ANALISIS DE DATOS Y LOGICA DE ENUNCIADOS

Yosu YURRAMENDI

ABSTRACT

We understand by data analysis the full set of methods which are used to point out the structural relations between certain objects and the characteristics observed in them. In the case where these characteristics admit a process of dichotomy, the relations can be expressed in terms of propositional logic.

In the present paper, we endeavour to make clear the opposition between the problems of propositional logic and those of certain model utilised in data analysis: in propositional logic, we look for a solution verified by a given set of propositions; as long as we are concerned with data analysis, we look for a solution in terms of propositions which include as proper parts the observed objects.

Finally, we draw the main lines of two very used methods which can be relevant with the aid of computer: the factor analysis and the automatic classification (cluster analysis).

SUMARIO

1. Introducción.

Frecuentemente el inicio de muchas investigaciones es una recogida de datos. Por lo que sabemos, no hay regla fija alguna para el investigador que le oriente en la construcción de un cuerpo teórico que responda a los datos recogidos. De todas formas, no es un trabajo yermo el insistir en esos quehaceres, pues los datos hay que describirlos, resumirlos y clasificarlos de alguna manera.

Hay un conjunto de técnicas que vale para ayudar a dar los primeros pasos para esos quehaceres: se denomina estadística descriptiva. Veamos lo que opinaba hace ya algún tiempo el renombrado estadístico John Tukey en un artículo titulado "The Future of Data Analysis" aparecido en los *Annals of Mathematical Statistics*, Vol. 33 (1962), p. 2:

> *"For a long time I have thought I was statistician, interested in inferences from the particular to the general.*

THEORIA - Segunda Época,
Año I (1985), Nº 1, pp. 275-291.

But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt".

Y un poco más adelante:

"I have come to feel that my central interest is in data analysis, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways for planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data".

Quizás en estas líneas no ha quedado muy claro de qué trata el análisis de datos. En general, los datos vienen de la observación de características en varios objetos: analizar los datos, por tanto, consiste en sacar a la luz las relaciones entre los objetos, las relaciones entre las características y las relaciones entre los objetos y las características.

Respecto a cómo deben ser las técnicas o procedimientos del análisis de datos, he aquí cómo opinaba John Tukey en el citado artículo (p. 6):

"Data analysis, and the parts of statistics which adhere to it, must then take on the characteristics of a science rather than those of mathematics, specifically:

(a-1) Data analysis must seek for scope and usefulness rather than security.

(a-2) Data analysis must be willing to err moderately often in order that inadequate evidence shall more often suggest the right answer.

(a-3) Data analysis must use mathematical argument and mathematical results as bases for judment rather than as bases for proof or stamps of validity".

Y un poco más tarde, añade:

"All sciences have much of art in their makeup (...). As well as teaching facts and well-established structures, all sciences must teach their apprentices how to think about things in the manner of that particular science,

and what are its current beliefs and practices. Data analysis must do the same".

Aquí entendemos la palabra "structure" como un conjunto de axiomas y definiciones. Las propiedades se deducen a través del razonamiento (lógico).

Desde los tiempos de este artículo hasta hoy en día, el desarrollo y difusión de las calculadoras han sido sorprendentes. Su capacidad de cálculo y memoria ha posibilitado proyectos que antes eran inimaginables, y de esta manera, los volúmenes de datos que son analizables son cada vez más grandes.

Veamos ahora cuáles son los principios del análisis de datos, unos diez años más tarde, en opinión del que ha sido precursor en la escuela francesa, Jean Paul Benzécri, en la obra titulada *L'Analyse des Données* (Tome I: *La Taxinomie*; Tome II: *L'Analyse des Correspondances*), París: Dunod, 1973, 4ª edición, 1982 (realizada por un extenso equipo de colaboradores. bajo la dirección de Benzécri), t. II, pp. 3-17:

"1er Principe: Statistique n'est pas probabilité. Sous le nom de statistique mathématique, des auteurs (qui, je vous le dis en français, n'écrivent guère dans notre langue. ..) ont édifié une pompeuse discipline, riche en hypothèses qui ne sont jamais satisfaites dans la pratique. Ce n'est pas de ces auteurs qu'il faut attendre la solution de nos problèmes typologiques.

