

Clustering



Ramírez Morales Luz Janet
Victoria Benavides Isaac
Díaz Maldonado Jesús Renato

Clustering

El clustering o agrupamiento es una técnica de minería de datos que consiste en dividir un conjunto de datos en grupos o clusters, de tal forma que los elementos dentro de cada grupo sean similares entre sí y diferentes a los de otros grupos. El objetivo del clustering es encontrar patrones o estructuras ocultas en los datos que puedan ser útiles para la clasificación, la segmentación, la compresión o la visualización de los mismos.

Clustering

Existen diferentes tipos de clustering según el criterio que se utilice para definir la similitud entre los elementos, el algoritmo que se aplique para formar los grupos, o la forma que tengan los grupos resultantes. Algunos ejemplos de tipos de clustering son: jerárquico, particional, basado en densidad, basado en centroides, basado en modelos, etc. Cada tipo de clustering tiene sus ventajas y desventajas, y se adapta mejor a ciertos tipos de datos o problemas.

EJEMPLO

Ejemplo simple de agrupamiento, Dunham, M.H.(2002)

Una empresa internacional de catálogos en línea desea agrupar a sus clientes en función de características comunes. La administración de la empresa no tiene etiquetas predefinidas para estos grupos. Según el resultado de la agrupación, dirigirán las campañas de marketing y publicidad a los diferentes grupos. La información que tienen sobre los clientes incluye grupos con mayores ingresos y al menos un título universitario.' El último grupo tiene hijos. Se habrían encontrado diferentes agrupamientos al examinar la edad o el estado del manto, ingresos, edad, número de hijos, estado civil y educación. (p.125)

TABLE 5.1: Sample Data for Example 5.1

Income	Age	Children	Marital Status	Education
\$25,000	35	3	Single	High school
\$15,000	25	1	Married	High school
\$20,000	40	0	Single	High school
\$30,000	20	0	Divorced	High school
\$20,000	25	3	Divorced	College
\$70,000	60	0	Married	College
\$90,000	30	0	Married	Graduate school
\$200,000	45	5	Married	Graduate school
\$100,000	50	2	Divorced	College

Fuente: Dunham, M.H.(2002)

