Algoritmos de Clustering

Abdelmalik Moujahid

Grupo de Inteligencia Computacional
Universidad del País Vasco UPV/EHU
Curso 2014-2015

Índice

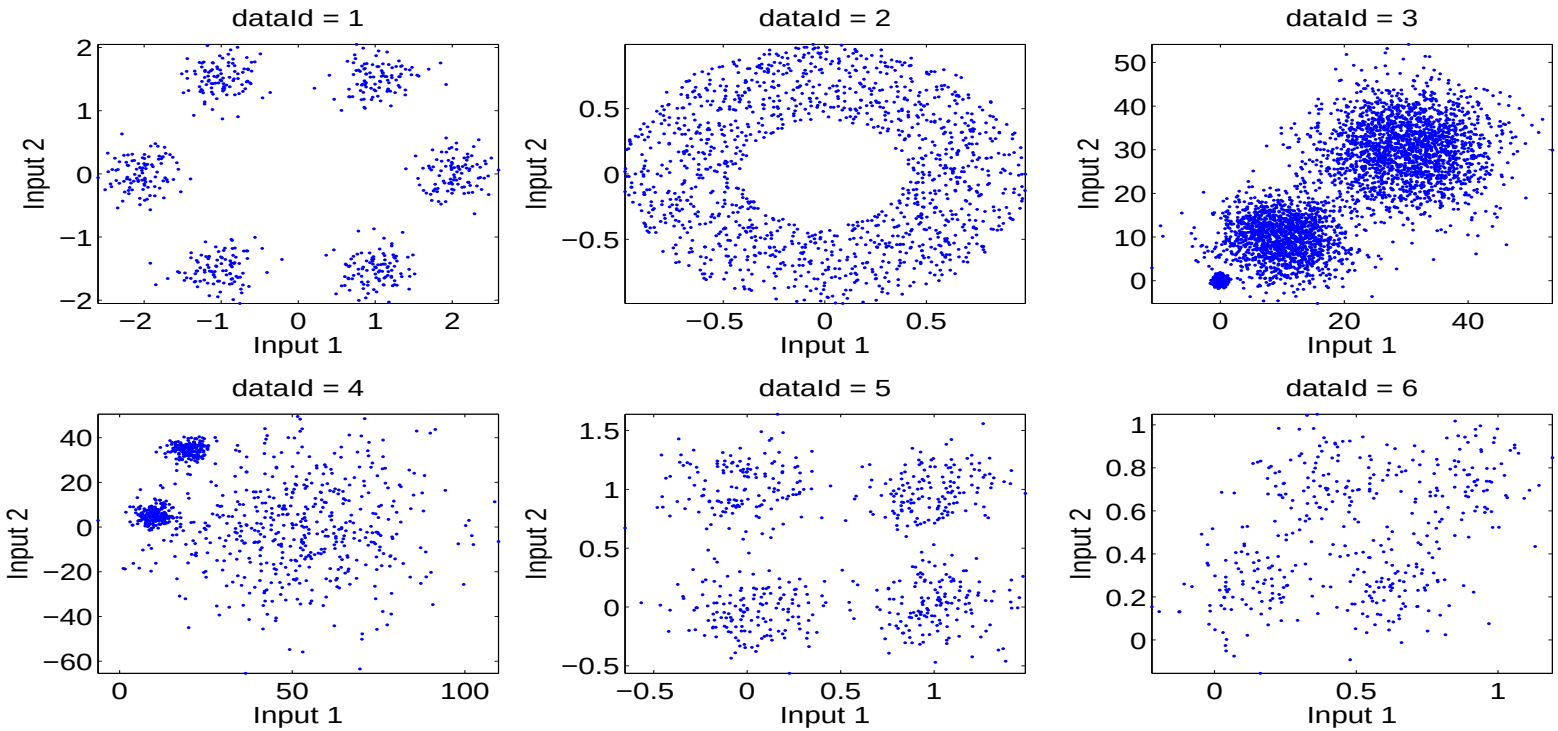
Índice

Problema de clasificación no supervisada

	a_1	...	a_j	...	a_d
X_1	x_{11}	...	x_{1j}	...	x_{1d}
...	...	...	...	...	...
X_i	x_{i1}	...	x_{ij}	...	x_{id}
...	...	...	...	...	...
X_N	x_{N1}	...	x_{Nj}	...	x_{Nd}

- Estos objetos (o vectores) son vistos como puntos en un espacio d-dimensional
- Los clusters son descritos como **regiones continuas** de este espacio formados por puntos relativamente densos
- A estos clusters se les conoce como **clusters naturales**

Estructuras de clustering



Análisis de clusters: ejemplos de distribución de los datos

Definición del Clustering

Definición1

A cluster is a set of entities which are alike, and entities from different clusters are not alike.

Definición2

A cluster is an aggregate of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it.

Everitt, B. (1980) Cluster Analysis, 2nd edition. London Social Science Research Council.

Definición del Clustering

Dado un conjunto de datos $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_j, ..., \mathbf{x}_N\}$, donde $\mathbf{x}_j = (\mathbf{x}_{j1}, \mathbf{x}_{j2}, ..., \mathbf{x}_{jd}) \in R^d$.

