Revista de Matemática: Teoría y Aplicaciones 1(1): 87–100 (1994)

Generación de reglas estadísticas a partir de grandes bases de datos

Yves Schektman[1] – Javier Trejos Zelaya[2] – Marylène Troupé[3]

Abstract

Dado un conjunto de variables cualitativas, queremos predecir una o varias de ellas mediante reglas. Proponemos un algoritmo que (*i*) es guiado por resultados estadísticos en el marco de una geometría relacional, dentro de la cual se utilizan índices de asociación disimétricos, y (*ii*) efectúa aproximaciones estadísticas y euclidianas. El método iterativo propuesto puede obtener muchas reglas sin tener que introducir *a priori* sus premisas en el conjunto de conjunciones explicativas que el generador analiza en cada etapa. El algoritmo es de complejidad lineal respecto al número de individuos, por lo que sería particularmente bien adaptado a las grandes bases de datos. Se presentan resultados sobre ejemplos de datos.

Palabras clave: reglas de producción, asociación disimétrica, adquisición de conocimientos, distancia relacional, número equivalente.

Given a set of categorical variables, we want to predict one or more of them by the way of rules. We propose an algorithm that (*i*) is guided by statistical results in a relationnal geometry where we use assymetrical association indices, and (*ii*) makes statistical and euclidian approximations. The iterative method we propose can obtain rules without introducing a priori their premises in the set of independent conjonctions analized by the generator at each step. The algorithm has a linear complexity with regard to the number of individuals; this property makes it suitable for large data sets. We present results over data examples.

1. Problemática

Se han observado variables cualitativas sobre un gran número de individuos y se quiere predecir una de ellas mediante reglas del tipo: $C^j \to y^k \quad (P[C^j], P[y^k], P[y^k/C^j])$, donde C^j es una conjunción de modalidades de las variables explicativas, y^k una modalidad de la variable a explicar, $P[C^j]$ (resp. $P[y^k]$) el porcentaje observado de C^j (resp. y^k) y $P[y^k/C^j]$ el porcentaje observado de y^k sabiendo C^j.

[1]Universidad de Toulouse Le Mirail, Francia
[2]Escuela de Matemática, Universidad de Costa Rica
[3]Universidad de las Antillas-Guyana, Francia

	x^1	x^2	x^3	x^4
y^1	20	0	10	10
y^2	0	20	10	10
y^3	10	10	10	10
y^4	10	10	10	10

Cuadro 1: *Tabla de contingencia para x y y de modalidades $\{x^j\}$ y $\{y^k\}$*

Denotamos $\mathcal{C}$ el conjunto de conjunciones de modalidades de variables explicativas. Como el cardinal de $\mathcal{C}$, y por tanto el número de premisas posibles, crece rápidamente en función del número de variables, no es razonable tratar de extraer todas las reglas haciendo una exploración exhaustiva de $\mathcal{C}$: por ejemplo, en presencia de sólo 30 variables explicativas teniendo cada una 4 modalidades, tendríamos que $card(\mathcal{C}) = \sum_{j=1}^{30} \mathsf{C}_{30}^{j} 4^j > 10^{20}$, lo cual nos llevaría a considerar algoritmos no eficaces.

Durante la exploración de $\mathcal{C}$, el generador propuesto es guiado por resultados estadísticos, dentro del marco de una geometría relacional [14] enriquecida por la utilización de índices de asociación disimétricos [**?**, **?**]. Entonces el árbol de $\mathcal{C}$ podrá ser "podado"mediante aproximaciones estadísticas y euclidianas. Este punto de vista parece razonable dado que el algoritmo presenta una complejidad lineal respecto al número de individuos. Este resultado, presentado en la sección 3.3, proviene del hecho que los índices estadísticos usados, para ser calculados, necesitan un sólo paso sobre los individuos. Por lo tanto, esta visión [**?**, **?**, **?**, **?**] se distingue de los métodos basados en el aprendizaje [**?**, **?**, **?**], que necesitan pasar varias veces sobre el conjunto de individuos ("patterns.º ejemplos) y pueden llevar a considerar algoritmos de complejidad al menos cuadrática (*cf.* por ejemplo [**?**]).

2. Generalidades

2.1. Algunas observaciones fundamentales

Se sabe que las medidas de asociación entre variables cualitativas dan a menudo resúmenes demasiado sintéticos. Por ejemplo, sobre los datos presentados en la tabla **??**, se han medido el T^2 de Chuprov y el tau de Goodman-Kruskal (τ), que miden respectivamente la asociación simétrica y disimétrica entre las variables x et y, y dan por valor de la asociación $T^2 = 0{,}028$ y $\tau = 0{,}083$. Sin embargo se puede ver que hay una fuerte asociación entre algunas modalidades: por ejemplo, $P[\neg y^1/x^2] = P[\neg y^2/x^1] = 1$, donde $\neg y^1$ (resp. $\neg y^2$) es la negación de la modalidad y^1 (resp. y^2). Por lo tanto, el generador propuesto trabajará a nivel de las indicatrices asociadas a las modalidades de las variables. Además –cuando esto tenga sentido para la aplicación que se estudie–, se introducirán las indicatrices de las negaciones de las modalidades a explicar.

El ejemplo presentado en la tabla **??** muestra que también puede ser útil el introducir las negaciones de las modalidades explicativas. Se puede apreciar que
$P[y/x] \approx P[\neg y/x] \approx 0{,}5 = P[y] = P[\neg y]$, mientras que $P[y/\neg x] = 0$ y $P[\neg y/\neg x] = 1$.

