Capítulo 7

Análisis Cluster

CLASIFICACIÓN

- Asignar objetos en su lugar correspondiente dentro de un conjunto de categorías establecidas o no.

PROBLEMA

- Dado un conjunto de m objetos (animales, plantas, minerales...), cada uno de los cuales viene descrito por un conjunto de p características o variables, deducir una división útil en un número de clases. Se han de determinar tanto el número de clases como las propiedades de dichas clases.

SOLUCIÓN

- Partición de los m objetos en un conjunto de grupos donde un objeto pertenezca a un grupo sólo y el conjunto de dichos grupos contenga a todos los objetos.

PLANTEAMIENTO DEL PROBLEMA

PUNTO DE PARTIDA

X es una muestra de m individuos sobre los que se miden p variables

$$\Downarrow$$

X es un conjunto de valores numéricos que se pueden ordenar en una matriz

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \dots & x_{mp} \end{pmatrix}$$

- x_{11} : El primer elemento de la matriz, x_{11} , es el valor que presenta el primer individuo en la primera variable.
- x_{12} : El elemento x_{12} corresponde al valor que presenta el primer individuo en la segunda variable.
- x_{1j} : Los valores de la primera fila son los valores que presenta el primer individuo para cada una de las variables.
- x_{2j} : Los valores de la segunda fila se refieren al segundo individuo y así, cada fila referida a uno de los individuos que se estudian.
- $x_{i1}, x_{i2}, \cdots$ Cada columna contiene los valores que toman todos los individuos para cada variable que se estudia.

EJEMPLO

Se realiza un estudio sobre 10 flores distintas a las cuales se les miden 5 variables. Tras hacer este estudio y obtener los valores de las variables, se ordenan los datos en una matriz:

$$X = \begin{pmatrix} 12 & 4 & 7 & 3 & 9 \\ 22 & 3 & 3 & 12 & 7 \\ 5 & 1 & 5 & 7 & 12 \\ 8 & 12 & 10 & 5 & 7 \\ 5 & 4 & 15 & 12 & 13 \\ 6 & 9 & 7 & 1 & 3 \\ 9 & 13 & 9 & 6 & 7 \\ 9 & 12 & 11 & 5 & 7 \\ 10 & 3 & 6 & 2 & 8 \\ 11 & 5 & 8 & 4 & 10 \end{pmatrix}$$

¿Qué significado tiene la segunda fila de la matriz?

Se trata de los valores que toman todas las variables medidas en la segunda flor que se está estudiando.

¿Qué significado tiene la cuarta columna de la matriz?

Se trata de los valores que toma la cuarta variable en todas las flores que se estudian.

OBJETIVO

Encontrar una partición de los m individuos en c grupos de forma que cada individuo pertenezca a un grupo y solamente a uno.

SOLUCIÓN

¿Por qué no hacer todas las particiones posibles y, de entre las resultantes, elegir la más atrayente?

Este método lleva a construir un número de particiones muy elevado.
Ejemplo: 4 Objetos A, B, C y D

Num. de grupos	Clasificaciones
1 Grupo	$ABCD$
2 Grupos	$A-BCD$ $B-ACD$ $C-ABD$ $D-ABC$ $AB-CD$ $AC-BD$ $AD-BC$
3 Grupos	$A-B-CD$ $A-C-BD$ $A-D-BC$ $B-C-AD$ $B-D-AC$ $C-D-AB$
4 Grupos	$A-B-C-D$

Total: 15 posibles clasificaciones.

ANÁLISIS CLUSTER

Es un procedimiento estadístico que parte de un conjunto de datos que contiene información sobre una muestra de entidades e intenta reorganizarlas en grupos relativamente homogéneos a los que se llama clusters.

