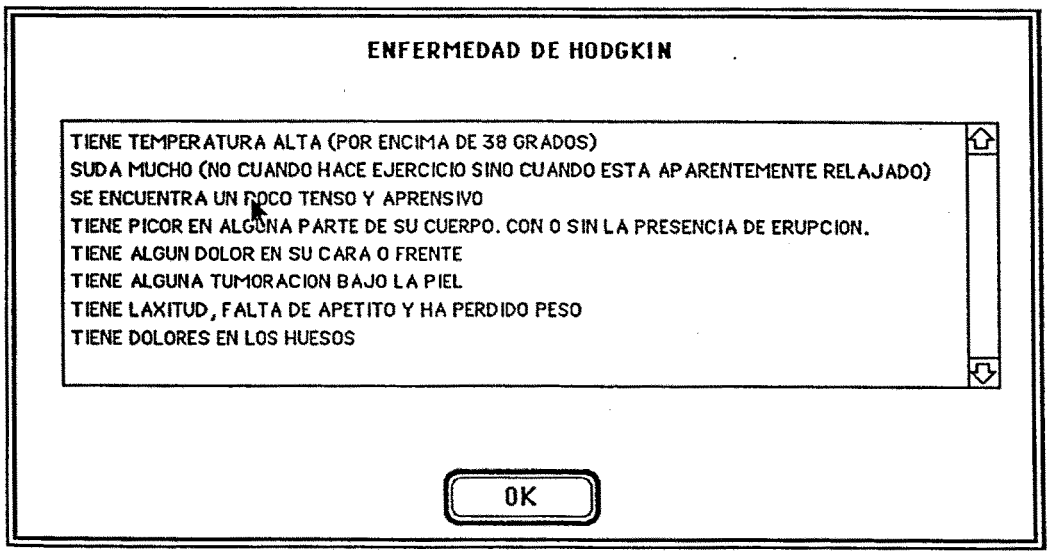


Figura 2.7.- Síntomas de una enfermedad

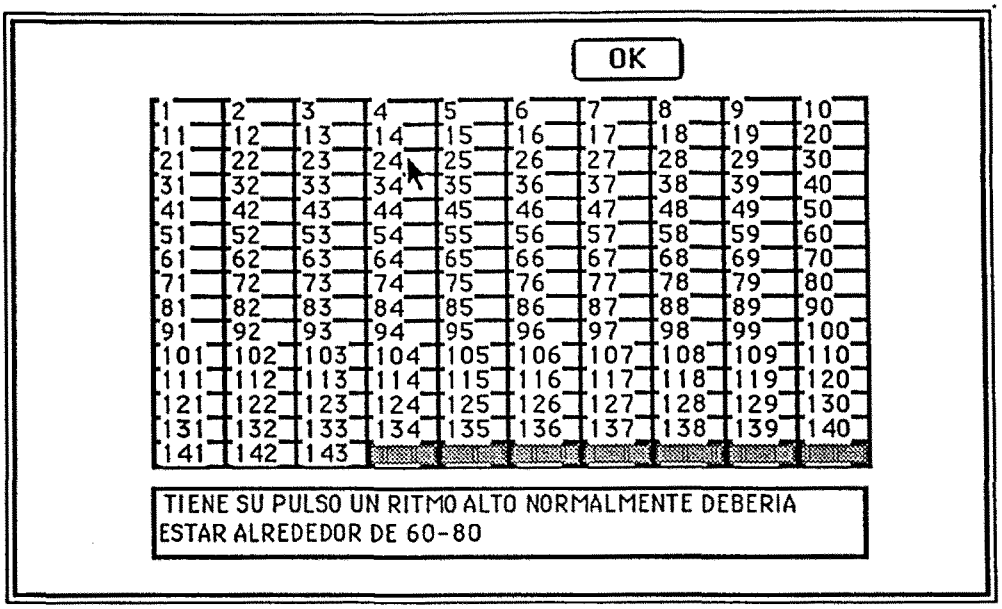


Figura 2.8.- Acceso a los síntomas

Análogamente, por desplazamiento del ratón sobre la malla en la que se muestran los números de los síntomas, puede accederse a cualquiera de ellos (figura 2.8) y, pulsando el ratón, a las enfermedades de las que éstos son relevantes (figura 2.9). Esta información indirecta es también muy útil al usuario, pues puede conocer qué enfermedades tienen un síntoma dado como relevante. Esta información indirecta es muy valiosa para el diagnóstico.

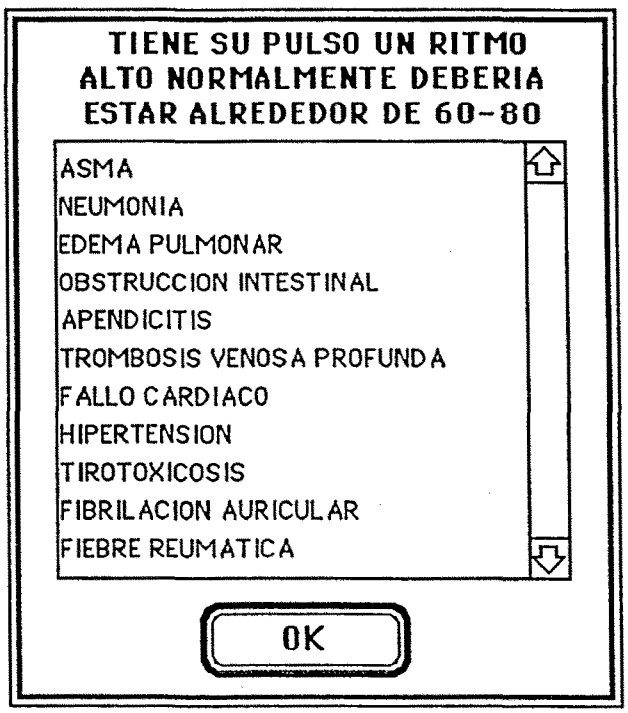


Figura 2.9.- Enfermedades con un síntoma dado

En fase de diagnóstico de un cierto enfermo, la información sobre los síntomas que posee éste puede darse pulsando el ratón sucesivamente sobre los números de los síntomas. Por permutación circular se pueden elegir las respuestas "sí", "no", "no sabe" o "desconocido" (ver figura 2.10).

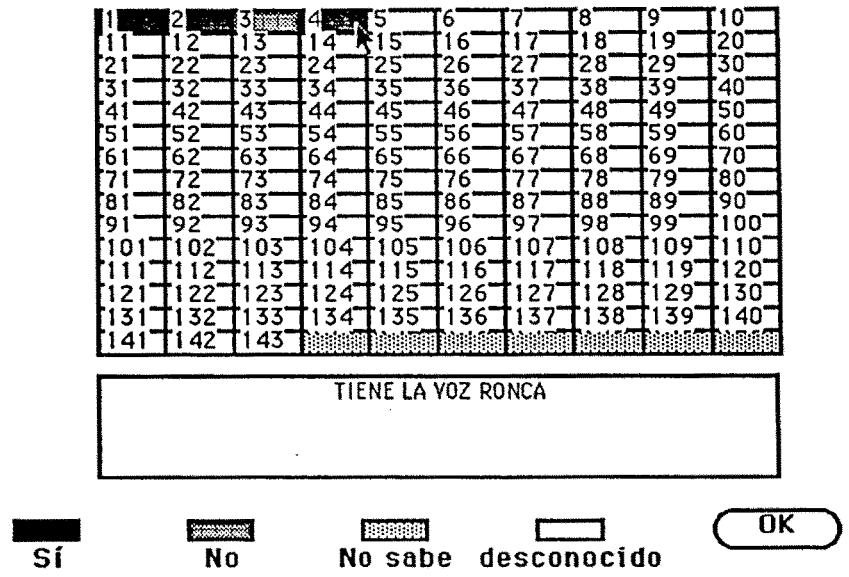


Figura 2.10.- Definición de síntomas del enfermo

Una vez conocidos algunos de estos síntomas, el sistema puede realizar un diagnóstico automático, dando la lista (ordenada de más a menos probable) de las 5 enfermedades más probables para los síntomas dados junto con sus probabilidades asociadas (ver figura 2.11).

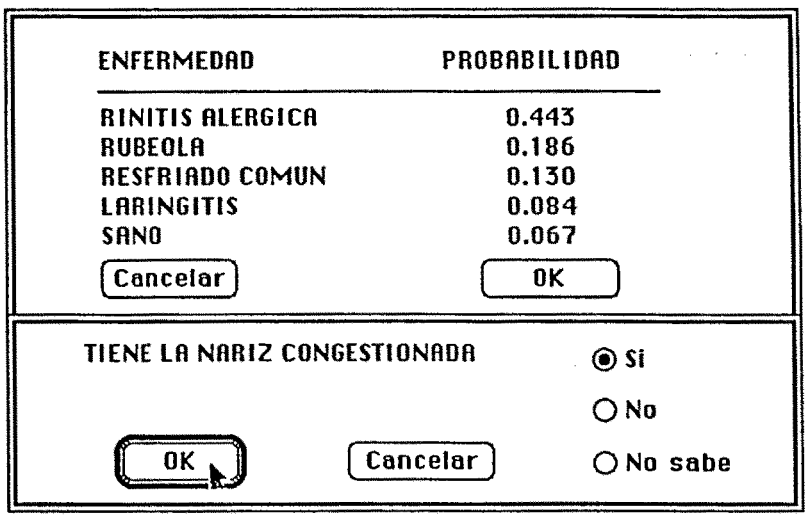


Figura 2.11.- Diagnóstico automático

Si la probabilidad de la enfermedad que ocupa el primer puesto en la lista está lejana del valor 1, el sistema experto puede preguntar más síntomas y actualizar las probabilidades hasta conseguir dicha condición, en cuyo caso puede darse un diagnóstico fiable. Al terminar éste, puede solicitarse del sistema una explicación de la decisión tomada (ver figura 2.12).

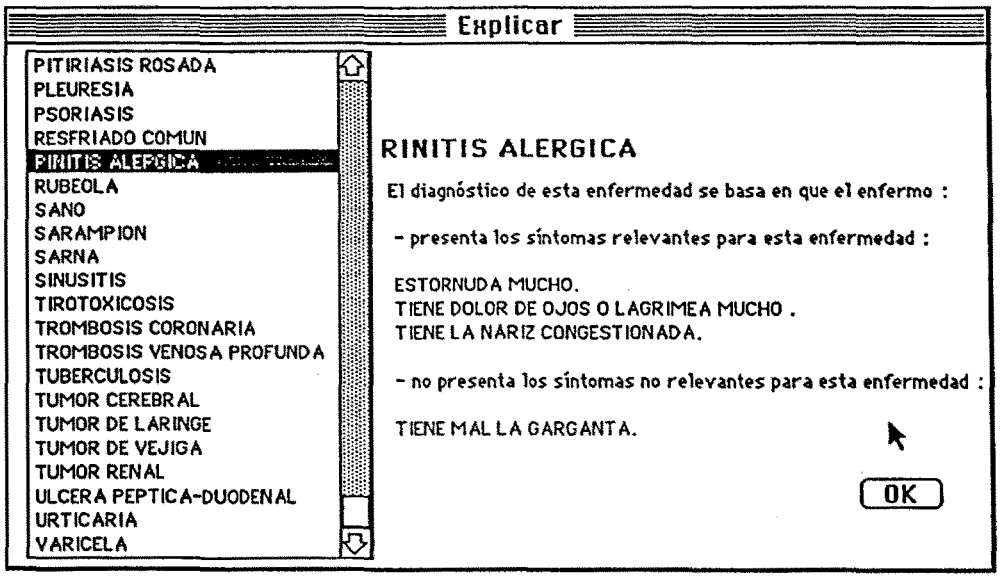


Figura 2.12.- Explicación que sigue al diagnóstico

Con objeto de posibilitar la enseñanza mediante el sistema experto, se ha incluido un simulador de pacientes. Una vez simulado un paciente, el médico, alumno o enfermera puede consultar sus síntomas, mediante la opción "síntomas enfermo" del menú "diagnóstico" y tratar de adivinar la enfermedad simulada; después, puede solicitarse el diagnóstico automático y, finalmente, preguntar al sistema la enfermedad realmente simulada (figura 2.13). Una comparación entre éstas tres permite contrastar la bondad del usuario, el sistema experto o la calidad de la base de datos utilizada.

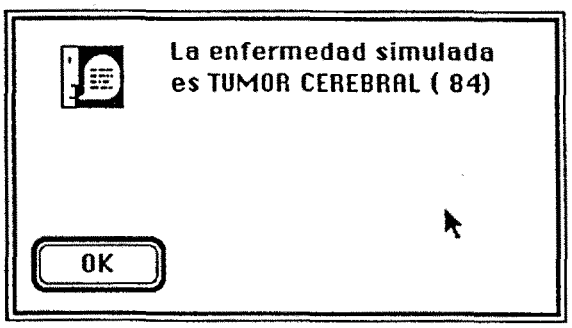


Figura 2.13.- Enfermedad simulada

El acceso y modificación de la base de datos, puede hacerse muy fácilmente, ya que se da una interfase de usuario que lo permite. Para acceder a los síntomas, bien para consulta o actualización, puede utilizarse la opción "síntomas" del menú "archivo" (ver figura 2.14). En esta fase pueden modificarse los nombres de los síntomas, borrar o añadir síntomas y acceder a ellos de uno en uno, mediante los botones "anterior" y "siguiente", o tecleando su número en la casilla superior derecha de la figura 2.14.

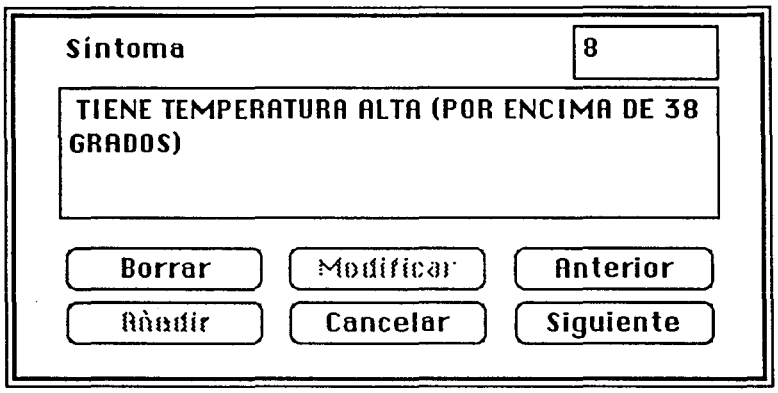


Figura 2.14.- Acceso a los síntomas

Análogamente, el acceso a los datos de las enfermedades y sus correspondientes síntomas relevantes, puede hacerse mediante la opción "enfermedades" del menú "archivo" (ver figura 2.15). Además de borrar, modificar o añadir enfermedades, pueden incluirse y quitarse síntomas relevantes a cada una de las enfermedades.

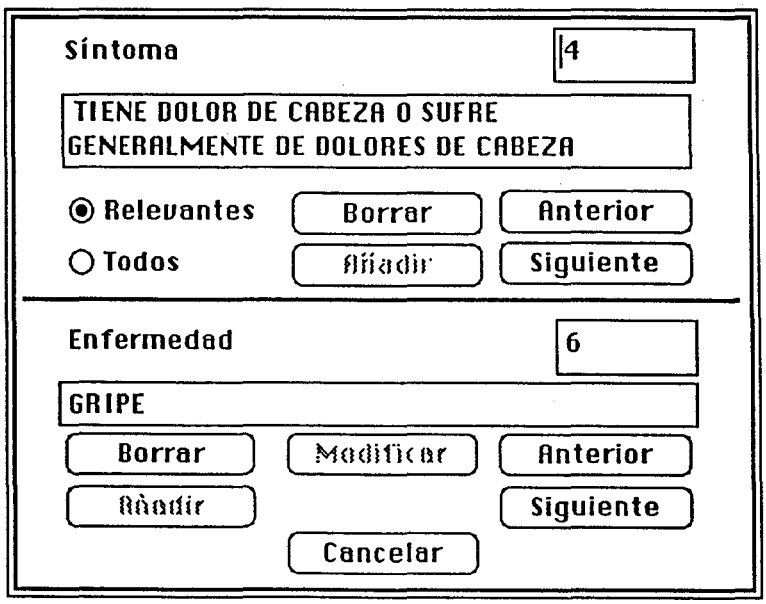


Figura 2.15.- Acceso a las enfermedades

La opción "frecuencias" del menú "archivo" permite conocer el número de pacientes que, teniendo una cierta enfermedad, poseen un conjunto de síntomas dado, que se selecciona mediante un conjunto de botones de radio (ver figura 2.16). En los tres casos, la información puede ser modificada, añadida o borrada sin más que utilizar los cuadros de diálogo en la forma habitual.

Finalmente, el sistema incluye aprendizaje paramétrico ya que los parámetros pueden ser actualizados al conocerse datos de nuevos pacientes. También puede considerarse un cierto tipo de aprendizaje estructural cambiando algunos parámetros desde valores límites, de cero o uno, a valores intermedios; así, si se sospecha sobre la relevancia de cierto síntoma, puede considerarse como relevante e inicialmente darle probabilidad nula o con la hipótesis de independencia y, luego, dejar que se actualice a la luz de los datos de los pacientes.

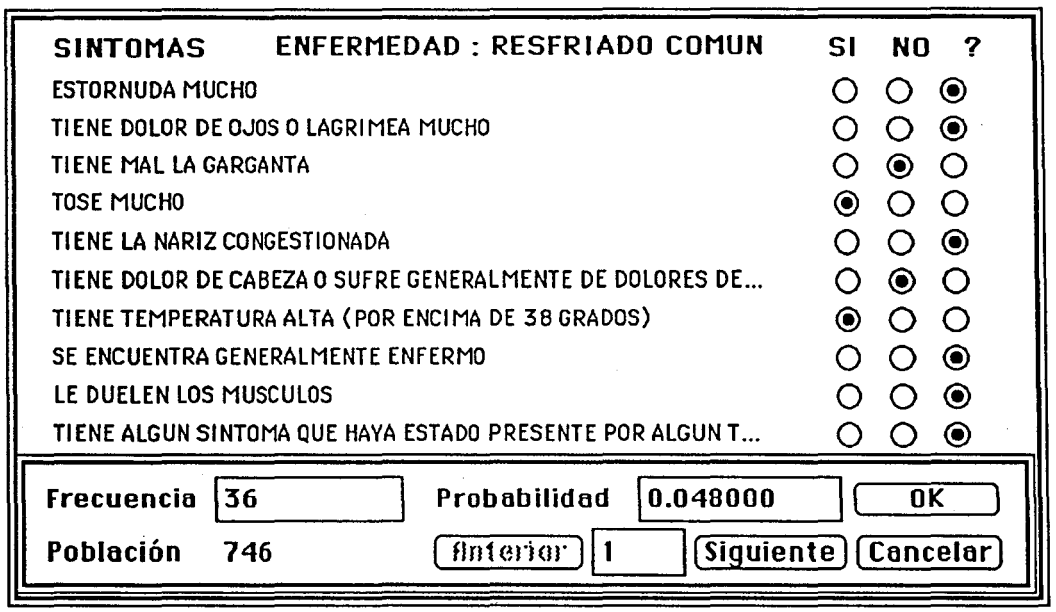


Figura 2.16.- Acceso a las frecuencias

CONTROL DE LA COHERENCIA

3.-CONTROL DE LA COHERENCIA

La conservación de la coherencia de los modelos es un requisito necesario (Smith (1961)). Un sistema experto debe de sintetizar el conocimiento de un experto o de una base de datos pero nunca sus imperfecciones. Por ello es necesario una estrategia de control que nos permita validar el modelo. En este capítulo se verá que si el conocimiento se almacena mediante reglas al combinar estas con los hechos podremos llegar a contradicciones. Por otro lado, si se opta por una estructura probabilística, no se puede ya actuar con total libertad, sino que se está limitado por los axiomas de la probabilidad. En consecuencia, las probabilidades asignadas por el experto humano o las calculadas por el subsistema de propagación de incertidumbre deben respetar dichos axiomas.

3.1.-MODELOS BASADOS EN REGLAS

Los sistemas expertos basados en reglas poseen, como ya se ha comentado tres partes claramente diferenciadas: las reglas, la base de hechos y el motor de inferencia. El sistema de reglas permite establecer una relación lógica entre los objetos, mientras que el motor de inferencia se encarga de aplicar las reglas a los hechos conocidos (base de datos) para obtener nuevos hechos (deducidos).

Es de sumo interés caracterizar los conjuntos que pueden ser generados mediante ciertos tipos de reglas, ya que ello nos permite conocer las limitaciones de cada uno de ellos. A este fin dedicaremos el apartado 3.1.1.

En el apartado 3.1.2 se tratará de determinar el estado de cada objeto mediante una sola regla que combine toda la información que sobre dicho objeto exista en la base del conocimiento del sistema experto. Para ello se necesita definir el concepto de reglas compiladas. De esta manera será posible conocer el valor de un objeto aplicando una vez la regla compilada, sin necesidad de encadenar las reglas. Se analizará también cuando la combinación de reglas y hechos produce nuevos hechos contradictorios con los anteriores.

3.1.1.-Formas booleanas

El problema que se plantea en este apartado es el de caracterizar los conjuntos asociados a expresiones escritas mediante la utilización limitada de operadores lógicos "o" e "y" aplicados a variables booleanas. Este problema es equivalente al de caracterizar los conjuntos que resultan mediante uniones e intersecciones, en número limitado, de conjuntos dados. En primer lugar se

introducirán algunas nociones ya conocidas sobre expresiones booleanas (MacLane y Birkhoff (1967), Abott (1969), Halmos (1967), Gill (1976), etc).

Se partirá de una clase $\mathcal{A}=\{A_1, A_2,..., A_r\}$, que llamaremos clase generadora, de r subconjuntos de uno universal Ω.

Definición 3.1.- *(Conjunto generado por una clase de conjuntos)*- Se dirá que S es un conjunto generado por la clase $\mathcal{A}$ sii se puede escribir utilizando las operaciones unión, intersección y/o complementación de conjuntos de dicha clase. Llamaremos clase $\mathcal{G}$ generada por la clase $\mathcal{A}$ a la constituida por todos los conjuntos generados por $\mathcal{A}$.

Se supondrá además que se trata del caso más general, es decir, que esta clase genera por intersecciones 2^r conjuntos diferentes no vacios.

Definición 3.2.- *(Minset)* - Dada la clase $\mathcal{A}$, todo conjunto de la forma

$$\bigcap_{i=1}^{r} A_i^{\delta_i} \quad \text{donde} \quad A_i^{\delta_i} = \begin{cases} \overline{A_i} & \text{si } \delta_i = 0 \\ A_i & \text{si } \delta_i = 1 \end{cases}$$

se dirá que es un minset. A estos conjuntos se les llamará también conjuntos básicos.

La clase de los minsets, que se denotará por $\mathcal{B}$, constituye una partición del conjunto Ω.

Proposición 3.1.- La intersección de dos minsets distintos es vacia.

Demostración.-

Se consideran dos minsets $M_1 = \bigcap_{i=1}^{r} A_i^{\delta_i}$ y $M_2 = \bigcap_{i=1}^{r} A_i^{\delta_i^*}$.

Si M_1 y M_2 son distintos existe un índice K, $1 \le K \le r$, de forma que $\delta_k \neq \delta_k^*$ y, por tanto, $A_k^{\delta_k} \cap A_k^{\delta_k^*} = \Phi$.

Proposición 3.2.- La unión de todos los minsets es el conjunto universal Ω.

Demostración.-

Se tiene

$$\Omega = \bigcap_{i=1}^{r} (A_i^0 \cup A_i^1) = \bigcup_{\{\delta_i \in \{0,1\} / i = 1,\ldots r\}} \left(A_1^{\delta_1} \cap .. \cap A_r^{\delta_r} \right)$$

donde la segunda igualdad se obtiene por inducción sobre r utilizando la propiedad distributiva de la intersección de conjuntos respecto de la unión.

Teorema 3.1.-*(Forma normal tipo minset)* -Todo conjunto S generado por la clase $\mathcal{A}$ es igual al conjunto vacío o se puede expresar mediante uniones de distintos minsets. A esta última expresión se le llama forma normal tipo minset.

Demostración.-

En primer lugar supongamos que el conjunto S es un elemento de la clase $\mathcal{A}$, es decir $S=A_i$ para algún i.

Por la proposición 3.2, Ω se puede escribir como $\bigcup_{i=1}^{l} M_i$ siendo cada M_i un minset. Intersecando con A_i ambos miembros de la igualdad se tiene

$$A_i = \bigcup_{j=1}^{l} (M_j \cap A_i)$$

Para cada j,

$$M_j \cap A_i = \begin{cases} \Phi \text{ si } M_j \subset A_i^0 \\ \\ M_j \text{ si } M_j \subset A_i^1 \end{cases}$$

luego A_i es la unión de aquellos minsets que contienen a A_i^1 como elemento de la intersección.

Para demostrar el teorema es suficiente probar ahora que dados S_1 y S_2 expresables como uniones de distintos minsets entonces los conjuntos $S_1 \cap S_2$, $S_1 \cup S_2$ y $\overline{S_1}$ verifican la tesis del teorema.

Supongamos

$$S_1 = \bigcup_{i=1}^{m} M_{1i} \quad \text{y} \quad S_2 = \bigcup_{i=1}^{n} M_{2i}$$

* $S_1 \cap S_2 = \left(\bigcup_{i=1}^{m} M_{1i}\right) \cap \left(\bigcup_{i=1}^{n} M_{2i}\right) = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n} (M_{1i} \cap M_{2j})$

Como por la proposición 3.1 la intersección de dos minsets distintos es vacía se tiene trivialmente que $S_1 \cap S_2$ es o bien el conjunto vacío o puede escribirse como unión de minsets distintos.

* $S_1 \cup S_2 = \left(\bigcup_{i=1}^{m} M_{1i}\right) \cup \left(\bigcup_{i=1}^{n} M_{2i}\right)$

* $\overline{S_1} = \bigcap_{i=1}^{m} \overline{M_{1i}} = \bigcap_{i=1}^{m} \left(\overline{\bigcap_{j=1}^{r} A_j^{\delta_{ji}}}\right) = \bigcap_{i=1}^{m} \left(\bigcup_{j=1}^{r} \overline{A_j^{\delta_{ji}}}\right)$

Realizando operaciones conjuntistas y teniendo en cuenta que $A_i^0 \cup A_i^1 = \Omega$ y añadiendo el término $A_i^0 \cup A_i^1$ a aquellas intersecciones en las que no aparece el índice i se obtiene que S_1 se puede poner como unión de minsets distintos.

Definición 3.3.-*(Maxset)* - Dada la clase $\mathcal{A}$, todo conjunto de la forma

$$\bigcup_{i=1}^{r} A_i^{\delta_i} \quad \text{donde} \quad A_i^{\delta_i} = \begin{cases} A_i & \text{si } \delta_i = 0 \\ A_i & \text{si } \delta_i = 1 \end{cases}$$

se dirá que es un maxset.

Teorema 3.2.-*(Forma normal tipo maxset)* -Todo conjunto S generado por la clase $\mathcal{A}$ es igual al conjunto Ω o se puede expresar mediante intersecciones de distintos maxsets. A esta última expresión se le llama forma normal tipo maxset.

Demostración.-

En efecto, si S es distinto de Ω entonces su complementario es distinto de vacío y por el teorema 3.1 se puede escribir como unión de minsets, $\overline{S} = \bigcup_{i=1}^{l} M_i$. Por lo tanto,

$$S=\bigcap_{i=1}^{l}\overline{M_i}=\bigcap_{i=1}^{l}\overline{\left(\bigcap_{j=1}^{r}A_j^{\delta_{ji}}\right)}=\bigcap_{i=1}^{l}\left(\bigcup_{j=1}^{r}A_j^{\overline{\delta_{ji}}}\right)=\bigcap_{i=1}^{l}\left(\bigcup_{j=1}^{r}A_j^{\delta^*_{ij}}\right)$$

donde

$$\delta^*_{ij}=\begin{cases}1 \text{ si } \delta_{ij}=0\\ \\ 0 \text{ si } \delta_{ij}=1\end{cases}$$

Teorema 3.3.-*(Unicidad de formas normales)* - Las expresiones en forma normal son únicas salvo una permutación de sus elementos. Además, el número de términos de la forma normal tipo minset más el número de términos de la forma normal tipo maxset para un mismo conjunto S es 2^r.

Demostración.-

* Unicidad forma normal tipo minset- Supongamos que S se puede escribir como

$$S=\bigcup_{i=s_1}^{l_1}M_i=\bigcup_{i=s_2}^{l_2}M_i^*$$

y consideramos los conjuntos

$$C_{S1}=\{M_i/i=s_1,\ldots,l_1\}\ \text{ y }\ C_{S2}=\{M_i^*/i=s_2,\ldots,l_2\}$$

que están formados por minsets distintos y no vacíos, entonces los conjuntos C_{S1} y C_{S2} coinciden.

En efecto, sea M_i^* en C_{S2} entonces

$$M_i^*=S\cap M_i^*=\bigcup_{j=s_1}^{l_1}\left(M_j\cap M_i^*\right)$$

Por la proposición 3.1 y teniendo en cuenta que M_i^* no es vacío se tiene que debe existir un índice j, $s_1 \le j \le l_1$, de manera que $M_i^*=M_j$.

Análogamente, para cada M_i en C_{S1} existe M_j^* elemento de C_{S2} que coincide con él.

* Unicidad forma normal tipo maxset.- Supongamos que S se puede escribir como

$$S = \bigcap_{i=s_1}^{l_1} m_i = \bigcap_{i=s_2}^{l_2} m_i^*$$

entonces,

$$\overline{S} = \bigcup_{i=s_1}^{l_1} M_i = \bigcup_{i=s_2}^{l_2} M_i^*$$

donde $M_i = \overline{m_i}$ y $M_i^* = \overline{m_i^*}$ son minsets.

Aplicando la unicidad tipo minset al complementario de S se tiene que la descomposición en forma normal tipo maxset es única salvo ordenación de sus elementos.

Nótese que se ha probado además que el número de términos de la forma normal tipo maxset de un conjunto S coincide con el número de términos de la forma normal tipo minset de $\overline{S}$. Un razonamiento análogo nos permitiría probar que el número de términos de la forma normal tipo minset de S coincide con el número de términos de la forma normal tipo maxset de $\overline{S}$.

* Dado un conjunto arbitrario S podemos descomponer Ω como la unión de S y su complementario. Teniendo en cuenta además la proposición 3.2 se tiene

$$\Omega = S \cup \overline{S} = \left(\bigcup_{i=s_1}^{l_1} M_i \right) \cup \left(\bigcup_{i=s_2}^{l_2} M_i \right) = \bigcup_{i=1}^{2^r} M_i$$

La intersección de un minset correspondiente a la descomposición de S y uno correspondiente a $\overline{S}$ es vacía ya que $S \cap \overline{S} = \phi$. Por tanto, el conjunto Ω se puede escribir de dos formas como unión de minsets distintos. Aplicando la unicidad de la descomposición y la observación anterior se concluye que para S la suma del número de términos de las formas normales tipo minset y maxset resulta 2^r.

Definición 3.4.-*(Conjunto pseudobásico)* - Llamaremos conjunto pseudobásico a todo aquel conjunto que se puede expresar como

$$\bigcap_{i=1}^{r} A_i^{\delta_i} \quad \text{donde} \quad A_i^{\delta_i} = \begin{cases} \overline{A_i} & \text{si } \delta_i = 0 \\ A_i & \text{si } \delta_i = 1 \\ \Omega & \text{si } \delta_i = 2 \end{cases}$$

A la clase de los conjuntos pseudo-básicos la llamaremos $\mathfrak{P}$.

Definición 3.5.-*(Orden de un conjunto pseudobásico)* - Diremos que un conjunto pseudobásico es de orden k sii k de los r elementos $A_i^{\delta_i}$ son diferentes de Ω.

Nótese que los minsets son conjuntos pseudo-básicos de orden r y, por tanto, estos conjuntos son una generalización de los minsets.

Teorema 3.4.-*(Cardinal de la clase de los conjuntos pseudo-básicos)* - Existen 3^r conjuntos pseudo-básicos.

Demostración.-

En efecto, se tienen tres grados de libertad por cada uno de los términos i-ésimos, resultando, por tanto, 3^r conjuntos. Como comprobación, puede utilizarse el teorema 3.1, para sumar todos los conjuntos pseudo-básicos de todos los órdenes, resultando

$$\sum_{k=0}^{r} \binom{r}{k} 2^k = (1+2)^r = 3^r$$

Teorema 3.5.-*(Cardinal de la clase de los conjuntos pseudo-básicos de orden k)*- Existen $\binom{r}{k}2^k$ conjuntos pseudo-básicos diferentes de orden k.

Demostración.-

En efecto, el número de combinaciones de r elementos tomados k a k es el número de elementos distintos de Ω en el conjunto pseudobásico. Una vez elegidos estos k elementos, por poder tomarse ya sólo dos opciones, resultan 2^k posibles soluciones distintas.

Teorema 3.6.-*(Cota del cardinal de la clase de los conjuntos generados por $\mathcal{A}$)* - El cardinal del conjunto generado por la clase $\mathcal{A}$ es a lo sumo 2^{2^r}. Esta cota es alcanzable cuando todos los conjuntos básicos o minsets son diferentes.

Demostración.-

Todo conjunto generado por $\mathcal{A}$ se puede expresar como unión de los 2^r minsets y, por constituir éstos una partición, resulta que su número es

$$\binom{2^r}{1}+\binom{2^r}{2}+\ldots+\binom{2^r}{2^r}=(1+1)^{2^r}=2^{2^r}$$

Nótese que los sumandos del miembro de la izquierda representan los conjuntos que se obtienen mediante uniones de uno, dos, ... 2^r conjuntos.

Corolario 3.1.-(*Limitación de los conjuntos pseudo-básicos*) - No todo conjunto generado por la clase $\mathcal{A}$ es un conjunto pseudobásico.

Dado que, según el teorema 3.6, el número de conjuntos generados por $\mathcal{A}$ diferentes posibles es 2^{2^r}, existen el mismo número de funciones booleanas distintas basadas en la clase $\mathcal{A}$.