Para medir las asociaciones entre las modalidades explicativas (o conjunciones de ellas) y las variables a explicar, es preferible utilizar índices de asociación disimétricos. Por ejemplo, para las variables de la tabla **??**, se obtiene que $\tau(x, y) = 1$ mientras que $T^2(x, y) = 0{,}091$.

	x	$\neg x$
y	60	0
$\neg y$	59	1

Cuadro 2: *Tabla de contingencia que muestra la utilidad de introducir las negaciones de las modalidades de las variables explicativas*

	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8	x^9	x^{10}	x^{11}	x^{12}
y^1	1	1	1	1	0	0	0	0	0	0	0	0
y^2	0	0	0	0	1	1	1	1	0	0	0	0
y^3	0	0	0	0	0	0	0	0	1	1	1	1

Cuadro 3: *Tabla de contingencia para medir la asociación simétrica y disimétrica entre x y y*

Esta observación es importante en nuestro contexto por el papel diferente que juegan las variables explicativas y a explicar.

Cuando el generador construye *a priori* conjunciones de modalidades explicativas (*cf.* sección 3.2), esta operación se lleva a cabo sólo cuando las medidas de asociación ya no nos pueden guiar. Esto puede ilustrarse con el ejemplo compuesto de cuatro individuos presentados en la tabla **??**, que muestra que la construcción *a priori* de conjunciones de modalidades explicativas es un problema abierto: se tiene que

$$T^2(C^i, C^j) = 0 \text{ y } \tau(C^i, y^k) = \tau(C^j, y^k) = 0,$$

$$\begin{aligned}
\text{mientras que } \tau(C^i \& C^j, y^k) &= \tau(\neg C^i \& \neg C^j, y^k) = \tau(\neg C^i \& C^j, \neg y^k) \\
&= \tau(C^i \& \neg C^j, \neg y^k) = 1/3 \\
\text{y } P[y^k = 1 | (C^i = 1) \& (C^j = 1)] &= P[y^k = 1 | (C^i = 0) \& (C^j = 0)] \\
&= P[y^k = 1 | (C^i = 0) \& (C^j = 1)] \\
&= P[y^k = 1 | (C^i = 1) \& (C^j = 0)] = 1,
\end{aligned}$$

donde "&"simboliza el operador de intersección entre dos conjunciones de modalidades explicativas.

En lo que sigue, el término modalidad explicativa significará modalidad de una variable explicativa o negación de una modalidad explicativa si se ha juzgado útil el introducir las negaciones; y el término conjunción explicativa será la intersección de modalidades explicativas y/o de negaciones de modalidades explicativas.

Individuos	C^i	C^j	y^k
1	1	1	1
2	1	0	0
3	0	1	0
4	0	0	1

Cuadro 4: *Construcción de conjunciones explicativas*

2.2. Herramientas

Utilizamos el Número equivalente (*Neq*) de G. Der Megreditchian [**?**] para limitar la longitud de las conjunciones explicativas creadas *a priori* (*cf.* sección 3.2). La definición original del *Neq* es probabilista, pero para una tabla de datos puede expresarse bajo la forma:

$$Neq = \frac{(traza\ VM)^2}{traza(VM)^2} = \frac{\left(\sum_j \lambda_j\right)^2}{\sum_j \lambda_j^2}$$

donde V es la matriz de covarianzas de las variables de la tabla de datos, M es la matriz de la distancia en el espacio de los individuos y λ_j es el j-ésimo valor propio de VM. *Neq* es una medida de la cantidad de información no redundante, relativamente a M, aportada por un conjunto de variables: a manera de ilustración, se puede ver fácilmente que si VM tiene r valores propios iguales no nulos entonces $Neq = 1 + 2C_r^2/r = r$, o incluso que si VM sólo tiene 2 valores propios no nulos y diferentes entonces $Neq \in]1, 2[$. Si $M = diag(1/\sigma_j^2)$, donde σ_j^2 es la varianza de la j-ésima variable x^j, entonces el cálculo de *Neq* se puede simplificar a [**?**, **?**]:

$$Neq = p^2 / \sum_{j,k=1}^{p} \rho^2(x^j, x^k)$$

donde p es el número de variables y ρ es el coeficiente de correlación lineal de Bravais-Pearson.

También utilizamos las distancias llamadas relacionales [12,14]. En el espacio euclidiano de los individuos $E = \oplus_r E_r$, donde E_r está asociado por dualidad a un subconjunto G_r de variables, estas distancias están caracterizadas geométricamente por $\cos[u_j(r), u_k(s)] = \cos[U^j(r), U^k(s)]$, donde los $u_j(r) \in E_r$ son los vectores axiales principales de la nube $\mathcal{N}_r$, proyección cartesiana de la nube de individuos sobre E_r y los $U^j(r) \in F_r$ son las componentes principales de $\mathcal{N}_r$, donde F_r es el subespacio engendrado por las variables de G_r.
Algunas expresiones algebraicas y propiedades de estas distancias son dadas en [**?**, **?**, **?**, **?**]. Si M_r y M_s son los bloques diagonales de la matriz de la distancia relacional asociados a G_r y G_s, una expresión del bloque extradiagonal M_{rs} es la siguiente:

$$M_{rs} = M_r[(V_{rr}M_r)^{1/2}]^{\dagger} V_{rs} M_s[(V_{ss}M_s)^{1/2}]^{\dagger},$$

donde V_{rs} es la matriz de covarianzas entre las variables de los grupos G_r y G_s y $\dagger$ simboliza la inversa generalizada ponderada relativamente a M_r o a M_s. En particular, en el generador de reglas propuesto, introducir los puntos-conjunciones explicativas C^i y C^j en un espacio euclidiano relacional puede reducirse a posicionar estos puntos de tal manera que $\cos(C^i, C^j) = \rho(C^i, C^j)$.

Se debe notar que, escogiendo convenientemente las distancias relacionales [15], la inercia de la proyección ortogonal de $\mathcal{N}_r$ sobre E_s es igual al numerador de los índices de asociación simétrica clásicos (Bravais-Pearson, phi cuadrado, suma de los cuadrados de las correlaciones canónicas,...) y disimétrica (Goodman-Kruskal, Stewart-Love) entre las variables de los grupos G_r y G_s.

