William Castillo - Javier Trejos
Editores científicos

Métodos Matemáticos aplicados a las Ciencias

VII y VIII Simposios

EDITORIAL DE LA UNIVERSIDAD DE COSTA RICA

Contenido

Presentación

Una de las actividades más importantes del quehacer científico es la divulgación de los trabajos de investigación. Las dos formas más extendidas para ello son: la publicación de artículos en revistas especializadas y la celebración de foros donde se concentren durante un tiempo los científicos que trabajan en cierta área o especialidad, y donde expongan y compartan sus experiencias. Ambas actividades tienen igual importancia y son complementarias.

Desde 1978 se desarrolla en Costa Rica, cada dos años, un simposio sobre Métodos Matemáticos Aplicados a las Ciencias (SMMAC). Los ocho simposios realizados hasta la fecha han sido auspiciados por la Escuela de Matemática de la Universidad de Costa Rica, la Embajada de Francia en Costa Rica, la Universidad Paul Sabatier de Toulouse y el Instituto Nacional de Investigaciones Agronómicas de Francia.

Esta actividad ha sido fundamental para el desarrollo en nuestro país de una disciplina que ha mostrado su utilidad y su eficacia, como es el Análisis de Datos. En Costa Rica han estado algunos de los más destacados investigadores franceses en este campo, quienes tuvieron la visión de darle a nuestro simposio la importancia que se merecía.

El impacto que han tenido en nuestro país los SMMAC ha ido mucho más allá de la realización misma de la actividad. En efecto, un número grande de profesores de la Escuela de Matemática se interesó desde el inicio en los aspectos teóricos y prácticos de las técnicas que se exponían. Es así como se ha constituido en nuestro país un grupo de unos quince investigadores que conocen a fondo estas técnicas y que han hecho numerosas aplicaciones en diversos campos. Son muchas las ramas de investigación que se estudian, entre las que podemos citar las encuestas de opinión pública, la clasificación automática, el análisis de datos temporales, la estadística computacional, la modelación matemática, la optimización estocástica, el análisis de datos simbólicos. Además, esta formación de costarricenses ha repercutido en otras instituciones nacionales, como el Instituto Tecnológico de Costa Rica y la Caja Costarricense de Seguro Social, donde también hay colegas que trabajan en temas de investigación relacionados con el Análisis de Datos.

El Análisis de Datos (o *Analyse des Données* como es conocido en la comunidad científica internacional) tiene una diferencia sustancial con el Análisis Multivariado de Datos clásico, al estilo anglosajón: no se hacen hipótesis *a priori* sobre las distribuciones de probabilidad de las variables observadas sobre la población. Este punto

de vista filosófico fue compartido a mediados de los años sesenta por una serie de investigadores franceses en Estadística (citamos los nombres de Benzécri, Pagès, Escoufier, Schektman, entre otros), quienes no veían sentido en asumir una distribución de probabilidad teórica (la más usada era la normal) para una población que nunca la cumplía. Esto hizo que se emplearan entonces las herramientas de la Geometría. Euclídea, principalmente. Así, el punto de vista del *Analyse des Données* es esencialmente geométrico. Diremos de paso que varios de estos destacados colegas han estado en Costa Rica con motivo de la celebración de los SMMAC.

El lector que consulte las Memorias de los primeros simposios, se dará cuenta que las principales contribuciones de nuestros visitantes franceses usaban ampliamente la Geometría. Sin embargo, en estas visitas nuestros colegas siempre han procurado presentarnos lo más reciente de sus investigaciones. Así hemos podido apreciar la evolución de la disciplina en Francia, y cómo se han ido tratando y resolviendo distintos problemas que estaban planteados.

En estas Memorias recogemos las principales contribuciones de los exponentes en los VII y VIII SMMAC realizados en 1990 y 1992 en nuestro país. Destacan los artículos de nuestros colegas Edwin Diday, Yves Lechevallier y Gilbert Saporta. Se trata no sólo de artículos en que se resumen las principales contribuciones del autor a la disciplina, sino también las contribuciones más actuales y que aún son un campo fecundo de investigación.

El primer artículo es una extensa descripción de los llamados *objetos simbólicos* que ha introducido el profesor Edwin Diday. Su objetivo principal es definir los objetos sobre los que se podría hacer análisis multivariado pero de tal manera que se pueda manipular la información simbólica, en contraste con el manejo numérico que se ha dado tradicionalmente en Análisis de Datos.

Enseguida, el artículo de Yves Lechevallier muestra la construcción en la práctica de una red neuronal. Se muestra un lado útil de una técnica de Análisis de Datos, como es la segmentación, para obtener una red que permite resolver problemas de discriminación y clasificación.

Los dos artículos que siguen son de Gilbert Saporta. En el primero de ellos se hace una exposición detallada de las diferentes técnicas de discriminación. En el segundo, se presenta una técnica original del autor para hacer discriminación sobre variables cualitativas y se presenta la aplicación de la técnica al puntaje en créditos (*credit-scoring*).

El profesor José Pastrana de la Escuela de Estadística contribuye, en la misma línea de la discriminación, con un artículo sobre los aspectos prácticos de la discrimi-

nación bayesiana. Las redes neuronales, en sus versiones más usadas, son presentadas por Javier Trejos; el artículo está enfocado a la aplicación de las redes a resolver los problemas planteados en Análisis de Datos. William Castillo y Carlos Arce muestran la construcción de índices de asociación para la clasificación automática y se hace una aplicación a una encuesta de opinión pública, en la que se clasificaron los temas de conflicto.

Los dos artículos siguientes muestran el impacto que ha tenido en nuestro país la investigación en las encuestas de opinión pública. Olga Prieto y Marta López, de la Escuela de Sociología y Antropología, contribuyen con dos estudios hechos a partir de la encuesta de opinión pública de 1992 que tratan respectivamente sobre la seguridad ciudadana y el V centenario de la llegada de los españoles a América.

Jaime Lobo presenta sus estudios sobre al análisis de supervivencia y los procesos de conteo. Finalmente, los profesores rumanos Ioan Muntean y Neculae Vornicescu nos presentan sus más recientes investigaciones sobre estrategias de aprendizaje en tiempo mínimo.

Para terminar, deseamos agradecer a todas las instituciones y personas, tanto nacionales como francesas, que hicieron posible la realización de los VII y VIII simposios, así como la publicación de estas memorias.

William Castillo Elizondo
Javier Trejos Zelaya

Editores científicos

San José, abril de 1994

Objetos Probabilísticos, Posibilísticos y Creencia para el Análisis de Conocimientos

Edwin Diday*

Resumen

La idea principal del enfoque simbólico en análisis de datos es extender los problemas, métodos y algoritmos usados sobre datos clásicos a datos más complejos llamados "objetos simbólicos" los cuales se adaptan bien a la representación del conocimiento y además son genéricos, cualidad que no poseen las observaciones usuales que caracterizan "cosas individuales". Introducimos varias clases de objetos simbólicos: booleanos, posibilísticos, probabilísticos y creencia. Presentamos brevemente algunas de sus cualidades y propiedades; tres teoremas muestran cómo las teorías de Probabilidad, Posibilidad y Creencia pueden ser extendidas a estos objetos. Finalmente, cuatro clases de problemas de análisis de datos, incluyendo la extensión simbólica, son ilustrados por medio de varios algoritmos que inducen conocimiento desde datos clásicos o desde un conjunto de objetos simbólicos.

Palabras clave: Análisis de conocimiento, análisis de datos simbólicos, metadatos, metaconocimiento, Teoría de la Probabilidad, de la Posibilidad y de la Creencia, lógica de la incertidumbre.

1 Introducción

Si deseamos describir la producción de frutas de una determinada región por las características "**el peso es entre 300 y 400 gramos, su color es blanco o rojo y si el color es blanco entonces el peso es menor que 350 gramos**", no es posible colocar esta clase de información en una tabla de datos clásica donde las filas representan regiones y las columnas los descriptores de las frutas. Esto es debido a que no habrá un solo valor en cada celda de la tabla (por ejemplo, para el peso) y también porque no será fácil representar reglas (`si ...entonces ...`) en esta tabla. Es más fácil representar esta clase de información por medio de una expresión lógica tal como:

$$a_i = [\text{peso} = [300, 400]] \wedge [\text{color} = \{\text{rojo, blanco}\}] \wedge [\text{si}[\text{color} = \text{blanco}]\text{entonces}[\text{peso} \leq 350]]$$

*Universidad Paris IX-Dauphine; Institut National de Recherche en Informatique et Automatique - Rocquencourt, Francia

donde a_i, que representa la i-ésima región, es una aplicación definida sobre el conjunto de frutas Ω tal que para cada fruta $\omega \in \Omega$, $a_i(\omega)$ = verdadero si el peso de ω pertenece al intervalo [300, 400], su color es rojo o blanco y si su color es blanco entonces su peso es menor o igual que 350 gramos.

Siguiendo la terminología de este artículo a_i es una clase de objeto simbólico. Simbólico porque a_i es descrito por una expresión que contiene operadores diferentes de los usados con números clásicos, "objeto" porque se considera como un objeto individual para una estadística de nivel una unidad más alto.

Si tenemos un conjunto de 1000 regiones representadas por un conjunto de 1000 objetos simbólicos $a_1, \ldots, a_{1000}$, un problema importante es saber cómo aplicar el análisis de datos o los métodos estadísticos a esta información. Por ejemplo, ¿qué es un histograma, una clasificación o una ley de probabilidad para ese conjunto de objetos? La idea del análisis de datos simbólicos [7,8] es proveer herramientas para responder a este problema.

En algunos campos la representación booleana del conocimiento

$$(a_i(\omega) = \text{verdadero o falso})$$

es suficiente para conseguir la información principal, pero en muchos casos necesitamos incluir la incertidumbre para representar el mundo real con más eficiencia. Por ejemplo, si decimos que en la región i-ésima "el color de las frutas es frecuentemente rojo y rara vez blanco", podemos representar esta información por medio de

$$a_i = [\text{color} = \text{frecuentemente rojo, rara vez blanco}]$$

Más generalmente, en el caso de objetos booleanos u objetos con frecuencia de aparición, podemos escribir $a_i = [\text{color} = q_i]$ donde q_i es una función característica en el caso booleano y una medida de probabilidad en el segundo caso. Más precisamente, en el caso booleano si $a_i = [\text{color} = \{\text{rojo, blanco}\}]$ tenemos $q_i(\text{rojo}) = q_i(\text{blanco}) = 1$ y $q_i = 0$ para los otros colores. En el caso probabilístico, si $a_i = [\text{color} = 0.9\ \text{rojo}, 0.1\ \text{blanco}]$ tenemos $q_i(\text{rojo}) = 0.9$ y $q_i(\text{blanco}) = 0.1$.

Si un experto dice que las frutas de una cierta región son rojas, podemos representar esta información por un objeto simbólico $a_i = [\text{color} = q_i]$ donde q_i es una función posibilista en el sentido de Dubois y Prade [11]. Tendremos por ejemplo, $q_i(\text{blanco}) = 0$, $q_i(\text{rosado}) = 0.5$ y $q_i(\text{rojo}) = 1$.

Ahora, si a partir de una muestra representativa de frutas de la i-ésima región un experto dice que el 60% son rojas, el 30% son blancas y el color del 10% restante es irreconocible por el mal estado de la fruta, podemos representar esta información por $a_i = [\text{color} = q_i]$ donde q_i es una función creencia definida por $q_i(\text{rojo}) = 0.6$, $q_i(\text{blanco}) = 0.3$ y $q_i(O) = 1$, donde O es el conjunto de colores posibles.

El objeto simbólico a_i se llamará boleano, probabilista, posibilista o creencia dependiendo de la clase de aplicación que sea la q_i asociada. En todos los casos anteriores q_i es una aplicación de Ω (el conjunto de frutas) a $[0, 1]$. Ahora, el problema es conocer cómo

calcular $a_i(\omega)$; si hay duda sobre el color de una fruta dada ω porque el experto dice que puede ser "roja o rosada" entonces esta información puede ser descrita por medio de una función característica r y representada por medio de un objeto simbólico $\omega^S = [\text{color} = r]$ tal que $r(\text{rojo}) = r(\text{rosado}) = 1$ y $r = 0$ para los otros colores. Entonces, dependiendo de la clase de conocimiento que el usuario desee representar, r puede ser una función probabilista, posibilista o creencia. Teniendo $a_i = [\text{color} = q_i]$ y $\omega^S = [\text{color} = r]$, para calcular $a_i(\omega)$ introducimos una función de comparación g tal que $a_i(\omega) = g(q_i, r)$ que mide el ajuste entre q_i y r. ¿Cuál es el significado de $a_i(\omega)$? ¿Podemos decir que $a_i(\omega)$ mide un tipo de probabilidad, posibilidad o creencia de que ω pertenezca a la clase de frutas descritas por a_i cuando –dependiendo del conocimiento previo– q_i y r sean funciones características, probabilistas, posibilistas o creencias respectivamente? Para responder a esta pregunta necesitamos extender a_x (donde x representa una clase de conocimiento previo) a a_x^* definido sobre un conjunto de objetos simbólicos $\mathcal{A}$ y definir un conjunto de operadores $OP_x = \{\cup_x, \cap_x, c_x\}$ en $\mathcal{A}$, adaptado a x. Si decimos que los conjuntos clásicos representan un nivel de conocimiento de orden 0; probabilidad, posibilidad y creencia un nivel de conocimiento de orden 1, la pregunta es ahora conocer si a_x^* representa un nivel conocimiento de orden 2. En otras palabras, si a_x^* es una probabilidad de probabilidad, una posibilidad de posibilidad o una creencia de creencia respectivamente asociada con los correspondientes operadores OP_x. Los teoremas 1, 2 y 3 muestran que este es el caso si OP_x y ciertas funciones g_x y f_x son bien escogidas.

En teoría de probabilidad muy poco se dice acerca de eventos que son generalmente identificados como partes de un espacio muestral Ω. En Ciencias de la Computación los lenguajes orientados a objetos consideran enventos más generales llamados objetos o "frames" definidos por intensión. En análisis de datos (ajuste multidimensional, clasificación automática, análisis exploratorio de datos, etc.) se da más importancia a los objetos elementales, los cuales pertenecen a una muestra Ω, que en estadística clásica, donde se enfatiza la atenció sobre las leyes de probabilidad de Ω. Sin embargo, los objetos en análisis de datos son generalmente identificados con puntos en R^p y por tanto son inadecuados para tratar objetos complejos provenientes de grandes bases de datos y de bases de conocimientos.

Nuestro objetivo es definir objetos complejos llamados "objetos simbólicos" inspirados por aquéllos de los lenguajes orientados a objetos de modo tal que el análisis de datos sea generalizado a un análisis de conocimientos. Los objetos pueden ser definidos intensionalmente por las propiedades de un elemento genérico de la clase que ellos representan. Distinguimos esta clase de objetos como "objetos elementales observados" que caracterizan "cosas individuales": por ejemplo "los clientes de mi tienda" en lugar de "un cliente de mi tienda", una "especie de hongos" en lugar de "el hongo que tengo en mi mano".

El concepto de objeto simbólico extiende el de objeto clásico de dos formas:

- primero, en el caso de "objetos elementales" los cuales representan cosas individuales, dando la posibilidad de introducir en su definición, información estructurada (ver el caso de "horda" en §2 para la descripción de una imagen), probabilidades (subjetivas

u objetivas), posibilidades (en caso de vaguedad e imprecisión, por ejemplo), creencia (en caso de probabilidades sólo conocidas sobre partes y para expresar ignorancia);

- segundo, en caso de objetos que son descritos intensionalmente por las mismas posibilidades que en el caso de objetos elementales, más la posibilidad de expresar variación de los valores tomados por cada variable entre los miembros de su extensión ([color = rojo,blanco]) y también expresando restricciones entre estos valores con reglas (**si** [color = blanco] **entonces** [peso $\leq$ 350]).

Extendiendo los métodos de análisis de datos a objetos simbólicos, este artículo establece un puente entre varios campos: "análisis de datos y estadística" (que, como se ha mostrado, tiene un interés limitado en el tratamiento de esta clase de objetos), "bases de datos estadísticos" (donde los objetos simbólicos pueden ser considerados como metadatos, lo cual significa datos sobre datos), administración de la incertidumbre en sistemas de bases de conocimientos (donde el énfasis es ahora más sobre representación del conocimiento y razonamiento que sobre análisis de datos), aprendizaje por medio de máquinas o *learning machine* (donde han sido despreciados esta clase de objetos como entradas del sistema, así como los métodos clásicos de análisis de datos) y, más generalmente en Inteligencia Artificial (donde los resultados aquí obtenidos, en los teoremas 1, 2 y 3, conciernen a metaconocimiento o conocimiento sobre conocimiento).

No hemos usado el concepto de "predicados" de la lógica clásica; primero porque usando sólo aplicaciones o funciones, las cosas parecen más entendibles, especialmente para los estadísticos; segundo, porque los predicados no pueden ser usados facilmente en el caso de objetos probabilísticos, posibilísticos y creencia donde la incertidumbre está presente.

2 Objetos simbólicos

2.1 Definición de objetos simbólicos

Empezamos introduciendo alguna notación. Denotamos Ω un conjunto de cosas elementales llamadas "objetos individuales", Δ un conjunto de descriptores posibles de Ω, y una aplicación $\Omega \to \Delta$ que asocia a cada $\omega \in \Omega$ su descripción $\delta = y(\omega)$.
Y_Ω es una aplicación $P(\Omega) \to D$ que asocia a cada $\Omega' \subseteq \Omega$ su descripción $d \in D$, donde D es un conjunto de descripción de subconjuntos de Δ y $P(\Omega)$ es el conjunto de partes de Ω; Y es una aplicación $P(\Omega) \to P(\Delta)$ tal que

$$Y(\Omega') = \Delta' \text{ si y sólo si } \Delta' = \{y(\omega) \mid \omega \in \Omega'\};$$

Y_Δ es una aplicación $P(\Delta) \to D$ que asocia a cada $\Delta' \subseteq \Delta$ una descripción $d \in D$ que satisface al menos la siguiente propiedad:

$$Y_\Delta(\Delta') \subseteq D.$$

En esta sección A es un conjunto de aplicaciones $\Delta \to L$ donde $L = \{\text{verdadero, falso}\}$ (más generalmente $L = [0, 1]$ en la sección §3).
h_Ω es una aplicación $D \to A$ tal que $h_\Omega(d) = a$
donde a es la aplicación $\Omega \to \{\text{verdadero, falso}\}$ tal que

$$a(\omega) = \text{verdadero si y sólo si } y(\omega) = \delta \in d.$$

B es un conjunto de aplicaciones $D \to \{\text{verdadero, falso}\}$ tal que $h_\Delta(d) = b$ donde b es la aplicación $\Delta \to \{\text{verdadero, falso}\}$ tal que

$$b(\delta) = \text{verdadero si y sólo si } \delta \in d.$$

Z es una aplicación $\mathcal{B} \Rightarrow \mathcal{A}$ tal que

$$Z(b) = a \text{ si y sólo si } a = \text{niño},$$

donde $\mathcal{A} = h_\Omega(D)$ y $\mathcal{B} = h_\Delta(D)$.

Una *intensión* de un conjunto de objetos individuales $\Omega' \subseteq \Omega$ puede ser definida por $d = Y_\Omega(\Omega')$, $a = h_\Omega(Y_\Omega(\Omega'))$ o $b = h_\Delta(Y_\Omega(\Omega'))$, en §2.4 comparamos estas diferentes clases de intensión. La extensión de a en Ω es un subconjunto de Ω denotado $Ext(a/\Delta)$ y definido por $Ext(a/\Omega) = \{\omega \in \Omega \mid a(\omega) = \text{verdadero}\}$. La extensión de b es un subconjunto de Δ definido por $Ext(b/\Delta) = \{\delta \in \Delta \mid b(\delta) = \text{verdadero}\}$. Por último, la extensión de $d \in D$ en X es denotada por $Ext(d/X)$; por definición, ponemos $Ext(d/\Omega) = Ext(a/\Omega)$ y $Ext(d/\Delta) = Ext(b/\Delta)$.

E_Δ es la aplicación $\mathcal{B} \to P(\Delta)$ tal que $E_\Delta(b) = Ext(b/\Omega)$.

E_Ω es la aplicación $\mathcal{A} \to P(\Omega)$ tal que $E_\Omega(a) = Ext(a/\Omega)$.

Todas las aplicaciones definidas anteriormente son resumidas en la figura 1.

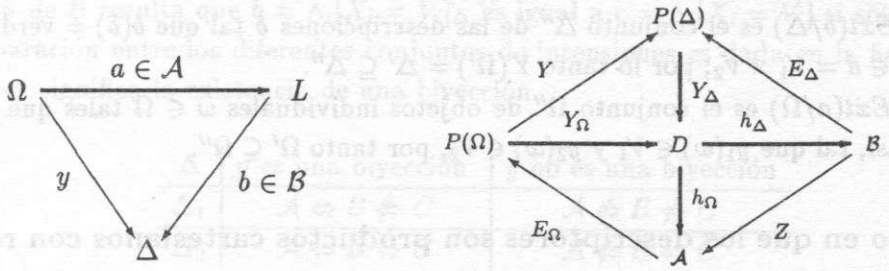


Figura 1: Cualquier elemento de D, $\mathcal{B}$ ó $\mathcal{A}$ puede ser considerado como un objeto simbólico

En estadística o en análisis de datos clásico estudiamos una base de conocimientos definida por el par (Ω, Δ) tal que las unidades son parejas (ω, δ) donde $\omega \in \Omega$ es un objeto individual descrito por $\delta \in D$. Por otra parte en análisis de datos simbólicos estudiamos una base de conocimientos (W, X) donde W es un subconjunto de $P(\Omega)$ y X es un espacio intensión incluido en D, $\mathcal{B}$ o $\mathcal{A}$.

Un objeto simbólico es un conjunto de propiedades concernientes a un subconjunto de Ω. Todo elemento de D, $\mathcal{B}$ o $\mathcal{A}$ puede ser considerado como un objeto simbólico; en la próxima sección damos un ejemplo que ilustra las aplicaciones y conjuntos que han sido definidos en esta sección.

2.2 El caso en que los descriptores son productos cartesianos

En este caso especial asumimos que Ω es descrito por $\Delta = O_1 \times \ldots \times O_p$ donde O_i es un dominio que contiene un conjunto de valores posibles (el color de las frutas, por ejemplo) y $D = P(O_1) \times \ldots \times P(O_p)$; en el caso finito resulta que $|P(\Delta)| = |2^{\Pi_i O_i}|$ y $D = 2^{\sum_i |O_i|}$; por tanto D que es incluido en $P(\Delta)$ es generalmente mucho más pequeño que $P(\Delta)$.

En este caso si $d = (V_1, \ldots, V_p)$ donde $V_i \subseteq O_i$ y $h_\Delta(d) = b$, entonces denotamos $b = \wedge_i [X_i = V_i]_\Delta$ lo cual significa que si $\omega = (x_1, \ldots, x_p)$ $b(\omega) =$ verdadero si y sólo si las proposiciones $x_i \in V_i$ son verdaderas. Más aún, si $h_\Omega(d) = a$ tenemos $a(\omega) = \wedge_i [y(\omega) \in V_i]_\Omega$ lo cual puede ser escrito como $a(.) = \wedge_i [y(.) \in V_i]_\Omega$ que se simplifica en $a = \wedge_i [y = V_i]_\Omega$.

Ejemplo:
Ω es un conjunto de frutas, Δ es el conjunto de todas las posibles descripciones de las frutas por su color y su peso; por lo tanto si O_1 es el conjunto de todos los pesos posibles y O_2 es el conjunto de todos los colores posibles, tenemos $\Delta = O_1 \times O_2$; W es el conjunto cuyos elementos son las frutas producidas por una región; Y_Ω asocia al conjunto de frutas $\Omega' \subseteq \Omega$ de una región, el más pequeño intervalo V_1 de pesos posibles para las frutas y la unión de su color V_2; por tanto tenemos $Y_\Omega(\Omega') = V_1 \times V_2 = d$, $a = h_\Omega(d) = [y_1 = V_1]_\Omega \wedge [y_2 = V_2]_\Omega$ y $b = h_\Delta(d) = [x_1 = V_1]_\Delta \wedge [x_2 = V_2]_\Delta$ donde, como en el ejemplo de la introducción, $V_1 = [300, 400]$ y $V_2 = \{$rojo, blanco$\}$; $Y(\Omega')$ es el conjunto de descripciones Δ' de las frutas de la región (o sea de Ω') y $Y_\Delta(\Delta') = (Y_{\Delta \circ Y)(\Omega')}) = V_1 \times V_2$.

$E_\Delta(b) = Ext(b/\Delta)$ es el conjunto Δ'' de las descripciones δ tal que $b(\delta) =$ verdadero y así, tal que $\delta \in d = V_1 \times V_2$; por lo tanto $Y(\Omega') = \Delta' \subseteq \Delta''$.

$E_\Omega(a) = Ext(a/\Omega)$ es el conjunto Ω'' de objetos individuales $\omega \in \Omega$ tales que $a(\omega) =$ verdadero y asi, tal que $y_1(\omega) \in V_1$ y $y_2(\omega) \in V_2$, por tanto $\Omega' \subseteq \Omega''$.

2.3 El caso en que los descriptores son productos cartesianos con restricciones

Las restricciones pueden aparecer con el fin de describir más precisamente un conjunto $\Omega' \subseteq \Omega$ de objetos individuales; en el ejemplo de la introducción a la descripción

$$a = [y = [300, 400]] \wedge [\text{color} = \{\text{rojo, blanco}\}$$

agregamos la restricción

$$[\textbf{si}[\text{color} = \text{blanco}]\textbf{entonces}\ [\text{peso} \leq 350]].$$

Otras clases de restricciones pueden aparecer para evitar incoherencias en la descripción de un conjunto $\Omega' \subseteq \Omega$. Por ejemplo, si Ω' es un conjunto de hongos con o sin sombrero y una de las descripciones se refiere al color del sombrero, debemos agregar la condición que no hay color de sombrero cuando no hay sombrero.

2.4 Comparaciones entre los conjuntos de intensiones $D, \mathcal{A}, \mathcal{B}, C$

Estas comparaciones dependen de la escogencia de y y Δ. A fin de simplificar, asumimos que $D \subseteq P(\Delta)$. Es entonces fácil mostrar que h_Δ es una biyección (lo que no es el caso para h_Ω si y no es biyectiva). Si y es sobreyectiva es fácil probar que Z es inyectiva y si y es inyectiva que Z es sobreyectiva. Por lo tanto, si y es una biyección entonces Z lo es entre $\mathcal{B}$ y $\mathcal{A}$.

Dos escogencias naturales para Δ son las siguientes: la primera denotada por Δ_1 es el conjunto de descripciones con restricciones (por ejemplo, descripciones coherentes); la segunda, denotada por Δ_2 es el conjunto de todas las posibles descripciones (realizables o no). Cuando y es biyectiva y $\Delta = \Delta_1$, entonces $\Omega = \Omega_1$ es el conjunto de todos los objetos individuales coherentes u "observables". Cuando y es biyectiva y $\Delta = \Delta_2$, entonces $\Omega = \Omega_2$ es el conjunto de los objetos individuales "posibles" (realizables o no); Ω_2 se llama el conjuto de "posibilidades". En la práctica tenemos $\Omega = \Omega_0$ el conjunto de objetos individuales "observados" el cual no está en biyección con los conjuntos Δ_1 o Δ_2 así como varios objetos individuales pueden tener la misma descripción y alguna descripción de Δ_1 ó Δ_2 puede no corresponder a ningún objeto individual de Ω_0. Por tanto debemos considerar también el caso donde y no es biyectiva. Denotamos C el conjunto de ℓ-complejos introducido por Michalski *et al.* [14], cuyos elementos son expresiones lógicas del tipo $c = \wedge_i[X_i = V_i]$ donde la proposición $[X_i = V_i]$ significa "valor de X_i es uno de los elementos de V_i"; de la definición de $\mathcal{B}$ resulta que $b = \wedge_i[X_i = V_i]_\Delta$ es igual a $c = \wedge_i[X_i = V_i]$ si sólo si $\Delta = \Delta_2$. La comparación entre los diferentes conjuntos de intensiones es dada en la figura 2 donde el signo $\Leftrightarrow$ significa la existencia de una biyección.

Δ	y es una biyección	y no es una biyección
Δ_1	$\mathcal{A} \Leftrightarrow \mathcal{B} \not\Leftrightarrow C$	$\mathcal{A} \not\Leftrightarrow \mathcal{B} \not\Leftrightarrow C$
Δ_2	$\mathcal{A} \Leftrightarrow \mathcal{B} \Leftrightarrow C$	$\mathcal{A} \not\Leftrightarrow \mathcal{B} \Leftrightarrow C$

Figura 2 Comparación entre conjuntos de intensiones; en cualquier caso $\mathcal{B} \Leftrightarrow D$; C es el conjunto de ℓ-complejos de Michalski.

2.5 Objetos simbólicos completos y retículos sobre $\mathcal{A}$, $\mathcal{B}$ y C

Cuando asociamos a un elemento $\Omega' \in P(\Omega)$ un objeto simbólico $Y_\Omega(\Omega') = d \in D$ la extensión de d en Ω la cual es $E_\Omega(h_\Omega(d))$ y contiene a Ω' pues es el conjunto de los $\omega \in \Omega$ tales que $y(\omega) \in d$, en otras palabras tenemos $\Omega' \subseteq E_\Omega(a)$ con $a = h_\Omega(Y_\Omega(\Omega'))$; en particular cuando $\Omega' = E_\Omega(a)$, decimos que a es un *objeto simbólico completo*; similarmente decimos

que b es un *objeto simbólico completo* si $\Omega' = E_\Delta(b)$ con $b = h_\Delta(Y_\Omega(\Omega'))$. Denotamos $\mathcal{A}_c$ (resp. $\mathcal{B}_c$) el conjunto de objetos simbólicos completos incluidos en $\mathcal{A}$ (resp. $\mathcal{B}$).

Definimos un orden parcial sobre un conjunto de objetos simbólicos estableciendo que un objeto simbólico s_1 es menor o igual que un objeto simbólico s_2 si la extensión de s_1 está contenida en la extensión de s_2. Si definimos el supremo (resp. ínfimo) de dos objetos simbólicos s_1, s_2, cuyas descripciones son respectivamente $d_1 = O'_1 \times \ldots \times O'_p$ y $d_2 = O''_1 \times \ldots \times O''_p$, por $d_1 \cup d_2 = O'_1 \cup O''_1 \times \ldots \times O'_p \cup O''_p$ (resp. $d_1 \cap d_2 = O'_1 \cap O''_1 \times \ldots \times O'_p \cap O''_p$.

La más pequeña descripción de $\Omega' \subseteq \Omega$ es la intersección de todas las descripciones $d \in D$, tales que $E_\Omega(h_\Omega(d)) = \Omega'$. Se puede mostrar que $\mathcal{A}$, $\mathcal{A}_c$ (ver [10]), y $\mathcal{B}_c$ (ver [2]) constituyen un retículo.

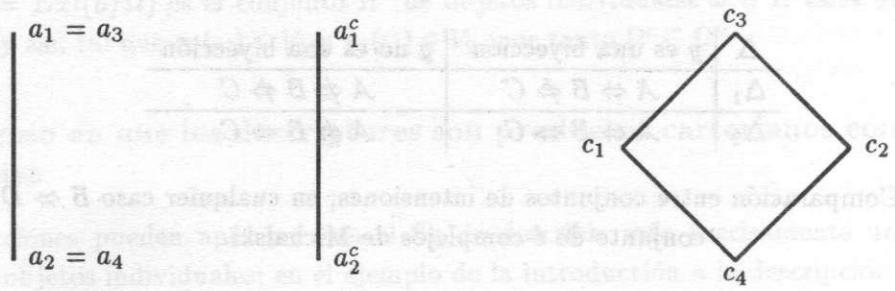
Ejemplo:
Sea $\Omega_3^c = \{\omega_1, \omega_2\}$ descrito por $y : \Omega_3^c \to O = \{1,2\}$ tal que $y(\omega_1) = 1$, $y(\omega_2) = 2$; por tanto $y(\Omega_3^c) = \Delta_3 = \{\delta_1, \delta_2\}$ donde $\delta_1 = 1$ y $\delta_2 = 2$; se sigue que $D = P(O) = \{\{1\}, \{2\}, \{1,2\}, \emptyset\}$.

Definimos los objetos simbólicos de $\mathcal{A}$ por:

$$a_1 = [y = 1]_\Omega, a_2 = [y = 2]_\Omega, a_3 = [y = \{1,2\}]_\Omega \text{ y } a_4 = [y = \emptyset]_\Omega;$$

escogemos $\Omega = \{\omega_1\}$, por tanto $a_1 = a_3$ y $a_2 = a_4$.

Definimos también las aplicaciones $b_i \in \mathcal{B}$ representadas por los ℓ-complejos $c_i = Z(b_i)$: $c_1 = [X = 1]$, $c_2 = [X = 2]$, $c_3 = [X = \{1,2\}]$ y $c_4 = [X = \emptyset]$. En este caso es fácil ver que el conjunto de objetos completos es $\mathcal{A}_c = \{a_1^c, a_2^c\}$ con $a_1^c = [y = 1]$ y $a_2^c = [y = \emptyset]$. En la figura 3 representamos tres retículos respectivamente asociados a $\mathcal{A} = \{a_1 = a_3, a_2 = a_4\}$, $\mathcal{A}_c = \{a_1^c, a_2^c\}$ y $l_c = \{c_1, c_2, c_3, c_4\}$.



(a) (b) (c)

Figura 3 (a), (b) y (c) representan respectivamente los retículos de $\mathcal{A}, \mathcal{A}_c$ y l_c.
En (a) representamos el orden $a_2 = a_4 < a_1 = a_3$;
en (b) el orden $a_2^c < a_1^c$
y en (c) $c_4 < c_1, c_4 < c_2, c_1 < c_3, c_2 < c_3$

2.6 Escogencia de la base de conocimientos para un análisis de datos simbólicos

Hemos visto en §2.1 que una base de conocimientos es un par (W, X) donde $X = \mathcal{A}$ ó $\mathcal{B}$ ó C. Así, una pregunta natural es: ¿en qué caso debemos usar $\mathcal{A}$, $\mathcal{B}$ ó C en la práctica? Si deseamos tener en cuenta sólo el conjunto de descripciones Δ_2 entonces, la mejor escogencia es $X = \mathcal{B}$. Esto ocurre por ejemplo cuando las descripciones de los subconjuntos Ω' de Ω (es decir, $\Omega' \in W$) tienen restricciones y no dependen de ninguna muestra Ω. Este tipo de base de conocimientos se usa cuando se quiere estudiar especies en biología, escenarios de accidentes en transporte, equipos en una compañía (cada especie, escenario o equipo es un elemento de W), independientemente de cualquier conjunto muestral.

Si deseamos estudiar un conjunto W descrito sin restricciones e independientemente de Ω, la mejor escogencia es $X = C$. Si deseamos tener en cuenta las informaciones estadísticas contenidas en Ω, la mejor escogencia es $\mathcal{A}$. Más aún, $\mathcal{A}$ permite la posibilidad de calcular retículos más simples (ver el ejemplo en la sección §2.5) y distancias entre objetos simbólicos cuando las descripciones varían. Este caso puede ocurrir por ejemplo cuando varios sensores dan diferentes medidas sobre el mismo conjunto Ω, o cuando Ω es descrito por variables cuyos valores varían con el tiempo.

Si Ω es descrito por dos aplicaciones y_1 y y_2 tales que $y_i(\Omega) = \Delta_i = O_i$, entonces las apliciaciones $a_i \in \mathcal{A}_i$ definidas por $h_\Omega : P(O_i) \to \mathcal{A}_i$ cuando i varía, son comparables usando una disimilitud (por ejemplo $s(a_1, a_2) = \sum\{|a_1(\omega) - a_2(\omega)| \mid \omega \in \Omega\}$) mientras las aplicaciones $b_i \in \mathcal{B}_i$ definidas por $h_{\Delta_i} : \Delta_i \to \mathcal{B}_i$ no son comparables cuando i varía.

Ejemplo:
Sea $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ un conjunto descrito por las aplicaciones $y_1 : \Omega \to O_1 = \{1, 2\}$ y $y_2 : \Omega \to O_2 = \{1, 2, 3\}$ dadas en la figura 4. Sea $a_1 = [y_1 = 1]_\Omega$ y $a_2 = [y_2 = 1]_\Omega$.

Ω	y_1
ω_1	1
ω_2	2
ω_3	2
ω_4	1

Ω	y_2
ω_1	1
ω_2	2
ω_3	2
ω_4	3

Δ_1	O_1
δ_1^1	1
δ_2^1	2

Δ_2	O_2
δ_1^2	1
δ_2^2	2
δ_3^2	3

Figura 4

Podemos calcular $s(a_1, a_2) = \sum_{\omega \in \Omega} |a_1(\omega) - a_2(\omega)| = 2$, mientras que $c_1 = [X_1 = 1]$ y $c_2 = [X_2 = 1]$ no son comparables pues están definidas sobre conjuntos de objetos diferentes: c_1 lo está sobre Δ_1 y c_2 sobre Δ_2.

En este artículo enfatizamos sobre la base de conocimientos $(P(\Omega), \mathcal{A})$ pues $\mathcal{A}$ es el único conjunto del cual se pueden tener en cuenta las informaciones estadísticas contenidas en Ω cuando y no es inyectiva, y también se pueden tomar en cuenta sólo las descripciones cuando y es biyectiva. Sobre este punto P. Brito [2] enfatiza sobre la base de conocimientos $(W, \mathcal{B})$

cuando y no es biyectiva y $\Delta = \Delta_2$; mientras que F. De Carvalho [5] lo hace sobre $(W, \mathcal{A})$ cuando y es biyectiva y $\Delta = \Delta_1$. En su disertación J. Lebbe y R. Vignes [13] enfatizan sobre (W, D) con $\Delta = \Delta_1$ y y no biyectiva.

3 Objetos simbólicos booleanos

En esta sección las descripciones son productos cartesianos. Así tenemos $\Delta = O_1 \times \ldots \times O_p$ y $D = P(O_1) \times \ldots \times P(O_p)$; $y = (y_1, \ldots, y_p)$ es una aplicación $\Omega \to \Delta$ tal que $y(\omega) = (y_1(\omega), \ldots, y_p(\omega))$, donde y_i es una aplicación $\Omega \to O_i$. Los objetos simbólicos booleanos son objetos simbólicos considerados cuando L es booleano (es decir, $L = \{\text{verdadero, falso}\}$). Seguidamente se definen las diversas clases de objetos booleanos.

3.1 Eventos

Sea $D_i = P(O_i)$ y h^i_Ω la aplicación $D_i \to \mathcal{A}$ tal que $h^i_\Omega(V_i) = e_i$ donde e_i es la aplicación $\Omega \to \{\text{verdadero, falso}\}$ tal que $e_i(\omega) =$ verdadero si y sólo si $y_i(\omega) \in V_i$. Por analogía con la terminología usada en teoría de probabilidad (donde un "evento" es un subconjunto $V_i \subseteq \Omega$), el objeto simbólico básico e_i se llama *evento*. En términos lógicos podemos escribir

$$e_i(\omega) = [y_i(\omega) \in V_i]_\Omega$$

donde $[y_i(\omega) \in V_i]_\Omega$ es la proposición lógica que es verdadera si y sólo si $y_i(\omega) \in V$.

Para expresar el objeto simbólico e_i, a fin de simplificar las notaciones, en lugar de escribir $\{\forall\omega, e_i(\omega) = [y_i(\omega) \in V]_\Omega\}$ ó $e_i(.) = [y_i(.) \in V]_\Omega$ escribiremos $e_i = [y_i = V_i]_\Omega$, o más simplemente

$$e_i = [y_i = V_i]$$

omitiendo Ω cuando no hay ambigüedad. Por ejemplo si $e_i = [\text{color} = \{\text{rojo, blanco}\}]$, entonces $e_i(\omega) =$ verdadero si el color de ω es rojo o blanco. Cuando $y_i(\omega)$ no tiene sentido (por ejemplo, el tipo de computadora usada por una empresa que no tiene computadoras) entonces $V_i = \emptyset$ y cuando tiene sentido pero éste es desconocido entonces $V_i = O_i$.

La *extensión* de e_i en Ω denotada por $ext(e_i/\Omega)$ es el conjunto de elementos ω de Ω tales que $e_i(\omega) =$ verdadero.

3.2 Aserciones

Una *aserción* es una conjunción de eventos, es decir una aplicación

$$h_\Omega : D = D_1 \times \ldots \times D_p \to \mathcal{A}$$

tal que si $V = (V_1, \ldots, V_p)$ donde $V_i \subseteq O_i$ entonces $h_\Omega(V) = a$ tal que $a(\omega) =$ verdadero si y sólo si $y(\omega) \in V$.

En términos de la lógica podemos escribir

$$a(\omega) = \wedge_i [y_i(\omega) \in V_i] = \wedge_i e_i(\omega).$$

Conforme con la notación para un evento, una aserción a es denotada

$$a = \wedge_i [y_i = V_i].$$

Por ejemplo, si $a = [\text{color} = \{\text{rojo, blanco}\}] \wedge [\text{altura} = [0, 15]]$, $a(\omega) =$ verdadero si y sólo si ω es rojo o blanco y su altura está entre 0 y 15.

La *extensión* de una aserción denotada $ext(a/\Omega)$ es el conjunto de elementos de Ω tales que $\forall i, y_i(\omega) \in V_i$.

3.3 Objetos horda y síntesis

Una "horda" es un objeto simbólico que es usado cuando necesitamos describir una estructura compuesta por varios elementos de Ω relacionados entre ellos; por ejemplo, cuando necesitamos expresar relaciones entre elementos de una imagen que deseamos describir.

Una *horda* es definida por la aplicación $h_\Omega : D \rightarrow H$ donde H es el conjunto de aplicaciones $\Omega^p \rightarrow \{\text{verdadero, falso}\}$, tal que $h_\Omega(V) = H$ donde $V = (V_1, \ldots, V_p)$ y si $u = (u_1, \ldots, u_p)$ $H(u) =$ verdadero si y sólo si $y_i(\omega) \in V_i$.

Una horda es denotada por

$$H = \wedge_i [y_i(u_i) = V_i].$$

Note que si agregamos la restricción $u_1 = u_2 = \ldots = u_p$ una horda se convierte en una aserción.

La *extensión* de H en Ω^p es $Ext(H/\Omega^p) = \{\omega \in \Omega^p \mid H(\omega) = \text{verdadero}\}$.

Por ejemplo si Ω es un conjunto de personas en una ciudad, entonces

$$H = [y_1(u_1) = 1] \wedge [y_2(u_2) = 2] \wedge [y_3(u_1) = [30, 35] \wedge [\text{vecinos}(u_1, u_2) = \text{si}]$$

significa que u_1 es un hombre, u_2 es una mujer y ambos son vecinos.

Un *objeto síntesis* es una conjunción o una relación semántica entre hordas denotada, en el caso de la conjunción, por

$$s = \wedge_i h_i$$

donde cada horda puede estar definida sobre un conjunto diferente Ω_i por descriptores diferentes.

Por ejemplo Ω_1 puede ser individuos, Ω_2 localización, Ω_3 clase de trabajo, etc. Todos estos objetos son detallados en [8].

Ejemplo:
Sea Ω un conjunto de hongos descritos por su color y su longitud. Son representados por dos variables $\text{col}_t : \Omega \rightarrow O_c$ y $\ell_t : \Omega \rightarrow O_l$ que dependen del tiempo. Para simplificar supongamos que en cualquier tiempo ellos pueden tomar unicamente dos colores y dos clases de longitud tal que $O_{\text{col}} = \{1, 2\}$ y $O_l = \{1, 2\}$. En los tiempos t_1 y t_2 obtenemos las tablas (a) y (b) dadas abajo para el caso de un conjunto de dos hongos $\Omega_1 = \{\omega_1, \omega_2\}$;

la tabla (c) representa los valores tomados por los elementos de un conjunto de objetos Ω en un tiempo dado.

Ω_1	col_{t_1}	ℓ_{t_1}
ω_1	1	1
ω_2	2	1

Tabla (a)

Ω_1	col_{t_2}	ℓ_{t_2}
ω_1	2	1
ω_2	1	2

Tabla (b)

O	O_1	O_2
x_1	1	1
x_2	2	1
x_3	1	2
x_4	2	2

Tabla (c)

Sean a_{t_1}, a_{t_2}, c tres aserciones donde c es un ℓ-complejo tales que

$$a_{t_1} = [\text{col}_{t_1} = 1] \wedge [\ell_{t_1} = 1,2]$$

$$a_{t_2} = [\text{col}_{t_2} = 1] \wedge [\ell_{t_2} = 1,2]$$

$$c = [X_1 = 1] \wedge [X_2 = 1,2].$$

Por definición a_{t_1} y a_{t_2} son aplicaciones $\Omega \to \{\text{verdadero, falso}\}$ tales que

$$a_{t_1}(\omega_1) = [\text{col}_{t_1}(\omega_1) \in \{1\}] \wedge [l_{t_1}(\omega_1) \in \{1,2\}] = \text{verdadero;}$$

similarmente obtenemos

$$a_{t_1}(\omega_2) = \text{falso}$$

$$a_{t_2}(\omega_1) = \text{falso}$$

$$a_{t_2}(\omega_2) = \text{verdadero}$$

$$c(x_1) = c(x_3) = \text{verdadero y}$$

$$c(x_2) = c(x_4) = \text{falso.}$$

De lo anterior sigue que

$$\text{ext}(a_{t_1}/\Omega) = \{\omega_1\},$$

$$\text{ext}(a_{t_2}/\Omega) = \{\omega_2\} \text{ y}$$

$$\text{ext}(c/O) = \{x_1, x_3\}.$$

También podemos definir tres hordas como sigue:

$$h_1 = [\text{col}_{t_1}(u_1) = 1] \wedge [\ell_{t_2}(u_2) = 1,2],$$

$$h_2 = [\text{col}_{t_2}(u_1) = 1] \wedge [\ell_{t_2}(u_2) = 1,2] \text{ donde } u_i \in \Omega,$$

$$h_c = [X_1(u_1) = 1] \wedge [X_2(u_2) = 1,2] \text{ donde } u_i \in O.$$

Por tanto es fácil ver que

$$Ext(h_1/\Omega) = \{(\omega_1,\omega_1),(\omega_1,\omega_2)\},$$

$$Ext(h_2/\Omega) = \{(\omega_2,\omega_1),(\omega_2,\omega_2)\},$$

$$Ext(c,O) = \{(x_1,x_1),(x_1,x_2),(x_1,x_3),(x_1,x_4),(x_3,x_1),(x_3,x_2),(x_3,x_3),(x_4,x_4)\}.$$

4 Objetos modales

4.1 Objetos modales externos e internos

Supongamos que deseamos usar un objeto simbólico para representar los individuos de un conjunto que satisface la siguiente oración: "Es posible que su peso varíe entre 300 y 500 gramos y su color sea frecuentemente rojo y rara vez blanco"; esta oración contiene dos eventos e_1 = [peso = [300,500]] y e_2 = [color = {rojo, blanco}] en los cuales aparecen los modos *posible, frecuentemente* y *rara vez*. Una nueva clase de eventos denotados f_1 y f_2 se necesitan si deseamos introducir estos modos: f_1 = *posible*[peso = [300,500] y f_2 = [color = {*frecuentemente* rojo, *rara vez* blanco}]. Podemos ver que f_1 contiene un modo *posible* externo que afecta e_1 mientras que f_2 contiene modos internos que afectan los valores contenidos en e_2. Por tanto es posible describir la oración informalmente por un objeto aserción modal denotado $a = f_1 \wedge_x f_2$ donde $\wedge_x$ representa una clase de conjunción relacionada con el conocimiento previo del dominio. El caso de aserciones modales de la clase $a = \wedge_i F_i$ donde todos los f_i son eventos con modos externos han sido estudiados, por ejemplo en [7]. Este artículo comprende sólo el caso de todos los f_i que contienen unicamente modos internos.

4.2 Definición formal de objetos modales internos

Sean x el conocimiento previo y:

- M^x un conjunto de modos, por ejemplo
  M^x = {frecuentemente, algunas veces, rara vez, nunca} ó $M^x = [0,1]$.

- $Q_i = \{q_i^j\}_j$ un conjunto de aplicaciones de O_i a M^x, por ejemplo

  $$O_i = \{\text{rojo, amarillo, verde}\}$$

  $M^x = [0,1]$ y q_i^j(rojo) = 0.1; q_i^j(amarillo) = 0.3; q_i^j(verde) = 1, donde el significado de los valores 0.1, 0.3 y 1 depende del conocimiento previo (por ejemplo q_i^j puede expresar una posibilidad, ver §5.1)

- y_i es un descriptor o aplicación de Ω a Q_i (el *color*, por ejemplo). Note que en el caso de objetos booleanos y_i fue una aplicación de Ω a O_i y no Q_i.
  Ejemplo: Si O_i y M^x son escogidos como en el ejemplo previo y el color de ω es rojo entonces $y_i(\omega) = r$ significa que $r \in Q_i$ es definido por una aplicación característica tal que r(rojo) = 1, r(amarillo) = 0 y r(verde) = 0.

- $OP_x = \{\cup_x, \cap_x, c_x\}$ donde $\cup_x, \cap_x$ expresan una clase de unión e intersección entre subconjuntos de Q_i, y $c_x(q_i)$ (algunas veces denotado $\overline{q}_i$), es el complementario de $q_i \in Q_i$. Para ganar claridad en el concepto de unión $\cup_x$ podemos decir que $q_1 \cup_x q_2$ es una "generalización" de la observación q_1, q_2 dada, por ejemplo por dos expertos o dos sensores.

Denotamos Q_i^x el más pequeño conjunto estable para OP_x. Es decir Q_i^x es el conjunto de cualquier $*_x$ ó c_x combinación de elementos $q_i^j \in Q_i$.

Si $Q_x \subseteq Q_i^x$, denotamos Q la aplicación $Q = \cup_x\{q \mid q \in Q_x\}$. El complementario de Q_x en Q_i^x es $c(Q_x) = 1 - Q$.
Ejemplo: Si $q_i^j \in Q_i$ y $Q_i^j \subseteq Q_i$ entonces

$$q_i^1 \cup_x q_i^2 = q_i^1 + q_i^2 - q_i^1 q_i^2$$

$$\text{y } q_i^1 \cap_x q_i^2 = q_i^1 q_i^2 \text{ donde } q_i^1 q_i^2(v) = q_i^1(v) q_i^2(v) \text{ y } c_x(q_i) = 1 - q_i$$

Intuitivamente , si q_i^j es la distribución de probabilidad de las palabras contenidas en un texto T_i^j, entonces $q_i^1 q_i^2(v)$ es la probabilidad de conseguir v entre dos palabras obtenidas independientemente una en T_i^1 y la otra en T_i^2. Si $P_2 > P_1$, es menos "general" sacar una palabra entre P_1 palabras sacadas entre P_1 textos independientemente, que sacar una palabra entre P_2 palabras sacadas independientemente en P_2 textos. Esta escogencia de OP_x es "arquimediana" puesto que satisface una familia de propiedades estudiadas por Shweizer y Sklar [18] y mencionada por Dubois y Prade [10]. En §6.2 usamos estos operadores con el fin de definir objetos probabilísticos.

- g_x^i es una aplicación de "comparación" de $Q_i^x \times Q_i^x$ en un espacio ordenado L^x. En este artículo g_x^i no dependerá de i y será denotado simplemente por g_x.
  Ejemplo: $L^x = M^x = [0,1]$ y $g_x(q_i^1, q_i^2) =< q_i^1, q_i^2 >$ el producto escalar.

- f_x es una aplicación de "agregación de $P(L^x)$, el conjunto potencia de L^x, a L^x. Por ejemplo $f_x(\{L_1, \ldots, L_n\}) = \text{Max} L_i$.

Sea $\{y_i\}$ un conjunto de descriptores y $\{q_i^j\}_j \subseteq Q_i^x$. Ahora somos capaces de dar una definición formal de un objeto modal interno (llamado objeto *im*). Se trata de un objeto simbólico con $D = P(Q_1^x) \times \ldots P(Q_p^x)$ y $h(d) = a$ donde $d = (\{q_1^j\}_j, \ldots, \{q_p^j\}_j)$ y a es una aserción *im* definida como sigue:

Definición de una aserción *im*:
dados OP_x, g_x y f_x, una aserción es una aplicación de Ω en un espacio ordenado L^x, denotado por $a = \wedge_i[y_i = \{q_i^j\}_j]$ tal que si $\omega \in \Omega$ es descrito para cualquier i por $y_i(\omega) = r_i$ entonces a es dado por: $\{\forall \omega \in \Omega, a(\omega) = f_x(g_x(\bigcup_j {}_x q_i^j, r_i)i)\}$.

Denotamos por $\mathcal{A}_x$ el conjunto de los objetos *im* asociados al conocimiento previo x, y por ϕ la aplicación de Ω a $\mathcal{A}_x$ tal que $\phi(\omega = \omega^S = \bigwedge_i {}_x[y_i = y_i(\omega)]$.

Por convención en todo este artículo un evento $[y_i = \{q_i^j\}_j]$ puede también ser denotado $[y_i = q_i^1, q_i^2, \ldots]$. Resulta de la definición que $[y_i = \{q_i^j\}_j]$ es equivalente al evento $[y_i = \bigcup_j {}_x q_i^j]$. En otras palabras, usando la notación anterior, el evento $[y_i = Q_x]$ será considerado equivalente a $[y_i = Q]$.

La x-unión de dos aserciones a_1 y a_2 denotada por $a_j = \bigwedge_i{}_x[y_i = \{q_i^j\}$ es definida por $a_1 \cup_x a_2 = \bigwedge_i{}_x[y_i = q_i^1 \cup_x q_i^2]$; más generalmente tenemos $\bigcup_j{}_x a_j = \bigwedge_i{}_x[y_i = \bigcup_j{}_x q_i^j]$. Por tanto, resulta con nuestra convención que $\bigcup_j{}_x a_j = \wedge_i[y_i = \{q_i^j\}_j]$. La intersección de aserciones es definida en la misma forma:$(\cap_x)_j a_j = \wedge_i[y_i = (\cap_x)_j q_i^j]$. Los operadores OP_x extendidos sobre $\mathcal{A}_x$ será mejor estudiado en §9.

Hay al menos dos formas de definir la extensión de un objeto *im* a. La primera consiste en considerar que cada elemento ω de Ω está en la extensión acorde con su peso dado por $a(\omega)$; en este caso la extensión de a denotada por $\text{Ext}(a/\Omega)$ será el conjunto de pares $\{(\omega, a(\omega)) \mid \omega \in \Omega\}$. La segunda requiere un umbral α dado y entonces la extensión será $\text{Ext}(a/\Omega, \alpha) = \{(\omega, a(\omega)) \mid \omega \in \Omega, a(\omega) \geq \alpha\}$.

4.3 Semánticas de los objetos *im*

Además de los modos, otros conceptos pueden ser expresados para un objeto *im* a:

a) Certidumbre: $a(\omega)$ no es verdadero o falso como en el caso de los objetos booleanos pero expresa un grado de certidumbre.

b) Variación: Aparece en dos niveles en un objeto *im* denotado $a = \bigwedge_i{}_x[y_i = \{q_i^j\}_j]$. Primero, en cada q_i^j, por ejemplo si y_i es el color y $q_i^1(\text{rojo}) = 0.5$ y $q_i^1(\text{verde}) = 0.3$ ello significa que una variación existe entre los objetos individuales que pertenecen a la extensión de a (por ejemplo una especie de hongos) donde algunos son rojos y otros verdes. Segundo, para una descripción y_i dada y $v \in O_i$ hay una variación entre los $q_i^j(v)$ cuando j varía (cada $q_i^j(v)$ expresa por ejemplo la variación del color v entre las diferentes clases de especies).

c) Duda: Si decimos que el color de una especie de hongo es rojo "o" verde, se trata de un "o" de variación, pero si decimos que el color del hongo que tengo en mi mano es rojo "o" verde, se trata de un "o" de duda. Por tanto, si describimos $\omega \in \Omega$ por $\phi(\omega) = \omega^S = \wedge_i[y_i = y_i(\omega)]$ donde $y_i(\omega) = \{r_i^j\}_j$ expresamos una vaguedad o una imprecisión en cada r_i^j y una duda entre los r_i^j provistos, por ejemplo, por varios expertos.

4.4 Un ejemplo de conocimiento previo expresando intensidad

Aquí el conocimiento previo x es denotado por i (de intensidad). Cada objeto individual $\omega \in \Omega$ es un objeto manufacturado descrito por dos características: y_1 la cual expresa el grado de redondez o lo plano que es un objeto, y y_1, la pesadez, donde $O_1 = \{\text{plano, redondo}\}$, $O_2 = \{\text{pesado}\}$ y $M^i = \{\textit{muy, bastante, un poco, muy poco, nil}\}$.

Sean a y ω^S definidos por:

$$\begin{aligned} a &= [y_1 = \textit{un poco} \text{ plano}, \textit{bastante} \text{ redondo}] \wedge_i [y_2 = \textit{un poco} \text{ pesado}] \\ \omega^S &= [y_1 = \textit{bastante} \text{ redondo}] \wedge_i [y_2 = \textit{muy} \text{ pesado}, \textit{bastante} \text{ pesado}] \end{aligned}$$

En el caso de ω el usuario tiene duda entre "*muy* pesado" y "*bastante* pesado". El problema es conocer si es aceptable decir que ω pertenece a la clase de objetos manufacturados descritos por a. Por tanto:
$q_1^1(\text{plano}) = \textit{un poco}$, $q_1^1(\text{redondo}) = \textit{bastante}$, $q_2^1(\text{pesado}) = \textit{un poco}$, $r_1^1(\text{plano}) = \textit{nil}$, $r_1^1(\text{redondo}) = \textit{bastante}$, $r_2^1(\text{pesado}) = \textit{muy}$, $r_2^2(\text{pesado}) = \textit{bastante}$.

Una taxonomía (Tax) dada que expresa el conocimiento previo sobre los valores de M^i hace posible decir que $\text{Tax}(\textit{muy}, \textit{bastante}) = \textit{algo}$, por tanto si establecemos que $r_2^1 \cup_i r_2^2(v) = \text{Tax}(r_2^1(v), r_2^2(v))$ tenemos que $r_2^1 \cup_i r_2^2(\text{pesado}) = \text{Tax}(\textit{muy}, \textit{bastante}) = \textit{algo}$.

Definimos L_j por L_1 = no aceptable, L_2 = aceptable, L_3 = completamente aceptable. Suponemos que la aplicación de comparación g_i es dada por una tabla T_{g_i} tal que:

$$\begin{aligned} g_i(q_1^1, r_1^1) &= T_{g_i}((\textit{un poco} \text{ plano}, \textit{bastante} \text{ redondo}), \\ &\qquad (\textit{nil} \text{ plano}, \textit{bastante} \text{ redondo})) \\ &= \text{aceptable} \end{aligned}$$

$$\begin{aligned} g_i(q_2^1, r_2^1 \cup_i r_2^2) &= T_{g_i}(\textit{un poco} \text{ pesado}, \textit{algo} \text{ pesado}) \\ &= \text{no aceptable} \end{aligned}$$

Finalmente, si ponemos $f(L_j) = \min L_j$ y $L_1 < L_2 < L_3$ obtenemos $a(\omega) = f_i(g_i(q_1^1, r_1^1), g_i(q_2^1, r_2^1 \cup_i r_2^2)) = f_i(\text{aceptable}, \text{no aceptable}) = \text{no aceptable}$.

Note que objetos más complejos puden ocurrir cuando en lugar de sólo uno, como en la definición precedente, varios eventos conciernen a la misma variable. Por ejemplo, si tenemos $a = \wedge_i a_i$ con $a_i = \bigwedge_{\ell} {}_x[y_i = a_i^\ell]$; en este caso es necesario introducir una tercera aplicación h de $P(L^x)$ a L^x tal que $a_i(\omega) = h(\{g(q_i^\ell, r_i)\}_\ell)$. Por tanto, más generalmente, si $a = \bigwedge_i {}_x a_i = \bigwedge_i {}_x \bigwedge_\ell {}_x[y_i = q_i^\ell]$ entonces $a(\omega) = f_x(\{a_i(\omega)\}_i) = f_x(\{h_x(\{g_x(q_i^\ell, r_i)\}_\ell)\}_i)$.

El siguiente ejemplo puede ser omitido en una primera lectura. Su propósito es construir una aserción a_i formada por una conjunción de los eventos para los cuales la extensión al nivel $\frac{1}{2}$ contiene un $\omega \in \Omega$ dado.

Ejemplo:
Sea $M_i^x = [0,1]$, $O_i = \{v_1, v_2\}$ y Q_i el conjunto de medidas de probabilidad $P(O_i) \longrightarrow [0,1]$; y es una aplicación de un conjunto Ω a Q_i y $\omega \in \Omega$ es descrito por $\omega^S = [y_i = r]$ tal que

$r(v_1) = r(v_2) = \frac{1}{2}$. El conjunto de eventos $im\ e_i = [y = q_i]$ tal que $a_i(\omega) \geq \frac{1}{2}$ se define por el conjunto de medidas de probabilidad q_i que satisfacen la desigualdad $e_i(\omega) = f_x(g_x(q_i, r)) \geq \frac{1}{2}$. Si f_x es la media y g_x es el producto escalar obtenemos $e_i(\omega) = \text{Media}(\{\langle q_i, r\rangle\}) = \langle q_i r\rangle$ cuando hay sólo una variable. Por tanto, q_i debe satisfacer la siguiente desigualdad:

$$e_i(\omega) = \langle q_i, r\rangle = q_i(v_1)r(v_1) + q_i(v_2)r(v_2) \geq \frac{1}{2}$$

lo cual es equivalente a:

$$\frac{1}{2}q_i(v_1) + \frac{1}{2}q_i(v_2) \geq \frac{1}{2}$$

lo cual es satisfecho por cualquier evento e_i, como $q(v_1)+q(v_2) = 1$ para cualquier medida de probabilidad q definida sobre O_i. Si $a_i = \bigwedge_{\ell} {}_x\{e_i^\ell | e_i^\ell(\omega) \geq \frac{1}{2}\}$ entonces $a_i(\omega) = h_x(\{e_i^\ell(\omega)\}_\ell)$. Si $h_x = \text{mín}$ entonces $a_i(\omega) = \text{mín}(\{e_i^\ell(\omega)\}_\ell) = \frac{1}{2}$.

5 Objetos posibilísticos

5.1 El enfoque posibilista

Para dar la idea principal de este enfoque, seguimos a Dubois y Prade [10].

Definición de una medida de posibilidad y de necesidad:
Una medida de posibilidad es una aplicación Π *de* $P(\Omega)$ *a* $[0,1]$ *tal que:*

(1) $\Pi(\Omega) = 1$ *y* $\Pi(\emptyset) = 0$.

(2) $\forall A, B \subseteq \Omega \quad \Pi(A \cup B) = \textit{máx}(\Pi(A), \Pi(B))$.

Una medida de necesidad es una aplicación de $P(\Omega)$ *a* $[0,1]$ *tal que:*

(3) $\forall A \subseteq \Omega$ *se tiene* $N(A) = 1 - \Pi(\overline{A})$.

A partir de las definiciones anteriores se pueden probar las siguientes propiedades:

- $N(\emptyset) = 0$, $N(A \cap B) = \text{mín}(N(A), N(B)$, $\Pi(\cup_i A_i) = \text{máx}_i(\Pi(A_i))$
- $N(\cap A_i) = \text{mín} N(A_i)$, $\Pi(A) \leq \Pi(B)$ si $A \subseteq B$, $\text{máx}(\Pi(A), \Pi(\overline{A})) = 1$
- $\text{mín}(N(A), N(\overline{A})) = 0$, $\Pi(A) \geq N(A)$, si $N(A) > 0$ entonces $\Pi(A) = 1$
- si $\Pi(A) < 1$ entonces $N(A) = 0$, $\Pi(A) + \Pi(\overline{A}) \geq 0$ y $N(A) + N(\overline{A}) \leq 1$

Ejemplo:
Definimos $\Pi_E(A)$ (respectivamente $N_E(A)$) como la posibilidad (respectivamente la necesidad) de obtener $\omega \in A$ cuando $\omega \in E$. Decimos que $\Pi_E(A) = 1$ si esta posibilidad es cierta

y $\Pi_E(A) = 0$ si no. Por tanto, Π_E y N_E son aplicaciones de $P(\Omega)$ a $\{0,1\}$. Es entonces fácil probar que estas aplicaciones satisfacen las tres condiciones de su definición.

La teoría de posibilidad modela varias clases de semánticas. Generalmente posibilidades valuadas en observaciones vagas de características inaccesibles, por ejemplo:

i) La posibilidad física: expresa la dificultad material de una acción por ocurrir. Por ejemplo, si varios expertos han descrito que un atleta tiene la posibilidad $\Pi(\{200\}) = 0,8$ de cargar 200 Kg y la posibilidad $\Pi(\{250\}) = 0,5$ de cargar 250 Kg, entonces, para estos expertos, la posibilidad del atleta de cargar 200 ó 250 Kg será:

$$\Pi(\{200 \cup_p \{250\}) = \text{máx}(\Pi(\{200\}), \{250\})) = 0,8$$

ii) La posibilidad como una concordancia con conocimiento actual: "es posible que llueva o nieve hoy".

iii) La no sorpresa: "la 'tipicalidad' del color de una flor de ser amarilla o café".

