# Tema 11. Clustering

*Abdelmalik Moujahid, Iñaki Inza, Pedro Larrañaga*
*Departamento de Ciencias de la Computación e Inteligencia Artificial*
*Universidad del País Vasco–Euskal Herriko Unibertsitatea*

## 1 Introducción

Un problema de clasificación no supervisada –véase Figura 1– parte de un conjunto de casos u objetos cada uno de los cuales está caracterizado por varias variables, y a partir de dicha información trata de obtener grupos de objetos, de tal manera que los objetos que pertenecen a un grupo sean muy homogéneos entre sí y por otra parte la heterogeneidad entre los distintos grupos sea muy elevada. Expresado en términos de variabilidad hablaríamos de minimizar la variabilidad dentro de los grupos para al mismo tiempo maximizar la variabilidad entre los distintos grupos.

Si bien los objetos pueden estar caracterizados por variable nominales –obteniéndose

| | $X_1$ | $\ldots$ | $X_i$ | $\ldots$ | $X_n$ |
|---|---|---|---|---|---|
| $O_1$ | $x_1^1$ | $\ldots$ | $x_i^1$ | $\ldots$ | $x_n^1$ |
| $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| $O_j$ | $x_1^j$ | $\ldots$ | $x_i^j$ | $\ldots$ | $x_n^j$ |
| $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ | $\ldots$ |
| $O_N$ | $x_1^N$ | $\ldots$ | $x_i^N$ | $\ldots$ | $x_n^N$ |

Figura 1: Fichero de casos de partida para la clasificación no supervisada

entonces un *clustering conceptual*– en este tema vamos a considerar que todas las variables descriptoras son cuantitativas. Se van a presentar métodos básicos para llevar a cabo dos tipos de clustering numérico: el *clustering particional* y el *clustering ascendente jerárquico*.

El apartado siguiente se dedica al clustering particional, mientras que el clustering ascendente jerárquico es tratado en un apartado posterior.

## 2 Clustering Particional

### 2.1 Introducción

En el clustering particional el objetivo es obtener una partición de los objetos en grupos o clusters de tal forma que todos los objetos pertenezcan a alguno de los $k$ clusters posibles y que por otra parte los clusters sean disjuntos.

Si denotamos por $\mathcal{O} = \{O_1, \ldots, O_N\}$ al conjunto de $N$ objetos, se trata de dividir $\mathcal{O}$ en $k$ grupos o clusters, $Cl_1, \ldots Cl_k$ de tal forma que:

- $\bigcup_{j=1}^{k} Cl_j = \mathcal{O}$
- $Cl_j \cap Cl_i = \emptyset$ para $i \neq j$

Uno de los problemas con los que uno se enfrenta en aplicaciones prácticas es el desconocimiento del valor de $k$ adecuado. Si bien existen algoritmos –más o menos sofisticados– que son capaces de adaptar el valor de $k$ a medida que se lleva a cabo la

búsqueda, los métodos que vamos a exponer consideran que el valor de $k$ está fijado de antemano.

### 2.2 Número de Conglomerados

El número de posibles aprupaciones que se pueden formar con $N$ objetos con los cuales se pretende crear $k$ grupos, lo denotamos por $S(N,k)$ y verifica la siguiente ecuación en diferencias:

$$S(N,k) = kS(N-1,k) + S(N-1,k-1) \tag{1}$$

con condiciones iniciales:

$$S(N,1) = S(N,N) = 1 \tag{2}$$

Para obtener la primera de las fórmulas anteriores, basta con tener en cuenta que partiendo de $N-1$ objetos ya clasificados, podremos obtener una agrupación de $N$ objetos en $k$ clusters, siempre y cuando los $N-1$ objetos anteriores estén agrupados bien en $k$ clusters o bien en $k-1$ clusters. Si partimos de que los $N-1$ objetos están agrupados en $k$ clusters, tenemos $k$ posibilidades para el $N$-ésimo objeto, ya que puede ir a cualquiera de los $k$ clusters. En el caso de los $N-1$ objetos estén agrupados en $k-1$ clusters, sólo tenemos una posibilidad para el objeto $N$-ésimo, y es que forme el $k$-ésimo cluster. Finalmente las fórmulas de las condiciones iniciales son obvias, ya que tan sólo existe una posible agrupación de los $N$ objetos en un sólo grupo, al igual que es única la posible agrupación de los $N$ objetos en $N$ grupos.

La resolución de la anterior ecuación en diferencias nos lleva a que:

$$S(N,k) = \frac{1}{k!}\sum_{i=1}^{k}(-1)^{k-i}\binom{k}{i}i^N \tag{3}$$

El número anterior es un número de Stirling de segunda clase, y se hace muy grande a poco que aumentemos el valor de $N$ y $k$ deje de ser igual a 1 o 2.