3. El generador de reglas

El algoritmo es iterativo y a cada iteración produce reglas limitando su análisis a un Subconjunto de Conjunciones explicativas (S_c) de $\mathcal{C}$. Las conjunciones explicativas y modalidades muy poco o demasiado representadas podrán ser eliminadas.

El algoritmo parte de la situación más sencilla: así, durante la primera iteración, S_c está formado sólo por las modalidades de las variables explicativas; por ello, únicamente las asociaciones de segundo orden son analizadas. Asociaciones de órdenes superiores son progresivamente introducidas en el análisis, enriqueciendo S_c con conjunciones explicativas cuya construcción se precisa en la sección 3.2.

3.1. Producción de reglas

En cada iteración, un algoritmo de acumulación no exclusiva construye, a partir de las asociaciones medidas entre todas las conjunciones explicativas de S_c, Clases no disjuntas de Conjunciones explicativas Bien Asociadas (CCBA), es decir, tales que los valores de las asociaciones entre todas las parejas de conjunciones explicativas de una misma clase sean superiores a un umbral fijado (aquí por el usuario). Cada CCBA es tratada separadamente en el proceso de producción de reglas. Esta escogencia es fundamental en nuestro tratamiento estadístico porque permite poner en evidencia premisas de reglas de longitud superior a la longitud máxima de las conjunciones de la CCBA. Esto ocurre cuando varias conjunciones explicativas de una misma CCBA están bien asociadas a una misma modalidad a explicar: así, si dos conjunciones explicativas C^i y C^j pertenecen a una misma CCBA y están bien asociadas a una misma modalidad a explicar y^k, entonces a menudo también lo estará la conjunción explicativa $C^i\&C^j$ que de esta manera se convertiría en una premisa de regla de longitud superior a la de C^i y la de C^j. El tratamiento estadístico propuesto será entonces eficaz si se observan grandes asociaciones entre las conjunciones explicativas, por una parte, y entre conjunciones explicativas y modalidades a explicar por otra parte. En este caso, es razonable pensar que el número de premisas obtenidas sin haber sido introducidas en el S_c tratado, será grande. Por lo tanto, el algoritmo efectuará numerosos descensos selectivos en el árbol de $\mathcal{C}$.

Para generar las reglas a partir de una CCBA, se construye, en el marco de una geometría relacional, la nube reunión de los puntos-conjunciones explicativas (C^j) de la CCBA y de los puntos-modalidades a explicar (y^k):

a) Se sumergen las modalidades a explicar en un espacio euclidiano de dimensión igual a *card*(CCBA) y provisto de una distancia relacional de matriz M_c definida por:

$$[M_c]_{ij} = \rho(C^i, C^j),\ \forall (i,j) \in \{1, \ldots, card(CCBA)\}^2.$$

Debe notarse que los $[M_c]_{ii}$ podrían tomarse diferentes de uno y depender de los efectivos de los C^j. Si n es el número total de individuos y n_j, n_k, n_{jk} son respectivamente el número de individuos que tienen las categorías C^j, y^k y $C^j\&y^k$, se posicionan los puntos-modalidades a explicar y^k de tal manera que la j-ésima coordenada de y^k, en

la base canónica, sea igual a:

$$y^k(C^j) = \begin{cases} \left(\frac{n_{jk}}{n_j} - \frac{n_k}{n}\right) / \left(1 - \frac{n_k}{n}\right) & \textit{si este valor es positivo} \\ 0 & \textit{si no.} \end{cases}$$

El denominador de esta cantidad es un factor de normalización tal que entre mayor sea el efectivo de la modalidad a explicar, mayores serán sus coordenadas.

b) Se sumergen las conjunciones explicativas C^j en el subespacio engendrado por los y^k de tal manera que para toda pareja (j,k), la k-ésima coordenada $C^j(y^k)$ de C^j, relativamente a la base de los y^k, se escribe:

$$C^j(y^k) = \begin{cases} \dfrac{n_{jk}}{n_j} - \dfrac{n_k}{n} & \textit{si este valor es positivo} \\ 0 & \textit{si no.} \end{cases}$$

En este subespacio la matriz del producto escalar M_y en la base de los y^k es tal que:

$$[M_y]_{kl} = \frac{\cos(y^k, y^l)}{\|y^k\|_{M_c}\ \|y^l\|_M}.$$

Procediendo así, en cada espacio asociado a un CCBA, las proximidades entre los C^j y los y^k son función de las asociaciones simétricas entre los C^j y de las asociaciones disimétricas positivas entre los C^j y los y^k.

Para proceder a la búsqueda de las reglas a partir de la nube de puntos-conjunciones explicativas y puntos-modalidades a explicar asociados a un CCBA, aplicamos las operaciones siguientes:

o_1) Se hace un Análisis en Componentes Principales no centrado del conjunto de los y^k, con pesos iguales. Este análisis permite disminuir la dimensión del espacio asociado a cada CCBA, por proyección de la nube de puntos C^j y y^k sobre un subespacio principal y se ponen en evidencia los ejes principales cuya importancia mide la calidad de la predicción.

o_2) En el subespacio principal se construyen sectores no disjuntos cuyos lados son paralelos a los planos principales. A partir de la condición $P[y^k/C^j]/P[y^k] = \frac{n_{jk}}{n_j}/\frac{n_k}{n} \geq \alpha$ equivalente a $C^j(y^k) \geq \frac{n_k}{n}(\alpha - 1)$ (*cf.* o_4) y de ciertas aproximaciones, se calculan los límites de los sectores.

o_3) En cada sector, las reglas posibles se ponen en evidencia por proximidad, haciendo previamente una clasificación mediante un algoritmo de accumulación non exclusiva con líder, tomando como líderes las modalidades a explicar. Por ejemplo, en la figura 1, donde los rectángulos simbolizan los sectores, las reglas que se sugieren son:

