REPRESENTACION DEL CONOCIMIENTO Y RAZONAMIENTO AUTOMATICO

VICENTE MORET BONILLO

Profesor Titular de Universidad. Senior Member, IEEE.

Departamento de Computación. Facultad de Informática.

UNIVERSIDAD DE A CORUÑA

2014

Lógica proposicional

La **lógica proposicional** o **lógica de orden cero** es un sistema formal cuyos elementos más simples representan proposiciones, y cuyas constantes lógicas, llamadas *conectivas*, representan operaciones sobre proposiciones, capaces de formar otras proposiciones de mayor complejidad.

La lógica proposicional trata con sistemas lógicos que carecen de cuantificadores, o variables interpretables como entidades. En lógica proposicional si bien no hay signos para variables de tipo entidad, sí existen signos para variables proposicionales (es decir, que pueden ser interpretadas como proposiciones con un valor de verdad de definido), de ahí el nombre proposicional. La lógica proposicional incluye además de variables interpretables como proposiciones simples signos para conectivas lógicas, por lo que dentro de este tipo de lógica puede analizarse la inferencia lógica de proposiciones a partir de proposiciones, pero sin tener en cuenta la estructura interna de las proposiciones más simples.

Introducción

Considérese el siguiente argumento:

1. Mañana es miércoles o mañana es jueves.
2. Mañana no es jueves.
3. Por lo tanto, mañana es miércoles.

Es un argumento válido. Quiere decir que es imposible que las premisas sean verdaderas y la conclusión falsa. Esto no quiere decir que la conclusión sea verdadera. Si las premisas son falsas, entonces la conclusión también podría serlo. Pero si las premisas son verdaderas, entonces la conclusión también lo es. La validez de este argumento no se debe al significado de las expresiones «mañana es miércoles» y «mañana es jueves», porque éstas podrían cambiarse por otras y el argumento permanecer válido. Por ejemplo:

1. Está soleado o está nublado.
2. No está nublado.
3. Por lo tanto, está soleado.

En cambio, la validez de estos dos argumentos depende del significado de las expresiones «o» y «no». Si alguna de estas expresiones se cambiara por otra, entonces podría ser que los argumentos dejaran de ser válidos. Por ejemplo:

1. Ni está soleado ni está nublado.
2. No está nublado.
3. Por lo tanto, está soleado.

Las expresiones de las que depende la validez de los argumentos se llaman constantes lógicas. La lógica proposicional estudia el comportamiento de algunas de estas expresiones, llamadas conectivas lógicas. En cuanto a las expresiones como "está nublado" o "mañana es jueves", lo único que importa de ellas es que tengan un valor de verdad. Es por esto que se las reemplaza por simples letras, cuya intención es simbolizar una expresión con valor de verdad cualquiera. A estas letras se las llama *variables proposicionales*, y en general se toman del alfabeto latino, empezando por la letra *p*, luego *q*, *r*, *s*, etc. Así, los dos primeros argumentos de esta sección podrían reescribirse así:

1. *p* o *q*
2. No *q*
3. Por lo tanto, *p*

Y el tercer argumento, a pesar de no ser válido, puede reescribirse así:

1. Ni *p* ni *q*
2. No *q*
3. Por lo tanto, *p*

Conectivas lógicas

A continuación hay una tabla que despliega todas las conectivas lógicas que ocupan a la lógica proposicional, incluyendo ejemplos de su uso en el lenguaje natural y los símbolos que se utilizan para representarlas en lenguaje formal.

Conectiva	Expresión en el lenguaje natural	Ejemplo	Símbolo en este artículo	Símbolos alternativos
Negación	no	**No** está lloviendo.	$\neg$	$\sim$
Conjunción	y	Está lloviendo **y** está nublado.	$\wedge$	$\&$ $\cdot$
Disyunción	o	Está lloviendo **o** está soleado.	$\vee$	
Condicional material	si... entonces	**Si** está soleado, **entonces** es de día.	$\rightarrow$	$\supset$
Bicondicional	si y sólo si	Está nublado **si y sólo si** hay nubes visibles.	$\leftrightarrow$	$\equiv$
Negación conjunta	ni... ni	**Ni** está soleado **ni** está nublado.	$\downarrow$	
Disyunción excluyente	o bien... o bien	**O bien** está soleado, **o bien** está nublado.	$\nleftrightarrow$	$\oplus, \not\equiv, W$

En la lógica proposicional, las conectivas lógicas se tratan como funciones de verdad. Es decir, como funciones que toman conjuntos de valores de verdad y devuelven valores de verdad. Por ejemplo, la conectiva lógica «no» es una función que si toma el valor de verdad V, devuelve F, y si toma el valor de verdad F, devuelve V. Por lo tanto, si se aplica la función «no» a una letra que represente una proposición falsa, el resultado será algo verdadero. Si es falso que «está lloviendo», entonces será verdadero que «no está lloviendo».

El significado de las conectivas lógicas no es nada más que su comportamiento como funciones de verdad. Cada conectiva lógica se distingue de las otras por los valores de verdad que devuelve frente a las distintas combinaciones de valores de verdad que puede recibir. Esto quiere decir que el significado de cada conectiva lógica puede ilustrarse mediante una tabla que despliegue los valores de verdad que la función devuelve frente a todas las combinaciones posibles de valores de verdad que puede recibir.

Negación

ϕ	$\neg\phi$
V	F
F	V

Conjunción

ϕ	ψ	$\phi \wedge \psi$
V	V	V
V	F	F
F	V	F
F	F	F

Disyunción

ϕ	ψ	$\phi \vee \psi$
V	V	V
V	F	V
F	V	V
F	F	F

Condicional

ϕ	ψ	$\phi \rightarrow \psi$
V	V	V
V	F	F
F	V	V
F	F	V

Bicondicional

ϕ	ψ	$\phi \leftrightarrow \psi$
V	V	V
V	F	F
F	V	F
F	F	V

Leyes notables en lógica

Entre las reglas de la lógica proposiconal clásica algunas de la más notables son listadas a continuación:

1. Ley de doble negación
2. Leyes de idempotencia
3. Leyes asociativas
4. Leyes conmutativas
5. Leyes distributivas
6. Leyes de De Morgan

Otras leyes como el principio del tercero excluido son admisibles en lógica clásica, pero en lógica intuicionista y con fines a sus aplicaciones matemáticas no existe un equivalente del tercero excluido, por ejemplo.

Límites de la lógica proposicional

La maquinaria de la lógica proposicional permite formalizar y teorizar sobre la validez de una gran cantidad de argumentos. Sin embargo, también existen argumentos que son intuitivamente válidos, pero cuya validez no puede ser probada por la lógica proposicional. Por ejemplo, considérese el siguiente argumento:

1. Todos los hombres son mortales.
2. Sócrates es un hombre.
3. Por lo tanto, Sócrates es mortal.

Como este argumento no contiene ninguna de las conectivas «no», «y», «o», etc., según la lógica proposicional, su formalización será la siguiente:

1. *p*
2. *q*
3. Por lo tanto, *r*

Pero esta es una forma de argumento inválida, y eso contradice nuestra intuición de que el argumento es válido. Para teorizar sobre la validez de este tipo de argumentos, se necesita investigar la estructura interna de las variables proposicionales. De esto se ocupa la lógica de primer orden. Otros sistemas formales permiten teorizar sobre otros tipos de argumentos. Por ejemplo la lógica de segundo orden, la lógica modal y la lógica temporal.

Dos sistemas formales de lógica proposicional

A continuación se presentan dos sistemas formales estándar para la lógica proposicional. El primero es un sistema axiomático simple, y el segundo es un sistema sin axiomas, de deducción natural.

Sistema axiomático

Alfabeto

El alfabeto de un sistema formal es el conjunto de símbolos que pertenecen al lenguaje del sistema. Si L es el nombre de este sistema axiomático de lógica proposicional, entonces el alfabeto de L consiste en:

- Una cantidad finita pero arbitrariamente grande de variables proposicionales. En general se las toma del alfabeto latino, empezando por la letra *p*, luego *q*, *r*, etc., y utilizando subíndices cuando es necesario o conveniente. Las variables proposicionales representan proposiciones como "está lloviendo" o "los metales se expanden con el calor".
- Un conjunto de operadores lógicos: $\neg, \wedge, \vee, \rightarrow, \leftrightarrow$
- Dos signos de puntuación: los paréntesis izquierdo y derecho. Su única función es desambiguar ciertas expresiones ambiguas, en exactamente el mismo sentido en que desambiguan la expresión 2 + 2 ÷ 2, que puede significar tanto (2 + 2) ÷ 2, como 2 + (2 ÷ 2).

Gramática

Una vez definido el alfabeto, el siguiente paso es determinar qué combinaciones de símbolos pertenecen al lenguaje del sistema. Esto se logra mediante una gramática formal. La misma consiste en un conjunto de reglas que definen recursivamente las cadenas de caracteres que pertenecen al lenguaje. A las cadenas de caracteres construidas según estas reglas se las llama fórmulas bien formadas. Las reglas del sistema L son:

1. Las variables proposicionales del alfabeto de L son fórmulas bien formadas.
2. Si ϕ es una fórmula bien formada de L, entonces $\neg\phi$ también lo es.
3. Si ϕ y ψ son fórmulas bien formadas de L, entonces $(\phi \wedge \psi)$, $(\phi \vee \psi)$, $(\phi \rightarrow \psi)$ y $(\phi \leftrightarrow \psi)$ también lo son.

4. Sólo las expresiones que pueden ser generadas mediante las cláusulas 1 a 3 en un número finito de pasos son fórmulas bien formadas de L.

Según estas reglas, las siguientes cadenas de caracteres son ejemplos de fórmulas bien formadas:

$$p$$
$$\neg\neg\neg q$$
$$(p \wedge q)$$
$$\neg(p \wedge q)$$
$$(p \leftrightarrow \neg p)$$
$$((p \rightarrow q) \wedge p)$$
$$(\neg(p \wedge (q \vee r)) \vee s)$$

Y los siguientes son ejempos de fórmulas mal formadas[*cita requerida*]:

Fórmula	Error	Corrección
(p)	Sobran paréntesis	p
$\neg(p)$	Sobran paréntesis	$\neg p$
$(\neg p)$	Sobran paréntesis	$\neg p$
$p \rightarrow q$	Faltan paréntesis	$(p \rightarrow q)$
$(p \wedge q \rightarrow r)$	Faltan paréntesis	$((p \wedge q) \rightarrow r)$

Por otra parte, dado que la única función de los paréntesis es desambiguar las fórmulas, en general se acostumbra omitir los paréntesis *externos* de cada fórmula, ya que estos no cumplen ninguna función. Así por ejemplo, las siguientes fórmulas generalmente se consideran bien formadas:

$$p \wedge q$$
$$\neg p \rightarrow q$$
$$(p \wedge q) \vee \neg q$$
$$(p \leftrightarrow q) \leftrightarrow (q \leftrightarrow p)$$

Otra convención acerca del uso de los paréntesis es que las conjunciones y las disyunciones tienen «menor jerarquía» que los condicionales materiales y los bicondicionales. Esto significa que dada una fórmula sin paréntesis, las conjunciones y las disyunciones deben agruparse *antes* que los condicionales materiales y los bicondicionales. Por ejemplo:

Fórmula	Lectura correcta	Lectura incorrecta
$p \wedge q \rightarrow r$	$(p \wedge q) \rightarrow r$	$p \wedge (q \rightarrow r)$
$\neg p \leftrightarrow q \vee r$	$\neg p \leftrightarrow (q \vee r)$	$(\neg p \leftrightarrow q) \vee r$
$p \wedge q \leftrightarrow r \vee s$	$(p \wedge q) \leftrightarrow (r \vee s)$	$(p \wedge (q \leftrightarrow r)) \vee s$

Estas convenciones son análogas a las que existen en el álgebra elemental, donde la multiplicación y la división siempre deben resolverse antes que la suma y la resta. Así por ejemplo, la ecuación $2 + 2 \times 2$ podría interpretarse como $(2 + 2) \times 2$ o como $2 + (2 \times 2)$. En el primer caso el resultado sería 8, y en el segundo caso sería 6. Pero como la multiplicación siempre debe resolverse antes que la suma, el resultado correcto en este caso es 6, no 8.

Axiomas

Los axiomas de un sistema formal son un conjunto de fórmulas bien formadas que se toman como punto de partida para demostraciones ulteriores. Un conjunto de axiomas estándar es el que descubrió Jan Łukasiewicz:

- $(\phi \to (\psi \to \phi))$
- $((\phi \to (\psi \to \chi)) \to ((\phi \to \psi) \to (\phi \to \chi)))$
- $((\neg\phi \to \neg\psi) \to (\psi \to \phi))$

Reglas de inferencia

Una regla de inferencia es una función que va de conjuntos de fórmulas a fórmulas. Al conjunto de fórmulas que la función toma como argumento se lo llama *premisas*, mientras que a la fórmula que devuelve como valor se la llama *conclusión*. En general se busca que las reglas de inferencia transmitan la verdad de las premisas a la conclusión. Es decir, que sea imposible que las premisas sean verdaderas y la conclusión falsa. En el caso de L, la única regla de inferencia es el modus ponens, el cual dice:

$$(\phi \to \psi), \phi \vdash \psi$$

Recordando que ϕ y ψ no son fórmulas, sino metavariables que pueden ser reemplazadas por cualquier fórmula bien formada.

Ejemplo de una demostración

A demostrar: $\phi \to \phi$		
Paso	**Fórmula**	**Razón**
1	$\phi \to (\phi \to \phi)$	Instancia del primer axioma.
2	$\phi \to ((\phi \to \phi) \to \phi)$	Instancia del primer axioma.
3	$\Big(\phi \to ((\phi \to \phi) \to \phi)\Big) \to \Big((\phi \to (\phi \to \phi)) \to (\phi \to \phi)\Big)$	Instancia del segundo axioma.
4	$\Big((\phi \to (\phi \to \phi)) \to (\phi \to \phi)\Big)$	Desde (2) y (3) por modus ponens.
5	$\phi \to \phi$	Desde (1) y (4) por modus ponens. QED

Deducción natural

Un sistema de lógica proposicional también puede construirse a partir de un conjunto vacío de axiomas. Para ello se especifican una serie de reglas de inferencia que intentan capturar el modo en que naturalmente razonamos acerca de las conectivas lógicas.

- Introducción de la negación:

  Si suponer ϕ lleva a una contradicción, entonces se puede inferir que $\neg\phi$ (reducción al absurdo).

- Eliminación de la negación:

  $$\neg\neg\phi \vdash \phi$$

- Introducción de la conjunción:

  $$\phi, \psi \vdash (\phi \wedge \psi)$$
  $$\phi, \psi \vdash (\psi \wedge \phi)$$

- Eliminación de la conjunción:

  $$(\phi \wedge \psi) \vdash \phi$$
  $$(\phi \wedge \psi) \vdash \psi$$

- Introducción de la disyunción:

$$\phi \vdash (\phi \vee \psi)$$
$$\phi \vdash (\psi \vee \phi)$$

- Eliminación de la disyunción:

$$(\phi \vee \psi), (\phi \rightarrow \chi), (\psi \rightarrow \chi) \vdash \chi$$

- Introducción del condicional (véase el teorema de la deducción):

  Si suponer ϕ lleva a una prueba de ψ , entonces se puede inferir que $(\phi \rightarrow \psi)$.

- Eliminación del condicional (modus ponens):

$$(\phi \rightarrow \psi), \phi \vdash \psi$$

- Introducción del bicondicional:

$$(\phi \rightarrow \psi), (\psi \rightarrow \phi) \vdash (\phi \leftrightarrow \psi)$$
$$(\phi \rightarrow \psi), (\psi \rightarrow \phi) \vdash (\psi \leftrightarrow \phi)$$

- Eliminación del bicondicional:

$$(\phi \leftrightarrow \psi) \vdash (\phi \rightarrow \psi)$$
$$(\phi \leftrightarrow \psi) \vdash (\psi \rightarrow \phi)$$

Ejemplo de una demostración

A demostrar: $\phi \rightarrow \phi$		
Paso	**Fórmula**	**Razón**
1	ϕ	Supuesto.
2	$\phi \vee \phi$	Desde (1) por introducción de la disyunción.
3	$(\phi \vee \phi) \wedge \phi$	Desde (1) y (2) por introducción de la conjunción.
4	ϕ	Desde (3) por eliminación de la conjunción.
5	$\phi \vdash \phi$	Resumen de (1) hasta (4).
6	$\vdash \phi \rightarrow \phi$	Desde (5) por introducción del condicional. QED

Lenguaje formal en la notación BNF

El lenguaje formal de la lógica proposicional se puede generar con la gramática formal descrita usando la notación BNF como sigue:

$$\begin{aligned}
\langle Bicondicional \rangle &::= \langle Condicional \rangle \leftrightarrow \langle Bicondicional \rangle \mid \langle Condicional \rangle \\
\langle Condicional \rangle &::= \langle Conjuncion \rangle \leftrightarrow \langle Condicional \rangle \mid \langle Conjuncion \rangle \\
\langle Conjuncion \rangle &::= \langle Disyuncion \rangle \vee \langle Conjuncion \rangle \mid \langle Disyuncion \rangle \\
\langle Disyuncion \rangle &::= \langle Literal \rangle \wedge \langle Disyuncion \rangle \mid \langle Literal \rangle \\
\langle Literal \rangle &::= \langle Atomo \rangle \mid \neg \langle Atomo \rangle \\
\langle Atomo \rangle &::= \top \mid \bot \mid \langle Letra \rangle \mid \langle Agrupacion \rangle \\
\langle Agrupacion \rangle &::= (\langle Bicondicional \rangle) \mid [\langle Bicondicional \rangle] \mid \{\langle Bicondicional \rangle\}
\end{aligned}$$

La gramática anterior define la precedencia de operadores de la siguiente manera:

1. Negación ($\neg$)
2. Conjunción ($\wedge$)
3. Disyunción ($\vee$)
4. Condicional material ($\rightarrow$)
5. Bicondicional ($\leftrightarrow$)

Semántica

Una interpretación para un sistema de lógica proposicional es una asignación de valores de verdad para cada variable proposicional, sumada a la asignación usual de significados para los operadores lógicos. A cada variable proposicional se le asigna uno de dos posibles valores de verdad: o V (verdadero) o F (falso). Esto quiere decir que si hay *n* variables proposicionales en el sistema, el número de interpretaciones distintas es de 2^n.

Partiendo de esto es posible definir una cantidad de nociones semánticas. Si A y B son fórmulas cualquiera de un lenguaje L, Γ es un conjunto de fórmulas de L, y M es una interpretación de L, entonces:

- A es verdadera bajo la interpretación M si y sólo si M asigna el valor de verdad V a A.
- A es falsa bajo la interpretación M si y sólo si M asigna el valor de verdad F a A.
- A es una tautología (o una verdad lógica) si y sólo si para toda interpretación M, M asigna el valor de verdad V a A.
- A es una contradicción si y sólo si para toda interpretación M, M asigna el valor de verdad F a A.
- A es satisfacible (o consistente) si y sólo si existe al menos una interpretación M que asigne el valor de verdad V a A.
- Γ es consistente si y sólo si existe al menos una interpretación que haga verdaderas a todas las fórmulas en Γ.
- A es una consecuencia semántica de un conjunto de fórmulas Γ si y sólo si para toda fórmula B que pertenezca a Γ, no hay ninguna interpretación en que B sea verdadera y A falsa. Cuando A es una consecuencia semántica de Γ en un lenguaje L, se escribe: $\Gamma \models_L A$.
- A es una verdad lógica si y sólo si A es una consecuencia semántica del conjunto vacío. Cuando A es una verdad lógica de un lenguaje L, se escribe: $\models_L A$.

Tablas de verdad

La tabla de verdad de una fórmula es una tabla en la que se presentan todas las posibles interpretaciones de las variables proposicionales que constituye la fórmula y el valor de verdad de la fórmula completa para cada interpretación. Por ejemplo, la tabla de verdad para la fórmula $\neg(p \vee q) \rightarrow (p \rightarrow r)$ sería:

p	q	r	$(p \vee q)$	$\neg(p \vee q)$	$(p \rightarrow r)$	$\neg(p \vee q) \rightarrow (p \rightarrow r)$
V	V	V	V	F	V	V
V	V	F	V	F	F	V
V	F	V	V	F	V	V
V	F	F	V	F	F	V
F	V	V	V	F	V	V
F	V	F	V	F	V	V
F	F	V	F	V	V	V
F	F	F	F	V	V	V

Como se ve, esta fórmula tiene 2^n interpretaciones posibles —una por cada línea de la tabla—, donde *n* es el número de variables proposicionales (en este caso 3, es decir p, q, r) , y resulta ser una tautología, es decir que bajo todas las interpretaciones posibles de las variables proposicionales, el valor de verdad de la fórmula completa termina siendo V.

Formas normales

A menudo es necesario transformar una fórmula en otra, sobre todo transformar una fórmula a su forma normal. Esto se consigue transformando la fórmula en otra equivalente y repitiendo el proceso hasta conseguir una fórmula que sólo use los conectivos básicos ($\wedge, \vee, \neg$). Para lograr esto se utilizan las equivalencias lógicas:

$$(p \rightarrow q) \leftrightarrow (\neg p \vee q)$$
$$(p \leftrightarrow q) \leftrightarrow [(\neg p \vee q) \wedge (\neg q \vee p)]$$

Por ejemplo, considérese la siguiente fórmula:

$$(p \rightarrow q) \wedge (\neg q \leftrightarrow p)$$

La misma puede desarrollarse así:

$$(\neg p \vee q) \wedge (q \vee p) \wedge (\neg p \vee \neg q)$$

Se dice que una fórmula está en *forma normal disyuntiva* (FND) si y sólo si tiene la siguiente forma:

$$A_1 \vee A_2 \vee ... \vee A_n$$

donde cada A es una conjunción de fórmulas. Por ejemplo, la siguiente fórmula está en forma normal disyuntiva:

$$p \vee (q \wedge s) \vee (\neg q \wedge p)$$

Se dice que una fórmula está en *forma normal conjuntiva* (FNC) si y sólo si tiene la siguiente forma:

$$A_1 \wedge A_2 \wedge ... \wedge A_n$$

donde cada A es una disjunción de fórmulas. Por ejemplo, la siguiente fórmula está en forma normal conjuntiva:

$$p \wedge (q \vee s) \wedge (\neg q \vee p)$$

Por las leyes de De Morgan, es posible pasar de una forma normal disyuntiva a una forma normal conjuntiva y viceversa:

$$\neg(A \vee B) \leftrightarrow (\neg A \wedge \neg B)$$
$$\neg(A \wedge B) \leftrightarrow (\neg A \vee \neg B)$$

Las FNC y FND son mutuamente duales. La demostración hace uso de las leyes de De Morgan y de la propiedad distributiva de la conjunción y la disyunción. Se debe cumplir que:

$$\neg[(A_1 \wedge B_1) \vee (A_2 \wedge B_2) \vee ... \vee (A_n \wedge B_n)] \leftrightarrow [(\neg A_1 \vee \neg B_1) \wedge (\neg A_2 \vee \neg B_2) \wedge ... \wedge (\neg A_n \vee \neg B_n)]$$

Y viceversa:

$$\neg[(A_1 \vee B_1) \wedge (A_2 \vee B_2) \wedge ... \wedge (A_n \vee B_n)] \leftrightarrow [(\neg A_1 \wedge \neg B_1) \vee (\neg A_2 \wedge \neg B_2) \vee ... \vee (\neg A_n \wedge \neg B_n)]$$

La lógica proposicional y la computación

Debido a que los computadores trabajan con información binaria, la herramienta matemática adecuada para el análisis y diseño de su funcionamiento es el Álgebra de Boole. El Álgebra de Boole fue desarrollada inicialmente para el estudio de la lógica. Ha sido a partir de 1938, fecha en que Claude Shannon publicó un libro llamado "Análisis simbólico de circuitos con relés", estableciendo los primeros conceptos de la actual teoría de la conmutación, cuando se ha producido un aumento considerable en el número de trabajos de aplicación del Álgebra de Boole a los computadores digitales. Hoy en día, esta herramienta resulta fundamental para el desarrollo de los computadores ya que, con su ayuda, el análisis y síntesis de combinaciones complejas de circuitos lógicos puede realizarse con rapidez.

Aristóteles con respecto al estudio de la lógica

La lógica es conocida como una de las ciencias más antiguas, tanto es así que se le atribuye a Aristóteles la paternidad de esta disciplina. Partiendo de que corresponde a Aristóteles haber sido el primero en tratar con todo detalle la lógica, se le considera pues ser su fundador. En un principio se llamó Analítica, en virtud del título de las obras en que trató los problemas lógicos. Más tarde los escritos de Aristóteles relativos a estos eventos fueron recopilados por sus discípulos con el título de Organon, por considerar que la lógica era un instrumento para el conocimiento de la verdad.

Aristóteles se planteó cómo es posible probar y demostrar que un conocimiento es verdadero, es decir, que tiene una validez universal. Aristóteles encuentra el fundamento de la demostración en la deducción, procedimiento que consiste en derivar un hecho particular de algo universal. La forma en que se afecta esa derivación es el silogismo, por cuya razón la silogística llega a ser el centro de la lógica aristotélica.

Notas y referencias

Bibliografía

- Enderton, H. B. (1972). *A Mathematical Introduction to Logic*. Academic Press.
- Hamilton, A. G. (1981). *Lógica para matemáticos*. Paraningo.
- Mendelson, E. (1997). *Introduction to Mathematical Logic* (4ª edición). Chapman and May.
- Pla, J. (1991). *Lliçons de lógica matemática*. P.P.U..
- Badesa, C.; Jané, I.; Jansana, R. (1998). *Elementos de lógica formal*. Ariel.
- Barnes, D. W.; Mack, J. M. (1978). *Una introducción algebraica a la lógica matemática*. Eunibar.
- Bridge, J. (1977). *Beginning Model Theory*. Oxford University Pres.
- Ershov, Y.; Paliutin, E. (1990). *Lógica matemática*. Mir.
- Hofstadter, D. (1987). *Gödel, Escher, Bach: un Eterno y Grácil Bucle*. Tusquets Editores.
- Jané, I. (1989). *Álgebras de Boole y lógica*. Publicaciones U.B..
- Monk, J. D. (1976). *Mathematical Logic*. Springer-Verlag.
- Nidditch, P. H. (1978). *El desarrollo de la lógica matemática*. Cátedra.
- Van Dalen, D. (1983). *Logic and Structure* (2ª edición). Universitext, Springer-Verlag.

Enlaces externos

- Introducción a la lógica proposicional (http://portales.educared.net/wikiEducared/index.php?title=LÃ³gica_proposicional)

Lógica de primer orden

La **lógica de primer orden**, también llamada **lógica de predicados** o **cálculo de predicados**, es un sistema formal diseñado para estudiar la inferencia en los lenguajes de primer orden. Los lenguajes de primer orden son, a su vez, lenguajes formales con cuantificadores que alcanzan sólo a variables de individuo, y con predicados y funciones cuyos argumentos son sólo constantes o variables de individuo.

La lógica de primer orden tiene el poder expresivo suficiente para definir a prácticamente todas las matemáticas.

Introducción

Como el desarrollo histórico y las aplicaciones de la lógica de primer orden están muy ligados a la matemática, en lo que sigue se hará una introducción que contemple e ilustre esta relación, tomando ejemplos tanto de la matemática como del lenguaje natural. Primero se introducen cada uno de los conceptos básicos del sistema, y luego se muestra cómo utilizarlos para analizar argumentos.

Predicados

Un predicado es una expresión lingüística que puede conectarse con una o varias otras expresiones para formar una oración. Por ejemplo, en la oración «Marte es un planeta», la expresión «es un planeta» es un predicado que se conecta con la expresión «Marte» para formar una oración. Y en la oración «Júpiter es más grande que Marte», la expresión «es más grande que» es un predicado que se conecta con dos expresiones, «Júpiter» y «Marte», para formar una oración.

En lógica matemática, cuando un predicado se conecta con *una* expresión, se dice que expresa una *propiedad* (como la propiedad de *ser un planeta*), y cuando se conecta con dos o más expresiones, se dice que expresa una *relación* (como la relación de *ser más grande que*). La lógica de primer orden no hace ningún supuesto, sin embargo, sobre si existen o no las propiedades o las relaciones. Sólo se ocupa de estudiar el modo en que hablamos y razonamos con expresiones lingúisticas.

En la lógica de primer orden, los predicados son tratados como funciones. Una función es, metafóricamente hablando, una máquina que recibe un conjunto de cosas, las procesa, y devuelve como resultado una *única* cosa. A las cosas que entran a las funciones se las llama *argumentos*,[1] y a las cosas que salen, *valores* o *imágenes*. Considérese por ejemplo la siguiente función matemática:

$f(x) = 2x$

Esta función toma números como argumentos y devuelve más números como valores. Por ejemplo, si toma el número 1, devuelve el número 2, y si toma el 5, devuelve el 10. En la lógica de primer orden, se propone tratar a los predicados como funciones que no sólo toman números como argumentos, sino expresiones como «Marte», «Mercurio» y otras que se verán más adelante. De este modo, la oración «Marte es un planeta» puede transcribirse, siguiendo la notación propia de las funciones, de la siguiente manera:

$Planeta(Marte)$

O, más abreviadamente:

$P(m)$

En la matemática existen además funciones que toman varios argumentos. Por ejemplo:

$f(x,y) = x + y$

Esta función, si toma los números 1 y 2, devuelve el número 3, y si toma el -5 y el -3, devuelve el -8. Siguiendo esta idea, la lógica de primer orden trata a los predicados que expresan relaciones, como funciones que toman *dos* o más argumentos. Por ejemplo, la oración «Caín mató a Abel» puede formalizarse así:

$Mató(Caín,Abel)$

O abreviando:

$M(c,a)$

Este procedimiento puede extenderse para tratar con predicados que expresan relaciones entre muchas entidades. Por ejemplo, la oración «Ana está sentada entre Bruno y Carlos» puede formalizarse:

$S(a,b,c)$

Constantes de individuo

Una constante de individuo es una expresión lingüística que refiere a una entidad. Por ejemplo «Marte», «Júpiter», «Caín» y «Abel» son constantes de individuo. También lo son las expresiones «1», «2», etc., que refieren a números. Una entidad no tiene que existir para que se pueda hablar acerca de ella, de modo que la lógica de primer orden tampoco hace supuestos acerca de la existencia o no de las entidades a las que refieren sus constantes de individuo.

Variables de individuo

Además de las constantes de individuo que hacen referencia a entidades determinadas, la lógica de primer orden cuenta con otras expresiones, las *variables*, cuya referencia no está determinada. Su función es similar a la de las expresiones del lenguaje natural como «él», «ella», «esto», «eso» y «aquello», cuyo referente varía con el contexto. Las variables generalmente se representan con letras minúsculas cerca del final del alfabeto latino, principalmente la x, y y z. Del mismo modo, en la matemática, la x en la función $f(x) = 2x$ no representa ningún número en particular, sino que es algo así como un espacio vacío donde pueden insertarse distintos números. En conclusión, podemos representar una expresión como «esto es antiguo» con la expresión:

$Antiguo(x)$

O abreviadamente:

$A(x)$

Es evidente, sin embargo, que hasta que no se determine a qué refiere la x, no es posible asignar un valor de verdad a la expresión «esto es antiguo», del mismo modo que hasta que no se determine un número para la x en la función $f(x) = 2x$, no será posible calcular ningún valor para la función.

Por supuesto, al igual que con las constantes de individuo, las variables sirven también para formalizar relaciones. Por ejemplo, la oración «esto es más grande que aquello» se formaliza:

$G(x,y)$

Y también pueden combinarse constantes de individuo con variables. Por ejemplo en la oración «ella está sentada entre Bruno y Carlos»:

$S(x,b,c)$

Cuantificadores

Considérese ahora la siguiente expresión matemática:

$x > 3$

Esta expresión no es ni verdadera ni falsa, y parece que no lo será hasta que no reemplacemos a la x por algún número cualquiera. Sin embargo, también es posible dar un valor de verdad a la expresión si se le antepone un cuantificador. Un cuantificador es una expresión que afirma que una condición se cumple para un cierto número de individuos. En la lógica clásica, los dos cuantificadores más estudiados son el cuantificador universal y el cuantificador existencial. El primero afirma que una condición se cumple para *todos* los individuos de los que se está hablando, y el segundo que se cumple para *al menos uno* de los individuos. Por ejemplo, la expresión "para todo x" es un cuantificador universal, que antepuesto a "$x < 3$", produce:

Para todo x, $x < 3$

Esta es una expresión con valor de verdad, en particular, una expresión falsa, pues existen muchos números (muchos *x*) que son *mayores* que tres. Anteponiendo en cambio la expresión "para al menos un *x*", un cuantificador existencial, se obtiene:

Para al menos un *x*, $x < 3$

La cual resulta ser una expresión verdadera.

Adviértase ahora, sin embargo, que el valor de verdad de las dos expresiones anteriores depende de qué números se esté hablando. Si cuando se afirma "para todo *x*, $x < 3$", se está hablando sólo de los números negativos, por ejemplo, entonces la afirmación es verdadera. Y si al afirmar "para al menos un *x*, $x < 3$" se está hablando solamente de los números 3, 4 y 5, entonces la afirmación es falsa. En lógica, a aquello de lo que se está hablando cuando se usa algún cuantificador, se lo llama el dominio de discurso.

Esta maquinaria puede adaptarse fácilmente para formalizar oraciones con cuantificadores del lenguaje natural. Tómese por caso la afirmación "todos son amigables". Esta oración puede traducirse así:

Para todo *x*, *x* es amigable.

Y una oración como "alguien está mintiendo" puede traducirse:

Para al menos un *x*, *x* está mintiendo.

También es frecuente traducir esta última oración así:

Existe al menos un *x*, tal que *x* está mintiendo.

A continuación se formalizan ambas oraciones, introduciendo a la vez la notación especial para los cuantificadores:

Para todo *x*, *x* es amigable.	$\forall x\, A(x)$
Existe al menos un *x*, tal que *x* está mintiendo.	$\exists x\, M(x)$

Conectivas

La lógica de primer orden incorpora además las conectivas de la lógica proposicional. Combinando las conectivas con los predicados, constantes, variables y cuantificadores, es posible formalizar oraciones como las siguientes:

Oración	**Formalización**
Sócrates es sabio y prudente.	$Ss \wedge Ps$
Si Sócrates es sabio, entonces también es prudente.	$Ss \rightarrow Ps$
Nadie es sabio y además prudente.	$\neg\exists x\, (Sx \wedge Px)$
Todos los sabios son prudentes.	$\forall x\, (Sx \rightarrow Px)$

Argumentos

Considérese el siguiente argumento clásico:

1. Todos los hombres son mortales.
2. Sócrates es un hombre.
3. Por lo tanto, Sócrates es mortal.

La tarea de la lógica de primer orden consiste en determinar por qué los argumentos como éste resultan válidos. Para eso, el primer paso es traducirlos a un lenguaje más preciso, que pueda ser analizado mediante métodos formales. Según lo visto más arriba, la formalización de este argumento es la siguiente:

1. $\forall x\, (Hx \rightarrow Mx)$
2. Hs
3. $\therefore Ms$

Sistema formal

A continuación se define un lenguaje formal, Q, y luego se definen axiomas y reglas de inferencia sobre ese lenguaje que dan como resultado el sistema lógico SQ.

Sintaxis

El alfabeto del lenguaje formal Q consta de los siguientes símbolos:

$a \quad x \quad f \quad P \quad * \quad ' \quad \neg \quad \wedge \quad \vee \quad \rightarrow \quad \leftrightarrow \quad \forall \quad \exists \quad (\quad)$

A partir de estos símbolos, se definen las siguientes nociones:

Un **nombre** (o **constante de individuo**) es una a seguida de una o más comillas. Por ejemplo, a', a'' y a'''''' son nombres. Para facilitar la lectura, se suelen omitir las comillas y utilizar distintas letras cerca del comienzo del alfabeto latino, con o sin subíndices, para distinguir nombres distintos: a, b, c, d, e, a_1, a_3, c_9, etc.

Una **variable** (o **variable de individuo**) es una x seguida de una o más comillas. Por ejemplo, x', x'' y x'''''' son variables. Para facilitar la lectura, se suelen omitir las comillas y utilizar distintas letras cerca del final del alfabeto latino, con o sin subíndices, para distinguir variables distintas: x, y, z, x_1, x_3, z_9, etc.

Un **functor** es una f seguida de uno o más asteriscos, y luego de una o más comillas. Por ejemplo, $f*'$, $f**''''$ y $f****''$ son functores. El número de asteriscos indica la aridad del functor. Para facilitar la lectura, se suelen omitir los asteriscos y las comillas y utilizar distintas letras del alfabeto latino cerca de la f, con o sin subíndices, para distinguir functores distintos: f, g, h, f_1, f_3, h_9, etc.

Un **predicado** es una P seguida de uno o más asteriscos, y luego de una o más comillas. Por ejemplo, $P*'$, $P**''''$ y $P****''$ son predicados. El número de asteriscos indica la aridad del predicado. Para facilitar la lectura, se suelen omitir los asteriscos y las comillas y utilizar distintas letras en mayúscula a lo largo del alfabeto latino para distinguir predicados distintos: P, A, B, C, S, T, etc.

La noción de **término** se define recursivamente mediante las siguientes cláusulas:

1. Todos los nombres son términos.
2. Todas las variables son términos.
3. Si f es un functor de aridad $n \geq 1$ y $t_1,...,t_n$ son términos, entonces $f(t_1,...,t_n)$ es un término.
4. Nada más es un término.

Según esta definición, las siguientes cadenas de caracteres son términos:

Cadena	**Simplificación**	**Posible interpretación**
a'	a	Aristóteles
x''''''	y	
$f*'''(a''')$	$h(c)$	El hermano de Caín
$f*''(f*''(f*''(a')))$	$f(f(f(b)))$	El padre del padre del padre de Beatriz

Y en cambio, las siguientes cadenas de caracteres no son términos:

Cadena	Error
a	Faltan comillas.
$x*''''$	Sobra el asterisco.
f'	Faltan asteriscos y argumentos.
$f**$	Faltan comillas y argumentos.
$f*'(f*')$	Falta el argumento del functor más anidado.
$f*'(a',a'')$	El functor es de aridad 1 pero tiene dos argumentos.