La Tabla 3.1 muestra la cardinalidad de las clases $\mathfrak{B}$, $\mathfrak{P}$ y $\mathfrak{G}$ para distintos valores de r.

Tamaño del problema	Cardinalidad		
r	$\lvert\mathfrak{B}\rvert=2^r$	$\lvert\mathfrak{P}\rvert=3^r$	$\lvert\mathfrak{G}\rvert=2^{2^r}$
2	4	9	16
3	8	27	256
4	16	81	65536
10	1024	59049	—

Tabla 3.1- Cardinalidad de las clases de conjuntos $\mathfrak{B}$, $\mathfrak{P}$ y $\mathfrak{G}$.

A continuación se estudiará las representaciones mínimas de los conjuntos generados por $\mathcal{A}$. En principio sabemos, según los teoremas 3.1, 3.2 y 3.3 que todo conjunto generado por la clase $\mathcal{A}$ se puede expresar mediante uniones o intersecciones de a lo sumo 2^r minsets o maxsets, respectivamente. Se trata, por tanto, de buscar representaciones que utilicen menos conjuntos. El siguiente teorema original nos da una cota del número mínimo de conjuntos pseudo-básicos

que son necesarios para obtener cualquier elemento de la clase $\mathcal{A}$. En lo que sigue utilizaremos la clase de los conjuntos pseudo-básicos.

Teorema 3.7.-*(Cota del número mínimo de conjuntos pseudo-básicos necesarios para la representación)* - El número mínimo de conjuntos pseudo-básicos que es necesario unir para obtener cualquier conjunto de la clase generada por $\mathcal{A}$ es 2^{r-1} para $r \geq 2$.

Demostración.-

* En primer lugar se probará que todo conjunto generado por A se puede poner como unión de a lo sumo 2^{r-1} conjuntos pseudo-básicos.Se aplicará inducción sobre r.

Sea $S \in \mathfrak{A}$. Por el teorema 3.1, S se puede descomponer en unión de a lo sumo 2^r conjuntos básicos,

$$S = \bigcup_{i=1}^{2^r} \left(\bigcap_{j=1}^{r} A_{ij} \right) \quad \text{donde} \quad A_{ij} \in \left\{ A_j, \overline{A_j} \right\}, i = 1,..,2^r$$

Podemos suponer sin pérdida de generalidad que los L_s primeros minsets de la descomposición de S están contenidos en A_1 y el resto en su complementario. De este modo

$$S = \left[\bigcup_{i=1}^{L_s} (A_1 \cap A_{i2} \cap .. \cap A_{ir}) \right] \cup \left[\bigcup_{i=L_s+1}^{2^r} (\overline{A_1} \cap A_{i2} \cap .. \cap A_{ir}) \right] =$$

$$= \left[A_1 \cap \left\{ \bigcup_{i=1}^{L_s} (A_{i2} \cap .. \cap A_{ir}) \right\} \right] \cup \left[\overline{A_1} \cap \left\{ \bigcup_{i=L_s+1}^{2^r} (A_{i2} \cap .. \cap A_{ir}) \right\} \right]$$

Caso r=2.

La unión que aparece entre llaves se reduce a un único conjunto, A_2, $\overline{A_2}$ u Ω, pudiendo cada uno de ellos combinarse con A_1 o su complementario. Por lo tanto, cualquier conjunto S de la clase $\mathfrak{A}$ se puede poner como unión de 2 conjuntos pseudo-básicos.

Caso general.
En este caso, la primera llave que aparece en la fórmula anterior es la unión de L_S conjuntos básicos en la clase $\mathcal{A}' = \{A_2, A_3, ..., A_r\}$. Por hipótesis de inducción esta unión se puede escribir como 2^{r-2} conjuntos pseudo-básicos

formados a partir de $\mathcal{A}'$. Cada uno de estos conjuntos pseudo-básicos se puede combinar con A_1 o su complementario luego uniendo $2^*2^{r-2}= 2^{r-1}$ conjuntos pseudo-básicos pueden obtenerse todos los conjuntos de $\mathcal{G}$.

* Existen 2^{r-1} conjuntos básicos que no pueden ponerse como unión de L conjuntos pseudo-básicos con $L<2^{r-1}$.Se aplicará inducción sobre r.

Caso r=2.
Se consideran las clases de conjuntos

$$G_1=\left\{A_1\cap\overline{A_2}\ ,\ \overline{A_1}\cap A_2\right\}\ \text{y}\ G_2=\mathcal{B}\setminus G_1$$

Los conjuntos de G_1 y G_2 son básicos y no pueden escribirse como unión de un número de pseudo-básicos menor que dos.

Caso general.
Sea $\mathcal{A}_r=\{A_1, A_2,..., A_r\}$ la clase generadora. Por hipótesis de inducción se pueden considerar las clases de conjuntos

$$G_1^{r-1}=\left\{M_1^{r-1},...,M_{2^{r-2}}^{r-1}\right\}\ \text{y}\ G_2^{r-1}=\mathcal{B}^{r-1}\setminus G_1^{r-1}$$

donde $G_1{}^{r-1}$ está formado por 2^{r-2} conjuntos básicos cuya unión no se reduce a un número inferior de pseudo-básicos y $\mathcal{B}^{r-1}$ la clase de todos los conjuntos básicos para $\mathcal{A}_{r-1}=\{A_1, A_2,..., A_{r-1}\}$

Como en $\mathcal{A}_r$ la unión de dos conjuntos básicos se reduce a un pseudobásico unicamente si tienen r-1 A_i iguales, no existen $L<2^{r-1}$ pseudo-básicos cuya unión sea la unión de los dos 2^{r-1} minsets en $G_1{}^r$

$$G_1^r=\left\{M\cap A_r\,/\,M\in G_1^{r-1}\right\}\cup\left\{M\cap\overline{A_r}\,/\,M\in G_2^{r-1}\right\}$$

Ejemplo 3.1.- Consideramos r=4 y la clase generadora de Ω está formada por $\mathcal{A}_4=\{A,B,C,D\}$. Los conjuntos G_1^3 y G_2^3 están formados por

$$G_1^3=\{A\cap\overline{B}\cap C,\overline{A}\cap B\cap C,A\cap B\cap\overline{C},\overline{A}\cap\overline{B}\cap\overline{C}\}$$
$$G_2^3=\mathcal{B}^3\setminus G_1^3=\{A\cap\overline{B}\cap\overline{C},\overline{A}\cap B\cap\overline{C},\overline{A}\cap\overline{B}\cap\overline{C},A\cap B\cap C\}$$

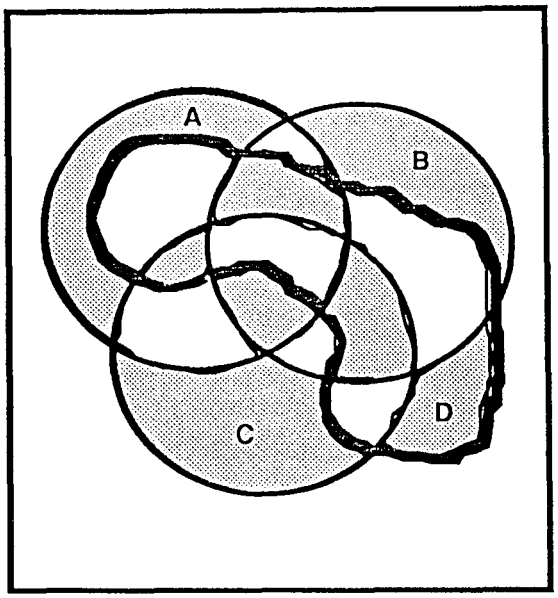


Figura 3.1.- Representación de G_1^4 y G_2^4

y a partir de ellos se obtiene G_1^4 intersecando cada uno de los minsets de G_1^3 con D y los de G_2^3 con $\overline{D}$ (Figura 3.1)

$$G_1^4 = \{A \cap \overline{B} \cap C \cap D, \overline{A} \cap B \cap C \cap D, A \cap B \cap \overline{C} \cap D, \overline{A} \cap \overline{B} \cap \overline{C} \cap D\}$$
$$\cup \{A \cap \overline{B} \cap \overline{C} \cap \overline{D}, \overline{A} \cap B \cap \overline{C} \cap \overline{D}, \overline{A} \cap \overline{B} \cap \overline{C} \cap \overline{D}, A \cap B \cap C \cap \overline{D}\}$$

¤

3.1.1.1.-Ordenes en $\mathfrak{B}$, $\mathfrak{P}$ y $\mathfrak{G}$

Desde el punto de vista computacional tiene interés la representación de conjuntos de cualquiera de las clases $\mathfrak{B}$, $\mathfrak{P}$ y $\mathfrak{G}$ mediante uniones de conjuntos de cualesquiera de las otras dos (Figura 3.2).

Se estudiarán los algoritmos que permiten los cambios de representación basados en diferentes ordenaciones que se definirán en las clases $\mathfrak{B}$, $\mathfrak{P}$ y $\mathfrak{G}$.

El apartado 7.2 incluye el listado del programa con los algoritmos anteriores para cambio de representación.

* Ordenación en $\mathfrak{B}$.

Puesto que un conjunto básico se expresa como intersección de los conjuntos $A_i^{\delta_i}$ para δ_i cero o uno, cada conjunto básico se puede expresar como un número en base dos, formado por r dígitos correspondiendo el i-ésimo al conjunto A_i y tomando el valor δ_i .

De esta manera, dado $S \in \mathfrak{B}$, $S = \bigcap_{i=1}^{r} A_i^{\delta_i}$ con $\delta_i \in \{0,1\}, i = 1,..,r$
Entonces el número de orden de S según la ordenación en $\mathfrak{B}$ es:

$$|S|_1 = \sum_{i=1}^{r} \delta_i 2^{i-1} \in \{0,..,2^r - 1\}$$

La clase de los conjuntos básicos es un conjunto totalmente ordenado según el orden implícito por esta ordenación.

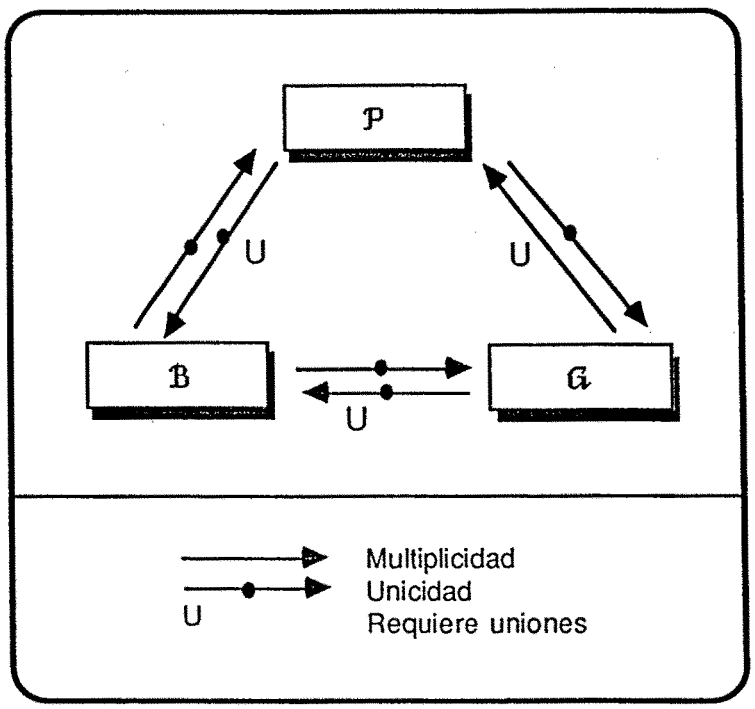


Figura 3.2.- Esquema de los cambios de representaciones en las clases $\mathfrak{B}$, $\mathfrak{P}$ y $\mathfrak{G}$.

* Ordenación en $\mathfrak{P}$.

Análogamente a la ordenación en la clase de conjuntos básicos, la clase $\mathfrak{P}$ puede ordenarse utilizando la numeración en base tres, asociaremos a cada

conjunto pseudobásico, formado por la intersección de los conjuntos $A_i^{\delta_i}$ con δ_i cero, uno o dos, un número de r dígitos correspondiendo el i-ésimo al conjunto A_i y tomando el valor δ_i.

De esta manera, dado $S \in \mathcal{P}$, $S = \bigcap_{i=1}^{r} A_i^{\delta_i}$ con $\delta_i \in \{0,1,2\}$, $i = 1,..,r$.
Entonces el número de orden de S según la ordenación en $\mathcal{P}$ es:

$$|S|_2 = \sum_{i=1}^{r} \delta_i 3^{i-1} \in \{0,..,3^r - 1\}$$

La clase de los conjuntos pseudo-básicos es un conjunto totalmente ordenado según el orden implicito por esta ordenación.

* Ordenación en $\mathcal{G}$.

Puesto que todo conjunto generado por $\mathcal{A}$ se puede representar de forma única mediante la unión de conjuntos básicos, existe una correspondencia biunívoca entre este conjunto y el conjunto de los números de 2^r dígitos en base 2, representando el i-ésimo dígito el i-ésimo conjunto básico según el orden dado en $\mathcal{B}$.

Supondremos para ello los conjuntos básicos ordenados según el orden en $\mathcal{B}$, es decir, si i<j entonces $|M_i|_1 < |M_j|_1$. Sea $S \in \mathcal{G}$,

$$S = \bigcup_{i=1}^{2^r} M_{i-1}^{\delta_i} \quad \text{con} \quad M_{i-1}^{\delta_i} = \left\{ \begin{array}{ll} \phi & \text{si } \delta_i = 0 \\ M_{i-1} & \text{si } \delta_i = 1 \end{array} \right\}, \; M_{i-1} \in \mathcal{B}, \; i = 1,..,2^r$$

Entonces el número de orden de S según el orden de $\mathcal{G}$ es

$$|S|_3 = \sum_{i=1}^{2^r} \delta_i 2^{i-1} \in \left\{0,..,2^{2^r} - 1\right\}$$

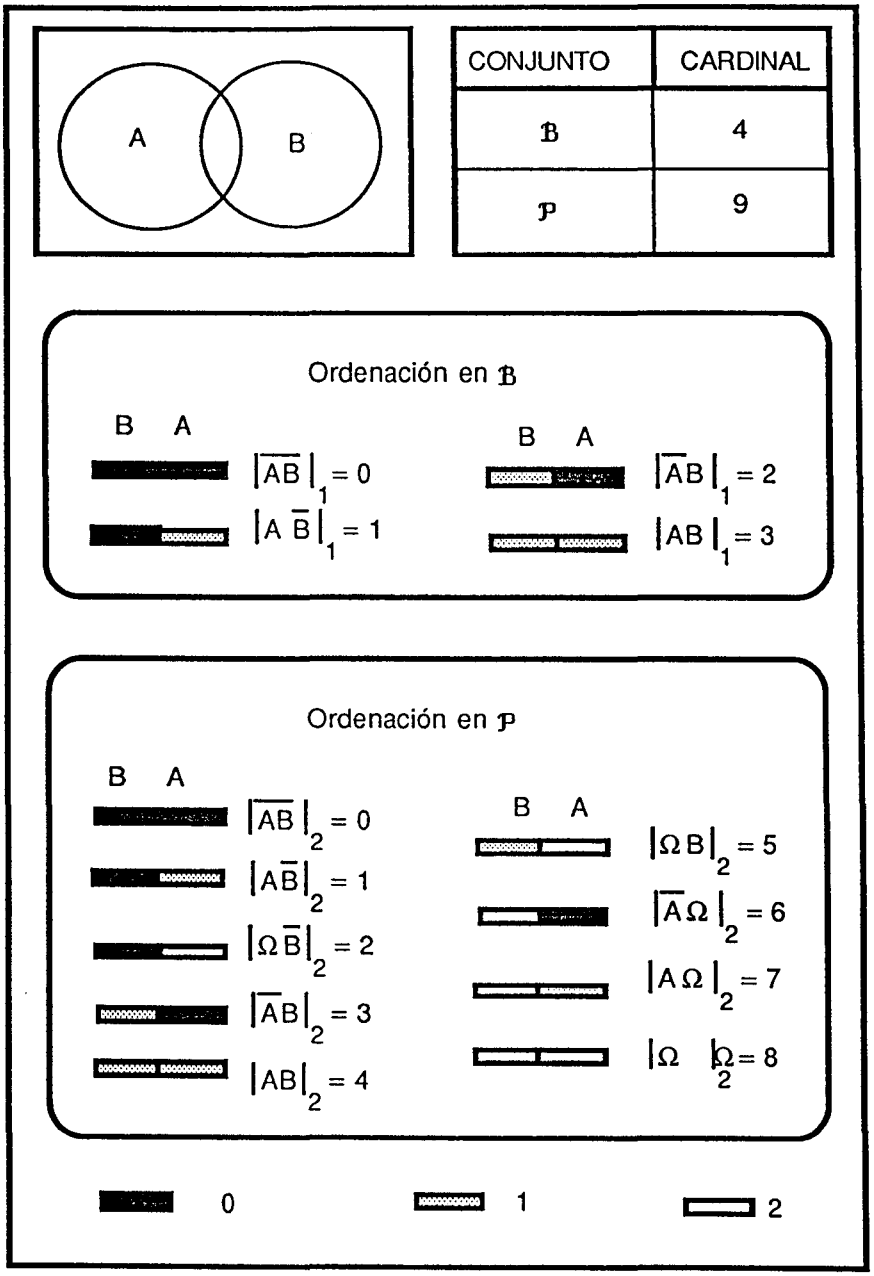


Figura 3.3.- Ordenación en $\mathfrak{B}$ y $\mathfrak{P}$.

Ejemplo 3.2.- Consideremos r=2, y $\mathcal{A}$={A, B }. La figura 3.3 muestra los conjuntos de las clases $\mathfrak{B}$ y $\mathfrak{P}$ así como los números de orden según las ordenaciones en estas clases.

Si tomamos el conjunto en $\mathfrak{G}$ siguiente

$$S = A\overline{B} \cup AB = M_1 \cup M_3$$

entonces el número de orden según la ordenación en $\mathfrak{G}$ es

$$|S|_3 = \sum_{i=1}^{2^r} \delta_i 2^{i-1} = 10 \text{ , ya que } \delta_i = \begin{cases} 0 \text{ si } i \in \{1,3\} \\ 1 \text{ si } i \in \{2,4\} \end{cases}$$

¤

Algoritmo 1.- *(Paso de $\mathcal{B}$ a $\mathcal{P}$)* - Sea $S \in \mathcal{B}$, el algoritmo consta de las siguientes etapas:

* Escribir $|S|_1$ en base 2, $|S|_1 = \sum_{i=1}^{r} \delta_i 2^{i-1}$, con $\delta_i = \{0, 1\}$, $i = 1,..,r$

* Leer los dígitos que resulten del paso anterior en base 3 para obtener el número de orden según la ordenación en $\mathcal{P}$, $|S|_2 = \sum_{i=1}^{r} \delta_i 3^{i-1}$, con $\delta_i = \{0, 1\}$, $i = 1,..,r$.

Corolario 3.2.- Dado $S \in \mathcal{B}$ se tiene $|S|_1 \leq |S|_2$.

Algoritmo 2.- *(Paso de $\mathcal{P}$ a $\mathcal{B}$)* - Sea $S \in \mathcal{P}$, el algoritmo consta de las siguientes etapas:

* Escribir $|S|_2$ en base 3, $|S|_2 = \sum_{i=1}^{r} \delta_i 3^{i-1}$, con $\delta_i = \{0, 1, 2\}$, $i = 1,..,r$.

* Sea $I = \{1 \leq i \leq r / \delta_i = 2\}$, entonces S se puede escribir como unión de conjuntos básicos de la forma siguiente:

$$S = \bigcup_{\mu_i \in \{0,1\}} \left[\left(\bigcap_{i \notin I} A_i^{\delta_i} \right) \cap \left(\bigcap_{i \in I} A_i^{\mu_i} \right) \right]$$

Algoritmo 3.- *(Paso de $\mathcal{B}$ a $\mathcal{G}$)* - Sea $S \in \mathcal{B}$, se tiene que $|S|_3 = 2^{|S|_1}$.

Algoritmo 4.- *(Paso de $\mathcal{G}$ a $\mathcal{B}$)* - Sea $S \in \mathcal{B}$, el algoritmo consta de las siguientes etapas:

* Escribir $|S|_3$ en base 2, $|S|_3 = \sum_{i=1}^{2^r} \delta_i 2^{i-1}$, con $\delta_i = \{0,1\}$, $i = 1,..,r$

* Entonces si suponemos la clase de conjuntos básicos ordenados según el orden en $\mathfrak{B}$, S se puede escribir como

$$S = \bigcup_{\left\{1 \le i \le 2^r \,/\, \delta_i = 1\right\}} M_{i-1}$$

3.1.2.-Compilación de reglas

En muchas aplicaciones se parte de un conjunto de datos y se pretende llegar mediante la aplicación de las reglas a concluir determinados objetivos. Cuando tanto los datos como los objetivos son conocidos, puede tener interés compilar las reglas, es decir, obtener los objetivos en función de los datos mediante ecuaciones. Con ello, tanto el encadenamiento hacia atrás como el encadenamiento hacia adelante se simplifican notablemente. En efecto, el encadenamiento hacia adelante consiste en evaluar las expresiones cuyos términos son hechos conocidos y el encadenamiento hacia atrás consiste en evaluar la expresión correspondiente al objetivo elegido, preguntando por los valores o hechos desconocidos.

Se considera entonces un sistema constituido por N objetos, A_1, A_2,..., A_N. Se admitirá que cada objeto A_i puede tomar los siguientes valores: cierto (A_i), falso ($\overline{A_i}$) o desconocido (Ω), representándolo de la forma

$$A_i^{\delta_i} = \left\{ \begin{array}{ll} A_i & \text{si } \delta_i = 1 \\ \overline{A_i} & \text{si } \delta_i = 0 \\ \Omega & \text{si } \delta_i = 2 \end{array} \right\}$$

Se considerará que las reglas que relacionan estos objetos verifican las tres hipótesis siguientes

(H1) Las premisas se expresan mediante uniones de intersecciones de objetos que toman valores ciertos o falsos. Por tanto, serán de la forma

$$\bigcup_j \left(\bigcap_{k=1}^{N} A_k^{\delta_k^j} \right)$$

(H2) Las conclusiones serán simples, es decir, se concluirá unicamente sobre el valor (cierto o falso) de un único objeto.

(H3) Las reglas que concluyen un determinado valor del objeto se pueden disponer en forma de árbol.

Definición 3.6.- *(Compilación de reglas*)- Dado un conjunto de reglas se dice que está compilado cuando el valor de cada objeto queda determinado mediante una sóla regla del mismo tipo.

3.1.2.1-Modelo de equivalencia de reglas

En este modelo se considerará además de las hipótesis (H1), (H2) y (H3) que las premisas de la regla determinan biunívocamente el estado del objeto de la conclusión. Por tanto, se considerará que una regla es de la forma

$$(3.1) \qquad R_s = \left\{ a_{ij}^s \in \{0, 1, -1\} \,/\, i = 1, 2, \ldots, k_s;\ j = 1, 2, \ldots, N \right\} \text{ con } s \in \{1, 2, \ldots, N\}$$

cuyo significado es el siguiente

$$(3.2) \qquad A_s = \bigcup_{i=1}^{k_s} \left[\left(\bigcap_{j \in C_i^s} A_j \right) \cap \left(\bigcap_{j \in F_i^s} \overline{A}_j \right) \right]$$

con

$$(3.3) \qquad C_i^s = \left\{ j \in \{1, 2, \ldots, N\} \,/\, a_{ij}^s = 1 \right\} \text{ y } F_i^s = \left\{ j \in \{1, 2, \ldots, N\} \,/\, a_{ij}^s = -1 \right\}$$

Nótese que definida una regla que concluya el valor del objeto A_s como cierto queda perfectamente definida la regla que concluye A_s como falso de la forma

$$(3.4) \qquad \overline{A}_s = \bigcap_{i=1}^{k_s} \left[\left(\bigcup_{j \in C_i^s} \overline{A}_j \right) \cup \left(\bigcup_{j \in F_i^s} A_j \right) \right]$$

Definimos una transformación del conjunto de reglas llamado "sustitución", que consiste en sustituir los términos de la izquierda de (3.2) en su derecha,

resultando el compilado de A_S, sin más que aplicar la propiedad distributiva de la unión respecto de la intersección

$$(3.5)\quad A_S^* = \bigcup_{i=1}^{k_S} \left\{ \left(\bigcap_{j \in C_i^S} \left[\bigcup_{k=1}^{k_j} \left[\left(\bigcap_{l \in C_k^j} A_l \right) \cap \left(\bigcap_{l \in F_k^j} \overline{A_l} \right) \right] \right] \right) \cap \right.$$

$$\left. \cap \left(\bigcap_{j \in F_i^S} \left[\bigcap_{k=1}^{k_j} \left[\left(\bigcup_{l \in C_k^j} \overline{A_l} \right) \cup \left(\bigcup_{l \in F_k^j} A_l \right) \right] \right] \right) \right\} =$$

operando

$$= \bigcup_{i=1}^{k_S} \left\{ \left(\bigcup_{k=1}^{k_j} \left[\bigcap_{j \in C_i^S} \left[\left(\bigcap_{l \in C_k^j} A_l \right) \cap \left(\bigcap_{l \in F_k^j} \overline{A_l} \right) \right] \right] \right) \cap \right.$$

$$\left. \cap \left(\bigcap_{j \in F_i^S} \left[\left(\bigcup_{l \in C_k^j} \left[\bigcap_{k=1}^{k_j} \overline{A_l} \right] \right) \cup \left(\bigcup_{l \in F_k^j} \left[\bigcap_{k=1}^{k_j} A_l \right] \right) \right] \right) \right\} =$$

$$= \bigcup_{i=1}^{k_S} \left\{ \left(\bigcup_{k=1}^{k_j} \left[\left(\bigcap_{j \in C_i^S} \bigcap_{l \in C_k^j} A_l \right) \cap \left(\bigcap_{j \in C_i^S} \bigcap_{l \in F_k^j} \overline{A_l} \right) \right] \right) \cap \right.$$

$$\left. \cap \left(\left[\bigcup_{l \in C_k^j} \left[\bigcap_{j \in F_i^S} \left[\bigcap_{k=1}^{k_j} \overline{A_l} \right] \right] \right] \cup \left[\bigcup_{l \in F_k^j} \left[\bigcap_{j \in F_i^S} \left[\bigcap_{k=1}^{k_j} A_l \right] \right] \right] \right) \right\} =$$

$$= \bigcup_{i=1}^{k_s} \left\{ \left[\left[\left(\bigcup_{k=1}^{k_j} \left[\left(\bigcap_{j \in C_i^s} \bigcap_{l \in C_k^j} A_l \right) \cap \left(\bigcap_{j \in C_i^s} \bigcap_{l \in F_k^j} \overline{A_l} \right) \right] \right) \cap \left[\bigcup_{l \in C_k^j} \left[\bigcap_{j \in F_i^s} \left[\bigcap_{k=1}^{k_j} \overline{A_l} \right] \right] \right] \right] \cup \right.$$

$$\left. \cup \left[\left(\bigcup_{k=1}^{k_j} \left[\left(\bigcap_{j \in C_i^s} \bigcap_{l \in C_k^j} A_l \right) \cap \left(\bigcap_{j \in C_i^s} \bigcap_{l \in F_k^j} \overline{A_l} \right) \right] \right) \cap \left[\bigcup_{l \in F_k^j} \left[\bigcap_{j \in F_i^s} \left[\bigcap_{k=1}^{k_j} A_l \right] \right] \right] \right] \right\} =$$

volviendo a utilizar la propiedad distributiva de la unión respecto de la intersección

$$= \left\{ \bigcup_{i=1}^{k_s} \bigcup_{k=1}^{k_j} \bigcup_{l \in C_k^j} \left[\left(\bigcap_{j \in C_i^s} \bigcap_{l \in C_k^j} A_l \right) \cap \left(\bigcap_{j \in C_i^s} \bigcap_{l \in F_k^j} \overline{A_l} \right) \cap \left(\bigcap_{j \in F_i^s} \bigcap_{k=1}^{k_j} \overline{A_l} \right) \right] \right\} \cup$$

$$\cup \left\{ \bigcup_{i=1}^{k_s} \bigcup_{k=1}^{k_j} \bigcup_{l \in F_k^j} \left[\left(\bigcap_{j \in C_i^s} \bigcap_{l \in C_k^j} A_l \right) \cap \left(\bigcap_{j \in C_i^s} \bigcap_{l \in F_k^j} \overline{A_l} \right) \cap \left(\bigcap_{j \in F_i^s} \bigcap_{k=1}^{k_j} A_l \right) \right] \right\} =$$

3.1.2.2-Modelo de implicaciones de reglas

En este modelo se considerará que la verificación de las premisas de una regla es suficiente para establecer la conclusión pero no tiene porqué darse el recíproco. Se utilizará la representación

$$(3.6) \qquad \bigcup_{j \in I_i} \left(\bigcap_{k=1}^{N} A_k^{\delta_k^j} \right) \subseteq A_i \subseteq \bigcup_{j \in D_i} \left(\bigcap_{k=1}^{N} A_k^{\beta_k^j} \right)$$

cuyo significado es,

$$\forall j \in I_i \quad \bigcap_{k=1}^{N} A_k^{\delta_k^j} \Rightarrow A_i$$

$$\forall j \in D_i \quad A_i \Rightarrow \bigcap_{k=1}^{N} A_k^{\beta_k^j} \text{ , es decir, } \bigcup_{k=1}^{N} A_k^{\beta_k^{*j}} \Rightarrow \overline{A_i}$$

para representar en una única "regla" todas aquellas que concluyen un determinado objeto bien como cierto o como falso. Nótese que los conjuntos de índices I_i y D_i pueden ser para algún i el vacío.

Definición 3.7.-(*Centro de una regla*).- Se llama centro de la regla dada como (3.6) al objeto A_i.

Se define una transformación del conjunto de reglas llamado "compilación", que consiste en incluir a la izquierda (derecha) de todas las reglas que no tienen como centro A_i los conjuntos pseudobásicos que resultan de sustituir los términos de la izquierda (derecha) en los que aparezca A_i ó $\overline{A_i}$ por la parte izquierda o derecha de la regla cuyo centro es A_i. Para obtener las transformadas en la forma dada en (3.6) basta aplicar posteriormente la propiedad distributiva de la unión respecto de la intersección. El proceso de compilación permitirá obtener las reglas compiladas cuando al repetir esta transformación no varía ninguna regla.

Supongamos que tenemos dos reglas

$$\bigcup_{j \in I_l} \left(\bigcap_{K=1}^{N} A_k^{\delta_k^j} \right) \subseteq A_l \subseteq \bigcup_{j \in D_l} \left(\bigcap_{K=1}^{N} A_k^{\beta_k^j} \right)$$

(3.7)

$$\bigcup_{j \in I_i} \left(\bigcap_{K=1}^{N} A_k^{\gamma_k^j} \right) \subseteq A_i \subseteq \bigcup_{j \in D_i} \left(\bigcap_{K=1}^{N} A_k^{\eta_k^j} \right)$$

y queremos compilarlas. Para ello definimos previamente los conjuntos siguientes

$$I_{il}^1 = \{ j \in I_i / \delta_l = 1 \} \qquad D_{il}^1 = \{ j \in D_i / \delta_l = 1 \}$$

$$I_{il}^0 = \{ j \in I_i / \delta_l = 0 \} \qquad D_{il}^0 = \{ j \in D_i / \delta_l = 0 \}$$

$$I_{il}^{2} = \{ j \in I_i / \delta_l = 2 \} \qquad D_{il}^{2} = \{ j \in D_i / \delta_l = 2 \}$$

Teniendo en cuenta estos conjuntos podemos expresar la regla asociada al objeto A_i de la forma

$$\bigcup_{j \in I_{il}^{1}} \left(\bigcap_{K=1}^{N} A_k^{\gamma_k^j} \right) \cup \bigcup_{j \in I_{il}^{0}} \left(\bigcap_{K=1}^{N} A_k^{\gamma_k^j} \right) \cup \bigcup_{j \in I_{il}^{2}} \left(\bigcap_{K=1}^{N} A_k^{\gamma_k^j} \right) \subseteq A_i$$

$$A_i \subseteq \bigcup_{j \in D_{il}^{1}} \left(\bigcap_{K=1}^{N} A_k^{\eta_k^j} \right) \cup \bigcup_{j \in D_{il}^{0}} \left(\bigcap_{K=1}^{N} A_k^{\eta_k^j} \right) \cup \bigcup_{j \in D_{il}^{2}} \left(\bigcap_{K=1}^{N} A_k^{\eta_k^j} \right)$$

Distinguimos los siguientes casos a la hora de transformar la parte izquierda de la regla

Caso 1.- Supongamos $I_l \neq \phi$ y $D_l \neq \phi$ entonces

$$\bigcup_{j \in I_{il}^{1}} \left(\bigcap_{K=1}^{N} A_k^{\gamma_k^j} \right) \supseteq \bigcup_{j \in I_{il}^{1}} \left[\left(\bigcap_{\substack{K=1 \\ k \neq l}}^{N} A_k^{\gamma_k^j} \right) \cap \left(\bigcup_{j \in I_l} \left(\bigcap_{K=1}^{N} A_k^{\delta_k^j} \right) \right) \right] =$$

(3. 8)

$$= \bigcup_{j \in I_{il}^{1}} \left[\bigcup_{j \in I_l} \left[\left(\bigcap_{\substack{K=1 \\ k \neq l}}^{N} A_k^{\gamma_k^j} \right) \cap \left(\bigcap_{K=1}^{N} A_k^{\delta_k^j} \right) \right] \right]$$

del mismo modo

$$\bigcup_{j \in I_{il}^{0}} \left(\bigcap_{K=1}^{N} A_k^{\gamma_k^j} \right) \supseteq \bigcup_{j \in I_{il}^{0}} \left[\left(\bigcap_{\substack{K=1 \\ k \neq l}}^{N} A_k^{\gamma_k^j} \right) \cap \left(\bigcap_{j \in D_l} \left(\bigcup_{k=1}^{N} A_k^{\beta_k^{*j}} \right) \right) \right] =$$

(3. 9)

$$= \bigcup_{j \in I_{il}^{0}} \left[\bigcup_{j \in I_l} \left[\bigcup_{k=1}^{N} \ldots \bigcup_{k=1}^{N} \left(A_k^{\gamma_k^j} \cap A_k^{\beta_k^{*j1}} \ldots \cap A_k^{\beta_k^{*js}} \right) \right] \right]$$

y, por tanto, la parte izquierda de la regla transformada es

$$(3.10)\quad \left\{ \bigcup_{j\in I_{il}^{1}} \left(\bigcap_{K=1}^{N} A_k^{\gamma_k^j} \right) \right\} \cup \left\{ \bigcup_{j\in I_{il}^{1}} \bigcup_{j\in I_l} \left[\left(\bigcap_{\substack{K=1 \\ k\neq l}}^{N} A_k^{\gamma_k^j} \right) \cap \left(\bigcap_{K=1}^{N} A_k^{\delta_k^j} \right) \right] \right\} \cup$$
$$\cup \left\{ \bigcup_{j\in I_{il}^{0}} \left(\bigcap_{K=1}^{N} A_k^{\gamma_k^j} \right) \right\} \cup \left\{ \bigcup_{j\in I_{il}^{0}} \bigcup_{j\in I_l} \bigcup_{k=1}^{N} \dots \bigcup_{k=1}^{N} \left(A_k^{\gamma_k^j} \cap A_k^{\beta_k^{*j1}} \dots \cap A_k^{\beta_k^{*js}} \right) \right\} \cup$$
$$\cup \left\{ \bigcup_{j\in I_{il}^{2}} \left(\bigcap_{K=1}^{N} A_k^{\gamma_k^j} \right) \right\}$$

Caso 2.- Supongamos $I_l \neq \phi$ y $D_l = \phi$ entonces los únicos pseudobásicos que hay que incluir a la izquierda de la regla de centro A_i es (3.8).