1. El clustering **particional** consiste en encontrar la K-partición de X, $C = \{C_1, ..., C_K\}$, con $K \leq N$, tal que:
   - $C_i \neq \emptyset, i = 1, ..., K$
   - $\bigcup_{i=1}^{K} C_i = \mathbf{X}$
   - $C_i \cap C_j = \emptyset$ para $i, j = 1, ..., K$ y $i \neq j$
2. El clustering **jerárquico** trata de construir una jerárquica de particiones de $\mathbf{X}$, $H = \{H_1, ..., H_Q\}$ con $Q \leq N$, tal que:
   - $C_i \in H_m, C_j \in H_l$, y $m > l$ implica que $C_i \subset C_j$ ó $C_i \cap C_j = \emptyset$ para cada $i, j \neq i, m, l = 1, ..., Q$

El clustering como un problema de búsqueda

El 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)

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

Fases del proceso de análisis de clusters

1. Selección o extracción de atributos
2. Diseño del algoritmo de clustering
3. Validación del clustering

1- Selección o extracción de atributos

- Distinguir los atributos relevante y redundantes.
- Construir nuevos atributos a partir de los atributos originales.
- Reducir la dimensionalidad del problema (Principal component analysis, independent component analysis, multidimensional scaling, etc).
- Mejorar la visualización de los datos y la interpretación de los resultados.

2- Diseño del algoritmo de clustering

- Determinar una medida de proximidad apropiadada.
- Transformar el problema de clustering en un problema de optimización definiendo una función de coste específica.
- Hay que tener en cuenta que los clusters obtenidos son dependiente de la función de proximidad elegida.

2- Diseño del algoritmo de clustering

Medidas de proximidad

Dado un conjunto de datos, X, con N objetos definidos en un espacio d-dimensional.
Una **función de distancia (o disimilitud)** satisface las siguientes condiciones:

1. Simétrica: $D(\mathbf{x}_i, \mathbf{x}_j) = D(\mathbf{x}_j, \mathbf{x}_i)$
2. Positiva: $D(\mathbf{x}_i, \mathbf{x}_j) \geq 0 \quad \forall \mathbf{x}_i, \mathbf{x}_j$
   Si además satisface las dos siguientes condiciones,
3. $D(\mathbf{x}_i, \mathbf{x}_j) \leq D(\mathbf{x}_i, \mathbf{x}_k) + D(\mathbf{x}_k, \mathbf{x}_j), \quad \forall \mathbf{x}_i, \mathbf{x}_j \quad y \quad \mathbf{x}_k$
4. $D(\mathbf{x}_i, \mathbf{x}_j) = 0$ sii $\mathbf{x}_i = \mathbf{x}_j$

se dice que D es una **métrica**.

2- Diseño del algoritmo de clustering

Medidas de proximidad

Una **función de similitud** satisface las siguientes condiciones:

1. Simétrica: $S(\mathbf{x}_i, \mathbf{x}_j) = S(\mathbf{x}_j, \mathbf{x}_i)$
2. Positiva: $0 \leq S(\mathbf{x}_i, \mathbf{x}_j) \leq 1 \quad \forall \mathbf{x}_i, \mathbf{x}_j$
   Si además satisface las dos siguientes condiciones,
3. $S(\mathbf{x}_i, \mathbf{x}_j)S(\mathbf{x}_j, \mathbf{x}_k) \leq [S(\mathbf{x}_i, \mathbf{x}_j) + S(\mathbf{x}_j, \mathbf{x}_k)]S(\mathbf{x}_i, \mathbf{x}_k)$,
   $\forall \mathbf{x}_i, \mathbf{x}_j \quad y \quad \mathbf{x}_k$
4. $S(\mathbf{x}_i, \mathbf{x}_j) = 1$ sii $\mathbf{x}_i = \mathbf{x}_j$

se dice que S es una **métrica**.

2- Diseño del algoritmo de clustering

Medidas de disimilitud

1. Distancia Euclidea (L_2): $D(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{l=1}^{d} \left|\mathbf{x}_{il} - \mathbf{x}_{jl}\right|^2\right]^{\frac{1}{2}}$
2. Distancia de Minkowski (L_p): $D(\mathbf{x}_i, \mathbf{x}_j) = \left[\sum_{l=1}^{d} \left|\mathbf{x}_{il} - \mathbf{x}_{jl}\right|^p\right]^{\frac{1}{p}}$
3. Distancia de Manhattan (L_1): $D(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{d} \left|\mathbf{x}_{il} - \mathbf{x}_{jl}\right|$
4. Distancia de Chebyshev (L_∞): $D(\mathbf{x}_i, \mathbf{x}_j) = max_{1 \leqslant l \leqslant d} \left|\mathbf{x}_{il} - \mathbf{x}_{jl}\right|$
5. Distancia de Mahalanobis: $D(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T S^{-1} (\mathbf{x}_i - \mathbf{x}_j)}$