La noción de fórmula bien formada de Q se define a través de las siguientes cláusulas:

1. Si P es un predicado de aridad $n \geq 1$ y $t_1,...,t_n$ son términos, entonces $P(t_1,...,t_n)$ es una fórmula bien formada.
2. Si A es una fórmula bien formada, entonces ¬A también lo es.
3. Si A y B son fórmulas bien formadas, entonces (A ∧ B), (A ∨ B), (A → B) y (A ↔ B) también lo son.
4. Si A es una fórmula bien formada y x es una variable, entonces $\forall x$ A y $\exists x$ A son fórmulas bien formadas.
5. Nada más es una fórmula bien formada.

Según esta definición, las siguientes cadenas de caracteres son fórmulas bien formadas:

Cadena	Simplificación	Posible interpretación
$P*'(a')$	Pa	Abel es pastor.
$P**''''(a'',a''')$	Aae	Abelardo ama a Eloísa.
$\neg P*'(f*'(a'))$	$\neg P(h(a))$	El hermano de Abel no es pastor.
$(P*'''(a'') \rightarrow \neg P*''''''(a''))$	$Pv \rightarrow \neg Ev$	Si Venus es un planeta, entonces no es una estrella.
$\forall x''\ P*'''(x'')$	$\forall x\ Mx$	Todos son mentirosos.
$\forall x''\ \exists x''''\ P**'(x'',x'''')$	$\forall x\ \exists y\ Axy$	Todos aman a alguien.
$\exists x''\ \forall x''''\ P**'(x'',x'''')$	$\exists x\ \forall y\ Axy$	Alguien ama a todos.

Y en cambio, las siguientes cadenas de caracteres no son fórmulas bien formadas:

Cadena	Error
$P*'$	El predicado es de aridad 1 pero no tiene argumentos.
$P***'(a')$	El predicado es de aridad 3 pero tiene un sólo argumento.
$P*'(a') \rightarrow P*'(a''')$	Faltan los paréntesis externos.
$(P*'(a'))$	Sobran los paréntesis externos.
$\forall a'\ P*'(a')$	El cuantificador está seguido de un nombre en vez de una variable.

Para ciertos predicados muy utilizados, la notación estándar puede tener la forma $a\ R\ b$ en vez de $R(a,b)$. Por ejemplo, se escribe 2 > 1 en vez de >(2,1), y 4 = 4 en vez de =(4,4). Análogamente, si f es un functor de aridad 2, a veces se escribe $a\ f\ b$ en vez de $f(a,b)$. Por ejemplo, se escribe 1 + 2 en vez de +(1,2).

Observaciones

- El símbolo de identidad a veces se incluye entre los símbolos primitivos del alfabeto y se comporta sintácticamente como un predicado binario. A una lógica de primer orden que incluye el símbolo de identidad se la llama, justamente, *lógica de primer orden con identidad*.
- Los nombres pueden ser definidos como functores de aridad 0, de modo que es posible omtir a la *a* de entre los símbolos primitivos.
- En la definición anterior se requiere que los predicados tengan aridad mayor o igual que 1. Es posible permitir predicados de aridad 0, considerándolos como variables proposicionales de la lógica proposicional.
- Es posible reducir el número de símbolos primitivos hasta quedarse con sólo nueve: x f P $*$ $'$ $\downarrow$ $\forall$ $($ $)$
- Hay diferentes convenciones acerca de dónde poner los paréntesis. Por ejemplo, algunos escriben ($\forall x$) en vez de $\forall x$. A veces se usan dos puntos (:) o un punto (.) en vez de paréntesis para desambiguar fórmulas. Una notación interesante pero poco usual es la notación polaca, donde se omiten todos los paréntesis y se escribe $\wedge$, $\vee$, delante de los argumentos en vez de entre ellos. La notación polaca es compacta pero poco común por ser difícil para ser leída por los humanos.
- Una observación técnica es que si existe un símbolo de función de aridad 2 representando el par ordenado (o símbolo de predicado de aridad 2 representando la relación) no se necesitan funciones y predicados de aridad mayor que 2.
- Usualmente se considera que el conjunto de constantes, funciones y relaciones forman un *lenguaje*, mientras que las variables, los operadores lógicos y cuantificadores se los considera pertenecientes a la lógica. Por ejemplo, el lenguaje de la teoría de grupos consiste de una constante (el elemento identidad), una función de aridad 1 (la inversa), una función de aridad 2 (el producto), y una relación de aridad 2 (la igualdad), omitida por los autores que incluyen la igualdad en la lógica subyacente.

Substitución de variables libres

Las nociones de variable libre y variable ligada se introducen para evitar un posible error en el proceso de substitución. Supongamos por un momento la fórmula $\forall x(x \leq y)$. Intuitivamente, esta fórmula dice que para todo *x*, *x* es menor o igual que *y* (es decir, que *y* es máximo). En esta fórmula, *y* es una variable libre, o sea que no está bajo el alcance de ningún cuantificador. Si substituimos *y* por cualquier otro término *t*, entonces la fórmula pasará a decir que *t* es máximo. Pero supongamos ahora que substituimos a *y* por *x* mismo (a fin de cuentas, *x* es un término). En ese caso, *y* pasa a estar ligada por un cuantificador universal, porque la nueva fórmula es: $\forall x(x \leq x)$. Pero esta fórmula ya no dice de un término que es máximo, sino algo muy distinto. Para evitar este tipo de desplazamiento de significado, convenimos que al substituir una variable libre por un término cualquiera, hay que evitar que las variables libres en el nuevo término queden ligadas por algún cuantificador. Es decir, que permanezcan libres.
Dicho de una manera más general, si *t* es un término y $\phi(x)$es una fórmula que posiblemente contiene a *x* como una variable libre, entonces $\phi(t)$es el resultado de substituir todas las apariciones libres de *x* por *t*, *suponiendo que ninguna variable libre en t se vuelva ligada en este proceso*. Si alguna variable libre de *t* se volviera ligada, entonces para substituir *t* por *x* se necesita cambiar los nombres de las variables ligadas de $\phi(x)$por otros que no coincidan con las variables libres de *t*.

Identidad

Hay varias maneras diferentes de introducir la noción de identidad en la lógica de primer orden, pero todas con esencialmente las mismas consecuencias. Esta sección resume las principales:

- La manera más común de introducir a la identidad es incluyendo al símbolo entre los primitivos, y agregando axiomas que definan el comportamiento del mismo. Estos son:

$$\forall x(x = x)$$
$$\forall x \forall y \Big((x = y) \rightarrow \forall f \Big((f(...x...) = f(...y...)) \Big) \Big)$$
$$\forall x \forall y \Big((x = y) \rightarrow \forall P \Big((P(...x...) \leftrightarrow P(...y...)) \Big) \Big)$$

- Otra manera es incluir al símbolo de identidad como una de las relaciones de la teoría y agregar los axiomas de identidad a la teoría. En la práctica esta convención es casi indistinguible de la anterior, salvo en el caso inusual de las teorías sin noción de identidad. Los axiomas son los mismos. La única diferencia es que unos se llaman axiomas lógicos y los otros axiomas de la teoría.
- En las teorías sin funciones y con un número finito de relaciones, es posible definir la identidad en términos de las relaciones. Esto se hace definiendo que dos términos *a* y *b* son iguales si y sólo si ninguna relación presenta cambios reemplazando *a* por *b* en cualquier argumento. Por ejemplo, en teoría de conjuntos con una relación de pertenencia (∈), definiríamos $a = b$ como una abreviación para $\forall x\, [(a \in x) \leftrightarrow (b \in x)] \wedge [(x \in a) \leftrightarrow (x \in b)]$. Esta definición de identidad automáticamente satisface los axiomas de identidad.
- En algunas teorías es posible dar definiciones ad hoc para la identidad. Por ejemplo, en una teoría de órdenes parciales con una relación de menor o igual (≤) podríamos definir $a = b$ como una abreviación para $(a \leq b) \wedge (b \leq a)$.

Reglas de inferencia

La lógica de primer orden tiene dos reglas de inferencia. La primera es el modus ponens, heredada de la lógica proposicional. La segunda es la regla de *Generalización universal*, que es característica de la lógica de primer orden. La misma dice:

$$\frac{A}{\forall x A}$$

O en la notación del cálculo de secuentes:

$$A \vdash \forall x A$$

Es decir: a partir de A es posible concluir que $\forall x$ A.

Nótese que la regla de generalización universal es análoga a la regla de Necesitación de la lógica modal.

Axiomas

Los axiomas considerados aquí son los axiomas *lógicos* los cuales son parte del cálculo de predicados. Al formalizar teorías de primer orden particulares (como la aritmética de Peano) se agregan axiomas *no-lógicos* específicos, es decir axiomas que no se consideran verdades de la lógica pero sí verdades de una teoría particular.

Cuando el conjunto de axiomas es infinito, se requiere de un algoritmo que pueda decidir para una fórmula bien formada si es un axioma o no. Más aún, debería existir un algoritmo que pueda decidir si la aplicación de una regla de inferencia es correcta o no.

Es importante notar que el cálculo de predicados puede ser axiomatizado de varias formas diferentes. No existe nada canónico sobre los axiomas y reglas de inferencia aquí dadas, pero cualquier formalización produce los mismos teoremas de la lógica (y permite deducir los mismos teoremas de cualquier conjunto de axiomas no-lógicos).

Los siguientes tres axiomas son heredados de la lógica proposicional y se incorporan a la lógica de primer orden. Sean A, B y C fórmulas bien formadas de Q. Luego, los siguientes son axiomas lógicos:

Ax1: $A \rightarrow (B \rightarrow A)$

Ax2: $(A \rightarrow (B \rightarrow C)) \rightarrow ((A \rightarrow B) \rightarrow (A \rightarrow C))$

Ax3: $(\neg A \rightarrow \neg B) \rightarrow (B \rightarrow A)$

Los dos axiomas siguientes son característicos de la lógica de primer orden. Sean A y B fórmulas bien formadas de Q con *como máximo* una variable libre, *x*. Sea *t* un término cerrado y A(*x*/*t*) el resultado de reemplazar toda aparición de *x* en A por *t*. Luego, los siguientes son axiomas lógicos:

Ax4: $\forall x\ A \rightarrow A(x/t)$

Ax5: $\forall x\ (A \rightarrow B) \rightarrow (\forall x\ A \rightarrow \forall x\ B)$

Intuitivamente, el cuarto axioma dice que lo que vale para todos vale para cualquiera. Por ejemplo, un caso particular del axioma podría ser: «Si todos son mortales, entonces Abel es mortal»; o también: «Si todos son mortales, entonces el padre de Mateo es mortal» El quinto axioma es análogo al axioma K de la lógica modal, y un caso particular del mismo podría ser: «Si todos los humanos son mortales, entonces, si todos son humanos, todos son mortales.»

Semántica

Una interpretación es un par **<D,I>**, donde D es un conjunto no vacío llamado el dominio de discurso e **I** es una función llamada la *función de interpretación* definida como sigue:

1. Si *a* es un nombre, entonces **I** le asigna un elemento del dominio.
2. Si *f* es un functor de aridad *n*, entonces **I** le asigna una función de *n* argumentos que toma elementos del dominio y devuelve elementos del dominio.
3. Si *P* es un predicado de aridad *n*, entonces **I** le asigna un conjunto de *n*-tuplas construidas a partir del dominio.

Luego es posible definir la noción de **verdad** para una interpretación (para las *oraciones* de Q):[2]

1. $P(t_1,...,t_n)$ es verdadera para la interpretación M si y sólo si la *n*-tupla formada por las interpretaciones de $t_1,...,t_n$ es un elemento de la interpretación de *P*.
2. $\neg A$ es verdadera para la interpretación M si y sólo si A es falsa bajo esa interpretación.
3. $(A \wedge B)$ es verdadera para la interpretación M si y sólo si A es verdadera y B es verdadera bajo esa interpretación.
4. $(A \vee B)$ es verdadera para la interpretación M si y sólo si A es verdadera o B es verdadera bajo esa interpretación.
5. $(A \rightarrow B)$ es verdadera para la interpretación M si y sólo si A es falsa o B es verdadera bajo esa interpretación.
6. $(A \leftrightarrow B)$ es verdadera para la interpretación M si y sólo si A y B son ambas verdaderas o ambas falsas bajo esa interpretación.

Para dar las definiciones de verdad para fórmulas con la forma $\forall x$ A o $\exists x$ A, primero son necesarias algunas definiciones preliminares: Sea A(*x*/*a*) el resultado de reemplazar toda aparición de *x* en A por un nombre *a* (que no haya sido utilizado en la fórmula). Además, si M y M' son interpretaciones y *a* un nombre, entonces M' es una *a*-variante de M si y sólo si M' es idéntica a M o difiere sólo en el elemento del dominio que le asigna al nombre *a*.[3]

1. $\forall x$ A es verdadera para M si y sólo si A(*x*/*a*) es verdadera para toda *a*-variante de M.
2. $\exists x$ A es verdadera para M si y sólo si A(*x*/*a*) es verdadera para al menos una *a*-variante de M.

Una fórmula es **falsa** bajo una interpretación si y sólo si no es verdadera bajo esa interpretación.

A partir de esto pueden definirse varias otras nociones semánticas:

- Una fórmula es una verdad lógica si y sólo si es verdadera para toda interpretación.
- Una fórmula es una **contradicción** si y sólo si es falsa para toda interpretación.
- Una fórmula es **consistente** si y sólo si existe al menos una interpretación que la haga verdadera.
- Una fórmula A es una **consecuencia semántica** de un conjunto de fórmulas Γ si y sólo si no hay ninguna interpretación que haga verdaderas a todas las fórmulas en Γ y falsa a A. Cuando A es una consecuencia

semántica de Γ en un lenguaje Q, se escribe: $\Gamma \models_Q A$

- Una fórmula A es **lógicamente válida** si y sólo si es una consecuencia semántica del conjunto vacío. Cuando A es una fórmula lógicamente válida de un lenguaje Q, se escribe: $\models_Q A$

Metalógica

La lógica de primer orden es uno de los sistemas lógicos con propiedades metalógicas mejor conocidas. A continuación se introducen algunas de las más importantes.

Completitud

El teorema de completitud de Gödel, demostrado por Kurt Gödel en 1929, establece que existen sistemas de primer orden en los que todas las fórmulas lógicamente válidas son demostrables. Esto quiere decir que dado un lenguaje de primer orden Q, es posible seleccionar algunas fórmulas como axiomas, y algunas reglas de inferencia, de modo tal que todas las fórmulas lógicamente válidas (verdaderas bajo cualquier interpretación) sean demostrables a partir de los axiomas y las reglas de inferencia. Un ejemplo de axiomas y reglas de inferencia que permiten demostrar completitud son los que se dieron más arriba en este artículo.

Decidibilidad

Un sistema es decidible cuando existe al menos un método efectivo (un algoritmo) para decidir si una fórmula cualquiera del lenguaje del sistema es lógicamente válida o no. Por ejemplo, en la lógica proposicional, la evaluación de las fórmulas mediante tablas de verdad es un método efectivo para decidir si una fórmula cualquiera es lógicamente válida (una tautología). En este sentido, la lógica de primer orden es indecidible, siempre y cuando tenga al menos un predicado de aridad 2 o más (distinto de la identidad). Este resultado fue alcanzado de manera independiente por Alonzo Church en 1936 y por Alan Turing en 1937, dando así una respuesta negativa al Entscheidungsproblem planteado por David Hilbert en 1928. Por otra parte, la lógica de primer orden monádica (con o sin identidad) es decidible, como lo demostró Leopold Löwenheim en 1915.

El teorema de Löwenheim-Skolem

El teorema de Löwenheim-Skolem establece que si una teoría de primer orden numerable tiene un modelo infinito, entonces para cualquier número cardinal K, la teoría tiene un modelo de cardinalidad K.

En este contexto, una teoría de primer orden es simplemente un conjunto de fórmulas en un lenguaje de primer orden. Una teoría es numerable si sus fórmulas pueden ser puestas en correspondencia biunívoca con algún subconjunto (finito o infinito) de los números naturales. Y una teoría tiene un modelo infinto si tiene al menos una interpretación con un dominio infinito que hace verdaderas a todas las fórmulas de la teoría. Lo que el teorema de Löwenheim-Skolem afirma, entonces, es que si una teoría tiene una interpretación con un dominio infinito que hace verdaderas a todas las fórmulas de la teoría, entonces también tiene interpretaciones con dominios de *cualquier* cardinalidad que hacen verdaderas a todas las fórmulas de la teoría.

Esto significa que las lógicas de primer orden son incapaces de controlar la cardinalidad de sus modelos infinitos: si una teoría tiene un modelo infinito, entonces también tiene modelos infinitos de todas las cardinalidades. Una consecuencia de esto es que por ejemplo, la aritmética de Peano, que es una teoría de primer orden, tendrá como modelo no sólo al conjunto de los números naturales (que sería lo deseable), sino también al conjunto de los números reales e infinitos otros conjuntos de mayor cardinalidad.

El teorema de compacidad

El teorema de compacidad afirma que un conjunto de fórmulas de primer orden tiene un modelo si y sólo si todo subconjunto finito de ese conjunto tiene un modelo. Esto implica que si una fórmula es una consecuencia lógica de un conjunto infinito de axiomas, entonces es una consecuencia lógica de algún subconjunto finito de ellos.

El teorema fue demostrado por primera vez por Kurt Gödel como una consecuencia del teorema de completitud, pero con el tiempo se han encontrado varias demostraciones adicionales. El teorema es una herramienta central en teoría de modelos, ya que provee un método fundamental para construir modelos.

El teorema de Lindström

El teorema de Lindström establece que la lógica de primer orden es el sistema lógico más fuerte que cumple con el teorema de compacidad y el teorema descendente de Löwenheim-Skolem. Esto significa que el cumplimiento de esos dos teoremas *caracteriza* a la lógica de primer orden. Fue demostrado por Per Lindström, quien también definió la clase de los sistemas lógicos abstractos, permitiendo así la comparación entre sistemas.

Historia

Dónde ubicar los orígenes de la lógica de primer orden depende de lo que se entienda por lógica de primer orden. Si se entiende cualquier sistema lógico en torno a la cuantificación sobre individuos, entonces la lógica de primer orden es tan antigua como la lógica misma, y sus orígenes se remontan al *Órganon* de Aristóteles. Aristóteles realizó una gran cantidad de observaciones y contribuciones acerca del comportamiento de los cuantificadores «todos», «algunos», «ningún», etc. Construyó, por ejemplo, el famoso cuadro de oposición de los juicios, y ofreció una influyente clasificación para los distintos juicios con cuantificadores.

Sin embargo, si por lógica de primer orden se entiende un sistema lógico similar al expuesto en este artículo, entonces los orígenes de la lógica de primer orden deben buscarse recién en el siglo XIX, en la obra de Gottlob Frege. En 1879, Frege publicó su *Conceptografía (Begriffsschrift)*, donde presentó el primer sistema de lógica de predicados tal como lo entendemos hoy (aunque con una notación muy diferente a la actual). Luego lo refinaría en un trabajo de 1893 (y reeditado en 1903) titulado *Los fundamentos de la aritmética (Grundgesetze der Arithmetik)*. Sin embargo, la notación de Frege era difícil de entender, y sus revolucionarias contribuciones permanecieron desconocidas por varios años.

Entre 1910 y 1913, Bertrand Russell y Alfred North Whitehead publicaron *Principia Mathematica*, una monumental obra directamente influida por los trabajos de Frege. Con ella la lógica de predicados en general, y la lógica de primer orden en particular, cobraron una forma más familiar y alcanzaron una mayor audiencia.

Luego de *Principia Mathematica* comenzó una fértil época de resultados metalógicos para la lógica de primer orden (y otras). En 1915, Leopold Löwenheim demostró la consistencia, completitud semántica y decidibilidad de la lógica de primer orden monádica. En 1928, David Hilbert y Wilhelm Ackermann demostraron la consistencia de la lógica de primer orden. En 1929, Kurt Gödel demostró la completitud semántica de la lógica de primer orden. Y en 1936, Alonzo Church y Alan Turing demostraron, de manera independiente, la indecibilidad de la lógica de primer orden (no monádica).

En 1933, Alfred Tarski abrió otro capítulo en la historia de la lógica de primer orden (y de la lógica en general), con la publicación de sus definiciones de verdad para lenguajes formales. Las mismas permitieron el surgimiento de la teoría de modelos. En su trabajo, Tarski ofreció una definición de verdad para el lenguaje de la lógica de primer orden (entre otros) que todavía se utiliza. Dicha definición permitió refinar las demostraciones de consistencia y completitud semántica para la lógica de primer orden.

En 1934-1935, Gerhard Gentzen publicó *Investigaciones sobre la inferencia lógica (Untersuchungen über das logische Schliessen)*, donde introdujo una alternativa a la construcción axiomática de los sistemas lógicos (incluyendo la lógica de primer orden), conocida como la deducción natural.[4] Gentzen pronto desarrollaría la

deducción natural hasta llegar al cálculo de secuentes, y con la demostración del teorema de corte-eliminación (*cut-elimination theorem*), proveyó una nueva aproximación a la teoría de la demostración.

Notas y referencias

[1] No deben confundirse con los argumentos que estudia la lógica.

[2] Esta definición de verdad sólo sirve para las fórmulas bien formadas *cerradas* (oraciones) de Q. Es posible dar una definición para todas las fórmulas bien formadas, pero dicha definición involucra muchas complicaciones que no convienen a este artículo. Para la definición más general, véase

[3] Esta estrategia está tomada de

[4] Véase la sección «Natural deduction and sequent calculus» en

1. RAZONAMIENTO CATEGÓRICO Y CORRECCIÓN BAYESIANA

La inteligencia artificial no sólo se ocupa de mecanismos generales relacionados con la búsqueda de soluciones en un espacio dado, o de cómo representar y utilizar el conocimiento de un determinado dominio de discurso. Otro aspecto, hasta ahora sólo esbozado en el capítulo anterior, es el que corresponde a los mecanismos y/o procesos inferenciales, que consideraremos como el punto de partida de los llamados *modelos de razonamiento*.

En cualquier dominio, la propagación del conocimiento[44] por medio de programas de IA se efectúa siempre siguiendo un modelo de razonamiento bien definido. Estos modelos de razonamiento forman parte del motor de inferencias, si hablamos de sistemas de producción, o de las estructuras de control del conocimiento, si hablamos de cualquier otro tipo de sistemas de IA, y contribuyen de manera decisiva a organizar correctamente la búsqueda de soluciones.

Normalmente, las características del dominio, y las características de los problemas que deben resolverse, condicionan el tipo de modelo de razonamiento que debemos emplear. Así:

- Hay dominios de naturaleza marcadamente simbólica, en los que las soluciones pueden establecerse "con total seguridad". En estos casos emplearemos modelos categóricos de razonamiento.

- Por otra parte, hay dominios de naturaleza estadística, en los que las soluciones no pueden ser unívocamente obtenidas y en los que, además, tendremos que averiguar cuál de las posibles soluciones encontradas es la más probable. En estos casos preferiremos razonar con modelos de naturaleza estadística, de los cuales, dadas las peculiaridades de los procesos inferenciales que trata la IA, el esquema bayesiano es el más utilizado.

- Hay otros dominios en los que aparece el concepto de *incertidumbre*, que puede ser inherente a los datos del problema y a los hechos del dominio, o a los propios mecanismos inferenciales. En estos casos elegiremos modelos de razonamiento que sean capaces de manipular correctamente dicha incertidumbre.

- Por último[45], hay dominios en los que los elementos inferenciales incluyen matices de carácter lingüístico, entre los que pueden establecerse jerarquías y

[44] O lo que es lo mismo, el establecimiento de circuitos inferenciales apropiados.

[45] Aunque hay más tipos de dominios diferentes, los modelos de razonamiento que desarrollaremos bastarán para poder abordar gran cantidad de problemas interesantes. Más aún si tenemos en cuenta el carácter introductorio de este texto.

clasificaciones. En estos casos es conveniente emplear modelos de razonamiento basados en *conjuntos difusos*[46].

Como es obvio, la clasificación que acabamos de establecer es sólo eso, una clasificación, y habrá dominios que participen de varias de las características mencionadas. En estos casos podrá optarse por la combinación de diferentes modelos, o por la implementación del modelo de razonamiento que más se ajuste a las características del dominio.

1.1. Interpretación Diferencial

Una de las grandes cuestiones en la resolución de problemas de inteligencia artificial es cómo utilizar los datos y las verdades demostradas, según un procedimiento encadenado y lógico, al objeto de poder discriminar entre las posibles "soluciones", inicialmente candidatas, hasta encontrar la verdadera respuesta del problema planteado.

Cuando el dominio es de naturaleza simbólica, ya se ha comentado que el proceso de razonamiento adecuado debe seguir una aproximación categórica. Uno de tales procedimientos categóricos es el de la *interpretación diferencial*. Supongamos que a un experto de un universo de discurso concreto le preguntamos:

-¿Cómo interpretaría usted estos datos y esta información en este contexto?-

Una posible respuesta del experto a nuestra pregunta podría ser la siguiente:

"En primer lugar analizo los datos disponibles tratando de evaluar la importancia relativa de los mismos. Parte de la información puede ser de vital importancia. El resto de la información puede ser simplemente una consecuencia menor del problema principal, o puede estar relacionada con el problema, pero ser... digamos, información de segundo orden. A continuación, una vez clasificada la información de partida, trato de establecer un conjunto de posibles interpretaciones que sean compatibles con los datos iniciales. Una vez establecido este conjunto inicial de hipótesis realizo un análisis más profundo de los datos, y busco información complementaria que me permita ir descartando una a una las hipótesis iniciales. Este proceso lo continúo hasta que finalmente encuentro la solución. Pero a veces mi conjunto inicial de hipótesis está mal planteado, o mi conocimiento sobre el caso a veces no es completo, lo que se suele traducir en que no soy capaz de resolver el problema; es decir, me quedo sin hipótesis. Otras veces, por el contrario, el problema está en los datos, que no son suficientes. En estos casos suelo encontrarme con más de una hipótesis que es perfectamente compatible con la información de partida y con mi conocimiento sobre el tema. En otras palabras... Soy capaz de "acotar" la solución, pero al no ser única, mi interpretación puede no bastar, puede no decidir...-"

Nuestro hipotético experto, que es de una modestia más que notable, nos acaba de describir un caso típico de interpretación diferencial que nos va a servir para ilustrar

[46] O borrosos - *fuzzy* en inglés-.

el modelo de razonamiento categórico, sus ventajas, sus inconvenientes, y el verdadero alcance de este procedimiento en el contexto de los procesos inferenciales que trata la IA.

Tratemos de formalizar este complejo proceso de razonamiento que nos ha descrito nuestro hipotético experto. En primer lugar, nuestro experto ha tenido que reunir todo un conjunto de información *relevante*. Luego ha efectuado una ponderación de la importancia relativa de dicha información. A continuación ha tratado de relacionar la información disponible con un conjunto de interpretaciones inicialmente posibles. Finalmente ha ido eliminando de su "lista" inicial de hipótesis todas aquellas que no se podían concatenar lógicamente con los datos. Por último, tras el proceso de eliminación de hipótesis potencialmente relevantes, nuestro experto se ha dado cuenta de que: (a) ha encontrado la solución del problema, o (b) el conjunto de hipótesis formulado inicialmente no es consistente con los datos, o (c) hay varias hipótesis que se corresponden con los datos, entre las cuales no es posible discriminar. Así, el proceso global podría seguir el siguiente esquema:

- Recopilación de información.
- Análisis de la importancia relativa de las manifestaciones del problema.
- Análisis de las posibles causas del problema tras considerar, conjunta y razonablemente, todas las manifestaciones del problema. Ello implica el establecimiento tentativo de relaciones causa-efecto.
- Exclusión una a una de todas aquellas interpretaciones que no pueden ser explicadas completa y razonablemente por los datos.
- Fin del proceso con alguno de los siguientes resultados:
    - Existe una única solución
    - No hay ninguna solución
    - Hay varias soluciones posibles entre las que no se puede discriminar.

Este proceso de razonamiento -sistemático pero complejo- puede simplificarse en función del verdadero grado de experiencia del experto. Así, muchos pasos pueden ser claramente innecesarios para un profesional consolidado, que aplicará su "sentido común" y su conocimiento heurístico para restringir al máximo su conjunto inicial de hipótesis. También, su intuición, matizada y depurada a través de largos años de profesión, llevan al experto a efectuar el proceso de establecimiento de relaciones causa-efecto de una manera eficaz y eficiente.

El novato (o simplemente, el "menos experto"), tiende a aplicar metodologías completas que le permitan resolver adecuadamente el problema. El resultado final puede ser el mismo, incluso puede ser correcto, pero el verdadero experto suele optimizar mucho más sus recursos que el novato. En resumen, nosotros asumiremos el

papel de novatos y trataremos de definir un modelo de razonamiento que nos permita resolver problemas de naturaleza categórica. Exigiremos además que nuestro modelo:

- Sea sistemático
- Utilice algún sucedáneo del "sentido común"
- Incorpore algo parecido a la "intuición"

1.2. Elementos del Razonamiento Categórico

Puesto que una de las tareas que hay que realizar en un proceso de razonamiento categórico es la de establecer un conjunto de relaciones causales[47], comenzaremos describiendo nuestro dominio de discurso a partir de dos entidades diferentes, entre las cuales deberemos ser capaces de establecer relaciones. Estas entidades son:

- las manifestaciones posibles en el dominio de discurso
- las interpretaciones posibles en el dominio de discurso

Cualquier dominio estará perfectamente descrito cuando hayamos especificado:

- Todas las posibles manifestaciones de los problemas que puedan darse en el dominio.
- Todos los posibles problemas del dominio.
- Todas las relaciones causales que puedan establecerse entre problemas y manifestaciones.

Dicho de este modo la cuestión puede parecer algo abstracta. Trataremos ahora de clarificar todo este embrollo con un ejemplo: Supongamos que queremos construir un sistema inteligente para efectuar predicciones meteorológicas. En este caso, el dominio será el de las PREDICCIONES METEOROLOGICAS. El conjunto de manifestaciones podría estar formado por una serie de observaciones como *el color del cielo, la presencia de nubes, el grado de humedad,...* El conjunto de interpretaciones podría estar formado por hipótesis como *posibilidad de lluvia, posibilidad de tormentas, posibilidad de vientos fuertes,...* Finalmente, las relaciones causales serán las que nos permitan inferir interpretaciones a partir de manifestaciones. Por ejemplo: *Si el crepúsculo es de color rojizo, y no hay nubes en el cielo, entonces el pronóstico es de buen tiempo*.

Formalmente, para construir el modelo necesitamos definir una serie de funciones de carácter booleano[48] que nos sirvan para describir el dominio. Si $x_1,..., x_n$ es el conjunto completo de todas las manifestaciones posibles del universo de discurso, entonces la función $f(x_1,..., x_n)$, de carácter booleano, le asignará el valor "1" a la manifestación x_i, si la manifestación x_i está presente en nuestro problema concreto, o le asignará el valor "0", si dicha manifestación está ausente.

---

[47] Es decir, relaciones causa-efecto.

[48] Las funciones son de carácter booleano puesto que el modelo es categórico. En tales modelos, algo está presente o está ausente, algo es posible o es imposible.

Del mismo modo, si $y_1,..., y_m$ representa el conjunto completo de todas las posibles interpretaciones que se pueden dar a los problemas del dominio, para un problema concreto la función $g(y_1,..., y_m)$ le asignará el valor "1" a la interpretación y_j si la interpretación y_j es posible[49], o le asignará el valor "0" si dicha interpretación es imposible.

Necesitaremos una tercera función, que denominaremos función *E* que representa el conjunto de todas las posibles relaciones causales que se pueden establecer en nuestro dominio de discurso, entre manifestaciones e interpretaciones. La función *E* es la *función de conocimiento*. Con estos tres elementos nuestro problema se reduce a encontrar, ante un conjunto de manifestaciones relacionadas con un problema, en un dominio, la función *g* que satisface:

$E: (f \rightarrow g)$

Expresión que podemos leer del siguiente modo: "... encontrar el conjunto de interpretaciones que es compatible con las observaciones y datos de que disponemos, tras la aplicación de nuestro conocimiento sobre el dominio de discurso..."

Claramente, puesto que *E* es la función de conocimiento que nos permite establecer relaciones de tipo causal entre manifestaciones e interpretaciones, su expresión formal será del tipo:

$E\ (x_1, ..., x_n, y_1, ..., y_m)$

Para ilustrar cómo podemos utilizar estas tres funciones en un modelo categórico de razonamiento trataremos de resolver un ejemplo sencillo:

Sea un dominio *D* en el que podemos identificar dos posibles manifestaciones que están relacionadas con dos posibles interpretaciones. Llamaremos:

- *D* = Dominio de discurso
- *M* = Conjunto de manifestaciones = [*m*(1), *m*(2)]
- *I* = Conjunto de interpretaciones = [*i*(1), *i*(2)]

Supongamos que, tras estudiar el dominio, hemos sido capaces de identificar el siguiente conocimiento:

- Para que la interpretación *i*(2) sea cierta, la manifestación *m*(1) debe estar presente.

- Para que la interpretación *i*(1) sea cierta y, al mismo tiempo, podamos descartar la interpretación *i*(2), entonces la manifestación *m*(2) debe estar presente.

- Para que la interpretación *i*(2) sea cierta y, al mismo tiempo, podamos descartar la interpretación *i*(1), entonces la manifestación *m*(2) no debe estar presente.

---

[49] ... a la vista de las manifestaciones del problema.

- Si alguna de las manifestaciones está presente es porque alguna de las interpretaciones puede establecerse.

Con las declaraciones anteriores, que expresan en lenguaje natural nuestro conocimiento sobre el dominio, vamos a construir la función *E*:

$$
\begin{aligned}
E = \{ \; & [\, i(2) \rightarrow m(1)] \quad \text{and} \\
& [\, i(1) * \neg i(2) \rightarrow m(2)] \quad \text{and} \\
& [\, \neg i(1) * i(2) \rightarrow \neg m(2)] \quad \text{and} \\
& [\, m(1) + m(2) \rightarrow i(1) + i(2)] \\
\} &
\end{aligned}
$$

En esta expresión, los símbolos “*” y “+” deben leerse, respectivamente como “and” y como “or inclusivo”.

Una vez hemos sido capaces de formalizar nuestro conocimiento, trataremos de resolver un problema que presenta las siguientes manifestaciones:

- la manifestación $m(2)$ está presente en el problema
- hay evidencia total de que la manifestación $m(1)$ no está presente

Con esta información la función booleana *f* será:

$$f = \neg m(1) * m(2)$$

Nótese que, según nuestro planteamiento, la función *f* debe incluir toda la evidencia positiva y toda la evidencia negativa. Así, dicha función sería diferente si no dispusiéramos de información acerca de una manifestación dada[50].

Según hemos mencionado ya, nuestro problema de razonamiento es encontrar la función *g* que verifique:

$$E: (f \rightarrow g)$$

En lenguaje natural, la pregunta que deberíamos ser capaces de responder es la siguiente: ¿Qué “conjunto” de interpretaciones verifica simultáneamente la ausencia de la manifestación $m(1)$ y la presencia de la manifestación $m(2)$?

En este sencillo ejemplo es fácil ver que la solución es[51]:

$$g = i(1) * \neg i(2)$$

En este caso la aplicación de los métodos y procedimientos lógicos habituales conducen rápidamente a la solución. La situación, no obstante, cambia drásticamente si

---

[50] Más concretamente, si cambiamos la declaración (b) de nuestro ejemplo por esta otra declaración: -No tenemos información acerca de $m(1)$-, entonces la función *f* sería: $f = m(2)$

[51] Los autores animan al lector a que traten de resolver por los procedimientos “clásicos” el problema planteado.

consideramos -pongamos por caso- un dominio en el que fuesen posibles 30 manifestaciones, relacionadas con 600 posibles interpretaciones a través de una función *E* en la que se resumen 125 relaciones causales[52]. En este contraejemplo, los procedimientos lógicos clasicos, la complejidad de los procesos de resolución, y la necesidad de efectuar muchas sustituciones, podrían hacer inviable la resolución del problema. Aparece así la necesidad de encontrar alternativas mejores, conceptualmente correctas, y computacionalmente eficaces. Una de tales alternativas es la que se describe a continuación.

1.3. Un Procedimiento Sistemático para el Razonamiento Categórico

Vamos a plantearnos la cuestión anterior desde una perspectiva algo diferente. Decíamos que, para un dominio concreto, habíamos sido capaces de encontrar todas las posibles manifestaciones de los problemas que pudieran aparecer, y que estas manifestaciones estaban relacionadas con un conjunto de interpretaciones. Por otra parte, las relaciones entre manifestaciones e interpretaciones se establecían a través de nuestro conocimiento sobre el dominio. En este contexto, el procedimiento sistemático que proponemos para razonar categóricamente consta de las siguientes fases:

- construcción del conjunto de todas las combinaciones que se pueden establecer entre las manifestaciones del dominio. A este conjunto lo denominamos *conjunto de complejos de manifestaciones*.

- construcción del conjunto de todas las combinaciones que se pueden establecer entre las interpretaciones del dominio. A este conjunto lo denominamos *conjunto de complejos de interpretaciones*.

- construcción del conjunto completo de todas las combinaciones posibles entre complejos de manifestaciones y complejos de interpretaciones. A este conjunto lo denominamos *conjunto de complejos manifestación-interpretación*.

De este modo, si en el dominio considerado hemos identificado n manifestaciones (i.e., $m(1)$, ..., $m(n)$) y t interpretaciones (i.e., $i(1)$, ..., $i(t)$), entonces el número de complejos de manifestaciones será 2^n, el número de complejos de interpretaciones será 2^t, y el número de complejos manifestación-interpretación será $2^{(n+t)}$.