5.2 Una definición formal de objetos posibilísticos

Aquí el conocimiento previo x es denotado por p (de posibilidad).

Definición: Una aserción posibilística denotada $a_p = \wedge_i[y_i = \{q_i^j\}_j]$ es una aserción *im* que toma sus valores en $L^p = [0,1]$ tal que

- Para todo i, Q_i es un conjunto de medidas de posibilidad.
- OP_p: para todo i, $q_i^1, q_i^2 \in Q_i$ se define $q_i^1 \cup_p q_i^2 = \text{máx}(q_i^1, q_i^2)$, $q_i^1 \cap_p q_i^2 = \text{mín}(q_i^1, q_i^2)$.
- $c_p(q) = 1$ denotado también $\overline{q}$.
- $g_p(q_i^1, q_i^2) = \sup\{\text{mín}(q_i^1(v), q_i^2(v)) | v \in O_i\}$.
- f_p se define por: para todo $L \subseteq [0,1]$, $f_p(L) = \text{máx}\{\ell | \ell \in L\}$.

Note que OP_p es definido como en conjuntos difusos y g_p ha sido también propuesto por Zadeh [19]. Note además que $q_i^1 \cap_p q_i^2$ no necesariamente es una medida de posibilidad.

Es también posible definir una aserción "necesitista" a_n (agradezco a M.O. Menessier, D. Dubois y H. Prade, por sus útiles observaciones, las cuales me han permitido mejorar este punto) definiendo: $a_n = 1 - \overline{a_p}$ donde $\overline{a_p} = \bigwedge_{i p}[y_i = \overline{q_i}]$ y $\overline{q_i} = c_p(q_i) = 1 - q_i$.

Esto resulta en $a_n(\omega) = 1 - f_p(\{g_p(\overline{q_i}, r_i\}_i)$ y entonces:

$$
\begin{aligned}
a_n(\omega) &= 1 - \mathrm{Max}_i g_p(\overline{q_i}, r_i) \\
&= 1 - \text{máx}\{\sup\{\text{mín}(\overline{q_i}(v), r_i(v)) | v \in O_i\}\}_i \\
&= \text{mín}\{1 - \sup \text{mín}(\overline{q_i}(v), r_i(v)) | v \in O_i\}\}_i \\
&= \text{mín}\{\inf\{1 - \text{mín}(\overline{q_i}(v), r_i(v)) | v \in O_i\}\}_i \\
&= \text{mín} \inf\{\text{máx}(q_i(v), 1 - r_i(v)) | v \in O_i\}
\end{aligned}
$$

Finalmente: $a_n(\omega) = \text{mín} g_n(q_i, r_i)$

Resulta de lo anterior que un objeto necesitista es definido por $OP_n = \{\cup_n, \cap_n, c_n\}$ donde $\cup_n$ es $\cup_p$, $\cap_n$ es $\cap_p$ y c_n es c_p; $g_n(q_i, r_i) = \inf\{\text{máx}(q_i(v), \overline{r_i}(v)) | v \in O_i\}$ y $f_n = \text{mín}$.

Ejemplo:
Un experto describe una clase de objetos por la siguiente aserción posibilística (limitada, para simplificar, a un solo evento):

$$e_p = [\text{altura} = [\text{alrededor de}[12, 15], \text{aproximadamente}\{18\}]]$$

Un objeto elemental ω es definido por $\omega^S = [\text{altura} = \text{cerca de } 16]$.
El problema es encontrar la posibilidad y la necesidad de ω si se conoce e_p, en el caso donde e_p y ω^S se escriben: $e_p = [\text{altura} = q_1, q_2]$ y $\omega^S = [\text{altura} = r_1]$ donde q_1, q_2 y r_1 son aplicaciones posibilísticas de $O = [0, 20]$ a $[0, 1]$ definidas por el conocimiento previo de la figura 5. Esto significa que un objeto de altura 14 (respectivamente 10), tiene una posibilidad 1 (respectivamente 0,3).

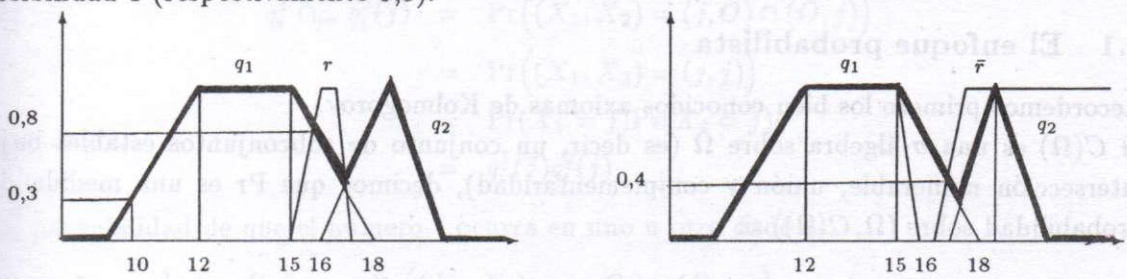


Figura 5
(a) $q_1 \cup q_2 = \max(q_1, q_2)$
(b) $\bar{r}_i = 1 - r_i$

Es entonces posible calcular la posibilidad de ω por:

$$e_p(\omega) = g_p(q_1 \cup_p q_2, r_1) = \sup\ \{\text{mín}(q_1 \cup_p q_2(v), r_1(v))|v \in O\} = 0,8$$

La necesidad de ω es dada por:

$$e_n(\omega) = g_n(q_1 \cup_p q_2, r_1) = \inf\{\text{máx}(q_1 \cup_p q_2(v), \overline{r}_1(v))|v \in O\} = 0,4$$

Este ejemplo muestra que los objetos posibilísticos pueden representar no sólo certidumbre, variación y duda, sino también inexactitud (alrededor de, aproximadamente, cerca de). Es también posible usar vaguedad al representar "alto" o "pesado" por una medida de posibilidad.

5.3 El caso particular de los objetos booleanos

Un objeto booleano $a = \wedge_i[y_i = V_i]$ es un objeto *im* $a_b = \wedge_i[y_i = q_i]$ donde q_i es una aplicación característica de $V_i \subseteq O_i$, $OP_b = \{\cup_b, \cap_b, c_b\}$ es tal que $q_i \cup_b q_2 = \text{máx}(q_1, q_2)$, $q_1 \cap_b q_2 = \text{mín}(q_1, q_2)$ y $c_b(q) = 1 - q$. Hay dos escogencias para g_b y f_b: $(g_b, f_b)) = (g_p, f_p)$ o $(g_b, f_b) = (g_n, f_n)$.

Si $\omega^S = \wedge_i[y_i = r_i]$ donde r_i es la aplicación característica de $y_i(\omega) \subseteq O_i$ (hay duda si $y_i(\omega)$ no es reducido a un sólo elemento), es entonces fácil mostrar que $y_i(\omega) \cap V_i \neq \emptyset \iff a_b(\omega) = 1$ en la escogencia posibilística y, $y_i(\omega) \subseteq V_i \iff a_b(\omega) = 1$ en la escogencia necesitista.

Si denotamos $|a|_\Omega$ el conjunto de elementos de Ω tales que $a(\omega) =$ verdadero tenemos $|a|_\Omega = \text{Ext}(a_b/\Omega, \alpha) \quad \forall \alpha \in]0, 1]$, para ambas escogencias.

6 Objetos probabilísticos

6.1 El enfoque probabilista

Recordemos primero los bien conocidos axiomas de Kolmogorov.
Si $C(\Omega)$ es una σ-álgebra sobre Ω (es decir, un conjunto de subconjuntos estables bajo intersección numerable, unión y complementaridad), decimos que Pr es una medida de probabilidad sobre $(\Omega, C(\Omega))$ si:

i) $\Pr(\Omega) = 1$

ii) $\Pr(\cup_i A_i) = \sum \Pr(A_i)$ si $A_i \in C(\Omega)$ y $A_i \cap A_j = \emptyset$

Hay varias semánticas que respetan estos axiomas: por ejemplo el azar en juegos, frecuencias, algunas clases de incertidumbre en probabilidad subjetiva. Sea Q_i un conjunto de medidas de probabilidades definidas sobre $(O_i, C(O_i))$. Suponemos que los $\omega^S = [y_i = y_i(\omega)]$ son tales que $y_i(\omega) \in Q_i$. Recordemos que Q_i^x fue definido en §4.1.

6.2 Una definición formal de objetos probabilísticos

Definición: Una aserción probabilística es una aserción *im* que toma sus valores en $L^{pr} = [0,1]$:

$$
\begin{aligned}
OP_{pr}: \quad & \forall q_i^1, q_i^2 \in Q_i \quad q_i^1 \cup_{pr} q_i^2 = q_i^1 + q_i^2 - q_i^1 q_i^2 \ ; \\
& q_i^1 \cap_{pr} q_i^2 = q_i^1 q_i^2 \text{ que es la aplicación que asocia a } v \in O_i \quad q_i^1(v) q_i^2(v); \\
& c_{pr}(q) = \overline{q} = 1 - q \\
g_{pr}: \quad & \forall \{q_i^1, q_i^2\} \in Q_i^x \times Q_i \quad g_{pr}(q_i^1, q_i^2) = \langle q_i^1, q_i^2 \rangle = \sum \{q_i^1(v) q_i^2(v) | v \in Q_i\} \\
f_{pr}: \quad & f_{pr}(\{L_i\}) = \text{ media de los } L_i
\end{aligned}
$$

Note que si hubiesen algunas dependencias características entre variables, entonces un evento de la forma $[y_i = q_i]$ puede representarlas; por ejemplo, si el experto desea describir las dependencias entre y_1, y_3, y_7, entonces esta información puede ser representada por el evento denotado $y_{137} = \text{pr}(y_1, y_3, y_7)]$ donde $\text{pr}(y_1, y_3, y_7)$ representa la probabilidad conjunta de y_1, y_3 y y_7. Este evento es de la forma $y_i = q_i]$ donde $y_i = y_{137}$ y $q_i = \text{pr}(y_1, y_3, y_7)$. En casos donde "causalidades" o "influencias" entre conjuntos de variables son dadas por el experto para describir un objeto simbólico, técnicas de propagación (ver [12] o [15]), pueden ser usadas para inducir otras aplicaciones g_{pr} y f_{pr}.

Para dar una idea intuitiva del concepto de unión y de intersección de medidas de probabilidad es fácil ver que si q_i^1 y q_i^2 son las medidas de probabilidad asociadas a dos dados, $q_i^1 \cup_{pr} q_i^2(v)$, con $v \in O_i$, es la probabilidad de que el evento v ocurra, para un dado o (no exclusivo) para el otro, $q_i^1 \cap_{pr} q_i^2$ es la probabilidad de que el evento v ocurra para ambos dados cuando los dos dados son tirados independientemente. Esto resulta del hecho que si (X_1, X_2) es un par de variables aleatorias $\Omega \longrightarrow O_i \times O_i$ donde $O_i = \{1, 2, \ldots, 6\}$ con probabilidad $(q_i^1 q_i^2)$, entonces la probabilidad de que el número j ocurra en ambos dados tirados independientemente es

$$
\begin{aligned}
q_i^1 \cap_{pr} q_i^2(j) &= \Pr\Big((X_1, X_2) = (j, O) \cap (O, j)\Big) \\
&= \Pr\Big((X_1, X_2) = (j, j)\Big) \\
&= \Pr(X_1 = j) \Pr(X_2 = j) \\
&= q_i^1(j) q_i^2(j)
\end{aligned}
$$

La probabilidad de que el número j ocurra en uno u otro dado es:

$$
\begin{aligned}
q_i^1 \cup_{pr} q_i^2(j) &= \Pr\Big((X_1, X_2) = (j, O_i) \cup (O_i, j)\Big) \\
&= \Pr\Big((X_1, X_2) = (j, O_i)\Big) + \Pr\Big((X_1, X_2) = (O_i, j)\Big) \\
&\quad - \Pr\Big((X_1, X_2) = (O_i, j) \cap (j, O_i)\Big) \\
&= q_i^1(j) q_i^2(O_i) + q_i^1(O_i) q_i^2(j) - q_i^1(j) q_i^2(j) \\
&= (q_i^1 + q_i^2 - q_i^1 q_i^2)(j)
\end{aligned}
$$

Note que $q_i^1 \cup_{pr} q_i^2$ no es una medida de probabilidad puesto que aun si $q_i^1 \cup_{pr} q_i^2(v) \in [0,1]$ la suma de los $q_i^1 \cup_{pr} q_i^2(v)$ sobre O_i es más grande que 1. Tampoco $q_i^1 \cap_{pr} q_i^2$ es una medida de probabilidad puesto que la suma de los $q_i^1 \cap_{pr} q_i^2(v)$ sobre O_i puede ser menor que 1. Hemos definido g sobre $Q_i^x \times Q_i$ y no sobre $Q_i^x \times Q_i^x$ como para un objeto *im* general, porque por ejemplo $g(q_i^1 \cup_{pr} q_i^2, \cup_j q_i^j)$ se puede volver más grande que 1. Pero, en este caso es fácil transformar $q_i^1 \cup_{pr} q_i^2$ en una medida de probabilidad dividiéndola por la suma de los $q_i^1 \cup_{pr} q_i^2(v)$ sobre O_i.

Ejemplo:
Un objeto ω es descrito por su color $y_1(\omega)$ el cual puede ser rojo o azul y por su redondez $y_2(\omega)$ la cual puede ser redonda o plana. Sea $a = [y_i = q_1^1, q_1^2] \wedge_{pr} [y_2 = q_2]$ y $\omega^S = [y_1 = r_1] \wedge_{pr} [y_2 = r_2]$ donde $q_1^1(\text{rojo}) = 0.9$; $q_1^1(\text{azul}) = 0.1$; $q_1^2(\text{rojo}) = 0.5$; $q_1^2(\text{azul}) = 0.5$; $q_2(\text{redondo}) = 0.2$; $q^2(\text{plano}) = 0.8$. De lo cual resulta que a es descrito por dos clases de objetos: cualquiera de "a menudo rojo y raramente azul" o "rojo o azul con probabilidad igual".

Usando $q_1^3 = q_1^1 \cup_{pr} q_1^2 = q_1^1 + q_1^2 - q_1^1 q_1^2$ obtenemos

- $q_1^3(\text{rojo}) = 0.9 + 0.5 - 0.9 \times 0.5 = 0.95$,
- $q_1^3(\text{azul}) = 0.1 + 0.5 - 0.1 \times 0.5 = 0.55$.

Si r_1 y r_2 son definidos como sigue:
$r_1(\text{rojo}) = 1$, $r_1(\text{azul}) = 0$, $r_2(\text{rojo}) = 1$, $r_2(\text{plano}) = 0$, entonces

$$\begin{aligned} a(\omega) &= g_{pr}(q_1^3, r_1) \wedge_{pr} g_{pr}(q_2, r_2) \\ &= (0.95 \times 1 + 0.55 \times 0) \wedge_{pr} (0.2 \times 1 + 0.8 \times 0) \\ &= 0.95 \wedge_{pr} 0.2 \\ &= 1/2(0.95 + 0.2) \\ &= 0.57 \end{aligned}$$

lo cual representa la probabilidad promedio de que una instancia de la clase de objetos descrita por a sea ω y pueda ser interpretada como una clase de grado de pertenencia para ω al objeto *im* definido por a.

7 Objetos creencia

7.1 El formalismo de la función creencia

Entre los trabajos que originaron esta teoría podemos mencionar al menos el de Choquet [4] sobre "capacidades de orden 2"y el de Dempster [6] sobre "probabilidades superior e inferior

inducidas por una aplicación multivaluada". Las nociones básicas de este formalismo se encuentran en el libro de Schafer [17]: "A Mathematical Theory of Evidence", una referencia estándar para esta teoría.

Una **función de asignación de probabilidad** m, de $P(\Omega)$ (Ω finito) en $[0,1]$ es definida por: $\sum\{m(V)|V \in P(\Omega)\} = 1$ y $m(\emptyset) = 0$.

Una **función de creencia** Cre: $P(\Omega) \longrightarrow [0,1]$ se define por:

$$\mathrm{Cre}(A) = \sum\{m(V)|V \in P(\Omega), V \subseteq A\}$$

Un **cuerpo de evidencia** es visto como un par $(\mathcal{F}, m)$ donde m es una función de asignación de probabilidad y $\mathcal{F} = \{V \in P(\Omega)|m(V) \neq 0\}$ es el conjunto de elementos "focales". Dado un cuerpo de evidencia es posible definir exactamente una función de creencia; también es posible definir una función *plausibilista* $\mathrm{Pl} : P(\Omega) \longrightarrow [0,1]$ tal que

$$\mathrm{Pl}(A) = \sum\{m(V)|V \in P(\Omega), V \cap A \neq \emptyset\}$$

y entonces tenemos: $\mathrm{Cre}(A) = 1 - \mathrm{Pl}(\overline{A})$.

Se pueden probar [17] que Cre es una función de creencia si y sólo si:

i) $\mathrm{Cre}(\Omega) = 1$

ii) $\mathrm{Cre}(\emptyset) = 0$

iii) $$\begin{aligned}\mathrm{Cre}(A_1 \cup \ldots \cup A_n) &\geq \sum_i \mathrm{Cre}(A_i) - \sum_{i<j} \mathrm{Cre}(A_i \cap A_j) + \ldots \\ &= \sum_{\substack{I \subseteq \{1,\ldots,n\} \\ I \neq \emptyset}} (-1)^{|I|+1} \mathrm{Cre}(\cap_{i \in I} A_i)\end{aligned}$$

donde $|I|$ denota la cardinalidad de I.

Como una consecuencia de iii) obtenemos:

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

Dada una función de creencia Cre, la función básica de asignación de probabilidad m relacionada con Cre es obtenida por:

$$\forall A \subseteq P(\Omega) \quad m(A) = \sum_{B \subseteq A} (-1)^{|A-B|} \mathrm{Cre}(B)$$

Dadas dos funciones de creencia Cre_1 y Cre_2, su suma ortogonal $\mathrm{Cre}_1 \oplus \mathrm{Cre}_2$, también conocida como regla de combinación de Dempster, es definida por sus asignaciones de probabilidad:

$$m_1 \oplus m_2(A) = \frac{\sum_{V_1 \cap V_2 = A} m_1(V_1) m_2(V_2)}{\sum_{V_1 \cap V_2 \neq \emptyset} m_1(V_1) m_2(V_2)}$$

Como un caso especial obtenemos una generalización de la regla de condicionamiento de Bayes, la cual es conocida como condicionamiento de Dempster:

$$\mathrm{Cre}(A/B) = \frac{\mathrm{Cre}(A \cup \overline{B}) - \mathrm{Cre}(\overline{B})}{(1 - \mathrm{Cre}(\overline{B}))}$$

Tenemos la siguiente relación entre las teorías de probabilidad y posibilidad: se puede mostrar que si $\mathcal{F}$ contiene sólo conjuntos unitarios, entonces Cre es una medida de probabilidad clásica. Dempster [6] dice que Pl y Cre pueden ser vistos como probabilidades superior e inferior. Schafer [17] mostró que si $\mathcal{F}$ contiene sólo una secuencia anidada de subconjuntos $V_1 \subseteq V_2 \subseteq \ldots \subseteq V_n$ entonces tenemos: $\mathrm{Cre}(A \cap B) = \mathrm{Min}(\mathrm{Cre}(A), \mathrm{Cre}(B))$ y $\mathrm{Pl}(A \cup B) = \mathrm{Max}(\mathrm{Pl}(A), \mathrm{Pl}(B))$ y por tanto, en este caso, Cre y Pl satisfacen respectivamente las propiedades de las medidas de necesidad y posibilidad. Dada una medida de probabilidad Pr, se puede mostrar que existen funciones de posibilidad, de necesidad, de creencia y de plausibilidad, denotas respectivamente pos, nec, cre y pla, tales que nec $\leq$ cre $\leq$ pr $\leq$ pla $\leq$ pos.

La teoría de creencia modela varias clases de conocimiento:

i) Probabilidad: dice J. Pearl [15] "las funciones de creencia resultan de las probabilidades de asignación a conjuntos más que a puntos individuales".

Ejemplo:
Una máquina es capaz de calcular el promedio de vehículos cuyas velocidades varían en un intervalo dado *a priori*, por ejemplo $V_1 = [0, 110]$. Algunas veces la máquina falla al dar la velocidad pero, es capaz de dar el número de vehículos que pasan sobre la carretera. Si la máquina indica, por ejemplo, 40% para velocidades en V_1, 50% en $V_2 = \{\text{velocidad} > 110\}$ y 10% cuando la velocidad es desconocida, podemos representar esta información por una función de creencia q con cuerpo de evidencia $(\mathcal{F}, m)$ tal que:

$$\mathcal{F} = \{V_1, V_2, I\!R^+\}, m(V_1) = 0.4, m(V_2) = 0.5 \text{ y } m(I\!R^+) = 0.1$$

Entonces obtenemos, por ejemplo $\mathrm{cre}([0, 130]) = 0.4$ y $\mathrm{pla}([0, 130]) = 0.4 + 0.5 = 0.9$.

ii) Testimonio: si dos testigos observan el mismo evento A, entonces usando la regla de Dempster se puede probar que la creencia en A crece. Si uno observa A y el otro B

con $A \neq B$ y $A \cap B \neq \emptyset$ entonces se puede probar que la creencia en A y B decrece. Si $A \cap B = \emptyset$ la creencia en A y B decrece más que en el caso anterior, y cuanto mayor sea la creencia en B menor será la creencia en A.

Ejemplo:

Después de un accidente observado por dos testigos, el primero está casi seguro que el auto viajaba a no más de 100 Km por hora ($V_1 =]0, 100]$) y el segundo testigo, quien estaba más lejos, piensa lo mismo pero está menos seguro. Por tanto cada testigo puede ser representado por una función de creencia: el primero por q_1, con cuerpo de evidencia $(\mathcal{F}_1, m_1)$ tal que $\mathcal{F}_1 = [V_1, \mathbb{R}^+]$, $m_1(V_1) = 0.9$ y q_2 definido por $(\mathcal{F}_2, m_2)$ tal que $\mathcal{F}_1 = \mathcal{F}_2$ y $m_2(V_1) = 0.7$. Entonces usando la regla de Dempster se obtiene:

$$q_1 \oplus q_2(V_1) = q_1(V_1) + q_2(V_1) - q_1(V_1)q_2(V_1) = 0.9 + 0.7 - 0.63 = 0.97$$

7.2 Una definición formal de objetos creencia

De acuerdo con Dubois y Prade [11], definimos la unión e intersección de dos cuerpos de evidencia $(\mathcal{F}_1, m_1)$ y $(\mathcal{F}_2, m_2)$ como sigue:
$\forall A \in P(\Omega)$,

- $m_1 \cup_{\text{cre}} m_2(A) = \sum_{V_1 \cup V_2 = A} m_1(V_1) m_2(V_2)$.
- $m_1 \cap_{\text{cre}} m_2(A) = \sum_{V_1 \cap V_2 = A} m_1(V_1) m_2(V_2)$.

La definición anterior es consistente con la regla de Dempster si el término $m_1 \cap m_2(\emptyset)$ (que refleja la cantidad de desacuerdo entre las fuentes o su independencia) es eliminado. En la siguiente definición denotamos por q_i^j una función de creencia con cuerpo de evidencia $(\mathcal{F}_i^j, m_i^j)$.

Definición

Una aserción creencia denotada $a_{\text{cre}} = \bigwedge_i {}_{\text{cre}}[y_i = \{q_i^j\}_j]$ es una aserción *im* que toma sus valores en $L^{\text{cre}} = [0, 1]$ tal que:

- $\forall i \quad Q_i$ es un conjunto de funciones creencia definidas sobre O_i.
- OP_{cre}: $\forall i \quad q_i^1, q_i^2 \in Q_i$, definimos:
  $q_i^1 \cup_{\text{cre}} q_i^2(V) = \sum_{A \subseteq V} m_i^1 \cap_{\text{cre}} m_i^2(A)$;
  $q_i^1 \cap_{\text{cre}} q_i^2(V) = \sum_{A \subseteq V} m_i^1 \cup_{\text{cre}} m_i^2(A)$;
  $c_{\text{cre}}(q_i^j)(V) = \overline{q}_i(V) = \sum_{A \subseteq V} \overline{m}_i^j(A)$ donde $\overline{m}_i^j(A) = m_i^j(\overline{A})$.
- g_{cre}: $g_{\text{cre}}(q_i^1, q_i^2) = \sum\{m_i^1 \cap_{\text{cre}} m_i^2(V_2) | V_2 \subseteq V_1, (V_1, V_2) \in \mathcal{F}_1 \times \mathcal{F}_2\}$.

- f es la media.

Note que la unión e intersección de funciones creencia son funciones creencia (al contrario del caso de probabilidades como de posibilidades). Igual que en el caso de los objetos probabilísticos, la escogencia de la función f puede ser más general, aquí hemos escogido la media a fin de simplificar.

Es también posible definir un *objeto plausibilista* por:

- $OP_{p\ell}$:
  $q_i^1 \cup_{p\ell} q_i^2(V) = \sum_{A \cap V \neq \emptyset} m_i^1 \cap m_i^2(A)$;
  $q_i^1 \cap_{p\ell} q_i^2(V) = \sum_{A \cap V \neq \emptyset} m_i^1 \cup m_i^2(A)$;
  $c_{p\ell}(q_i) = \overline{q}_i$ es definida como en el caso creencia.
- $g_{p\ell}$: $g_{p\ell}(q_i^1, q_i^2) = \sum\{m_i^1(V_1)m_i^2(V_2) | V_2 \cap V_1 \neq \emptyset, (V_1, V_2) \in \mathcal{F}_1 \times \mathcal{F}_2\}$;
- f es la media.

Se pueden mostrar las siguientes propiedades:

1. $q_i^1 \cap_{cre} q_i^2 = q_i^1 q_i^2$, puesto que

$$\begin{aligned} q_i^1 \cap_{cre} q_i^2(V) &= \sum_{A \subseteq V} m_i^1 \cap_{cre} m_i^2(A) \\ &= \sum_{V_1 \cup V_2 = A \subseteq V} m_i^1(V_1) m_i^2(V_2) \\ &= \sum_{V_1 \subseteq V} m_i^1(V_1) \sum_{V_2 \subseteq V} m_i^2(V_2) \end{aligned}$$

2. $g_{cre}(q_i^1, q_i^2) = \sum_{V_1 \in \mathcal{F}_1} m_i^1(V_1) q_i^2(V_1)$

3. $g_{p\ell}(q_i^1, q_i^2) = \sum_{V_2 \in \mathcal{F}_2} m_i^2(V_2) p\ell_i^2(V_2) = \sum_{V_1 \in \mathcal{F}_1} m_i^1(V_1) p\ell_i^1(V_1)$
   donde $p\ell_i^j(V_j) = \sum_{V \cap V_j \neq \emptyset} q_i^j(V)$; por tanto $g_{p\ell}$ es simétrica mientras g_{cre} no lo es.

4. $\forall A \in P(\Omega) \quad q_i^1 *_{cre} q_i^2(A) = 1 - q_i^1 *_{p\ell} q_i^2(\overline{A})$.

5. Si dos expertos observan el mismo evento A y son asociados a las funciones creencia q_i^1 y q_i^2 con $\mathcal{F}_i^1 = \mathcal{F}_i^2 = \{A, O\}$; entonces se puede mostrar que $q_i^1 \cup_{cre} q_i^2 = q_i^1 + q_i^2 - q_i^1 q_i^2$.

Ejemplo:

Varios expertos en transporte definen el escenario de un accidente entre un carro y una bicicleta por una función creencia q_1 concerniente a la velocidad del carro. Conociendo q_1 se puede definir el objeto creencia a = [velocidad = q_1] donde el cuerpo de evidencia de q_1

es $\{\mathcal{F}_1, m_1\}$ tal que $\mathcal{F}_1 = \{V_1, O\}$, donde O es el conjunto de posibles velocidades y $V_1 \subseteq O$ es un intervalo de velocidad (por ejemplo $V_1 = [100, 120]$ kilómetros por hora). Ahora supongamos que un testigo observa un accidente y dice que es definido por una función creencia q_2 con cuerpo de evidencia $\{\mathcal{F}_2, m_2\}$ tal que $\mathcal{F}_2 = \{V_2, O\}$. Si deseamos conocer qué tanto un accidente definido por w^S = [velocidad = q_2] satisface el escenario definido por a, debemos calcular $a(\omega)$. Como a es un objeto creencia, por definición tenemos: $a(\omega) = \sum_{V \subseteq \mathcal{F}_1} m_1(V_1)q_2(V) = m_1(V_1)q_2(V_1) + m_1(O)q_2(O) = m_1(V_1)q_2(V_1) + m_1(O)$.
Por tanto si $V_2 \subseteq V_1$ entonces $a(\omega) = m_1(V_1)m_2(V_2) + m_1(O)$ y entre más alta es la creencia del testigo en V_2, más ω satisface el escenario definido por a. Si $V_1 \subseteq V_2$ entonces $a(\omega) = m_1(O)$ y entre más grande es la ignorancia del experto que ha definido el escenerio, más ω lo satisface.

8 Algunas propiedades y cualidades de los objetos simbólicos

8.1 Orden, unión e intersección entre objetos *im*

Es posible definir un preorden parcial $\leq_\alpha$ sobre los objetos *im* estableciendo que:

$$a_1 \leq_\alpha a_2 \text{ si y sólo si } \forall \omega \in \Omega \quad \alpha \leq a_1(\omega) \leq a_2(\omega).$$

Deducimos de este preorden una relación de equivalencia $\mathcal{R}$ por

$$a_1 \mathcal{R} a_2 \text{ si y sólo si } \mathrm{Ext}(a_1/\Omega, \alpha) = \mathrm{Ext}(a_2/\Omega, \alpha)$$

y un orden parcial denotado $\leq_\alpha$ y llamado "orden simbólico" sobre las clases de equivalencia inducidas por $\mathcal{R}$.

Decimos que a_1 *hereda* de a_2 o que a_2 es *más general* que a_1 en el nivel α si y sólo si $a_1 \leq_\alpha a_2$ (lo cual implica que $\mathrm{Ext}_\alpha(a_1/\Omega, \alpha) \subseteq \mathrm{Ext}_\alpha(a_2/\Omega, \alpha)$).

Llamamos *intensión al nivel* α de un subconjunto $\Omega_1 \subseteq \Omega$ al objeto simbólico b definido por la conjunción de eventos cuya extensión al nivel contiene a Ω_1.

La unión simbólica $a_1 \cup_{x,\alpha} a_2$ (resp. intersección $a_1 \cap_{x,\alpha} a_2$) al nivel α es la intensión de $\mathrm{Ext}(a_1/\Omega, \alpha) \cup \mathrm{Ext}(a_2/\Omega, \alpha)$ (resp. $\mathrm{Ext}(a_1/\Omega, \alpha) \cap \mathrm{Ext}(a_2/\Omega, \alpha)$).

8.2 Algunas cualidades de los objetos simbólicos

Como en el caso booleano [1],[8], es posible definir diferentes clases de cualidades de objetos simbólicos (refinamiento, simplicidad, completitud, etc). Por ejemplo decimos que un objeto simbólico s es "completo" si y sólo si las propiedades que caracterizan su extensión son exactamente aquellas cuya conjunción define el objeto. En otras palabras s es un objeto simbólico completo si es la intensión de su extensión. Más intuitivamente, si puedo ver algunos perros blancos y establezco "Puedo ver algunos perros", mi proposición no describe los perros en una forma completa, puesto que no estoy diciendo que ellos son blancos.

Por otro lado, la simplicidad al nivel α de un *im* objeto es el número más pequeño de eventos elementales cuya extensión al nivel α coincide con la extensión de s al mismo nivel.

Cuando un operador $\cup_x$ debe ser definido en un dominio relacionado con una semántica específica la cual induce el concepto de similitud entre objetos simbólicos, parece natural que satisfaga las siguientes propiedades específicas:

a) La unión de dos objetos simbólicos es más general que cada uno de ellos. Es decir, la extensión de la unión de dos objetos simbólicos contiene la extensión de cada uno de ellos.

b) La unión de un objeto simbólico consigo mismo tiene una extensión que contiene la de este objeto.

c) Entre más dos objetos son similares, menos ellos son generales.

d) Los objetos más opuestos (opuestos en todas las variables que los definen) tienen una unión cuya extensión contiene a cada uno.

e) La unión de dos objetos similares debe rechazar, en su extensión, objetos que no les sean similares.

En el caso de intersección, se pueden definir condiciones "naturales" análogas que expresan condiciones inversas. Por ejemplo la intersección de dos objetos simbólicos es menos general que cada uno de ellos.

En el caso de objetos probabilísticos o posibilísticos, es fácil ver que la condición a) es satisfecha, puesto que cuando q_1 y q_2 son dos medidas de probabilidad, tenemos:

$$q_1 \cup_{\mathrm{pr}} q_2 = q_1 + q_2 - q_1 q_2 \geq q_k$$

para $k = 1, 2$. Cuando q_1 y q_2 son medidas de posibilidad tenemos

$$q_1 \cup_p q_2 = \max(q_1, q_2) \geq q_k$$

para $k = 1, 2$.

Si $a_j = \wedge_i[y_1 = q_i^j]$ obtenemos $a_1 \cup_x a_2 = \wedge_i[y_i = q_i^1 \cup_x q_i^2]$ y $\forall \omega \in \Omega$ tal que $\omega^S = \wedge_i[y_i = r_i]$ tenemos:

$$a_1 \cup_x a_2(\omega) = f_x(\{g_x(q_i^1 \cup_x q_i^2, r_i)\}_i),$$

por tanto,

$$a_1 \cup_{\mathrm{pr}} a_2(\omega) = \mathrm{Media}\{\sum_{v \in O_i} q_i^1 \cup_{\mathrm{pr}} q_i^2(v) r_i(v)\}_i \geq \mathrm{Media}\{\sum_{v \in O_i} q_i^k(v) r_i(v)\}_i = a_k(\omega)$$

con $k = 1, 2$.

De la misma forma, en el caso de posibilidades obtenemos:

$$\begin{aligned} a_1 \cup_p a_2(\omega) &= \max_i\{\max_{v\in O_i} \min(q_i^1 \cup_{\mathrm{pr}} q_i^2(v), r_i(v))\} \\ &\geq \max_i\{\max_{v\in O_i} \min(q_i^k(v), r_i(v))\} \\ &= a_k(\omega) \text{ con } k = 1,2 \end{aligned}$$

Es fácil ver, de la misma forma, que la intersección posibilística y probabilística satisfacen la condición inversa.

La condición b) es probada en el caso de objetos probabilísticos, de la siguiente forma, si se trata de una aserción reducida a un evento, y puede ser fácilmente generalizada (tomando la media) al caso de una conjunción de varios eventos:
sea $a = [y = p]$, tenemos por definición $a \cup_{\mathrm{pr}} a = [y = p \cup_{\mathrm{pr}} p] = [y = 2p - p^2]$; por tanto $\forall\omega^S = [y = r]$ obtenemos

$$a \cup_{\mathrm{pr}} a(\omega) = \sum_{v\in O}(2p - p^2)r(v) \geq \sum_{v\in O} p(v)r(v)$$

y así $a \cup_{\mathrm{pr}} a(\omega) \geq a(\omega)$; por tanto $a \cup_{\mathrm{pr}} a \geq a$.

En caso de objetos posibilísticos es fácil ver que $a \cup_p a = a$ pues si q es una medida de posibilidad y $a = [y = q]$ obtenemos $a \cup_{\mathrm{pr}} a = [y = q \cup_p q] = [y = \max(q, q)] = a$.

Las condiciones c) y e) dependen de la escogencia de la similitud. Con la similitud propuesta en §10.1 se puede mostrar que la tercera condición no es satisfecha por objetos probabilísticos. Es fácil mostrar que la condición d) es satisfecha por objetos probabilísticos y posibilísticos. Sea $a_i = [y = p_i]$ con $p_i(v_i) = 1$ y por tanto $p_i(v_j) = 0$ si $v_i \neq v_j$. Resulta que en el caso probabilístico obtenemos $\cup_i p_i = \mathbf{1} \in Q_i^{\mathrm{pr}}$ donde $\mathbf{1}$ es la aplicación tal que $\mathbf{1}(v) = 1$ para todo v. De lo cual resulta que para cualquier $\omega^s = [y = r]$ donde r es una medida de probabilidad, $\bigcup_{\mathrm{pr}_i} a_i(\omega) = 1$. En el caso en que los p_i son posibilidades obtenemos también $\cup_i p_i = \mathbf{1}$ (que es una posibilidad). Resulta también que para cualquier $\omega^s = [y = r]$ donde p es una medida de posibilidad, $\bigcup_{p_i} a_i(\omega) = 1$. Por tanto en ambos casos la unión de los objetos más opuestos son iguales a Ω^S, el objeto pleno cuya extensión contiene todos los elementos de Ω.

8.3 Algunas propiedades de los objetos *im* : retículo y completitud

Se puede mostrar [9] que dado un nivel α, el conjunto de objetos *im* es una retículo para el orden simbólico, y que la unión e intersección simbólica definen el supremo y el ínfimo de cualquier par. Para ello, f_x, g_x y h_x (ver §3.1) deben ser bien escogidas e introducimos un "pleno" y un "vacío" (los cuales también podrían ser llamados "superior" e "inferior") porque son respectivamente el más y menos general de los objetos simbólicos, denotados respectivamente Ω^S y ϕ, y tales que para todo $\omega \in \Omega$, $\Omega^S(\omega) = 1$ y $\phi(\omega) = 0$. Es entonces fácil ver que la extensión de Ω^S contiene todos los elementos de Ω (es decir, es "pleno") y la extensión de ϕ no contiene ninguno (es decir, es "vacío").

Se puede también mostrar que la unión y la intersección simbólica de objetos *im* completos son objetos *im* completos y por tanto que el conjunto de objetos *im* completos es también un retículo.

9 Una extensión de las aserciones posibilísticas, probabilísticas y creencia, sobre objetos simbólicos

9.1 Aserciones duales

Varias clases de valuaciones de objetos simbólicos pueden ser estudiadas, por ejemplo, en el caso de objetos booleanos las obtuvimos definiendo $\forall A \subseteq \mathcal{A}_b, a^*(A) = \frac{|A|}{|\mathcal{A}_b|}$ (en este caso los O_i deben ser finitos y satisfacer los axiomas de Kolmogorov; $a^*(A)$ también puede ser calculada teniendo en cuenta las restricciones que pueden existir entre las variables (ver [5] para más detalles sobre restricciones). Otra posibilidad puede ser considerar la x-unión de subconjuntos de $\mathcal{A}_x$ mediante la siguiente definición donde

$$*_x \in \{\cup_x, \cap_x\} \quad \forall A_x^1, A_x^2 \subseteq \mathcal{A}_x, \quad A_x^1 *_x A_x^2 = \{a_1 *_x a_2 | (a_1, a_2) \in A_x^1 \times A_x^2\}$$

y entonces estudiando la relación entre $a^*(A_x^1 \cup_x A_x^2)$, $a^*(A_x^1 \cap_x A_x^2)$, $a^*(A_x^1)$ y $a^*(A_x^2)$ (donde, por ejemplo, $a^*(A_x) = \sum\{a^*(a_i) | a_i \in A_x\}$).

En este artículo, nuestra idea es extender una aserción *im* $a = \bigwedge_i [y_i = q_i]$ a una aserción *im* dual, denotada a^* y definida sobre subconjuntos de $\mathcal{A}_x$, donde q_i depende de la escogencia de x que puede ser, por ejemplo, una función posibilística, probabilística o creencia y, $\mathcal{A}_x$ es el conjunto de aserciones asociadas a x. Más generalmente, a^* puede ser definida sobre "$*_x$-combinaciones" de tales subconjuntos de la clase $A *_x B$ donde $*_x \in \{\cup_x, \cap_x\}$ y mostrar que a^* es en sí misma un tipo de función posibilística, probabilística o creencia dependiente de x.

Más precisamente: dado $A_x \subseteq \mathcal{A}_x$, tenemos $A_x = \{a | a \in A_x\}$ y para definir $A = \cup_x \{a | a \in A_x\}$ usamos el conjunto $Q_i^{A_x} \subseteq Q_i^x$ tal que $Q_i^{A_x} = \{q_i | a = \bigwedge_j [y_j = q_j] \in A_x\}$.

Denotamos: $q_i^A = \cup_x \{q_i | q_i \in Q_i^{A_x}\}$.

Definimos $\cup_x$ para aserciones *im* por

$$\cup_x \{a | a \in A_x\} = \bigwedge_i [y_i = q_i^A]$$

por tanto tenemos $A = \bigwedge_i [y_i = q_i^A]$.

Definimos a_ℓ^* como la medida "dual" de $a_\ell = \bigwedge_i [y_i = q_i^\ell]$ por

$$a_\ell^*(a_j) = f_x(\{g_x(q_i^\ell, q_i^j)\}_i)$$

por tanto, dado $A_x^k \subseteq \mathcal{A}_x$, denotamos $A_k = \cup_x\{a|a \in A_x^k\}$ y obtenemos

$$a_\ell^*(A_k) = f_x(\{g_x(q_i^\ell, q_i^{A_k})\}_i).$$

Más generalmente

$$a_\ell^*(A_1 *_x A_2) = f_x(\{g_x(q_i^\ell, q_i^{A_1} *_x q_i^{A_2})\}_i)$$

donde $*_x \in \{\cup_x, \cap_x\}$ y $q_i^{A_k} = *_x\{q_i|q_i \in Q_i^{A_x^k}\}$.

9.2 Tres teoremas de metaconocimiento

Los tres resultados siguientes [9], prueban la existencia de los objetos probabilísticos, posibilísticos y creencia, definidos respectivamente sobre objetos probabilísticos, posibilísticos y creencia, ellos mismos definidos sobre Ω.

Teorema 1 (caso de objetos posibilísticos)

1. $a^*(\mathcal{A}_p) = 1$ y $a^*(\emptyset) = 0$.

2. $\forall A_1, A_2 \subseteq \mathcal{A}_p \quad a^*(A_1 \cup_p A_2) = \max(a^*(A_1), a^*(A_2))$.

Teorema 2 (caso de objetos probabilísticos)

1. $a^*(\mathcal{A}_{pr}) = 1$ y $a^*(\emptyset) = 0$.

2. $\forall A_1, A_2 \subseteq \mathcal{A}_{pr} \quad a^*(A_1 \cup_{pr} A_2) = a^*(A_1) + a^*(A_2) - a^*(A_1 \cap_{pr} A_2)$.

En el caso de objetos creencia, decimos que hay *independencia* entre el cuerpo de evidencia de dos objetos creencia a_1 y a_2 si y sólo si para todo i los cuerpos de evidencia $(\mathcal{F}_i^j, m_i^j)$ asociados a q_i^j para $j = 1, 2$ son tales que $m_i^1 \cap_{\text{cre}} m_i^2(\emptyset) = 0$. En otras palabras, los elementos focales $V_i^1 \in \mathcal{F}_i^1$ y $V_i^2 \in \mathcal{F}_i^2$ son tales que $V_i^1 \cap V_i^2 \neq \emptyset$.

Los cuerpos de evidencia de dos subconjuntos A_1 y A_2 de $\mathcal{A}_{\text{cre}}$ se dicen independientes si y sólo si para $j = 1, 2$ y para todo i tales que $Q_i^j = \cup_{\text{cre}}\{q_i^j|q_i^j \in Q_i^{A_j}\}$, los cuerpos de evidencia de Q_i^1 y Q_i^2 son independientes.

Teorema 3 (caso de objetos creencia)

1. $a^*(\mathcal{A}_{\text{cre}}) = 1$ y $a^*(\emptyset) = 0$.

2. Si $\forall i \;\; A_i \subseteq \mathcal{A}_{\text{cre}}$ los cuerpos de evidencia de los A_i son independientes, entonces

$$a^*\Big(\bigcup_{i=1}^{n}{}_{\text{cre}} A_i\Big) \geq \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{|I|+1} a^*\Big(\bigcap_{i \in I}{}_{\text{cre}} A_i\Big)$$

3. $\forall A \subseteq \mathcal{A}_{\text{cre}}$ se tiene:

$$m^*(A) = \frac{a^*_{\text{cre}}(A)}{a^*_{\text{cre}}(h(A))} \sum_{B \subseteq A} (-1)^{|A-B|} a^*_{\text{cre}}(h(B))$$

donde $h(B) = \cap_{\text{cre}}\{A_i | A_i = A - \{a_i\}, a_i \in A - B, B \neq A\}$
$h(A) = \cup_{\text{cre}}\{A_i | A_i = A - \{a_i\}, a_i \in A\}$
entonces m^* es una función de asignación de probabilidad sobre $\mathcal{A}_{\text{cre}}$.
En otras palabras, $m^* : P(\mathcal{A}_{\text{cre}}) \longrightarrow [0,1]$ es tal que $m^*(\emptyset) = 0$, $\sum_{A \subseteq \mathcal{A}_{\text{cre}}} m^*(A) = 1$ y

$\forall A \subseteq \mathcal{A}_{\text{cre}} \quad a^*(A) = \sum_{B \subseteq A} m^*(B)$.

Usando m^* es entonces posible extender la regla de Dempster y el condicionamiento de Dempster sobre el conjunto de aserciones creencia.

9.3 Semántica de a^* en el caso de objetos probabilísticos

En caso de probabilidades $a_1^*(a_2)$ representa intuitivamente la probabilidad promedio de que la misma instancia ocurra en ambas entidades descritas por a_1 y a_2. Será una probabilidad alta si y sólo si $g(q_i^1, q_i^2) = \sum_v q_i^1(v) q_i^2(v)$ es grande para todo i. Dicho de otro modo, entre más grandes o más pequenãs sean $q_i^1(v)$ y $q_i^2(v)$ a la vez, y sus valores grandes se concentren en pocos elementos $v \in O_i$, más grande será $g(q_i^1, q_i^2)$. Si $q_i^1(v)$ es grande cuando $q_i^2(v)$ es pequeño para cualquier i, entonces $g(q_i^1, q_i^2)$ será pequeño.

Nótese también que si consideramos que $a^*(A_1 \cap_{pr} A_2)$ es una medida de especialización probabilística y $a^*(A_1 \cup_{pr} A_2)$ una medida de generalización probabilística entre A_1 y A_2, entonces el teorema 2 muestra que, cuando $a^*(A_1) + a^*(A_2)$ es constante, entre más A_1 y A_2 son especializados (por ejemplo $a^*(A_1 \cap_{pr} A_2)$ grande), menos son generales (por ejemplo $a^*(A_1 \cup_{pr} A_2)$ pequeño).

9.4 Semántica de a^* en el caso de objetos posibilísticos

Si a_1 y a_2 son objetos posibilísticos, $a_1^*(a_2)$ representa intuitivamente la "posibilidad" que el algún objeto individual "posible" para a_2 sea "posible" para a_1. En el caso extremo donde a_1 y a_2 sean además aserciones booleanas $a_1^*(a_2)$ mide la posibilidad que un objerto individual satisfaga simultáneamente a_1 y a_2. Más precisamente, si a_j es un objeto posibilístico

booleano, se puede escribir como $a_j = \wedge_i[y_i = q_i^j]$, donde q_i^j es una aplicación característica tal que $q_i^j(v) = 1$ si $v \in V_i^j$. Así a_j puede ser escrito como un objeto simbólico booleano: $a_j = \wedge_i[y_i = V_i^j]$, de lo cual resulta que (ver §5.3)

$$a_1^*(a_2) = \max_i \left(\sup_{v \in O_i} \{\min (q_i^1(v), q_i^2(v))\} \right) = 1$$

si y sólo si $\forall i \quad V_i^1 \cap V_i^2 \neq \emptyset$, lo cual expresa el hecho que es **posible** que un valor tomado en V_i^2 sea tomado en V_i^1.

Si a_1 es un objeto necesitista booleano, obtenemos en el caso booleano

$$a_1^*(a_2) = \min_i \left(\inf_{v \in O_i} \{\max (q_i(v), \bar{r}_i(v))\} \right) = 1$$

si y sólo si $\forall i \quad V_i^2 \subseteq V_i^1$, lo cual expresa el hecho que un valor tomado en V_i^2 es **necesariamente** tomado en V_i^1.

Note, además, que es necesario y suficiente que al menos para un $v \in O_i$, $q_i^1(v)$ y $q_i^1(v)$ sean ambos grandes para obtener un valor grande de

$$g_{\text{pos}}(q_i^1, q_i^2) = \sup_v \inf \{(q_i^1(v), q_i^2(v))\}.$$

Ejemplo:
debemos clasificar varios documentos que son caracterizados por la frecuencia de algunas palabras dadas.
Objetos probabilísticos: usando la frecuencia asociamos a cada documento d_i una medida de probabilidad q_i y una aserción probabilística a_i. Es entonces fácil ver que $a_i^*(a_j)$ es la probabilidad de que la misma palabra ocurra en ambos documentos d_i y d_j, ésta será alta si en los documentos d_i y d_j las frecuencias están concentradas sobre pocas palabras y son grandes para las mismas palabras.
Objetos posibilísticos: algunas palabras pueden aparecer pero con contrasentido, y algunas otras, importantes para algunos documentos, pueden no aparecer. Así, tomando en cuenta el contexto, un experto asocia a cada palabra una medida de posibilidad; por tanto, cada documento d_i puede ser representado por una aserción posibilística a_i y $a_i^*(a_j)$ será grande si al menos para una palabra, las posibilidades son simultáneamente grandes para ambos documentos d_i y d_j.

9.5 Semántica de a^* en el caso de objetos creencia

El significado de $a_1^*(a_2)$ puede ser interpretado como una "creencia de creencia" o la "convicción" de algo, denotado E_1, cuya creencia es representada por a_1, de la creencia de alguna otra cosa denotada E_2, cuya creencia es representada por a_2.

Ejemplo:
para $i = 1, 2$, sea $a_i = [y = q_i]$ donde q_i es una función creencia de $\mathcal{O}$ a $[0, 1]$, con cuerpo de

evidencia $(\mathcal{F}_i, m_i)$ y $\mathcal{F}_1 = \mathcal{F}_2 = \{A, B, \mathcal{O}\}$ con $A \cap B = \emptyset$; entonces obtenemos

$$a_1^*(a_2) = g_{\text{cre}}(q_1, q_2) = \sum_{V \in \mathcal{F}_1} m_1(V) q_2(V) = m_1(A)m_2(A) + m_1(B)m_2(B) + m_1(\mathcal{O}) \quad (1)$$

Siguiendo un ejemplo clásico de Schafer [16], supongamos que: Betty es una experto E_2 y yo soy un experto E_1, A = "una rama de árbol cayó sobre mi carro", B = "ninguna rama de árbol cayó sobre mi carro". Supongamos que Betty me dice que una rama de árbol cayó sobre mi carro, esto es $m_2(A) = 1$ y $m_2(B) = 0$. Sabiendo que mi probabilidad subjetiva de que se puede creer en Betty es $p = 0.9$, digo que su testimonio justifica un 0.9 del grado de creencia de que una rama de árbol cayó sobre mi carro, por tanto $m_1(A) = 0.9$, $m_1(B) = 0$ y $m_1(\mathcal{O}) = 0.1$. Resulta de (1) que mi creencia sobre su creencia es $a_1^*(a_2) = 1$. Esto es justificado puesto que mi creencia no me da razon para rechazar la creencia de Betty dado que $m(B) = 0$. Si tengo alguna razón para confiar en B, entonces $m_1(B) \neq 0$ y mi creencia sobre su creencia $a_1^*(a_2) = m_1(A) + m_1(O)$ se vuelve más pequeña que 1, dado que $m_1(A) + m_1(B) + m_1(\mathcal{O}) = 1$.

Notemos que "mi probabilidad subjetiva de que Betty dice la verdad" es igual a mi creencia sobre su creencia, es decir $a_1^*(a_2) = 0.9$, en los dos siguientes casos:

1. $m_1(A) = 0.9$, $m_1(B) = 0.1$ y $m_2(A) = 1$,

2. $m_1(A) = 1$ y $m_2(A) = 0.9$,

lo cual corresponde a la intuición.

Más generalmente, podemos ver que la convicción de E_1 concerniente a la creencia de E_2 será máxima, es decir $a_1^*(a_2) = 1$, si:

- E_1 es totalmente ignorante de las creencias de A y B, puesto que en tal caso $m_1(A) = m_1(B) = 0$ y $m_1(\mathcal{O}) = 1$,
- E_1 y E_2 confían totalmente en la misma creencia puesto que $m_1(A) = m_2(A) = 1$ o $m_1(B) = m_2(B) = 1$.

Si $m_1(B) = 1$ y E_1 tiene alguna ignorancia de A (es decir $m_1(\mathcal{O}) \in]0, 1[$), entonces su convicción de la creencia de E_2 sobre A (es decir $q_2(A)$) será más grande que $q_2(A)$; por ejemplo si $m_1(A) = m_2(A) = 1/2$ entonces $m_1(\mathcal{O}) = 1/2$ y la convicción de E_1 será $a_1^*(a_2) = 0.75$. Si E_1 confía totalmente en A ($m_1(A) = 1$, $m_1(B) = m_1(\mathcal{O}) = 0$) y E_2 confía totalmente en B ($m_2(B) = 1$ y $m_1(A) = 0$) entonces la convicción de E_1 de la creencia de E_2 será 0. Si E_2 es completamente ignorante, es decir $m_2(A) = m_2(B) = 0$, entonces la convicción de E_1 de la creencia de E_2 será baja si su creencia es fuerte (es decir, su ignorancia medida por $m_1(\mathcal{O})$ es baja).

Ejemplo

Varios sensores, en posiciones diferentes, tienen una creencia de un evento A. Este conocimiento induce una creencia de cada sensor de la creencia de los otros sensores cuando ellos

están en la misma situación. En la figura 6 presentamos 4 situaciones que permiten a 4 sensores obtener una creencia de la creencia del sensor número 5.

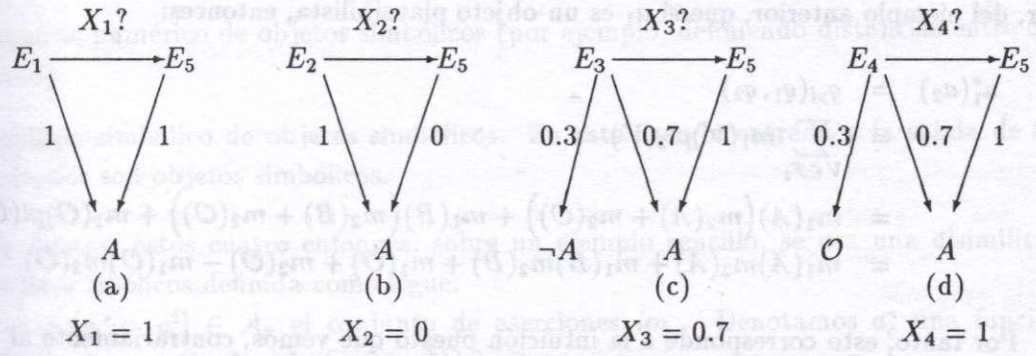


Figura 6: $X_i = a_i^*(a_5)$ es la creencia de E_i de la creencia de E_5, es calculada según la ecuación (1)

En esta figura, si denotamos $a_i = [y_i = q_i]$ la aserción creencia asociada al sensor i, y $\mathcal{F}_i$ el elemento focal de la función creencia q_i, tenemos:

a) $\mathcal{F}_1 = \mathcal{F}_5 = \{A\}$,
por lo tanto $m_1(A) = m_5(A) = 1$. Resulta de (1) que $X_1 = a_1^*(a_5) = 1$.

b) $\mathcal{F}_2 = \{A\}$, $\mathcal{F}_5$ no contiene a A,
por lo tanto $a_2^*(a_5) = 0$.

c) $\mathcal{F}_3 = \{A, \neg A\}$, $\mathcal{F}_5 = \{A\}$, $m_3(A) = 0.7$ y $m_3(\neg A) = 0.3$.
Por lo tanto $a_3^*(a_5) = m_3(A)m_5(A) + m_3(\neg A)m_5(\neg A) + m_3(A)m_5(A) + m_3(\mathcal{O}) = 0.7$.

d) $\mathcal{F}_4 = \{A, \mathcal{O}\}$, $\mathcal{F}_5 = \{A\}$, $m_4(A) = 0.7$, $m_4(\mathcal{O}) = 0.3$ y $m_5(A) = 1$,
por lo tanto $a_4^*(a_5) = m_4(A)m_5(A) + m_4(\mathcal{O}) = 1$.

Si una gran mayoría de sensores (por ejemplo, al menos el 75%) tienen una creencia sobre un sensor dado, menor o igual que un umbral α, este sensor puede ser rechazado para el reconocimiento de A. En este ejemplo, si $\alpha = 1/2$, el sensor número 5 no es rechazado y si $\alpha = 0.8$, entonces es rechazado. Note que si un sensor es completamente ignorante ($m_i(\mathcal{O}) = 1$ y por tanto $m_i(A) = 0$ para todo A) confiará en cualquier sensor sin importar la creencia de este sensor. Por tanto podemos rechazar el juicio de sensores que son muy ignorantes.

En lugar de usar una regla mayoritaria, es posible usar la regla de Dempster (al segundo nivel) aplicada a la creencia de creencia, de un conjunto de sensores, de un sensor dado. De esa forma el sensor representado por a_5 es rechazado si $\oplus_{i=1}^4 a_i(a_5) < \alpha$. La creencia en A, si ningún sensor es rechazado, es medida por la regla clásica de Dempster (al nivel 1): $\oplus_{i=1}^4 a_i(A)$.

Hay un teorema análogo si a_1 es una aserción plausibilista y $a_1^*(a_2)$ puede ser interpretado como la "no-discordancia" mutua entre lo que E_1 y E_2 confían. Para ilustrar eso, podemos ver, del ejemplo anterior, que si a_1 es un objeto plausibilista, entonces:

$$\begin{aligned} a_1^*(a_2) &= g_{pl}(q_1, q_2) \\ &= \sum_{V \in \mathcal{F}_1} m_1(V) pl_2(V) \\ &= m_1(A)\Big(m_2(A) + m_2(\mathcal{O})\Big) + m_1(B)\Big(m_2(B) + m_2(\mathcal{O})\Big) + m_1(\mathcal{O}) pl(\mathcal{O}) \\ &= m_1(A)m_2(A) + m_1(B)m_2(B) + m_1(\mathcal{O}) + m_2(\mathcal{O}) - m_1(\mathcal{O})m_2(\mathcal{O}). \end{aligned}$$

Por tanto, esto corresponde a la intuición puesto que vemos, contrariamente al caso de convicción, que la no discordancia entre lo que E_1 y E_2 confían se conserva alto cuando E_2 es totalmente ignorante. Es decir, $m_2(A) = m_2(B) = 0$ aun si la creencia de E_1 es fuerte, esto es $m_1(\mathcal{O}) = 0$.

Otro tipo de interpretación de $a_1^*(a_2)$ puede ser obtenida en términos de "ajuste". Si consideramos la clase C_i (de frutas producidas en una región, por ejemplo) descrita por el objeto creencia a_i, podemos decir, cuando a_1 es un objeto creencia, que $a_1^*(a_2)$ mide cuánto C_2 "ajusta" a C_1; cuando a_1 es un objeto plausibilista, podemos decir que $a_1^*(a_2)$ mide el "no-desacuerdo" entre C_1 y C_2.

Por ejemplo, si y expresa el color y si las frutas de ambas regiones tienen el mismo color denotado A (es decir, $m_1(A) = m_2(A) = 1$, $m_1(B) = m_2(B) = 0$ y $m_1(\mathcal{O}) = m_2(\mathcal{O}) = 0$) entonces $a_1^*(a_2) = 1$ mide qué tanto C_2 "ajusta" a C_1 y también el "no-desacuerdo", para el color, entre C_1 y C_2. Si el color de las frutas de la segunda región es totalmente ignorado, es decir $m_2(A) = m_2(B) = 0$ y $m_2(\mathcal{O}) = 1$, y el color de las frutas de la primera región es A, es decir $m_1(A) = 1$ y $m_1(\mathcal{O}) = 0$, entonces, cuando a_1 es un objeto creencia, obtenemos $a_1^*(a_2) = 0$ lo cual mide qué tanto C_2 ajusta C_1; cuando a_1 es un objeto plausibilista, obtenemos $a_1^*(a_2) = 1$ lo cual mide el no-desacuerdo entre C_1 y C_2.

10 Análisis de datos de objetos simbólicos

10.1 Los cuatro enfoques

Varios estudios han sido realizados recientemente sobre este tópico: para histogramas de objetos simbólicos ver [5]; para generadores de reglas por grafos de decisión sobre objetos *im* en el caso de objetos posibilísticos con "tipicalidades" como nodos, ver [13]; para generadores de clases con intersección por pirámides sobre objetos simbólicos, ver [1].

Más generalmente, cuatro clases de análisis de datos pueden, aproximativamente, ser definidos dependiendo de la entrada y la salida:

a) Análisis numérico de tablas clásicas de datos.

b) Análisis simbólico de tablas clásicas de datos (por ejemplo, obtener un análisis factorial o una clasificación automáticamente interpretadas mediante objetos simbólicos).

c) Análisis numérico de objetos simbólicos (por ejemplo, definiendo distancias entre objetos).

d) Análisis simbólico de objetos simbólicos. En este caso la entrada y la salida de los métodos son objetos simbólicos.

Para ilustrar estos cuatro enfoques, sobre un ejemplo sencillo, se usa una disimilitud entre objetos sibólicos definida como sigue:
sea $a_\ell = \wedge_i[y_i = q_i^\ell] \in \mathcal{A}_x$ el conjunto de aserciones *im* . Denotamos a_ℓ^* una función $\mathcal{A}_x \to [0,1]$ tal que $a_\ell^*(a_k) = f_x(\{g_x(q_i^\ell, q_i^k)\}_i)$. Entonces ponemos:

$$s(a_\ell, a_k) = \frac{1}{2}\left(\frac{a_\ell^*(a_k) + a_k^*(a_\ell)}{\sqrt{a_\ell^*(a_\ell)a_k^*(a_k)}}\right) \tag{2}$$

en el caso en que g_x sea simétrica (lo cual sucede cuando tenemos aserciones probabilísticas, posibilísticas o plausiblistas), s puede escribirse:

$$s(a_\ell, a_k) = \frac{a_\ell^*(a_k)}{\sqrt{a_\ell^*(a_\ell)a_k^*(a_k)}} = \frac{a_k^*(a_\ell)}{\sqrt{a_\ell^*(a_\ell)a_k^*(a_k)}}$$

Ejemplos:
Sean a_1, a_2 dos objetos probabilísticos tales que:

$$a_1 = [y = 0.7v_1, 0.3v_2] \qquad a_2 = [y = 0.3v_1, 0.7v_2]$$

Obtenemos:

$$s(a_1, a_2) = \frac{a_2^*(a_1)}{\sqrt{a_1^*(a_1)a_2^*(a_2)}} = \frac{0.7 \times 0.3 + 0.3 \times 0.7}{\sqrt{(0.7^2 + 0.3^2)(0.3^2 + 0.7^2)}} = 0.724$$

A partir de este ejemplo, se ve que los objetos probabilísticos no satisfacen la condición c) dada en §8.2 cuando se define $a = [y = 1v_1, 0v_2] = [y = v_1]$ obtenemos $a_1 \cup_{\text{pr}} a_2\ (a) = 0.79$ y $a_1 \cup_{\text{pr}} a_1\ (a) = 0.91$; por lo tanto $a_1 \cup_{\text{pr}} a_2$ no deberían ser considerados más generales que $a_1 \cup_{\text{pr}} a_1$, aún si el par (a_1, a_2) puede ser considerado tan similar como el par (a_1, a_1), pues $s(a_1, a_1) = 1$ y $s(a_1, a_2) = 0.724$.

Sean a_1, a_2 dos objetos posibilísticos tales que $a_1 = [y = 1v_1, xv_2]$ y $a_2 = [y = xv_1, 1v_2]$, tenemos:

$$s(a_1, a_2) = \frac{\max\Big(\min(1, x), \min(x, 1)\Big)}{\sqrt{\max\Big(\min(1, 1), \min(x, x)\Big)}} = x$$

Por lo tanto cuanto más pequeño sea x lo más disímiles serán a_1 y a_2. Por lo tanto, $a_1 \cup_p a_2 = [y = 1v_1, 1v_2]$ es el objeto entero pues $\forall a$, $a_1 \cup_p a_2^*(a) = 1$ y por lo tanto, contrariamente al caso probabilístico, en este ejemplo el caso posibilístico satisface la condición c) §8.2.

Ilustramos estos cuatro puntos de vista aplicando tres métodos de análisis de datos: componentes principales, clasificación jerárquica y clasificación piramidal.

Sea T la siguiente tabla de datos donde el conjunto de objetos individuos es $\Omega = \{\omega_1, \ldots, \omega_5\}$ que son cinco compañías descritas por dos variables: y_1 es la tasa de empleo y y_2 la ganancia.

	ω_1	ω_2	ω_3	ω_4	ω_5
y_1	$-1/2$	$1/2$	2	1	2
y_2	$-1/2$	$1/2$	1	2	2

Tabla T

Esta tabla está representada en la figura 7.