El anterior número puede ser interpretado como la cardinalidad del espacio de búsqueda si consideramos el problema de la obtención del mejor cluster como un problema de optimización combinatorial. Dada la ingente cantidad de posibles soluciones al problema del clustering particional, una búsqueda exhaustiva dentro del espacio de búsqueda no es computacionalmente posible, de ahi que se han venido utilizando algoritmos heurísticos para tal fin. Se remite al lector al apartado de notas bibliográficas para obtener referencias de la utilización de algoritmos estocásticos de optimización –enfriamiento estadístico, algoritmos genéticos y algoritmos de estimación de distribuciones– para este problema. En este apartado vamos a presentar dos heurísticos de optimización local –método de Forgy y método de McQueen– que están basados en una mejora iterativa de una solución inicial.

### 2.3 Método de Forgy

El método de Forgy se propuso en 1965 y constituye la aproximación mas simple al clustering particional. Al igual que el método de McQueen utiliza el concepto de centroide. El centroide de un cluster se define como el punto equidistante de los objetos pertenecientes a dicho cluster.

Denotamos por:

- $Cl_j$ al $j$-ésimo cluster

- $\mathbf{cl}_j = (cl_{j1}, \ldots, cl_{jn})$ al centroide de $Cl_j$
- $cl_{jr} = \frac{1}{|Cl_j|} \sum_{O_m \in Cl_j} O_{mr}$ con $O_m$ denotando un objeto agrupado en $Cl_j$. Las coordenadas de dicho objeto serán: $O_m = (o_{m1}, \ldots, o_{mn})$.

| | |
|---|---|
| Paso 1: | Comenzar con cualquier configuración inicial<br>Ir al Paso 2 si se comienza por un conjunto de $k$ centroides<br>Ir al Paso 3 si se comienza por una partición del conjunto de objetos en $k$ grupos |
| Paso 2: | Asignar cada objeto a clasificar al centroide mas próximo.<br>Los centroides permanecen fijos en este paso |
| Paso 3: | Computar los nuevos $k$ centroides como los baricentros de los $k$ conglomerados obtenidos |
| Paso 4: | Alternar los Pasos 2 y 3 hasta que se alcance un determinado criterio de convergencia |

Figura 2: Algoritmo de Forgy ($k$ fijo)

