Tema 11: Clustering

Abdelmalik Moujahid, Iñaki Inza, Pedro Larrañaga

Departamento de Ciencias de la Computación e Inteligencia Artificial

Universidad del País Vasco

http://www.sc.ehu.es/isg/

Contenido

- Introducción
- Clustering particional
  - Método de Forgy (1965)
  - Método de $k - medias$ de McQueen (1967)
- Clustering ascendente jerárquico

Introducción

	X_1	...	X_i	...	X_n
O_1	x_1^1	...	x_i^1	...	x_n^1
...	...	...	...	...	...
O_j	x_1^j	...	x_i^j	...	x_n^j
...	...	...	...	...	...
O_N	x_1^N	...	x_i^N	...	x_n^N

Fichero de casos de partida para la clasificación no supervisada

- Homogeneidad dentro de las clases y heterogeneidad entre las distintas clases
- Dos tipos de métodos: particional y jerárquico

Clustering particional

- Objetivo: partición de los objetos en grupos o clusters. Todos los objetos pertenecen a alguno de los k clusters. Los clusters son disjuntos
- $\mathcal{O} = \{O_1, \ldots, O_N\}$ conjunto de N objetos. Se trata de particionar $\mathcal{O}$ 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 $i \neq j$
- Desconocimiento del valor de k adecuado. Consideramos que el valor de k está fijado de antemano

Clustering particional. Número de conglomerados

- $S(N,k)$ número de posibles aprupaciones con N objetos en k grupos
- $S(N,k)$ verifica la siguiente ecuación en diferencias:

$$S(N,k) = kS(N-1,k) + S(N-1,k-1)$$

$$S(N,1) = S(N,N) = 1$$

- Cardinalidad del espacio de búsqueda (número de Stirling de segunda clase)

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

Clustering particional. Método de Forgy (1965)

Denotamos por:

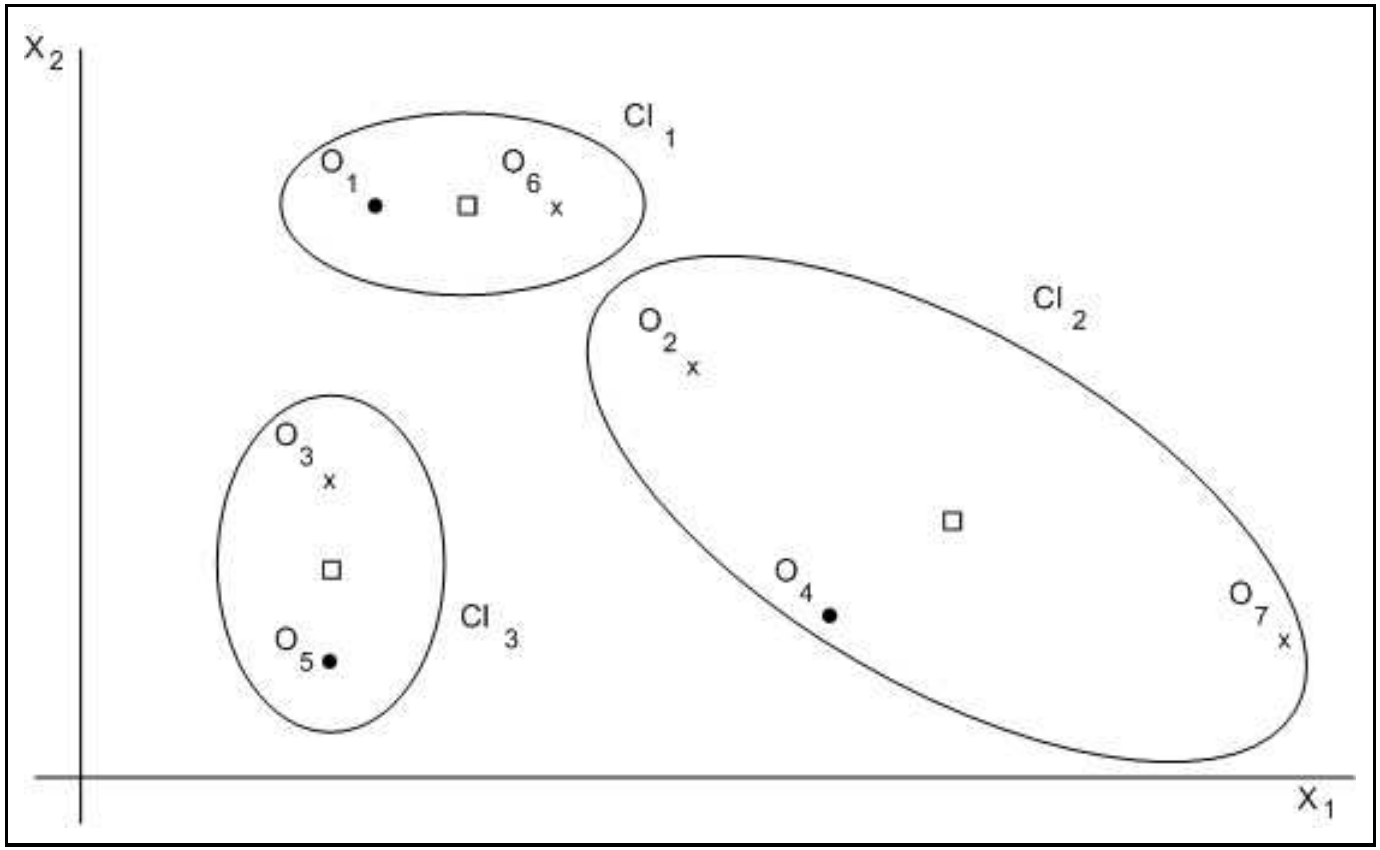
- Cl_j al j-ésimo cluster
- $\mathbf{cl}_j = (cl_{j1}, \ldots, cl_{jn})$ al centroide (punto equidistante de los objetos del cluster) de Cl_j
- $cl_{jr} = \frac{1}{|Cl_j|} \sum_{O_m \in Cl_j} O_{mr}$ con O_m denotando un objeto agrupado en Cl_j. Las coordenadas de dicho objeto serán: $O_m = (o_{m1},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