ETAPAS DEL ANÁLISIS CLUSTER

1) **Elección de las variables**
2) **Elección de la medida de asociación**
3) **Elección de la técnica Cluster**
4) **Validación de los resultados**

(1) ELECCIÓN DE LAS VARIABLES

Dependiendo del problema que se plantee

- **Variables:**
  - ◊ **Variables cualitativas**
    - ★ **Ordinales** (ej: nivel de estudios)
    - ★ **Nominales** (ej: nacionalidad)
  - ◊ **Variables cuantitativas**
    - ★ **Variables discretas** (ej: número de hermanos)
    - ★ **Variables continuas** (ej: peso)

ANÁLISIS CLUSTER POR VARIABLES O POR INDIVIDUOS

- Si se pretende agrupar a los individuos en grupos se ha de realizar un análisis cluster de los individuos

- Si se pretende agrupar las variables más parecidas se debe realizar un análisis cluster de las variables, para ello basta considerar la matriz de datos inicial X traspuesta[1] (girada).

(2) ELECCIÓN DE LA MEDIDA DE ASOCIACIÓN

Para poder unir variables o individuos es necesario tener algunas medidas numéricas que caractericen las relaciones entre las variables o los individuos. Cada medida refleja asociación en un sentido particular y es necesario elegir una medida apropiada para el problema concreto que se esté tratando. La medida de asociación puede ser una **distancia** o una **similaridad**.

- Cuando se elige una distancia como medida de asociación (por ejemplo la distancia euclídea) los grupos formados contendrán individuos parecidos de forma que la distancia entre ellos ha de ser pequeña.

- Cuando se elige una medida de similaridad (por ejemplo el coeficiente de correlación) los grupos formados contendrán individuos con una similaridad alta entre ellos.

[1] Trasponer una matriz es cambiar las filas por las columnas. (Ver Apéndice)

DISTANCIA MÉTRICA

Una función $d : U \times U \longrightarrow R$ se llama una **distancia métrica** si $\forall x, y \in U$

a) $d(x, y) \geq 0$

b) $d(x, y) = 0 \Leftrightarrow \ x = y$

c) $d(x, y) = d(y, x)$

d) $d(x, z) \leq d(x, y) + d(y, z)\ ,\ \forall z \in U$

SIMILARIDAD

Una función $s : U \times U \longrightarrow R$ se llama **similaridad** si $\forall x, y \in U$

a) $s(x, y) \leq s_0$

b) $s(x, x) = s_0$

c) $s(x, y) = s(y, x)$

donde s_0 es un número real finito arbitrario.

SIMILARIDAD MÉTRICA

Una función s, verificando las condiciones de la definición anterior, se llama **similaridad métrica** si, además, verifica:

a) $s(x, y) = s_0 \Longrightarrow \ x = y$

b) $|s(x, y) + s(y, z)|s(x, z) \geq s(x, y)s(y, z)\ ,\ \forall z \in U$

Dependiendo del tipo de análisis (por variables o por individuos) que se realiza, existen distintas medidas de asociación aunque, técnicamente, todas las medidas pueden utilizarse en ambos casos.

MEDIDAS DE ASOCIACIÓN PARA VARIABLES

1. **Coseno del ángulo de dos vectores** (invarianza, salvo signo, frente a homotecias)

2. **Coeficiente de correlación** (invarianza frente a traslaciones y salvo signo frente a homotecias)

3. **Medidas para datos dicotómicos**

X_i/X_j	1	0	Totales
1	a	b	$a+b$
0	c	d	$c+d$
Totales	$a+c$	$b+d$	$m=a+b+c+d$

- **Ochiai** $\rightarrowtail \dfrac{a}{[(a+b)(a+c)]^{\frac{1}{2}}}$

- **Medida ϕ** $\rightarrowtail \dfrac{ad-bc}{[(a+b)(c+d)(a+c)(b+d)]^{\frac{1}{2}}}$

- **Medida de Russell y Rao** $\rightarrowtail \dfrac{a}{a+b+c+d} = \dfrac{a}{m}$

- **Medida de Parejas simples** $\rightarrowtail \dfrac{a+d}{a+b+c+d} = \dfrac{a+d}{m}$

