BIBLIOTECA DE PROGRAMAS

ANALISIS DE TRES PROGRAMAS SOBRE CONGLOMERADOS

Introducción

El gran impulso que está recibiendo la investigación, en campos tales como la sociología, psicología, medicina, - biología y agricultura; donde las variables en estudio son- en numerosas ocasiones cualitativas, e incluso cuando son - cuantitativas, no se ajustan con fiabilidad en un gran núme ro de casos a las hipótesis de los modelos estadísticos li- neales, tradicionalmente empleados en su tratamiento; ha - llevado a la necesidad de encontrar otros métodos para tra- tar esta información, siendo uno de estos el análisis de - conglomerados o clustering; herramienta básica para las téc nicas de ordenación y clasificación.

El objeto fundamental de las técnicas de ordenación- y clasificación, es determinar estructuras básicas y funda- mentales de sistemas multivariables. El estudio se comienza con un universo simple compuesto de individuos. Estos indi- viduos son caracterizados de acuerdo con un conjunto de cua lidades, (atributos, parámetros, propiedades, caracteres,- etc.) y son representados en un espacio multidimensional - cuyos ejes coordenados son las cualidades.

Las técnicas de ordenación intentan reducir la dimen

sionalidad del espacio, reemplazando el conjunto original de cualidades por un nuevo conjunto de menor dimensión. La importancia que estos nuevos tratamientos pueden tener para los estudios arriba indicados; nos han llevado a tratar de encontrar una serie de programas que agrupen un conjunto de individuos en clases o conglomerados, de acuerdo con ciertas reglas, y los tres programas que discutiremos a continuación cumplen estecometido.

La referencia de la procedencia de estos programas se dará al final de este trabajo.

I. Breves ideas sobre conglomerados

Una gran parte de los métodos de conglomerados, tienen como fundamento el establecer un coeficiente dedesemejanza entre los individuos.

Un coeficiente de desemejanza no es más que una aplicación del producto cartesiano del conjunto de individuos por sí mismo en los números reales positivos.

Esquemáticamente, podemos expresar la idea anterior como sigue. Sea P el conjunto de individuos; y d el coeficiente de desemejanza; entonces d es una aplicación:

$$d: P \times P \longrightarrow R^{+}$$

El coeficiente de desemejanza, no pretende otra cosa que ser una medida de lo distintos que son dos individuos, respecto de las cualidades que los caracterizan.

Paralelamente al coeficiente de desemajanza, se puede definir un coeficiente de semejanza que en algún sentido sea complementario de d. Así si tenemos

$$\underset{p_1 \in P \quad p_2 \in P}{\text{Max}} \quad d\,(p_1, p_2) = M$$

podemos definir un coeficiente de semejanza como

$$S\,(p_1, p_2) = M - d\,(p_1, p_2)$$

y dos individuos serán tanto más semejantes cuanto mayor sea el valor que asigna S a este par de individuos. Sobre un mismo conjunto de individuos, y un mismo conjunto de cualidades, se pueden establecer un número prácticamente infinito de coeficientes de desemejanza; (aunque muchos de ellos equivalentes en cuanto a la formación de conglomerados); de ahí la gran variedad de conglomerados que se van a poder formar sobre un mismo conjunto, y por tanto es de gran importancia para el investigador el elegir el coeficiente de desemejanza que se adapte en buena medida a su problema.

Para construir conglomerados, los métodos que se siguen a partir de un coeficiente de desemejanza son los siguientes por regla general:

1°) Se fija un número real $\rho > 0$; y se buscan los siguientes conjuntos:

$$A_\rho(p) = \{p_1 \mid d(p, p_1) \leq \rho\} \qquad p \in P$$

quedándonos con aquel conglomerado $A_\rho(p)$ que tuviere más individuos.

Claramente $p \in A_\rho(p)$; pues todo buen coeficiente de desemejanza debe cumplir que $d(p,p)=0$ para todo $p \in P$ También todo coeficiente de desemejanza debe ser simétrico.

