Revista de Matemática: Teoría y Aplicaciones 2(2): 75–86 (1995)

EL NÚMERO EQUIVALENTE COMO MEDIDA DE LA INFORMACIÓN EN ANÁLISIS DE DATOS

Javier Trejos Zelaya[1]

Resumen

Se recuerda la definición del número equivalente, propuesto para medir el número de unidades independientes de información y lo adaptamos al contexto del análisis multivariado de datos. Proponemos algunas propiedades en el contexto euclídeo usual y estudiamos algunas aplicaciones: en la determinación del número de factores en un análisis factorial, del número de clases en clasificación automática y del número de componentes de una conjunción de modalidades.

Palabras–clave: Análisis multivariado de datos, unidades de información independientes, análisis en componentes principales, clasificación automática, generación de reglas.

1 Introducción

El número equivalente fue estudiado por G. Der Mégréditchian [2, 3, 4] en un contexto probabilístico para calcular el número de estaciones independientes en la previsión meteorológica. Para una tabla de datos X definida por p variables cuantitativas $x^1, \ldots, x^p$, si se introduce una métrica M en el espacio de los individuos $E = \mathbb{R}^p$, podemos adaptar la definición del número equivalente (*Neq*), al contexto euclídeo, de la manera siguiente:

Definición 1 *El* **número equivalente** *asociado a la matriz X, respecto a la métrica M, es: $Neq(X, M) = \dfrac{(traza VM)^2}{traza(VM)^2}$, donde V es la matriz de varianzas–covarianzas de las p variables x^j.*

Se deduce claramente que $Neq(X, M) = \frac{(\sum_{j=1}^{p} \lambda_j)^2}{\sum_{j=1}^{p} \lambda_j^2}$ donde λ_j es el j-ésimo valor propio no nulo de VM.

La siguiente propiedad, debida a M. Troupé [16], precisa el sentido que damos en este contexto al *Neq* como medida de la cantidad de información no redundante aportada por un conjunto de variables cuantitativas (respecto a M).

[1]PIMAD, Escuela de Matemática, Universidad de Costa Rica

Proposición 1

a) $Neq(X, M) \geq 1$

b) $Neq(X, M) = 1$ *si y sólo si hay solamente un valor propio no nulo de* VM.

c) *Si* VM *tiene al menos dos valores propios no nulos distintos, entonces* $Neq(X, M) < rang\, X$.

d) $Neq(X, M) = rang\, X$ *si y sólo si todos los valores propios no nulos de* VM *son iguales.*

La demostración de esta propiedad está basada en la observación que:

$$Neq(X, M) = 1 + \frac{\sum_{j=1}^{rang\ X} \sum_{\substack{k=1 \\ k \neq j}}^{rang\ X} \lambda_j \lambda_k}{\sum_{j=1}^{rang\ X} \lambda_j^2}$$

y que $Neq(X, M) \leq rang\, X \Leftrightarrow rang\, X \sum_{j=1}^{p} \lambda_j^2 - (\sum_{j=1}^{p} \lambda_j)^2 \geq 0$.

Se puede ilustrar el carácter de medida de la información no redundante que posee el *Neq*, considerando el caso en que se está en presencia de dos variables ($p = 2$). Si λ_1 y λ_2, los dos valores propios de VM, son iguales, entonces el número equivalente calculado a partir de los λ es: $Neq_\lambda(X, M) = (2\lambda_1)^2/2\lambda_1^2 = 2$, lo cual indica claramente el hecho que hay dos medidas de la información independientes. Si se hace una variación sobre estos valores propios, manteniendo constante el valor de la inercia de la nube de puntos (dada por $traza(VM) = \lambda_1 + \lambda_2$): sea $\alpha > 0$ y sean $\gamma_1 = \lambda_1 + \alpha$, $\gamma_2 = \lambda_2 - \alpha$ los nuevos valores propios de VM, entonces el número equivalente calculado a partir de los γ sería: $Neq_\gamma(X, M) = Neq_\lambda(X, M)\frac{\lambda_1^2}{\lambda_1^2+\alpha^2}$, por lo que $Neq_\gamma(X, M) < Neq_\lambda(X, M)$. Es decir, entre mayor sea la separación entre los valores propios, menor será el valor de *Neq* puesto que el mayor de los valores propios tiene mayor parte de la inercia explicada por la componente principal asociada.

