Introducción al Análisis Clúster

El análisis clúster es una técnica multivariante cuya idea básica es clasificar objetos formando grupos/conglomerados (clúster) que sean lo más homogéneos posible dentro de si mismos y heterogéneos entre sí.

Surge ante la necesidad de diseñar una estrategia que permita definir grupos de objetos homogéneos. Este agrupamiento se basa en la idea de distancia o similitud entre las observaciones y la obtención de dichos clusters depende del criterio o distancia considerados, por ejemplo, una baraja de carta española se podría dividir de distintas formas: en dos clusters (figuras y números), en cuatro clusters (los cuatro palos), en ocho clusters (los cuatro palos y según sean figuras o números). Es decir, el número de clusters depende de lo que consideremos como similar.

El análisis clúster es una tarea de clasificación. Por ejemplo

- Clasificar **grupos de consumidores** respecto a sus preferencias en nuevos productos
- Clasificar las **entidades bancarias** donde sería más rentable invertir
- Clasificar las estrellas del cosmos en función de su luminosidad
- Identificar si hay grupos de municipios en una determinada comunidad con una tendencia similar en el consumo de agua con el fin de identificar buenas prácticas para la sostenibilidad y zonas problemáticas por alto consumo.

Como se puede comprender fácilmente el análisis clúster tiene una extraordinaria importancia en la investigación científica, en cualquier rama del saber. La clasificación es uno de los objetivos fundamentales de la Ciencia y en la medida en que el análisis clúster nos proporciona los medios técnicos para realizarla, se nos hará imprescindible en cualquier investigación.

Planteamiento del problema

Consideremos una muestra X formada por n individuos sobre los que se miden p variables, $X_1,\ldots,X_p$ (p variables numéricas observadas en n objetos). Sea x_{ij} el valor de la variable X_j en el i-ésimo objeto $i = 1,\ldots,n; j = 1,\ldots,p$.

Este conjunto X de valores numéricos se pueden ordenar en una matriz

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{pmatrix}$$

La i-ésima fila de la matriz X contiene los valores de cada variable para el i-ésimo individuo, mientras que la j-ésima columna muestra los valores pertenecientes a la j-ésima variable a lo largo de todos los individuos de la muestra.

Se trata, fundamentalmente, de resolver el siguiente problema: Dado un conjunto de n individuos caracterizados por la información de p variables X_j, ($j = 1,2,\ldots, p$), nos planteamos clasificarlos de manera que los individuos pertenecientes a un grupo (clúster) (y siempre con respecto a la información disponible de las variables) sean lo más similares posibles entre sí y los distintos grupos sean entre ellos tan disimilares como sea posible.

El proceso completo puede estructurarse de acuerdo con el siguiente esquema:

- Partimos de un conjunto de n individuos de los que se dispone de una información cifrada por un conjunto de p variables (una matriz de datos de n individuos y p variables).
- Establecemos un criterio de similaridad y construimos una matriz de similaridades que nos permita relacionar la semejanza de los individuos entre sí. Para medir lo similares (o disimilares) que son los individuos existe una gran cantidad de índices de similaridad y de disimilaridad o divergencia. Todos ellos tienen propiedades y utilidades distintas y habrá que ser consciente de ellas para su correcta aplicación.
- Elegimos un algoritmo de clasificación para determinar la estructura de agrupación de los individuos.
- Especificamos esa estructura mediante diagramas arbóreos.

Resumiendo

- El objetivo del Análisis Clúster es obtener grupos de objetos de forma que, por un lado, los objetos pertenecientes a un mismo grupo sean muy semejantes entre sí y, por el otro, los objetos pertenecientes a grupos diferentes tengan un comportamiento distinto con respecto a las variables analizadas.
- Es una técnica exploratoria puesto que la mayor parte de las veces no utiliza ningún tipo de modelo estadístico para llevar a cabo el proceso de clasificación.
- Conviene estar siempre alerta ante el peligro de obtener, como resultado del análisis, no una *clasificación* de los datos sino una *disección* de los mismos en distintos grupos. ***El conocimiento que el analista tenga acerca del problema decidirá que grupos obtenidos son significativos y cuáles no***.