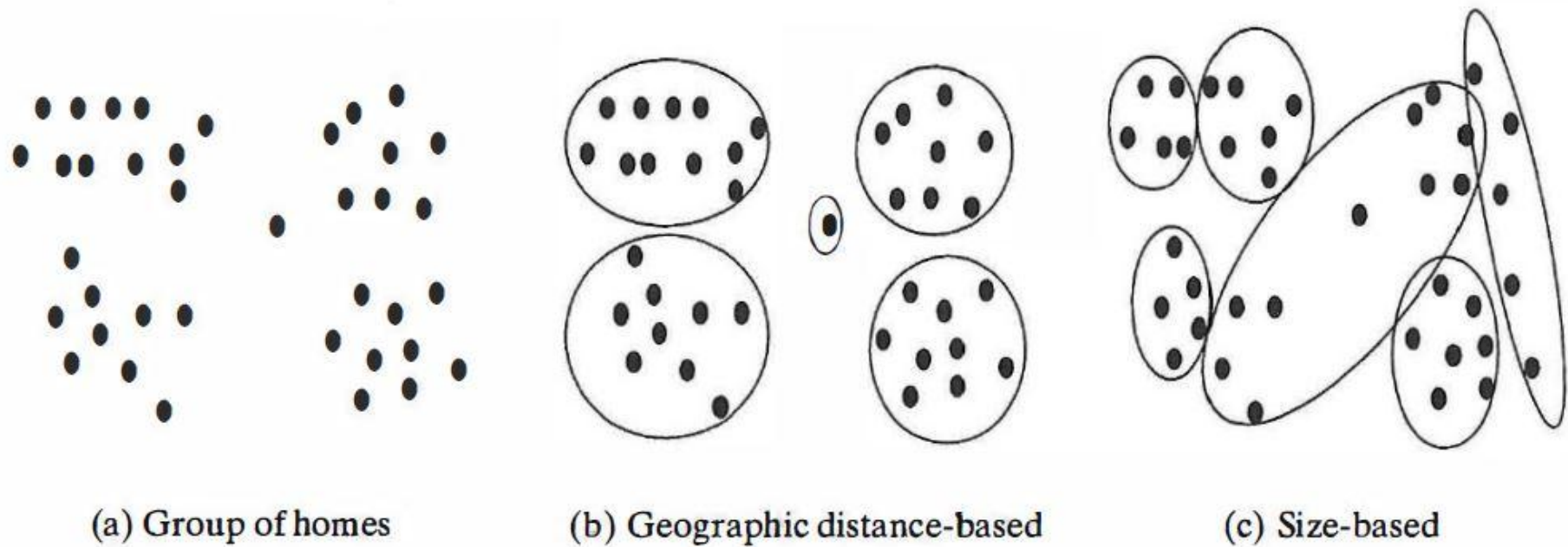



FIGURE 5.1: Different clustering attributes.

Fuente: Dunham, M.H.(2002)

Algunos problemas interesantes que ocurren al aplicar la agrupación en clústeres a una base de datos pueden ser los siguientes:



El manejo de valores atípicos es difícil.



Los datos dinámicos en la base de datos implican que la membresía del clúster puede cambiar con el tiempo.



Interpretar el significado semántico de cada grupo puede ser difícil.



No hay una respuesta correcta para un problema de agrupamiento.

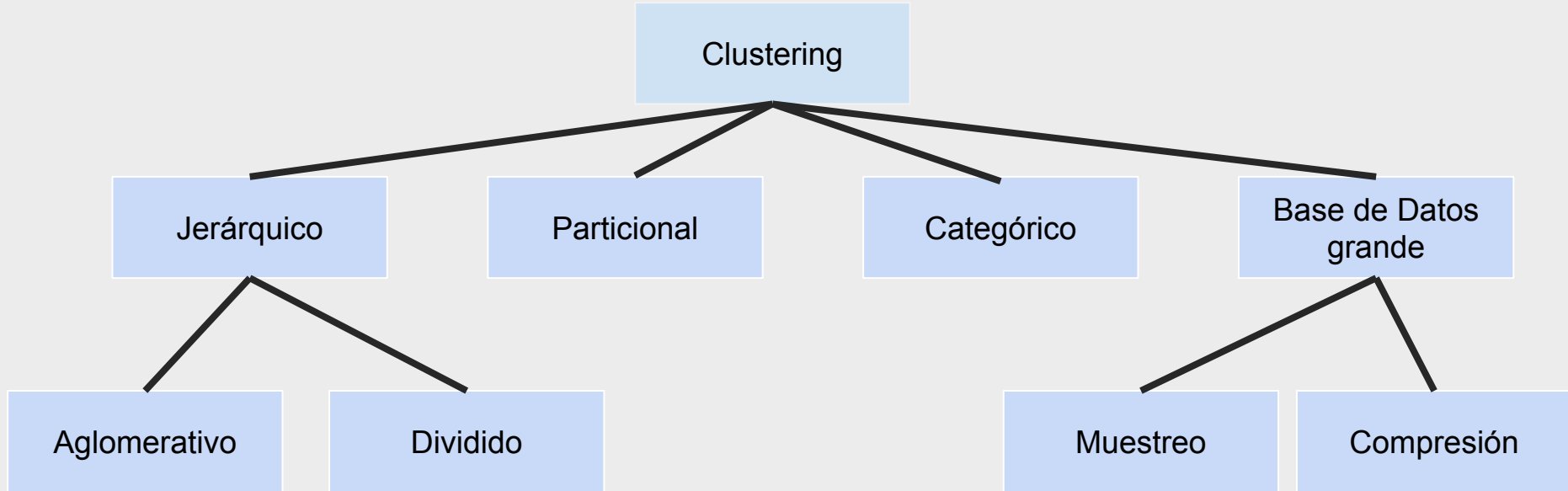


Saber qué datos se deben usar para la agrupación.