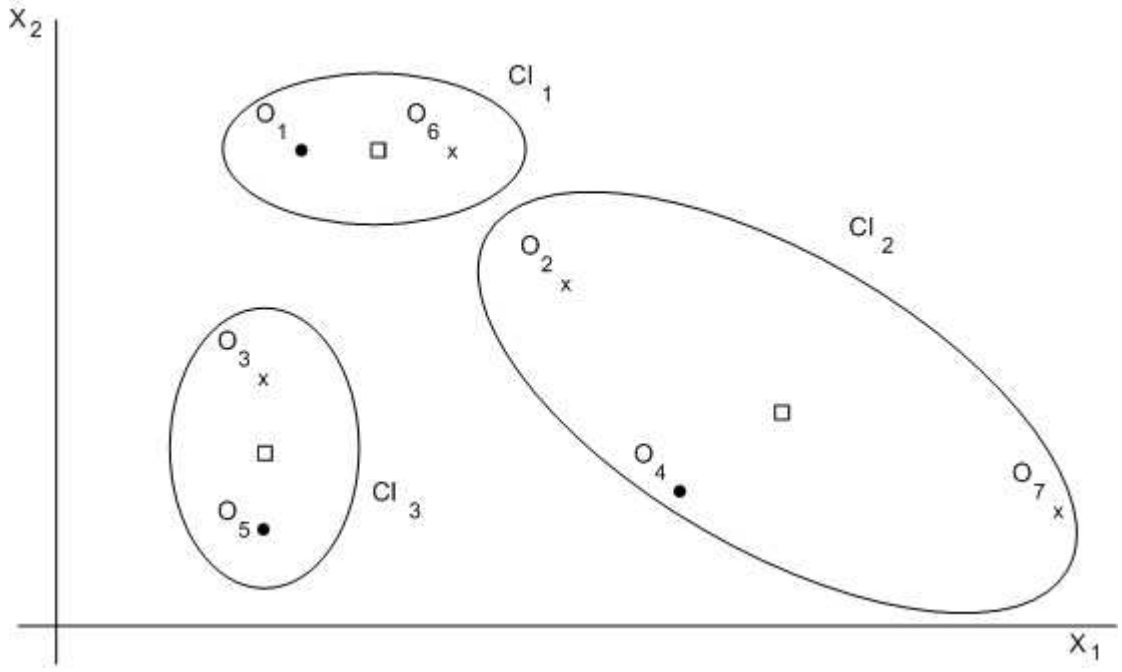


Figura 3: Partición en la iteración inicial. Método de Forgy

Tal y como se expresa en el pseudocódigo de la Figura 2, el algoritmo de Forgy comienza con cualquier configuración inicial, entendiéndose por la misma bien un conjunto de $k$ centroides –escogidos bien al azar o bien con una estrategia en la cual dichos centroides se encuentren lo suficientemente separados unos de otros en el espacio $n$-dimensional– o por medio de una partición del conjunto de objetos en $k$ clusters. En el caso de que se haya comenzado por un conjunto de $k$ centroides, obtendremos la partición correspondiente sin mas que asignar cada objeto que se pretende clasificar al centroide mas cercano. Una vez que se haya llevado a cabo la asignación de todos los objetos a clasificar, debemos de computar los $k$ nuevos centroides. Dichos $k$ nuevos centroides se utilizarán de nuevo como *atractores*, es decir asignaremos cada uno de los $N$ objetos que pretendemos clasificar al centroide mas cercano, y volveremos a recalcular los centroides cuando se hayan llevado a

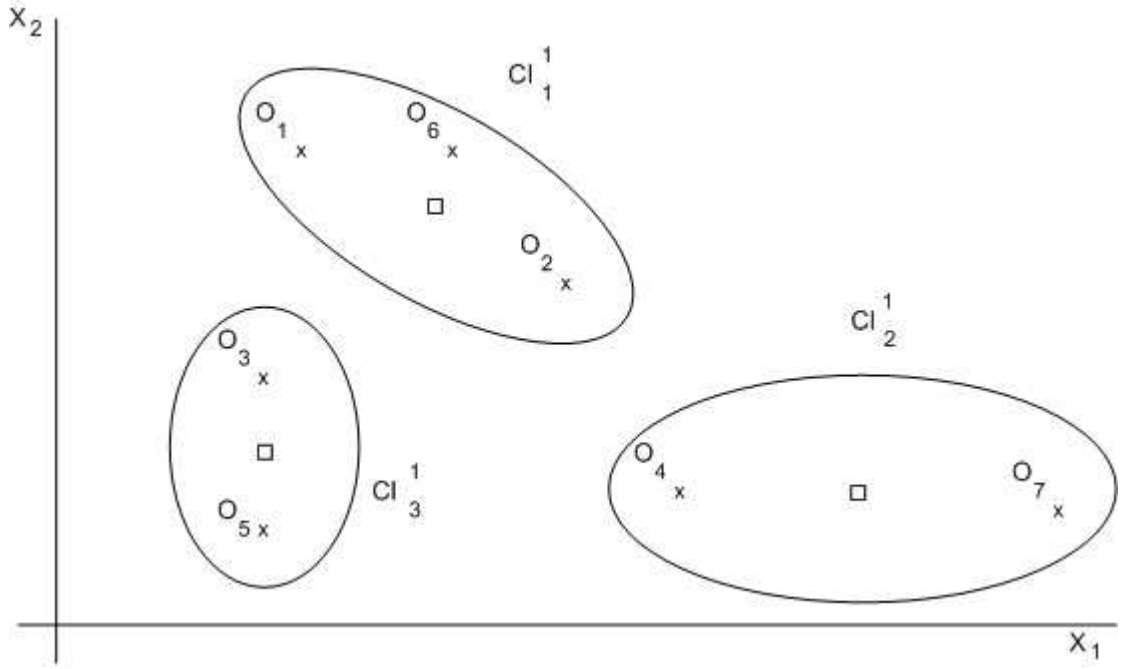


Figura 4: Partición en la iteración 1. Método de Forgy

cabo la asignación de todos los objetos. La alternación de los 2 pasos anteriores se mantendrá hasta que un determinado criterio de parada se verifique. Ejemplos de criterios de parada pueden ser: un número máximo de iteraciones, el que no se produzcan cambios en los clusters de dos iteraciones sucesivas, el que los nuevos centroides disten de los centroides obtenidos en la iteración previa menos que una determinada distancia, etc.

### 2.3.1 Ejemplo

Supongamos el ejemplo representado en la Figura 3. Tenemos $N = 7$ objetos a agrupar, caracterizados en un espacio bidimensional $(X_1, X_2)$, en $k = 3$ grupos. Los 7 objetos están representados por $O_1, O_2, O_3, O_4, O_5, O_6$ y $O_7$. Supongamos que los objetos $O_1, O_4$ y $O_5$ son seleccionados como centroides en el paso 0. Tal y como se observa en la Figura 3, en la iteración inicial del algoritmo de Forgy obtenemos:

$Cl_1^0 = \{O_1, O_6\}$
$Cl_2^0 = \{O_2, O_4, O_7\}$
$Cl_3^0 = \{O_3, O_5\}$

Representamos, por medio del símbolo □ los centroides de los clusters obtenidos, los cuales van a actuar como "atractores" para los 7 objetos, tal y como se muestra en la Figura 4.