Métodos de clasificación

Se distinguen dos grandes categorías de métodos clusters: Métodos jerárquicos y Métodos no-jerárquicos

- **Métodos Jerárquicos:** En cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo. La clasificación resultante tiene un número creciente de clases anidadas.
- **Métodos No jerárquico o Repartición**: Comienzan con una solución inicial, un número de grupos **g** fijado de antemano y agrupa los objetos para obtener los **g** grupos.

Proceso que se debe seguir en un análisis clúster

Paso 1: Selección de variables

La clasificación dependerá de las variables elegidas. Introducir variables irrelevantes aumenta la posibilidad de errores. Hay que utilizar algún criterio de selección:

- Seleccionar sólo aquellas variables que caracterizan los objetos que se van agrupando, y referentes a los objetivos del análisis clúster que se va a realizar
- Si el número de variables es muy grande se puede realizar previamente un Análisis de Componentes Principales y resumir el conjunto de variables.

Paso 2: Detección de valores atípicos. El análisis clúster es muy sensible a la presencia de objetos muy diferentes del resto (valores atípicos).

Paso 3. Seleccionar la forma de medir la distancia/disimilitud entre objetos dependiendo de si los datos con cuantitativos o cualitativos

- Datos métricos: Medidas de correlación y medidas de distancia
- Datos no métricos: Medidas de asociación.

Paso 4: Estandarización de los datos (Decidir si se trabaja con los datos según se miden o estandarizados). El orden de las similitudes puede cambiar bastante con sólo un cambio de escala de una de las variables por lo que sólo se realizará una tipificación cuando resulte necesario.

Paso 5: Obtención de los clusters y valoración de la clasificación realizada

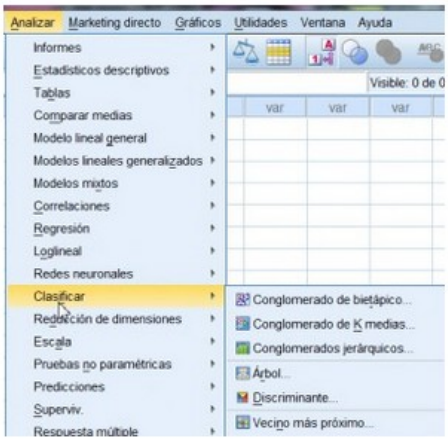
- Elegir el algoritmo para la formación de clúster (Procedimientos jerárquicos o procedimientos no jerárquicos)
- Número de clusters: Regla de parada. Existen diversos métodos de determinación del número de clusters, algunos están basados en reconstruir la matriz de distancias original, otros en los coeficientes de concordancia de Kendall y otros realizan análisis de la varianza entre los grupos obtenidos. No existe un criterio universalmente aceptado. Dado que la mayor parte de los paquetes estadísticos proporciona las distancias de aglomeración, es decir, las distancias a las que se forma cada clúster, una forma de determinar el número de grupos consiste en localizar en qué iteraciones del método utilizado dichas distancias dan grandes saltos
- Adecuación del modelo. Comprobar que el modelo no ha definido clúster con un solo objeto, clúster con tamaños desiguales,...

Análisis clúster en SPSS

El programa SPSS dispone de tres tipos de análisis clúster:

- Análisis de conglomerados de bietápico
- Análisis de conglomerados de *K medias*
- Análisis de conglomerados *jerárquicos.*

Cada uno de estos procedimientos utiliza un algoritmo distinto en la creación de clusters y contiene opciones que no están disponibles en los otros.