- **Medida de Jaccard** $\rightarrowtail \dfrac{a}{a+b+c}$

- **Medida de Dice** $\rightarrowtail \dfrac{2a}{2a+b+c}$

- **Medida de Rogers-Tanimoto** $\rightarrowtail \dfrac{a+d}{a+d+2(b+c)}$

MEDIDAS DE ASOCIACIÓN PARA INDIVIDUOS

■ **Distancia euclídea:**

$$d_2(x_i, x_j) = \|x_i - x_j\|_2 = \sqrt{\sum_{l=1}^{p}(x_{il} - x_{jl})^2}$$

■ **Distancia de Minkowski** (Invariante frente a traslaciones)

$$d_q(x_i, x_j) = \|x_i - x_j\|_q = \left(\sum_{l=1}^{p} |x_{il} - x_{jl}|^q\right)^{\frac{1}{q}} \quad ; \quad q \geq 1$$

★ **Distancia d_1 o ciudad (City Block o bloque)** $(q = 1)$

$$d_1(x_i, x_j) = \|x_i - x_j\|_1 = \sum_{l=1}^{p} |x_{il} - x_{jl}|$$

★ **Distancia euclídea** $(q = 2\ )$

★ **Distancia de Tchebychev o del máximo** $(\ q = \infty\ )$

$$d_\infty e(x_i, x_j) = \text{máx}\ \{l = 1, \ldots, p\}\ |x_{li} - x_{lj}|$$

■ **Distancia de Mahalanobis**

$$D_S(x_i, x_j) = \sqrt{(x_i - x_j)' S^{-1} (x_i - x_j)}$$

■ **Distancia χ^2**

$$\chi^2 = m\left[\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_{ij}^2}{m_{i.} m_{.j}} - 1\right]$$

(3) ELECCIÓN DE LA TÉCNICA CLUSTER

MÉTODOS JERÁRQUICOS

OBJETIVO

Agrupar cluster para formar uno nuevo o separar alguno ya existente para dar origen a otros dos de forma que se maximice una medida de similaridad o se minimice alguna distancia.

CLASIFICACIÓN

- **Asociativos o Aglomerativos**: Se parte de tantos grupos como individuos hay en el estudio y se van agrupando hasta llegar a tener todos los casos en un mismo grupo.

- **Disociativos**: Se parte de un solo grupo que contiene todos los casos y a través de sucesivas divisiones se forman grupos cada vez mas pequeños.

Los métodos jerárquicos permiten construir un árbol de clasificación o **dendograma**

MÉTODOS NO-JERÁRQUICOS

Están diseñados para la clasificación de individuos (no de variables) en K grupos. El procedimiento es elegir una partición de los individuos en K grupos e intercambiar los miembros de los clusters para tener una partición mejor.

MÉTODOS DE ANÁLISIS CLUSTER

- **Jerárquicos**
  - **Aglomerativos**
    - Linkage Simple
    - Linkage Completo
    - Promedio entre grupos
    - Método del Centroide
    - Método de la Mediana
    - Método de Ward
  - **Disociativos**
    - Linkage Simple
    - Linkage Completo
    - Promedio entre grupos
    - Método del Centroide
    - Método de la Mediana
    - Método de Ward
    - Análisis de Asociación

- **No Jerárquicos**
  - **Reasignación**
    - K − Medias
    - Nubes Dinámicas
  - **Búsqueda de densidad**
    - Aproximación Tipológica
    - ♦ (Análisis Modal)
    - ♦ (Métodos Taxap)
    - ♦ (Método de Fortin)
    - Aproximación Probabilística
    - ♦ (Método de Wolf)
  - Métodos Directos
  - ♦ Block − Clustering
  - Metodos Reductivos
  - ♦ (Análisis Factorial tipo Q)

MÉTODO LINKAGE SIMPLE AGLOMERATIVO

- Una vez se conocen las distancias existentes entre cada dos individuos se observa cuáles son los individuos más próximos en cuanto a esta distancia o similaridad (qué dos individuos tienen menor distancia o mayor similaridad). Estos formarán un grupo que no vuelve a separarse durante el proceso. Se mide la distancia o similaridad entre todos los individuos de nuevo (tomando el grupo ya formado como si de un solo individuo se tratara) de la siguiente forma:

- Cuando se mide la distancia entre el grupo formado y un individuo, se toma la distancia mínima de los individuos del grupo al nuevo individuo.

