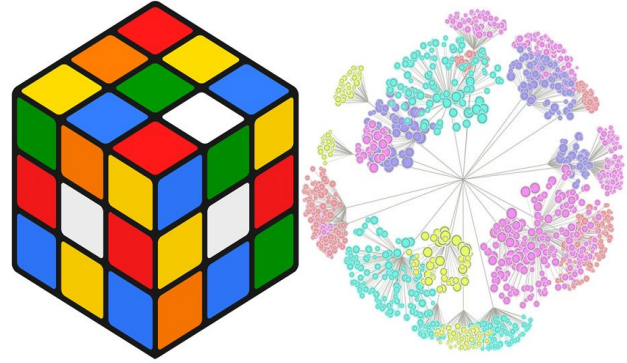
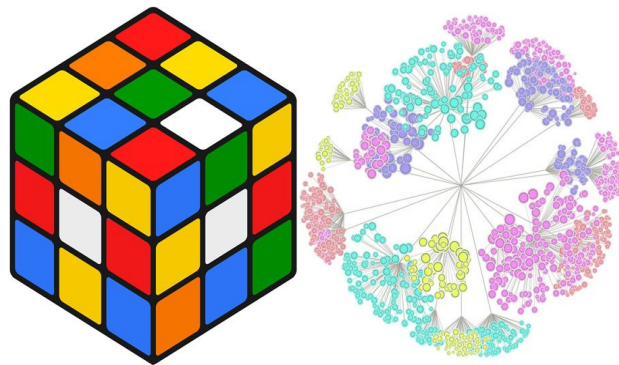


CLUSTERING



Equipo 5 y 9

5-1



¿QUE ES ?

EL Clustering es similar a la clasificación en el sentido que los datos se agrupan, a diferencia que en la clasificación los datos, los grupos no están predefinidos.

Se han propuesto muchas definiciones para los clusters:

- Conjunto de elementos similares. Los elementos de diferentes clusters no son iguales.
- La distancia entre los puntos de un clúster es menor que la distancia entre un punto en el clúster y cualquier punto fuera de él.

EJEMPLO 5.1

Una empresa internacional de catálogos en línea desea agrupar a sus clientes en base a características comunes.

La dirección de la empresa no tiene ninguna etiqueta predefinida para estos grupos.

En función del resultado de la agrupación, dirigirán las campañas de marketing y publicidad a los distintos grupos. La información que tienen sobre los clientes incluye:

Ingresos	Edad	Hijos	Estado civil	Educación
\$25,000	35	3	Soltero	Preparatoria
\$15,000	25	1	Casado	Preparatoria
\$20,000	40	0	Soltero	Preparatoria
\$30,000	20	0	Divorciado	Preparatoria
\$20,000	25	3	Divorciado	Universidad
\$70,000	60	0	Casado	Universidad
\$90,000	30	0	Casado	Maestría
\$200,000	45	5	Casado	Maestría
\$100,00	50	2	Divorciado	Universidad

AGRUPACIÓN

Ingresos, edad, número de hijos, estado civil y educación. Sin la muestra algunas tuplas de esta base de datos para clientes de Estados Unidos. Dependiendo del tipo de publicidad no todos los atributos son importantes. Por ejemplo, supongamos que el anuncio es para una oferta especial de ropa para niños.

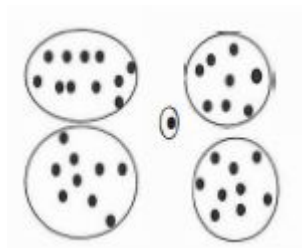
Podríamos dirigir el anuncio sólo a las personas con niños.

- El primer grupo de personas tiene hijos pequeños y un nivel de estudios superior.
- El segundo grupo es similar pero no tiene hijos.
- El tercer grupo tiene tanto hijos como un título universitario.
- Los dos últimos grupos tienen mayores ingresos y al menos un título universitario.
- El último grupo tiene hijos. Se habrían encontrado diferentes agrupaciones por la edad o el estado del manto.



a) Grupos de hogares

Aquí se muestra un grupo de viviendas en una zona geográfica



(b) Basado en la distancia geográfica

En el segundo tipo de agrupación, las viviendas se agrupan en función de el tamaño de la casa.