- **Análisis de conglomerados de bietápico**. El clúster en dos etapas está pensado para minería de datos, es decir para estudios con un **número de individuos grande** que pueden tener problemas de clasificación con los otros procedimientos. Se puede utilizar tanto cuando el **número de clúster es conocido a priori y cuando es desconocido**. Permite trabajar conjuntamente con **variables de tipo mixto (cualitativas y cuantitativas**).
- **Análisis de conglomerados de K medias.** Es un método de clasificación **No Jerárquico (Repartición).** El número de clusters que se van a formar es fijado de antemano (**requiere conocer el número de clusters a priori)** y se agrupan los objetos para obtener esos grupos. Comienzan con una solución inicial y los objetos se reagrupan de acuerdo con algún criterio de optimalidad. El clúster no jerárquico sólo puede ser aplicado a **variables cuantitativas**. Este procedimiento puede analizar **archivos de datos grandes**.
- **Análisis de conglomerados jerárquicos.** En el método de clasificación **Jerárquico** en cada paso del algoritmo sólo un objeto cambia de grupo y los grupos están anidados en los pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo. El método *jerárquico* es idóneo para determinar el número óptimo de conglomerados existente en los datos y el contenido de los mismos. Se utiliza cuando **no se conoce el número de clusters a priori y cuando el número de objetos no es muy grande.** Permite trabajar conjuntamente con variables de tipo mixto (**cualitativas y cuantitativas**). Siempre que todas las variables sean del mismo tipo, el procedimiento Análisis de Conglomerados Jerárquico podrá analizar variables de intervalo (continuas), de recuento o binarias.

Decisiones que hay que tomar para hacer un clúster

1. Elegir el método clúster que se va a utilizar
2. Decidir si se estandarizan los datos
3. Seleccionar la forma de medir la distancia/disimilitud entre los individuos
4. Elegir un criterio para unir grupos, distancia entre grupos.

Proceso que se debe seguir en un Análisis Clúster Jerárquico Aglomerativo

Paso 1: **Selección de las variables**. Se recomienda que las variables sean del mismo tipo (continuas, categóricas,..)

Paso 2: **Detección de valores atípicos**. El análisis clúster es muy sensible a la presencia de objetos muy diferentes del resto (valores atípicos).

Paso 3: **Elección de una medida de similitud entre objetos y obtención de la matriz de distancias.** Mediante estas medidas se determinan los clusters iniciales.

Paso 4: **Buscar los clusters más similares**

Paso 5: **Unir estos dos clusters en un nuevo clúster** que tenga al menos dos objetos, de forma que el número de clúster decrece en una unidad.

Paso 6: **Calcular la distancia entre este clúster y el resto**. Los distintos métodos para el cálculo de las distancias entre los clusters producen distintas agrupaciones, por lo que no existe una agrupación única.

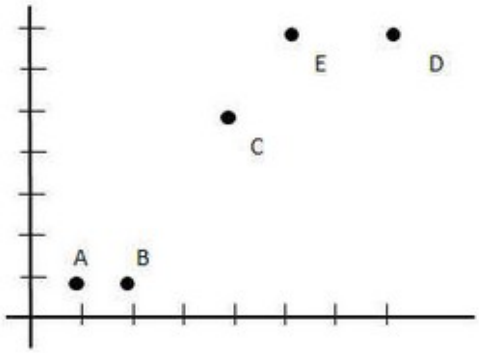
Paso 7: **Repetir desde el paso 4** hasta que todos los objetos estén en un único clúster.

Veamos un ejemplo sencillo de este proceso.

Tenemos 5 objetos o individuos (A,B, C,D, E) 2 variables (X_1, X_2). Los datos están en la siguiente tabla.

Objetos/individuos	X_1	X_2
A	1	1
B	2	1
C	4	5
D	7	7
E	5	7

Paso 1 y 2: Para detectar valores atípicos podemos representar los puntos en el plano



No detectamos valores atípicos.

Paso 3: La medida de distancia que vamos a tomar entre los objetos va a ser la distancia euclídea cuya expresión es:

$$d(P_1,P_2)=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$$

La distancia entre el grupo A y el B (comenzamos pensado que cada individuo es un grupo) es

$$d(A,B)=\sqrt{(2-1)^2+(1-1)^2}=1$$

$$d(A,B)=\sqrt{(2-1)^2+(1-1)^2}=1$$

Realizamos la distancia euclídea entre todos los puntos y obtenemos la siguiente matriz de distancias euclídeas entre los objetos

	A	B	C	D	E
A	0				
B	1	0			
C	5	4.5	0		
D	8.5	7.8	3.6	0	
E	7.2	6.7	2.2	2	0

Nota: Estamos realizado el método jerárquico aglomerativo, por lo qie inicialmente tenemos 5 cluster, uno por cada objeto.

