Capítulo 3

ANÁLISIS CLUSTER

3.1 ¿Qué es el Análisis Cluster?

El Análisis Cluster (Análisis de Conglomerados) es un conjunto de técnicas que se utilizan para clasificar los objetos o casos en grupos relativamente homogéneos llamados conglomerados (clusters). Los objetos en cada grupo tienden a ser similares entre sí (alta homogeneidad interna, dentro del cluster) y diferentes a los objetos de los otros grupos (alta heterogeneidad externa, ente clusters) con respecto a algún criterio de selección predeterminado.

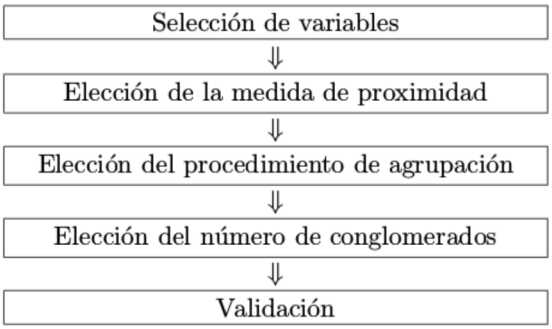
Aunque esta técnica tiene su origen en Biología, ciencia en la que el problema de clasificación de las especies adquiere gran relieve, después ha sido aplicada con éxito a muchos otros campos, incluido el campo de la Sociología, pero especialmente en Medicina, Psiquiatría, Arqueología, Antropología, etc. En todos estos ejemplos, el analista trata de encontrar una estructura **natural** a través de las observaciones basándose en un perfil multivariado.

Este análisis se conoce también como análisis de *clasificación automática o no supervisada*, *reconocimiento de patrones sin supervisión* o *taxonomía numérica*.

3.2 Etapas en el Análisis Cluster

La matriz inicial de datos es una matriz $N \times p$, siendo N el número de individuos y p el número de variables que se han tenido en cuenta, que se transforma en una tabla de distancias $N \times N$ que recoge las disimilaridades entre todos los individuos.

En el siguiente cuadro se reflejan las etapas fundamentales del análisis:



El primer paso consiste en formular el problema de agrupación al definir las variables en las que se basa ésta. Después, debe seleccionarse una medida de proximidad para determinar cómo de similares o diferentes son los objetos que se agrupan. Se han desarrollado varios procedimientos de agrupación y el investigador debe seleccionar uno apropiado para el problema que se maneja. La decisión del número de conglomerados requiere del criterio del investigador que deberá evaluar la consistencia de los grupos que se han formado y determinar las características de cada uno de ellos. Por último, es conveniente que los conglomerados derivados se interpreten en términos de las variables utilizadas para formarlos.

3.3 Formulación del problema

Un aspecto fundamental de la formulación del problema del Análisis Cluster es la selección de las variables en las que se basa la agrupación. La inclusión de una o más variables irrelevantes puede distorsionar una solución de agrupación que de otra forma podría ser útil. Las variables deben seleccionarse con base en la investigación que se va a llevar a cabo.

3.4 Selección de una medida de distancia o similaridad

Se conoce con el nombre genérico de proximidades a un conjunto de medidas que nos indican si dos o más elementos son cercanos o lejanos entre sí. Entre estas medidas podemos encontrar distancias o similaridades. Mientras las similaridades tienen el valor máximo si los elementos son cercanos y disminuye si son lejanos, las distancias alcanzan valores mínimos para casos cercanos y valores grandes para casos lejanos.

Existe una gran variedad de estas medidas y la elección de una concreta debe hacerse teniendo en cuenta la naturaleza de las variables y los objetivos de la agrupación. El uso de distintas medidas puede llevar a diversos resultados de conglomerado. Por consiguiente, se recomienda utilizar medidas diferentes y comparar los resultados.

Entre las medidas de proximidad más comunes se pueden citar:

Variables cuantitativas. Se consideran dos individuos x e y con valores observados en p variables.

Ind. \ Var.	1	...	i	...	p
x	x_1	...	x_i	...	x_p
y	y_1	...	y_i	...	y_p

Distancia euclídea,

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2}$$

Distancia City block o Manhattan,

$$d(x,y) = |x_1 - y_1| + \cdots + |x_p - y_p|$$

Distancia de Mahalanobis,

$$d(x,y) = (\vec{x} - \vec{y})' \Sigma^{-1} (\vec{x} - \vec{y})$$

donde Σ es la matriz de varianzas-covarianzas de las variables consideradas.

Si las variables se miden en unidades muy diferentes, el investigador debe plantearse una última cuestión... ¿deben tipificarse los datos antes de calcular las distancias?. Para poder responder a esta pregunta de forma adecuada, el investigador debe tener en cuenta que la mayoría de las medidas de distancia son bastante sensibles a las diferencias de escalas o de magnitudes hechas entre las variables. En general, las variables con una gran dispersión (valores grandes de sus desviaciones típicas) tienen más impacto en el valor final de la distancia.