- $L_\infty < L_2 < L_1$
- S *es la matriz de covarianza entre variables.*

2- Diseño del algoritmo de clustering

Medidas de similitud

1. Similitud Coseno: $S(\mathbf{x}_i, \mathbf{x}_j) = cos\alpha = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|}$
2. Coeficiente de correlación de Pearson:
$$r_{ij} = \frac{\sum_{l=1}^{d} (\mathbf{x}_{il} - \bar{\mathbf{x}}_i)(\mathbf{x}_{jl} - \bar{\mathbf{x}}_j)}{\sqrt{\sum_{l=1}^{d} (\mathbf{x}_{il} - \bar{\mathbf{x}}_i)^2 \sum_{l=1}^{d} (\mathbf{x}_{jl} - \bar{\mathbf{x}}_j)^2}}$$

- *Disimilitud basada en la similitud coseno:*
$$D(\mathbf{x}_i, \mathbf{x}_j) = 1 - S(\mathbf{x}_i, \mathbf{x}_j)$$
- *Disimilitud basada en el coeficiente de Pearson:*
$$D(\mathbf{x}_i, \mathbf{x}_j) = (1 - r_{ij})/2.$$

2- Diseño del algoritmo de clustering

Función de coste

El análisis de cluster consiste en encontrar una partición de los datos en donde los objetos del mismo cluster sean similares, mientras que los objetos de diferentes clusters sean muy disimilares.

- Función de coste basado en el error cuadrático medio (Duda et al., 2001)

$$J_s(\Gamma, \mathbf{M}) = \sum_{i=1}^{K} \sum_{j=1}^{N} \gamma_{ij} ||\mathbf{x}_j - \mathbf{m}_i||^2$$

$$J_s(\Gamma, \mathbf{M}) = \sum_{i=1}^{K} \sum_{j=1}^{N} \gamma_{ij} (\mathbf{x}_j - \mathbf{m}_i)^T (\mathbf{x}_j - \mathbf{m}_i)$$

2- Diseño del algoritmo de clustering

Función de coste

- $\Gamma = \{\gamma_{i,j}\}$ es la matriz de partición,
  $$\gamma_{i,j} = \begin{cases} 1 & \text{si } \mathbf{x}_j \in \text{cluster } i \\ 0 & \text{en otro caso} \end{cases} \quad \text{con } \sum_{i=1}^{K} \gamma_{i,j} = 1 \forall j$$
- $\mathbf{M} = [\mathbf{m}_1, ..., \mathbf{m}_K]$ es el vector de los centroides, $\mathbf{m}_i = \frac{1}{N_i} \sum_{j=1}^{N} \gamma_{ij} \mathbf{x}_j$ es el centroide del cluster i con N_i objetos.

2- Diseño del algoritmo de clustering

Función de coste

- La partición que minimiza la función de coste basandose en el error cuadrático medio se considera óptima y se llama **partición de varianza mínima**
- El error cuadrático medio es un criterio apropiado cuando los clusters son compactos y bien separados
- Sin embargo, es un criterio muy sensible a la existencia de objetos atípicos (**outliers**), y induce a dividir de manera incorrecta grandes clusters en pequeños conglomerados (Duda et al.2001)

3- Validación del clustering

- Independientemente de si los datos tienen o no una estructura de clustering, los algoritmos de clustering siempre producen una partición de los datos en clusters.
- Diferentes estrategias de clustering generalmente revelan diferentes estructuras de clustering.
- La efectividad de los criterios y estandares de evaluación del clustering es crucial.

3- Validación del clustering

- **Indices externos:** son basados en alguna estructura específica que refleja la información a priori sobre la estructura del clustering en los datos (Rand index, Jaccard coefficient, etc).
- **Indices internos:** evaluan la estructura del clustering exclusivamente a partir de los datos sin ninguna información externa (Matriz de proximidad).
- **Indices relativos:** comparan el clustering con otras estructuras de clustering obtenidas a partir de la aplicación de diferentes algoritmos de clustering.