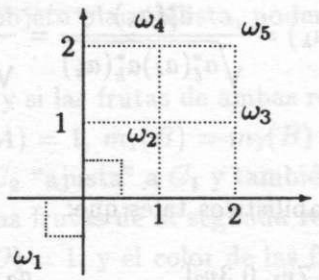


Figura 7: Representación gráfica de la tabla T

10.2 Análisis numérico de una tabla de datos clásica

• Análisis en componentes principales de la tabla T

De la matriz de covarianzas

$$V = \begin{pmatrix} 0.9 & 0.7 \\ 0.7 & 0.9 \end{pmatrix}$$

deducimos los valores propios $\lambda_1 = 1.6$ y $\lambda_2 = 0.2$ y los vectores propios $u_1^t = \frac{1}{\sqrt{2}}(1\ 1)$, $u_2^t = \frac{1}{\sqrt{2}}(1\ -1)$. Finalmente obtenemos la representación de las componentes principales dada en la figura 8, donde la proyección de ω_j sobre el eje i es dada por $F_i(\omega_j) = u_i^t x_j$, donde $x_j^t = (y_1(\omega_1)) - \bar{y}_1, y_2(\omega_2) - \bar{y}_2$ y $\bar{y}_i = 1$ es la media de y_i; por ejemplo, $F_1(\omega_1) = \frac{1}{\sqrt{2}}(1\ 1)\binom{-3/2}{-3/2}$.

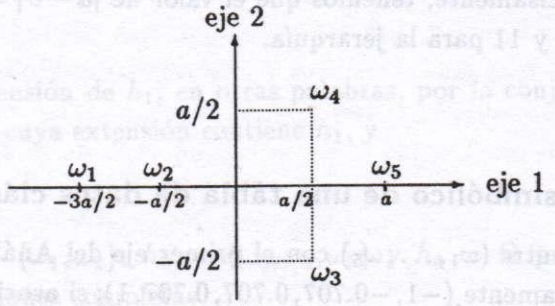


Figura 8: Análisis en componentes principales de la tabla T con $a = \sqrt{2}$

• **Clasificaciones jerárquica y piramidal de la tabla T**

Hacemos la "jerarquía del ligamen completo" basada en la distancia del "city–block" definida por:

$$d(\omega_\ell, \omega_k) = \sum_{j=1}^{2} |y_j(\omega_\ell) - y_j(\omega_k)|$$

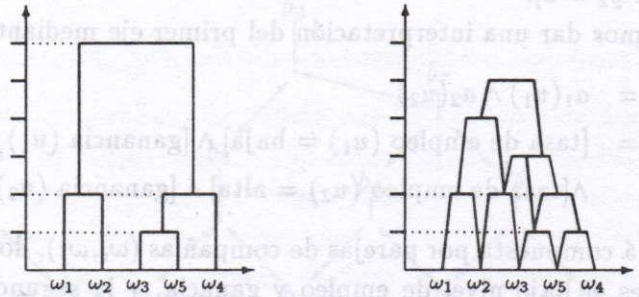
El algoritmo es el siguiente: empezando con 5 clases $C_i = \{\omega_i\}$, donde en cada paso se unen los $\omega_i \in \Omega$ de las dos clases que minimizan $\delta(C_i, C_j)$:

$$\delta(C_i, C_j) = \max\{d(\omega_i, \omega_j)/\omega_i \in C_i, \omega_j \in C_j\}$$

Cuando dos clases se unen, sus elementos son suprimidos del conjunto a clasificar y el proceso continúa hasta que una sóla clase quede.

Para obtener una pirámide, podemos usar un algoritmo similar en el que las clases pueden ser agrupadas dos veces (en lugar de sólo una vez, como en el caso de las jerarquías) si respetan un orden común (para más detalles, puede consultarse [1]).

Usando estos algoritmos obtenemos la jerarquía y la pirámide dadas en la figura 9.



(a) La jerarquía de la tabla T (b) La pirámide de la tabla T

Figura 9

Observación: si asociamos una disimilitud σ inducida por una jerarquía y una pirámide, al poner: $\sigma(\omega_i, \omega_j) = \{$altura del menor nivel que contiene ω_i y $\omega_j\}$, entonces, es fácil ver

que σ es más cercana a la distancia inicial d en el caso de la pirámide que en el caso de la jerarquía; más precisamente, tenemos que el valor de $|d - \sigma| = \sum |d(\omega_i, \omega_j) - \sigma(\omega_i, \omega_j)|$ es 3 para la pirámide y 11 para la jerarquía.

10.3 Análisis simbólico de una tabla de datos clásica

Las correlaciones entre $(\omega_1, \ldots, \omega_5)$ con el primer eje del Análisis en Componentes Principales son respectivamente $(-1, -0.707, 0.707, 0.707, 1)$; si asociamos a cada lado del primer eje los objetos cuya correlación es mayor que 0.707 o menor que -0.707, obtenemos dos clases de objetos: la primera clase, $C_1 = \{\omega_1, \omega_2\}$, explica el lado izquierdo del eje, y la segunda $C_2 = \{\omega_3, \omega_4, \omega_5\}$ explica el lado derecho. Usando estas clases, obtenemos dos tipos de interpretaciones simbólicas del primer eje, pues podemos decir que el lado izquierdo es explicado por: $a_1 = [y_1 = -1/2, 1/2] \wedge [y_2 = -1/2, 1/2]$ y el lado derecho es explicado por $a_2 = [y_1 = 1, 2] \wedge [y_2 = 1, 2]$. Si en la entrada se da una taxonomía que diga que la tasa de empleo y la ganancia son bajas cuando son menores que $1/2$ y altas cuando son mayores que 1, podemos usar las aserciones a_1 y a_2 para obtener la siguiente explicación del primer eje: es explicado por dos aserciones opuestas que caracterizan dos clases de compañías:

$$a_1 = [\text{tasa de empleo} = \text{baja}] \wedge [\text{ganancia} = \text{baja}]$$

$$a_1 = [\text{tasa de empleo} = \text{alta}] \wedge [\text{ganancia} = \text{alta}]$$

Por supuesto, en los ejemplos reales las cosas son mucho más complicadas; por ejemplo, para tener más precisión cuando las dos clases contienen numerosos objetos, cada lado del eje puede ser explicado por una disyunción de aserciones obtenidas por una interpretación simbólica de una clasificación hecha en cada clase. Podemos también enriquecer la interpretación añadiendo ciertas propiedades; por ejemplo, podemos añadir a a_1 las siguientes reglas: [si $y_1 = -1/2$ entonces $y_2 = -1/2$] $\wedge$ [si $y_1 = 1/2$ entonces $y_2 = 1/2$] y a a_2 la regla [si $y_1 = 1$ entonces $y_2 = 2$].

También podemos dar una interpretación del primer eje mediante un objeto horda h:

$$\begin{aligned} h &= a_1(u_1) \wedge a_2(u_2) \\ &= [\text{tasa de empleo } (u_1) = \text{baja}] \wedge [\text{ganancia } (u_1) = \text{baja}] \\ &\quad \wedge [\text{tasa de empleo } (u_2) = \text{alta}] \wedge [\text{ganancia } (u_2) = \text{alta}] \end{aligned}$$

cuya extensión está compuesta por parejas de compañías (ω_i, ω_j), donde el primer elemento de la pareja, ω_i, es el bajo nivel de empleo y ganacia, y la segunda, ω_j, de alta tasa de empleo y ganancia. Si una variable externa indica la edad de las compañías, el objeto horda h se convertiría en: $h = a_1(u_1) \wedge a_2(u_2) \wedge [\text{edad } (u_1) < \text{ edad } (u_2)]$.

Un análisis simbólico de una tabla de datos clásica también puede ser una interpretación automática de una clasificación mediante objetos simbólicos: por ejemplo, es posible asociar

a cada nivel de la jerarquía un objeto simbólico completo (ver § 8.2); más preciosamente, si denotamos $h_1 = \{\omega_3, \omega_5\}$ entonces, podemos asociar a h_1, la aserción $a_1 = [y_1 = 2] \wedge [y_2 = 1,2]$; a_1 es completo pues:

- es definido por la intensión de h_1, en otras palabras, por la conjunción de todos los eventos $e_i = [y_i = V_i]$ cuya extensión contiene h_1, y
- su extensión es h_1.

De la misma forma, $h_2 = \{\omega_1, \omega_2\}$, $h_3 = \{\omega_3, \omega_4, \omega_5\}$ y $h_4 = \Omega$ pueden ser asociados respectivamente a las aserciones completas

$$a_2 = [y_1 = -1/2, 1/2] \wedge [y_2 = -1/2, 1/2]$$

$$a_3 = [y_1 = 1,2] \wedge [y_2 = 1,2]$$

$$a_4 = [y_1 = O_1] \wedge [y_2 = O_2]$$

donde O_1 y O_2 son los conjuntos de todos los valores que toman y_1 y y_2 en la tabla T. Usando el hecho de que cada nivel es representado por una aserción completa deducimos de cualquier nivel $h_\ell = h_i \cup h_k$ la regla $a_\ell \rightarrow a_i \vee a_k$. Por lo tanto, de la jerarquía obtenemos las siguientes reglas:
$R_1 : a_4 \rightarrow a_2 \vee a_3$ y $R_2 : a_3 \rightarrow a_1 \vee \omega_4^s$ donde $\omega_4^s = [y_1 = 1] \wedge [y_2 = 2]$ es el objeto simbólico asociado a ω_4. Todas las reglas de la parte superior tales como $a_1 \rightarrow a_3$ son válidas pues a_i y b_i son objetos completos. Finalmente hemos inducido de la jerarquía dada en a) un grafo (ver figura 10 (a)) cuyos nodos son aserciones y las reglas son expresadas entre ellas por las direcciones. En la figura 10 (c), (c_1) expresa la regla $r_1 : x \rightarrow y \vee z$; (c_2) expresa la regla $r_2(y \rightarrow x) \wedge (z \rightarrow x)$ y (c_3) expresa la regla $r_1 \wedge r_2$.

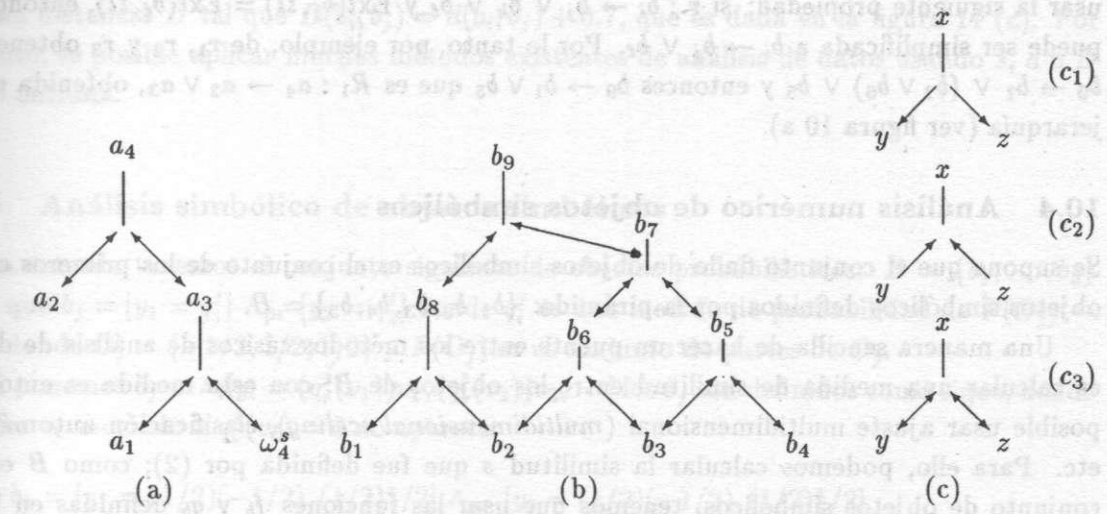


Figura 10: Grafo inducido de reglas entre aserciones: (a) a partir de la jerarquía, (b) a partir de la pirámide, donde los arcos con doble cabeza son explicados por (c)

El mismo tipo de interpretación simbólica puede hacerse empezando a partir de la pirámide dada en la figura 9; por lo tanto, tenemos el grafo dado en la figura 10 (b); de esta forma, obtenemos más aserciones y más reglas entre ellas.
Si denotamos $h_1 = \{\omega_1, \omega_2\}$, $h_2 = \{\omega_2, \omega_3\}$, $h_3 = \{\omega_3, \omega_5\}$, $h_4 = \{\omega_4, \omega_5\}$, $h_5 = \{\omega_3, \omega_4, \omega_5\}$, $h_6 = \{\omega_2, \omega_3, \omega_5\}$, $h_7 = \Omega - \{\omega_1\}$, $h_8 = \{\omega_1, \omega_2, \omega_3\}$, $h_9 = \Omega$, las aserciones completas asociadas son:

$$b_1 = [y_1 = -1/2, 1/2] \wedge [y_2 = -1/2, 1/2]$$

$$b_2 = [y_1 = 1/2, 2] \wedge [y_2 = 1/2, 1]$$

$$b_3 = [y_1 = 2] \wedge [y_2 = 1, 2]$$

$$b_4 = [y_1 = 1, 2] \wedge [y_2 = 2]$$

$$b_5 = [y_1 = 1, 2] \wedge [y_2 = 1, 2]$$

$$b_6 = [y_1 = 1/2, 2] \wedge [y_2 = 1/2, 1, 2]$$

$$b_7 = [y_1 = 1/2, 1, 2] \wedge [y_2 = 1/2, 1, 2]$$

$$b_8 = [y_1 = -1/2, 1/2, 2] \wedge [y_2 = -1/2, 1/2, 1]$$

$$b_9 = [y_1 = O_1] \wedge [y_2 = O_2]$$

Podemos entonces inducir las reglas siguientes:

$r_1 : b_9 \rightarrow b_8 \vee b_7$ $\quad r_2 : b_7 \rightarrow b_6 \vee b_5$ $\quad r_3 : b_8 \rightarrow b_1 \vee b_2$

$r_4 : b_6 \rightarrow b_2 \vee b_3$ $\quad r_5 : b_5 \rightarrow b_3 \vee b_4$

Tenemos $b_1 = a_2$, $b_3 = a_1$, $b_5 = a_3$ y $b_9 = a_4$; por lo tanto, es posible deducir a partir de las reglas r_i dadas por la pirámide, las reglas dadas por la jerarquía; para ello, necesitamos usar la siguiente propiedad: si $r : b_i \rightarrow b_j \vee b_k \vee b_\ell$ y $\mathrm{Ext}(b_j, \Omega) = \mathrm{Ext}(b_\ell, \Omega)$, entonces r puede ser simplificada a $b_i \rightarrow b_j \vee b_\ell$. Por lo tanto, por ejemplo, de r_1, r_2 y r_3 obtenemos $b_9 \rightarrow b_1 \vee (b_2 \vee b_6) \vee b_5$ y entonces $b_9 \rightarrow b_1 \vee b_5$ que es $R_1 : a_4 \rightarrow a_2 \vee a_3$, obtenida de la jerarquía (ver figura 10 a).

10.4 Análisis numérico de objetos simbólicos

Se supone que el conjunto dado de objetos simbólicos es el conjunto de los primeros cinco objetos simbólicos definidos por la pirámide: $\{b_1, b_2, b_3, b_4, b_5\} = B$.

Una manera sencilla de hacer un puente entre los métodos cásicos de análisis de datos es calcular una medida de similitud entre los objetos de B; con esta medida es entonces posible usar ajuste multidimensional (*multidimensional scaling*), clasificación automática, etc. Para ello, podemos calcular la similitud s que fue definida por (2); como B es un conjunto de objetos simbólicos, tenemos que usar las funciones f_b y g_b definidas en §5.3. Tenemos, por ejemplo:

$$s_b(b_1, b_2) = \frac{b_1^*(b_2)}{\sqrt{b_1^*(b_1) b_2^*(b_2)}}$$

con $b_1 = [y_1 = q_1^1] \wedge_b [y_2 = q_2^1]$ donde q_1^1 y q_2^1 son funciones características tales que: $q_1^1(-1/2) = q_1^1(1/2) = 1$ y $q_2^1(-1/2) = q_2^1(1/2) = 1$, y $q_1^1(v) = q_2^1(v) = 0$ para $v \neq 1/2, -1/2$.

Tenemos $b_2 = [y_1 = q_1^2] \wedge_b [y_2 = q_2^2]$ y $q_1^2(v) = 1$ si $v \in \{1/2, 1/2\}$, $q_1^2(v) = 0$ sino, $q_2^2(v) = 1$ si $v \in \{1/2, 1\}$ y $q_2^2(v) = 0$ sino.

Como tenemos (ver §7)

$$\begin{aligned} b_1^*(b_2) &= f_b(\{g_b(q_i^1, q_i^2)\}_i) \\ &= \min(\langle q_1^1, q_1^2\rangle, \langle q_2^1, q_2^2\rangle) \\ &= \min\Big(\sum_{v\in O_1} q_1^1(v)q_1^2(v), \sum_{v\in O_2} q_2^1(v)q_2^2(v)\Big) \\ &= \min\Big(q_1^1(1/2)q_1^2(1/2), q_2^1(1/2)q_2^2(1/2)\Big) \\ &= \min(1,1) = 1, \end{aligned}$$

entonces obtenemos $b_1^*(b_1) = \min(\langle q_1^1, q_1^1\rangle, \langle q_2^1, q_2^1\rangle) = \min(2,2) = 2$
y también $b_2^*(b_2) = 2$; por lo tanto

$$s_b(b_1, b_2) = \frac{1}{\sqrt{2 \times 2}} = 1/2.$$

Calculando de la misma forma todas la similitudes $s_b(b^{(i)}, b^{(j)})$ finalmente obtenemos la tabla simétrica de similitudes dada en la figura 11 (a).

La similitud s_b es transformada en la disimilitud $d = 1 - s_b$ dada en la figura 11 b). Si escogemos $c = \max d(b_i, b_j) - M$, donde M es la suma de las dos parejas (b_i, b_j) de menor disimilitud $d(b_i, b_j)$, entonces $c \geq \max(d(b_i, b_j) - d(b_i, b_k) - d(b_k, b_j))$ y D tal que $D(b_i, b_j) = d(b_i, b_j)$, $D(b_i, b_i) = 0$, es una distancia pues $\forall\, i, j, k$, $d(b_i, b_j) + c \leq d(b_i, b_k) + c + d(b_k, b_j) + c$. Es fácil ver que $M = 0 + 0.3$, y $c = 1 - 0.3$; es entonces posible cambiar d en una distancia D tal que $D(b_i, b_j) = d(b_i, b_j) + 0.7$, que es dada en la figura 11 (c). Por lo tanto, es posible aplicar muchos métodos existentes de análisis de datos usando s, d o D como entrada.

10.5 Análisis simbólico de objetos simbólicos

Como entrada tenemos el conjunto siguiente de objetos probabilísticos: $B = \{b_1, \ldots, b_5\}$ tales que $b_j = [y_1 = q_1^j] \wedge_{\text{pr}} [y_2 = q_2^j]$ donde q_i^j es una medida de probabilidad de $P(O_j) \to [0, 1]$ donde $O_j = \{-1/2, 1/2, 1, 2\}$ y $P(O_j)$ es el conjunto de partes de O_j.

Si ponemos $b_j = \wedge_i [y_i = (q_i^j(v_1))v_1, (q_i^j(v_2))v_2, \ldots]$ los b_j son definidos como sigue, donde el valor v_ℓ asociado a $q_i^j(v_\ell) = 0$ no aparece:

- $b_1 = [y_1 = (1/2)(-1/2), (1/2)1/2] \wedge_{\text{pr}} [y_2 = (1/2)(-1/2), (1/2)1/2]$
- $b_2 = [y_1 = (1/2)1/2, (1/2)2] \wedge_{\text{pr}} [y_2 = (1/2)1/2, (1/2)1]$
- $b_3 = [y_1 = (1)2] \wedge_{\text{pr}} [y_2 = (1/2)1, (1/2), 2]$

s	b_1	b_2	b_3	b_4	b_5
b_1	1	1/2	0	0	0
b_2		1	$\frac{\sqrt{2}}{2}$	0	1/2
b_3			1	1	$\frac{\sqrt{2}}{2}$
b_4				1	$\frac{\sqrt{2}}{2}$
b_5					1

(a)

d	b_1	b_2	b_3	b_4	b_5
b_1	0	0.5	1	1	1
b_2		0	0.3	1	0.5
b_3			0	0	0.3
b_4				0	0.3
b_5					0

(b)

D	b_1	b_2	b_3	b_4	b_5
b_1	0	1.2	1.7	1.7	1.7
b_2		0	1	1.7	1.2
b_3			0	0.7	1
b_4				0	1
b_5					0

(c)

Figura 11

- $b_4 = [y_1 = (1/2)1, (1/2)2] \wedge_{\text{pr}} [y_2 = (1)2]$
- $b_5 = [y_1 = (1/2)1, (1/2)2] \wedge_{\text{pr}} [y_2 = (1/2)1, (1/2)2]$

Para tratar este conjunto de objetos probabilísticos, la primera manera sería calcular la similitud

$$s_{\text{pr}}(b_i, b_j) = \frac{b_i^*(b_j)}{\sqrt{b_i^*(b_i) b_j^*(b_j)}}$$

y entonces, usar por ejemplo, análisis en componentes principales o métodos de clasificación automática interpretados por objetos simbólicos tal como ya indicamos.

Por ejemplo, para la pareja (b_1, b_2), se calcula $b_1^*(b_2) = f_{\text{pr}}(\{g_{\text{pr}}(q_i^1, q_i^2)\}_i)$ como sigue:

$$b_1^*(b_2) = \text{ Media } \left(\langle q_1^1, q_1^2\rangle, \langle q_2^1, q_2^2\rangle\right)$$

por lo tanto:

$$b_1^*(b_2) = \text{ Media } \left(\sum_{v \in O_1} q_1^1(v) \cdot q_1^2(v), \sum_{v \in O_2} q_2^1(v) \cdot q_2^2(v)\right)$$

Por lo tanto

$$\begin{aligned} b_1^*(b_2) &= \text{ Media } (1/2 \cdot 0 + 1/2 \cdot 1/2 + 0 \cdot 0 + 0 \cdot 1/2, 1/2 \cdot 0 + 1/2 \cdot 1/2 + 0 \cdot 1/2 + 0 \cdot 0) \\ &= \text{ Media } (1/4, 1/4) = (1/4 + 1/4)1/2 \\ &= 1/4. \end{aligned}$$

Así $b_1^*(b_1) = \text{ Media } (1/2, 1/2) = 1/2$ y $b_2^*(b_2) = \text{ Media } (1/2, 1/2) = 1/2$.

Finalmente, poniendo $\alpha = \sqrt{\frac{3}{2}}$ obtenemos la siguiente similitud:

$$\{s_{pr}(b_i, b_j)\} = \begin{bmatrix} 1 & 1/2 & 0 & 0 & 0 \\ & 1 & \frac{\alpha}{2} & \frac{\alpha}{6} & 1/2 \\ & & 1 & \frac{2}{3} & \frac{2\alpha}{3} \\ & & & 1 & \frac{2\alpha}{3} \\ & & & & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.5 & 0 & 0 & 0 \\ & 1 & 0.6 & 0.2 & 0.5 \\ & & 1 & 0.7 & 0.8 \\ & & & 1 & 0.8 \\ & & & & 1 \end{bmatrix}$$

Para tratar a B, otra manera es obtener directamente de B clases de objetos simbólicos representados por una jerarquía de "herencia", en la que cada nodo es expresado por una aserción probabilística completa a_{jk}, o una aproximación de ella tal que si $a_{jk} = a_j \cup_{x,\alpha} a_k$ entonces $a_{jk} \geq_\alpha \max(a_j, a_k)$ donde $\cup_{\text{pr},\alpha}$ y $\geq_\alpha$ han sido definidos en §9.1. Para hacerlo, podemos usar el algoritmo siguiente sobre un conjunto de objetos simbólicos A:

- `Primer paso:` $a_{jk} = \cup_{x,\alpha} a_k$ es calculado $\forall a_j, a_k \in A$.
- `Segundo paso:` los a_{jk} de menor extensión constituyen los primeros niveles de la jerarquía, su altura es la cardinalidad de su extensión.

- **Tercer paso:** los a_{jk} retenidos en el paso 2 son añadidos a A y a_j, a_k son suprimidos de A; entonces regresamos al primer paso hasta que la cardinalidad de A sea 1.

En la práctica, ¿cómo podemos calcular $a_{jk} = a_j \cup_{\mathrm{pr},\alpha} a_k$? Por definición, a_{jk} es la conjunción de eventos elementales $a^i_{jk} = [y_i = q_i]$ tales que Ext $(a^i_{jk} / \Omega, \alpha)$ contiene $\Omega_1 =$ Ext $(a_j / \Omega, \alpha) \cup$ Ext $(a_k / \Omega, \alpha)$.

Por lo tanto, para cualquier $\omega \in \Omega_1$ tal que $\omega^s = \wedge_i [y_i = r_i]$ tenemos $a_{jk}(\omega) \geq \alpha$; esta condición es satisfecha si tenemos $\forall i$, $g(q_i, r_i) \geq \alpha$ pues $a_{jk}(\omega) = f(\{g(q_i, r_i)\}_i)$ y, por definición de f, es la media de números mayores que α; por lo tanto, si denotamos $x^i_j = q_i(v_j)$, tenemos la inecuación

$$g(q_i, r_i) = \sum_{v_j \in O_j} x^\ell_j \cdot r_i(v_j) \geq \alpha$$

por lo tanto, tenemos que resolver un sistema de $|\Omega_1|$ inecuaciones donde las incógnitas son las x^ℓ_j. Si este sistema tiene muchas soluciones, para cada i las denotamos $[y_i = q^\ell_i]$; por lo tanto, obtenemos $a_{jk} = \bigwedge_i {}_{\mathrm{pr}} \Big(\bigwedge_\ell {}_{\mathrm{pr}} [y_i = q^\ell_i] \Big)$; escogiendo $h_{\mathrm{pr}} = \min$ (ver §3.1) la extensión de a_{jk} al nivel α es $\Omega_2 = \{\omega / a_{jk}(\omega) = f(\{\min_\ell \{g(q^\ell_i, r_i)\}\}_i) \geq \alpha\}$.

Para obtener la jerarquía de herencia sobre B dada por el algoritmo, el primer paso consite en calcular los $a_{jk} = b_j \cup_{\mathrm{pr},\alpha} b_k$ cuya extensión es de cardinalidad mínima; escogemos $\alpha = 1/2$ y para calcular por ejemplo $a_{12} = b_1 \cup_{\mathrm{pr},\alpha} b_2$ hacemos lo siguiente:
primero ponemos $a_{jk} = a^1_{jk} \wedge a^2_{jk}$ donde $a^\ell_{jk} = [y_\ell = q_\ell]$ es tal que $a^\ell_{12}(b_1) \geq 1/2$ y $a^\ell_{12}(b_2) \geq 1/2$. Por lo tanto $x^\ell_j = q_\ell(v_j)$ donde $\{v_1, \ldots, v_4\} = O_1 = O_2 = \{-1/2, 1/2, 1, 2\}$, tenemos que resolver las siguientes inecuaciones, donde las x^ℓ_j son las incógnitas, con la restricción $\sum_j q_\ell(v_j) = \sum_j x^\ell_j = 1$:

$$a^\ell_{12}(b_1) = g_{\mathrm{pr}}(q_\ell, r^\ell_1) = \sum_{v_i \in O_\ell} q_\ell(v_i) r^1_\ell(v_i)$$

por lo tanto, obtenemos:

$$a^1_{12}(b_1) = 1/2x^1_1 + 1/2x^1_2 \geq 1/2; \quad a^1_{12}(b_2) = 1/2x^1_2 + 1/2x^1_4 \geq 1/2$$

de las que deducimos que $x^1_1 + x^1_2 = 1$, por lo tanto (como $\sum_{i=1}^{4} x^1_i = 1$) obtenemos $x^1_4 = 0$ y $x^1_2 = 1$, $x^1_i = 0$ si $i \neq 2$.

Se tiene

$$a^2_{12}(b_1) = 1/2x^2_1 + 1/2x^2_2 \geq 1/2; \quad a^1_{12}(b_2) = 1/2x^2_2 + 1/2x^2_3 \geq 1/2$$

de lo que resulta que $x^2_2 = 1$ y $x^2_i = 0$ si $i \neq 2$.

Finalmente, obtenemos:

$$a_{12} = a^1_{12} \wedge_{\mathrm{pr}} a^2_{12} = [y_1 = (1)1/2] \wedge_{\mathrm{pr}} [y_2 = (1)1/2]$$

(que es equivalente al objeto booleano $[y_1 = 1/2] \wedge_b [y_2 = 1/2]$).

De la misma forma, obtenemos:

$$a^1_{13}(b_1) = 1/2x^1_1 + 1/2x^1_2 \geq 1/2 \quad \text{y} \quad a^1_{13}(b_3) = x^1_4 \geq 1/2$$

esto es contradictorio pues la primera ecuación implica $x^1_4 = 0$. Por lo tanto el único objeto simbólico cuya extensión contiene b_1 y b_3 es el objeto pleno Ω^S cuya extensión es Ω; $\Omega^S = \wedge_i[y_i = q_i]$ es definido en el casi de objetos probabilísticos por funciones $q_i : P(O_i) \to \{1\}$ (¡que no son, por supuesto, probabilidades!), entonces es fácil ver que $\Omega^S(\omega) = 1, \forall \omega \in \Omega$.

De la misma forma obtenemos:

$$\begin{aligned} a_{14} &= a_{15} = a_{24} = \Omega^S \\ a_{23} &= [y_1 = (1)2] \wedge_{\text{pr}} [y_2 = (1)1] \\ a_{25} &= a_{23} \\ a_{34} &= [y_1 = (1)2] \wedge_{\text{pr}} [y_2 = 1(2)] \end{aligned}$$

a_{35} es calculado como sigue:

$$\begin{aligned} a^1_{35}(b_3) &= x^1_4 \geq 1/2 \\ a^1_{35}(b_5) &= 1/2x^1_3 + 1/2x^1_4 \geq 1/2 \text{ implica } x^1_4 = 1 \text{ y } x^1_i = 0 \text{ si } i \neq 4 \\ a^2_{35}(b_3) &= 1/2x^2_3 + 1/2x^2_4 \geq 1/2 \\ a^2_{35}(b_5) &= 1/2x^1_3 + 1/2x^2_4 \geq 1/2 \end{aligned}$$

Tenemos tres soluciones:

i) $x^2_3 = x^2_4 = 1/2$

ii) $x^2_3 = 1, x^1_i = 0$ para $i \neq 3$

iii) $x^2_4 = 1, x^2_i = 0$ para $i \neq 4$

Por lo tanto:

$$a_{35} = [y_1 = (1)2] \wedge_{\text{pr}} [y_2 = (1/2)1, (1/2)2] \wedge_{\text{pr}} [y_2 = (1)1] \wedge_{\text{pr}} [y_2 = (1)2]$$

De manera similar, tenemos:

$$a_{45} = [y_1 = (1/2)1, (1/2)2] \wedge_{\text{pr}} [y_1 = (1)1] \wedge_{\text{pr}} [y_1 = (1)2] \wedge_{\text{pr}} [y_2 = (1)2]$$

En la siguiente tabla damos en la entrada correspondiente a la i–ésima fila y la j–ésima columna la extensión de $a_{ij} = b_i \cup_{\text{pr},1/2} b_j$:

$$\text{Ext } (a_{ij}B, 1/2) =$$

	1	2	3	4	5
1		b_1b_2	Ω	Ω	Ω
2			$b_2b_3b_5$	Ω	$b_2b_3b_5$
3				$b_3b_4b_5$	b_3b_5
4					b_4b_5
5					

Usando esta tabla es fácil construir la jerarquía de herencia, uniendo en cada paso la pareja de menor extensión. Por lo tanto, las primeras parejas son (b_1, b_2), (b_3, b_5), (b_4, b_5); para obtener la jerarquía no es posible retener simultáneamente (b_3, b_5) y (b_4, b_5), por lo tanto si no hay restricciones externas en las clases (por ejemplo, restricciones de proximidad geográfica) tenemos que escoger una de ellas al azar; si retenemos por ejemplo (b_3, b_5) las primeras parejas a ser unidas son finalmente (b_1, b_2) y (b_3, b_5); por lo tanto, obtenemos los dos primeros niveles de la jerarquía caracterizados por $a_{12} = b_1 \cup_{\text{pr},1/2} b_2$ y $a_{35} = b_3 \cup_{\text{pr},1/2} b_2$. Por lo tanto, queda b_4 para ser fusionado con (b_1, b_2) o con (b_3, b_5). Es entonces fácil ver que $a^2_{124}(b_1) = 1/2x_1^2 + 1/2x_2^2 \geq 1/2$ y $a^2_{124}(b_4) = 1/2x_2^4 \geq 1/2$ lo que no da solución tal que $\sum_{i=1,4} x_i^2 = 1$; por lo tanto, $a_{124} = \Omega^S$ cuya extensión es B. Ya vimos que Ext $(a_{34}/B, 1/2) = \{b_3, b_4, b_5\}$, por lo tanto $a_{345} = a_{34}$; así, la próxima pareja para ser fusionada será $(b_4, (b_3, b_5))$ lo que da el tercer nivel representedo por $a_{345} = a_{34}$; el último nivel une (b_1, b_2) con (b_3, b_4, b_5) y es representado por el objeto pleno Ω^S.

Para resumir, hemos obtenido finalmente cuatro niveles cuya representación y extensión son dadas en la tabla 2.

Nivel	Representación	Extensión
1	$a_{12} = [y_1 = (1)1/2] \wedge_{\text{pr}} [y_2 = (1)1/2]$	$\{b_1, b_2\}$
2	$a_{35} = [y_1 = (1)2] \wedge_{\text{pr}} [y_2 = (1/2)1, (1/2)2] \wedge_{\text{pr}} [y_2 = (1)1] \wedge_{\text{pr}} [y_2 = (1)2]$	$\{b_3, b_4\}$
3	$a_{345} = [y_1 = (1)2] \wedge_{\text{pr}} [y_2 = (1)2]$	$\{b_3, b_4, b_5\}$
4	$a_{12345} = \sum_{i=1,2} [y_i = (1)(-1/2), (1)1/2, (1)1, (1)2] = \Omega^S$	B

Tabla 2

Usando el hecho de que la altura de cada nivel es la cardinalidad de la extensión de su aserción probabilística asociada, es fácil construir la jerarquía de herencia asociada al conjunto B de objetos probabilísticos, representada en la figura 12.

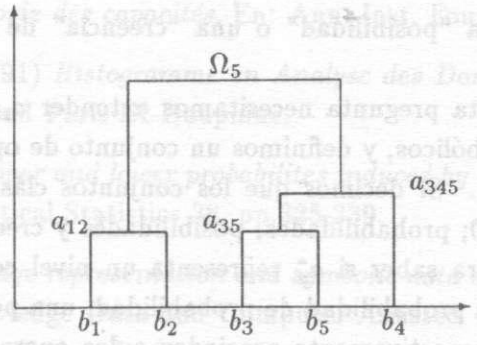


Figura 12: Jerarquía de herencia sobre objetos probabilísticos

Note que el mismo algoritmo puede ser usado con la unión probabilista, posibilista y creencia definidas respectivamente en §5, §6 y §7, en lugar de la unión simbólica definida en §9 que ha sido usada aquí. La ventaja de la unión simbólica (ver §8.1) es que define el supremo del retículo asociado al orden simbólico. La ventaja de la unión probabilista, posibilista y creencia es que permiten usar los teoremas 1, 2 y 3; en este caso si la alttura de un nivel definido por $a_3 = a_1 \cup_x a_2$ es dada por $a_3^*(a_1 \cup_x a_2)$, obtenemos en el caso de objetos probabilísticos

$$a_3^*(a_1 \cup_{\text{pr}} a_2) = a_3^*(a_1) + a_3^*(a_2) - a_3^*(a_1 \cap_{\text{pr}} a_2) \geq a_1^*(a_1) + a_2^*(a_2) - a_3^*(a_1 \cap_{\text{pr}} a_2)$$

Resulta que la jerarquía obtenida no tendrá inversiones, pues puede probarse que $a_3^*(a_3) \geq a_1^*(a_1)$ y $a_3^*(a_3) \geq a_2^*(a_2)$, y cuanto más independientes sean a_1 y a_2 (*i.e.* $a_1 \cap_{\text{pr}} a_2$ cercano a 0) más tenderá la altura de a_3 a ser grande).

Decimos que tenemos una regla entre dos aserciones probabilísticas a_1 y a_2 al nivel (α_1, α_2) denotada por $R : a_1 \overset{(\alpha_1,\alpha_2)}{\longrightarrow} a_2$ cuando Ext $(a_1/B, \alpha_1) \subseteq$ Ext $(a_2/B, \alpha_2)$; en otras palabras, la regla R es verdadera si, cuando b está en la extensión de a_1 al nivel α_1, entonces, está en la extensión de a_2 al nivel α_2; cuando $\alpha_1 = \alpha_2 = \alpha$ esta regla es denotada por $a_1 \overset{\alpha}{\rightarrow} a_2$. Usando esta notación es fácil inducir a partir de la jerarquía de herencia de la figura 6, yendo de abajo hacia arriba, la regla: $a_{35} \overset{1/2}{\rightarrow} a_{345}$; también es posible inducir de arriba hacia abajo la regla siguiente: $\Omega^S \overset{(1,1/2)}{\longrightarrow} a_{12} \vee a_{345}$ que significa que si b está en la extensión de Ω^S al nivel 1, está también en la extensión de a_{12} o de a_{345}, al nivel 1/2; de la misma forma también obtenemos $a_{345} \overset{1/2}{\rightarrow} b_4 \vee a_{35}$.

Conclusión

Empezando por las clases de objetos individuales definidos por intensión (en contraste con las clases sólo definidas por extensión) hemos dado varias maneras de definirlas por medio de una función a_x sobre Ω (el conjunto de los objetos individuales) dependiendo del conocimiento x. Surge naturalmente una pregunta: ¿Es posible decir que $a_x(\omega)$ mide

una "probabilidad", una "posibilidad" o una "creencia" de que ω pertenezca a la clase representada por a_x?

Para responder a esta pregunta necesitamos extender a_x a a_x^* definido sobre $\mathcal{A}_x$, un conjunto de objetos simbólicos, y definimos un conjunto de operadores $OP_x = \{\cup_x, \cap_x, c_x\}$ sobre $\mathcal{A}_x$ adaptado a x. Si decimos que los conjuntos clásicos representan un nivel de conocimiento de orden 0; probabilidades, posibilidades y creencias, un nivel conocimiento de orden 1, el asunto era saber si a_x^* representa un nivel conocimiento de orden 2. En otras palabras, si es una probabilidad de probabilidad, una posibilidad de posibilidad, una creencia de creencia, respectivamente asociados a los operadores correspondientes OP_x. Los teoremas 1,2 y 3 muestran que este es el caso si OP_x y las funciones f y g están bien escogidas.

En Reconocimiento de Patrones, el Análisis Simbólico de Datos (ASD) permite la representación y el análisis de patrones complejos; en "Procesamiento de Imágenes", el ASD puede ser usado por ejemplo con el fin de comparar varios sensores, para fusión de datos, o para entender imágenes por clasificación de objetos de alto nivel (casas, árboles, carreteras, ...) representadas por medio de objetos simbólicos sobre los cuales ASD puede ser usado.

En Aprendizaje de Inteligencia Artificial ("Learning Machine"), ASD permite la posibilidad de extender algoritmos inteligentes (donde las entradas son usualmente objetos individuales) a objetos simbólicos; más aún, definiendo objetos simbólicos sobre el conjunto Ω de muestras y no sobre el conjunto de descripción Δ, ADS permite hacer un puente entre Estadística y Máquinas Inteligentes.

Al contrario de la mayoría de los trabajos en Inteligencia Artificial, el Análisis Simbólico de Datos constituye una "crítica al razonamiento puro", dándole menos importancia al razonamiento y más importancia al estudio estadístico de bases de conocimiento, consideradas como "objetos simbólicos".

Un vasto campo de investigación se abre extendiendo la Estadística clásica a la Estadística de intensiones, y, especialmente, extendiendo problemas, métodos y algoritmos del Análisis de Datos a los objetos simbólicos.

Bibliografía

[1] Brito P., Diday E. (1990). *Pyramidal representation of symbolic objects.* En NATO ASI Series, Vol. F 61, Knowledge Data and Computer-Assisted Decisions, M. Schader y W. Gaul (eds.), Springer Verlag, Berlín.

[2] Brito P. (1991) *Analyse de Données Symboliques, Pyramides d'Héritage.* Tesis Doctoral de la Universidad Paris IX-Dauphine.

[3] Celeux G., Diday E., Govaert G., Lechevallier Y., Ralambondrainy H. (1989) *Classification Automatique: Environement Statistique et Informatique.* Ed. Dunod, París.

[4] Choquet G. (1953) *Théorie des capacités.* En: Ann. Inst. Fourier 5, pp.131-295.

[5] De Carvalho F.A.T. (1991) *Histogramme en Analyse des Données Symboliques.* Tesis Doctoral de la Universidad Paris IX-Dauphine.

[6] Dempster A.P.(1967) *Upper and lower probabilites induced by a multivaluated mapping* En: Annals of Mathematical Statistics 38, pp.325-339.

[7] Diday E. (1990) *Knowledge representation and symbolic data analysis.* En: NATO ASI Series, Vol. F 61 , Knowledge Data and Computer-Assisted Decisions, M. Schader y W. Gaul (eds.), Springer Verlag, Berlín.

[8] Diday E. (1991) *Des objets de l'analyse des donées a ceux de l'analyse des connaissances.* En "Induction Symbolique-Numérique à Partir de Données", Y. Kodratoff y E. Diday (eds), Cépaduès, Toulouse.

[9] Diday E. (1992) *From data to knowledge: new objects for a statistical analysis.* En: Conference on New Techniques and Technologies for Statistics, GMD, Bonn.

[10] Dubois D., Prade H. (1988) *Possibility Theory.* Plenum, New York.

[11] Dubois D., Prade H. (1986) *A set-theoretic view of belief funtions.* En: International Journal General Systems, Vol. 12, pp.193-226.

[12] Lauritzen S.L., Spiegelhalter D.J. (1990) *Local computation with probabilities on graphical structures and their application to expert systems.* En: Readings in Uncertain Reasoning, G. Shafer y J. Pearl (eds.), Morgan Kaufman Publ., San Mateo.

[13] Lebbe J., Vignes R., Darmoni S. (1989) *Symbolic numeric approach for biological knowledge representation: a medical example with creation of identification graphs.* En: Proc. of Conf. on Data Analysis, Learning Symbolic and Numerical Knowledge, Antibes, E. Diday (ed.), Nova-Science Publ., New York.

[14] Michalski R.S., Diday E., Stepp R.E. (1982) *Recent advances in data analysis: clustering objects into classes characterized by conjontive concepts.* Progress in Pattern Recognition, Vol. 1, L. Kanal y A. Rosenfeld (eds.).

[15] Pearl J. (1988) *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufman Publ., San Mateo.

[16] Schafer G. (1990) *Perspectives on the theory and practice of belief functions.* International Journal of Approximate Reasoning. Vol 4, Nos. 5 y 6.

[17] Schafer G. (1976) *A Mathematical Theory of Evidence.* Princeton University Press.

[18] Shweizer B., Sklar A. (1960) *Statistical metric spaces.* Pacific J. Math 10: 313-334.

[19] Zadeh L.A. (1971) *Quantitative fuzzy semantics.* Information Sciences, 159-176.

Construcción Eficaz de una Red Neuronal a partir de un Arbol de Decisión

Yves Lechevallier *

Resumen

Proponemos determinar la arquitectura de una red de neuronas a partir de un árbol de decisión. En este contexto, las conexiones nunca son completas entre las capas de la red. Después de la inicialización de los pesos, el algoritmo de retropropagación del gradiente es usado para optimizar la función de costo del error cuadrático medido sobre la capa de salida. La simulación muestra una ganancia interesante sobre el tiempo de aprendizaje de esta red. No obstante, debe tenerse en cuenta que el tiempo de aprendizaje, los parámetros medidos y las codificaciones, contribuyen grandemente al éxito de este enfoque.

Palabras clave: árbol de decisión, segmentación, clasificación por árboles, red multicapas.

1 Introducción

La construcción de funciones de decisión eficaces para resolver problemas de diagnóstico industrial, debe ser un compromiso entre dos objetivos opuestos. Uno es construir reglas de decisión fácilmente interpretables por el usuario; el otro, es obtener una tasa de buenas clasificaciones lo más elevada posible. Los métodos de segmentación realizan el primer objetivo generando una función de decisión que puede ser representada en la forma de un árbol de decisión, o bien, de manera equivalente, bajo la forma de una lista de reglas de producción. Contrariamente, las redes neuronales son cajas negras particularmente eficaces cuando son bien inicializadas. Utilizando estos dos enfoques, proponemos una metodología eficaz para resolver problemas de identificación en el mundo industrial.

A partir del árbol de decisión, se propone una representación en la forma de red con tres capas. La primera capa constituye un particionamiento del espacio de las variables seleccionadas. La segunda representa las regiones de decisión asociadas a cada nodo terminal del árbol. La tercera capa es una combinación disjunta de estas regiones y la decisión se realiza sobre la neurona más activa. La ventaja de este enfoque es que se determina, por medio del árbol de decisión, el número de capas y el número de neuronas por capa. Una de sus principales dificultades, es escoger las conexiones activas entre las capas.

*Institut National de Recherche en Informatique et Automatique – Rocquencourt, Francia

La primera solución propuesta por I. K. Sethi [17] es activar todas las conexiones entre dos capas sucesivas. Nosotros proponemos activar un subconjunto de estas conexiones teniendo en cuenta las relaciones entre los nodos no terminales del árbol y escoger unas ponderaciones iniciales de modo que la red simule perfectamente, sobre el conjunto de aprendizaje, el árbol de decisión. Las ponderaciones entre las neuronas son luego modificadas de manera incremental.

1.1 Árbol de decisión, segmentación

En el mundo industrial, las informaciones son adquiridas por medio de señales o de imágenes. Con esta información, el usuario puede calcular numerosos parámetros. También el método de discriminación debe jugar un papel explicativo y de ayuda para la selección de los parámetros. De los árboles de decisión se obtiene un árbol binario que da una representación jerárquica de las zonas del espacio de los datos y las funciones de decisión asociadas pueden ser escritas en forma de reglas de produción.

El desarrollo de estos métodos fue iniciado por Morgan y Sonquist [13]. El libro de Breiman, Friedman, Olsen y Stone [1] es una referencia importante, así como el artículo de Quinlan [14] en el dominio del aprendizaje. El algoritmo asociado a estos métodos es muy simple; se trata de construir de manera recursiva un árbol binario, seleccionando en cada nodo de este árbol la mejor pregunta binaria. Las herramientas necesarias para la construcción de un árbol de decisión son:

- la definición de un conjunto de preguntas binarias,
- un criterio de evaluación,
- una regla para detener la construcción del árbol,
- una regla de asignación de cada segmento terminal a una clase a priori.

Los métodos de segmentación tienen dos representaciones principales:

- Una es en la forma de árbol binario, por ejemplo: el árbol de la figura 1 está constituido de dos preguntas binarias ¿$(x_5 < 2)$? y ¿$(x_2 = 1)$?, con tres segmentos terminales. Uno de estos segmentos está asociado a la clase a priori número 1 y, los otros dos, a la clase a priori 2.

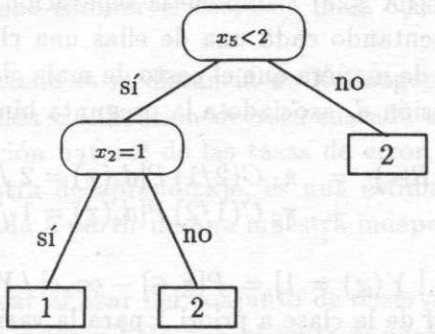


Figura 1: Árbol de decisión

- La otra es en la forma de reglas de producción, por ejemplo:

SI	$(x_5 < 2)$
Y	$(x_2 = 1)$
ENTONCES	Clase a priori 1
SI	$(x_5 \geq 2)$
O	$((x_5 < 2)$ Y $(x_2 \neq 1))$
ENTONCES	Clase a priori 2

Durante la fase de reconocimiento, la claridad y la simplicidad de esta representación hacen posible un diálogo y da al usuario una interpretación del poder discriminante de cada parámetro.

1.2 Escogencia de las preguntas binarias y de la regla de asignación

Los parámetros calculados son siempre variables continuas. Así, el conjunto de las preguntas binarias es el conjunto de todos los cortes posibles sobre las variables continuas seleccionadas. Hemos escogido como regla de asignación de cada segmento terminal a una clase a priori, la regla bayesiana cuyo principio es el que sigue: se denotan π_j las probabilidades a priori asociadas a las k clases a priori y $C(i/j)$ el costo de mala clasificación de una observación de la clase a priori j, en la clase de asignación i. Para una observación representada por sus p valores $\underline{x} = (x_1, \ldots, x_p)$ y perteneciente a la clase a priori j (denotada por $Y(\underline{x}) = j$) decidimos su asignación a la clase a priori i utilizando la función de decisión d (denotada por $d(\underline{x}) = i$). El costo de mala clasificación asociado a esta función es igual a:

$$R(d) = \sum_{j=1}^{k} \pi_j \left[\sum_{i \neq j} C(i/j)\, P\left[d(\underline{x}) = i \,/\, Y(\underline{x}) = j\right] \right]$$

La función de decisión d se llama **regla de Bayes** si su costo asociado es el más pequeño posible.

1.3 Escogencia del criterio de evaluación

Cada pregunta binaria ¿$(X \leq c)$? impone la separación de la variable X seleccionada, en dos semirrectas, representando cada una de ellas una clase a priori. En este contexto se desea escoger el corte c de manera que el costo de mala clasificación sea mínimo. Este costo R de la función de decisión d_c asociado a la pregunta binaria ¿$(X \leq c)$? se escribe así:

$$\begin{aligned} R(c) &= \pi_1 C(2/1) P[d_c(\underline{x}) = 2 \,/\, Y(\underline{x}) = 1] \\ &+ \pi_2 C(1/2) P[d_c(\underline{x}) = 1 \,/\, Y(\underline{x}) = 2] \end{aligned}$$

Como $P[d_c(\underline{x}) = 2 \mid Y(\underline{x}) = 1] = P[\underline{x} \in]-\infty, c] \,/\, Y(\underline{x}) = 1] = F_1(c)$ donde F_1 es la función de distribuciónf de la clase a priori 1 para la variable X seleccionada, se obtiene:

$$R(c) = \pi_1 C(2/1) F_1(c) + \pi_2 C(1/2)(1 - F_2(c))$$

Si se supone que $\pi_1 C(2/1) = \pi_2 C(1/2) = \alpha$ entonces se tiene:

$$R(c) = \alpha + F_1(c) - F_2(c)$$

Se observará que $\text{Mín}_c R(c)$ se alcanza en el mismo punto que $\text{Máx}_c |F_1(c) - F_2(c)|$ que es la distancia de Kolmogorov–Smirnov entre las dos clases a priori. Se escoge la región discriminante $R_1 =]-\infty, c]$ si $F_1(c) > F_2(c)$, si no $R_1 =]c, +\infty[$. Se estima $D(c)$ por:

$$\hat{D}(\hat{c}) = \text{Sup}_x |\widehat{F_1}(x) - \widehat{F_2}(x)|$$

donde $\widehat{F_i}$ son las estimaciones de las funciones de distribución F_i (Friedman [6], Celeux y Lechevallier [2]).

Cuando el número de clases a priori es superior a 2, la primera solución es considerar la discriminación de k clases a priori como una sucesión de discriminaciones de dos clases, oponiendo cada clase a priori a las otras. Pero esto genera k árboles de decisión y la definición de una regla mayoritaria difícilmente utilizable. En Celeux y Lechevallier [2] se propone otra solución que permite construir sólo un árbol. Se trata, para cada variable y para cada corte, de buscar el mejor reagrupamiento A de las clases a priori. De donde

$$\hat{D}(\hat{c}) = \sup_x \sup_{A \in \mathcal{A}} \left| \widehat{F_A}(x) - \widehat{F_{\overline{A}}}(x) \right|$$

donde $\overline{A}$ es el complemento de A, $\mathcal{A}$ es el conjunto de todos los reagrupamientos posibles de las clases a priori y $\widehat{F_A}$ es la estimación de la función de distribución de A dada por:

$$\widehat{F_A}(x) = \frac{1}{\pi_A} \sum_{i \in A} \pi_i \hat{F}(x)$$

donde $\pi_A = \sum_{i \in A} \pi_i$.

Se muestra que la solución pertenece a un subconjunto de $\mathcal{A}$ con $k - 1$ elementos, lo cual tiene la ventaja de que elimina, en este caso, el carácter combinatorio de la búsqueda.

1.4 *Estimación del costo de mala clasificación*

Adoptando esta estrategia, se debe estimar el costo de mala clasifcación asociado a este método.

El criterio de evaluación utilizado es la distancia de Kolmogorov–Smirnov. Esta escogencia implica que el costo de mala clasificación decrece en cada segmentación de nuestra población. Por tanto, la estimación natural de las tasas de error, que consiste en aplicar esta regla de decisión a la muestra de aprendizaje, es una estimación muy optimista de dicha tasa. Ésta debe ser calculada a partir de una muestra independiente del conjunto de aprendizaje.

La estrategia adoptada es sacar al azar del conjunto de observaciones, un subconjunto que representa el 20% de la población. Este subconjunto servirá de conjunto test y las observaciones restantes constituirán el conjunto de aprendizaje. La ventaja de esta técnica es la ausencia de sesgo en la estimación de las tasas de error y su facilidad de implementación. Otras posibilidades de estimar esta tasa de error son descritas y propuestas en Breiman y coautores [1], Mingers [11] y en el libro "Analyse discriminante sur variables continues" Celeux [4].

2 Aplicación al análisis de las encuestas psico-sociales

2.1 Contexto del análisis

Algunas encuestas psico–sociales se refieren al uso y abuso de las drogas. Ante este fenómeno plurifactorial, donde el individuo, la droga y el contexto social interactúan, los estudios epidemiológicos son organizados para ilustrar ciertas etapas de la génesis del desarrollo de la toxicomanía [12]. Las encuestas epidemiológicas se basan en cuestionarios cerrados con tres paquetes de datos: los datos socio–demográficos y legales, el consumo de droga y los datos médicos. Estos tres paquetes son caracterizados esencialmente por unas variables cualitativas.

La encuesta que se analiza fue realizada entre 1988 y 1989 a 3099 personas en prisión identificadas como toxicómanos. Estos pacientes han sido descritos individualmente y anónimamente de acuerdo con un cuestionario que aborda los factores socio–demográficos y las características de los productos (drogas). Así se constituyó una muestra de 3040 expedientes. El objetivo del análisis es predecir los datos socio–demográficos y legales, a partir de los datos médicos.

2.2 Selección de los grupos a priori

El objetivo de los métodos de discriminación es predecir o reconocer el grupo a priori de un individuo en función de un conjunto de predictores. En el caso de la encuesta que nos ocupa, los predictores son 9 variables cualitativas asociadas a los datos médicos. Como los datos socio–demográficos y legales son caracterizados por varias variables cualitativas, no podemos aplicar directamente los métodos de discriminación. Usando métodos de clasificación automática con los datos penales y sociales, se obtiene una tipología de la muestra en 5 clases, la cual constituye la variable a predecir. Sin embargo, dos clases son de escaso

interés pues una tiene muy poco efectivo (6%) y la otra comprende cuestionarios con muchas "no respuestas".

A partir de las clases 1 (32% efectivo), 2 (27% efectivo) y 3 (30% efectivo) se realiza una discriminación. Después de estos análisis, se verá que es difícil reconocer estas tres clases. Las clases 2 y 3 forman una población homogénea en función de la edad. Se hará una discriminación entre ellas para analizar si existe diferencia de comportamiento.

2.3 Resultados de la discriminación en tres clases

La variable a explicar representa las tres clases de la tipología. Las variables explicativas conciernen a los datos médicos. Después de la aplicación del método de clasificación, las concordancias son analizadas por clase para las muestras de aprendizaje y test. En las tablas 1, 2, 3, 4, 5 y 6 que siguen, se cruzan las clases de asignación (por filas) con las clases a priori (por columnas).

	1	2	3
1	338	232	159
2	138	167	97
3	284	286	339

Tabla 1: Tabla de clasificación para el conjunto de aprendizaje.

Las tasas de buena clasificación aparente son de 44,5% para la clase 1, de 24,4% para la clase 2 y 57% para la clase 3.

	1	2	3
1	77	70	41
2	29	44	29
3	56	63	101

Tabla 2: Tabla de clasificación para el conjunto test.

Las tasas de buena clasificación de la clase 1 es 47,5%, de la clase 2 es 24,9% y de la clase 3 es 59,1%. Utilizando la estrategia de la validación cruzada obtenemos:

	1	2	3
1	407	293	200
2	154	153	116
3	379	367	481

Tabla 3: Tabla de clasificación por validación cruzada.

Las tasas de buena clasificación de la clase 1 es ahora 43,3%, de la clase 2 es 18,8% y de la clase 3 es 60,3%. Estas tasas son muy parecidas a las obtenidas mediante la estrategia del conjunto test. Es claro que hay un muy mal reconocimiento de la clase 2. Por el contrario, las clases 1 y 3 son muy bien reconocidas.

2.4 Resultados de la discriminación en dos clases

La variable a explicar comprende ahora sólo las clases 2 y 3 de la tipología. El conjunto de variables explicativas es idéntico al conjunto anterior. Las tablas 4 y 5 presentan las clasificaciones sobre los conjuntos de aprendizaje y test, respectivamente.

	2	3
2	362	206
3	325	395

Tabla 4: Tabla de clasificación para el conjunto de aprendizaje.

La tasa de buena clasificación aparente de la clase 2 es 52,7% y la de la clase 3 es 65,7%. Por otra parte, sobre el conjunto test la tasa de buena clasificación de la clase 2 es 49,7% y de la clase 3 es 62,6%.

	2	3
2	87	55
3	88	92

Tabla 5: Tabla de clasificación para el conjunto test.

Utilizando la estrategia de la validación cruzada, se obtiene de la tabla 6 que la tasa de buena clasificación de la clase 2 es igual a 50,4%, mientras que de la clase 3 es 64%.

	2	3
2	410	287
3	403	510

Tabla 6: Tabla de clasificación para la validación cruzada.

Como para la discriminación anterior las tasas obtenidas sobre el conjunto test y las obtenidas por la validación cruzada son muy próximas, en este contexto tenemos una buena separación entre las dos clases. Sin embargo, no podemos explicar la pequeña magnitud de la tasa de reconocimiento de la clase 2 en la primera discriminación.

3 Los principales resultados de la segmentación

Seleccionando el criterio asociado a la distancia de Kolmogorov–Smirnov hemos transformado todas las variables cualitativas en variables binarias utilizando una codificación disyuntiva completa. Esta codificación crea una variable binaria por cada modalidad de la variable cualitativa por codificar.

Hemos construido dos árboles debidamente podados, a partir de la discriminación con tres y dos clases respectivamente.

3.1 Resultados de la discriminación con tres clases

En este caso, la función de decisión construye 8 segmentos terminales correspondientes al siguiente árbol de decisión, debidamente podado, donde NR simboliza "no responde":

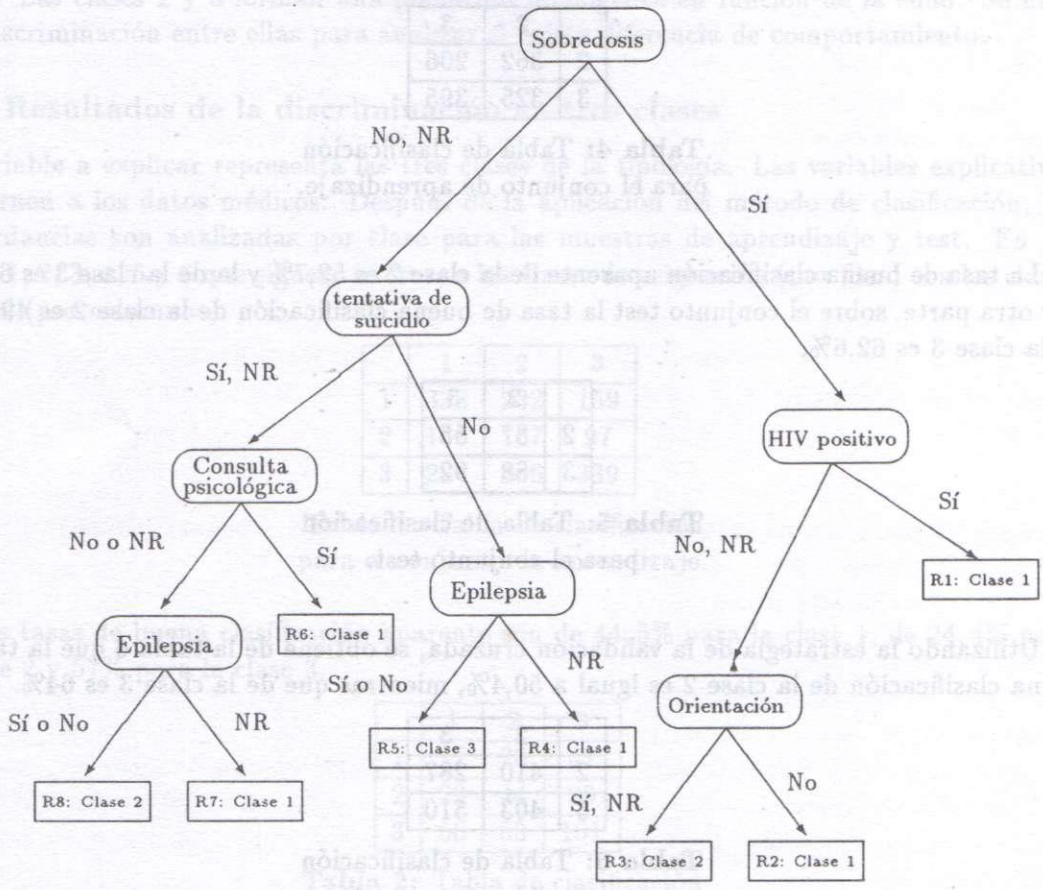


Figura 2: Árbol binario de la discriminación en tres clases

Los segmentos R_1, R_2, R_4, R_6 y R_7 caracterizan la clase 1; los segmentos R_3 y R_8 la clase 2 y el segmento R_5 la clase 3.

Para la clase 1 tenemos las siguientes reglas de producción:

- REGLA 1 asociada al segmento R_1:
  [Sobredosis=Sí] y
  [HIV positivo=Sí]
  Esta regla se verifica para el 14% de los individuos de la clase 1.
- REGLA 2 asociada al segmento R_2:
  [Sobredosis=Sí y [[HIV positivo=No] o [HIV positivo=No Responde]] y

[Orientación=No]

Esta regla se verifica para el 12% de los individuos de la clase 1.

- REGLA 4 asociada al segmento R_4:

  [Sobredosis=No] o [Sobredosis=No Responde] y

  [Tentativa de suicidio=No] y [Epilepsia=No Responde].

- REGLA 6 asociada al segmento R_6:

  [[Sobredosis=No] o [Sobredosis=No Responde]] y

  [[Tentativa de suicidio=Sí] o [Tentativa de suicidio=No Responde]] y [Consulta=Sí].

- REGLA 7 asociada al segmento R_7:

  [[Sobredosis=No] o [Sobredosis=No Responde]] y

  [[Tentativa de suicidio=Sí] o [Tentativa de suicidio=No Responde]] y [[Consulta=No] o [Consulta=No Responde]] y

  [Epilepsia=No Responde].

Las reglas 4, 6 y 7 son cada una verificadas para aproximadamente el 10% de los individuos de la clase 1.

Para la clase 2 tenemos las reglas siguientes:

- REGLA 3 asociada al segmento R_3:

  [Sobredosis=Sí] y [[HIV positivo=No] o [HIV positivo=No Responde]] y

  [[Orientación=Sí] o [Orientación=No Responde]].

- REGLA 8 asociada al segmento R_8:

  [[Sobredosis=No] o [Sobredosis=No Responde]] y

  [[Tentativa e suicidio=Sí] o [Tentativa de suicidio=No Responde]] y [Consulta=No] o [Consulta=No Responde]] y

  [[Epilepsia=Sí] o [Epilepsia=No]].

Los reglas 3 y 8 son cada una verificadas para aproximadamente 10% de los individuos de la clase 2.

Para la clase 3 tenemos la siguiente regla:

- REGLA 5 asociada al segmento R_5:

  [[Sobredosis=No] o [Sobredosis=No Responde]] y [Tentativa de suicidio=No] y

  [[Epilepsia=Sí] o [Epilepsia=No]]

  Esta regla representa la clase de asignación más importante.

El rendimiento del método es medido por las tablas siguientes:

	1	2	3
1	411	281	206
2	133	151	122
3	202	235	299

Tabla 7: Tabla de clasificación para el conjunto de aprendizaje.

La tasa de buena clasificación aparente de la clase 1 es 55%, de la clase 2 es 22,6% y de la clase 3 es 47,7%.

	1	2	3
1	101	62	62
2	40	23	24
3	53	61	84

Tabla 8: Tabla de clasificación para el conjunto test.

Las tasas de buena clasificación de la clase 1 es 52,1%, de la clase 2 es 15,7% y el de la clase 3 es 49,4%.

El rendimiento por este método es idéntico al de la discriminación. Sin embargo, la segmentación da una explicación al no reconocimiento de la clase 6. La población del segmento R_6, obtenida por medio de la muestra de aprendizaje como por la muestra test, es compuesta por individuos de las clases 1 y 2 en la misma proporción. La proporción de la clase 1 siendo ligeramente superior, el segmento es asignado a la clase 1. La segmentación de R_6 no es posible a partir de las variables del paquete de datos médicos. Este particionamiento sería fácil introduciendo la edad.