(c) Basado en el tamaño

Primer tipo de agrupación se basa en la ubicación de la vivienda. Las viviendas que están geográficamente cerca

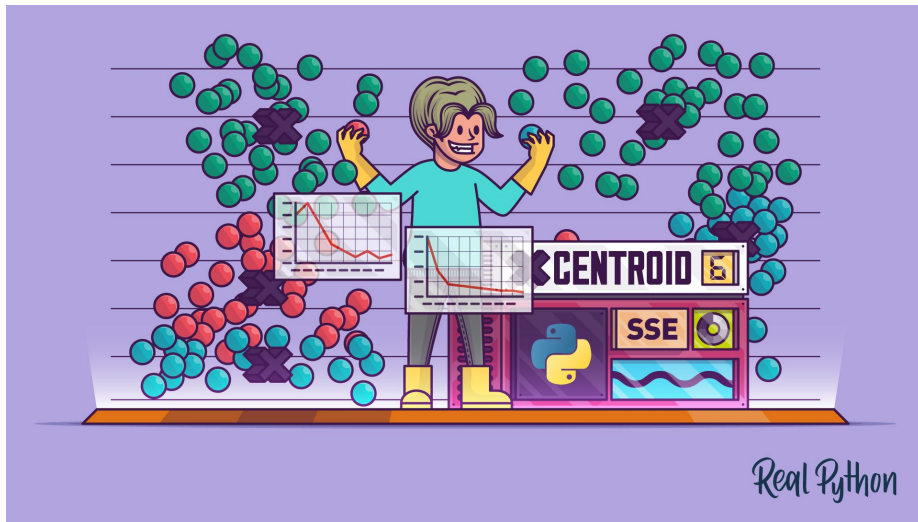
¿PARA QUÉ SIRVE ?

El clustering se ha utilizado en muchos ámbitos de aplicación, como la biología, la medicina, antropología, marketing y economía. Las aplicaciones de la agrupación incluyen las plantas y los animales.

Clasificación de plantas y animales, clasificación de enfermedades, procesamiento de imágenes, reconocimiento de patrones y recuperación de documentos.

Uno de los primeros ámbitos en los que se utilizó la agrupación fue la taxonomía biológica.

Los usos más recientes incluyen el examen de los datos de registro de la web para detectar patrones de uso.



CUANDO LA AGRUPACIÓN SE APLICA A UNA BASE DE DATOS DEL MUNDO REAL, SURGEN MUCHOS PROBLEMAS INTERESANTES COMO:

- La gestión de los valores atípicos es difícil. En este caso, los elementos no caen naturalmente en ningún clúster.
- Los datos dinámicos de la base de datos implican que la pertenencia a un clúster puede cambiar con el tiempo.
- Interpretar el significado semántico de cada clúster puede ser difícil. Con la clasificación, el etiquetado de las clases se conoce de antemano. Sin embargo, con el clustering, puede no ser el caso. Cuando el proceso de clustering termina de crear un conjunto de clusters, el significado exacto de cada cluster puede no ser obvio. Aquí es donde un experto en la materia es necesario para asignar una etiqueta o interpretación a cada cluster.
- No existe una única respuesta correcta a un problema de clustering.
- Otra cuestión relacionada es qué datos deben utilizarse para la agrupación. A diferencia del aprendizaje.