Caso 3- Supongamos $I_l = \phi$ y $D_l \neq \phi$ entonces los únicos pseudobásicos que hay que incluir a la izquierda de la regla de centro A_i es (3.9).

Análogamente habría que distinguir tres casos a la hora de realizar la transformación a la derecha de la regla.

Nótese que el proceso de compilación de dos reglas es cerrado, obteniendo como regla transformada una regla del tipo definido en (3.6). Este proceso de compilación termina cuando al compilar cada una de las reglas con el resto no se produce ningún cambio en las reglas. En este momento se dirá que las reglas están compiladas.

Ejemplo 3.3.- Supongamos el conjunto de reglas de la figura 3. 4.

Estas reglas pueden escribirse de la siguiente forma

$$\begin{array}{l} A \cap B \subseteq E \\ C \cap D \cap E \subseteq I \\ (F \cap G) \cup H \subseteq J \\ I \cap J \subseteq K \end{array} \qquad (3.11)$$

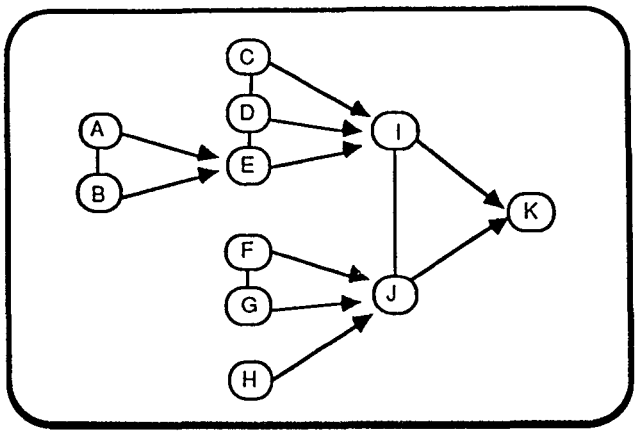


Figura 3.4.- Ejemplo de reglas

Se compila cada una de las reglas con las otras dos obteniendo

$$
\begin{aligned}
& A \cap B \subseteq E \\
& (C \cap D \cap E) \cup (A \cap B \cap C \cap D) \subseteq I \\
& (F \cap G) \cup H \subseteq J \\
& (I \cap J) \cup (C \cap D \cap E \cap J) \cup (F \cap G \cap I) \cup (I \cap H) \subseteq K
\end{aligned}
\qquad (3.12)
$$

volviendo a realizar la misma trasformación

$$
\begin{aligned}
& A \cap B \subseteq E \\
& (C \cap D \cap E) \cup (A \cap B \cap C \cap D) \subseteq I \\
& (F \cap G) \cup H \subseteq J \\
& (I \cap J) \cup (C \cap D \cap E \cap J) \cup (A \cap B \cap C \cap D \cap J) \cup \\
& \quad \cup (C \cap D \cap E \cap F \cap G) \cup (C \cap D \cap E \cap H) \cup (F \cap G \cap I) \cup \\
& \quad \cup (C \cap D \cap E \cap F \cap G) \cup (A \cap B \cap C \cap D \cap F \cap G) \cup (I \cap H) \cup \\
& \quad \cup (C \cap D \cap E \cap H) \cup (A \cap B \cap C \cap D \cap H) \subseteq K
\end{aligned}
\qquad (3.13)
$$

que dan lugar al conjunto compilado de reglas de la figura 3.4. Nótese que las reglas compiladas nos permite conocer nuevos hechos sin necesidad de realizar un encadenamiento. Así, por ejemplo, si nuestra base de datos estuviera formada por los siguientes hechos A cierto, B falso, F cierto y G cierto se tendría por la tercera y cuarta regla compilada que J y k toman el valor cierto.

¤

3.1.3.-Coherencia de reglas

El conjunto de reglas de un sistema experto pretenden describir la realidad si la modelización es correcta entonces las reglas no son nunca contradictorias.

Sin embargo, al aplicar las reglas dada una base de hechos se pueden obtener valores contradictorios para ciertos objetos. Esto significa que, en este caso, las premisas de las reglas que nos conducen a estas contradiciones no pueden darse simultáneamente.

Dado un conjunto de reglas y unos hechos se puede analizar si son coherentes comprobando, después de realizar la compilación del conjunto de reglas en sí mismo, si dada una regla compilada para todo conjunto pseudobásico que permite concluir el objeto de dicha regla existe o no un pseudobásico de la parte derecha de la misma regla que no lo contradice.

Definición 3.8.-*(Pseudobásicos contradictorios).-* Dados dos conjuntos pseudobásicos

$$M_1 = \bigcap_{i=1}^{N} A_i^{\delta_i} \quad y \quad M_2 = \bigcap_{i=1}^{N} A_i^{\beta_i}$$

diremos que M_2 no contradice a M_1 si para cada $i \in \{1,..., N\}$ con $\delta_i \neq 2$ se cumple $\beta_i = 2$ ó $\beta_i = \delta_i$.

Ejemplo 3.4.- Supongamos que tenemos las reglas

$$\begin{array}{l} (A \cap B) \subseteq C \\ (D \cap E \cap F) \subseteq G \\ (C \cup G) \subseteq I \subseteq \overline{G} \cup \overline{H} \end{array} \qquad (3.14)$$

y los hechos A, B, G y H ciertos. La compilación de las reglas (3.14) es

$$\begin{array}{l} (A \cap B) \subseteq C \\ (D \cap E \cap F) \subseteq G \\ (C \cap G) \cup (A \cap B \cap \overline{G}) \cup (C \cap D \cap E \cap \overline{F}) \cup (A \cap B \cap D \cap E \cap F) \subseteq I \\ \qquad \subseteq (\overline{G} \cup \overline{H}) \cup (\overline{D} \cup \overline{E} \cup \overline{F} \cup \overline{H}) \end{array} \qquad (3.15)$$

El pseudobásico $(A \cap B \cap G \cap H)$ permite concluir el objeto I que se contradice con los dos pseudobásicos de la derecha de la regla. ¤

3.2.-MODELOS PROBABILÍSTICOS

Con objeto de ilustrar las incoherencias que pueden surgir al dar probabilidades al modelo, se presentan a continuación unos gráficos aclaratorios. La figura 3.5 muestra dos conjuntos A y B y las restricciones que deben satisfacer las probabilidades de la unión e intersección de ambos. Puede ocurrir que se

pregunte al experto humano por los valores de P(A) y de P(B) inicialmente, y una vez dados éstos se solicite el valor de P(A∪B) o P(A∩B). Si el experto desconoce estas restricciones, puede ocurrir que suministre valores que las violen, con lo que se habrán violado también los axiomas de la probabilidad. Nótese que estas violaciones pueden conducir a graves problemas en el mecanismo de propagación de la incertidumbre.

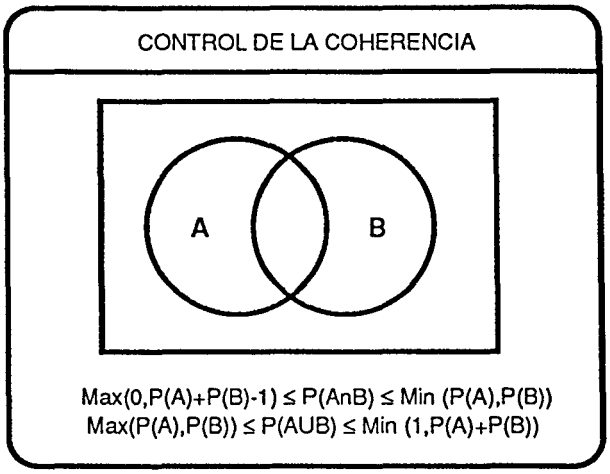


Figura 3.5.- Restricciones impuestas a las probabilidadesde la unión e intersección de dos conjuntos

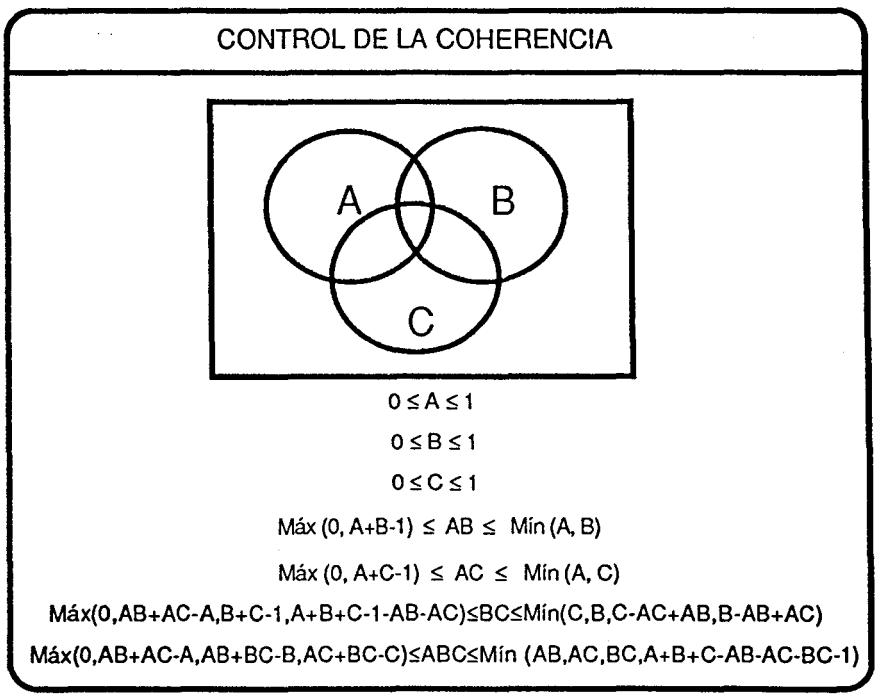


Figura 3.6.- Restricciones impuestas a las probabilidades en el caso de tres conjuntos

En el caso de dos conjuntos A y B, las restricciones que resultan son triviales, pero si el número de conjuntos es mayor, es casi imposible que un

experto humano pueda controlarlas, pues son bastante complejas, como lo demuestra la figura 3.6, en la que se dan las restricciones que se han de cumplir al preguntar sucesivamente por los valores de P(A), P(B), P(C), P(A∩B), P(A∩C), P(B∩C) y P(A∩B∩C). En la figura se ha denotado a estos valores por A, B, C, AB, AC, BC y ABC, respectivamente. A medida que el número de síntomas va aumentando la complejidad de estas acotaciones también crece. Por todo ello, es conveniente que se controle la coherencia (Smith (1961)), para lo cual el subsistema de adquisición de conocimiento puede suministrar al experto humano el intervalo $[P_{min}(A),P_{max}(A)]$ antes de que éste asigne valores a P(A) y comprobar luego que el valor suministrado está dentro del intervalo, no permitiendo incoherencias. El apartado 7.3 incluye el listado del programa que da las restricciones a las que debe someterse la probabilidad para uniones e intersecciones de conjuntos.

3.2.1.-Validación del modelo mediante ordenación de pseudobásicos

En este apartado se supone que el experto humano es preguntado por la probabilidad asociada a ciertos conjuntos pseudobásicos de una clase generadora $\mathcal{A}$. Se requerirá del experto la probabilidad de los conjuntos pseudobásicos en un orden determinado, suministrándole previamente un intervalo dentro del cual debe dar esta medida para evitar la incoherencia con los valores ya dados anterioremente. Se considera la clase $\mathcal{A}$ formada por r subconjuntos de uno universal Ω, $\mathcal{A}=\{A_1, A_2,..., A_r\}$, y llamemos

$$
\begin{aligned}
&M_0 = \bigcap_{s=1}^{r} A_s \qquad \text{con } \delta_s = 1 \quad s = 1,..,r \\
&M_i = \bigcap_{s=1}^{r} A_s^{\delta_s} \qquad \text{con } \delta_i = 0 \quad \delta_j = 1 \quad j \neq i \\
&M_{ij} = \bigcap_{s=1}^{r} A_s^{\delta_s} \qquad \text{con } \delta_i = \delta_j = 0 \quad \delta_k = 1 \quad k \notin \{i,j\} \quad i < j \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&M_{i_1 i_2 \ldots i_k} = \bigcap_{s=1}^{r} A_s^{\delta_s} \qquad \text{con } \delta_{i_1} = \delta_{i_2} = \ldots = \delta_{i_k} = 0 \quad \delta_k = 1 \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad k \notin \{i_1, i_2,..., i_k\} \quad i_1 < i_2 < \ldots < i_k \\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&M_{123\ldots r} = \bigcap_{s=1}^{r} A_s^{\delta_s} \qquad \text{con } \delta_s = 0 \quad s = 1,..,r
\end{aligned}
\tag{3.16}
$$

Nótese que los conjuntos básicos definidos en (3.18) constituyen los 2^r que se pueden generar dada $\mathcal{A}$.

Se considerarán estos conjuntos básicos en el mismo orden en el que se han definido anteriormente, es decir,

$$(3.17)\qquad \mathfrak{M} = \{ M_0, M_1, M_2, \ldots, M_r, M_{12}, M_{13}, \ldots, M_{1r}, M_{23}, \ldots, M_{2r}, \ldots, M_{12\ldots r} \}$$

que los podemos escribir como un vector columna

$$(3.18)\qquad B = (M_0 M_1 M_2 \ldots M_r M_{12} M_{13} \ldots, M_{1r} M_{23} \ldots M_{2r} \ldots M_{12\ldots r})^T$$

y denotaremos por $\overline{M}_j$ al j-ésimo minset en la ordenación dada en (3.17).

Análogamente consideramos 2^r conjuntos pseudobásicos definidos como

$$(3.19)\qquad \begin{aligned}
&P_0 = \bigcap_{s=1}^{r} A_s \qquad \text{con } \delta_s = 1 \quad s = 1,\ldots,r \\
&P_i = \bigcap_{s=1}^{r} A_s^{\delta_s} \qquad \text{con } \delta_i = 2 \quad \delta_j = 1 \quad j \neq i \\
&P_{ij} = \bigcap_{s=1}^{r} A_s^{\delta_s} \qquad \text{con } \delta_i = \delta_j = 2 \quad \delta_k = 1 \quad k \notin \{i,j\} \quad i < j \\
&\text{---} \\
&P_{i_1 i_2 \ldots i_k} = \bigcap_{s=1}^{r} A_s^{\delta_s} \qquad \text{con } \delta_{i_1} = \delta_{i_2} = \ldots = \delta_{i_k} = 2 \quad \delta_k = 1 \\
&\qquad\qquad\qquad\qquad k \notin \{i_1, i_2, \ldots, i_k\} \quad i_1 < i_2 < \ldots < i_k \\
&\text{---} \\
&P_{123\ldots r} = \bigcap_{s=1}^{r} A_s^{\delta_s} \qquad \text{con } \delta_s = 0 \quad s = 1,\ldots,r
\end{aligned}$$

la ordenación que consideramos es

$$(3.20)\qquad \mathfrak{P} = \{ P_0, P_1, P_2, \ldots, P_r, P_{12}, P_{13}, \ldots, P_{1r}, P_{23}, \ldots, P_{2r}, \ldots, P_{12\ldots r} \}$$

y los escribiremos en forma vectorial como

(3.21) $$C = (P_0 P_1 P_2 \ldots P_r P_{12} P_{13} \ldots P_{1r} P_{23} \ldots P_{2r} \ldots P_{12\ldots r})^T$$

o

(3.22) $$C = \left(\overline{P_1}\ \overline{P_2}\ \overline{P_2} \ldots \overline{P_{2^r}}\right)^T$$

donde $\overline{P_j}$ es el j-ésimo conjunto pseudobásico según la ordenación dada en (3.20).

Dado un elemento en $\mathcal{P}$, por el teorema 3.1, sabemos que se puede expresar como unión de conjuntos básicos, es decir, como unión de elementos de $\mathcal{M}$. Para $P_{i_1 i_2 \ldots i_k} \in \mathcal{P}$ esta expresión es

$$P_{i_1 i_2 \ldots i_k} = \bigcup_{\substack{\delta_s = 1 \ s \notin I \\ \delta_s \in \{0,1\} \ s \in I}} \left(\bigcap_{s=1}^{r} A_s^{\delta_s} \right) =$$

(3.23)

$$= M_0 \cup \left(\bigcup_{s \in I} M_s \right) \cup \left(\bigcup_{\substack{s_1, s_2 \in I \\ s_1 < s_2}} M_{s_1 s_2} \right) \cup \ldots \cup M_{i_1 i_2 \ldots i_k}$$

donde $I=\{i_1, i_2, \ldots, i_k\}$. Por tanto, $\overline{P_j} = P_{i_1 i_2 \ldots i_k}$ se puede expresar como uniones de los primeros j minsets en el orden establecido apareciendo en esta expresión M_j. De esta manera cuando el experto asigna un valor de probabilidad a P_1 queda determinado M_1, cuando en una segunda etapa da la probabilidad de $\overline{P_2}$ queda fijado el valor de M_2, y así sucesivamente.

Nótese que por el teorema 3.3 la expresión (3.23) es única.

Supongamos que el experto ya sido preguntado por la probabilidad de $P_1, P_2, \ldots, P_j$ y que se pretende obtener las cotas superior e inferior de la

probabilidad del pseudobásico $\overline{P}_{j+1} = P_{i_1 i_2 \ldots i_k}$. En este momento las probabilidades de los j primeros minsets están ya fijadas y, por tanto, la probabilidad del conjunto $\overline{P}_{j+1}$ es un número real perteneciente al intervalo

(3.24) $$\left(\sum_{i \in I_{j1}} P(\overline{M}_i), 1 - \sum_{i \in I_{j2}} P(\overline{M}_i) \right)$$

donde

(3.25) $$I_{j1} = \left\{ i \in \{1,..,j\} / \overline{M}_i \cap \overline{P}_j \neq \phi \right\} \qquad I_{j2} = \left\{ i \in \{1,..,j\} / \overline{M}_i \cap \overline{P}_j = \phi \right\}$$

El extremo inferior representa la suma de las probabilidades de todos aquellos minsets contenidos en $\overline{P}_{j+1}$ cuyo valor está fijado por los datos previos dados por el experto y el extremo superior es la diferencia entre las probabilidad total y la suma de probabilidades de aquellos minsets cuyo valor está determinado y no interseca a $\overline{P}_{j+1}$.

Nótese que al final de este proceso, esto es, una vez que el experto asigne probabilidades a todos los conjuntos de $\mathcal{P}$, dentro de los intervalos correspondientes, se tienen fijadas las probabilidades de cada conjunto básico y teniendo en cuenta el teorema 3.1 también de cada conjunto generado por la clase $\mathcal{A}$. En efecto, la relación dada en (3.23) se puede expresar en forma matricial para todos los conjuntos de $\mathcal{P}$ como C=AB, donde A es una matriz de orden 2^r cuyos coeficientes son ceros o unos, que además es triangular inferior y que posee unos en la diagonal principal. Esto permite expresar todos los minsets en términos de elementos de $\mathcal{P}$.

Ejemplo 3.5.- Consideremos el caso en que r=3 y $\mathcal{A}=\{A_1,A_2,A_3\}$. Los conjuntos básicos y pseudobásicos ordenados son

$$\overline{M}_1 = M_0 = A_1^1 \cap A_2^1 \cap A_3^1 \qquad \overline{P}_1 = P_0 = A_1^1 \cap A_2^1 \cap A_3^1$$
$$\overline{M}_2 = M_1 = A_1^0 \cap A_2^1 \cap A_3^1 \qquad \overline{P}_2 = P_1 = A_1^2 \cap A_2^1 \cap A_3^1$$
$$\overline{M}_3 = M_2 = A_1^1 \cap A_2^0 \cap A_3^1 \qquad \overline{P}_3 = P_2 = A_1^1 \cap A_2^2 \cap A_3^1$$
$$\overline{M}_4 = M_3 = A_1^1 \cap A_2^1 \cap A_3^0 \qquad \overline{P}_4 = P_3 = A_1^1 \cap A_2^1 \cap A_3^2$$

$$\overline{M_5} = M_{12} = A_1^0 \cap A_2^0 \cap A_3^1 \qquad \overline{P_5} = P_{12} = A_1^2 \cap A_2^2 \cap A_3^1$$
$$\overline{M_6} = M_{13} = A_1^0 \cap A_2^1 \cap A_3^0 \qquad \overline{P_6} = P_{13} = A_1^2 \cap A_2^1 \cap A_3^2$$
$$\overline{M_7} = M_{23} = A_1^1 \cap A_2^0 \cap A_3^0 \qquad \overline{P_7} = P_{23} = A_1^1 \cap A_2^2 \cap A_3^2$$
$$\overline{M_8} = M_{123} = A_1^0 \cap A_2^0 \cap A_3^0 \qquad \overline{P_8} = P_{123} = A_1^0 \cap A_2^0 \cap A_3^0$$

Si se denota por

$$P(A_1^{\delta_1} \cap A_2^{\delta_2} \cap A_3^{\delta_3}) = A_1^{\delta_1} A_2^{\delta_2} A_3^{\delta_3}$$

entonces los intervalos que nos proporciona este método en cada etapa vienen dados en la Tabla 3.2

Pseudobásico	Intervalo
$\overline{P_1}$	$(0,1)$
$\overline{P_2}$	$(A_1^1 A_2^1 A_3^1, 1)$
$\overline{P_3}$	$(A_1^1 A_2^1 A_3^1, 1 - A_1^2 A_2^1 A_3^1 + A_1^1 A_2^1 A_3^1)$
$\overline{P_4}$	$(A_1^1 A_2^1 A_3^1, 1 - A_1^2 A_2^1 A_3^1 - A_1^1 A_2^2 A_3^1 + 2^* A_1^1 A_2^1 A_3^1)$
$\overline{P_5}$	$(A_1^1 A_2^2 A_3^1 + A_1^2 A_2^1 A_3^1 - A_1^1 A_2^2 A_3^1, 1 - A_1^1 A_2^1 A_3^2 + A_1^1 A_2^1 A_3^1)$
$\overline{P_6}$	$(A_1^1 A_2^1 A_3^2 + A_1^2 A_2^1 A_3^1 - A_1^1 A_2^2 A_3^1, 1 - A_1^2 A_2^2 A_3^1 + A_1^2 A_2^1 A_3^1)$
$\overline{P_7}$	$(A_1^1 A_2^1 A_3^2 + A_1^1 A_2^2 A_3^1 - A_1^1 A_2^2 A_3^1, 1 - A_1^2 A_2^1 A_3^2 - A_1^2 A_2^2 A_3^1 + A_1^1 A_2^1 A_3^2 + A_1^1 A_2^2 A_3^1 + A_1^2 A_2^1 A_3^1 - A_1^1 A_2^1 A_3^1)$

Tabla 3.2.- Intervalos de acotación de las probabilidades de los pseudobásicos para r=3.

¤

3.2.2.-Validación del modelo mediante programación lineal

Este apartado se centrará en los modelos GD y GR y se describirá una herramienta para ayudar a los expertos a mantener la coherencia a la hora de dar los parámetros o reglas iniciales.

Suponiendo que se ha seleccionado uno de los modelos de dependencia del capítulo 2, el objetivo del ingeniero del conocimiento será obtener de los expertos la máxima información posible sobre los parámetros. La forma más directa de hacerlo sería preguntando al experto sobre sus valores; sin embargo, en muchas ocasiones es díficil responder a ello, y el experto prefiere contestar a cuestiones alternativas más sencillas. Se supone en lo que sigue que el experto es preguntado por el valor exacto o el intervalo en el que se encuentre el valor exacto de $P(E_i \cap A)$ ó $P(E_i \cap A) / P(E_i \cap B)$ donde A y B pertenecen a la clase $\mathbf{C}$.

Nótese que estos valores son combinaciones lineales de los valores de los parámetros, y que las probabilidades condicionadas son de la segunda forma, de esta manera son incluidas como posibles preguntas. Por tanto, el conocimiento del experto, en forma de probabilidades, puede ser almacenado como un conjunto de restricciones del tipo

(3. 26) $$b \leq P(E_i \cap A) \leq c$$

o

(3. 27) $$b \leq P(E_i \cap A) / P(E_i \cap B) \leq c$$

donde b y c son constantes reales en el intervalo [0,1], y se supone que $P(E_i \cap B) > 0$. Nótese que para el modelo GD las desigualdades anteriores se pueden escribir como desigualdades de combinaciones lineales de los parámetros del modelo, es decir, (3.26) puede expresarse como

(3. 28) $$b \leq \sum_{D \in \mathbf{D}} \propto(D) P(E_i \cap D) \leq c$$

donde

$$\propto(D) = \begin{cases} 1 & \text{si} \quad A \cap D \neq \phi \\ 0 & \text{en otro caso} \end{cases}$$

y para (3.27) las desigualdades son

$$\sum_{D \in \mathfrak{D}} \gamma(D) P(E_i \cap D) \leq 0$$

(3.29)

$$\sum_{D \in \mathfrak{D}} \delta(D) P(E_i \cap D) \leq 0$$

donde

(3.30)
$$\gamma(D) = \begin{cases} 1 & \text{si } A \cap D \neq \phi \text{ y } B \cap D = \phi \\ 1-c & \text{si } A \cap D \neq \phi \text{ y } B \cap D \neq \phi \\ -c & \text{si } A \cap D = \phi \text{ y } B \cap D \neq \phi \\ 0 & \text{en otro caso} \end{cases}$$

(3.31)
$$\delta(D) = \begin{cases} -1 & \text{si } A \cap D \neq \phi \text{ y } B \cap D = \phi \\ b-1 & \text{si } A \cap D \neq \phi \text{ y } B \cap D \neq \phi \\ b & \text{si } A \cap D = \phi \text{ y } B \cap D \neq \phi \\ 0 & \text{en otro caso} \end{cases}$$

respectivamente.

El conjunto de restricciones anteriores, junto con las incluidas en (2.4) ó (2.12) según que el modelo sea GD o DR respectivamente, (restricciones iniciales), definen el conjunto de valores posibles para los parámetros una vez que se ha obtenido el conocimiento del experto.

Si el experto no es coherente, las restricciones anteriores conducen a un problema sin solución, por lo que deben redefinirse las restricciones. Para evitar este problema, antes de dar cada pieza de información, el experto es informado de los valores máximos y mínimos que conducen a soluciones factibles. Si se requiere del experto la probabilidad $P(E_i \cap A)$, esto puede hacerse mediante la resolución de los dos problemas de programación siguientes:

$$\underset{P(E_i \cap D)}{\text{Maximizar}} \sum_{D \in \mathfrak{D}} \alpha(D) P(E_i \cap D)$$

y

$$\underset{P(E_i \cap D)}{\text{Minimizar}} \sum_{D \in \mathfrak{D}} \alpha(D) P(E_i \cap D)$$

condicionados a las restricciones iniciales y las que resulten de la información previa del experto. Si, por el contrario, el usuario requiere el valor de

$P(E_i \cap A) / P(E_i \cap B)$, se necesita resolver los dos problemas no lineales siguientes:

$$\underset{P(E_i \cap D)}{\text{Maximizar}} \quad \sum_{D \in \mathfrak{D}} \alpha(D) P(E_i \cap D) / \sum_{D \in \mathfrak{D}} \beta(D) P(E_i \cap D)$$

y

$$\underset{P(E_i \cap D)}{\text{Minimizar}} \quad \sum_{D \in \mathfrak{D}} \alpha(D) P(E_i \cap D) / \sum_{D \in \mathfrak{D}} \beta(D) P(E_i \cap D)$$

donde

$$\beta(D) = \begin{cases} 1 & \text{si} \quad B \cap D \neq \phi \\ 0 & \text{en otro caso} \end{cases}$$

sometidos a las mismas restricciones.

Aunque estos dos últimos problemas no son lineales, pueden reducirse a la resolución iterativa de problemas lineales (programación lineal paramétrica) pues equivale a

$$\underset{\lambda}{\text{Max}} \left[\begin{array}{l} \text{Max} \sum_{D \in \mathfrak{D}} \alpha(D) P(E_i \cap D) / \lambda \quad \text{sometido a} \\ \text{las restricciones iniciales y a} \sum_{D \in \mathfrak{D}} \beta(D) P(E_i \cap D) = \lambda \end{array} \right]$$

(3.32)

$$\underset{\lambda}{\text{Min}} \left[\begin{array}{l} \text{Min} \sum_{D \in \mathfrak{D}} \alpha(D) P(E_i \cap D) / \lambda \quad \text{sometido a} \\ \text{las restricciones iniciales y a} \sum_{D \in \mathfrak{D}} \beta(D) P(E_i \cap D) = \lambda \end{array} \right]$$

Los dos problemas anteriores pueden resolverse mediante el método de la bisección tomando como intervalo inicial $[\lambda_{min}, \lambda_{max}]$ donde

$$\lambda_{min} = \text{Min} \sum_{D \in \mathfrak{D}} \beta(D) P(E_i \cap D) \text{ sujeto a las restricciones iniciales}$$

$$\lambda_{max} = \text{Max} \sum_{D \in \mathfrak{D}} \beta(D) P(E_i \cap D) \text{ sujeto a las restricciones iniciales}$$

Todo lo expuesto está referido al modelo GD. Para el caso del modelo GR, las expresiones (3.28) y (3.29) no son lineales en todos los parámetros. Pero si se

supone fijados p_{ijk} y que al experto no se le permite cambiarlos, para el modelo GR se tiene también un problema de programación lineal.

La herramienta descrita anteriormente es un verdadero motor de inferencia ya que el diagnóstico automático del sistema experto se reduce simplemente a calcular las probabilidades $P(E_i \cap A)$ que pueden ser formuladas como combinación lineal de los parámetros. Por tanto, el sistema experto daría un máximo y un mínimo para las probabilidades de cada problema, y, en función de éstos, se haría el diagnóstico. Si el usuario prefiere estimaciones puntuales, pueden ser utilizados intervalos degenerados (de amplitud cero).

El tamaño del problema de programación lineal asociado está relacionado estrechamente con el número de parámetros, así, si el último es alto también lo será el primero.

Nótese que las cotas superior e inferior así obtenidas son extremadamente útiles porque dan información sobre la incertidumbre y la ignorancia de $P(E_i \cap A)$ (ó $P(E_i \cap A) / P(E_i \cap B)$) en forma análoga a como lo hacía el intervalo [Cr(A),Pl(A)] en teoría de la evidencia. La incertidumbre se manifiesta si tanto la cota superior como la inferior son valores lejanos de la unidad o a cero y la ignorancia si la diferencia entre ambas cotas es grande.

Un diagnóstico correcto requiere certeza (cotas cercanas a uno o a cero) y ausencia de ignorancia (valores máximo y mínimo próximos entre sí). La aparición de intervalos de amplitud grande durante el proceso de inferencia nos indica una definición pobre de la medida de probabilidad, aunque este hecho puede ser inevitable en muchos casos prácticos y el que toma las decisiones debe acostumbrarse a trabajar con esta situación. Si los valores de la probabilidad que aparecen son lejanos a uno queda reflejado que no se tiene suficiente información (no se conocen suficientes síntomas para el diagnóstico), o que se carece de un adecuado conjunto de síntomas (los síntomas seleccionados no permiten una correcta distinción entre los problemas).

3.2.3.-El problema de la programación lineal visto como región admisible degenerada

El problema general de programación lineal (Gass (1969), Wu y Coppins (1981)) consiste en encontrar un vector $(x_1,x_2,...,x_n)$ que minimiza la expresión lineal

(3.33) $$c_1x_1 + c_2x_2 + \dots + c_nx_n$$

sometido a las restricciones lineales

(3.34) $$x_j \geq 0 \quad ; \quad j = 1, 2, \dots, n$$

y

(3.35)
$$\begin{aligned} a_{11}\, x_1 + a_{12}\, x_2 &+ \dots + a_{1n}\, x_n \leq b_1 \\ a_{21}\, x_1 + a_{22}\, x_2 &+ \dots + a_{2n}\, x_n \leq b_2 \\ \dots\ \dots\dots &\dots\dots\dots\dots \\ a_{m1}\, x_1 + a_{m2}\, x_2 &+ \dots + a_{mn}\, x_n \leq b_m \end{aligned}$$

donde a_{ij}, b_i y c_j (i=1,2,...,m ; j=1,2,...,n) son constantes dadas.

La figura 3.7 muestra un ejemplo, para n=2, en el que se muestran las restricciones y las curvas de nivel de la función a minimizar.