Tal y como se observa en la Figura 4, el objeto $O_2$ ha cambiado de grupo, ya que se encuentra más cercano al centroide del primer cluster que al del segundo.

## 2.4 Método de $k$-medias de McQueen

El método que McQueen propuso en el año 1967 es conocido como $k$-medias y es el método de clustering particional más utilizado.

| | |
|---|---|
| Paso 1: | Considerar los $k$ primeros elementos del fichero como $k$ conglomerados con un único elemento |
| Paso 2: | Asignar en el orden del fichero cada uno de los objetos al centroide más próximo. Después de cada asignación se recalculará el nuevo centroide |
| Paso 3: | Después de que todos los objetos hayan sido asignados en el paso anterior, calcular los centroides de los conglomerados obtenidos, y reasignar cada objeto al centroide más cercano |
| Paso 4: | Repetir los pasos 2 y 3 hasta que se alcance un determinado criterio de parada |

Figura 5: Algoritmo de McQueen ($k$ medias)

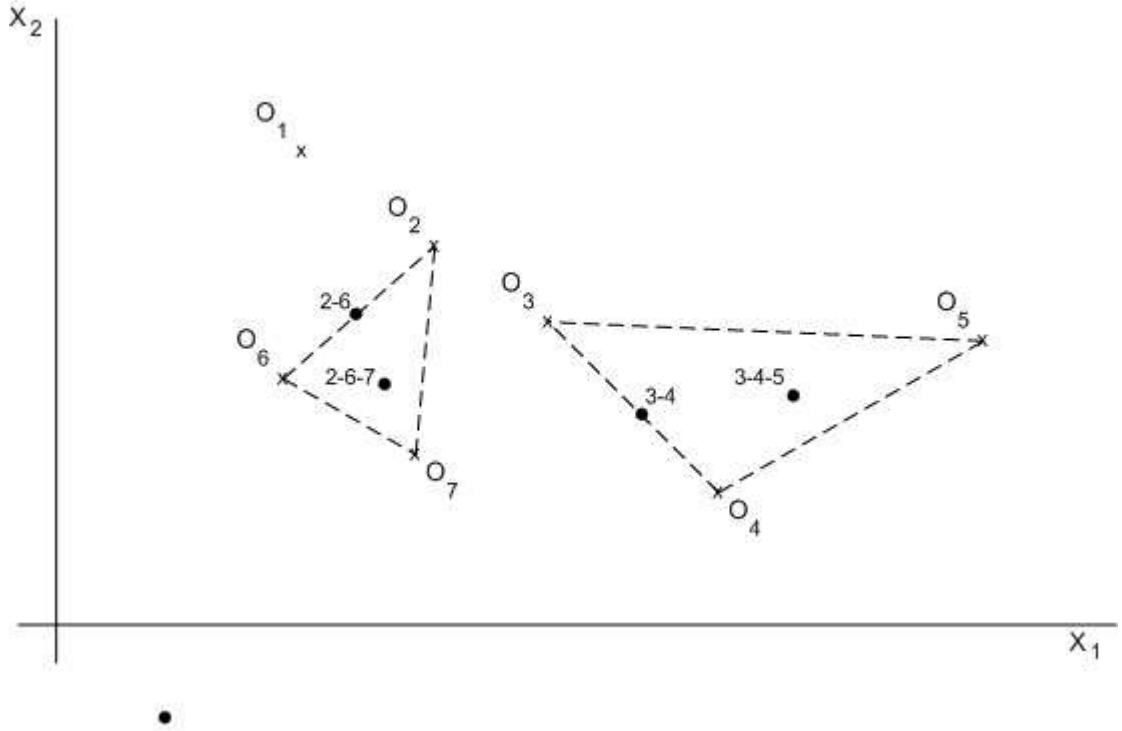


Figura 6: Partición en la iteración 1. Paso 2. Método de McQueen.

Tal y como puede verse en la Figura 5, en el algoritmo propuesto por McQueen se comienza considerando los $k$ primeros elementos del fichero de casos como los $k$ centroides iniciales, o dicho de forma equivalente como conglomerados con un único elemento. A continuación, y siguiendo el orden establecido en el fichero, cada uno de los objetos se va asignando al conglomerado con centroide más próximo, con la característica de que al efectuar cada asignación se recalculan las coordenadas del nuevo centroide. Esta característica es la que básicamente distingue al método de McQueen del método de Forgy. Finalmente, una vez asignados todos los objetos, se calculan los centroides para cada uno de los conglomerados y se reasigna cada objeto al centroide más cercano sin que en este paso se lleve a cabo una recalculación del centroide para cada asignación. Los pasos anteriores se iteran hasta que se verifique un determinado criterio de parada (ejemplos de los cuales han sido comentados con anterioridad).

