Analisis de conglomerados

El objetivo es estudiar técnicas para determinar grupos o subconjuntos de individuos en una muestra de datos. Para lo anterior se puede usar un criterio que determine cuando algunos de los individuos en la muestra son "similares" ó bien cuando no lo son. Desde un punto de vista práctico, la situación a la cual se debería llegar, es cuando los subconjuntos ó conglomerados son lo más homogéneos que se pueda (dos individuos en un grupo se parecen), mientras que las diferencias entre dos grupos diferentes, son lo más grandes que se pueda.

Hay dos pasos fundamentales para hacer análisis de conglomerados

1 Selección de una "medida de proximidad", una función de dos argumentos que permita determinar cuando estos son "parecidos o cercanos".
2 Selección de un algoritmo de agrupamiento

Este algoritmo sienta las bases ó los pasos a seguir, para que usando la medida de proximidad, se asignen individuos a los grupos.

Para una matriz de datos $\mathcal{X}$ con n renglones (individuos) y p columnas (variables), la proximidad entre los individuos se puede describir usando una matriz $\mathbf{D}$ de dimensiones $n \times n$

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix}$$

la entrada i,j contiene el valor de la medida de similaridad ó proximidad (podría ser medida de disimilaridad[1], por ejemplo distancias) entre los individuos i y j $i,j = 1,2,\ldots,n$.

Distancia y similaridad son conceptos duales, ya que

(1) Una distancia mide disimilaridad, ya que entre mayor sea la distancia entre dos individuos, estos serían menos similares.

si d_{ij} es una distancia, entonces $d'_{ij} = \max_{i,j \in A}\{d_{ij}\} - d_{ij}$ es una medida de proximidad ($A = \{1,2,\dots,n\}$).

La elección de una medida de proximidad, depende de la naturaleza de los datos. Por ejemplo, para datos que provienen de variables binarias, conviene usar medidas de similaridad, en general lo anterior funciona cuando las escalas de medición de las variables son Nominales. Cuando la escala de medicion de las variables es continua, en general D es una matriz de distancias.

SIMILARIDAD ENTRE INDIVIDUOS CON COMPONENTES BINARIAS

Si x_{G_i} y x_{G_j} son observaciones $x'_{G_i} = (x_{i1},\dots,x_{ip})$ $x'_{G_j} = (x_{j1},\dots,x_{jp})$, donde $x_{ik}, x_{jk} \in \{0,1\}$, $\forall k=1,\dots,p$.

Entonces pueden suceder cuatro casos:

$$x_{ik} = x_{jk} = 1,$$
$$x_{ik} = 0 \; ; \; x_{jk} = 1,$$
$$x_{ik} = 1 \; ; \; x_{jk} = 0,$$
$$x_{ik} = x_{jk} = 0.$$

Se definen entonces a_1, a_2, a_3 y a_4 como

$$a_1 = \sum_{k=1}^{p} \mathbb{1}(x_{ik} = x_{jk} = 1),$$

$$a_2 = \sum_{k=1}^{p} \mathbb{1}(x_{ik} = 0; x_{jk} = 1),$$

$$a_3 = \sum_{k=1}^{p} \mathbb{1}(x_{ik} = 1; x_{jk} = 0),$$

$$a_4 = \sum_{k=1}^{p} \mathbb{1}(x_{ik} = x_{jk} = 0)$$

Cada a_l es función de (x_{Gi}, x_{Gj}) ; $l = 1,2,3,4$.

Se puede definir una familia paramétrica de medidas de proximidad como

$$d_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)},$$

donde los parámetros δ y λ son pesos. La siguiente tabla muestra a varios elementos de esta familia paramétrica, dependiendo de los valores de δ y λ

Nombre	δ	λ	Definición
Jaccard	0	1	$\frac{a_1}{a_1+a_2+a_3}$
Simple matching	1	1	$\frac{a_1+a_4}{p}$
Kulczynski	—	—	$\frac{a_1}{a_2+a_3}$

Nombre	δ	λ	Definición
Tanimoto	1	2	$\frac{a_1 + a_4}{a_1 + 2(a_2 + a_3) + a_4}$
Russel and Rao	—	—	$\frac{a_1}{p}$
Dice	0	$\frac{1}{2}$	$\frac{2a_1}{2a_1 + (a_2 + a_3)}$

Estas medidas tienen diferentes formas de ponderar discrepancias, así como coincidencias positivas (presencia de caracteres comúnes) ó coincidencias negativas (ausencia de caracteres comúnes).

En el capítulo 3 del libro de Everitt et. al. (2011)[1] "Cluster Analysis", Wiley, se discute con mayor profundidad sobre el tipo de datos binarios y las circunstancias o contextos en los cuales, algunas de estas medidas resultan adecuadas.

(1) Everitt, B.S., Landau, S., Leese, M., y Stahl, D. (2011) "Cluster Analysis", Wiley