Aun cuando la tipificación puede eliminar la influencia de la unidad de medición, también es probable que reduzca las diferencias entre los grupos en las variables que pueden discriminar mejor los grupos o conglomerados.

Variables de frecuencia. Se consideran dos variables cualitativas x e y de las que se tienen las frecuencias observadas en p modalidades.

Var. \ Modal.	1	...	i	...	p	
x	x_1	...	x_i	...	x_p	n_x
y	y_1	...	y_i	...	y_p	n_y

Distancia Chi-cuadrado,

$$d(x,y) = \sqrt{\sum_{i=1}^{p} \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum_{i=1}^{p} \frac{(y_i - E(y_i))^2}{E(y_i)}}$$

Distancia Phi-cuadrado,

$$d(x,y) = \sqrt{\frac{\sum_{i=1}^{p} \frac{(x_i - E(x_i))^2}{E(x_i)} + \sum_{i=1}^{p} \frac{(y_i - E(y_i))^2}{E(y_i)}}{n_x + n_y}}$$

donde $E(\cdot)$ representa la frecuencia esperada.

Variables dicotómicas. A partir de los valores de presencia/ausencia de dos individuos x e y en p variables, se construye una tabla de asociación de la siguiente forma:

	Individuo x	
Individuo y	Presencia	Ausencia
Presencia	a	b
Ausencia	c	d

Distancia euclídea, $d(x,y) = \sqrt{b+c}$

Distancia de Lance-Williams, $d(x,y) = \dfrac{b+c}{2a+b+c}$

Similaridad de Jaccard, $S(x,y) = \dfrac{a}{a+b+c}$

Similaridad de Russel-Rao, $S(x,y) = \dfrac{a}{a+b+c+d}$

Similaridad concordancia simple, $S(x,y) = \dfrac{a+d}{a+b+c+d}$

3.5 Selección del procedimiento de agrupación

Existen dos grandes tipos de Análisis Cluster: los jerárquicos y los no jerárquicos.

Los métodos **jerárquicos** forman los grupos en pasos sucesivos, de forma que los clusters de niveles más bajos van siendo englobados en otros de niveles superiores. Se pueden analizar en cada paso las distancias entre los grupos formados.

Los métodos **no jerárquicos** realizan una sola partición de los casos iniciales en un número de grupos fijado de antemano, sin que unos dependan de otros.

Clusters jerárquicos

Los algoritmos existentes funcionan de manera que los elementos son sucesivamente asignados a los grupos, pero la asignación es irrevocable, es decir, una vez hecha, no se cuestiona nunca más. Los algoritmos son de dos tipos: de aglomeración o de división. El conglomerado por aglomeración empieza con cada objeto en un grupo separado. Los conglomerados se forman al agrupar los objetos en conjuntos cada vez más grandes. Este proceso continúa hasta que todos los objetos formen parte de un solo grupo. El conglomerado por división comienza con todos los objetos agrupados en un solo conjunto. Los conglomerados se dividen hasta que cada objeto sea un grupo independiente. La representación gráfica de estas etapas de formación de grupos, a modo de árbol se denomina **dendograma**.

Evidentemente la decisión de todas estas agrupaciones ha de tomarse en función de la similitud proporcionada por el conjunto de variables estudiadas, ya que en cada nivel se unen los clusters más cercanos. Se presentan a continuación los principales métodos de unión entre grupos.

- Método de **enlace sencillo** o regla del vecino más próximo. Considera como distancia entre dos grupos la separación que existe entre los individuos más próximos de uno y otro grupo. Si bien no es adecuado para la obtención de grupos compactos, resulta de utilidad para clusters irregulares.
- Método del **enlace completo** o regla del vecino más lejano. La distancia entre dos grupos es la de los individuos más separados de ambos grupos. Presenta una excesiva tendencia a producir grupos de igual diámetro y se ve muy distorsionado por valores atípicos.
- Método del **enlace promedio entre grupos**. En este caso, la distancia entre dos conglomerados se define como el promedio de las distancias de cada caso de un grupo con todos los casos del otro grupo. Esta operación se realiza con todos los grupos, asociando en el paso siguiente a los dos grupos con valor promedio menor.
- Método del **enlace promedio dentro de grupos**. Con este método se agrupan de dos en dos

los grupos previos, calculando a continuación el promedio de las distancias de todos los miembros del grupo. Así se agrupan en ese paso, de forma definitiva, los dos grupos cuya unión tenga el promedio menor.

- Método **centroide**. Calcula la distancia entre dos grupos como la distancia entre sus centroides (medias para todas las variables). Cada vez que se agrupan los objetos, se calcula un centroide nuevo como el promedio de los centroides de los grupos iniciales, ponderado por el tamaño de los mismos.
- **Método de Ward**. Para cada conglomerado, suma las distancias euclídeas al cuadrado de cada individuo a la media de su grupo y une aquellos grupos que hacen esta suma menor.