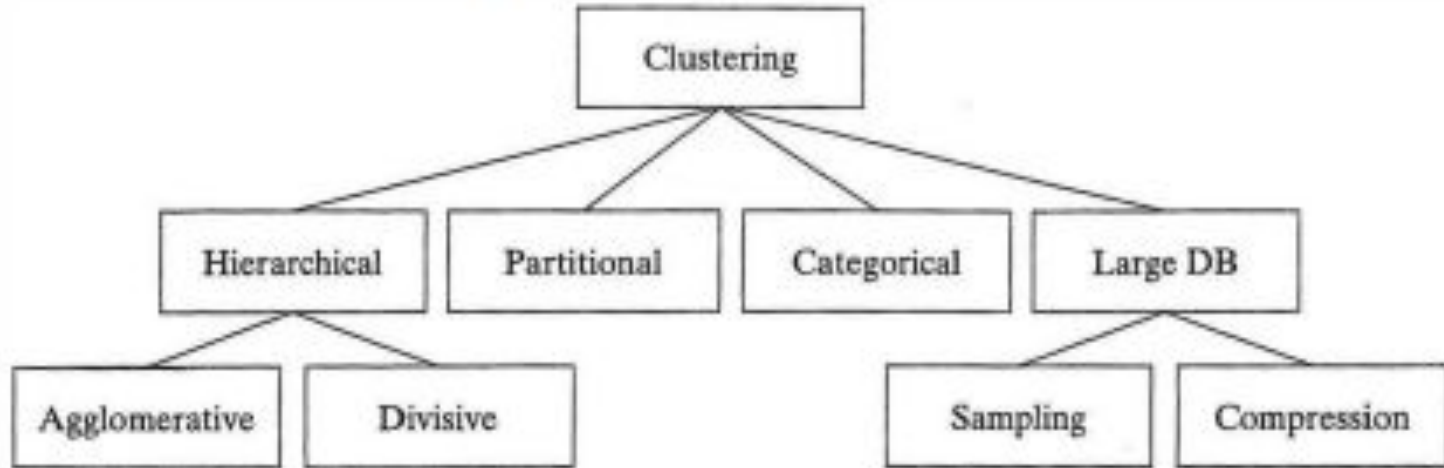
EN RESUMEN

Podemos resumir algunas características básicas de la agrupación (en contraposición a la clasificación):

- No se conoce el (mejor) número de clusters.
- Puede que no haya ningún conocimiento a priori sobre los clusters.
- Los resultados de los clusters son dinámicos.

CLASIFICACIÓN DE LOS ALGORITMOS DE CLUSTERING.

128 Chapter 5 Clustering



JERÁRQUICO

Cada nivel de la jerarquía tiene un conjunto separado de grupos.

- En el nivel más bajo, cada elemento está en su propio grupo único.
- En el nivel más alto, todos los elementos pertenecen al mismo grupo.
- Con agrupamiento jerárquico, el deseado no se ingresa el número de clústeres.

AGLOMERATIVO Y DIVISIVO

Los algoritmos jerárquicos se pueden clasificar como aglomerativos o divisivo.

- "Aglomerativo" implica que los clústeres se crean de forma ascendente.
- Los divisivos funcionan de arriba hacia abajo.

PARTICIONAL

Con la agrupación en clústeres particional, el algoritmo crea solo un conjunto de grupos. Estos enfoques utilizan el número deseado de clústeres para impulsar la forma en que se crea el conjunto final.

Los algoritmos de agrupamiento tradicionales tienden a estar dirigidos a pequeños números bases de datos que caben en la memoria.

CATEGÓRICOS

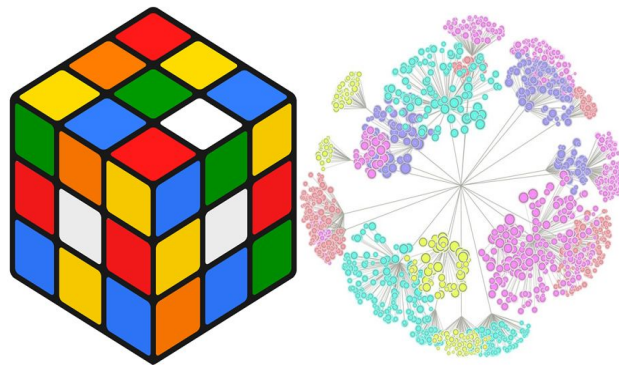
Están dirigidos a bases de datos más grandes, quizás dinámicas. Algoritmos dirigidos a bases de datos más grandes pueden adaptarse a las limitaciones de memoria mediante base de datos o el uso de estructuras de datos, que se pueden comprimir o podar para adaptarse al almacenamiento de memoria independientemente del tamaño de la base de datos.

ALGORITMOS INTRÍNSECOS Y EXTRÍNSECOS

Los algoritmos de agrupación en clústeres también pueden diferir según si producen agrupaciones superpuestas o no superpuestas. Sucesivamente, los grupos que no se superponen pueden considerarse extrínsecos o intrínsecos.