.....

2ème Principe: Le modèle doit suivre les données, non l'inverse.' (...) Mais ce dont nous avons besoin c'est d'une méthode rigoureuse qui extraie des structures à partir des données.

.....

3ème Principe: Il convient de traiter simultanément des informations concernant le plus grand nombre possible de dimensions.

.....

4ème Principe: Pour l'analyse des faits complexes et notamment de faits sociaux, l'ordinateur est indispensable.

Principe évidemment vrai ... mais qu'en eussent pensé nos pères les Gaulois il y a 15 ans?

.....

5ème Principe: Utiliser un ordinateur implique d'abondonner toutes techniques conçues avant l'avènement du calcul automatique. Je dis technique, non science: (...)".

Por lo que se puede apreciar, y a pesar de los estilos diferentes de ambos profesores, hay puntos en común ... Parece ser que en opinión de ambos el análisis de datos es la configuración nueva de la estadística descriptiva clásica.

En el presente artículo presentaremos la construcción de un modelo que responde a estos principios, que es completamente general y que tiene ligazón con la lógica de enunciados. Plantearemos en él diversos problemas.

2. El modelo del cuestionario.

Al ir a construir un modelo ante un montón de datos, hay que dar, desde algún enfoque, la posibilidad de reconstrucción de los datos. Sin embargo, la mayoría de las veces difícilmente se puede cumplir esta intención, pues la codificación misma de los datos no suele recoger más que una parte de la información.

El modelo que presentaremos aquí se denomina cuestionario, pues recuerda el que a menudo se ha utilizado en las ciencias sociales.

Sea Ω el conjunto de los objetos que han de analizarse. Supondremos que Ω es finito:

$$\Omega = \{w_1, w_2, ..., w_n\}$$

Sea X el conjunto de modalidades de una característica del conjunto Ω, o, siguiendo la comparación anterior, el conjunto de las respuestas que pueden dar a una pregunta los pertenecientes a Ω. Supondremos que también X es finito:

$$X = \{x_1, x_2, ..., x_p\}$$

Sea X^*: $\Omega \to X$ una pregunta hecha a los objetos de Ω, ésto es, una aplicación o variable de Ω.

Sean:

$\{X_n \mid n=1, \ldots, r\}$ y $\{X^*_n \mid n=1, \ldots, r\}$

respectivamente una colección de conjuntos de las respuestas que pueden darse a r preguntas y su correspondiente colección de aplicaciones.

Así, lo que da por supuesto el modelo es que todos los objetos de Ω responden de una única manera a cada una de las preguntas.

Sea:

$\{e^*_{nk} \mid n=1, \ldots, r \text{ y } k=1, \ldots, p_n\}$

una colección de aplicaciones de Ω en el conjunto $\{0, 1\}$ definida de la siguiente manera:

para todo $w \in \Omega$, todo $n=1, \ldots, r$ y todo $k=1, \ldots, p_n$,

si $X^*_n(w)=X_{nk}$,

entonces $e^*_{nk}(w)=1$,

y, si no, $e^*_{nk}(w)=0$.

Puede notarse que este conjunto es el de los identificadores de las respuestas, es decir, el de las funciones características. Una propiedad de este conjunto es la siguiente:

para todo $w \in \Omega$ y todo $n=1, \ldots, r$:

$$\sum_{k=1}^{p_n} e^*_{nk}(w) = 1$$

Este conjunto de aplicaciones será designado por el símbolo 'E'.

Evidentemente, hay una correspondencia biunívoca entre los conjuntos:

$$\bigcup_{n=1}^{r} X_n \quad \text{y} \quad E$$

Al construir este modelo, como ha quedado señalado, la intención

no es otra que la de descubrir las posibles relaciones notables (en un sentido que habrá que especificar) entre los elementos de Ω (los objetos), entre las respuestas, y entre los objetos y las respuestas, a través de los identificadores correspondientes a las respuestas.