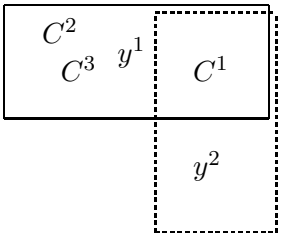

Figura 1: *Una configuración de conjunciones explicativas* (C^j) *y de modalidades a explicar* (y^k) *en dos sectores no disjuntos*

$$\begin{array}{lll} C^1 \to y^1 & C^1\&C^2 \to y^1 & C^1\&C^2\&C^3 \to y^1 \\ C^2 \to y^1 & C^1\&C^3 \to y^1 & \\ C^3 \to y^1 & C^2\&C^3 \to y^1 & \\ C^1 \to y^2 & & \end{array}$$

o_4) Entre las reglas sugeridas (para un sector), sólo aquéllas que satisfagan las condiciones fijadas por el usuario son generadas. Así, en el caso de la figura 2, la regla $C^1 \to y^1$ es generada si:

$$P[y^1/C^1]/P[y^1] \geq \alpha.$$

Además, si la regla $C^1 \to y^1$ ha sido generada, entonces la regla $C^1\&C^2 \to y^1$ es generada sólo si:

$$P[y^1/C^1\&C^2]/P[y^1/C^1] \geq \beta.$$

α y β son constantes (mayores que 1) fijadas por el usuario: α es proporcional a la calidad de la predicción deseada y β permite rechazar el crecimiento en la longitud de una premisa si la ganancia en la predicción observada es insuficiente. Veamos otra ilustración de la utilización de las constantes α y β: si las reglas $C^1 \to y^1$ y $C^2 \to y^1$ no han sido generadas, entonces la regla $C^1\&C^2 \to y^1$ es generada si $P[y^1/C^1\&C^2]/P[y^1] \geq \alpha$ y la regla $C^1\&C^2\&C^3 \to y^1$ es generada sólo si además $P[y^1/C^1\&C^2\&C^3]/P[y^1/C^1\&C^2] \geq \beta$. Para la presentación de las reglas, además de las probabilidades marginales y condicionales arriba usadas, también se muestra el valor del índice de previsión $prev = P[y^k/C^j]/P[y^k] = n_{jk}/(n_j n_k)$.

Se introducen en S_c las premisas de reglas generadas; luego las operaciones descritas arriba se repiten, haciendo ciclos, hasta que no se puedan generar más reglas. No es sino hasta este momento que se pasa a la etapa siguiente, descrita en la próxima sección, que es una etapa de preparación necesaria para la iteración siguiente.

3.2. Enriquecimiento del conjunto de conjunciones explicativas analizadas

Una vez que se ha sacado el máximo provecho de las asociaciones elevadas observadas entre las conjunciones de S_c, debemos hacer aparecer nuevas asociaciones introduciendo en

S_c nuevas conjunciones explicativas obtenidas por intersección de conjunciones explicativas existentes.

Para proceder al enriquecimiento de S_c, varias estrategias podrían considerarse sin que se pueda decir *a priori* cuál es la mejor (*cf.* el último ejemplo de la sección 2.1). Podemos, por ejemplo, ser guiados por el usuario quien podría sugerir intersecciones de modalidades a hacer o a evitar, proceder por aprendizaje sobre una muestra, ...

En este momento, hemos decidido construir familias de CCBA y cruzar las conjunciones que pertenecen a CCBA diferentes de una misma familia. La heurística implementada está basada en una clasificación por acumulación no exclusiva. La escogencia de este método de clasificación se hizo porque la primitiva para llevarla a cabo se encuentra en la biblioteca de programas usada para programar nuestro algortimo. Si denotamos N el número de CCBA, entonces el número de familias será la parte entera de $N/2$. La distancia entre CCBA se calcula a partir de la asociación simétrica entre los dos primeros elementos de cada CCBA.

La última operación consiste en cruzar las conjunciones explicativas que pertenecen a una misma familia pero provenientes de CCBA diferentes.

No se construirán conjunciones de longitud superior a $Neq{+}1$: en efecto, consideramos que las conjunciones más largas que $Neq{+}1$ contienen información redundante.

El conjunto S_c usado en la iteración siguiente contendrá por lo tanto las conjunciones del S_c anterior, al que se añaden las nuevas conjunciones construidas según el método aquí descrito y las conjunciones que han formado premisas de las reglas generadas durante la presente iteración.

El algoritmo se detiene cuando ya no se pueda enriquecer S_c. Sin embargo, el usuario puede decidir detenerlo imponiendo un número máximo de reglas a generar, de iteraciones a efectuar, o si el número de reglas generadas es mucho menor que el número de reglas sugeridas,...

3.3. Estudio de la complejidad del algoritmo

Como ya hemos indicado, el algoritmo es lineal respecto al número de individuos. La complejidad para realizar la primera iteración es de $\mathcal{O}(n, q^2, p^4)$, donde n es el número de individuos, q el número de modalidades a explicar y p el número de modalidades explicativas. A continuación damos en detalle la complejidad de cada operación implementada.

Para el cálculo de Neq utilizando la fórmula simplificada dada en la sección 2.2, se calculan primero las correlaciones entre las modalidades explicativas, lo cual tiene una complejidad de $\mathcal{O}(np^2)$. Si a medida que se calculan las correlaciones se introduce su cuadrado en una sumatoria, entonces el cálculo sigue siendo del orden $\mathcal{O}(np^2)$ pues sólo se hace la división de p^2 entre la suma de los cuadrados de las correlaciones. Debe notarse que para un producto escalar arbitrario esta complejidad sería del orden $\mathcal{O}(np^3)$.

En la primera iteración, para construir los CCBA mediante el algoritmo de acumulación no exclusiva, se consideran secuencialmente las modalidades explicativas que se van a clasificar. La primera modalidad x^1 forma ella sola una clase. La segunda entra dentro de esa clase si su correlación con x^1 es mayor al umbral fijado s, sino forma una nueva clase. En

la etapa j, hay a lo sumo $j-1$ clases ya construidas y la variable x^j entrará en aquéllas donde su correlación con todas las variables de la clase sea menor a s; si no existen tales clases, x^j forma una nueva clase. Según las condiciones de este algoritmo [**?**, **?**], en la etapa j se debe construir el conjunto $\{x^i/\rho(x^i, x^j) \geq s\}$ y podrían llegar a redefinirse a lo sumo $j-1$ clases. Por lo tanto la complejidad del algoritmo de acumulación no exclusiva es de $\mathcal{O}(p^2)$ en el peor de los casos.

