Capítulo 6

Análisis Cluster

6.1. INTRODUCCIÓN

El Análisis Cluster consiste en agrupar un conjunto de individuos u objetos en grupos homogéneos, en función de los valores observados de p variables. Se parte de una tabla de datos como la siguiente

Unidad	X_1	X_2	$\cdots$	X_p
1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Clasificación de los métodos de agrupación existentes:

- *Métodos jerárquicos:* son algoritmos secuenciales en los que cada etapa del algoritmo consiste en unir grupos, o bien en separar grupos.
  - *Métodos jerárquicos aglomerativos:* En el primer paso cada unidad forma un grupo. En cada etapa del algoritmo se unen dos grupos, y así sucesivamente hasta que se llega a un único grupo con todas las unidades.
  - *Métodos jerárquicos divisivos:* En este caso el algoritmo comienza con un único grupo que contiene a todas las unidades. En cada etapa se divide un grupo en dos, hasta

que se llega a tantos grupos como unidades. Estos métodos no están implementados en SPSS.

- *Métodos partitivos:* En estos métodos, se parte de un número de grupos fijado de antemano, y se van clasificando de forma secuencial los individuos en algún grupo en función de cierto criterio, hasta que no queden unidades y los grupos sean estables.

Vamos a estudiar solamente los métodos jerárquicos aglomerativos más comunes, y el método de las k-medias, que se considera como un método partitivo.

6.2. MÉTODOS JERÁRQUICOS AGLOMERATIVOS

(1) Se comienza con n grupos, cada uno consistente en una unidad, y una matriz $n \times n$ de disimilaridades $\Delta = (\delta_{ij})$ simétrica y con ceros en la diagonal.

(2) Se busca en la matriz de disimilaridades el par de grupos más próximos. Sean U y V los grupos más próximos, y d_{UV} su distancia.

(3) Se unen los grupos U y V, y se etiqueta el nuevo grupo como (UV). Se actualiza la matriz de disimilaridades, de la siguiente forma:

   (a) se borran las filas y columnas correspondientes a los grupos U y V.

   (b) se añade una fila y una columna con las distancias entre el grupo (UV) y los grupos restantes.

(4) Repetir los pasos (2) y (3) $n - 1$ veces, y al final todas las unidades estarán incluidas en un único grupo. Se deben guardar

las etiquetas de los grupos que se han unido, así como las disimilaridades con las que se unieron.

Para la realización del segundo paso, es necesario la definición de una medida de disimilaridad entre grupos. La medida de disimilaridad que se defina determina el tipo de método aglomerativo. Las disimilaridades más comunes son las siguientes:

- *Vecino más próximo (nearest neighbor o simple link):* En este método, la disimilaridad entre dos grupos es la disimilaridad entre sus miembros más próximos, es decir, si U y V son dos grupos, entonces

$$d_{UV} = \text{mín}\{d_{ij} : i \in U,\ j \in V\}\,.$$

- *Vecino más lejano (complete linkage):* la disimilaridad entre dos grupos es la disimilaridad entre sus miembros más alejados, es decir,

$$d_{UV} = \text{máx}\{d_{ij} : i \in U,\ j \in V\}\,.$$

- *Promedio entre grupos (average linkage):* la disimilaridad entre dos grupos es la disimilaridad media entre todos los pares de unidades, donde un elemento del par es de un grupo, y el otro elemento pertenece al otro grupo, es decir, si n_u es el número de unidades en U, y n_v es el número de unidades en V, entonces

$$d_{UV} = \frac{1}{n_u n_v} \sum_{i \in U} \sum_{j \in V} d_{ij}\,.$$

- *Incremento en suma de cuadrados (método de Ward):* La disimilaridad entre los grupos U y V se calcula de la manera siguiente: supongamos que los valores observados de las p variables en los individuos de ambos grupos son

	Unidad	X_1	X_2	$\cdots$	X_p
U	1	x_{u11}	x_{u12}	$\cdots$	x_{u1p}
	2	x_{u21}	x_{u22}	$\cdots$	x_{u2p}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	n_u	x_{un_u1}	x_{un_u2}	$\cdots$	x_{un_up}
		$\overline{x}_{u1}$	$\overline{x}_{u2}$	$\cdots$	$\overline{x}_{up}$
V	1	x_{v11}	x_{v12}	$\cdots$	x_{v1p}
	2	x_{v21}	x_{v22}	$\cdots$	x_{v2p}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	n_v	x_{vn_v1}	x_{vn_v2}	$\cdots$	x_{vn_vp}
		$\overline{x}_{v1}$	$\overline{x}_{v2}$	$\cdots$	$\overline{x}_{vp}$