2°) A cada $p \in P$, se le asigna $\bar{p}(p)$ de tal forma que

$$d(p,\bar{p}(p)) = \min_{\bar{p} \in P} \{d(p,\bar{p})\}$$

y agrupamos en un solo conglomerado aquel par de individuos que hagan mínimo el coeficiente de desemejanza.

$$d(p_1,\bar{p}(p_1)) = \min_{p \in P} d(p,\bar{p}(p))$$

Para continuar formando conglomerados de más individuos necesitamos definir el coeficiente de desemejanza entre conglomerados; y mediante él se continuaría el 2° método.

Como un ejemplo para aclarar las ideas anteriores, tenemos el siguiente:

Sea P un conjunto de personas sometidas a una cierta encuesta; y sean $V_1, V_2 \ldots, V_k$ las cuestiones sobre las que se pide información. Podemos entonces definir un coeficiente de desemejanza como sigue:

$d(p_1,p_2)$ = número de cuestiones respondidas de distinta forma por las personas p_1 y p_2

Si seguimos el primer método de formar conglomerados y fijamos $\rho = 3$ tenemos que $A_\rho(p)$

es el conjunto de personas que contestan de igual forma que p a todas las cuestiones excepto como máximo a 3.

Si por el contrario seguimos el segundo método; tenemos que $\bar{p}$ (p), es aquella persona que ha respondido mayor número de cuestiones en la misma forma que p;- siendo (p_1, $\bar{p}$ (p_1)) las dos personas que más cues - tiones han respondido en común.

A veces no quedan determinados los conglomera - dos de forma única siguiendo estos métodos, de ahí la - necesidad de dar una regla de elección cuando suceda - una ambiguedad por los métodos 1° y 2°.

ENTRADA DE DATOS PARA LOS POSTERIORES PROGRAMAS

Los datos de entrada se organizarán según la siguiente matríz X:

	X_1	X_2	X_3		X_n
p_1	X_{11}	X_{12}	X_{13}	———	X_{1n}
p_2	X_{21}	X_{22}	X_{23}	———	X_{2n}
p_3	X_{31}	X_{32}	X_{33}	———	X_{3n}
p_k	X_{k1}	X_{k2}	X_{k3}	———	X_{kn}

Donde p_1, p_2,...., p_k son individuos; X_1, X_2...., X_n son variables; y x_{ij} nos representa la respuesta del individuo i a la variable j.

Puesto que los programas fueron escritos para el estudio de ecosistemas en biología; necesitan que los datos - x_{ij} sean numéricos. Como a su vez están en fase de incorporación no se dan todavía las restricciones en dimensiones y características de uso, limitándonos a hacer un breve diseño - de su cometido. Estos programas reciben los siguientes nom - bres: MINFO, M-DISP y CLUSTER.

II. Programa MINFO

Este programa parte de un conjunto de K individuos,- y en cada ciclo del programa reduce en uno el número de conglomerados existentes en el ciclo anterior. Así en el primer ciclo, construye un conglomerado con dos individuos; y por - tanto la entrada del segundo ciclo constará únicamente de K-1 conglomerados.

Si en el comienzo de un ciclo los conglomerados son- A_1, A_2 ....., A_m, en el ciclo siguiente calcula los siguientes números reales, llamando Aij = Ai$\cup$ Aj:

$$I_{Aij} = \sum_{\ell=1}^{n} \sum_{K \in Aij} x_{kl} \ln \left\{ \frac{S_A \; x_{kl}}{\sum_{S=1}^{n} x_{kS} \cdot \sum_{t \in Aij} xtl} \right\}$$

donde

$$S_{Aij} = \sum_{\ell=1}^{n} \sum_{k \in Aij} x_{kl} = S_A$$

Halla a continuación el mínimo de I_{Aij}; y en el siguiente paso, existirán por tanto m-1 conglomerados, ya que si $IAi_1j_1 = \min IAij$; entonces en el ciclo co - rriente se forma un conglomerado uniendo Ai_1 y Aj_1.