Sea:

$E^*: \Omega \to \{0, 1\}^p$, donde $p = \sum_{n=1}^{r} p_n$, definida de esta manera:

para todo $w \in \Omega$, $E^*(w) = \{e^*_{nk}(w) \mid n=1, ..., r; k=1, ..., p_n\}$.

Restringiremos el modelo construido al caso de las preguntas dicotómicas, ya que será suficiente para el objetivo perseguido en este artículo. Así, para todo n=1, ..., r, el conjunto X_n está compuesto por dos elementos:

$$X_n = \{x_n^+, x_n^-\},$$

y sus respectivos identificadores son designados por e^{*+}_n y e^{*-}_n.

La propiedad de los identificadores, mencionada más arriba, queda convertida en la siguiente:

para todo $w \in \Omega$ y todo n=1, ..., r, $e^*_n{}^-(w) = 1 - e^*_n{}^+(w)$.

Por tanto, y para todo lo que sigue, tomaremos en cuenta la siguiente aplicación, que es pareja a E^*:

$E^{*+}: \Omega \to \{0, 1\}^r$, donde para todo $w \in \Omega$,

$E^{*+}(w) = \{e^*_n{}^+(w) \mid n=1, ..., r\}$.

3. La lógica de enunciados.

Al definir los conjuntos X_n (n=1, ..., r), tal y como hemos dicho, definimos las posibles respuestas a r preguntas. Si las respuestas posibles son dos, es fácilmente imaginable que una de ellas puede expresar la presencia de una cualidad, y la otra su ausencia.

ANALISIS DE DATOS Y LOGICA DE ENUNCIADOS

En la lógica de enunciados (el nivel más simple de la lógica), las relaciones entre dos enunciados X_k, X_l son los subconjuntos del conjunto $X_k \times X_l$. El producto cartesiano $X_k \times X_l$ está compuesto de 4 elementos y, por lo que se sabe, se pueden definir 2^4 relaciones diferentes. Las más usuales son las siguientes:

a) Conjunción: $X_k \& X_l = \{(x_k^+, x_l^+)\}$

b) Disyunción inclusiva: $X_k v X_l = \{(x_k^+, x_l^+), (x_k^+, x_l^-), (x_k^-, x_l^+)\}$
$= \{(x_k^-, x_l^-)\}^c$

c) Condicional: $X_k \rightarrow X_l = \{(x_k^+, x_l^+), (x_k^-, x_l^+), (x_k^-, x_l^-)\}$
$= \{(x_k^+, x_l^-)\}^c$

d) Bicondicional: $X_k \leftrightarrow X_l = X_k \rightarrow X_l \cap X_l \rightarrow X_k$
$= \{(x_k^+, x_l^+), (x_k^-, x_l^-)\}$

e) Disyunción exclusiva: [1] $X_k \not\leftrightarrow X_l = \{(x_k^+, x_l^-), (x_k^-, x_l^+)\}$

Por otra parte, las aplicaciones de $\{0, 1\}$ en $\{0, 1\}$ son también 2^4; por tanto, se puede establecer una correspondencia biunívoca entre las relaciones y las aplicaciones tras el siguiente emparejamiento:

(x_k^+, x_l^+) y $(1, 1)$;

(x_k^+, x_l^-) y $(1, 0)$;

(x_k^-, x_l^+) y $(0, 1)$:

(x_k^-, x_l^-) y $(0, 0)$.