Dado el carácter exhaustivo del procedimiento, los tres conjuntos que acabamos de definir tienen las siguientes propiedades:

- Sus elementos son mutuamente excluyentes
- Son conjuntos completos

El conjunto de complejos manifestación-interpretación representa el total de situaciones idealmente posibles en nuestro universo de discurso, pero es evidente que

[52] En cualquier caso, un dominio relativamente reducido.

no todas ellas van a poder darse en la realidad. Es más, muchas de ellas serán claramente absurdas. El papel del conocimiento será el de restringir el conjunto total de situaciones idealmente posibles a un conjunto de situaciones realmente posibles (i.e., permitidas por el conocimiento disponible).

Para ilustrar mejor el procedimiento, aplicaremos este esquema al problema que nos ha servido de ejemplo en la sección anterior, y en el que habíamos identificado dos posibles manifestaciones, $m(1)$ y $m(2)$, posiblemente relacionadas con dos interpretaciones, $i(1)$ e $i(2)$, a través del siguiente conocimiento:

(a) $i(2) \rightarrow m(1)$
(b) $i(1) * \neg i(2) \rightarrow m(2)$
(c) $\neg i(1) * i(2) \rightarrow \neg m(2)$
(d) $m(1) + m(2) \rightarrow i(1) + i(2)$

de forma que E = [(a) and (b) and (c) and (d)]

Si M es el conjunto de complejos de manifestaciones, entonces, en función de los valores lógicos de cada manifestación particular, y de acuerdo con el criterio que se muestra a continuación:

$m(1)$	0	0	1	1
$m(2)$	0	1	0	1
	m_1	m_2	m_3	m_4

el conjunto M estará formado por los siguientes elementos:

$M = [m_1, m_2, m_3, m_4]$, donde:

$m_1 = \neg m(1) * \neg m(2)$ -primera columna-
$m_2 = \neg m(1) * m(2)$ -segunda columna-
$m_3 = m(1) * \neg m(2)$ -tercera columna-
$m_4 = m(1) * m(2)$ -cuarta columna-

Nótese que los valores de los complejos de M dependen del criterio elegido para asignar valores lógicos a las manifestaciones individuales. El resultado hubiese sido distinto si decidiésemos emplear el siguiente criterio:

$m(1)$	0	1	0	1
$m(2)$	0	0	1	1
	m_1	m_2	m_3	m_4

Obsérvese también que las manifestaciones de un problema concreto del dominio vendrán representadas exclusivamente por uno cualquiera de los complejos de M[53].

[53] Evidentemente, siempre que todos los datos sean conocidos.

Análogamente, si *I* es el conjunto de complejos de interpretaciones, entonces, en función de los valores lógicos de cada interpretación particular, y de acuerdo con el criterio que se muestra a continuación:

$i(1)$	0	0	1	1
$i(2)$	0	1	0	1
	i_1	i_2	i_3	i_4

el conjunto *I* estará formado por los siguientes elementos:

$I = [i_1, i_2, i_3, i_4]$, donde:

$i_1 = \neg i(1) * \neg i(2)$ -primera columna-
$i_2 = \neg i(1) * i(2)$ -segunda columna-
$i_3 = i(1) * \neg i(2)$ -tercera columna-
$i_4 = i(1) * i(2)$ -cuarta columna-

A continuación debemos construir el conjunto completo de todas las posibles combinaciones entre complejos de manifestaciones y complejos de interpretaciones:

	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$	$\boldsymbol{m_3}$	$\boldsymbol{m_4}$	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$	$\boldsymbol{m_3}$	$\boldsymbol{m_4}$	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$	$\boldsymbol{m_3}$	$\boldsymbol{m_4}$	$\boldsymbol{m_1}$	$\boldsymbol{m_2}$	$\boldsymbol{m_3}$	$\boldsymbol{m_4}$
***m*(1)**	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
***m*(2)**	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
***i*(1)**	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
***i*(2)**	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
	i_1				i_2				i_3				i_4			

El resultado es una lista exhaustiva de complejos manifestación-interpretación, que se denomina *base lógica expandida*, y que -en este caso- estará formada por 16 complejos siguientes[54]:

Base Lógica Expandida = BLE = [m_1i_1, m_2i_1, m_3i_1, m_4i_1,
m_1i_2, m_2i_2, m_3i_2, m_4i_2,
m_1i_3, m_2i_3, m_3i_3, m_4i_3,
m_1i_4, m_2i_4, m_3i_4, m_4i_4,]

Pero no todos los complejos manifestación-interpretación de la base lógica expandida van a ser "realmente" posibles en el dominio de discurso. Por el contrario, muchos de tales complejos van a estar prohibidos por el conocimiento y, por lo tanto, habrá que descartarlos. Así, la aplicación de la función de conocimiento E sobre la base lógica expandida nos genera la llamada "Base Lógica Reducida" -BLR-, en la que sólo figuran aquellos complejos manifestación-interpretación que son compatibles con el conocimiento que se tiene sobre el dominio en cuestión:

E: BLE $\rightarrow$ BLR

---

[54] Recuerde que para *n* manifestaciones y *t* interpretaciones, el número máximo de complejos manifestación-interpretación es $2^{(n+t)}$. Aquí tenemos 2 manifestaciones y 2 interpretaciones, por lo que hay 16 complejos.

De este modo, la declaración (a) de la función *E* del ejemplo que estamos considerando, elimina de la BLE los complejos:

- $m_1 i_2$
- $m_1 i_4$
- $m_2 i_2$
- $m_2 i_4$

Esta eliminación se explica considerando que, con independencia de lo que pueda pasar con $i(1)$, para que $i(2)$ sea cierta (i.e., en términos de complejos: i_2 ó i_4), la manifestación $m(1)$ debe estar presente. De ello se deduce inmediatamente que m_1 y m_2 no tienen sentido en el contexto de i_2 y de i_4.

Siguiendo el mismo razonamiento, la declaración (b) del ejemplo elimina los complejos:

- $m_1 i_3$
- $m_3 i_3$

Análogamente, la declaración (c) elimina los complejos:

- $m_2 i_2$ (ya eliminado en un paso anterior)
- $m_4 i_2$

Por último, la declaración (d) elimina los complejos:

- $m_2 i_1$
- $m_3 i_1$
- $m_4 i_1$

Al final, la aplicación del conocimiento del dominio sobre la BLE se traduce en la siguiente BLR:

$$\text{BLR} = [m_1 i_1, m_3 i_2, m_2 i_3, m_4 i_3, m_3 i_4, m_4 i_4]$$

Si nuestro conocimiento sobre el dominio es completo, y si el dominio está descrito correctamente, la solución a cualquier problema que podamos plantearnos está en BLR.

Volvamos al problema que nos sirve de ejemplo, y en el cual la cuestión planteada es encontrar una solución para el problema representado por la función:

$f = \neg m(1) * m(2)$

Nótese que, de acuerdo con el criterio elegido para el establecimiento de los correspondientes valores lógicos:

$f = \neg m(1) * m(2) = m_2$

La cuestión hay que formularla ahora en estos términos: ¿Qué complejos de BLR contienen al complejo de manifestaciones m_2?

En este caso sólo hay uno: $m_2 i_3$. Por lo tanto la solución al problema planteado es:

$g = i_3 = i(1) * \neg i(2)$

Esta solución es, precisamente, la que habríamos encontrado si hubiésemos resuelto el problema siguiendo un procedimiento "lógico tradicional".

Ahora somos capaces de asegurar que la interpretación $i(1)$ es absolutamente cierta, y que la interpretación $i(2)$ es falsa o imposible, conclusión a la que hemos llegado tras la aplicación sistemática de un procedimiento que es consistente con la lógica.

Supongamos ahora que revisamos nuestro conocimiento, y encontramos una nueva declaración que es cierta:

(e) $\neg i(1) * \neg i(2) \rightarrow m(1) + m(2)$

Esta nueva declaración debería ser añadida a E con lo que obtendríamos una nueva función de conocimiento: [$E' = E$ and (e)], y su aplicación sobre la BLE nos generaría la siguiente BLR:

BLR' = [$m_3 i_2$, $m_2 i_3$, $m_4 i_3$, $m_3 i_4$, $m_4 i_4$]

ya que la declaración (e) elimina al complejo $m_1 i_1$. ¿Qué ocurriría ahora si tuviésemos que resolver un problema con las manifestaciones: $f' = \neg m(1)$ and $\neg m(2)$? En este caso, $f' = m_1$, y no hay ningún complejo en BLR que contenga a dicho complejo de manifestaciones. Esta situación puede darse por una cualquiera (o todas) de las siguientes razones:

- Las manifestaciones no son realmente esas.
- El conocimiento no es correcto. Hay algún error en la función E'.
- El dominio no está bien construido.

El primer problema se resuelve efectuando una nueva recogida de datos, al objeto de comprobar que no hemos cometido errores al construir la función f'. Si el conjunto de manifestaciones sigue siendo el mismo, entonces hay que sospechar que el error está, bien en la función de conocimiento (que puede ser incompleta, demasiado restrictiva, o simplemente falsa), bien en la construcción del dominio, en el pudiera haber manifestaciones y/o posibles interpretaciones que no hayan sido tenidas en cuenta.

En este ejemplo, el lector atento se habrá dado cuenta de que la declaración (e) es contradictoria con la declaración (d), por lo que la función E' es incorrecta. Además, aún considerando que el dominio está bien construido, el nuevo problema planteado, f',

indica que no hay manifestaciones, y si no hay manifestaciones no hay problemas, y si no hay problemas... ¿para qué queremos interpretar algo que no existe?[55]

Analicemos a continuación otra posibilidad que puede surgir. Supongamos ahora que el problema se manifiesta del siguiente modo:

$f'' = m(1) * \neg m(2) = m_3$

En este caso aparecen en BLR dos complejos manifestación-interpretación:

- $m_3 i_2$
- $m_3 i_4$

Por lo tanto, de la BLR derivamos que:

$g'' = i_2 + i_4 = \neg i(1) * i(2) + i(1) * i(2)$

En otras palabras, la aplicación del método nos permite afirmar que la interpretación $i(2)$ es cierta; sin embargo, no nos permite afirmar nada acerca de $i(1)$. Estamos ante uno de los problemas que nos comentaba nuestro hipotético experto: hemos podido "acotar" la solución, pero no hemos podido resolver la cuestión completamente. Como regla general, esta situación no es aceptable, y constituye uno de los problemas más serios del modelo que acabamos de desarrollar.

Otro problema importante del esquema categórico es el de la casi siempre inevitable explosión combinatoria. En efecto, con sólo 7 manifestaciones y 24 posibles interpretaciones, nuestra BLE estaría constituida por 2.147.483.600 complejos manifestación-interpretación.

Estas, y otras deficiencias más sutiles, aconsejan la puesta a punto de un modelo alternativo.

1.4. La Corrección Bayesiana

Las interpretaciones categóricas son más bien infrecuentes en el mundo real. Además:

- ¿Podemos afirmar siempre, y sin equivocarnos, que todos los problemas se manifiestan? Ya hemos comentado que no. Al contrario, hay problemas que no se manifiestan nunca, y otros que tardan mucho en manifestarse.

- ¿La presencia de una manifestación dada es siempre indicativa de algún problema? En principio, parece que si hay algún tipo de manifestación es porque hay algo que

[55] Los autores son conscientes de que esta última afirmación es peligrosa. En efecto, por ejemplo en medicina existen muchas enfermedades, algunas de ellas muy serias, que son asintomáticas, y por lo tanto no se manifiestan, o tardan mucho en manifestarse. En tales casos, este modelo podría no ser apropiado.

hace que tal manifestación se produzca, pero no siempre una manifestación dada es indicativa de lo que podríamos llamar “un problema importante”

Con estas consideraciones vamos a replantearnos la cuestión desde otro punto de vista. Para ello trataremos de encontrar una respuesta a la pregunta siguiente:

- Dado un universo y un conjunto de atributos... ¿cual es la probabilidad de que un determinado elemento del universo presente ciertos atributos del conjunto total?

En términos estadísticos, si N es una determinada población, y x_1, x_2,..., x_n es el conjunto de todos los atributos posibles, la función booleana $f(x_1,..., x_n)$ genera un subconjunto de atributos. De este modo, si $N(f)$ es el subconjunto de los elementos de N que presentan tales atributos, la probabilidad total de f será:

$$P(f) = \frac{N(f)}{N}$$

El concepto de probabilidad total, sin embargo, no es suficiente para construir un modelo de razonamiento. Necesitamos introducir el concepto de *probabilidad condicional*, que la excelente enciclopedia Espasa define como “... la probabilidad de las causas”.

En la probabilidad condicional aparecen involucrados dos sucesos, de forma que la ocurrencia del segundo depende de la ocurrencia del primero. Veamos un ejemplo:

Para tratar de resolver un problema concreto sabemos que podemos ejecutar dos acciones potencialmente eficaces, *A* ó *B*. Después de ejecutar una cualquiera de tales acciones, sabemos que nuestro problema puede evolucionar positivamente (*E* = éxito), o puede evolucionar mal (*F* = fracaso). En cualquier caso sabemos que la evolución del problema depende de la acción[56] y, por lo tanto, el problema no evoluciona solo. Elijamos la acción por el método de la moneda[57]. Así:

- $p(A) = 0.5$
- $p(B) = 0.5$

Sabemos además que, para este problema concreto, y después de muchos ensayos, la evolución del problema tras ejecutar la acción A fue positiva el 20% de las veces. En el caso de la acción *B* se obtuvo una evolución positiva el 60% de las veces. Según este planteamiento, la probabilidad “a priori” de resolver el problema es:

$$p(E) = p(E/A)\,p(A) + p(E/B)\,p(B) = (0.2 \times 0.5) + (0.6 \times 0.5) = 0.4$$

en donde:

- $p(E)$ = probabilidad total de éxito “a priori”

---

[56] Es decir, hay una relación causal más o menos evidente.

[57] Un método como otro cualquiera, en ocasiones muy utilizado.

- $p(A)$ = probabilidad de ejecutar la acción *A*
- $p(B)$ = probabilidad de ejecutar la acción *B*
- $p(E/A)$ = probabilidad de éxito si ejecutamos la acción *A*
- $p(E/B)$ = probabilidad de éxito si ejecutamos la acción *B*

Claramente, $p(E/A)$ y $p(E/B)$ son probabilidades condicionales "a priori", puesto que proceden de estudios previos.

Este planteamiento se acerca algo más a los aspectos relativos al *razonamiento*. Sin embargo, todavía nos falta algo. En realidad lo que a nosotros nos interesa, en principio, es interpretar cosas que ya han pasado, para lo cual necesitamos introducir algún mecanismo para razonar "a posteriori". Para ello, vamos a variar algo las premisas del ejemplo:

Sabemos que hemos tenido un problema concreto para cuya resolución podíamos ejecutar dos acciones, *A* ó *B*, cada una de las cuales llevaba asociada una determinada probabilidad de éxito. Sabemos que, efectivamente, una de tales acciones fue ejecutada, a consecuencia de lo cual el problema evolucionó positivamente. ¿Cuál es la probabilidad de que la acción ejecutada haya sido la acción *A*?

Este fue, precisamente, el planteamiento del reverendo Bayes, que finalmente se tradujo en el establecimiento de su famoso *teorema*[58]. En concreto, en nuestro caso, el problema planteado es encontrar $p(A/E)$. La solución que se deriva del teorema de Bayes, y que desarrollaremos a continuación, es la siguiente:

$$p(A/E) = \frac{p(E/A)\,p(A)}{p(E)}$$

En esta expresión, como cabía esperar, la probabilidad condicional "a posteriori" se obtiene a partir de la probabilidad condicional "a priori" y de las probabilidades totales.

Para ilustrar la obtención de la ecuación del teorema de Bayes que acabamos de utilizar, consideraremos una población sobre parte de cuyos elementos ha sido ejecutada la acción *A*, de entre un número arbitrario de acciones posibles. Sobre todos los elementos de dicha población se ha ejecutado una acción (*A* u otra cualquiera de las posibles), a consecuencia de lo cual ha habido una determinada respuesta (*E*, *F*, *G*,...) Supongamos también que nos interesa investigar la relación causal *A*-*E*. De este modo:

- N = población
- Acciones posibles = A, $\neg A$
- Respuestas posibles = E, $\neg E$
- $n(A)$ = nº de elementos de *N* sobre los que se ejecutó *A*
- $n(\neg A)$ = nº de elementos de *N* sobre los que no se ejecutó *A*
- $n(E)$ = nº de elementos de *N* cuya respuesta fue *E*

[58] Nombre quizás demasiado pretencioso para lo que en realidad es una simple ecuación.

- $n(\neg E)$ = nº de elementos de N cuya respuesta no fue E
- $n(A \cap E)$ = nº de elementos de N sobre los que se ejecutó A y se obtuvo como respuesta E

Por la definición de probabilidad condicional:

$$p(E/A) = \frac{n(A \cap E)}{n(A)}$$

Si dividimos numerador y denominador por N obtenemos:

$$p(E/A) = \frac{\frac{n(A \cap E)}{N}}{\frac{n(A)}{N}} = \frac{p(A \cap E)}{p(A)}$$

De donde:

$$p(A \cap E) = p(E/A)\,p(A)$$

Análogamente,

$$p(A/E) = \frac{\frac{n(A \cap E)}{N}}{\frac{n(E)}{N}} = \frac{p(A \cap E)}{p(E)}$$

De donde:

$$p(A \cap E) = p(A/E)\,p(E)$$

Igualando términos:

$$p(E/A)\,p(A) = p(A/E)\,p(E)$$

Expresión que representa la ecuación del teorema de Bayes utilizada en el ejemplo anterior.

Esta ecuación debe poder generalizarse para el análisis de problemas más complicados. La forma de obtener una expresión generalizada para el teorema de Bayes se ilustra a continuación.

Sea una característica cualquiera x y denotemos por A al conjunto de individuos de una población estadísticamente significativa en los que la característica x está presente. Obviamente, $\neg A$ es el conjunto de individuos de la misma población estadística en los que x está ausente. Sea P una prueba potencialmente resolutiva para investigar la característica x, de forma que E es el conjunto de elementos de la población para los cuales P ha dado resultados positivos, y $\neg E$ es el conjunto de

elementos de la población para los cuales *P* ha dado resultados negativos. Sean ahora los datos representados en la Tabla 1.1, donde las letras minúsculas representan números.

	A	**¬*A***	**Totales**
E	*a*	*b*	*a* + *b*
¬*E*	*c*	*d*	*c* + *d*
Totales	*a* + *c*	*b* + *d*	*N*

Tabla 1.1

Del análisis de la Tabla 1.1, fácilmente se observa que:

- *a* = nº de positivos reales
- *b* = nº de falsos positivos
- *c* = nº de falsos negativos
- *d* = nº de negativos reales

A partir de los datos de la tabla se pueden establecer las siguientes probabilidades totales o *prevalencias*:

$$p(E) = \frac{(a+b)}{N}$$
$$p(\neg E) = \frac{(c+d)}{N}$$
$$p(A) = \frac{(a+c)}{N}$$
$$p(\neg A) = \frac{(b+d)}{N}$$

También a partir de los datos de la tabla podemos establecer las siguientes probabilidades condicionales (que numeramos para referenciarlas en el curso de la demostración):

(1) $p(E/A) = \frac{a}{a+c}$

(2) $p(E/\neg A) = \frac{b}{b+d}$

(3) $p(\neg E/A) = \frac{c}{a+c}$

(4) $p(\neg E/\neg A) = \frac{d}{b+d}$

(5) $p(A/E) = \frac{a}{a+b}$

(6) $p(A/\neg E) = \frac{c}{c+d}$

(7) $p(\neg A/E) = \frac{b}{a+b}$

(8) $p(\neg A/\neg E) = \frac{d}{c+d}$

Con esta información ¿cómo podríamos conocer la probabilidad condicional “a posteriori” de la característica *x* dado un resultado positivo de la prueba *P*?

Según el teorema de Bayes que acabamos de demostrar:

$$p(A/E) = \frac{p(E/A)\,p(A)}{p(E)}$$

pero,

$$p(E) = \frac{(a+b)}{N}$$

de la expresión (1) se deduce que:

$$a = (a+c)\,p(E/A)$$

y de la expresión (2) se deduce que:

$$b = (b+d)\,p(E/\neg A)$$

por otra parte,

$$p(A) = \frac{(a+c)}{N}$$
$$p(\neg A) = \frac{(b+d)}{N}$$

por lo que,

$$(a+c) = N \cdot p(A)$$
$$(b+d) = N \cdot p(\neg A)$$

y efectuando las sustituciones oportunas,

$$a = N \cdot p(A) \cdot p(E/A)$$
$$b = N \cdot p(\neg A) \cdot p(E/\neg A)$$

Si sustituimos ambas expresiones en la ecuación de $p(E)$ resulta:

$$p(E) = p(A)\,p(E/A) + p(\neg A)\,p(E/\neg A)$$

Llevando este resultado a la expresión original del teorema de Bayes, obtenemos que:

$$p(A/E) = \frac{p(E/A)\,p(A)}{p(E/A)\,p(A) + p(E/\neg A)\,p(\neg A)}$$

Esta última ecuación es directamente generalizable. Efectivamente, si contemplamos más de dos posibilidades (en lugar de simplemente A y ¬A), resulta:

$$p(A_0 / E) = \frac{p(E / A_0)\, p(A_0)}{\sum_i p(E / A_i)\, p(A_i)}$$

que es la expresión generalizada del teorema de Bayes que emplearemos en este texto[59].

Esta ecuación podemos tratar de aplicarla al ejemplo que habíamos utilizado para ilustrar el razonamiento categórico, y para el cual, recordemos, habíamos obtenido la siguiente BLR:

$$\text{BLR} = [m_1 i_1,\ m_3 i_2,\ m_2 i_3,\ m_4 i_3,\ m_3 i_4,\ m_4 i_4]$$

La cuestión que entonces se nos planteaba era la siguiente:

Dado el conjunto de manifestaciones $f = m(1) \times \neg m(2) = m_3$, ¿cuál es la función g de las interpretaciones permitidas?

Para este problema, directamente de BLR, concluíamos que $g = i_2 + i_4$ o, lo que es lo mismo: $g = \neg i(1)\ i(2) + i(1)\ i(2)$, lo cual traducíamos del siguiente modo: -la interpretación $i(2)$ es segura, pero de $i(1)$ no podemos asegurar nada-.

Según la corrección bayesiana que acabamos de sugerir, el problema se reduce a resolver las siguientes expresiones:

- ¿$p(i_2/m_3)$?
- ¿$p(i_4/m_3)$?

Para ello, y como ya hemos demostrado:

$$p(i_2 / m_3) = \frac{p(m_3 / i_2)\, p(i_2)}{p(m_3 / i_1)\, p(i_1) + p(m_3 / i_2)\, p(i_2) + p(m_3 / i_3)\, p(i_3) + p(m_3 / i_4)\, p(i_4)}$$

y, análogamente:

$$p(i_4 / m_3) = \frac{p(m_3 / i_4)\, p(i_4)}{p(m_3 / i_1)\, p(i_1) + p(m_3 / i_2)\, p(i_2) + p(m_3 / i_3)\, p(i_3) + p(m_3 / i_4)\, p(i_4)}$$

Evidentemente, si queremos conocer los valores absolutos de tales probabilidades condicionales desde una perspectiva totalmente general, necesitaremos conocer los valores de:

$p(i_1)$,	$p(i_2)$,	$p(i_3)$,	$p(i_4)$
$p(m_1/i_1)$,	$p(m_1/i_2)$,	$p(m_1/i_3)$,	$p(m_1/i_4)$
$p(m_2/i_1)$,	$p(m_2/i_2)$,	$p(m_2/i_3)$,	$p(m_2/i_4)$
$p(m_3/i_1)$,	$p(m_3/i_2)$,	$p(m_3/i_3)$,	$p(m_3/i_4)$
$p(m_4/i_1)$,	$p(m_4/i_2)$,	$p(m_4/i_3)$,	$p(m_4/i_4)$

[59] Hay expresiones todavía más generales.

Sin embargo, si nuestro tratamiento lógico del problema es correcto, si los conjunto de manifestaciones e interpretaciones son completos, si las manifestaciones y las interpretaciones cumplen el requisito de independencia exigido por el teorema de Bayes[60], si la función de conocimiento está bien construida, y si el tratamiento estadístico efectuado es correcto, entonces sólo las siguientes probabilidades condicionales tendrán valores distintos de cero:

$$p(m_1/i_1),\ p(m_3/i_2),\ p(m_2/i_3),\ p(m_4/i_3),\ p(m_3/i_4),\ p(m_4/i_4)$$

ya que son las únicas probabilidades condicionales relativas a complejos manifestación-interpretación que aparecen en BLR. Dicho de otro modo: si después de un análisis estadístico encontramos valores distintos de cero, para las probabilidades condicionales de complejos manifestación-interpretación que no aparecen en BLR (e.g., $p[m_2/i_2] \neq 0$), entonces tendremos que pensar en alguna (o todas) de las siguientes deficiencias:

- nuestro planteamiento lógico no es correcto
- nuestra función de conocimiento no es correcta, bien porque sea incompleta o, simplemente, porque esté mal construida
- la estadística no ha sido bien realizada

Además, para asegurar la consistencia matemática del modelo se tiene que cumplir que:

$$\sum_i P(m_i/i_0) = 1$$

Asumamos para nuestro ejemplo los siguientes valores de las prevalencias y de las probabilidades condicionales:

$$p(i_1) = \frac{910}{1000} \qquad p(m_1 / i_1) = 1$$
$$p(i_2) = \frac{50}{1000} \qquad p(m_3 / i_2) = \frac{3}{5}$$
$$p(i_3) = \frac{25}{1000} \qquad p(m_2 / i_3) = 1$$
$$p(i_4) = \frac{15}{1000} \qquad p(m_4 / i_3) = \frac{2}{3}$$
$$p(m_3 / i_4) = \frac{2}{5}$$
$$p(m_4 / i_4) = \frac{1}{3}$$

Con estos valores:

---

[60] Volveremos sobre esta cuestión un poco más adelante.

$$p(i_2 / m_3) = \frac{\frac{3}{5} \times \frac{50}{1000}}{\left(\frac{3}{5} \times \frac{50}{1000}\right) + \left(\frac{2}{5} \times \frac{15}{1000}\right)}$$

$$p(i_4 / m_3) = \frac{\frac{2}{5} \times \frac{15}{1000}}{\left(\frac{3}{5} \times \frac{50}{1000}\right) + \left(\frac{2}{5} \times \frac{15}{1000}\right)}$$

Nótese que éstos son valores absolutos de probabilidad condicional. Si sólo hubiésemos deseado averiguar cuál de ambos complejos de interpretaciones es más probable[61], dado que el denominador es el mismo, simplemente dividiendo ambas expresiones entre sí obtendríamos:

$$\frac{p(i_2 / m_3)}{p(i_4 / m_3)} = \frac{\frac{3}{5} \times \frac{50}{1000}}{\frac{2}{5} \times \frac{15}{1000}} = \frac{5}{1}$$

lo que indica que el complejo de interpretaciones $i_2 = \neg i(1) \times i(2)$ es cinco veces más probable que el complejo de interpretaciones $i_4 = i(1) \times i(2)$.

A pesar de lo que acabamos de comentar, la corrección bayesiana también presenta problemas. Así, si manifestaciones e interpretaciones no son independientes, el modelo bayesiano fracasa. No pretendemos abordar una demostración rigurosa y formal; sin embargo, sí trataremos de ilustrar este hecho por medio de un ejemplo.

Sean tres posibles interpretaciones para tres manifestaciones posibles en un universo de discurso dado. Manifestaciones e interpretaciones están relacionadas según los datos de la tabla que se muestra a continuación, y en la que no se cumple el criterio de independencia:

	INTERPRETACIONES				
MANIFESTACIONES	**sólo A_1**	**A_1 y A_2**	**sólo A_2**	**A_3**	**totales**
sólo m_1	$b1$	$c1$	$d1$	$e1$	$n1$
m_1 y m_2	$b2$	$c2$	$d2$	$e2$	$n2$
sólo m_2	$b3$	$c3$	$d3$	$e3$	$n3$
m_3	$b4$	$c4$	$d4$	$e4$	$n4$
totales	B	C	D	E	N

Tabla 1.2

Trataremos de encontrar $p(A_1/m_1)$, con los datos de la tabla, de dos formas diferentes:

---

[61] Dadas las manifestaciones del problema, claro está.

(a) utilizando el teorema de Bayes
(b) directamente a partir del concepto de probabilidad condicional

Para calcular $p(A_1/m_1)$ a partir del teorema de Bayes necesitamos:

$$p(A_1) = \frac{B + C}{N}$$
$$p(A_2) = \frac{C + D}{N}$$
$$p(A_3) = \frac{E}{N}$$
$$p(m_1 / A_1) = \frac{b1 + b2 + c1 + c2}{B + C}$$
$$p(m_1 / A_2) = \frac{c1 + c2 + d1 + d2}{C + D}$$
$$p(m_1 / A_3) = \frac{e1 + e2}{E}$$

Aplicando directamente la ecuación del teorema de Bayes:

$$p(A_1 / m_1) = \frac{p(m_1 / A_1)\, p(A_1)}{p(m_1 / A_1)\, p(A_1) + p(m_1 / A_2)\, p(A_2) + p(m_1 / A_3)\, p(A_3)}$$

Sustituyendo valores:

$$p(m_1 / A_1)\, p(A_1) = \frac{b1 + b2 + c1 + c2}{N}$$
$$p(m_1 / A_2)\, p(A_2) = \frac{c1 + c2 + d1 + d2}{N}$$
$$p(m_1 / A_3)\, p(A_3) = \frac{e1 + e2}{N}$$

y, en consecuencia:

$$p(A_1 / m_1) = \frac{\dfrac{b1 + b2 + c1 + c2}{N}}{\dfrac{b1 + b2 + c1 + c2}{N} + \dfrac{c1 + c2 + d1 + d2}{N} + \dfrac{e1 + e2}{N}}$$

... operando:

$$p(A_1 / m_1) = \frac{b1 + b2 + c1 + c2}{b1 + b2 + 2c1 + 2c2 + d1 + d2 + e1 + e2} = \frac{b1 + b2 + c1 + c2}{n1 + n2 + c1 + c2}$$

Este resultado es manifiestamente falso. Aplicando el concepto de probabilidad condicional directamente sobre los datos de la tabla obtenemos que:

$$p(A_1 / m_1) = \frac{b1 + b2 + c1 + c2}{n1 + n2}$$

Esta limitación del modelo plantea problemas cuando se pretende su aplicación en dominios del mundo real, en los que los requisitos de independencia casi nunca se cumplen.

Pero ésta no es la única deficiencia del modelo bayesiano. En los problemas interesantes para la aplicación de técnicas de inteligencia artificial, la información suele ir apareciendo progresivamente, secuencialmente y, generalmente, de forma poco ordenada. En estos casos, adecuar la aproximación bayesiana a la interpretación secuencial supone considerar que la información factual aparece incrementalmente y, por lo tanto, habrá que adaptar las ecuaciones correspondientes:

Sea $E1$ el conjunto de toda la información disponible en un momento dado, y sea $S1$ un nuevo dato (i.e., un nuevo elemento de información que acaba de aparecer), entonces E será el nuevo conjunto formado por la información de $E1$ y el nuevo dato $S1$. Con estas consideraciones, la ecuación del teorema de Bayes debe reescribirse del siguiente modo:

$$p(I_i / E) = \frac{p(S1 / I_i \text{ y } E1)\, p(I_i / E1)}{\sum_j p(S1 / I_j \text{ y } E1)\, p(I_j / E1)}$$

En estas condiciones, y dado que la aplicación del modelo bayesiano tiene que estar avalada por un estudio estadístico correcto y fiable, la reformulación de la ecuación de Bayes para su utilización en la interpretación secuencial complica, aún más, la estadística asociada.

Otro problema frecuente de este modelo procede de su aplicación poco cuidadosa. Ilustraremos este tipo de situaciones con un ejemplo, muy exagerado pero ilustrativo:

Es claro que existe una relación entre el cáncer de pulmón y la hemoptisis[62]. Supongamos que a un determinado hospital llega un paciente con hemoptisis. El médico que lo atiende, que es de la escuela bayesiana, trata de cuantificar la probabilidad de que dicha manifestación sea consecuencia de un cáncer de pulmón. Para ello se dirige a los archivos del hospital y observa que de los 15315 pacientes ingresados durante los últimos tres años, 320 padecían cáncer de pulmón. Por otra parte, de esos 15315 pacientes, 231 ingresaron a causa de una hemoptisis. Por último, 150 pacientes de los 320 diagnosticados de cáncer de pulmón manifestaban hemoptisis. Con estos datos, nuestro médico bayesiano realiza el siguiente análisis:

---

[62] Efectivamente, muchos pacientes que padecen esta terrible enfermedad ingresan en el hospital escupiendo sangre por la boca.

$$p(\text{hemoptisis}) = \frac{231}{15315}$$

$$p(\text{cancer de pulmón}) = \frac{320}{15315}$$

$$p(\text{hemoptisis / cancer de pulmón}) = \frac{150}{320}$$

Aplicando Bayes, y dado que el paciente presenta una hemoptisis:

$$p(\text{cancer de pulmón / hemoptisis}) = \frac{\frac{150}{320} \times \frac{320}{15315}}{\frac{231}{15315}} = 0.67$$

El médico, apesadumbrado, le anuncia al paciente:

"- Tengo malas noticias para usted, tiene un 67% de probabilidades de padecer un cáncer de pulmón."

En esto se abre la puerta de la sala y entra una enfermera que le pregunta al médico:

"- Doctor, ¿le sacamos ya al paciente la viga de hierro que lleva incrustada en el pecho?"

Evidentemente, este ejemplo no es más que un esperpento de lo que teóricamente podría suceder cuando un modelo, en este caso el bayesiano, se aplica mal.

El último gran problema del modelo bayesiano es consecuencia de su consistencia matemática. Al respecto, y por definición, siempre se tiene que cumplir que:

$$p(A) + p(\neg A) = 1$$

Sin embargo, cuando tratamos con problemas del mundo real, los expertos difícilmente asumen esta consistencia[63]. Así, si estamos investigando una hipótesis *H*, hasta el momento avalada, por ejemplo, por las observaciones *O*1, *O*2 y *O*3, la aplicación de un esquema bayesiano podría traducirse del siguiente modo:

$$p(H/O1 \text{ y } O2 \text{ y } O3) = x;\ 0 \leq x \leq 1$$

lo que inmediatamente nos conduciría a que:

$$p(\neg H/O1 \text{ y } O2 \text{ y } O3) = 1 - x$$

[63] ... y no olvidemos que estamos tratando de diseñar programas que "razonen" como expertos.

Es decir, que un mismo conjunto de evidencias apoya simultáneamente (aunque en distinto grado), a una hipótesis y a su negación. Este es uno de los puntos más débiles de los modelos estadísticos[64] cuando tratamos con "conocimiento" en lugar de tratar con "datos". En problemas del mundo real, dada una proposición basada en la experiencia, la consistencia matemática no tiene por qué mantenerse. Necesitamos la puesta a punto de un nuevo modelo.

1.5. Resumen

En este tema abordamos por primera vez los problemas de razonamiento desde la perspectiva de la inteligencia artificial. Introducimos el tema identificando distintos tipos de dominios para las cuales unos esquemas de razonamiento son más apropiados que otros. Planteamos a continuación el problema de la interpretación diferencial como uno de los métodos disponibles para formalizar el razonamiento categórico. Según este modelo, el proceso consiste en construir todas las posibles combinaciones entre todas las evidencias posibles y todas las interpretaciones. Ello da lugar a la llamada Base Lógica Expandida que está formada por todos los complejos manifestación-interpretación idealmente posibles. Sobre esta Base Lógica Expandida tendremos que aplicar el conocimiento del dominio para eliminar relaciones entre manifestaciones e interpretaciones que sean absurdas o imposibles. El resultado es una Base Lógica Reducida en la que, previsiblemente, estará la solución de cualquier problema que pudiéramos plantear en el dominio. No obstante, nuestro conocimiento puede no ser completo, o simplemente, nuestras observaciones y evidencias pueden señalar a más de una interpretación. En este último caso, la propia naturaleza del modelo no nos permite discriminar. Esta razón, y la inevitable explosión combinatoria, nos llevan a considerar a los modelos estadísticos como una alternativa potencialmente resolutiva en lo que a los esquemas de razonamiento se refiere. Así, planteamos el esquema bayesiano que utiliza las llamadas probabilidades condicionales -en las que aparecen involucrados dos sucesos, de forma que la ocurrencia del segundo depende de la ocurrencia del primero-. Después de analizar algunas propiedades del modelo bayesiano, aplicamos dicho esquema en la resolución -con éxito- de un problema no resuelto mediante el uso de la aproximación categórica. No obstante, los requisitos de independencia exigidos por el modelo bayesiano, la necesidad de efectuar previamente estudios estadísticos amplios, y el hecho de que, de acuerdo con la consistencia matemática del esquema una misma observación apoye simultáneamente a una hipótesis y a su negación, nos llevan a buscar nuevas soluciones alternativas.

1.6. Textos básicos

- Dude, Hart, Nilsson, "Subjective Bayesian Methods for Rule-Based Inference Systems", National Computer Conference, 1976.

[64] No sólo de los modelos bayesianos.

- Grosof, “Non-monotonicity in Probabilistic Reasoning”, Uncertainty in Artificial Intelligence, vol.2, 1988.

- Ledley, Lusted, “Reasoning Foundations in Medical Diagnosis”, Science, vol.130, 1959.

1. RAZONAMIENTO PROBABILÍSTICO

En el capítulo anterior veíamos como el modelo de razonamiento categórico podía ser corregido a través del teorema de Bayes, evitando de esta forma el uso de interpretaciones categóricas poco frecuentes en el mundo real. Sin embargo, los esquemas probabilísticos también pueden utilizarse por sí mismos como modelos de razonamiento, sin aparecer como una corrección del modelo categórico.

En este capítulo presentaremos dos tipos de modelos de razonamiento probabilístico: el modelo bayesiano y el modelo basado en redes de creencia, y comentaremos sus ventajas e inconvenientes.