2 Caso de la métrica diagonal de las inversas de las varianzas

A continuación estudiamos el comportamiento del número equivalente en el caso en que $M = D_{1/\sigma^2}$, la diagonal de las inversas de las varianzas [13]. Recuérdese que este es el caso usual en Análisis en Componentes Principales cuando las variables están centradas y estandarizadas. Tenemos $M = D_{1/\sigma^2} = diag(1/\text{var } x^j)$, donde var x^j es la varianza de la variable x^j. Supondremos en lo que sigue que las variables están centradas.

Proposición 2 *Si $M = D_{1/\sigma^2}$ entonces:*

$$Neq(X, D_{1/\sigma^2}) = \frac{p^2}{\sum_{j=1}^{p}\sum_{k=1}^{p}\rho^2(x^j, x^k)} \tag{1}$$

donde ρ es el coeficiente de correlación lineal.

DEMOSTRACIÓN: Como $M = D_{1/\sigma^2}$ y V es la matriz de varianzas–covarianzas de las variables $x^1, \ldots, x^p$ se tiene $[VM]_{jk} = \frac{\mathrm{cov}(x^j,x^k)}{\mathrm{var}\ x^k}$, donde $\mathrm{cov}(x^j, x^k)$ es la covarianza entre x^j y x^k. Luego, se tiene $traza(VM) = p$. Además, el j-ésimo elemento de la diagonal de $(VM)^2$ es: $\sum_{k=1}^{p} \frac{\mathrm{cov}(x^j,x^k)}{\mathrm{var}\ x^k}\frac{\mathrm{cov}(x^k,x^j)}{\mathrm{var}\ x^j} = \sum_{k=1}^{p}\rho^2(x^j,x^k)$, por lo que $traza(VM)^2 = \sum_{j=1}^{p}\sum_{k=1}^{p}\rho^2(x^j,x^k)$, y finalmente $Neq(X, D_{1/\sigma^2}) = \frac{p^2}{\sum_{j=1}^{p}\sum_{k=1}^{p}\rho^2(x^j,x^k)}$. ■

Para $M = D_{1/\sigma^2}$, la igualdad 1 permite reducir la complejidad del cálculo del *Neq*: en efecto, como el cálculo de cada correlación es en $\mathcal{O}(n)$, la suma de los cuadrados de las p^2 correlaciones –y por consiguiente el cálculo del *Neq*– es de complejidad $\mathcal{O}(np^2)$, mientras que con la definición general 1, la complejidad del cálculo del *Neq* es en al menos $\mathcal{O}(np^3)$.

Observación: De la demostración anterior se deduce inmediatamente que si I_p es la matriz identidad de orden p, entonces $Neq(X, I_p) = (\sum_{j=1}^{p}\mathrm{var}\ x^j)^2/\sum_{j=1}^{p}\sum_{k=1}^{p}\mathrm{cov}^2(x^j,x^k)$. En este caso, la complejidad del cálculo del *Neq* es también en $\mathcal{O}(np^2)$. □

La siguiente propiedad establece claramente el sentido de "número de unidades de información independientes" que posee el número equivalente en este contexto.

Corolario 3 *Sea $M = D_{1/\sigma^2}$. Si se tienen m clases de variables $K_1, \ldots, K_m$ de mismo cardinal s y tales que $\forall(x^j, x^{j'}) \in K_\ell \times K_{\ell'}$ $\rho^2(x^j, x^{j'}) = \delta_{\ell\ell'}$, entonces $Neq(X, M) = m$.*

DEMOSTRACIÓN: Se tiene $\sum_{j=1}^{p}\sum_{k=1}^{p}\rho^2(x^j,x^k) = \sum_{\ell=1}^{m} card^2(K_\ell) = ms^2$ y $p^2 = (\sum_{\ell=1}^{m} card(K_\ell))^2 = (ms)^2$. Por lo tanto $Neq(M) = \frac{m^2s^2}{ms^2} = m$. ■

En presencia de grupos de variables con correlaciones intra elevadas y correlaciones inter bajas, el *Neq* tendrá un valor vecino al número de grupos: este resultado es una ilustración suplementaria del poder de medida de redundancia de la información que hemos mencionado que posee el *Neq*.