Sea $(\overline{x}_{u1}, \ldots, \overline{x}_{up})'$ el vector de medias del grupo U, y $(\overline{x}_{v1}, \ldots, \overline{x}_{vp})'$ el del grupo V. Se define la suma de cuadrados dentro de los grupos U y V de la forma

$$SCD_u = \sum_{j=1}^{p} \sum_{m \in U} (x_{ujm} - \overline{x}_{uj})^2, \quad SCD_v = \sum_{j=1}^{p} \sum_{m \in V} (x_{vjm} - \overline{x}_{vj})^2 .$$

Si se uniesen ambos grupos en otro llamado W, con vector de medias $(\overline{x}_{w1}, \ldots, \overline{x}_{wp})'$, la suma de cuadrados sería

$$SCD_w = \sum_{j=1}^{p} \sum_{m \in U \cup V} (x_{wjm} - \overline{x}_{wj})^2 .$$

La disimilaridad entre U y V es el incremento producido en la suma de cuadrados al unir ambos grupos, es decir,

$$d_{UV} = SCD_w - (SCD_u + SCD_v) .$$

Es decir, el método de Ward consiste en unir los grupos con menor incremento en suma de cuadrados, con lo que se unen los grupos más homogéneos.

Ejemplo 6.1 (Método del vecino más próximo)
Se han medido ciertas variables cuantitativas a 10 empresas. Utilizando la distancia euclídea al cuadrado, la matriz de distancias es

	E1	E2	E3	E4	E5	E6	E7	E8	E9
E2	1.55								
E3	7.34	4.92							
E4	6.05	7.03	9.14						
E5	15.80	17.84	29.34	8.45					
E6	18.54	16.78	13.78	6.57	12.83				
E7	11.69	10.42	15.63	13.85	14.20	13.29			
E8	10.77	10.17	10.90	11.18	27.32	20.07	13.91		
E9	7.27	7.42	17.63	4.06	3.71	13.42	12.27	15.95	
E10	15.77	13.82	16.55	5.98	11.66	8.17	11.56	8.05	7.36

Se desea aplicar el método del vecino más próximo para agrupar las empresas en función de sus analogías.

6.3. MÉTODO DE LAS k-MEDIAS

El algoritmo de las k medias consta de las siguientes etapas:

(1) Se dividen las unidades en k grupos iniciales, calculando los centroides (medias) de cada grupo.

(2) Se selecciona una unidad, y se asigna al grupo más cercano (normalmente se utiliza la distancia euclídea). Se recalcula el centroide para el grupo que recibe la unidad, y para el grupo que la pierde.

(3) Se repite el Paso (2) hasta que no hay más reasignaciones.

Si se utiliza la distancia euclídea y las variables tienen unidades de medida o rangos dispares, entonces es necesario estandarizarlas.

Ejemplo 6.2 Supongamos que se han medido dos variables X_1 y X_2 para cuatro unidades A, B, C, y D. Los datos aparecen en la tabla siguiente

Unidad	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Se desea dividir estas unidades en $K = 2$ grupos, de manera que las unidades dentro del mismo grupo sean más parecidas que las unidades de grupos distintos.

(1) Comenzamos dividiendo arbitrariamente las unidades en dos grupos (AB) y (CD), y calculamos las coordenadas $(\overline{x}_1, \overline{x}_2)$ del centroide (media) de cada grupo.

Grupo	$\overline{x}_1$	$\overline{x}_2$
(AB)	$\frac{5+(-1)}{2} = 2$	$\frac{3+1}{2} = 2$
(CD)	$\frac{1+(-3)}{2} = -1$	$\frac{(-2)+(-2)}{2} = -2$

(2) Seleccionamos una unidad cualquiera, por ejemplo A, y calculamos su distancia (euclídea al cuadrado) a los centroides de cada grupo.

$$d(A,(AB)) = (5-2)^2 + (3-2)^2 = 10\,,$$
$$d(A,(CD)) = (5+1)^2 + (3+2)^2 = 61\,.$$

Como A está más cerca del grupo (AB) que de (CD), no se reasigna. Los centroides no cambian.

(2) Seleccionamos ahora la unidad B. Calculamos su distancia a los centroides de ambos grupos.

$$d(B,(AB)) = (-1-2)^2 + (1-2)^2 = 10\,,$$
$$d(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9\,.$$

Como B está más cerca de (CD) que de (AB), entonces B se reasigna a (CD), obteniéndose el cluster (BCD). Los centroides son ahora

Grupo	$\overline{x}_1$	$\overline{x}_2$
A	5	3
(BCD)	-1	-1

(2) Dado que han cambiado los centroides, se comprueba de nuevo si cada unidad puede ser reasignada. Las distancias euclídeas al cuadrado de las unidades a los dos grupos aparecen en la tabla siguiente

	Unidad			
Grupo	A	B	C	D
A	0	40	41	89
(BCD)	52	4	5	5

Podemos observar que cada unidad está asignada al grupo cuyo centroide es el más cercano. Por tanto, el proceso se para y los dos grupos finales son A y (BCD).