A continuación se hace un tratamiento para cada CCBA (hay a lo sumo p de ellas con a lo sumo p elementos cada una). El cálculo de las coordenadas $y^k(C^j)$ en el espacio engendrado por las conjunciones explicativas C^j es del orden $\mathcal{O}(nqp)$, el cálculo de la distancia M_c (*cf.* a en 3.1) es del orden $\mathcal{O}(np^2)$ puesto que se trata de calcular las correlaciones en el CCBA. El cálculo del producto escalar M_y en el espacio engendrado por los y^k (*cf.* b en 3.1) es del orden $\mathcal{O}(q^2p^2)$ pues $M_y = Y^T M_c Y$ donde $[Y]_{jk} = y^k(C^j)$ y el cálculo de las coordenadas $C^j(y^k)$ es de $\mathcal{O}(qp)$. La proyección sobre el espacio principal es del orden $\mathcal{O}(q^2p)$ puesto que se trata de diagonalizar M_y que es una matriz $q \times q$ y proyectar los puntos de la CCBA. La construcción de los sectores (*cf.* o_2 en 3.1) tiene complejidad $\mathcal{O}(q^2p)$ pues se deben calcular los límites de cada sector y determinar las conjunciones explicativas que están en cada uno. Para detectar las reglas se recorren los sectores; en cada uno puede haber a lo sumo qp reglas y la verificación de las condiciones dadas por el usuario (*cf.* o_4 en 3.1) tiene un orden de $\mathcal{O}(nq^2p^2)$.

Por lo tanto, las operaciones para tratar cada CCBA son del orden $\mathcal{O}(nq^2p^2)$, y como hay a lo sumo p de ellos, la etapa de generación de reglas tiene una complejidad de $\mathcal{O}(nq^2p^3)$ en el peor de los casos.

La etapa de enriquecimiento de S_c, tal como está implementada actualmente, requiere de calcular las correlaciones entre las conjunciones (o modalidades) explicativas de los CCBA para calcular las distancias entre los CCBA, lo que tiene un orden de $\mathcal{O}(np^2)$, y la clasificación por acumulación exclusiva es del orden $\mathcal{O}(p)$. Finalmente, los productos cartesianos entre conjunciones explicativas de clases distintas es de $\mathcal{O}(np^2)$. Por lo tanto, esta etapa tiene una complejidad de $\mathcal{O}(np^2)$.

La primera iteración tiene entonces une complejidad de $\mathcal{O}(nq^2, p^3)$.

Para llevar a cabo las iteraciones sucesivas, el conjunto de conjunciones explicativas crece debido al enriquecimiento de S_c. Por ejemplo, en la segunda iteración, el número de conjunciones explicativas a considerar será a lo sumo de $p(p+1)/2$, esto es, las p modalidades simples y las $p(p-1)/2$ conjunciones explicativas de longitud 2 que se pueden construir. Por lo tanto, la complejidad sólo es modificada en el término correspondiente a p en una función de p, por lo que finalmente la complejidad del algoritmo propuesto en el peor de los casos es de $\mathcal{O}\left(nq^2[f(p)]^3\right)$, donde $f(p)$ es un polinomio en pp de grado 3 por lo menos.
Esta complejidad podrá reducirse puesto que en general no se harán p iteraciones, sino un número mucho menor cuando hay bastante asociación en los datos, ya que es en este contexto que el algoritmo es interesante.

3.4. Características del programa

El programa del algoritmo aquí presentado está constituido por 76 sub-programas que constituyen una biblioteca de 121.280 bytes y el tamaño de su módulo ejecutable sobre PC

compatible es de 454.477 bytes.

La programación se hizo bajo el sistema operativo MS-DOS versión 5.00 en Fortran 77 versión 4.01, con ayuda de una Biblioteca Matemática para el Análisis de Datos (BMAD) [**?**]. Actualmente disponible sobre PC compatible, la BMAD es portátil, necesita 443.904 bytes y está programada en Fortran. Bajo esta versión de MS-DOS, el tamaño delos segmentos es de 64 Kb, es decir 16 K palabras de 32 bits. Ahora bien, la zona de trabajo de una tabla de la BMAD está implantada en un `common` y el compilador Fortran atribuye a lo sumo 1 segmento a cada `common`,lo que limita las dimensiones de las tablas que podemos tratar. Estas restricciones deberían desaparecer con las nuevas versiones de MS-DOS y Fortran.

4. Ejemplos

4.1. Datos simulados

Se dispone de una tabla de datos simulados, con 20 individuos sobre los que se han medido 5 variables explicativas x^1, x^2, x^3, x^4, x^5 y una variable a explicar y. Todas estas variables son de presencia-ausencia. La tabla **??** muestra los datos: en ella, la primera columna indica cuántos individuos presentan cada patrón. Por ejemplo, la primera línea significa que 7 individuos presentan la misma combinación de variables $(x^1, x^2, x^3x^4, x^5, y) = (1, 1, 1, 0, 1, 1)$. Las conjunciones explicativas cuyos porcentajes observados son inferiores a 1 % o superiores a 98 % son eliminadas. El umbral de agrupamiento del algoritmo de acumulación no exclusiva es 0.35, y las reglas son aceptadas sólo si $P[C^j] \geq 5\,\%$, $P[y^k] \geq 5\,\%$ y $P[y^k|C^j] \geq 20\,\%$; además se tomó $\alpha = 1{,}3$, $\beta = 1{,}05$ (*cf.* o_4 en 3.1) y un máximo de 7 conjunciones explicativas por grupo dentro de una región rectangular.