Índice

Clustering particional

Objetivo

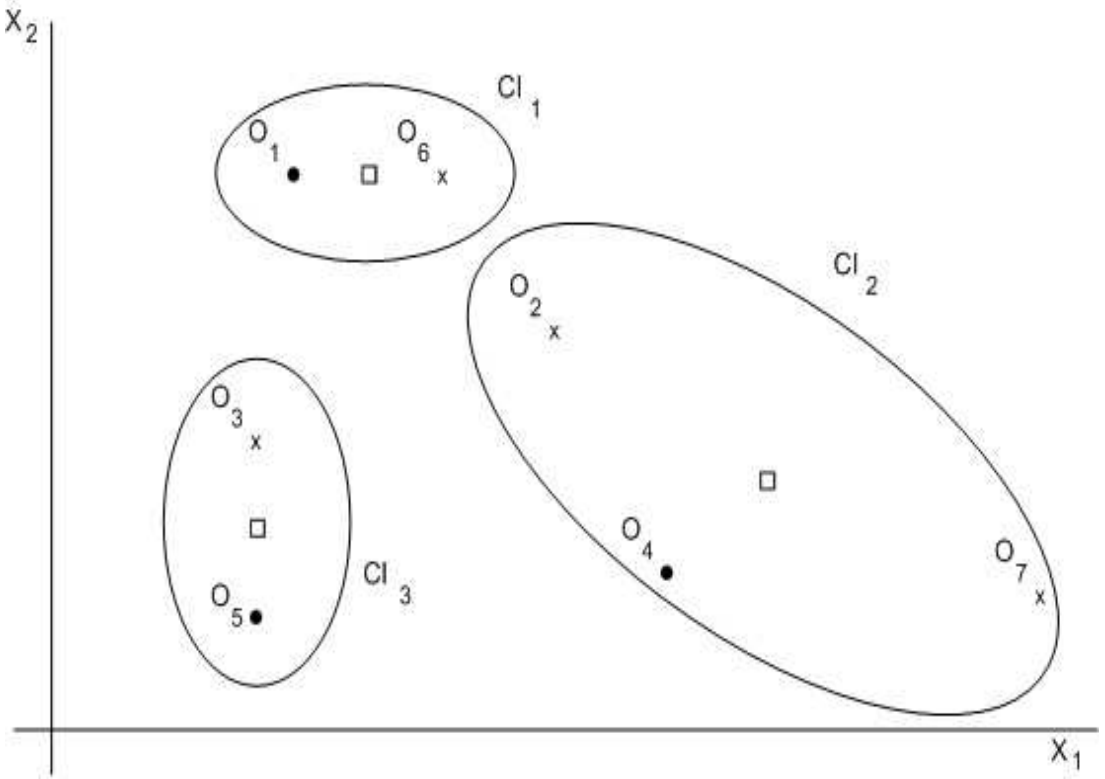
- **Obtener una partición** de los objetos en k grupos o clusters,
- Cada uno de los k **clusters posibles** debe tener por lo menos un objeto,
- Los clusters obtenidos deben de ser disjuntos.

Clustering particional

Método de Forgy (1965)

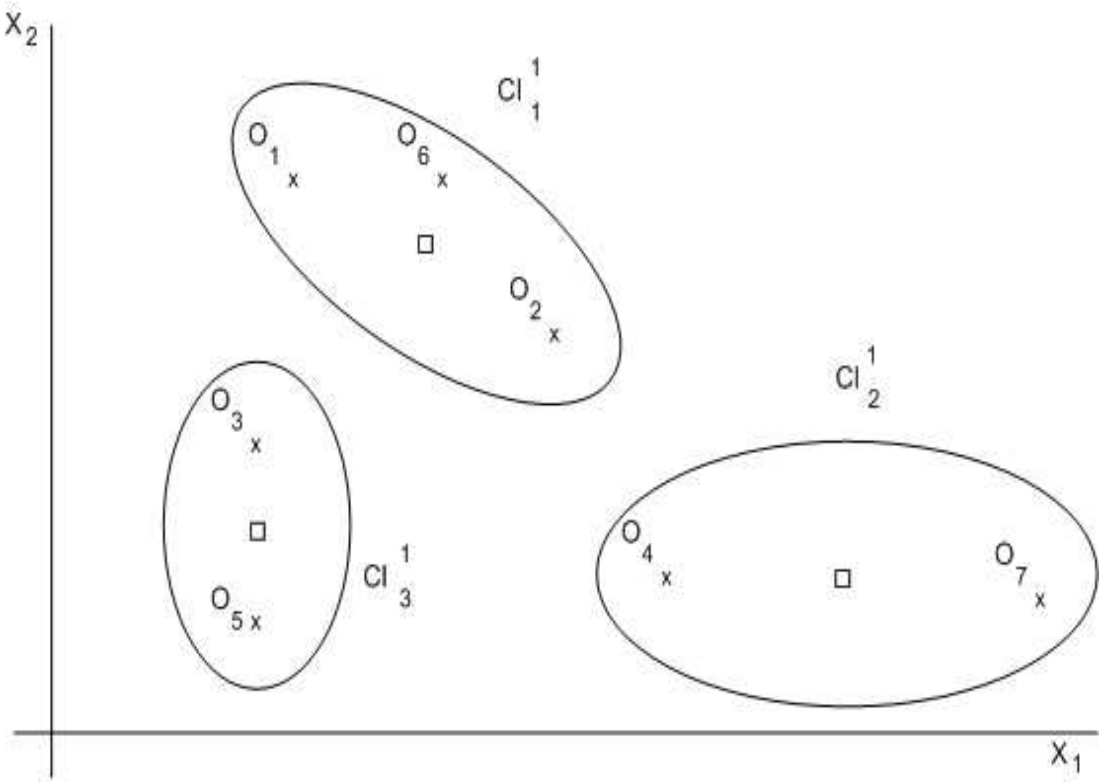
1. Comenzar con cualquier configuración inicial
   Ir al Paso 2 si se comienza por un conjunto de k centroides
   Ir al Paso 3 si se comienza por una partición del conjunto de objetos en k grupos
2. Asignar cada objeto a clasificar al centroide más próximo. Los **centroides permanecen fijos** en este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