Es importante tener en cuenta que el algoritmo de McQueen es sensible al orden

con el que se encuentran los objetos en el fichero de casos, y fundamentalmente es sensible a los objetos que se encuentran en las $K$ primeras posiciones.

### 2.4.1 Ejemplo

Supongamos que queremos obtener 3 grupos a partir de los 7 objetos, $O_1, O_2, O_3, O_4, O_5$, $O_6, O_7$ representados en la Figura 6.

Al considerar $O_1, O_2$ y $O_3$ como centroide en la iteración 0, e ir asignando el resto de los objetos en el orden en que se encuentran en el fichero de casos, obtenemos que $O_4$ se integra en el cluster con centroide $O_3$, a continuación $O_5$ se integra en el cluster con centroide $3-4$, $O_6$ se agrupa en el cluster con centroide $O_2$, y finalmente $O_7$ se agrupa en el segundo cluster, ahora con centroide denotado por $2-6$.

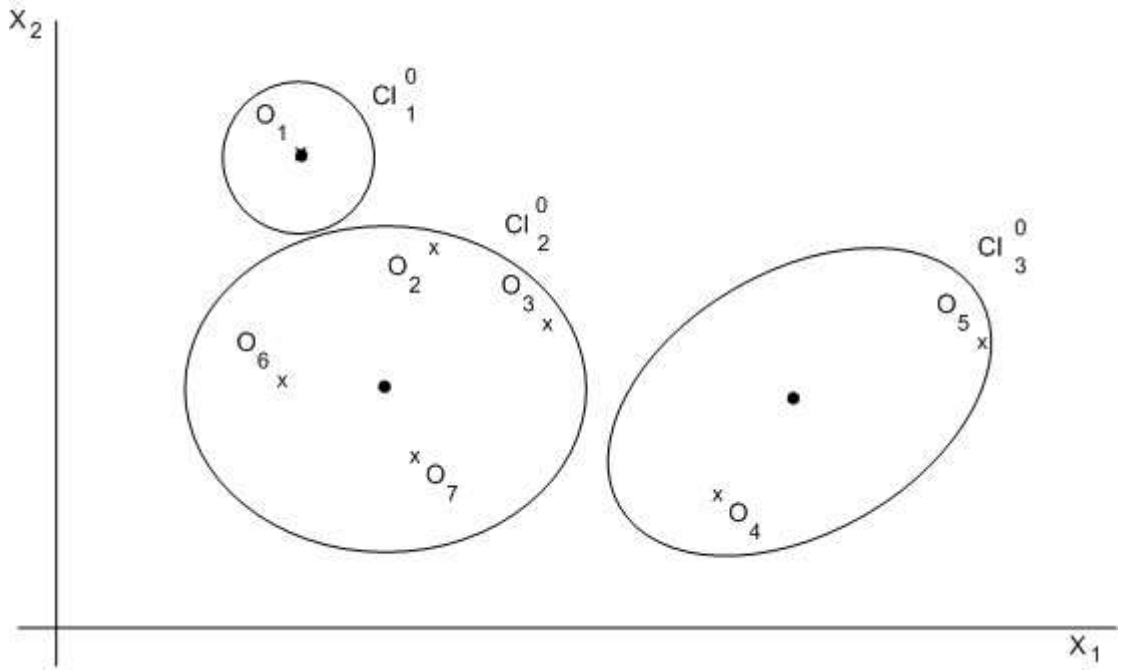


Figura 7: Partición en la iteración 1. Paso 3. Método de McQueen

Para llevar a cabo el Paso 3 del algoritmo de McQueen, supongamos que el objeto $O_3$ se encuentra más cerca del punto $2-6-7$ que del $3-4-5$. En tal caso obtenemos que al final de la iteración 0, los clusters son:

$$Cl_1^0 = \{O_1\}, Cl_2^0 = \{O_2, O_3, O_6, O_7\}, Cl_3^0 = \{O_4, O_5\}$$

## 3 Clustering Ascendente Jerárquico

### 3.1 Introducción

En el clustering ascendente jerárquico se pretende ir agrupando en cada paso aquellos 2 objetos (o conglomerados) más cercanos, para de esta forma ir construyendo una estructura conocida como dendograma, la cual parte en su base de tantos conglomerados como objetos a clasificar, los cuales son agrupados finalmente en un único grupo

conteniendo todos los objetos.