3 Aplicaciones

En sus trabajos originales, G. Der Mégréditchian estudió la aplicación del *Neq* para determinar el número de estaciones de observación meteorológica necesarias para tener toda la información pertinente, de manera tal que no se repita la información aportada por dos estaciones diferentes. Como hemos dicho, estos trabajos estaban enmarcados en un contexto probabilístico. Nosotros hemos encontrado, a partir de los desarrollos de la

sección anterior, algunas aplicaciones que pueden ser interesantes en el Análisis Multivariado de Datos según la Escuela Francesa, es decir, sin asumir distribuciones de probabilidad teóricas *a priori* en los datos.

3.1 Análisis en componentes principales: determinación del número de factores

El análisis en componentes principales (A.C.P.) trata de encontrar un conjunto de q variables sintéticas C^j a partir de una tabla de datos descrita por p variables cuantitativas $x^1, \ldots, x^p$, tales que las C^j sean no correlacionadas y con inercia máxima, en el sentido que la proyección de la nube de puntos–individuos en $\mathbb{R}^p$ sobre el espacio generado por las C^j tenga inercia máxima [1, 5, 6, 7, 8, 14]. En el caso usual, las variables están centradas y se estandarizan, por lo que la métrica en $\mathbb{R}^p$ es $M = D_{1/\sigma^2}$. Es sabido que la solución de este problema se obtiene a partir de la diagonalización de la matriz VM, producto de la matriz V de varianzas–covarianzas y la métrica M sobre $\mathbb{R}^p$, por lo que cuando $M = D_{1/\sigma^2}$ los valores y vectores propios de VM se obtienen a partir de la matriz de correlaciones.

Uno de los problemas ligados a la práctica del A.C.P. es el de la determinación del número q de componentes principales (es claro que $q < p$ para que tenga sentido hacer el análisis). Diversos autores [1, 6, 7, 8] han propuesto algunos criterios empíricos, tales como:

1. Tomar q tal que la inercia explicada por $C^1, \ldots, C^q$ sobrepase un umbral (porcentual, por ejemplo 70% u 80%) de la inercia total de la nube de puntos–individuos.

2. Tomar q tal que el diagrama de los valores propios de VM, ordenados en orden decreciente, muestre el punto donde el decrecimiento se aprecie como estable (este método es conocido como el método del "codo").

3. En el caso usual de la métrica $M = D_{1/\sigma^2}$, tomar q como el número de valores propios de VM mayores que 1; este criterio está basado en el hecho que, para variables estandarizadas, las variables originales tienen varianza 1, y como la varianza de una componente principal es el valor propio al que está asociada, entonces no tiene sentido tomar componentes principales con varianza menor que la de las variables originales.

4. Tomar tantas componentes principales como sean necesarias, en el sentido que una componente principal C^j es "interpretable" cuando hay por lo menos un individuo tal que el coseno cuadrado entre su vector en $\mathbb{R}^p$ y su proyección sobre C^j es mayor que 0.5, o bien cuando la correlación entre al menos una variable original y C^j es 0.7.

Ninguno de estos criterios es un criterio absoluto, antes bien se pregoniza la utilización conjunta de varios de ellos para decidir, lo mejor posible, la escogencia de q, y se llega incluso a afirmar que esta escogencia depende en mucho de la experiencia del analista. ¿Puede entonces darse una herramienta confiable que pueda servir al usuario, lego en la materia, para la determinación de q? Nosotros pensamos que el número equivalente

puede ayudar a responder a esta cuestión. En efecto, por tratarse de una medida de la información independiente contenida en una tabla de datos, es posible que ayude a decidir cuántos factores guardar de un A.C.P.

Con el fin de estudiar esta posibilidad, calculamos el *Neq* sobre varias tablas de datos, y comparamos el resultado con los criterios 1 y 3 mencionados arriba. Los resultados para varias tablas de datos se dan en la tabla 1. Los datos de las tablas correspondientes se pueden solicitar al autor.