Clustering particional: Algoritmo de Forgy



Partiendo de los objetos O_1, O_4, O_5 como centroides iniciales, se obtiene el clustering dado por los clusters: $Cl_1^0 = \{O_1, O_6\}$, $Cl_2^0 = \{O_2, O_4, O_7\}$,$Cl_3^0 = \{O_3, O_5\}$

Clustering particional: Algoritmo de Forgy



Clustering obtenido en la iteración 1: $Cl_1^1 = \{O_1, O_2, O_6\}$, $Cl_2^1 = \{O_4, O_7\}$,$Cl_3^1 = \{O_3, O_5\}$

K-means (Forgy, 1965; MacQueen, 1967)

- El algoritmo K-means es uno de los algoritmos de clustering más populares (Duda et al.,2001)
- El K-means busca la partición óptima de los datos en K clusters utilizando un procedimiento iterativo basado en minimizar el error cuadrático medio

Clustering particional

K-means (Forgy, 1965; MacQueen, 1967)

1. Considerar los k primeros elementos del fichero como k conglomerados con un único elemento
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**
3. Después de que todos los objetos hayan sido asignados en el paso anterior, calcular los centroides de los conglomerados y reasignar cada objeto al centroide más cercano
4. Repetir los pasos 2 y 3 hasta que se alcance un determinado criterio de parada

K-means (Forgy, 1965; MacQueen, 1967)

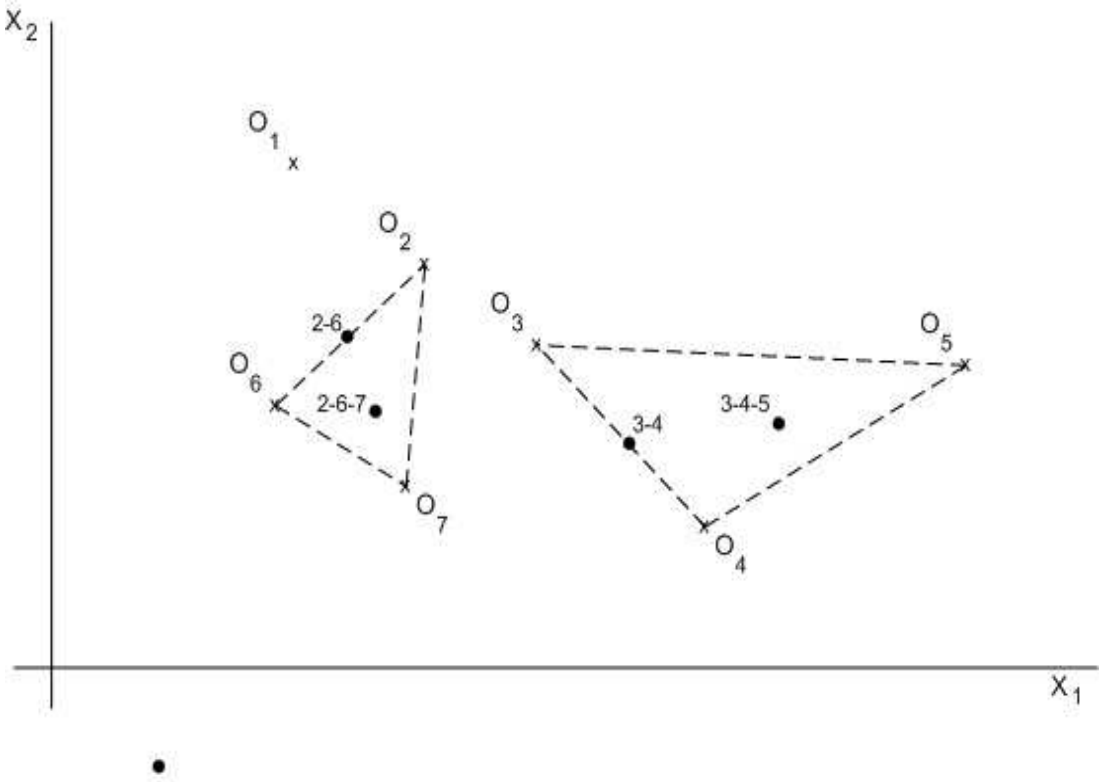
1. Partir de una partición inicial (aleatoria o basada en algun criterio) de los datos en K clusters. Calcular los prototipos de los K clusters $M = [\mathbf{m}_1, ..., \mathbf{m}_K]$;
2. Asignar cada objeto del conjunto de datos al cluster más cercano:

$$\mathbf{x}_j \in C_l, \quad si \|\mathbf{x}_j - \mathbf{m}_l\| \leq \|\mathbf{x}_j - \mathbf{m}_i\|,$$