DEFINICIÓN 5.1.

“Dada una base de datos $D = \{t_1, t_2, \dots, t_n\}$ de tuplas y un valor entero k , el problema de agrupamiento es definir un mapeo $f : D \rightarrow \{1, \dots, k\}$ donde cada t_i se asigna a un grupo K_j , $1 \leq j \leq k$. Un clúster, K_j , contiene precisamente esas tuplas asignadas a él; es decir, $K_j = \{t_i \mid f(t_i) = j, 1 \leq i \leq n, \text{ y } t_i \in D\}$.”(Fuente: Dunham, M.H, 2002, p. 127)

Clasificación de algoritmos de agrupamiento



5.2 MEDIDAS DE SIMILITUD Y DISTANCIA

¿Qué es la similitud?

$$D = \{t_1, t_2, \dots, t_n\}$$

Dada una base de datos de n-tuplas, se define un mapeo $f : D \rightarrow \{1, \dots, k\}$, una medida de $\text{sim}(t_i, t_j)$ entre dos tuplas cualesquiera es deseable que una tupla de un clúster se parezca más a las tuplas de ese clúster que a las tuplas de fuera de él.

$$K_j, \forall t_{jl}, t_{jm} \in K_j \text{ y } t_i \notin K_j, \text{sim}(t_{jl}, t_{jm}) > \text{sim}(t_{jl}, t_i)$$

Dado un clúster K_m de N puntos $\{t_{m1}, t_{m2}, \dots, t_{mN}\}$

Centroide (C_m)

$$C_m = \frac{\sum_{i=1}^N (t_{mi})}{N}$$

Centro del clúster.

Algunos algoritmos fijan un centro único llamado “Medoide”.

M_m

Radio (R_m)

$$R_m = \sqrt{\frac{\sum_{i=1}^N (t_{mi} - C_m)^2}{N}}$$

Distancia media promedio entre cualquier punto y el centroide

Diámetro (D_m)

$$D_m = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^N (t_{mi} - t_{mj})^2}{(N)(N-1)}}$$

Distancia media entre pares dentro de un grupo.

Dado un clúster K_m de N puntos $\{t_{m1}, t_{m2}, \dots, t_{mN}\}$

Centroide (C_m)

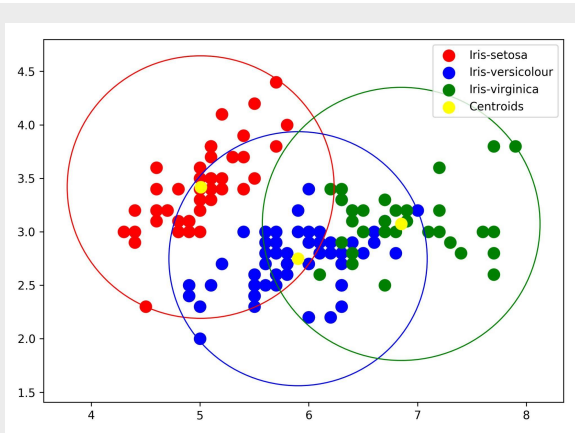


Imagen tomada de: [Find Cluster Diameter and Associated Cluster Points with KMeans Clustering \(scikit learn\)](#)

Radio (R_m)

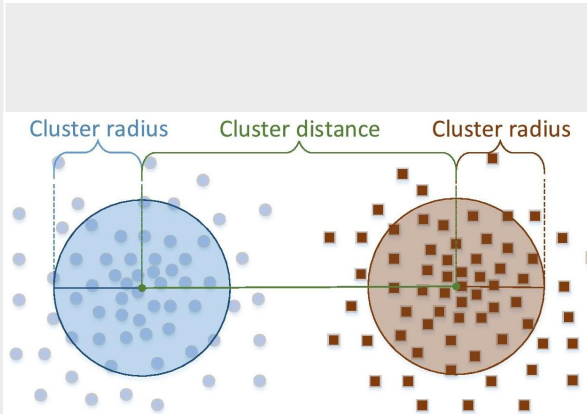


Imagen tomada de: [Variations on the Clustering Algorithm BIRCH](#)