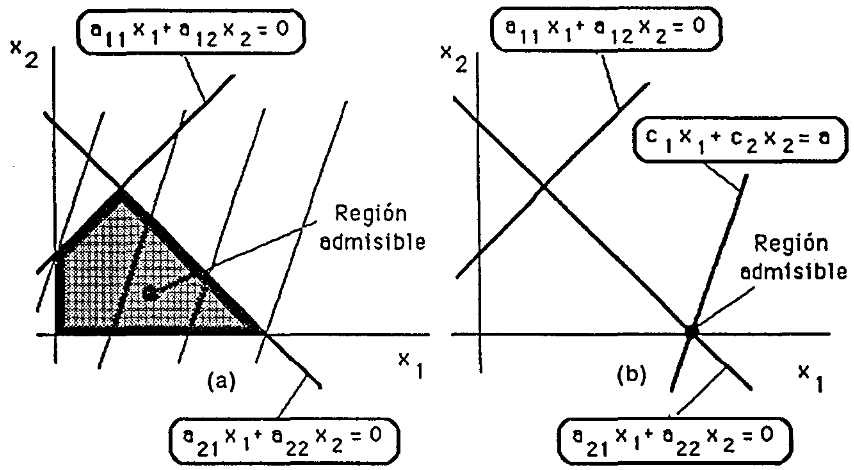


Figura 3.7.- Curvas de nivel de la función a minimizar y restricciones

Este problema se reduce normalmente a un problema equivalente en el que las desigualdades de (3.34) se convierten en igualdades mediante la introducción de las llamadas variables de holgura y se resuelve por métodos varios, como el simplex, por ejemplo. En lo que sigue supondremos que ya se tienen igualdades y que m<n.

De especial interés es el caso en el que uno o varios de los coeficientes que aparecen en la función a minimizar o en las restricciones son parámetros. A este problema, ya clásico en programación lineal, se le llama programación

paramétrica y tiene mucho interés en especial para estudiar la sensibilidad de los coeficientes que en él intervienen en la solución.

En este apartado tratamos de resaltar no los aspectos teóricos del método, sino los computacionales. Es sabido de todos la importancia que está adquiriendo el cálculo simbólico en la Matemática y la Inteligencia Artificial. Basta hacer referencia a las nuevas posibilidades que se abren con la utilización de paquetes de programas como el Macsyma, el Reduce o el Maple por citar dos de los más importantes. Sin embargo, dichos paquetes tienen aún grandes limitaciones, entre las que se encuentra la de no poder trabajar con desigualdades.

El problema anterior ((3.33) a (3.35)) puede ser planteado de forma equivalente en la siguiente forma: Elegir el valor del parámetro "a" que reduce a un punto la región:

$$
\begin{aligned}
& a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \le b_1 \\
& a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \le b_2 \\
& \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
& a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \le b_m \\
& c_1x_1 + c_2x_2 + \ldots + c_nx_n \le a \\
& \\
& x_j \ge 0 \quad ; \quad j = 1, 2, \ldots, n
\end{aligned}
\qquad (3.36)
$$

De esta forma, el enfoque del problema es totalmente diferente aún cuando se trata del mismo problema.

El método del simplex consta de dos etapas claramente diferenciadas: en la primera se busca una solución factible básica y en la segunda se mejora la solución hasta alcanzar la óptima. El planteamiento anterior reduce el problema anterior a la primera fase, si bien es a cambio de exigir trabajar con un parámetro desconocido, por lo que habrá que discutir diferentes alternativas.

Este método no pretende ser más sencillo ni siquiera emplear menos esfuerzo computacional, sino mostrar como puede hacerse el tratamiento paramétrico de forma automática. De forma análoga podría resolverse cualquier otro problema de programación lineal paramétrica de los clásicos.

La solución al problema planteado se ha realizado mediante el algoritmo del método simplex en su primera fase. Para ello, se utiliza la tabla siguiente

$$\begin{array}{cccc} c_1+\sum_i^m a_{i1} & c_2+\sum_i^m a_{i2} & \cdots & c_n+\sum_i^m a_{in} \quad a+\sum_i^m b_i \\ a_{11} & a_{12} & \cdots & a_{1n} \quad b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} \quad b_2 \\ & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \quad b_m \\ c_1 & c_2 & \cdots & c_n \quad a \end{array}$$

y se procede, como es costumbre a:

(a) determinar si todos los elementos de la fila primera, salvo el último, son menores o iguales que cero.
(b) si el resultado del apartado (a) es negativo, se selecciona la columna $K \leq n$ con mayor valor en la primera fila de la tabla.
(c) se selecciona la fila L tal que $a_{LK}>0$ y que da el menor cociente entre los valores de la última columna y la K.
(d) se procede al proceso de eliminación con pivote el elemento de la fila L y la columna K.

Este proceso se repite hasta que ocurre uno de estos casos:

1.- Se han eliminado todos los elementos de la base artificial (normalmente m iteraciones) o todos los elementos de la fila primera son no positivos y el último es nulo, en cuyo caso se comprueba que los elementos de la última columna son no negativos
2.- Todos los elementos de la primera fila son no positivos salvo el último, en cuyo caso el problema no admite solución admisible
3.- En la etapa (c) todos los elementos de la columna K son no positivos, en cuyo caso la solución no está acotada

Durante este proceso, los elementos de las n primeras columnas son constantes, pero los de la última columna dependen del valor del parámetro "a". Por ello, las etapas (a) y (b) no presentan ningún problema y la etapa (d) presenta el problema de que la eliminación puede conducir a expresiones lineales en el parámetro "a" (polinomio de grado 1 en a), por lo que hay que almacenar su coeficiente y el término independiente. Finalmente, la etapa (c) es más complicada, ya que hay que discutir los valores de "a" y considerar todas las alternativas posibles.

Ejemplo 3.6.- Con objeto de ilustrar el método se resuelve completamente el siguiente ejemplo:

Minimizar $2x_1 + 3x_2$

sometido a las restricciones

$$x_1 \geq 0 \;\; ; \;\; x_2 \geq 0$$
$$-x_1 + 2x_2 \leq 4$$
$$x_1 + x_2 \leq 6$$
$$x_1 + 3x_2 \geq 4$$

La figura 3.8 muestra los casos que van surgiendo al aplicar el algoritmo del simplex y si se optiene o no, solución.

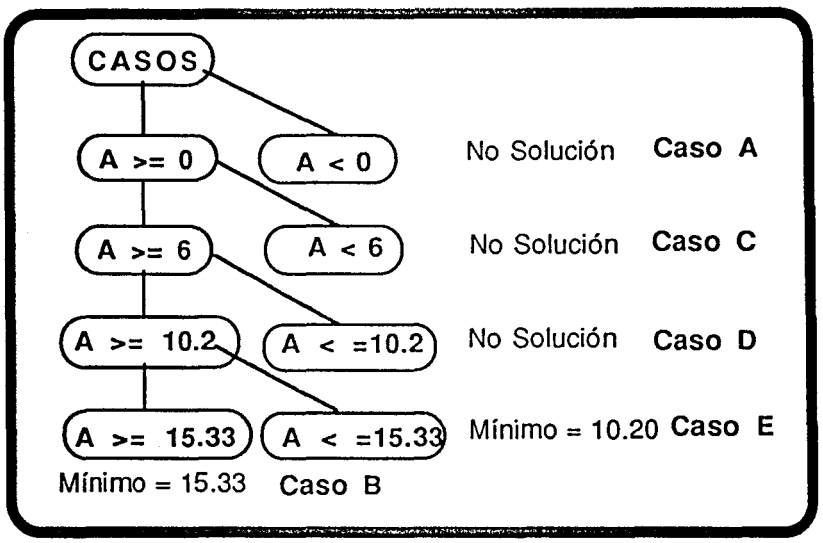


Figura 3.8.- Casos que surgen al resolver el ejemplo

A continuación se muestran las tablas que van surgiendo en el proceso y las restricciones que van acumulándose para el parámetro "a".

El apartado 7.4 incluye el listado del programa correspondiente a este ejemplo aplicando el método anteriormente explicado.

CASO A

$A \in (-\infty, 0)$

-1.00	3.00	1.00	1.00	-1.00 =	19.00 - A	
-1.00	2.00	1.00	0.00	0.00 =	4.00	
1.00	1.00	0.00	1.00	0.00 =	6.00	
1.00	3.00	0.00	0.00	-1.00 =	9.00	
-2.00	-3.00	-0.00	-0.00	-0.00 =	-0.00 - A	

$A \in (-\infty, 0)$

0.50	0.00	-0.50	1.00	-1.00 =	13.00 - A
-0.50	1.00	0.50	0.00	0.00 =	2.00
1.50	0.00	-0.50	1.00	0.00 =	4.00
2.50	0.00	-1.50	0.00	-1.00 =	3.00
-3.50	0.00	1.50	0.00	0.00 =	6.00 - A

$A \in (-\infty, 0)$

-1.00	0.00	0.00	0.00	-1.00 =	9.00 - A
-0.50	1.00	0.50	0.00	0.00 =	2.00
1.50	0.00	-0.50	1.00	0.00 =	4.00
2.50	0.00	-1.50	0.00	-1.00 =	3.00
-3.50	0.00	1.50	0.00	0.00 =	6.00 - A

Solución = 15.333para $X_1 = 2.66$, $X_2 = 3.33$

CASO B

$A \in [0, \infty)$

3.00	9.00	1.00	1.00	-1.00 =	19.00 + A
-1.00	2.00	1.00	0.00	0.00 =	4.00
1.00	1.00	0.00	1.00	0.00 =	6.00
1.00	3.00	0.00	0.00	-1.00 =	9.00
2.00	3.00	0.00	0.00	0.00 =	0.00 + A

$A \in [6, \infty)$

7.50	0.00	-3.50	1.00	-1.00 =	1.00 + A
-0.50	1.00	0.50	0.00	0.00 =	2.00
1.50	0.00	-0.50	1.00	0.00 =	4.00
2.50	0.00	-1.50	0.00	-1.00 =	3.00
3.50	0.00	-1.50	0.00	0.00 =	-6.00 + A

$A \in [10.2, \infty)$

0.00	0.00	1.00	1.00	2.00 =	-8.00 + A
0.00	1.00	0.20	0.00	-0.20 =	2.60
0.00	0.00	0.40	1.00	0.60 =	2.20
1.00	0.00	-0.60	0.00	-0.40 =	1.20
0.00	0.00	0.60	0.00	1.40 =	-10.20 + A

$A \in [15.333, \infty)$

0.00	0.00	-0.33	-2.33	0.00 =	-15.33 + A
0.00	1.00	0.33	0.33	0.00 =	3.33
0.00	0.00	0.67	1.67	1.00 =	3.67
1.00	0.00	-0.33	0.67	0.00 =	2.67
0.00	0.00	-0.33	-2.33	0.00 =	-15.33 + A

CASO C

$A \in [0, 6)$

3.00	9.00	1.00	1.00	-1.00 =	19.00 + A
-1.00	2.00	1.00	0.00	0.00 =	4.00
1.00	1.00	0.00	1.00	0.00 =	6.00
1.00	3.00	0.00	0.00	-1.00 =	9.00
2.00	3.00	0.00	0.00	0.00 =	0.00 + A

$A \in [0, 6)$

-3.00	0.00	1.00	1.00	-1.00 =	19.0 - 2A
-2.33	0.00	1.00	0.00	0.00 =	4.0 - 0.67 A
0.33	0.00	0.00	1.00	0.00 =	6.0 - 0.33 A
-1.00	0.00	0.00	0.00	-1.00 =	9.0 - A
0.67	1.00	0.00	0.00	0.00 =	0.0+0.33 A

$A \in [0, 6)$

-0.67	0.00	0.00	1.00	-1.00 =	15 - 1.33 A
-2.33	0.00	1.00	0.00	0.00 =	4 - 0.67 A
0.33	0.00	0.00	1.00	0.00 =	6 - 0.33 A
-1.00	0.00	0.00	0.00	-1.00 =	9 - A
0.67	1.00	0.00	0.00	0.00 =	0 + 0.33 A

$A \in [0, 6)$

-1.00	0.00	0.00	0.00	-1.00 =	9 - A
-2.33	0.00	1.00	0.00	0.00 =	4 - 0.67 A
0.33	0.00	0.00	1.00	0.00 =	6 - 0.33 A
-1.00	0.00	0.00	0.00	-1.00 =	9 - A
0.67	1.00	0.00	0.00	0.00 =	0 + 0.33 A

CASO D

$A \in [6, 10.20)$

7.50	0.00	-3.50	1.00	-1.00 =	1.00 + A
-0.50	1.00	0.50	0.00	0.00 =	2.00
1.50	0.00	-0.50	1.00	0.00 =	4.00
2.50	0.00	-1.50	0.00	-1.00 =	3.00
3.50	0.00	-1.50	0.00	0.00 =	-6.00 + A

$A \in [6, 10.20)$

0.00	0.00	-0.29	1.00	-1.00 =	13.9 - 1.14 A
0.00	1.00	0.29	0.00	0.00 =	1.14 + 0.14 A
0.00	0.00	0.14	1.00	0.00 =	6.57 - 0.43 A
0.00	0.00	-0.43	0.00	-1.00 =	7.29 - 0.71 A
1.00	0.00	-0.43	0.00	0.00 =	-1.71 + 0.29A

$A \in [6, 10.20)$

0.00	0.00	-0.43	0.00	-1.00 =	7.29 - 0.71 A
0.00	1.00	0.29	0.00	0.00 =	1.14 + 0.14 A
0.00	0.00	0.14	1.00	0.00 =	6.57 -0 .43 A
0.00	0.00	-0.43	0.00	-1.00 =	7.29 - 0.71 A
1.00	0.00	-0.43	0.00	0.00 =	-1.71 + 0.29A

CASO E

$A \in [10.20, 15.333)$

0.00	0.00	1.00	1.00	2.00 = -8.00+A
0.00	1.00	0.20	0.00	-0.20= 2.60
0.00	0.00	0.40	1.00	0.60= 2.20
1.00	0.00	-0.60	0.00	-0.40= 1.20
0.00	0.00	0.60	0.00	1.40= -10.20+A

$A \in [10.20, 15.333)$

0.00	0.00	0.14	1.00	0.00=6.57-0.43A
0.00	1.00	0.29	0.00	0.00=1.14+0.14A
0.00	0.00	0.14	1.00	0.00=6.57-0.43A
1.00	0.00	-0.43	0.00	0.00=-1.71+0.29A
0.00	0.00	0.43	0.00	1.00= -7.29+0.71A

$A \in [10.20, 15.333)$

0.00	0.00	-0.00	0.00	0.00 = 0.00 - 0.00 A
0.00	1.00	0.29	0.00	0.00 = 1.14 +0.14 A
0.00	0.00	0.14	1.00	0.00 = 6.57 - 0.43 A
1.00	0.00	-0.43	0.00	0.00 = -1.71 + 0.29 A
0.00	0.00	0.43	0.00	1.00 = -7.29 + 0.71 A

Solución =10.20 para $X_1 = 1.2$, $X_2 = 2.6$

MODELOS ESTADÍSTICOS

4.-MODELOS ESTADÍSTICOS

En capítulos anteriores se analizó el papel de la probabilidad en los sistemas expertos y se vió como puede utilizarse, no sólo para propagar incertidumbre, sino también para servir de base al motor de inferencia. En este capítulo se analizarán otras aportaciones de la Estadística a los sistemas expertos. En especial se estudiarán los modelos logarítmico-lineales, los métodos de regresión, los del análisis discriminante. Los sistemas expertos basados en estas técnicas estadísticas son diferentes de los ya vistos (los basados en reglas o en la probabilidad) y disponen de elementos distintos de los de éstos. En particular esta diferencia es muy notable en la forma de almacenar el conocimiento y en el motor de inferencia.

4.1.- MODELOS LOGARITMICO LINEALES

En el capítulo 2 se veía como el modelo de dependencia general, que es el más perfecto para reproducir la realidad, era prácticamente inviable por el elevado número de parámetros a que daba lugar. Poco después, el modelo de independencia se mostraba como inaceptable en muchos casos y, como solución intermedia para dar salida al problema, nacía el modelo de dependencia con síntomas relevantes. Sin embargo, para el caso de un problema con 10 síntomas relevantes binarios, el número de parámetros resultante es ya 1024, que es todavía bastante elevado. Por ello, sería bueno conseguir algún modelo que, sin llegar a ese número de parámetros, reprodujese la estructura de la probabilidad con suficiente fidelidad. Esto es lo que pretende el modelo logarítmico lineal.

El modelo logarítmico lineal (Bishop et al. (1975), Fienberg (1977), Darroch et al. (1980)) parte de una población dividida en clases mutuamente exclusivas (ningún individuo de la población pertenece a más de una clase) y exhaustivas (todo individuo de la población pertenece a una de ellas). Ello quiere decir que todo individuo pertenece a una y sólo a una de las clases. En el caso del diagnóstico, las clases son los diferentes conjuntos producidos por las intersecciones de las problemas y los síntomas, es decir, por los problemas y por las categorías de los síntomas (presencia o ausencia de un síntoma, o su nivel: alto, medio, bajo, etc.). Se denotará por $c_{ijk...}$ a la clase definida por el problema i-ésimo, el nivel j del primer síntoma, el nivel k del segundo síntoma, etc.

En la figura 4.1 se muestra un ejemplo de un problema, E, descrito mediante tres síntomas S_1, S_2 y S_3, con las 16 clases a que da lugar y sus notaciones

correspondientes (i=1: E, i=2: $\overline{E}$, j=1: S_1, j=2: $\overline{S_1}$, k=1: S_2, k=2: $\overline{S_2}$, l=1: S_3, l=2: $\overline{S_3}$).

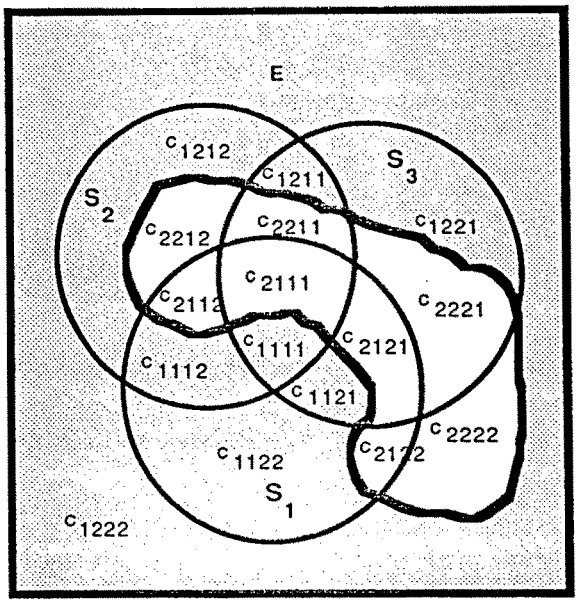


Figura 4.1.- Clases definidas por el problema E y los síntomas S_1, S_2 y S_3

Sea $p_{ijk...}$ la probabilidad de que un individuo pertenezca a la clase $c_{ijk...}$ y $x_{ijk...}$ su frecuencia. Entonces debe ser:

(4. 1) $$\sum_{ijk..} p_{ijk..} = 1$$

(4. 2) $$\sum_{ijk..} x_{ijk..} = N$$

donde N es el número total de individuos de la población.

Si se extrae una muestra de la población, las esperanzas de los números de individuos $m_{ijk...}$ pertenecientes a cada clase de la muestra resultan

(4. 3) $$m_{ijk..} = E(x_{ijk...}) = Np_{ijk..}$$

En la figura 4.2 se reproducen dos casos, a y b, con frecuencias asociadas a las 16 clases c_{ijkl} en el caso del problema E y los síntomas S_1, S_2 y S_3.

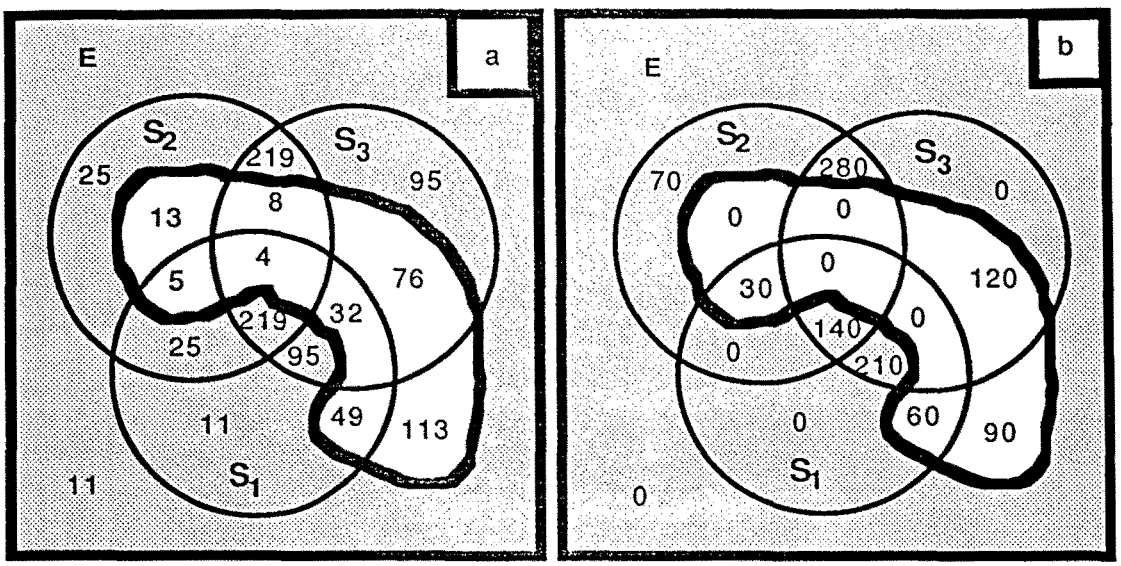


Figura 4.2.- Dos distribuciones de frecuencias para E.

Las frecuencias x_{ijkl} de esta figura pueden tomar valores libremente o estar condicionadas por el tipo de muestreo. Así, estos valores podrían haber resultado de los tres experimentos siguientes:

(a) toma de datos de los casos estudiados durante un tiempo determinado

(b) toma de datos de los 1000 casos

(c) toma de datos hasta conseguir 700 casos con el problema y 300 sin él

En el caso (a) las frecuencias son totalmente libres y no deben satisfacer ninguna restricción. En el caso (b), las frecuencias deben satisfacer la restricción

$$\sum_{ijkl} x_{ijkl} = 1000$$

Finalmente, en el caso (c) las restricciones son

$$\sum_{i=1} x_{ijkl} = 700 \quad ; \quad \sum_{i=2} x_{ijkl} = 300$$

Estas restricciones deben conocerse en cada caso concreto y son necesarias para estimar los parámetros del modelo logarítmico lineal.

El modelo logarítmico lineal general es de la forma (Bishop et al. (1975)):

$$\log m_{ijk..r} = u + u_1(i) + u_2(j) + u_3(k) + ... + u_s(r) + u_{12}(ij) + u_{13}(ik) + ... +$$

(4.4)
$$+ u_{(s-1)s}(qr) + u_{123}(ijk) + u_{124}(ijl) + ... +$$
$$+ u_{(s-2)(s-1)s}(pqr) + ... + u_{12..s}(ij..r)$$

donde los parámetros deben satisfacer las restricciones

(4.5)

$$\sum_i u_1(i) = \sum_j u_2(j) = \sum_k u_3(k) = ... = \sum_r u_s(r) = 0$$

$$\sum_i u_{12}(ij) = \sum_j u_{12}(ij) = \sum_i u_{13}(ik) = \sum_k u_{13}(ik) = ... = \sum_q u_{(s-1)s}(qr) = \sum_r u_{(s-1)s}(qr) = 0$$

$$\sum_i u_{123}(ijk) = \sum_j u_{123}(ijk) = \sum_k u_{123}(ijk) = \sum_i u_{124}(ijl) = \sum_j u_{124}(ijl) = \sum_l u_{124}(ijl) = 0$$

$$\sum_i u_{134}(ikl) = ... = \sum_p u_{(s-2)(s-1)s}(pqr) = \sum_q u_{(s-2)(s-1)s}(pqr) \sum_r u_{(s-2)(s-1)s}(pqr) = 0$$

$$\sum_j u_{1234}(ijkl) = ... = \sum_r u_{12...s}(ijk..r) = 0$$

y los índices varian entre 1 y el número de niveles de cada síntoma.

Los parámetros "u" reciben el nombre de acciones o interacciones dependiendo de si tienen 1 ó más subíndices, respectivamente. Las restricciones anteriores hacen que la interpretación de estos parámetros sea muy sencilla: los valores positivos de una acción o interacción hay que interpretarlos como que los síntomas asociados a sus subíndices, en bloque, contribuyen a indicar la presencia del problema a que hace referencia el índice i. Por el contrario, los valores negativos de una acción o interacción contribuyen a indicar la ausencia del mismo problema. Finalmente, los valores nulos indican que el bloque de síntomas a que se refieren los subíndices no aclara nada respecto del citado problema.

El modelo (4.4) es equivalente al de síntomas relevantes y tiene el mismo número de parámetros libres. Sin embargo, la ventaja está en que pueden suponerse muchos de ellos nulos, con lo que se consigue una notable reducción en su número. Este es precisamente el objetivo perseguido con los modelos logarítmico-lineales. Recuérdese que la crítica de los modelos de tipo probabilístico se basa en el elevado número de parámetros que requieren.

El modelo (4.4) puede expresarse también en función de sus parámetros libres directamente, en cuyo caso las restricciones (4.5) desaparecen. Esto se consigue eliminando los parámetros superfluos mediante estas restricciones. El

modelo en esta forma es de muy fácil manejo, pero su interpretación física es mucho más compleja.

4.1.1.- Modelos jerárquicos

Entre las familias de modelos existentes destaca la de los modelos jerárquicos. Se dice que un modelo es de tipo jerárquico si cuando está presente una interacción de un cierto orden entre varias variables están también presentes todas las interacciones de menor orden entre esas mismas variables. Así, por ejemplo, son jerárquicos los siguientes modelos:

$$\log m_{ijkl} = u + u_1(i) + u_4(l) + u_{14}(il)$$

$$\log m_{ijkl} = u + u_1(i) + u_2(j) + u_3(k) + u_4(l) + u_{12}(ij) + u_{13}(ik) + u_{14}(il)$$

$$\log m_{ijkl} = u + u_2(j) + u_3(k) + u_4(l) + u_{23}(jk) + u_{24}(jl) + u_{34}(kl) + u_{234}(jkl)$$

y no lo son

$$\log m_{ijkl} = u + u_1(i) + u_2(j) + u_3(k) + u_{12}(ij) + u_{13}(ik) + u_{123}(ijk)$$

$$\log m_{ijkl} = u + u_1(i) + u_3(k) + u_{12}(ij)$$

Los modelos jerárquicos no siempre son adecuados para su utilización en sistemas expertos. De hecho, en el caso de diagnóstico médico suelen aparecer los grupos de síntomas (síndromes) que siempre aparecen juntos, lo cual conduce a un modelo con una interacción entre esos síntomas y la ausencia de las interacciones de menor orden.

En muchos casos prácticos se plantea la pregunta siguiente: ¿influye o no influye la inclusión de una variable en el valor de los términos del modelo que no la incluyen?, o, dicho de otra forma, si se incluye una variable ¿se obtendrán diferentes parámetros para las otras variables que los que se obtendrían si no se incluyera esa variable?

La respuesta a esta pregunta está íntimamente ligada al concepto de colapsibilidad. Se dice que un conjunto de variables es colapsible respecto de un conjunto de parámetros referentes a otro conjunto de variables si las primeras no influyen en sus valores. Esto ocurre si y solamente si los dos grupos de variables son independientes.

La implicación práctica más importante de la colapsibilidad es la posibilidad de reducir y simplificar los cálculos, pues los valores de los términos del modelo no cambian y se pueden utilizar menos variables para su cálculo.

4.1.2.- Tipos de muestreo

En este apartado se considerarán dos tipos de muestreo que serán necesarios posteriormente al considerar la estimación de los parámetros del modelo mediante el método de la máxima verosimilitud.

4.1.2.1.- Muestreo Poissoniano

Se supone que cada celda tiene una distribución de Poisson independiente con parámetro $m_{ij\ldots r}$. Por ello, la función de probabilidad conjunta de las frecuencias de todas las celdas viene dada por

$$p(\{x_{ij\ldots r}\}) = \prod_{ij\ldots r} \frac{m_{ij\ldots r}^{x_{ij\ldots r}} \exp(-m_{ij\ldots r})}{x_{ij\ldots r}!} \tag{4.6}$$

Este tipo de muestreo resulta al observar durante un periodo de tiempo y, por tanto, el número total de observaciones no está prefijado. A veces puede también fijarse el tamaño de la muestra como criterio de finalización del muestreo. En este caso hay que añadir esa restricción.

4.1.2.2.- Muestreo multinomial

En este caso se supone que el tamaño de la muestra, N, está fijado, por lo que la función de probabilidad conjunta de las frecuencias resulta

$$\frac{N!}{\prod_{ij\ldots r} x_{ij\ldots r}!} \prod_{ij\ldots r} \left(\frac{m_{ij\ldots r}}{N}\right)^{x_{ij\ldots r}} \tag{4.7}$$

4.1.3.- Estimación de los modelos logarítmico-lineales

La aplicación de la fórmula (4.4) da unas frecuencias que en general no coincidirán exáctamente con las reales. Nótese que al ajustar un modelo sencillo (con pocos parámetros) no se tienen suficientes grados de libertad para conseguir la coincidencia con todos los valores. Esto sí se consigue con el modelo saturado

(aquel que tiene todos los parámetros). Para medir la bondad del ajuste se emplean algunos estadísticos, de los cuales se dan a continuación los dos más importantes (Rao (1973)):

$$(4.8) \qquad \chi^2 = \sum_{ijk\ldots r} \frac{(x_{ijk\ldots r} - m_{ijk\ldots r})^2}{m_{ijk\ldots r}}$$

$$(4.9) \qquad G^2 = 2 \sum_{ijk\ldots r} x_{ijk\ldots r} \log \left(\frac{x_{ijk\ldots r}}{m_{ijk\ldots r}}\right)$$

donde $x_{ijk...r}$ representan las frecuencias reales y $m_{ijk...r}$ las frecuencias estimadas.

Nótese que si se da la coincidencia perfecta (todos los $x_{ijk...r}=m_{ijk...r}$), el valor de estos estadísticos es cero y que a mayores discrepancias mayores valores de los estadísticos.

Ambos estadísticos, en el caso de que el modelo sea correcto, se distribuyen según una ley χ^2 con p-s-1 grados de libertad, donde p es el número de parámetros del modelo saturado y s el número de parámetros del modelo que se estima. Esto permite contrastar la hipótesis de si las discrepancias son debidas a la aleatoriedad del problema o hay que atribuirlas a que el modelo estimado es incorrecto. Si el nivel de significación que resulta es bajo, se rechaza el modelo estimado. También se utilizan estos dos estadísticos para elegir el modelo óptimo, como se verá más adelante.

4.1.3.1.- Estimación por el método de la máxima verosimilitud

Uno de los principales problemas a resolver antes de la utilización de un modelo logarítmico-lineal es la estimación de sus parámetros. En este etapa juega un papel preponderante el método de la máxima verosimilitud que se describe a continuación.

El método de la máxima verosimilitud se basa en elegir los valores de los parámetros que maximizan la función de verosimilitud de la muestra. Para el caso del muestreo Poissoniano se maximiza la función (4.6), o lo que es equivalente su logaritmo, con respecto a $m_{ij...r}$.

De esta forma para el muestreo Poissoniano de (4.6) tomando logaritmos se obtiene

$$(4.10) \qquad L = \sum_{ij\ldots r} x_{ij\ldots r} \log (m_{ij\ldots r}) - \sum_{ij\ldots r} m_{ij\ldots r} - \sum_{ij\ldots r} \log (x_{ij\ldots r}!)$$

pero como dados los valores de $x_{ij\ldots r}$ estos resultan constantes, y dado que

$$(4.11) \qquad \sum_{ij\ldots r} m_{ij\ldots r} = N$$

la maximización de (4.10) equivale a la de

$$(4.12) \qquad L = \sum_{ij\ldots r} x_{ij\ldots r} \log m_{ij\ldots r}$$

Análogamente, para el caso de muestreo multinomial, tomando logaritmos en (4.7) resulta la función

$$(4.13) \qquad L = \log \left(\frac{N!}{\prod_{ij\ldots r} x_{ij\ldots r}}\right) + \sum_{ij\ldots r} x_{ij\ldots r} \log m_{ij\ldots r} - \log N \sum_{ij\ldots r} x_{ij\ldots r}$$

La maximización de la expresión (4.13) es también equivalente a la maximización de la expresión (4.12), pues en ella dos de los sumandos no dependen de $m_{ij\ldots r}$.

Entre las ventajas más importantes de este método destacan las siguientes:

(a) los estimadores de máxima verosimilitud son muy fáciles de calcular e incluso tienen formas explícitas en muchos casos.
(b) los estimadores de máxima verosimilitud satisfacen ciertas restricciones intuitivas no satisfechas por otros métodos.
(c) el método de la máxima verosimilitud puede aplicarse directamente a datos multivariados con varias celdas nulas sin que se produzcan ceros como estimaciones de dichas celdas, lo cual es una propiedad deseable en muchos casos prácticos.

Para la estimación de los parámetros del modelo se dispone de la tabla de frecuencias observadas. Sin embargo, el número de datos que figura en la tabla es excesivamente alto y no es necesaria tanta información. Sin pérdida alguna de la información que dichos datos poseen sobre los parámetros del modelo, puede sustituirse ésta por los llamados estadísticos suficientes, que precisamente se definen como aquellos estadísticos que recogen toda la información que posee la muestra sobre los parámetros a estimar.

En el caso de modelos jerárquicos la obtención de los estadísticos suficientes es sencilla y se realiza mediante la regla que sigue (Bishop (1975)):

(a) se seleccionan los términos de orden k (interacción de mayor orden) que aparecen en el modelo
(b) si aparecen todos los términos posibles de orden k se para y en caso contrario se seleccionan aquellos términos de orden k-1 que no son hijos (sus índices no están incluidos en ninguno de los de orden superior) de los términos anteriores
(c) se procede análogamente con los de órdenes inferiores hasta analizar todos los órdenes y sin incluir términos hijos de los anteriores.