1.1. Elementos Básicos del Modelo Bayesiano

Los esquemas básicos del modelo de razonamiento bayesiano ya han sido introducidos en el capítulo anterior. En este apartado veremos de nuevo estos conceptos desde una óptica más formal y veremos cómo el modelo bayesiano puede utilizarse como modelo de razonamiento dentro de los sistemas expertos.

En primer lugar examinaremos algunos conceptos fundamentales de la teoría de las probabilidades. Sea A un evento, la colección de todos los posibles eventos elementales Ω se conoce como espacio de eventos. La probabilidad de un evento A se denota como $p(A)$, y toda función de probabilidad p debe satisfacer tres axiomas:

(1) La probabilidad de cualquier evento elemental A es no negativa, es decir, $\forall A \in \Omega : p(A) \geq 0$.

(2) La probabilidad del espacio de eventos es uno, es decir, $p(\Omega) = 1$.

(3) Si k eventos $A_1, A_2, ..., A_k$ son mutuamente excluyentes (es decir, no pueden ocurrir simultáneamente), entonces la probabilidad de que al menos uno de esos eventos ocurra es la suma de las probabilidades individuales $p(A_1 \cup A_2 \cup ... \cup A_k) = \sum_{i=1}^{k} p(A_i)$

A partir de estos axiomas podemos deducir otros. Así, de los axiomas 1 y 2 podemos obtener el siguiente resultado:

$$\forall A \in \Omega : 0 \leq p(A) \leq 1$$

Esta ecuación muestra que la probabilidad de cualquier evento se encuentra entre cero y uno. Por definición si $p(A)=0$ el evento A nunca ocurre, y cuando $p(A)=1$ el evento A debe ocurrir. El complementario de A (representado como $\neg A$) contiene la colección de todos los eventos elementales en Ω excepto A. Como A y $\neg A$ son mutuamente excluyentes y $A \cup \neg A = \Omega$, por el axioma 3 podemos obtener el siguiente resultado:

$$p(A) + p(\neg A) = p(A \cup \neg A) = p(\Omega) = 1$$

Reescribiendo esta ecuación como $p(\neg A) = 1\text{-}p(A)$, obtenemos un método sencillo para calcular $p(\neg A)$ a partir de $p(A)$.

Si suponemos que $B \in \Omega$ es otro evento, la probabilidad de que A ocurra sabiendo que B ocurre, representado por $p(A/B)$ se conoce como la probabilidad condicional de A dado B. Como vimos en el apartado anterior esta probabilidad se define como:

$$p(A/B) = \frac{p(A \cap B)}{p(B)} \tag{1.1}$$

A partir de esta ecuación es sencillo obtener la regla de Bayes, base de los modelos bayesianos, de la siguiente forma: la probabilidad de B dado A es $p(\text{B/A}) = p(B \cap A)\ p(A)$, como la probabilidad conjunta es conmutativa entonces $p(A \cap B) = p(B \cap A)$, con lo que podemos escribir $p(B \cap A) = p(A \cap B) = p(B/A)\ p(A)$. Substituyendo esta expresión en la definición de probabilidad condicionada de A dado B podemos obtener la regla de Bayes:

$$p(A/B) = \frac{p(B/A)\,p(A)}{p(B)} \tag{1.2}$$

La ecuación básica de Bayes puede generalizarse teniendo en cuenta que

$$\begin{aligned} p(B) &= p((B \cap A) \cup (B \cap \neg A)) \\ &= p(B \cap A) + p(B \cap \neg A) \\ &= p(B/A)\ p(A) + p(B/\neg A)\ p(\neg A) \end{aligned}$$

Sustituyendo esta expresión de $p(B)$ en la regla de Bayes podemos obtener la siguiente ecuación:

$$p(A/B) = \frac{p(B/A)\,p(A)}{p(B/A)\,p(A) + p(B/\neg A)\,p(\neg A)} \tag{1.3}$$

que nos permite evitar valorar la probabilidad a priori p(B) a cambio de valorar la probabilidad condicional p(B/¬A). Como habíamos visto en el capítulo esta expresión puede generalizarse como:

$$p(A_0/B) = \frac{p(B/A_0)\,p(A_0)}{\sum_i p(B/A_i)\,p(A_i)} \tag{1.4}$$

1.2. Aplicación Elemental de la Regla de Bayes

Aparentemente, la regla básica de Bayes no parece ser de mucha utilidad. Se necesitan tres términos (una probabilidad condicional y dos probabilidades incondicionales) sólo para poder calcular una probabilidad condicional. Sin embargo, la regla de Bayes es útil en la práctica gracias a que hay muchos casos en los que se dispone de buenas estimaciones de probabilidad para estos tres números y es necesario calcular el cuarto.

Un ejemplo sencillo de la aplicación de la regla básica de Bayes se puede encontrar en el campo del diagnóstico médico. Supongamos que conocemos la probabilidad a priori (o prevalencia) de una determinada enfermedad *E*. También conocemos la probabilidad a priori de aparición de un determinado síntoma *S* y la probabilidad condicionada de *S* dado *E*. En tal caso, si nos aparece por la consulta un paciente con el síntoma *S*, la regla de Bayes nos sirve para obtener la probabilidad de que sufra de la enfermedad *E*.

$$p(E/S) = \frac{p(S/E)\,p(E)}{p(S)}$$

Sin embargo una pregunta que, probablemente, le ha surgido a todo aquel que analiza por primera vez la regla de Bayes es ¿por qué la probabilidad $p(E/S)$ es necesario calcularla y la probabilidad $p(S/E)$ se dispone de antemano? ¿No podría obtenerse también de antemano la probabilidad $p(E/S)$? La respuesta a estas preguntas es que el conocimiento obtenido por diagnóstico es más débil que el conocimiento causal.

La probabilidad $p(S/E)$ es información causal, es decir, es la probabilidad de que una determinada enfermedad cause un determinado síntoma. Estas probabilidades suelen ser bastante bien conocidas y no se ven afectadas por condicionantes externos.

Sin embargo, la probabilidad de tener una determinada enfermedad sabiendo que se tiene un determinado síntoma, $p(E/S)$, no es tan fácil de calcular ya que depende de condicionantes externos (por ejemplo, una epidemia de la enfermedad *E*). La regla de Bayes nos permite obtener la probabilidad $p(E/S)$ a través de una probabilidad condicional que puede ser fácil de fijar, $p(E/S)$, y dos probabilidades a priori, $p(E)$ y $p(S)$, que reflejan la relación del domino con la enfermedad y el síntoma.

1.3. El Modelo Bayesiano en Sistemas Expertos

La regla de Bayes expresada como la ecuación (1.4) es la base para la utilización de la teoría de la probabilidad en el manejo de la incertidumbre.

La tarea de un sistema experto es deducir qué hipótesis H es cierta dado un determinado conjunto de evidencias E. La ecuación (1.4), reescrita ahora en base a *H* y *E*, representa el caso en el que existen varias hipótesis pero una sola evidencia.

$$p(H_0 / E) = \frac{p(E / H_0)\, p(H_0)}{\sum_i p(E / H_i)\, p(H_i)} \tag{1.5}$$

Como vemos la regla de Bayes nos permite obtener la probabilidad de una hipótesis dada, en base a las probabilidades condicionales de observar una evidencia dada una hipótesis, *p*(*E*/*H*), y a las probabilidades a priori de todas las hipótesis, *p*(*H*). Estas probabilidades deberán residir en la base de conocimientos del sistema experto.

Sin embargo, la ecuación (1.5) debe generalizarse para contemplar la posibilidad de aparición de múltiples evidencias:

$$p(H_0 / E_1, E_2, \ldots, E_n) = \frac{p(E_1, E_2, \ldots, E_n / H_0)\, p(H_0)}{\sum_i p(E_1, E_2, \ldots, E_n / H_i)\, p(H_i)} \tag{1.6}$$

El problema con esta nueva versión de la regla de Bayes es la explosión combinatoria que se produce a medida que el número de evidencias crece. Así el número de probabilidades condicionales que debemos almacenar es igual a (nº H) × $2^{(\text{nºE})}$ siempre teniendo en cuenta que las evidencias sólo pueden tomar dos valores, ser ciertas o falsas. Así para un sencillo problema en el que 10 evidencias apuntan a 3 posibles hipótesis nos encontramos con que el número de probabilidades condicionales que tenemos que almacenar es de $3 \times 2^{10} = 3 \times 1024 = 3072$, un número claramente elevado para un problema claramente sencillo.

La solución al problema de la explosión combinatoria consiste en suponer que las evidencias son condicionalmente independientes. Dos evidencias E_1 y E_2 son condicionalmene independientes cuando su probabilidad conjunta dada una determinada hipótesis *H* equivale al producto de las probabilidades condicionales de cada evento dado *H*, o lo que es lo mismo $p(E1, E2/H) = p(E_1/H)\ p(E_2/H)$. De esta forma podemos reescribir la ecuación (1.6) como:

$$p(H_0 / E_1, E_2, \ldots, E_n) = \frac{p(E_1 / H_0)\ p(E_2 / H_0) \cdots p(E_n / H_0)\ p(H_0)}{\sum_i p(E_1 / H_i)\ p(E_2 / H_i) \cdots p(E_n / H_i)\ p(H_i)} \tag{1.7}$$

Suponiendo independencia condicional, el número de probabilidades condicionales a almacenar pasa a ser (nº H) × (nº E). Para el ejemplo con 10 evidencias y 3 hipótesis el resultado es 30. Número claramente inferior al resultado de 3072 obtenido suponiendo dependencia entre la hipótesis.

Para ilustrar la aplicación de la regla de Bayes podemos partir del siguiente ejemplo. Supongamos que tenemos un conjunto exhaustivo formado por tres hipótesis mutuamente excluyentes (H_1, H_2 y H_3) y dos evidencias independientes (E_1 y E_2). Las probabilidades condicionadas y a priori se muestran en la Tabla 1.1:

	i=1	i=2	i=3
$p(H_i)$	0.5	0.3	0.2
$p(E_1 / H_i)$	0.4	0.8	0.3
$p(E_2 / H_i)$	0.7	0.9	0.0

Tabla 1.1

A priori, la hipótesis que más probabilidad tiene de ser cierta es H_1. Sin embargo, a medida que van apareciendo evidencias, la creencia en las hipótesis aumentará o disminuirá consecuentemente. Por ejemplo imaginemos que observamos la evidencia E_1, si calculamos la probabilidad a posteriori de las hipótesis basándonos en la ecuación (1.7) obtenemos los siguientes valores:

$$p(H_1 / E_1) = \frac{(0.4 \times 0.5)}{(0.4 \times 0.5) + (0.8 \times 0.3) + (0.3 \times 0.2)} = 0.40$$

$$p(H_2 / E_1) = \frac{(0.8 \times 0.3)}{(0.4 \times 0.5) + (0.8 \times 0.3) + (0.3 \times 0.2)} = 0.48$$

$$p(H_3 / E_1) = \frac{(0.3 \times 0.2)}{(0.4 \times 0.5) + (0.8 \times 0.3) + (0.3 \times 0.2)} = 0.12$$

Como vemos, después de la aparición de la evidencia E_1 la creencia en las hipótesis H_1 y H_3 ha decrecido mientras que la creencia en H_2 ha aumentado. Si observamos ahora la evidencia E_2 debemos volver a calcular las probabilidades a posteriori:

$$p(H_1 / E_1 E_2) = \frac{(0.4 \times 0.7 \times 0.5)}{(0.4 \times 0.7 \times 0.5) + (0.8 \times 0.9 \times 0.3) + (0.3 \times 0.0 \times 0.2)} = 0.39$$

$$p(H_2 / E_1 E_2) = \frac{(0.8 \times 0.9 \times 0.3)}{(0.4 \times 0.7 \times 0.5) + (0.8 \times 0.9 \times 0.3) + (0.3 \times 0.0 \times 0.2)} = 0.61$$

$$p(H_1 / E_1 E_2) = \frac{(0.3 \times 0.0 \times 0.2)}{(0.4 \times 0.7 \times 0.5) + (0.8 \times 0.9 \times 0.3) + (0.3 \times 0.0 \times 0.2)} = 0.00$$

Así, después de la aparición de las evidencias E_1 y E_2 sólo son relevantes dos hipótesis, H_1 y H_2, siendo H_2 la más probable.

1.4. Ventajas e Inconvenientes del Modelo Bayesiano

La principal ventaja de los métodos bayesianos reside en que están fuertemente fundados en la teoría de la probabilidad, sin embargo su principal dificultad estriba en la gran cantidad de probabilidades que es necesario obtener para construir una base de conocimientos.

Así, aun suponiendo hipótesis mutuamente excluyentes, evidencias condicionalmente independientes y variables restringidas a dos valores (verdadero y falso); si un diagnóstico médico implica a 50 enfermedades posibles basadas en 100 síntomas es necesario disponer de 5000 probabilidades condicionales y 50 probabilidades a priori.

Desafortunadamente, la suposición de independencia condicional raramente es válida y la suposición de mutua exclusividad y exhaustividad de las hipótesis suele ser falsa, siendo lo más corriente la aparición de hipótesis concurrentes y superpuestas. Además, como veíamos en el capítulo 7, los métodos bayesianos no permiten una explicación clara de sus conclusiones y permiten que una misma evidencia apoye, al mismo tiempo, a una hipótesis y a su negación.

Una aproximación para resolver estos problemas son las redes de creencia (belief networks). Una red de creencia es un tipo especial de diagrama de influencia en el que los nodos representan variables aleatorias. Pearl (1988) demostró que el uso de redes de creencia permite construir bases de conocimiento probabilisticas consistentes, sin imponer innecesarias asunciones de independencia condicional. Estas redes también aseguran que la evidencia a favor de una hipótesis no será construida por soporte parcial de su negación, y que explicaciones consistentes pueden ser obtenidas mediante el rastreo de las creencias hasta los puntos iniciales de la red.

1.5. Introducción a la Teoría de Grafos

Antes de entrar en detalle con los modelos de redes de creencia vamos a realizar una breve introducción a la teoría de grafos.

Un grafo (o red) G se compone de un par de conjuntos $G = (X, L)$ donde $X = \{X_1, X_2, ..., X_n\}$ es un conjunto finito de elementos (nodos) y L es un conjunto de aristas, es decir, un subconjunto de pares ordenados de elementos distintos de X. Así, si una arista une a los nodos X_i y X_j se denotará mediante L_{ij}.

Se dice que la arista que une los nodos X_i y X_j es dirigida cuando $L_{ij} \in L$ y $L_{ji} \notin L$, y se denota como $X_i \rightarrow X_j$. Por otro lado la arista entre los nodos X_i y X_j es no dirigida cuando $L_{ij} \in L$ y $L_{ji} \in L$ y se representa como X_i — X_j. Un grafo en el cual todas las aristas son dirigidas se denomina grafo dirigido, y un grafo en el que todas sus aristas son no dirigidas se denomina grafo no dirigido.

Un camino que va del nodo X_i al nodo X_j es una sucesión de nodos $(X_{i1}, ..., X_{ir})$, comenzando en $X_{i1} = X_i$ y finalizando en $X_{ir} = X_j$, de forma que existe una arista del

nodo X_{ik} al nodo X_{ik+1} desde $k = 1$ hasta $k = r$-1. Un camino es cerrado si el nodo inicial coincide con el final. Un bucle es un camino cerrado en un grafo no dirigido, y un ciclo es un camino cerrado en un grafo dirigido.

Los gráficos no dirigidos pueden dividirse en varias subclases tal y como se muestra en la Figura 1.1. Un grafo no dirigido se denomina conexo si existe al menos un camino entre cada par de nodos. En caso contrario se denomina inconexo. Un árbol es un grafo conexo en el que existe un único camino entre cada par de nodos (es decir, no tiene bucles), mientras que un grafo se denomina múltiplemente conexo si contiene al menos un par de nodos que estén unidos por más de un camino (o equivalentemente, si contiene al menos un bucle).

Los grafos dirigidos también se dividen en varias subclases, tal y como se muestra en la

Figura 1.2. Un grafo dirigido se considera conexo si el grafo no dirigido asociado (el que se obtiene al sustituir cada arista dirigida del grafo por la correspondiente arista no dirigida) también es conexo. Un grafo dirigido se denomina cíclico si contiene al menos un ciclo. Un grafo dirigido conexo se denomina árbol si el grafo no dirigido asociado es un árbol, en caso contrario se denomina múltiplemente conexo. Un árbol simple es aquel en el que cada nodo tiene como máximo un padre, en caso contrario se denomina poliárbol.

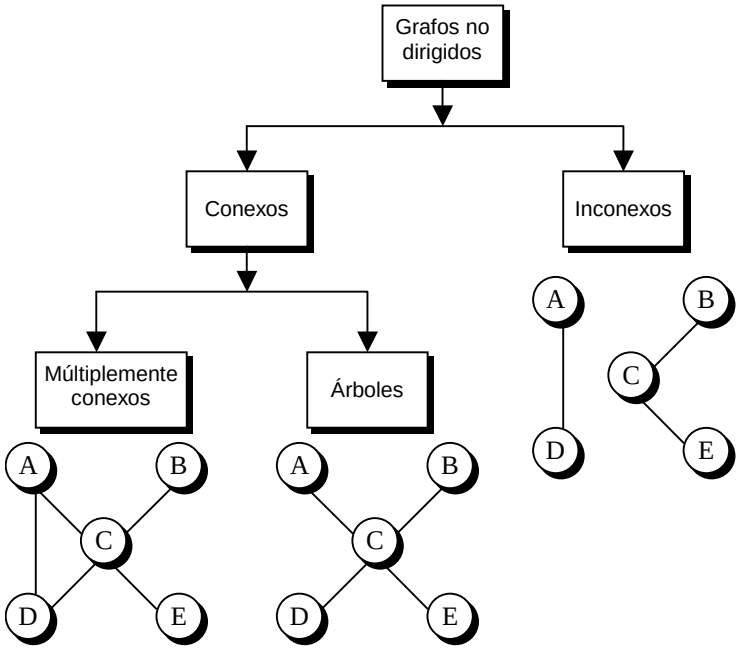

Figura 1.1 Tipos de grafos no dirigidos

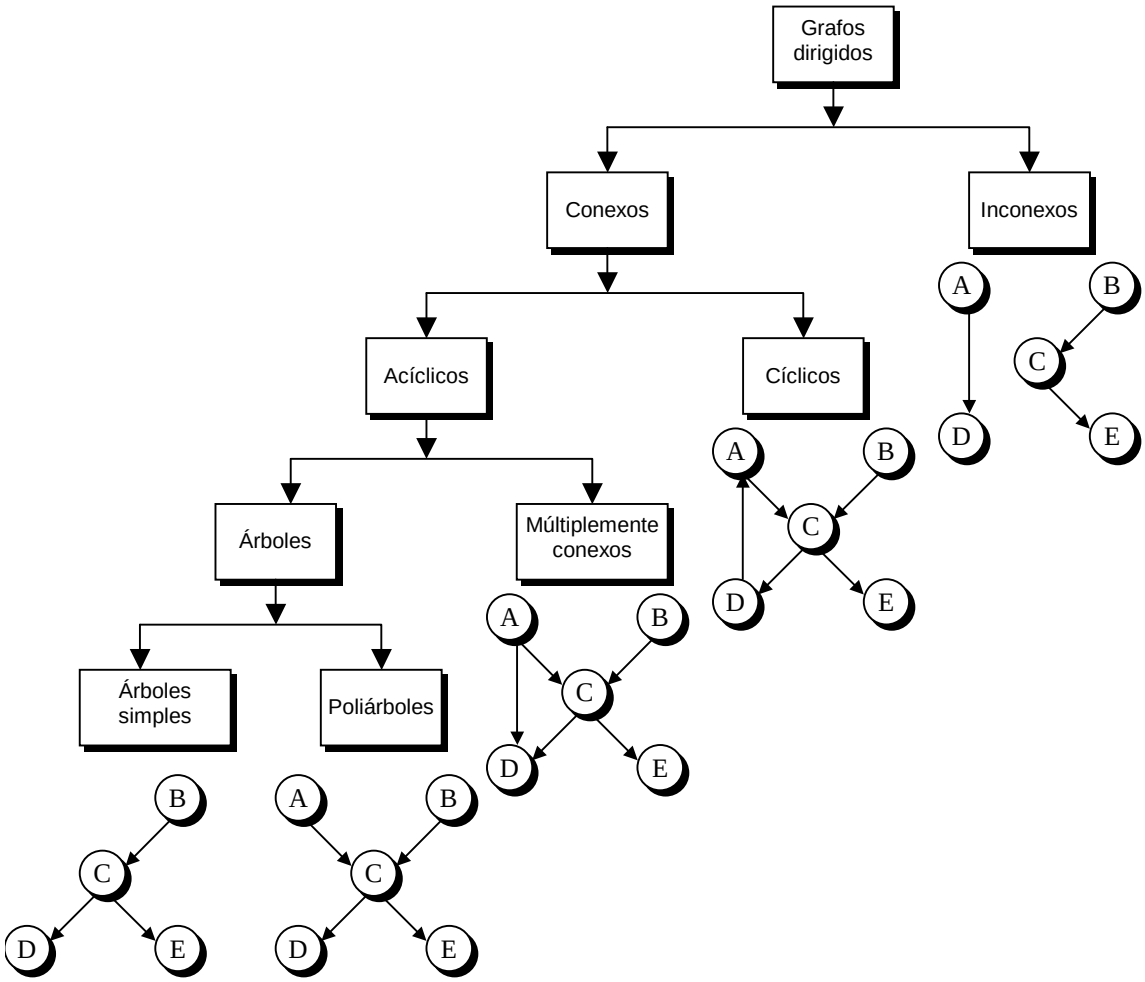

Figura 1.2 Tipos de grafos dirigidos

1.6. Elementos Fundamentales de las Redes de Creencia

Las personas solemos tener dificultades en estimar los valores de las probabilidades condicionales $p(x_i/x_j)$, sin embargo no tenemos tantos problemas en determinar si dos proposiciones x_i y x_j son dependientes o independientes. De la misma forma, las personas también tienden a evaluar con claridad, convicción y consistencia las relaciones de dependencia a tres, es decir, x_i influye sobre x_j a través de x_k. Esto sugiere que las nociones de dependencia e independencia condicional son más básicas para el razonamiento humano que los valores numéricos asociados a los juicios de probabilidad.

La naturaleza de las dependencias probabilísticas entre proposiciones se estructura en forma de grafos. Esta metáfora gráfica sugiere que la estructura fundamental del conocimiento humano puede ser representada por grafos de dependencia, y que un rastreo en esos grafos puede utilizarse para interrogar o actualizar ese conocimiento.

La idea central que subyace bajo los modelos gráficos es la representación de las relaciones de dependencia a través de grafos en los que los nodos son variables aleatorias y los arcos reflejan la estructura de las relaciones de dependencia.

Los modelos gráficos se clasifican en dos categorías: grafos dirigidos acíclicos (Directed Acyclic Graphs, DAGs) y grafos no dirigidos (Undirected Graphs, UG). También suelen utilizarse modelos mixtos que comparten características con ambas categorías. Los grafos dirigidos acíclicos son usados fundamentalmente en inteligencia artificial y en estadística, donde las relaciones causa-efecto cobran importancia fundamental. Estas relaciones pueden hacerse explícitas mediante el uso de arcos dirigidos en el grafo. Los grafos dirigidos acíclicos son a menudo referenciados como redes bayesianas, redes de creencia, modelos gráficos recursivos y, menos frecuentemente, como redes causales, redes de Markov dirigidas, y redes probabilísticas.

Los grafos no dirigidos son populares en la física estadística, y en el procesado de imágenes, en donde la asociación entre las variables es considerada una correlación y no una causalidad. Los grafos no dirigidos suelen referirse en la literatura como mapas aleatorios de Markov, redes de Markov, máquinas de Boltzmann y modelos log-lineales.

Redes de creencia como grafos dirigidos acíclicos

En este apartado nos ocuparemos de las redes de creencia, que están basadas en grafos dirigidos acíclicos, ya que son las de más común aplicación en el campo de la inteligencia artificial. Las redes de creencia se componen de los siguientes elementos:

(1) Nodos.- que representan proposiciones o variables aleatorias y que representaremos como $\{x_1, x_2, ..., x_n\}$.

(2) Arcos dirigidos.- que representan la existencia de influencias causales directas entre las proposiciones que une. El significado implícito de una flecha que vaya del nodo xi al nodo x_j es el de que x_i ejerce una influencia directa sobre x_j (se suele decir que xi es un nodo padre de x_j).

(3) Tablas de probabilidad condicional.- que miden la potencia de las influencias mediante el uso de probabilidades condicionadas $p(x_i/\text{Padres}(x_j))$. Por cada nodo hay una tabla de probabilidad condicional que sirve para cuantificar los efectos de los padres sobre dicho nodo.

Una típica red bayesiana sería la que se muestra en la Figura 1.3. Esta red representa la situación en la que una serie de circustancias condicionantes (representadas como C_1 y C_2 y que podrían ser la edad, el clima, la presencia de ciertas enfermedades, etc.) influyen en la aparición de una determinada enfermedad E. Esta enfermedad a su vez es causa de distintos síntomas S_1 y S_2.

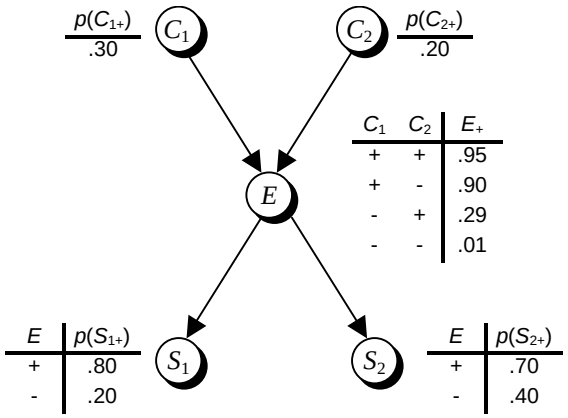

Figura 1.3 Ejemplo de red bayesiana

Como vemos, cada nodo de la red representan un conjunto de proposiciones exhaustivas y mutuamente excluyentes. Así, el nodo C_1 representa a las proposiciones C_{1+} y C_{1-}, que representan respectivamente "C_1 = True" y "C_1 = False". Cada nodo tiene una distribución de probabilidad por cada una de las combinaciones de sus nodos condicionantes, E está condicionada por C_1 y C_2 por ello hay cuatro posibles distribuciones de probabilidad de E, que corresponden con las cuatro posibles combinaciones de C_1 y C_2. Se representa sólo la probabilidad de E_+, ya que la probabilidad de E_- puede obtenerse como $1 - p(E_+)$. Todo esto siempre y cuando las proposiciones sean binarias, es decir, que sólo tomen dos posibles valores. También es posible la existencia de proposiciones multivaluadas y proposiciones que toman valores continuos. Las variables que toman valores continuos pueden tener funciones de densidad de probabilidad como, por ejemplo, la curva o campana de Gauss. De esta forma en vez de tener una tabla de probabilidades en el nodo tendríamos los parámetros (media y varianza) que caracterizarían a la curva gaussiana. De todas formas lo más corriente suele ser discretizar los valores continuos mediante intervalos, por ejemplo la edad podría tomar los valores "<18", "18 – 65" y ">65".

La ausencia de arcos refleja aserciones de independencia condicional. Así si no hay ningún arco entre C_1 y C_2 quiere decir que la probabilidad de C_2 no depende de C_1. De la misma forma S_1 no depende directamente de C_1, sino que están conectados a través del nodo E. En este caso se dice que S_1 es independiente de C_1 dado E. En este contexto Pearl describe con exactitud la semántica de la ausencia de arcos y su relación con la independencia mediante el concepto de separación dependiente de la dirección o d-separación.

Separación dependiente de la dirección (d-separación)

El concepto de d-separación es un criterio gráfico que se utiliza para saber qué relaciones de independencia condicional están contenidas en un grafo, y se define de la siguiente manera:

Sean X, Y, y Z tres subconjuntos disjuntos de nodos en un grafo dirigido acíclico D; entonces se dice que Z d-separa X e Y si y sólo si, a lo largo de todo el camino no

dirigido entre cualquier nodo de *X* y cualquier nodo de *Y*, existe un nodo intermedio *A* tal que, o bien:

- *A* no es un nodo de aristas convergentes en el camino y *A* está en *Z*, o bien
- *A* es un nodo de aristas convergentes en el camino y ni *A* ni sus descendientes está en *Z*.

Cuando *Z* d-separa *X* e *Y*, se escribe *I* (*X*, *Y* / *Z*) para indicar que *X* e *Y* son condicionalmente independientes dado *Z*. Esto expresado gráficamente sería de la siguiente forma:

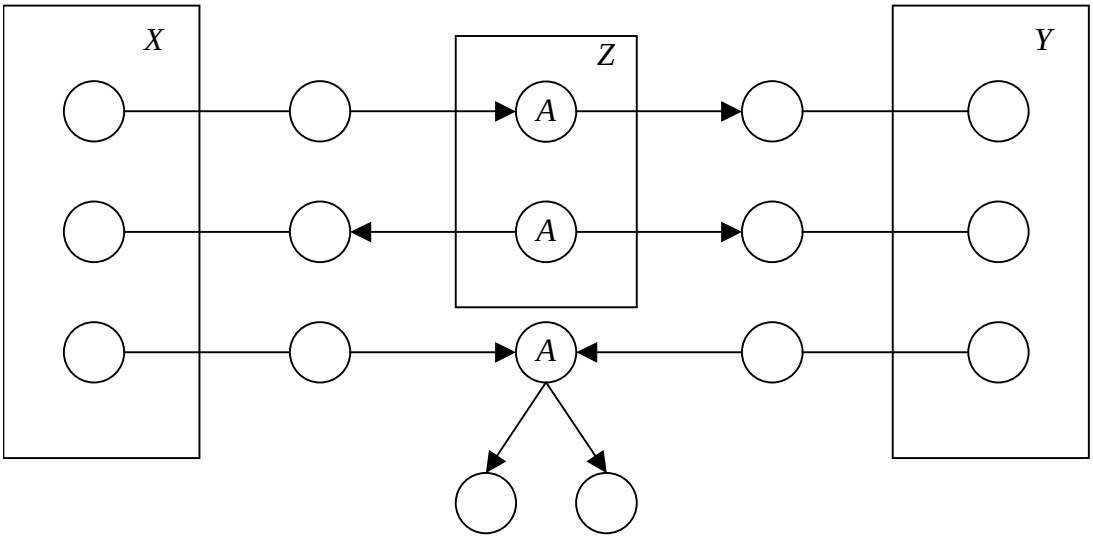

Figura 1.4 Criterio de d-separación, los dos primeros casos corresponden a la primera definición de d-separación y el tercero a la segunda definición.

Así, por ejemplo, en la red de la Figura 1.3 vemos como los nodos C_1 y C_2 son condicionalmente independientes cuando no se conoce ninguna evidencia, esto es, *I* (C_1, C_2 / ∅). Esto es debido a que en el camino no dirigido que los une ($C_1 - E - C_2$) existe un nodo convergente (*E*) que no pertenece al subconjunto de referencia (que en este caso es el conjunto vacío).

Sin embargo, si el valor de *E* está ejemplificado (es decir, posee un valor), C_1 y C_2 son dependientes dado *E*, expresado como *D* (C_1, C_2 / *E*), ya que no se cumple ninguna de las condiciones de d-separación. Para entender esto un poco mejor podemos suponer que C_1 indica "drogadicción", C_2 indica "hemofilia" y *E* indica "sida". Si sabemos que el paciente tiene sida, el hecho de que no sea hemofílico hace que aumente la probabilidad de que sea drogadicto.

Distribución de probabilidad conjunta

Existe una regla en estadística, conocida como la regla de la cadena o *chain-rule*, que dice que: dada una serie de variables aleatorias {x_1, x_2, ..., x_n} ordenadas

arbitrariamente, podemos calcular la probabilidad conjunta de las variables, $p(x_1, x_2, ..., x_n)$, como:

$$p(x_1, x_2, ..., x_n) = p(x_n / x_{n-1}, ..., x_1) \ldots p(x_3 / x_2, x_1)\, p(x_2 / x_1)\, p(x_1)$$

Podemos aplicar esta regla para calcular la distribución de probabilidad conjunta de las redes bayesianas. Así, si utilizamos la red de la Figura 1.3 ordenando los nodos de padres a hijos (C_1, C_2, E, S_1 y S_2) obtenemos:

$$p(C_1, C_2, E, S_1, S_2) = p(S_2 / S_1, E, C_2, C_1)\, p(S_1 / E, C_2, C_1)\, p(E / C_2, C_1)\, p(C_2 / C_1)\, p(C_1)$$

Sin embargo, esta ecuación puede simplificarse si tenemos en cuenta que el conjunto de nodos padre de un determinado nodo x_i, representado como $Padres(x_i)$ "protege" a x_i de la influencia de sus otros predecesores. Es decir, si tenemos un conjunto de nodos $\{y_1, ..., y_n\}$ predecesores de x_i, pero que no pertenecen a $Padres(x_i)$, entonces

$$p(x_i / Padres(x_i), y_1, ..., y_n) = p(x_i / Padres(x_i))$$

De esta forma podemos expresar la distribución de probabilidad conjunta de la red de creencia como:

$$p(x_1,..., x_n) = \prod_{i=1}^{n} p(x_i / Padres(x_i))$$

Así la expresión $p(C_1, C_2, E, S_1, S_2)$ calculada anteriormente puede simplificarse ahora de la siguiente forma:

$$p(C_1, C_2, E, S_1, S_2) = p(S_2 / E)\, p(S_1 / E)\, p(E / C_2, C_1)\, p(C_2)\, p(C_1)$$

Por ejemplo, si queremos conocer la probabilidad de la siguiente situación C_1 = True, C_2 = False, E = True, S_1 = True y S_2 = False, deberemos calcular

$$\begin{aligned} p(C_{1+}, C_{2-}, E_+, S_{1+}, S_{2-}) &= p(S_{1+}/E_+)\, p(S_{2-}/E_+)\, p(E+/C_{1+}, C_{2-})\, p(C_{1+})\, p(C_{2-}) \\ &= 0.80 \cdot 0.30 \cdot 0.90 \cdot 0.30 \cdot 0.80 \\ &= 0.052 \end{aligned}$$

De este modo podemos calcular todos los posibles valores de probabilidad conjunta, que en este caso suman $2^5 = 32$.

1.7. Inferencia en Redes de Creencia

La forma más simple de inferencia que puede aparecer en una red de creencia es aquella que se refiere al cálculo de las probabilidades $p(Q=q_0)$ para una determinada variable de consulta Q y un determinado valor q_0. Sin embargo la frase "inferencia en redes de creencia" generalmente se refiere al cálculo de las probabilidades $p(Q=q_0 /E=e_0)$, en donde Q es la variable de consulta, E es una lista de evidencias observadas y e_0 son sus correspondientes valores.

Según la posición de Q y E en la red de creencia podemos distinguir cuatro tipos de inferencia (Figura 1.5):

(1) **Inferencias por diagnóstico**. Que van de las evidencias a las causas, por ejemplo siguiendo con el ejemplo de la Figura 1.3, $p(E/S_1)$. También se conoce como razonamiento abductivo ya que intenta encontrar las causas que mejor explican los efectos observados.

(2) **Inferencias causales**. Que van de las causas a los efectos, $p(S_1/E)$. También se conoce como razonamiento deductivo o razonamiento predictivo cuando se refiere a efectos que se van a manifestar en el futuro.

(3) **Inferencias intercausales**. Cuando las inferencias se realizan entre las causas de un efecto común. Así hemos visto antes como cambiaba la probabilidad $p(C_1/E)$ cuando se añade una nueva evidencia que también apunta a E, $p(S_1/E, C_2)$, es decir, sabiendo que se tiene sida y hemofilia la probabilidad de drogadicción disminuye. Es lo que se conoce como "explaining away" o disculparse dando explicaciones.

(4) **Inferencias mixtas**. Son una combinación de una o varias de las inferencias anteriores. $p(E/S_1, C_1)$.

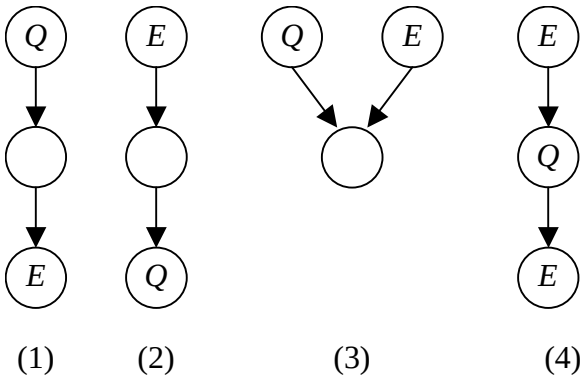

Figura 1.5 Tipos de inferencia (1) diagnóstico, (2) causal, (3) intercausal y (4) mixta.

Como podemos ver, las redes de creencia permiten realizar inferencias en una dirección contraria a la dirección en la que las distintas influencias fueron evaluadas.

En las redes de creencia existen distintos métodos para calcular las inferencias que, generalmente, suelen agruparse en dos clases. Por un lado tenemos los **métodos exactos**. Un algoritmo se denomina exacto si calcula las probabilidades de los nodos sin otro error que el resultante del redondeo producido por las limitaciones de cálculo del ordenador. Por otro lado tenemos los **métodos aproximados**, que utilizan distintas técnicas de simulación para obtener valores aproximados de las probabilidades. Los métodos aproximados se utilizan en aquellos casos en los que los algoritmos exactos no son aplicables, o son computacionalmente costosos.