3.2 Resultados de la discriminación con dos clases

De la figura 3 se ve que los segmentos S_1, S_2 y S_3 caracterizan la clase 2 y el segmento S_4 la clase 3. %2 es la frecuencia con que un individuo que pertenece a la clase 2 es asociado al segmento S_2. Igualmente, %3 en relación con la clase 3.

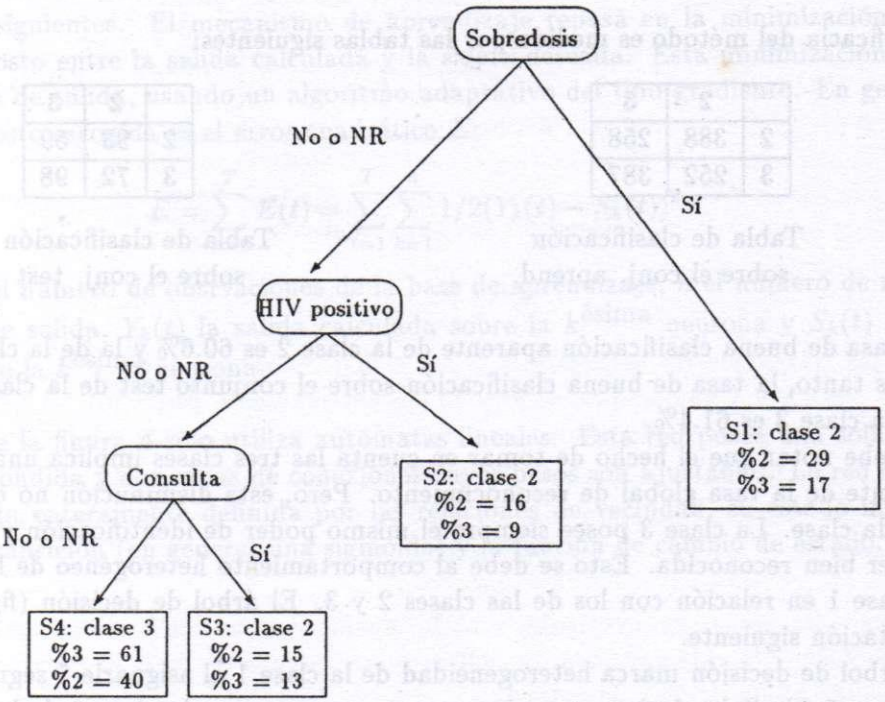


Figura 3: Árbol binario de la discriminación en dos clases

Para la clase 2 tenemos las siguientes reglas:

- REGLA 1 asociada al segmento S_1:
  [Sobredosis = Sí].
- REGLA 2 asociada al segmento S_2:
  [[Sobredosis=No] o [Sobredosis=No Responde]] y
  [HIV=Positivo].
- REGLA 3 asociada al segmento S_3:
  [[Sobredosis=No] o [Sobredosis=No Responde]] y
  [[HIV=No Positivo] o [HIV=No Responde]] y [Consulta=Sí].

Para la clase 3 tenemos la regla siguiente:

- REGLA 4 asociada al segmento S_4:

[[Sobredosis=No] o [Sobredosis=No Responde]] y
[[HIV=No Positivo] o [HIV=No Responde]] y
[[Consulta=No] o [Consulta=No Responde]].

La eficacia del método es medida por las tablas siguientes:

	2	3
2	388	258
3	252	387

Tabla de clasificación sobre el conj. aprend.

	2	3
2	93	59
3	72	98

Tabla de clasificación sobre el conj. test

La tasa de buena clasificación aparente de la clase 2 es 60.6% y la de la clase 3 es 60%. Mientras tanto, la tasa de buena clasificación sobre el conjunto test de la clase 2 es 56.4% y el de la clase 3 es 61.4%.

Se debe notar que el hecho de tomar en cuenta las tres clases implica una disminución importante de la tasa global de reconocimiento. Pero, esta disminución no es equivalente para cada clase. La clase 3 posee siempre el mismo poder de identificación y la clase 2 no puede ser bien reconocida. Esto se debe al comportamiento heterogéneo de los individuos de la clase 1 en relación con los de las clases 2 y 3. El árbol de decisión (figura 2) da la interpretación siguiente.

El árbol de decisión marca heterogeneidad de la clase 1 al asignarle 5 segmentos terminales; la variable Sobredosis es muy importante para separar la clase 2 de la clase 3. Por el contrario, ella no caracteriza la clase 1 puesto que los numerosos segmentos asociados a esta clase se reparten equitativamente entre las partes izquierda y derecha del árbol de decisión. El conjunto de los individuos que respondieron positivamente a la pregunta Sobredosis (parte derecha del árbol) de la clase 1 son difícilmente separables de los individuos de la clase 2 que respondieron de la misma manera a esta pregunta.

En la parte izquierda del árbol, formada por los individuos que respondieron negativamente a la pregunta Sobredosis, se encuentran dos conjuntos de individuos de la clase 1. El primer conjunto, representado por el segmento R_4, se opone a la clase 3 (segmento R_5) utilizando la pregunta concerniente a la epilepsia. El segundo conjunto, representado por los segmentos R_6 y R_7, es difícilmente discriminado de los elementos de la clase 2 (segemento R_8) por las dos preguntas asociadas a la consulta y a la epilepsia.

4 Red multicapas

El algoritmo de retropropagación del gradiente tiene su origen en el "Perceptron" que es una red con una capa presentada por Rosenblatt [15]. Este trabajo se apoyaba sobre el de Mc Culloch y Pitts [9]. Rumelhart [16] y Le Cun [8] extienden la arquitectura a las redes multicapas y resuelven el problema de la propagación de los errores en las capas escondidas. Esta red pertenece a la familia de los algoritmos supervisados. Su funcionamiento es así: se le presentan secuencialmente las observaciones, el algoritmo evoluciona hasta llegar a

un cierto estado el cual es comparado con una respuesta deseada S (la clase a priori de la observación). La red adapta entonces los pesos W_ij de las conexiones de las neuronas para realizar la correspondencia deseada. La base de aprendizaje es presentada indefinidamente hasta la obtención de un mínimo aceptable de la función de costo.

Las neuronas de una misma capa no están conectadas entre ellas, pero están conectadas a las capas siguientes. El mecanismo de aprendizaje reposa en la minimización de una función de costo entre la salida calculada y la salida deseada. Esta minimización se hace sobre la capa de salida, usando un algoritmo adaptativo del tipo gradiente. En general, la función de costo escogida es el error cuadrático E:

$$E = \sum_{t=1}^{T} E(t) = \sum_{t=1}^{T} \sum_{k=1}^{n} 1/2(Y_k(t) - S_k(t))^2$$

donde T es el número de obsrvaciones de la base de aprendizaje, n el número de neuronas de la capa de salida, $Y_k(t)$ la salida calculada sobre la $k^{\text{ésima}}$ neurona y $S_k(t)$ la salida deseada sobre la $k^{\text{ésima}}$ neurona.

La red de la figura 4 sólo utiliza autómatas lineales. Esta red posee una sola capa de neuronas escondida y dos capas de conexiones cuyos pesos son ajustables. La red es de dos capas. Queda enteramente definida por las relaciones de vecindad, su estado interno, la función de transición (en general una sigmoide) y la función de cambio de estado.

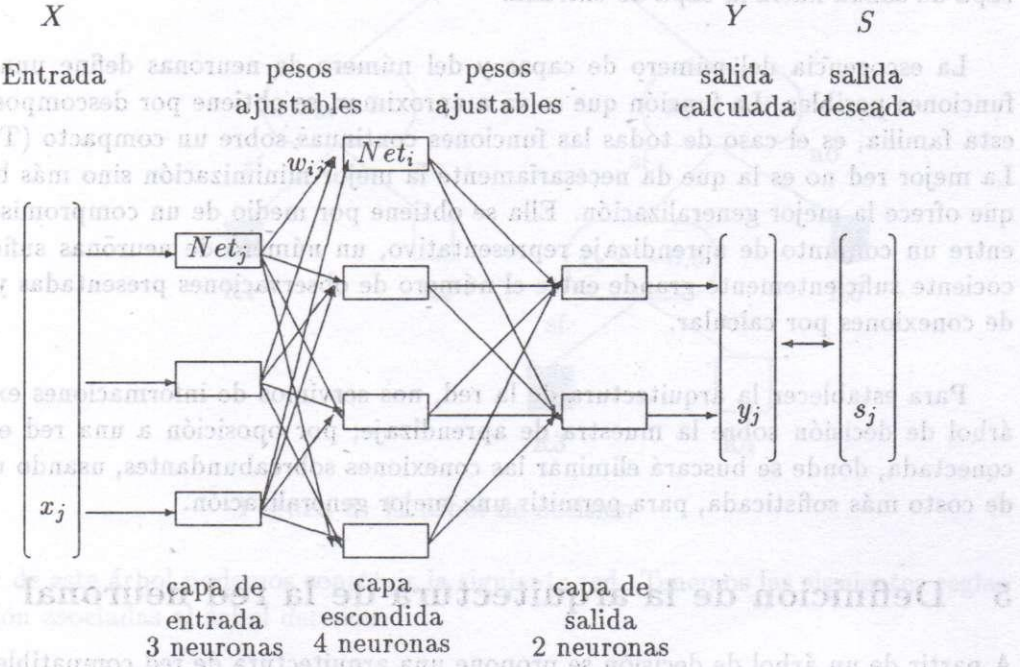


Figura 4: Una red multicapas

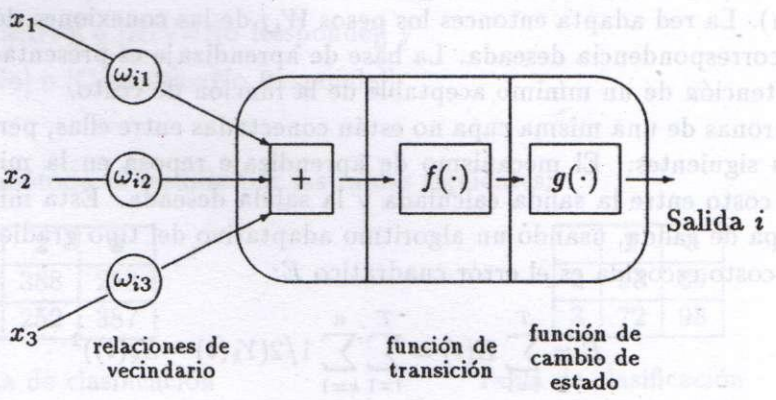


Figura 5: Una neurona

Con las relaciones de vecindad X_1, X_2 y X_3: el estado interno se caracteriza por $Net_i = \sum_j W_{ij}X_j$. Cuando Net_i es grande (respectivamente pequeño), $f(Net_i)$ es grande (respectivamente pequeño) hay entonces activación (respectivamente inhibición). La función de transición es la función sigmoide: $f(x) = b\left(\frac{e^{ax}+1}{e^{ax}-1}\right)$. La función de cambio de estado g_i es usualmente tomada en cuenta como la función identidad. Las ecuaciones de esta red son desarrolladas en el artículo de Chabanon–Dubuisson que aparece en [4].

Las dos características principales son: la función de transición es continua y derivable (una sigmoide) y la retropropagación del error se calcula de manera recurrente desde la capa de salida hacia la capa de entrada.

La escogencia del número de capas y del número de neuronas define una familia de funciones posibles. La función que se va a aproximar, se obtiene por descomposición sobre esta familia, es el caso de todas las funciones continuas sobre un compacto (Thiria, [18]). La mejor red no es la que da necesariamente la mejor minimización sino más bien aquélla que ofrece la mejor generalización. Ella se obtiene por medio de un compromiso aceptable entre un conjunto de aprendizaje representativo, un número de neuronas suficiente y un cociente suficientemente grande entre el número de observaciones presentadas y el número de conexiones por calcular.

Para establecer la arquitectura de la red, nos servimos de informaciones extraidas del árbol de decisión sobre la muestra de aprendizaje; por oposición a una red enteramente conectada, donde se buscará eliminar las conexiones sobreabundantes, usando una función de costo más sofisticada, para permitir una mejor generalización.

5 Definición de la arquitectura de la red neuronal

A partir de un árbol de decisión se propone una arquitectura de red compatible con aquél. Como para I.K. Sethi [17], la red es multicapas y su número de capas de conexiones es tres. La primera capa es la capa PARTICIONAMIENTO, que constituye una codificación del espacio

de variales. La segunda capa es la capa ET que representa las regiones de las decisiones asociadas a cada segmento terminal del árbol. La tercera capa es la OU, una combinación disyuntiva de estas regiones y la decisión se realiza sobre la neurona más activada.

El número de neuronas de la primera capa escondida es igual al número de nodos del árbol de decisión, es decir, al número de segmentos no terminales. El número de neuronas de la segunda capa escondida es igual al número de segmentos terminales del árbol. Esta red multicapas puede ser enteramente conectada. No obstante, para beneficiar las capacidades de generalización de una red neuronal, se debe reducir al máximo el número de estas conexiones. *Ahora se debe escoger el algoritmo que permita calcular los pesos asociados a las* conexiones seleccionadas. I. K. Sethi [18] propone un algoritmo muy próximo al algoritmo del perceptron. Nosotros escogimos el algoritmo de retropropagación porque minimiza un criterio basado sobre el error cuadrático. El ejemplo siguiente ilustra el paso de un árbol de decisión a una red de neuronas. Al inicio, nuestro conjunto de aprendizaje es definido por nueve variables y sus elementos pertenecientes a dos clases a priori. El árbol de decisión ha seleccionado las variables X_4, X_8, X_1, X_7 y está constituido por cuatro nodos y cinco segmentos terminales.

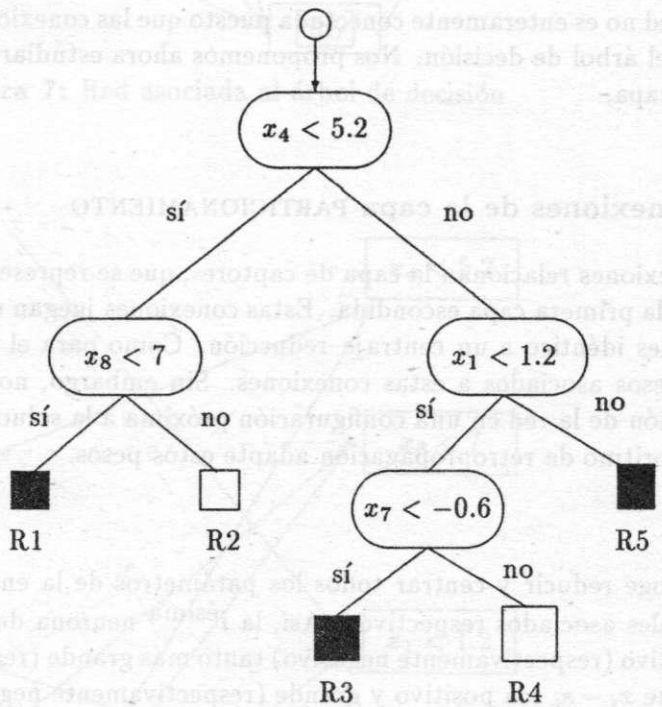


Figura 6: El árbol de decisión

A partir de este árbol podemos construir la siguiente red. Tenemos las siguientes reglas de producción asociadas al árbol decisión:

- Regla R_1

  SI $(X_4 < 5.2)$ Y $(X_8 < 7)$ ENTONCES clase 1

- Regla R_2

  SI $(X_4 < 5.2)$ Y $(X_8 \geq 7)$ ENTONCES clase 2

- Regla R_3

  SI $(X_4 \geq 5.2)$ Y $(X_1 < 1.2)$ Y $(X_7 < -0.6)$ ENTONCES clase 1

- Regla R_4

  SI $(X_4 \geq 5.2)$ Y $(X_1 < 1.2)$ Y $(X_7 \geq -0.6)$ ENTONCES clase 2

- Regla R_5

  SI $(X_4 \geq 5.2)$ Y $(X_1 \geq 1.2)$ ENTONCES clase 1

Esta red no es enteramente conectada puesto que las conexiones activadas se han definido mediante el árbol de decisión. Nos proponemos ahora estudiar esta selección de conexiones capa por capa.

5.1 Conexiones de la capa PARTICIONAMIENTO

Estas conexiones relacionan la capa de captores, que se representa por las variables descriptivas, con la primera capa escondida. Estas conexiones juegan un papel de normalización de las variables idéntico a un centraje–reducción. Como para el perceptron, se puede decidir fijar los pesos asociados a estas conexiones. Sin embargo, nos proponemos más bien una inicialización de la red en una configuración próxima a la solución dada por el árbol y dejar que el algoritmo de retropropagación adapte estos pesos.

Se escoge reducir y centrar todos los parámetros de la entrada de la red alrededor de sus umbrales asociados respectivos. Así, la $i^{\text{ésima}}$ neurona de la primera capa recibirá un valor positivo (respectivamente negativo) tanto más grande (respectivamente pequeño) en la medida que $x_i - s_i$ sea positivo y grande (respectivamente negativo y pequeño). Por tanto, cuanto más el valor se aleje del umbral, más la neurona es activada, lo que corresponde perfectamente a la estrategia del árbol de decisión donde el alejamiento del corte determina una asigación no ambigua de una clase a priori. Si se desea permitir a la red neuronal modificar el valor del corte, se debe, en este caso, agregar un sesgo inicializado en el valor cero.

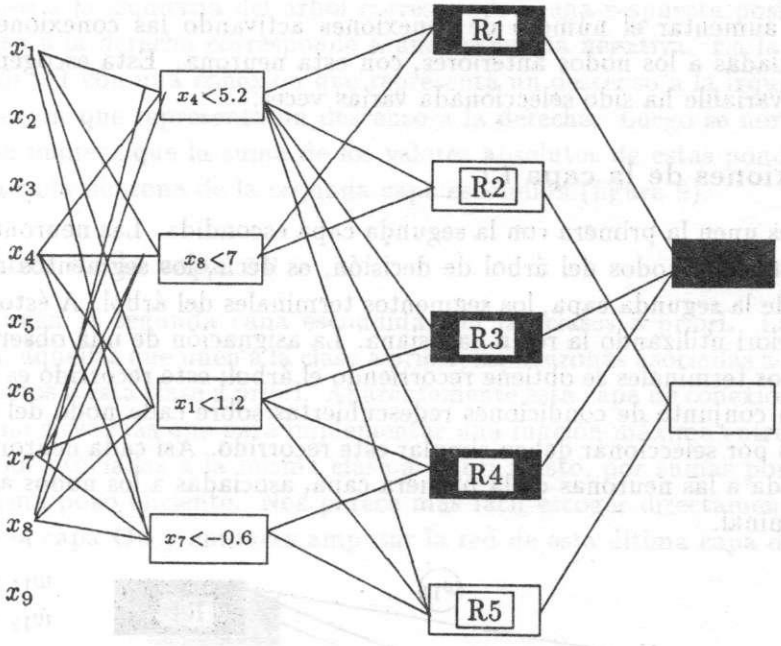


Figura 7: Red asociada al árbol de decisión

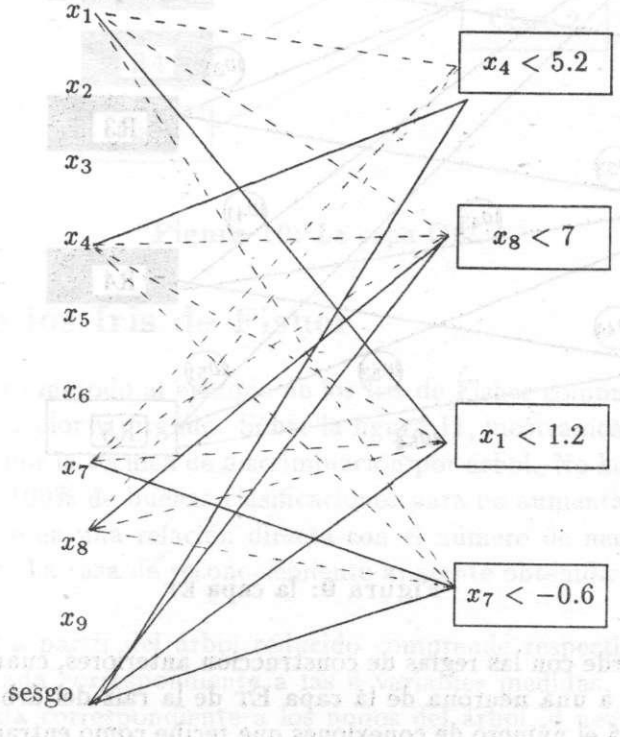


Figura 8: La capa PARTICIONAMIENTO

Se puede aumentar el número de conexiones activando las conexiones que ligan las variables asociadas a los nodos anteriores, con esta neurona. Esta escogencia es difícil de tomar si una variable ha sido seleccionada varias veces.

5.2 Conexiones de la capa ET

Las conexiones unen la primera con la segunda capa escondida. Las neuronas de la primera capa representan los nodos del árbol de decisión, es decir, los segmentos no terminales, y las neuronas de la segunda capa, los segmentos terminales del árbol. A éstos se les asignará una clase a priori utilizando la regla bayesiana. La asignación de una observación a uno de estos segmentos terminales se obtiene recorriendo el árbol; este recorrido es una verificación en serie de un conjunto de condiciones redescubiertas sobre cada nodo del árbol. Además, las conexiones por seleccionar deben simular este recorrido. Así cada neurona de la segunda capa está ligada a las neuronas de la primera capa, asociadas a los nodos anteriores por su segmento terminal.

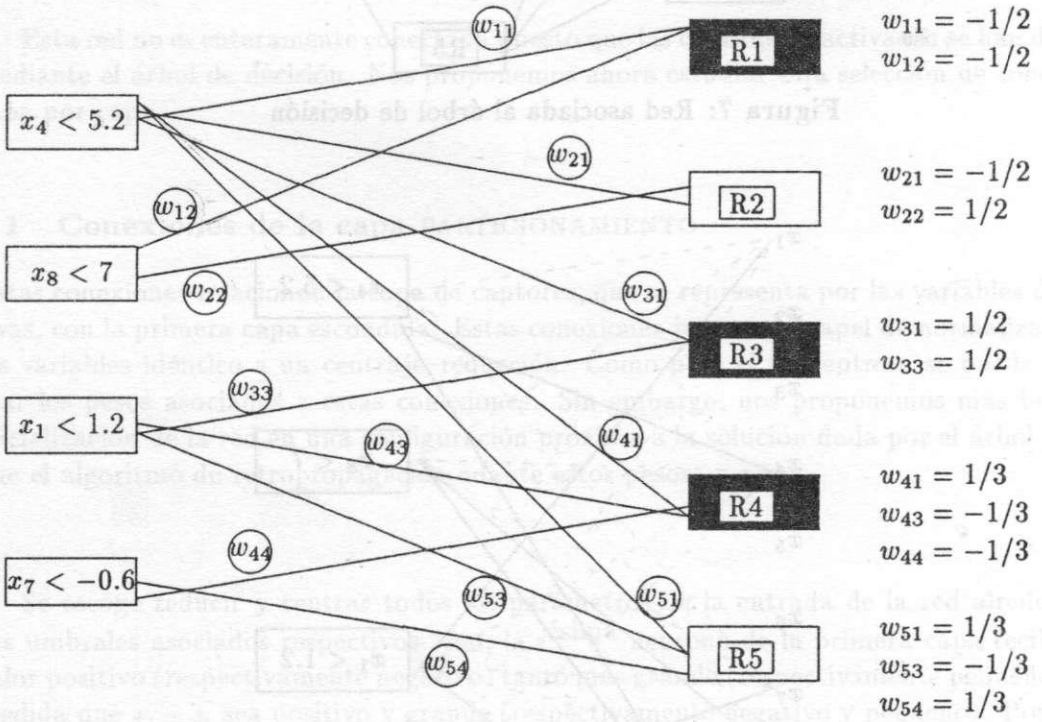


Figura 9: la capa ET

Sin embargo, acorde con las reglas de construcción anteriores, cuanto más grande sea la distancia que separa a una neurona de la capa ET de la raíz del árbol de decisión correspondiente, mayor será el número de conexiones que recibe como entrada. Una escogencia al azar de los pesos, al inicio del aprendizaje, favorece las neuronas que reciben una cantidad

importante de conexiones. Además, las conexiones caracterizan dos tipos de descenso en el árbol. El descenso a la izquierda del árbol corresponde a una respuesta positiva a la pregunta, el descenso a la derecha corresponde a una respuesta negativa. En la inicialización se efectúa el valor -1 con una conexión que representa un descenso a la izquierda, el valor $+1$ con una conexión que representa un descenso a la derecha. Luego se normalizan estas ponderaciones de manera que la suma de los valores absolutos de estas ponderaciones sea igual a uno para cada neurona de la segunda capa escondida (figura 9).

5.3 Conexiones de la capa OU

Las conexiones unen la segunda capa escondida con las clases a priori. Las conexiones seleccionadas son aquéllas que unen a la clase a priori, las neuronas asociadas a los segmentos terminales asignados a esta clase a priori. Aparentemente esta capa de conexiones no es muy útil, puesto que no sirve más que para implementar una función máxima entre las neuronas de la capa anterior asociadas a la misma clase a priori y esto, por sumas ponderadas, por tanto de una forma poco eficiente. Nos parece más fácil escoger directamente la neurona más activada de la capa OU y entonces amputar la red de esta última capa de conexiones.

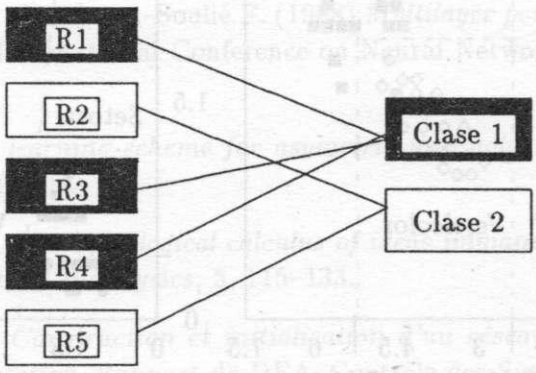


Figura 10: La capa OU

6 Ejemplo de los Iris de Fisher

Hemos aplicado nuestro método al ejemplo de los Iris de Fisher compuesto de tres especies diferentes: setosa, versicolor y virginia. Sobre la figura 11, mostramos las fronteras (líneas punteadas) obtenidas por la técnica de discriminación por árbol. No hemos intentado hacer una separación con el 100% de buenas clasificaciones para no aumentar la profundidad del árbol, puesto que crece en una relación directa con el número de neuronas por capa y el número de conexiones. La tasa de reconociemiento aparente obtenida es de 93,33%.

La red construida a partir del árbol reducido comprende respectivamente 4 neuronas sobre la capa de entrada correspondiente a las 4 variables medidas, 3 neuronas sobre la primera capa escondida correspondiente a los nodos del árbol, 4 neuronas de la segunda capa escondida correspondiente a los 4 nodos terminales y 3 neuronas sobre la capa de salida que caracteriza las tres especies a reconocer.

Con una red enteramente conectada que posea el mismo número de neuronas que la red anteriormente construida, se necesitan unas 200 iteraciones para alcanzar una tasa de reconocimiento de 93,33%. El peso inicial de las conexiones se toma al azar. Con nuestro método, una tasa equivalente se obtiene con solamente unas 30 iteraciones. Las fronteras se muestran sobre la figura 11: se constata que el tiempo de aprendizaje es considerablemente reducido. En las dos redes, varios ensayos fueron necesarios para fijar el paso del gradiente y la pendiente de la sigmoide.

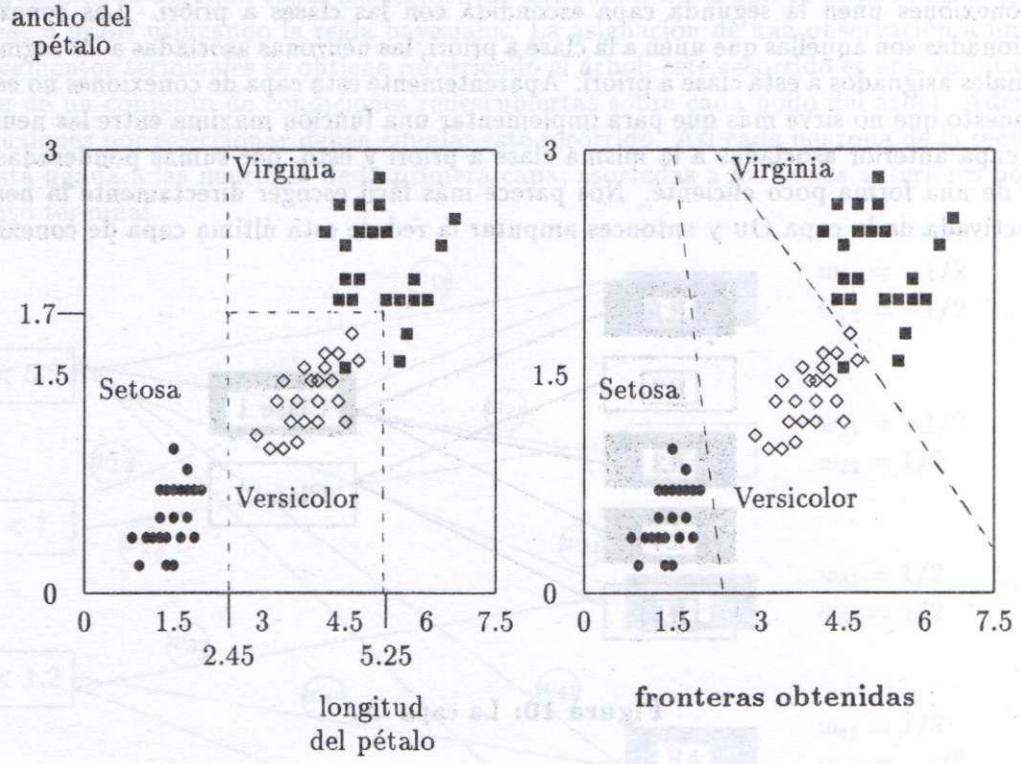


Figura 11: Iris de Fisher

7 Conclusión

Explotando las ideas del artículo de Sethi, hemos desarrollado un método que permite construir automáticamente una red de neuronas a partir de un árbol de decisión. El tamaño, la estructura y la inicialización de los pesos de las conexiones se determinan a partir del árbol. El cálculo de las conexiones reposa sobre un algoritmo de tipo retropropagación del gradiente. Esta escogencia permite reducir considerablemente el tiempo de aprendizaje. Otros ejemplos, no desarrollados en este artículo, muestran que se puede esperar una ganancia de 5% aproximado de la tasa de reconocimiento alcanzada por una técnica de árbol de decisión.

Bibliografía

[1] Breiman L., Friedman J.H., Ohlsen R.A. y Stone C.J. (1984) *Classification and Regression Trees.* Wadsworth.

[2] Celeux G., Lechevallier Y. (1982) *Méthodes de segmentation non paramétriques.* Revue de Statistique Appliquée, 30(4): 39–53.

[3] Celeux G., Diday E., Govaert G., Lechevallier Y., Ralambondrainy, H. (1989) *Classification Automatique des Donneés: Environnement Statistique et Informatique.* Dunod, París.

[4] Celeux G. et al. (1990) *Analyse Discriminante sur Variables Continues.* Collection Didactique INRIA, Rocquencourt.

[5] Dubuisson B. (1990) *Diagnostic et Reconnaissance de Formes*, Hermès, París.

[6] Friedman J.H. (1977) *A recursive partitionning decision rule for non parametric classification.* IEEE Trans on Comp., C 26–4: 404–408.

[7] Gallinari P., Thiria S., Fogelman–Soulié F. (1988) *Multilayer perceptron and data analysis. Second Annual International Conference on Neural Networks.* San Diego, I, 391–401.

[8] Le Cun Y. (1985) *A learning scheme for asymetric threshold network*, Cognitiva 85, Eds. Cesta–Afcet, 599–604.

[9] Mc Culloch, Pitts W. (1943) *A logical calculus of ideas immanent in nervous activity.* Bulletin of Mathematical Biophysics, 5, 115–133.

[10] Millemann S. (1991) *Construction et initialisation d'un réseau de neurones à partir d'un arbre de segmentation.* Rapport de DEA–Contrôle des Systèmes, UTC, 1991.

[11] Mingers J. (1989) *An empirical comparison of pruning methods for decision tree induction.* Machine Learning, 227–243.

[12] Ministère de la santé. *SESI–Enquête toxicomanie*, Nov. 1989.

[13] Morgan J.N., Sonquist J.A. (1963) *Problems in the analysis of survey data and a proposal.* J. Amer. Statist. Assoc., 58:415–435.

[14] Quinlan J.R. (1983) *Learning efficient classification procedures and their applications to chess and games.* Machine Learning: An Artificial Intelligence Approach, Morgan Kaufman.

[15] Rosenblatt F. (1958) *The Perceptron: a probabilistic model for information storage and organisation in brain*, Psychological Review, 65, 386–408.

[16] Rumelhart D., Hinton, G.E., Williams R.J. (1986) *Learning internal representations by error retropropagation*, Parallel Distributed Processing: Explorations in the Micro–Structure of Cognition, The MIT Press, Cambridge Mass.

[17] Sethi, I.K. (1990) *Entropy nets: from decision trees to neural networks.* Proceedings of IEEE, Vol. 78.

[18] Thiria, S. (1992) *Statistique et réseaux de neurones: points de recontre dans la recherche et dans les résultats*, Séminaire Analyse des Données, BURO–AFCET.

El Análisis Discriminante

Gilbert Saporta*

El objetivo de los métodos de discriminación consiste en predecir una variable cualitativa con k categorías con la ayuda de p predictores, generalmente numéricos.

Se puede considerar el análisis discriminante como una extensión del problema de regresión en el caso donde la variable a explicar es cualitativa; veremos por otra parte que en el caso de dos categorías, podemos referirnos exactamente a una regresión lineal múltiple.

Los datos están constituidos por n observaciones repartidas en k clases y descritas por p variables explivativas.

Se distinguen clásicamente dos aspectos en análisis discriminante:

a) **descriptivo:** buscar cuales son las combinaciones lineales de variables que permiten separar lo mejor posible las k categorías y dar una representación gráfica (así como en análisis factorial), que da cuenta lo mejor posible de esta separación;

b) **decisional:** un nuevo individuo se presenta para el que se conocen los valores de los predictores. Se trata entonces de decidir en cual categoría hay que asignarlo. Es un problema de clasificación (pero no en el sentido de clasificación automática).

Estos dos aspectos corresponden, *grosso modo,* a la distinción entre métodos geométricos y métodos probabilísticos.

Entre las innumerables aplicaciones del análisis discriminante citamos algunos campos:

- ayuda a la decisión en medicina: a partir de medidas de laboratorio, se busca una función que permita predecir lo mejor posible el tipo de afección de un enfermo, o su evolución probable con el fin de orientar el tratamiento;

- meteorología: prevención de avalanchas a partir de variables ligadas a la atmósfera y a la nieve;

- finanzas: prevención del comportamiento de los demandantes de crédito.

*Département de Mathématiques et Informatique, Centre National d'Arts et Métiers, París

1 Métodos geométricos

Estos métodos, esencialmente descriptivos, se basan solamente en los conceptos de distancia y no hacen intervenir hipótesis probabilísticas.

1.1 Datos y notaciones

Los n individuos e_i de la muestra constituyen una nube E de $\mathbb{R}^p$ dividida en k subnubes $E_1, E_2, \ldots, E_k$ con centros de gravedad $g_1, g_2, \ldots, g_k$ y matrices de varianzas $V_1, V_2, \ldots, V_k$ (figura 1).

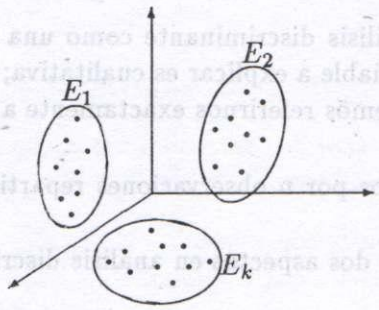


Figura 1

Sea g el centro de gravedad y V la matriz de varianza de E. Si a los n individuos $e_1, e_2, \ldots, e_n$ les son afectados los pesos $p_1, p_2, \cdots, p_n$, entonces los pesos $q_1, q_2, \cdots, q_k$ de cada subnube son entonces

$$q_j = \sum_{e_i \in E_j} p_i$$

Se tiene:

$$g_j = \frac{1}{q_j} \sum_{e_i \in E_j} p_i e_i$$

$$g = \sum_{j=1}^{k} q_j g_j \quad \text{y} \quad V_j = \frac{1}{q_j} \sum_{e_i \in E_j} p_i (e_i - g_j)(e_i - g_j)^t$$

Llamemos *matriz de varianza interclase*, la matriz de varianza B de los k centros de gravedad provistos de los pesos q_j:

$$B = \sum_{j=1}^{k} q_j (g_j - g)(g_j - g)^t$$

y matriz de varianza intraclase W, el promedio de las matrices V_j:

$$W = \sum_{j=1}^{k} q_j V_j$$

En general, W es invertible mientras que B no lo es, porque los k centros de gravedad están en un subespacio de dimensión $k-1$ de $\mathbb{R}^p$ (si $p > k-1$ lo que es generalmente el caso), entonces la matriz B es de tamaño p.

Tenenemos entonces la relación siguiente:

$$V = W + B$$

que se demuestra fácilmente y constituye una generalización de la relación clásica:

varianza total = promedio de las varianzas + varianza de las medias

Supondremos en adelante que $g = 0$, es decir, que las variables explicativas están centradas.

Si consideramos que la tabla de datos a estudiar se pone bajo la forma:

$$[A \mid X]$$

donde X es la matriz de p variables explicativas y A la tabla lógica asociada a la variable cualitativa, los k centros de gravedad $g_1, g_2, \cdots g_k$ son los filas de la matriz $(A^tDA)^{-1}(A^tDX)$.

(A^tDA) es la matriz diagonal de pesos q_j de las subnubes:

$$A^tDA = D_q = \begin{bmatrix} q_1 & & 0 \\ & q_2 & \\ & & \ddots \\ 0 & & q_k \end{bmatrix}$$

La matriz de varianza interclases se escribe entonces, si $g = 0$:

$$\begin{aligned} D &= ((A^tDA)^{-1}A^tDX)^tA^tDA((A^tDA)^{-1}A^tDX) \\ &= X^tDA(A^tDA)^{-1}A^tDX = (X^tDA)D_q^{-1}(A^tDX) \end{aligned}$$

En el caso donde $p_i = \frac{1}{n}$ las expresiones anteriores se simplifican e introduciendo los efectivos $n_1, n_2, \cdots, n_k$ de las k subnubes, tenemos:

$$B = \frac{1}{n}\sum_j n_j g_j g_j^t; \qquad g_j = \frac{1}{n_j}\sum_{e_i \in E_j} e_i; \qquad W = \frac{1}{n}\sum_j n_j V_j$$

Supondremos en adelante estar en este caso.

1.2 El análisis factorial discriminante (AFD)

A. Los ejes y variables discriminantes

El AFD consiste en la búsqueda de nuevas variables (las variables discriminantes) correspondientes a las direcciones de $\mathbb{R}^p$ que separan lo mejor posible en proyección a los k grupos de observaciones.

El eje 1 de la figura 2 posee un buen poder discriminante mientras que el eje 2 (que está en el eje principal usual) no permite separar en proyección los dos grupos.

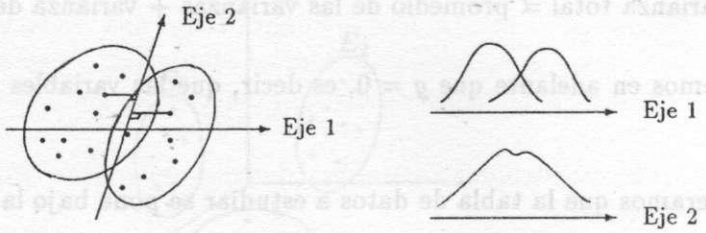


Figura 2

Supongamos que $\mathbb{R}^p$ está dotado de una métrica M. Notaremos a al eje discriminante, u al factor asociado $u = Ma$, la variable discriminante será Xu.

En proyección sobre el eje a, los k centros de gravedad deben ser tan separados como sea posible, mientras que cada subnube debe ser proyectada de manera agrupada alrededor de la proyección de su centro de gravedad.

En otras palabras, la inercia de la nube de los g_j proyectados sobre a debe ser máxima. La matriz de inercia de la nube de los g es MBM y la inercia de la nube proyectada sobre a es a^tMBMa si a es de M-norma igual a 1.

Es necesario también que en proyección sobre a, cada subnube debe ser bien agrupada, y por tanto que a^tMV_jMa sea pequeño para $j = 1, 2, \cdots, k$.

Se buscará por lo tanto minimizar el promedio $\sum_{j=1}^{k} q_j a^tMV_jMa$, sea a^tMWMa.

Ahora bien, la relación $V = B + W$ implica que $MVM = MBM + MWM$ y, por consiguiente:

$$a^tMVMa = a^tMBMa + a^tMWMa$$

Tomaremos entonces como criterio a maximizar, la razón de la inercia interclases entre la inercia total. Es decir,

$$\max_a \frac{a^t MBMa}{a^t MVMa}$$

Sabemos que el máximo se alcanza si a es vector propio de $(MVM)^{-1}MBM$ asociado a su mayor valor propio λ_1:

$$M^{-1}V^{-1}BMa = \lambda a$$

Al eje discriminante a es entonces asociado el factor discriminante u tal que $u = Ma$, de modo que $V^{-1}Bu = \lambda_1 u$.

Los factores discriminantes y, por consiguiente, las variables discriminantes Xu, son independientes de la métrica M. Se escogerá por comodidad $M = V^{-1}$ que da $BV^{-1}a = \lambda a$ y $V^{-1}Bu = \lambda u$.

Se tiene siempre $0 \leq \lambda_1 \leq 1$ pues $0 \leq \dfrac{a^t MBMa}{a^t MVMa} \leq 1$.

$\lambda_1 = 1$ corresponde al caso siguiente: en proyección sobre a las dispersiones intraclases son nulas, las k nubes están por lo tanto en un hiperplano ortogonal a a (ver figura 3).

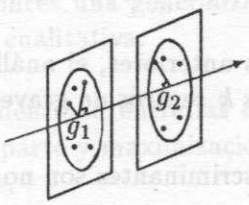


Figura 3

Hay evidentemente discriminación perfecta si los centros de gravedad se proyectaran en puntos diferentes.

$\lambda_1 = 0$ corresponde al caso en que el mejor eje no permite separar los centros de gravedad g_i, es el caso donde están confundidos y las nubes son, por lo tanto, concéntricas y ninguna separación lineal es posible (er figura 4).

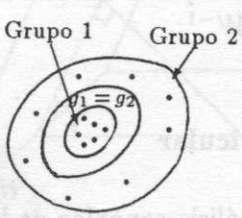


Figura 4

Puede ocurrir, sin embargo, que existe la posibilidad de discriminación no lineal: la distancia al centro permite separar los grupos, pero se trata de una función cuadrática de las variables.

El valor propio λ es una medida pesimista del poder discrimiante de un eje. La figura 5 muestra que se puede dicriminar perfectamente pues los grupos están bien separados a pesar que $\lambda < 1$.

El número de los valores propios no nulos, y por tanto el de ejes discriminantes, es igual a $k-1$ en el caso habitual $n > p > k$ y donde las variables no están ligadas por relaciones lineales.

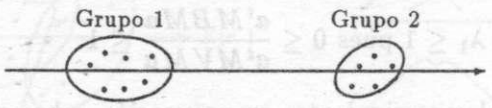


Figura 5

B. Un análisis en componentes principales (ACP) particular

De acuerdo con las ecuaciones anteriores, el análisis factorial discriminante no es otra cosa que el ACP de la nube de los k centros de gravedad con la métrica V^{-1}.

Se deduce que las variables discriminantes son no correlacionadas 2 a 2.

Así como en ACP, se podrá interpretar las variables discriminantes por medio del círculo de las correlaciones.

Representación gráfica

Si existe un segundo eje discriminante, es posible representar la nube de las n observaciones en proyección sobre el plano definido por estos dos ejes, este plano es entonces el que permite visualizar mejor la separación de las observaciones en clases.

Nosotros veremos más adelante que el análisis factorial discriminante equivale también al ACP de los g_i con la métrica W^{-1}.

C. Un análisis canónico particular

El análisis discriminante es el análisis canónico de las tablas A y X.

En efecto, la ecuación del análisis canónico de A y X que da las variables canónicas asociadas a X, se escribe:

$$(X^tDX)^{-1}X^tDA(A^tDA)^{-1}A^tDXu = \lambda u$$

esto es idéntico a $V^{-1}Bu = \lambda u$ de acuerdo con el párrafo 1. Esto es una nueva prueba de que las variables discriminantes son no correlacionadas dos a dos.

Si se designa por Aa la primera variable canónica asociada a A, solución de la otra ecuación del análisis canónico:

$$(A^tDA)^{-1}A^tDX(X^tDX)^{-1}X^tDAa = \lambda a$$

normada de tal forma que su proyección sobre el subespacio de $\mathbb{R}^n$ generado por las p variables explicativas sea indéntica a Xu, se puede presentar al análisis discriminante como la búsqueda de la codificación de la variable cualitativa que la vuelve más próxima del espacio generado por las columnas de X. Si las p variables explicativas son centradas, entonces la variable codificada lo es también y u es el vector de los coeficientes de regresión de Aa sobre X.

El primer valor propio λ_1 es entonces el cuadrado del coeficiente de correlación múltiple.

El análisis discriminante es entonces una generalización de la regresión múltiple en el caso donde la variable a explicar es cualitativa.

La figura 6 en $\mathbb{R}^n$ muestra la identidad entre las dos concepciones del análisis discriminante: análisis canónico por una parte y maximización de la varianza interclase dividida por la varianza total, por otra parte.

W_X es el espacio generado por las columnas de X; W_A es el espacio generado por las indicatrices de la variable a explicar.

Si se proyecta D-ortogonalmente la variable discriminante ξ sobre W_A en Aa, el teorema de Pitágoras se escribe:

$$\|\xi\|^2 = \|Aa\|^2 + \|Aa - \xi\|^2$$

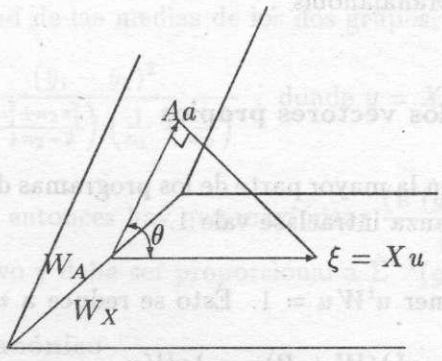


Figura 6

Varianza total de ξ = varianza interclase + varianza intraclase.

La maximización del cociente $\dfrac{\text{varianza interclase}}{\text{varianza total}}$ no es otra cosa que la maximización de $\cos^2\theta$, donde θ es el ángulo formado por Aa y ξ, lo que es el criterio del análisis canónico.

Los autores de habla inglesa llaman a este método **análisis discriminante canónico.**

D. Análisis de varianza y métrica W^{-1}

Si solamente hubiera una variable explicativa, se mediría la eficacia de su poder separador sobre la variable de grupo por medio de un análisis de varianza ordinario con un factor. La estadística F valdría entonces $\dfrac{\text{varianza inter}/k-1}{\text{varianza intra }/n-k}$.

Como hay p variables se puede buscar la combinanción lineal definida por unos coeficientes u dando el valor maximal para la estadística de test, lo cual se reduce a maximizar:

$$\frac{u^tBu}{u^tWu}$$

La solución es dada por la ecuación:

$$W^{-1}Bu = \mu u \quad \text{con } \mu \text{ maximal}$$

Los vectores propios de $W^{-1}B$ son los mismos que los de $V^{-1}B$ con $\mu = \dfrac{\lambda}{1-\lambda}$.

En efecto, $Bu = \lambda Vu$ es equivalente a:

$$Bu = \lambda(W+B)u, \quad \text{es decir} \quad (1-\lambda)Bu = \lambda Wu$$

de donde $W^{-1}Bu = \dfrac{\lambda u}{1-\lambda}$.

Si $0 \leq \lambda \leq 1$ se tiene en cambio $0 \leq \mu \leq \infty$ y $\lambda = \dfrac{\mu}{1+\mu}$.

La utilización de V^{-1} o de W^{-1} como métrica es por tanto indiferente. La métrica W^{-1} es llamada "métrica de Mahalanobis".

E. Normalización de los vectores propios

La convención usual en la mayor parte de los programas de computador es tener variables discriminantes cuya varianza intraclase vale 1.

Se debe por tanto tener $u^tWu = 1$. Esto se reduce a $u^tBu = \dfrac{\lambda}{1-\lambda} = \mu$ y a $u^tVu = \dfrac{1}{1-\lambda}$ puesto que $u^tBu = u^t\lambda(W+B)u = \lambda u^tVu$.

1.3 El caso de dos grupos

A. La función de Fisher

Solamente hay una variable discriminante puesto que $k - 1 = 1$.

El eje discriminante es entonces necesariamente la recta que une los dos centros de gravedad g_1 y g_2:

$$a = (g_1 - g_2)$$

La variable discriminante d se obtiene entonces de la proyección sobre a según la métrica V^{-1} o W^{-1} que tiene en cuenta la orientación de las nubes respecto a la recta de los centros.

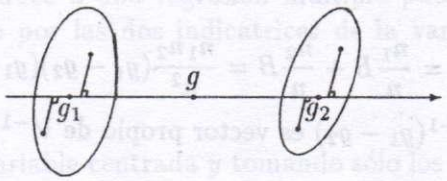


Figura 7

El factor discriminante u vale por tanto:

$$u = V^{-1}(g_1 - g_2) \quad \text{o} \quad u = W^{-1}(g_1 - g_2)$$

que son proporcionales.

$W^{-1}(g_1 - g_2)$ es la función de Fisher (1936)

Para razones de estimación habitualmente no se toma W^{-1}, sino:

$$\frac{n_1 + n_2 - 2}{n_1 + n_2} W^{-1}$$

Se puede en efecto reencontrar el procedimiento de Fisher por el razonamiento siguiente: se busca la combinanción lineal de las variables explicativas tales que el cuadrado de la estadística del test T de igualdad de las medias de los dos grupos, toma un valor maximal:

$$\max \frac{(\bar{y}_1 - \bar{y}_2)^2}{\left(\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad \text{donde } y = Xu$$

si tomamos $\hat{\Sigma} = \dfrac{n_1 + n_2}{n_1 + n_2 - 2} W$ entonces hay que maximizar $\dfrac{(u^t(g_1 - g_2))^2}{u^t \hat{\Sigma} u}$. u es definido salvo por un factor multiplicativo y debe ser proporcional a $\hat{\Sigma}^{-1}(g_1 - g_2)$.

B. Aplicación del análisis canónico

Se puede encontrar el único valor propio de $V^{-1}B$ observando que para dos grupos:

$$B = \frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)^t$$

En efecto: $B = \frac{n_1}{n} g_1 g_1^t + \frac{n_2}{n} g_2 g_2^t$; ahora bien:

$$g = \frac{n_1}{n} g_1 + \frac{n_2}{n} g_2 = 0$$

De donde $B = \frac{n_1}{n} g_1 g_1^t - \frac{n_1}{n} g_1 g_2^t = \frac{n_1}{n} g_1 (g_1^t - g_2^t)$ y simétricamente:

$$B = \frac{n_2}{n} g_2 (g_1^t - g_2^t)$$

luego:

$$B = \frac{n_1}{n} B + \frac{n_2}{n} B = \frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)^t$$

Se verifica que $u = V^{-1}(g_1 - g_2)$ es vector propio de $V^{-1}B$:

$$V^{-1} \frac{n_1 n_2}{n^2}(g_1 - g_2)(g_1 - g_2)^t V^{-1}(g_1 - g_2) = \lambda V^{-1}(g_1 - g_2)$$

con:

$$\lambda = \frac{n_1 n_2}{n^2}(g_1 - g_2)^t V^{-1}(g_1 - g_2)$$

y:

$$\mu = \frac{\lambda}{1 - \lambda} = \frac{n_1 n_2}{n^2}(g_1 - g_2)^t W^{-1}(g_1 - g_2)$$

μ es, por lo tanto, proporcional a la D_p^2 de Mahalanobis estimada entre los dos grupos ([1], capítulo 15).

Se tiene exactamente:

$$\mu = \frac{n_1 n_2}{n(n-2)} D_p^2 \text{ pues } D_p^2 = \frac{n-2}{n}(g_1 - g_2)^t W^{-1}(g_1 - g_2)$$

Se puede hallar entonces:

$$W^{-1}(g_1 - g_2) = \left(1 + \frac{n_1 n_2}{n(n-2)} D_p^2\right) V^{-1}(g_1 - g_2)$$

El uso del convenio de normalización $u^t W u = 1$ presenta la ventaja siguiente:

Las coordenadas de los dos centros de gravedad sobre el eje discriminante tienen una diferencia igual a la distancia de Mahalanobis D_p.

En efecto, $g_1^t u$ y $g_2^t u$ son estas coordenadas donde u es el factor canónico normalizado. éste es proporcional a $W^{-1}(g_1 - g_2)$, la constante de proporcionalidad α es tal que $u^t W u = 1$, es decir:

$$[\alpha W^{-1}(g_1 - g_2)]^t W [\alpha W^{-1}(g_1 - g_2)] = \alpha^2 (g_1 - g_2)^t W^{-1}(g_1 - g_2)$$

Despreciando la corrección por $\frac{n}{n-2}$ (o utilizando $\hat{\Sigma}$ en lugar de W) sigue que $|\alpha| = \frac{1}{D_p}$.

De donde:

$$|g_1^t u - g_2^t u| = |(g_1 - g_2)^t u| = |\alpha|(g_1 - g_2)^t W^{-1}(g_1 - g_2) = \frac{D_p^2}{D_p} = D_p$$

C. Equivalencia con una regresión múltiple

El análisis canónico se reduce a una regresión múltiple puesto que después de haber centrado, el espacio generado por las dos indicatrices de la variable de los grupos es de dimensión 1.

Es suficiente definir una variable centrada y tomando sólo los dos valores a y b sobre los grupos 1 y 2 respectivamente ($n_1 a + n_2 b = 0$).

Se obtendrá entonces un vector de coeficientes de regresión proporcional a la función de Fisher para cualquier escogencia de a.

La escogencia $a = \frac{n}{n_1}$, $b = -\frac{n}{n_2}$ conduce entonces a $b = (X^t X)^{-1} X^t y = V^{-1}(g_1 - g_2)$.

Se tiene:

$$R^2 = \frac{D_p^2}{\frac{n(n-2)}{n_1 n_2} + D_p^2}$$

Se tendrá cuidado en el hecho que las hipótesis habituales de la regresión no son verificables. Por el contrario; aquí y no es aleatorio y X sí lo es. Las estadísticas usuales que provee un programa de regresión, se deben utilizar únicamente a título indicativo.

D. Un ejemplo

Los datos de la tabla 1 (suministrados por J. P. Nakache) concernientes a 101 víctimas de infartos del miocardo (51 fallecieron, 50 sobrevivieron) sobre los cuales han sido medidos en el momento de la admisión, 7 variables (frecuencia cardiaca, índice cardiaco, índice sistólico, presión diastólica, presión arterial pulmonar, presión ventricular y resistencia pulmonaria). La tabla 2 da las estadísticas elementales por grupo.

Tabla 1

FRCAR	INCAR	INSYS	PRDIA	PAPUL	PVENT	REPUL	PRONOSTICO
90	1.71	19.0	16	19.5	16.0	912	supervivencia
90	1.68	18.7	24	31.0	14.0	1476	fallecimiento
120	1.40	11.7	23	29.0	8.0	1657	fallecimiento
82	1.79	21.8	14	17.5	10.0	782	supervivencia
80	1.58	19.7	21	28.0	18.5	1418	fallecimiento
80	1.13	14.1	18	23.5	9.0	1664	fallecimiento
94	2.04	21.7	23	27.0	10.0	1059	supervivencia
80	1.19	14.9	16	21.0	16.5	1412	supervivencia
78	2.16	27.7	15	20.5	11.5	759	supervivencia
100	2.28	22.8	16	23.0	4.0	807	supervivencia
90	2.79	31.0	16	25.0	8.0	717	supervivencia
86	2.70	31.4	15	23.0	9.5	681	supervivencia
80	2.61	32.6	8	15.0	1.0	460	supervivencia
61	2.84	47.3	11	17.0	12.0	479	supervivencia
99	3.12	31.8	15	20.0	11.0	513	supervivencia
92	2.47	26.8	12	19.0	11.0	615	supervivencia
96	1.88	19.6	12	19.0	3.0	809	supervivencia
86	1.70	19.8	10	14.0	10.5	659	supervivencia
125	3.37	26.9	18	28.0	6.0	665	supervivencia
80	2.01	25.0	15	20.0	6.0	796	supervivencia
82	3.15	38.4	13	20.0	6.0	508	supervivencia
110	1.66	15.1	23	31.0	6.5	1494	fallecimiento
80	1.50	18.7	13	17.0	12.0	907	fallecimiento
118	1.03	8.7	19	27.0	10.0	2097	fallecimiento
95	1.89	19.9	25	27.0	20.0	1143	fallecimiento
80	1.45	17.1	19	23.0	15.0	1269	fallecimiento
85	1.30	15.1	13	18.0	10.0	1108	fallecimiento
105	1.84	17.5	18	22.0	10.0	957	fallecimiento
122	2.79	22.9	25	36.0	10.0	1032	supervivencia
81	1.77	21.9	18	27.0	11.0	1220	supervivencia
118	2.31	19.6	22	27.0	10.0	935	supervivencia
87	1.20	13.8	34	41.0	20.0	2733	fallecimiento
65	1.19	18.3	15	18.0	13.0	1210	fallecimiento
84	2.15	25.6	27	37.0	10.0	1377	supervivencia
103	0.91	8.8	30	33.5	10.0	2945	fallecimiento
75	2.54	33.9	24	31.0	16.0	976	supervivencia
90	2.08	23.1	20	28.0	6.0	1077	supervivencia
90	1.93	21.4	11	18.0	10.0	746	supervivencia
90	0.95	10.6	20	24.0	6.0	2021	fallecimiento
65	2.38	36.6	16	22.0	12.0	739	supervivencia
95	0.99	10.4	20	27.5	8.0	2222	fallecimiento
95	0.85	8.9	19	22.0	15.5	2071	fallecimiento
86	2.05	23.8	21	28.0	10.0	1093	supervivencia
82	2.02	24.6	16	22.0	14.0	871	supervivencia
70	1.44	20.6	19	26.5	11.0	1472	fallecimiento
92	3.06	33.3	10	15.0	6.0	392	supervivencia

FRCAR	INCAR	INSYS	PRDIA	PAPUL	PVENT	REPUL	PRONOSTICO
94	1.31	13.9	26	40.0	15.0	2443	fallecimiento
79	1.29	16.3	24	31.0	10.0	1922	fallecimiento
67	1.47	21.9	15	18.0	16.0	980	supervivencia
75	1.21	16.1	19	24.0	4.0	1587	fallecimiento
80	2.41	30.9	19	24.0	7.0	797	supervivencia
61	3.28	54.0	12	16.0	7.0	390	supervivencia
110	1.24	11.3	22	27.5	11.0	1774	fallecimiento
116	1.85	15.9	33	42.0	13.0	1816	fallecimiento
75	2.00	26.7	16	22.0	5.0	880	supervivencia
92	1.97	21.4	18.0	27.0	3.0	1096	fallecimiento
110	0.96	8.8	15.0	19.0	16.0	1583	supervivencia
95	2.56	26.9	8.0	13.0	3.0	406	supervivencia
75	2.32	30.9	8.0	10.0	6.00	345	supervivencia
80	2.65	33.1	13.0	19.0	9.0	574	supervivencia
102	1.60	15.7	24.0	31.0	16.0	1550	fallecimiento
86	1.67	19.4	18.0	23.0	8.5	1102	supervivencia
60	0.82	13.7	22.0	32.0	13.05	3122	fallecimiento
100	1.76	17.6	23.0	33.0	2.0	1500	supervivencia
80	3.28	41.0	12.0	17.0	2.0	415	supervivencia
108	2.96	27.4	24.0	35.0	6.5	946	supervivencia
92	1.37	14.8	25.0	46.0	11.0	2686	fallecimiento
100	1.38	13.8	20.0	31.0	11.0	1797	fallecimiento
80	2.85	35.6	25.0	32.0	7.0	898	supervivencia
87	2.51	28.8	16.0	24.0	20.0	765	fallecimiento
100	2.31	23.1	8.0	12.0	0.0	416	supervivencia
120	1.18	9.9	25.0	36.0	8.0	2441	fallecimiento
115	1.83	15.9	25.0	30.0	8.0	1311	fallecimiento
101	2.55	25.2	23.2	30.5	9.0	957	supervivencia
100	2.31	23.1	8.0	12.0	0.0	416	supervivencia
120	1.18	9.9	25.0	36.0	8.0	2441	fallecimiento
115	1.83	15.9	25.0	30.0	8.0	1311	fallecimiento
101	2.55	25.2	23.2	30.5	9.0	957	supervivencia
92	2.17	23.5	19.0	24.0	3.0	885	supervivencia
87	1.42	16.1	20.0	26.0	10.0	1465	fallecimiento
80	1.59	19.9	13.0	20.5	4.0	1031	supervivencia
88	1.47	16.7	23.0	32.5	10.0	1769	fallecimiento
104	1.23	11.8	27.0	33.0	11.0	2146	fallecimiento
90	1.45	16.1	17.0	24.0	8.5	1324	supervivencia
67	0.85	12.7	26.0	33.0	11.0	3106	fallecimiento
87	2.37	27.2	15.0	22.0	10.0	743	supervivencia
108	2.40	22.2	26.0	31.0	4.0	1033	supervivencia
120	1.91	15.9	18.0	27.0	15.0	1131	fallecimiento
108	1050	13.9	28.0	43.0	16.0	1813	fallecimiento
86	2.36	27.4	24.0	34.0	8.0	1153	supervivencia
112	1.56	13.9	24.0	29.0	4.0	1487	fallecimiento
80	1.34	17.0	16.0	25.0	16.0	1493	fallecimiento
95	1.65	17.4	20.0	33.0	7.0	1600	fallecimiento
90	2.04	22.7	28.0	41.0	10.0	1608	fallecimiento

FRCAR	INCAR	INSYS	PRDIA	PAPUL	PVENT	REPUL	PRONO
90	2.04	22.7	28.0	41.0	10.0	1608	fallecimiento
90	3.03	33.6	17.0	23.5	7.0	620	supervivencia
94	1.21	12.9	17.0	22.0	3.0	1455	fallecimiento
51	1.34	26.3	11.0	17.0	6.0	1015	fallecimiento
110	1.17	10.6	29.0	35.0	10.5	2393	fallecimiento
96	1.74	18.1	24.0	29.0	6.0	1333	fallecimiento
132	1.31	9.9	23.0	28.0	12.0	1710	fallecimiento
135	0.95	7.0	15.0	20.0	7.0	1684	fallecimiento
105	1.92	18.3	18.0	24.0	3.0	1000	fallecimiento
99	0.83	8.4	23.0	27.0	8.0	2602	fallecimiento
116	0.60	5.2	33.0	38.0	10.0	5067	fallecimiento
112	1.54	13.8	25.0	31.0	8.0	1610	fallecimiento

Un análisis factorial discriminante entre los sobrevivientes y los fallecidos dan los resultados siguientes: la distancia de Mahalanobis al cuadrado vale:

$$D_7^2 = 4.942 \text{ de donde } D_7 = 2.223$$

Tabla 2

Variable	N	Media	Desviación estándar
FRCAR	51	95.90196078	17.97693511
INCAR	51	1.39470588	0.37619332
INSYS	51	14.99607843	4.63900682
PRDIA	51	21.09803922	5.14183152
PAPUL	51	29.09803922	6.81910523
PVENT	51	10.64705882	4.34429985
REPUL	51	1797.27450980	739.87296419
FRCAR	50	88.34000000	13.84109527
INCAR	50	2.30580000	0.56055035
INSYS	50	26.75200000	8.08319597
PRDIA	50	16.50400000	5.15304388
PAPUL	50	22.84000000	6.46532352
PVENT	50	8.33000000	4.05398519
REPUL	50	841.38000000	303.68256050

Bajo las hipótesis de multinormalidad ([1], capítulo 15), es el valor correspondiente a $F = 16476$:

$$\frac{n_1 n_2}{n} \frac{n-p-1}{p(n-2)} D_p^2 = F$$

El valor crítico al 1% para un valor $F(7;93)$ es 2.84, el D^2 es significativo de una diferencia neta entre los dos grupos.

Se halla $R^2 = \lambda = 0.5576$ y $\mu = 1.2604$.

La variable discriminante se obtiene entonces por la combinación lineal de las 7 variables centradas sobre la media general de los dos grupos (var tabla 3).

Tabla 3

FRCAR	-0.026445290
INCAR	2.768181397
INSYS	-0.075037835
PRDIA	0.009115031
PAPUL	-0.074211897
PVENT	-0.021086258
REPUL	0.000084078

Si no se centra se suma la constante 1.22816 a la combinación lineal anterior de los datos brutos.

Los coeficientes de correlación lineales de la variable discriminante con las 7 variables (los dos grupos confundidos), son idénticos sobre la tabla 4.