De esta forma se selecciona el conjunto

$$C = \{u_{\theta_1}, u_{\theta_2}, \ldots, u_{\theta_t}\}$$

que conduce a los siguientes estadísticos suficientes

$$x_{\theta_j} = \sum_{i_k \notin \theta_j} x_{i_1 i_2 \ldots i_s} \quad ; \quad j = 1, 2, \ldots, t \tag{4.14}$$

Una vez que se conoce un conjunto de estadísticos suficientes la estimación puede hacerse de varias formas. En particular resulta interesante la estimación directa (mediante fórmula explícita) cuando ésta existe. Sin embargo, esto no siempre es posible y, por ello, resulta más sencillo y mucho más práctico para su programación mediante ordenador el método iterativo que se describe, en detalle, a continuación.

(a) se obtiene una estimación inicial cualquiera que presente sólo las interacciones del modelo
(b) se procede a aplicar iterativamente las fórmulas siguientes hasta conseguir el error deseado:

$$\hat{m}_{\theta}^{(i_{sr})} = \hat{m}_{\theta}^{(i_{sr}-1)} \frac{x_{\theta_q}}{\hat{m}_{\theta_q}^{(i_{sr}-1)}} \quad ; \quad s = 1, 2, \ldots, t \tag{4.15}$$

donde r representa el número de ciclos (veces que se han recorrido todos los estadísticos suficientes), i_{sr} representa el número de la

iteración s-ésima dentro del ciclo r-ésimo y se supone que la estimación inicial de un ciclo es la última del ciclo anterior.

La estimación mediante este método tiene, entre otras, las siguientes ventajas:

(a) converge a la solución dada por el método de la máxima verosimilitud
(b) se puede utilizar un criterio de parada que garantiza las estimaciones de las frecuencias de las celdas hasta una precisión prefijada
(c) puede elegirse cualquier estimación inicial sin más que exigirle que posea sólo las interacciones del modelo a ajustar
(d) si existen estimaciones directas (fórmulas explícitas) el método da las estimaciones exactas en sólo un ciclo (para modelos de muchas dimensiones hay que elegir un orden conveniente en los estadísticos suficientes)
(e) da las estimaciones de las celdas directamente sin necesidad de pasar por los parámetros (nótese sin embargo, que el número de celdas es más elevado que el de parámetros)

A continuación se da un ejemplo para ilustrar el método anterior.

Ejemplo 4.1.- El modelo logarítmico lineal general para el caso del ejemplo de la figura 4.2

$$\begin{aligned}\log m_{ijkl} = u + u_1(i) + u_2(j) + u_3(k) + u_4(l) + u_{12}(ij) + u_{13}(ik) + u_{14}(il) + \\ + u_{23}(jk) + u_{24}(jl) + u_{34}(kl) + u_{123}(ijk) + u_{124}(ijl) + u_{134}(ikl) + \\ + u_{234}(jkl) + u_{1234}(ijkl)\end{aligned}$$

donde los parámetros u deben satisfacer las restricciones

$$\sum_i u_1(i) = \sum_j u_2(j) = \sum_k u_3(k) = \sum_l u_4(l) = 0$$

$$\sum_i u_{12}(ij) = \sum_j u_{12}(ij) = \sum_i u_{13}(ik) = \sum_k u_{13}(ik) = \sum_i u_{14}(il) = \sum_l u_{14}(il) = 0$$

$$\sum_j u_{23}(jk) = \sum_k u_{23}(jk) = \sum_j u_{24}(jl) = \sum_l u_{24}(jl) = \sum_k u_{34}(kl) = \sum_l u_{34}(kl) = 0$$

$$\sum_i u_{123}(ijk) = \sum_j u_{123}(ijk) = \sum_k u_{123}(ijk) = \sum_i u_{124}(ijl) = \sum_j u_{124}(ijl) = \sum_l u_{124}(ijl) = 0$$

$$\sum_i u_{134}(ikl) = \sum_k u_{134}(ikl) = \sum_l u_{134}(ikl) = \sum_j u_{234}(jkl) = \sum_k u_{234}(jkl) = \sum_l u_{234}(jkl) = 0$$

$$\sum_j u_{1234}(ijkl) = \sum_j u_{1234}(ijkl) = \sum_k u_{1234}(ijkl) = \sum_l u_{1234}(ijkl) = 0$$

los índices i, j, k y l se refieren a E, S_1, S_2 y S_3, respectivamente, y los valores 1 y 2 de los niveles a la presencia o ausencia del problema o del síntoma respectivo.

Si se ajusta el modelo anterior a los datos de frecuencias de la figura 4.2a resulta el modelo logarítmico-lineal

$$\log m_{ijkl} = u + u_1(i) + u_2(j) + u_3(k) + u_4(l) + u_{12}(ij) + u_{13}(ik) + u_{14}(il)$$

con los siguientes valores de los parámetros libres

$$u = 3.4474 ; u_1(1) = 0.441; u_2(1) = -0.212; u_3(1) = -0.341; u_4(1) = 0.440$$
$$u_{12}(1,1) = 0.212 ; u_{13}(1,1) = 0.758 ; u_{14}(1,1) = 0.643$$

Este modelo tiene 8 grados de libertad (8 parámetros libres) frente a los 16 del modelo de síntomas relevantes. La Tabla 4.1 muestra los valores reales de las frecuencias y las dadas por el modelo anterior (entre paréntesis).

S_1	S_2	S_3	E	$\bar{E}$
SI	SI	SI	219 (218.85)	4 (3.60)
Si	Si	NO	25 (25.09)	5 (5.40)
SI	NO	SI	95 (95.05)	32 (32.41)
SI	NO	NO	11 (10.90)	49 (48.64)
NO	SI	SI	219 (218.85)	8 (8.40)
NO	SI	NO	25 (25.09)	13 (12.61)
NO	NO	SI	95 (95.05)	76 (75.67)
NO	NO	NO	11 (10.90)	113 (113.57)

Tabla 4.1.- Valores reales y predicciones de las frecuencias para el caso de la figura 4.2a

Si se ajustan ahora los datos de la figura 4.2b, resulta el modelo

$$\log m_{ijkl} = u + u_1(i) + u_2(j) + u_3(k) + u_4(l) + u_{12}(ij) + u_{13}(ik) + u_{14}(il) + u_{23}(jk) + u_{24}(jl) + u_{123}(ijk) + u_{124}(ijl)$$

con los valores de los parámetros libres

$$u = 1.9248 ; u_1(1) = 0.190 ; u_2(1) = -0.067 ; u_3(1) = -0.072; u_4(1) = 0.378$$
$$u_{12}(1,1) = 0.168 ; u_{13}(1,1) = 1.436; u_{14}(1,1) = 1.432 ; u_{23}(1,1) = -0.200 ;$$
$$u_{24}(1,1) = -0.039; u_{123}(1,1,1) = -1.366 ; u_{124}(1,1,1) = 1.158 ;$$

Este modelo tiene 12 grados de libertad, lo cual supone un ahorro de 4 parámetros frente al modelo general.

La Tabla 4.2 muestra los valores reales de las frecuencias y las dadas por el modelo anterior (entre paréntesis).

S_1	S_2	S_3	E	$\bar{E}$
SI	SI	SI	140 (140.16)	0 (0.34)
SI	SI	NO	0 (0.40)	30 (30.23)
SI	NO	SI	210 (209.94)	0 (0.66)
SI	NO	NO	0 (0.60)	60 (59.91)
NO	SI	SI	280 (280.00)	0 (0.57)
NO	SI	NO	70 (70.3)	0 (0.43)
NO	NO	SI	0 (0.80)	120 (119.92)
NO	NO	NO	0 (0.20)	90 (90.09)

Tabla 4.2.- Valores reales y predicciones de las frecuencias para el caso de la figura 4.2b

¤

4.1.4.- Base de conocimiento y motor de inferencia

La base de conocimiento en el caso de los sistemas expertos basados en los modelos logarítmico-lineales está constituida por la estructura de los modelos asociados a cada problema junto con sus parámetros.

El motor de inferencia se reduce a calcular las probabilidades asociadas a los síntomas disponibles en cada instante. Como se tratará normalmente de información parcial, es decir, no serán conocidos todos los síntomas que aparecen

en los modelos logarítmico-lineales, será necesario sumar las frecuencias asociadas a todos los casos de información total que corresponden a la parcial dada. En otras palabras hay que descomponer la situación de información parcial en unión de casos de información total.

4.2.- MODELOS DE REGRESION

Los modelos logarítmico lineales de la sección anterior son útiles para síntomas o variables de tipo discreto, es decir, que toman un conjunto discreto (finito) de valores, pero no tienen validez para variables o síntomas continuos, a menos que se discreticen sus valores por subdivisión en intervalos. Dada la frecuencia de aparición de este tipo de síntomas o variables, ésta es una limitación notable que da lugar a ciertas críticas de los modelos logarítmico-lineales.

Con objeto de resolver este problema surgen los modelos de regresión, que generalizan los modelos anteriores.

El modelo de regresión que se empleará en esta sección es de tipo logit, es decir, se supone para cada problema un modelo del tipo

$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=1}^{r} A_{ij}\beta_j \qquad (4.16)$$

donde los términos A_{ij} y β_j son de la forma

$$\beta_j \equiv u_{i_1 i_2 \ldots i_s}(j_1 j_2 \ldots j_s) \quad ; \quad 1 \le j_k \le I_k$$

$$A_{ij} = f_{i_1 i_2 \ldots i_s}(x_1 x_2 \ldots x_t) \qquad (4.17)$$

p_i es la probabilidad del problema condicionado por los síntomas dados, I_j es el número de niveles del síntoma discreto j-ésimo y las funciones $f(x_1,x_2,\ldots,x_t)$, que son conocidas, pueden ser constantes, valer la unidad, en el caso de que el término represente la influencia de un grupo de síntomas discretos únicamente, o ser nulas; análogamente, los parámetros "u" pueden degenerar en la unidad si el término refleja la influencia de un grupo de síntomas continuos únicamente.

Nótese que (4.16) equivale a suponer

(4.18)
$$p_i = \frac{\exp\left(\sum_{j=1}^{r} A_{ij}\beta_j\right)}{1+\exp\left(\sum_{j=1}^{r} A_{ij}\beta_j\right)}$$

y que ahora el modelo de independencia no es directo, es decir, no corresponde a un modelo sencillo en el modelo logit. Por el contrario, se puede suponer la llamada independencia en logit para la que la probabilidad de la intersección ya no es el producto de las probabilidades sino que el cociente p/(1-p) de la intersección es el producto de los mismos cocientes de los componentes.

4.2.1.- Estimación de los parámetros del modelo de regresión

Si se tienen suficientes datos de los casos con sus síntomas y sus problemas, los parámetros del modelo (4.16) pueden estimarse por el método de la máxima verosimilitud (Luceño (1988)). En efecto, supóngase que se trata de modelar un determinado problema y que se tienen datos de k casos en los que se ha observado la variable Z_i (i=1,2,..,k), que toma el valor 1 si el caso tiene el problema y 0 si no lo tiene, junto con sus síntomas, que conducen a conocer los valores de A_{ij}. La variable Z_i es pues una variable de Bernouilli (toma sólo los valores 0 y 1) de parámetro p_i y, por tanto, la función de verosimilitud de la muestra es

(4.19)
$$V = \prod_{i=1}^{k} p_i^{z_i} (1-p_i)^{1-z_i}$$

y su logaritmo

(4.20)
$$L = \log(V) = \sum_{i=1}^{k} z_i \log(p_i) + (1-z_i)\log(1-p_i)$$

Derivando respecto a β_j, sustituyendo el valor de p_i de (4.18) e igualando a cero resulta

(4.21)
$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{k} \left(\frac{z_i}{p_i} - \frac{1-z_i}{1-p_i}\right)\frac{\partial p_i}{\partial \beta_j} = \sum_{i=1}^{k} \frac{z_i - p_i}{p_i(1-p_i)}\, p_i(1-p_i)A_{ij} = \sum_{i=1}^{k} (z_i - p_i)A_{ij} = 0$$

El sistema (4.21) es no lineal respecto de los parámetros $\beta_1,...,\beta_r$, sin embargo, puede resolverse iterativamente, ya que puede escribirse como

$$(4.22) \qquad \sum_{l=1}^{r} \left(\sum_{i=1}^{k} w_i A_{il} A_{ij} \right) \beta_l = \sum_{i=1}^{k} w_i Y_i A_{ij} \quad ; \quad j = 1, 2,..., r$$

donde

$$w_i = p_i(1 - p_i)$$

$$(4.23) \qquad Y_i = \sum_{l=1}^{r} A_{il}\beta_l + \frac{z_i - p_i}{p_i(1-p_i)}$$

El sistema (4.22) puede resolverse mediante el algoritmo que sigue

1.- elegir unos valores iniciales de $\beta_1,...,\beta_r$. Estos valores sólo influyen en la velocidad de convergencia, pero no, en la convergencia misma. En otras palabras, la convergencia está asegurada para cualesquiera valores de partida
2.- para cada dato, calcular z_i, Y_i y w_i mediante la expresión (4.23)
3.- estimar $\beta_1,...,\beta_r$ resolviendo el sistema lineal (4.22)
4.- repetir los pasos 2 y 3 hasta conseguir el error deseado.

Ejemplo 4.2.- Supóngase que se trata de diagnosticar entre los 4 problemas: E_1 (1), E_2 (2), E_3 (3) y E_4 (4), y que las diferencias entre ellos se basan en los siguientes síntomas:

1.- S_1 (síntoma discreto)
2.- S_2 (síntoma discreto)
3.- S_3 (síntoma discreto)
4.- S_4 (síntoma continuo)

Supóngase también que el estado del conocimiento consiste en la base de datos que figura en la tabla 4.3, en la que se dan los datos de 99 casos con los problemas y los síntomas indicados (código 0 = ausencia del síntoma, código 1 = presencia del síntoma) (el número 63, entre paréntesis, indica las veces que se repite ese mismo dato, es decir, que hay 63 casos con S_4=36.5, sin S_1, sin S_2 y sin S_3).

De los datos de la tabla 4.3 se deduce la información contenida en la tabla 4.4, que muestra las frecuencias de los síntomas discretos y las medias del

síntoma continuo para los casos de cada problema. La tabla 4.5 da una descripción cualitativa de los problemas.

	Síntoma					Síntoma			
Problema	1	2	3	4	Problema	1	2	3	4
1	0	1	0	38.0	1	0	1	0	38.5
1	0	1	0	39.0	1	0	1	0	38.0
1	0	1	0	38.7	1	0	1	0	38.9
1	0	1	0	38.8	1	0	1	0	39.5
1	0	0	0	39.2	1	0	0	0	39.0
2	0	1	1	39.0	2	0	1	1	38.5
2	0	1	0	39.5	2	0	1	1	36.5
2	0	1	1	37.0	2	0	1	0	36.5
2	0	0	1	39.2	2	0	0	1	38.7
2	0	0	0	38.8	2	0	0	1	36.5
2	0	0	1	36.7	2	0	0	0	36.6
3	1	0	1	39.2	3	1	0	1	38.7
3	1	0	1	39.3	3	1	0	1	38.5
3	0	0	1	38.0	3	1	0	0	38.2
3	1	0	1	38.3	4	0	0	0	36.6
4	0	0	0	36.7	4	0	0	0	36.7
4	0	0	0	36.7	4	0	0	0	36.6
4	0	0	0	36.7	4	0	0	0	36.3
(63) 4	0	0	0	36.5					

Tabla 4.3.- Base de datos

	FRECUENCIA RELATIVA			
	S_1	S_2	S_3	Valor medio de S_4
E_1	0.00	0.80	0.00	38.76
E_2	0.00	0.50	0.66	37.79
E_3	0.86	0.00	0.86	38.60
E_4	0.00	0.00	0.00	36.51

Tabla 4.4.- Frecuencias y valores medios de los síntomas por problemas

	S_1	S_2	S_3	Valor medio de S_4
E_1	NO	SI	NO	ALTA
E_2	NO	-	SI	MEDIA
E_3	SI	NO	SI	ALTA
E_4	NO	NO	NO	NO

Tabla 4.5.- Descripción cualitativa de los problemas

Con objeto de diseñar un motor de inferencia para el diagnóstico de estos problemas, y basado en las tablas anteriores, se ha elegido, para cada uno de los 4 problemas, un modelo del tipo

$$\log\left(\frac{p_i}{1-p_i}\right) = u + u_{123}(121)\,X_4 + u_{123}(211)\,X_4 + u_{123}(212)\,X_4 + \\ + u_{123}(221)\,X_4 + u_{123}(222)\,X_4$$

en el que X_4 representa el síntoma S_4, los subíndices 1, 2 y 3 se refieren a los tres síntomas S_1, S_2 y S_3, respectivamente, y los niveles 1 y 2 se refieren a la presencia o a la ausencia del síntoma, respectivamente.

La estimación de los parámetros se ha realizado por el método de la máxima verosimilitud, descrito anteriormente, y siguiendo el algoritmo indicado con valores iniciales nulos de los parámetros, resultando los valores que se muestran en la tabla 4.6.

Problema	u	$u_{123}(121)$	$u_{123}(211)$	$u_{123}(212)$	$u_{123}(221)$	$u_{123}(222)$
1	-85.257	1.955	1.988	2.263	1.983	2.191
2	-15.757	0.195	0.639	0.372	0.455	0.333
3	4.737	0.141	-0.397	-0.389	-0.162	-0.408
4	-1.400	-0.175	-0.181	-0.177	-0.180	0.115

Tabla 4.6.- Valores estimados de los parámetros

Finalmente, la tabla 4.7 muestra las probabilidades de cada uno de los problemas (P_1,P_2,P_3 y P_4) condicionadas por los síntomas que se indican. Nótese que al dar casos con los síntomas típicos de los problemas, es decir, los que figuran en la tabla 4.5, los problemas que resultan, según los valores de P_1, P_2, P_3

y P_4, coinciden con las reales. Así, un caso con $S_1 \cap \overline{S_2} \cap S_3$ y S_4=40 da lugar a los valores P_1=0.0009, P_2=0.0003, P_3=1.0 y P_4=0.0002, lo que indica que tiene E_3.

S_1	S_2	S_3	S_4	P1	P2	P3	P4
Sí	No	Sí	36.5	0.0000	0.0002	0.9999	0.0004
Sí	No	Sí	40.0	0.0009	0.0003	1.0000	0.0002
No	Sí	Sí	36.5	0.0000	0.9995	0.0001	0.0003
No	Sí	Sí	40.0	0.0032	0.9999	0.0000	0.0002
No	Sí	No	36.5	0.0646	0.1000	0.0001	0.0004
No	Sí	No	40.0	0.9948	0.2897	0.0000	0.0002
No	No	Sí	36.5	0.0000	0.6976	0.2367	0.0003
No	No	Sí	40.0	0.0026	0.9189	0.1496	0.0002
No	No	No	36.5	0.0051	0.0263	0.0000	0.9425
No	No	No	40.0	0.9168	0.0798	0.0000	0.9608
O T R O S			36.5	0.0000	0.0000	0.9913	0.1978
O T R O S			40.0	0.0000	0.0000	0.9913	0.1978

Tabla 4.7.- Predicción de probabilidades de problemas dados por el modelo

4.2.2.- Base de conocimiento y motor de inferencia

La base de conocimiento en el caso de los sistemas expertos basados en modelos de regresión está constituida por la estructura de los modelos asociados a cada problema junto con sus parámetros.

El motor de inferencia, al igual que en el caso de los modelos logarítmico-lineales se reduce a calcular las probabilidades asociadas a los síntomas disponibles en cada instante. Para ello, por tratarse normalmente de información parcial, será necesario sumar las frecuencias asociadas a todos los casos de información total que corresponden a la parcial dada. El cálculo de las frecuencias para casos de información total se hace como en los modelos logarítmicos.

4.3.- MODELOS DEL ANALISIS DISCRIMINANTE

En este apartado se dan algunos métodos del análisis discriminante (Lachenbruch (1975)) que tienen relación con el problema del diagnóstico y que

sirven para diseñar motores de inferencia sencillos de implantación. Los métodos que se describen a continuación son también válidos para que el sistema experto aprenda e incluso para que optimice, mediante la reducción del número de variables que intervienen. Sin embargo, el análisis de estos aspectos se postponen para el capítulo siguiente. Se comenzará con el estudio del problema de clasificación en dos grupos y luego se extrapolará al caso de n grupos.

4.3.1.- Clasificación en dos grupos

El problema que se plantea en este apartado es el de clasificar individuos sabiendo que pertenecen a uno de dos grupos dados y teniendo como datos los valores que toman un conjunto de variables $(X_1,X_2,...,X_k)$. Con objeto de aclarar las ideas se comienza con un ejemplo.

Ejemplo 4.3.- Supónganse dos grupos de casos E_1 y E_2 que se diferencian por el síntoma S_1 (en este caso se ha elegido k=1, por simplicidad), de forma tal que las del grupo E_1 tienen una distribución normal N(38,0.7) y las del grupo E_2, una distribución normal N(40,0.7), como se muestra en la figura 4.3. Este gráfico ilustra como los casos del grupo E_1 tienen valores de S_1 más bajos que los del grupo E_2, entendiendo esta afirmación en sentido estadístico, es decir, que aunque pueden encontrarse casos de E_1 con valores de S_1 más altos que los de E_2, esto es poco probable.

Si se extrae una muestra de tamaño 10 de cada población y se observan los valores del síntoma S_1 de los 20 individuos que resultan, se obtienen los resultados de la figura 4.3, en la que con círculos negros se muestran los individuos del grupo E_1 y con círculos blancos los del grupo E_2. Conviene notar que los individuos del grupo E_1 se agrupan principalmente en la zona izquierda y los del grupo E_2, en la zona derecha. En la zona intermedia, sin embargo, los individuos de ambos grupos se solapan. La idea del análisis discriminante es diferenciar a estos dos grupos por los valores de S_1, es decir, definir dos regiones R_1 y R_2, tales como las de la figura 4.3, de forma que si el valor de S_1 para un caso está en R_1 se le asigna a E_1 y si está en R_2, a E_2. La elección de estas dos regiones se hace de varias maneras; una de ellas se basa en la idea de hacer mínimas las probabilidades de equivocarse al decir que un caso de E_2, con valores de S_1 en R_1, pertenece al grupo E_1 y que uno de E_1, con valores de S_1 en R_2, pertenece a E_2. Nótese que en el ejemplo de la figura el error es con un individuo de cada grupo, es decir, hay un círculo blanco en R_1 y uno negro en R_2.

Otra forma elemental de clasificar a un individuo en uno de los grupos es calcular las distancias de su punto representativo a los centros de gravedad de los grupos y asignarlo al grupo más cercano. Sin embargo, la idea más utilizada en

análisis discriminante es, como ya se ha indicado, la de minimizar la probabilidad total de clasificación errónea o su coste.

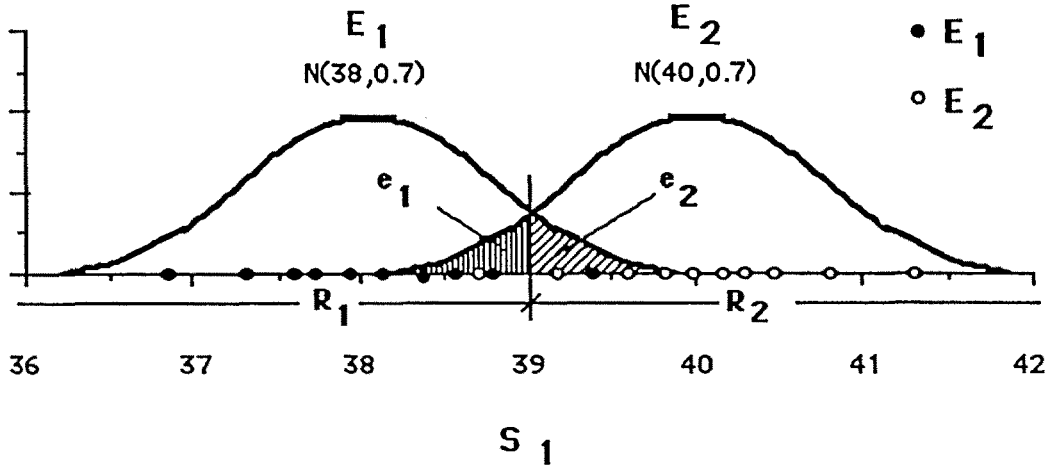


Figura 4.3.- Distribuciones de S_1 en dos grupos

En el ejemplo 4.3, la probabilidad de que un caso, elegido al azar, sea de E_1 y tenga el valor de S_1 en R_2 es

$$Q_1 = p_1 \int_{R_2} f_{N(38,\,0.7)}(x)\,dx$$

donde p_1 es la probabilidad "a priori" de que el individuo pertenezca al grupo E_1 y f es la función de densidad de la ley normal indicada. El valor e_2, sombreado en la figura 4.3, que es la integral que aparece en la expresión anterior, es una medida de este error.

Análogamente, la probabilidad de que un individuo sea de E_2 y tenga valor de S_1 en R_1 es

$$Q_2 = p_2 \int_{R_1} f_{N(40,\,0.7)}(x)\,dx$$

donde ahora p_2 es la probabilidad "a priori" de que el individuo pertenezca al grupo E_2 y e_1 es también la integral que aparece en la expresión.

Entre los diferentes métodos posibles para elegir R_1 y R_2, se analizan a continuación los siguientes:

(a) minimizar la probabilidad total de clasificación errónea (Q_1+Q_2)(Welch (1939))

(b) minimizar el coste total esperado de los errores de clasificación

(c) minimizar el máximo de los dos costes de clasificación errónea ($Max(C_1Q_1, C_2Q_2)$) (Kendall (1961))

(d) encontrar una nueva variable transformada que sea lineal en las variables X y tal que la razón de la distancia al cuadrado de los centros de gravedad de esta variable en los grupos a su varianza sea máxima (Fisher).

4.3.1.1.- Minimización de la probabilidad de clasificación errónea

En este caso se trata de minimizar la siguiente expresión, que da la probabilidad total de clasificación errónea:

$$(4.24) \qquad Q(R_1, R_2, f_1, f_2) = p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \quad ,$$

donde $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ son las funciones de densidad de $\mathbf{X}$ en E_1 y E_2, respectivamente y $\mathbf{X}$ es el vector $(X_1, X_2, \ldots, X_k)$.

Q puede escribirse también como

$$(4.25) \qquad Q = p_1 \left[1 - \int_{R_1} f_1(\mathbf{x})\, d\mathbf{x} \right] + p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} =$$

$$= p_1 + \int_{R_1} [\, p_2 f_2(\mathbf{x}) - p_1 f_1(\mathbf{x})]\, d\mathbf{x}$$

de donde se deduce que Q se minimiza para las regiones definidas por

$$(4.26) \qquad R_1 \equiv \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \quad ; \quad R_2 \equiv \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1}$$

Un caso particular muy importante es el de las distribuciones normales. Si $\mathbf{X}$ tiene una distribución normal con medias $\mathbf{m_1}$ y $\mathbf{m_2}$ en los grupos E_1 y E_2, respectivamente y la matriz de varianzas-covarianzas, $\mathbf{S}$, es común en ambos, entonces las funciones de densidad son

$$f_i(\mathbf{x}) = (2\pi)^{-k/2} \left| \mathbf{S}^{-1/2} \right| \exp\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m}_i)' \mathbf{S}^{-1} (\mathbf{x} - \mathbf{m}_i) \} \quad ; \quad i = 1, 2$$

y la expresión (4.26) resulta

$$\frac{f_1(x)}{f_2(x)} = \exp\{-\frac{1}{2}(x-m_1)'S^{-1}(x-m_1)+\frac{1}{2}(x-m_2)'S^{-1}(x-m_2)\} =$$

$$= \exp\{[x-\frac{1}{2}(m_1+m_2)]'S^{-1}(m_1-m_2)\}$$

por lo que la región R_1 resulta ser

$$(4.27) \qquad R_1 \equiv [x-\frac{1}{2}(m_1+m_2)]'S^{-1}(m_1-m_2) > \log\frac{p_2}{p_1}$$

La expresión

$$(4.28) \qquad R(x) = [x-\frac{1}{2}(m_1+m_2)]'S^{-1}(m_1-m_2)$$

depende de las medias m_1 y m_2 y de la matriz de varianzas-covarianzas, **S**, de la población, que normalmente son desconocidas. Su equivalente, si se dispone de una muestra es

$$(4.29) \qquad R_S(x) = [x-\frac{1}{2}(\bar{x}_1+\bar{x}_2)]'S^{*-1}(\bar{x}_1-\bar{x}_2)$$

donde $\bar{x}_1$ y $\bar{x}_2$ representan las medias muestrales de **X** en los grupos 1 y 2, respectivamente y **S*** es

$$(4.30) \qquad S^* = \frac{1}{n-2}\sum_{i=1}^{n}(x_i-\bar{x}_1)(x_i-\bar{x}_2)'$$

siendo n el número total de individuos de los dos grupos.

El estadístico R(x), que es lineal en las variables X, se distribuye en el grupo i, según una ley normal de media y varianza

$$(4.31) \qquad E[R(x)] = \frac{1}{2}(-1)^{i+1}(m_1-m_2)'S^{-1}(m_1-m_2) = \frac{1}{2}(-1)^{i+1}\delta^2$$

y

$$(4.32) \qquad E[R(x)-R(m_i)]^2 = \delta^2$$

donde δ^2 se obtiene de (4.31). Una estimación de δ^2 es su equivalente en la muestra

(4.33) $$D^2 = (\bar{x}_1 - \bar{x}_2)' S^{*-1} (\bar{x}_1 - \bar{x}_2)$$

x	F(x)	x	F(x)	x	F(x)	x	F(x)	x	F(x)
0.00	0.5000	0.02	0.5080	0.04	0.5160	0.06	0.5239	0.08	0.5319
0.10	0.5398	0.12	0.5478	0.14	0.5557	0.16	0.5636	0.18	0.5714
0.20	0.5793	0.22	0.5871	0.24	0.5948	0.26	0.6026	0.28	0.6103
0.30	0.6179	0.32	0.6255	0.34	0.6331	0.36	0.6406	0.38	0.6480
0.40	0.6554	0.42	0.6628	0.44	0.6700	0.46	0.6772	0.48	0.6844
0.50	0.6915	0.52	0.6985	0.54	0.7054	0.56	0.7123	0.58	0.7190
0.60	0.7257	0.62	0.7324	0.64	0.7389	0.66	0.7454	0.68	0.7517
0.70	0.7580	0.72	0.7642	0.74	0.7704	0.76	0.7764	0.78	0.7823
0.80	0.7881	0.82	0.7939	0.84	0.7995	0.86	0.8051	0.88	0.8106
0.90	0.8159	0.92	0.8212	0.94	0.8264	0.96	0.8315	0.98	0.8365
1.00	0.8413	1.02	0.8461	1.04	0.8508	1.06	0.8554	1.08	0.8599
1.10	0.8643	1.12	0.8686	1.14	0.8729	1.16	0.8770	1.18	0.8810
1.20	0.8849	1.22	0.8888	1.24	0.8925	1.26	0.8962	1.28	0.8997
1.30	0.9032	1.32	0.9066	1.34	0.9099	1.36	0.9131	1.38	0.9162
1.40	0.9192	1.42	0.9222	1.44	0.9251	1.46	0.9279	1.48	0.9306
1.50	0.9332	1.52	0.9357	1.54	0.9382	1.56	0.9406	1.58	0.9429
1.60	0.9452	1.62	0.9474	1.64	0.9495	1.66	0.9515	1.68	0.9535
1.70	0.9554	1.72	0.9573	1.74	0.9591	1.76	0.9608	1.78	0.9625
1.80	0.9641	1.82	0.9656	1.84	0.9671	1.86	0.9686	1.88	0.9699
1.90	0.9713	1.92	0.9726	1.94	0.9738	1.96	0.9750	1.98	0.9761
2.00	0.9772	2.02	0.9783	2.04	0.9793	2.06	0.9803	2.08	0.9812
2.10	0.9821	2.12	0.9830	2.14	0.9838	2.16	0.9846	2.18	0.9854
2.20	0.9861	2.22	0.9868	2.24	0.9875	2.26	0.9881	2.28	0.9887
2.30	0.9893	2.32	0.9898	2.34	0.9904	2.36	0.9909	2.38	0.9913

Tabla 4.8.- Función de distribución de la ley normal

De acuerdo con esto, las probabilidades de clasificación errónea pueden escribirse como

(4.34) $$P_1 = \Pr[R(x) < \log \frac{p_2}{p_1}] = \Phi\left(\frac{\log(\frac{p_2}{p_1}) - \frac{\delta^2}{2}}{\delta}\right)$$

$$(4.35) \qquad P_2 = \Phi\left(-\frac{\log\left(\frac{p_2}{p_1}\right)+\frac{\delta^2}{2}}{\delta}\right)$$

donde $\Phi(x)$ es la función de distribución de la ley normal N(0,1), que se da en la tabla 4.8.

Ejemplo 4.4.- Dos grupos de casos se caracterizan por sus tensiones arteriales máxima y mínima, que se distribuyen conjuntamente según una ley normal de medias $\mathbf{m}_1=(6,12)$ y $\mathbf{m}_2=(10,16)$ y matriz de varianzas-covarianzas

$$S=\begin{pmatrix}4 & 5\\ 5 & 9\end{pmatrix}$$

Las probabilidades "a priori" de que un enfermo, elegido al azar, pertenezca a cada uno de los grupos son $p_1=0.70$ y $p_2=0.30$.