Las relaciones aparecidas en los ejemplos pueden ser expresadas así por medio de las llamadas funciones lógicas:

Para todo $(e_1, e_2) \in \{0, 1\}^2$:

a) $f_{\&}(e_1, e_2) = e_1 e_2$

b) $f_v(e_1, e_2) = 1-(1-e_1)(1-e_2)$

c) $f_{\rightarrow}(e_1, e_2) = 1-(e_1(1-e_2)$

d) $f_{\leftrightarrow}(e_1, e_2) = 1-(e_1-e_2)^2$

e) $f_{\not\leftrightarrow}(e_1, e_2) = (e_1-e_2)^2$

Los subconjuntos del producto cartesiano

$\prod_{n=1}^{r} X_n$ expresan las relaciones entre los r enunciados, algunas de las cuales tienen sentido en el lenguaje. Las relaciones de los ejemplos anteriores están, por decirlo así, sumergidas en este nuevo conjunto:

a) $X_k \& X_l = \{(x_1, \ldots, x_r) | x_n \in X_n \ (n=1, \ldots, r) \ \& \ x_k = x_k^+ \text{ y } x_l = x_l^+\}$

b) $X_k \vee X_l = \{(x_1, \ldots, x_r) | x_n \in X_n \ (n=1, \ldots, r) \ \& \ x_k = x_k^+ \text{ o } x_l = x_l^+\}$

c) $X_k \rightarrow X_l = \{(x_1, \ldots, x_r) | x_n \in X_n \ (n=1, \ldots, r) \ \& \ x_k = x_k^- \text{ o } x_l = x_l^+\}$

d) $X_k \leftrightarrow X_l = (X_k \rightarrow X_l) \cap (X_l \rightarrow X_k)$

e) $X_k \not\leftrightarrow X_l = (X_k \leftrightarrow X_l)^c$

Si a los elementos de $\{0, 1\}^r$ les corresponde un emparejamiento similar al efectuado en dos dimensiones, las relaciones de los ejemplos pueden ser expresadas por estas funciones lógicas:

para todo $(e_1, \ldots, e_r) \in \{0, 1\}^r$,

a) $f(e_1, \ldots, e_r) = e_k e_l$

b) $f(e_1, \ldots, e_r) = 1-(1-e_k)(1-e_l)$

c) $f(e_1, \ldots, e_r) = 1-e_k(1-e_l)$

d) $f(e_1, \ldots, e_r) = 1-(e_k-e_l)^2$

e) $f(e_1, \ldots, e_r) = (e_k-e_l)^2$

Desde un enfoque lógico, cuando se plantea un problema al nivel de los enunciados, siempre se puede encontrar una función lógica f que lo refleje (en el sentido de la correspondencia biunívoca definida). La solución es el subconjunto $f^{-1}(1)$. Por lo que se sabe, si $f^{-1}(1) = \emptyset$, ésto se dice que la relación es una contradicción, y si $f^{-1}(1) = \{0, 1\}^r$, que es una tautología.

En cambio, desde el enfoque del análisis de datos, el problema es otro. Por medio de los identificadores de respuestas llegamos a un subconjunto de $\{0,1\}^r$, y el problema consiste en saber de qué funciones lógicas es solución, así como de qué funciones lógicas son solución algunas partes especiales (en un sentido que será especificado más adelante) de ese subconjunto.

ANALISIS DE DATOS Y LOGICA DE ENUNCIADOS

Cuando tanto el tamaño (cardinal) de Ω es suficientemente grande (digamos N = 500) como también el número de características (r = 30), la imagen de Ω a través de la aplicación E^{*+} suele ser bastante grande como para que se esclarezcan fácilmente relaciones entre dos o tres características (es decir, para aquélla sea solución de éstas) otras que las triviales como $X_k \times X_l$, a pesar de que al lado de 2^{30} (el cardinal de $\{0,1\}^{30}$) sea muy pequeño; por otra parte, la comprensión de los datos pide relaciones simples (esto es, conjunciones, disyunciones, condicionales de términos simples, etc.).

Entre estos dos polos contrapuestos hay que buscar una solución. Desde la estadística se ha hecho una aproximación: se buscan relaciones simples para la mayoría de los objetos de Ω o de algunas de sus partes especiales (habrá que precisar el sentido de la palabra "mayoría").