Tabla 4

FRCAR	-0.3097
INCAR	0.9303
INSYS	0.8976
PRDIA	-0.6321
PAPUL	-0.5751
PVENT	-0.3591
REPUL	-0.8676

Las medias de los dos grupos sobre las variables discriminantes son:

Fallecimientos -1.1005

Sobrevivencia 1.1225

Se vuelve a encontrar $D_7 = +1.1005 + 1.1225 = 2.2230$

1.4 Reglas geométricas de asignación

Habiendo encontrado la mejor representación de la separación en k clases de n individuos, se puede entonces intentar asignar una observación e a uno de los grupos.

La regla natural consiste en calcular las distancias de la observación a clasificar, a cada uno de los k centros de gravedad y asignar según la distancia más pequeña. Falta definir la métrica que se va a utilzar.

A. Regla de Mahalanobis–Fisher

Consiste en utilizar la métrica W^{-1} (o V^{-1} que es equivalente):

$$d^2(e, g_i) = (e - g_i)^t W^{-1}(e - g_i)$$

Desarrollando esta cantidad se encuentra:

$$d^2(e, g_i) = e^t W^{-1} e + g_i^t W^{-1} g_i - 2e^t W^{-1} g_i$$

Puesto que $e^t W^{-1} e$ no depende del grupo i, la regla consiste en buscar el mínimo de $g_i^t W^{-1} g_i - 2e^t W^{-1} g_i$ o el máximo de $e^t W^{-1} g_i - \frac{g_i^t W^{-1} g_i}{2}$.

Se ve que esta regla es lineal respecto a las coordenadas de e.

Se debe calcular para cada individuo, k funciones lineales de sus coordenadas y buscar el valor maximal.

La tabla 5 da para el ejemplo de los infartos las dos funciones de clasificación.

Tabla 5

	deceso	supervivencia
CONSTANTE	-91.57481116	-89.97134555
FRCAR	1.53609883	1.47730875
INCAR	-52.09444392	-45.94054613
INSYS	5.44165359	5.27483824
PRDIA	-0.64815662	-0.62789315
PAPUL	0.70738671	0.54240748
PVENT	0.85037707	0.80350057
REPUL	0.00638975	0.00657667

En el caso de los dos grupos se decidirá asignar al grupo 1 si:

$$e^t W^{-1} g_1 - \frac{1}{2}(g_1^t W^{-1} g_1) > e^t W^{-1} g_2 - \frac{1}{2} g_2^t W^{-1} g_1$$

o sea

$$e^t W^{-1}(g_1 - g_2) > \frac{1}{2}(g_1 + g_2)^t W^{-1}(g_1 - g_2)$$

Puesto que $W^{-1}(g_1 - g_2)$ es la función de Fisher, la regla consiste en asignar al grupo 1 si el valor de la función discriminante es superior al umbral:

$$\frac{1}{2}(g_1 + g_2)^t W^{-1}(g_1 - g_2)$$

Cuando los dos grupos son de igual efectivo $g_1 + g_2 = 0$; se asigna al grupo 1 si la función $e^t W^{-1}(g_1 - g_2)$ es positiva.

En el ejemplo de los infartos es poco más o menos el caso puesto que $n_1 = 50$ y $n_2 = 51$. Se predecirá la supervivencia si la función discriminante es positiva.

Observemos que la aplicación de la regla geométrica se puede hacer indiferentemente en el espacio $\mathbb{R}^p$ o en el espacio factorial $\mathbb{R}^{k-1}$.

En particular si $k = 3$, las fronteras de asignación a los grupos son hiperplanos ortogonales al plano de los tres centros de gravedad. Se pueden leer directamente las distancias de Mahalanobis a g_1, g_2, g_3 utilizando el gráfico de las dos variables canónicas discriminantes normalizadas a 1 (en el sentido de la varianza interclase).

B. Insuficiencia de las reglas geométricas

La utilización de la regla anterior conduce a asignaciones incorrectas cuando las dispersiones de los grupos son muy diferentes entre ellos, nada justifica entonces el uso de la misma métrica para los diferentes grupos.

En efecto, si se considera la figura 8, aunque e sea más próximo de g_1 que de g_2 en el sentido habitual, es más natural asignar e a la segunda clase que a la primera cuyo "poder de atracción" es más pequeño.

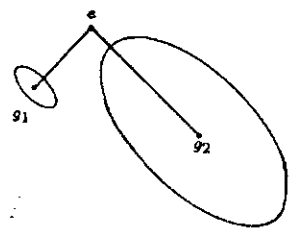


Figura 8

Diversas soluciones utilizan métricas locales M_i tales como:

$$d^2(e, g_i) = (e - g_i)^t M_i (e - g_i)$$

que han sido propuestas, tomando en general, M_i proporcional a V_i^{-1}.

El problema de la optimalidad de la regla de decisión geométrica no puede ser resuelto sin referencia a un modelo probabilístico. En efecto, el problema es de saber cómo se comportará esta regla frente a nuevas observaciones, lo que lleva a hacer hipótesis distribucionales sobre la distribución en el espacio de estas nuevas observaciones. Se alcanzan, por lo tanto, los límites de los métodos descriptivos. Veremos más adelante en qué condiciones ellas conducen a unas reglas optimales.

1.5 Un método de discriminación sobre variables cualitativas: el método Disqual

Cuando los predictores son p variables cualitativas $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$ con $m_1, m_2 \ldots, m_p$ modalidades respectivamente, se puede utilizar el procedimiento siguiente: se efectúa *en una primera etapa el análisis* de correspondencias múltiples de las variables $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_p$, es decir el análisis de correspondencias de la tabla disyuntiva $X = (X_1|X_2 \ldots |X_p)$.

Se reemplazan entonces las p variables cualitativas por las q coordenadas sobre los ejes factoriales y se efectúa luego un análisis factorial discriminante sobre estas q variables numéricas $z_1, z_2, \ldots, z_q$.

Un factor discriminante d es una combinación lineal de los z_j que son combinaciones lineales de las indicatrices de los $\mathcal{X}_i$. Se expresa entonces directamente d como combinación lineal de las indicatrices de los $\mathcal{X}_i$, o lo que se reduce a atribuir a cada categoría de cada variable un valor numérico *puntaje*. d es entonces simplemente igual a la suma de los puntajes obtenidos en las categorías de las p varibles. Esto se reduce a trasformar cada *variable cualitativa en una variable* numérica discreta con m valores.

Cuando $k = 2$ este método es óptimo en el sentido siguiente: tomando todos los factores posibles del Análisis de las Correspondencias Múltiples (ACM) $\left(\sum_i m_i - p\right)$ la cuantificación de las variables $\mathcal{X}_i$ es la que da la distancia de Mahalanobis más grande entre los dos grupos.

En la práctica, sin embargo, se utilizan únicamente los factores que presentan a la vez una inercia y en poder separar entre las clases, suficientes.

Igualmente, se podrían utilizar técnicas parecidas a las expuestas en el capítulo 17 de [1] (modelo lineal general y análisis de varianza), consistentes en anular los coeficientes de ciertas variables indicatrices y efectuar un análisis discriminante sobre $\sum_i m_i - p$ columnas de X.

La ventaja del análisis de correspondencias es proveer además de una descripción de las relaciones entre las variables explicativas, unas componentes ortogonales (el D_q^2 es entonces la suma de los D_1^2 sobre las diversas componentes con la métrica V^{-1}).

En estas mismas memorias, se detalla un poco más este método y se hace una aplicación al puntaje en crédito (*credit-scoring*).

2 Métodos probabilísticos

2.1 La regla Bayesiana

Se supone que los k grupos están en proporción $p_1, p_2, \ldots, p_k$ en la población total y que la distribución de probabilidad del vector observación $x = (x_1, \ldots, x_p)$ es dado para cada grupo j por una densidad (o una ley discreta) $f_j(x)$.

Observando un punto de coordenadas $(x_1, x_2, \ldots, x_p)$ la probabilidad de que provenga del grupo j es dada por la fórmula de Bayes:

$$P(G_j/x) = \frac{p_j f_j(x)}{\sum_{j=1}^{k} p_j f_j(x)}$$

La regla bayesiana consiste entonces en asignar la observación x al grupo que tiene la probabilidad *a posteriori* máxima.

Los denominadores siendo los mismos para los k grupos, se debe entonces buscar el máximo de:

$$p_j f_j(x)$$

Es entonces necesario conocer o estimar $f_j(x)$. Diversas posibilidades existen.

A. Métodos no paramétricos

No se hace hipótesis específica sobre la familia de leyes de probabilidad. Variantes multidimensionales del método del núcleo permiten estimar $f_j(x)$ por

$$\hat{f}_j(x) = \frac{1}{n_j h} \sum_{i=1}^{n_j} K\left(\frac{x - x_i}{h}\right)$$

donde K es una densidad multidimensional.

La discriminación "por bolas" es un caso particular: se traza alrededor de x una bola en $\mathbb{R}^p$ de radio ρ dado y se cuenta el número de observaciones k_j del grupo j en esta bola. Se estimará directamente $P(G_j/x)$ por:

$$\frac{k_j}{\sum_j k_j}$$

(*Observación:* la bola puede ser vacía si ρ es muy pequeño).

Uno de los métodos más utilizados es el método de los k vecinos más cercanos. Se buscan los k puntos más próximos de x en el sentido de una métrica que se precisa y se clasifica x en el grupo más representado: la probabilidad *a posteriori* se obtiene igual que para la discriminación por bolas, pero no tiene mucho sentido si k es pequeño.

B. Métodos paramétricos

Se da una familia parametrizada de leyes de probabilidad para $f_j(x)$ y se utiliza la muestra para estimar los parámetros. El caso normal p-dimensional es el más clásico y será desarrollado más adelante.

Se puede igualmente dar una expresión parametrizada de la probabilidad *a posteriori* y estimarla directamente: será el caso de la regresión logística tratada en la sección §2.4.

2.2 El modelo normal multidimensional

Se supondrá que x sigue una ley $N_p(\mu, \Sigma_j)$ para cada grupo:

$$f_j(x) = \frac{1}{(2\pi)^{p/2}(\det \Sigma_j)^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_j)^t\Sigma_j^{-1}(x-\mu_j)\right]$$

A. El caso general

La regla bayesiana $\max p_j f_j(x)$ se reduce, pasando a logaritmos, a minimizar:

$$(x-\mu_j)^t\Sigma_j^{-1}(x-\mu_j) - 2\ln p_j + \ln(det\Sigma_j)$$

Cuando los Σ_j son diferentes esta regla es cuadrática y se debe comparar k funciones cuadráticas de x. Σ_j es en general estimado por $\frac{n}{n-1}V_j$ y μ_j por g_j.

B. El caso de igualdad de matrices de varianza–covarianza

Si $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k = \Sigma$, la regla se vuelve lineal. En efecto $\ln(\det \Sigma_j)$ es una constante y $(x-\mu_j)^t\Sigma^{-1}(x-\mu_j)$ es entonces igual a la distancia de Mahalanobis teórica de x a μ_j: $\Delta^2(x, \mu_j)$.

Desarrollando y eliminando $x^t\Sigma^{-1}x$ que no depende del grupo, se tiene:

$$\text{máx}\left\{x^t\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^t\Sigma^{-1}\mu_j + \ln p_j\right\}$$

Si Σ es estimada por $\frac{n}{n-k}W$, la regla bayesiana corresponde a la regla geométrica cuando hay igualdad de probabilidades *a priori*. La regla geométrica es entonces óptima.

La probabilidad *a posteriori* de pertenencia al grupo j es proporcional a:

$$p_j \exp\left(-\frac{1}{2}\Delta^2(x, \mu_j)\right)$$

C. Dos grupos con igualdad de las matrices de varianza

Se asignará x al grupo 1 si:

$$x^t\Sigma^{-1}(\mu_1 - \mu_2) > \frac{1}{2}(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) + \ln\frac{p_2}{p_1}$$

Si $p_1 = p_2 = 0.5$ se encuentra la regla Fisher estimando Σ por $\frac{n}{n-2}W$.

Es decir:

$$S(x) = x^t\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)\Sigma^{-1}(\mu_1 - \mu_2) + \ln\frac{p_2}{p_1}$$

Se asignará a x al grupo 1 si $S(x) > 0$ y al grupo 2 si $S(x) < 0$.

La función $S(x)$ llamada *puntaje* o estadística de Anderson está ligada simplemente a la probabilidad *a posteriori* de pertenencia al grupo 1.

Tenemos en efecto

$$P(G_1/x) = P = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)}$$

de donde

$$\frac{1}{P} = 1 + \frac{p_2 f_2(x)}{p_1 f_1(x)} = 1 + \frac{p_2}{p_1} \exp\left[-\frac{1}{2}(x-\mu_2)\Sigma^{-1}(x-\mu_2) + \frac{1}{2}(x-\mu_1)\Sigma^{-1}(x-\mu_1)\right]$$

$$\frac{1}{P} - 1 = \frac{p_2}{p_1} \exp\left[\frac{1}{2}\Delta^2(x,\mu_1) - \frac{1}{2}\Delta^2(x,\mu_2)\right]$$

de donde $\ln\left(\frac{1}{P} - 1\right) = -S(x)$.

Es decir:

$$P = \frac{1}{1 + \exp(-S(x))} = \frac{\exp(S(x))}{1 + \exp(S(x))}$$

Se dice que P es función "logística" del *puntaje*.

Cuando $p_1 = p_2 = \frac{1}{2}$:

$$P = \frac{1}{1 + \exp\left(-\frac{1}{2}(\Delta^2(x,\mu_1) - \Delta^2(x,\mu_2))\right)}$$

He aquí a manera de ejemplo la tabla 6 que da las asignaciones de las 45 primeras observaciones de los datos de infartos según la regla anterior. El asterisco indica un error de clasificación.

Tabla 6

	Grupo real	Grupo atribuido		$P(G_1/x)$	$P(G_2/x)$
1	sobreviviente	sobreviviente		0.4515	0.5485
2	fallecido	fallecido		0.8140	0.1860
3	fallecido	fallecido		0.9597	0.0403
4	sobreviviente	sobreviviente		0.2250	0.7750
5	fallecido	fallecido		0.8112	0.1888
6	fallecido	fallecido		0.8928	0.1072
7	sobreviviente	sobreviviente		0.3202	0.6798
8	sobreviviente	fallecido	*	0.8711	0.1289
9	sobreviviente	sobreviviente		0.0984	0.9016
10	sobreviviente	sobreviviente		0.0797	0.9203
11	sobreviviente	sobreviviente		0.0138	0.9862
12	sobreviviente	sobreviviente		0.0160	0.9840
13	sobreviviente	sobreviviente		0.0052	0.9948
14	sobreviviente	sobreviviente		0.0105	0.9895
15	sobreviviente	sobreviviente		0.0019	0.9981
16	sobreviviente	sobreviviente		0.0258	0.9742
17	sobreviviente	sobreviviente		0.2011	0.7989
18	sobreviviente	sobreviviente		0.2260	0.7740
19	sobreviviente	sobreviviente		0.0022	0.9978
20	sobreviviente	sobreviviente		0.1222	0.8778
21	sobreviviente	sobreviviente		0.0014	0.9986
22	fallecido	fallecido		0.8629	0.1371
23	fallecido	sobreviviente	*	0.4804	0.5196
24	fallecido	fallecido		0.9900	0.0100
25	fallecido	fallecido		0.5845	0.4155
26	fallecido	fallecido		0.7447	0.2553
27	fallecido	fallecido		0.7067	0.2933
28	fallecido	sobreviviente	*	0.4303	0.5697
29	sobreviviente	sobreviviente		0.1118	0.8882
30	sobreviviente	fallecido	*	0.5734	0.4266
31	sobreviviente	sobreviviente		0.2124	0.7876
32	fallecido	fallecido		0.9928	0.0072
33	fallecido	fallecido		0.7301	0.2699
34	sobreviviente	fallecido	*	0.5354	0.4646
35	fallecido	fallecido		0.9943	0.0057
36	sobreviviente	sobreviviente		0.1218	0.8782
37	sobreviviente	sobreviviente		0.2757	0.7243
38	sobreviviente	sobreviviente		0.1759	0.8241
39	fallecido	fallecido		0.9555	0.0445
40	sobreviviente	sobreviviente		0.0695	0.9305
41	fallecido	fallecido		0.9762	0.0238
42	fallecido	fallecido		0.9785	0.0215
43	sobreviviente	sobreviviente		0.3240	0.6760
44	sobreviviente	sobreviviente		0.2121	0.7879
45	fallecido	fallecido		0.7880	0.2120

Bajo reserva del carácter realista de la hipótesis de multinormalidad, sus resultados son

entonces más precisos que una simple decisión según la distancia más corta. El cálculo de probabilidades *a posteriori* muestra aquí que 4 clasificaciones erróneas sobre 5 se produjeron en una zona de incertidumbre.

D. Acerca de ciertas pruebas

La hipótesis de igualdad de las matrices Σ_i puede ser probada por medio de una prueba de Box que generaliza la de Bartlett para el caso unidimensional.

Si la hipótesis $\Sigma_1 = \Sigma_2 = \ldots = \Sigma_k$ *es verdadera, la cantidad:*

$$\left(1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right)\left[\left(\sum_i \frac{1}{n_i - 1} - \frac{1}{n-k}\right)(n-k)\ln\left|\frac{n}{n-k}W\right| - \sum_i (n_i - 1)\ln\left|\frac{n_i}{n_i - 1}V_i\right|\right]$$

sigue aproximadamente una ley χ^2 *con* $\dfrac{p(p+1)(k-1)}{2}$ *grados de libertad.*

Si se rechaza la hipótesis de igualdad, ¿debemos utilizar las reglas cuadráticas? Esto no es seguro en todos los casos. Para empezar, el test de Box no es perfectamente confiable, además que el uso de las reglas cuadráticas implica la estimación de más parámetros que la regla lineal porque se debe estimar cada Σ_j. Cuando las muestras son pequeñas las funciones obtenidas son muy poco robustas y es mejor utilizar una regla lineal a pesar de todo.

Para dos grupos el resultado siguiente está en el origen de los métodos clásicos de selección de variables:

Sea un subgrupo de l *variables entre las* p *componentes de* x. *Supongamos que* $\Delta_p^2 = \Delta_l^2$, *en otros términos las* $p-1$ *variables restantes no dan ninguna información para separar las dos poblaciones; entonces:*

$$\frac{(n_1 + n_2 - p - 1)n_1 n_2 (D_p^2 - D_l^2)}{(p-l)(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2 D_l^2} = F(p - l; n_1 + n_2 - p - 1)$$

Se puede así probar el crecimiento de la distancia de Mahalanobis aportada por una nueva variable a un grupo ya constituido tomando $l = p - 1$.

Cuando se hace la discriminación entre más de dos grupos, las pruebas son las que utilizan el Λ de Wilks.

El test de igualdad de las k esperanzas $\mu_1 = \mu_2 = \cdots = \mu_k$ es el siguiente

$$\Lambda = \frac{|W|}{|V|} = \frac{|W|}{|W + B|} = \frac{1}{|W^{-1}B + I|}$$

sigue la ley de Wilks de parámetros p, $n-k$, $k-1$ *bajo* $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.

Porque nV, nW, nB siguen respectivamente las leyes de Wishart con $n-1, n-k, k-1$ grados de libertad.

Si $k = 3$ se utilizará la ley exacta de Λ y no una aproximación

$$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} = \frac{p}{n-k-p+1} F(2p; 2(n-k-p+1))$$

Si $k = 2$, la prueba de Wilks y el test de la distancia de Mahalanobis ($H_0 : \Delta_p^2 = 0$) son idénticos porque B siendo de rango 1 se tiene:

$$\Lambda = \frac{1}{1 + D_p^2 \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)}} = \frac{1}{\mu+1} = 1-\lambda$$

La prueba $H_0 : \mu_i = \mu \ \ \forall i$ puede efectuarse igualmente utilizando como estadística de prueba la traza de $W^{-1}B$ llamada estadística de Lawley-Hotelling que sigue la ley T_0^2 generalizada de Hotelling aproximable por un $\chi^2_{p(k-1)}$.

La traza de $V^{-1}B$ es llamada traza de Pillai. Para la introducción paso a paso de variables discriminando en k grupos, se utiliza usualmente la prueba de variación Λ medida por:

$$\frac{n-k-p}{k-1}\left(\frac{\Lambda_p}{\Lambda_{p+1}} - 1\right)$$

que se compara con un $F_{k-1;n-k-p}$

2.3 Medidas de eficacia de las reglas de clasificación

El criterio usual es la probabilidad de clasificar bien una observación cualquiera. Se comparará la eficacia de los diversos métodos de clasificación en términos de tasas de error.

A. Tasa de error para dos grupos con $\Sigma_1 = \Sigma_2$ y distribución normal

Cuando $p_1 = p_2$, la regla de clasificación teórica es asignar al grupo 1 si:

$$S(x) = x^t\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1+\mu_2)^t\Sigma^{-1}(\mu_1-\mu_2) > 0$$

La probabilidad de error de clasificación es entonces:

$$P(S(x) > 0 / x \in N_p(\mu_2; \Sigma))$$

La ley de $S(x)$ es una ley de Gauss de una dimensión como combinación lineal de componentes de x.

$$E(S(x)) = \mu_2^t\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}(\mu_1+\mu_2)^t\Sigma^{-1}(\mu_1-\mu_2)$$

$$= \frac{1}{2}(\mu_1-\mu_2)^t\Sigma^{-1}(\mu_1-\mu_2) = -\frac{1}{2}\Delta_p^2$$

$$V(S(x)) = (\mu_1-\mu_2)^t\Sigma^{-1}\Sigma\Sigma^{-1}(\mu_1-\mu_2) = \Delta_p^2$$

de donde

$$S(x) \quad \text{sigue una} \quad LG\left(-\frac{1}{2}\Delta_p^2;\Delta_p\right) \quad \text{si} \quad x \in G_2.$$

La probabilidad de clasificar en el grupo 1 una observación del grupo 2 es:

$$P(1/2) = P\left(U > \frac{\Delta_p}{2}\right)$$

que es igual a $P(2/1)$. Esta relación da una interpretación concreta de la distancia de Mahalanobis.

Si $p_1 \neq p_2$ se tiene:

$$P(1/2) = P\left(U > \frac{\Delta_p}{2} + \frac{1}{\Delta_p}\ln\frac{p_2}{p_1}\right)$$

$$P(2/1) = P\left(U > \frac{\Delta_p}{2} - \frac{1}{\Delta_p}\ln\frac{p_2}{p_1}\right)$$

Cuando μ_1, μ_2 y Σ son estimados $S(x)$ no sigue una ley normal y utilizar D_p como estimación del Δ_p conduce a una estimación sesgada de las probabilidades de error de clasificación: hay en promedio, subestimación de la probabilidad global de error, $p_1P(2/1)+p_2P(1/2)$ debido entre otras razones al hecho que D_p^2 sobreestima Δ_p^2 ([1] capítulo 15, sección 15.5.6c).

Para el ejemplo de los infartos, como $D_p = 2.223$ se llega a una estimación de tasas de error igual a $P(U > 1.11) = 0.13$.

La utilización de la estimación sin sesgo de Δ^2, $\frac{n-p-1}{n-2}D^2 - p\frac{n}{n_1n_2} = 4.37$ conduce a una estimación de la tasa de error cercana al 15%.

B. El método de resustitución

éste consiste en reasignar las n observaciones según las funciones discriminantes encontradas. En el ejemplo de los infartos se obtienen los resultados dados en la tabla 7, es decir, con una tasa de error de 13%.

Tabla 7

grupo atribuido / grupo real	deceso	sobrevivencia
deceso	46	5
sobrevivencia	8	42

Este método tiene un gran defecto: subestima sistemáticamente la tasa de error porque se utilizan las mismas observaciones que sirvieron para encontrar las funciones discriminantes. La regla óptima para la muestra da buenos resultados si se aplica sobre ella.

C. Los métodos de validación cruzada

Para evitar el defecto del método de resustitución se aconseja partir la muestra en dos submuestras: una servirá para la elaboración de reglas de clasificación (muestra de base o de aprendizaje), la otra para la aplicación de las reglas de clasificación (muestra–prueba).

La tasa de error medida sobre la muestra–prueba será entonces una estimación sin sesgo de la tasa verdadera. Esto supone que se tienen numerosos datos para poder sustraer sin riesgo una parte considerable de los datos (25% es aconsejado).

En el caso de muestras pequeñas la técnica siguiente de Lachenbruch y Mickey (comparable al *press* en regresión) permite obtener una estimación realista de la tasa de error.

Se efectúan n análisis discriminantes sobre cada una de las n muestras de las $n-1$ observaciones obtenidas poniendo de lado cada vez una de las observaciones. Se clasifica entonces la observación que fue dejada aparte y se cuenta el porcentaje de error de clasificación.

Observación: Los procedimientos usuales de selección de variables optimisan criterios probabilísticos: ley de Λ Wilks o distancia de Mahalanobis, que no necesariamente optimizan el porcentaje correctamente.

2.4 La regresión logística

Cuando sólo hay dos grupos, bajo la hipótesis de normalidad e igualdad de las matrices de varianza, se ha visto que la probabilidad *a posteriori* era una función logística del *puntaje*, el cual era una función lineal de las variables explicativas.

Se tiene pues:

$$\ln\left(\frac{f_1(x)}{f_2(x)}\right) = \beta_0 + \beta^t x \quad \text{donde} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

El modelo de regresión logística consiste en partir de la relación anterior y estimar los $p+1$ parámetros según el máximo de verosimilitud.

Respecto a la discriminación lineal usual, el modelo logístico implica menos parámetros y cubre una amplia gama de leyes de probabilidades (las variables explicativas pueden ser binarias).

Se tiene, por tanto:

$$P(G_1/x) = \frac{p_1 f_1(x)}{p_1 f_1(x) + p_2 f_2(x)} = \frac{\frac{p_1}{p_2}\frac{f_1(x)}{f_2(x)}}{1 + \frac{p_1}{p_2}\frac{f_1(x)}{f_2(x)}}$$

$$= \frac{\exp\left(\beta_0 + \ln\frac{p_1}{p_2} + \beta^t x\right)}{1 + \exp\left(\beta_0 + \ln\frac{p_1}{p_2}\beta^t x\right)}$$

$$P(G_2/x) = \frac{1}{1 + \exp\left(\beta_0 + \ln\frac{p_1}{p_2} + \beta^t x\right)}$$

La verosimilitud de las β (suponiendo n_1 y n_2 fijos y no aleatorios) es:

$$L = \prod_{i \in G_1} f_1(x_i) \prod_{i \in G_2} f_2(x_i)$$

Como:

$$f_1(x) = \frac{P(G_1/x) f(x)}{p_1} \quad \text{y} \quad f_2(x) = \frac{P(G_2/x) f(x)}{p_2}$$

Con $f(x) = p_1 f_1(x) + p_2 f_2(x)$, se tiene:

$$L = \frac{1}{p_1^{n_1} p_2^{n_2}} \prod_{i \in G_1} P(G_1/x_i) \prod_{i \in G_2} P(G_2/x_i) \prod_{i=1} f(x_i)$$

$$L = \frac{L_1 L_2}{p_1^{n_1} p_2^{n_2}}$$

donde L_1 es la verosimilitud condicional de los parámetros, conociendo las x_i y la densidad incondicional de las x_i, L_2.

No siendo conocida f se estimarán $\beta_0, \beta_1 \cdots \beta_p$ por un método de máxima verosimilitud condicional:

$$\max_{\beta} \prod_{i \in G_1} \frac{\exp\left(\beta_0 + \ln \frac{p_1}{p_2} + \beta^t x_i\right)}{1 + \exp\left(\beta_0 + \ln \frac{p_1}{p_2} + \beta^t x_i\right)} \prod_{i \in G_2} \frac{1}{1 + \exp\left(\beta_0 + \ln \frac{p_1}{p_2} + \beta^t x_i\right)}$$

Para esto hay que utilizar un método numérico, puesto que no hay solución análitica a la ecuación de la verosimilitud.

Siendo estimadas las β, la regla Bayesiana puede ser aplicada para las clasificaciones. Como:

$$\ln \frac{P(G_1/x)}{P(G_2/x)} = \beta_0 + \ln \frac{p_1}{p_2} + \beta^t x$$

se asignará al grupo 1 si $\beta_0 + \ln \frac{p_1}{p_2} + \beta^t x > 0$.

Cuando los datos provienen de dos pobalciones normales con $\Sigma_1 = \Sigma_2$ la regresión logística es menos adecuada que el análisis discriminante clásico, pues la solución dada por $S(x)$ corresponde a un máximo de verosimilitud verdadera y no a un máximo de verosimilitud condicional (se utiliza menos información en la regresión logísta puesto que sólo f_1/f_2 se supone conocido y no f_1 y f_2).

Parece que la regresión logística da resultados verdaderamente mejores que la regla geométrica, sólo para poblaciones claramente no normales o con Σ_1 muy diferente de Σ_2, pero al precio de un procedimiento de cálculo mucho más complejo que la simple inversión de la matriz W.

Bibliografía

[1] Saporta, G. (1988) *Théorie et Méthodes de la Statistique*, 2a. edición. Technip, París.

Los métodos y las aplicaciones del *credit-scoring*

Gilbert Saporta*

La técnica de discriminación sobre variables cualitativas propuesta por el autor [5], fue inicialmente concebida para calcular puntajes de riesgos bancarios. Desde entonces y con la implementación computacional del proceso para calcular los puntajes, más de una centena de estudios han sido realizados en Francia.

En una primera parte se presenta un resumen de las diferentes técnicas estadísticas utilizables y evaluamos su interés. En la segunda parte se presenta un resumen de los trabajos operacionales. En particular se muestra que surgen numerosos problemas (muestreo, reajuste, acceso a los datos, robustez, fiabilidad, ...) y que la realización de un puntaje operacional depende no sólo de la escogencia de un buen algoritmo, sino además, de un *savoir faire* adquirido gracias a la experiencia. Algunos perfeccionamientos estadísticos son igualmente evocados.

1 Metodología estadística

1.1 Planteamiento del problema

Esquematicamente un problema de *credit-scoring* se puede describir como la búsqueda de un procedimento de separación entre dos grupos: los "malos pagadores" y los "buenos pagadores", conociendo un conjunto de descriptores. Las variables explicativas son casi siempre cualitativas (categoría profesional, estado matrimonial, geotipo, etc.) o se transforman en cualitativas por descomposición en clases (antigüedad en el empleo, por ejemplo). Esta situación excluye el uso clásico de los procedimientos de análisis discriminante. Los datos son por tanto constituidos por dos muestras de tamañ generalmente elevado (varias centenas de observaciones) descritas por p variables cualitativas X_i con m_i modalidades; $i = 1, 2, \ldots, p$ donde p es geenralmente del orden de varias decenas.

1.2 Técnicas estadísticas utilizables

El objetivo es prever la calidad de un expediente a partir de las variables explicativas. Dos grandes categorías de procedimientos se pueden proponer: los fundamentados en la estimación de la probabilidad de pertenencia a una de las clases y los fundamentados en la estimación de una regla numérica de clasificación: "el puntaje". A continuación se describen algunas de esas técnicas estadísticas.

*Département de Mathématiques et Informatique, Centre National d'Arts et Métiers, París

1.2.1 Estimación de la probabilidad

El esquema multinomial es utilizable puesto que hay $2q\Pi_{i=1}^{p} m_i$ casos posibles y este número es en general excesivo. Se debe, por tanto, reducir el nivel de las interacciones útiles y tomar por ejemplo unos modelos del tipo loglineal (ver [2,4] por ejemplo). Se puede también utilizar un método de segmentación. Sin embargo, estas técnicas son muy pesadas para implementar y mal adaptadas al tamaño de las bases de datos que tratamos.

1.2.2 Función de puntaje

Esta enfoque es el más usado: se atribuye a cada categoría de cada variable un valor numérico. El puntaje es la suma de los valores de las categorías a las cuales pertenece un individuo y es la variable numérica de decisión: se puede fijar ya sea un umbral de aceptación o bien decidir en función de las curvas dando los porcentajes de la demanda y el porcentaje de malos expedientes en función del puntaje. La regla es de muy fácil aplicación e interpretación.

El puntaje no es más que la variable discriminate en el sentido de Fisher, combinación lineal de las $\sum_i^p m_i$ variables indicatrices de las categorías de las p variables. El método del puntaje es por lo tanto un análisis discriminante sobre las variables cualitativas, y una forma de cuantificar las variables cualitativas.

El conjunto de las variables indicatrices no es de rango pleno puesto que la suma de las indicatrices de una misma variable cualitativa vale 1. Soluciones técnicas elementales como la supresión de la última categoría de cada variable, permitirían utilizar un programa ordinario de análisis discriminante. Sin embargo, esta no es la técnica que hemos propuesto [5] y que es utilizada bajo el nombre de Disqual.

1.2.3 Disqual

Este método consiste en hacer un análisis factorial (de correspondencias múltiples) del conjunto de las variables cualitativas explicativas. El conjunto de todos los factores engendra el mismo espacio que el conjunto de las indicatrices. Se eliminan los factores de poca inercia y los que no separan suficientemente los dos grupos. (Una de las ventajas de este procedimiento es que permite usar un métrica diagonal ya que los factores son dos a dos ortogonales. Además los resultados de un análisis de correspondencias múltiples son muy robustos, lo que da una seguridad).

Las k coordenadas factoriales retenidas tienen la función de variables explicativas numéricas para un análisis discriminante. Se calcula luego la función de Fisher, que se expresa como una combinación lineal de las indicatrices, lo que da la función de puntaje (una variante de este método, pero sin la selección de los factores, figura en el programa SPAD, etapa DIS2G).

En efecto, si d es la función de Fisher; $d = (d_1 \quad d_k)$ donde $d_j = \dfrac{\overline{z}_1^j - \overline{z}_2^j}{\lambda_j}$ y $\overline{z}_i^j$ es la media de las coordenadas del grupo i sobre el eje factorial j de varianza λ_j.
El puntaje s de un individuo se obtiene por: $s = \sum_{j=1}^{k} d_j z^j$.

Como $z^j = \sum_{l=1}^{m} \alpha_l^j x_l$ se tiene $s = \sum_{j=1}^{k} \sum_{l=1}^{m} d_j \alpha_l^j x_l$ o sea

$$s = \sum_l (\sum_j d_j \alpha_l^j) x_l$$

donde x_l es la indicatriz de la categoría l. l valor del puntaje de esta categoría vale por lo tanto $\sum_{j=1}^{k} d_j \alpha_l^j$.

Por construcción de la función de puntaje, esto corresponde a un modelo puramente aditivo sin interacción. Tomando en cuenta las interacciones entre las variables explicativas mejoraría claramente el comportamiento del puntaje. Esto se puede hacer de la manera siguiente: se detecta por medio de un modelo log-lineal o por medio de una segmentación, que X_1 y X_2 interacctúan. Se reemplazan estas variables por la variable $X_1 \times X_2$ con $m_1 m_2$ categorías y se aplica el método Disqual. Si $m_1 m_2$ es muy grande (normalmente lo es) se pueden agrupar categorías.

2 Las aplicaciones del *credit-scoring*: riesgos de fracaso y condiciones de éxito

2.1 Algunos problemas metodológicos

Mencionamos aquí algunos aspectos de la implementación del método mencionado al *credit-scoring*. Aparecen varias dificultades ligadas ya sea a problemas metodológicos o bien a problemas prácticos. Para que un puntaje sea operacional y utilizable, no sólo hay que escoger un buen algoritmo aino que también se debe adquirir un gran *savoir faire*.

2.1.1 La población, el muestreo y los reajustes

El primer problema es quizás el más delicado para resolver. Es el que lleva a un gran número de fracasos. Para que un puntaje sea operacional debe ser aplicable al conjunto de la demanda. Esta es asimilada a una población numerable P. Para construir un puntaje lo ideal sería observar el comportamiento pagador de una muestra aleatoria representativa de P. Ahora bien, en la práctica, se realiza una preselección y algunos expedientes son rechazados. Consecuentemente la muestra observada M_1, es sacada de una parte P_1 de P, compuesta de los expedientes aceptables. Si no se tienen los cuidados necesarios, el puntaje se aplicará por tanto a P_1 y no a P.

Durante la implementación del procedimiento ello significa que los expedientes deberán ser primero seleccionados con la ayuda de los métodos tradicionales, y en segundo término, con la ayuda del puntaje. Pero esto, en la práctica, es inaceptable.

Para solucionar el inconveniente, varios enfoques son posibles. Se les puede clasificar en dos categorías las cuales hacen referencia a una muestra M_2 de expedientes rechazados sacados de $P_2 = P - P_1$. En el primer caso, se busca rehacer M_1 para darle la estructura de P sobre las principales variables de control. En el segundo caso, se simula el comportamiento pagador (desconocido) de las observaciones pertenecientes a M_2.

Un segundo problema de muestreo se presenta al tener en cuenta la dimensión temporal de la demanda. Se observan frecuentemente fenómenos estacionales. Por ejemplo, en Francia junio es más malo que octubre. Además, la observación del comportamiento pagador de

un expediente rechazado se debe hacer en un período fijo (por ejemplo 24 meses), teniendo en cuenta las leyes de aparición de las cuentas morosas. Los expedientes técnicamente utilizables tienen cuando menos 24 meses de antigüedad. Mientras tanto la estructura de la demanda, la reglamentación y las condiciones de aceptación pueden evolucionar y el puntaje corre el riesgo de ser obsoleto antes de ser utilizado. Un plan de sondeo bien construido permite crear una muestra completa de expedientes suficientemente antiguos, además de una muestra representativa de la demanda reciente sobre la cual buscará ajustarse.

En conclusión, dos tipos de reajustes se imponen. El primero permite corregir la estructura de la submuestra observada para volverle a dar la estructura de la demanda. El segundo permite corregir los efectos de envejecimiento de la muestra para volverle a dar la estructura de la demanda reciente.

2.1.2 La escogencia de las variables y de las categorías

Otros problemas surgen cuando se trata de escoger las variables que se van a introducir en el modelo y su descomposición en categorías. Un algoritmo eficaz de selección que no necesitaría realizar el análisis discriminante en cada paso, no ha sido aun implementado. El método que he propuesto con la ayuda de los coeficientes parciales de Chuprow, no ha aportado los resultados deseados. Una dificultad proviene del hecho que las variables indicatrices de una variable cualitativa no pueden ser disociadas y por tanto se debe proceder por bloques de variables.

Sin embargo, los métodos automáticos de selección del tipo "stepwise" conducen usualmente a soluciones inaceptables para el usuario. Una variable "clásica" como la tasa de endeudamiento, por ejemplo, puede ser rechazada puesto que su poder explicativo es ligeramente inferior que una variable "exótica" y poco familiar para el usuario. De la misma manera un método de descomposición optimal en categorías o de reagrupamiento optimal, puede conducir a la construcción de clases que sorprenden al usuario. Además ciertas escogencias "optimales" sobre una muestra pueden ser engañosas. A este nivel, el verdadero *savoir faire* no reside en la escogencia de las técnicas de optimización, sino esencialmente en el conocimiento del campo y la acumulación de experiencias. Los métodos de optimización proveen una primera solución "estandarizada", luego es necesario buscar las soluciones operacionales.

Lo mismo ocurre con la escogencia de las variables cruzadas. La experiencia nos ha mostrado que, frecuentemente resulta preferible sustituir las dos variables cualitativas por una sola variable obtenida por producto cartesiano. En este caso, es además, necesario reagrupar entre ellas algunas categorías de la variable cruzada. La búsqueda de mejores cruzamientos se puede hacer con la ayuda de modelos tales como el análisis de la varianza cualitativo o el modelo loglineal. Este tipo de modelo puede conducir también hacia soluciones artificiosas, y, a menudo, la experiencia guiará al análisis en sus escogencias *a priori*, de cruzamientos por probar.

2.1.3 Estabilidad e interpretación de los coeficientes

Las técnicas de puntaje se parecen a las de los modelos lineales generalizados: son proclives a aumentar artificialmente el número de variables y de categorías en el modelo. Ello permite, en apariencia, mejorar la calidad de la discriminación, pero, el problema de la estabilidad

de los estimadores de los puntajes es crucial. Además no se dispone en este caso del arsenal de tests utilizados en el caso del modelo lineal. Los métodos de remuestreo del tipo "bootstrap" deberían permitir determinar intervalos de confianza para los puntajes de las categorías y probar las hipótesis concernientes en ausencia de toda hipótesis distribucional. En la práctica es deseable poder interpretar los pesos de los diferentes puntajes en relación con la tasa de no pago en la categoría. Una inversión, aún cuando sea estadísticamente justificada es, para el usuario, difícil de explicar. Un modelo más "económico" en variables y en reagrupamientos juiciosos de categorías permitiría eliminar tales inversiones.

2.1.4 Estimación de las probabilidades a posteriori

Si las distribuciones condicionales de las coordenadas factoriales son normales se pueden aplicar los resultados usuales del análisis discriminante. Esta hipótesis no se verifica siempre y una estimación precisa de las probabilidades se puede efectuar por medio de un método no paramétrico del tipo núcleos de Parzen [3]. Su integración en un programa del tipo Disqual está por hacer.

2.1.5 Desviaciones entre las previsiones y las realizaciones

Una vez el modelo elaborado su utilización es doble. Por supuesto, las políticas de selección fundadas sobre la red de puntaje son implementadas. Por otra parte, el modelo es utilizado para hacer previsiones de no pagadores, en número y en volumen, teniendo en cuenta las políticas de selección escogidas. Rápidamente pueden aparecer desviaciones entre las previsiones y las realizaciones. Su análisis es un punto cardinal y permite orientar las acciones. Las desviaciones pueden surgir por causa de una evolución del ambiente: contexto reglamentario o política comercial del organismo. También pueden deberse a una evolución de la demanda, lo que se detecta facilmente siguiendo el perfil de la clientela y comparándolo con el de la muestra. Los sesgos no controlados en el momento de la construcción de la muestra también pueden contribuir a las desviaciones.
La detección de todas estas desviaciones requiere la puesta a punto de herramientas de control. ¿Cómo combatir el problema? Normalmente un simple ajuste de umbrales de aceptación permite corregir la desviación. Si ello no es suficiente y si las causas son estructurales, la única solución es ajustar un nuevo modelo sobre una muestra actualizada.

2.2 Algunos problemas prácticos

2.2.1 El acceso a los datos pertinentes

Usualmente las informaciones contenidads en los expedientes aceptados y rechazados no son almacenados sobre soporte informático. Los expedientes rechazados normalmente no son conservados. Las estadísticas de demanda y de aceptación que permiten elaborar el plan de sondeo no siempre existen. Es por lo tanto necesario realizar una fase previa de conteo, de codificación y de digitación. En lo sucesivo se deben implementar los procedimientos para la adquisición de datos de manera sistemática, que permita alimentar una base de datos. Esta base servirá para realizar los trabajos de seguimiento de las desviaciones y sacar las muestras de futuros estudios de puntaje.

Otro problema que se presenta frecuentemente es que lo histórico no es utilizable, ya sea porque se trata de una actividad nueva o reciente, o bien porque las reglas de selección eran tan severas que ningún expediente deriva en moroso. Es entonces posible implementar técnicas totalmente diferentes, haciendo simulaciones de la demanda, de la aceptación y del comportamiento.

2.2.2 La modificación de los procedimentos

Otros problemas están ligados a las modificaciones de los procedimentos administrativos e informáticos, y éstos son aún más delicados. Para resolverlos es necesario evolucionar las aplicaciones informáticas e implementar acciones de formación y motivación de los usuarios. Por el momento no desarrollamos estos aspectos, solamente llamamos la atención sobre esas dificultades cuya subevaluación lleva, la mayor parte de las veces, a un fracaso.

3 Conclusión

Pese a todas las dificultades metodológicas y prácticas y a los riesgos de fracaso derivados, 10 años de consultoría y de realización me han permitido poner en evidencia las ventajas decisivas del *credit-scoring*. Comparado con otros procedimentos tradicionales, el puntaje posee tres cualidades fundamentales:

1. La **simplicidad**: la recolección de un número mínimo de informes permite tomar una decisión rápida.
2. La **homogeneidad**: la política de aceptación es aplicada uniformente en la red.
3. La **flexibilidad**: la política de selección está basada sobre bases objetivas. Toda inflexión puede ser decidida rápidamente midiendo las consecuencias de las nuevas escogencias.

Aparte de estas ventajas, específicas del método, citamos las principales consecuencias de un puntaje:

1. La **disminución del número de morosos**: una selección optimizada permite reducir el volumen de los morosos.
2. La **previsión de los morosos**: el modelo permite simular con precisión los futuros no pagadores y por tanto optimizar el cálculo de las provisiones.
3. La **productividad**: solamente los expedientes con riesgo son objeto de un examen cuidadoso. En la mayoría delos casos la decisión puede ser delegada y el análisis acelerado.

Para obtener todas estas ventajas, los bancos y los organismos financieros deben ser conscientes que la implementación de un puntaje necesita un conjunto de competencias estadísticas y prácticas y que, por consiguiente, se trata de una decisión estratégica cuya realización necesita tiempo e inversión.

Bibliografía

[1] Bouroche J.M., Saporta G., Tenenhaus M. (1977) *Some methods of qualitative data analysis*, in Recent Developments in Statistics, J.R. Barra (ed.), pp 749-755, North Holland, Amsterdam.

[2] Daudin J.J. (1980) *Régression qualitative: choix de l'espace prédicteur*, in Data Analysis and Informatics, E. Diday (ed.), pp 329-345, North Holland, Amsterdam.

[3] Gautier J.M., Saporta G. (1984) *Méthodes non paramétriques en analyse discriminante; quelques propositions nouvelles*, in Data Analisis and Informatics III, E. Diday et al. (eds.), pp 591-605, North Holland, Amsterdam.

[4] Goldstein M., Dillon W.R. (1978) *Discrete Discriminant Analysis*, John Wiley & Sons, Nueva York.

[5] Saporta G. (1976) *Discriminant Analysis when the variables are nominal*, Spring Meeting of the Psychometric Society, Murray Hill, Nueva York.

[6] Saporta G. (1977) *Une méthode et un programme d'analyse discriminante pas à pas sur variables qualitatives*, 1eres Journées Analyse des Donées et Informatique, INRIA, pp 201-210, París.

[7] Sireyjol N. (1987) *Les apports du crédit-scoring*, Banque, N^0 457, pp 788-794.

Enfoque bayesiano del análisis discriminante

José Pastrana Z.*

Resumen

En este artículo se presenta la **Inferencia Bayesiana** dentro del marco del **Análisis de Decisiones**, como el tipo de inferencia ideal para asegurar la optimalidad de la regla de clasificación de un nuevo individuo o caso a una de varias poblaciones preestablecidas. En el contexto de la **Inferencia Bayesiana**, optimalidad consiste en la cualidad que posee la regla de clasificación, de minimizar el costo esperado.

En él se argumenta que la regla de clasificación establecida de tal manera que maximiza la variabilidad entre poblaciones relativa a la variabilidad dentro de poblaciones y que está basada en el **Análisis Discriminante**, es un caso particular de la regla de Bayes, siempre que se usen todas las funciones discriminantes, haya igualdad de matrices de covariancias, las poblaciones sean (multi)normales y no haya diferencias entre los costos de clasificación errónea, ni entre las probabilidades previas de pertenencia de población a población. Bajo otras condiciones (diferentes matrices de covariancia, desigualdad de costos de clasificación errónea, etc.), esta regla posee como único atributo óptimo el de separar (discriminar) al máximo las poblaciones, a través de los puntajes asignados a cada uno de sus miembros, empleando las funciones discriminantes, siempre que las matrices de covariancia sean iguales. Cuando la regla separa al máximo no las poblaciones, sino muestras aleatorias observadas independientemente de cada población (asumiendo de nuevo igualdad de matrices de covariancia), se espera que la regla también separe al máximo las poblaciones, lo cual será cierto sobretodo si las muestras son suficientemente grandes y de verdad las matrices de covariancia son iguales. A este respecto, la conveniencia en cuanto a la reducción de dimensiones del empleo de unas pocas funciones discriminantes en la regla de clasificación, dejando por fuera aquéllas que son funciones discriminantes marginales, en términos del reducido porcentaje de la dispersión poblacional que explican, tiene las desventajas de prescindir de información presente en la muestra original (empleo de una menor distancia al clasificar) y de excluir la posibilidad de que la regla sea óptima en el sentido de Bayes.

1 Introducción

Los problemas de **separación** de distintos grupos de objetos u observaciones y de **ubicación** de nuevos objetos u observaciones a poblaciones preestablecidas, son a menudo

*Escuela de Estadística, Universidad de Costa Rica

resueltos empleando respectivamente, las técnicas de análisis multivariable conocidas como **Análisis Discriminante** y **Clasificación**. Ambas técnicas están estrechamente relacionadas, porque una o varias funciones (discriminantes) que se empleen para separar varios grupos de objetos, puede(n) emplearse a su vez, para ubicar nuevos objetos a esos grupos y también una regla de clasificación, puede sugerir un procedimiento para separar grupos. El **Análisis Discriminante** es una técnica básicamente exploratoria, para investigar diferencias observadas entre los objetos de grupos preestablecidos, cuando las relaciones causales no se conocen bien. La **Clasificación** es una técnica menos exploratoria que el **Análisis Discriminante**, porque ésta conduce a reglas bien definidas para ubicar nuevos objetos a grupos definidos. Como parte del **Análisis Discriminante**, las observaciones correspondientes a cada objeto son reducidas a puntajes obtenidos mediante la aplicación de funciones generadas siguiendo el criterio de separar al máximo los vectores de puntajes medios o centroides. En esta tarea, se busca utilizar el menor número de funciones para así reducir la dimensión del espacio de observaciones. Está claro que la técnica del **Análisis Discriminante**, no tiene como fin primordial el ubicar nuevos objetos a los grupos separados mediante las funciones de discriminación y que su empleo para ese fin, constituye una fuente de anomalías en cuanto a carencia de atributos óptimos y bajo rendimiento (alta tasa de clasificación errónea), por parte de la regla de clasificación respectiva. En este artículo se recomienda el empleo del **Análisis Discriminante** para fines de clasificación, exclusivamente en aquellos casos en que rijan condiciones que hacen que la regla de clasificación respectiva, sea óptima o cuasi–óptima desde el punto de vista de la minimización del costo esperado de clasificación errónea ((multi)normalidad de las poblaciones, igualdad de matrices de covariancia, igualdad de costos de clasificación errónea e igualdad de probabilidades de pertenencia previa). En cuanto a referencias de estudios aplicados dentro del ánimo de esta recomendación, las siguientes tres referencias, podrían ser útiles: estudio para establecer la financiabilidad de un proyecto de investigación sometido a la Vicerrectoría de Investigación de la UCR [3]; estudio para ubicar estudiantes potencialmente desertores en el Centro Regional de la UNED en San Carlos [4] y estudio para evaluar la persistencia potencial en el pago de primas, por parte de un individuo que desea comprar un seguro de vida en el INS [6].

La técnica de **Clasificación** incluye la aplicación de un procedimiento o regla, para ubicar nuevos objetos a grupos preestablecidos. La regla no es derivada arbitrariamente, sino siguiendo algún criterio de idoneidad. Una buena regla de clasificación debe incluir la probabilidad previa P de pertenencia a un grupo pues, en efecto, existe una probabilidad previa mayor de pertenencia al grupo más numeroso en comparación con los grupos más pequeños. Análogamente, la regla debe incluir el costo C de clasificación errónea, pues hay errores de clasificación mucho más costosos que otros, como por ejemplo, cuando se compara el costo de clasificar erróneamente un enfermo dentro del grupo de pacientes que tienen SIDA, con el costo de clasificar erróneamente un enfermo dentro del grupo de pacientes que tienen una enfermedad curable y no contagiosa.

Dentro del marco de la **Inferencia Bayesiana**, la regla de clasificación responde al criterio de minimización del costo esperado de clasificación errónea e incluye los costos de errores de clasificación y las probabilidades de pertenencia a priori. En situaciones en que no existe ninguna idea sobre los costos de errores de clasificación, ni sobre las probabilidades previas, lo usual es asumir igualdad de costos entre sí e igualdad de probabilidades previas entre sí. Claro está, la solución obtenida asumiendo igualdad entre costos de clasificación errónea y entre probabilidades previas, es solamente una dentro de muchas soluciones que podrían generarse, asumiendo diversos escenarios de costo de errores, probabilidades previas y, cuando proceda, distribuciones alternativas de las variables.

La regla de Bayes, que es discutida en la próxima sección de **Desarrollo teórico**, también incluye las funciones de densidad o probabilidad f_i de cada grupo o población. Ante el desconocimiento de esas funciones, lo que procede es estimarlas mediante pruebas de bondad de ajuste. La regla de Bayes no requiere ninguna forma específica de las funciones de densidad o probabilidad. En particular, la mayoría de las funciones podrían ser funciones de probabilidad correspondientes a variables categóricas, medidas a escala nominal, sin que ese hecho repercuta en la cualidad óptima de la regla. Como es lógico, las estimaciones de funciones de densidad o probabilidad y de parámetros que se empleen dentro de la regla de Bayes, deben ser establecidas con base en muestras representativas observadas suficientemente grandes (tanto como el grado de asimetría del grupo lo demande).

2 Desarrollo teórico

El marco natural para analizar el problema de clasificación de un nuevo sujeto a una de varias poblaciones $Z_1, Z_2, \ldots, Z_g$ es el de decisiones múltiples [1]. En efecto, existen g decisiones posibles, $D_1, D_2, \ldots, D_g$, en donde D_i consiste en la decisión de ubicar un nuevo individuo en la población Z_i, $i = 1, 2, \ldots, g$. Las probabilidades previas P de pertenencia a las poblaciones $Z_1, Z_2, \ldots, Z_g$ vienen dadas respectivamente por $P_1, P_2, \ldots, P_g$. El costo de clasificar un nuevo individuo en la i–ésima población, esto es, de optar por la decisión D_i, cuando él realmente pertenece a la j–ésima población, viene dado por $C(i/j)$, para todo $i, j = 1, 2, \ldots, g$. Note que $C(i/j) = 0$ si $i = j$, pues en tal caso no hay error de clasificación, mientras que $C(i/j) > 0$, para $i \neq j$.

El marco de decisiones múltiples es realmente el marco de la metodología llamada **Análisis de Decisiones**. Los autores Pastrana, Gold y Swartzel (1992), han descrito la naturaleza de un problema de **Análisis de Decisiones**, en términos de los elementos que debe tener presentes. En un problema de **Clasificación** todos los elementos citados por esos autores están presentes, a saber: sistema subyacente con respecto al cual se desea adoptar una decisión; un conjunto de posibles decisiones D; un conjunto de consecuencias que son resumidas en la función de costo C; una función de valor (la función identidad)

que asigna un grado de deseabilidad relativa V a cada consecuencia C (en nuestro caso, $V = C$); una distribución de probabilidad sobre V y, por lo tanto, sobre C que corresponde a la distribución de probabilidades previas P (cuando no hay casos observados) o a la distribución de probabilidades de pertenencia a posteriori, P/x (cuando hay una matriz de datos x observada).

Dados los elementos mencionados en el párrafo anterior, del problema de **Clasificación**, enfocado dentro del marco del **Análisis de Decisiones** con reglas de inferencia [1] correspondientes a la Inferencia Bayesiana, la regla de Bayes establece que la decisión D debe ser aquélla que minimice el costo esperado. En otras palabras, asígnese el nuevo elemento a la población Z_k, k en $\{1, 2, \ldots, g\}$, siempre que:

$$P_1 C(k/1) f_1 + \ldots + P_{k-1} C(k/k-1) f_{k-1} + P_{k+1} C(k/k+1) f_{k+1} + \ldots + P_g C(k/g) f_g$$

sea mínima sobre todas las sumas de productos similares correspondientes a cada una de las otras poblaciones [2] Para el caso de igualdad entre costos, esta regla se reduce a asignar el nuevo elemento a la población Z_k si el producto $P_k \cdot f_k$ es máximo sobre todos los demás productos similares para cada una de las otras poblaciones.

3 Recomendaciones

Además de la recomendación presentada en la **Introducción**, en el sentido de cuando es más seguro emplear las funciones discriminantes con fines clasificatorios, a continuación se sugiere una serie de pasos para resolver un problema de clasificación aplicando lo más fielmemente posible, la regla de Bayes de minimización del costo esperado de clasificación errónea:

1. Efectúe un estudio de las distribuciones marginales y conjuntas de las variables. Obtenga distribuciones acumulativas empíricas e incluya pruebas de bondad de ajuste, según corresponda. Además, efectúe transformación de variables conforme se considere necesario.

2. Establezca un escenario inicial realista de probabilidades de pertenencia previas y de costos de clasificación errónea. Este escenario podría ser uno de igualdad entre costos e igualdad entre probabilidades de pertenencia previas.

3. Aplique la regla de Bayes presentada en la sección de **Desarrollo teórico**, para ubicar nuevos individuos a las diversas poblaciones.

4. Establezca otros escenarios de probabilidades previas y costos de clasificación errónea y aplique la regla de Bayes para clasificar los individuos correspondientes.

[1] Proceso inductivo mediante el cual se generaliza de los casos individuales (muestra), a la población correspondiente.

5. Determine la sensibilidad de la clasificación de los nuevos casos a los cambios en probabilidades previas y costos de los errores de clasificación. Aquí conviene evaluar el rendimiento de la regla de Bayes mediante la consideración de índices de clasificación riesgosa [2].

6. Escoja la clasificación de nuevos casos correspondiente al escenario de probabilidades previas y costos de errores que, en su opinión, mejor refleja la realidad. Note que los pasos 4 y 5 permiten evaluar el riesgo de estar errado en la percepción de las probabilidades previas y costos de clasificación errónea que se consideran realistas.

Si se desea emplear las funciones discriminantes con fines de clasificación de nuevos casos, de tal manera que la regla correspondiente se comporte de manera similar a la regla de Bayes cuando hay igualdad entre costos de clasificación errónea e igualdad entre probabilidades de pertenencia previas, entonces se recomienda seguir los siguientes pasos:

a) Transforme las variables para inducir (multi)normalidad e igualdad de matrices de covariancia.

b) Pruebe la hipótesis de igualdad de matrices de covariancia. El rechazo de la hipótesis debe conducir al ejercicio de gran cautela en la aplicación de la regla de clasificación, por ejemplo, la efectividad de la regla debe evaluarse a posteriori mediante el cálculo de la tasa de clasificación errónea.

c) Pruebe la hipótesis de igualdad de medias poblacionales. Si la hipótesis es mantenida, cuando hay igualdad de matrices de covariancia, entonces no tiene sentido proseguir con la clasificación; los nuevos casos pueden asignarse a cualquier población.

d) Aplique la regla correspondiente y evalúe el desempeño de la misma, mediante del cálculo de la tasa de clasificación errónea.

Bibliografía

[1] DeGroot Morris H. (1975) *Probability and Statistics*, Capítulo 8. Addison–Wesley Publishing Company, Nueva York.

[2] Johnson Richard A.; Wichern Dean W. (1988) *Applied Multivariate Statistical Analysis*, Capítulo 11. Segunda Edición. Prentice Hall, Englewood-Cliffs, Nueva Jersey.

[3] Mora Oconitrillo Belia, Ramos de Anaya Alfaro Pilar (1983) *Técnica de Análisis Discriminante para la Selección de Proyectos en la Vicerrectoría de Investigación de la UCR.* Proyecto curso Técnicas de Investigación, a cargo del Prof. F. Pastrana.

[2] Un índice de clasificación riesgosa difiere de la tasa de clasificación errónea usual, en que el índice al costo promedio (y no al porcentaje) de los errores de clasificación.

[4] Mora Oconitrillo Belia, Romero Jiménez Dulia (1983) *Análisis Discriminante Aplicado a la Deserción de los Estudiantes del Centro Regional de San Carlos de la UNED.* Proyecto curso Tópicos de Teoría Estadística, a cargo del Prof. F. Pastrana.

[5] Pastrana José F., Gold Harvey J. y Swartzel Kenneth R. (1992) *Escogencia del Sistema de Procesamiento para un Producto Alimenticio, Aplicando la Metodología del Análisis de Decisiones.* Artículo a ser presentado por el Prof. F. Pastrana a la III Jornada de Análisis Estadístico de Datos, del LINCE, Escuela de Estadística, UCR.

[6] Segura Brenes Eddy y Castro Rojas Orlando (1988) *Análisis Discriminante Lineal: Una Alternativa para la Construcción de un Calificador de Persistencia para Seguros de Vida.* Presentado a la I Jornada de Análisis Estadístico de Datos del LINCE y basado en el trabajo de graduación de estos profesionales, dirigido por el Prof. F. Pastrana.

Presentación de las Redes Neuronales: Aplicaciones al Análisis de Datos

Javier Trejos Zelaya*

Resumen

Presentamos las principales nociones que giran alrededor de las redes neuronales: neuronas formales, pesos sinápticos, leyes de aprendizaje, tipos de redes.
Enseguida presentamos en detalle las redes multicapas con aprendizaje supervisado, basadas en la retropropagación del gradiente. También se presentan las redes de Kohonen y de Hopfield y otros modelos vecinos.
En todo el artículo se mencionan aplicaciones de las redes neuronales al Análisis de Datos: clasificación, discriminación, análisis en componentes principales, regresión lineal, etc.

Palabras clave: retropropagación del gradiente, aprendizaje competitivo, discriminación, clasificación, análisis en componentes principales, regresión lineal.

1 Introducción

Las redes neuronales constituyen uno de los temas de investigación más populares actualmente, que se encuentra en el cruce de caminos entre las Matemáticas, la Computación, la Ingeniería, la Biología, las Ciencias Cognoscitivas, la Física y otras ciencias.

Basados en investigaciones en Neurobiología, se han propuesto modelos de físicos que simulan el funcionamiento de las neuronas del cerebro humano. Los distintos enfoques asumidos, han conducido a la formulación de varios modelos.

Existen también redes neuronales que sólo han utilizado las neuronas formales como un instrumento de cálculo, sin referencia a ningún modelo biológico propiamente dicho, sino simplemente por su utilidad. A partir de allí, se han planteado varios tipos de redes neuronales.

Lo importante, desde el punto de vista de las Matemáticas Aplicadas, es que las redes neuronales sirven para resolver problemas concretos. En este sentido, el analista de datos y el estadístico encontrarán que este campo de investigación es muy fecundo. Y el usuario del Análisis de Datos y la Estadística, ya sea en Medicina, en Economía, en las Ciencias Sociales, en Agronomía u otras disciplinas, encontrará herramientas muy útiles y eficaces.

*Escuela de Matemática, Universidad de Costa Rica

Las redes neuronales (también llamadas modelos conexionistas o de Inteligencia Artificial de bajo nivel) presentan una ventaja fundamental respecto a los modelos clásicos: el paralelismo. Por su estructura misma, las redes neuronales son paralelas. En vista de que las computadoras tienden a ser capaces de hacer grandes cálculos en paralelo, a corto plazo esta ventaja se verá plasmada en términos de economía de tiempo.

El número de congresos y de revistas internacionales especializadas, en este campo, son cada vez más numerosos. Esto muestra el interés y la actualidad de este tema de investigación.

2 Las Redes Neuronales

Una red de neuronas formales o, más simplemente, una **red neuronal** es un grafo orientado cuyos nodos son llamados neuronas formales, y cuyas aristas ponderadas son llamadas **sinapsis**.

La idea inspiradora es la teoría neurobiológica del cerebro humano [6, 7, 9, 28]; así, una red neuronal simulará el funcionamiento de éste, en el sentido que:

- habrá un gran número de neuronas formales que van a recibir información y a hacer cálculos;
- habrá un gran número de interconexiones entre las neuronas.

Debe notarse que, como en el cerebro, muchos de esos cálculos podrán hacerse simultáneamente. Por lo tanto, las redes neuronales serán muy útiles cuando se disponga de una computadora capaz de hacer cálculos en paralelo.

Una neurona formal, o más simplemente, una **neurona**, es una unidad de cálculo (denotada neurona j) que recibe n *entradas* $x_1, \ldots, x_n$ por medio de n sinapsis o arcos ponderados por ciertos **pesos sinápticos** (o simplemente pesos) $\omega_{j1}, \ldots, \omega_{jn}$. La neurona aplica una función $h(x_1, \ldots, x_n)$, que generalmente es del tipo suma ponderada más un parámetro de control:

$$h(x_1, \ldots, x_n) = \sigma_j = \sum_{i=1}^{n} \omega_{ji} x_i + \theta_j$$

donde θ_j es el parámetro de control de la neurona j. Sin embargo, h también podría ser una función booleana, polinomial, etc. Sobre $h(x_1, \ldots, x_n)$ se aplica una *función de transferencia* real f, cuya acción denotamos $f(\sigma_j)$, que determina el **estado** de la neurona. Esta función es generalmente una función característica, escalera, sigmoide o estocástica. Cuando los estados de la neurona son 0 ó 1, se dice que la neurona está *inhibida* o *excitada*, respectivamente.

La salida de la neurona es dada por una *función de salida* g, que se aplica sobre f y que va a actuar sobre otras neuronas de la red. Denotaremos $s_j = g(f(\sigma_j))$ la salida de la neurona j.

Las definiciones y conceptos anteriores son ilustrados en la figura 1.

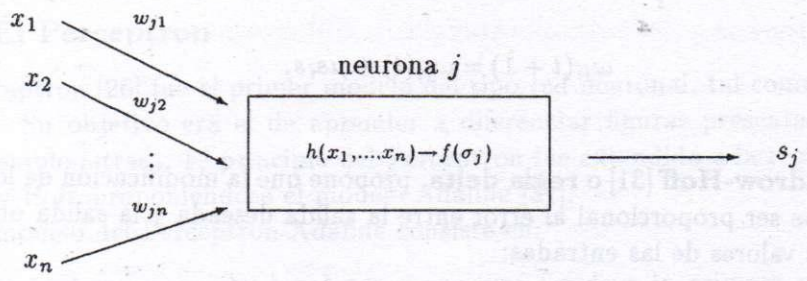


Figura 1: Representación de una neurona formal

El **estado** de una red neuronal está dado por el conjunto de los pesos sinápticos y el conjunto de los estados de las neuronas.