Si bien el costo computacional asociado a un clustering ascendente jerárquico es superior al que se relaciona con un clustering particional, el dendograma que se obtiene con el primer método es más rico que una simple partición, ya que posibilita la obtención de distintas particiones, simplemente variando el nivel de corte de dicha estructura, tal y como se observa en la Figura 8. De esta forma la problemática a la que se aludía en el apartado del clustering particional relativa a la determinación a priori del valor de $k$ (número de grupos) queda solventada.

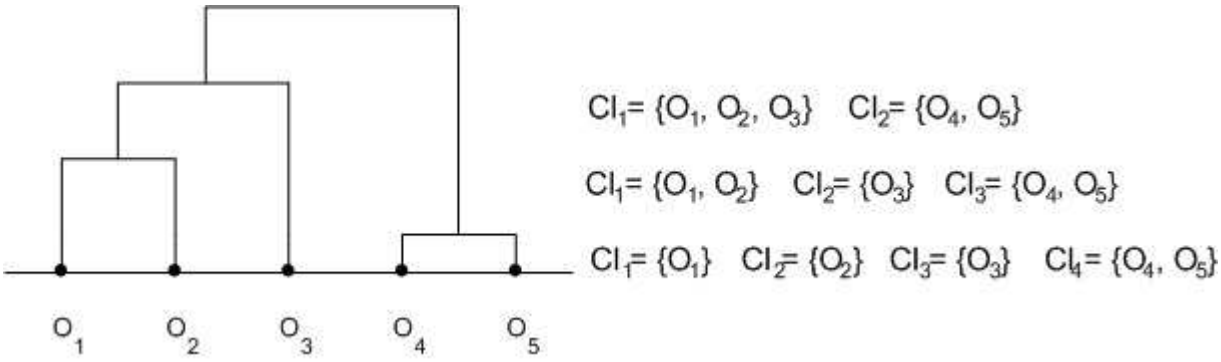


Figura 8: Dendograma resultado de la clasificación ascendente jerárquica

### 3.2 Ilustración de los pasos básicos por medio de un ejemplo

En este apartado vamos a ilustrar los pasos básicos de la clasificación ascendente jerárquica por medio de un ejemplo. Supongamos que tratamos de obtener el dendograma correspondiente a los 5 objetos $O_1, O_2, O_3, O_4, O_5$ cuyas características aparecen reflejadas en la Figura 9

En primer lugar necesitamos definir una distancia entre objetos, con la finalidad de

| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| $O_1$ | 2 | 4 | 6 |
| $O_2$ | 3 | 5 | 7 |
| $O_3$ | 1 | 1 | 4 |
| $O_4$ | 3 | 10 | 1 |
| $O_5$ | 3 | 9 | 2 |

Figura 9: Características de los 5 objetos sobre los que se va a efectuar una clasificación ascendente jerárquica

determinar que dos objetos deben de agruparse en el primer paso por encontrarse a distancia mínima. Supongamos que para tal final escogemos la distancia euclídea. Es decir:

$$d\left(O_i, O_j\right) = \sum_{w=1}^{3} \left(O_i^w - O_j^w\right)^2$$

Representamos en la matriz $D_0 \in M(5,5)$ –véase Figura 10– las distancias entre pares de objetos. Obviamente se verifica:

$d_{ii} = 0$
$d_{ij} = d_{ji}$
para todo $i$, $j \in \{1, \ldots, N\}$

Por tal motivo se presentan tan sólo los elementos de la matriz triangular superior.

| | $O_1$ | $O_2$ | $O_3$ | $O_4$ | $O_5$ |
|---|---|---|---|---|---|
| $O_1$ | | 3 | 14 | 62 | 40 |
| $O_2$ | | | 29 | 61 | 41 |
| $O_3$ | | | | 94 | 72 |
| $O_4$ | | | | | 2 |
| $O_5$ | | | | | |

Figura 10: Matriz $D_0 \in M(5, 5)$

A la vista de los resultados de la Figura 10, el par de objetos que se encuentran más cercanos son $O_2$ y $O_3$. Esto lo representamos en la Figura 11.

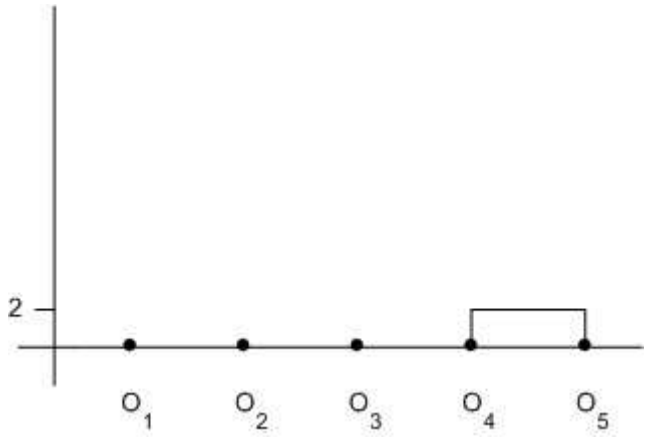


Figura 11: Dendograma parcial, resultado de la primera agrupación

Para llevar a cabo el siguiente paso necesitamos definir una distancia, no sólo entre dos objetos, sino entre conglomerados, ya que en este momento tenemos agrupados los objetos $O_4$ y $O_5$. Usamos el símbolo $h_i$ para denotar un conglomerado genérico. Así, tendremos que:

$h_1 = \{O_1\}$
$h_2 = \{O_2\}$
$h_3 = \{O_3\}$
$h_4 = \{O_4\}$
$h_5 = \{O_5\}$
$h_6 = \{O_4, O_5\}$

Existen distintas posibilidades para definir distancias entre dos conglomerados, que sean generalizaciones de las distancias definidas entre dos objetos, en el sentido de que particularizando al caso en que cada conglomerado tiene tan sólo un objeto coincidan con la distancia entre dos objetos.

Para seguir con el ejemplo, vamos a utilizar la distancia entre conglomerados denominada "enlace medio entre conglomerados", definida como:

$$d\left(h_i, h_j\right) = \frac{d_{ij}}{N_i N_i}$$

donde:

$$d_{ij} = \sum_{\substack{O_m \in h_i \\ O_l \in h_j}} d\left(O_m, O_l\right)$$

$$N_i = |h_i|$$

$$N_j = |h_j|$$

Es decir, el enlace medio entre conglomerados calcula la distancia media entre los objetos de ambos conglomerados.

La Figura 12 contiene los elementos de la matriz $D_1 \in M(4,4)$, cuyos elementos se

| | | $h_1$ | $h_2$ | $h_3$ | $h_6$ |
|---|---|---|---|---|---|
| | | $O_1$ | $O_2$ | $O_3$ | $O_4, O_5$ |
| $h_1$ | $O_1$ | | 3 | 14 | 51 |
| $h_2$ | $O_2$ | | | 29 | 51 |
| $h_3$ | $O_3$ | | | | 83 |
| $h_6$ | $O_4, O_5$ | | | | |

Figura 12: Matriz $D_1 \in M(4,4)$

refieren a enlaces medios entre conglomerados. Así por ejemplo, el valor 51 se ha obtenido a partir de:

$$d\left(h_1, h_6\right) = \frac{d\left(O_1, O_4\right) + d\left(O_1, O_5\right)}{1 \cdot 2} = \frac{62 + 40}{2} = 51$$

De dicha tabla se desprende que el siguiente agrupamiento debe hacerse entre $h_1$ y $h_2$, obteniéndose como resultado gráfico la Figura 13.

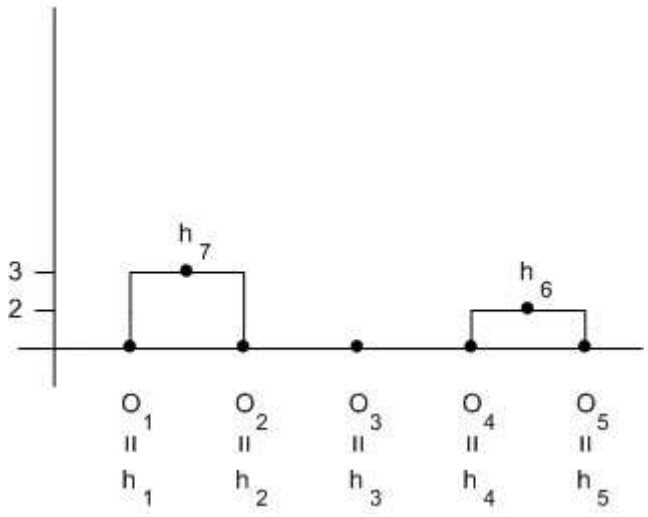


Figura 13: Dendograma parcial, resultado de las dos primeras agrupaciones

La Figura 14 representa las distancias entre los conglomerados $h_7$, $h_3$ y $h_6$ por medio

| | | $h_7$ | $h_3$ | $h_6$ |
|---|---|---|---|---|
| | | $O_1, O_2$ | $O_3$ | $O_4, O_5$ |
| $h_7$ | $O_1, O_2$ | | 21,5 | 51 |
| $h_3$ | $O_3$ | | | 83 |
| $h_6$ | $O_4, O_5$ | | | |

Figura 14: Matriz $D_2 \in M(3,3)$

de la matriz $D_2$. Los elementos de dicha matriz se han obtenido siguiendo el mismo criterio anterior. A la vista de la misma se juntan en el siguiente paso los conglomerados $h_7$ y $h_3$, según se muestra en la Figura 15.