El tipo de ligazones que se van a analizar tiene que estar en el corazón mismo del método. Por ejemplo, si se quisiera analizar las relaciones condicionales entre dos preguntas, no sería descabellado el calcular para cada par de preguntas cantidades como

$$\sum_{w \in \Omega} (1-e^*_k{}^+(w)(1-e^*_k{}^+(w)))$$

(si esa relación se diese en todo Ω, la suma valdría N), y tomar en cuenta aquéllas que fuesen las mayores; por ejemplo, la búsqueda de escalas de Guttman[2] ha sido una práctica muy común hasta hace poco tiempo.

Otro ejemplo es el de la relación conjuntiva entre dos preguntas. Se podría basar el nivel de asociación entre dos preguntas en la cantidad de objetos que han respondido afirmativamente a las dos:

$$n_{kl}{}^{++} = \sum_{w \in \Omega} e^*_k{}^+(w)e^*_l{}^+(w)$$

Muchas veces sin embargo, no es adecuado devaluar las respuestas negativas: si una de las preguntas se hubiera planteado al revés, el nivel de asociación sería distinto. Ante casos como éste, el procedimiento más propio es el de tomar en cuenta las dos respuestas. Así, los cuatro posibles sucesos correspondientes a dos preguntas, es decir, los

elementos de $X_k \times X_l$ tienen el mismo rol en el análisis. Por otra parte, no se pierde más información.

Se sabe que

$$n_{kl}^{++} + n_{kl}^{+-} + n_{kl}^{-+} + n_{kl}^{--} = N$$

	x_l^+	x_l^-
x_k^+	n_{kl}^{++}	n_{kl}^{+-}
x_k^-	n_{kl}^{-+}	n_{kl}^{--}

Los métodos de análisis de datos tienen que ser potentes para dar a conocer las relaciones existentes al investigador: tienen que encaminarse hacia un análisis general de las características, y tienen que dejar al descubierto sus ligazones, similitudes y diferencias.

4. Representación geométrica.

En este capítulo ubicaremos el modelo construido en un espacio geométrico. Esta forma de ubicación es generalizable a otros muchos modelos.

Sea

$$X = \bigcup_{n=1}^{r} X_n \quad \text{el conjunto de respuestas.}$$

El mencionado nivel de asociación es la aplicación $k: J \times J \to N$, donde, para todo $(j,j') \in J \times J$,

$$k(j,j') = \sum_{w \in \Omega} e^*_j(w) e^*_{j'}(w)$$

De esta definición se deducen las siguientes propiedades:

a) La aplicación es simétrica: para todo $(j,j') \in J \times J$, $k(j.j')=k(j',j)$

b) Para todo $j \in J$, $k(j,j')$ es su frecuencia en Ω

c) Para todo $j \in J$ y todo $n=1, \ldots, r$, $k(x_n^+,j)+k(x_n^-,j)=k(j,j)$

d) Para todo $n=1, \ldots, r$, $k(x_n^+,x_n^-)=0$

e) Para todo $n=1, \ldots, r$, $k(x_n^+,x_n^+)+k(x_n^-,x_n^-)=N$

Para entender mejor los niveles de asociación, es más adecuado expresarlos proporcionalmente:

Para todo $j \in J$ y todo $j' \in J$, $f^j_{j'} = k(j,j')/k(j,j)$ (por tanto, $f^j_{j'} \geq 0$).

Así:

a) Para todo $j \in J$ y todo $n=1, \ldots, r$, $f^j_{x_n^+}+f^j_{x_n^-} = 1$

b) Para todo $n=1, \ldots, r$: $f^{x_n^+}_{x_n^+}=1$; $f^{x_n^+}_{x_n^-}=0$; $f^{x_n^-}_{x_n^+}=0$; $f^{x_n^-}_{x_n^-}=1$.

El <u>perfil</u> de una respuesta es este punto de $|R^{2r}$:

para todo $j \in J$, $f^j_J = (f^j_{j'} | j \cdot J) \in |R^{2r}$

La <u>masa</u> de una respuesta es este número positivo:

para todo $j \in J$, $f_j=k(j,j)/N$.