- Las técnicas extrínsecas etiquetan los artículos para ayudar en el proceso de clasificación. Estos algoritmos son los algoritmos de aprendizaje supervisado de clasificación tradicional en los que una formación de entrada especial se utiliza el conjunto.
- Los algoritmos intrínsecos no utilizan etiquetas de categoría a priori, sino que dependen solo en la matriz de adyacencia que contiene la distancia entre los objetos.

5-2



MEDIDAS DE SIMILITUD Y DISTANCIA

Hay muchas propiedades deseables para los clusters creados por una solución a un problema de clustering específico. La más importante es que una tupla dentro de un clúster se parezca más a las tuplas dentro de ese clúster que a las tuplas fuera de él.

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

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

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

La figura muestra seis elementos, {A, B, C, D, E, F}, que deben agruparse. Las partes (a) a (e) de la

figura muestran cinco conjuntos diferentes de clusters.

En la parte (a) se ve que cada cluster está formado por un solo elemento.

En la parte (b) se muestran cuatro conjuntos. Aquí hay dos conjuntos de dos elementos

de dos elementos. Estos racimos se forman en este nivel porque estos dos elementos están más cerca

La parte (c) muestra un nuevo grupo formado por la adición de un elemento cercano a uno de los elementos de la lista.

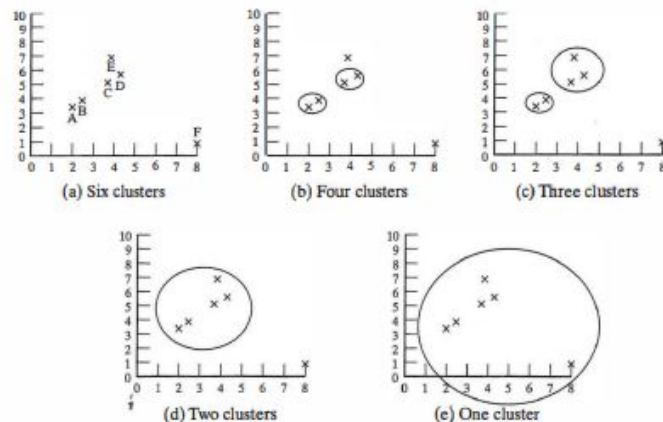
añadiendo un elemento cercano a uno de los racimos de dos elementos.

En la parte (d), las agrupaciones de dos elementos

y de tres elementos se fusionan para dar lugar a un clúster de cinco elementos. Esto se hace porque

Estos dos clusters están más cerca el uno del otro que del cluster de elementos remotos, {F}. En la

última etapa, la parte (e), se fusionan los seis elementos.



EJEMPLO 5.2

La Figura 5.5 muestra seis elementos, {A, B, C, D, E, F}, para agrupar. Partes (a) a (e) de la figura muestra cinco conjuntos diferentes de grupos.

- A. Se considera que cada grupo consta de un solo elemento.
- B. Ilustra cuatro grupos. Aquí hay dos conjuntos de dos elementos plumeros. Estos grupos se forman en este nivel porque estos dos elementos están más cerca entre sí que cualquiera de los otros elementos.
- C. Muestra un nuevo grupo formado por elementos cercanos agregados a uno de los clústeres de dos elementos.
- D. Los dos elementos y los grupos de tres elementos se fusionan para dar un grupo de cinco elementos. Esto se hace porque estos dos grupos están más cerca uno del otro que del grupo de elementos remotos, {F}.
- E. Se fusionan los seis elementos.