Paso 4: Observamos en la matriz de distancias cuales son los objetos más similares, en nuestro ejemplo son el A y B que tienen la distancia menor (1).

	A	B	C	D	E
A	0				
B	1	0			
C	5	4.5	0		
D	8.5	7.8	3.6	0	
E	7.2	6.7	2.2	2	0

Paso 5: Fusionamos los clusters más similares construyendo un nuevo clúster que contiene A y B. Se han formado los clusters: AB, C, D y E.

Paso 6: Calculamos la distancia entre el clúster AB y los objetos C, D y E. Para medir esta distancia tomamos como representante del clúster AB el centroide, es decir, el punto que tiene como coordenadas las medias de los valores de las componentes de las variables, es decir, las coordenadas de AB son: ((1+2)/2 , (1+1)/2) = (1.5, 1). La tabla de datos es la siguiente

Clúster	X_1	X_2
AB	1.5	1
C	4	5
D	7	7
E	5	7

Paso 7: Repetimos desde el paso 4 hasta que todos los objetos estén en un único clúster

Paso 4: A partir de estos datos calculamos de nuevo la matriz de distancias

	AB	C	D	E
AB	0			
C	4.7	0		
D	8.1	3.6	0	
E	6.9	2.2	2	0

Paso 5: Los clusters más similares son el D y E con una distancia de 2, que se fusionan en un nuevo clúster DE. Se han formado tres clusters AB, C, DE

Paso 6: Calculamos el centroide del nuevo clúster que es el punto (6,7) y formamos de nuevo la tabla de datos

Clúster	X_1	X_2
AB	1.5	1
C	4	5
DE	6	7

Paso 4: A partir de estos datos calculamos de nuevo la matriz de distancias

	AB	C	DE
AB	0		
C	4.7	0	
DE	7.5	2.8	0

Paso 5: Los clusters más similares son el C y DE con una distancia de 2.8, que se fusionan en un nuevo clúster CDE. Se han formado dos clusters AB y CDE

Paso 6. Calculamos el centroide del nuevo clúster ((4+5+7)/3 , (5+7+7)/3) = (5.3, 6.3) y formamos de nuevo la tabla de datos