Clustering particional. Método de Forgy (1965)

Paso 1:	Comenzar con cualquier configuración inicial Ir al Paso 2 si se comienza por un conjunto de k centroides 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 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

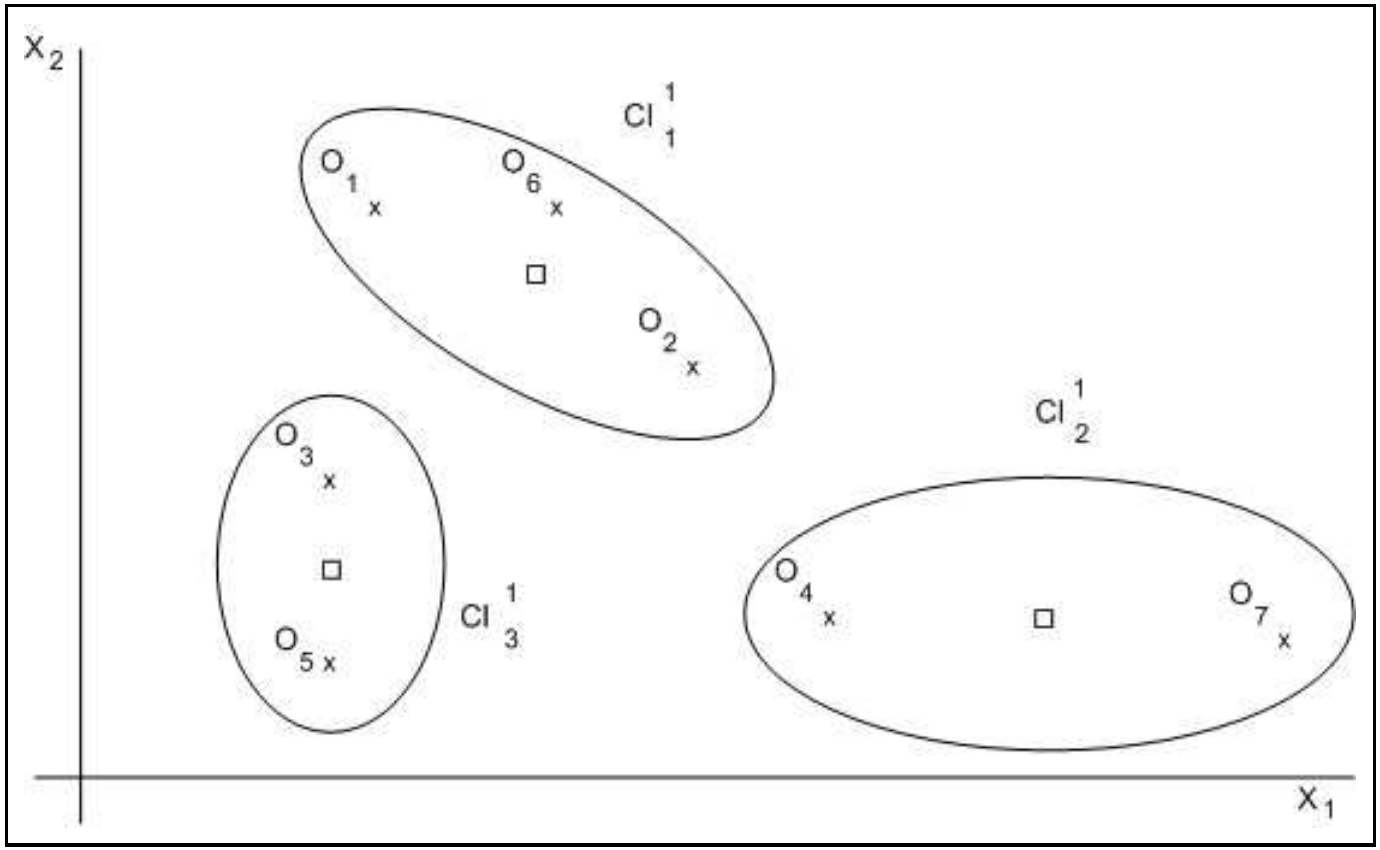
Algoritmo de Forgy

Clustering particional. Método de Forgy (1965)



Partición en la iteración inicial

Clustering particional. Método de Forgy (1965)