Diámetro (D_m)

$$D_m = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^N (t_{mi} - t_{mj})^2}{(N)(N-1)}}$$

El punto más distante da inicio a un nuevo cluster, para ciertos algoritmos como “divisive”

¿Y si hay más de un clúster? Dados los clusters K_i, K_j

Enlace simple

- Distancia más pequeña entre un elemento en un clúster y otro elemento de otro
$$dis(K_i, K_j) = \min \left(dis(t_{il}, t_{jm}) \right) \forall t_{il} \in K_i \notin K_j, \forall t_{jm} \in K_j \notin K_i$$

Enlace completo

- Distancia más larga entre un elemento de un clúster y otro
$$dis(K_i, K_j) = \max \left(dis(t_{il}, t_{jm}) \right) \forall t_{il} \in K_i \notin K_j, \forall t_{jm} \in K_j \notin K_i$$

Promedio

- Distancia promedio entre un elemento de un clúster y otro
$$dis(K_i, K_j) = \text{mean} \left(dis(t_{il}, t_{jm}) \right) \forall t_{il} \in K_i \notin K_j, \forall t_{jm} \in K_j \notin K_i$$

Centroide y medoide

- La distancia del centroide se define como la distancia entre los centroides $dis(K_i, K_j) = dis(C_i, C_j)$, donde C_i es el centroide para K_i y así con C_j
- La distancia entre clusters se puede definir como la distancia entre los medoides $dis(K_i, K_j) = dis(M_i, M_j)$

5.3 OUTLIERS

Valores atípicos

- Puntos de muestra con valores muy diferentes a los valores restantes del conjunto de datos
- Representan errores en los datos o simplemente sean datos correctos que simplemente son muy diferentes del resto de los datos

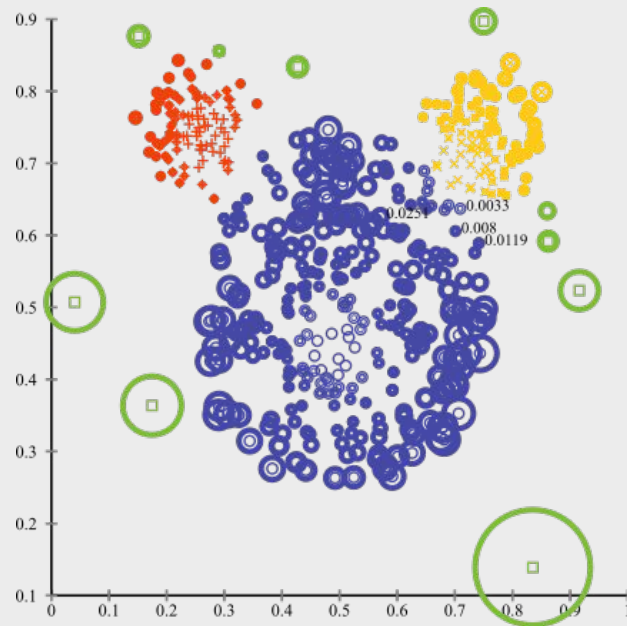


Imagen tomada de: [Anomaly detection based on clustering](#)

Ejemplo

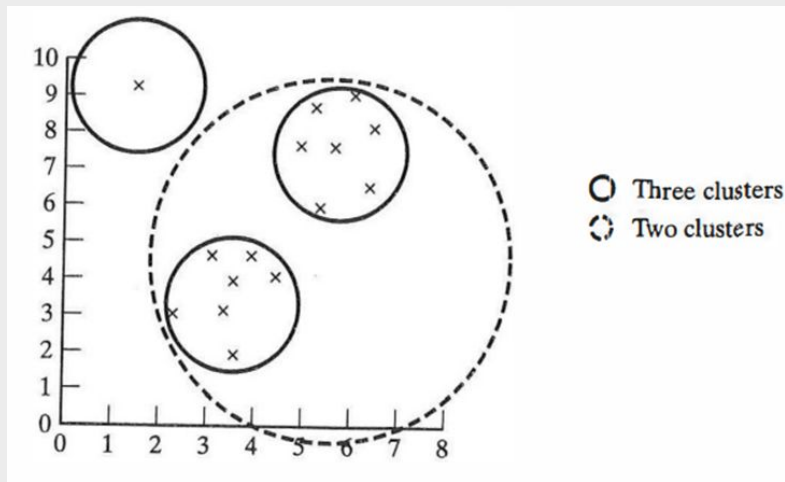
Analizando la altura de un conjunto de personas, si una persona mide 2.5m, probablemente sea visto como un “outlier”.