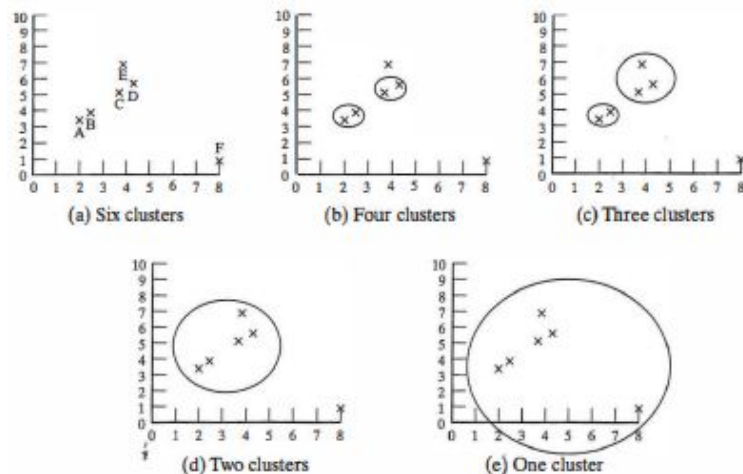
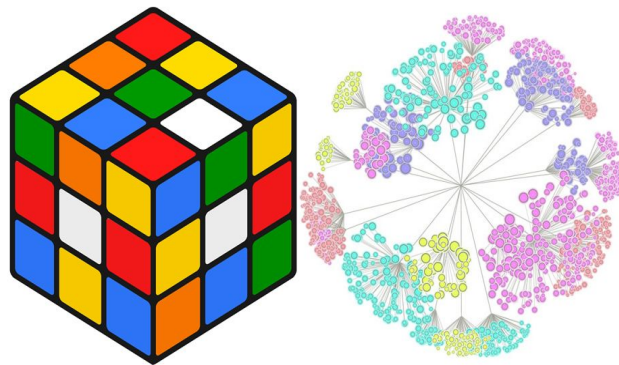


FIGURE 5.5: Five levels of clustering for Example 5.2.

5-3



VALORES ATÍPICOS

Son puntos muestrales con valores muy diferentes de los del conjunto de datos restante. Los valores atípicos pueden representar errores en los datos (quizás un sensor que funciona mal registró un valor de datos incorrecto) o podrían ser valores de datos correctos que son simplemente muy diferentes de los datos restantes.

AGRUPACIONES

Algunas técnicas de agrupación no funcionan bien con la presencia de valores atípicos.

Si se encuentran tres grupos (línea continua), el valor atípico ocurrirá en un clúster por sí mismo.

Si se encuentran dos grupos (línea discontinua), los dos diferentes conjuntos de datos se colocarán en un grupo porque están más cerca juntos que el valor atípico.

Este problema se complica por el hecho de que muchas agrupaciones.

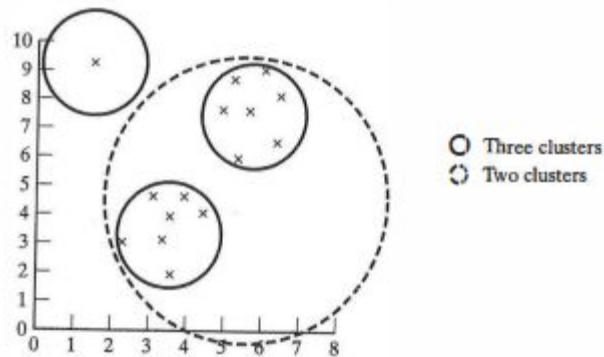


FIGURE 5.3: Outlier clustering problem.

EJEMPLO 5.3

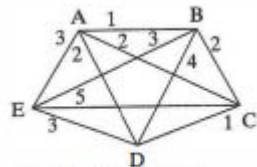
La tabla 5.2 contiene cinco elementos de datos de muestra con la distancia entre los elementos indicados en las entradas de la tabla.

Cuando se ve como un problema de gráfico, (a) muestra el gráfico con todas las ecuaciones etiquetadas con las distancias respectivas.

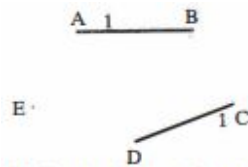
(b) muestra solo aquellos bordes con una distancia de 1 o menos. El primer nivel de agrupación en clúster de un solo enlace combinará los grupos, lo que da tres grupos: {A, B}, {C, D} y {E}.

TABLE 5.2: Sample Data for Example 5.3

Item	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	3	5	3	0



(a) Graph with all distances



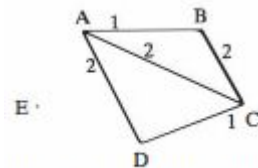
(b) Graph with threshold of 1

EJEMPLO 5.3

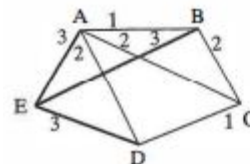
Durante el siguiente nivel de agrupamiento, miramos los bordes con una longitud de 2 o menos. El gráfico que representa este umbral de distancia (c). Esto para tener una ventaja entre los dos grupos $\{A, B\}$ y $\{C, D\}$.