Tabla de datos	n	p	*Neq*	r	Valores propios	Inercia
Notas escolares F	9	5	2.36	2	$\lambda_1 = 2.87$	56%
					$\lambda_2 = 1.13$	80%
					$\lambda_3 = 0.98$	99%
Notas escolares CR	10	5	2.24	2	$\lambda_1 = 2.89$	58%
					$\lambda_2 = 1.62$	90%
Peces de Amiard	23	16	3.43	3	$\lambda_1 = 7.52$	46%
					$\lambda_2 = 3.69$	70%
					$\lambda_3 = 1.52$	80%
					$\lambda_4 = 0.94$	86%
Sociomatriz de Thomas	24	24	7.42	7	$\lambda_1 = 5.25$	22%
					$\lambda_2 = 4.72$	42%
					$\lambda_3 = 3.92$	58%
					$\vdots$	$\vdots$
					$\lambda_8 = 0.84$	87%
Iris de Fisher	150	4	1.70	1	$\lambda_1 = 2.50$	62%
					$\lambda_2 = 0.91$	85%
Proteínas	25	9	3.80	3	$\lambda_1 = 4.00$	44%
					$\lambda_2 = 1.63$	63%
					$\lambda_3 = 1.12$	75%
					$\lambda_4 = 0.95$	85%
Pintores	24	4	2.52	1	$\lambda_1 = 2.27$	57%
					$\lambda_2 = 0.98$	81%

Tabla 1: Comparación entre el número equivalente (*Neq*) y el número r de valores propios mayores que 1, para varias tablas de datos de dimensiones n (número de individuos) por p (número de variables).

Puede verse en la tabla que el *Neq* tiende a ser superior al número de valores propios mayores que uno. Por lo tanto, es posible que el número equivalente tienda a sobreestimar el número de factores importantes de un A.C.P. Esta observación puede ser de utilidad para el usuario nuevo en el campo, que puede tener cierta aprehensión a dejar de lado información que puede ser útil para su estudio. Por ello, el número equivalente podría servirle como número de componentes principales suficientes para tomar en cuenta.

3.2 Clasificación por particiones: determinación del número de clases

En clasificación automática, los métodos de particionamiento tratan de obtener una partición de un conjunto de objetos sobre los que se han observado una serie de variables, de manera tal que los elementos de una misma clase sean lo más parecidos posible, y los elementos de clases distintas sean bastante diferentes [1, 5, 7, 14]. Usualmente, se aplican métodos que fijan *a priori* el número de clases, tales como los métodos de nubes dinámicas, de las k-medias, de transferencias, etc., al contrario de métodos como Isodata que estiman el número de clases pero con base en un gran número de parámetros difíciles de controlar para un usuario poco experimentado. Sería por lo tanto útil contar con un método que estime el número de clases antes de implementar la metodología de particionamiento.

Para abordar esta cuestión, hemos pensado en usar una adaptación del número equivalente que presentamos anteriormente. En efecto, los métodos de particionamiento buscan tipologías de los *individuos*, mientras que los métodos factoriales hacen tipologías de las *variables*. Las medidas del "parecido" entre individuos generalmente están basadas en criterios de *disimilitud* o *distancia*: entre menor sea el índice más parecidos son los objetos, mientras que las medidas del "parecido" entre variables están basadas en criterios de asociación estadística, tales como la correlación lineal: entre mayor sea el índice de asociación más parecido es el comportamiento de las variables.

Sea Ω un conjunto de n individuos, sobre los que se dispone de una medida de disimilitud $d : \Omega \times \Omega \to \mathbb{R}^+$ (d puede ser una distancia). Sea d^* el máximo valor que alcanza d sobre $\Omega \times \Omega$, entonces se define la similitud s:

$$s(i,j) = \frac{(d^*)^2 - d^2(i,j)}{(d^*)^2}$$

Obsérvese que así el valor máximo de $s(i,j)$ es 1, y corresponde al caso en que $i = j$. Se denota S la matriz de similitudes calculadas sobre los elementos de Ω.

Definición 2 *Dado un conjunto Ω con una medida de similitud $s : \Omega \times \Omega \to [0,1]$, se define el* **número equivalente** *$Neq(\Omega, S)$ por:*

$$Neq(\Omega, S) = \frac{(traza S)^2}{traza(S^2)} = \frac{n^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} s^2(i,j)}$$

Adaptando la proposición 3 a la definición anterior, se tiene el resultado enunciado en la proposición 4.

Proposición 4 *Si existe una partición $C_1, \ldots, C_k$ de Ω en k clases de mismo cardinal π, tales que $\forall (i,j) \in C_\ell \times C_{\ell'} \quad s(i,j) = \delta_{\ell\ell'}$, entonces $Neq(\Omega, s) = k$.*

El resultado anterior sugiere que, si se tienen k clases bastante homogéneas y de cardinal similar, el número equivalente puede dar una aproximación de ese número de clases. En caso que las clases no tengan mismo cardinal, entonces $Neq(\Omega, s) = (\sum_{\ell=1}^{k} \pi_\ell)^2 / \sum_{\ell=1}^{k} \pi_\ell^2$, donde π_ℓ es el cardinal de la clase C_ℓ.