Clúster	X_1	X_2
AB	1.5	1
CDE	5.3	6.3

Paso 4 : A partir de estos datos calculamos de nuevo la matriz de distancias

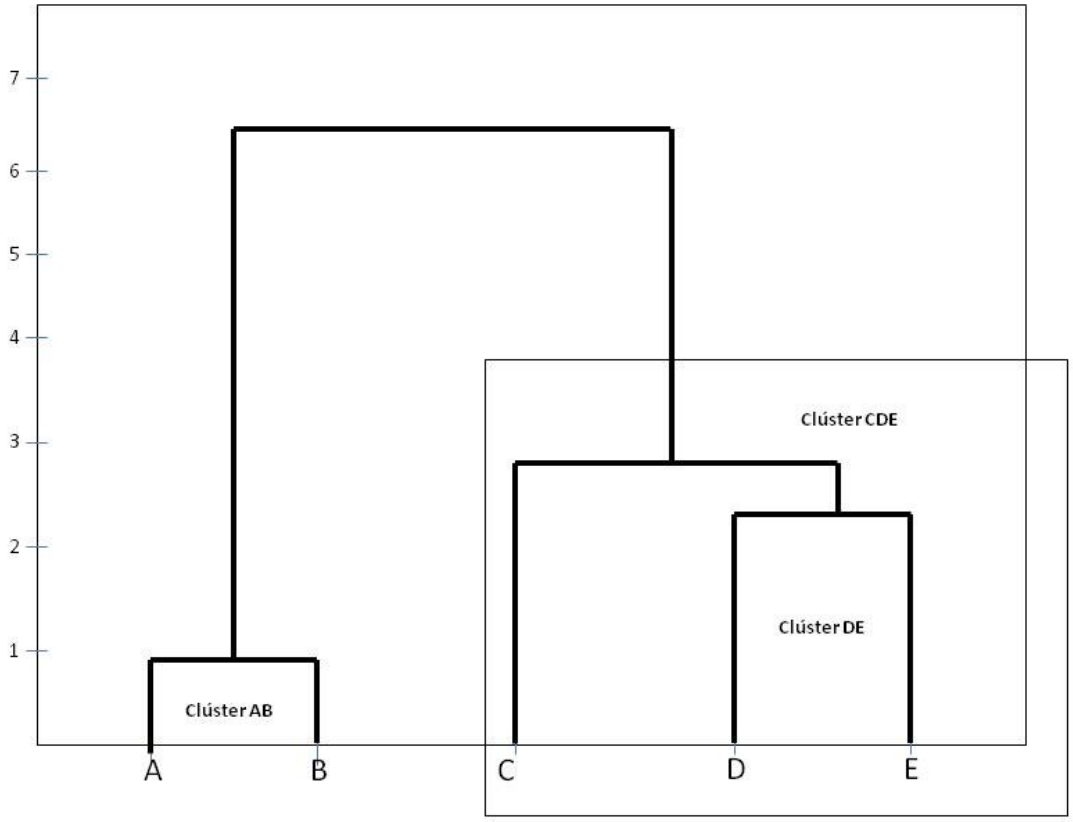
	AB	CDE
AB	0	
CDE	6.4	0

En este último paso tenemos solamente dos clusters con distancia 6.4 que se fusionarán en un único clúster en el paso siguiente terminando el proceso.

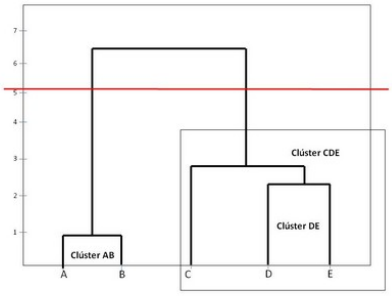
El último Cluster contiene ya a todos los elementos.

Dendrograma (δένδρον *déndron* 'árbol')

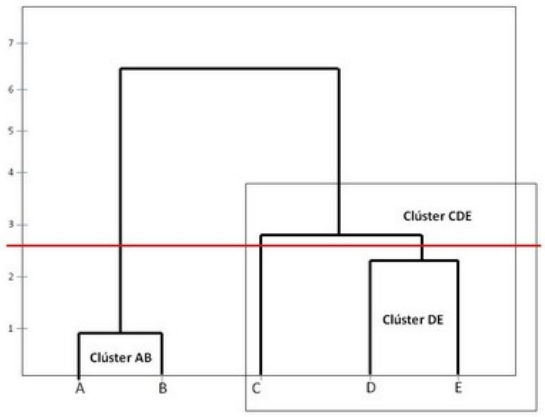
A continuación vamos a representar gráficamente el proceso de fusión mediante un dendrograma



A continuación mostramos varias soluciones, para ello cortamos el dendograma por medio de líneas horizontales, así por ejemplo



Se muestran 2 cluster o grupos AB y CDE en una distancia un poco mayor a 5.



En esta figura la línea de corte nos muestra 3 clusters: AB, C y DE

El número de clusters depende del sitio donde cortemos el dendograma, por lo tanto la decisión sobre el número óptimo de clusters es subjetiva. Es conveniente elegir un número de clusters que sepamos interpretar. Para interpretar los clúster podemos utilizar:

- ANOVA
- Análisis factorial
- Análisis discriminante
- ...
- Sentido común

Para decidir el número de clusters nos puede ser de gran utilidad representar los distintos pasos del algoritmo y las distancias a la que se produce la fusión de los clusters. En los primeros pasos el salto de las distancias es pequeño, mientras que esas diferencias van aumentando en los sucesivos pasos. Podemos elegir como punto de corte aquel donde comienzan a producirse saltos más bruscos. En nuestro ejemplo, el salto brusco se produce entre etapas 3 y 4, por lo tanto son dos el número de clusters óptimo.