Por lo tanto, en este nivel del algoritmo de agrupación en clústeres de enlace único, fusionamos estos dos clústeres para obtener un total de dos grupos: $\{A, B, C, D\}$ y $\{E\}$ (d).

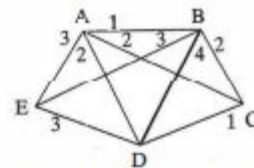
El último nivel se fusiona en un gran grupo que contiene todos los elementos (e). El etiquetado de la derecha muestra la distancia de umbral utilizada para fusionar los grupos en cada nivel.



(c) Graph with threshold of 2

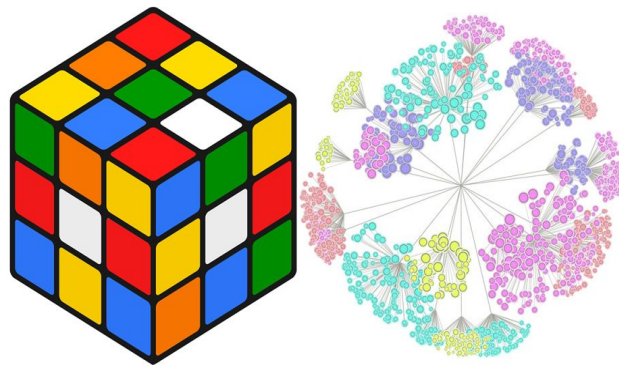


(d) Graph with threshold of 3



(e) Graph with threshold of 4

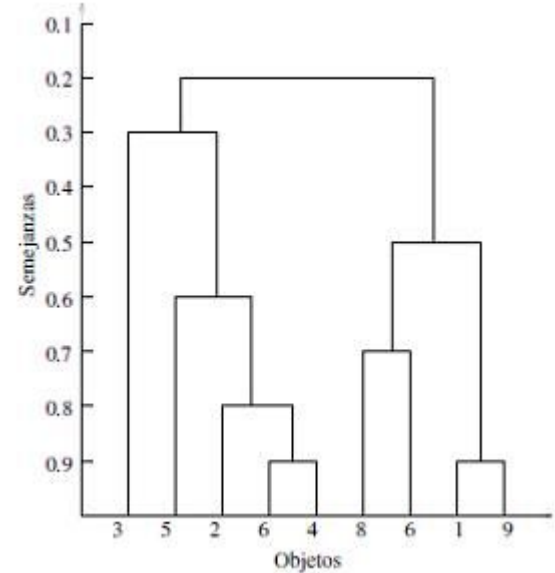
5-4



ALGORITMOS JERÁRQUICOS

Los algoritmos jerárquicos difieren en la forma de crear los conjuntos. Se puede utilizar una estructura de datos en forma de árbol, llamada dendrograma, para ilustrar la técnica de agrupación jerárquica y los conjuntos de diferentes grupos.

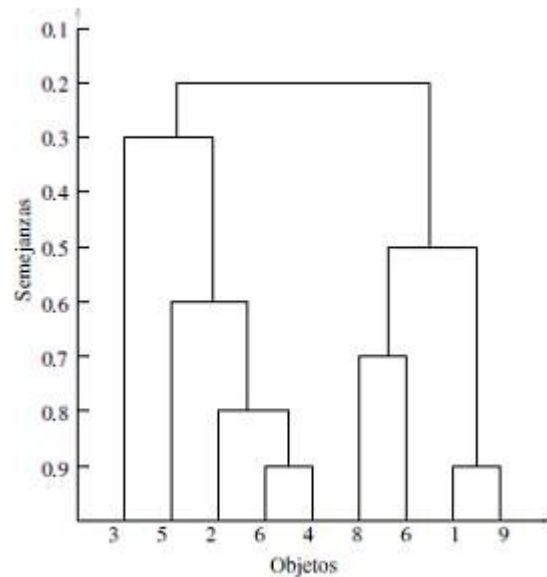
de la técnica de agrupación jerárquica y los conjuntos de los diferentes clusters. La raíz de un dendrograma



Las hojas del dendrograma consisten cada una en un clúster de un solo elemento. Los nodos internos del dendrograma representan nuevos clusters formados por la fusión de los clusters que aparecen como sus hijos en el árbol. Cada nivel del árbol está asociado a la medida de distancia que se utilizó para fusionar los clusters.