para $j = 1, ..., N, i \neq l, \quad i = 1, ..., K$

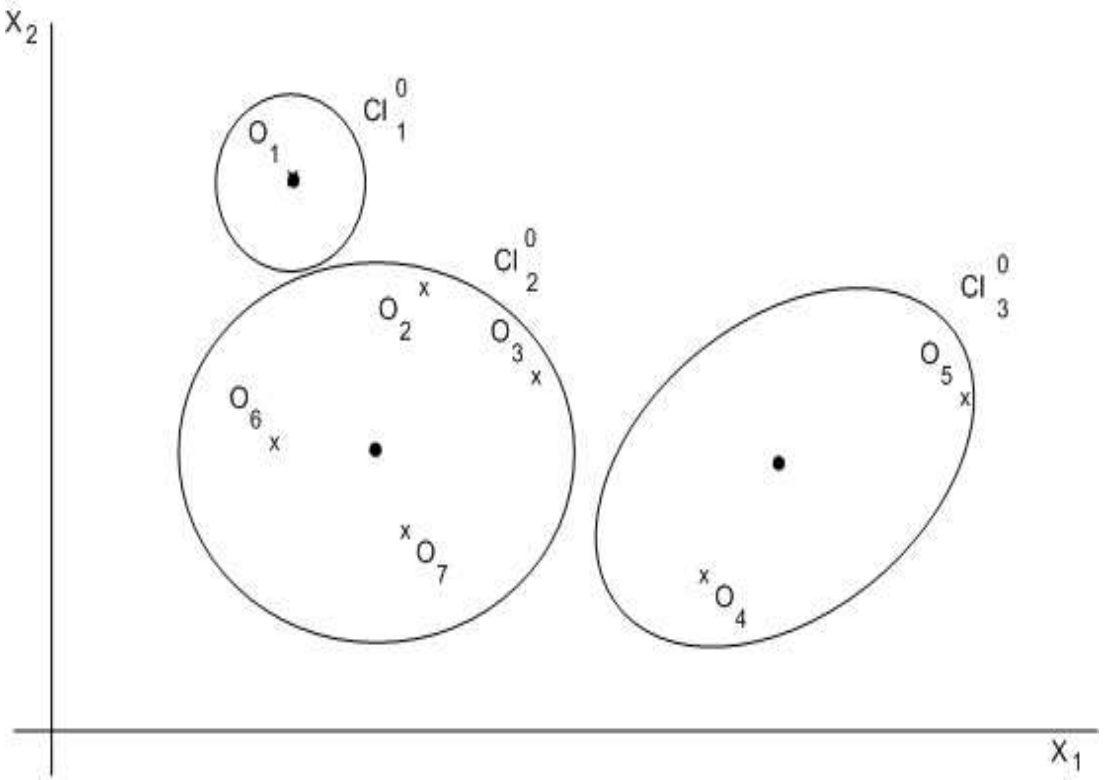
3. Recalcular los prototipos de los clusters basandose en la partición actual: $\mathbf{m}_i = \frac{1}{N_i} \sum_{x_j \in C_i} x_j$
4. Repetir los pasos 2 y 3 hasta que se alcance un determinado criterio de parada

Algoritmo de k-medias de McQueen



Partición en la iteración 1. Paso 2

Algoritmo de k-medias de McQueen



Partición en la iteración 1. Paso 3

K-means: ventajas

- El K-means es un algoritmo facil de implementar
- Funciona bien en muchos problemas prácticos, en particular, cuando los clusters son compactos hiperesféricos
- La complejidad temporal del K-means es del orden $O(NKdT)$, donde T es el número de iteraciones.
- El K-means es una buena opción para abordar problemas de clustering en grandes conjuntos de datos

K-means: inconvenientes

- El mayor inconveniente que afecta el K-means es inherente al método de optimización utilizado para minimizar la función de coste.
- El procedimiento iterativo que usa el K-means no garantiza la convergencia hacia un óptimo global
- Diferentes puntos iniciales generalmente conducen a diferentes centroides finales (clusters initialization)
- No hay un método universal para determinar la partición inicial ni para identificar el número de clusters