Métodos exactos de inferencia

La distribución de probabilidad conjunta puede utilizarse para computar cualquier probabilidad de interés de la red. Así, por ejemplo, supongamos que estamos interesados en conocer la probabilidad a priori de padecer la enfermedad *E*, es decir, estamos interesados en conocer la probabilidad $p(E_+)$. Esta probabilidad puede ser obtenida de la distribución de probabilidad conjunta de la siguiente forma:

$$p(E_+) = \sum_{C_{1i}, C_{2i}, S_{1i}, S_{2i}} p(C_{1i}, C_{2i}, E_+, S_{1i}, S_{2i})$$

Si lo que queremos es conocer la probabilidad de padecer la enfermedad *E* sabiendo que tenemos los síntomas S_1 y S_2, es decir, estamos interesados en conocer la probabilidad $p(E_+/S_{1+}, S_{2+})$. La propia definición de probabilidad condicional nos permite decir que $p(E_+/S_{1+}, S_{2+}) = p(E_+, S_{1+}, S_{2+}) / p(S_{1+}, S_{2+})$ y estas últimas probabilidades pueden ser obtenidas también obtenidas de la distribución de probabilidad conjunta:

$$p(E_+ / S_{1+}, S_{2+}) = \frac{p(E_+, S_{1+}, S_{2+})}{p(S_{1+}, S_{2+})} = \frac{\sum_{C_{1i}, C_{2i}} p(C_{1i}, C_{2i}, E_+, S_{1+}, S_{2+})}{\sum_{C_{1i}, C_{2i}, E_i} p(C_{1i}, C_{2i}, E_i, S_{1+}, S_{2+})}$$

Evidentemente este método de “fuerza bruta” es muy poco eficiente ya que a medida que crezca la red el número de cálculos a realizar aumentará de forma exponencial. Sin embargo una forma más eficiente de calcular estas probabilidades es utilizar la estructura de independencia contenida en la propia red de creencia. Existen muchos algoritmos para hacer esto pero sería muy extenso entrar en profundidad en los mismos, en vez de eso mostraremos un ejemplo sencillo de cómo los cálculos pueden reducirse drásticamente utilizando la estructura de independencia de la red, y citaremos los principales algoritmos de inferencia desarrollados.

Habíamos visto con antelación como la función de probabilidad conjunta puede expresarse de forma factorizada atendiendo a las relaciones de independencia entre nodos. Para el caso del cálculo de $p(E_+)$ obtenemos la siguiente expresión:

$$p(E_+) = \sum_{C_{1i}, C_{2i}, S_{1i}, S_{2i}} p(S_{1i} / E_+) p(S_{2i} / E_+) p(E_+ / C_{1i}, C_{2i}) p(C_{1i}) p(C_{2i})$$

El número de operaciones en esta nueva expresión puede ser reducido agrupando los términos dentro del sumatorio de la siguiente forma:

$$p(E_+) = \sum_{S_{1i}, S_{2i}} p(S_{1i} / E_+) p(S_{2i} / E_+) \sum_{C_{1i}, C_{2i}} p(E_+ / C_{1i}, C_{2i}) p(C_{1i}) p(C_{2i})$$

De esta forma podemos calcular cada una de las sumas por separado y reducir en número de sumas de $2^4 = 16$ a $(2^2 + 2^2) = 8$. Puede obtenerse una reducción adicional reordenando los términos dentro de los sumatorios de la siguiente forma:

$$p(E_+) = \sum_{S_{1i}} \left[p(S_{1i} / E_+) \sum_{S_{2i}} p(S_{2i} / E_+) \right] \sum_{C_{1i}} \left[p(C_{1i}) \sum_{C_{2i}} \left[p(E_+ / C_{1i}, C_{2i}) p(C_{2i}) \right] \right]$$

Para el caso de esta red en concreto es fácil comprobar que $\sum_{S_{1i}, S_{2i}} p(S_{1i} / E_+) p(S_{2i} / E_+) = 1$ con lo que el cálculo de $p(E_+)$ se reduce a

$$p(E_+) = \sum_{C_{1i}} \left[p(C_{1i}) \sum_{C_{2i}} \left[p(E_+ / C_{1i}, C_{2i}) p(C_{2i}) \right] \right] =$$

$$= p(C_{1+}) \left[p(E_+ / C_{1+}, C_{2+}) p(C_{2+}) + p(E_+ / C_{1+}, C_{2-}) p(C_{2-}) \right] +$$
$$+ p(C_{1-}) \left[p(E_+ / C_{1-}, C_{2+}) p(C_{2+}) + p(E_+ / C_{1-}, C_{2-}) p(C_{2-}) \right] =$$

$$= 0.30 \times \left[0.95 \times 0.20 + 0.90 \times 0.80 \right] + 0.70 \times \left[0.29 \times 0.20 + 0.01 \times 0.80 \right] =$$
$$= 0.30 \times 0.91 + 0.70 \times 0.066 = 0.319$$

Actuando de la misma forma podemos calcular las probabilidades a priori de S_{1+} y S_{2+}

$$P(S_{1+}) = \sum_{E_i} p(S_{1+} / E_i) p(E_i) = p(S_{1+} / E_+) p(E_+) + p(S_{1+} / E_-) p(E_-) =$$
$$= 0.80 \times 0.319 + 0.20 \times 0.681 = 0.392$$

$$P(S_{21+}) = \sum_{E_i} p(S_{2+} / E_i) p(E_i) = p(S_{21+} / E_+) p(E_+) + p(S_{2+} / E_-) p(E_-) =$$
$$= 0.70 \times 0.319 + 0.4 \times 0.681 = 0.496$$

De este modo obtenemos las probabilidades a priori de cada nodo. Sin embargo lo que suele ser más interesante en las redes de creencia es el cálculo de las probabilidades a posteriori, es decir, cual es la probabilidad de tener la enfermedad *E*, si se tiene un determinado síntoma S_1 o S_2. Esto puede hacerse aplicando la regla de Bayes entre los nodos de la red de la siguiente forma:

$$p(E_+ / S_{1+}) = \frac{p(S_{1+} / E_+) p(E_+)}{p(S_{1+})} = \frac{0.8 \times 0.319}{0.392} = 0.652$$

$$p(E_+ / S_{2+}) = \frac{p(S_{2+} / E_+) p(E_+)}{p(S_{2+})} = \frac{0.7 \times 0.319}{0.496} = 0.451$$

Si los dos síntomas aparecen simultáneamente el cálculo debería ser

$$p(E_+ / S_{1+}, S_{2+}) = \frac{p(S_{1+}, S_{2+} / E_+) p(E_+)}{p(S_{1+}, S_{2+})}$$

El valor de $p(S_{1+}, S_{2+} / E_+)$ no puede extraerse directamente de las tablas de probabilidades condicionales, sin embargo podemos aplicar la separación condicional para obtener una expresión más sencilla

$$p(S_{1+}, S_{2+} / E_+) = p(S_{1+} / E_+) p(S_{2+} / E_+)$$

con lo que llegamos a

$$p(E_+ / S_{1+}, S_{2+}) = \frac{p(S_{1+} / E_+) p(S_{2+} / E_+) p(E_+)}{p(S_{1+}, S_{2+})} = \alpha \times p(S_{1+} / E_+) p(S_{2+} / E_+) p(E_+)$$

En este caso se ha representado $1/p(S_{1+}, S_{2+})$ como α para indicar que no se necesita calcular este valor a partir de los datos de la red, sino que puede obtener por normalización. Así tenemos que

$$p(E_+ / S_{1+}, S_{2+}) = \alpha \times p(S_{1+} / E_+) p(S_{2+} / E_+) p(E_+) = \alpha \times 0.8 \times 0.7 \times 0.319 = \alpha \times 0.179$$

$$p(E_- / S_{1+}, S_{2+}) = \alpha \times p(S_{1+} / E_-) p(S_{2+} / E_-) p(E_-) = \alpha \times 0.2 \times 0.4 \times 0.681 = \alpha \times 0.054$$

Como las probabilidades deben sumar la unidad obtenemos que

$$\alpha \times 0.179 + \alpha \times 0.054 = 1 \quad \Rightarrow \quad \alpha \times 0.233 = 1 \quad \Rightarrow \quad \alpha = \frac{1}{0.233} = 4.288$$

por lo tanto

$$p(E_+ / S_{1+}, S_{2+}) = \alpha \times 0.179 = 0.766$$
$$p(E_- / S_{1+}, S_{2+}) = \alpha \times 0.054 = 0.234$$

Evidentemente, al darse los dos síntomas a la vez la probabilidad de sufrir la enfermedad *E* es mayor que si sólo aparece un síntoma.

El ejemplo visto ilustra las simplificaciones que pueden obtenerse utilizando la estructura de independencia contenida en el modelo probabilístico. El principal problema de los métodos de inferencia exacta en redes probabilísticas es que son del tipo NP-Completos, lo que los hace intratables computacionalmente. El método más popular para resolver este problema es limitar la topología de la red a un tipo restringido, como los poliárboles (árboles en los que cada nodo puede tener más de una raíz).

Kim y Pearl (1983) desarrollaron un algoritmo eficiente para la inferencia en redes con topología de poliárbol. La principal característica de este algoritmo es que su complejidad es lineal en el tamaño de la red (es decir en el número de nodos y aristas que la componen), a diferencia del método de fuerza bruta que requiere un número exponencial de operaciones para realizar la propagación.

El algoritmo de propagación en poliárboles se basa en que cada nodo divide a la red en dos poliárboles inconexos: uno contiene a sus padres y a los nodos a los que está conectado pasando por sus padres y otro incluye a sus hijos y a los nodos a los que está conectado pasando por sus hijos. El proceso de propagación puede realizarse en este tipo de grafos de un modo eficiente combinando la información procedente de los distintos subgrafos mediante el envío de mensajes (cálculos locales) de un subgrafo a otro.

El problema de los poliárboles es que no son aplicables en numerosas situaciones prácticas. En estos casos es necesario trabajar con grafos múltiplemente conexos en los que, al existir más de un camino entre dos nodos, los cálculos resultan más complejos. Por ejemplo imaginemos la red de la Figura 1.6. Supongamos que sabemos que el valor del nodo *D* es *cierto* y queremos conocer el valor de las probabilidades condicionales en el nodo *C*. En esta red el cambio en *D* afectará a *C* de distintas formas, no solo *C* tiene que tener en cuenta el cambio en *D*, sino también el cambio en *A* que es causado por *D* a través de *B*. Desgraciadamente estos cambios no son separables con claridad.

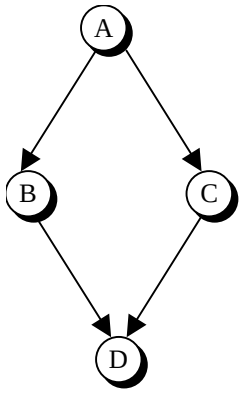

Figura 1.6 Una red múltiplemente conexa.

Una estrategia comúnmente adoptada es transformar las redes múltiplemente conexas en otras equivalentes pero en las cuales pueda aplicarse el algoritmo de inferencia en poliárboles. Dentro de esta estrategia podemos clasificar dos tipos de métodos: los *métodos de condicionamiento* y los *métodos de agrupamiento*.

La idea fundamental del método de propagación por condicionamiento es cortar los múltiples caminos entre los nodos mediante la asignación de valores a un conjunto reducido de variables contenidas en los bucles (Pearl (1988) y Suedermont y Cooper (1991)). De esta forma se obtendrá un poliárbol en el cual se podrá aplicar el algoritmo de propagación para poliárboles.

El método de agrupamiento construye representaciones auxiliares, de estructura más simple, uniendo conjuntos de nodos del grafo original. De esta forma se puede obtener un grafo con estructura de poliárbol en el que puede aplicarse el algoritmo de inferencia en poliárboles (Lauritzen y Spiegelhalter (1988), Jensen, Olesen y Andersen (1990)).

Sin embargo, no todos los algoritmos de inferencia requieren poliárboles. Shachter (1988) y otros investigadores han desarrollado un algoritmo que invierte los arcos en la red, aplicando el teorema de Bayes en cada inversión, hasta que se han computado las probabilidades de interés. D'Ambrosio (1991) desarrolló un método de inferencia denominado inferencia probabilística simbólica (o SPI de sus siglas en inglés). El SPI sólo realiza aquellos cálculos que son necesarios para responder a una determinada consulta, y no como a los métodos vistos hasta ahora, en los que se calculan todas las distribuciones locales sin tener en cuenta su relevancia para la consulta formulada.

Finalmente, Shachter, Andersen y Szolovitz (1993) desarrollaron el *Algoritmo de Clustering* como un marco unificado de los métodos exactos de inferencia, en el cual se demostraba que todos los algoritmos de inferencia desarrollados podían derivarse de dicho marco unificado. El algoritmo de clustering demostró que la esencia de una inferencia probabilística eficiente es la factorización de la función de distribución de probabilidad conjunta en base a asunciones de independencia condicional.

Métodos aproximados de inferencia

Los métodos aproximados de inferencia son aquellos métodos que calculan las probabilidades condicionales de los nodos de forma aproximada, y se utilizan en aquellos casos en los que los algoritmos exactos no son aplicables, o son computacionalmente costosos.

Un estudio realizado sobre los métodos aproximados de inferencia (Dagum y Luby, 1993) ha demostrado que el problema general de inferencia probabilística aproximada en redes de creencia también es NP-Completa. Sin embargo, a pesar de estos resultados, los métodos aproximados pueden ser muy útiles en distintas situaciones y siguen siendo una importante área de investigación.

La idea básica de los métodos aproximados consiste en generar una muestra de posibles escenarios de la red (en donde cada escenario define una red completamente ejemplificada) y luego utilizar la muestra generada para calcular valores aproximados de las probabilidades de ciertos sucesos dada una determinada evidencia. Los métodos de propagación aproximada pueden clasificarse en dos tipos: *métodos de simulación estocástica*, que generan la muestra usando mecanismos aleatorios, y *métodos de búsqueda determinista*, que generan la muestra de forma sistemática.

El método más conocido de simulación estocástica es el método del muestreo lógico probabilístico propuesto por Henrion (1988). Este método se basa en que las redes de creencia son acíclicas, por lo que siempre va a existir un conjunto de nodos raíz cuya probabilidad a priori es conocida. Los nodos raíz pueden ejemplificarse en base a su probabilidad a priori, y estos valores puede transmitirse al resto de nodos a través de las probabilidades condicionadas que existen en los arcos de la red. Una red ejemplificada constituye un punto en el espacio probabilístico, una vez que tenemos un número elevado de puntos podemos aproximar la distribución de probabilidad de la red.

El problema de esta aproximación es que no tiene en cuenta si ya existe de antemano algún nodo ejemplificado, es decir, que forma parte de las evidencias conocidas del problema. Para resolver esto se han propuesto nuevos métodos como el método de integración de evidencias de Ching y Cooper (1989) que invierte los arcos de las evidencias conocidas mediante la aplicación iterativa de la regla de Bayes para convertir los nodos observados en nodos raíz, el método de muestreo de Markov propuesto por Pearl (1987), el método de ponderación de evidencias de Fung y Chang (1990) y Shachter y Peot (1990), etc. Por otro lado, entre los métodos de muestreo sistemático podemos destacar el algoritmo propuesto por Bouckaert (1994).

1.8. Aprendizaje en Redes de Creencia

La mayor parte de la investigación en redes de creencia se ha centrado fundamentalmente en desarrollar algoritmos de inferencia que nos permitan extraer probabilidades de interés de la red. Estos algoritmos de inferencia trabajan sobre una red ya desarrollada que esta compuesta por dos elementos principales: una estructura de dependencias en forma de grafo dirigido acíclico y un conjunto de probabilidades condicionadas. Sin embargo, en muchas situaciones prácticas, el investigador no contará con alguno de estos elementos por lo que será necesario estimarlos a partir de un conjunto de datos. Este proceso de estimación de las características de la red es lo que se conoce como aprendizaje.

Existen dos tipos de aprendizaje: aprendizaje paramétrico, cuyo objetivo es determinar las probabilidades a incluir en la red de creencia; y aprendizaje estructural, cuyo objetivo es determinar la estructura gráfica de dependencia.

La complejidad del aprendizaje estructural es mayor que la complejidad del aprendizaje paramétrico y ambos necesitan de muestras con un tamaño razonable para evitar problemas como el sobreaprendizaje. El sobreaprendizaje es muy conocido en tareas de aprendizaje supervisado y consiste, básicamente, en que el esquema resultante del aprendizaje se adapte muy bien a la muestra suministrada pero que pierda la capacidad de generalización cuando se le pasa un dato no incluido en la muestra original.

Aprendizaje paramétrico

El aprendizaje paramétrico se basa fundamentalmente en encontrar los valores de los parámetros de las distribuciones de probabilidad condicional que maximizan la verosimilitud de los datos suministrados. Estos datos se suelen denominar datos de entrenamiento y se dividen en una serie de casos que se suponen son independientes entre si.

Por ejemplo, consideremos la estimación de la tabla de probabilidades condicionales del nodo E de la red de la Figura 1.3. A partir de los datos de entrenamiento podemos contar el número de casos en los que el valor de E es True, para cada una de las combinaciones de valores de C_1 y C_2, y que representamos como $N(E_+, C_{1+}, C_{2+})$, $N(E_+, C_{1+}, C_{2-})$, $N(E_+, C_{1-}, C_{2+})$ y $N(E_+, C_{1-}, C_{2-})$.

Calculados estos valores podemos dar una estimación de la probabilidad $p(E_+ / C_{1+}, C_{2+})$ de la siguiente forma:

$$p(E_+ / C_{1+}, C_{2+}) \approx \frac{N(E_+, C_{1+}, C_{2+})}{N(C_{1+}, C_{2+})}$$

Actuando de la misma forma para el resto de valores de C_{1+} y C_{2+} podemos obtener la tabla de probabilidades condicionales del nodo E.

Otra aproximación es el algoritmo bayesiano MAP (Maximum a posteriori) que extiende la aproximación de máxima verosimilitud mediante la introducción de probabilidades a priori (Kass y Raftery, 1995).

El proceso de encontrar la distribución de probabilidades de máxima verosimilitud tiene una importante restricción y es que los datos suministrados deben ser completos, no puede haber ningún caso en el que existan valores perdidos o valores desconocidos. Esta asunción puede no ser realista porque en diversas ocasiones es común que existan valores perdidos (por ejemplo, en bases de datos médicas algunas medidas que son caras de realizar no se llevan a cabo si no son críticas para el diagnóstico).

Para el caso en el que haya valores perdidos el algoritmo más comúnmente empleado es el algoritmo EM (Expectation Maximization) que es un método general de aprendizaje con datos incompletos. El algoritmo EM consta de dos etapas: la primera es la etapa de cálculo de los valores esperados, en la que se calcula la esperanza de los datos incompletos o funciones de ellos; la segunda etapa es una etapa de maximización, en la que se maximiza una cierta función. Las dos etapas se iteran hasta conseguir la convergencia del proceso, que está garantizado bajo ciertas condiciones de regularidad (ver Dempster, Laird y Rubin, 1977).

Además de los algoritmos citados existen otros (Laplace, IPF, Gibbs, MCMC, etc.) que han sido desarrollados para resolver distintos problemas específicos encontrados en la estimación de parámetros. Una lista de referencias sobre los mismos puede encontrarse en Buntine (1996). Es de destacar que el aprendizaje de parámetros en una red bayesiana guarda bastantes similitudes con el entrenamiento de redes de neuronas artificiales.

Aprendizaje estructural

El proceso de aprendizaje estructural aparece como paso previo al aprendizaje paramétrico y consiste en decidir que nodos están unidos por un arco y cual es la dirección de dicho arco. En este proceso se pueden cometer dos tipos de errores: añadir un arco de más, con lo que estaremos aumentando el número de parámetros a estimar y estaremos realizando asunciones equívocas acerca de causalidad y la estructura del dominio y; omitir un arco, cuyo efecto no podrá ser compensado por la estimación de los demás parámetros además de perder causalidad y afectar a la estructura del dominio (Figura 1.7).

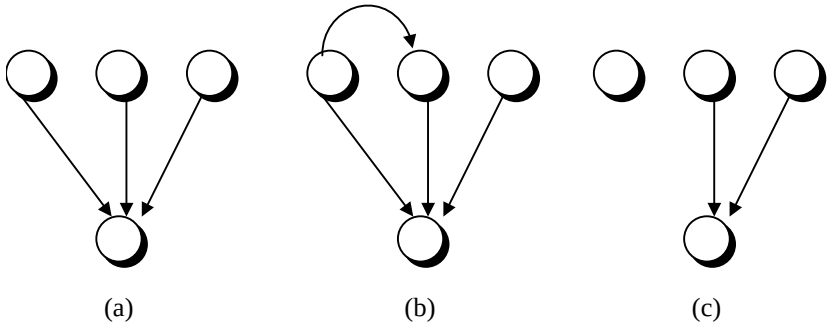

Figura 1.7 Aprendizaje estructural: (a) red a estimar, (b) error por adición de arcos, (c) error por eliminación de arcos.

El aprendizaje estructural, al menos para el caso de variables discretas, está muy relacionado con el problema de aprendizaje de árboles de clasificación, de los que podemos destacar el algoritmo CART en estadística y los algoritmos ID3 y C4 en la inteligencia artificial.

El problema de los árboles de clasificación tiene una larga historia y ha sido estudiado desde distintas perspectivas: estadística aplicada, inteligencia artificial, estadística bayesiana, métodos MDL (Minimum Description Length), algoritmos genéticos y la teoría del aprendizaje computacional. Una adaptación de un algoritmo para construir árboles de clasificación en un algoritmo para el aprendizaje de redes bayesianas puede consultarse en (Buntine, 1991).

Un aspecto que deben tener en cuenta los algoritmos de aprendizaje estructural es que existen redes que son equivalentes. Decimos que dos redes son equivalentes cuando representan las mismas declaraciones de independencia. Por ejemplo supongamos las redes mostradas en la Figura 1.8.

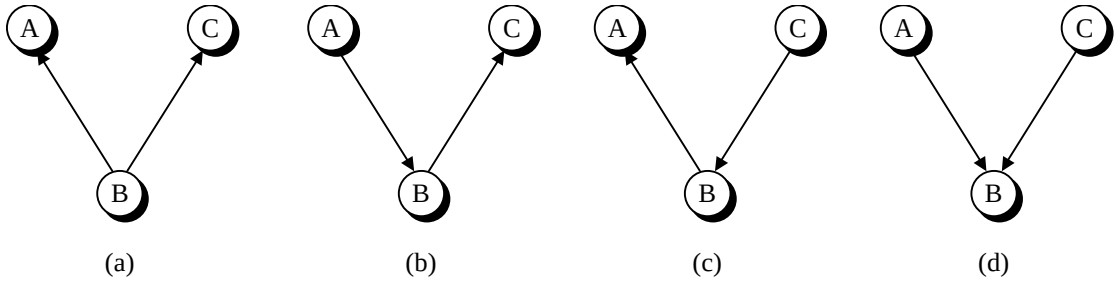

Figura 1.8 Distintos tipos de redes con tres nodos y dos arcos.

Estas tres primeras redes (a), (b) y (c) son equivalentes porque representan la misma estructura de independencia. Para probar esto podemos ver que la función de probabilidad conjunta de estas redes es:

$$p_a(A,B,C) = p(B)\,p(A/B)\,p(C/B)$$

$$p_b(A,B,C) = p(A)\,p(B/A)\,p(C/B)$$

$$p_c(A,B,C) = p(C)\,p(B/C)\,p(A/B)$$

Realizando sencillas operaciones sobre las probabilidades conjuntas p_b (A, B, C) y p_c (A, B, C) aplicando la regla de Bayes podemos ver como resultan ser equivalentes a p_a (A, B, C).

$$p_b(A,B,C) = p(A)\,p(B/A)\,p(C/B) \xrightarrow{Bayes} = p(A)\frac{p(A/B)\,p(B)}{p(A)}\,p(C/B)$$
$$\xrightarrow{simplificando} = p(B)\,p(A/B)\,p(C/B) = p_a(A,B,C)$$

$$p_c(A,B,C) = p(C)\,p(B/C)\,p(A/B) \xrightarrow{Bayes} = p(C)\frac{p(C/B)\,p(B)}{p(C)}\,p(A/B)$$
$$\xrightarrow{simplificando} = p(B)\,p(A/B)\,p(C/B) = p_a(A,B,C)$$

Sin embargo la red de la Figura 1.8 (d) no es equivalente a las demás ya que su distribución de probabilidad conjunta es

$$p_d(A,B,C) = p(A)\,p(C)\,p(B/A,C)$$

En los algoritmos de aprendizaje estructural también puede aparecer el problema de los valores perdidos. Una táctica muy común para solucionarlo es aplicar una versión del algoritmo EM visto para el aprendizaje paramétrico y que se conoce como SEM (Structural EM).

Además de los valores perdidos también aparece el problema de las variables ocultas, es decir, variables que no son observadas en la muestra de datos de entrenamiento. La correcta identificación de estas variables es importante ya que permite realizar una importante reducción en el número de parámetros a estimar. Por ejemplo, si atendemos a la red representada en la Figura 1.9 (a), el número de parámetros a estimar si todas las variables son binarias es de 17. Sin embargo, en la red de la Figura 1.9 (b), que no incluye la variable oculta pero que captura la misma distribución que la red anterior, el número de parámetros a estimar es de 59, lo que significa que la existencia de la variable oculta ha permitido una reducción del orden del 70%.

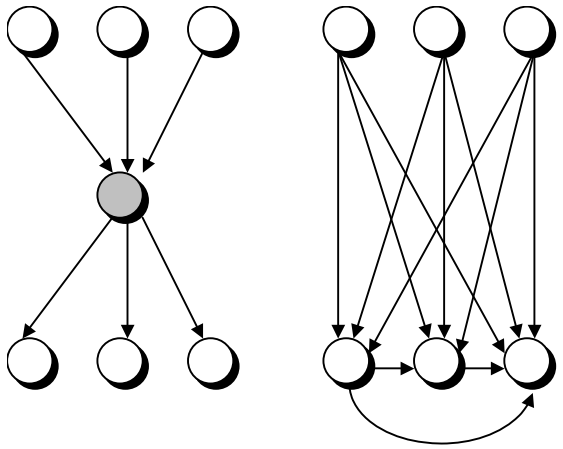

Figura 1.9 (a) Red con una variable oculta, (b) la red más sencilla que puede capturar la misma distribución sin usar la variable oculta.

Aunque la inclusión de variables ocultas permite una reducción importante en el número de parámetros a estimar también conllevan una serie de problemas a resolver como: ¿cómo reconocer la necesidad de una nueva variable oculta?, ¿dónde introducir la variable oculta en la estructura actual?, determinar su cardinalidad (número de posibles valores) y su tipo de distribución condicional, etc. Entre los algoritmos desarrollados para actuar con valores perdidos y variables ocultas podemos destacar el algoritmo MS-EM (Model Selection EM) desarrollado por Friedman (1997).

1.9. Resumen

Seguramente, el reverendo presbiteriano Thomas Bayes (1702-1761) jamás sospechó que su trabajo "Essay Towards Solving a Problem in the Doctrine of Chances" en el que explicaba sus teorías estadísticas, y que fue publicado póstumamente en 1763 en los *Philosophical Transactions of the Royal Society of London*, tendría la repercusión mundial que finalmente ha tenido. Tal es la devoción que sienten los estadísticos hacia sus teorías que en 1969 su tumba fue restaurada con contribuciones recibidas de estadísticos de todo el mundo.

Debido a estar fuertemente fundados en la teoría de la probabilidad, los métodos bayesianos se han hecho muy populares en el mundo de la inteligencia artificial como métodos para tratar la incertidumbre. Sin embargo, durante muchos años estos métodos han sido denostados porque se creía que eran impracticables en problemas reales debido a la gran cantidad de probabilidades que es necesario obtener para construir una base de conocimientos. Esta situación ha cambiado con la llegada de las redes de creencia o redes bayesianas.

Las redes de creencia representan un matrimonio entre la teoría de la probabilidad y la teoría de grafos. La idea fundamental de las redes de creencia es la noción de modularidad, un sistema complejo es construido combinando partes más simples. La teoría de la probabilidad es el pegamento que une a las partes, asegurando

que el sistema es consistente, mientras que la teoría de grafos nos provee, por un lado, de una interfaz intuitiva en la que modelar conjuntos de variables que interactúan entre si y, por otro lado, una estructura de datos que permite el diseño de algoritmos eficientes de propósito general.

Las redes de creencia se estructuran en forma de grafos dirigidos acíclicos en los que los nodos son variables aleatorias y los arcos reflejan la existencia de influencias causales entre los nodos. Los algoritmos de inferencia de conocimiento en redes de creencia explotan los mecanismos de independencia reflejados en la estructura de la red para simplificar los procesos de razonamiento. En la actualidad los avances en el aprendizaje de redes a partir de datos han minimizado la influencia de los expertos humanos en su construcción.

Las redes de creencia se han hecho muy populares en los últimos años y prueba de esta popularidad es que el gigante informático Microsoft decidió integrar en 1993 dentro de su plantilla al grupo de formado por Heckerman, Horvitz y Breese especializados en la investigación en redes de creencia. Según el propio Bill Gates, la experiencia de Microsoft en redes de creencia es una ventaja competitiva que ya ha empezado a utilizar en sus productos (sobre todo en sistemas de ayuda a la resolución de problemas en el ordenador). Como curiosidad decir que el articulo de *Los Angeles Times* que daba cuenta de esta noticia tenía el siguiente y sensacionalista titular: “El futuro del software puede depender de las oscuras teorías de un clérigo del siglo XVIII llamado Thomas Bayes”. No creemos que las teorías de Bayes sean oscuras y mucho menos que de ellas dependa el futuro del software, pero es evidente que su importancia ha crecido en los últimos años.

1.10. Textos básicos

- Castillo, Gutiérrez, Hadi, “Sistemas expertos y modelos de redes probabilísticas”, Monografías de la Academia de Ingeniería, 1996.

- Jensen “An introduction to Bayesian Networks”, UCL Press, 1996.

- Pearl “Probabilistic reasoning in intelligent systems: networks of plausible inference”, Morgan Kauffman Publishers, 1988.

- Russell, Norvig “Inteligencia Artificial: un enfoque moderno”, Prentice-Hall Hispanoamericana, 1996.

- Juez Martel, Díez Vegas “Probabilidad y estadística en medicina”, Díaz de Santos, 1997.

1. FACTORES DE CERTIDUMBRE

Parte de los inconvenientes encontrados en los modelos descritos en capítulos anteriores podrían ser resueltos empleando conocimiento heurístico. Si más concretamente nos referimos a los problemas estadísticos derivados de la inevitable y exhaustiva recolección de datos, una posible solución podría pasar por el establecimiento de un nuevo concepto, el de *probabilidad condicional subjetiva*, que definimos como una medida numérica que relaciona dos sucesos, de forma que la ocurrencia de uno está condicionada por la ocurrencia del otro, pero en donde la relación no está avalada por amplios estudios estadísticos. Así, la expresión:

$$p(I_i / S_k) = x$$

podría traducirse como:

SI: La manifestación S_k está presente,
ENTONCES: Según mi experiencia hay, digamos, una probabilidad (subjetiva) x, de que la interpretación sea I_i

Este enfoque resuelve el problema de la recogida masiva de información y datos, pero sigue presentando problemas. Así, tras un amplio diálogo con uno o varios expertos, es posible que si I_1, ..., I_n son todas las posibles interpretaciones relativas al conjunto de manifestaciones S_k, la suma de las correspondientes probabilidades condicionales subjetivas sea distinta de la unidad. En otras palabras:

$$\sum_i p(I_i / S_k) \neq 1$$

Esta situación, que no es compatible con la estadística tradicional, se resuelve fácilmente *normalizando* a "uno" las correspondientes probabilidades condicionales subjetivas, con el único objetivo de mantener la consistencia matemática del modelo. Otros problemas, sin embargo, no son tan fáciles de resolver. Y es que cuando trabajamos con información de carácter simbólico, en lugar de trabajar con datos numéricos, aparecen conceptos tales como *imprecisión, incertidumbre, falta de información, credibilidad, ...* que son difíciles de definir. La necesidad de construir programas inteligentes, que sean capaces de manipular estos tipos diferentes de información, sugiere la conveniencia de formalizar nuevas aproximaciones y modelos.

1.1. El Modelo de los Factores de Certidumbre

La ya familiar expresión "$p(I_i / S_k) = x$" puede interpretarse en términos de implicación del siguiente modo:

$$S_k \xrightarrow{x} I_i$$

Expresión en la que x define la llamada *potencia evidencial* de la implicación correspondiente. Si cada vez que S_k se manifiesta, podemos concluir I_i con total

seguridad (i.e., sin ninguna incertidumbre), entonces decimos que la relación existente entre S_k e *Ii* es *patognómica*, y x vale 1. En cualquier otro caso, para $x \geq 0$, y para $x < 1$, la implicación viene afectada de alguna incertidumbre y, por lo tanto, no siempre que se da S_k se puede concluir I_i, pero se puede establecer una relación causal entre S_k e I_i que será tanto más categórica cuanto más se acerque x a 1. De este modo la *intensidad* de la relación causal viene establecida a través de la potencia evidencial x.

A lo largo de un proceso completo de razonamiento, y manteniendo I_i como hipótesis de trabajo, seguramente aparecerá evidencia a favor de la hipótesis considerada, pero también aparecerá evidencia en contra de dicha hipótesis. Por otra parte, si aplicamos un esquema bayesiano al problema planteado, deberíamos concluir también que:

$$S_k \xrightarrow{1-x} \neg I_i$$

lo cual, como ya hemos visto, cuando trabajamos con conocimiento, es inaceptable.

En 1975, y para tratar de resolver este tipo de problemas, Shortliffe y Buchanan plantearon un modelo de razonamiento que sacudió los cimientos del entonces incipiente mundo de la inteligencia artificial. El modelo de Shortliffe y Buchanan es de naturaleza “ad hoc”, y por consiguiente carece de una base teórica fuerte. No obstante, dicho modelo fue inmediatamente aceptado debido a su fácil comprensión y a la calidad de los resultados obtenidos tras su aplicación. Las ideas básicas del modelo pueden resumirse en los siguientes puntos:

- Dada una hipótesis que está siendo considerada, la potencia evidencial de una declaración se debe representar a través de dos medidas diferentes: la *medida de confianza creciente MB*, y la *medida de desconfianza creciente MD*.

- *MB* y *MD* son, en realidad, índices dinámicos que representan incrementos asociados a evidencias nuevas.

- Si *h* es una hipótesis, y *e* una evidencia, la misma evidencia *e* no puede, simultáneamente, incrementar la confianza en *h* y disminuir la confianza en *h*.

- *MB*(*h*,*e*) representa el incremento de la confianza en *h* dada la evidencia *e*.

- *MD*(*h*,*e*) representa el incremento de la desconfianza en *h* dada la evidencia *e*

Con estas premisas podemos establecer el siguiente formalismo y casos particulares:

- Sea *p*(*h*) la confianza previa en *h* antes de *e*

- Sea *p*(*h*/*e*) la confianza en *h* tras la aparición de *e*

- Sea 1 - *p*(*h*) la desconfianza previa en *h* antes de *e*

CASO 1

Si $p(h/e) > p(h)$, entonces la nueva evidencia produce un aumento de confianza en la hipótesis considerada. En este caso:

- $MB\ (h,e) > 0$
- $MD\ (h,e) = 0$

y $MB\ (h,e)$ se define del siguiente modo:

- $$MB\ (h,e) = \frac{p(h/e) - p(h)}{1 - p(h)}$$

De acuerdo con esta expresión $MB\ (h,e)$ representa el incremento relativo de la confianza en la hipótesis *h* tras la aparición de la evidencia *e*, que coincide con la disminución relativa de la desconfianza en *h* tras la aparición de la evidencia *e*.

CASO 2

Si $p(h) > p(h/e)$, entonces la nueva evidencia produce una disminución de la confianza depositada en la hipótesis considerada. En este caso:

- $MB\ (h,e) = 0$
- $MD\ (h,e) > 0$

y $MD\ (h,e)$ se define del siguiente modo:

- $$MD\ (h,e) = \frac{p(h) - p(h/e)}{p(h)}$$

De acuerdo con esta expresión, $MD\ (h,e)$ representa el incremento relativo de la desconfianza en la hipótesis *h* tras la aparición de la evidencia *e*, que coincide con la disminución relativa de la confianza en *h* tras la aparición de la evidencia *e*.

CASO 3

Si $p(h/e) = p(h)$, entonces la nueva evidencia es independiente de la hipótesis considerada, ya que no aumenta ni la confianza ni la desconfianza. En este caso:

$$MB\ (h,e) = MD\ (h,e) = 0$$

Claramente, si $p(h)$ simboliza una probabilidad a priori en sentido clásico, podemos establecer los valores límite de las correspondientes medidas de confianza y desconfianza crecientes, según las expresiones siguientes:

- $MB\,(h,e) = 1 \Leftrightarrow p(h) = 1$

$$= \frac{\max\,[p(h/e), p(h)] - p(h)}{\max\,[1,0] - p(h)} \text{ en otro caso}$$

- $MD\,(h,e) = 1 \Leftrightarrow p(h) = 0$

$$= \frac{\min\,[p(h/e), p(h)] - p(h)}{\min\,[1,0] - p(h)} \text{ en otro caso}$$

Ambas expresiones no son más que representaciones formales y simétricas de las medidas de confianza y desconfianza crecientes, expresadas en términos de probabilidades condicionales y de probabilidades a priori.

Además de estas dos medidas, Shortliffe y Buchanan definen un tercer índice, denominado *factor de certidumbre, CF*, que combina las dos medidas anteriores según la expresión:

- $CF(h,e) = MB\,(h,e) - MD\,(h,e)$

expresión que también es de carácter formal, ya que una misma evidencia nunca puede incrementar, simultáneamente, la confianza y la desconfianza en la misma hipótesis.

Shortliffe y Buchanan justifican la introducción de este factor de certidumbre como un medio para facilitar la comparación entre potencias evidenciales de hipótesis alternativas (i.e., h1, ..., hn), en relación a una misma evidencia *e*.