- Cuando se mide la similitud o similaridad entre el grupo formado y un individuo, se toma la máxima de los individuos del grupo al nuevo individuo.

EJEMPLO

Se tienen las siguientes distancias entre individuos [1]

Distancia	A	B	C	D
A	0			
B	9	0		
C	4	5	0	
D	7	3	11	0

- Distancia mínima:
  - ★ Entre B y D (distancia 3)
  - ★ B y D forman un grupo.
    - ♠ Se miden las distancias de nuevo:

[1] La tabla es simétrica puesto que la distancia entre A y B es la misma que la distancia entre B y A.

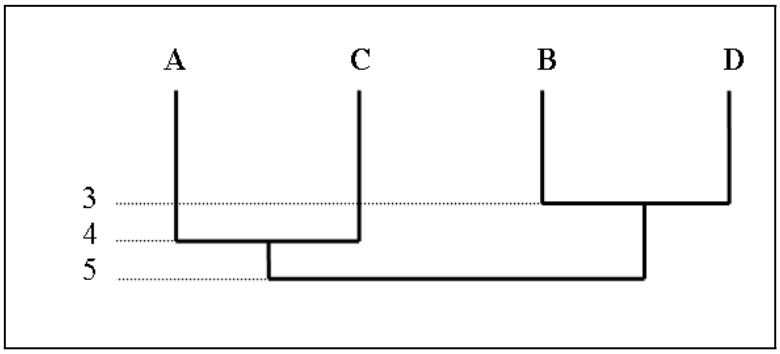
Distancia	A	$B-D$	C
A	0		
$B-D$	7	0	
C	4	5	0

- Distancia mínima:
  - ★ Entre C y A (distancia 4)
  - ★ C y A forman un grupo.
    - ♠ Se miden las distancias de nuevo:

Distancia	$A-C$	$B-D$
$A-C$	0	
$B-D$	5	0

- Último paso:
  - ★ Se unen los grupos $A-C$ y $B-D$ formando el grupo $A-B-C-D$.

El proceso seguido se representa en un árbol de clasificación llamado dendograma



- En la representación se puede ver qué elementos se unen y la distancia a la que lo hacen.
- El número de grupos se puede decidir a posteriori.
- Si se desea clasificar estos elementos en dos grupos la clasificación resultante es: $B-D$ y $A-C$.
- Si se quieren tres grupos, se toma la clasificación en el paso anterior: $B-D$, A y C.

MÉTODO LINKAGE COMPLETO AGLOMERATIVO

- Conocidas las distancias o similaridades existentes entre cada dos individuos se observa cuáles son los individuos más próximos en cuanto a esta distancia o similaridad (qué dos individuos tienen menor distancia o mayor similaridad). Estos formarán un grupo que no vuelve a separarse durante el proceso. Se mide la distancia o similaridad entre todos los individuos de nuevo de la siguiente forma:

- Cuando se mide la distancia entre el grupo formado y un individuo, se toma la distancia máxima de los individuos del grupo al nuevo individuo.

- Cuando se mide la similitud o similaridad entre el grupo formado y un individuo, se toma la mínima de los individuos del grupo al nuevo individuo.