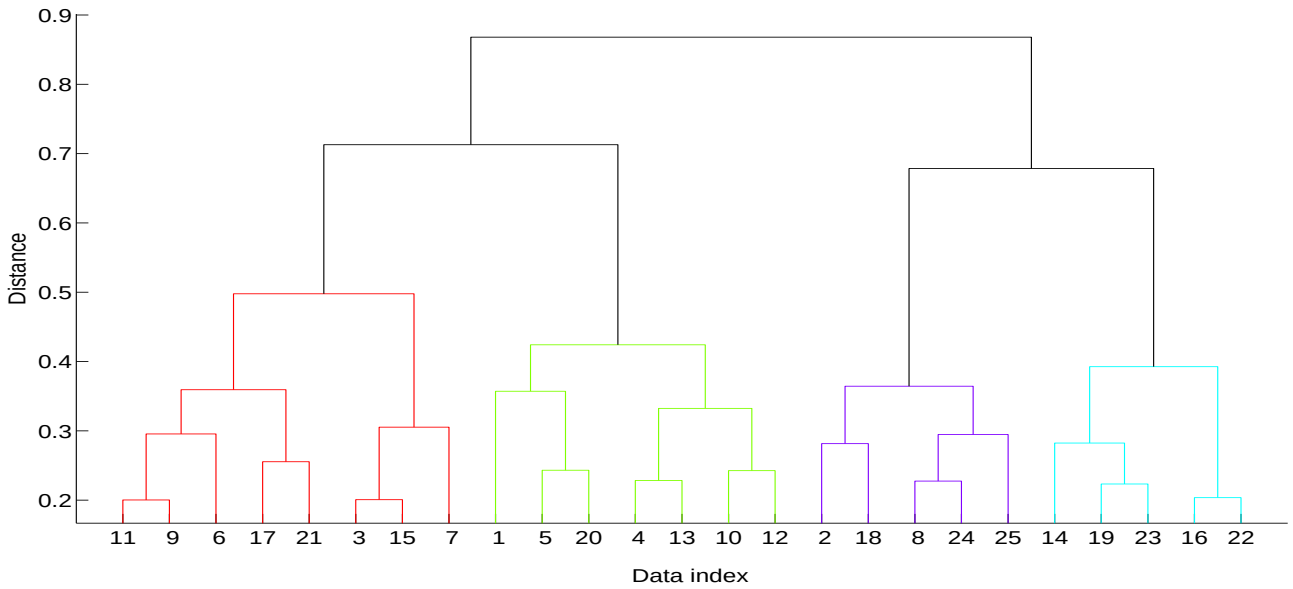
Índice

Clustering Jerárquico

Introducción

- Los algoritmos de Clustering jerárquico son diferentes de los algoritmos de clustering particional
- En vez de producir un único clustering (una partición única), producen una **jerárquia de clusterings** o particiones basandose en la matriz de proximidad.
- Los resultados del clustering son generalmente representados mediante un árbol binario o dendrograma.

Clustering Jerárquico



Ejemplo de un dendrograma correspondiente a un clustering jerárquico

Clustering Jerárquico

- Los algortimos se componen de tantos pasos como instancias en la base de datos
- En cada paso t un nuevo clustering es obtenido basandose en el clustering obtenido en el paso anterior $t - 1$
- Existen dos categorías principales de estos algoritmos: los **aglomerativos o ascendentes** y los **divisivos o descendentes**

Clustering Jerárquico

- Los algortimos aglomerativos producen una secuencia de clusterings, $\Re_0, \Re_1, ..., \Re_{N-1}$, con un número descendiente de clusters en cada paso,
- El clustering producido en cada paso es el resultado de juntar en uno dos clusters del clustering anterior

Clustering jerárquico aglomerativo

Esquema aglomerativo general

1. Partir de N clusters. Calcular la matriz de proximidad (función de distancia) para los N clusters;
2. En la matriz de proximidad, buscar la distancia mínima $D(C_i, C_j) = minD(C_m, C_l) \quad 1 \leqslant m, l \leqslant N \quad m \neq l$, y combinar los cluster C_i y C_j para formar un nuevo cluster C_{ij};
3. Actualizar la matriz de proximidad calculando la distancia entre el cluster C_{ij} y los demás clusters;
4. Repetir los pasos 2 y 3 hasta obtener un único cluster.

Enlaces entre clusters (clustering linkage)

Lance and Williams (1967)

Las posibles medidas de distancia entre un cluster C_l y un nuevo cluster C_{ij} obtenido agrupando los clusters C_i y C_j pueden resumirse mediante la siguiente ecuación:

$$D(C_l, (C_i, C_j)) = \alpha_i D(C_l, C_i) + \alpha_j D(C_l, C_j) + \beta D(C_i, C_j) + \gamma |D(C_l, C_i) - D(C_l, C_j)|$$

donde $\alpha_i, \alpha_j, \beta$ and γ son parámetros cuyos valores determinan la función de disimilitud. n_i, n_j and n_l representan el número de objetos en los clusters C_i, C_j y C_l respectivamente. La ecuación anterior es una definición recursiva de la noción de distancia entre dos clusters.

Enlaces entre clusters (clustering linkage)

Clustering algorithms	α_i	α_j	β	γ
Single linkage (nearest neighbor)	1/2	1/2	0	-1/2
Complete linkage (farthest neighbor)	1/2	1/2	0	1/2
Group average linkage (UPGMA)	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	0	0
Weighted average linkage (WPGMA)	1/2	1/2	0	0
Median linkage (WPGMC)	1/2	1/2	-1/4	0
Centroid linkage (UPGMC)	$\frac{n_i}{n_i+n_j}$	$\frac{n_j}{n_i+n_j}$	$\frac{-n_i n_j}{(n_i+n_j)^2}$	0
Ward's method (Minimun variance method)	$\frac{n_i+n_l}{n_i+n_j+n_l}$	$\frac{n_j+n_l}{n_i+n_j+n_l}$	$\frac{-n_l}{n_i+n_j+n_l}$	0

U: unweighted; W: weighted; PGM: pair group method; A: average; C: centroid

Clustering jerárquico aglomerativo

Ejemplo1: Clustering jerárquico basado en el enlace medio.