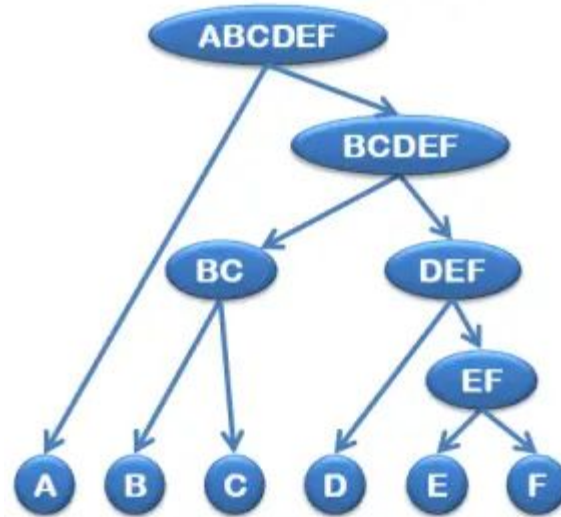
Todos los conglomerados creados en un nivel determinado se combinaron porque los conglomerados hijos tenían

una distancia entre ellos menor que el valor de la distancia asociada a este nivel en el árbol.



CLUSTERING DIVISIVO

Con la agrupación divisiva, todos los elementos se colocan inicialmente en un clúster y los clústeres repetidamente en dos hasta que todos los artículos estén en su propio clúster. La idea es dividir clústeres en los que algunos elementos no están lo suficientemente cerca de otros elementos.



EJEMPLO 5.4

Supongamos que nos dan los siguientes elementos para agrupar: {2, 4, 10, 12, 3, 20, 30, 11, 25}

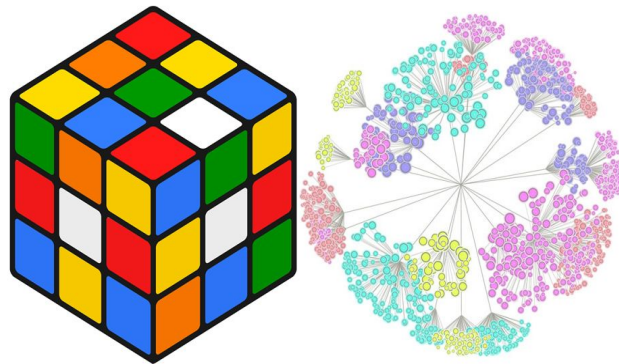
Supongamos que $k = 2$. Inicialmente asignamos las medias a los dos primeros valores: $m_1 = 2$ y $m_2 = 4$.

Utilizando la distancia euclidiana, encontramos que inicialmente $K_1 = \{2, 3\}$ y $K_2 = \{4, 10, 12, 20, 30, 11, 25\}$. El valor 3 está igualmente cerca de ambas medias, por lo que arbitrariamente elegiremos K_1 . Se podría utilizar cualquier asignación deseada en el caso de los empates. A continuación, volvemos a calcular las medias para obtener $m_1 = 2.5$ y $m_2 = 16$. Volvemos a hacer las asignaciones a los clusters para obtener $K_1 = \{2, 3, 4\}$ y $K_2 = \{10, 12, 20, 30, 11, 25\}$. Siguiendo así, obtenemos lo siguiente:

m_1	m_2	K_1	K_2
3	18	{2, 3, 4, 10}	{12, 20, 30, 11, 25}
4.75	19.6	{2, 3, 4, 10, 11, 12}	{20, 30, 25}
7	25	{2, 3, 4, 10, 11, 12}	{20, 30, 25}

Obsérvese que los conglomerados de los dos últimos pasos son idénticos. Esto dará lugar a medias idénticas, y por lo tanto las medias han convergido. Nuestra respuesta es, pues, $K_1 = \{2, 3, 4, 10, 11, 12\}$ y $K_2 = \{20, 30, 25\}$.