Por otra parte, tienen esta propiedad:

para todo $n=1, \ldots, r$: $f_{x_n^+}+f_{x_n^-}=1$; por tanto: $\sum_{j \in J} f_j = r$

De esta manera, al conjunto de respuestas se le ha hecho corresponder un sistema de puntos de $|R^{2r}$:

$$N(J) = \{(f^j_J, f_j) \mid j \in J\},$$

donde f^j_J son las coordenadas de $j \in J$ y f_j su masa o importancia en ese sistema (en razón de su frecuencia).

El centro de gravedad (interpretación mecánica) o perfil medio (interpretación estadística) de este sistema de puntos tiene las siguientes coordenadas:

para todo $j \in J$,

$$1/r \times \sum_{j' \in J} f_{j'} \times f_j^{j'} = 1/r \times \sum_{j' \in J} \frac{k(j',j')}{N} \times \frac{k(j',j)}{k(j',j')} = 1/r \times \frac{rk(j,j)}{N} = f_j$$

o, resumiendo, $f_j = (f_j \mid j \in J) \in |R^{2r}$ es el perfil medio de $N(J)$.

Además, el perfil medio de las dos respuestas a una misma pregunta es el mismo:

para todo $j \in J$ y todo $n=1, \ldots, r$:

$$f_{x_n^+} \times f_j^{x_n^+} + f_{x_n^-} \times f_j^{x_n^-} = \frac{k(x_n^+,x_n^+)}{N} \times \frac{k(x_n^+,j)}{k(x_n^+,x_n^+)} + \frac{k(x_n^-,x_n^-)}{N} \times \frac{k(x_n^-,j)}{k(x_n^-,x_n^-)}$$

$$= \frac{k(j,j)}{N} = f_j$$

así, pues, una propiedad de este sistema es la siguiente: los perfiles de las dos respuestas a una pregunta y el perfil medio están en una recta.

El esclarecimiento de las relaciones entre las respuestas, desde dentro de este sistema, es uno de los objetivos.

Se hace notar que, si la relación entre dos preguntas es la bicondicional o la disyuntiva exclusiva, entonces los perfiles de sus respuestas, tomados dos a dos de alguna manera, son iguales. E inversamente, si los perfiles de dos respuestas son iguales, también lo serán los de sus contrarias y, por tanto, la correspondiente relación entre las preguntas será la bicondicional (en el caso de que sean iguales las afirmativas y/o las negativas) o la disyunción exclusiva (en otro caso).

Así podemos decir lo siguiente: la proximidad de dos respuestas en este espacio está concebida en relación a su equivalencia. En cambio, la proximidad no está todavía totalmente definida: es preciso definir la distancia entre dos perfiles para introducir las respuestas en un espacio métrico.

A una distancia tal se le puede pedir que cumpla la siguiente propiedad: si dos preguntas están en relación bicondicional (o en relación disyuntiva exclusiva), entonces el aunar ambas (reduciendo la dimensión del espacio) no debe alterar las distancias entre las demás respuestas.

Si a esta demanda le agregamos la de ser euclidea (para sujetarnos a una realidad que nos es conocida y, además, por comodidad matemática), entonces se define esa distancia del modo siguiente:

para todo $(j',j'') \in J \times J$:

$$d^2(j',j'') = \sum_{j \in J} (f_j^{j'} - f_j^{j''})/f_j$$

Se observa que los sumandos están sopesados por la importancia de las respuestas. A esta distancia se le denomina distancia de χ^2, ya que recuerda, en el caso de dos preguntas, la fórmula utilizada en el test de χ^2 de la estadística clásica.

Una vez representadas geométricamente las respuestas, en lo que viene presentaremos dos métodos generales para su análisis: el análisis factorial y la clasificación automática.