En cada una de las tres medidas desarrolladas podemos identificar las características siguientes:

RANGOS

$$0 \leq MB\,(h,e) \leq 1$$
$$0 \leq MD\,(h,e) \leq 1$$
$$-1 \leq CF\,(h,e) \leq 1$$

COMPORTAMIENTO EN CASOS EXTREMOS E HIPOTESIS MUTUAMENTE EXCLUYENTES

Si *h* es cierta, y por lo tanto $p(h/e) = 1$, entonces:

$$MB\,(h,e) = \frac{1 - p(h)}{1 - p(h)} = 1$$
$$MD\,(h,e) = 0$$
$$CF(h,e) = 1$$

Si la negación de *h* es cierta, y por lo tanto $p(\neg h/e) = 1$, entonces:

$MB\ (h,e) \quad = \quad 0$

$MD\ (h,e) \quad = \quad \dfrac{0 - p(h)}{0 - p(h)} = 1$

$CF(h,e) \quad = \quad -1$

Este planteamiento conduce a que $MB\ (\neg h,e) = 1$, si y sólo si $MD(h,e) = 1$, resultado que se obtiene sin más que recordar cómo se definen *MB* y *MD*.

Por otra parte, si $h1$ y $h2$ son hipótesis mútuamente excluyentes, y sabemos que $MB\ (h1,e) = 1$, entonces podemos afirmar rotundamente que $MD\ (h2,e) = 1$.

EVIDENCIAS INDEPENDIENTES DE LA HIPOTESIS

Sea *h* la hipótesis considerada, y sea e una evidencia. Si la evidencia es independiente de la hipótesis (i.e., ni la apoya, ni va en contra de ella), entonces:

$p(h/e) = p(h)$

$MB(h,e) = 0$

$MD(h,e) = 0$

$CF(h,e) = 0$

DIFERENCIA ENTRE FACTORES DE CERTIDUMBRE Y PROBABILIDADES CONDICIONALES

Recordemos, una vez más, que uno de los puntos más débiles de los modelos probabilísticos era el hecho de que una misma evidencia apoyaba, simultáneamente, a una hipótesis y a su negación. Ello era consecuencia de la consistencia matemática de tales modelos (i.e., $p(h/e) + p(\neg h/e) = 1$).

Este inconveniente no aparece en el modelo de los factores de certidumbre de Shortliffe y Buchanan, quienes textualmente[44] afirman: "... los factores de certidumbre de las hipótesis *h* y ¬*h* no son complementarios a la unidad, son opuestos entre sí. Así, si el apoyo que una evidencia presta a una hipótesis es bajo, no debería ser alto el apoyo a la negación de tal hipótesis, sobre todo en el caso de que la información no sea completa. En este caso, el apoyo a ambas, hipótesis y negación, debe ser bajo..."

Lo que realmente enfatizan Shortliffe y Buchanan en esta cita es un hecho muy concreto: si "algo" tiene muy poco que ver con "una cosa", no es razonable pensar que ese mismo "algo" tenga "mucho que ver con la negación de esa "misma cosa".

Shortliffe y Buchanan demuestran tal afirmación analizando casos extremos:

Si $P(h/e) > P(h)$, entonces $[P(\neg h/e) = 1 - P(h/e)] < [1 - P(h) = P(\neg h)]$

---

[44] ... aunque con una traducción más o menos libre,...

De este modo:

$MB\ (\neg h,e) = 0$
$MD\ (\neg h,e) > 0$
$MB\ (h,e) > 0$
$MD\ (h,e) = 0$

pero:

$$CF\ (\neg h,e) = MB\ (\neg h,e) - MD\ (\neg h,e) =$$

$$= \frac{0 - [p(\neg h) - p(\neg h/e)]}{p(\neg h)} = \frac{[p(\neg h) - p(\neg h/e)]}{p(\neg h)} =$$

$$= \frac{1 - p(h/e) - 1 + p(h)}{1 - p(h)} = \frac{p(h) - p(h/e)}{1 - p(h)}$$

$$= - MB\ (h,e) = - CF(h,e)$$

por lo que $CF(\neg h,e) = - CF(h,e)$

El mismo resultado se obtiene si consideramos el otro caso extremo, en el que $P(h) > P(h/e)$.

1.2. Combinación de Evidencias

¿Cómo debe manejar un experto los factores de certidumbre...? Para el caso concreto de una única evidencia la respuesta es clara:

- El experto indicará un valor mayor que cero y menor o igual a uno, si la evidencia en cuestión apoya a la hipótesis.

- El experto indicará un valor menor que cero, pero mayor o igual a menos uno, si la evidencia en cuestión va en contra de la hipótesis.

- El experto indicará un valor igual cero, si estima que la evidencia encontrada no tiene nada que ver con la hipótesis considerada.

El problema, no obstante, se complica cuando hay más de una evidencia relativa a una misma hipótesis. En este caso hablamos de combinación de evidencias que afectan a una misma hipótesis. El problema planteado podemos formularlo en los siguientes términos:

Sea un conjunto de reglas, todas ellas con la misma conclusión, cada una de las cuales viene afectada de un factor de certidumbre diferente, ¿cual es el factor de certidumbre resultante, considerando toda la evidencia?

IF: $e1$ THEN: H with: $CF(H,e1)$
IF: $e2$ THEN: H with: $CF(H,e2)$
. . .
IF: en THEN: H with: $CF(H,en)$

En este caso, los factores de certidumbre de las distintas reglas pueden interpretarse como las potencias evidenciales de las relaciones causales correspondientes:

$$e1 \xrightarrow{CF(H,e1)} H$$
$$e2 \xrightarrow{CF(H,e2)} H$$
...
$$en \xrightarrow{CF(H,en)} H$$

Es evidente que cada una de tales evidencias contribuye, favorable o desfavorablemente, al establecimiento de la veracidad de la hipótesis considerada. El problema estriba en encontrar una formulación adecuada que nos permita evaluar $CF(H,E)$, donde $E = [e1, e2, ..., en]$

Shortliffe y Buchanan proponen una primera aproximación para la combinación entre pares de evidencias que se refieren a la misma hipótesis. Esta primera aproximación, en términos de factores de certidumbre, puede expresarse del siguiente modo:

Caso A

Si $e1$ y $e2$ contribuyen positivamente a la veracidad de la hipótesis H, entonces:

- $CF(H, e1) > 0$
- $CF(H, e2) > 0$
- $CF(H, e1 \text{ y } e2) = CF(H, e1) + CF(H, e2) - [CF(H, e1) \times CF(H, e2)]$

Caso B

Si $e1$ y $e2$ contribuyen negativamente a la veracidad de la hipótesis H, entonces:

- $CF(H, e1) < 0$
- $CF(H, e2) < 0$
- $CF(H, e1 \text{ y } e2) = CF(H, e1) + CF(H, e2) + [CF(H, e1) \times CF(H, e2)]$

Caso C

Si $e1$ contribuye positivamente a la veracidad de la hipótesis H, y $e2$ contribuye negativamente a la veracidad de la hipótesis H, entonces:

- $CF(H, e1) > 0$
- $CF(H, e2) < 0$
- $CF(H, e1) \times CF(H, e2) < 0$
- $CF(H, e1 \text{ y } e2) = CF(H, e1) + CF(H, e2)$

Esta primera aproximación es coherente con la idea de que, ante la posibilidad de información incompleta, el efecto conjunto (supongamos positivo), de dos evidencias, debe ser igual a la suma de sus efectos por separado menos su efecto conjunto. En otras palabras, nos previene de la hipotética situación de que ambas evidencias pudieran no ser completamente independientes. Además, y siempre en el caso de que ambas evidencias contribuyan positivamente a la veracidad de la hipótesis[45], la expresión es generalizable directamente.

En efecto, si en lugar de dos evidencias tenemos *n* evidencias, todas ellas con CFs mayores que cero, y considerando que *CF*1, ... *CFn* son los factores de certidumbre de las implicaciones correspondientes, el efecto conjunto de todas ellas sobre la hipótesis *H* responde a la expresión:

$$CF\ (H,E)=\sum_{i}^{n} CFi - \sum_{\substack{i \\ i<j}}^{n} CFi\ \ CFj\ + \sum_{\substack{i \\ i<j<k}}^{n} CFi\ \ CFj\ \ CFk - ...$$

Obsérvese la alternancia de signos en la expresión anterior, y obsérvese también que la segunda condición de las sumas evita duplicar los productos de las factores de certidumbre involucrados.

Esta aproximación parece razonable y, de hecho, no tiene ninguna fisura teórica. No obstante, fueron los propios Shortliffe y Buchanan quienes inmediatamente propusieron un modelo alternativo. La razón de ello es la falta de asociatividad de la formulación y las consecuencias que de dicha falta de asociatividad se derivan.

En primer lugar, el orden en que aparecen las evidencias modifica considerablemente el resultado final. Sin embargo, en opinión de los autores de este texto, Shortliffe y Buchanan fueron demasiado rigurosos consigo mismos. Más concretamente, la falta de asociatividad de un modelo puede ser hasta una ventaja... ¿qué ocurre en aquellos dominios en los que hay relaciones temporales y para los cuales el orden de aparición de las evidencias es realmente importante?

La segunda objeción, que también está relacionada con la falta de asociatividad, tiene que ver con la gran sensibilidad de la formulación ante la aparición de evidencias contradictorias en estados avanzados del proceso de razonamiento.

Vamos a ilustrar este problema con un ejemplo:

---

[45] Los otros dos casos son susceptibles de un tratamiento similar.

Sean 8 reglas que apoyan a una misma hipótesis *H*, cuyos factores certidumbre están entre 0.4 y 0.8, y sea una evidencia, esta vez en contra de la hipótesis, con un factor de certidumbre, pongamos por caso de -0.6.

En este caso, debido a la rapidez con que el modelo converge asintóticamente a uno, la aparición sucesiva de 8 "éxitos" genera un factor de certidumbre $CF_{12345678} = 0.99$, sin embargo, la aparición tardía de una evidencia "moderadamente contradictoria" ($CF_9 = -0.6$), hace caer la certidumbre total final en la hipótesis hasta $CF_{123456789} = 0.39$.

Por el contrario, si la evidencia negativa hubiese aparecido al principio del proceso, el resultado final hubiese sido muy distinto. Por ejemplo:

$$CF_{192345678} = 0.99$$

Ni Shortliffe ni Buchanan consideraron aceptable esta situación, y fueron ellos mismos quienes propusieron una segunda aproximación que paliaba estas "deficiencias". La nueva formulación que ambos propusieron fue la siguiente:

- Si $CF(h, e1) > 0$, y $CF(h, e2) > 0$, entonces:

$$CF(h, e1 \text{ y } e2) = CF(h, e1) + CF(h, e2) - [CF(h, e1) \times CF(h, e2)]$$

- Si $CF(h, e1) < 0$, y $CF(h, e2) < 0$, entonces:

$$CF(h, e1 \text{ y } e2) = CF(h, e1) + CF(h, e2) + [CF(h, e1) \times CF(h, e2)]$$

- Si $CF(h, e1) \times CF(h, e2) < 0$, entonces

$$CF(h, e1 \text{ y } e2) = \frac{CF(h, e1) + CF(h, e2)}{1 - \min\{|CF(h, e1)|, |CF(h, e2)|\}}$$

Esta nueva forma de combinar evidencias referidas a una misma hipótesis sí que es asociativa y, por lo tanto, las evidencias se pueden considerar en cualquier orden sin que el resultado final se vea afectado. Además presenta una ventaja (en la que no profundizaremos por exceder la cuestión las pretensiones de este texto), y es que nos permite modelizar procesos de razonamiento, sin tener que almacenar explícitamente los *MB*s y los *MD*s.

Quedan, no obstante, algunos problemas por resolver. Así, el modelo supone implícitamente independencia condicional de las evidencias. Por ello, si "*e*1" implica lógicamente a "*e*2", entonces $CF(h, e1 \text{ y } e2)$ DEBERÍA ser igual a $CF(h, e1)$, pero de la aplicación del modelo no se deduce este resultado. Esta situación constituye un problema no resuelto de la combinación de evidencias. Al respecto, Shortliffe y Buchanan proponen algunas alternativas como estructurar muy bien las bases de conocimientos, o agrupar en una sola regla cláusulas con evidencias condicionalmente dependientes. En todo caso, tales soluciones pertenecen más al ámbito de la ingeniería del conocimiento que al de la inteligencia artificial.

1.3. Propagación de Incertidumbre

Hasta ahora se ha considerado que la evidencia *e*, que estaba relacionada con la hipótesis, era un hecho (a favor o en contra de la hipótesis), que no venía afectado de incertidumbre. Por ello, el factor de certidumbre correspondiente *CF*(*h*, *e*) podía ser interpretado como la potencia evidencial de la implicación: $e \rightarrow h$, de tal forma que:

- si *CF*(*h*,*e*) = 1, entonces la evidencia *e* implica lógicamente a la hipótesis *h*
- si *CF*(*h*, *e*) = -1, la evidencia implica lógicamente a la negación de *h*
- si *CF*(*h*, *e*) = 0, la evidencia es independiente de la hipótesis

Sin embargo, cuando tratamos de representar conocimiento para luego llevar a cabo procesos de razonamiento, lo más normal, y también lo más correcto, es tratar de establecer relaciones causales basadas en "hechos inciertos". Ilustraremos la idea con un ejemplo:

Si establecemos la relación causal imprecisa

SI *e* ENTONCES *h*, con *CF*(*h*,*e*)

la imprecisión asociada está referida a la implicación subyacente. En otras palabras, siempre que tenemos *e* podemos concluir *h* con una incertidumbre asociada que viene dada por el factor de certidumbre. Lo normal, sin embargo, es que en los problemas reales la propia evidencia venga afectada de una cierta imprecisión, que en el curso de un proceso inferencial va a contribuir a modificar la incertidumbre de las conclusiones alcanzadas.

Nótese que hay una sutil diferencia entre "imprecisión" e "incertidumbre". La imprecisión es una característica que afecta a las entidades, hechos y/o datos del dominio, mientras que la incertidumbre es una característica ligada a los procesos de razonamiento. En cualquier caso, si basamos nuestro proceso de razonamiento en hechos imprecisos, las conclusiones que obtengamos serán inciertas, o no podrán establecerse con total certidumbre. Nos acabamos de topar con uno de los problemas más interesantes de la IA: la propagación de incertidumbre, problema que está ligado al concepto de *entropía de la información*[46].

Este problema surge, fundamentalmente, por dos circunstancias que pueden darse aislada o conjuntamente:

- La propia evidencia es imprecisa

- La evidencia considerada es, en realidad, consecuencia de otra regla, y forma parte de un proceso de razonamiento que supone varias inferencias

Un ejemplo que puede ilustrar la primera situación podría ser el siguiente:

---

[46] ... según el cual la información tiende a degradarse en la medida que es utilizada.

Sea la siguiente regla:

SIEMPRE QUE HACE MUY BUEN DIA VOY A LA PLAYA

La implicación subyacente puede, en este caso, traducirse del siguiente modo:

E = Hacer muy buen día
H = Ir a la playa
CF(*H*,*E*) = 1.00 (la implicación no tiene incertidumbre)

El problema que se nos plantea ahora es ¿de qué manera afectaría a la hipótesis *H* el hecho de no tener *E*, sino "algo parecido" a *E*, pero que no es exactamente *E*? Por ejemplo, ¿cual sería nuestra decisión si *el día es bastante bueno* [47]?

La segunda circunstancia puede evidenciarse con este otro ejemplo:

Sean las siguientes reglas:

primera

CUANDO LLUEVE MUCHO CASI SIEMPRE ME QUEDO EN CASA

segunda

CUANDO ME QUEDO EN CASA SUELO LEER

En ambos casos podemos representar -arbitrariamente- el conocimiento del siguiente modo:

*E*1 = Llover mucho
*H*1 = Quedarse en casa
CF(*H*1,*E*1) = 0.95 -arbitrario- (hay incertidumbre en la implicación)

*E*2 = Quedarse en casa
*H*2 = Leer
CF(*H*2,*E*2) = 0.75 -arbitrario- (hay incertidumbre en la implicación)

En este caso ¿cuál es la incertidumbre final que afecta a la hipótesis *H*2 -leer- sabiendo que se da la evidencia *E*1= llueve mucho?

Nótese que ahora la situación es más complicada, ya que hay un efecto de "arrastre" de la incertidumbre.

La situación combinada, según la cual aparecen simultáneamente imprecisión e incertidumbre por arrastre (situación que es, con diferencia, la más común), puede

---

[47] El problema de asignar etiquetas lingüísticas a números se denomina procesado simbólico, sobre el cual volveremos más adelante.

visualizarse en el ejemplo anterior suponiendo que en lugar de *llover mucho* la situación es que *llueve bastante*[48].

Shortliffe y Buchanan proponen un esquema en el que la primera circunstancia comentada -imprecisión-, puede considerarse como un caso particular de la segunda. Así:

- Si E' Entonces E con $CF(E, E') = x$
- Si E Entonces H con $CF(H, E) = y$

genera el siguiente circuito inferencial:

$$E' \rightarrow E \rightarrow H$$

y para resolverlo postulan:

$$CF(H, E') = CF(H, E) \times \max [0, CF(E, E')]$$

Esta formulación presenta una dificultad no prevista. Cuando $CF(E,E')$ es menor que cero, y por lo tanto E' apoya a la negación de E, la ecuación proporciona un $CF(H,E') = 0$, que como sabemos impide afirmar nada, ni positivo ni negativo, sobre la hipótesis considerada.

Este inconveniente del modelo fue estudiado por Heckerman, discípulo de Shortliffe y Buchanan, quien propuso una formulación alternativa basándose en el siguiente razonamiento:

Dado que $CF(\neg E, E') = -CF(E, E')$, la línea de razonamiento de nuestro circuito inferencial debería poder modificarse en los siguientes términos:

$$E' \rightarrow \neg E \rightarrow H$$

en lugar de

$$E' \rightarrow E \rightarrow H$$

para aquellos casos en los que $CF(E', E) < 0$, así, directamente podemos obtener el factor de certidumbre buscado:

$$CF(H, E') = CF(H, \neg E) \times \max [0, CF(\neg E, E')]$$

La modificación de Heckerman puede resumirse del siguiente modo:

$CF(H, E') = CF(h, E) \times CF(E, E') \Leftrightarrow CF(E, E') \geq 0$
$CF(H, E') = CF(h, \neg E) \times CF(\neg E, E') \Leftrightarrow CF(E, E') < 0$

---

[48] Esta cuestión nos remite otra vez al procesado simbólico, según el cual podríamos considerar que *llueve bastante* = *llueve mucho* con *CF*(llueve mucho) = 0.75 -arbitrario-

El esquema supuestamente alternativo propuesto por Heckerman no es más que una aplicación estricta del modelo de Shortliffe y Buchanan. No obstante este nuevo planteamiento adolece del mismo problema que hemos comentado en relación al esquema bayesiano. Precisamente Shortliffe y Buchanan establecieron que $CF(H, E) = -CF(\neg H, E)$ como un argumento en contra de la consistencia matemática de los modelos estadísticos cuando, en lugar de datos, empleamos conocimiento.

No obstante, ni Shortliffe ni Buchanan le dan demasiada importancia a esta peculiaridad de su modelo, y siguen pensando que cuando un experto afirma un hecho, no debe derivarse que su afirmación sirva también para cuantificar la negación del hecho afirmado. Por esta razón los autores de este texto prefieren seguir utilizando la formulación inicial de Shortliffe y Buchanan.

Un último aspecto relacionado con la propagación de incertidumbre es el relativo a la *combinación lógica de evidencias*. Recordemos que en los sistemas de producción[49], las cláusulas de los antecedentes de las reglas solían estar anidadas a través de los operadores lógicos AND, OR y NOT. El problema ahora es ¿cómo obtener los valores correspondientes a las expresiones:

- $CF(H1 \text{ and } H2, E)$
- $CF(H1 \text{ or } H2, E)$

a partir de $CF(H1, E)$ y de $CF(H2, E)$, en donde E es toda la evidencia del antecedente?

Sea por ejemplo:

$$\text{IF } \{E1 \text{ and } (E2 \text{ or } E3) \text{ and } E4\} \text{ THEN } H \text{ WITH } CF = x$$

Evidentemente el procedimiento seguido debe pasar por evaluar primero el antecedente, e inferir luego la conclusión teniendo en cuenta la potencia evidencial de la declaración correspondiente.

Así, si E representa la evidencia disponible (que puede no coincidir exactamente con la de la declaración), y teniendo en cuenta la propia regla, entonces:

$$CF(H, E) = CF(H, [E1 \text{ and } (E2 \text{ or } E3) \text{ and } E4]) \times CF([E1 \text{ and } (E2 \text{ or } E3) \text{ and } E4], E)$$

En esta expresión se tienen en cuenta simultáneamente la potencia evidencial de la declaración $-CF(H, [E1 \text{ and } (E2 \text{ or } E3) \text{ and } E4])-$, y la imprecisión en la información disponible de acuerdo con la estructura del antecedente de la regla $-CF([E1 \text{ and } (E2 \text{ or } E3) \text{ and } E4], E)-$.

Las funciones propuestas para evaluar los efectos de las conjunciones y de las disyunciones son las siguientes:

---

[49] ...y el modelo de los factores de certidumbre fue diseñado “ad hoc” para construir el sistema de producción MYCIN.

- $CF(H1 \text{ and } H2, E) = \min \{ CF(H1, E), CF(H2, E) \}$
- $CF(H1 \text{ or } H2, E) = \max \{ CF(H1, E), CF(H2, E) \}$

Ilustraremos el proceso global de razonamiento según el modelo de Shortliffe y Buchanan con el siguiente ejemplo: "- Dadas las siguientes reglas que vienen afectadas de incertidumbre:

R1:	IF { (*e*31 or *e*32) and *e*21}	THEN *H*	WITH *CF*r1 = 0.9
R2:	IF { *e*33 and *e*34 }	THEN *e*11	WITH *CF*r2 = 0.6
R3:	IF { *e*11 }	THEN *H*	WITH *CF*r3 = 0.8
R4:	IF { *e*22 }	THEN *e*12	WITH *CF*r4 = 0.8
R5:	IF { *e*23 }	THEN *e*12	WITH *CF*r5 = -0.6
R6:	IF { *e*12 }	THEN *H*	WITH *CF*r6 = 0.8

y dadas las siguientes evidencias imprecisas:

eA = *e*21 with *CF*(*e*21) = 1.0
eB = *e*31 with *CF*(*e*31) = 0.7
eC = *e*32 with *CF*(*e*32) = 0.8
eD = *e*33 with *CF*(*e*33) = -1.0
eE = *e*34 with *CF*(*e*34) = 0.8
eF = *e*22 with *CF*(*e*22) = 0.7
eG = *e*23 with *CF*(*e*23) = 0.6

calcular el factor de certidumbre de la hipótesis *H* considerando toda la evidencia disponible."

En este ejemplo es fácil ver que el circuito inferencial, resultante de considerar conjuntamente todas las reglas, puede descomponerse en tres ramas diferentes:

<u>rama 1:</u>

[*e*21 and (*e*31 or *e*32)] $\rightarrow$ *H*

<u>rama 2:</u>

(*e*33 and *e*34) $\rightarrow$ *e*11 $\rightarrow$ *H*

<u>rama 3:</u>

*e*22 $\rightarrow$ *e*12 $\rightarrow$ *H*

*e*23 ——————↑

Resolviendo cada rama por separado obtenemos que

<u>Solución para la rama 1:</u>

Debemos evaluar *CF*(*H*, evidencia_rama1) que, de acuerdo con la formulación inicial de Shortliffe y Buchanan:

CF(*H*, evidencia_rama1) =
= *CF*[*H*, *e*21 and (*e*31 or *e*32)] ×
× max { 0, *CF*[*e*21 and (*e*31 or *e*32), evidencia_rama1] }

Evidentemente, *CF*[*H*, *e*21 and (*e*31 or *e*32)] no es más que la potencia evidencial de la regla R1 y, por lo tanto:

$$CF[H, e21 \text{ and } (e31 \text{ or } e32)] = 0.9$$

Por otra parte, *CF*[*e*21 and (*e*31 or *e*32), evidencia_rama1] se evalúa teniendo en cuenta lo anteriormente mencionado para resolver el problema de la combinación lógica de evidencias. Así:

CF[*e*21 and (*e*31 or *e*32), evidencia_rama1] =
= min { *CF*(*e*21, eA), max [*CF*(*e*31, eB), *CF*(*e*32, eC)] } =
= min { 1.0, max (0.7, 0.8) } = min {1.0, 0.8 } = 0.8

por lo que:

$$CF(H, \text{evidencia\_rama1}) = 0.9 \times \max \{ 0, 0.8 \} = 0.72$$

<u>Solución para la rama 2:</u>

Aquí el problema es algo diferente, ya que aparece propagación. De acuerdo con la propuesta de Shortliffe y Buchanan:

$$CF(H, \text{evidencia\_rama2}) = CF(H, e11) \times \max \{ 0, CF(e11, \text{evidencia\_rama2}) \}$$

como antes, *CF*(*H*, *e*11) es la potencia evidencial de la regla R3, y vale

$$CF(H, e11) = 0.8$$

El problema es ahora calcular *CF*(*e*11, evidencia_rama2) que, de acuerdo con nuestro modelo:

CF(*e*11, evidencia_rama2) =
= *CF*(*e*11, *e*33 and *e*34) ×
× max { 0, *CF*(*e*33 and *e*34, evidencia_rama2) }

Nuevamente, *CF*(*e*11, *e*33 and *e*34) es la potencia evidencial de una regla, en este caso R2, y su valor es:

$$CF(e11, e33 \text{ and } e34) = 0.6$$

Continuando el proceso, nos queda por evaluar: *CF*(*e*33 and *e*34, evidencia_rama2) que, de acuerdo con la formulación para la combinación lógica de evidencias:

CF(*e*33 and *e*34, evidencia_rama2) =
= min [*CF*(*e*33, eD), *CF*(*e*34, eE)] =
= min [-1.0, 0.8] = -1.0

Efectuando ahora las sustituciones oportunas:

$$CF(e11, \text{evidencia\_rama2}) = 0.6 \times \max \{ 0, -1.0 \} = 0$$

y, evidentemente:

$$CF(H, \text{evidencia\_rama2}) = 0$$

Por lo que se puede concluir que la rama 2 no contribuye a la veracidad de la hipótesis *H*.

Solución para la rama 3:

Ahora aparece un proceso de combinación de evidencias que se refieren ambas a la misma hipótesis, y uno posterior de propagación.

CF(*H*, evidencia_rama3) =
= *CF*(*H*, *e*12) × max { 0, *CF*(*e*12, evidencia_rama3) }

CF(*H*, *e*12) es la potencia evidencial de la regla R6, y su valor es 0.8.

Por otra parte, *CF*(*e*12, evidencia_rama3) es una combinación de dos evidencias que apuntan a la misma hipótesis.

De este modo:

CF(*e*12, eF) = *CF*(*e*12, e22) × max { 0, *CF*(*e*22, eF) } =
= 0.8 × max [0, 0.7] = 0.56

Además:

$$CF(e12, eG) = CF(e12, e23) \times \max \{ 0, CF(e23, eG) \} = -0.6 \times \max [0, 0.7] = -0.36$$

El resultado parcial en esta rama 3 es una combinación de evidencias de signo contrario que, según la formulación asociativa del modelo, se resuelve del siguiente modo:

CF(*e*12, evidencia_rama3) =

$$= \frac{CF(e12, eF) + CF(e12, eG)}{1 - \min\{|CF(e12, eF)|, |CF(e12, eG)|\}} =$$

$$= \frac{(0.56 - 0.36)}{(1 - 0.36)} = \frac{0.20}{0.64}$$

y por lo tanto,

$$CF(H, \text{evidencia\_rama3}) = 0.8 \times \max \{ 0, [0.20/0.64]\} = 0.25$$

Solución final

Dado que:

- $CF(H, \text{evidencia\_rama1}) = 0.72$
- $CF(H, \text{evidencia\_rama2}) = 0.00$
- $CF(H, \text{evidencia\_rama3}) = 0.25$

estamos ante un nuevo caso de combinación de evidencias referidas todas a la misma hipótesis -*H*-, en donde dos de las evidencias contribuyen positivamente a la veracidad de la hipótesis, mientras que la otra no contribuye en absoluto, ni favorable ni desfavorablemente.

En este caso:

$CF(H, \text{evidencia\_total}) =$
$CF(H, \text{evidencia\_rama1}) + CF(H, \text{evidencia\_rama3}) -$
$- [CF(H, \text{evidencia\_rama1}) \times CF(H, \text{evidencia\_rama3})] =$
$= 0.72 + 0.25 - (0.72 \times 0.25) = 0.79$

Aunque ya se ha comentado que el modelo carece de base teórica, la realidad es que su utilización suele conducir a excelentes resultados. No obstante, el mayor o menor éxito del modelo puede estar condicionado por la labor de ingeniería del conocimiento, como también hemos mencionado.

1.4. Resumen

En este tema se desarrolla un modelo que permite el razonamiento con conocimiento impreciso, incierto o inexacto. El modelo de los factores de certidumbre de Shortliffe y Buchanan es un modelo diseñado ad hoc para la construcción del sistema experto MYCIN, y aborda el problema de la incertidumbre y del conocimiento impreciso definiendo dos medidas independientes: la medida de confianza creciente -índice dinámico que representa el incremento en la confianza de una hiipótesis dada una evidencia, y la medida de desconfianza creciente -índice dinámico que representa el incremento en la desconfianza de una hipótesis dada una evidencia-. Ambas medidas se resumen formalmente en el llamado factor de certidumbre. Tras discutir algunas propiedades de los tres índices que acabamos de mencionar, se demuestra que los factores de certidumbre son conceptualmente diferentes a las probabilidades condicionadas. A continuación se discute el modo según el cual se combinan evidencias y se propaga la incertidumbre en el modelo de Shortliffe y Buchanan. Este último aspecto se ilustra con un ejemplo.

1.5. Textos básicos

- Heckerman, “Probabilistic Interpretation for MYCIN’s Certainty Factors”, Uncertainty in Artificial Intelligence, 1986.

- Shortliffe, Buchanan “A Model of Inexact Reasoning in Medicine”, Mathematical Biosciencies, vol.23, 1975.

1. TEORÍA EVIDENCIAL

A diferencia del modelo de los factores de certidumbre de Shortliffe y Buchanan, el esquema de razonamiento propuesto en su día por Dempster y Shafer sí tiene una fuerte base teórica, hasta el punto que, lo que inicialmente era un modelo de razonamiento propuesto por Dempster, se convirtió en una verdadera teoría tras la formalización de Shafer.

Este esquema de razonamiento es atractivo, entre otras razones, porque:

- permite modelizar de forma sencilla la incertidumbre asociada a evidencias e hipótesis.
- permite considerar conjuntos de hipótesis sin que la confianza depositada en cada uno de ellos tenga que ser distribuida de ningún modo entre cada una de las hipótesis individuales del conjunto.
- permite reflejar de forma elegante la falta de conocimiento tan frecuentemente ligada a los procesos de razonamiento.
- contiene a la teoría de la probabilidad como un caso particular.
- contiene a algunas de las funciones combinatorias de evidencias del modelo de los factores de certidumbre.

Pero... ¿cómo se puede manejar el conocimiento inexacto y la falta de conocimiento en el modelo de Dempster y Shafer?

1.1. La Teoría Evidencial de Dempster y Shafer

En primer lugar, y dado un universo de discurso cualquiera, Dempster y Shafer introducen el concepto de *marco de discernimiento*, que definen como "... el conjunto finito de todas las hipótesis que se pueden establecer en el dominio del problema". El marco de discernimiento debe formar un conjunto completo, y por tanto exhaustivo, de hipótesis mútuamente excluyentes.

Por otra parte, el efecto de una determinada evidencia sobre el conjunto global de hipótesis no viene determinado por la contribución de la confianza depositada en las hipótesis individuales. Por el contrario, el efecto de cada evidencia afecta generalmente a un subconjunto de hipótesis del marco de discernimiento. Este planteamiento es coherente con la realidad de casi todos los problemas cotidianos[44]. En problemas reales, lo más normal es que las evidencias permitan discriminar entre grupos de hipótesis

[44] ... recuérdese que ya hemos comentado lo difícil que resulta encontrar evidencias que confirmen a una y sólo una hipótesis.

alternativas, manteniéndose, sin embargo, la incertidumbre entre las hipótesis individuales. Según este planteamiento:

- θ es el marco de discernimiento
- A es un subconjunto cualquiera del marco
- $h1, ..., hn$ son las hipótesis del marco

Ahora podemos establecer fácilmente el conjunto $\Gamma\theta$ de todos los subconjuntos posibles del marco. En este contexto, la aparición de una determinada evidencia e favorecerá a un determinado subconjunto A de θ, de forma que el grado en que A se vea favorecido se representa por $m(A)$, en donde m es indicativo de la confianza que la evidencia e permite depositar en A. m se denomina *función básica de asignación de verosimilitud*, y toma valores en el intervalo cerrado [0, 1]. Al respecto, utilizaremos la siguiente notación:

$$e: A = \{ha, hb, hc\} \rightarrow m(A) = x, \text{ con } x \in [0, 1]$$

El hecho de que la evidencia e apoye al subconjunto A no implica, como ya hemos dicho, que las hipótesis individuales se repartan de forma explícita la confianza depositada en la propia A. Esto constituye una diferencia notable con respecto a la teoría clásica de la probabilidad, según la cual, si $h1$, $h2$, $h3$ y $h4$ son las cuatro únicas hipótesis posibles en un dominio dado:

Si $p(h1, h2) = 0.80$

de alguna manera estamos afirmando que:

$p(h1) = 0.40$

$p(h2) = 0.40$

Este mismo ejemplo contemplado bajo la óptica de la teoría evidencial tendría el siguiente tratamiento:

$\theta = \{ h1, h2, h3, h4 \}$

$\Gamma\theta = \{ø, (h1), (h2), (h3), (h4), (h1, h2), (h1, h3), (h1, h4), (h2, h3), (h2, h4), (h3, h4), (h1, h2, h3), (h1, h2, h4), (h1, h3, h4), (h2, h3, h4), (h1, h2, h3, h4) \}$

$\Gamma\theta$ contiene a los 16 subconjuntos posibles que se pueden establecer con las cuatro hipótesis iniciales del marco de discernimiento. Nótese que en $\Gamma\theta$ están incluidos el conjunto vacío $\{ø\}$, y el propio marco $\{ \theta = (h1, h2, h3, h4) \}$

Si elegimos un subconjunto A del marco, por ejemplo $A = (h1, h2, h3)$, de tal manera que:

$$e: A = (h1, h2, h3) \rightarrow m(A) = 0.75,$$

lo único que se afirma es que, dada la evidencia *e*, la verosimilitud de *A* es 0.75. Es claro que ninguna de las hipótesis individuales de *A* se ve afectada por esta asignación de verosimilitud, ya que cualquier otra hipótesis, o cualquier otro conjunto de hipótesis son, en realidad, subconjuntos diferentes de Γθ, independientes -en principio- de la evidencia *e*.

Todo subconjunto del marco de discernimiento para el cual, dada una evidencia *e*, se verifique que $m(A) \neq 0$, se denomina *elemento focal*.

Volviendo por un instante a la función básica de asignación de verosimilitud, Dempster y Shafer definen las siguientes condiciones para *m*:

- $\sum_{A \subset \phi} m(A) = 1$
- $m(\text{ø}) = 0$

Ambas condiciones son consecuencia de las restricciones impuestas al marco de discernimiento[45].

Decíamos también que la teoría evidencial proporciona un medio elegante para tratar la falta de conocimiento asociada a los procesos de razonamiento. Supongamos un marco de discernimiento θ y una evidencia tal que:

$$e: A \subset \theta \rightarrow m(A) = s, \text{ con } 0 \leq s \leq 1$$

La primera condición exigida a *m* establece que $\sum_{A \subset \phi} m(A) = 1$, entonces ...¿qué pasa con el resto de confianza que no ha sido asignada al elemento focal *A*?

Al respecto, Dempster y Shafer postulan que:

Si: $e: A \subset \theta \rightarrow m(A) = s, \text{ con } 0 \leq s \leq 1$
Entonces: $m(\theta) = 1 - m(A) = 1 - s$

Esta formulación debe interpretarse del siguiente modo: puesto que la evidencia *e* supone la asignación de una confianza dada a un determinado elemento focal *A* del marco, el resto de la confianza no asignada representa "falta de conocimiento" y, por lo tanto, debe ser asignada al propio marco de discernimiento. Con estas premisas podemos reflexionar del siguiente modo:

- La confianza no asignada es ignorancia o falta de conocimiento sobre el grado de importancia de la evidencia en relación al elemento focal considerado. En otras palabras, se sabe que la evidencia apoya al elemento focal en un grado *s*; sin embargo, la confianza no asignada (1-*s*), no sabemos si contribuye o no a *A* (o a cualquier otro subconjunto del marco.)

---

[45] Recuérdese que el marco debe ser un conjunto completo de hipótesis mutuamente excluyentes.

- La confianza no asignada (1-*s*) debe asignarse al marco ya que, por construcción del esquema, lo que sí sabemos es que la solución está en el marco.

La formulación completa de la aproximación es la siguiente:

θ = marco de discernimiento = $\{h1,..., hn\}$
A = elemento focal, $A \subset \theta$
e = evidencia referida a A
$m(A)$ = medida de asignación básica de verosimilitud de A, dado e
e: $A \rightarrow m(A) = s$
$m(\theta) = 1 - s$
$m(B) = 0 \qquad \forall B \subset \theta, B \neq \theta, B \neq A$

Si el planteamiento fuese probabilístico, la misma evidencia apoyaría al elemento focal A y al complementario del elemento focal:

$$p(A) = s \rightarrow p(\neg A) = 1 - s$$

y recordemos que éste era uno de los aspectos más débiles de los modelos probabilísticos. Con esta nueva teoría:

Si: $\theta = \{h1, h2, h3, h4\}$
y: $A = \{h1, h2\}$
con: e: $A \rightarrow m(\{h1, h2\}) = s$
entonces: $m(\{h1, h2, h3, h4\}) = 1 - s$

Generalizando un poco estas ideas puede afirmarse, sin equivocarnos demasiado, que el procedimiento según el cual se maneja la falta de información en la teoría evidencial corrige las carencias de los modelos probabilísticos.

Siguiendo el mismo esquema de presentación que en temas anteriores, consideraremos ahora el modo en que las evidencias aparecen en los problemas del mundo real, y cómo la teoría evidencial trata este tipo de situaciones.