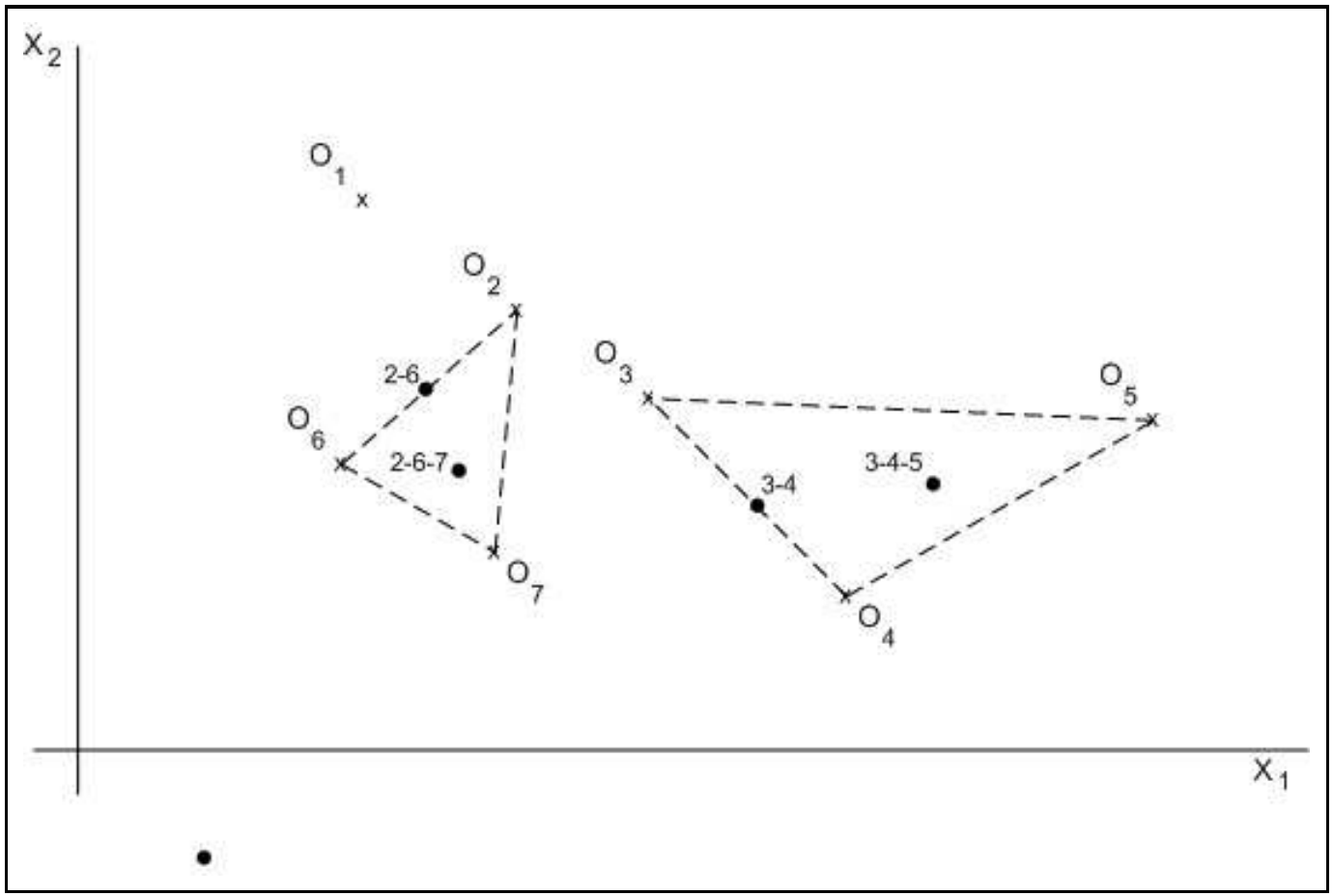
Partición en la iteración 1

Clustering particional. Método de k-medias de McQueen (1967)

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 obt 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