5. El análisis factorial.

El método del análisis factorial de un sistema de puntos fue concebido en este siglo en el campo de la psicometría, al formalizar Thurstone (a partir de 1930) la teoría de Spearman sobre la inteligencia general.

En este campo ha habido al respecto muchos puntos de vista, pero desde el análisis de datos podría ser presentado así: buscar en el espacio $\mathbb{R}^{2^r}$ el eje principal del sistema de puntos $N(J)$ es decir, proyectar (en el sentido de la distancia definida) las respuestas sobre cualquier eje y buscar aquél que tenga mayor inercia (interpretación mecánica) o mayor varianza (interpretación estadística); después, plantear el mismo problema entre las perpendiculares al eje elegido, y así sucesivamente hasta agotar la dimensión del espacio.

A los valores que toman las respuestas en estos ejes se les llama factores, y son, por así decirlo, sus nuevas coordenadas en el espacio

$\mathbb{R}^r$. Esas coordenadas cartesianas tienen ciertas propiedades; por ejemplo, la mayoría de las veces casi toda la información (en el sentido de la varianza) se concentra en los primeros factores.

Los elementos de Ω pueden ser también proyectados en este espacio (por tanto, en los ejes principales), definiéndoles estos perfiles:

para todo $w \in \Omega$ y todo $j \in J$, $f_j^w = e^*_j(w)$.

Si en el modelo del cuestionario, en lugar de basar el análisis factorial en la aplicación k: $J \times J \to N$, se hubiera basado en la aplicación k': $\Omega \times J \to \{0, 1\}$, donde:

para todo $(w,j) \in \Omega \times J$, $k'(w,j) = e^*_j(w)$,

siendo definidos los perfiles, las masas y las distancias entre los elementos de J de similar manera a la anterior (en relación a los elementos de Ω) así como los correspondientes a los de Ω (en relación a los de J), parece que se procedería a otro análisis factorial. De hecho, los dos análisis son parecidos y pueden establecerse fórmulas matemáticas que los relacionen. Además, los factores de los elementos de Ω en los dos análisis son los mismos.

Se demuestra también que en el segundo análisis los ejes principales de los sistemas de puntos $N(\Omega)$ y $N(J)$ son los mismos, y por tanto, que pueden representarse simultáneamente las respuestas y los objetos.

En la práctica estos resultados son muy importantes: por un lado, se podría observar en cada uno de los ejes qué objeto está cerca (en el sentido de la distancia del χ^2 de qué respuesta, y así medir el nivel de asociación relativo; por otra parte, se constata que para la obtención de los resultados es suficiente proceder con el primer análisis, ya que las dimensiones de los cálculos son mucho más reducidas.

Puesto que el sistema de puntos de las respuestas se proyecta en un eje principal (o en el plano definido por dos ejes principales), las similitudes entre las respuestas pueden aparecer adecuadamente. Por ejemplo, si entre las preguntas existiese una estructura de escala de Guttman, el primer factor los ordenaría y los demás factores serían funciones polinómicas del primero.

Para el cálculo de los factores, salvo en los casos en que r es

muy pequeño, el uso de las calculadoras resulta ineludible. El hacerlo a mano resultaría un trabajo excesivamente largo y pesado.

6. Clasificación.

El objetivo de los métodos de clasificación es el de agrupar en pocas clases homogéneas los elementos de un conjunto.

En el presente apartado, nos limitaremos a los casos en que cada elemento pertenece a una sola clase, es decir a los casos de particiones[3].

Estos métodos pueden dividirse en dos: los jerárquicos y los no jerárquicos.

Los no jerárquicos forman las particiones directamente, una vez que el número de clases ha sido determinado. En cambio, los jerárquicos forman una sucesión de particiones que están ordenadas por la relación ser más fina que[4].

Se puede decir que éstos últimos fueron concebidos por los taxonomistas de los siglos XVII y XVIII.