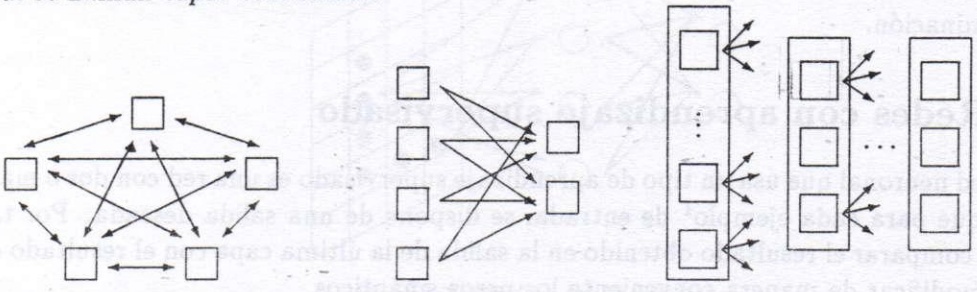
La arquitectura o estructura del grafo que define a una red neuronales es lo que determina la *estructura* de la red. Las estructuras más usadas son:

a) red totalmente conectada (figura 2.a)

b) red con dos capas (figura 2.b)

c) red multicapas o red con más de dos capas (figura 2.c)

Cada capa es un conjunto de neuronas que pueden ser conectadas o no entre ellas. En el caso de una red multicapas, las capas intermedias, es decir, todas, excepto la primera y la última, se llaman *capas escondidas*.



(a) Red totalmente conectada (b) Red con dos capas (c) Red multicapas

Figura 2: Tipos de estructuras de redes de neuronas

Dependiendo de la aplicación que se estudie, se escogerá la estructura de la red y la dinámica de las conexiones, es decir, la forma cómo cambian los pesos sinápticos en el tiempo conforme se introduce la información. Así, se puede definir una forma de *aprendizaje* que determina la modificación de los pesos.

Las dos formas de aprendizaje más usadas están inspiradas en la neurobiología y son:

- La **ley de Hebb** [10], que propone que cuando dos neuronas están excitadas al mismo tiempo, hay que modificar los pesos sinápticos para reforzar esta excitación simultánea. Por ejemplo, si $\omega_{ji}(t)$ (resp. $\omega_{ji}(t+1)$) denota el peso sináptico entre la neurona i y la neurona j en el tiempo t (resp. $t+1$) y si s_i y s_j son las salidas respectivas de las neuronas i y j con $s_i, s_j \in \{0.1\}$ entonces:

$$\omega_{ji}(t+1) = \omega_{ji}(t) + \mu s_i s_j$$

donde $\mu > 0$.

- La **ley de Widrow-Hoff** [31] o **regla delta**, propone que la modificación de los pesos sinápticos debe ser proporcional al error entre la salida deseada y la salida obtenida, así como a los valores de las entradas:

$$\omega_{ji}(t+1) = \omega_{ji}(t) + \eta \frac{\partial(\text{error})}{\partial(\omega_{ji})} x_i$$

donde $\eta > 0$, error $= 1/2$ (salida deseada - salida obtenida)^2 y ∂ denota la derivada parcial. Debe notarse que esta ley es aplicable cuando se dispone de un conjunto de salidas deseadas para cada patrón de entrada (en las neuronas de la capa de salida en una estructura multicapas), como es el caso en discriminación o reconocimiento de patrones.

El resto de esta exposición la dividiremos en dos partes. Primero estudiaremos las redes neuronales que usan leyes de aprendizaje supervisado, como la regla delta, por ser las redes más estudiadas hasta ahora gracias a los trabajos sobre retropropagación del gradiente. Luego se presentarán otros modelos de redes, como los de Kohonen y Hopfield, pioneros en el estudio de las redes neuronales. Estos modelos nos permitirán introducir algunas redes empleadas por diversos autores para abordar problemas de Análisis de Datos, como la determinación de componentes principales, la clasificación automática de objetos y la discriminación.

3 Redes con aprendizaje supervisado

Una red neuronal que usa un tipo de aprendizaje supervisado es una red con dos o más capas, en la que para cada ejemplo[1] de entrada, se dispone de una salida deseada. Por tanto, se puede comparar el resultado obtenido en la salida de la última capa con el resultado deseado y así modificar de manera conveniente los pesos sinápticos.

Este tipo de situaciones se encuentran frecuentemente en discriminación: la pertenencia de los individuos (por ejemplo, enfermos) a ciertas clases (por ejemplo, enfermedades) de acuerdo a lo observado en ciertas variables (por ejemplo, síntomas). Pero también es una situación típica del reconocimiento de patrones: reconocimiento de voz, de caracteres, de trazos, de imágenes.

Estas redes neuronales se usan frecuentemente para fijar los pesos sinápticos sobre un conjunto de aprendizaje, y luego hacer la predicción sobre los otros elementos de la población. Es el caso típico del diagnóstico médico o de la previsión meteorológica.

Si la red tiene sólo dos capas, se utiliza generalmente la regla de aprendizaje de Widrow-Hoff. Si tiene más capas, se utiliza la retropropagación del gradiente que veremos en la sección 3.2

[1] También llamado unidad estadística, individuo, patrón, objeto observado, estímulo, etc. según diferentes terminologías en distintas áreas de aplicación.

3.1 El Perceptron

El Perceptron [26] fue el primer modelo del tipo red neuronal, tal como las conocemos hoy en día. Su objetivo era el de aprender a diferenciar figuras presentadas en dimensión 2 (por ejemplo letras). El principio del Perceptron fue extendido a la regla de aprendizaje de Widrow-Hoff, proponiéndose el modelo Adaline [31].

El modelo del Perceptron-Adaline consiste en:

- una *retina* que recibe los datos y que está ligada a la primera capa mediante pesos fijos;
- una primera capa, llamada *capa de asociación*, con pesos ω_{ji} que la ligan a la capa de salida;
- una *capa de salida*, que da el estado obtenido para cada ejemplo;
- una regla de aprendizaje del tipo regla delta:

$$\omega_{ji}(t+1) = \omega_{ji}(t) + \eta(d_j - s_j)$$

donde d_j y s_j son respectivamente la salida deseada y la salida obtenida para la neurona j y $\eta > 0$;

- no hay conexiones al interior de una capa y las conexiones están orientadas de una capa a la siguiente (ver figura 3).

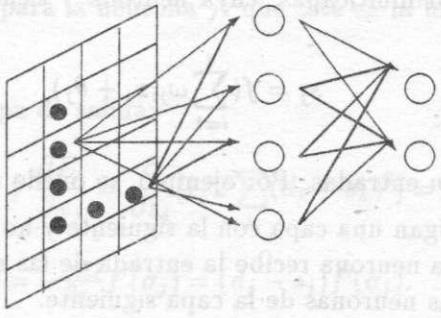


Figura 3: Modelo del Perceptron

Realmente, el Perceptron es un discriminador lineal que converge hacia la solución del problema *solamente si* ésta existe, es decir, si existe un conjunto de pesos que pueden discriminar linealmente entre los elementos de un conjunto dado.

En 1969, Minsky & Papert [19] estudiaron las limitaciones del Perceptron. Desde entonces, varios autores han abordado este problema [6, 21, 23, 28], especialmente usando el ejemplo del XOR (llamado también "o exclusivo"). La principal conclusión obtenida es que el Perceptron tiene muchas limitaciones para representar el conocimiento, lo que se traduce en el hecho que cuando no hay un hiperplano de separación el Perceptron no puede encontrar otro hipervolumen capaz de llevar a cabo esta separación.

Para paliar esta deficiencia, las redes multicapas y el aprendizaje por retropropagación del gradiente [16, 21, 28] han sido propuestos de manera alternativa.

3.2 Redes multicapas y la retropropagación del gradiente

Las redes multicapas necesitan de un algoritmo adaptado para el aprendizaje de los pesos sinápticos. Es decir, con la regla delta, sólo los pesos entre la penúltima y la última capa son modificados. Por ello, la regla delta generalizada o ley de retropropagación del gradiente fue propuesta por Le Cun [16] y poco tiempo después, pero de manera independiente, por Rumelhart y su equipo PDP [28]. Para ser operacional, necesita que la función de transferencia en cada neurona sea continua: por ello se usan las funciones sigmoides que dan una aproximación continua y diferenciable de las funciones características (ver figura 4).

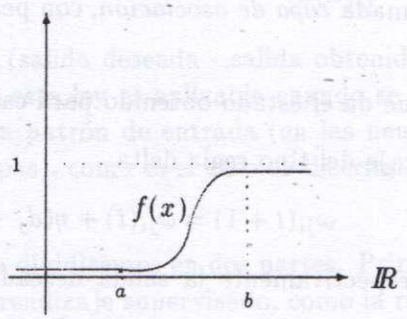


Figura 4: Ejemplo de función sigmoide

Se dispone de una red multicapas, cuya neurona j tiene por función de transferencia una función sigmoide:

$$s_j = f(\sum_{i=1}^{n} \omega_{ji} x_i + \theta_j)$$

donde $x_1 \ldots, x_n$ son las n entradas. Por ejemplo, se puede usar $f(x) = \dfrac{1}{1+e^{-x/\alpha}}$.

Los pesos sinápticos ligan una capa con la siguiente y no hay interconexiones al interior de una misma capa. Cada neurona recibe la entrada de las neuronas de la capa precedente y transmite su salida a las neuronas de la capa siguiente.

Durante la presentación de un ejemplo en el instante t, la señal pasa de una capa a la siguiente *hacia adelante*. El error es calculado en la capa de salida: si d_j es el resultado deseado para la neurona j y s_j el obtenido, el error es entonces:

$$E_t = 1/2 \sum_j (d_j - s_j)^2.$$

Efectuando una aproximación del descenso del gradiente respecto a los pesos, se obtiene que la ley de aprendizaje para los pesos sinápticos es:

$$\omega_{ji}(t+1) = \omega_{ji}(t) + \eta \delta_j x_i$$

donde ω_{ji} representa el peso de la neurona i hacia la neurona j.

Si j pertenece a la capa de salida, entonces

$$\delta_j = (d_j - s_j) f'(\sigma_j)$$

y si no,

$$\delta_j = f'(\sigma_j) \sum_k \delta_k \omega_{kj}$$

donde $\sigma_j = \sum_{i=1}^{n} \omega_{ji} x_i + \theta_j$ y k varía entre las neuronas de la capa que sigue. Se ve entonces que el aprendizaje se hace *hacia atrás*: la modificación de los pesos de las capas intermedias toma en cuenta las modificaciones de los pesos de las capas posteriores, hecho simbolizado por el término $\sum_k \delta_k \omega_{kj}$.

DEMOSTRACIÓN: Se propone que el cambio de cada peso, durante la presentación del ejemplo t, $\Delta_t \omega_{ji}$, sea proporcional a $-\dfrac{\partial E_t}{\partial \omega_{ji}}$, donde $E_t = 1/2 \sum_j (d_j - s_j)^2$.

Escribiendo $\dfrac{\partial E_t}{\partial \omega_{ji}} = \dfrac{\partial E_t}{\partial \sigma_j} \dfrac{\partial \sigma_j}{\partial \omega_{ji}}$ y como $\sigma_j = \sum_i \omega_{ji} x_i + \theta_j$ (donde x_i puede ser una entrada de la red si i está en la primera capa, o bien la salida de una neurona de la capa anterior si i está en alguna capa posterior), entonces $\dfrac{\partial \sigma_j}{\partial \omega_{ji}} = x_i$, y si denotamos $\delta_j = -\dfrac{\partial E_t}{\partial \sigma_j}$ entonces $-\dfrac{\partial E_t}{\partial \omega_{ji}} = \delta_j x_i$.

Por tanto $\Delta_t \omega_{ji} = \eta \delta_j x_i$, con $\eta \in \mathbb{R}^+$.

Ahora bien, si escribimos $\delta_j = -\dfrac{\partial E_t}{\partial \sigma_j} = -\dfrac{\partial E_t}{\partial x_j} \dfrac{\partial x_j}{\partial \sigma_j}$ el segundo término es $\dfrac{\partial x_j}{\partial \sigma_j} = f'(\sigma_j)$ pues j no es una neurona de la primera capa y $x_j = s_j = f(\sigma_j)$.

Se pueden presentar dos casos para la neurona j: que esté en la última capa o en una capa escondida.

i) 1er. caso: j está en la capa de salida:

$$x_j = s_j \text{ y } \frac{\partial E_t}{\partial x_j} = \frac{\partial}{\partial s_j}[1/2 \sum_k (d_k - s_k)^2] = s_j - d_j$$

lo que implica $\delta_j = -\frac{\partial E_t}{\partial \sigma_j} = -\frac{\partial E_t}{\partial x_j} f'(\sigma_j) = (d_j - s_j) f'(\sigma_j)$.

ii) 2o caso: j está en una capa escondida:

sean los σ_k asociados a las neuronas de la capa que sigue, entonces $\dfrac{\partial \sigma_k}{\partial s_j} = \omega_{kj}$ y podemos escribir:

$$-\frac{\partial E_t}{\partial x_j} = -\frac{\partial E_t}{\partial s_j} = -\sum_k \frac{\partial E_t}{\partial \sigma_k} \cdot \frac{\partial \sigma_k}{\partial s_j} = -\sum_k \delta_k \omega_{kj}$$

donde k varía entre las neuronas de la capa siguiente.

Así $\delta_j = f'(\sigma_j) \sum_k \delta_k \omega_{kj}$.

Se obtiene entonces el resultado anunciado. □

Esta fórmula de aprendizaje permite calcular los pesos sinápticos de atrás hacia adelante, y provee un descenso del gradiente. Evidentemente, es posible que el método converja hacia un mínimo local.

En general, la aplicación del algoritmo de retropropagación del gradiente necesita un gran número de pasos sobre el conjunto de los individuos o ejemplos. Sin embargo, debido a las representaciones internas creadas por las capas intermedias, el algoritmo permite resolver satisfactoriamente varios problemas de discriminación [6, 16, 21, 23, 25, 28, 29]. En efecto, en muchos casos las capas intermedias tienen interpretaciones interesantes y permiten hacer discriminaciones cuadráticas, cúbicas, etc.

Por el momento, aún no hay una teoría definitiva que permita escoger *a priori* una arquitectura apropiada para cada caso a tratar; son más bien la heurística y la experiencia las que dan una idea de la escogencia del número de capas y del número de neuronas en cada capa. Sin embargo, puede consultarse el artículo de Yves Lechevallier en estas mismas Memorias, sobre la construcción de una red neuronal utilizando la segmentación.

3.3 Aplicaciones

a) Análisis en Componentes Principales

Muller y Radoui [20] propusieron una red con tres capas para calcular las componentes principales de una tabla de datos de n individuos y p variables, utilizando la retropropagación del gradiente.

La tabla de datos puede ser considerada como una nube de n puntos en el espacio euclidiano de dimensión p, provisto de un producto escalar o métrica M, y tal que los individuos están equiponderados. El Análisis en Componentes Principales (A.C.P.) consiste en buscar un espacio de dimensión q ($q < p$) que reconstruya lo mejor posible la configuración de la nube inicial. Se sabe que la solución de este problema está dada por el espacio generado por q vectores propios $v_1, \ldots, v_q$ (llamados vectores principales) asociados a los primeros q valores propios de VM, donde V es la matriz de varianzas-covarianzas de las p variables.

La red de tres capas propuesta para resolver este problema tendrá p neuronas en las capas de entrada y de salida y q neuronas en la capa escondida.

Las neuronas de la capa de entrada reciben las líneas x_i de la tabla de datos, y la salida deseada será x_i mismo. Es decir, queremos verificar si los vectores principales reconstruyen efectivamente la nube de puntos.

Los pesos sinápticos entre la capa de entrada y la capa escondida serán combinaciones lineales de las coordenadas de los q primeros vectores principales.

b) Discriminación

El problema de la discriminación puede ser presentado como el de encontrar criterios o reglas para decidir si un individuo medido por un conjunto de variables, pertenece o no a alguna de las clases de la población dadas a priori.

Es claro que se pueden idear algoritmos basados en el aprendizaje supervisado para tratar los problemas de discriminación. Las neuronas de la capa de salida representarían las clases que se quieren discriminar y tendrán valor 0 ó 1. Varias aplicaciones [5, 20, 21, 23, 25] de soluciones a problemas de discriminación lineal (usando un modelo similar al Perceptron) o no lineal, han sido propuestas.

c) Regresión lineal

Se puede ver una gran analogía entre el aprendizaje supervisado usando la regla de Widrow-Hoff y la regresión lineal múltiple (contribución de S.Stone en [28]).

En la regresión lineal múltiple se quiere prever una variable cuantitativa y por un conjunto de variables cuantitativas $x^1, \ldots, x^q$ en $\mathbb{R}^n$. Para ello, se quiere encontrar un conjunto de coeficientes $b = (b_1, \ldots, b_n)^t$ tal que $\hat{y}_k = b^t x_k$ donde $\hat{y}$ minimiza

$$\|y - \hat{y}\|^2 = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2.$$

Se sabe que el estimador de mínimos cuadrados de b está dado por $\hat{b} = (X^t X)^{-1} X^t y$, donde los x^i son las columnas de X.

Tomemos:

- el valor de la variable a explicar y_k como el valor deseado d_k asociado a un ejemplo k;
- los valores alcanzados para un ejemplo k sobre las variables explicativas $x_k^1, \ldots, x_k^q$ como las entradas de una red neuronal;
- los coeficientes $b = (\omega_{j1}, \ldots, \omega_{jn})^T$ como los pesos sinápticos;
- la función de transferencia f como la identidad.

Si denotamos $W(t)$ la matriz que contiene los pesos cuya entrada (j, i) es ω_{ji}, y si se usa la regla de aprendizaje de Widrow-Hoff, se prueba [28] que $W(t)$ converge efectivamente hacia el estimador $\hat{b}$, es decir,

$$\lim_{t \to \infty} W(t) = \hat{b} = (X^T X)^{-1} X^T y$$

si las variables x^j son linealmente independientes.

d) Mercadeo

En una compañía de venta por correspondencia, se dispone de una tipología de 400 individuos o clientes, y de una tabla de datos que contiene las características de estos clientes así como las compras que han efectuado en la compañía.

Con el fin de identificar diferentes comportamientos de consumo, se proponen tres tipos de redes neuronales [29] con fines de discriminación. Se trata de redes con dos capas que usan la ley de Widrow-Hoff y como función f de transferencia una función característica. Los estados de las neuronas de salida son 1 ó 0, según el individuo pertenezca o no a la clase de la tipología representada por la neurona.

i) Primer caso: se quiere identificar un tipo de cliente.

- entrada: una neurona por característica de compra (139)
- salida: una neurona por clase (10)

ii) Segundo caso: se quieren identificar los comportamientos asociados a clientes existentes.

- entrada: una neurona por característica de compra (139)
- salida: una neurona por individuo (400)

iii) Tercer caso: se quiere determinar el posible comportamiento de compra de un nuevo cliente.

- entrada: una neurona por característica (85)
- salida: una neurona por comportamiento de compra(47).

e) Otros modelos

Fukushima [8] propuso un modelo, el Neocognitron, que consiste en varias capas con distintas funciones. Es un modelo que ha tenido mucho éxito en el reconocimiento de imágenes por sus cualidades de representación del conocimiento en las capas intermedias.

Pao [23] propone implementar, sobre una red multicapas, un funcional que amplíe la representación inicial de la información. La nueva representación será en un espacio de dimensión mayor y no sólo las combinaciones lineales entre las salidas de las neuronas serán permitidas, sino también las multiplicaciones, las funciones trigonométricas, logarítmicas, exponenciales u otras. El autor ha obtenido resultados promisorios pero aún no es posible decidir *a priori* el modelo funcional ideal para cada caso a tratar.

4 Otros tipos de redes neuronales

Hay muchos otros tipos de redes neuronales que han sido estudiados, entre los que destacan los modelos de Kohonen, de Hopfield y de aprendizaje competitivo. Nuestro mayor interés será el de presentar aquellos que conduzcan a aplicaciones en el Análisis de Datos, especialmente en el cálculo de componentes principales, en la determinación de clases para una clasificación automática o en discriminación.

4.1 El modelo de Kohonen

Basado en observaciones neurobiológicas, las redes propuestas por Kohonen [14] tienen las innovaciones siguientes:

- se introduce un término de olvido en la ley de aprendizaje;
- hay conexión lateral entre neuronas de una misma capa;
- hay inhibición lateral del tipo "sombrero mexicano", es decir, las neuronas más excitadas por otra neurona son aquéllas que le son más cercanas, mientras que las más lejanas son inhibidas.

Así la ley de aprendizaje es:

$$\omega_{ji}(t+1) = \omega_{ji}(t) + \alpha s_j x_i - \beta(s_i)\omega_{ji}(t)$$

donde $\alpha \geq 0$, $\beta(s_i) \geq 0$ es el término de olvido y $s_j = f(\sum_i \omega_{ji} x_i)$ es el estado de la neurona j, siendo f una función sigmoide.

El modelo de Kohonen ha sido utilizado con éxito para dar soluciones prácticas al problema del viajero de comercio. Asimismo, ha sido utilizado para resolver problemas que admiten un "mapa topológico" del problema (representado por las conexiones entre las neuronas).

a) Aplicación a la clasificación

Este modelo consiste en escoger la neurona que tenga la respuesta más fuerte, ante la presencia de un estímulo. Dados los x_i que forman el vector x, se procede como sigue:

1) encontrar c tal que $\|x - \omega_c\| = \min_{j=1,\ldots,n} \|x - \omega_j\|$ donde $\omega_j = (\omega_{j1}, \ldots, \omega_{jn})^T$ son los pesos de las neuronas que se conectan hacia la neurona j;

2) calcular el vecindario V_c alrededor de c tal que: si $j \in V_c \Rightarrow s_j = 1$
si $j \notin V_c \Rightarrow s_j = 0$

3) actualizar los pesos: $\omega_{ji}(t+1) = \omega_{ji}(t) + k(t)[x_i - \omega_{ji}(t)]$
donde $k(t) \neq 0$ si $i \in V_c$ y $k(t) = 0$ si $i \notin V_c$.

Este procedimiento permite localizar la neurona adecuada a cada entrada y refuerza esta adecuación. En el sentido en que el sistema escoge la neurona que responde más fuerte, este modelo puede ser visto como un clasificador automático.

b) Aplicación a la discriminación

Una idea similar puede ser usada en discriminación. Sobre el conjunto de aprendizaje, hacer lo siguiente:

1) representar la clase C_j por un número de neuronas proporcional a la probabilidad *a priori* de esta clase;

2) inicializar los pesos con los primeros ejemplos: $\omega_j = x$ si $x \in C_j$;

3) presentar el ejemplo x y escoger C tal que:

$$\|\omega_c - x\| = \min_j \|\omega_j - x\|;$$

4) actualizar los pesos mediante el aprendizaje siguiente:

- si $x \in C : \omega_c(t+1) = \omega_c(t) + k(t)[x - \omega_c(t)]$
- si $x \notin C : \omega_c(t+1) = \omega_c(t) - k(t)[x - \omega_c(t)]$
- si $i \neq C : \omega_i(t+1) = \omega_i(t)$

donde $k(t) \geq 0$ y k decrece conforme t crece.

Puede verse que se trata de un caso de aprendizaje supervisado, pero distinto a los modelos presentados en la sección 2.

c) Cálculo de componentes principales

Oja [22] propuso una red del tipo de Kohonen para el cálculo de vectores propios de una matriz de correlaciones. Al sistema se introducen vectores que representan los valores que toma una variable cuantitativa sobre los individuos, y se considera la ley de aprendizaje

$$\omega_{ji}(t+1) = \omega_{ji}(t) + \gamma\eta(t)[x_i(t) - \eta(t)\omega_{ji}(t)$$

con $\gamma > 0$, cuyo término de olvido es $\eta(t)\omega_{ji}(t)$. Se demuestra [22] que los pesos sinápticos convergen, cuando t crece, al primer vector propio de la matriz de correlaciones.

Basado en las ideas de Oja y en el algoritmo de aproximación estocástica [3, 15] para el cálculo de los elementos principales de un ACP o un Análisis de Correspondencias, Lelu [17] propuso un algoritmo que permite calcular todas las componentes principales.

El algoritmo de aproximación estocástica consiste en dar una fórmula de actualización de los elementos principales ante la presentación de un nuevo individuo. Por tanto, permite el cálculo de los elementos principales conforme entran los individuos, lo que hace posible la realización de un ACP de una población muy grande pues no es necesario almacenar ninguna tabla de datos en memoria central. Se demuestra [15] que la sucesión dada por la fórmula de actualización converge al primer vector propio de la matriz a diagonalizar.

El algoritmo neuronal de Lelu usa por ley de aprendizaje:

$$\omega_{ji}(t+1) = \omega_{ji}(t) + \eta(t)(\sum_k \omega_{jk}x_k)[x_i(t) - x_i(t)\omega_{ji}(t)/\|m\|^2]$$

donde $\|m\|$ es la norma deseada para los pesos ω_{ji}.

El autor propone además un método para simular el proceso de ortogonalización de Gram-Schmidt con el fin de poder calcular otros vectores principales, además del primero. Las razones que da Lelu para explicar por qué este algoritmo converge no son muy claras. Sin embargo, su método tiene el mérito de poder extenderse al cálculo de semi-ejes principales, de ejes oblícuos y de clases similares a las que da el método de nubes dinámicas. Finalmente, esta red neuronal es generalizable al caso de variables continuas que varían con el tiempo [27]. Una curiosidad de este modelo es que se pueden crear o eliminar neuronas conforme se avanza en el algoritmo.

4.2 El modelo de Hopfield

Hopfield, un reconocido físico, propuso [12] una red neuronal original y que dió una segunda vida a la investigación en este campo, luego de que Minsky y Papert pusieran al descubierto las debilidades del Perceptron.

La red de Hopfield es una red totalmente conectada y los estados de las neuronas pueden ser 0 ó 1. Si $s_i(t)$ denota el estado de la neurona i en el tiempo t, sea $\sigma_j = \sum_i \omega_{ji}\, s_j(t)$ la entrada total a la neurona j, donde ω_{ji} representa el peso sináptico entre i y j. Entonces el nuevo estado de la neurona será $s_j(t+1) = \begin{cases} 1 & \text{si} \quad \sigma_j > \tau \\ 0 & \text{si} \quad \sigma_j \leq \tau \end{cases}$

donde τ es un umbral fijado.

El tiempo para realizar los cambios de estado de las neuronas, puede ser aleatorio o a intervalos regulares. Además, las neuronas pueden cambiar de estado simultáneamente, secuencialmente o mediante una escogencia al azar.

Este modelo es útil para memorizar un conjunto de estados o prototipos; es decir, estados estables del sistema que tengan un papel de atracción sobre los otros estados. Así, ante la presentación de m prototipos, el peso ω_{ji} es recalculado de tal forma que:

$$\omega_{ji} = \sum_{k=1}^{m} (2x_i(k) - 1)(2x_j(k) - 1)$$

donde $x_j(k)$ es el valor que toma el k-ésimo prototipo en la neurona j.

Puede verse que esta ley de aprendizaje es del tipo de Hebb y no depende de los pesos en el tiempo anterior.

Gracias a una modelización física [6], se ha estimado que el número de prototipos que puede memorizar una red de n neuronas es $m \approx (0.14)n$.

El modelo de Hopfield permite realizar memorias auto-asociativas: una entrada ligeramente modificada o incompleta converge en general al estado estable esperado. Los principales defectos de este tipo de red son que puede encontrar estados parásitos (que no son ninguno de los prototipos esperados) y que en ocasiones cae en un "olvido catastrófico": si se trata de memorizar más estados que la proporción arriba indicada, el sistema puede olvidar todos los estados estables.

a) Optimización de una forma cuadrática

El modelo de red neuronal de Hopfield ha sido utilizado para optimizar una función que se pueda describir mediante una forma cuadrática.

Si se introduce la función de energía H tal que para todo estado $x = (x_1, \ldots, x_n)^T$,

$$H(x) = -1/2 \sum_i \sum_j \omega_{ji} x_j x_i$$

se demuestra que H es decreciente [6] siguiendo la regla de evolución antes descrita. Por lo tanto, todos los estados estables de la red son mínimos locales de H. Ahora bien, la función de energía arriba definida es una forma cuadrática. Así, si en un sistema parametrado por n variables binarias de estado se define una función de costo H que sea una forma cuadrática simétrica, entonces la minimización de H puede resolverse mediante una red de Hopfield. Para ello basta construir una red cuyos pesos sinápticos ω_{ji} sean los coeficientes de la forma cuadrática H.

Puede verse que la solución al problema puede hacerse mediante un algoritmo totalmente paralelo. La convergencia hacia un mínimo global puede hacerse con la implementación del método del sobrecalentamiento simulado, del que hablaremos enseguida.

Varios problemas de gran complejidad han sido tratados de esta forma (por ejemplo, el problema del viajero de comercio), obteniendo resultados satisfactorios.

El sobrecalentamiento simulado

La técnica del sobrecalentamiento simulado, es usada para definir la Máquina de Boltzmann, un tipo de red neuronal que veremos a continuación. Esta técnica fue propuesta en 1983 por Kirkpatrick, Gelatt y Vecchi [13] con el fin de calcular el óptimo global de un problema de optimización combinatoria. Está basado en una analogía con la termodinámica: el método

de sobrecalentamiento (*annealing* en inglés o *recuit* en francés), en metalurgia, consiste en aumentar fuertemente la temperatura de un cristal y hacerla decrecer poco a poco, con el fin de obtener un cristal muy puro (es decir, muy estable).

Desde el punto de vista de la minimización de una función real f, la analogía consiste en introducir un parámetro T que juegue el papel de la temperatura. Entonces, en vez de sólo aceptar nuevas soluciones que hagan disminuir f (como en el caso del descenso del gradiente clásico), se permitirán algunos aumentos de f, con una probabilidad que dependerá de T: entre mayor sea T más aumentos de f serán aceptados, conforme T decrezca, estos aumentos serán cada vez más raros. Esto hará posible que, si el estado inicial x_0 estaba en un "valle" definido por un mínimo local x_1, se podrá subir la "colina" para, eventualmente, caer en el "valle" que define el mínimo global x^* (ver figura 5)

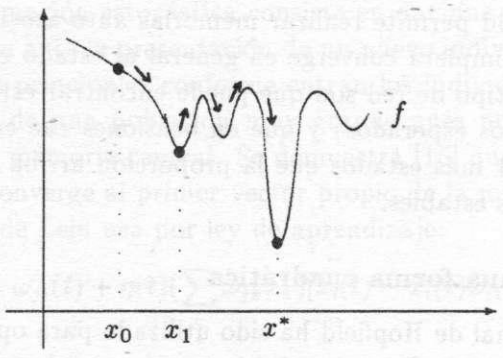


Figura 5: Minimización mediante el sobrecalentamiento simulado

Se ha demostrado [1] que, asintóticamente, el sobrecalentamiento simulado converge a un óptimo global de la función a optimizar. Tal demostración se basa en la teoría de las cadenas de Markov, homogéneas y no homogéneas.

Esta técnica de optimización permite entonces mejorar la dinámica de las conexiones de una red de Hopfield, pues con ello se evitan los mínimos locales.

b) La máquina de Boltzmann

La máquina de Boltzmann consiste en una implementación particular de las redes de Hopfield, usando el sobrecalentamiento simulado, con el fin de mejorar la dinámica de las conexiones (es decir, de los pesos sinápticos).

Hinton [11] propuso un algoritmo muy original que usa el sobrecalentamiento simulado, cuya idea es construir una red tal que la distribución de probabilidad definida por los estados estables de la red sea igual a la distribución de probabilidad definida por los prototipos.

Así, dados p prototipos o vectores binarios de dimensión n, se quiere memorizarlos en una red de Hopfield. Los prototipos definen una distribución de probabilidad (cada uno con probabilidad $1/p$) en el conjunto de 2^n posibles prototipos. Se considera una red de Hopfield con n neuronas que funcionan según el principio del sobrecalentamiento simulado: es decir, no sólo disminuciones de la energía son aceptadas sino también aumentos, con una probabilidad que depende de un parámetro de temperatura.

Si se modifican los pesos sinápticos habrá una modificación de la energía y por lo tanto, de la probabilidad del estado. Ahora bien, para forzar el estado a seguir la distribución dada por los prototipos, es necesario introducir nuevas neuronas. Esto se puede ver pues al modificar un peso sináptico ω_{ji}, sólo se puede modificar la *correlación* entre la actividad de i y j: es decir, la probabilidad de que i y j estén simultáneamente activas.

Se añaden entonces m nuevas neuronas (llamadas neuronas escondidas) y se introduce un modo de *funcionamiento forzado*, que ayudará a estimar la distribución sobre las $n + m$ neuronas: se fijan las primeras n neuronas y se varían los estados de las m nuevas, hasta que haya estabilidad. Luego, se deja el sistema en *funcionamiento libre*: todas las neuronas varían. Finalmente, luego de iterar varias veces los dos tipos de funcionamiento, se trata de igualar las distribuciones de probabilidad de los estados, en modo libre y forzado, minimizando la distancia entre ambas distribuciones.

Puede verse que la implementación de una máquina de Boltzmann requiere de muchos parámetros y de una gran inversión en términos de tiempo de cálculo. Es un método original, que fue utilizado después de su introducción en reconocimiento de patrones, pero que ha ido cayendo en desuso debido a su alto costo de implementación.

4.3 Modelos para Clasificación Automática

Muchos han sido los autores que han propuesto modelos de redes neuronales con el fin de encontrar particiones o tipologías en una población. A continuación presentamos tres de estos modelos, distintos de la red de Kohonen antes presentada.

a) Aprendizaje competitivo

La red propuesta por Rumelhart y Zipser (ver [28]) es una red multicapas y al interior de cada capa hay clases. Las neuronas de una clase inhiben a las otras neuronas de la misma clase. Esta red es ilustrada en la figura 6.

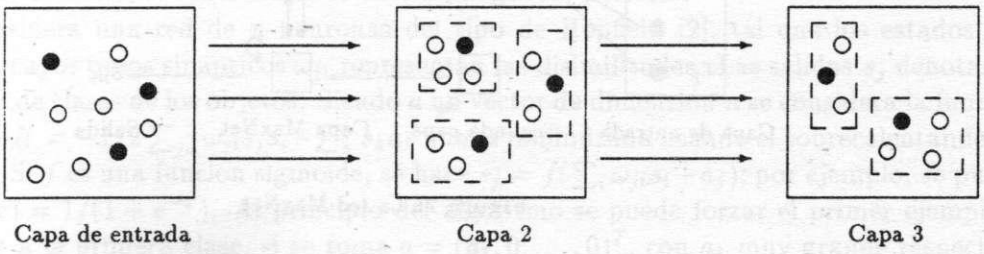


Figura 6: Red de aprendizaje competitivo

Los pesos ω_{ji} que conectan las neuronas i a la neurona j deben cumplir $\sum_i \omega_{ji} = 1$ y las entradas del sistema son sucesiones binarias. Si denotamos s_{ik} el estado de la neurona i durante la presentación del ejemplo x_k (1 si i está activa, 0 si no), entonces la competencia consiste en encontrar, para cada clase de una capa, la neurona j que maximiza $\alpha_{jk} = \sum_i \omega_{ji} s_{ik}$. Entonces el nuevo estado de j será 1 si j gana en su clase, 0 si no.

La ley de aprendizaje será entonces:

$$\omega_{ji}(t+1) = \begin{cases} \omega_{ji}(t) + \eta[\frac{s_{ik}}{n_k} - \omega_{ji}(t)] & \text{si } j \text{ gana cuando se presenta } x_k, \\ \omega_{ji}(t) & \text{si no.} \end{cases}$$

donde $\eta > 0$ y $n_k = \sum_i s_{ik}$.

Se ve que una neurona aprende traspasando parte del peso de las entradas inactivas a las activas. La neurona pierde una proporción η de su peso, que es luego distribuida entre los pesos de las entradas activas. Existe una variante de la ley de aprendizaje, que permite también modificar los pesos a las neuronas que pierden la competencia, con una proporción η_p mucho menor que la proporción η_g de las neuronas que ganan.

Este modelo ha sido utilizado con éxito para el reconocimiento de letras, palabras y trazos horizontales y verticales.

b) El modelo de Grossberg

Este modelo [9, 23] también usa un tipo de aprendizaje competitivo, pero la estructura de la red es distinta de la anterior. Además, se usa la inhibición lateral entre las neuronas. El modelo está basado en el funcionamiento del algoritmo MaxNet, que consiste en escoger la neurona que responde más fuerte a un estímulo. La capa de entrada tiene p neuronas y cada entrada $x = (x_1, \ldots, x_p)$ es un vector binario. Hay dos capas escondidas, cada una con m neuronas. La estructura de la red MaxNet es ilustrada en la figura 7.

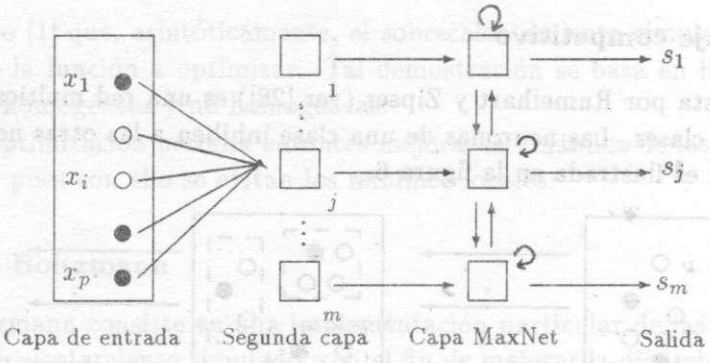


Figura 7: La red MaxNet

Si y es el prototipo (binario) de la clase C, el ejemplo x pertenecerá a la clase C^* que minimiza $d(x, y^*)$, donde d es la distancia de Hamming: $d(x, y) = p - \sum_i x_i y_i$.

Esto se implementa en la red MaxNet al definir la salida de las neuronas de la segunda capa como $\sigma_j = \sum_i \omega_{ji} x_i$. Si ω_{jk} es el peso entre las neuronas j, k de la capa MaxNet, se define $u_{jk} = 1$ si $j = k$ y $u_{jk} = -\epsilon$ si $j \neq k$, donde $\epsilon < 1/m$ (ver figura 8).

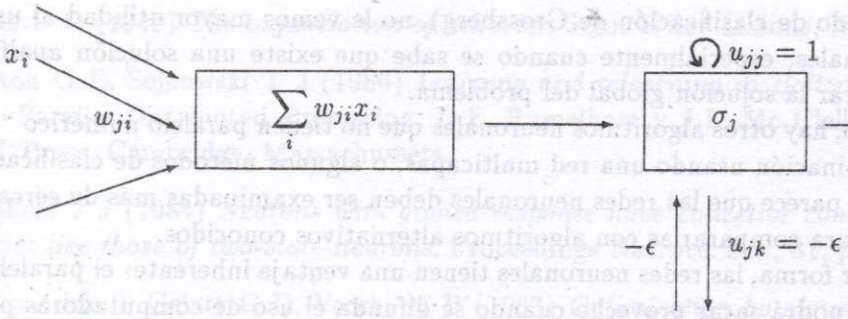


Figura 8: Neuronas de la red MaxNet

Se define entonces la salida de la neurona j como $s_j = f\left(\sigma_j - \epsilon \sum_{k \neq j} \sigma_k\right)$, donde f es una función sigmoide, y se dirá que hay convergencia cuando sólo un nodo tiene salida distinta de cero. Las entradas se presentan sucesivamente y reiteradamente hasta que se haya encontrado la clase a la que pertenece cada ejemplo. Esta clase será la dada por el prototipo más cercano según la distancia de Hamming.

El algoritmo dado por Grossberg para encontrar una partición está basado en la utilización de la red MaxNet. No daremos los detalles de este algoritmo, sólo haremos notar que, además de la inhibición lateral (simbolizada por los pesos u_{jk}) se definen pesos "hacia atrás" que van de la capa MaxNet a la capa de entrada y que permitirán verificar que un ejemplo pertenece a una clase, según un cierto grado de tolerancia τ. Además, τ jugará el papel de parámetro para determinar el número de clases (que no es fijado *a priori*). El problema con este método es que puede haber ejemplos que nunca se clasifiquen, por lo que habría que crear clases para ellos solos. Sin embargo el algoritmo se generaliza fácilmente al caso continuo, si se quiere usar una distancia euclídea, por ejemplo.

c) Modelo de Adorf-Murtagh

Se dispone de una matriz de distancias o disimilitudes y se quiere minimizar la disimilitud intraclases: es decir, obtener clases lo más homogéneas posibles.

Se considera una red de n neuronas del tipo de Hopfield [2], tal que los estados son binarios y cuyos pesos sinápticos ω_{ji} representan las disimilitudes. Las salidas s_j denotan la asignación de clases de los objetos. Siendo a un vector de dimensión n se considera la función de energía $H = -1/2 \sum_{j,i} \omega_{ji} s_j s_i - \sum_k s_k a_k$ que es minimizada usando el sobrecalentamiento simulado. Si f es una función sigmoide, se hace $s_j = f(\sum_i \omega_{ji} s_i + a_j)$; por ejemplo, se puede tomar $f(x) = 1/(1 + e^{-x})$. Al principio del algoritmo se puede forzar el primer ejemplo a pertenecer a la primera clase, si se toma $a = (a_1, 0, \ldots, 0)^T$, con a_1 muy grande respecto a las disimilitudes.

Conclusiones y expectaticas

En las diversas aplicaciones presentadas de las redes neuronales al Análisis de Datos, se ha visto que algunas de ellas son meras reformulaciones de los métodos clásicos conocidos. Para

lineal y el método de clasificación de Grossberg), no le vemos mayor utilidad al uso de los modelos neuronales, especialmente cuando se sabe que existe una solución analítica que permite encontrar la solución global del problema.

Sin embargo, hay otros algoritmos neuronales que no tienen paralelo numérico "clásico", como la discriminación usando una red multicapas, o algunos métodos de clasificación. En estos casos, nos parece que las redes neuronales deben ser examinadas más de cerca y hacer experimentos para compararlas con algoritmos alternativos conocidos.

De cualquier forma, las redes neuronales tienen una ventaja inherente: el paralelismo; de esta ventaja se podrá sacar provecho cuando se difunda el uso de computadoras paralelas. Sin embargo, también limitaciones evidentes; en especial, con el algoritmo de retropropagación del gradiente se obtiene un óptimo local. La consideración de métodos de optimización global, como el sobrecalentamiento simulado, los algoritmos genéticos o la búsqueda tabú podrían eventualmente considerarse. Por otra parte, hay un problema evidente con la escogencia del número de capas, o el número de neuronas en una capa. Quizás la investigación acerca de medidas de la redundancia de la información[2] pueda dar alguna luz a resolver esta cuestión.

Bibliografía

[1] Aarts E, Korst J (1989) *Simulated annealing and Boltzmann machines: a stochastic approach to combinatorial optimization and neural computing.* John Wiley & Sons, Chichester.

[2] Adorf H M, Murtagh F (1988) *Clustering based on neural network processing.* En: Compstat 88, IASC, Physica-Verlag, Heidelberg.

[3] Benzécri J P (1982) *L'approximation stochastique en Analyse des Correspondances.* En: Cahiers de l'Analyse des Données, vol. VII, n^o 4.

[4] Bourret P, Reggis J, Samuelidès M (1991) *Réseaux neuronaux: une approche connexioniste de l'Intelligence Artificielle.* Teknea, Toulouse.

[5] Chabanon C, Lechevallier Y, Milleman S (1992) *Proposition d'une construction efficace d'un réseau de neurones à partir d'un arbre de décision.* En: Actas de las III Journées Symbolique-Numérique. E. Diday & Y. Kodratoff (eds.). Univ. Paris-Dauphine.

[6] Davalo E, Naïm P (1990) *Des réseaux de neurones.* 2ª edición, Eyrolles, París.

[7] Delacour J, editor (1978) *Neurobiologie de l'apprentissage.* Masson, París.

[8] Fukushima K (1988) *Neural networks and visual pattern recognition.* En: Systems with Learning and Memory Abilities, J. Delacour y J.C.S. Levy (eds.) Elsevier Sc. Publ., Amsterdam.

[9] Grossberg S, editor (1988) *Neural networks and natural intelligence.* MIT Press, Cambridge, Massachussets.

[2] Puede por ejemplo considerarse los interesantes trabajos sobre el *número equivalente* de G. Der Mégreditcian en las Memorias del IV Simposio "Métodos Matemáticos Aplicados a las Ciencias".

[10] Hebb D O (1949) *The organization of behavior.* John Wiley & Sons, Nueva York.

[11] Hinton G E, Sejnowski T J (1986) *Learning and relearning in Boltzmann machines.* En: Parallel Distributed Processing, D.E. Rumelhart y J.L. Mc Clelland (eds.) The MIT Press, Cambridge, Massachussets.

[12] Hopfield J J (1984) *Neurons with graded response have collective computational properties like those of two-state neurons.* Proceedings Nat. Ac. Sci., 81, pp. 3088-3092.

[13] Kirkpatrick S, Gelatt C D, Vecchi M P (1983) *Optimization by simulated annealing.* En: Science, vol. 220, n^o 4598.

[14] Kohonen T (1984) *Self-organization and associative memory.* Springer-Verlag, Berlín.

[15] Lebart L (1976) *Sur les calculs impliqués par la description de certains grands tableaux.* En: Annales INSEE, n^o 22-23.

[16] Le Cun Y (1985) *Une procédure d'apprentissage pour réseau à seuil asymétrique.* Proceedings of Cognitiva 85, París, pp. 599-604.

[17] Lelu A (1989) *A neural model for highly multidimensional data analysis.* En: Data Analysis, Learning Symbolic and Numeric Knowledge, E. Diday (editor), INRIA - Nova Science, Nueva York.

[18] McCulloch W S, Pitts W (1942) *A logical calculus of the ideas immanent in nervous activity.* Bull. of Math. Biophysics, 5, pp. 115-133.

[19] Minsky M, Papert G (1969) *Perceptrons.* The MIT Press, Cambridge, Massachussets.

[20] Muller C, Radoui M (1990) *Réseaux neuromimétiques et Analyse des Données.* En: *XXII Journées de Statistique - ASU, Tours.*

[21] Murtagh F (1990) *Multilayer perceptrons for classification and regression.*

[22] Oja E (1982) *A simplified neuron model as a principal component analiser.* En: Journal of Math. Biology, vol. 15, pp. 267-273.

[23] Pao Y H (1989) *Adaptive pattern recognition and neural networks.* Addison-Wesley, Nueva York.

[24] Perez J C (1989) *De nouvelles voies vers l'Intelligence Artificielle: pluri-disciplinarité, auto-organisation, réseux neuronaux.* 2^a edición, Masson, París.

[25] Ripley B D (1993) *Statistical aspects of neural networks.* En: Proceedings of the SemStat. Chapman & Hall, Londres.

[26] Rosenblatt F (1958) *The Perceptron: a probabilistic model for information storage and organization in the brain.* En: Psychological Review, 65, pp. 386-408.

[27] Rosenblatt D, Lelu A, Georgel A (1989) *Learning in a single pass: a neural model for principal component analysis and linear regression.* En: 1^{st} IEEE Conference on Artificial Neural Networks and Applications.

[28] Rumelhart D E, McClelland J L. editores (1986) *Parallel distributed processing.* Vol. 1: *Foundations.* Vol. 2: *Exploration in the microstructure of cognition.* The MIT Press, Cambridge, Massachussets.

[29] Tsiknaki C, Graillat C, Berthet J (1990) *Segmentation marketing par les réseaux de neurones.* Rapport technique, ISI, Montpellier.

[30] White H (1989) *Some asymptotic results for learning in single hidden-layer feedforward network models.* Jour. Amer. Statistical Assoc., vol. 84, n^o 408.

[31] Widrow B, Hoff M E (1960) *Adaptive switching circuits.* IRE WESCON Connection Record, Nueva York, pp. 96-104.

Clasificación con Indices Probabilísticos: una aplicación a las encuestas de opinión pública

William Castillo E.* Carlos Arce S.*

Abstract

En este artículo se hace una presentación rápida del método de clasificación jerárquica de I.C. Lerman y se describen los resultados de una aplicación a las encuestas de opinión pública.

1 Introducción

Los métodos —o procedimientos— de análisis de datos conocidos con el nombre de Clasificación Automática Jerárquica Ascendente tienen como objetivo construir una o varias particiones —clasificaciones— de un conjunto $\mathcal{E}$ con n objetos. Esta construcción empieza con la partición discreta de $\mathcal{E}$ y utilizando un criterio de similitud entre clases de objetos, se fusionan paso a paso los pares de clases más próximas. Puntualizamos de la siguiente forma los elementos que intervienen en el proceso:

1. Los datos se ordenan en la forma de una matriz de datos rectangular $n \times p$ en la que cada fila corresponde a un individuo caracterizado por p variables. Las columnas se identifican con las variables. El conjunto $\mathcal{E}$, que va a ser escindido en partes disjuntas, es el conjunto de filas o el conjunto de las columnas de la matriz de datos.
2. Un criterio para cuantificar la similitud entre objetos que se define y calcula a partir de la tabla de datos. Este criterio se expresa como una función y se le denomina usualmente *índice de proximidad.*

*Escuela de Matemática, Universidad de Costa Rica

3. Un criterio para cuantificar la similitud entre grupos de objetos que se define y calcula a partir de las proximidades entre objetos. Este criterio se denomina *índice de agregación.*

4. Un algoritmo que permita fusionar paso a paso los grupos de objetos más próximos según el criterio de agregación. El algoritmo se inicializa con la partición discreta de $\mathcal{E}$ en la que cada clase consta de un sólo elemento y sucesivamente se van reuniendo los grupos de objetos más próximos, según el índice de agregación elegido, hasta que en el último paso la partición resultante se reduce a una sóla clase constituida por $\mathcal{E}$. Si en cada paso se fusionan exactamente dos grupos, diremos que se trata del algoritmo *usual.*

5. Los resultados del procedimiento anterior se representan por medio de un árbol jerárquico binario. Cada nivel de este árbol corresponde a una partición del conjunto $\mathcal{E}$.

6. Un criterio para escoger una o varias particiones "cortando horizontalmente" el árbol jerárquico a un cierto nivel.

7. Elaboración de un sistema computacional automático que permita la ejecución de los procesos anteriores.

8. Por último, corresponde dar una interpretación de los resultados.

En este contexto los aportes de I.C. Lerman se relacionan principalmente con los puntos 2, 3, 5 y 6. La cuestión fundamental en su modelo es la incorporación de una hipótesis de independencia a partir de la cual se construye un índice de proximidad normalizado ([1],[2],[4]).

Esta hipótesis no es un concepto idéntico a la hipótesis de independencia usual en estadística inferencial. Más bien, se trata de un planteamiento del análisis de datos que es "opuesto" al de los test de hipótesis. Dejemos que sea el mismo Lerman quien nos exponga su punto de vista [3]:

> "Pensamos que la filosofía del análisis de datos es en todo caso opuesta a la de los test de independencia o de ausencia de relaciones. En efecto, tomemos por ejemplo el problema de la investigación de las relaciones entre elementos de un conjunto V de variables definidas sobre una población $\mathcal{P}$. El segundo enfoque — los test de hipótesis — le da más importancia a la creencia en la inexistencia de relaciones, las cuales si realmente existiesen — sobre la base de una muestra $\Omega \subset \mathcal{P}$ y un umbral fijo —, no pueden ser realmente medidas y comparadas. Al contrario, para el análisis de datos

no hay ninguna duda en cuanto a la existencia de las relaciones entre las variables sobre la población $\mathcal{P}$. *Sin embargo, estas relaciones son más o* menos fuertes o más o menos tenues y se trata de evaluarlas de forma objetiva para organizarlas lo mejor posible."

2 Proximidad entre objetos

En la terminología matemática, un índice de proximidad entre objetos es una función $P : \mathcal{E}\times\mathcal{E} \longmapsto [0, r]$ con las siguientes propiedades:

- Es simétrica: $P(x, y) = P(y, x) \quad \forall\,(x, y) \in \mathcal{E}\times\mathcal{E}$.
- Entre más grande sea $P(a, b)$ más próximos han de estar a y b.
- La similitud de un objeto con sí mismo es máxima e igual para todos ellos, esto es,

$$P(x, x) = \max\{P(a, b) \mid (a, b) \in \mathcal{E}\times\mathcal{E}\} = r \quad \forall x \in \mathcal{E}$$

Para los fines del análisis de datos, un índice de proximidad se presenta bajo la forma de una matriz simétrica cuyas entradas son todas no negativas y su diagonal es constante.

I.C. Lerman ha desarrollado un procedimiento para construir índices de proximidad entre individuos y entre variables, el cual es expuesto con detalle, para individuos en [4] y para variables en [1]. Haremos un breve resumen del procedimiento limitándonos al caso de las variables. Simultáneamente se describe el proceso de construcción del índice de proximidad entre atributos que será utilizado en la aplicación. El esquema global de construcción comprende cuatro etapas que podemos resumir así:

1. Las variables —en nuestro caso los atributos— deben ser representadas por medio de estructuras matemáticas adecuadas. Piénsese por ejemplo en una variable nominal: una estructura matemática adecuada es la partición que induce sobre el conjunto de individuos. Un atributo descriptivo, como variable nominal con dos modalidades, se representa por una partición con dos clases: una clase formada por los individuos en los que se observa la primera modalidad —los individuos que poseen el atributo — y la otra, por los que tienen la segunda modalidad. Así, es suficiente considerar al atributo x representado por Ω_x, el conjunto de individuos que poseen el atributo x.

2. Se escoge un índice "bruto" de proximidad entre variables. Si por ejemplo, como es nuestro caso, a y b son dos atributos, entonces se considera la cantidad de individuos que poseen simultáneamente ambos atributos. Esto es, el índice de proximidad "bruto" sería: $s(a,b) = |\Omega_a \bigcap \Omega_b|$. En esta etapa de elaboración del índice de proximidad, el concepto de proximidad entre atributos se traduce así: cuanto mayor es el número de individuos que poseen simultáneamente los atributos a y b, mayor es su similitud.

3. Hipótesis de independencia. Para evaluar la proximidad entre dos variables a y b —no necesariamente atributos descriptivos—, mediante un índice $p(a,b)$, se le asocia a a un índice bruto aleatorio $s(a,x)$, escogiendo x al azar de un universo de variables a determinar. Esta escogencia de x engendra una "ausencia de asociación" entre a y x, de modo que si $s(a,b) \leq s(a,x)$ deberíamos concluir que la asociación entre a y b es pequeña.

Vamos a indicar brevemente el modelo probabilístico que se usará cuando el conjunto de objetos $\mathcal{E}$ a clasificar es un conjunto de atributos descriptivos, "observados" sobre un conjunto de individuos que denotaremos con la letra Ω.

Suponiendo que se han elegido $a, b \in \mathcal{E}$ y que $n_a = |\Omega_a|$ y $n_b = |\Omega_b|$, se dota al conjunto $\mathcal{B} = \{x \in \mathcal{E} \mid |\Omega_x| = n_b\}$, de una ley de probabilidad uniforme P, que en este caso asume la forma: $P(x) = 1/\binom{n}{n_b}$ donde $n = |\Omega|$.

Al índice bruto $s(a,b)$ se asocia el índice bruto aleatorio $S_a(a,x) = |\Omega_a \bigcap \Omega_x|$ escogiendo x al azar en $\mathcal{B}$. Se comprueba fácilmente que la ley de probabilidad de S_a es hipergeométrica dada por:

$$P(S_a = k) = \frac{|\{x \in \mathcal{B} \mid |\Omega_x \bigcap \Omega_a| = k\}|}{|\mathcal{B}|} = \frac{\binom{n_a}{k}\binom{n - n_a}{n_b - k}}{\binom{n}{n_b}} \tag{1}$$

con $k \in \{0, 1, \ldots, \min\{n_a, n_b\}\}$ y donde $\binom{n}{k}$ es el coeficiente binomial.

Análogamente, sea $\mathcal{A} = \{x \in \mathcal{E} \mid |\Omega_x| = n_a\}$, dotado de la probabilidad uniforme $P(x) = 1/\binom{n}{n_a}$. Si al índice bruto $s(a,b)$ se asocia el índice bruto aleatorio $S_b(x,b) = |\Omega_x \bigcap \Omega_b|$ escogiendo x al azar en $\mathcal{A}$, entonces es posible demostrar que $P(S_a = k) = P(S_b = k)$. Por tanto la estrategia de fijar a y escoger al azar $x \in \mathcal{B}$ es equivalente a fijar b y escoger al azar $x \in \mathcal{A}$. Y el criterio de proximidad es independiente de esta elección.

Los cómputos involucrados en la ecuación 1 se simplifican sin pérdida de información, si se aproxima la ley hipergeométrica asociada a S_a por la ley normal. Con ese propósito se considera la transformación o estandarización respecto de la ley hipergeométrica,

$$S_a^* = \frac{S_a - \mu}{\sigma}$$

que sigue asintóticamente una ley $N(0,1)$, donde $\mu = n_a n_b/n$ y

$$\sigma^2 = n_a n_b (n - n_a)(n - n_b)/n^2(n-1)$$

Por último, el índice de proximidad $p(a,b)$, se define como la probabilidad de que ocurra el evento $S_a^*(a,x) \leq S_a^*(a,b)$. Es decir,

$$p(a,b) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{S_a^*(a,b)} e^{-x^2/2} dx \approx P(S_a^* \leq S_a^*(a,b))$$

Finalmente observemos que $p(a,b)$ es un índice de proximidad:

1. $p(a,b) = p(b,a)$ para todo a, b ya que $S_a^*(a,b) = S_a^*(b,a)$.

2. $p(a,b) \in [0,1]$ y $p(x,x) = r$, donde

$$r = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{n-1}} e^{-x^2/2} dx$$

Además como $S_a^*(a,b) \leq S_a^*(a,a)$ entonces

$$p(x,x) = \max\{p(a,b) \mid (a,b) \in \mathcal{E}\times\mathcal{E}\}$$

3. Entre mayor sea $p(a,b)$ menos probable es que $S_a^*(a,b) \leq S_a^*(a,x)$ por tanto más asociados o próximos se encuentran a y b.

3 Proximidad entre grupos de objetos

El criterio para cuantificar la proximidad entre grupos de objetos se construye de la siguiente manera: Sean A y B dos partes de $\mathcal{E}$, disjuntas y no vacías. Se considera el conjunto $\{p(a,b) \mid (a,b) \in A\times B\}$ como una realización u observación de $|A||B|$ variables aleatorias independientes y uniformemente distribuidas en $[0,1]$. Esto tiene sentido porque, como se sabe, una distribución de probabilidad puede ser considerada como una variable aleatoria con distribuión uniforme en $[0,1]$. Sea ahora $d_1(A,B) =$

$\max\{p(a,b) \mid (a,b) \in A\times B\}$ y X_{ab} la variable aleatoria uniformente distribuida en $[0,1]$, tal que $p(a,b)$ es su realización. Observe que X_{ab} no es más que la función de distribución de S_a^* de manera que $X_{ab} = P(S_a^* \leq x)$. Consideremos la variable aleatoria $U = \max\{X_{ab} \mid (a,b) \in A\times B\}$ y su función de distribución $P(U \leq u) = u^{|A||B|}$.

Teniendo en cuenta lo anterior se define el índice de agregación entre A y B por medio de la fórmula:

$$\delta_1(A,B) = -\ln(-\ln D_1(A,B))$$

donde

$$D_1(A,B) = P(U \leq d_1(A,B)) = (d_1(A,B))^{|A||B|}$$

El árbol de clasificación construido con δ_1 es el mismo que se puede constuir con D_1, puesto que la función $f(x) = -\ln(-\ln x)$ es estrictamente creciente para $0 < x < 1$, y $\delta_1(A,B) = f(D_1(A,B))$.

Para evaluar $\delta_1(X,Y)$ en cada paso del algoritmo, se utiliza la siguiente fórmula de actualización de δ_1

$$\delta_1(A\cup B, C) = -ln|A\cup B| + \max\{\delta_1(A,C) + ln|A|, \delta_1(B,C) + ln|B|\}$$

4 Niveles y nodos significativos

Una vez construido el árbol jerárquico es necesario "cortarlo" en uno o varios niveles que correspondan a las mejores particiones en un cierto sentido. La idea es que la partición seleccionada maximice el número de veces que ocurre lo siguiente: todo par de objetos en una misma clase de la partición escogida, son más próximos — más asociados, en el caso de variables— que cualquier otro par de objetos en clases distintas.

Con base en este principio se ha establecido un índice para identificar las "mejores" particiones de una jerarquía binaria dada. Si la proximidad entre objetos es dada por una matriz cuyas entradas debajo de la diagonal son todas diferentes, entonces se construye un índice $S(k)$, llamado estadística global [5], de la forma siguiente: para cada partición P_k de la jerarquía,

$$S(k) = \frac{z_k - rs/2}{[rs(f+1)/12]^{1/2}}$$

donde

- $f = (n^2 - n)/2$.
- s es el número de pares de objetos que están en clases distintas de P_k.
- $r = f - s$.
- z_k es el número de veces que para P_k se cumple lo siguiente: sean a, b, c, d cuatro objetos distintos del conjunto a clasificar, entonces la proximidad entre a y b es menor que la proximidad entre c y d si y sólo si a y b están en distintas clases mientras que c y d están en la misma clase.

Si como ocurre normalmente, los valores $S(1), \ldots, S(n-1)$ muestran una tendencia global creciente, exhibiendo máximos locales; las particiones que corresponden a los "picos" más pronunciados son las "mejores". A estas particiones se les llama niveles significativos o particiones significativas. Así, el analista escogerá una partición teniendo en cuenta este criterio y el número de clases.

Empíricamente se ha observado que cuando se produce un crecimiento "acelerado" de los valores $S(i)$ hasta un nivel h y luego "desacelerado", la clase formada en ese nivel puede, con frecuencia, ser objeto de alguna interpretación. Por tanto, conviene calcular las primeras diferencias de la estadística global $D(i) = S(i) - S(i-1)$ y ubicar sus máximos locales. Los nodos correspondientes se denominan nodos significativos o clases significativas y la función $D(i)$ estadística local.

5 Aplicación a la encuesta de 1991

Los datos que se analizan corresponden a una submuestra de la encuesta de opinión pública realizada en julio 1991 por el grupo *Agorametría* (ver anexo A). Esta encuesta tiene como objetivo estudiar el fenómeno de la opinión púbica, para lo cual se elaboran frases o proposiciones cuyo contenido es polémico. El entrevistado debe responder cada proposición con una de las siguientes cinco opciones:

1. Totalmente de acuerdo
2. De acuerdo
3. Podría estar de acuerdo
4. En desacuerdo
5. Totalmente en desacuerdo

Las respuestas fueron recodificadas para convertirlas en atributos mediante el siguiente procedimiento: cada proposición se "desdobla" en dos atributos. Por ejemplo, la proposición "se puede confiar en la justicia" da lugar a:

1. El atributo "Sí; se puede confiar en la justicia", cuyas modalidades de respuesta son un uno si el entrevistado respondió la opción 1. o la 2., y cero si no.

2. El atributo "No; se puede confiar en la justicia", cuyos valores son un uno si el entrevistado respondió la opción 4. o la 5., y cero si no.

El conjunto de atributos a clasificar dió un gran total de 204. De ellos 174 se obtuvieron por el procedimiento antes indicado y el resto son atributos de señalización deducidos de las correspondientes variables, por desdoblamiento de cada una de sus modalidades. La matriz de datos de la cual sus columnas forman el conjunto a clasificar, es de tamaño 897×204. Todas sus entradas son ceros o unos.

Usando los criterios de proximidad expuestos en la secciones 2 y 3 y el algoritmo *usual*, se construyó una jerarquía binaria sobre el conjunto de las 204 columnas. Luego se escogió una partición significativa al nivel 187, formada por 17 clases o grupos de atributos.

Observando directamente en la jerarquía de clasificaciones construida [1], encontramos los atributos que fueron agregados en las primeras etapas del algoritmo y que, por tanto, son los más asociados. Enseguida se listan algunos de estos grupos de proposiciones fuertemente asociadas.

1. El gobierno debe rebajar salarios por paros y huelgas.

   Los trabajadores deben hacer paros y huelgas contra las alzas.

2. Hay que cooperativizar las clínicas del SS.

   El cooperativismo es una salida a la crisis.

3. Se deben privatizar los servicios eléctrico y telefónico.

   Se deben privatizar las instituciones como el ICE y RECOPE.

   Se debe privatizar el SS.

4. La policía irrespeta los derechos ciudadanos.

   En CR se torturan detenidos.

5. Los diputados no deben legislar en su propio beneficio.

   Los médicos no deben cobrar al usar el equipo del SS.

[1] El instrumento informático para representar en forma sintetizada grandes árboles jerárquicos binarios, no es aún disponible, lo cual impidió incluir nuestro árbol en este artículo

6. Padece dolor de espalda.

   Padece migraña.

7. Tiene automóvil.

   Tiene VHS.

En cuanto a la partición seleccionada, un primer examen indica que 12 de las 17 clases que la constituyen, poseen características que dan lugar a unos comentarios. Por motivos de espacio presentamos la estructura y las características de solamente dos clases.