Según (4.27) se tiene

$$R_1 \equiv (x_1-8, x_2-14)\begin{pmatrix}\frac{9}{11} & -\frac{5}{11}\\ -\frac{5}{11} & \frac{4}{11}\end{pmatrix}\begin{pmatrix}-4\\ -4\end{pmatrix} > \log\left[\frac{0.3}{0.7}\right]$$

que conduce a

$$R_1 \equiv 4x_1 - x_2 < 20.33$$
$$R_2 \equiv 4x_1 - x_2 \geq 20.33$$

El valor de δ^2 según (4.31) es

$$\delta^2 = (-4\ -4)\begin{pmatrix}\frac{9}{11} & \frac{-5}{11}\\ \frac{-5}{11} & \frac{4}{11}\end{pmatrix}\begin{pmatrix}-4\\ -4\end{pmatrix} = \frac{48}{11} = 4.3636$$

por lo que las probabilidades de clasificación errónea resultan (ver tabla 4.8):

$$P_1 = \Phi\left(\frac{\log\left(\frac{0.3}{0.7}\right)-2.1818}{\sqrt{4.3636}}\right) = \Phi(-1.45) = 0.0735$$

$$P_2 = \Phi\left(- \frac{\log\left(\frac{0.3}{0.7}\right) + 2.1818}{\sqrt{4.3636}} \right) = \Phi(-0.639) = 0.261$$

¤

En muchos casos, la hipótesis de que la matriz de varianzas-covarianzas coincide en ambos grupos no es cierta ni siquiera aproximadamente. En estos casos, la expresión (4.26) conduce a

(4.36)

$$R_1 \equiv \log \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \log \frac{p_2}{p_1} \equiv$$

$$\equiv \frac{1}{2} \log \frac{|S_2|}{|S_1|} - \frac{1}{2}(\mathbf{x}-\mathbf{m}_1)' S_1^{-1} (\mathbf{x}-\mathbf{m}_1) + \frac{1}{2}(\mathbf{x}-\mathbf{m}_2)' S_2^{-1} (\mathbf{x}-\mathbf{m}_2) =$$

$$= A_0 - \frac{1}{2}[\mathbf{x}'(S_1^{-1} - S_2^{-1})\mathbf{x} - 2\mathbf{x}'(S_1^{-1}\mathbf{m}_1 - S_2^{-1}\mathbf{m}_2) > \log \frac{p_2}{p_1}$$

donde A_0 es una constante.

Nótese que en (4.36) si $S_1 = S_2$, la región es lineal en las variables.

Ejemplo 4.5.- Si en el ejemplo 4.4 se supone que las matrices de varianzas-covarianzas de los dos grupos de enfermos son diferentes y toman los valores

$$S_1 = \begin{pmatrix} 4 & 5 \\ 5 & 9 \end{pmatrix} \quad y \quad S_2 = \begin{pmatrix} 9 & 10 \\ 10 & 16 \end{pmatrix},$$

la región R_1, según la expresión (4.36), resulta

$$R_1 \equiv \frac{1}{2} \log\left(\frac{44}{11}\right) - \frac{1}{2}(x_1 - 6, x_2 - 12) \begin{pmatrix} \frac{9}{11} & \frac{-5}{11} \\ \frac{-5}{11} & \frac{4}{11} \end{pmatrix} \begin{pmatrix} x_1 - 6 \\ x_2 - 12 \end{pmatrix} +$$

$$+ \frac{1}{2}(x_1 - 10, x_2 - 16) \begin{pmatrix} \frac{16}{44} & \frac{-10}{44} \\ \frac{-10}{44} & \frac{9}{44} \end{pmatrix} \begin{pmatrix} x_1 - 10 \\ x_2 - 16 \end{pmatrix} > \log\left(\frac{0.3}{0.7}\right)$$

que equivale a

$$R_1 \equiv 0.227\ x_1^2 + 0.0795\ x_2^2 - 0.227\ x_1 x_2 + 0.545\ x_1 - 0.636\ x_2 - 1.3583 < 0.0$$

Nótese como ahora hay términos en x_1^2 y x_2^2, es decir, se trata de una expresión no lineal.

¤

4.3.1.2.- Minimización del coste esperado

A veces el error de clasificar a un individuo de E_1 como E_2 y el de clasificar a uno de E_2 como de E_1 no tiene la misma gravedad y, por tanto, el mismo coste. Así, no es lo mismo diagnosticar cáncer a un individuo que no lo tiene que no diagnosticarlo a uno que lo tiene. En el primer caso el problema será que se dará al individuo un tratamiento innecesario junto con el daño producido por la impresión psicológica, mientras que en el segundo, el individuo pierde la oportunidad de ser tratado. Por ello, en vez de minimizar la probabilidad total de clasificación errónea es mejor (método b) minimizar el coste esperado, que puede escribirse como:

$$Q_C = C_1 p_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} + C_2 p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \tag{4.37}$$

donde C_1 y C_2 son los costes asociados a los dos tipos de errores.

Con ello, las regiones óptimas resultan

$$R_1 \equiv \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{C_2 p_2}{C_1 p_1} \quad ; \quad R_2 \equiv \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{C_2 p_2}{C_1 p_1} \tag{4.38}$$

Las regiones (4.26) son un caso particular de las (4.38), que resultan de hacer $c_1 = c_2$.

Ejemplo 4.6.- Si en el ejemplo 4.4 los costes de clasificación errónea de un individuo del grupo 1 en el 2 y uno del grupo 2 en el 1 son $c_1 = 1$ y $c_2 = 3$, respectivamente, la región R_1 se transforma en

$$R_1 \equiv (x_1 - 8, x_2 - 14) \begin{pmatrix} \frac{9}{11} & \frac{-5}{11} \\ \frac{-5}{11} & \frac{4}{11} \end{pmatrix} \begin{pmatrix} -4 \\ -4 \end{pmatrix} > \log\left(\frac{3 \times 0.3}{0.7}\right)$$

con lo que se obtiene

$$R_1 \equiv 4x_1 - x_2 - 17.309 < 0.0$$

$$R_2 \equiv 4x_1 - x_2 - 17.309 \geq 0.0$$

¤

4.3.1.3.- Método minimax

El método (c), conocido como el método minimax, es el que minimiza los costes de error manteniéndolos iguales. Por tanto, se minimiza el valor:

(4.39)

$$Q = C_2 P_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \quad \text{sometido a} \quad C_1 P_1 \int_{R_2} f_1(\mathbf{x})\, d\mathbf{x} = C_2 P_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x}$$

que equivale a minimizar

(4.40)

$$Q = C_2 p_2 \int_{R_1} f_2(\mathbf{x})\, d\mathbf{x} \quad \text{sometido a} \quad \int_{R_1} [\, C_1 p_1 f_1(\mathbf{x}) + C_2 p_2 f_2(\mathbf{x}) \,]\, d\mathbf{x} = C_1 p_1$$

El método de los multiplicadores de Lagrange conduce a la función auxiliar

(4.41) $$\int_{R_1} \{ C_2 p_2 f_2(\mathbf{x})\, d\mathbf{x} - \lambda [\, C_1 p_1 f_1(\mathbf{x}) + C_2 p_2 f_2(\mathbf{x}) \,] \}\, d\mathbf{x} + \lambda C_1 p_1$$

con lo que se obtienen las regiones óptimas

(4.42) $$R_1 \equiv \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \alpha \quad ; \quad R_2 \equiv \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \alpha$$

donde α se obtiene de la condición

(4.43) $$\int_{R_1} [\, C_1 p_1 f_1(\mathbf{x}) + C_2 p_2 f_2(\mathbf{x}) \,]\, d\mathbf{x} = C_1 p_1$$

Nótese que las regiones óptimas (4.26), del método (a) son del tipo (4.42).

4.3.1.4.- Método de Fisher

Finalmente, en el método de Fisher (método (d)) se trata de calcular para todo individuo una nueva variable Y lineal en las variables X, es decir, $Y = \lambda' X$, y

tal que diferencie lo mejor posible a los dos grupos. Esta variable tiene como medias $\lambda' m_1$ y $\lambda' m_2$ y como varianzas $\lambda' S_1 \lambda$ y $\lambda' S_2 \lambda$ en los grupos 1 y 2, respectivamente. Si se supone que $S_1 = S_2$, se busca maximizar el valor

$$V = \frac{(\lambda' m_1 - \lambda' m_2)^2}{\lambda' S \lambda} \qquad (4.44)$$

Este método, da exactamente las mismas soluciones que el método (a), a pesar de estar basado en una idea distinta.

4.3.1.5.- Medidas de la calidad de la clasificación

En el problema de clasificación que se ha abordado, tienen mucho interés las siguientes preguntas:

(a) ¿existe realmente una diferencia entre los grupos definida por las variables?, o, en otras palabras, ¿permiten estas variables clasificar a un individuo entre los dos grupos?
(b) ¿cuáles son las probabilidades de clasificación errónea?
(c) ¿pueden reducirse las variables? o, dicho de otra forma, ¿se podría clasificar con menos variables y con las mismas o parecidas probabilidades de error ?

Las dos primeras preguntas tratan de contestarse en los párrafos que siguen, mientras que la última, por ser un tema directamente ligado al de aprendizaje, se deja para el capítulo próximo.

El problema (a) se resuelve teniendo en cuenta que el estadístico

$$F = \frac{n_1 n_2 (n_1 + n_2 - k - 1)}{(n_1 + n_2)(n_1 + n_2 - 2)k} D^2 \qquad (4.45)$$

donde D^2 está dado en (4.33), n_1 y n_2 son los números de individuos en cada grupo y k es la dimensión de **X**, se distribuye según una F de Snedecor con k y n_1+n_2-k-1 grados de libertad.

Por ello, si el nivel de significación es menor que un valor crítico (normalmente 0.05) se concluye que la diferencia entre los grupos es significativa (existe realmente).

La Tabla 4.9 da los valores críticos correspondientes al nivel de significación 0.05 de la distribución F de Snedecor para realizar el contraste anterior.

Para responder a la pregunta (b) puede utilizarse la siguiente estimación del error (4.24), dada por Lachenbruch (1975)

$$\hat{Q}(R_1, R_2, f_1, f_2) = \Phi(-\frac{D}{2}) \qquad (4.46)$$

Grados de libertad del denominador	Grados de libertad del numerador (k)									
	1	2	3	4	5	6	7	8	9	10
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.97
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.80	2.76
14	4.60	3.74	3.34	3.11	2.96	2.85	2.77	2.70	2.65	2.60
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
20	4.35	3.49	3.10	2.87	2.71	2.60	2.52	2.45	2.40	2.35
25	4.24	3.38	2.99	2.76	2.60	2.49	2.41	2.34	2.28	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.34	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.07
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.02
60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.99
80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	2.05	1.99	1.95
100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.92
1000	3.85	3.00	2.61	2.38	2.22	2.10	2.02	1.95	1.89	1.84
∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.83

Tabla 4.9.- Distribución F de Snedecor

Ejemplo 4.7.- Se han registrado los valores de S_1 y S_2 para los problemas E_1 y E_2 obtenidas por 20 casos que poseen el problema E_1 y 20 el E_2, obteniéndose los valores de la Tabla 4.10.

De los datos de la Tabla 4.10 y de (4.30) resulta

$$\bar{x}_1 = \begin{pmatrix} 8.63 \\ 6.64 \end{pmatrix} \; ; \; \bar{x}_2 = \begin{pmatrix} 7.115 \\ 8.49 \end{pmatrix} \; ; \; S^* = \begin{pmatrix} 0.5560 & 0.0213 \\ 0.0213 & 0.8938 \end{pmatrix}$$

de donde

$$S^{*-1} = \begin{pmatrix} 1.8002 & -0.0429 \\ -0.0429 & 1.1198 \end{pmatrix}$$

por lo que, según (4.33), el valor del estadístico D^2 es

$$D^2 = (\bar{x}_1 - \bar{x}_2)' S^{*-1} (\bar{x}_1 - \bar{x}_2) =$$

$$= (1.515, -1.85) \begin{pmatrix} 1.8002 & -0.0429 \\ -0.0429 & 1.1198 \end{pmatrix} \begin{pmatrix} 1.515 \\ -1.85 \end{pmatrix} = 8.20$$

con lo que el estadístico F (ver expresión (4.45)) toma el valor

$$F = \frac{20x20x(20+20-2-1)}{(20+20)(20+20-2)2} 8.20 = 39.92$$

Entrando en la Tabla 4.9 con 2 y 37 grados de libertad se obtiene para α=0.05 el valor 3.26, por lo que el nivel de significación es mucho menor que 0.05 y puede concluirse que la diferencia entre los grupos es muy significativa.

Aplicando ahora la expresión (4.46) y consultando la Tabla 4.8 se obtiene una probabilidad de clasificación errónea

$$\hat{Q} = \Phi\left(-\frac{\sqrt{8.20}}{2}\right) = \Phi(-1.43) = 1 - \Phi(1.43) = 0.0764$$

Suponiendo que las probabilidades "a priori" de E_1 y E_2 son p_1=0.8 y p_2=0.2, respectivamente, una estimación de la región R_1, según (4.27) es

$$R_1 \equiv (x_1 - 7.8725, x_2 - 7.565) \begin{pmatrix} 1.8002 & -0.0429 \\ -0.0429 & 1.1198 \end{pmatrix} \begin{pmatrix} 1.515 \\ -1.85 \end{pmatrix} > \log\left(\frac{0.2}{0.8}\right)$$

que equivale a

$$R_1 \equiv 2.8066\, x_1 - 2.1366\, x_2 - 4.5452 > 0.0$$

$$R_2 \equiv 2.8066\, x_1 - 2.1366\, x_2 - 4.5452 \leq 0.0$$

¤

E_1		E_2	
S_1	S_2	S_1	S_2
8.7	6.2	7.8	9.2
8.2	8.2	6.5	8.9
7.9	5.4	6.8	7.5
7.8	6.8	7.9	8.2
8.5	5.2	6.3	9.3
9.2	6.5	6.9	8.5
9.8	7.2	6.5	8.6
8.6	6.5	8.2	8.5
8.7	5.7	9.1	8.2
7.9	6.7	6.2	9.2
9.3	8.7	8.1	7.9
8.7	7.2	7.7	8.8
8.3	5.3	6.5	6.9
9.2	5.4	6.1	8.4
8.2	8.9	7.2	9.4
9.1	6.5	7.7	8.7
8.6	7.8	7.1	8.9
7.9	5.3	6.0	9.6
9.5	6.5	7.8	7.9
8.5	6.8	5.9	7.2

Tabla 4.10.- Datos de S_1 y S_2 para E_1 y E_2

4.3.2.- Clasificación en más de dos grupos

Seguidamente se generaliza el problema anterior al caso de n grupos. Supóngase una población de individuos divididos en n grupos, que se diferencian por los valores que toman k variables cuantitativas $\mathbf{X}=(X_1,X_2,...,X_k)$. El problema del análisis discriminante se puede plantear como: dado un caso, definido por los valores $\mathbf{x}=(x_1, x_2, ..., x_k)$, determinar a qué grupo pertenece. Este es precisamente el problema general de diagnosis: dado un caso del que se conocen unos síntomas y se desconoce su problema, asignarle una de los n posibles problemas.

Lo que se busca es definir unas funciones de los datos que, en función de sus valores, permitan la asignación. Si para el individuo i-ésimo las variables anteriores toman los valores $\mathbf{X}_i=(x_{i1},x_{i2},...,x_{ik})$, éste podría representarse como un punto en el espacio k-dimensional R^k. Además, conociendo los individuos de cada grupo, resulta fácil calcular los centros de gravedad de la nube de puntos asociada a cada uno de ellos. Estos centros de gravedad dan una idea de donde se sitúan los diferentes grupos.

Ejemplo 4.8.- Considerese dos síntomas S_1 y S_2 de los problemas E_1,E_2,E_3 y E_4, y supongase que son normales, independientes y con las medias y desviaciones típicas que se muestran en la Tabla 4.11. En este caso, la nube de puntos de los casos, definidos por S_1 y S_2, tendrían el aspecto que se muestra en la figura 4.4.

Con objeto de calcular la probabilidad total de clasificación errónea, se supone que se conocen las funciones de densidad, $\mathbf{f_1(x)}$, $\mathbf{f_2(x)}$, ... , $\mathbf{f_n(x)}$ de la variable k-dimensional **X** en los n grupos y que se tiene una regla de decisión, constituida por una partición $\{R_1, R_2, ..., R_n\}$ del conjunto R^k, de forma que si el punto representativo del individuo pertenece a R_j se decide que pertenece al grupo j-ésimo.

PROBLEMA	S_1		S_2	
E_i (m_{i1})	Media (s_{i1})	Desviación típica (m_{i2})	Media (s_{i2})	Desviación típica
E_1	37.5	0.7	70	10.0
E_2	37.5	0.7	100	10.0
E_3	39.0	0.7	100	10.0
E_4	39.0	0.7	70	10.0

Tabla 4.11.- Medias y desviaciones típicas de S_1 y S_2 en los 4 problemas

El conocimiento de las funciones de densidad anteriores permite valorar la calidad de la regla de decisión. La probabilidad de clasificar erróneamente a un individuo del grupo i-ésimo en el grupo j-ésimo es

$$P(j/i) = \int_{R_j} f_i(\mathbf{x})\, d\mathbf{x} \qquad (4.47)$$

Por tanto, la probabilidad total de clasificación incorrecta, PCI, resulta

(4.48) $$PCI = \sum_j \int_{R_j} \sum_{i \neq j} p_i f_i(x)\, dx$$

donde p_i (i=1,2,...,n) son las probabilidades "a priori" de que un individuo pertenezca al grupo i-ésimo y el coste total de clasificación incorrecta es

(4.49) $$PCI = \sum_j \int_{R_j} \sum_{i \neq j} c_{ji}\, p_i\, f_i(x)\, dx$$

donde c_{ji} es el coste de clasificar erróneamente como grupo j lo que es grupo i.

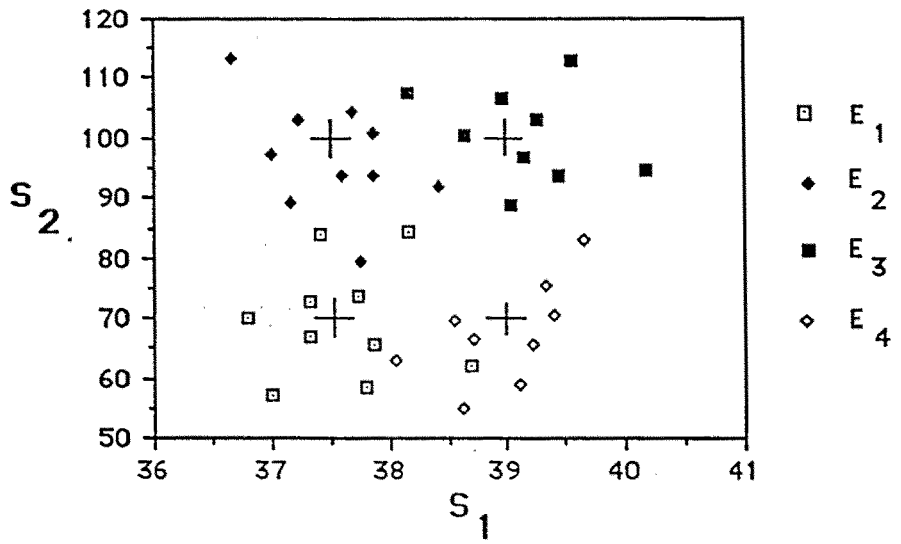


Figura 4.4.- Representación puntual de individuos en R_2

La regla de decisión óptima se obtiene minimizando la expresión anterior con respecto a $\{R_1, R_2, \ldots, R_n\}$, es decir, se trata de elegir la partición $\{R_1, R_2, \ldots, R_n\}$ que hace mínimo el valor de PCI. Este valor mínimo se alcanza para

(4.50) $$R_j = \{x / \min_m \sum_{i \neq m} p_i f_i(x) c_{ji} = \sum_{i \neq j} p_i f_i(x) c_{ji}\} \quad ; \quad j = 1, 2, \ldots, n$$

Para el caso particular de distribuciones normales multivariadas con $c_{ji}=1$ si $i \neq j$ y $c_{ii}=0$, medias $\mathbf{m}_i$ y matrices de varianzas-covarianzas $\mathbf{S}_i$, se tiene

(4.51) $$R_i \equiv p_i \frac{1}{(2\pi)^{k/2} |S_i|^{1/2}} \exp\left[-\frac{1}{2}(x - m_i)' S_i^{-1} (x - m_i)\right] =$$

$$= \max_j \left\{ p_j \frac{1}{(2\pi)^{k/2} |S_j|^{1/2}} \exp\left[-\frac{1}{2}(x - m_j)' S_j^{-1} (x - m_j)\right] \right\}$$

que equivale a

$$(4.52)\qquad R_i \equiv \log p_i - \frac{1}{2}\log\left|S_j\right| - \frac{1}{2}(x-m_i)'S_i^{-1}(x-m_i) = \max_j \left\{ \log p_j - \frac{1}{2}\log\left|S_j\right| - \frac{1}{2}(x-m_j)'S_j^{-1}(x-m_j) \right\}$$

Si los valores m_i y S_i no son conocidos pueden utilizarse, para muestras grandes, los correspondientes valores de la muestra.

	PUNTO			
F. Discriminante	(39.5,110)	(39.5,60)	(36.5,110)	(36.5,60)
D_1	0.002731	0.002673	0.001249	0.000007
D_2	0.002673	0.002731	0.000007	0.001249
D_3	0.000060	0.002730	0.001243	0.001249
D_4	0.002730	0.000060	0.001249	0.001243
PROBLEMA	E_3	E_4	E_2	E_1

Tabla 4.12.- Valores de D_j correspondientes a 4 puntos y problema asociado

Ejemplo 4.9.-Para saber a qué grupo pertenece un caso del ejemplo 4.8 con x_1 como valor de S_1 y x_2 para S_2, bastará calcular

$$D_j = \sum_{i \neq j} p_i f_i(x_1) g_i(x_2) c_{ji} \quad \text{para todo } j$$

donde

$$f_i(x_1) = \frac{\exp\{-(x_1-\mu_{i1})^2/2\sigma_{i1}^2\}}{\sigma_{i1}\sqrt{2\pi}} \quad ; \quad g_i(x_2) = \frac{\exp\{-(x_2-\mu_{i2})^2/2\sigma_{i2}^2\}}{\sigma_{i1}\sqrt{2\pi}}$$

y determinar para que valor de j se obtiene el mínimo. En la tabla 4.12 se dan los valores que corresponden a 4 puntos y la decisión que les corresponde suponiendo que las probabilidades "a priori" son idénticas para los cuatro problemas ($p_1=p_2=p_3=p_4=0.25$) y que los costes c_{ij} son idénticos e iguales a la unidad.

¤

4.3.3.- Datos incompletos

A veces se tienen individuos en los que algunos componentes del vector **X** son desconocidos (datos incompletos) y, en consecuencia, no pueden aplicarse directamente los métodos anteriores.

En estos casos, puede seguirse varias estrategias. Entre las más conocidas destacan:

1.- utilizar sólo los individuos con datos completos
2.- utilizar todas las variables conocidas para estimar medias y varianzas (este método no debe utilizarse si D^2 sale negativo, lo cual ocurre a veces)
3.- sustituir los valores desconocidos por sus medias
4.- estimar los valores desconocidos por los estimadores de regresión dadas las variables conocidas.

4.3.4.- Base de conocimiento y motor de inferencia

La base de conocimiento en el caso de los sistemas expertos basados en modelos de análisis discriminante está constituida por la estructura de las funciones discriminantes junto con sus parámetros.

El motor de inferencia, al igual que en el caso de los modelos logarítmico-lineales se reduce a utilizar las funciones discriminantes con valores medios en las variables o síntomas cuyo valor es desconocido en cada instante.

APRENDIZAJE

5.-APRENDIZAJE

En este capítulo se estudian los métodos para realizar el aprendizaje en los diferentes tipos de sistemas expertos analizados en los capítulos anteriores. Para este estudio se diferenciará entre el aprendizaje paramétrico y el estructural.

5.1.- APRENDIZAJE PARAMÉTRICO

El aprendizaje paramétrico es el que se refiere al conocimiento de los parámetros de la base de conocimiento. Tanto si se trabaja con reglas como si se trabaja con probabilidades, los modelos de incertidumbre dependen de parámetros, cuyo conocimiento preciso es necesario para conseguir un sistema experto fiable. Los mecanismos que permiten una mejor y progresiva estimación de estos parámetros constituyen la base del subsistema de aprendizaje paramétrico.

5.1.1- Aprendizaje en los modelos de probabilidad

En el capítulo 2 se dieron las fórmulas de actualización de los parámetros de los modelos probabilísticos GD, DR y GI cuando se tiene información sobre nuevos casos. En esta sección se dará la idea intuitiva básica que da lugar a ellas. Para ello se considerará el siguiente ejemplo.

Ejemplo 5.1.-Supongamos el ejemplo de la figura 4.2, que se reproduce en la figura 5.1. Imagínese que se conoce que un suceso con el problema E tiene los síntomas: S_1 y S_2 y que no tiene S_3, es decir, $E \cap S_1 \cap S_2 \cap \overline{S_3}$. Entonces este caso está en el mismo que los 25 que aparecen en un círculo en la figura 5.1a con los mismos síntomas. Por tanto, la actualización de los parámetros (frecuencias), es decir, el aprendizaje paramétrico, consiste en incrementar una unidad a esa frecuencia obteniendo el valor 26. Ahora bien, ¿qué pasaría si se conoce sólo que el caso tiene S_2 y no tiene S_3 pero se ignora si tiene S_1, es decir, $E \cap S_2 \cap \overline{S_3}$? En este caso no se sabría si está en el caso de los 25 con S_1 o en el de los 11 sin él (ver figura 5.1b), por lo que no se sabe si sumar una unidad a uno o al otro. La solución, salomónica, está en repartir esa unidad proporcionalmente a los valores anteriores, con lo que el valor 25 pasa a ser 25+25/(25+11) y el valor 11, a ser 11+11/(25+11). De esta forma se obtienen valores fraccionarios o racionales en vez de enteros, pero se consigue actualizar la información hasta el máximo posible.

¤

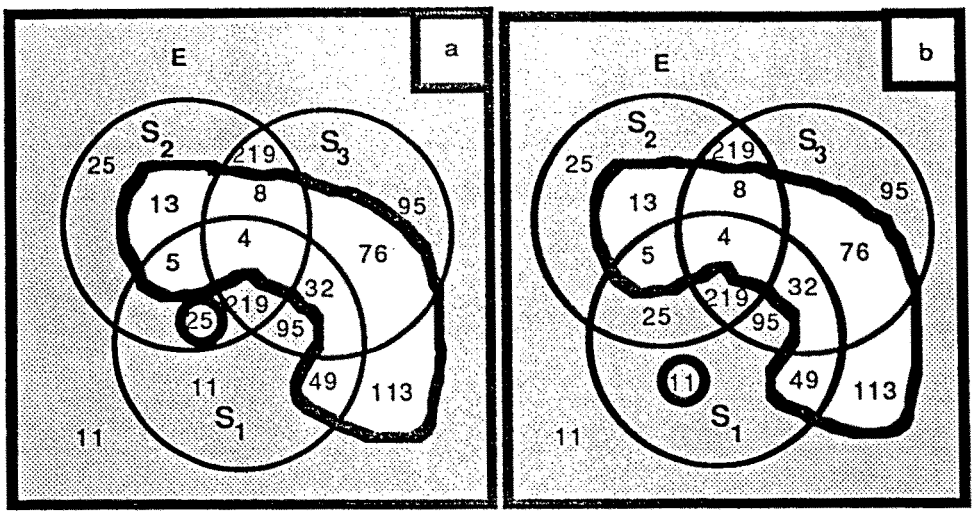


Figura 5.1.- Ilustración del mecanismo de actualización

Con este procedimiento de aprendizaje puede partirse de una base de conocimiento imperfecta e ir mejorándola sucesivamente con la experiencia.

5.1.2- Aprendizaje en modelos logarítmico-lineales o de regresión

En el caso de los modelos logarítmico-lineales o de regresión, el aprendizaje paramétrico se reduce simplemente a una nueva estimación de los parámetros, con la inclusión de los nuevos datos y sin modificar la estructura del modelo.

Puesto que la estimación en este caso es costosa, en tiempo de computación, esta actualización no tiene por qué hacerse después de cada nuevo dato, sino cuando un cierto número de ellos se incorporan a la base de datos.

Ejemplo 5.2.- Imagínese que, partiendo de los datos de la figura 4.2a, se estima inicialmente el modelo dado en el ejemplo 4.1 y que después se conocen los 100 nuevos casos con el problema E y los 40 sin él de la Tabla 5.1.

En este caso, el aprendizaje paramétrico consiste en incorporar estos nuevos datos a la base de datos y estimar el mismo modelo anterior con la totalidad de los datos (antiguos + nuevos), resultando el modelo

$$\log m_{ijkl} = u + u_1(i) + u_2(j) + u_3(k) + u_4(l) + u_{12}(ij) + u_{13}(ik) + u_{14}(il)$$

con los siguientes valores de los parámetros libres

$$u = 3.6442 ; u_1(1) = 0.459 ; u_2(1) = -0.193 ; u_3(1) = -0.321; u_4(1) = 0.375$$
$$u_{12}(1,1) = 0.196 ; u_{13}(1,1) = 0.715 ; u_{14}(1,1) = 0.615$$

S_1	S_2	S_3	con E	sin E
SI	SI	SI	20	1
SI	SI	NO	7	2
SI	NO	SI	15	4
SI	NO	NO	9	10
NO	SI	SI	30	2
NO	SI	NO	5	3
NO	NO	SI	10	3
NO	NO	NO	4	15

Tabla 5.1.- Nuevos datos

S_1	S_2	S_3	con E	sin E
SI	SI	SI	239 (242.3)	5(4.57)
Si	Si	NO	32 (33.45)	7 (7.39)
SI	NO	SI	110 (110.18)	36 (36.31)
SI	NO	NO	20 (15.21)	59 (58.68)
NO	SI	SI	249 (240.85)	10 (9.95)
NO	SI	NO	30 (33.25)	16 (16.09)
NO	NO	SI	105 (109.53)	79 (79.05)
NO	NO	NO	15(15.12)	128 (127.76)

Tabla 5.2.- Valores reales y predicciones de las frecuencias para el caso de la figura 4.2a

La Tabla 5.2 muestra los valores reales de las frecuencias y las dadas por el modelo anterior (entre paréntesis). Nótese que ahora el ajuste no es tan bueno como el anterior (el obtenido en la Tabla 4.1) pero es suficientemente bueno como para mantener la utilidad del modelo.

5.1.3- Aprendizaje en los modelos del análisis discriminante

Análogamente a los casos anteriores, el aprendizaje paramétrico consiste ahora en recalcular las funciones discriminantes o regiones R_i con la inclusión de los nuevos datos o información.

Ejemplo 5.3.- Supóngase ahora que se está en el caso del ejemplo 4.7 y que se conocen 10 nuevos casos con E_1 y 10 con E_2 que se dan en la Tabla 5.3.

E_1		E_2	
S_1	S_2	S_1	S_2
8.2	8.5	7.2	9.6
7.9	6.3	6.8	8.2
9.2	7.1	7.7	7.6
7.1	6.8	8.6	8.5
6.8	5.7	5.3	6.7
8.7	9.4	9.1	9.7
8.2	8.2	6.4	9.2
5.9	7.2	7.8	8.3
9.1	7.1	8.2	9.1
8.7	5.4	6.5	7.2

Tabla 5.3.- Nuevos datos con E_1 y E_2

Ahora, con los datos de las tablas 4.10 y 5.3 y la expresión (4.30) resulta

$$\bar{x}_1 = \begin{pmatrix} 8.413 \\ 6.817 \end{pmatrix} \quad ; \quad \bar{x}_2 = \begin{pmatrix} 7.197 \\ 8.463 \end{pmatrix} \quad ; \quad S^* = \begin{pmatrix} 0.8025 & 0.1290 \\ 0.1290 & 1.0202 \end{pmatrix}$$

de donde

$$S^{*-1} = \begin{pmatrix} 1.2720 & -0.1608 \\ -0.1608 & 1.0005 \end{pmatrix}$$

y el valor del estadístico D^2 es ahora

$$D^2 = (\bar{x}_1 - \bar{x}_2)' S^{*-1} (\bar{x}_1 - \bar{x}_2) =$$

$$= (1.216, -1.646) \begin{pmatrix} 1.272 & -0.1608 \\ -0.1608 & 1.0005 \end{pmatrix} \begin{pmatrix} 1.216 \\ -1.646 \end{pmatrix} = 5.235$$

y el estadístico F (ver expresión (4.45)) toma el valor

$$F = \frac{30x30x57}{60x58x2} 5.235 = 38.58$$

Entrando nuevamente en la Tabla 4.9 con 2 y 57 grados de libertad se obtiene para α=0.01 el valor 3.16, por lo que la diferencia entre los grupos es muy significativa.

Aplicando ahora la expresión (4.46) y consultando la Tabla 4.8 se obtiene una probabilidad de clasificación errónea

$$\hat{Q} = \Phi(-\frac{\sqrt{5.235}}{2}) = \Phi(-1.14) = 1-\Phi(1.14) = 0.127$$

Suponiendo como antes que las probabilidades "a priori" de E_1 y E_2 son p_1=0.8 y p_2=0.2, respectivamente, una estimación de la región R_1, según (4.27) es

$$R_1 \equiv (x_1 - 7.805, x_2 - 7.64)\begin{pmatrix} 1.272 & -0.1608 \\ -0.1608 & 1.0005 \end{pmatrix}\begin{pmatrix} 1.216 \\ -1.646 \end{pmatrix} > \log\left(\frac{0.2}{0.8}\right)$$

que equivale a

$$R_1 \equiv 1.8114\, x_1 - 1.84235\, x_2 + 1.32387 > 0.0$$

$$R_2 \equiv 1.8114\, x_1 - 1.84235\, x_2 + 1.32387 \leq 0.0$$

¤

5.2.- APRENDIZAJE ESTRUCTURAL

Toda modificación en la estructura de la base de conocimiento que dé lugar a una mejora de la misma, forma parte del aprendizaje estructural. Entre estas mejoras, la más corriente es la incorporación de nuevos parámetros, que conduce a una reproducción más fiel del conocimiento. En los apartados que siguen se analizarán varios métodos que permiten seleccionar o decidir, entre un conjunto de parámetros dados, cuales son los más convenientes para representar el conocimiento.