EJEMPLO

Se tienen las siguientes similaridades (ejemplo el coeficiente de correlación entre variables)

Distancia	A	B	C	D	E
A	1				
B	0.39	1			
C	0.75	0.24	1		
D	0.56	0.63	0.42	1	
E	0.81	0.72	0.12	0.93	1

- Similaridad máxima:

  - ★ Entre D y E (similaridad 0.93)
  - ★ D y E forman un grupo.
    - ♠ Se miden las similaridades de nuevo:

Distancia	A	B	C	$D-E$
A	1			
B	0.39	1		
C	0.75	0.24	1	
$D-E$	0.56	0.63	0.12	1

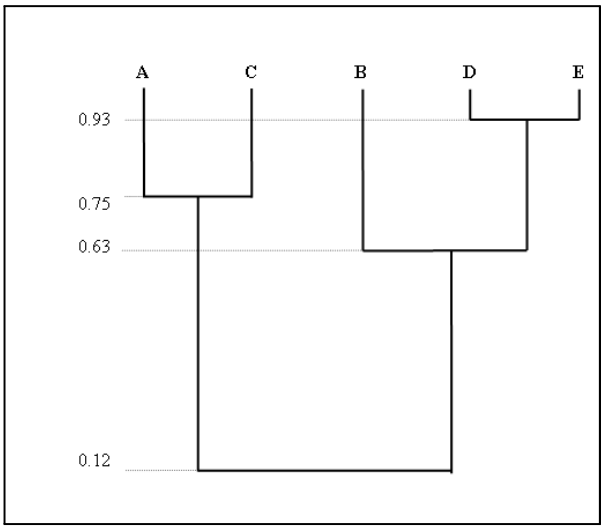
- Similaridad máxima:
  - ★ Entre C y A (similaridad 0.75)
  - ★ C y A forman un grupo.
    - ♠ Se miden las distancias de nuevo:

Distancia	A – C	B	D – E
A – C	1		
B	0.24	1	
D – E	0.12	0.63	1

- Similaridad máxima:
  - ★ Entre B y $D-E$ (similaridad 0.63)
  - ★ B y $D-E$ forman un grupo.
    - ♠ Se miden las distancias de nuevo:

Distancia	$A-C$	$B-D-E$
$A-C$	1	
$B-D-E$	0.12	1

- Último paso:
  - ★ Se unen los grupos $A-C$ y $B-D-E$ formando el grupo $A-B-C-D-E$.



MÉTODO DE LAS K−MEDIAS

- No es necesario medir distancias o similaridades puesto que no es un método jerárquico y por tanto la clasificación en K grupos se hará en un solo paso.

- Se toman los K primeros casos como grupos unitarios y se asignan el resto de casos a los grupos con el centroide más próximo.

  - ★ Se recalcula el centroide de cada grupo después de cada asignación.

  - ★ Tras la asignación de todos los individuos se toman los centroides de los grupos existentes como fijos y se vuelven a asignar los individuos al centroide más próximo.

- Este método puede ser iterado hasta que ningún individuo cambie de grupo en la reasignación. En ese caso se trata del método de las K−medias convergente.

APÉNDICE

Trasponer una matriz es cambiar las filas por las columnas. Ejemplo:

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & x_{16} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & x_{26} \\ x_{31} & x_{32} & x_{33} & x_{34} & x_{35} & x_{36} \\ x_{41} & x_{42} & x_{43} & x_{44} & x_{45} & x_{46} \end{pmatrix}$$

Trasponiendo nos queda un matriz, X' o X^t, donde las filas eran las columnas de la matriz anterior, y las columnas son las filas de la matriz anterior,

$$X' = \begin{pmatrix} x_{11} & x_{21} & x_{31} & x_{41} \\ x_{12} & x_{22} & x_{32} & x_{42} \\ x_{13} & x_{23} & x_{33} & x_{43} \\ x_{14} & x_{24} & x_{34} & x_{44} \\ x_{15} & x_{25} & x_{35} & x_{45} \\ x_{16} & x_{26} & x_{36} & x_{46} \end{pmatrix}$$

Bibliografía utilizada:

★ **R. Gutiérrez, A. González, F. Torres, J.A. Gallardo (1994).** *"Técnicas de Análisis de datos Multivariable. Tratamiento computacional". Universidad de Granada.*

★ **B. Visauta Vinacua (1998).** *"Análisis estadístico con SPSS para Windows, volumen II: Estadísitca multivariante". McGraw Hill .*

♦ **Temporalización:** Dos horas