1. Clase 2 (11): Globalmente se podría resumir diciendo que esta clase agrupa atributos que expresan aprobación a la política del gobierno y, complementariamente, una posición contraria a los sindicatos.

   - *Subclase* 1 (6): Contiene atributos favorables al gobierno y contrarios a los sindicatos.

     Sí; El gobierno actuó bien en el caso de las cocineras.

     Sí; El gobierno debe continuar con la movilidad laboral.

     Sí; El gobierno debe rebajar salarios por paros y huelgas.

     Sí; Los trabajadores no deben hacer paros y huelgas contra las alzas.

     Sí; Para el trabajador no es mejor sindicalismo que solidarismo.

     Sí; Las libertades sindicales no fortalecen la democracia costarricense.

   - *Subclase* 2 (5): Esta subclase es como una variable "pro privatización". Los atributos que la forman son:

     Sí; Se deben privatizar los servicios eléctrico y telefónico.

     Sí; Se deben privatizar instituciones como el ICE y RECOPE.

     Sí; El SS debe privatizarse.

     Sí; Hay que privatizar las lavanderías de los hospitales.

     Sí; Se debe privatizar el CNP.

2. Clase 3 (15): En esta clase se agrupan variables que denotan una actitud crítica y desacuerdo frente a ciertas políticas del gobierno. Podemos distinguir tres tipos de políticas correspondientes a la estructuración de la clase en tres subclases, como lo indicamos a continuación.

- *Subclase* 1 (5): Se observa que en esta subclase se agrupan variables que configuran una cierta política en el campo laboral relativa al derecho que asiste a los trabajadores a organizarse para luchar contra las "alzas". Asociadas aparecen también atributos que indican desaprobación a: el accionar del gobierno frente a los desastres naturales, la reducción del presupuesto en salud y del tratamiento que el gobierno dio al caso de las cocineras. Estos atributos son:

  No; El gobierno está preparado para enfrentar las catástrofes.

  No; La crisis justifica la reducción del presupuesto en salud.

  No; El gobierno debe rebajar salarios por paros y huelgas.

  Sí; Los trabajadores deben hacer paros y huelgas contra las alzas.

  No; El gobierno actuó bien en el caso de las cocineras.

- *Subclase* 2 (4): Está formada por variables que implican una valoración negativa sobre: el control de la inflación, los organismos financieros internacionales y la movilidad laboral. Estos atributos son:

  Sí; Las exigencias de los organismos financieros atentan contra la soberanía nacional.

  No; El FMI ayuda a resolver la crisis.

  No; El gobierno debe continuar con la movilidad laboral.

  No; El gobierno controló la inflación.

- *Subclase* 3 (6): Involucra atributos que valoran negativamente los resultados del bono alimentario y la atención, por parte del gobierno, de la salud pública y las demandas populares. Coherente con esto se asocian atributos indicadores de un deterioro de los servicios de las clínicas y hospitales. Los atributos son:

  No; Los servicios en las clínicas del SS son buenos.

  Sí; Los servicios de los hospitales se han deteriorado.

  Sí; Las universidades privadas son un negocio.

  Sí; El gobierno ha descuidado la salud.

  Sí; Las decisiones del gobierno no toman en cuenta a los sectores populares.

  Sí; El bono alimentario fue sólo una promesa.

6 Conclusiones y perspectivas

Los resultados obtenidos son, a nuestro juicio, satisfactorios. La mayor parte de las clases de la partición seleccionada poseen un patrón que las caracteriza. Eso reafirma

la idea que los criterios de proximidad usados recogen finamente las similitudes entre objetos y entre grupos de ellos.

El árbol jerárquico y los criterios para elegir la partición y detectar clases y subclases significativas son recursos invaluables en el proceso de investigación. Ello nos permitió primero, escoger una clasificación y luego ubicar las clases y subclases significativas, una guía fundamental en el proceso de análisis.

Las perspectivas de desarrollo en este campo de investigación, en varias direcciones, son amplias. Entre ellas mencionamos las siguientes:

- El sistema informático diseñado permite ampliarlo, por ejemplo, para que sea capaz de recibir otros tipos de datos o que implemente otros criterios de proximidad. Ello haría posible investigaciones para comparar diversos métodos de clasificación sobre la base de la experimentación.
- También el sistema puede ser ampliado para complementar este tipo de investigaciones con el estudio de la correspondencia de grupos de individuos (filas en la matriz de datos) con las clases o subclases de variables. En el caso de las encuestas este es un aspecto sumamente valioso puesto que facilitaría la ubicación de grupos sociales asociados a las características de alguna clase (s) o subclase (s) de variables.
- Es posible lograr una representación arborescente "reducida" utilizando principalmente los nodos significativos. Esto facilitaría el análisis de las clasificaciones y permitiría, eventualmente, reportar el árbol de clasificaciones.
- El tiempo de cálculo en computador puede ser reducido con la implementación de algoritmos más eficientes en caso de grandes tablas de datos.

Referencias

[1] Lerman, I.C. (1981) *Classification et Analyse Ordinale des Données.* Dunod, París.

[2] Lerman, I.C. (1973) *Etude distributionnelle de statistiques de proximité entre structures de même type, application à la classification automatique.* Cahiers du Bureau Universitaire de Recherche Opérationnelle. París.

[3] Lerman, I.C. (1984) *Justification et valité statistique d'une échelle [0,1] de fréquence mathématique pour une structure de proximité sur un ensemble de variables observées.* Rapport de Recherche No. 278, INRIA, Rocquencourt.

[4] Lerman, I.C.; Peter, P. (1985) *Elaboration et logiciel d'un indice de similarité entre objets d'un type quelconque. Application au problème de consensus en classification.* Publication Interne No. 262, IRISA, Rennes.

[5] Lerman, I.C.; Ghazzali, N. (1990) *Quoi retenir d'un arbre de classification? Un essai en quantification d'image numérisée.* Publication Interne No. 568, Rennes.

La Seguridad Ciudadana y la Opinión Pública en el Valle Central, 1992

Lic. Olga Prieto C.*

1 Introducción

El presente trabajo forma parte del Proyecto de Investigación sobre las Estructuras de Opinión Pública en Costa Rica, que están realizando conjuntamente las Escuelas de Antropología y Sociología y de Matemática.

La expresión "seguridad ciudadana", dentro del contexto que aquí se presenta, se refiere específicamente a la seguridad contra la delincuencia, a la seguridad policial. Sin embargo, este concepto debe ir mucho más allá y "considerar inseguridades actuales de la desplanificación urbana y de ciertos proyectos de vivienda que devienen en ghettos criminológicos; inseguridad laboral, inseguridad en el futuro socioeconómico y familiar, inseguridad en las posibilidades de estudio, etc." [1, p.14]. Es decir, hay que entender el concepto integralmente y no como un fenómeno aislado de nuestra sociedad.

La idea de hacer una encuesta de opinión sobre este tema surgió con base en dos aspectos:

a) En primer lugar, el tema fue propuesto por los estudiantes del grupo 36 del curso de Introducción a la Sociología, como tema a investigar en la parte práctica de dicho curso.

b) En segundo lugar, se consideró incluir este tema en el proyecto mencionado, por cuanto en la encuesta nacional que se realizó en 1991, el primer eje en el análisis de componentes principales, era precisamente el de la búsqueda de seguridad, tanto ciudadana como económica.

Metodológicamente, en la elaboración de la encuesta correspondiente, se siguieron los lineamientos que se han dado para la elaboración de la encuesta nacional desde que el

*Escuela de Antropología y Sociología, Universidad de Costa Rica

proyecto de investigación se inició. Se partió del análisis de los medios de comunicación (prensa escrita, radio y televisión), trabajo que se realizó conjuntamente con los estudiantes, para sacar los temas de conflicto y, posteriormente, se convocó a una reunión de expertos en el tema, cuyas sugerencias permitieron la elaboración final del cuestionario. Éste se pasó a una muestra de 500 ciudadanos costarricenses, mayores de 18 años, residentes en el Valle Central, en donde se concentra el 64% de la población total del país, lo que la hace la zona más propensa a la falta de seguridad ciudadana. La recolección de los datos la hicieron los estudiantes.

Cuando se escogió el tema de la seguridad ciudadana se puso de manifiesto el hecho de que, cotidianamente, durante los últimos años, los medios de comunicación social aluden a actos de delincuencia que se cometen en nuestra sociedad: robos, homicidios, violaciones a mujeres y niños, agresiones, narcotráfico y drogadicción, corrupción, etc. Asimismo, se habla sobre el aumento de la delincuencia, sobre el hecho de que algunos policías con que cuentan las instituciones públicas resultan ser delincuentes, sobre los bajos salarios que los policías tienen, su bajo nivel educativo, los riesgos con que se enfrentan y las malas condiciones de trabajo que tienen. También es común leer algunas noticias en las que entra en juego la Corte Suprema de Justicia, cuestionándose algunas de sus acciones.

2 La encuesta

Las proposiciones que se incluyeron en la encuesta se pueden dividir en los siguientes grupos temáticos:

1. Opinión sobre la Corte Suprema de Justicia.
2. Percepción sobre la delincuencia y los delincuentes.
3. Opinión sobre la delincuencia juvenil.
4. Opinión sobre los castigos que se imponen o se deben imponer, y la prevención del delito.
5. Opinión sobre la policía y acciones contra el delito.

2.1 Opinión sobre la Corte Suprema de Justicia

Las proposiciones que se formularon en la encuesta en relación con la Corte Suprema de Justicia y los resultados que se obtuvieron se pueden observar en el cuadro 1.

CUADRO 1

La opinión pública y la Corte Suprema de Justicia

1992 (%)

Temas	De acuerdo	Podría estar de acuerdo	En desacuerdo
Corte se caracteriza por eficiencia	29,3	34,3	36,4
Se puede confiar en la justicia	27,4	26,8	45,8
Justicia igual para todos ciudadanos	21,3	12,2	66,5
La Corte está exenta de corrupción	13,8	12,7	73,5
La Corte aplica penas muy benévolas	56,4	24,7	18,9

FUENTE: Encuesta de Seguridad Ciudadana, Agorametría de Costa Rica, 1992.

La tendencia aquí es la de cuestionar a la Corte, especialmente en lo que se refiere a la corrupción en el Tercer Poder de la República, a la aplicación equitativa de la justicia y a las penas que aplica. También se observa un porcentaje elevado de personas que desconfían de la justicia.

2.2 Percepción sobre la delincuencia y los delincuentes

Existe una tendencia, aparentemente estereotipada, a considerar que los delincuentes provienen de los sectores sociales socioeconómicos menos privilegiados de la sociedad, especialmente si se tiene en cuenta a los que cometen delitos considerados como "comunes" (robos callejeros, robos en viviendas). Esto es reforzado por los medios de comunicación social, como una manera de esconder otras formas de criminalidad no convencional, logrando atraer toda la agresión social hacia estos sectores sociales, que se convierten en los clientes predilectos del sistema penal y, sobre todo, de la institución carcelaria [1, p.8]. Sin embargo, en el Cuadro 2 se observa que no existen tantos estereotipos en relación con el delincuente, además de que se percibe la delincuencia como algo que va más allá del delito común.

CUADRO 2

La opinión pública y la acción del delito

1992 (%)

Temas	De acuerdo	Podría estar de acuerdo	En desacuerdo
Todos los delincuentes son pobres	10,8	9,1	80,1
Todos los presos son delincuentes	12,6	15,0	72,4
Delinc. se da sólo en calle y robos	10,8	3,9	85,4
Medios com incitan a la violencia	54,7	20,8	24,5

FUENTE: Encuesta de Seguridad Ciudadana, Agorametría de Costa Rica, 1992.

Por otro lado, es interesante observar que la mayoría de las personas encuestadas opina que los medios de comunicación incitan a la violencia.

2.3 Opinión sobre la delincuencia juvenil

CUADRO 3

La opinión pública y la delincuencia junvenil

1992 (%)

Temas	De acuerdo	Podría estar de acuerdo	En desacuerdo
Necesario policía espec. en menores	86,6	7,5	5,9
Menores tratados igual adultos delinc.	9,8	5,7	84,6
Jóvenes hacen más delitos que los adultos	43,0	27,6	29,4
Menores violentos víctimas violencia	81,6	11,8	6,5
Conciertos rock contra seguridad ciudadana	45,1	20,2	34,7
Hay instituciones protegen menor y prev. del.	62,3	23,7	14,0

FUENTE: Encuesta de Seguridad Ciudadana, Agorametría de Costa Rica, 1992.

Del cuadro 3 se desprende que hay una tendencia a creer que existe más delincuencia entre los jóvenes que entre los adultos, lo cual podría estar condicionado porque los jóvenes son víctimas de la violencia y por los conciertos de rock. Se observa además que hay desacuerdo en cuanto a que los menores infractores sean tratados de igual manera que los adultos delincuentes y se plantea la necesidad de que exista una policía especializada en menores.

2.4 Opinión sobre los castigos que se imponen o se deben imponer y sobre la prevención del delito

Dentro del marco ideológico, de corte funcionalista que impera en el país, hay una tendencia a atacar la delincuencia imponiendo castigos a los infractores, quizá no de la manera más justa y muchas veces pueden influir en ello los estereotipos que se tienen en relación con la delincuencia. Se tiende a atacar la delincuencia con violencia, sin que el problema se ataque directamente, al no buscar la causa o causas del mismo. Se le trata como un fenómeno aislado, lo que, en la práctica, hace que se tienda a confundir la prevención con el "tratamiento institucional", el cual, si se consideran las condiciones de las cárceles en nuestro país, funciona como escuela del crimen y sirve para crear estereotipos y estigmas sobre las personas recluidas [1].

El problema de la delincuencia y su prevención no se integran dentro de la planificación del desarrollo global del país. Más bien, cada día se eleva más la alarma social en relación

con el delito, lo que genera que la población tienda a ceder derechos por seguridad, es decir, la gente pide castigos más severos, con el riesgo de que éstos se vuelvan en su propia contra y se deja de lado el aspecto preventivo [1, p.8].

Con respecto a la prevención, cabe anotar que ésta se empieza a dar en nuestro país en la década de los años 80 y se crea la Dirección General de Prevención del Delito, como dependencia del Ministerio de Justicia. Con esto, a nivel institucional, se llega a considerar como delincuencia además del delito común, todo lo concerniente a la violencia, a la seguridad ciudadana, la violencia televisiva y otros medios de difusión masiva y los accidentes de tránsito [1, p.11], lo que en alguna medida viene a cuestionar el sistema penal vigente. De esta manera, al menos teóricamente, el delito se trata de explicar y enfrentarlo en el concurso con la educación, el empleo, el acceso a los servicios sociales, la planificación social de las ciudades y urbanizaciones, etc., además de dar un énfasis a la atención de menores de edad [1, p.12].

CUADRO 4

La opinión pública, la prevención y la penalización del delito

1992 (%)

Temas	De acuerdo	Podría estar de acuerdo	En desacuerdo
Restablecer la pena de muerte	26,6	13,2	60,2
Cárcel no resuelve delincuencia	48,1	13,4	38,5
En Costa Rica se torturan detenidos	32,1	27,5	40,4
Delinc. resuelve con más polic. y cárcel	15,5	11,7	72,8
Presos no juzgados, viola der. humanos	77,8	12,6	9,6
La Corte aplica penas muy benévolas	56,4	24,7	18,9
Hay programas rehabil. delincuentes	41,5	27,4	31,2
Prevención delito es sólo Gobierno	11,6	8,1	80,2
Hay instituc. protegen menor y prev. del.	62,3	23,7	14,0

FUENTE: Encuesta de Seguridad Ciudadana, Agorametría de Costa Rica, 1992.

La mayoría de las personas encuestadas se manifiesta en contra de la pena de muerte y está en desacuerdo con medidas de carácter represivo tales como una mayor cantidad de policías y más cárcel. Asimismo, se opina que se violan los derechos humanos cuando se encarcela durante mucho tiempo a las personas antes de ser juzgadas (lo cual es uno de los problemas de la Corte Suprema de Justicia que se denuncian en la prensa) y casi una tercera parte de las personas opina que se torturan detenidos. Sin embargo, contrariamente, se tiende a considerar que se aplican penas muy benévolas.

En el aspecto preventivo, un alto porcentaje de las personas entrevistadas opina que en el país se rehabilita al delincuente y que existen instituciones para prevenir el delito. Por

otro lado, es notable el hecho de que los encuestados opinan que la prevención no es tarea sólo del Gobierno.

2.5 Opinión sobre la policía y sobre acciones civiles contra el delito:

En relación con la policía, a ésta le corresponde la parte represiva de la seguridad, así como se supone que tiene un papel de servidor público, de guía comunal, de defensora de la niñez. Sin embargo, como se observa en el cuadro 5, la opinión de las personas encuestadas tiende a criticar a la policía.

CUADRO 5

La opinión pública y la policía

1992 (%)

Temas	De acuerdo	Podría estar de acuerdo	En desacuerdo
Seguridad es sólo de la policía	16,6	11,4	72,0
Más policías es militarizarnos	20,2	9,2	70,6
Se está implantando un ejército	7,4	11,2	81,5
Prohibir grupos armados privados	58,6	9,6	31,8
Civiles armados contra delincuencia	46,7	26,0	27,4
Policía irrespeta derechos ciudadanos	46,7	29,6	23,7
Los policías son muy ignorantes	51,3	28,8	19,9
La policía no cumple su misión	42,9	35,4	21,7
Muchos patrulleros son delincuentes	45,5	29,7	24,7
La policía abusa de su autoridad	62,0	25,2	12,9
Policía en desventaja respecto delinc.	71,8	16,4	11,9
Pol. deficiente para resolver delitos	55,0	22,7	22,3
Todos los delitos son denunciados	5,5	4,9	89,6

FUENTE: Encuesta de Seguridad Ciudadana, Agorametría de Costa Rica, 1992.

Con la introducción de la prevención en el país se habla, institucionalmente, de que se debe redefinir el papel de la policía, de que debe existir un equilibrio entre su función preventiva convencional, con la prevención integral [2, p.16].

De acuerdo con los datos de la encuesta, se puede decir que la crítica que se hace genera una búsqueda, por parte de la opinión pública, de la redefinición del papel de la policía en la sociedad –que podría ser extensivo al sistema de la seguridad total–, al punto de llegar a considerarse que la seguridad ciudadana no le compete únicamente a la policía, tanto así que un alto porcentaje de los encuestados opina que los civiles deben armarse contra la delincuencia.

3 Seguridad ciudadana y estructura de opinión pública

Al relacionar los resultados de la encuesta con la estructura de la opinión pública [3], en la figura 1 se observa que casi todos los temas de conflicto la encuesta se ubican el tercero y cuarto cuadrantes, denominados, respectivamente, de repliegue o conformismo y de búsqueda de seguridad.

En el primer cuadrante, al que se señala como de "inconformidad" y que se opone al conformismo del tercer cuadrante, se caracteriza porque en él se ubican las personas con edades productivas (25 a 44 años) y de más alto nivel educativo (ver figuras 2 y 3).

Por su ocupación (ver figura 4), dentro de este cuadrante se ubican los profesionales liberales, los patronos de empresa o comercio, clérigos, trabajadores de la computación y los desempleados. Si se toma en cuenta la condición de ocupación de las personas encuestadas, están en este cuadrante aquéllas que son patronos o socios activos.

En el plano opuesto a la inconformidad, está el III cuadrante del repliegue o conformismo, donde se observa una crítica nula al sistema, a lo actuado, es decir, hay un acuerdo con el estado actual de la sociedad en lo que a seguridad ciudadana se refiere. Aquí se encuentran ubicados temas de conflicto tales como (ver figura 1): la prevención del delito es tarea sólo del Gobierno, la delincuencia se resuelve con más policías y más cárcel, la Corte está exenta de corrupción, la justicia es igual para todos los ciudadanos, entre otros.

En este tercer cuadrante coinciden las personas de más edad (65 años y más) y las de más bajo nivel educativo (figuras 2 y 3), las personas encuestadas que son empleadas o servidoras domésticas y las amas de casa que son jefes de familia. Por condición de ocupación, están aquí los servidores domésticos y los trabajadores familiares no remunerados (figura 4).

En el segundo cuadrante, al que se le interpreta como de indiferencia, y que se opone a la búsqueda de seguridad, se ubican las personas con edades jóvenes (18 a 24 años) y aquellas personas con edades de 50 a 54 años, que puede decirse que tienden a pensionarse. En cuanto a ocupación, coinciden aquí los pequeños y medianos productores agropecuarios, los profesionales del sector público, mensajeros y misceláneos, cuadros de nivel gerencial, empleados del sector privado y estudiantes. También las personas que se autoperciben como de izquierda.

Finalmente, en el cuarto cuadrante, se ubican aquellos temas de conflicto que hacen una crítica al sistema de seguridad ciudadana del país, como por ejemplo: en Costa Rica se torturan detenidos, la policía no cumple su misión, los policías están en desventaja con respecto a los delincuentes, etc. Crítica que puede interpretarse como una búsqueda de seguridad.

Dentro de este cuadrante están las personas con edades entre los 45 y 49 y entre los 55

y 59 años, las personas que no tienen religión, las que no cuentan con seguro social ni con casa propia, además de aquellas cuyo estado civil es el de viudo, divorciado o separado, o sea, personas que han perdido el respaldo de su pareja. Por ocupación, están además los trabajadores especializados, personal de servicio y amas de casa en donde el jefe de familia es un profesional liberal. También están aquí los trabajadores por cuenta propia.

4 Reflexiones finales

1. Se constata que la opinión pública del Valle Central cuestiona a la Corte Suprema de Justicia. Se nota una tendencia a que, según aumente la edad, así aumenta este cuestionamiento.
2. Se da una diferenciación en las opiniones de los entrevistados según la variable sexo: los hombres parecen tener una actitud de inconformidad (I cuadrante), en tanto que las mujeres manifiestan un sentimiento de repliegue (III cuadrante).
3. Los estereotipos generalizados en la sociedad costarricense en cuanto a las características socio–económicas del delincuente, no son compartidos, en su totalidad, por las personas encuestadas.
4. Pareciera que los cuerpos policiales guían su conducta por los mitos y creencias sobre el delincuente y no por la realidad objetiva de éste.
5. Lo anterior hace que la acción policíaca y de la misma "justicia", centre sus acciones en el delito convencional.
6. El carácter comercial de los medios de comunicación social los incita, en apariencia, a dedicar grandes espacios al delito convencional. De donde surge la interrogante: ¿Realmente ha aumentado la delincuencia o el miedo de la ciudadanía a los actos delictivos fomentado por los medios?
7. Se constata que la opinión pública cree en la necesidad de que exista un trato diferente al infractor, según su edad.
8. Como opinión, las personas del Valle Central, consideran que la seguridad no compete únicamente a la policía y que la prevención no es tarea sólo del gobierno.
9. Puede observarse que existe un cuestionamiento por parte de la población, al sistema de seguridad imperante en el país.

Bibliografía

[1] Artavia, P. (1991) *Prevención y protección a la minoridad*, en curso sobre "Protección a menores", Comisión Interinstitucional de Capacitación a Autoridades.

[2] Artavia, P. (1991) *Urbanismo y Violencia.* Ponencia presentada en el Foro: El aumento de la violencia urbana y la criminalidad en Costa Rica, Colegio Federado de Ingenieros y Arquitectos, San José, octubre.

[3] Poltronieri, J.; Piza, E. (1989) *Estructuras de la Opinión Pública en Costa Rica.* Editorial Universidad de Costa Rica, San José.

Figura 1: Temas de conflicto

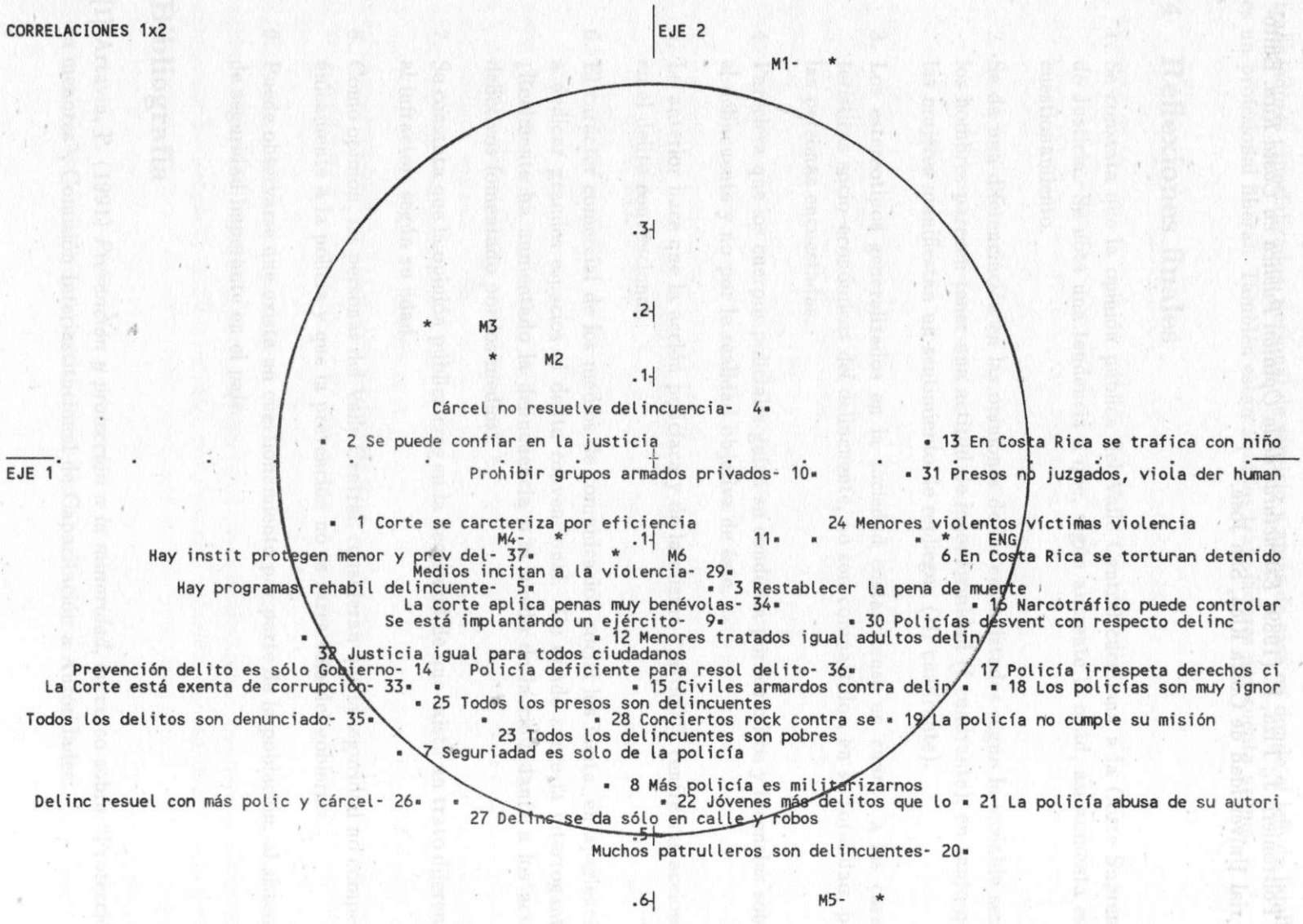


Figura 2: Edad y sexo

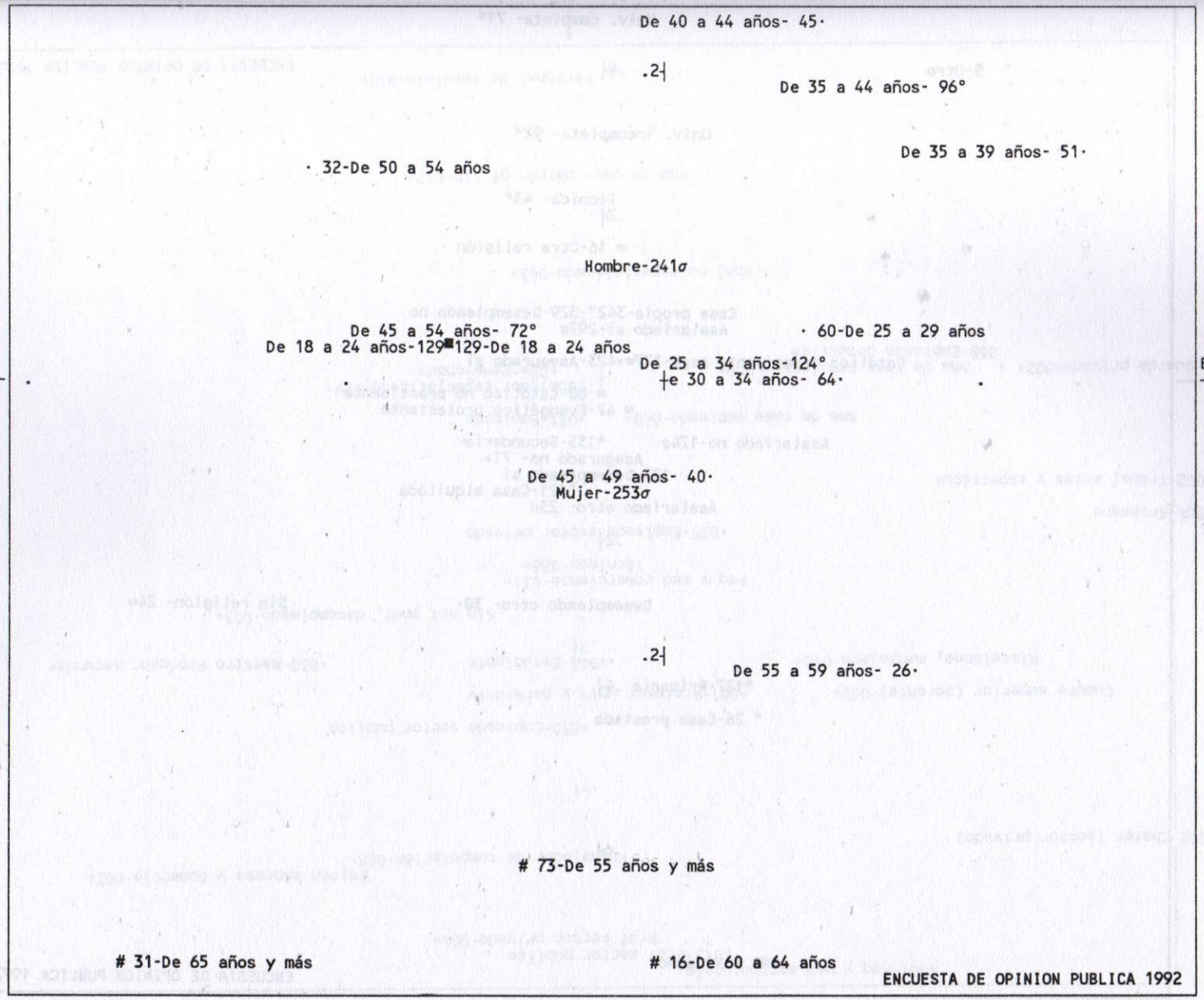


Figura 3: Nivel de Educación y religión

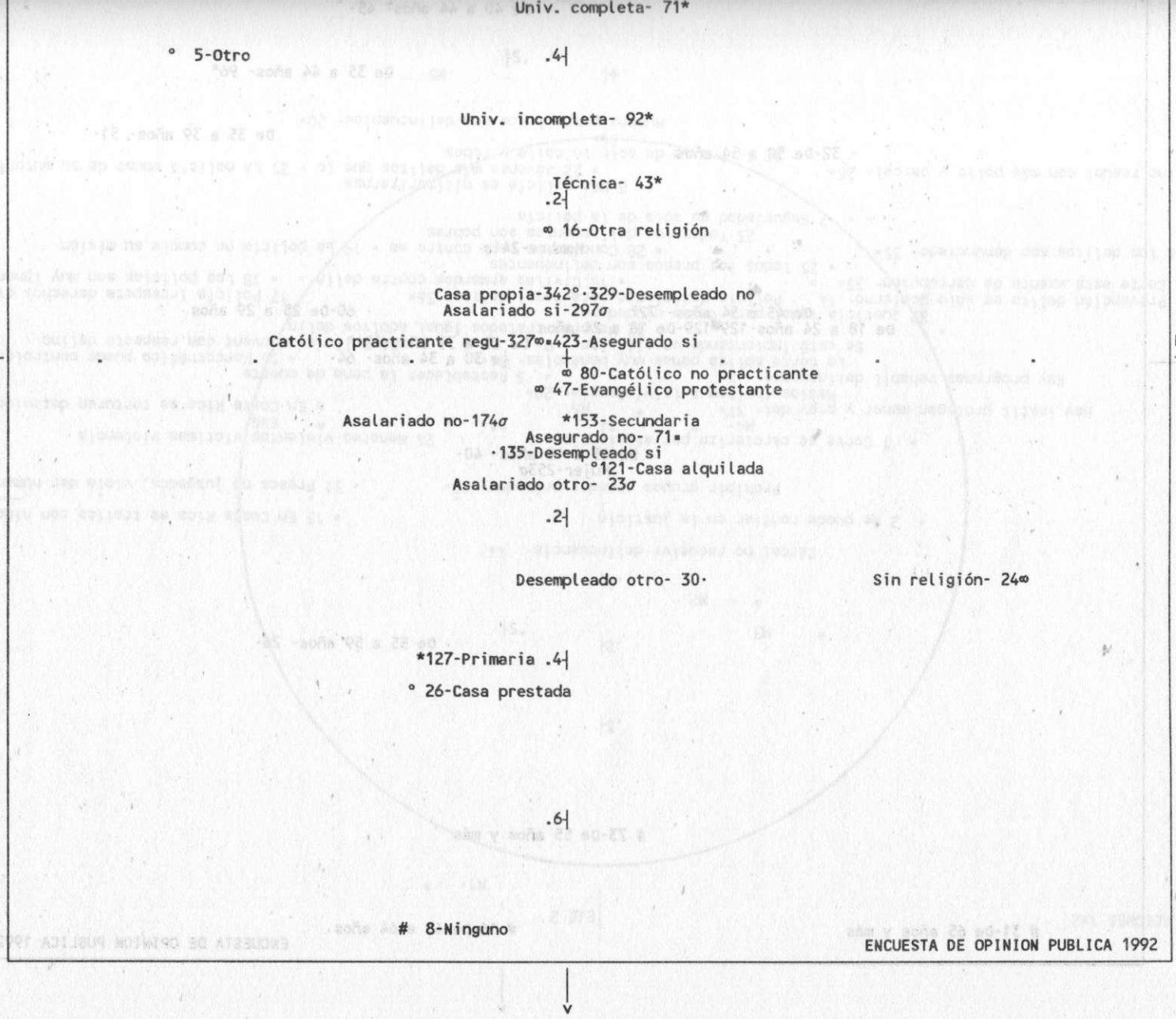
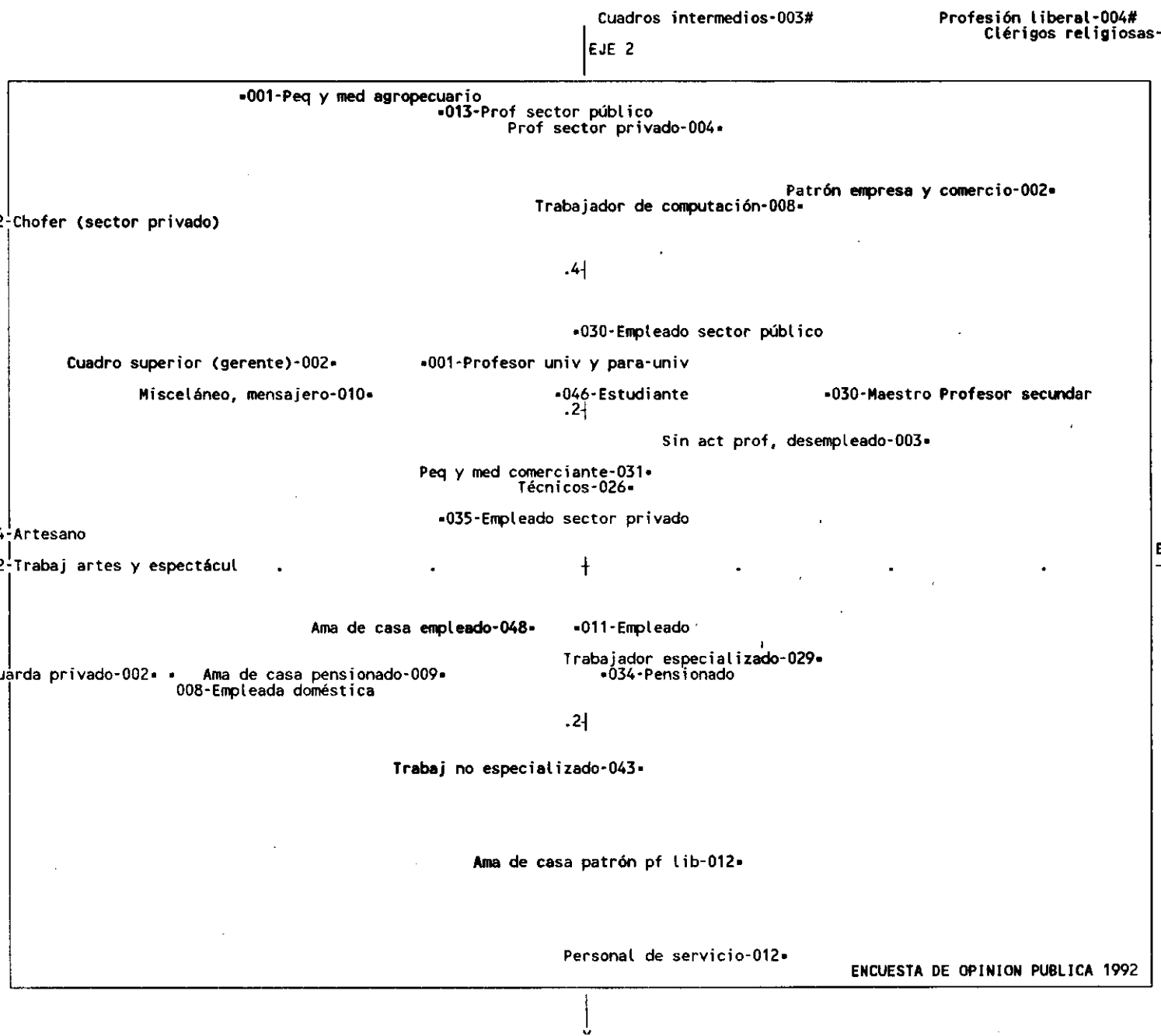


Figura 4: Ocupación



V Centenario y la Opinión Pública del Valle Central

Marta E. López Subirós*

Introducción

Existe mucha polémica en torno a los 500 años de la llegada de los españoles al Continente Americano. Se puede escuchar y leer las diferentes posiciones entorno al tema, pero la mayoría de éstas proviene de académicos y políticos.

Sin embargo, se desconoce qué piensa la población indígena de nuestro país, cuáles son sus inquietudes y cuál es la percepción del resto de la población nacional respecto de ellas.

Es así como nos propusimos intentar resolver dichas interrogantes a través de una encuesta de opinión en el Valle Central.

Para tal efecto se estableció contacto con Sejtko y se iniciaron largas sesiones de trabajo con el fin de:

- explicar la metodología a utilizarse
- explicitar, de parte de Sejtko, cuáles eran sus intereses e inquietudes respecto del V Centenario y que les interesaría conocer la percepción de la población del Valle Central. Es así como se logran determinar los primeros ítems susceptibles de conformar la encuesta.

Posteriormente, se realiza una reunión de expertos en la que participan delegados indígenas, psicólogas, sociólogas y antropólogos, quienes después de arduas discusiones determinan los temas definitivos de la encuesta.

El trabajo de campo es realizado por estudiantes de Introducción a la Sociología, grupos 01 y 15, de la profesora Pilar Ferlini el 18 y 19 de julio de 1992.

*Escuela de Antropología y Sociología, Universidad de Costa Rica

La muestra respeta la densidad poblacional de las provincias de San José, Heredia, Alajuela y Cartago, respecto de la población nacional, de los cantones respecto a la provincia y de los distritos respecto al cantón, cuotas de sexo y ocupación. La población entrevistada es costarricense y mayor de 18 años.

La encuesta

Para poder hacer una interpretación rigurosa de las encuestas de opinión, debe hacerse un ejercicio epistemológico sobre cada pregunta y aún más, sobre el sistema de preguntas planteadas. La relación teoría–método–técnicas debe estar presente en los estudios de opinión pública puesto que los simples resultados de una encuesta sólo sobrepasan el mero empirismo cuando algún cuerpo teórico permite su interpretación. Únicamente así las técnicas cuantitativas empleadas en la investigación de la opinión pública pueden aportar a la comprensión de la realidad social como totalidad. **(López y Garita, 1991:4).**

El análisis de la encuesta arroja las siguientes áreas temáticas de interés para los indígenas (entre otras):

1 El carácter de los 500 años y el Día de la Raza

"El carácter festivo de los 500 años ofende a los indios"
"Debemos celebrar con júbilo el V Centenario"
"Debe derogarse la ley que establece el Día de la Raza"
"El Día de la Raza festeja sólo a los blancos"
"Los españoles destruyeron la religión indígena"
"El Día de la Raza ofende a los pueblos indios"
"Por suerte llegaron los españoles, sino seríamos muy atrasados"
"América perdió mucho con la llega de los españoles"
"Entre más indios tenga un país, más atrasado es"

Se nota la preocupación de conocer la percepción de la población en lo referente a fechas festivas que a través de la historia les ha afectado directamente.

Para los indígenas, estas celebraciones no debieron nunca existir, pues festeja un etnocidio y se proclama un día para la raza vencedora. Tergivesan la realidad histórica y atentan contra su cultura y tradiciones.

Ante este bloque temático, la población del Valle Central se manifiesta de acuerdo con la realización de estas fechas festivas. Sin embargo, al preguntárseles si éstas ofenden a la población indígena, se da una distribución de las respuestas entre los desacuerdos y los

acuerdos. Una situación similar se da con los conflictos relacionados con la llegada española a nuestras tierras y sus consecuencias histórico–culturales:

- 42,1% de los entrevistados considera que América perdió con la llegada de los españoles, contra un 40,3
- 46,4% de las personas se muestra en desacuerdo con que sin la llegada española seríamos muy atrasados, contra un 36% que sí está de acuerdo. Esto se refuerza con la proposición "entre más indios tenga un país, más atrasado es", en donde el 60,1% se manifestó en desacuerdo.

2 La discriminación y los estereotipos hacia los indígenas

"Está bien que el gobierno prohíba a los indios hablar en sus propias lenguas"
"En Costa Rica hay discriminación hacia los indios"
"En nuestro país los indios son realmente iguales que los blancos"
"Los indios deben de cambiar su manera de pensar"
"Los indios tienen capacidad de resolver sus propios problemas"
"Los indios son muy creyenceros"
"Los indios son vagos"
"Los indios tienen poderes secretos"
"Los indios son buenos para curar"
"Los indios son unos salvajes"
"Entre más indios tenga un país, más atrasado es"
"Por suerte llegaron los españoles, sino seríamos muy atrasados"

Es interesante notar los supuestos que los indígenas tienen acerca de los estereotipos vigentes en el resto de la sociedad costarricense. Hay un elemento importante de citar: la preocupación manifiesta en torno al poder hablar en sus propias lenguas se deriva de situaciones vividas en su infancia, donde los maestros les pegaban si lo hacían bajo el pretexto de que "se están burlando de mí".

En cuanto a los estereotipos negativos hacia los indígenas, las personas del Valle Central se manifiestan claramente en contra de tales estereotipos. En los que se refieren a sus cualidades curativas y mágicas, cabe señalar que sólo un 16,5% de listados muestran un acuerdo en cuanto a que "Los indios son buenos para curar". Las respuestas se distribuyen en la proposición "Los indios tienen poderes ocultos" (31,1% de acuerdo, 25,6% podría estar de acuerdo y 43,3% en desacuerdo).

Se encuentra un manifiesto acuerdo de que en Costa Rica se da una discriminación contra los indígenas (78,1% en acuerdo) y que en nuestro país los indios no son realmente iguales

a los blancos (67,1% en acuerdo). Esta proposición permite diferentes interpretaciones, por ejemplo, que hay discriminación o que realmente los indígenas son diferentes (racismo o no).

Se considera abrumadoramente que los grupos indígenas deben mantener vigentes sus lenguas (91,4% en acuerdo), que no deben cambiar su manera de pensar y que los indígenas tienen capacidad para resolver sus propios asuntos (61,3%).

3 El gobierno y la población indígena

"El gobierno debe reducir el dinero que gasta en la salud de los indios"
"En Costa Rica los indios han recibido la ayuda financiera requerida"
"El Seguro Social atiende bien a los indios"
"El sistema educativo anula la identidad de los indios"
"Está bien que el gobierno otorgue permisos de explotación minera en las reservas indígenas"
"El subsuelo debe otorgarse legalmente a los indios"
"Hay organismos que se dicen indígenas para su beneficio económico"

En los tres primeros temas, la población entrevistada se dice estar en desacuerdo. Sólo en la última preposición un 38,3% se manifestó en desacuerdo.

Aquí se encuentra el reclamo sobre sus tierras y el derecho histórico sobre el subsuelo.

En este tema también se dan resultados sorpresivos: un 75,9% de la población está en desacuerdo con otorgar permisos de explotación minera en las reservas indígenas, actitud que se reafirma con el otorgamiento legal del subsuelo a los indios (76,5% de acuerdo). A la vez, parece existir el sentimiento de que hay organismos que se dicen indígenas para su provecho económico (51,2% de acuerdo).

4 Participación en el desarrollo del país y capacidad política del indígena

"Los indios participan en el desarrollo de la sociedad actual"
"Los indios deben dirigir sus propios asuntos"
"Me gustaría que un maestro indígena diera clases a mis hijos"
"En la Asamblea Legislativa debe haber al menos un diputado indio"
"Debe promoverse la profesionalización indígena"
"En Costa Rica un indio sería un buen presidente"
"Tenemos que aprender de los indios"

La posibilidad de que niños no indígenas tengan un maestro que sí lo es, es la única vía posible de contacto permanente e institucionalizado entre familias de otras razas y algún miembro de la población indígena, ya que en esta última, el problema de orfandad no existe: un niño indígena siempre tiene un hogar.

Muchas de estas preocupaciones parten de la hipótesis de que el resto de la población costarricense los subvalora. Son temas que también cabrían en el apartado II, sin embargo, el interés primordial es conocer hasta dónde son considerados autosuficientes y capaces.

En el apartado II se vio que los entrevistados consideran capaz a la población indígena ¿Hasta dónde llega esta percepción positiva? Un 76,9% considera que el resto de la población nacional debe aprender de los indios, a un 56,5% le gustaría que sus hijos tuvieran un maestro indio, un 71% considera que debe haber al menos un diputado indio. Pero si se trata de si un indio sería un buen Presidente, el 31,8% se muestra de acuerdo, un 33,6% podría estar de acuerdo y un 34,6% en desacuerdo.

Al mismo tiempo, sólo un 19,7% opina que los indios participan en el desarrollo de la sociedad actual y un 85,9% cree que debe promoverse la profesionalización del indio.

5 Existencia de los indios en Costa Rica

"Aún existen indios en Costa Rica"
"En Costa Rica sólo hay un grupo indígena"
"La mayoría de los costarricenses tenemos sangre india"
"Los indios siempre han sido reconocidos como costarricenses"

Este apartado muestra el interés de conocer el grado de información que tienen los entrevistados acerca de los indígenas.

Pareciera que existe una buena información acerca de la existencia de los indios en Costa Rica y que son más de un grupo (sólo un 6% y un 9,4% se declararon en desacuerdo, respectivamente). De igual manera, existe un buen nivel de conocimiento de que no siempre los indígenas han sido reconocidos como costarricenses (69,8% de acuerdo con la proposición).

Llaman la atención las respuestas al conflicto "La mayoría de los costarricenses tenemos sangre india": 61,1% dio una respuesta afirmativa, contra lo esperado si se parte del mito de que "Costa Rica es la Europa de Centroamérica".

La variable que ofrece mayor diferenciación en la opinión pública es el nivel educativo. Así por ejemplo, se encuentra que en conflictos tales como "Debemos celebrar con júbilo el V Centenario" y "Los indios tienen poderes secretos", conforme aumenta el nivel educativo,

aumenta el desacuerdo. Otro ejemplo es el tema "Por suerte llegaron los españoles, sino seríamos muy atrasados", los mayores desacuerdos se encuentran en las personas con educación universitaria completa e incompleta y por aquellas personas que declararon tener ninguna educación (un promedio de 62% en desacuerdo). Un segundo bloque lo conforman las personas con educación secundaria, técnica y primaria (36,6% de desacuerdos en promedio).

Cuadro I

La opinión pública / religión, 1992

Conflicto	Sin relig.		Católico no practic.		Católico pract. relig.		Evangel. protest.		Otra relig.	
	A	D	A	D	A	D	A	D	A	D
Asamblea 1 dip. indio	85,1	3,7	66,2	3,7	70,5	9,1	70,9	8,1	89,2	8,3
Celebrar V Cent.	37,0	48,1	47,4	37,5	36,3	23,9	31,0	42,5	58,2	33,2
Gob. prohíba lengua	23,0	76,8	2,5	97,8	6,8	90,0	4,8	93,4	0	100,0
Indios capc. prop.	52,0	28,0	71,2	21,2	60,0	26,9	57,3	19,5	66,6	8,3
Gob. reduc. gasto salud	15,3	72,9	3,7	92,4	9,7	87,9	8,1	90,1	0	100,0
Derogar Día Raza	59,2	25,9	22,5	54,9	27,8	53,5	18,0	63,8	18,1	47,3
Org. se dicen indios...	57,6	25,6	45,7	29,5	50,6	29,5	50,9	33,4	90,0	0
Esp. dest. relig. ind.	74,0	18,5	65,3	16,6	53,3	29,5	66,6	18,1	83,3	8,3
Día Raza ofende ind.	74,0	38,1	45,3	40,1	36,2	44,1	42,3	37,2	24,9	41,6
Indios son salvajes	33,3	66,6	8,7	90,2	8,2	88,5	11,2	85,3	0	99,9
Seg. Soc. atiende bien ind.	20,0	48,0	2,9	83,5	12,0	66,2	3,7	64,1	0	66,6
Am. perdió mucho	77,7	18,5	42,4	34,9	36,7	46,3	52,4	24,4	45,3	45,4

FUENTE: Encuesta V Centenario, 1992. Agorametría de Costa Rica. Elaborado por la autora.

Cabe señalar que una variable de diferenciación, constante, es la religión. Aparece como significativa en el 33% de los conflictos de la encuesta, constituyéndose en la segunda variable de diferenciación en importancia. En ella encontramos las siguientes proposiciones:

- En el tema de un indio como diputado se puede observar que son los católicos practicantes regulares, seguidos por los evangélicos–protestantes y los de otras religiones quienes manifiestan un mayor desacuerdo. Si se trata de la capacidad indígena aparecen los católicos no practicantes y los católicos regulares como los grupos que menos creen en la capacidad de la población indígena.
- En cuanto a las festividades oficiales, aparecen de nuevo los católicos regulares como el grupo que más las apoya y que menos cree que éstas ofenden a los indios.
- Las personas que se dicen sin religión son quienes consideran a los indios como salvajes, que el Día de la Raza no los ofende, que se les debe prohibir hablar sus propias lenguas

y que no tienen capacidad. Son quienes más apoyarían una reducción de los gastos en salud para los indios, así como que los españoles no destruyeron la religión indígena y que el Seguro Social los atiende bien.

La figura 1 representa el círculo de correlaciones asociado al Análisis en Componentes Principales [1, 2] de la encuesta de opinión pública de 1992, y las figuras 2 y 3 el plano principal asociado donde se representan respectivamentelas modalidades correspondientes a la religión, estado civil y nivel de educación, por una parte, y al sexo y los niveles de edad, por otra parte.

El primer cuadrante de la figura 1 (arriba a la derecha) está constituido por temas que se oponen a las festividades, al efecto negativo de los españoles y al abuso por parte de organismos que se dicen indígenas. Es un cuadrante que parece expresar un buen grado de información y por ende un sentimiento de INCONFORMIDAD. Es el cuadrante de las ofensivas que se desarrollan sus ataques contra lo prohibido.

¿Quiénes tienen ese sentimiento? Son los hombres y las personas con edades entre los 25 y 34 años, con educación universitaria, separados, sin religión o evangélicos protestantes. Por ocupación son los profesionales liberales, los cuadros intermedios, técnicos, los estudiantes, las amas de casa cuyo jefe de familia es patrono o de profesión liberal.

Opuesto a este cuadrante de inconformidad (abajo a la izquierda), se encuentra el RETO, definido por temas relacionados con los estereotipos, la inexistencia de discriminación y la real participación de los indios en la sociedad actual. Es una especie de fortaleza en donde se defiende, como en ningún otro lugar, los valores, las reglas, las normas o los estereotipos de que se ha dotado la sociedad ¿DESINFORMACIÓN? ¿REPLIEGUE?

Son las mujeres, las personas de 60 años y más, los viudos, con ninguna educación o educación primaria, católicos practicantes regulares, personas que viven en casas prestadas y se dicen cercanos al PUSC. Son las amas de casa, los choferes del sector privado, las empleadas domésticas y los artesanos, los trabajadores no especializados. Estas ocupaciones se repiten para los jefes de familia agregándose los jefes de familia.

La polémica que opone a los otros dos cuadrantes se centra en los valores positivos sobre el indio vrs. los negativos.

Al segundo cuadrante lo definen temas referentes a los estereotipos negativos, a la creencia de que a los indígenas se les debe reducir la ayuda pues son bien tratados, a que ellos son sinónimo de atraso y que deben cambiar, así como a que se deben otorgar permisos en las reservas indígenas; es la INTRANQUILIDAD (percepción negativa, la INDIFERENCIA).

Este sentimiento es representado por los jóvenes (18–24 años) y las personas entre 55 y 59 años, los que viven en unión libre y que tienen estudios de secundaria. Se definen políticamente como de extrema derecha o indiferentes. Son los pensionados, los patrones de empresa y comercio, el pequeño y mediano comerciante, el guarda del sector privado y

las amas de casa cuyo jefe de familia es pensionado, empleado o desempleado.

En oposición a éste, aparece la SEGURIDAD (abajo a la derecha), en donde se dan proposiciones que hacen referencia a la capacidad indígena, al reconocimiento de su existencia, su herencia y conocimientos, que se deben aprovechar, así como al reconocimiento de que existe discriminación. Son demandas simultáneas de más libertad, de igualdad y solidaridad que provienen de una reivindicación que parece legítima: se vuelve a las premisas de la Declaración de los Derechos Humanos, el DESAFÍO.

Son las personas con edades entre los 35 y 54 años, divorciados, de otra religión y católicos no practicantes. Son los profesionales del sector público, los maestros y profesores, los guardas públicos y los trabajadores especializados.

Conclusiones

- Existen posiciones diferentes entre los indígenas y el resto de la población del Valle Central en torno a las fechas festivas del Día de la Raza y el V Centenario.
- A pesar de la constatación anterior, un 41% de la población entrevistada concuerda con ellos en que dichas fiestas los ofenden.
- 44% de la población defiende la herencia cultural indígena y no la ve como factor de subdesarrollo.
- La existencia de indios no es percibida como factor de atraso, respuesta no esperada por los compañeros indios, si se toman en cuenta los estereotipos que se tienen de países como Guatemala y Bolivia.
- Los estereotipos negativos esperados no se dan.
- Existe un apoyo para salvaguardar la cultura indígena y que son ellos quienes deben establecer las vías, o al menos, ser sujeto en la toma de posiciones.
- Se da un claro reconocimiento de que la ayuda gubernamental para los pueblos indígenas ha sido insuficiente.
- Se da una defensa parcial al sistema educativo vigente en cuanto a su responsabilidad en la anulación de la identidad india.
- La población entrevistada considera capaz a los indígenas. Las frecuencias de las respuestas se convierten en resultados inesperados para los formuladores de dichas inquietudes.
- El conocimiento que se encuentra a través de la encuesta acerca de la realidad indígena no fue el esperado.

- El mito de que la población coatarricense es mayoritariamente blanca, sólo se da en el 19,6% de la población entrevistada.
- Existe un apoyo a las demandas indígenas de su derecho al subsuelo, demandas presentadas por Sejtko a la Asamblea Legislativa.
- Se encuentra una notable diferencia en las opiniones emitidas por las personas que se declararon católicos practicantes regulares y los no practicantes, siendo estos últimos un poco más abiertos a la realidad indígena que los primeros.
- Los católicos regulares se asemejan en casi la mitad de sus posiciones a los evangélicos–protestantes. Las diferencias se dan en los temas relacionados con los días festivos, en si América perdió con la llegada española y en la atención del Seguro Social, donde los católicos manifiestan una opinión más cercana al discurso oficial.
- Por edades, partiendo de los más jóvenes, parece que el camino de la estructura de opinión va de la indiferencia a la inconformidad, de ésta a la seguridad para terminar en el repliegue.
- Si se toma la educación, la trayectoria es repliegue–indiferencia–inconformidad, (de ninguna a educación universitaria completa).
- La posición política (de la izquierda a la extrema derecha) pasaría de la inconformidad al repliegue a la indiferencia. Es interesante notar que ninguna posición da el sentimiento de seguridad.
- Si lo que se trata de evidenciar es el comportamiento de los diferentes sectores productivos, se tiene que arranca de la indiferencia (sector informal y trabajador familiar no remunerado), pasa a la inconformidad (empleados y obreros del Estado) y termina en el repliegue (trabajador por cuenta propia, patrono o socio activo y empleados y obreros del sector privado).

Bibliografía

[1] López, M. (1989) *La opinión pública regional y la política económica en Costa Rica, 1988.* Revista de Ciencia y Tecnología, vol. XIII, Nos. 1 y 2.

[2] Poltronieri, J.; Piza, E. (1989) *Estructuras de la Opinión Pública en Costa Rica.* Editorial Universidad de Costa Rica, San José.

Figura 1: Temas de conflicto

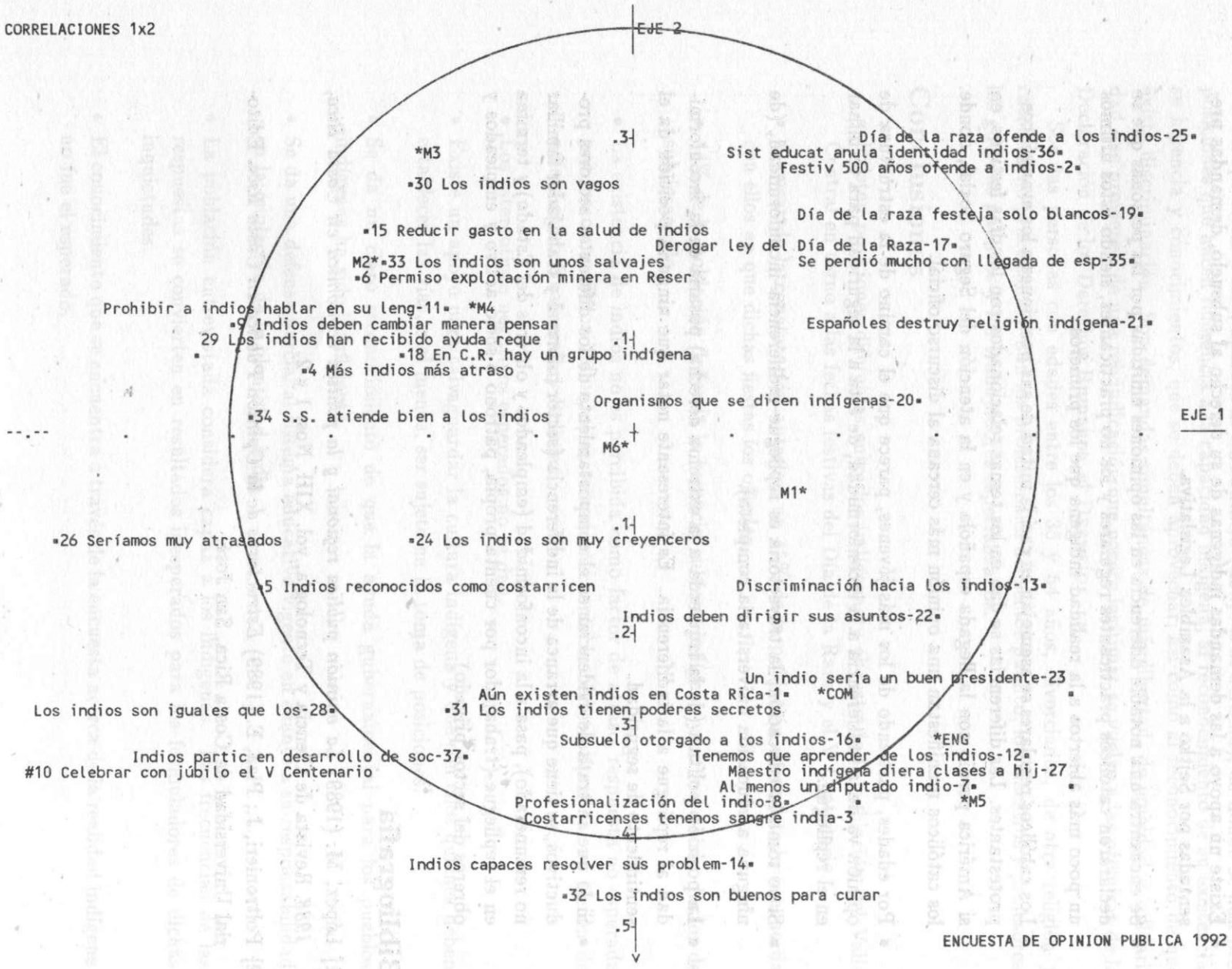
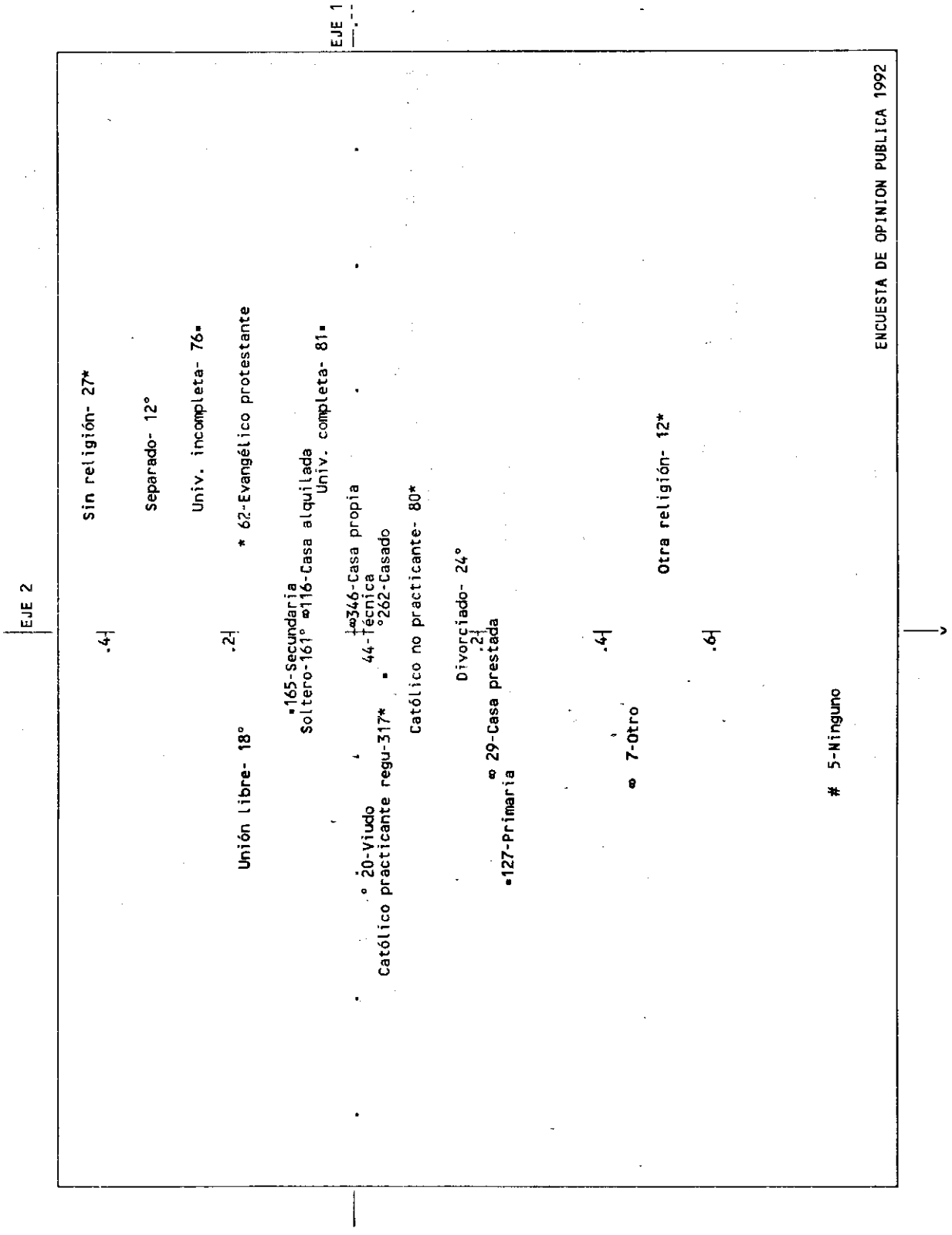


Figura 2: Religión, estado civil y nivel de educación



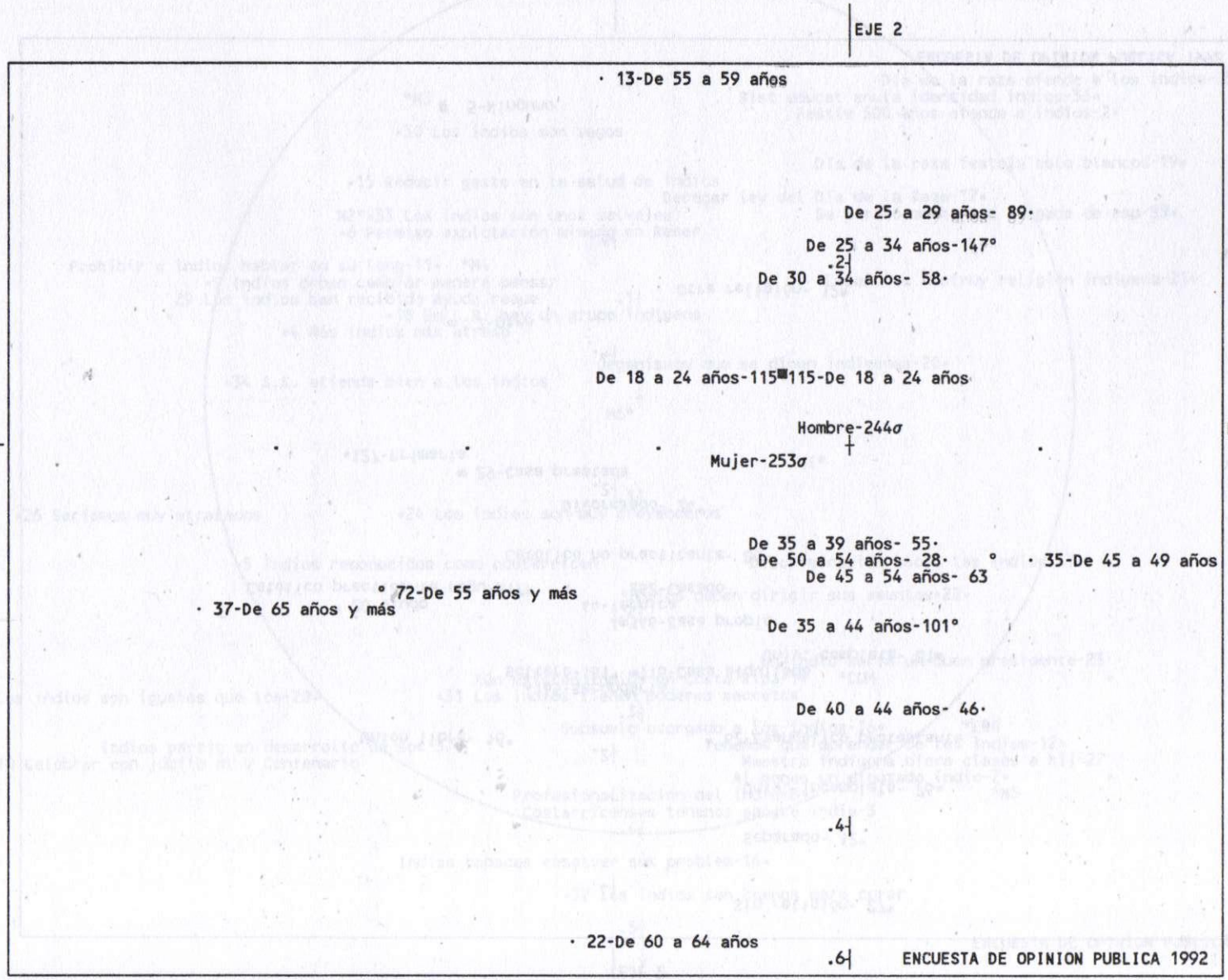


Figura 3: Sexo y edad

Procesos de Conteo y Análisis de Supervivencia

Jaime Lobo*

Resumen

En este artículo doy un recuento general del problema del análisis de supervivencia, y de los métodos matemáticos que se han elaborado al respecto. Daré mayor interés al vínculo con la teoría de los procesos de conteo, limitando la discusión al caso univariado.