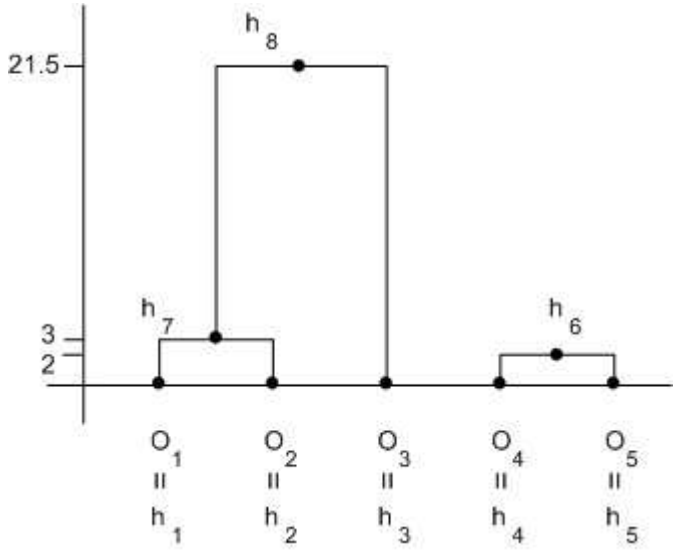


Figura 15: Dendograma parcial, resultado de las tres primeras agrupaciones

Finalmente calculamos $d(h_8, h_6)$ para obtener el dondograma global, obteniendo:

$$d(h_8, h_6) = \frac{d_{86}}{N_8 \cdot N_6} = \frac{1}{3 \cdot 2}\left[d(O_1, O_4) + d(O_1, O_5) + d(O_2, O_5) + d(O_2, O_4)\right.$$

$$\left. + d(O_3, O_4) + d(O_3, O_5)\right] = \frac{1}{3 \cdot 2}\left[62 + 40 + 61 + 91 + 94 + 72\right] = 61{,}6$$

La Figura 16 muestra el dendograma obtenido.

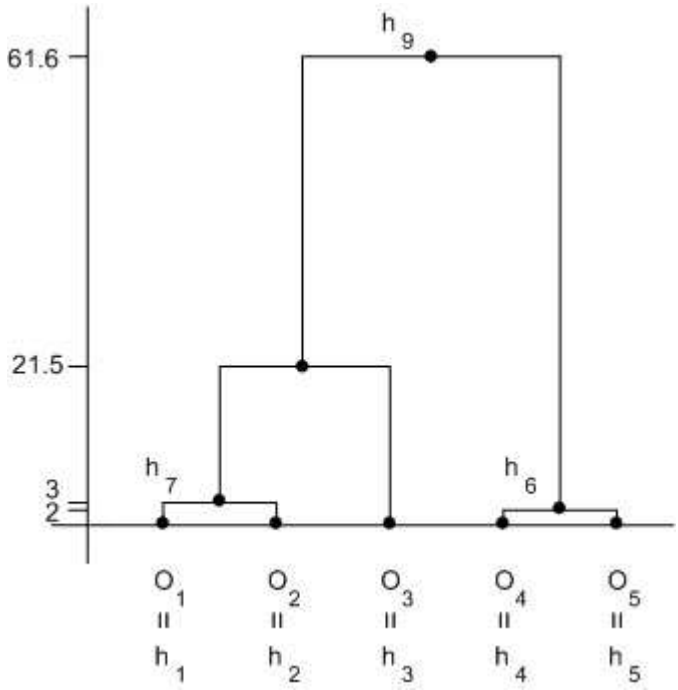


Figura 16: Dendograma global

## Referencias


1. E. Forgy (1965). Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications, *Biometrics* **21**, 768

2. J.B. McQueen (1967). Some methods for classification and analysis of multivariate observations, *Proceeding of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 281-297

3. A.D. Gordon (1987). A review of hierarchical classification journal of the Royal Statistical Society, *Sieries 4* **150(2)**, 119-137

4. J.M. Peña, J.A. Lozano, P. Larrañaga (1999). An empirical comparison of four initialization methods for the $k$-means algorithm, *Pattern Recognition Letters* **20(10)**, 1027-1040

5. J.A. Lozano, P. Larrañaga, M. Graña (1998). Partitional cluster analysis with genetic algorithms: searching for the numbert of clusters, *Data Science, Classification and Related Methods*, 117-125

6. J. Roure, P. Larrañaga, R. Sangüesa (2001). An Empirical Comparison Between $k$-Means, GAs and EDAs in Particional Clustering, *Estimation of Distribution Algorithms*, 339-356

7. J.A. Lozano, P. Larrañaga (1999). Applying genetic algorithms to search for the hierarchical clustering of a dataset. *Pattern Recognition Letters*, 20 (9), 911–918.