Para los clusters...

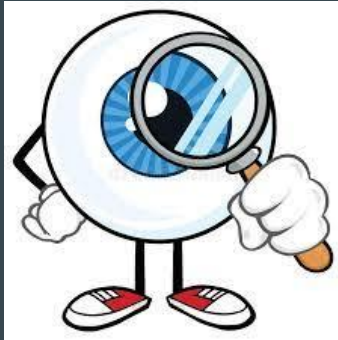
Algunas técnicas de agrupamiento no se desempeñan correctamente en la presencia de valores atípicos.

Los algoritmos de clustering pueden encontrar y eliminar los valores atípicos para asegurar que su desempeño sea mejor.



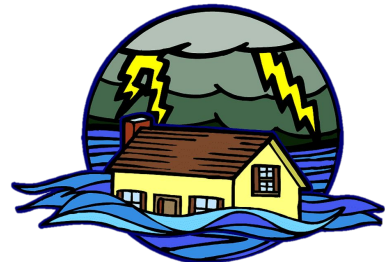
Fuente: Dunham, M.H.(2002)

¡CUIDADO!



Se debe de tener cuidado al remover los valores atípicos.

Ej.: Remover valores altos de niveles de agua no permitirían que los algoritmos de minería de datos trabajen efectivamente porque no habría datos que muestren que realmente hayan ocurrido inundaciones alguna vez.

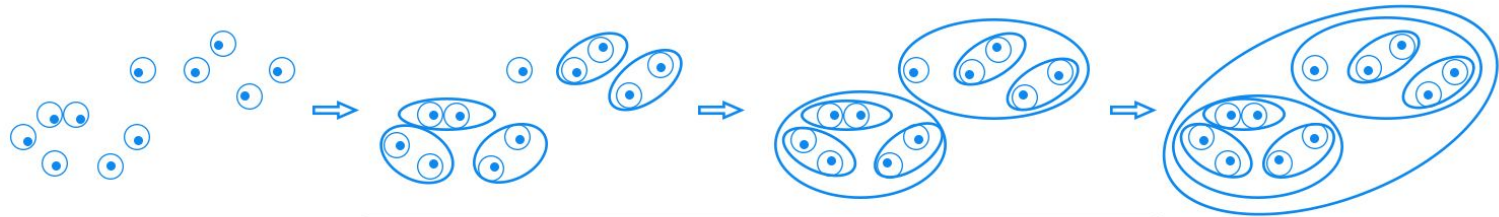


5.4 HIERARCHICAL ALGORITHMS

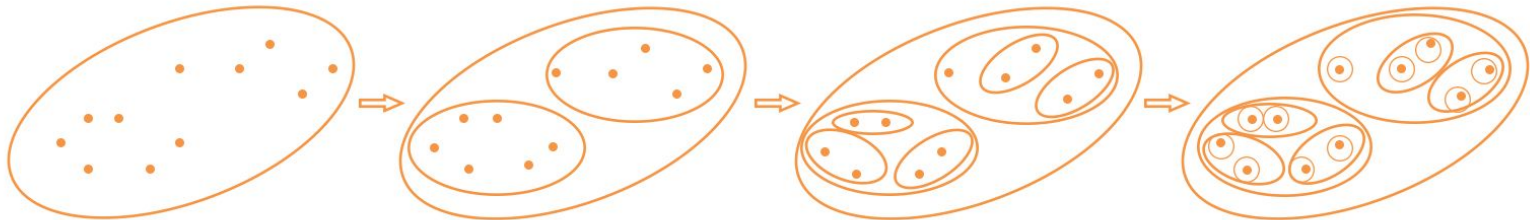
Algoritmos jerárquicos

Una de las técnicas más utilizadas para el clustering son los algoritmos jerárquicos que dividen el conjunto de objetos en subgrupos a través de un dendrograma, que es la representación gráfica de la estructura jerárquica resultante.