Parece claro que en problemas reales las evidencias no vienen solas. Más aún, distintas evidencias no necesariamente tienen por qué referirse a los mismos elementos focales. Así si $\theta = \{h1, h2, h3, h4\}$ es nuestro marco de discernimiento, o conjunto completo de hipótesis mútuamente excluyentes posibles en nuestro dominio, el conjunto de todos los subconjuntos posibles del marco será:

$\Gamma\theta$ = { ø, ($h1$), ($h2$), ($h3$), ($h4$), ($h1$, $h2$), ($h1$, $h3$), ($h1$, $h4$), ($h2$, $h3$),
($h2$, $h4$), ($h3$, $h4$), ($h1$, $h2$, $h3$), ($h1$, $h2$, $h4$), ($h1$, $h3$, $h4$),
($h2$, $h3$, $h4$), ($h1$, $h2$, $h3$, $h4$) }

Evidentemente, el número de elementos de $\Gamma\theta$ es 2 elevado al número de elementos del marco: $\#\Gamma\theta = 2^{\#\theta}$. En nuestro caso 16.

Sea ahora una evidencia $e1$ tal que:

$$e1: A1 = (h1, h2) \text{ con } m1(h1, h2) = x$$

por lo tanto, $m1(h1, h2, h3, h4) = 1 - x$

Sea ahora una nueva evidencia $e2$ tal que:

$$e2: A2 = (h2, h3, h4) \text{ con } m2(h2, h3, h4) = y$$

por lo tanto, $m2(h1, h2, h3, h4) = 1 - y$

Finalmente, sea una tercera evidencia $e3$ tal que:

$$e3: A3 = (h1) \text{ con } m3(h1) = z$$

por lo tanto, $m3(h1, h2, h3, h4) = 1 - z$

Nótese que cada elemento focal se refiere a cada evidencia concreta. La cuestión estriba ahora en considerar el efecto conjunto de todas las evidencias... pero antes tenemos que saber algo más de la teoría evidencial.

1.2. Combinación de Evidencias en la Teoría de Dempster y Shafer

Si dos (o más) fuentes de información proporcionan sendas evidencias (e.g., $e1$ y $e2$), relativas a dos elementos focales (e.g., $A1$ y $A2$) de un mismo marco de discernimiento, las funciones de asignación básica de verosimilitud -i.e., $m1(A1)$ y $m2(A2)$- se combinan para dar una nueva función de asignación básica de verosimilitud -$m12$- que representa el efecto conjunto de ambas evidencias sobre la intersección de los elementos focales correspondientes. De esta forma:

Si $e1: A1 \rightarrow m1(A1) = x$

Si $e2: A2 \rightarrow m2(A2) = y$

definimos $m12(C) = m1(A1) \times m2(A2)$, donde $C = A1 \cap A2$.

Esta expresión puede generalizarse directamente para distintas parejas de elementos focales. Así:

$$m12(C) = \sum_{C=Ai \cap Bj} m1(Ai)\, m2(Bj)$$

Formulación que coincide con la de asignación de probabilidad a la intersección de dos sucesos independientes en la teoría clásica de la probabilidad. Por este motivo se puede afirmar que la teoría evidencial asume implícitamente independencia entre las evidencias. Por otra parte, es claro que la primera condición exigida a la función de asignación básica de verosimilitud se cumple:

$$\sum_{C=Ai \cap Bj} m12(C) = 1$$

Sin embargo, puede ocurrir que distintas evidencias "señalen" a elementos focales muy distintos, tan distintos que no tengan ningún elemento en común, y por lo tanto su intersección sea nula. Este caso introduce una peculiaridad en el modelo:

Si: $e1$: $A1 \rightarrow m1(A1) = x$, $0 < x \leq 1$
Si: $e2$: $A2 \rightarrow m2(A2) = y$, $0 < y \leq 1$
Si: $C = A1 \cap A2 = \emptyset$
Entonces: $m12(C) = m12(\emptyset) = m1(A1) \times m2(A2) = xy \neq 0$

resultado que contradice la segunda condición exigida a la función básica de asignación de verosimilitud, según la cual la solución *tiene* que estar en el marco de discernimiento. ¿Estamos ante una contradicción de la teoría evidencial? Aparentemente sí. Sin embargo estas situaciones ocurren muchas veces en edificios tan bien estructurados como la Física o la Química y, además, tienen una solución muy sencilla... Si aparece una inconsistencia que no debía aparecer ¡hay que eliminarla! Para ello, *por decreto* $m12(\emptyset)=0$, y si hay que corregir algo se corrige.

Este procedimiento se denomina *Normalización* y no es, en absoluto, ni artificioso ni extraño. Concretamente, en la teoría evidencial la normalización de resultados debe conseguir mantener las funciones de asignación básica de verosimilitud entre los límites definidos, lo que supondrá corregir las asignaciones a elementos focales de intersección no nula, de forma que su suma siga siendo la unidad. La nueva expresión para la combinación de evidencias es la siguiente:

$$m12(C) = \frac{\sum_{C = Ai \cap Bj} m1(Ai)\ \ m2(Bj)}{1 - \sum_{Ai \cap Bj = \phi} m1(Ai)\ \ m2(Bj)} = \frac{\sum_{C = Ai \cap Bj} m1(Ai)\ \ m2(Bj)}{\sum_{Ai \cap Bj \neq \phi} m1(Ai)\ \ m2(Bj)}$$

Claramente, la necesidad de normalización aparece cuando las fuentes de información aportan evidencias sobre elementos focales muy diferentes, tan diferentes que no comparten hipótesis individuales.

La expresión $K = \sum_{Ai \cap Bj = \phi} m1(Ai)\, m2(Bj)$ del denominador de la expresión normalizada para $m12(C)$ se denomina *grado de conflicto,* y precisamente es una medida de la compatibilidad existente entre las evidencias que están siendo combinadas. Considerando la expresión para el grado de conflicto, la ecuación de $m12(C)$ se puede reescribir de la siguiente forma:

$$m12(C) = \frac{\sum_{C = Ai \cap Bj} m1(Ai)\, m2(Bj)}{1 - K}$$

expresión en la que el factor $\frac{1}{1-K}$ se denomina *factor de normalización*.

Nótese que cuando las distintas evidencias señalan a distintos elementos focales entre los cuales no hay intersecciones nulas, no podemos hablar de evidencias

contradictorias. En este caso no hay conflictos, K=0, y la expresión normalizada para $m12(C)$ coincide con la expresión sin normalizar. Por el contrario, cuando las evidencias son totalmente contradictorias y todos los elementos focales son disjuntos entre sí, el conflicto es total, K=1, y la combinación de evidencias no está definida.

Analizaremos todas estas ideas por medio de un ejemplo en el cual vamos a tratar de obtener la mejor interpretación para una manifestación dada, en un dominio cualquiera, en el que todas las posibles soluciones son {$h1$, $h2$, $h3$, $h4$}, y para lo cual disponemos de la siguiente información:

- La evidencia $e1$ apoya a las interpretaciones $h1$ y $h2$ con: $m1(h1, h2) = 0.6$
- La evidencia $e2$ apoya a las interpretaciones $h2$, $h3$, $h4$ con: $m2(h2, h3, h4) = 0.7$
- La evidencia $e3$ apoya a la interpretación $h1$ con: $m3(h1) = 0.8$

El primer paso consiste en formalizar el problema en los términos que la teoría propone:

$\theta = \{h1, h2, h3, h4\}$

$e1$: $A1 = \{h1, h2\}$, $m1(A1) = 0.6$
$m1(\theta) = 0.4$
$m1(B1) = 0.0$, con $B1 \subset \theta$, $B1 \neq A1$, $B1 \neq \theta$

$e2$: $A2 = \{h2, h3, h4\}$, $m2(A2) = 0.7$
$m2(\theta) = 0.3$
$m2(B2) = 0.0$, con $B2 \subset \theta$, $B2 \neq A2$, $B2 \neq \theta$

$e3$: $A3 = \{h1\}$, $m3(A3) = 0.8$
$m3(\theta) = 0.2$
$m3(B3) = 0.0$, con $B3 \subset \theta$, $B3 \neq A3$, $B3 \neq \theta$

Antes de proceder a combinar evidencias (y dado que tal combinación implica la intersección de conjuntos con medidas de asignación básica de verosimilitud relativas a distintas evidencias), al objeto de evitar confusiones utilizaremos la siguiente notación:

$e1$: $A1 = \{h1, h2\}$, $m1(A1) = 0.6$
$m1(\theta_1) = 0.4$

$e2$: $A2 = \{h2, h3, h4\}$, $m2(A2) = 0.7$
$m2(\theta_2) = 0.3$

$e3$: $A3 = \{h1\}$, $m3(A3) = 0.8$
$m1(\theta_3) = 0.2$

Nótese que hemos prescindido de aquellos subconjuntos para los cuales "mx = 0", independientemente de la evidencia considerada. Nótese, asimismo, que hemos señalado los respectivos marcos de discernimiento con un subíndice indicativo de la

evidencia de que procede la medida de verosimilitud asignada, y que corresponde a la confianza no asignada al elemento focal considerado. Evidentemente:

$$\theta_1 = \theta_2 = \theta_3 = \theta = \{h1, h2, h3, h4\}$$

Procederemos ahora a combinar evidencias:

*e*1 y *e*2:

$A1 \cap A2 = \{h2\}$,	$m12\{h2\} = 0.6 \times 0.7 = 0.42$
$A1 \cap \theta_2 = \{h1, h2\}$,	$m12\{h1, h2\} = 0.6 \times 0.3 = 0.18$
$\theta_1 \cap A2 = \{h2, h3, h4\}$,	$m12\{h2, h3, h4\} = 0.4 \times 0.7 = 0.28$
$\theta_1 \cap \theta_2 = \theta$,	$m12\{\theta\} = 0.4 \times 0.3 = 0.12$

Obsérvese que las intersecciones deben realizarse sobre todos los subconjuntos para los cuales, en función de cada evidencia, su función básica de asignación de verosimilitud es no nula.

Nótese también que si:

C1 = $A1 \cap A2$, con $m12\{h2\} = 0.42$
C2 = $A1 \cap \theta_2$, con $m12\{h1, h2\} = 0.18$
C3 = $\theta_1 \cap A2$, con $m12\{h2, h3, h4\} = 0.28$
C4 = $\theta_1 \cap \theta_2$, con $m12\{\theta\} = 0.12$

entonces: $\sum_{i=1}^{4} m12(Ci) = 1$

lo cual no hace más que confirmar la primera condición establecida para la función básica de asignación de verosimilitud.

Consideremos ahora la tercera evidencia, *e*3, según la cual:

*e*3: $A3 = \{h1\}$,	$m3\{h1\} = 0.8$
	$m3\ \{\theta\} = 0.2$

La combinación de esta nueva evidencia con el resultado de la combinación de las evidencias anteriores produce los siguientes resultados:

*e*1 y *e*2 y *e*3:

$C1 \cap A3 = \emptyset$	con $m123\{\emptyset\}$	= 0.336
$C1 \cap \theta_3 = \{h2\}$	con $m123\{h2\}$	= 0.084
$C2 \cap A3 = \{h1\}$	con $m123\{h1\}$	= 0.144
$C2 \cap \theta_3 = \{h1, h2\}$	con $m123\{h1, h2\}$	= 0.036
$C3 \cap A3 = \emptyset$	con $m123\{\emptyset\}$	= 0.224

$C3 \cap \theta_3 = \{h2,h3,h4\}$ con $m123\{h2,h3,h4\} = 0.056$
$C4 \cap A3 = \{h1\}$ con $m123\{h1\} = 0.096$
$C4 \cap \theta_3 = \{\theta\}$ con $m123\{\theta\} = 0.024$

La primera condición exigida para la función básica de asignación de verosimilitud sigue cumpliéndose (i.e., la suma de los $m123$ de las correspondientes intersecciones sigue siendo 1); sin embargo, la segunda condición se vulnera en dos ocasiones, por lo que hay que normalizar las expresiones. En este caso, y recordando que el grado de conflicto se calcula según la expresión:

$$K = \sum_{Ai \cap Bj = \varphi} m1(Ai)\ \ m2(Bj)\text{, resulta } K = 0.336 + 0.224 = 0.560$$

Llevando este resultado a la ecuación que nos permite calcular la medida de asignación básica de verosimilitud normalizada (o lo que es lo mismo, dividiendo los resultados anteriores por 1 - K), obtenemos los siguientes resultados:

$m123\{h2\} = 0.084/0.44 = 0.191$
$m123\{h1\} = 0.144/0.44 = 0.327$
$m123\{h1, h2\} = 0.036/0.44 = 0.082$
$m123\{h2,h3,h4\} = 0.056/0.44 = 0.127$
$m123\{h1\} = 0.096/0.44 = 0.218$
$m123\{\theta\} = 0.024/0.44 = 0.055$

Agrupando ahora los términos correspondientes obtenemos:

$m123\{h1\} = 0.545$
$m123\{h2\} = 0.191$
$m123\{h1, h2\} = 0.082$
$m123\{h2,h3,h4\} = 0.127$
$m123\{\theta\} = 0.055$

Este resultado final indica que, aunque ninguna evidencia por separado permite afirmar nada sobre ninguna hipótesis individual del marco, la concurrencia de evidencias permite que dos hipótesis concretas -$h1$ y $h2$- se destaquen del resto. Por otra parte, la combinación de evidencias hace que la confianza todavía no asignada disminuya drásticamente en relación a la obtenida cuando consideramos evidencias individuales.

1.3. Credibilidad, Plausibilidad e Intervalo de Confianza

La teoría evidencial permite el seguimiento de la evolución dinámica de la confianza depositada en los subconjuntos del marco de discernimiento a medida que aparecen nuevas evidencias. Para ello se definen dos nuevas medidas, la *credibilidad* y la *plausibilidad* que, respectivamente, son un indicador de la mínima y de la máxima confianza que podemos depositar en un elemento focal dado.

Formalmente la credibilidad se define según la ecuación:

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

en donde A es el elemento focal considerado, subconjunto del marco de discernimiento. La credibilidad es una medida de las contribuciones que todos los subconjuntos de A ejercen sobre el propio A. Consideremos el ejemplo del apartado anterior, y calculemos la evolución de la credibilidad en el elemento focal $\{h1, h2\}$, a medida que van apareciendo las evidencias $e1$, $e2$ y $e3$.

Con $e1$, la credibilidad de $\{h1, h2\}$ coincide exactamente con la medida de asignación básica de verosimilitud, ya que -aparte del propio elemento focal considerado- no existe ningún subconjunto de $\{h1, h2\}$ con valores de m distintos de 0. Así:

$$e1: \{h1, h2\} \rightarrow m1\{h1, h2\} = Cr(h1, h2) = 0.6$$

Consideremos ahora la segunda evidencia combinada con la primera, y analicemos el mismo elemento focal. En este caso obteníamos las siguientes intersecciones:

$(h2)$	$\rightarrow$	$m12\ (h2)$	$= 0.42$
$(h1, h2)$	$\rightarrow$	$m12(h1, h2)$	$= 0.18$
$(h2, h3, h4)$	$\rightarrow$	$m12(h2, h3, h4)$	$= 0.28$
$(h1, h2, h3, h4)$	$\rightarrow$	$m12(h1, h2, h3, h4)$	$= 0.12$

por lo que, de acuerdo con la expresión propuesta:

$$Cr(h1, h2) = m12(h2) + m12(h1, h2) = 0.42 + 0.18 = 0.60$$

Este resultado coincide con el anterior -una única evidencia-, por lo que la credibilidad asociada al elemento focal considerado no se ha visto modificada tras la aparición de $e2$. Consideremos, finalmente, el efecto conjunto de las tres evidencias, que nos producía los siguientes resultados:

$(h1)$	$\rightarrow$	$m123(h1)$	$= 0.545$
$(h2)$	$\rightarrow$	$m123(h2)$	$= 0.191$
$(h1, h2)$	$\rightarrow$	$m123(h1, h2)$	$= 0.082$
$(h2, h3, h4)$	$\rightarrow$	$m123(h2, h3, h4)$	$= 0.127$
$(h1, h2, h3, h4)$	$\rightarrow$	$m123(h1, h2, h3, h4)$	$= 0.055$

Si evaluamos ahora la credibilidad de $(h1, h2)$ obtenemos:

$$Cr(h1, h2) = m123(h1) + m123(h2) + m123(h1, h2) = 0.545 + 0.191 + 0.082 = 0.818$$

Nuestra credibilidad en el elemento focal considerado (es decir, la mínima confianza que podemos depositar en él) ha aumentado tras considerar las tres evidencias conjuntamente.

La otra medida importante, la plausibilidad, es un indicador de la máxima confianza que podemos depositar en un elemento focal dado, y se calcula considerando también las contribuciones de otros subconjuntos con intersección no nula, de acuerdo con la expresión:

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

en donde A es el elemento focal considerado, subconjunto del marco de discernimiento.

La plausibilidad no sólo tiene en cuenta los subconjuntos del propio elemento focal, sino también todas las contribuciones de todos aquellos subconjuntos que tienen algo que ver con dicho elemento focal. Volviendo a nuestro ejemplo:

$e1$: $Pl(h1, h2) = m1(h1, h2) + m1(h1, h2, h3, h4) = 0.6 + 0.4 = 1.0$

$e1$ y $e2$: $Pl(h1, h2) = m12(h2) + m12(h1,h2) + m12(h2, h3, h4)$
$+ m12(h1, h2, h3, h4) = 1.0$

$e1$ y $e2$ y $e3$: $Pl(h1, h2) = m123(h1) + m123(h2) + m123(h1, h2) +$
$+ m123(h2, h3, h4) + m123(h1, h2, h3, h4) = 1.0$

En este caso la plausibilidad del elemento focal considerado no varía, y su valor es siempre la unidad. Ello indica que, siendo optimistas, la confianza que podemos depositar en todas o alguna de las hipótesis de dicho elemento focal es completa (lo cual, por cierto, se refleja en los valores del resultado final del ejemplo). En otros casos, sin embargo, ambas medidas -*Cr* y *Pl*- varían durante el proceso inferencial[46].

Una tercera medida importante es el llamado *Intervalo de Confianza*, que se construye, para cada elemento focal, a partir de la credibilidad y de la plausibilidad. Así, en cada nivel del proceso de razonamiento, el intervalo de confianza es el segmento del espacio numérico [0,1] que tiene como valor mínimo el valor de la credibilidad del elemento focal, y como valor máximo el correspondiente valor de la plausibilidad. De este modo, y recurriendo de nuevo a nuestro ejemplo, los intervalos de confianza del elemento focal ($h1$, $h2$) varían del siguiente modo:

$e1$: IC($h1$, $h2$) = [0.600, 1.000]
$e1$ y $e2$: IC($h1$, $h2$) = [0.600, 1.000]
$e1$ y $e2$ y $e3$: IC($h1$, $h2$) = [0.818, 1.000]

Observamos aquí que, al considerar conjuntamente las tres evidencias, el intervalo de confianza se hace más estrecho, a la vez que se aproxima a la unidad. Conceptualmente, el intervalo de confianza representa la incertidumbre asociada al elemento focal considerado. Además, se puede demostrar que si A es un elemento focal:

$0 \leq Cr(A) \leq 1$

---

[46] Analícese, por ejemplo, el caso del elemento focal ($h1$)

$0 \leq Pl(A) \leq 1$
$Cr(A) \leq Pl(A)$
$Cr(A) \leq \text{Prob}(A) \leq Pl(A)$

en donde Prob(A) es la probabilidad estadística.

Claramente,

Si $Cr(A) = 0$ y $Pl(A) = 1$, entonces la ignorancia sobre A es total.
Si $Cr(A) = Pl(A) = 1$, entonces A es absolutamente cierto.
Si $Cr(A) = Pl(A) = 0$, entonces A es absolutamente falso.

1.4. Casos Particulares de la Teoría Evidencial

La teoría evidencial contiene, bajo ciertos supuestos y en ciertas situaciones, al modelo de los factores de certidumbre de Shortliffe y Buchanan.

Analicemos el caso de dos evidencias independientes que apoyen al mismo elemento focal A de un determinado marco de discernimiento. En este caso, el tratamiento que le daríamos a nuestro problema sería el siguiente:

$e1$: $A1$, $m1(A1) = s1$, $m1(\theta_1) = 1 - s1$
$e2$: $A2$, $m2(A2) = s2$, $m2(\theta_2) = 1 - s2$

con $A1 = A2$ y $\theta_1 = \theta_2$

Combinando ambas evidencias obtenemos el siguiente resultado:

$e1$ y $e2$:

$A1 \cap A2 = A$, $m12(A) = s1 \times s2$
$A1 \cap \theta_2 = A$, $m12(A) = s1 \times (1 - s2) = s1 - s1 \times s2$
$\theta_1 \cap A2 = A$, $m12(A) = (1 - s1) \times s2 = s2 - s1 \times s2$
$\theta_1 \cap \theta_2 = \theta$, $m12(\theta) = (1 - s1) \times (1 - s2)$

Agrupando ahora las expresiones para A obtenemos:

$$m12(A) = s1\ s2 + s1 - s1\ s2 + s2 - s1\ s2 = s1 + s2 - s1\ s2$$

pero este resultado es exactamente el que obtendríamos si aplicáramos el modelo de Shortliffe y Buchanan en el caso de dos evidencias independientes que apoyan a la misma hipótesis, cada una de las cuales con su correspondiente factor de certidumbre. Así:

Si: $e1$ Entonces A con $CF(A,e1) = s1$, $s1 > 0$

Si: $e2$ Entonces A con $CF(A,e2) = s2$, $s2 > 0$

el resultado que obtendríamos para CF(*A*, *e*1 y *e*2) = *s*1 + *s*2 - *s*1 *s*2

En otros casos, sin embargo, la teoría evidencial proporciona resultados diferentes de los que obtendríamos aplicando el modelo de los factores de certidumbre. Más concretamente, la teoría evidencial funciona igual que el modelo de Shortliffe y Buchanan en aquellos casos en los que este modelo funciona bien, y lo mejora notablemente cuando el modelo de los factores de certidumbre presenta carencias. Al respecto, los autores de este texto dejan como ejercicio para sus sufridos lectores el análisis del siguiente problema: ¿Qué resultados se obtendrían aplicando la teoría evidencial y el modelo de Shortliffe y Buchanan cuando dos evidencias *e*1 y *e*2 apoyan respectivamente a *A* y a ¬*A*?

1.5. Resumen

La teoría evidencial de Dempster y Shafer presenta un esqueleto formal para tratar convenientemente tanto el conocimiento inexacto como la falta de conocimiento. La teoría trabaja sobre conjuntos de hipótesis, sin que la confianza aportada por una determinada evidencia tenga que distribuirse entre las hipótesis individuales del conjunto considerado. Cuando existen evidencias que apoyan a grupos de hipótesis contradictorias, el modelo normaliza los resultados correspondientes. Para ello se define un índice -grado de conflicto-, que cuantifica las incompatibilidades encontradas. El problema de la incertidumbre se trata a partir del llamado intervalo de confianza, que se construye a partir de dos medidas, la credibilidad y la plausibilidad, indicativas respectivamente de la mínima y la máxima confianza que, en un momento dado, podemos depositar en un conjunto determinado de hipótesis. El intervalo de confianza evoluciona dinámicamente a medida que van apareciendo nuevas evidencias. Un resultado particularmente interesante de la teoría evidencial es que contiene al modelo de Shortliffe y Buchanan en aquellos casos en los que dicho modelo funciona bien, y sin embargo lo mejora en aquellos casos en los que dicho modelo presenta deficiencias.

1.6. Textos básicos

- Shafer, “A Mathematical Theory of Evidence”, Princeton University Press, eds.,1976.

1. CONJUNTOS DIFUSOS

El refranero español, rico como es en expresiones populares, nos ofrece un dicho que nos va a permitir introducir la idea de *conjunto difuso.* El refrán en cuestión es el siguiente: *En este mundo nada es verdad y nada es mentira, todo es según el color del cristal con que se mira*. Esta es precisamente la noción de conjuntos difusos, para los cuales la descripción de objetos y entidades del mundo real debe realizarse según los criterios lingüísticos propios de los seres humanos.

La mayoría de las declaraciones humanas son ambiguas, y esta ambigüedad es una característica esencial, no sólo del lenguaje, sino también de los procesos de clasificación, del establecimiento de taxonomías y jerarquías, y de los procesos de razonamiento en sí mismos.

Así, si definimos al "ser vivo" como una estructura molecular organizada, que nace, crece, se reproduce y muere, parece claro que una remolacha participa de todas y cada una de las características anteriores, y por lo tanto debe ser considerada como un ser vivo. Por el contrario, una lámina de mica no participa de todas las características anteriores, y en consecuencia no debe ser considerada como un ser vivo, pero... ¿cómo debemos considerar a un virus? Los virus tienen, a veces, una estructura molecular bien organizada -cuando encuentran un medio apropiado-, y en tal estado son capaces de reproducir su estructura molecular y hacer réplicas de sí mismos. En otras condiciones su estructura se desnaturaliza, y se convierten en conjuntos moleculares amorfos e inertes. Cuando vuelven a encontrar un medio apropiado, estos conjuntos moleculares vuelven a organizarse y, por decirlo de algún modo, *cobran vida*. En ningún caso debemos confundir el proceso -por supuesto muy simplificado- que hemos descrito en relación a los virus, con los procesos de *letargo* ni con los de *enquistamiento*. En los procesos de letargo simplemente se produce una disminución de la actividad metabólica del individuo, mientras que en los procesos de enquistamiento no tiene lugar una desorganización celular ni, por supuesto, una desorganización molecular. Los virus son pues entes ambiguos respecto al concepto de *ser vivo*: A veces actúan como tales, otras claramente se comportan como seres inertes. Si tuviésemos que responder categóricamente a la pregunta: ¿Considera usted al virus del mosaico del tabaco como un ser vivo?, probablemente la mejor respuesta sería un indeciso -y poco clarificador- "depende."

La dificultad para clasificar a un virus en el conjunto de los seres vivos surge de la propia definición del concepto de ser vivo. En otros casos, sin embargo, la dificultad aparece por cuestiones de carácter subjetivo. Así, caracterizar el conjunto de *mujeres guapas*[44] no es fácil -cada uno tendrá una idea distinta de los atributos que debe tener una mujer "abstracta" para ser considerada hermosa-, pero será todavía mucho más difícil decir si una mujer concreta es guapa o no. En este último caso aparecen matices subjetivos que desvirtúan la clasificación.

---

[44] Los autores no son en absoluto sexistas. Lo dicho para las mujeres puede ser igualmente aplicado a los hombres sin ninguna diferencia, a favor o en contra.

Por último, no sólo problemas de definición, o matices subjetivos, hacen difícil una clasificación categórica. En otras ocasiones incluso los contextos pueden modificar los criterios. Así, el concepto de *hombre alto* -que es intínsecamente ambiguo-, es notablemente diferente entre los habitantes de Estocolmo y entre los Pigmeos... ¡y lo grave del asunto es que probablemente ambos tengan razón!

De lo que acabamos de exponer se puede concluir rápidamente que los conjuntos ordinarios, en los que un elemento de un universo determinado pertenece o no pertenece al conjunto, no nos bastan para representar el conocimiento habitualmente empleado por los humanos, y mucho menos para *razonar* con él.

Las Matemáticas y la Inteligencia Artificial, atentas como siempre a problemas interesantes relacionados con las ciencias cognoscitivas, no podían quedar al margen de esta peculiaridad, y en 1965 Lofti Zadeh hizo públicos sus trabajos relacionados con este tema en su famoso artículo "Fuzzy Sets."

1.1. Aspectos Generales de los Conjuntos Difusos

Consideremos un universo cualquiera. Por ejemplo el universo formado por el conjunto *N* de los números naturales. Definamos un subconjunto *A* de *N* caracterizado por la siguiente descripción:

"*A* es el conjunto formado por los números naturales pares menores que diez"

El subconjunto *A* de *N* queda perfectamente definido del siguiente modo:

$$A = \{ 2, 4, 6, 8 \}$$

Claramente:

$2 \in A$
$3 \notin A$
$10 \notin A$

En este caso no tenemos ningún problema para establecer el grado de pertenencia de un elemento del universo de discurso con respecto al subconjunto considerado.

Consideremos ahora el universo *C* caracterizado por la siguiente descripción:

"*C* es el conjunto formado por todos los seres humanos vivos",

y sea *B* un subconjunto de *C* caracterizado por la descripción:

"*B* es el subconjunto de *C* de hombres morenos y altos"

En este caso sí tenemos problemas a la hora de establecer el grado de pertenencia de un elemento del universo al subconjunto *B* considerado.

Claramente, un conjunto ordinario puede definirse como una colección de elementos. Si un elemento del universo está representado en la colección, el elemento en cuestión pertenece a dicho conjunto. En estos casos se puede decir que el grado de pertenencia de un elemento cualquiera del referencial[45] tiene un valor booleano, de forma que:

- Si el elemento pertenece al conjunto, el valor booleano es "1"

- Si el elemento no pertenece al conjunto, el valor booleano es "0"

De este modo, podemos construir una función "f", que para conjuntos ordinarios es una función booleana, tal que dado un elemento "*x*" del referencial *U*, y dado un subconjunto *A* de *U*:

- $f_A(x) = 1 \Leftrightarrow x \in A$
- $f_A(x) = 0 \Leftrightarrow x \notin A$

Ampliaremos ahora la cuestión a ese tipo especial de conjuntos que hemos denominado *conjuntos difusos*. En este caso decíamos que matices de carácter lingüístico, subjetivo,... nos impedían establecer con claridad el grado de pertenencia de algunos elementos del referencial al conjunto difuso considerado. Así, habrá elementos del referencial que claramente pertenezcan al conjunto, habrá otros que claramente no pertenezcan, y habrá otros que *pertenezcan en cierto grado* -aunque no totalmente-.

Siguiendo con el planteamiento anteriormente esbozado, el problema es muy fácil de resolver si consideramos que la función *f* adopta los siguientes valores, dado un elemento *x* del referencial *U*, y un subconjunto difuso *A* de *U*:

- $f_A(x) = 1 \quad \Leftrightarrow \quad x \in A$
- $f_A(x) = 0 \quad \Leftrightarrow \quad x \notin A$
- $0 < f_A(x) < 1 \quad \Leftrightarrow$ *x* pertenece en cierto grado a *A*

La función *f* de algún modo cuantifica el grado de pertenencia de un elemento del referencial al conjunto difuso considerado. Así, un conjunto difuso es aquel en el que no existe una frontera clara entre la pertenecia y la no pertenencia de determinados elementos del referencial.

De todas formas, para establecer los "límites difusos" del conjunto correspondiente, vamos a necesitar un criterio, que casi siempre va a ser arbitrario. Analicemos el siguiente ejemplo:

Consideremos el referencial *U* de *personas vivas*, y consideremos también el subconjunto difuso *A* de *U* definido por la etiqueta "*A* es el conjunto de personas vivas jóvenes". Una propiedad que nos parece adecuada para caracterizar al subconjunto difuso *A* es la *edad* de los elementos del referencial, pero... ¿con qué criterio? Nos

---

[45] Referencial es el sinónimo de Universo de Discurso tradicionalmente empleado cuando se habla de conjuntos difusos.

encontramos ante el problema no trivial de la definición de criterios para la "fuzzyficación" de conjuntos. En nuestro caso, y continuando con el ejemplo, consideraremos "jóvenes" a todos aquellos elementos del referencial cuya edad les permita adquirir -legalmente- el "inter-rail"[46], y "no jóvenes" a todos aquellos del referencial que puedan beneficiarse -también legalmente- de la "tarjeta de la tercera edad" de RENFE (respectivamente, ≤ 25 años, y ≥ 65 años). De este modo:

- $f_A(x) = 1 \ \forall x$ / Edad$(x) \ \leq \ 25$
- $f_A(x) = 0 \ \forall x$ / Edad$(x) \ \geq \ 65$

Pero ¿qué pasa ahora con todos aquellos elementos del referencial con edades comprendidas entre 25 y 65 años? ¿cuál es su "grado de juventud" en base al criterio "edad"?... Estamos ante un nuevo problema: la caracterización de la zona difusa. Para salir del embrollo vamos a construir una función lineal del siguiente modo:

- $f_A(x) = \dfrac{65 - \text{Edad}\,(x)}{65\text{-}25} = \dfrac{65 - \text{Edad}\,(x)}{40} \qquad \forall x/\text{Edad}(x) \in [25,65]$

Con esta aproximación hemos sido capaces de segmentar nuestro espacio numérico [0 - 1] en tres zonas; dos de las cuales son no difusas y se refieren a aquellos elementos del referencial que pertenecen, o que no pertenecen, al subconjunto difuso considerado; y una tercera zona, de carácter difuso, que se refiere a aquellos elementos del referencial que pertenecen en cierto grado al subconjunto difuso considerado. La Figura 1.1 ilustra el tratamiento efectuado.

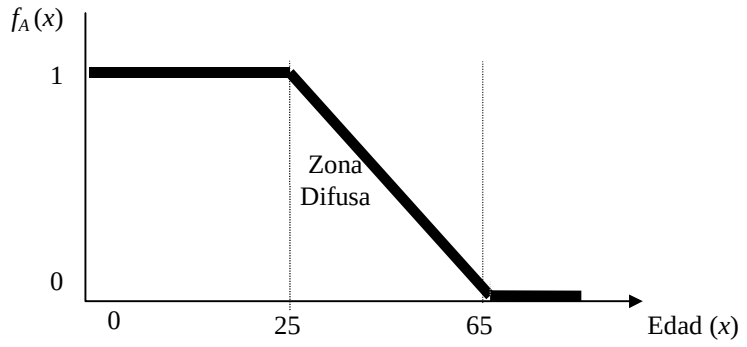

Figura 1.1

Planteémonos ahora la siguiente cuestión: sabiendo que Juan tiene 17 años, Marisa tiene 31 años, Joaquín tiene 47 años, Aurora tiene 57 años, y Marcial tiene 73 años, y sabiendo que todos ellos son personas vivas, ¿qué podemos decir acerca de su "juventud"?

[46] ... un excelente medio para viajar por toda Europa en tren, durante un mes y a un precio muy razonable.

De acuerdo con los criterios establecidos y efectuando las sustituciones oportunas, los valores de sus respectivas funciones de pertenencia al subconjunto difuso de "personas vivas jóvenes" serían los siguientes:

- $f_{\text{joven}}(\text{Juan}) = 1.00$
- $f_{\text{joven}}(\text{Marisa}) = 0.85$
- $f_{\text{joven}}(\text{Joaquín}) = 0.45$
- $f_{\text{joven}}(\text{Aurora}) = 0.20$
- $f_{\text{joven}}(\text{Marcial}) = 0.00$

Es claro que a medida que la edad de nuestros amigos aumenta, su grado de pertenencia al subconjunto difuso de *personas vivas jóvenes* disminuye. La aproximación es coherente, pero no es muy natural en términos lingüísticos. Para ilustrar lo que queremos decir analicemos el siguiente diálogo:

- Oye Luis ¿qué edad tiene Marisa?
- Creo que tiene 31 años
- ¡Ah, entonces Marisa es 0.85 joven!

¡Absurdo!, nadie se expresaría de tal modo. Estamos ante un nuevo problema, el de la *clasificación lingüística* con conjuntos difusos. Al respecto, la idea básica es que, una vez hemos sido capaces de segmentar el espacio numérico -indicativo de los grados de pertenencia de los elementos del referencial al subconjunto difuso considerado-, debemos segmentar también el espacio lingüístico, estableciendo un conjunto determinado de etiquetas dotadas de contenido semántico, y hacer corresponder a cada etiqueta lingüística un intervalo numérico concreto según un criterio mínimamente razonable. Existen estudios teóricos que tratan de demostrar que la máxima imprecisión lingüística puede conseguirse a través de una escala semántica formada por no más de nueve elementos literales[47].

Volvamos al caso de nuestro ejemplo, y definamos la siguiente escala lingüística, a la que asociaremos valores concretos de nuestra función grado de pertenenencia de los elementos del referencial al subconjunto difuso considerado. De este modo:

- $f_A(x) = 0.00 \rightarrow$ no es
- $0.00 < f_A(x) < 0.20 \rightarrow$ es muy poco
- $0.20 \leq f_A(x) \leq 0.40 \rightarrow$ es poco
- $0.40 < f_A(x) < 0.60 \rightarrow$ es algo
- $0.60 \leq f_A(x) \leq 0.80 \rightarrow$ es moderadamente
- $0.80 < f_A(x) < 1.00 \rightarrow$ es bastante
- $f_A(x) = 1.00 \rightarrow$ es

Según esta escala, y con los datos del ejemplo, ahora podríamos decir que:

---

[47] Sobre esta cuestión volveremos un poco más adelante.

- Juan es joven
- Marisa es bastante joven
- Joaquín es algo joven
- Aurora es poco joven
- Marcial no es joven

expresiones que se ajustan mucho más a la forma normal que utilizaríamos para emitir un juicio acerca de la juventud de nuestros amigos.

Llamaremos ahora la atención sobre algunos aspectos relacionados con este tratamiento:

- La función $f_A(x)$ definida para la zona difusa, que aquí hemos construido lineal, podría haber sido definida de cualquier otra forma[48].

- La escala lingüística asociada al espacio numérico -que es arbitraria- depende, no obstante, del tipo de clasificación que queremos obtener. Así, si queremos saber algo sobre la eventualidad de un determinado hecho, una escala semántica apropiada podría ser la siguiente:

  { imposible, casi imposible, muy improbable,..., casi seguro, seguro }

- El número de elementos semánticos de nuestra escala lingüística también es arbitrario.

- En algunos casos -como en nuestro ejemplo- es posible definir conjuntos difusos *lingüísticamente complementarios* que hagan más natural la expresión verbal. De este modo podríamos definir al subconjunto difuso *B* de "personas vivas viejas", -del referencial de "personas vivas"- de forma que alguien que sea "bastante joven" sea, simultáneamente "muy poco viejo", o lo que es más aceptable desde un punto de vista semántico, alguien que sea "muy poco joven" sea, al mismo tiempo "bastante viejo".