Número de patrones	x^1	x^2	x^3	x^4	x^5	y
7	1	1	1	0	1	1
1	1	0	1	0	1	1
1	0	0	0	1	0	1
1	0	1	1	0	0	1
6	0	0	0	1	1	0
1	1	1	1	1	1	0
1	1	1	0	0	1	0
2	1	0	1	0	0	0

Cuadro 5: *Tabla de datos simulados*

El valor de *Neq* es 3, y desde el primer ciclo de la primera iteración, el generador produce 13 reglas cuyas premisas tienen longitud entre 1 y 3; en el segundo ciclo se producen las reglas 14 y 15; en la segunda iteración aparece las reglas 16 a 19. Luego de ésto el algoritmo se detiene pues ya no pudo construir nuevas conjunciones explicativas para enriquecer S_c. Las reglas, según el orden de salida, se presentan en la tabla **??**.

4.2. Datos zoológicos

Consideramos un ejemplo de datos zoológicos que constituyen un conjunto de 101 individuos (especies zoológicas) descritos por 16 variables explicativas y una a explicar. Todas

regla	$P[y^k\|C^j]$	*prev*
$1)x^1 \longrightarrow y$	66,7 %	$prev = 1{,}3$
$2)x^2 \longrightarrow y$	80 %	$prev = 1{,}6$
$3)x^3 \longrightarrow y$	75 %	$prev = 1{,}5$
$4)\neg x^4 \longrightarrow y$	75 %	$prev = 1{,}5$
$5)x^2\&x^3 \longrightarrow y$	88,9 %	$prev = 1{,}8$
$6)x^2\&x^3 \longrightarrow y$	88,9 %	$prev = 1{,}8$
$7)x^3\&\neg x^4 \longrightarrow y$	81,8 %	$prev = 1{,}6$
$8)x^2\&x^3\&\neg x^4 \longrightarrow y$	100 %	$prev = 2$
$9)\neg x^1 \longrightarrow \neg y$	75 %	$prev = 1{,}5$
$10)\neg x^2 \longrightarrow \neg y$	80 %	$prev = 1{,}6$
$11)\neg x^3 \longrightarrow \neg y$	87,5 %	$prev = 1{,}8$
$12)x^4 \longrightarrow \neg y$	87,5 %	$prev = 1{,}8$
$13)\neg x^1\&\neg x^2 \longrightarrow \neg y$	85,7 %	$prev = 1{,}7$
$14)\neg x^1\&\neg x^5 \longrightarrow y$	100 %	$prev = 2$
$15)\neg x^3\&\neg x^5 \longrightarrow y$	100 %	$prev = 2$
$16)x^3\&\neg x^4\&x^5 \longrightarrow y$	100 %	$prev = 2$
$17)x^1\&x^5 \longrightarrow y$	80 %	$prev = 1{,}6$
$18)x^3\&x^5 \longrightarrow y$	88,9 %	$prev = 1{,}8$
$19)\neg x^4\&x^5 \longrightarrow y$	88,9 %	$prev = 1{,}8$

Cuadro 6: Reglas producidas para la tabla de datos simulados

las variables explicativas salvo una son binarias, en total se tienen 21 modalidades explicativas. Los datos pueden consultarse en [**?**]. Las variables explicativas binarias consideradas son:

(hair)	hair :	pelos	(feat)	feathers :	plumas
(eggs)	eggs :	pone huevos	(milk)	milk :	da leche
(airb)	airborne :	vuela	(aqua)	aquatic :	acuático
(pred)	predator :	depredador	(toot)	toothed :	tiene dientes
(back)	backbone :	posee columna vertebral	(brea)	breathes :	respira aire
(veno)	venomous :	venenoso	(fins)	fins :	posee aletas
(tail)	tail :	posee cola	(dome)	domestic :	doméstico
(cats)	catsize :	tamaño de un león			

La variable (legs) : número de patas, tiene 6 modalidades, indicadas por (legs = n), donde $n = 0, 2, 4, 5, 6, 8$. La variable a explicar tiene 7 modalidades, según el tipo de animal: (mamm) mammalian : mamífero, (bird) bird : ave, (rept) reptile : reptil, (fish) fish : pez, (batr) batrachia : batracio, (inse) insect : insecto, (moll) mollusc,crustacea : molusco,crustáceo

Le programa corrió con un umbral para la acumulación no exclusiva (construcción de las CCBA) de 0.46. Eliminamos las modalidades con u afectivo menor al 2 %, es decir (legs=8) et (legs=5). Las negaciones de las modalidades explicativas se introdujeron. Por lo tanto, la tabla era de dimensiones 101×38 para las modalidades explicativas y de 101×7 para las modalidades a explicar. El número equivalente asociado a la tabla de modalidades explicativas es 7. Las probabilidades marginales mínimas $P[C^j]$ y $P[y^k]$ son 0.05 y la probabilidad

condicional 0.5. Se fijó $\alpha = 1{,}01$ y $\beta = 1{,}03$. En la tabla **??** se muestran las 32 reglas obtenidas en la primera iteración. En las columnas de la derecha se presentan las probabilidades marginales y las probabilidades condicionales en forma de porcentaje, así como el índice de previsón *prev*.