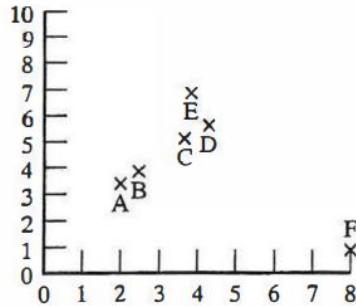
Agglomerative Hierarchical Clustering



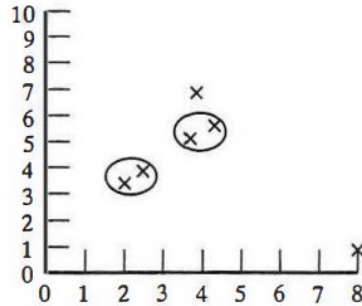
Divisive Hierarchical Clustering



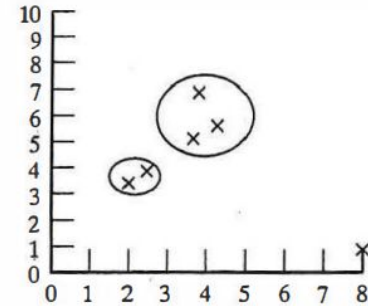
Ejemplo 5.2



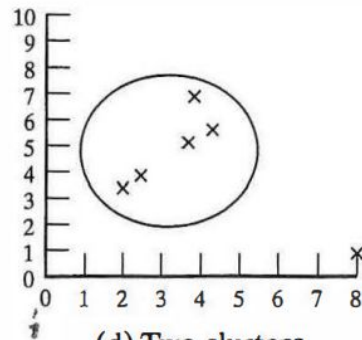
(a) Six clusters



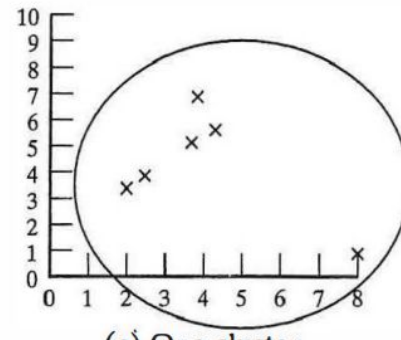
(b) Four clusters



(c) Three clusters



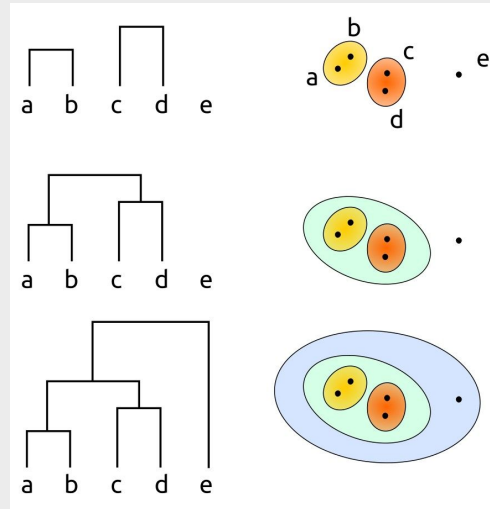
(d) Two clusters



(e) One cluster

Algoritmos aglomerativos

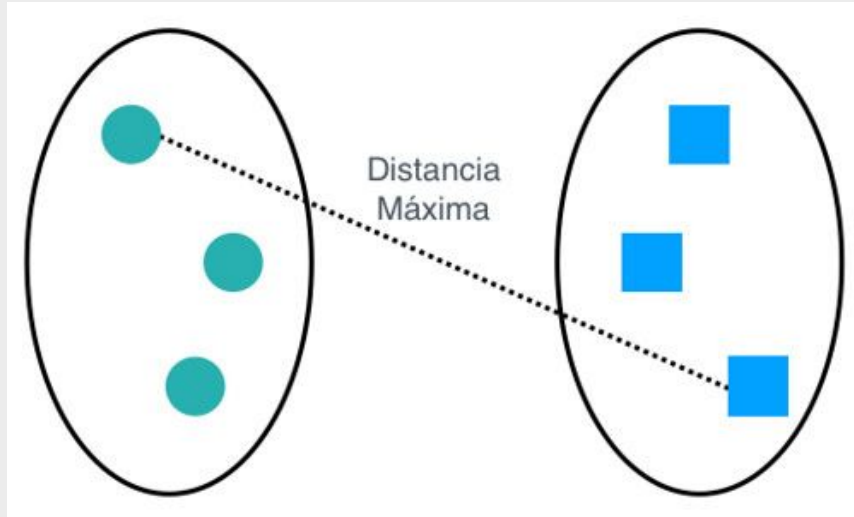
Los algoritmos aglomerativos son una técnica de agrupamiento en el aprendizaje automático no supervisado que comienza con cada elemento en su propio grupo y luego los fusiona iterativamente hasta que todos los elementos pertenecen a un solo grupo.



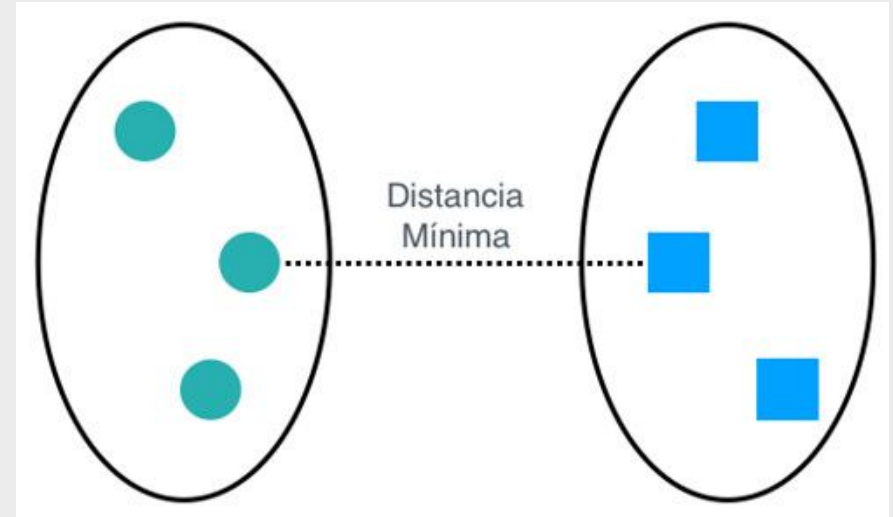
ALGORITHM 5.1

```
1  Input:
2    D = {t1, t2, ..., tn} // Conjunto de elementos
3    A // Matriz de adyacencia que muestra la distancia entre elementos
4
5  Output:
6    DE // Dendrograma representado como un conjunto de triples ordenados
7  Agglomerative algorithm:
8    d = 0; // Inicializa la distancia umbral en cero
9    k = n; // Inicializa el número de grupos en el número total de elementos
10   K = {{t1}, ..., {tn}}; // Inicializa cada elemento en su propio grupo
11   DE = {(d, k, K)}; // Inicializa el dendrograma con cada elemento en su propio grupo
12
13   repeat // Repetir hasta que todos los elementos pertenezcan a un solo grupo
14     oldk = k; // Guarda el número anterior de grupos
15     d = d + 1; // Aumenta la distancia umbral en 1
16     Ad = Vertex adjacency matrix for graph with threshold distance of d; // Calcula la matriz de adyacencia
17                                                //para la distancia umbral actual
18     (k, K) = NewClusters(Ad, D); // Calcula los nuevos grupos usando la matriz de adyacencia
19                                   //y el conjunto de elementos
20     if oldk != k then
21       DE = DE U (d, k, K); // Agrega los nuevos grupos al dendrograma
22     end if
23   until k = 1; // Todos los elementos pertenecen a un solo grupo
```