- Cualquier conjunto, sea cual sea su naturaleza, es "fuzzyficable"[49]; es decir, se puede establecer una gradación entre los niveles de pertenencia de distintos elementos de un referencial con respecto al conjunto considerado.

1.2. Caracterización y Nomenclatura de los Conjuntos Difusos

Cualquier conjunto, sea difuso o sea ordinario, tiene que poder ser descrito de manera conveniente. En el caso de los conjuntos ordinarios, y dado que se puede establecer sin ambigüedades la correspondiente relación de pertenencia de los elementos del referencial al conjunto considerado, resulta equivalente caracterizar al

---

[48] Dentro de un orden, por supuesto. ¡¡¡Una función helicoidal podría no ser del todo apropiada!!!

[49] Horrenda palabra para la que desgraciadamente no encontramos equivalente castellano... tal vez "difuminable".

conjunto en cuestión en función de su dominio (e.g., A es el conjunto de los números naturales pares menores que diez), o haciendo explícitos los elementos que lo constituyen (e.g., $A = \{ 2, 4, 6, 8 \}$)

Por otra parte ya hemos visto que, para cada elemento de un referencial dado, podemos definir una función f -que será de carácter booleano en el caso de conjuntos ordinarios-, tal que a cada elemento del referencial le asignará su valor lógico correspondiente, 0 ó 1, según el elemento en cuestión pertenezca, o no pertenezca, al conjunto. Así pues:

Dado un referencial U, y sea $A \subset U$,

$$\exists\, f_A(x) = 1 \Leftrightarrow x \in A$$
$$= 0 \Leftrightarrow x \notin A$$

Aplicando este criterio al conjunto ordinario A que nos sirve de ejemplo, A estará perfectamente determinado con la siguiente expresión:

$$f_A(x) = \{ f_A(1) = 0 + f_A(2) = 1 + f_A(3) = 0 + f_A(4) = 1 + f_A(5) = 0 + f_A(6) = 1 + f_A(7) = 0 + f_A(8) = 1 + f_A(9) = 0 + f_A(10) = 0 + ...\}$$

expresión en la que el signo "+" se lee "Y"

Una expresión equivalente a la anterior, pero algo más simplificada, es la siguiente:

$$f_A(x) = \{ 0/1 + ½ + 0/3 + ¼ + 0/5 + 1/6 + 0/7 + 1/8 + 0/9 + 0/10 + ...\}$$

en donde los numeradores de las fracciones representan los valores de la función de grado de pertenencia, y los denominadores son los elementos del referencial considerado. De este modo, un subconjunto ordinario A de un referencial U puede ser descrito:

- Implícitamente
- Explícitamente
- Mediante una $f_A(x)$, booleana, $\forall x \in U$

Por razones obvias, cuando trabajemos con conjuntos difusos preferiremos emplear descripciones implícitas o utilizar las funciones de grado de pertenencia. En este último caso tendremos en cuenta que se pierde el carácter booleano de la mencionada función. De esta forma:

$$\forall A \subset U \,/\, A \text{ difuso} \rightarrow \exists\, f_A(x) \,/\, f_A(x) : U \rightarrow [0,1] \quad \forall x \in U$$

es decir, la función $f_A(x)$ puede tomar cualquier valor en el intervalo [0,1]. Como veremos más adelante, esta forma de nombrar a los conjuntos difusos nos lleva directamente a establecer que los conjuntos ordinarios son un caso particular de los conjuntos difusos, aunque -como también veremos- no tienen la misma estructura algebraica.

Siguiendo con esta forma de denotar conjuntos, y suponiendo que queremos establecer una relación entre un conjunto *A* de un referencial *U*, y un conjunto *B* de un referencial *V*, también serán equivalentes expresiones del tipo:

A relación *B* = (a1,...,an) relación (b1,...,bm) = $f_A(x)$ relación $f_B(y)$,
con $A \subset U$, $B \subset V$, $x \in U$, y $\in V$

1.3. Estructura Algebraica de los Conjuntos Difusos

Para investigar la estructura algebraica de los conjuntos difusos se tienen que verificar un conjunto de propiedades, que trataremos de establecer y desarrollar a continuación. No obstante, el punto de enfoque de esta sección es tratar de demostrar que los conjuntos difusos no tienen estructura de álgebra de Boole. Al respecto, comenzaremos a establecer las propiedades correspondientes.

Conjunto vacío

- Sea un referencial *U*, y sea Z $\subset$ *U* tal que

$\exists\ f_Z(x) : U \rightarrow [0,1] \quad \forall x$ de U [50]

decimos que Z = ø $\Leftrightarrow\ f_Z(x) = 0 \quad \forall x \in U$

Identidad

- Sea un referencial *U*, y sean $A \subset U$ y $B \subset U$ tales que

$\exists\ f_A(x) : U \rightarrow [0,1] \quad \forall x \in U$
$\exists\ f_B(x) : U \rightarrow [0,1] \quad \forall x \in U$

decimos que $A = B \Leftrightarrow f_A(x) = f_B(x) \quad \forall x \in U$

Complementariedad

- Sea un referencial *U*, y sea $A \subset U$ tal que

$\exists\ f_A(x) : U \rightarrow [0,1] \quad \forall x \in U$

decimos que $A' = Ac$ = complementario de $A \Leftrightarrow f_A'(x) = 1 - f_A(x) \quad \forall x \in U$

Obviamente, $f_A'(x) : U \rightarrow [0,1]$

---

[50] Esta expresión define al subconjunto Z de *U* como un conjunto difuso.

Inclusión

- Sea un referencial U, y sean $A \subset U$ y $B \subset U$ tales que

$\exists\, f_A(x) : U \rightarrow [0,1] \quad \forall x \in U$
$\exists\, f_B(x) : U \rightarrow [0,1] \quad \forall x \in U$

decimos que $B \subset A \Leftrightarrow f_B(x) \leq f_A(x)\ \forall x \in U$

Esta caracterización es totalmente análoga a la que obtendríamos si considerásemos conjuntos ordinarios, y los describiésemos con notación difusa. Así:

- Sea U = {1, 2, 3, 4, 5, 6, 7, 8 }
- Sea A en U / A = { 1, 2, 3, 4 }
- Sea B en U / B = { 1, 3 }

evidentemente $B \subset A$. Si denotamos ahora a los mencionados subconjuntos de U por medio de sus funciones de grado de pertenencia resulta:

- $f_A(x)$ = { 1/1 + 1/2 + 1/3 + 1/4 + 0/5 + 0/6 + 0/7 + 0/8 }
- $f_B(x)$ = { 1/1 + 0/2 + 1/3 + 0/4 + 0/5 + 0/6 + 0/7 + 0/8 }

Efectivamente, podemos comprobar que $\forall x \in U$, $f_B(x) \leq f_A(x)$

Unión de conjuntos difusos

- Sea un referencial U, y sean $A \subset U$, $B \subset U$, y $C \subset U$ tales que

$\exists\, f_A(x) : U \rightarrow [0,1] \quad \forall x \in U$
$\exists\, f_B(x) : U \rightarrow [0,1] \quad \forall x \in U$
$\exists\, f_C(x) : U \rightarrow [0,1] \quad \forall x \in U$

decimos que $C = A \cup B \Leftrightarrow f_C(x) = \max \{f_A(x), f_B(x)\} \quad \forall x \in U$

Ilustraremos el concepto con un ejemplo -de naturaleza similar al empleado para describir la relación de *inclusión*-, en el que A, B y C son los conjuntos ordinarios que se muestran a continuación:

- A = { 1, 2 }
- B = { 1, 3, 5 }
- C = { 1, 2, 3, 5 }

en donde A, B, y C son subconjuntos del referencial U = { 1, 2, 3, 4, 5 } y, claramente, $C = A \cup B$.

Utilizando ahora la notación difusa:

- $f_A(x)$ = { 1/1 + 1/2 + 0/3 + 0/4 + 0/5 }

- $f_B(x)$ = { 1/1 + 0/2 + 1/3 + 0/4 + 1/5 }
- $f_C(x)$ = { 1/1 + 1/2 + 1/3 + 0/4 + 1/5 }

pero, según esta expresión, constatamos fácilmente que:

$$f_C(x) = \max \{ f_A(x), f_B(x) \} \ \forall x \in U$$

Al respecto, se puede demostrar que la unión de conjuntos difusos, que también puede denotarse del siguiente modo: $A \cup B = f_A(x)$ or $f_B(x) \ \ \forall x \in U$, o simplemente "$f_A$ or f_B", tiene la propiedad asociativa, por lo que:

- Dado un referencial U, y dados $A \subset U$, $B \subset U$, $C \subset U$, tales que

$\exists\, f_A(x) : U \rightarrow [0,1] \ \ \forall x \in U$
$\exists\, f_B(x) : U \rightarrow [0,1] \ \ \forall x \in U$
$\exists\, f_C(x) : U \rightarrow [0,1] \ \ \forall x \in U$

se verifica que $A \cup (B \cup C) = (A \cup B) \cup C$, expresión que equivale a la siguiente:

$$f_A \text{ or } (f_B \text{ or } f_C) = (f_A \text{ or } f_B) \text{ or } f_C$$

Intersección de conjuntos difusos

- Sea un referencial U, y sean $A \subset U$, $B \subset U$, y $C \subset U$ tales que

$\exists\, f_A(x) : U \rightarrow [0,1] \ \ \forall x \in U$
$\exists\, f_B(x) : U \rightarrow [0,1] \ \ \forall x \in U$
$\exists\, f_C(x) : U \rightarrow [0,1] \ \ \forall x \in U$

decimos que $C = A \cap B \Leftrightarrow f_C(x) = \min \{f_A(x), f_B(x)\} \ \ \forall x \in U$

El mismo tratamiento que hemos efectuado en el párrafo anterior para ilustrar la Unión, puede utilizarse para ilustrar la Intersección. La intersección de conjuntos difusos también puede denotarse del siguiente modo:

$$A \cap B = f_A(x) \text{ and } f_B(x) \ \ \forall x \in U, \text{ o simplemente "} f_A \text{ and } f_B \text{"}$$

Leyes de DeMorgan

- Sea un referencial U, y sean $A \subset U$, $B \subset U$ tales que

$\exists\, f_A(x) : U \rightarrow [0,1] \ \ \forall x \in U$
$\exists\, f_B(x) : U \rightarrow [0,1] \ \ \forall x \in U$

las leyes de DeMorgan establecen lo siguiente:

- Primera: El complementario de la Unión equivale a la intersección de los complementarios.

$$(A \cup B)' = A' \cap B'$$

$$(f_A \text{ or } f_B)' = (f_A' \text{ and } f_B')$$

- Segunda: El complementario de la Intersección equivale a la Unión de los complementarios.

$$(A \cap B)' = A' \cup B'$$

$$(f_A \text{ and } f_B)' = (f_A' \text{ or } f_B')$$

Analicemos brevemente la primera de estas leyes:

Sea un referencial U, y sean $A \subset U$, $B \subset U$ tales que

$\exists\, f_A(x) : U \to [0,1] \quad \forall x \in U$
$\exists\, f_B(x) : U \to [0,1] \quad \forall x \in U$

$A \cup B \quad \to \quad f_{\cup}(x) = \max \{ f_A(x), f_B(x) \} \qquad \forall x \in U$
$(A \cup B)' \to f_{\cup'}(x) = 1 - \max \{ f_A(x), f_B(x) \} \qquad \forall x \in U$

por otra parte

$A' \to f_A'(x) = 1 - f_A(x) \qquad \forall x \in U$
$B' \to f_B'(x) = 1 - f_B(x) \qquad \forall x \in U$

$A' \cap B' \to f'\cap'(x) = \min \{ 1 - f_A(x), 1 - f_B(x) \} \qquad \forall x \in U$

Lo que tendremos que demostrar ahora es que:

$$1 - \max \{ f_A(x), f_B(x) \} = \min \{ 1 - f_A(x), 1 - f_B(x) \} \qquad \forall x \text{ de } U$$

para lo cual utilizaremos una prueba basada en el análisis de casos extremos.

(a) Supongamos que $f_A(x) \geq f_B(x) \quad \forall x \in U$

en este caso se cumple que $1 - \max \{ f_A(x), f_B(x) \} = 1 - f_A(x) \quad \forall x$ de U y además,
$\min \{ 1 - f_A(x), 1 - f_B(x) \} = 1 - f_A(x) \quad \forall x$ de U

por lo tanto, bajo estas condiciones, la primera ley de DeMorgan se verifica.

(b) Sea ahora $f_A(x) \leq f_B(x) \quad \forall x \in U$

en este caso se cumple que $1 - \max \{ f_A(x), f_B(x) \} = 1 - f_B(x) \quad \forall x \in U$ y además,
$\min \{ 1 - f_A(x), 1 - f_B(x) \} = 1 - f_B(x) \quad \forall x \in U$

así que, bajo estas condiciones, la primera ley de DeMorgan se verifica también. Por lo tanto, si en ambas situaciones extremas dicha ley se cumple, también debe cumplirse para las situaciones intermedias[51].

Este mismo planteamiento puede seguirse en relación a la segunda ley de DeMorgan, y encontraríamos que también se verifica.

Leyes distributivas

Aunque no haremos un tratamiento formal y completo de las expresiones correspondientes[52], los conjuntos difusos también verifican las leyes distributivas.

- primera:

Dado un referencial U, y dados $A \subset U$, $B \subset U$, $C \subset U$, tales que

$\exists\, f_A(x) : U \to [0,1]\ \forall x \in U$
$\exists\, f_B(x) : U \to [0,1]\ \forall x \in U$
$\exists\, f_C(x) : U \to [0,1]\ \forall x \in U$

la primera ley distributiva establece que:

$$C \cap (A \cup B) = (C \cap A)\ \cup (C \cap B)$$

o, lo que es lo mismo,

$$f_C(x) \text{ and } \{f_A(x) \text{ or } f_B(x)\} = \{f_C(x) \text{ and } f_A(x)\} \text{ or } \{f_C(x) \text{ and } f_B(x)\}\ \forall x \in U$$

- segunda:

Dado un referencial U, y dados $A \subset U$, $B \subset U$, $C \subset U$, tales que

$\exists\, f_A(x) : U \to [0,1]\ \ \forall x \in U$
$\exists\, f_B(x) : U \to [0,1]\ \ \forall x \in U$
$\exists\, f_C(x) : U \to [0,1]\ \ \forall x \in U$

la segunda ley distributiva establece que:

$$C \cup (A \cap B) = (C \cup A)\ \cap (C \cup B)$$

o, lo que es lo mismo,

$$f_C(x) \text{ or } \{f_A(x) \text{ and } f_B(x)\} = \{f_C(x) \text{ or } f_A(x)\} \text{ and } \{f_C(x) \text{ or } f_B(x)\}\ \forall x \in U$$

---

[51] Esta no es una demostración formal. El lector interesado encontrará mejores demostraciones en los artículos de la bibliografía.

[52] Ejercicio que dejamos para el lector estudioso.

Según lo visto hasta ahora todo parece indicar que los conjuntos difusos tienen estructura de álgebra de Boole; sin embargo, hay dos leyes del álgebra de Boole que los conjuntos difusos no satisfacen. En concreto, los conjuntos difusos no cumplen ni el *principio de no contradicción*, ni la *ley del tercero excluido*.

Dado un referencial U, y dado $A \subset U$, en donde A es un conjunto ordinario, la ley del tercero excluido establece que la unión de un conjunto y su complementario es el referencial:

$$A \cup A' = U$$

Por otra parte, dado un referencial U, y dado $A \subset U$, en donde A es un conjunto ordinario, el principio de no contradicción establece que la intersección de un conjunto con su complementario es el conjunto vacío:

$$A \cap A' = \emptyset$$

En el caso de los conjuntos difusos, y en relación a la ley del tercero excluido, si U es un referencial, y A un subconjunto difuso del referencial -y por lo tanto $\exists\ f_A(x) : U \rightarrow [0,1] \ \ \forall x \in U$-, dado que $A' \rightarrow f_A'(x) = 1 - f_A(x)$ por lo tanto

$$A \cup A' \rightarrow f_{A \cup A'}(x) = \max \{ f_A(x), f_A'(x) \} = \max \{ f_A(x), 1 - f_A(x) \} \quad \forall x \in U$$

cantidad que es siempre $\geq 1/2$, pero que no necesariamente es “1”.

Análogamente, en el caso del principio de no contradicción, si U es un referencial, y A un subconjunto difuso del referencial, y por lo tanto $\exists\ f_A(x) : U \rightarrow [0,1]$, $\forall x \in U$, dado que $A' \rightarrow f_A'(x) = 1 - f_A(x)$ por lo tanto

$$A \cap A' \rightarrow \ f_{A \cap A'}(x) \quad = \ \min \{ f_A(x), f_A'(x) \} = \min \{ f_A(x), 1 - f_A(x) \} \ \forall x \in U$$

cantidad que es siempre $\leq 1/2$, pero que no necesariamente es “0”.

Claramente, los conjuntos difusos no verifican las dos leyes anteriormente mencionadas, y por lo tanto no tienen estructura de álgebra de Boole -lo cual va a traernos alguna sorpresa cuando trabajemos con ellos,... como veremos a continuación-

1.4. Operaciones Algebraicas con Conjuntos Difusos

El desarrollo efectuado hasta ahora nos permite describir algunas operaciones algebraicas que podemos realizar con conjuntos difusos. La descripción de tales operaciones se realizará a partir de las correspondientes funciones de grado de pertenencia.

Producto de conjuntos difusos

- Sea un referencial U, y sean $A \subset U$, $B \subset U$ tales que

$\exists\, f_A(x) : U \rightarrow [0,1] \quad \forall x \in U$
$\exists\, f_B(x) : U \rightarrow [0,1] \quad \forall x \in U$

definimos el producto de ambos subconjuntos difusos del siguiente modo:

$$A \; x \; B \rightarrow f_{AB}(x) = f_A(x) \cdot f_B(x) \qquad \forall x \in U$$

Para ilustrar el producto analizaremos el siguiente ejemplo: sea el referencial $U = \{1, 2, 3, 4\}$ y sean $A \subset U$ y $B \subset U$ tales que:

$f_A(x) = \{ 0/1 + 0.3/2 + 0.7/3 + 1/4\}$
$f_B(x) = \{ 0.5/1 + 0.4/2 + 1/3 + 0/4 \}$

De acuerdo con nuestra definición anterior, el producto $A \; x \; B$ estará caracterizado por la siguiente función de grado de pertenencia:

$f_{AB}(x) = \{ 0/1 + 0.12/2 + 0.7/3 + 0/4 \}$

Analicemos ahora un problema similar, pero con conjuntos ordinarios, de forma que las correspondientes funciones de grado de pertenencia le asignan el valor "1" a todos los elementos del referencial que tienen valores distintos de "0" en sus correspondientes homólogos difusos. Según este criterio:

A difuso $\rightarrow A$ ordinario $= A_o / f_{Ao}(x) = \{ 0/1 + 1/2 + 1/3 + 1/4 \}$
B difuso $\rightarrow B$ ordinario $= B_o / f_{Bo}(x) = \{ 1/1 + 1/2 + 1/3 + 0/4 \}$

El producto de los conjuntos ordinarios correspondientes vendrá descrito por la expresión:

$A_o \; x \; B_o \rightarrow f_{Ao \; x \; Bo}(x) = \{ 0/1 + 1/2 + 1/3 + 1/4 \}$

Vamos a calcular ahora la intersección de los conjuntos ordinarios *Ao* y *Bo*. En este caso:

$A_o \cap B_o \rightarrow f_{Ao \cap Bo}(x) = \min \{ f_{Ao}(x), f_{Bo}(x) \} \; \forall x \in U$

de donde: $f_{Ao \cap Bo}(x) = \{ 0/1 + 1/2 + 1/3 + 0/4 \}$

expresión que coincide exactamente con la obtenida para el producto. De este modo, podemos afirmar que en los conjuntos ordinarios el producto coincide con la intersección.

Observemos ahora qué ocurre en el caso de los conjuntos difusos, para ello recordemos que:

A difuso $= Ad \rightarrow f_{Ad}(x) = \{ 0/1 + 0.3/2 + 0.7/3 + 1/4\}$
B difuso $= Bd \rightarrow f_{Bd}(x) = \{ 0.5/1 + 0.4/2 + 1/3 + 0/4 \}$

y la intersección de ambos conjuntos difusos está caracterizada por la siguiente función de grado de pertenencia:

$Ad \cap Bd \rightarrow f_{Ad \cap Bd}(x) = \min \{f_{Ad}(x), f_{Bd}(x)\} \quad \forall x \in U$

de donde: $f_{Ad \cap Bd}(x) = \{ 0/1 + 0.3/2 + 0.7/3 + 0/4 \}$

Si comparamos el resultado obtenido para el producto con el resultado obtenido para la intersección, en el caso de conjuntos difusos observamos que:

$f_{AB}(x) \leq f_{A \cap B}(x) \qquad \forall x \in U$

pero esta expresión es, precisamente, la que define la relación de inclusión en los conjuntos difusos. Por lo tanto, el producto de conjuntos difusos está contenido en su intersección[53].

Suma y suma acotada

- Sea un referencial U, y sean $A \subset U$, $B \subset U$ tales que

$\exists\, f_A(x) : U \rightarrow [0,1] \qquad \forall x \in U$
$\exists\, f_B(x) : U \rightarrow [0,1] \qquad \forall x \in U$

definimos la suma de conjuntos difusos del siguiente modo:

$A + B \rightarrow f_{A+B}(x) = f_A(x) + f_B(x) \qquad \forall x \in U$

Claramente, la suma de conjuntos difusos sólo está definida cuando

$f_A(x) + f_B(x) \leq 1 \qquad \forall x \in U$

Para evitar este problema, definimos el concepto de *suma acotada* de conjuntos difusos del siguiente modo:

$A \,|+|\, B \rightarrow f_{A\,|+|\,B}(x) = \min \{ 1, f_A(x) + f_B(x) \} \quad \forall x \in U$

expresión que está definida siempre, de acuerdo con las restricciones impuestas a la función de grado de pertenencia.

Diferencia y diferencia absoluta

- Sea un referencial U, y sean A en U, B en U tales que

$\exists\, f_A(x) : U \rightarrow [0,1] \qquad \forall x \in U$
$\exists\, f_B(x) : U \rightarrow [0,1] \qquad \forall x \in U$

---

[53] No es trivial interpretar este resultado.

Definimos la diferencia de ambos conjuntos difusos del siguiente modo:

$A - B \rightarrow f_{A-B}(x) = f_A(x) - f_B(x) \qquad \forall x \in U$

Evidentemente, la diferencia de conjuntos difusos sólo está definida si $f_B(x) \leq f_A(x)$ $\forall x \in U$; es decir, la diferencia de conjuntos difusos sólo se puede establecer cuando $B \subset A$.

Para evitar este problema, en los modelos difusos se define el concepto de *diferencia absoluta* del siguiente modo:

$|A - B| \rightarrow f_{|A-B|}(x) = |f_A(x) - f_B(x)| \qquad \forall x \in U$

expresión que está definida siempre, de acuerdo con las restricciones impuestas a la función de grado de pertenencia.

Núcleo de un conjunto difuso

- Sea un referencial *U*, y sea *A* en *U*, tal que

$\exists\, f_A(x) : U \rightarrow [0,1] \qquad \forall x \in U$

definimos como *núcleo* del conjunto difuso *A*, a todo aquel elemento del referencial, para el cual $f_A(x) = 1$. Así:

$NA = \{\, x \in U \,/\, f_A(x) = 1 \,\}$

Un conjunto difuso se dice que está *normalizado* si tiene núcleo; es decir, si existe algún elemento del referencial que claramente pertenezca al conjunto difuso considerado.

Relación difusa

Dado un referencial *U*, definimos una relación difusa de orden "n" en *U*, como un conjunto difuso *A* en el espacio *U x U x ... x U* (n veces), caracterizado por una función de grado de pertenencia del tipo:

$f_A(x1, ..., xn) \qquad \forall x \in U$

Para visualizar el concepto consideremos el siguiente ejemplo: Sea el referencial $U = \{ 1, 2, 3, 4 \}$ y definamos una relación difusa de orden 2, caracterizada por la descripción "*A* es el conjunto difuso de los elementos que son aproximadamente iguales". Evidentemente, este conjunto difuso está definido en el espacio *U x U*, y su función de grado de pertenencia[54] podría ser la siguiente:

---

[54] Arbitraria...

	U2			
U1	**1**	**2**	**3**	**4**
1	1.0	0.8	0.2	0.0
2	0.8	1.0	0.8	0.2
3	0.2	0.8	1.0	0.8
4	0.0	0.2	0.8	1.0

Así la relación difusa establecida en los términos anteriores, relativa a los elementos 1 y 3 es la siguiente:

f_A (1, 3) = 0.2

1.5. Representación del Conocimiento y Razonamiento Difuso

Aunque un tratamiento amplio de los problemas derivados de la representación del conocimiento y del razonamiento difuso sobrepasa con mucho las pretensiones de este texto, sí parece conveniente iniciar al menos una aproximación a ambas cuestiones. Todos los conceptos vistos hasta ahora nos han permitido caracterizar a los conjuntos difusos y distinguirlos de los conjuntos ordinarios. Recordemos ahora que, desde la perspectiva de la inteligencia artificial, el modelo difuso debe permitirnos la representación de declaraciones como las siguientes:

- Normalmente se tarda alrededor de 45 minutos en llegar desde La Coruña a Santiago, por la autopista, y con tráfico ligero.

- No es previsible que el paro disminuya en España, al menos de manera drástica, en los próximos meses.

- La mayoría de los expertos opinan que la probabilidad de un terremoto serio en la zona del Caurel es muy pequeña en un futuro inmediato.

En las frases anteriores podemos reconocer predicados difusos, cuantificadores difusos y probabilidades difusas. Al respecto, otras aproximaciones más convencionales, usualmente empleadas para representar conocimiento, carecen de medios para representar eficazmente el significado de conceptos difusos. Así, los modelos que se basan en lógicas de primer orden, o los que se basan en teorías clásicas de la probabilidad, no nos permiten manipular correctamente el conocimiento de sentido común. Las causas evidentes son las siguientes:

- El conocimiento derivado del sentido común es léxicamente impreciso.
- El conocimiento derivado del sentido común es de naturaleza no categórica.

Por otra parte, las características ya estudiadas de los conjuntos difusos nos dan pistas sobre la manera de proceder, si lo que queremos es aplicar esquemas de representación del conocimiento, y modelos de razonamiento, basados en lógica difusa:

- En lógica difusa, el razonamiento categórico es un caso particular del razonamiento aproximado.

- En lógica difusa todo es una cuestión de grado.

- Cualquier sistema lógico puede ser “fuzzyficado”.

- En lógica difusa el conocimiento debe ser interpretado como una colección de *restricciones difusas* que operan sobre una colección de variables.

- En lógica difusa, los problemas de razonamiento -y por consiguiente los procesos inferenciales- deben interpretarse como *propagaciones* de las restricciones difusas mencionadas en el párrafo anterior.

Este último punto es de vital importancia, y merece que nos detengamos en una breve explicación, que nos llevará al establecimiento del llamado “modus ponens generalizado” como procedimiento inferencial básico en los sistemas difusos.

¿Cómo podríamos representar en un sistema difuso una declaración del tipo “Si *x* es *A*, Entonces y es *B*”, en donde *A* es un subconjunto difuso de un referencial *U*, *B* es un subconjunto difuso de un referencial *V* -que puede ser igual o distinto a *U*-, *x* es un elemento de *U*, e y es un elemento de *V*?

La respuesta a esta cuestión no es única, y varios autores proponen distintas soluciones. Al respecto, Zadeh propone que la función de grado de pertenencia de la declaración anterior puede calcularse del siguiente modo[55]:

Declaración : Si *x* es *A*, Entonces y es *B*

x es *A* : $f_A(x), x \in U$
y es *B* : $f_B(x), y \in V$

Si *x* es *A*, Entonces *y* es *B*:

$$f_A \rightarrow B(x,y) = A' \mid\!+\!\mid B = \min \{ 1, 1\text{-}f_A(x) + f_B(y) \}, x \in U, y \in V$$

Así, podemos introducir el mecanismo de inferencia conocido como “modus ponens”, según el cual:

Si $A \rightarrow B$
y A

Entonces B

Los sistemas difusos utilizan una generalización del “modus ponens”, que podemos representar del siguiente modo:

[55] No existe una verdadera justificación que avale la propuesta de Zadeh, pero tampoco la hay para las soluciones propuestas por Adlassnig y otros autores.

Si $A \rightarrow B$
y A

Entonces B

en donde A se parece a A, pero no es A, y en donde B se parece a B, pero no es B.

Este mecanismo inferencial se conoce con el nombre de "modus ponens generalizado" y, siempre según Zadeh, la expresión correspondiente para calcular $f_B(y)$ es la siguiente:

$$f_B(y) = \sup_v [A' |+| B] \cap A \,, \quad A \subset U, A \subset U, B \subset V, B \subset V$$

por lo que $f_B(y) = \sup_v [\min [\min 1, 1\text{-}f_A(x) + f_B(y)], f_A(x)]]$, $x \in U$, y $\in V$

Un ejemplo contribuirá a aclarar el proceso:

Sean $A \subset U$, $A \subset U$, $B \subset V$, $B \subset V$, donde $U = V = \{1, 2, 3, 4\}$, y sean:

$f_A(x) = \{0/1 + 0.6/2 + 1/3 + 0.5/4\}$
$f_A(x) = \{0/1 + 0.2/2 + 0.6/3 + 1/4\}$
$f_B(x) = \{0/1 + 1/2 + 0.6/3 + 0.2/4\}$

sea además la siguiente información:

- sabemos que: Si x es A, Entonces y es B
- sabemos que: x es A

(1) Encontrar A'
(2) Encontrar $A' |+| B$ en $U \times V$
(3) Evaluar $(A' |+| B) \cap A$ en $U \times V$
(4) Encontrar la expresión que caracteriza a B

Claramente este ejemplo no es más que el desarrollo "paso a paso" que nos permite caracterizar al subconjunto difuso B utilizando el modus ponens generalizado. De este modo:

$A' \rightarrow \quad f_A'(x) = \{1/1 + 0.4/2 + 0/3 + 0.5/4\}$

$A' |+| B$, en $U \times V =$

	V			
U	**1**	**2**	**3**	**4**
1	1.0	1.0	1.0	1.0
2	0.4	1.0	1.0	0.6
3	0.0	1.0	0.6	0.2
4	0.5	1.0	1.0	0.7

$(A' \,|+|\, B) \cap A$, en $U \times V =$

	V			
U	**1**	**2**	**3**	**4**
1	0.0	0.0	0.0	0.0
2	0.2	0.2	0.2	0.2
3	0.0	0.6	0.6	0.2
4	0.5	1.0	1.0	0.7

y, finalmente:

$B \rightarrow \quad f_B(y) = \sup_v \, [(A' \,|+|\, B) \cap A \,]$, en $U \times V$

$f_B(y) = \{ \, 0.5/1 + 1/1 + 1/1 + 0.7/1 \}$ con $y \in V$

El modus ponens generalizado es ya una primera diferencia en el razonamiento de los sistemas difusos, frente al razonamiento en sistemas más clásicos y convencionales. Pero además, también podemos encontrar otras diferencias -tanto en representación como en razonamiento-, que resumimos en los siguientes puntos:

Certeza:

En sistemas que utilizan lógica bivalente la verdad de una declaración sólo puede tener dos valores: la declaración es cierta, o la declaración es falsa. Por el contrario, en sistema multivaluados, la verdad de una declaración puede ser: un elemento de un conjunto finito, un intervalo (e.g.,, [0,1]), o un álgebra de Boole. En lógica difusa la verdad de una declaración puede ser un subconjunto difuso parcialmente ordenado, pero normalmente se asume la existencia de un subconjunto difuso del intervalo [0,1], o dicho de otro modo, un punto de dicho intervalo. Así, los denominados valores lingüísticos de la verdad de una declaración pueden expresarse por medio de etiquetas del tipo: cierto, muy cierto, no exactamente cierto,... que son etiquetas correspondientes a subconjuntos difusos del mencionado intervalo.

Predicados:

En sistemas bivalentes los predicados son categóricos -e.g., mortal, par, impar, más alto que,...- Por el contrario, en sistemas difusos los predicados son, precisamente, difusos -e.g., alto, pronto, mucho mayor que,...-

Modificadores:

En sistemas clásicos el único modificador realmente utilizado es la negación NOT. En sistemas difusos hay una gran variedad de modificadores -e.g., muy, más o menos, bastante,...- Estos modificadores son esenciales para generar los valores apropiados de la variables lingüísticas involucradas en un proceso -e.g., muy joven, no muy viejo,...-

Cuantificadores:

En los sistemas clásicos hay únicamente dos cuantificadores, el universal y el existencial. Por el contrario, en los sistemas difusos encontramos una gran variedad de cuantificadores -e.g., pocos, bastantes, normalmente, la mayoría,...-

Probabilidades:

En los sistemas lógicos clásicos, la probabilidad es numérica. En los sistemas difusos la probabilidad se expresa por medio de etiquetas lingüísticas (probabilidades difusas), del tipo: plausible, poco probable, alrededor de 0.8,... El manejo de tales probabilidades difusas debe efectuarse a través de la llamada aritmética difusa.

Posibilidades:

A diferencia de lo que ocurre con los sistema lógicos clásicos, el concepto de posibilidad en los sistemas difusos no es bivalente. De hecho, al igual que sucede con las probabilidades, las posibilidades pueden ser tratadas como variables lingüísticas que adoptan valores del tipo casi imposible, bastante posible,...

Los elementos que acabamos de describir, como componentes básicos de la representación del conocimiento y del razonamiento difuso, nos permiten definir una amplia variedad de modos de razonamiento, no necesariamente disjuntos, entre los que citaremos:

Razonamiento Categórico:

Este tipo de razonamiento utiliza declaraciones difusas, pero no emplea ni cuantificadores difusos ni probabilidades difusas. Un ejemplo sencillo podría ser el siguiente:

María es una chica delgada
María es una chica muy inteligente

María es una chica delgada y muy inteligente

En este ejemplo, las premisas “delgada” y “muy inteligente” deben interpretarse como predicados difusos. Por otra parte, el predicado difuso de la conclusión es la conjunción de las premisas anteriores.

Razonamiento Silogístico:

A diferencia de lo descrito al hablar del razonamiento categórico difuso, el razonamiento silogístico produce inferencias con premisas que incorporan cuantificadores difusos. Un ejemplo sencillo podría ser el siguiente:

La mayoría de los suecos son rubios
La mayoría de los suecos rubios son altos

(La mayoría)^2 de los suecos son rubios y altos

En este caso, el cuantificador difuso “la mayoría” debe interpretarse como una proporción difusa, y “(la mayoría)^2” es el cuadrado de “la mayoría” en aritmética difusa.

Razonamiento Disposicional:

En este tipo de razonamiento las premisas son disposiciones. La conclusión obtenida es una máxima que debe interpretarse como un mandato disposicional. Un ejemplo sencillo podría ser el siguiente:

Fumar mucho suele ser causa de abundante tos

Para evitar tos abundante evite fumar mucho

Razonamiento Cualitativo:

En sistemas difusos, el razonamiento cualitativo se define como un modo de razonamiento en el cual las relaciones entrada/salida de un sistema se representan por medio de una colección de reglas difusas -de tipo IF-THEN-, en las que los antecedentes y los consecuentes incluyen variables lingüísticas. Este tipo de razonamiento es el empleado habitualmente en las aplicaciones de la lógica difusa al análisis de sistemas y al control de procesos.

Actualmente la aplicación de los conjuntos difusos a los sistemas inteligentes es un tema de gran interés en investigación. De todas formas, aunque las bases teóricas del formalismo difuso están ya bastante claras, su aplicación a sistemas de naturaleza inferencial encuentra problemas que, hoy en día, siguen sin estar resueltos. Sí parece, no obstante, que los sistemas difusos aplicados a problemas de control están proporcionando soluciones alternativas, de gran brillantez y elegancia, frente a planteamientos más tradicionales.

1.6. Resumen

En este capítulo hemos descrito con cierto detalle los planteamientos de la lógica difusa, y su eventual aplicación a la inteligencia artificial como esquema de representación del conocimiento, como base de nuevos modelos de razonamiento, y también como medio eficaz de abordar el problema de la clasificación lingüística de variables. Tras una fugaz presentación de la naturaleza y alcance de los conjuntos difusos, entramos de lleno en su caracterización y nomenclatura. Como consecuencia del formalismo introducido aparecen los conjuntos ordinarios como un caso particular de los conjuntos difusos. No obstante, conjuntos ordinarios y conjuntos difusos no

tienen la misma estructura algebraica. De hecho los conjuntos difusos no tienen estructura de álgebra de Boole al no verificar ni el principio de no contradicción, ni la ley del tercero excluido. *A* continuación definimos y desarrollamos algunas operaciones algebraicas con conjuntos difusos. Los resultados de tales operaciones son, en ocasiones, difíciles de interpretar. Así, mientras el producto coincide con la intersección cuando hablamos de conjuntos ordinarios, cuando consideramos conjuntos difusos resulta que el producto está contenido en la intersección. Seguidamente abordamos el problema de las relaciones difusas, y proponemos la formulación de Zadeh para la representación de conocimiento del tipo: Si *x* es *A*, Entonces y es *B*. Ello nos lleva a definir el *modus ponens generalizado* como mecanismo inferencial estrella de los sistemas difusos. Finalmente se mencionan algunos de los modos de razonamiento que se pueden encontrar en los sistemas difusos. Estos modos de razonamiento surgen de las características diferenciales de la lógica difusa en relación a otras lógicas.

1.7. Textos Básicos

- Watson, Weiss and Donnell, "Fuzzy decision analysis", IEEE Trans. Systems, Man and Cybernetics, vol.9, 1979.

- Zadeh, "Fuzzy sets", Information and Control, vol.8, 1965.

- Zadeh, "Knowledge representation in fuzzy logic", IEEE Trans. Knowledge and Data Engineering, vol. 1, 1989.