Si el mínimo se alcanza para varios pares entonces se elige aleatoriamente el par que ha de unirse. El número de ciclos del programa debe ser K-1; siendo K el número de individuos; y el número de conglomerados después del ciclo i es k-i.

III. Programa MDISP

MDISP es un programa de clasificación que usa - un método de dispersión entre grupos sugerido por Orloci (1967). La distancia estandard usada como una medida de desemejanza.

Sea un subconjunto de P; e indicamos los individuos de A, por $1,2,\ldots,n_A$ respectivamente; entonces la dispersión entre grupos de A es:

$$Q_A = \frac{1}{n_A} \sum_{j=1}^{n_A-1} \sum_{j^1=j+1}^{n_A} d^2_{jj^1}$$

donde djj^1 es una medida de la distancia entre los individuos j y j^1. La distancia estandard, djj^1 es definida como

$$d^2_{jj^1} = \sum_{i=1}^{n} \left(\frac{x_{ji}}{v_j} - \frac{x_{j^1 i}}{v_{j^1}} \right)^2$$

donde $v^2_j = \sum_{i=1}^{n} x^2_{ji}$

Al comienzo del programa, cada individuo es considerado como un grupo separado.

Durante cada ciclo, el programa une dos grupos, mediante la hipótesis de que el incremento de la dispersión entre grupos, de los dos grupos unidos sea menor que si se unieran otros dos grupos cualesquiera.

Los pares reunidos pasan a formar un sólo grupo en el nuevo ciclo. El programa continúa uniendo grupos hasta que todos los individuos están unidos en un solo grupo. Al unirse dos grupos, el nuevo grupo formado se le indica con el menor índice de los dos formados.

IV. Programa CLUSTER

CLUSTER es un programa de clasificación que da al usuario la opción de o bien emplear un método entre pares de grupos pesado o no pesado. El usuario también tiene una elección entre dos medidas de desemejanza. La distancia estandard (definida en la sección precedente, programa MDISP), y el coeficiente de correlación inter-individuos. En orden a definir la matríz de correlación interindividuos, Q, una nueva matríz G, es computada centrando y estandarizando la matríz de datos X.

$$g_{ij} = \frac{X_{ji} - \frac{1}{n} \sum_{i=1}^{n} X_{ji}}{\sqrt{\sum_{i=1}^{n} (X_{ji} - \frac{1}{n} \sum_{i=1}^{n} X_{ji})^2}}$$

donde n es el número de variables o columnas de la matríz X.

Entonces $Q = G^T G$

donde qjj' = qj'j es el coeficiente de correlación entre los individuos j y j'; y el superíndice T indica la traspuesta de la matríz G.

Al comienzo del programa cada individuo es considerado como un grupo separado. Durante cada ciclo de clustering, el programa une el par de grupos que son más similares. Si hay más de un par que cumplen el criterio anterior, el programa deve elegir uno de estos pares como conglomerado. El par conglomerado llega a ser un nuevo grupo, que se identifica con el más pequeño número de identificación del par. Si la opción del par de grupo pesado es usada, entonces la posición del nuevo grupo se toma como el punto medio de la línea que une los dos grupos unidos. Si el grupo localizado como $\bar{r}a$ está compuesto de na individuos, entonces la posición del grupo c que resulta de la unión de los dos grupos a y b es

$$r_c = \frac{\bar{r}_a + \bar{r}_b}{2}$$

para el método del par de grupos no pesado, y

$$r_c = \frac{n_a \bar{r}_a + n_b \bar{r}_b}{n_a + n_b}$$

para el método del par de grupos pesado. El programa con-

tinúa uniendo grupos hasta que todos los individuos están en un grupo .

REFERENCIAS

Goldstein R.A. Grigal E.F.- " Computer programs for the coordination and classification of Ecosystems ". Enero-1972.

Miguel Sánchez García
Biblioteca de programas