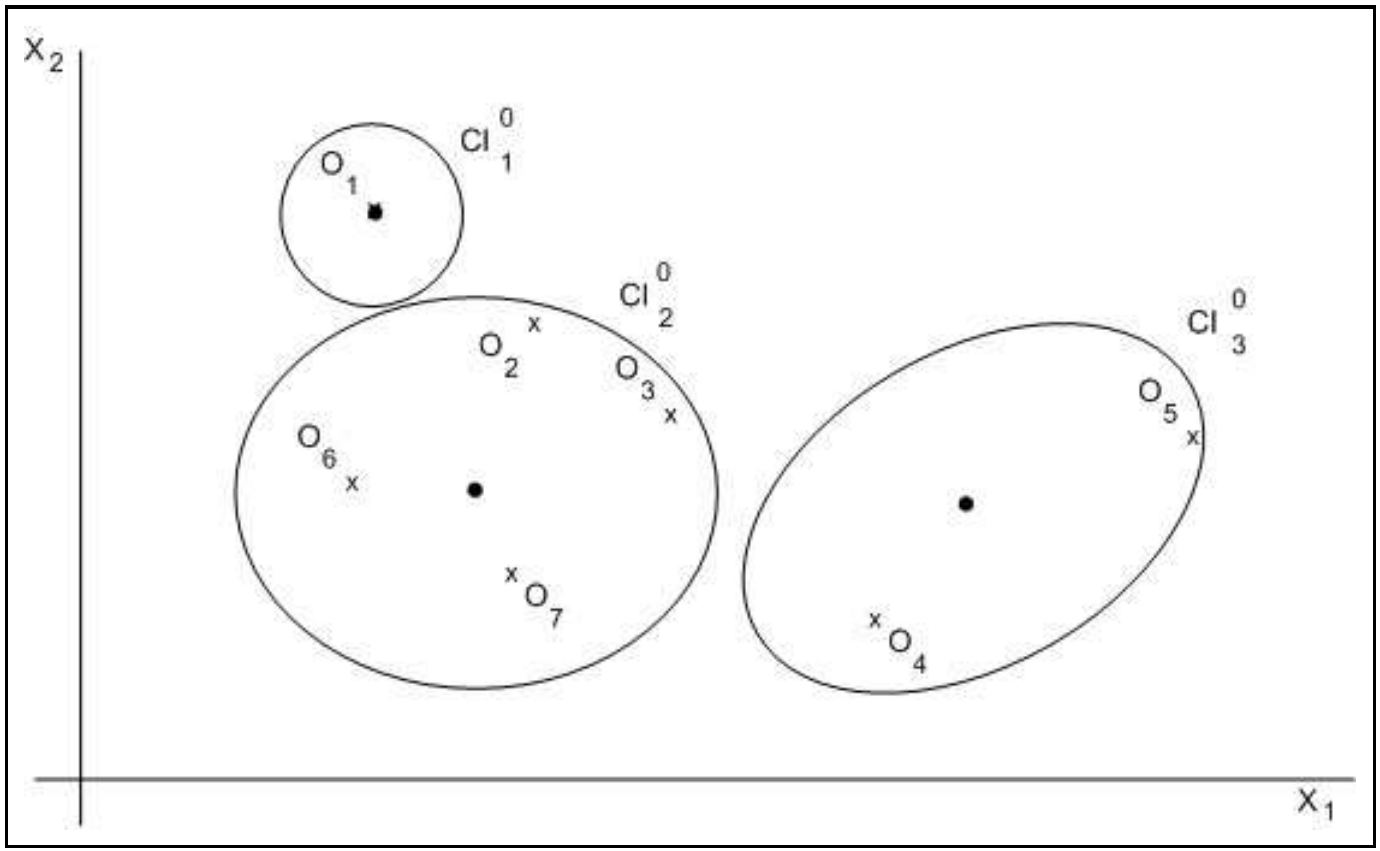
Algoritmo de McQueen (k medias)

Clustering particional. Método de k-medias de McQueen (1967)