1 El planteamiento clásico del análisis de supervivencia

Sea T una variable aleatoria positiva, tiempo de vida de un bombillo, por ejemplo, de distribución F. La función de supervivencia S se define por:

$$S(t) = 1 - F(t) = P(T > t)$$

Si S es positiva y diferenciable podemos definir la tasa de avería o tasa de intensidad ("hazard rate" en inglés) como :

$$l(t) = \lim_{h \to 0} P(T < t + h \mid T \geq t) = -\frac{d \log(S(t))}{dt}$$

y tenemos la relación :

$$S(t) = \exp\left(-\int_0^t l(u)\, du\right)$$

La medida L definida por $L(]0,t]) = \int_0^t l(u)\, du$ es llamada intensidad o medida de avería o medida de intensidad . Una extensa discusión de estos conceptos así como de sus aplicaciones en modelos particulares puede hallarse en el texto de Kalbfleisch y Prentice [3].

Se pueden generalizar estos conceptos sin restricciones sobre la función de supervivencia recurriendo a la noción de la integral producto (o integral de Volterra). Denotando por $\prod$ esta integral es posible demostrar que en un cierto intervalo las relaciones siguientes se cumplen :

*Escuela de Matemática, Universidad de Costa Rica

$$S(t) = \prod_{s \leq t} (1 - L(\{s\})) \exp\,(-L^c(t))$$

$$L(]0,t]) = -\int_0^t \frac{S(du)}{S(u-)}$$

siendo L^c la parte continua de L, con lo que se establece una relación biunívoca entre funciones S y L. Esta generalización es más reciente y ha sido sistematizada por ciertos autores para replantear la teoría de procesos de conteo en un marco funcional (ver Gill y Johansen [1], por ejemplo).

El problema general del análisis de supervivencia es el de construir y analizar modelos en diferentes situaciones para las funciones de supervivencia, o, lo que es equivalente, para las funciones de intensidad y el de establecer métodos estadísticos para estimar y probar hipótesis sobre S y L.

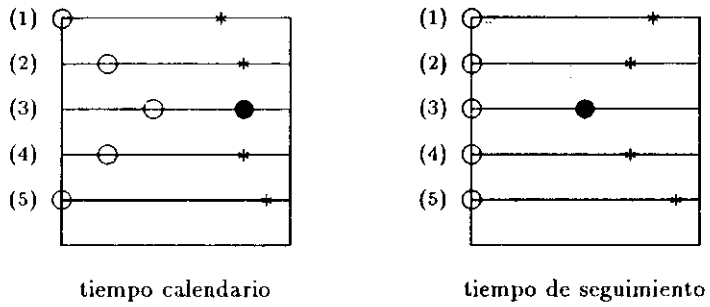
Para ello en las llamadas pruebas clínicas se efectúan experimentos siguiendo los tiempos de muerte (o averías) de diversos individuos a partir de su fecha de entrada (estudios longitudinales o de seguimiento , "follow–up studies"). La población en estudio es llamada generalmente cohorte. Cuando unidades experimentales (individuos) están todavía en operación (vivos) en el momento de clausura de la investigación y sus subsecuentes tiempos de avería son desconocidos, los datos de mortalidad son incompletos. Lo mismo si unidades de observación se pierden voluntaria o involuntariamente. La dificultad del análisis de supervivencia proviene precisamente de que, en general, los datos disponibles son incompletos, por lo que la teoría debe adaptarse a esta situación específica.

¿Cómo terminan las observaciones? Si el individuo no ha muerto todavía, se pueden presentar diversas situaciones:

- el tiempo de observación es preasignado en un tiempo fijo (fecha calendario); los datos resultantes son denominados truncados (censura tipo 1).
- el tiempo de observación termina cuando un determinado número de muertes han ocurrido. Los datos se llaman censurados (censura tipo 2).

Así entonces en la censura 1 el número de muertes es aleatorio mientras que en la censura 2 es el tiempo de término el que es aleatorio. En ambos casos los tiempos de muerte individuales son aleatorios.

Se distingue en este tipo de estudios entre tiempo calendario y tiempo de seguimiento de cada individuo, lo que gráficamente se representaría así:



o = entrada * = sale vivo • = muerto

A partir de las observaciones (truncadas, censuradas), se trata de estimar la función de supervivencia o la intensidad. Se trata de un problema no paramétrico donde los métodos clásicos como el de Kolmogorov–Smirnov o de Wilcoxon fallan, pues los datos son incompletos, aunque en algunos casos se pueden modificar.

Kaplan y Meier introducen en 1956 un estimador de la función de supervivencia, que actualmente se interpreta así : es, en la situación de observaciones censuradas de una distribución de vida, el producto límite de Volterra de la función de intensidad acumulada empírica. La función de intensidad acumulada empírica es el estimador dado por Nelson posteriormente y generalizado por Aalen (estimador de Nelson–Aalen). En [3] se puede hallar una presentación clásica de esta teoría, y en [1] un enfoque más moderno con la noción de integral producto.

Siguiendo a Gill se puede plantear así : en un modelo de censura clásico, se tienen $T_1, \ldots, T_n$ tiempos de vida independientes equidistribuidos de distribución F y supervivencia S y sean $C_1, \ldots, C_n$ variables aleatorias independientes equidistribuidas de censura, de distribución de supervivencia H.

Sean:

$$T_i^{\sim} = \min(T_i, C_i) \quad \text{y} \quad D_i = 1_{\{T_i \leq C_i\}};\ i = 1, \ldots, n$$

los datos realmente observados. Defina:

$$N_n(t) = \frac{1}{n} \# \{i : T_i^{\sim} \leq t,\ D_i = 1\}$$

$$Y_n(t) = \frac{1}{n} \# \{i : T_i^{\sim} \geq t\}$$

$$L_n(t) = \int_0^t Y_n^{-1} dN_n$$

$$S_n(t) = \prod_0^t (1 - dL_n)$$

Entonces S_n es el estimador de Kaplan-Meier de S y L_n el estimador de Nelson–Aalen de L. Se puede deducir de estas relaciones y de la ecuación de Duhamel que el proceso $S(t)/S_n(t) - 1$ es una martingala cuadrado integrable, lo que justifica el uso de las técnicas del cálculo estocástico en estos problemas.

2 Interpretación en términos de procesos de conteo

Un enfoque diferente al clásico del análisis de supervivencia consiste en modelar cada trayectoria o historia de vida individual como un proceso estocástico. Esto se logra de la manera de una manera muy sencilla: si T es el tiempo de avería (o muerte) de cada individuo se define el proceso en tiempo continuo definido como 1 en el intervalo $[0, T]$ y 0 fuera de él. Se trata pues de un proceso de conteo, en la terminología del análisis estocástico. El aporte decisivo para este planteamiento fue el dado por Aalen [4]. El marco matemático es el de los procesos de conteo multivariados y de integrales estocásticas , la teoría asintótica es basada en el teorema del límite central de martingalas como veremos a continuación.

Para entender el planteamiento de Aalen debemos introducir algunas nociones básicas sobre los procesos de conteo univariados . Una exposición clara de este tema es dada en [2]. Consideramos un espacio de probabilidad completo (Ω, F, P) y una familia creciente, continua a la derecha $(F_t, t \in [0,1])$ de subsigma álgebras de F. Un proceso estocástico $N = (N(t),\ t \in [0,1])$ es llamado un proceso de conteo si cada realización de N es una función escalonada continua a la derecha con un número finito de saltos, de altura 1. Se pone $N(0) = 1$ y además N es adaptado a (F_t) y suponemos $E(N(1)) < \infty$. Como N es una submartingala, tenemos de la descomposición de Doob–Meyer:

$$N = A + M$$

donde A es un proceso creciente predecible y M es una martingala.

Si se supone que A se escribe:

$$A(t) = \int_0^t L(s)ds$$

donde L es un proceso adaptado a (F_t) con límites a la derecha, entonces M es una martingala cuadrado integrable con proceso de varianza:

$$< M, M > (t) = \int_0^t L(s)ds$$

El proceso L es llamado intensidad de N. Se justifica por la propiedad:

$$\lim_{h\to 0} \frac{P(N(t+h)-N(t)=1 \mid F_t)}{h} = \lim_{h\to 0} \frac{E(N(t+h)-N(t) \mid F_t)}{h} = L(t+)$$

2.1 El modelo de intensidad multiplicativa: definiciones generales

Se considera un modelo estadístico $P = \{P_\theta, \theta \in \Theta\}$, familia de probabilidades en (Ω, F). Aalen supone que para cada $\theta \in \Theta$, el proceso intensidad L^θ con respecto a (F_t, P_θ) existe y que aún más existe un proceso Y $(F_t)-$ adaptado y funciones α^θ tales que:

$$L^\theta(t) = \alpha^\theta(t)Y(t)\ ,\ t \in [0,1]$$

donde α^θ es determinístico, mientras que Y no depende de θ. Se requiere que las trayectorias de Y sean no negativas y continuas a la izquierda con límites a la derecha; α^θ es también no negativo.

La inferencia estadística en el modelo de intensidad multiplicativa en el tiempo t es basada en la observación de $(N(s), Y(s), 0 \le s \le t \le 1)$, o más generalmente en la observación de la familia $(F_s, 0 \le s \le t)$. Se pueden encontrar discusiones sobre la clasificación de modelos con la estructura anterior y condiciones sobre α con el fin de parametrizar el modelo.

Ejemplo: ([2]): Si N está definido por $N(t) = \#\{i : X_i \le t\}$, donde $X_i, i = 1, \ldots, n$ son variables aleatorias independientes positivas de misma distribución con intensidad g. Entonces se puede calcular L directamente y obtener:

$$L(t) = g(t-)(n - N(t-))$$

y por lo tanto:

$$\alpha(t) = g(t-), \quad Y(t) = n - N(t-)$$

2.2 El modelo de intensidad multiplicativa: estimación

El razonamiento de Aalen para estimar α es el siguiente: simbólicamente se puede escribir:

$$dN = \alpha(t)Y(t)dt + \text{ ruido}$$

por lo tanto un estimador natural de:

$$\beta(t) = \int_0^t \alpha(s)ds$$

sería:

$$\int_0^t Y(s)^{-1} dN(s)$$

Sin embargo, como $Y = 0$ puede ocurrir, para contemplar esta posibilidad el problema es replanteado definiendo:

$$\beta^*(t) = \int_0^t \alpha(s)J(s)ds\,, \quad t \in [0,1]$$

donde:

$$J(s) = 1(Y(s) > 0) = \text{variable indicadora del evento } \{Y(s) > 0\}$$

Interpretando $J(t)/Y(t)$ como 0 siempre que $Y(t) = 0$ y suponiendo que existe una constante $c > 0$ tal que $Y(t) < c \Longrightarrow Y(t) = 0$ casi siempre, el estimador de β^* es definido por:

$$\beta\hat{}(t) = \int_0^t J(s)/Y(s)dN(s)$$

Se cumple ahora que $\beta\hat{} - \beta^*$ es una martingala cuadrado integrable con proceso de varianza:

$$< \beta\hat{} - \beta^* > (t) = \int_0^t \alpha(s)J(s)/Y(s)ds$$

El comportamiento asintótico de los estimadores y de los tests de hipótesis sobre α pueden entonces ser estudiados gracias a los resultados sobre comportamiento asintótico y distribuciones límites de martingalas.

2.3 El problema de la censura en el modelo de intensidad multiplicativa

Una ventaja importante de la formulación general del modelo de intensidad multiplicativa es que se adecua a muchos tipos diferentes de censura.

Supongamos que el intervalo de tiempo donde el proceso es observado es determinado por un proceso indicador $C(F_t)-$ adaptado, es decir C es la variable indicadora de un conjunto de $[0,1] \times \Omega$. Si C es predecible (en particular si es continuo a la izquierda), el proceso censurado:

$$N^C(t) = \int_0^t C(s)dN(s)$$

tiene como proceso intensidad:

$$\alpha(t)Y^C(t) = \alpha(t)Y(t)C(t)$$

Esto se ve observando que:

$$N^C(t) = \int_0^t C(s)\alpha(s)Y(s)ds + \int_0^t C(s)dM(s)$$

donde el último término es una martingala cuadrado integrable, siendo la integral estocástica de un proceso previsible con respecto a la martingala M. Un estudio de las "observaciones censuradas de N" es así equivalente al estudio del proceso de conteo N^C con intensidad multiplicativa αY^C.

Es importante hacer notar que no se restringe que el proceso de censura sea adaptado a la familia autoexcitada de sigma-álgebras (N_t) generada por el proceso N. Esto permite la dependencia del mecanismo de censura de influencias aleatorias externas fuera de los eventos del proceso de conteo mismo. Un ejemplo muy simple es el llamado modelo de censura aleatoria, donde el proceso de censura es estocásticamente independiente de N.

2.4 Prueba de una muestra

Sea $(N(t), t \in [0,1])$ un proceso de conteo unidimensional con proceso intensidad $\alpha(t)Y(t)$. Se quiere probar la hipótesis $H_0 : \alpha = \alpha_0$, donde α_0 es dada. Sean J y $\beta\hat{}$ definidos por N y defina:

$$\beta_0\hat{}(t) = \int_0^t \alpha_0(s)J(s)ds$$

Bajo H_0, $\beta\hat{} - \beta_0\hat{}$ es una martingala cuadrado integrable con proceso varianza:

$$< \beta\hat{} - \beta_0\hat{} >= \int_0^t \alpha_0(s)J(s)/Y(s)ds$$

Si K es un proceso previsible acotado c. s., entonces bajo H_0:

$$Z(t) = \int_0^t K(s)(d\beta\hat{}(s) - d\beta_0\hat{}(s))$$

es una martingala cuadrado integrable con:

$$< Z, Z > (t) = \int_0^t K^2(s)\alpha_0(s)J(s)/Y(s)ds$$

Resulta que es posible obtener como corolario del teorema del límite central de martingalas de Rebolledo [5] una ley límite para el proceso Z:

Teorema: Sea $N^{(n)}$, $n \in \mathbb{N}$, una sucesión de procesos y sean $Z^{(n)}$ los procesos definidos por cada proceso. Entonces si:

$$a_n = \left(< Z^{(n)}, Z^{(n)} > (1)\right)^{-1/2}$$

la sucesión de procesos $a_n Z^{(n)}$ converge en distribución al proceso de Wiener standard.

En particular, $U^{(n)} = Z^{(n)}(1)(< Z^{(n)}, Z^{(n)} > (1))^{-1/2}$ posee una distribución asintótica gaussiana standard y puede ser usada para probar H_0.

Observaciones finales

La breve exposición anterior muestra cómo el problema complejo del análisis de los datos de supervivencia puede inscribirse en el marco de los procesos de conteo, permitiendo así la aplicación de las técnicas del cálculo estocástico. En el modelo de Aalen se inscriben en efecto la mayoría de los modelos clásicos de datos censurados. Aunque lo he omitido en este informe, es posible también incluir en él los modelos multivariados, es decir poblaciones con funciones de supervivencia de diferentes tipos, en cuyo caso se recurre a procesos de conteo multivariados (ver [2] para una discusión general de este problema).

Como se pudo hacer notar, el estimador de Aalen, a diferencia de la mayoría de los estimadores clásicos, se obtiene de una manera casi heurística sin recurrir al principio de máxima verosimilitud. Sus buenas propiedades se derivan más bien del estudio de su comportamiento asintótico, para lo cual se emplean resultados muy fuertes de teoremas del límite central para martingalas. El punto de vista expuesto en Gill, por medio de la teoría de la integral producto, es muy cercano al de Aalen. Sin embargo, sus ideas están más ligadas al método de Von Mises, es decir a la teoría del análisis funcional.

Se puede entonces afirmar que sin lugar a dudas la teoría de los datos censurados ha logrado dar un gran salto en la década pasada con la introducción de este punto de vista.

Bibliografía

[1] R. Gill y S. Johansen (1987) *Product integrals and counting processes*, preprint.

[2] M. Jacobsen (1982) *Statistical Analysis of Counting Processes*, Springer Verlag, Lecture Notes in Statistics, 1982.

[3] J. Kalbfleisch, R. Prentice (1980) *The statistical analysis of failure time data*, Wiley, New York.

[4] Aalen (1976) *Statistical inference for a family of counting processes*, Impreso por el Institute of Mathematical Statistics, Universidad de Copenhague.

[5] R. Rebolledo (1980) *Central limit theorems for local martingales*, Zeitschrift für Wahrscheinlichkeitstheorie, 51.

Estrategias de aprendizaje en tiempo mínimo*

Ioan Muntean† Neculae Vornicescu‡

1 Introducción

Con el objeto de obtener conocimientos en un dominio dado en el transcurso de varias unidades fijas de tiempo (días, semanas, etc.), una persona toma para prepararse, un intervalo en cada unidad de tiempo. Se pone el problema de determinar estos intervalos de tal forma que al final se obtenga el mínimo necesario de conocimientos y la suma de los intervalos sea lo más pequeña posible.

G.F.Raggett, P.M. Hempson y K.A, Jukes [6] elaboraron un modelo matemático para el aprendizaje y resolvieron el problema de óptimo con ayuda del principio de máximo de Pontriaghin.

El análogo continuo del problema fue tratado por H. Bondi [2], con ayuda de métodos del cálculo variacional y W. Woodside [7] resolvió el problema con ayuda de técnicas de programación matemática. Finalmente M.S. Klamkin [4], dio recientemente una solución con métodos elementales basada en la desigualdad de Cauchy y en consideraciones geométricas.

En el presente trabajo se desarrolla y se fundamenta el método de M.S. Klamkin y se investiga el problema de óptimo en presencia de limitaciones superiores de los intervalos. En la sección 2 se describe el modelo matemático discreto del aprendizaje y se resuelve el problema de la existencia de las soluciones admisibles. La sección 3 contiene los resultados teóricos necesarios para el estudio de las estrategias ótimas. En las secciones 4 y 5 se deducen las fórmulas para cálculo efectivo de las estrategias óptimas en ausencia o en presencia de limitaciones de los intervalos. La última sección comprende ejemplos numéricos, conclusiones y comentarios.

*Presentado en el Simposio y traducido del rumano para su publicación por Edwin Castro.

†Universidad Babeş-Bolyai, Cluj-Napoca, Rumania

‡Universidad Técnica, Cluj-Napoca, Rumania

2 Modelo matemático del aprendizaje en tiempo mínimo

Se denota con $n \geq 1$ el número de unidades de tiempo en el transcurso de las cuales se va a desarrollar el programa de preparación: participación en cursos, seminarios, laboratorios, proyectos, exámenes, estudio individual, conferencias, etc. Se denota con c_0 el volumen inicial de conocimientos del dominio abordado con el cual el estudiante entra en el programa de preparación y $c > c_0$ el nivel mínimo de conocimientos que se desea alcanzar al final del estadio de preparación. Tomando en cada una de las n unidades de tiempo los intervalos de preparación $d_1 \geq 0, \ldots$, respectivamente $d_n \geq 0$; los experimentos ponen en evidencia las siguientes constataciones, admitidas aquí como hipótesis:

a) El volumen c_{i+1} de conocimientos alcanzados en las primeras $i+1$ unidades de tiempo es proporcional a la raíz cuadrada del intervalo d_{i+j} afectado por el estudiante para la preparación en el transcurso de la unidad de tiempo $i+1$.

b) El volumen c_{i+1} es proporcional con el volumen c_i de conocimientos alcanzados en las primeras i unidades de tiempo.

Las hipótesis de más arriba conducen a la relación de recurrencia:

$$c_{i+1} = mc_i + e\sqrt{d_{i+1}}\,, \quad i \in \{0, 1, \ldots, n-1\} \tag{1}$$

y a la restricción final:

$$c_n = c \tag{2}$$

Aquí $e \geq 0$ representa la eficiencia del tiempo afectado en la preparación y m, $0 \leq m \leq 1$ representa el coeficiente de memorización (existen situaciones cuando los parámetros e y m pueden depender de i).

Observemos que en las condiciones de más arriba, tenemos $e > 0$. En efecto si $e = 0$ de (1) (2) se llega a la contradicción:

$$c = c_n = mc_{n-1} = m^2 c_{n-2} = m^n c_0 \leq c_0 < c$$

Por este motivo, en lo que sigue vamos a admitir que $e > 0$.

De (1) podemos determinar el valor de c_n. La relación homogénea de recurrencia asociada a (1) es:

$$c_{i+1} = mc_i \quad i \in \{0, 1, \ldots, n-1\} \tag{3}$$

que conduce a $c_i = \alpha m^i$ con $\alpha \in \mathbb{R}$, lo cual sugiere que se puede buscar la solución de la relación de recurrencia (1) en la forma (3), es decir:

$$c_i = \alpha_i m^i \quad i \in \{0, 1, \ldots, n-1\} \tag{4}$$

Si suponemos al principio que $m > 0$ y sustituyendo (4) en (1) tenemos:

$$\alpha_{i+1} m^{i+1} = \alpha_i m^{i+1} + e\sqrt{d_{i+1}}$$

entonces:

$$\alpha_{i+1} = \alpha_i + em^{-i-1}\sqrt{d_{i+1}}, \quad i \in \{0, 1, \ldots, n-1\}$$

de donde después de sumar estas desigualdades se obtiene:

$$\alpha_n = \alpha_0 + e\sum_{i=1}^{n} m^{-i}\sqrt{d_i}$$

Como $\alpha_0 = c_0$, obtenemos:

$$c_n = m^n\left(c_0 + e\sum_{i=1}^{n} m^{-i}\sqrt{d_i}\right) \tag{5}$$

Cuando $m = 0$, de (1) con $i = n-1$ deducimos:

$$c_n = e\sqrt{d_n} \tag{6}$$

Así si $d_1, d_2, \ldots, d_n$ son números no negativos dados y $c_0, c_1, \ldots, c_n$ es una sucesión que satisface (1), entonces c_n se calcula según las fórmulas (5) ó (6) según sea $m > 0$, respectivamente $m = 0$. Recíprocamente de (5) y (6) resulta que $c_0, \ldots, c_n$ verifican la relación (1).

Denotamos con D el conjunto de todos los sistemas $d = (d_1, d_2, \ldots d_n)$ de n números $d_i \geq 0 \quad \forall i \in \{1, \ldots, n\}$ para los cuales se verifica (1) y (2).

Los elementos del conjunto D se llaman **estrategias (admisibles) de aprendizaje.**

Cuando $e > 0$ y $0 < m \leq 1$ de (2) resulta que (5) se puede escribir como:

$$\sum_{i=1}^{n} m^{-i}\sqrt{d_i} = r_n \quad \text{donde} \quad r_n = \frac{c - c_0 m^n}{em^n} \tag{7}$$

y el conjunto D toma la forma:

$$D = \{d = (d_1, \ldots, d_n) \in \mathbb{R}^n \mid 0 \leq d_i \quad \forall i \in \{1, \ldots, n\} \quad \text{y} \quad \sum_{i=1}^{n} m^{-i}\sqrt{d_i} = r_n\}$$

En relación con el modelo matamático de aprendizaje se proponen los siguientes problemas:

P_1 ¿Existe al menos una estrategia de aprendizaje?

P_2 ¿Existe una estrategia $d^* = (d_1^*, \ldots, d_n^*) \in D$ para la cual la suma de los intervalos destinados a la preparación sea mínima?, es decir:

$$\sum_{i=1}^{n} d_i^* \leq \sum_{i=1}^{n} d_i$$

para cualquier $d = (d_1, \ldots, d_n) \in D$

Las soluciones del problema P_2 se llaman **estrategias óptimas.**

En el modelo presentado más arriba se permite que los intervalos d_i sean grandes, esto no concuerda siempre con la realidad y con la "higiene del trabajo intelectual". Por consiguiente, se recomienda la modificación del modelo incluyendo la hipótesis de que los intervalos d_i no sobrepasen un límite preestablecido $d_0 > 0$ es decir d_i satisfaga también:

$$d_i \leq d_0 \quad \forall i \in \{1, 2, \ldots, n\} \tag{8}$$

El conjunto precedente D se sustituye ahora por:

$$D' = \{d = (d_1, \ldots, d_n) \in D | d_i \leq d_0 \quad \forall i \in \{1, 2, \ldots, n\}\}$$

y los problemas P_1 y P_2 se van transformar en los correspondientes P_1' y P_2' referentes al conjunto D' que incluye las restricciones (8) de limitación de los intervalos.

Los problemas P_1 y P_1' se resuelven inmediatamente. Se busca para ello una solución "uniforme" del problema P_1 es decir con intervalos iguales entre ellos $d_1 = \ldots = d_n = \delta$; si $m > 0$ de (2) y (5) se obtiene $c = c_n \leq m^n[c_0 + eS_n(m)\sqrt{\delta}]$ donde:

$$S_n(m) = \frac{1 - m^n}{m^n(1 - m)} \quad \text{cuando} \quad m < 1 \tag{9}$$

y $S_n(1) = n$.

Cuando $m = 0$, de (2) y (6) deducimos que $c = c_n \leq e\sqrt{\delta}$. Así pues en ambos casos es necesario que:

$$\delta \geq \left[\frac{c - c_0 m^n}{e m^n S_n(m)}\right]^2 > 0 \text{ con la convención } S_n(0) = 1 \tag{10}$$

y se tiene que $d_1 + \ldots + d_n = n\delta$. *Recíprocamente para cualquier* δ *verificando (10) tenemos* que $(\delta, \ldots, \delta) \in D$. Se observa luego que el problema P_1' tiene una solución si y sólo si

$$d_0 \geq \left[\frac{c - c_0 m^n}{e m^n S_n(m)}\right]^2 = \left[\frac{r_n}{S_n(m)}\right]^2 \tag{11}$$

En efecto si $(d_1, \ldots, d_n) \in D'$ de (2), (5) y (8) deducimos:

$$c = c_n \leq m^n[c_0 + eS_n(m)\sqrt{d_0}]$$

cuando $m > 0$ y de (2), (6) y (8) se deduce que:

$$c = c_n \leq e\sqrt{d_0}$$

cuando $m = 0$, es decir en ambos casos se llega a (11). Recíprocamente si admitimos (11), para cualquier número δ entre los números $\left[\frac{r_n}{S_n(m)}\right]^2$ y d_0 se tiene que $(\delta, \ldots, \delta) \in D'$ cuando $m = 0$ los problemas P_2 y p_2' se resuelve en forma semejante. En este caso para cualquier solución $(d_1^*, \ldots, d_n^*)$ del problema P_2, de (2) y (6) hallamos que $c = c_n = e\sqrt{d_n^*}$ de donde $d_n^* = \left(\frac{c}{e}\right)^2$ resulta $d_i^* = 0 \ \ \forall i \in \{1, \ldots, n-1\}$ y $d_n^* = \left(\frac{c}{e}\right)^2$. Recíprocamente $\left(0, \ldots, \left(\frac{c}{e}\right)^2\right) \in D'$ es la única solución de P_2.

Luego bajo la condición suplementaria $d_0 \geq \left(\frac{c}{e}\right)^2$ la solución precedente es la única del problema P_2'.

3 Condiciones de óptimo

En esta sección se presentan los resultados técnicos necesarios para la solución de los problemas P_2 y P_2' bajo la hipótesis $0 < e$, $0 < m \leq 1$ sobre los parámetros que aparecen en el modelo descrito en la sección precedente. La resolución del problema P_2 se basará en el caso de igualdad en la desigualdad bien conocida de Cauchy–Hölder [3].

3.1 Proposición:
Si $n \geq 1$ es un número natural p y q números mayores que 1 con

$$\frac{1}{p} + \frac{1}{q} = 1$$

a_i y b_i, $i \in \{1, \ldots, n\}$ son números reales no negativos, respectiva y estrictamente positivos, entonces tiene lugar la desigualdad:

$$\sum_{i=1}^{n} a_i b_i \geq \left(\sum_{i=1}^{n} a_i^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} b_i^q\right)^{\frac{1}{q}} \tag{12}$$

En (12) se cumple la desigualdad si y solo si existe $\lambda \geq 0$ tal que:

$$a_i^{p-1} = \lambda b_i \ , \ \ \forall i \in \{1, \ldots, n\}$$

Vamos a establecer ahora un resultado útil para resolver el problema P_2', para esto se da un número natural $n \geq 2$, un número real $t > 0$ y los números reales $b_1, \ldots, b_n$ verificando:

$$0 < b_1 \leq \ldots \leq b_n \tag{13}$$

y se considera el conjunto:

$$A = \{(a_1, \ldots, a_n) \in \mathbb{R}^n | 0 \leq a_j \leq t \cdot \sum_{i=1}^{n} a_i b_i \quad \forall j \in \{1, \ldots, n\}\}$$

Es claro que $(0, \ldots, 0) \in A$. Luego $A \neq \{(0, \ldots, 0)\}$ si y sólo si

$$1 \geq t \cdot \sum_{i=1}^{n} b_i \tag{14}$$

En efecto, si existe $(a_1, \ldots, a_n) \in A \setminus \{(0, \ldots, 0)\}$, entonces

$$\sum_{j=1}^{n} a_j b_j \leq t \sum_{i=1}^{n} a_i b_i \sum_{j=1}^{n} b_j$$

lo cual implica (14), luego, admitiendo (14) y poniendo $a_j = \left(\sum_{i=1}^{n} b_i\right)^{-1} \quad \forall j \in \{1, \ldots, n\}$ tenemos $(a_1, \ldots, a_n) \in A \setminus \{(0, \ldots, 0)\}$.

Suponiendo que se cumple (13) y denotando:

$$S_i = b_i(b_1^2 + \ldots + b_i^2 + b_i b_{i+1} + \ldots + b_i b_n)^{-1}, \; i \in \{1, \ldots, n\} \tag{15}$$

tenemos que $S_1 \leq S_2 \leq \ldots \leq S_n$ pues:

$$S_{i+1} - S_i = \frac{(b_1^2 + \ldots + b_i^2)(b_{i+1} - b_i)}{(b_1^2 + \ldots + b_i^2 + b_i b_{i+1} + \ldots + b_i b_n)(b_1^2 + \ldots + b_{i+1}^2 + b_{i+1} b_{i+2} + \ldots + b_{i+1} b_n)} \geq 0$$

para todo $i \in \{1, \ldots, n-1\}$.

A la par de (14) también introducimos la condición:

$$t \sum_{i=1}^{n} b_i^2 < b_n \tag{16}$$

y observamos que, de acuerdo con (14), (15) y (16), se tiene que:

$$S_1 = \frac{1}{b_1 + \ldots + b_n} \leq t < \frac{b_n}{b_1^2 + \ldots + b_n^2} = S_n$$

Por lo tanto, existe un único $k \in \{1, \ldots, n-1\}$ tal que:

$$S_k \leq t < S_{k+1} \tag{17}$$

3.2 Proposición:
Si tienen lugar las condiciones (13), (14) y (16), entonces $\forall (a_1, \ldots, a_n) \in A$ se verifica la desigualdad:

$$\sum_{i=1}^{n} a_i^2 \geq \left(\sum_{i=1}^{n} a_i b_i \right)^2 \left\{ (n-k)t^2 + [1 - t(b_{k+1} + \ldots + b_n)]^2 \left(\sum_{i=1}^{n} b_i^2 \right)^{-1} \right\} \tag{18}$$

Demostración: De (17) resulta:

$$t < S_{k+1} = \frac{1}{b_{k+1} + \ldots + b_n} \leq \frac{1}{b_{i+1} + \ldots + b_n}, \forall i \in \{k, \ldots, n-1\}$$

por lo tanto los números:

$$t_i = \frac{t}{1 - t(b_{i+1} + \ldots + b_n)} \quad i \in \{k, \ldots, n-1\} \tag{19}$$

están bien definidos y son positivos.

Para todo $(a_1, \ldots, a_n) \in A$ se tiene:

$$a_j \leq t_i \sum_{l=1}^{i} a_l b_l \quad \forall\, j \in \{1, \ldots, n\}, \forall\, i \in \{k, \ldots, n-1\} \tag{20}$$

En efecto denotando $S = t \sum_{l=1}^{n} a_l b_l$ tenemos $a_l \geq S \quad \forall\, l \in \{1, \ldots, n\}$ y $S \leq t \sum_{l=1}^{i} a_l b_l + tS \sum_{l=i+1}^{n} b_l$ lo cual conduce a (20).

$$a_j \leq S \leq \frac{t}{1 - t(b_{i+1} + \ldots + b_n)} \sum_{l=1}^{i} a_l b_l = t_i \sum_{l=1}^{i} a_l b_l$$

Observe que, si $\sum_{l=1}^{i} a_l b_l = 0 \quad \forall\, i \in \{k, \ldots, n-1\}$ de (20) sigue que $a_j \; \forall\, j \in \{1, \ldots, n\}$ y (18) está demostrado. Por consiguiente, se puede suponer que:

$$\sum_{l=1}^{i} a_l b_l \neq 0 \quad \forall\, i \in \{k, \ldots, n-1\} \tag{21}$$

Vamos a verificar (18) bajo la hipótesis (21) admitiendo en principio que $n \geq 3$.

En este caso los números t_i dados por (19) verifican

$$t_i < \frac{b_i}{b_1^2 + \ldots + b_i^2} \quad \forall\, i \in \{k+1, \ldots, n-1\} \tag{22}$$

Este hecho se ve claramente de las relaciones:

$$t < S_{k+1} \leq \frac{b_i}{b_1^2 + \ldots b_i^2 + b_i(b_{i+1} + \ldots + b_n)}$$

Asociamos a cada sistema $(a_1, \ldots, a_n) \in A$ la función $f : \{k, \ldots, n-1\} \to \mathbb{R}$ definida por la fórmula:

$$\begin{aligned} f(i) &= (n-i)t^2 + \frac{[1 + t(b_{i+1} + \ldots + b_n]^2}{(a_1b_1 + \ldots + a_ib_i)^2}(a_1^2 + \ldots + a_n^2) \\ &= t^2\left[n - i + \frac{a_1^2 + \ldots + a_1^2}{t^2(a_1b_1 + \ldots + a_ib_i)^2}\right] \end{aligned} \tag{23}$$

y probemos que:

$$f(i+1) \geq f(i) \ , \quad \forall i \in \{k, \ldots, n-1\} \tag{24}$$

Para simplificar la escritura, introducimos las notaciones (útiles también en lo que sigue):

$$\begin{aligned} \alpha_{i+1} &= a_1^2 + \ldots + a_{i+1}^2 \\ \beta_{i+1} &= b_1^2 + \ldots + b_{i+1}^2 \\ \gamma_{i+1} &= a_1b_1 + \ldots + a_{i+1}b_{i+1} \end{aligned} \tag{25}$$

y se constata que:

$$\begin{aligned} f(i+1) - f(i) &= t^2\left[n - i - 1 + \frac{\alpha_{i+1}}{t_{i+1}^2\gamma_{i+1}^2} - n + i - \frac{\alpha_{i+1} - \alpha_{i+1}^2}{t_i^2(\gamma_{i+1} - a_{i+1}b_{i+1})^2}\right] \\ &= t^2\left[\frac{\alpha_{i+1}}{t_{i+1}^2\gamma_{i+1}^2} - 1 - \frac{\alpha_{i+1} - \alpha_{i+1}^2}{t_i^2(\gamma_{i+1} - a_{i+1}b_{i+1})^2}\right] \end{aligned} \tag{26}$$

De (20) tenemos que $a_{i+1} \leq t_{i+1}\gamma_{i+1}$ y $0 \leq x \leq t_{i+1}\gamma_{i+1}$ y (22) implica que: $b_{i+1}x < \frac{\gamma_{i+1}}{\beta_{i+1}}b_{i+1}^2$ de donde $\gamma_{i+1} - b_{i+1}x > \gamma_{i+1} - \frac{\gamma_{i+1}}{\beta_{i+1}}b_{i+1}^2 = \frac{\gamma_{i+1}}{\beta_{i+1}}(b_1^2 + \ldots + b_i^2) > 0$.

Por consiguiente, la función $g : [a_{i+1}, t_{i+1}\gamma_{i+1}] \to \mathbb{R}$ está bien definida por la fórmula:

$$g(x) = -\frac{\alpha(i+1) - x^2}{(\gamma_{i+1} - b_{i+1}x)^2}$$

Su derivada verifica:

$$\begin{aligned} g'(x) &= 2\left(\frac{\gamma_{i+1}x - \alpha_{i+1}b_{i+1}}{(\gamma_{i+1} - b_{i+1}x)^3}\right) \\ &= \frac{2b_{i+1}}{(\gamma_{i+1} - b_{i+1}x)^3}\left(\frac{\gamma_{i+1}}{b_{i+1}}x - \alpha_{i+1}\right) \\ &< 0 \end{aligned}$$

pues $a_{i+1} \leq x \leq t_{i+1}\gamma_{i+1} < \dfrac{\gamma_{i+1}}{\beta_{i+1}}b_{i+1}$ implica $\dfrac{\gamma_{i+1}}{b_{i+1}}x < \dfrac{1}{\beta_{i+1}}\gamma_{i+1}^2 \leq \alpha_{i+1}$. Si se tiene en cuenta la desigualdad (19). Se deduce que g es estrictamente decreciente en $[a_{i+1}, t_{i+1}\gamma_{i+1}]$ y entonces $g(a_{i+1}) \geq g(t_{i+1}\gamma_{i+1})$.

Retomando ahora (26), de la última desigualdad resulta (24):

$$\begin{aligned} f(i+1) - f(i) &= t^2\left[\frac{\alpha_{i+1} - t_{i+1}^2\gamma_{i+1}^2}{t_{i+1}^2\alpha_{i+1}^2} + \frac{1}{t_i^2}g(a_{i+1})\right] \\ &\geq t^2\left[\frac{\alpha_{i+1} - t_{i+1}^2\gamma_{i+1}^2}{t_{i+1}^2\alpha_{i+1}^2} + \frac{1}{t_i^2}g(t_{i+1}\gamma_{i+1}^2)\right] \\ &= t^2\left[\frac{\alpha_{i+1} - t_{i+1}^2\gamma_{i+1}^2}{t_{i+1}^2\gamma_{i+1}^2} - \frac{\alpha_{i+1} - t_{i+1}^2\gamma_{i+1}^2}{t_i^2(\gamma_{i+1} - b_{i+1}t_{i+1}\gamma_{i+1})^2}\right] \\ &= t^2\frac{\alpha_{i+1} - t_{i+1}^2\gamma_{i+1}^2}{\gamma_{i+1}^2}\left[\frac{1}{t_{i+1}^2} - \frac{1}{t_i^2(1 - b_{i+1}t_{i+1})^2}\right] \\ &= 0. \end{aligned}$$

pues

$$\frac{1}{t_{i+1}} - \frac{1}{t_i} = b_{i+1} \quad t_i = \frac{t_{i+1}}{s - b_{i+1}t_{i+1}}$$

En particular de (24) deducimos que:

$$f(n-1) \geq f(k) \tag{27}$$

Vamos a probar que:

$$\frac{a_1^2 + a_2^2 + \ldots + a_n^2}{(a_1 b_1 + \ldots + a_n b_n)^2} \geq f(n-1) \tag{28}$$

lo cual con base en las notaciones (28) es igual a:

$$\frac{\alpha_n}{\gamma_n^2} \geq t^2 \left[\frac{1}{t_{n-1}^2} \cdot \frac{\alpha_n - a_n^2}{(\gamma_n - a_n b_n)^2} + 1 \right] \tag{29}$$

Para establecer (29) y por lo tanto (28), observamos que: $a_n \leq t\gamma_n$ y que $0 \leq x \leq t\gamma_n$ implica $b_n x \leq b_n \gamma_n t$ entonces $\gamma_n - b_n x \geq \gamma_n(1 - b_n t) > 0$, si se tiene en cuenta (16). Por consiguiente, la función: $h : [a_n, t\gamma_n] \to \mathbb{R}$ está bien definida por:

$$h(x) = \frac{\alpha_n - x^2}{(\gamma_n - b_n x)^2}$$

Tenemos:

$$h'(x) = -2\frac{\gamma_n x - \alpha_n b_n}{(\gamma_n - b_n x)^3} = \frac{2b_n}{(\gamma_n - b_n x)^3}\left(\alpha_n - \frac{\gamma_n}{b_n} x\right) > 0$$

pues $a_n \leq x \leq t\gamma_n < \dfrac{b_n \gamma_n}{b_1^2 + \ldots + b_n^2} = \dfrac{b_n \gamma_n}{\beta_n}$ implica: $\alpha_n \geq \dfrac{\gamma_n^2}{\beta_n} > t\dfrac{\gamma_n^2}{b_n} \geq \dfrac{\gamma_n}{b_n} x$, si se tiene cuenta (12). En consecuencia h es estrictamente creciente en $[a_n, t\gamma_n]$ y entonces $h(a_n) \leq h(t\gamma_n)$. De aquí resulta la desigualdad (29):

$$\begin{aligned}
\frac{\alpha_n}{\gamma_n^2} &= t^2 \left[\frac{\alpha_n t^2 \gamma_n^2}{(1 - tb_n)^2} \frac{1}{t^2(\gamma_n - t\gamma_n b_n)^2} + 1 \right] \\
&= t^2 \left[\frac{1}{\dfrac{1}{(1 - tb_n)^2} t^2} h(t\gamma_n) + 1 \right] \\
&\geq t^2 \left[\frac{1}{\dfrac{1}{(1 - tb_n)^2} t^2} h(a_n) + 1 \right] \\
&= t^2 \left[\frac{1}{t_{n-1}^2} \frac{\alpha_n - a_n^2}{(\gamma_n - a_n b_n)^2} + 1 \right]
\end{aligned} \tag{30}$$

De (23), (27), (28) y $\alpha_k \beta_k \geq \gamma_k^2$ resulta (18).

En el caso que queda $n = 2$ la condición (17) deviene $S_1 \leq t < S_2$ entonces $k = 1$ y (18) *deviene:*

$$\lambda_1^2 + \lambda_2^2 \geq (a_1 b_1 + a_2 b_2)^2 \left[t^2 + \frac{(1 - tb_2)^2}{b_1^2} \right] \tag{31}$$

Cuando $a_1 b_1 = 0$ de (20) resulta $a_1 = a_2 = 0$ y (31) es evidente. Se puede entonces suponer que $a_1 b_1 \neq 0$ y (31) se puede escribir bajo la forma:

$$\frac{a_1^2 + a_2^2}{(a_1 b_1 + a_2 b_2)^2} \geq t^2 + \frac{(1 - t b_2)^2}{b_1^2} \tag{32}$$

donde las notaciones son las de (25). Se utiliza luego el hecho que la función h, introducida más arriba, es estrictamente creciente en $[a_2, t\alpha_2]$ y se hace uso de (30) para obtener (32).

4 Estrategias óptimas sin límite de intervalos

Admiten las condiciones $0 < e$, $0 < m \leq 1$ y denotamos con $(d_1, d_2, \ldots, d_n)$ una solución del problema P_1.

Utilizando la Proposición 3.1 con $p = q = 2$, $a_i = \sqrt{d_i}$ y $b_i = m^{-i}$ obtenemos:

$$\left(\sum_{i=1}^{n} \sqrt{d_i}\, m_i^{-i} \right)^2 \leq \sum_{i=1}^{n} d_i \sum_{i=1}^{n} m^{-2i} \tag{33}$$

de donde, con base en (2) y (5) obtenemos:

$$\sum_{i=1}^{n} d_i \geq \left(\sum_{i=1}^{n} \sqrt{d_i}\, m_i^{-i} \right)^2 \left(\sum_{i=1}^{n} m^{-2i} \right) = E_n \tag{34}$$

$$E_n = \frac{1}{S_n(m^2)} r_n^2$$

las notaciones son las de (7) y (9).

Sea ahora $d^* = (d_1^*, \ldots, d_n^*)$ una solución del problema P_1, para el cual se realiza el signo de igualdad en (33) y por lo tanto en (34). Entonces para cualquier solución $(d_1, \ldots, d_n)$ de P_1 tenemos con base en (34).

$$\sum_{i=1}^{n} d_i \geq E_n = \sum_{i=1}^{n} d_i^*$$

es decir, d^* es una solución del problema P_2.

El hecho que d^* realiza el signo de igualdad en (33) equivale, de acuerdo con la Proposición 3.1, con la existencia de un número $\lambda \geq 0$ con:

$$\sqrt{d_i^*} = \lambda m^{-i}, \quad \forall\, i \in \{1, \ldots, n\} \tag{35}$$

De (2), (5) y (35) obtenemos:

$$\frac{c - c_0 m^n}{m^n} = e \sum_{i=1}^{n} \lambda m^{-2i} = e\lambda S_n(m^2)$$

de donde, con la notación de (7) se halla:

$$\lambda = \frac{c - c_0 m^n}{e m^n S_n(m^2)} = \frac{r_n}{S_n(m^2)} \tag{36}$$

y entonces, de acuerdo con (36):

$$d_i = \lambda^2 m^{-2i} = \left[\frac{r_n}{S_n(m^2)}\right]^2 m^{-2i}, \quad \forall\, i \in \{1, \ldots, n\} \tag{37}$$

Recíprocamente por verificación directa se constata que los números $d_1^*, \ldots, d_n^*$ dados por (37) son positivos y satisfacen $\sum_{i=1}^{n} m^{-i}\sqrt{d_i} = r_n$, entonces

$$d^* = (d_1^*, \ldots, d_n^*) \in D$$

más aún como los números $d_1^*, \ldots, d_n^*$ verifican (35), de lo establecido en el párrafo 3 se sigue que d^* es la única solución del problema P_2 y entonces d^* es la única estrategia óptima.

Las igualdades (37) prueban que la estrategia óptima $d^* = (d_1^*, \ldots, d_n^*)$ es una progresión geométrica con primer término $\left(\frac{\lambda}{m}\right)^2$ y con razón m^{-2}.

Resumiendo las consideraciones precedentes llegamos al resultado que sigue:

4.1 Teorema:
Si $0 < e$, $0 < m \leq 1$ entonces el problema P_2 tiene una única solución $d^* = (d_1^*, \ldots, d_n^*)$ y está dada por las fórmulas (37). Así pues el intervalo más corto σ^* destinado al aprendizaje es:

$$\sigma^* = \sum_{i=1}^{n} d_i^* = \left[\frac{r_n}{S_n(m^2)}\right]^2 \sum_{i=1}^{n} m^{-2i} = \frac{r_n^2}{S_n(m^2)} \tag{38}$$

5 Estrategias óptimas en presencia de la limitación de los intervalos

En presencia de la condición (11) y de la restricción (8) de limitación de los intervalos vamos a poner en evidencia las estrategias óptimas para el problema P_2'.

Podemos admitir que $n \geq 2$ (cuando $n = 1$ y $(\frac{c - c_0}{e})^2 \leq d_0$ entonces $d_1^* = \left(\frac{c - c_0}{e}\right)^2$ es la única solución del problema P_2'). Los números

$$b_i = m^{-i}, \quad i \in \{1, \ldots, n\}$$

verifican (13) y el número $t = \frac{\sqrt{d_0}}{r_n} > 0$ verifica por (11) la desigualdad (14). Admitimos en principio que t satisface también (16) y entonces:

$$t < S_n = \frac{b_n}{b_1^2 + \ldots + b_n^2} = \frac{1}{m^n S_n(m^2)}$$

Entonces existe un único $k \in \{1, \ldots, n-1\}$ que verifica (17).

Si $d = (d_1, \ldots, d_n) \in D'$ se tiene que $0 \leq d_j \leq 0, j \in \{1, \ldots, n\}$ y: $\sum_{i=1}^{n} m^{-1}\sqrt{d_i} = r_n$, entonces:

$$0 \leq d_j \leq d_0 = t^2 r_n^2 = t^2 \left(\sum_{i=1}^{n} m^{-i}\sqrt{d_i}\right)^2, \quad \forall\, j \in \{1, \ldots, n\}$$

lo cual prueba que $(a_1, \ldots a_n) \in A$ si se pone $a_j = \sqrt{d_j}$.

De la Proposición 3.2 resulta ahora que:

$$\begin{aligned}
\sum_{i=1}^{n} d_i &\geq \left(\sum_{i=1}^{n} m^{-i}\sqrt{d_i}\right)^2 \left\{(n-k)t^2 + \frac{\left[1 - t(m^{-k-1} + \ldots + m^{-n})\right]^2}{m^{-2} + \ldots + m^{-2k}}\right\} \\
&= r_n^2 \left\{(n-k)\frac{d_0}{r_n^2} + \frac{\left[1 - \sqrt{d_0}\frac{1}{r_n}(m^{-k-1} + \ldots + m^{-n})\right]^2}{m^{-2} + \ldots + m^{-2k}}\right\} \qquad (39) \\
&= (n-k)d_0 + \frac{r_n^2}{S_k(m^2)}\left[1 - \frac{\sqrt{d_0}}{r_n}P_k(m)\right]^2 \\
&= (n-k)d_0 + \frac{q_k^2}{S_k(m^2)}
\end{aligned}$$

donde $P_k(m) = \frac{1 - m^{n-k}}{(1-m)m^n}$ cuando $m < 1$, $P_k(1) = n - k$, $q_k = r_n - \sqrt{d_0}\, P_k(m)$.

Definimos ahora un sistema $d^* = (d_1^*, \ldots, d_n^*)$ por:

$$d_i^* = \begin{cases} \left[\dfrac{q_k}{S_k(m^2)}\right]^2 m^{-2i} & \text{si } i \in \{1, \ldots, k\} \\ d_0 & \text{si } i \in \{k+1, \ldots, n\} \end{cases} \tag{40}$$

Vamos a probar que $d^* \in D'$ y que d^* es una estrategia óptima para P_2'.

Para verificar $d_i^* \leq d_0 \quad \forall\, i \in \{1, \ldots, n\}$ será suficiente constatar que $\sqrt{d_k^*} \leq \sqrt{d_0}$ es decir:

$$\frac{q_k}{S_k(m^2)} m^{-k} \leq \sqrt{d_0} \text{ ó } \frac{1}{m^k S_k(m^2)} \left[r_n - \sqrt{d_0}\ P_k(m)\right] \leq \sqrt{d_0}$$

o:

$$r_n \leq \sqrt{d_0}\ [m^k S_k(m^2) + P_k(m)] \tag{41}$$

En esta forma la desigualdad (41) es inmediata:

$$\begin{aligned} \frac{\sqrt{d_0}}{r_n} &= t \geq S_k \\ &= \frac{m^{-k}}{m^{-2} + \ldots + m^{-2k} + m^{-k}(m^{-k-1} + \ldots + m^{-n})} \\ &= \frac{m^{-k}}{S_k(m^2) + m^{-k} P_k(m)} \end{aligned}$$

Finalmente para cualquier $(d_1, \ldots, d_n) \in D'$ de (39) y (40) obtenemos:

$$\begin{aligned} \sum_{i=1}^{n} d_i^* &= \sum_{i=1}^{k} \left[\frac{q_k}{S_k(m^2)}\right]^2 m^{-2i} + (n-k) d_0 \\ &= \left[\frac{q_k}{S_k(m^2)}\right]^2 S_k(m^2) + (n-k) d_0 \\ &\leq \sum_{i=1}^{n} d_i \end{aligned}$$

Cuando el número $t = \dfrac{\sqrt{d_0}}{r_n}$ no verifica (16), es decir:

$$t \geq S_n = \frac{1}{m^n S_n(m^2)}$$

entonces:

$$d_0 = r_n^2 t^2 \geq \left[\frac{r_n}{m^n S_n(m^2)}\right]^2 \tag{42}$$

y el sistema $(d_1^*, \ldots, d_n^*)$ dado de (40) con $k = n$ entonces, como de (37) va a ser una estrategia óptima para el problema P_2', en efecto para $i \in \{1, \ldots, n\}$ de (42) y $q_n = r_n$ deducimos que:

$$\begin{aligned} d_i^* &= \left[\frac{r_n}{S_n(m^2)}\right]^2 m^{-2i} \\ &\leq \left[\frac{r_n}{S_n(m^2)}\right]^2 m^{-2n} \\ &\leq d_0. \end{aligned}$$

Luego para cualquier $(d_1, \ldots, d_n) \in D'$ de (34) obtenemos:

$$\begin{aligned} \sum_{i=1}^{n} d_i^* &= \sum_{i=1}^{n} \left[\frac{r_n}{S_n(m^2)}\right]^2 m^{-2i} \\ &\leq \left[\frac{r_n}{S_n(m^2)}\right]^2 S_n(m^2) \\ &= E_n \\ &\leq \sum_{i=1}^{n} d_i \end{aligned}$$

Resumiendo las consideraciones anteriores llegamos al resultado:

5.1 Teorema:
Si $0 < e, \quad 0 < m \leq 1$ y $\dfrac{r_n}{S_n(m)} \leq \sqrt{d_0}$ entonces el problema P_2' tiene una solución $(d_1^*, \ldots, d_n^*)$ y está dada por las fórmulas (43) cuando existe $k \in \{1, \ldots, n-1\}$, verificando $S_k \leq \dfrac{\sqrt{d_0}}{r_n} \leq S_{k+1}$ y por las fórmulas (40) cuando $\dfrac{\sqrt{d_0}}{r_n} \geq S_n$. En el primer caso el intervalo más corto total σ^*, destinado al aprendizaje es:

$$\sigma^* = \sum_{i=1}^{n} d_i^* = (n-k)d_0 + \frac{1}{S_k(m^2)}\left[r_n - \sqrt{d_0}\ P_k(m)\right]^2 \tag{43}$$

Se observa que la resolución del problema P_2' con ayuda del Principio discreto de máximo ([1] pág. 409-410), necesita precauciones, pues la derivada de la función $d_i \mapsto \sqrt{d_i}$ es infinita en $d_i = 0$.

5.2 Observación:
Nos proponemos determinar el número $k \in \{1, \ldots, n-1\}$ del Teorema 5.1 en el caso que exista. Para esto denotamos $x = m^{-k}$ y se observa que:

$$\begin{aligned} t &= \frac{\sqrt{d_0}}{r_n} \geq S_k \\ &= \frac{m^{-k}}{m^{-2} + \ldots + m^{-2k} + m^{-k}(m^{-k-1} + \ldots + m^{-n})} \\ &= \frac{(m^2-1)x}{1 - x^2 + (m+1)(x^2 - m^{-n}x)} \\ &> 0 \end{aligned}$$

de donde resulta que x satisface la desigualdad $E(x) \leq 0$ en la cual:

$$E(x) = x^2 - \frac{(1+m)(m^{-n}t - 1 + m)}{mt}x + \frac{1}{m} \tag{44}$$

Denotando con x_2 la mayor raíz (positiva) de la ecuación $E(x) = 0$ el número k va a ser el mayor número del conjunto $\{1, \ldots, n-1\}$ para el cual $m^{-k} \leq x_2$.

6 Ejemplos, conclusiones y comentarios

6.1 Ejemplos

Los dos ejemplos presentados más abajo se refieren a los siguientes datos numéricos comunes:
$n = 10$ días, $c_0 = 1$ volumen (= 100 páginas)
$c = 4$ volúmenes (= 400 páginas), $e = 0.25$,
$m = 0.85$

y los intervalos se expresan todos en horas.

En el primer ejemplo falta la restricción de limitación de los intervalos y en el segundo, esta restricción está presente con $d_0 = 9$ horas.

Tenemos: $m^{-2} = 1.3841$, $S_n(m^2) = 89.3697$, $\lambda = 0.7476$; luego: $r_n = 77.2719$ y $t = \dfrac{\sqrt{d_0}}{r_n} = 0.0388$.

La ecuación $E(x)$ de la Observación 5.2 se escribe como:

$$x^2 - 2.6464x + 1.1765 = 0$$

Con la mayor raíz $x^2 = 2.0812$ y entonces $k = 4$. Con base en los teoremas 4.1 y 5.1, las estrategias óptimas existen y junto con el volumen de conocimiento correspondientes adquiridos en el programa se resumen en la tabla que sigue:

i	Ejem.1: sin restricc.		Ejem.2: con restricc.	
	d_i	c_i	d_i	c_i
0		1		1
1	1.0347	1.1042	2.9287	1.2778
2	1.4321	1.2378	4.0536	1.5894
3	1.9822	1.4041	5.6106	1.9432
4	2.7436	1.6076	7.7657	2.3484
5	3.7974	1.8536	9.0000	2.7463
6	5.2560	2.1487	9.0000	3.0843
7	7.2747	2.5001	9.0000	3.3717
8	10.0690	2.9189	9.0000	3.6159
9	13.9364	3.4143	9.0000	3.8236
10	19.2894	4.0002	9.0000	4.0001
	66.8155		74.3586	

6.2 Conclusiones y Comentarios

a) Se ha visto que si $e = 0$ es decir la eficiencia del tiempo afectado en la preparación es nula, entonces los problemas P_1 y P_1' y entonces con mayor razón P_2 y P_2' no tienen solución. Vamos a suponer que $e > 0$.

b) Si $m = 0$ es decir el coeficiente de memorización es nulo entonces la solución del problema P_2 con $n \geq 2$ está dada por:

$$d_1^* = \ldots = d_{n-1}^* = 0, \quad d_n^* = \left(\frac{c}{e}\right)^2$$

en otras palabras la cantidad total de tiempo destinado a la preparación se coloca en la última unidad de tiempo del estadio.

Cuando $m = 1$ es decir el estudiante asimila todo lo que aprende, de (37) se sigue que los intervalos óptimos en preparación en la unidad de tiempo para el problema P_2 son iguales entre ellos:

$$d_1^* = \ldots = d_n^* = \left(\frac{c - c_0}{ne}\right)^2$$

c) Cuando $0 < m < 1$ de (37) se sigue que: $d_1^* < d_2^* < \ldots < d_n^*$ y entonces los intervalos son cada vez mayores: ellos forman una progresión geométrica estrictamente creciente.

La estrategia admisible dada en la sección 2 para P_1:

$$d_1 = \ldots = d_n = \delta = \left[\frac{c - c_0 m^n}{e m^n S_n(m)}\right]^2$$

no es óptima, sin embargo tiene la ventaja que provee una dosis uniforme, más racional y más fácil de soportar.

d) Los problemas P_1 y P_2 en los cuales falta la restricción de limitación superior de los intervalos se pueden ver como casos límites de los problemas P_1' y P_2 en los cuales $d_0 = +\infty$, observamos luego que las primeras k componentes de la estrategia óptima $(d_1^*, \ldots, d_k^*, d_{k+1}^*, \ldots d_n^*)$ están dadas por las fórmulas (40) similares a las fórmulas (37). De hecho las primeras se transforman en las últimas cuando $k = n$. ¿Es única la estrategia óptima puesta en evidencia por el Teorema 5.1?

e) W. Woodside [7], investigó la dependencia de los parámetros n y m de la razón entre la duración mínima total $\sigma^* = \sum_{i=1}^{n} d_i^*$ dado por la fórmula (38) y la uniforme total $\sigma = n\delta$ dada por (10), para los problemas P_2 y P_1:

$$\begin{aligned} W(n,m) &= \frac{\sigma^*}{\sigma} \\ &= \frac{1}{n}\left[\frac{S_n(m)}{r_n}\right]^2 \frac{r_n}{S_n(m^2)} \\ &= \frac{1}{n}\,\frac{(1+m)}{(1-m)}\,\frac{(1-m^n)}{(1+m^n)} \quad \text{con } m < 1 \end{aligned}$$

M. S. Klamkin ([4]), demostró en forma elemental que la función $n \mapsto W(n,m)$ es decreciente respecto a $n \in \mathbb{N}$ para cada $m \in [0,1[$ y que la función $m \mapsto W(n,m)$ es creciente respecto a m para cada $n \in \mathbb{N}$. De aquí se desprenden las conclusiones que siguen:

- Cuanto más grande es el número de unidades de tiempo destinadas a la preparación, así la duración mínima total se aparta más de la uniforme $\sigma = n\delta$ y cuanto más grande es el coeficiente de memorización, así la duración mínima se acerca más a la uniforme.
- Un estudio análogo de la monotonía del cociente $\frac{\sigma^*}{\delta}$ sería deseable en el caso de los problemas P_2' y P_1'.

f) W. Woodside [7] y M.S. Klamkin [4] consideraron modelos de aprendizaje en los cuales el término $e\sqrt{d_{i+1}}$ de la fórmula (1) se sustituye con el término de forma más general ed_{i+1}^r donde $0 < r < 1$. En este caso la Proposición 3.1 se aplica aun con $p = \frac{1}{r}$.

Los resultados en este trabajo se van a extender en [5] para los casos mencionados y para modelos "continuos" de aprendizaje.

Bibliografía

[1] Bolstianskii, V.G. (1973) *Control Óptimo de Sistemas Discretos* (en lengua rusa) Izdat. Nauka, Moscú. (Hay traducción al Francés).

[2] Bondi, H. (1982) *Note on "A Student Related Optimal Control Problem" by Raggett,* Hempson and Jukes. Bull. Inst. Math. Appl. 18, 10-11.

[3] Gussi, G. (1988) *Stănăsibo si Stica, T. Elemente de Analiză Matematică. Manual Pentru clasa a XI-1.* Editura didactică si pedagogică, Bucureşti.

[4] Klamkin, M.S.(1985) *Mathematical Modelling: A student Optimal Control Problem and extensions.* Math Modelling 6, 49-64.

[5] Muntean, I. Şi Vornicescu, N. *Extensiones del Modelo de Aprendizaje en Tiempo Mínimo.* (en preparación).

[6] Raggett, G.F; Hempson, P.M. and Jukes, K.A (1981) *A Student Related Optimal Control Problem.* Bull. Inst. Math. Appl. 17, 133-136.

[7] Woodside, W (1982) *A student Optimal Control Problem or How to Pass Courses with the Minimun Expenditure of Effort.* Appl. Math. Notes, Canad. Math. Soc. 7, 2-13.