	regla	$P[C^j]$	$P[y^k]$	$P[y^k\|C^j]$	*prev*	$P[C^j\|y^k]$
1)	hair ⟶ mamm	42.6 %	40.6 %	90.7 %	2.2	95 %
2)	¬ aqua ⟶ mamm	64.4 %	40.6 %	53.8 %	1.3	85.4 %
3)	¬ feat ⟶ mamm	80.2 %	40.6 %	50.6 %	1.2	100.0 %
4)	¬ airb ⟶ mamm	76.2 %	40.6 %	50.6 %	1.2	95 %
5)	toot ⟶ mamm	60.4 %	40.6 %	65.6 %	1.6	97.6 %
6)	¬ feat & ¬ airb ⟶ mamm	72.3 %	40.6 %	53.4 %	1.3	95 %
7)	brea ⟶ mamm	79.2 %	40.6 %	51.3 %	1.3	100.0 %
8)	¬ aqua & ¬ (legs=0) ⟶ mamm	59.4 %	40.6 %	58.3 %	1.4	85.4 %
9)	¬ eggs ⟶ mamm	41.6 %	40.6 %	95.2 %	2.3	97.6 %
10)	milk ⟶ mamm	40.6 %	40.6 %	100.0 %	2.5	100.0 %
11)	(legs=4) ⟶ mamm	37.6 %	40.6 %	81.6 %	2.0	75.6 %
12)	hair & ¬ eggs ⟶ mamm	37.6 %	40.6 %	100.0 %	2.5	92.7 %
13)	hair & toot ⟶ mamm	37.6 %	40.6 %	100.0 %	2.5	92.7 %
14)	hair & (legs=4) ⟶ mamm	30.7 %	40.6 %	100.0 %	2.5	75.6 %
15)	¬ eggs & (legs=4) ⟶ mamm	29.7 %	40.6 %	100.0 %	2.5	73.2 %
16)	toot & (legs=4) ⟶ mamm	34.7 %	40.6 %	85.7 %	2.0	73.2 %
17)	dome ⟶ mamm	12.9 %	40.6 %	61.5 %	1.5	19.5 %
18)	cats ⟶ mamm	43.6 %	40.6 %	72.7 %	1.8	78.0 %
19)	¬ eggs & cats ⟶ mamm	30.7 %	40.6 %	100.0 %	2.5	75.6 %
20)	feat ⟶ bird	19.8 %	19.8 %	100.0 %	5.0	100.0 %
21)	¬ toot ⟶ bird	39.6 %	19.8 %	50.0 %	2.5	100.0 %
22)	(legs=2) ⟶ bird	26.7 %	19.8 %	74 %	3.7	100.0 %
23)	¬ hair & ¬ toot ⟶ bird	34.7 %	19.8 %	57 %	2.9	100.0 %
24)	¬ toot & ¬ (legs=4) ⟶ bird	36.6 %	19.8 %	54 %	2.7	100.0 %
25)	¬ hair & ¬ toot & ¬ (legs=4) ⟶ bird	32.7 %	19.8 %	60.6 %	3.0	100.0 %
26)	eggs & ¬ milk & ¬ toot ⟶ bird	37.6 %	19.8 %	52.6 %	2.7	100.0 %
27)	¬ hair & eggs & ¬ toot & ¬ (legs=4) ⟶ bird	31.7 %	19.8 %	62.5 %	3.2	100.0 %
28)	fins ⟶ fish	16.8 %	12.9 %	76.5 %	5.9	100.0 %
29)	(legs=0) ⟶ fish	22.8 %	12.9 %	56.5 %	4.4	100.0 %
30)	fins & (legs=0) ⟶ fish	15.8 %	12.9 %	81.3 %	6.3	100.0 %
31)	(legs=6) ⟶ inse	9.9 %	7.9 %	80.0 %	10.0	100.0 %
32)	¬ back ⟶ moll	17.8 %	9.9 %	55.6 %	5.6	100.0 %

Cuadro 7: *Reglas generadas para la tabla de datos zoológicos*

Notemos que muchas reglas generadas tienen premisas con más de un elemento. De hecho hay 11 de reglas de con remisa de longitud 2, 3 reglas con premisa de longitud 3 y una regla de 4. Todas estas conjunciones explicativas deberían de introducirse en S_c con tal de hacer un nuevo ciclo o una nueva iteración.

4.3. Datos médicos

Con el objetivo de hacer comparaciones del rendimiento de nuestro método respecto a otros métodos, consideramos un archivo de datos médicos que conciernen la hepatitis

[4]. El archivoi contiene las observaciones de 19 variables explicativas y una variable a explicar medidas sobre 155 pacientes. El número total de modalidades explicativas es 81. las modalidades a explicar son *die*(muerte) y *live* (sobrevivencia). En [**?**] se pueden encontrar comparaciones entre varios métodos con este archivo de datos.

Dadas las dimensiones de la tabla a analizar (155×81) y las limitaciones informáticas de nuestra maqueta de programa, agrupamos modalidades de ciertas variables y procesamos el archivo por subconjuntos de variables; estos agrupamientos fueron hechos a partir de las asociaciones observadas entre las modalidades explicativas. Por este procedimiento, obtuvimos un subconjunto del conjunto de reglas que hubiéramos obtenido en caso de haber podido procesar los datos en su totalidad. Por lo tanto, los resultados conmparativos que presentamos son solamente parciales. En efecto, desde el momento en que nuestras limitaciones informáticas sean superadas, haremos comparaciones más profundas.

Se llama *tasa de reconocimiento* de una base de reglas [**?**] el porcentaje de individuos que fueron bien clasificados por la base de reglas, es decir los que cuando se aplican las modalidades explicativas que los describen activan en la base mayoritariamente la modalidad a explicar que les está asociada.
La tasa de reconocimiento del subconjunto de regals obtenido es de 87.1 % (135 individuos sobre 155 son bien clasificados).

A manera de comparación, en [**?**] se dan las tasas de reconocimiento de otros métodos similares:

	base de test	base de aprendizaje
Bayes	84 %	
Assistant 86	83 %	
Multicapas optimal con 3 capas	86 %	98 %
Multicapas optimal con 2 capas	84 %	95 %
Multicapas optimal con 1 capa	84 %	94 %

Conclusión

El generador propuesto parece particularmente adaptado al tratamiento de bases de datos voluminosas, en virtud de su linealidad respecto al número de individuos. Mediante una exploración selectiva en profundidad del árbol del conjunto de conjunciones explicativas, presenta igualmente la ventaja de poder producir rápidamente reglas cuyas premisas pueden ser bastante largas. La exploración a lo ancho del árbol es parcial, pero puede ser localmente sistemática si el usuario lo pide. En efecto, el algoritmo no construye a priori todas las intersecciones de conjunciones explicativas de S_c.