Hemos medido el número equivalente definido sobre una matriz de similitudes, para algunas de las tablas de datos estudiadas anteriormente. Estos resultados se dan en la tabla 2. Para algunas de las tablas mostradas, el número equivalente da una idea del número de clases que podrían tomarse en una clasificación. Recuérdese que los árboles de clasificación jerárquica construidos ascendentemente, normalmente dan buenas agrupaciones en las partes bajas del árbol pero la calidad de la clasificación disminuye conforme se asceinde en la construcción. Contrariamente, los árboles construidos descendentemente dan una mejor calidad en las partes superiores la calidad dismunuye en las partes inferiores. Estas comparaciones deben ser ampliadas, con diversos métodos de clasificación, así como con diversos criterios de estimación del número de clases. En [9] se propone un índice para "cortar" un árbol de clasificación jerárquica, basado en conjuntos difusos. Además se hacen comparaciones entre 8 ídices para estimar el número de clases, entre ellos el que aquí proponemos basado en el número equivalente. Una próxima publicación dará cuenta de estas comparaciones.

Tabla de datos	n	p	*Neq*
Notas escolares F	9	5	3.00
Notas escolares CR	10	5	3.21
Peces de Amiard	23	16	2.24
Sociomatriz de Thomas	24	24	4.68
Iris de Fisher	150	4	2.14
Proteínas	25	9	2.76
Pintores	24	4	2.92

Tabla 2: Comparación entre el número equivalente (*Neq*) y el número de clases sugeridas por el árbol de clasificación jerárquica, para varias tablas de datos de dimensiones n (número de individuos) por p (número de variables).

A manera de ilustración, presentamos en las figuras 1 y 2 los árboles de clasificación jerárquica para las tablas de notas escolares, construidos usando el índice de agregación de Ward.

Hemos de decir que, a pesar de que según el árbol jerárquico correspondiente a las notas escolares costarricenses, aparentemente hay dos "clases naturales" entre los individuos, en realidad hay tres clases naturales. Para ello puede verse el primer plano principal obtenido a partir del A.C.P., que se muestra en la figura 3.

3.3 Generación de reglas: determinación del número de componentes de una conjunción

En el diseño de un sistema experto, en ocasiones se recurre a métodos automáticos para la elaboración de una base de conocimiento formada por reglas; estos métodos se llaman usualmente *generadores de reglas*. Nosotros hemos trabajado [10, 11, 13, 16] sobre un método basado en principios estadísticos y euclídeos.

Se dispone de un conjunto de p variables cualitativas explicativas $x^1, \ldots, x^p$ y de una variable cualitativa a explicar y, y se quiere obtener reglas del tipo $C^j \rightarrow y^k$ sabiendo

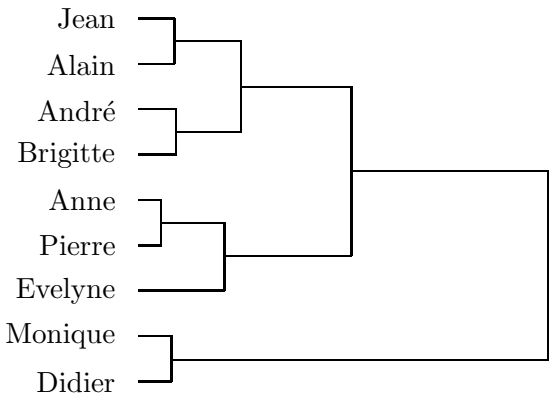

Figura 1: Arbol de clasificación jerárquica según Ward para la tabla de n otas escolares francesas

$I[y^k|C^j]$, donde C^j es una conjunción de modalidades de las variables explicativas, y^k es una modalidad de la variable a explicar, e I es una medida de la incertidumbre. El método está basado en el uso de medidas de asociación simétricas entre las modalidades explicativas y disimétricas entre éstas y las modalidades a explicar, y hace uso de estas medidas para representar las modalidades en espacios euclídeos apropiados con el fin de detectar las reglas mediante heurísticas de reducción de la dimensión y de agrupamiento. El método es iterativo, y en la primera iteración solamente analiza las modalidades, esto es, las conjunciones de longitud uno. Para la segunda iteración, se añaden las premisas de las reglas encontradas, y se hace una etapa de búsqueda de nuevas conjunciones de longitud dos, mediante el cruce de modalidades explicativas (esta etapa también está basada en técnicas de clasificación). Con el nuevo conjunto de conjunciones se buscan las reglas y se reiteran las dos operaciones: creación de nuevas conjunciones explicativas y búsqueda de reglas.