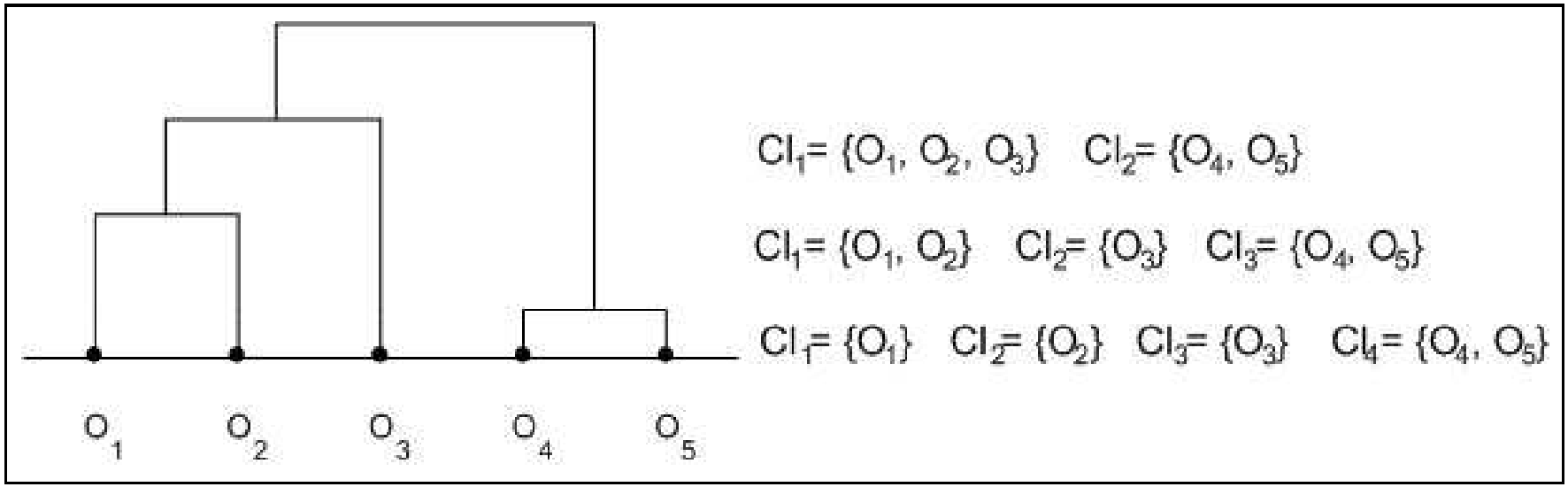
Partición en la iteración 1. Paso 2

Clustering particional. Método de k-medias de McQueen (1967)



Partición en la iteración 1. Paso 3

Clustering ascendente jerárquico



Dendograma resultado de la clasificación ascendente jerárquica

Clustering ascendente jerárquico

	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

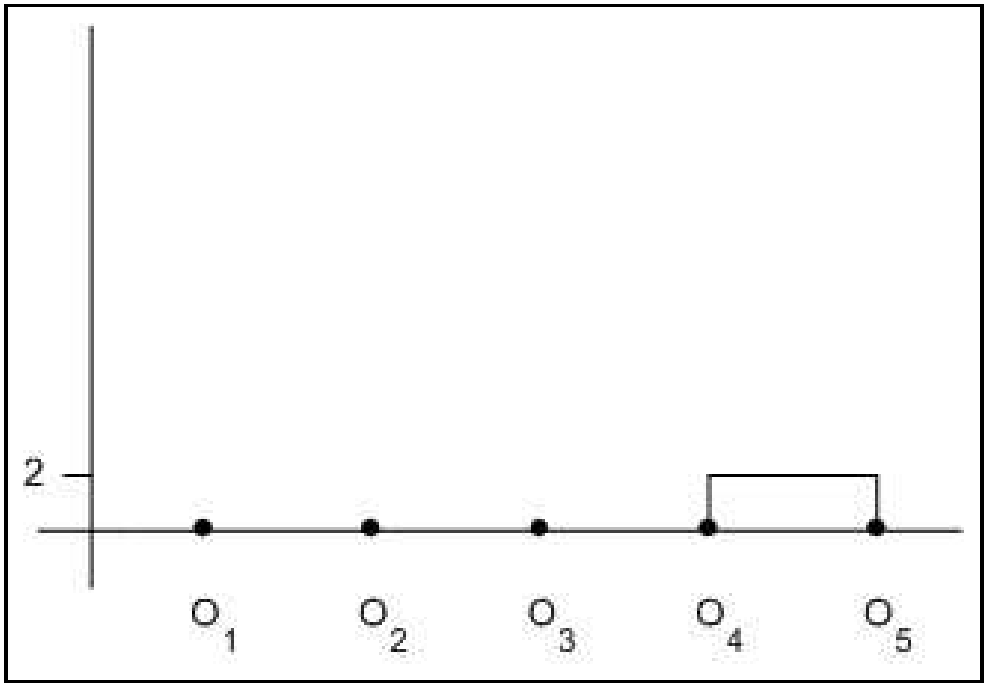
Características de los 5 objetos sobre los que se va a efectuar una clasificación ascendente jerárquica

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

	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					

Matriz $D_0 \in M(5,5)$

Clustering ascendente jerárquico



Dendograma parcial, resultado de la primera agrupación

Clustering ascendente jerárquico

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

Enlace medio entre conglomerados: $d(h_i, h_j) = \dfrac{d_{ij}}{N_i N_i}; \quad N_i = |h_i| \quad N_j = |h_j|$