Sin embargo, aún se pueden proponer algunas mejoras. Sería por ejemplo interesante experimentar las diferentes sugerencias enumeradas en la sección 3.2 sobre datos simulados o sobre datos propuestos por otros investigadores. La simplificación de las reglas y de sus premisas está estudiándose; se podrían utilizar índices de proximidad semántica entre las reglas generadas debido a que éstas pueden ser clasificadas según sus similitudes semánticas, lo que facilitaría la interpretación de los resultados. En efecto, es razonable pensar que entre más cercanos sean dos sectores las reglas generadas tendrán semánticas más parecidas. Para

[4]Archivo elaborado por par G. Gong, Carnagie-Mellon University y B. Cestnik, Jozef Stefan Institute, Ljubljana, Eslovenia

cada CCBA, se podría entonces construir un árbol jerárquico de sectores. Así mismo, podría asociarse a cada iteración del algoritmo una jerarquía de CCBA. Estas jerarquías pueden ser muy útiles para decidir acerca del orden de presentación de las reglas generadas.

El método propuesto es generalizable sin ninguna dificultad al caso de varias variables a explicar y se piensa adaptarlo al caso en que no se conozcan *a priori* las variables a explicar.

Finalmente, se piensa generalizar la metodología propuesta al caso en que se disponga de datos simbólicos mediante una adecuada definición de los índices de asociación y de los espacios de representación de los objetos.

Referencias

[1] Abdesselam, R.; Schektman, Y. (1989) *Dissymetrical association analysis between two qualitative variables*. En: Data Analysis, Learning Symbolic and Numeric Knowledge, E. Diday (ed.), Nova Science – INRIA, New York, 39-46.

[2] Cohen, P.R.; Feigenbaum, E.A. (1982) *The Handbook of Artificial Intelligence, Vol. III*. Pitman, London.

[3] Croquette, A. (1980) *Quelques résultats synthétiques en analyse des données multidimensionnelles. Optimalité et métriques à effets relationnels*. Tesis de 3er ciclo, Universidad Paul Sabatier, Toulouse.

[4] Der Megreditchian, G. (1979) *L'optimisation des réseaux d'observation des champs météorologiques*. La Météorologie, VI, n^o 17, Paris.

[5] Grau, D. (1983) *Mesure des effets relationnels*. Tesis de 3er ciclo, Universidad Paul Sabatier, Toulouse.

[6] Hammad, A; Jockin, J.; Sadeg, B.; Schektman, Y.; Vielle, D. (1987) *Bibliothèque Mathématique pour l'Analyse des Données (BMAD)*. En: Data Analysis and Informatics, E. Diday *et al.* (eds.), INRIA, North-Holland, Amsterdam, 5-13.

[7] Ibrahim, A.; Schektman, Y. (1984) *Analyse en partitions principales. Algorithme et exemples*. Journées de Classification, La Grande–Motte, Publ. CNET, 61-89.

[8] Kodratoff, Y; Diday, E. (1991) *Induction symbolique et numérique à partir de données*. Cépaduès-Editions, Toulouse.

[9] Labrèche, S.; Schektman, Y.; Trejos, J.; Troupé, M. (1992) *Les distances relationnelles: deux applications récentes*. Actes de Distancia'92; S. Joly, G. Le Calvé (eds.), Rennes, 369-372.

[10] Ralambondrainy, H. (1987) *GENREG un générateur de règles combinant techniques d'apprentissage et techniques d'Analyse des Données*. En: Actes des I Journées Symbolique–Numérique, Universidad Paris IX - Dauphine, 40-44.

[11] Schektman, Y. (1978) *Contribution à la mesure en facteurs dans les sciences expérimentales et à la mise en œuvre automatique dans les calculs statistiques*. Tesis de Estado, Universidad Paul Sabatier, Toulouse.

[12] Schektman, Y. (1987) *A general euclidean approach for measuring and describing associations between several sets of variables*. En: Recent Developments in Clustering and Data Analysis. Proc. of the 1^{st} French-Japanese Sem., Inst. Stat., Tokyo, 37-48.

[13] Schektman, Y. (1989) *Euclidean approach and statistical approximations for generating weighted knowledge rules from large sets of data*. En: Klassifikation und Ordnung, Gesellschaft für Klassifikation, Indeks-Verlag, Frankfurt, 328-330.

[14] Schektman, Y; Trejos, J.; Troupé, M. (1992) *Un générateur de règles floues à partir de bases de données volumineuses*. En: Actes des III Journées Symbolique–Numérique, Universidad Paris IX - Dauphine, 121-130.

[15] Schektman, Y; Trejos, J; Troupé, M. (1992) *Une approche relationnelle en prédiction par génération de règles en présence de bases de données volumineuses*. En: Actes des XXIV Journées de Statistique, ASU, Bruxelles, 441-443.

[16] Sebag, M. (1991) it Une approche symbolique-numérique pour la discrimination à partir d'exemples et de règles: l'apprentisage multi-couches. Tesis doctoral, Universidad Paris IX-Dauphine.

[17] Thanh Huyen, T.T.; Bao, H.T. (1991) *A method for generating rules from examples and its application*. En: Symbolic–Numeric Data Analysis and Learning; E. Diday, Y. Lechevallier (eds.), Nova Science – INRIA, New York, 493-504.

[18] Trejos, J.; Troupé, M. (1993) *Generating statistical rules for large volumes of data*. En: Proceedings First Panamerican Workshop in Applied and Computational Mathematics, Caracas, Venezuela, pp. 1.25-1.26.

[19] Trejos, J. (1994) *Contribution à l'acquisition automatique de connaisances à partir de données qualitatives*. Tesis doctoral, Universidad Paul Sabatier, Toulouse.

[20] Troupé, M. (1994) *Contribution à la protection de la régression multiple multidimensionnelle et à la génération de règles prévisionnelles*. Tesis doctoral, Universidad Paul Sabatier, Toulouse.