Se plantea entonces un problema: ¿Cuántas iteraciones hacer? ¿Será necesario hacer p iteraciones o bastará con hacer un cierto número, menor que p, a partir del cual la información obtenida será redundante? Para limitar la longitud de las conjunciones explicativas, hemos usado el número equivalente, calculado sobre la tabla de contingencia definida por las indicatrices de las modalidades explicativas, usando la métrica de chi–cuadrado. Para ilustrar las aplicaciones de este generador de reglas, damos algunos resultados obtenidos en la tabla 3. Estos resultados sugieren de nuevo que el número equivalente da una idea del número de unidades de información independientes.

Introducción de nuevas variables

El algoritmo de generación de reglas construye, a cada etapa, nuevas conjunciones explicativas mediante la intersección de conjunciones explicativas existentes; por lo tanto nos propusimos estudiar la evolución del *Neq* cuando se añaden nuevas variables en una tabla de datos.

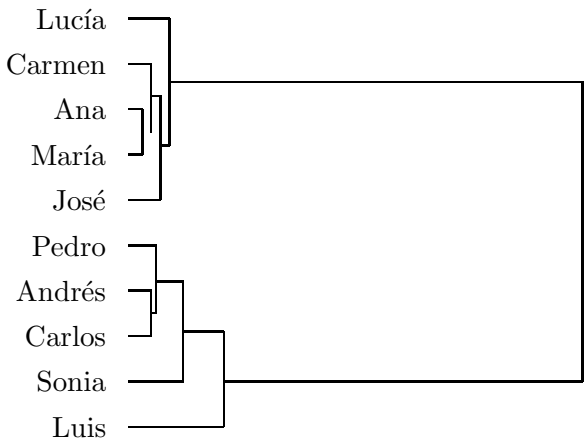

Figura 2: Arbol de clasificación jerárquica según Ward para la tabla de notas escolares costarricenses

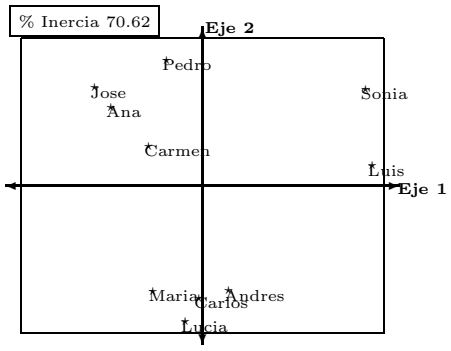

Figura 3: Primer plano principal del A.C.P. para la tabla de notas escolares costarricenses

Proposición 5 *Si* $Neq_p(X, D_{1/\sigma^2})$ *es el número equivalente asociado a las* p *variables* $x^1, \ldots, x^p$ *que definen la tabla* X*, para* $M = D_{1/\sigma^2}$*, entonces el número equivalente* $Neq_{p+q}(\tilde{X}, D_{1/\sigma^2})$ *asociado a las* $p+q$ *variables* $x^1, \ldots, x^p, x^{p+1}, \ldots, x^{p+q}$ *que definen la tabla* $\tilde{X}$*, para* $M = D_{1/\sigma^2}$*, es igual a:*

$$Neq_{p+q}(\tilde{X}, D_{1/\sigma^2}) = \frac{(p+q)^2}{\frac{p^2}{Neq_p(D_{1/\sigma^2})} + 2\rho_{Inter(p,q)} + \rho_{Intra(q)}}$$

donde $\rho_{Inter(p,q)} = \sum_{j=1}^{p} \sum_{k=1}^{q} \rho^2(x^j, x^{p+k})$ *es la suma de las correlaciones* inter *los grupos* $\{x^1, \ldots, x^p\}$ *y* $\{x^{p+1}, \ldots, x^{p+q}\}$*, y* $\rho_{Intra(q)} = \sum_{j=1}^{q} \sum_{k=1}^{q} \rho^2(x^{p+j}, x^{p+k})$*, la suma de las correlaciones* intra *el grupo* $\{x^{p+1}, \ldots, x^{p+q}\}$.