Dado el siguiente conjunto de datos:

	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

La matriz de proximidad a tener en cuenta para formar el primer clustering viene dada por:

$$P_0 = \begin{bmatrix} 0 & 3 & 14 & 62 & 40 \\ 3 & 0 & 29 & 61 & 41 \\ 14 & 29 & 0 & 94 & 72 \\ 62 & 61 & 94 & 0 & 2 \\ 40 & 41 & 72 & 2 & 0 \end{bmatrix}$$

y da lugar al clustering $C_1 = \{1, 2, 3, \{4, 5\}\}$ que define una partición de los datos en 4 clusters.

Ejemplo1: Clustering jerárquico basado en el enlace medio.

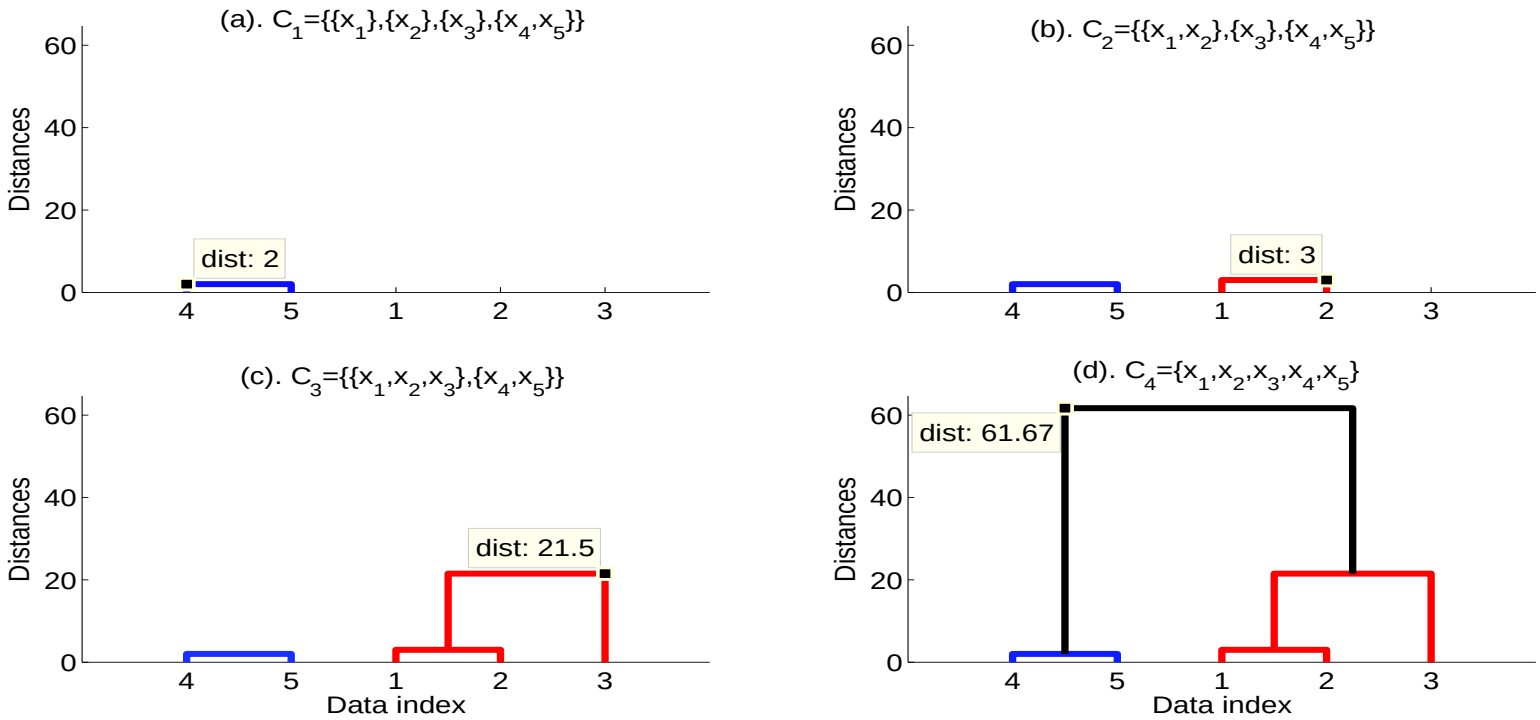
Matrices de proximidad a tener en cuenta para generar los clustering: C_2, C_3 y C_4 respectivamente

$$P_1 = \begin{bmatrix} 0 & 3 & 14 & 51 \\ 3 & 0 & 29 & 51 \\ 14 & 29 & 0 & 83 \\ 51 & 51 & 83 & 0 \end{bmatrix}$$

$$P_2 = \begin{bmatrix} 0 & 21{,}5 & 51 \\ 21{,}5 & 0 & 83 \\ 51 & 83 & 0 \end{bmatrix}$$

$$P_3 = \begin{bmatrix} 0 & 61{,}67 \\ 61{,}67 & 0 \end{bmatrix}$$

Dendrogramas correspondientes a las distintas particiones.



(a) Clustering C_1 compuesto de $(N - 1 = 4)$ clusters. (b) Clustering C_2 compuesto de $(N - 2 = 3)$ clusters. (c) Clustering C_3 compuesto de $(N - 3 = 2)$ clusters. (d) Clustering C_4 compuesto de $(N - 4 = 1)$ cluster.