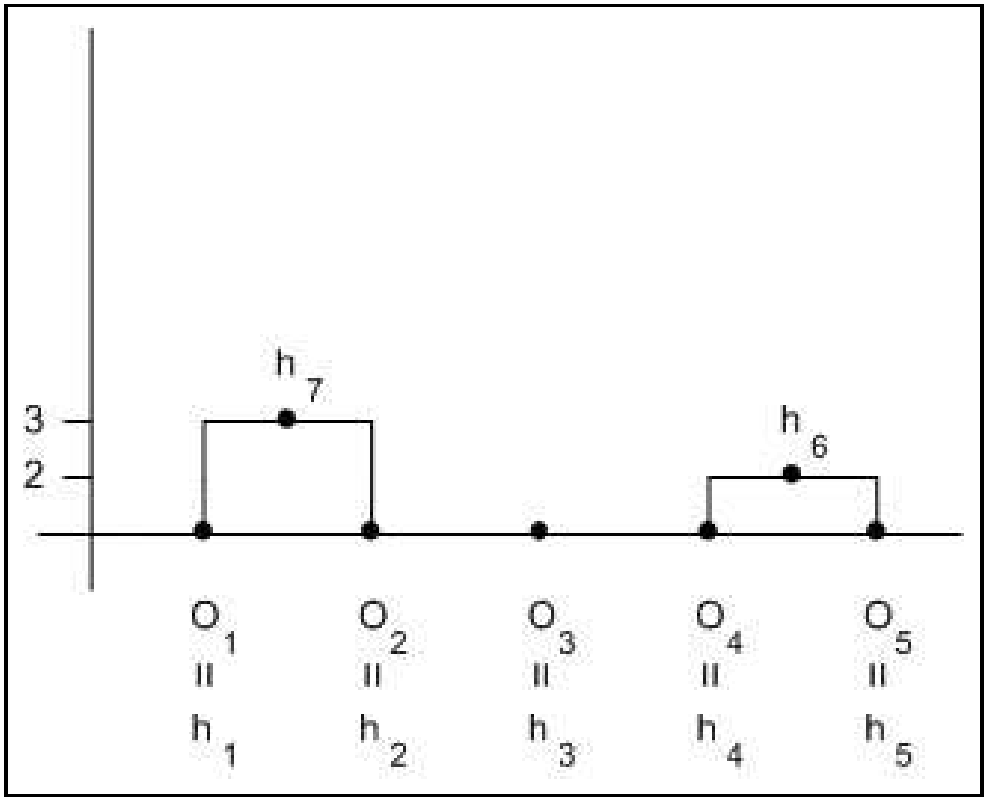
$$d_{ij} = \sum_{\substack{O_m \in h_i \\ O_l \in h_j}} d(O_m, O_l)$$

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

		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				

Matriz $D_1 \in M(4,4)$

Clustering ascendente jerárquico

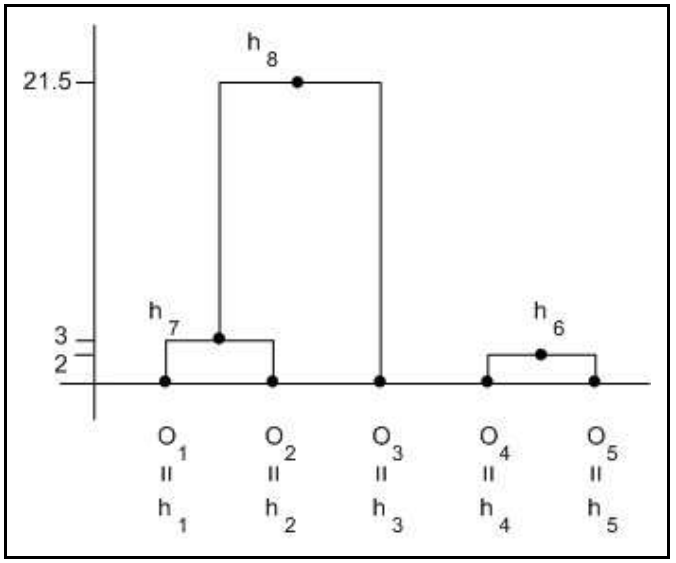


Dendograma parcial, resultado de las dos primeras agrupaciones

Clustering ascendente jerárquico

		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			

Matriz $D_2 \in M(3,3)$



Dendograma parcial, resultado de las tres primeras agrupaciones