Tabla de datos	n	p	Neq	número de reglas	número de reglas con premisa de longitud $> Neq$
Datos simulados	100	3	1.64348	9	3 (de longitud 2)
Datos simulados	20	5	2.134	19	2 (de longitud 3)
Datos zoológicos	101	16	7.0	185	0 (hay una de longitud 7)

Tabla 3: Resultados del uso de número equivalente respecto al método de generación de reglas, donde n es el número de individuos y p el número de variables explicativas.

DEMOSTRACIÓN: Según la proposición 2, se tiene: $Neq_p(X, D_{1/\sigma^2}) = \frac{p^2}{\sum_{j=1}^{p}\sum_{k=1}^{p}\rho^2(x^j,x^k)}$ y $Neq_{p+q}(\tilde{X}, D_{1/\sigma^2}) = \frac{(p+q)^2}{\sum_{j=1}^{p+q}\sum_{k=1}^{p+q}\rho^2(x^j,x^k)}$. Además:

$$\begin{aligned}\sum_{j=1}^{p+q}\sum_{k=1}^{p+q}\rho^2(x^j,x^k) &= \sum_{j=1}^{p}\sum_{k=1}^{p}\rho^2(x^j,x^k) + 2\sum_{j=1}^{p}\sum_{k=1}^{q}\rho^2(x^j,x^{p+k}) + \sum_{j=1}^{q}\sum_{k=1}^{q}\rho^2(x^{p+j},x^{p+k}) \\ &= \frac{p^2}{Neq_p(X, D_{1/\sigma^2})} + 2\rho_{Inter(p,q)} + \rho_{Intra(q)}\end{aligned}$$

de donde se deduce el resultado. ■

Puede observarse que cuando se introduce una sola variable nueva x^{p+1} (*i.e.* $q = 1$), se tiene:

$$Neq_{p+1}(\tilde{X}, D_{1/\sigma^2}) = \frac{(p+1)^2}{\frac{p^2}{Neq_p(X,D_{1/\sigma^2})} + 2\sum_{j=1}^{p}\rho^2(x^j,x^{p+1}) + 1} \tag{2}$$

Una condición necesaria y suficiente para que el número equivalente $Neq_{p+1}(\tilde{X}, D_{1/\sigma^2})$ asociado a las $p+1$ variables $x^1, \ldots, x^p, x^{p+1}$, sea superior a $Neq_p(X, D_{1/\sigma^2})$, el número equivalente asociado a las p variables $x^1, \ldots, x^p$, es que

$$Neq_p(X, D_{1/\sigma^2}) < \frac{2p+1}{2\sum_{j=1}^{p}\rho^2(x^j,x^{p+1}) + 1}$$

En efecto, por la igualdad 2 se tiene $Neq_{p+1}(\tilde{X}, D_{1/\sigma^2}) > Neq_p(X, D_{1/\sigma^2})$ si y sólo si $\frac{(p+1)^2}{\frac{p^2}{Neq_p(X,D_{1/\sigma^2})}+2\sum_{j=1}^{p}\rho^2(x^j,x^{p+1})+1} > Neq_p(X, D_{1/\sigma^2})$ lo que es equivalente a $(p+1)^2 > p^2 + Neq_p(X, D_{1/\sigma^2})\left[2\sum_{j=1}^{p}\rho^2(x^j,x^{p+1}) + 1\right]$, puesto que el denominador de $Neq_{p+1}(\tilde{X}, D_{1/\sigma^2})$ es positivo

4 Conclusiones y perspectivas

El número equivalente tiene propiedades interesantes que pueden explotarse en análisis de datos. Las aplicaciones mostradas han ayudado a abordar problemas abiertos que tiene el

análisis de datos, pudiéndose aun profundizar en algunas propiedades teóricas que podrían ayudar a esclarecer mejor los problemas planteados.

Sin embargo, las investigaciones deben continuarse para hacer comparaciones con métodos y criterios existentes para la determinación del número de factores en un análisis factorial o el número de clases en clasificación automática.

También debe tratar de generalizarse al caso en que se tenga una tabla con variables cualitativas, o cuando se tiene una tabla de contingencia. Este último caso sería particularmente útil para estimar el número de componentes en un Análisis de Correspondencias.