Métodos para medir la similitud entre grupos de datos

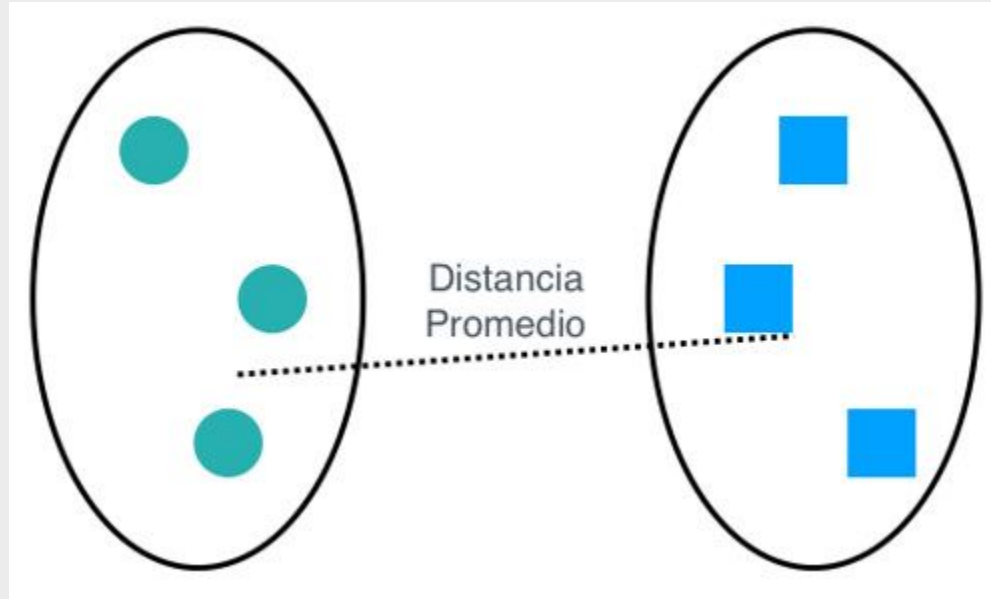


Complete Link



Single Link Technique

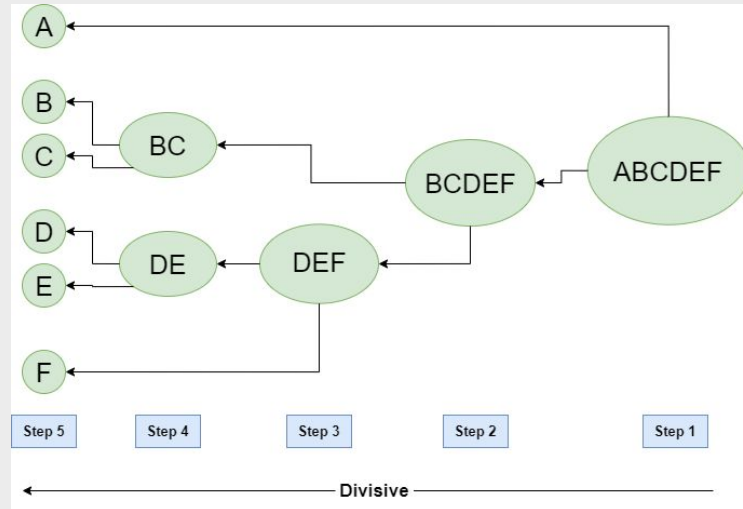
Métodos para medir la similitud entre grupos de datos



Average Link.

Agrupación divisiva

Es un enfoque de clustering jerárquico en el que todos los elementos se colocan inicialmente en un solo cluster y los clusters se dividen repetidamente en dos hasta que cada elemento se encuentra en su propio cluster. El objetivo es dividir los clusters en los que algunos elementos no están lo suficientemente cerca de otros elementos.



Referencias

Dunham, M. H. (2002). Data mining: introductory and advanced topics. Prentice Hall . Capítulo 5 (125-163)