Clusters no jerárquicos

En esta clase de procedimientos es necesario que el investigador fije de antemano el número de clusters en los que quiere agrupar sus datos.

Parece lógico que una clasificación correcta debe ser aquella en la que la dispersión dentro de cada grupo sea la menor posible. Esta condición se denomina *criterio de la varianza* y lleva a seleccionar una configuración cuando la suma de varianzas dentro de cada grupo sea mínima. Existen varios algoritmos de clasificación no jerárquica, basados en minimizar progresivamente esta varianza, el más utilizado es el **algoritmo de las k-medias**. Las etapas básicas del análisis son las siguientes:

1. Se seleccionan, generalmente al azar, los centros provisionales de cada clase.
2. Se asigna cada uno de los N individuos de la muestra a la clase más cercana. Posteriormente se recalculan los centros de cada grupo.
3. Se repite el proceso de asignación de los individuos a las clases hasta que se verifica algún criterio de parada; por ejemplo la clasificación de los individuos apenas ha sufrido variaciones entre dos iteraciones consecutivas.

Los procedimientos no jerárquicos tienen dos desventajas importantes frente a los jerárquicos que son que el número de grupos debe especificarse previamente y que la selección de los centros de grupo es arbitraria. Además, los resultados del conglomerado pueden depender de la forma en que se seleccionan los centros. Muchos programas no jerárquicos eligen los primeros k (k= número de grupos) casos sin valores faltantes como los centros de grupo iniciales. De manera que, los resultados del conglomerado pueden depender del orden de las observaciones en los datos. No obstante, el conglomerado no jerárquico es más rápido (coste computacional, velocidad de cálculo) que los métodos jerárquicos y es apropiado cuando el número de objetos u observaciones es alto. Se ha sugerido que los métodos jerárquicos y no jerárquicos se utilicen uno después del otro. Primero, una solución de conglomerado inicial se obtiene con el uso de un procedimiento jerárquico, como el enlace promedio o el de Ward. Las cantidades de grupos y centroides de grupo que se obtienen de esta forma se utilizan como entradas para el método de división para la optimización.

La elección de un método de conglomerado y la elección de una medida de distancia están interrelacionadas. Por ejemplo, las distancias euclidianas cuadradas deben utilizarse con los métodos de Ward y centroide. Varios procedimientos no jerárquicos emplean también las distancias euclidianas cuadradas.

3.6 Decisión del número de clusters

Un aspecto importante en el análisis de conglomerados es decidir el número de éstos. A pesar de que no existe ninguna regla general y rápida,

- Las consideraciones teóricas, conceptuales o prácticas pueden sugerir un número determinado de grupos. Por ejemplo, si el propósito de la agrupación es identificar los segmentos del mercado, es probable que la gerencia quiera un número de grupos en particular.
- En el conglomerado jerárquico, las distancias en las que los grupos se combinan pueden utilizarse como criterios. Esta información puede obtenerse del programa de aglomeración o del dendrograma.

3.7 Validación de clusters obtenidos

Dados los criterios generales que comprende el Análisis Cluster, no debe aceptarse ninguna solución de agrupación sin una evaluación de su confianza y validez. La validación es el intento por parte del analista de asegurar que los clusters obtenidos sean representativos de la población original y que sean generalizables a otros objetos y estables a lo largo del tiempo. Los siguientes procedimientos ofrecen revisiones adecuadas de la calidad de los resultados de la agrupación:

- Realizar el Análisis Cluster con los mismos datos y utilizar distintas medidas de distancia. Comparar los resultados con todas las medidas a fin de determinar la estabilidad de las soluciones.
- Utilizar diversos métodos de conglomerado y comparar los resultados.
- Dividir los datos a la mitad de forma aleatoria. Realizar el Análisis Cluster por separado en cada mitad (submuestra). Comparar las soluciones de los dos análisis y evaluar la correspondencia de los resultados o bien comparar los centroides de grupo de las dos submuestras.
- Eliminar las variables de forma aleatoria. Realizar la agrupación basándose en el conjunto reducido de variables. Comparar los resultados basados en el conjunto completo con los que se obtuvieron al realizar el conglomerado.
- En el conglomerado no jerárquico la solución puede depender del orden de los casos en el conjunto de datos. Para estudiar esto, es recomendable utilizar distintos órdenes de los casos hasta estabilizar la solución.

3.8 Interpretación y elaboración del perfil de los clusters

La interpretación y elaboración del perfil de los grupos comprende la descripción de los casos que forman cada grupo. Los centroides representan los valores medios de los objetos que contiene el grupo en cada una de las variables. Resulta útil elaborar el perfil de los grupos en términos de las variables utilizadas para el conglomerado.

3.9 Ejemplo