5.2.1.- Aprendizaje en los modelos de probabilidad

Con objeto de contrastar la conveniencia de un modelo dado frente a uno más general, es suficiente estimar, por el método de máxima verosimilitud, los parámetros de ambos modelos y calcular la razón de verosimilitudes. Si M_1 y M_2 son dos modelos con r_1 y r_2 parámetros respectivamente de forma que M_2 generaliza a M_1, se calcula la razón

$$V = \frac{\underset{M_2}{\text{Max}} \prod_{j=1}^{n} P(E_j \cap A_{1j} \cap A_{2j} \cap .. \cap A_{mj})}{\underset{M_1}{\text{Max}} \prod_{j=1}^{n} P(E_j \cap A_{1j} \cap A_{2j} \cap .. \cap A_{mj})}$$

donde n es el tamaño de la muestra (número de casos cuyos síntomas y problemas son conocidos en el momento en que tiene lugar la actualización o el aprendizaje) y la maximización hay que entenderla con respecto al conjunto de parámetros de los modelos M_1 y M_2 respectivamente, y sujetos a sus respectivas restricciones.

El nivel de significación puede ser calculado teniendo en cuenta que el estadístico -2logV converge en probabilidad a $\chi^2 (r_2 - r_1)$.

5.2.2.- Aprendizaje en modelos logarítmico-lineales o de regresión

El aprendizaje estructural en los modelos logarítmico-lineales consiste en elegir un modelo lo más sencillo posible y que reproduzca suficientemente bien las frecuencias reales. Por tanto, se trata de elegir los términos a incluir en el modelo.

Para esta selección existen dos métodos:

(1) partir del modelo saturado (con el máximo de parámetros) y proceder a eliminar términos hasta que el modelo se deteriore
(2) partir de un modelo sencillo y añadir términos hasta que no se logre una mejora sustancial en la calidad del modelo, en el sentido de representar fielmente la realidad mostrada por los datos.

El proceso consiste en calcular los valores de los estadísticos G^2 y χ^2 para un modelo determinado y para el mismo modelo con algún o algunos términos menos y calcular sus incrementos, que se distribuyen estadísticamente como una variable χ^2 con un número de grados de libertad igual a la diferencia del número de términos en ambos modelos.

Los estadísticos G^2 y χ^2 asociados a un modelo dado miden, de dos maneras diferentes, el error global que se comete al utilizar ese modelo en la predicción de las frecuencias reales. Toman valores altos cuando la predicción es mala y valen cero cuando esa predicción es exacta. Por ello, podría pensarse que los incrementos de estos estadísticos correspondientes a dos modelos (uno

extensión de otro) pueden ser utilizados para decidir sobre la conveniencia de incluir los nuevos términos. Sin embargo, puede ocurrir que un modelo dé lugar a una mayor reducción de G^2 y χ^2 que otro y ello sea debido a que tenga más términos. Por ello, con objeto de corregir el efecto del número de grados de libertad, deben utilizarse los niveles de significación (probabilidades) asociados a cada uno de los modelos alternativos a uno dado en vez de los valores de esos estadísticos.

En el método (1) se parte inicialmente del modelo saturado (con todos los grados de libertad), que reproduce perfectamente la realidad de los datos (G^2 y χ^2 son nulos para este modelo). En cada etapa se ajustan todos los modelos que resultan de quitar un término al modelo seleccionado en la etapa anterior, se evaluan los estadísticos G^2 y χ^2, sus incrementos, los niveles de significación asociados a esos incrementos (contribuciones de cada componente) y se elige el mejor de ellos, es decir, el que da el mayor nivel de significación para el incremento. Este proceso se repite hasta que el nivel de significación máximo es menor que un cierto valor crítico (normalmente 0.05), quedándose con el modelo que resulte en la etapa anterior a aquella en la que se produce esta última situación.

En el método (2) se parte inicialmente de un modelo sencillo elegido por el experto (si se desea puede ser el que incluye un solo parámetro) normalmente este reproducirá muy mal la realidad de los datos (G^2 y χ^2 serán muy grandes para este modelo). En cada etapa se ajustan todos los modelos que resultan de añadir un término al modelo seleccionado en la etapa anterior, se evaluan los estadísticos G^2 y χ^2, sus decrementos, los niveles de significación asociados a esos decrementos (contribuciones de cada componente) y se elige el mejor de ellos, es decir, el que da el menor nivel de significación para el decremento. Este proceso se repite hasta que el nivel de significación mínimo es mayor que un cierto valor crítico (normalmente 0.05), quedándose con el modelo que resulte en la etapa anterior a aquella en la que se produce esta última situación. Nótese que este último modelo ya no es satisfactorio.

Para realizar este proceso pueden utilizarse paquetes estadísticos estándar como pueden ser los BMDP, SPSS, SAS, etc. o diseñar un paquete de programas a medida.

Ejemplo 5.4.- Supóngase el ejemplo de la figura 4.2a asociado al problema E y que se busca el mejor modelo jerárquico que reproduce los datos.

Para designar los modelos se usará la notación del BMDP, es decir, combinaciones de letras de forma que se incluyen en el mismo todos los términos

(ES_1S_2,S_3) es el modelo que resulta de incluir todas las interacciones triples E-S_1-S_2, las dobles E-S_1, E-S_2 y S_1-S_2 y las acciones simples E, S_1, S_2 y S_3. El modelo (ES_1S_2, ES_3) incluye las interacciones triples E-S_1-S_2, las dobles E-S_1, E-S_2, S_1-S_2 y E-S_3 y las acciones simples E, S_1, S_2 y S_3. Finalmente, el modelo (ES_1,ES_2,S_3) incluye las dos interacciones dobles E-S_2 y E-S_1 y las acciones simples E, S_1, S_2 y S_3.

En los cuadros que siguen se muestran las etapas que corresponden al método (1). Inicialmente se ajusta el modelo saturado $ES_1S_2S_3$, al que corresponde un valor de G^2 nulo. En la etapa 2 se prescinde de las interacciones de cuarto orden y se obtiene el modelo ($S_1S_2S_3$,ES_2S_3,ES_1S_3,ES_1S_2) con un valor de G^2 de 0.0800, por lo que la contribución del término de cuarto orden es también 0.0800 al que corresponde un nivel de significación de 0.7725. Como éste es mayor del valor crítico 0.05 debe seguirse con la supresión de términos. En la etapa 3 se calculan los modelos que proceden de suprimir un término más al modelo anterior y se calculan los incrementos del estadístico G^2 debidos a ellos, eligiendo el que tiene un nivel de significación (probabilidad) asociado máximo. Este es el modelo ($S_1S_2S_3$,ES_1S_3,ES_1S_2) que surge después de quitarle el término ES_2S_3 al que corresponde un nivel de significación de 0.9963. A partir de este modelo se continua, con el mismo proceso, quitando términos al modelo seleccionado en la etapa anterior hasta la etapa 10 en la que ya se obtiene un nivel de significación de 0.0000, que es menor que el crítico y por tanto se suspende el proceso, quedándose con el modelo final de la etapa 9, que es el ES_3, ES_1, ES_2.

En los cuadros se muestran en negrita los valores máximos de las probabilidades asociadas a los términos suprimidos y los modelos que resultan en cada etapa. También se incluyen los grados de libertad asociados a cada caso.

Obsérvese como el estadístico G^2 crece en cada etapa desde el valor cero (para el mejor modelo posible) a un valor inadmisible de 35.05, en la etapa 10, en la que se acaba el proceso.

Es interesante hacer notar que el modelo (ES_3, ES_1, ES_2) no es otro que el modelo de independencia. Recuérdese que los datos de la figura 4.2a eran precisamente los correspondientes a este caso. Por tanto, la elección del modelo es la correcta.

MODELOS FORMADOS QUITANDO TERMINOS AL MODELO $ES_1S_2S_3$

ETAPA 1			
MODELO	G.L.	G^2	Probabilidad
$ES_1S_2S_3$	0	0.0000	—

ETAPA 2			
MODELO	G.L.	G^2	Probabilidad
$S_1S_2S_3, ES_2S_3, ES_1S_3, ES_1S_2$	1	0.0800	0.7725
Diferencia debida a $ES_1S_2S_3$	1	0.0800	**0.7725**

ETAPA 3			
MODELO	G.L.	G^2	Probabilidad
$S_1S_2S_3, ES_2S_3, ES_1S_3$	2	0.0900	0.9559
Diferencia debida a ES_1S_2	1	0.0100	0.9349
$S_1S_2S_3, ES_2S_3, ES_1S_2$	2	0.1000	0.9532
Diferencia debida a ES_1S_3	1	0.0200	0.9120
$S_1S_2S_3, ES_1S_3, ES_1S_2$	2	0.0800	0.9591
Diferencia debida a ES_2S_3	1	0.0000	**0.9963**
$ES_2S_3, ES_1S_3, ES_1S_2$	2	0.1200	0.9432
Diferencia debida a $S_1S_2S_3$	1	0.0400	0.8550

ETAPA 4			
MODELO	G.L.	G^2	Probabilidad
$S_1S_2S_3, ES_1S_3, ES_2$	3	0.0900	0.9930
Diferencia debida a ES_1S_2	1	0.0100	**0.9350**
$S_1S_2S_3, ES_3, ES_1S_2$	3	0.1000	0.9923
Diferencia debida a ES_1S_3	1	0.0200	0.9119
S_2P, ES_1S_3, ES_1S_2	3	0.1200	0.9896
Diferencia debida a $S_1S_2S_3$	1	0.0400	0.8531

ETAPA 5			
MODELO	G.L.	G^2	Probabilidad
S_1S_2P,ES_1S_3	4	239.94	0.0000
Diferencia debida a ES_2	1	239.85	0.0000
$S_1S_2S_3,ES_3,ES_1,ES_2$	4	0.1000	0.9988
Diferencia debida a ES_1S_3	1	0.0100	**0.9262**
$S_2S_3,S_1S_2,ES_1S_3,ES_2$	4	0.1200	0.9983
Diferencia debida a $S_1S_2S_3$	1	0.0300	0.8680

ETAPA 6			
MODELO	G.L.	G^2	Probabilidad
$S_1S_2S_3,ES_3,ES_2$	5	19.26	0.0017
Diferencia debida a ES_1	1	19.16	0.0000
$S_1S_2S_3,ES_3,ES_1$	5	239.94	0.0000
Diferencia debida a ES_2	1	239.84	0.0000
$S_1S_2S_3,ES_1,ES_2$	5	170.69	0.0000
Diferencia debida a ES_3	1	170.60	0.0000
$S_2S_3,S_1S_3,S_1S_2,ES_2,ES_1,ES_2$	5	0.1200	0.9998
Diferencia debida a $S_1S_2S_3$	1	0.0200	**0.8902**

ETAPA 7			
MODELO	G.L.	G^2	Probabilidad
$S_2S_3,S_1S_3,S_1S_2,ES_3,ES_2$	6	19.67	0.0032
Diferencia debida a ES_1	1	19.55	0.0000
$S_2S_3,S_1S_3,S_1S_2,ES_3,ES_1$	6	240.35	0.0000
Diferencia debida a ES_2	1	240.23	0.0000
$S_2S_3,S_1S_3,S_1S_2,ES_1,ES_2$	6	171.11	0.0000
Diferencia debida a ES_3	1	170.99	0.0000
$S_2S_3,S_1S_3,ES_3,ES_1,ES_2$	6	0.1200	1.0000
Diferencia debida a S_1S_2	1	0.0000	**0.9997**
$S_2S_3,S_1S_2,ES_3,ES_1,ES_2$	6	0.1200	1.0000
Diferencia debida a S_1S_3	1	0.0000	0.9996
$S_1S_3,S_1S_2,ES_3,ES_1,ES_2$	6	0.1200	1.0000
Diferencia debida a S_2S_3	1	0.0000	0.9655

ETAPA 8			
MODELO	G.L.	G^2	Probabilidad
S_2S_3,S_1S_3,ES_3,ES_2	7	25.54	0.0006
Diferencia debida a ES_1	1	25.43	0.0000
S_2S_3,S_1S_3,ES_3,ES_1	7	246.23	0.0000
Diferencia debida a ES_2	1	246.11	0.0000
S_2S_3,S_1S_3,ES_1,ES_2	7	171.40	0.0000
Diferencia debida a ES_3	1	171.28	0.0000
S_2S_3,ES_3,ES_1,ES_2	7	0.1200	1.0000
Diferencia debida a S_1S_3	1	0.0000	**0.9997**
S_1S_3,ES_3,ES_1,ES_2	7	0.5700	0.9992
Diferencia debida a S_2S_3	1	0.4500	0.5017

ETAPA 9			
MODELO	G.L.	G^2	Probabilidad
S_2S_3,ES_3,S_1,ES_2	8	35.05	0.0000
Diferencia debida a ES_1	1	34.93	0.0000
S_2S_3,ES_3,ES_1	8	246.23	0.0000
Diferencia debida a ES_2	1	246.11	0.0000
S_2S_3,ES_1,ES_2	8	176.24	0.0000
Diferencia debida a ES_3	1	176.13	0.0000
ES_3,ES_1,ES_2	8	0.1200	1.0000
Diferencia debida a S_2S_3	1	0.0000	**0.9655**

ETAPA 10			
MODELO	G.L.	G^2	Probabilidad
ES_3,S_1,ES_2	9	35.05	0.0001
Diferencia debida a ES_1	1	34.93	**0.0000**
ES_3,ES_1,S_2	9	331.50	0.0000
Diferencia debida a ES_2	1	331.38	0.0000
S_3,ES_1,ES_2	9	261.51	0.0000
Diferencia debida a ES_3	1	261.39	0.0000

Análogamente al caso del método (1), los cuadros que siguen muestran las etapas que corresponden al método (2). Inicialmente se ajusta el modelo (E, S_1,

S_2, S_3), al que corresponde un valor de G^2 de 627.83. En la etapa 2 se añade una interacción de segundo orden a cada modelo, se calculan los valores de los estadísticos G^2 para todos ellos, los incrementos correspondientes a cada término añadido y se determina el término al que corresponde el menor valor, que resulta ser 0.0000 (término (ES_2)) (de hecho hay varios con valor 0.0000 pero se diferencian en las cifras de menor orden y se ha seleccionado éste por tener menor valor). Como éste es menor que el valor crítico 0.05 debe seguirse con la adición de términos. A partir de este modelo se continua con el mismo proceso hasta la etapa 5 en la que ya se obtiene un nivel de significación de 0.9357, que es mayor que el crítico y por tanto se suspende el proceso, quedándose con el modelo final de la etapa 4, que es el (ES_3,ES_1,ES_2). Nótese que este modelo coincide con el obtenido mediante el método anterior. Sin embargo, no siempre se da esta circunstancia, pues en muchos casos se obtienen modelos distintos, aunque todos son válidos y suficientemente aproximados.

MODELOS FORMADOS AÑADIENDO TERMINOS AL MODELO E, S_1, S_2, S_3

ETAPA 1			
MODELO	G.L.	G^2	Probabilidad
E, S_1, S_2, S_3	11	627.83	0.0000

ETAPA 2			
MODELO	G.L.	G^2	Probabilidad
ES_1,S_2,S_3	10	592.89	0.0000
Diferencia debida a AS_1	1	34.93	0.0000
ES_2,S_1,S_3	10	296.45	0.0000
Diferencia debida a ES_2	1	331.38	**0.0000**
ES_3,S_1,S_2	10	366.43	0.0000
Diferencia debida a AS_3	1	261.39	0.0000
E,S_1S_2,S_3	10	617.58	0.0000
Diferencia debida a S_1S_2	1	10.25	0.0014
E,S_1S_3,S_2	10	618.32	0.0000
Diferencia debida a S_1S_3	1	9.51	0.0020
E,S_1,S_2S_3	10	542.55	0.0000
Diferencia debida a S_2S_3	1	85.27	0.0000

ETAPA 3			
MODELO	G.L.	G^2	Probabilidad
ES_1,ES_2,S_3	9	261.51	0.0000
Diferencia debida a ES_1	1	34.93	0.0000
ES_3,ES_2,S_1	9	35.05	0.0001
Diferencia debida a ES_3	1	261.39	**0.0000**
S_1S_2,ES_2,S_3	9	286.20	0.0000
Diferencia debida a S_1S_2	1	10.25	0.0014
ES_2,S_1S_3	9	286.94	0.0000
Diferencia debida a S_1S_3	1	9.51	0.0020
S_2S_3,ES_2,S_1	9	211.18	0.0000
Diferencia debida a S_2S_3	1	85.27	0.0000

	ETAPA 4		
MODELO	G.L.	G^2	Probabilidad
ES_3,ES_1,ES_2	8	0.1200	1.0000
Diferencia debida a ES_1	1	34.93	**0.0000**
ES_3,S_1S_2,ES_2	8	24.81	0.0017
Diferencia debida a S_1S_2	1	10.25	0.0014
S_1S_3,ES_3,ES_2	8	25.55	0.0013
Diferencia debida a S_1S_3	1	9.51	0.0020
S_2S_3,ES_3,ES_2,S_1	8	36.47	0.0000
Diferencia debida a S_2S_3	1	1.41	0.2348

ETAPA 5			
MODELO	G.L.	G^2	Probabilidad
S_1S_2,ES_1,ES_3,ES_2	7	0.1200	1.0000
Diferencia debida a S_1S_2	1	0.0000	0.9986
S_1S_3,ES_1,ES_3,ES_2	7	0.1300	1.0000
Diferencia debida a VP	1	0.0100	**0.9357**
EV,DP,EP,ED	7	0.1200	1.0000
Diferencia debida a DP	1	0.0000	0.9655

Para el caso de los modelos de regresión existen técnicas análogas que permiten relacionar las variables a incluir mediante técnicas paso a paso ya sea

añadiendolas o suprimiendolas de idéntica manera a los modelos logarítmico lineales (ver paquetes BMDP, SPSS, SAS, etc.).

5.2.3.- Aprendizaje en los modelos del análisis discriminante

El aprendizaje estructural de los modelos basados en el análisis discriminante se basa en la selección de variables a incluir en las funciones discriminantes. El objetivo es encontrar un conjunto mínimo de variables que discrimine suficientemente bien los grupos deseados. Para ello existen técnicas que, paso a paso, aumentan el número de variables o lo disminuyen. En una de las versiones se selecciona inicialmente la variable que mejor discrimina y se van añadiendo variables hasta que la discriminación es suficientemente buena. En la otra, se parte de todas las variables y se van quitando las que menos discriminan hasta que la discriminación se deteriora hasta niveles inaceptables. Aquí se describirá sólo la primera de ellas. En la etapa inicial se calculan los estadísticos

$$F_k = \frac{\sum_i n_i (\bar{x}_{ik} - \bar{\bar{x}}_k)^2 / (g-1)}{\sum_i \sum_j (x_{ijk} - \bar{x}_{ik})^2 / (N-g)}$$

donde i es el número del grupo, j es el número del caso dentro del grupo, k es la variable, g es el número de grupos, N es el tamaño total de la muestra, n_i es el número de casos del grupo i-ésimo, $\bar{x}_{ik}$ es la media de la componente k-ésima en el grupo i-ésimo, $\bar{\bar{x}}_k$ es la media de la componente k-ésima y x_{ijk} es el valor observado de la componente k-ésima en el caso j-ésimo del grupo i-ésimo. Este estadístico, que se distribuye asintóticamente como una F de Snedecor con (g-1) y (N-g) grados de libertad, mide la capacidad de la componente k-ésima para discriminar los grupos.

En la segunda etapa se elige la componente a la que corresponde el máximo valor de F_K y se calcula la función discriminante en función de ella. En las siguientes etapas se repiten las dos anteriores pero utilizando los residuos de la variable, y como covariables las que ya forman parte de la función discriminante. Este proceso se repite hasta que ya se han introducido todas las variables o hasta que ninguna variable discrimina los grupos significativamente (valor del nivel de significación menor que el valor crítico (normalmente 0.05)). Este método es el que utilizan la mayor parte de los paquetes de programas estadísticos conocidos, como el BMDP, SPSS, SAS, etc., que pueden utilizarse para realizar el aprendizaje.

GRUPO 1			GRUPO 2		
S_1	S_2	S_3	S_1	S_2	S_3
1.94	80	116	1.94	110	114
1.92	90	106	1.90	112	114
1.89	80	102	1.88	100	108
1.87	88	104	1.84	106	108
1.86	74	98	1.82	98	92
1.82	82	92	1.80	90	104
1.79	72	98	1.78	94	94
1.76	66	88	1.75	94	87
1.74	74	98	1.72	82	90
1.71	67	85	1.70	86	92
1.66	64	84	1.68	80	90
1.62	60	82	1.64	74	78

Tabla 5.4.- Datos de S_1, S_1 y S_3 para dos grupos

Etapa 1	**(grados de libertad 1 y 22)**
S_1	F=0.0730
S_2	F=18.561
S_3	F=0.1130
Se introduce la variable S_2	
Función de clasificación:	Grupo 1: -24.42+0.63496 S_2
	Grupo 2: -38.09+0.79707 S_2

Etapa 2	**(grados de libertad 1 y 21)**
S_1	F=46.006
S_3	F=18.065
Se introduce la variable S_1	
Función de clasificación: Grupo 1: -406.29 -4.574 S_2+641.20 S_1	
	Grupo 2: -321.00-3.687 S_2+551.91 S_1

Etapa 3	**(grados de libertad 1 y 20)**	
S_3	F=0.008	Fin del proceso

Figura 5.2.- Etapas en la selección de variables

Ejemplo 5.5.- La Tabla 5.4 muestra los valores de S_1, S_1 y S_3 para dos grupos de personas y la figura 5.2 ilustra las etapas que resultan al utilizar el método descrito anteriormente. Las funciones de clasificación resultantes se utilizan de la forma siguiente:

(1) se evaluan las funciones de los dos grupos con los valores del individuo
(2) el individuo se clasifica en el grupo que da mayor valor

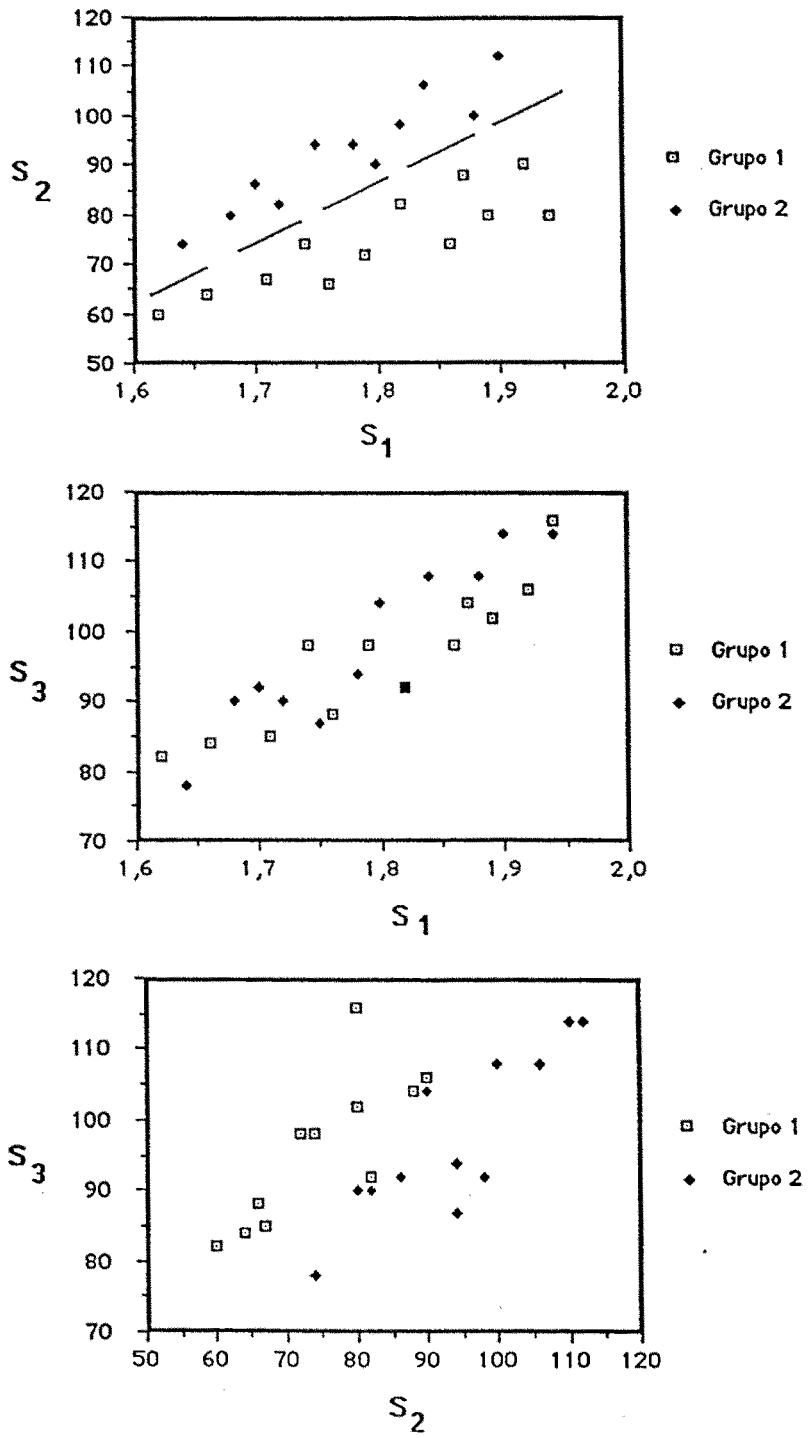


Figura 5.3.- Representación gráfica de los individuos del ejemplo

En este caso, sólo dos variables (S_2 y S_1) son suficientes para conseguir una clasificación perfecta (sin error) de los datos de la Tabla 5.4, como puede comprobarse sin más que calcular las funciones discriminantes para todos los sucesos de la Tabla 5.4 y seguir los dos pasos indicados. No es este el caso de las variablesS_3 y S_1.

En la figura 5.3 se muestran los 24 casos de esta tabla en los diagramas S_2-S_1, S_3-S_1 y S_3-S_2. En ella puede comprobarse claramente que las variables S_2 y S_1 clasifican linealmente (mediante expresiones lineales) y sin error a estos casos.

5.3.- APRENDIZAJE EN EL MODELO DE REDES CAUSALES

En esta sección se da un método de aprendizaje para el modelo de redes causales propuesto por Lauritzen y Spiegelhalter, descrito en el apartado 1.7., basado en el principio de máxima verosimilitud. Se demostrará que este aprendizaje puede tener lugar localmente por ser equivalente, para este método, al aprendizaje global.

Supongamos en primer lugar que tratamos de estimar la función de probabilidad de la variable aleatoria discreta $(X_1,X_2,...,X_n)$ donde los X_i $(i=1,...,n)$ son binarios (X_i=0 ó 1).

Sea $n(x_1,x_2,...,x_n)$ el número de elementos en la muestra tales que $X_i=x_i$ $(i=1,...,n)$, entonces, la función de verosimilitud de la muestra es

$$V = \prod_{(x_1, x_2, \ldots, x_n)} P(x_1, x_2, \ldots, x_n)^{n(x_1, x_2, \ldots, x_n)} \tag{5.1}$$

de donde

$$L = \log V = \sum_{n(x_1, x_2, \ldots, x_n)} n(x_1, x_2, \ldots, x_n) \log P(x_1, x_2, \ldots, x_n) \tag{5.2}$$

y como debe ser

$$\sum_{(x_1, x_2, \ldots, x_n)} P(x_1, x_2, \ldots, x_n) = 1 \tag{5.3}$$

la función de Lagrange a maximizar por el método de la máxima verosimilitud resulta

$$(5.4)\quad Q = \sum_{(x_1, x_2, \ldots, x_n)} \left[n(x_1, x_2, \ldots, x_n) \log P(x_1, x_2, \ldots, x_n) + \lambda P(x_1, x_2, \ldots, x_n) \right] - \lambda$$

de donde

$$(5.5)\quad \frac{\partial Q}{\partial P(x_1, x_2, \ldots, x_n)} = \frac{n(x_1, x_2, \ldots, x_n)}{\hat{P}(x_1, x_2, \ldots, x_n)} + \lambda = 0$$

luego

$$(5.6)\quad \hat{P}(x_1, x_2, \ldots, x_n) = \frac{n(x_1, x_2, \ldots, x_n)}{-\lambda}$$

y de (5.3) y (5.6) queda

$$(5.7)\quad \hat{P}(x_1, x_2, \ldots, x_n) = \frac{n(x_1, x_2, \ldots, x_n)}{\sum_{(x_1, x_2, \ldots, x_n)} n(x_1, x_2, \ldots, x_n)}$$

Con estas estimaciones el valor de

$$(5.8)\quad \hat{P}(x_k / x_1, x_2, \ldots, x_{k-1}) = \frac{\hat{P}(x_1, x_2, \ldots, x_{k-1}, x_k)}{\hat{P}(x_1, x_2, \ldots, x_{k-1})} =$$

$$= \frac{\sum_{x_{k+1} \ldots x_n} n(x_1, x_2, \ldots, x_n)}{\sum_{x_k \ldots x_n} n(x_1, x_2, \ldots, x_n)}$$

Mediante esta fórmula podemos estimar las tablas de probabilidades condicionadas a partir de los potenciales de evidencia. A partir de ahora vamos a demostrar que es equivalente a estimarlos directamente.

En efecto, supongamos ahora que estimamos a partir de la expresión

$$(5.9)\quad P(x_1, x_2, \ldots, x_n) = P(x_1)\, P(x_2 / x_1)\, P(x_3 / x_1\, x_2) \ldots P(x_n / x_1\, x_2, \ldots, x_n)$$

En este caso la función de verosimilitud de la muestra es

$$(5.10)\qquad V = \prod_{(x_1, x_2, \ldots, x_n)} \left[P(x_1)\, P(x_2/x_1) \ldots P(x_n/x_1\, x_2, \ldots, x_n) \right]^{n(x_1, x_2, \ldots, x_n)}$$

de donde

$$(5.11)\qquad L = \log V =$$

$$= \sum_{(x_1, \ldots, x_n)} n(x_1, \ldots, x_n) \left[\log P(x_1) + \log P(x_2/x_1) + \ldots + \log P(x_n/x_1, \ldots, x_n) \right]$$

y como debe ser

$$(5.12)\qquad \begin{cases} \sum_{x_1} P(x_1) = 1 \\ \sum_{x_2} P(x_2/x_1) = 1 \\ \sum_{x_n} P(x_n/x_1 x_2 \ldots x_{n-1}) = 1 \end{cases}$$

la función auxiliar de Lagrange queda

$$(5.13)\qquad Q = \sum_{(x_1, x_2, \ldots, x_n)} n(x_1, x_2, \ldots, x_n) \left[\sum_{i=1}^{n} \log P(x_i/x_1, \ldots, x_{i-1}) \right] +$$

$$+ \sum_{i=1}^{n} \left\{ \lambda_i \left[\sum_{x_i} P(x_i/x_1, \ldots, x_{i-1}) - 1 \right] \right\}$$

de donde

$$(5.14)\qquad \frac{\partial Q}{\partial P(x_k/x_1, \ldots, x_{k-1})} = \frac{\sum_{x_{k+1}, \ldots, x_n} n(x_1, \ldots, x_n)}{\hat{P}(x_k/x_1, \ldots, x_{k-1})} + \lambda_k = 0 \qquad k = 1, \ldots, n$$

y despejando

$$(5.15)\qquad \hat{P}(x_k / x_1,\ldots,x_{k-1}) = \frac{\sum_{x_{k+1},\ldots,x_n} n(x_1,\ldots,x_n)}{-\lambda_k}$$

y finalmente de (5.12) y (5.15) resulta

$$(5.16)\qquad \hat{P}(x_k / x_1,\ldots,x_{k-1}) = \frac{\sum_{x_{k+1},\ldots,x_n} n(x_1,\ldots,x_n)}{\sum_{x_k,\ldots,x_n} n(x_1,\ldots,x_n))} \qquad k = 1,\ldots, n$$

que coincide con (5.8) como se pretendía probar.

5.4.- APRENDIZAJE DE CONCEPTOS PROBABILÍSTICOS

En este apartado se analizará el problema de aprendizaje de clases de conceptos probabilísticos que son una extensión de los utilizados por Valiant (ver apartado 1.8.3). Para ello se utilizará un algoritmo basado en el principio de máxima verosimilitud y se considerarán EJEMPLOS neutros. Se caracterizarán las clases de conceptos aprendibles y no aprendibles a partir de resultados asintóticos basados en el delta método. Para aclarar este método se darán, finalmente, dos ejemplos de aplicación.

En la teoría de aprendizaje de conceptos booleanos (deterministas) (Valiant (1984), Pitt and Valiant (1988)), se supone que cada objeto se representa por una asignación de un conjunto de variables $\{X_i\}$ que toman valores 0 ó 1. Así, cada objeto es un vector $x \in \Omega = \{0, 1\}^t$. Un concepto C es un subconjunto de los 2^t posibles vectores. Sin embargo, en muchos problemas prácticos, existe incertidumbre: la observación de las variables no determina totalmente si el objeto pertenece o no al concepto. Esto es así porque existe otro conjunto de variables no observables, $\{Y_i\}$, i=1,2, . . . , k. En este caso, cada objeto se representa por el vector que resulta de concatenar estas variables $(x, y) \in \{0, 1\}^{t+k}$, donde únicamente X es observable, y la descripción del concepto viene dada por $D^+ \subset \{0, 1\}^{t+k}$.

Siguiendo el enfoque de Valiant, suponemos que los EJEMPLOS son generados por una distribución de probabilidad fija pero no conocida $\pi(x, y)$ en $\{0,1\}^{t+k}$. La distribución marginal de $X \in \{0, 1\}^t$ se denota por

$$m(x) = \sum_{y \in \{0,1\}^k} \pi(x, y)$$

y la probabilidad de que el objeto pertenezca al concepto después de observar x es

$$p(x) = \frac{\sum_{(x,y) \in D^+} \pi(x, y)}{m(x)}$$

En todo el desarrollo posterior se supondrá que el conjunto Ω está particionado en los tres conjuntos siguientes

$$\Omega_0 = \{x \in \Omega / p(x) = 0 \ \ \acute{o} \ \ p(x) = 1\}$$
$$\Omega_1 = \{x \in \Omega / 0 < p(x) < 1\}$$
$$\Omega_2 = \{x \in \Omega / p(x) \ \text{ conocido}\}$$

Nótese que los conceptos deterministas corresponden al caso especial en el que $\Omega_1 = \Phi$.