Por otro lado, es posible que el número equivalente encuentre aplicaciones en otros campos del análisis de datos, como en regresión y en discriminación. En efecto, uno podría pensar en abordar el problema del número de variables explicativas necesarias para un problema de regresión (no necesariamente lineal, y sin suponer ninguna distribución de probabilidad, ni en las variables activas ni en los residuos); así mismo, se podría pensar en que el número equivalente puede ser útil para la determinación del número de variables explicativas significativas en discriminación (de nuevo sin hacer hipótesis de probabilidad). Por otra parte, el conocido problema de la determinación del número de neuronas en una red neuronal con una capa escondida (para la aplicación del método de retropropagación del gradiente), podría encontrar alguna luz desde el punto de vista del número equivalente, adaptando su definición al uso de los pesos sinápticos entre las neuronas. Estas cuestiones serán estudiadas en futuras investigaciones dentro del Programa de Investigación en Modelos y Análisis de Datos de la Universidad de Costa Rica.

Referencias

[1] Cailliez, F.; Pagès, J.P. (1976) *Introduction à l'Analyse des Données*. Société de Mathématiques Appliquées et de Sciences Humaines, Paris.

[2] Der Mégréditchian, G. (1979) "L'optimisation des réseaux d'observation des champs météorologiques", *La Météorologie*, 6(17): 51–66.

[3] Der Mégreditchian, G. (1988) "Análisis espacial de los campos meteorológicos y aplicación a la optimización de redes de medida". En: Memorias *IV Simposio Métodos Matemáticos Aplicados a las Ciencias*, B. Montero & J. Poltronieri (eds.), 1984, Editorial de la Universidad de Costa Rica, pp. 1–34.

[4] Der Mégreditchian, G. (1988) "Condensación óptima de la información meteorológica por medio del análisis en componentes principales". En: Memorias *IV Simposio Métodos Matemáticos Aplicados a las Ciencias*, B. Montero & J. Poltronieri (eds.), 1984, Editorial de la Universidad de Costa Rica, pp. 35–61.

[5] Diday, E.; Lemaire, J.; Pouget, J.; Testu, F. (1982) *Eléments d'Analyse de Données*. Dunod, Paris.

[6] Escofier, B.; Pagès, J. (1988) *Analyses Factorielles Simples et Multiples: Objectifs, Méthodes et Interprétation*. Dunod, Paris.

[7] Jambu, M. (1989) *Exploration Informatique et Statistique des Données*. Dunod, Paris.

[8] Lebart, L.; Morineau, A.; Fénelon, J.-P. (1985) *Tratamiento Estadístico de Datos. Métodos y Programas*. Marcombo, Barcelona.

[9] Murillo, A. (1996) *Proposición de un índice para la interpretación de árboles de clasificación basado en conjuntos difusos*. Tesis para optar al grado de Magister Scientiæ en Computación, Instituto Tecnológico de Costa Rica, Cartago.

[10] Schektman, Y.; Trejos, J.; Troupé, M. (1992) "Un générateur de règles floues à partir de bases de données volumineuses". En: Actes des *3-èmes Journées Symbolique–Numérique*, Université Paris IX–Dauphine, pp. 121–130.

[11] Schektman, Y.; Trejos, J.; Troupé, M. (1994) "Generación de reglas estadísticas a partir de grandes bases de datos", *Revista de Matemática: Teoría y Aplicaciones*, 1(1): 87-100.

[12] Trejos, J. (1994) "Generación de reglas: un enfoque estadístico y euclídeo". En Memorias del *II Encuentro Centroamericano de Investigadores en Matemáticas*, I parte, G. Mora (ed.), 19-28.

[13] Trejos, J. (1994) *Contribution à l'Acquisition Automatique de Connaissances à Partir de Données Qualitatives*. Thèse de doctorat, Université Paul Sabatier, Toulouse.

[14] Trejos, J. (1995) *Principios de Análisis Multivariados de Datos*. Notas de curso, Universidad de Costa Rica, San Pedro.

[15] Trejos, J. (1996) "Propiedades y aplicaciones del número equivalente en análisis de datos", it IV Encuentro Centroamericano de Investigadores en Matemática, Antigua Guatemala, 17–19 enero.

[16] Troupé, M. (1994) *Contribution à la Régression Multiple Multidimensionnelle et à la Génération de Règles Incertaines*. Thèse de doctorat, Université Paul Sabatier, Toulouse.