Un ejemplo de estos métodos se basa en un algoritmo que es tan simple como productivo:

1. Tomar la partición de las clases singulares, es decir, la formada por las clases de un único elemento;

2. Juntar dos clases siguiendo un criterio y, por tanto, construir una nueva partición (evidentemente, menos fina que la anterior);

3. Si la nueva partición es la formada por el conjunto entero sólo, parar; si no, volver al paso 2...

Al aplicar este algoritmo (u otro análogo), la discusión se presenta en la elección del criterio de agregación de clases.

Teniendo en cuenta que ante una partición se tiene, por un lado, la inercia o varianza interior a cada clase, y por otro, la inercia o varianza entre las clases (una vez elegido un representante de cada una), un criterio (denominado "de Ward" (1963)) podría ser éste: agregar o juntar las dos clases que diesen la menor varianza interna (o aquéllas que contribuyesen menormente a la varianza entre las clases).

Una vez que se han construido sendas clasificaciones sobre los conjuntos Ω y J, la explicitación de los niveles de asociación entre los representantes de las clases supone una condensación de los primeros (relativos a las particiones de las clases singulares), y la información aparecerá resumida. Las clases de uno y otro conjunto (al principio de este artículo las hemos denominado "especiales") pueden ser conjuntamente comparadas e interpretadas. Así, si todos los elementos de una parte o clase de Ω responden de la misma manera a una cuantas preguntas (formando dichas respuestas una clase o parte "especial" de J), entonces puede decirse que queda cumplida una relación de conjunción en esa parte de Ω.

Los problemas de las relaciones condicionales ocupan un lugar privilegiado entre los métodos de clasificación. Una vez definida una partición de Ω por medio de una pregunta (o varias), la búsqueda de las relaciones entre las clases de esta partición y las respuestas al resto del cuestionario supone la búsqueda de las relaciones condicionales de una manera directa. En estadística este procedimiento es conocido con el nombre de discriminación.

Por último, tenemos la ligazón entre el análisis factorial y la clasificación. Si en la aplicación de ambos métodos sobre el sistema de puntos $N(J)$ se ha usado la misma distancia, ambos resultan complementarios, es decir, se conocen índices que sirven para interpretar mutuamente los factores (las relaciones más notables) y las clases. El lector que quisiera entrar en estos detalles así como en otros que no han sido mencionados a lo largo de este artículo, encontrará valiosos útiles en la bibliografía siguiente.

NOTAS

[1] Si se plantea una de las preguntas al revés, y sólo una, aparece la relación bicondicional.

[2] Una escala de Guttman sobre r cuestiones está formada por la intersección de (r-1) relaciones condicionales simples $(X_i \rightarrow X_{i+1})$ (i=1, ..., r-1), del tipo siguiente:

$$(X_1 \rightarrow X_2) \cap (X_2 \rightarrow X_3) \cap \ldots \cap (X_{r-1} \rightarrow X_r).$$

[3] Sean A y |P dos conjuntos. |P es una partición de A si y sólo si:

a) para todo $P \in |P$, $P \subset A$ y $P \neq \emptyset$;

b) $\sum_{P \in |P} P = A$;

c) para todo $P \in |P$ y todo $P' \in |P$:

si $P \neq P'$, entonces $P \cap P' = \emptyset$.

[4] Sean P y P' dos particiones del conjunto A. P es más fina que P' si y sólo si para todo P P existe P' P' tal que P P'.

BIBLIOGRAFIA

BENZECRI, J. P. *Histoire et Préhistoire de L'Analyse des Données*, París: Dunod, 1982.

BENZECRI, J. P. et al. *L'Analyse des Données* (Tome I: *La Taxinomie*; Tome II: *L'Analyse des Correspondances*), París: Dunod, 1ère éd. 1973; 4ème éd. 1982.

TUKEY, J. W. "The Future of Data Analysis", *Annals of Mathematical Statistics*, Vol. 33 (1962).

TUKEY, J. W. *Exploratory Data Analysis*, Addison Wesley, 1977.

Facultad de Filosofía y CCEE
Universidad del País Vasco, San Sebastián