Supondremos que m es conocido y que el objetivo que se persigue, a la hora de aprender un concepto probabilístico, es la estimación de la destribución p(x). Se dirá entonces que P es un concepto probabilístico si viene dado por una función $\left\{p(x), x \in \Omega = \{0, 1\}^t\right\}$, donde p(x) representa la probabilidad de que un objeto con vector asociado x sea un EJEMPLO positivo de P.

Definición 5.1.- *(Vector parcial y total).*- Un vector $x^* \in \Omega^* = \{0,1,*\}^t$ se dice que es total si cada variable está determinada, es decir, si $x^* \in \Omega$, en caso contrario se dirá que es un vector parcial.

Durante el proceso de aprendizaje una observación es un vector, completo o incompleto, $x^* \in \Omega^* = \{0,1,*\}^t$, y un valor h que puede ser h="+" si el EJEMPLO observado pertenece al concepto o h="-" en caso contrario. Nótese que cada $x^* \in \Omega^*$ representa el subconjunto

$$C_{x^*} = \{x \in \Omega / x_i = x_i^* \ \text{ si } \ x_i^* \neq "*", \forall i\}$$

Las funciones m y p pueden extenderse a Ω^* de la forma siguiente:

(5.17) $$m(x^*) = \sum_{x \in C_{x^*}} m(x) \quad ; \quad \forall\, x^* \in \Omega^*$$

y

(5.18) $$p(x^*) = \frac{\sum_{x \in C_{x^*}} m(x)\, p(x)}{m(x^*)} \quad ; \quad \forall\, x^* \in \Omega^*$$

Durante el proceso de aprendizaje se obtienen diferentes observaciones. Despues de un número de ellas, n, o después de un tiempo τ en el que se han producido n_τ observaciones, el proceso de aprendizaje termina y deduce un programa que obtiene una función q(x) para cada x en Ω. La función q(x) (una estimación de p(x)) es el concepto aprendido, y permite calcular la probabilidad de que un objeto con vector x pertenezca al concepto.

La bondad de q(x) se mide por el valor esperado de una función de pérdida r(p(x),q(x)), dada por

(5.19) $$L(q) = E[r(p(x), q(x))] = \sum_{x \in \Omega} m(x)\, r(p(x), q(x))$$

Diremos que la precisión de q(x) es ε si

(5.20) $$\sum_{x \in \Omega} m(x)\, r(p(x), q(x)) \leq \varepsilon$$

Como el proceso de observación es aleatorio no se puede asegurar una precisión ε con probabilidad 1 para ε suficientemente pequeño, únicamente podemos dar un nivel α, y asegurar que

(5.21) $$Pr\left(\sum_{x \in \Omega} m(x)\, r(p(x), q(x)) \leq \varepsilon \right) = 1 - \alpha$$

donde Pr representa la distribución de probabilidad de un proceso muestral.

Una de las funciones de pérdida más importante es:

$$r_1(z) = \begin{cases} 0 & , \quad \text{if } z^2 \leq \delta^2 \\ 1 & , \quad \text{if } z^2 > \delta^2 \end{cases}$$

y entonces

(5.22) $$L_1(q) = \sum_{(p(x)-q(x))^2 > \delta^2} m(x)$$

El principal problema en la teoría del aprendizaje (Valiant (1984)) es encontrar clases de conceptos que sean aprendibles en tiempo razonable. Extenderemos la definición de Valiant para conceptos probabilísticos en el sentido siguiente

Definición 5.2.-*(Concepto aprendible).*-Sea $\mathfrak{F}$ una clase de conceptos aprendibles. Sea r una función de pérdida, con parámetro asociado $\delta>0$, y $\varepsilon>0$ y $\alpha<0$ la precisión y el nivel de aprendizaje. Diremos que $\mathfrak{F}$ es aprendible a partir de EJEMPLOS sii existe un polinomio G(u,v,w,z) y un algoritmo de aprendizaje A tales que $\forall P \in \mathfrak{F}$, con función asociada p, el algoritmo A para en tiempo, o después de observar un número de EJEMPLOS, $G(T(P), \delta^{-1}, \varepsilon^{-1}, \alpha^{-1})$, donde T(P)=t es una medida del tamaño de P, y A devuelve una función q, tal que el concepto Q, dado por q pertenece a $\mathfrak{F}$, y

$$Pr\left(\sum_{x \in \Omega} m(x)\, r(p(x), q(x), \delta) \leq \varepsilon\right) = 1 - \alpha$$

5.4.1.- Algoritmos de aprendizaje

A continuación daremos dos algoritmos para aprender un concepto P basado en muestreos multinomial y poissoniano.

5.4.1.1.- Muestreo multinomial

Supongamos que durante el proceso de muestreo las observaciones son vectores aleatorios independientes, identicamente distribuidos y que la probabilidad de las observaciones {x,+} y {x,-} son m(x)p(x) y m(x)[1-p(x)], respectivamente. Supongamos también que tenemos una muestra de n objetos neutros no necesariamente totales. Es decir, se tiene el siguiente conjunto dado por la muestra

(5.23) $$S = \{n_x^h / x \in \Omega^* ; h \in \{+,-\}\} \quad ; \quad \sum_{n_x^h \in S} n_x^h = n$$

donde n_x^h es el número de objetos en la muestra del tipo {x,h}.

La función de verosimilitud de la muestra, que depende del conjunto de parámetros $\{p(x) / x \in \Omega\}$, viene dada por

$$(5.24) \qquad V = \prod_{y \in \Omega} \{ [p(y)]^{n_y^+} [1 - p(y)]^{n_y^-} \} \times$$

$$\times \prod_{y \in \Omega^* - \Omega} \left\{ \left(\sum_{z \in C_y} m(z)\, p(z) \right)^{n_y^+} \left(\sum_{z \in C_y} m(z)\, [1 - p(z)] \right)^{n_y^-} \right\}$$

y su logaritmo es

$$(5.25) \qquad L = \log V = \sum_{y \in \Omega} \{ n_y^+ \log p(y) + n_y^- \log [1 - p(y)] \} +$$

$$+ \sum_{y \in \Omega^* - \Omega} \{ n_y^+ \log [\sum_{z \in C_y} m(z)\, p(z)] + n_y^- \log [\sum_{z \in C_y} m(z)\, (1 - p(z))] \}$$

Derivando se obtienen las ecuaciones de verosimilitud:

$$(5.26) \qquad \frac{\partial L(q)}{\partial p(x)} = \frac{n_x^+}{q(x)} - \frac{n_x^-}{1 - q(x)} +$$

$$+ \sum_{y \in S_x} \left[\frac{n_y^+ m(x)}{\sum_{z \in C_y} m(z)\, q(z)} - \frac{n_y^- m(x)}{\sum_{z \in C_y} m(z)[1 - q(z)]} \right] = 0 \quad ; \quad x \in \Omega$$

donde

$$(5.27) \qquad S_x = \{ x^* \in \Omega^* - \Omega / x \in C_{x^*} \}$$

y q(x) es el estimador de máxima verosimilitud de p(x).

5.4.1.2.- Muestreo poissoniano

Supongamos ahora un muestreo poissoniano de duración τ, es decir, que las variables aleatorias n_x^+ y n_x^- son variables aleatorias independientes de Poisson con valores medios $km(x)p(x)\tau$ y $km(x)[1-p(x)]\tau$, respectivamente. En este caso se tiene el siguiente conjunto dado por la muestra

(5.28) $$S=\{n_x^h / x\in \Omega^* \ ; h\in\{+,-\}\}$$

donde, como antes, n_x^h es el número de objetos en la muestra del tipo {x,h}.

La función de verosimilitud de la muestra es

(5.29) $$V=\prod_{y\in\Omega}\{\frac{\exp\{-k\tau\, m(y)\}[m^+(y)]^{n_y^+}[m^-(y)]^{n_y^-}(k\tau)^{n_y^+ + n_y^-}}{n_y^+!\,n_y^-!}\}\ x$$

$$x\prod_{y\in\Omega^*-\Omega}\left\{\frac{\exp\{-k\tau\sum_{z\in C_y} m(z)\}[\sum_{z\in C_y} m^+(z)]^{n_y^+}[\sum_{z\in C_y} m^-(z)]]^{n_y^-}(k\tau)^{n_y^+ + n_y^-}}{n_y^+!\,n_y^-!}\right\}$$

donde

(5.30) $$m^+(y)=m(y)p(y) \quad ; \quad m^-(y)=m(y)[1-p(y)].$$

Eliminando los términos constantes en (5.29) obtenemos (5.24). Por tanto, el estimador q(x) coincide en ambos tipos de muestreo.

5.4.1.3.- Estimaciones de vectores totales de la muestra

Si todos los vectores en la muestra son totales el sistema (5.26) es

(5.31) $$\frac{n_x^+}{q(x)}-\frac{n_x^-}{1-q(x)}=0 \quad ; \quad x\in\Omega \quad \Rightarrow \quad q(x)=\frac{n_x^+}{n_x^+ + n_x^-}$$

Nótese que q(x) está bien definido salvo cuando $n_x^+ + n_x^- = 0$, en este caso se define q(x)=1/2. Es decir,

(5.32) $$q(x)=\begin{cases}\left.\begin{cases}\frac{n_x^+}{n_x^+ + n_x^-} & \text{si } n_x^+ + n_x^- > 0\\ \frac{1}{2} & \text{si } n_x^+ + n_x^- = 0\end{cases}\right\} & \text{si } x\notin\Omega_2\\ p(x) & \text{si } x\in\Omega_2\end{cases}$$

Las probabilidades asociadas a los dos casos diferentes en la expresión (5.32) cuando $x \notin \Omega_2$ son 1- $[1-m(x)]^n$ y $[1-m(x)]^n$, respectivamente, si el muestreo es multinomial y 1- exp [- $k\tau m(x)$] y exp [- $k\tau m(x)$], respectivamente, si el muestreo es poissoniano. Por ello, podemos manejar la variable aleatoria q(x) como una combinación lineal convexa de las dos variables obvias con las probabilidades anteriores como pesos.

5.4.2.-Familias de conceptos aprendibles

Con objeto de calcular la probabilidad en la expresión (5.21), para la función de pérdida (5.22) consideramos las variables aleatorias

$$(5.33) \qquad Z_x = \begin{cases} m(x) & \text{si } [q(x)-p(x)]^2 > \delta^2 \\ 0 & \text{en otro caso} \end{cases} \quad ; \quad x \in \Omega$$

y llamamos

$$(5.34) \qquad s(x) = Pr[Z_x = m(x)] = 1 - Pr[-\delta \le q(x) - p(x) \le \delta]$$

Nótese que

$$(5.35) \qquad L_1(q) = \sum_{x \in \Omega} Z_x$$

Los valores medios de Z_x y $L_1(q)$ son

$$(5.36) \qquad E[Z_x] = m(x)\,s(x) \quad ; \quad E[L_1(q)] = \sum_{x \in \Omega} m(x)\,s(x)$$

Entonces (5.21) puede escribirse

$$(5.37) \qquad Pr\ [L_1(q) \le \varepsilon\ ] = 1 - \alpha$$

5.4.2.1.- Caso multinomial

Si estamos en el caso de una muestreo multinomial el teorema central del límite permite afirmar que

$$(5.38)\qquad \sqrt{n}\left[\begin{pmatrix} \frac{n_x^+}{n} \\ \frac{n_x^+}{n} \end{pmatrix} - \begin{pmatrix} m^+ \\ m^- \end{pmatrix}\right] \xrightarrow{D} N\left(0\,,\begin{pmatrix} m^+(1-m^+) & -m^+m^- \\ -m^+m^- & m^-(1-m^-) \end{pmatrix}\right)$$

Teniendo en cuenta en la expresión (5.32) la primera parte para el caso en que x no pertenezca a Ω_2 y que

$$(5.39)\qquad \frac{\partial q(x)}{\partial n_x^+} = \frac{n_x^-}{(n_x^+ + n_x^-)^2} \quad ; \quad \frac{\partial q(x)}{\partial n_x^-} = \frac{-n_x^+}{(n_x^+ + n_x^-)^2}$$

aplicando el delta método (Bishop et al (1975))

$$(5.40)\qquad \sqrt{n}\,[q(x)-p(x)] \xrightarrow{D}$$

$$\xrightarrow{D} N\left(\mathbf{0},\begin{pmatrix} \frac{m^-}{m^2} & \frac{-m^+}{m^2} \end{pmatrix}\begin{pmatrix} m^+(1-m^+) & -m^+m^- \\ -m^+m^- & m^-(1-m^-) \end{pmatrix}\begin{pmatrix} \frac{m^-}{m^2} \\ \frac{-m^+}{m^2} \end{pmatrix}\right) = N\left(\mathbf{0}\,,\frac{m^-m^+}{m^3}\right)$$

se tiene

$$(5.41)\qquad [q(x)-p(x)] \xrightarrow{D} N\left(0\,,\frac{m^-m^+}{m^3 n}\right)$$

y entonces, por la expresión (5.34) tipificando la variable

$$(5.42)\qquad 1-F_{N\left(0,\frac{m^-m^+}{m^3 n}\right)}(\delta) = \frac{s(x)}{2} \iff 1-F_{N(0,1)}\left(\sqrt{\frac{\delta^2 m^3 n}{m^+ m^-}}\right) = \frac{s(x)}{2}$$

y teniendo en cuenta que (Galambos (1987))

$$(5.43)\qquad \lim_{x\to\infty} x\,[1-F_{N(0,1)}(x)]\;\exp\left(\frac{x^2}{2}\right) = (2\pi)^{-\frac{1}{2}}$$

resulta

$$(5.44)\qquad \lim_{n\to\infty} \sqrt{\frac{\delta^2 m^3 n}{m^+ m^-}}\;\frac{s(x)}{2}\;\exp\left(\frac{\delta^2 m^3 n}{2\,m^+ m^-}\right) = (2\pi)^{-\frac{1}{2}}$$

y entonces, para $n \to \infty$ podemos escribir

$$(5.45) \qquad s(x) \approx \sqrt{\frac{2\, m^- m^+}{\pi\, \delta^2\, m^3\, n}} \quad \exp\left(-\frac{\delta^2\, m^3\, n}{2\, m^- m^+}\right) \leq \sqrt{\frac{2\, m^- m^+}{\pi\, \delta^2\, m^3\, n}}$$

La ecuación (5.37), teniendo en cuenta la desigualdad de Markov

$$(5.46) \qquad 1-\alpha = P\,[L_1\,(q) \leq \varepsilon\,] \geq 1 - \frac{E\,[L_1\,(q)\,]}{\varepsilon}$$

y las expresiones (5.36) y (5.45), es

$$(5.47) \quad \varepsilon\,\alpha \leq \sum_{x \in \Omega} m(x)\, s(x) \leq$$

$$\leq \sum_{x \in \Omega_1} [1-(1-m(x))^n] \sqrt{\frac{2\, m(x)\, p(x)\,(1-p(x))}{\pi\, \delta^2\, n}} + \sum_{x \in \Omega_0 \cup \Omega_1} (1-m(x))^n m(x) \leq$$

$$\leq [1-(1-\frac{1}{|\Omega_1|})^n] \sqrt{\frac{|\Omega_1|}{2\pi\delta^2 n}} + (1-\frac{1}{|\Omega_1|+|\Omega_0|})^n$$

donde en la última desigualdad hemos considerado que $\frac{m^-(x)\, m^+(x)}{m(x)} \leq \frac{m(x)}{4}$ y se ha maximizado respecto a m(x).

Se pueden distinguir los siguientes casos:

Caso (i) : $|\Omega_1| = 0$

En este caso, la desigualdad (5.47) se convierte

$$(5.48) \qquad \varepsilon\,\alpha \leq (1-\frac{1}{|\Omega_0|})^n$$

de lo que se deduce

$$(5.49) \qquad n \leq \frac{\log\,(\varepsilon^{-1}) + \log\,(\alpha^{-1})}{\left|\log\,(1-\frac{1}{|\Omega_0|})\right|} \approx |\Omega_0|\,[\log\,(\varepsilon^{-1}) + \log\,(\alpha^{-1})\,]$$

Caso (ii) : $|\Omega_0| = 0$

En este caso, la desigualdad (5.47) se convierte

$$(5.50) \qquad \varepsilon\,\alpha \le [1-(1-\frac{1}{|\Omega_1|})^n]\sqrt{\frac{|\Omega_1|}{2\pi\delta^2 n}} + (1-\frac{1}{|\Omega_1|})^n \approx \sqrt{\frac{|\Omega_1|}{2\pi\delta^2 n}}$$

de lo que se deduce

$$(5.51) \qquad n \le \frac{|\Omega_1|}{2\pi\,\delta^2\alpha^2\,\varepsilon^2}$$

Caso (iii) : $|\Omega_0| \neq 0$ y $|\Omega_1| \neq 0$

En este caso, la desigualdad (5.47) se convierte

$$(5.52) \qquad \varepsilon\,\alpha \le [1-(1-\frac{1}{L})^n]\sqrt{\frac{L}{2\pi\,\delta^2 n}} + (1-\frac{1}{2L})^n \approx \sqrt{\frac{L}{2\pi\delta^2 n}}$$

donde $L = \max(|\Omega_0|, |\Omega_1|)$. Y, por tanto, se tiene

$$(5.53) \qquad n \le \frac{\max(|\Omega_0|, |\Omega_1|)}{2\pi\,\delta^2\alpha^2\,\varepsilon^2}$$

5.4.2.2.- Caso poissoniano

Si consideramos el caso de muestreo poissoniano, por el teorema central del límite, se tiene

$$(5.54) \qquad \sqrt{t}\left[\begin{pmatrix}\frac{n_x^+}{n}\\ \frac{n_x^+}{n}\end{pmatrix} - \begin{pmatrix}m^+\\ m^-\end{pmatrix}\right] \xrightarrow{D} N\left(\mathbf{0}, \begin{pmatrix}m^+ & 0\\ 0 & m^-\end{pmatrix}\right)$$

Y aplicando el delta método en este caso

$$(5.55)\qquad \sqrt{\tau}\,[q(x)-p(x)] \xrightarrow{D} N\left(0, \begin{pmatrix} \frac{m^-}{m^2} & \frac{-m^+}{m^2} \end{pmatrix} \begin{pmatrix} m^+ & 0 \\ 0 & m^- \end{pmatrix} \begin{pmatrix} \frac{m^-}{m^2} \\ \frac{-m^+}{m^2} \end{pmatrix}\right) =$$

que implica

$$(5.56)\qquad [q(x)-p(x)] \xrightarrow{D} N\left(0, \frac{m^- m^+}{m^3 \tau}\right)$$

Notese que la expresión (5.56) es la misma que (5.41) donde n pasa a ser τ. Así, (5.45) vuelve a verificarse con la sustitución obvia y la desigualdad (5.47) es

$$(5.57)\quad \varepsilon\,\alpha \le \sum_{x \in \Omega} m(x)\, s(x) \le$$

$$\le \sum_{x \in \Omega_1} \{1-\exp[-k\tau\, m(x)]\} \sqrt{\frac{2\, m(x)\, p(x)(1-p(x))}{\pi\, \delta^2\, \tau}} +$$

$$+ \sum_{x \in \Omega_0 \cup \Omega_1} \exp[-k\tau m(x)]\, m(x) \le$$

$$\le [1-\exp(-\frac{k\tau}{|\Omega_1|})] \sqrt{\frac{|\Omega_1|}{2\pi\delta^2\tau}} + \exp[-\frac{k\tau}{|\Omega_0|+|\Omega_1|}]$$

Se pueden distinguir los siguientes casos:

Caso (i) : $|\Omega_1| = 0$

En este caso la desiguadad (5.57) es

$$(5.58)\qquad \varepsilon\,\alpha \le \exp[-\frac{k\tau}{|\Omega_0|}]$$

de lo que se deduce

$$(5.59)\qquad \tau \le \frac{|\Omega_0|\,[\log(\varepsilon^{-1}) + \log(\alpha^{-1})]}{k}$$

Caso (ii) : $|\Omega_0| = 0$

En este caso la desiguadad (5.57) es

(5.60) $$\varepsilon\,\alpha \le [1-\exp(-\frac{k\tau}{|\Omega_1|})]\sqrt{\frac{|\Omega_1|}{2\pi\delta^2\tau}}+\exp[-\frac{k\tau}{|\Omega_1|}]\approx\sqrt{\frac{|\Omega_1|}{2\pi\delta^2\tau}}$$

y, por tanto,

(5.61) $$\tau \le \frac{|\Omega_1|}{2\pi\,\delta^2\alpha^2\,\varepsilon^2}$$

Caso (iii) : $|\Omega_0| \neq 0$ y $|\Omega_1| \neq 0$

En este caso la desiguadad (5.57) es

(5.62) $$\varepsilon\,\alpha \le [1-\exp(-\frac{k\tau}{L})]\sqrt{\frac{L}{2\pi\,\delta^2\tau}}+\exp[-\frac{k\tau}{2L}]\approx\sqrt{\frac{L}{2\pi\,\delta^2\tau}}$$

donde de nuevo $L=\max(|\Omega_0|,|\Omega_1|)$. Y, por tanto,

(5.63) $$\tau \le \frac{\max(|\Omega_0|,|\Omega_1|)}{2\pi\,\delta^2\alpha^2\,\varepsilon^2}$$

Las expresiones (5.49), (5.51), (5.53), (5.59), (5.61) y (5.63) demuestran que la siguiente familia de conceptos

$$\mathcal{F}=\{P\,/\,|\Omega_0|=Q(t)\ \text{ and }\ |\Omega_1|=R(t)\}$$

donde Q(t) y R(t) son polinomios en t, son aprendibles.

Con objeto de ilustrar los resultados anteriores se darán a continuación dos ejemplos de aplicación.

Ejemplo 5.6.- Supongamos una población de objetos que vienen definidos a través de tres síntomas binarios distintos: S_1, S_2 y S_3. Definimos el concepto E y observamos que las probabilidades m(x) y p(x) son las dadas en la Tabla 5.5. En la figura 5.4 aparece representado el concepto E y los valores de m(x)p(x) y m(x)[1-p(x)] asociados a las diferentes combinaciones de síntomas.

$x=(S_1, S_2, S_3)$	m(x)	p(x)
(0,0,0)	0.06	0.001
(0,0,1)	0.14	0.002
(0,1,0)	0.06	0.002
(0,1,1)	0.14	0.998
(1,0,0)	0.09	0.002
(1,0,1)	0.21	0.998
(1,1,0)	0.09	0.997
(1,1,1)	0.21	0.999

Tabla 5.5.- Valores de m(x) y p(x) para el concepto E

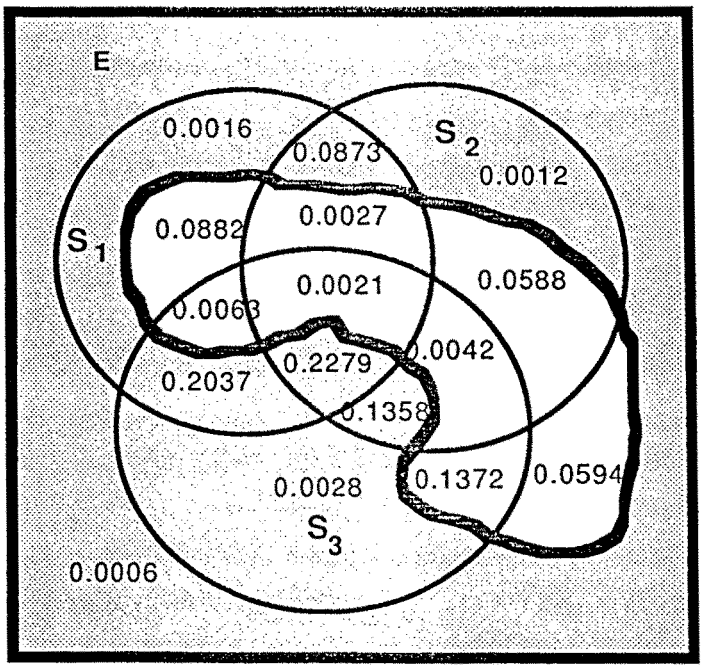


Figura 5.4.- Concepto E y valores de m(x)p(x) y m(x)[1-p(x)]

En este caso $|\Omega_0| = 0$ y $|\Omega_1| = 8$ y supondremos α=0.10, ε=0.10 y δ=0.10. Utilizando ahora la desiguadad (5.46):

$$\varepsilon\,\alpha \leq \sum_{x \in \Omega_1} [1-(1-m(x))^n] \sqrt{\frac{2\, m^-(x)\, m^+(x)}{\pi\, \delta^2\, m(x)\, n}} + \sum_{x \in \Omega_0 \cup \Omega_1} (1-m(x))^n m(x)$$

el método de la bisección nos conduce a que el tamaño de la muestra es n=8228.

¤

Ejemplo 5.7.- La figura 5.5.a muestra un mecanismo de seguridad de un recinto que consta a su vez de dos subsistemas. El primero, C, consiste en una cámara de video que trasmite la imagen a un computador para su análisis. Después del análisis, el computador decide si activa un relé que cierra un circuito eléctrico con una batería que activa una alarma. El segundo subsistema, F, consiste en una célula fotoeléctrica, D, que cierra otro circuito eléctrico E con una alarma activada por una batería. La figura 5.5.b muestra las reglas asociadas con el sistema de alarma. Nótese que el primer sistema ha sido simplificado como hardware, A, y software, B, y las reglas son interpretadas en un sentido estricto (las conclusiones son muy probables pero no seguras).

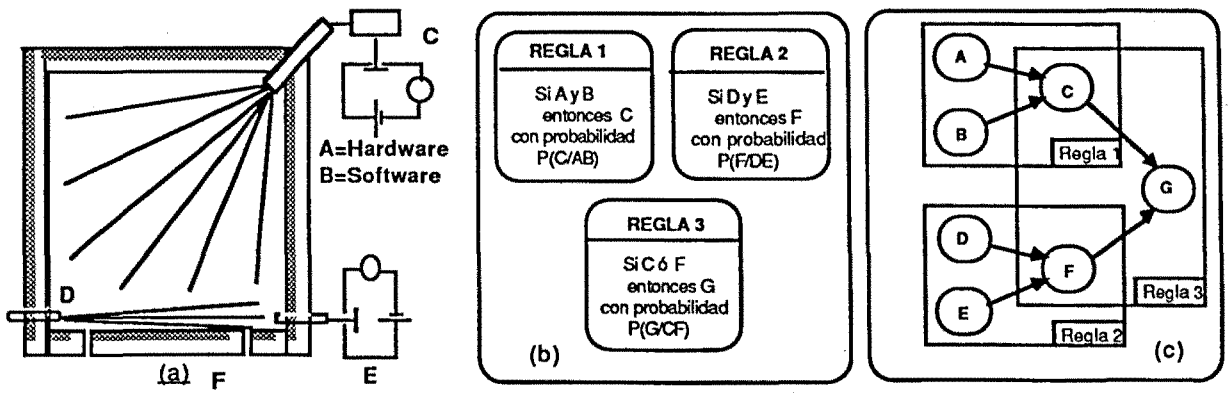


Figura 5.5.- Sistema de seguridad: reglas y diagrama de influencias

La Tabla 5.6 la distribución de probabilidad m(x) donde $x \in \{0, 1\}^4$. Las componentes del vector x están asociadas a las componentes A, B, D y E. El valor 1 indica que el elemento asociado funciona correctamente y el valor 0 que existe una anomalía. Las probabilidades $p_1(x)$, $p_2(x)$ y $p_3(x)$ definen los conceptos C, F y G, respectivamente.

Los tamaños de las muestras requeridos para aprender los conceptos C, F y G con α=0.10, β=0.10 y δ=0.10, obtenidos por el mismo método descrito en el ejemplo 5.6, son 11998, 2497 y 110, respectivamente.

Nótese que se tiene $\{|\Omega_0| = 0, |\Omega_1| = 16\}$, $\{|\Omega_0| = 10, |\Omega_1| = 6\}$ y $\{|\Omega_0| = 16, |\Omega_1| = 0\}$, para los conceptos C, F y G, respectivamente. Obsérvese también como el tamaño de la muestra aumenta cuando crece $|\Omega_1|$.

x=(A,B,D,E)	m(x)	$p_1(x)$	$p_2(x)$	$p_3(x)$
(0,0,0,0)	0.0144	0.001	0.000	0.00
(0,0,0,1)	0.0216	0.001	0.002	0.00
(0,0,1,0)	0.0336	0.002	0.000	0.00
(0,0,1,1)	0.0504	0.001	0.998	1.00
(0,1,0,0)	0.0336	0.001	0.000	0.00
(0,1,0,1)	0.0504	0.002	0.000	0.00
(0,1,1,0)	0.0784	0.001	0.001	0.00
(0,1,1,1)	0.1176	0.001	1.000	1.00
(1,0,0,0)	0.0216	0.001	0.000	0.00
(1,0,0,1)	0.0324	0.001	0.000	0.00
(1,0,1,0)	0.0504	0.001	0.001	0.00
(1,0,1,1)	0.0756	0.001	1.000	1.00
(1,1,0,0)	0.0504	0.998	0.001	1.00
(1,1,0,1)	0.0756	0.998	0.000	1.00
(1,1,1,0)	0.1176	0.999	0.000	1.00
(1,1,1,1)	0.1764	0.998	0.997	1.00

Tabla 5.6.- Valores de m(x) y p(x) para tres conceptos diferentes

CONCLUSIONES

6.-CONCLUSIONES

Entre las conclusiones más importantes de esta tesis destacan las siguientes:

1. - El modelo probabilístico de independencia no es adecuado para reproducir la realidad y el modelo de dependencia general conduce a tan elevado número de parámetros que no es posible implementarlo en los ordenadores actuales. El modelo de síntomas relevantes, propuesto en esta tesis, resuelve satisfactoriamente el problema, al considerar las dependencias más importantes y despreciar las irrelevantes.

2.- El modelo de síntomas relevantes no es sólo un modelo teórico, sino que puede implementarse prácticamente sin requerir medios extraordinarios de recursos informáticos (memoria, tiempo de cpu, etc.), tal como lo demuestra la concha implementada como parte de esta tesis.

3.- El método de los factores de certeza exige la existencia de una probabilidad y, por tanto, no la evita. Además, el concepto de factor de certeza no es tan intuitivo como el de probabilidad y presenta una falta de biunivocidad en su definición.

4.- Los métodos de propagación de incertidumbre utilizados en teoría de la evidencia, lógica difusa y en los modelos de factores de certeza son inadecuados para reproducir muchas situaciones de la realidad y pueden conducir a conclusiones erróneas. La elección que se hace en estos métodos de las fórmulas de propagación son tan arbitrarias y criticables como la hipótesis de independencia en los métodos probabilísticos.

5.- La representación de reglas mediante conjuntos pseudobásicos es más compacta que la correspondiente mediante conjuntos básicos. El número mínimo de conjuntos pseudobásicos que es necesario unir para obtener cualquier conjunto que pueda expresarse como unión, intersección y/o complementación de r conjuntos es 2^{r-1}.

6.- La compilación de reglas reduce enormemente el tiempo de respuesta de los motores de inferencia y facilita notablemente el estudio de la incoherencia entre hechos y reglas. En la tesis se da un algoritmo que permite compilar reglas mediante la agrupación de todas las reglas que concluyen un objeto o su complementario.

7.- No existen sistemas de reglas incoherentes, sino hechos que contradicen un sistema de reglas dado. El método descrito en esta tesis permite detectar estas incoherencias.

8.- El experto humano no está generalmente capacitado para suministrar un conocimiento de tipo probabilístico coherente. Es prácticamente imposible suministrar este conocimiento en casos de cierta complejidad, sin la ayuda de una herramienta adecuada.

9.- La aportación ordenada de unidades de información en forma de probabilidades facilita notablemente el control de la coherencia, como lo prueba el método de tipo conjuntista descrito en la tesis.

10.- El método descrito en este trabajo y basado en la programación lineal permite controlar también la incoherencia en la base del conocimiento en términos de probabilidades. Este método puede ser utilizado para asistir al experto humano al aportar su conocimiento.

11.- Una base de conocimiento basada en intervalos de probabilidad es notablemente superior a una basada en valores puntuales, si bien es mucho más compleja. Esta tesis da modelos que permiten desarrollar un verdadero motor de inferencia basado en el método de la programación lineal.

12.- La resolución de problemas de programación lineal con parámetros en los términos independientes de las desigualdades aparece como consecuencia de acotar las probabilidades del modelo anterior. Este problema puede resolverse mediante métodos de programación simbólica análogos a los utilizados por paquetes conocidos, como REDUCE, MACSYMA, MAPLE, etc., en otros problemas matemáticos, tal como se demuestra en esta tesis.

13- Muchas técnicas estadísticas, tales como los modelos logarítmico-lineales, los de regresión, análisis multivariante, analisis cluster, etc. son aplicables a los sistemas expertos. En particular permiten desarrollar bases de conocimiento, subsistemas de aprendizaje, subsistemas de explicación y motores de inferencia, como ha quedado indicado en los capítulos anteriores. Estas técnicas son buenas alternativas, especialmente para trabajar con síntomas u objetos de tipo continuo.

14.- Es posible el aprendizaje en modelos de tipo probabilístico. En particular, el modelo de aprendizaje en sistemas de tipo probabilístico descrito en los capítulos 2 y 5, se ha mostrado capaz de tratar satisfactoriamente no sólo el caso de información total, sino también el caso de información parcial.

15.- Los modelos de redes causales pueden ser fácilmente dotados de un subsistema de aprendizaje. El método de aprendizaje global, basado en el método de la máxima verosimilitud, que se ha descrito en el capítulo 5, es equivalente a uno local. Por tanto, no sólo éste se simplifica notablemente sino que puede realizarse mediante computación en paralelo.

16.- La teoría de aprendizaje de conceptos de tipo determinista es generalizable al caso de conceptos de tipo probabilístico, tal como se ha demostrado en este trabajo. La orden de complejidad del aprendizaje no varía; sin embargo el incremento de recursos necesarios es muy notable.

17.- La complejidad del aprendizaje está ligada a los cardinales de los conjuntos Ω_0 y Ω_1 descritos en esta tesis. El aprendizaje en el conjunto Ω_1 implica una complejidad mayor que la del conjunto Ω_0, pues en el primer caso esta depende de la precisión y el nivel, mientras que en el segundo sólo de sus logaritmos.