5-5



ALGORITMOS PARTICIONALES - AGRUPACIÓN DE CLÚSTERES

La agrupación en clústeres no jerárquica o particional crea los clústeres en un paso en lugar de varios pasos. Solo se crea un conjunto de clústeres, aunque varios conjuntos diferentes de clústeres puede crearse internamente dentro de los diversos algoritmos. Esta medida de calidad podría ser la distancia promedio entre grupos o alguna otra métrica.

La solución con el mejor valor para la función de criterio es la solución de agrupamiento utilizada. Una medida común es una métrica de error al cuadrado, que mide la distancia al cuadrado desde cada punto hasta el centroide del cluster asociado:

$$\sum_{m=1}^k \sum_{t_{mi} \in K_m} \text{dis}(C_m, t_{mi})^2$$

Un problema con los algoritmos particionales es que sufren de una combinación explosión debido al número de posibles soluciones. Claramente, la búsqueda de todas las posibles alternativas de agrupación en clústeres generalmente no sería factible.

ALGORITMO DE AGRUPACIÓN POR ERROR CUADRADO

El algoritmo de clustering de error cuadrado minimiza el error cuadrado.

El error al cuadrado de un cluster es la suma de las distancias euclidianas al cuadrado entre cada elemento del cluster y el centroide del cluster, C_k .

Dado un clúster K_i , dejemos que el conjunto de elementos asignados a ese cluster sea $\{t_{i1}, t_{i2}, \dots, t_{im}\}$.

El error al cuadrado se define como :

$$se_{K_i} = \sum_{j=1}^m \|t_{ij} - C_k\|^2$$

Dado un conjunto de clusters $K = \{K_1, K_2, \dots, K_k\}$, el error al cuadrado de K se define como

$$se_K = \sum_{j=1}^k se_{K_j}$$

AGRUPACIÓN DE K-MEANS

K-means es un algoritmo de clustering iterativo en el que los elementos se mueven entre conjuntos de clusters hasta que se alcanza el conjunto deseado.

Como tal, puede considerarse un tipo de algoritmo de error cuadrado del error cuadrático, aunque el criterio de convergencia no tiene por qué definirse en función del error cuadrático.

Se obtiene un alto grado de similitud entre los elementos de los clusters, mientras que se consigue simultáneamente un alto grado de disimilitud entre los elementos de los diferentes clusters.

La media de los clusters de $K_i = \{t_{i1}, t_{i2}, \dots, t_{im}\}$ se define como

$$m_i = \frac{1}{m} \sum_{j=1}^m t_{ij}$$

Esta definición supone que cada tupla tiene un solo valor numérico, a diferencia de una tupla con muchos valores de atributos. El algoritmo K-means requiere que exista alguna definición de la media de los clusters, pero no tiene por qué ser ésta en particular. En este caso, la media se define de forma idéntica a nuestra anterior definición de centroide. Este algoritmo asume que el número deseado de clusters, k , es un parámetro de entrada

AGRUPACIÓN POR VECINO MÁS CERCANO

Con este algoritmo en serie, los elementos se fusionan iterativamente en los clústeres existentes más cercanos.

En este algoritmo, se utiliza un umbral, t , para determinar si los elementos se agregarán a los clústeres existentes o si se creará un nuevo clúster.

La complejidad del algoritmo depende en realidad del número de elementos. Para cada iteración, cada elemento debe compararse con cada elemento que ya está en un grupo.

ALGORITHM 5.7

Input:

$D = \{t_1, t_2, \dots, t_n\}$ //Set of elements

A //Adjacency matrix showing distance between elements

Output:

K //Set of clusters

Nearest neighbor algorithm:

$K_1 = \{t_1\};$

$K = \{K_1\};$

$k = 1;$

for $i = 2$ **to** n **do**

find the t_m in some cluster K_m in K such that $\text{dis}(t_i, t_m)$ is the smallest;

if $\text{dis}(t_i, t_m) \leq t$ **then**

$K_m = K_m \cup t_i$

else

$k = k + 1;$

$K_k = \{t_i\};$

ALGORITMO PAM (PARTICIÓN ALREDEDOR DE MEDOIDES)

También llamado algoritmo K-medoides, representa un grupo por un medoide.

Usar un medoide es un enfoque que maneja bien los valores atípicos. Inicialmente, se considera que un conjunto aleatorio de k elementos es el conjunto de medoides.

Luego, en cada paso, todos los elementos del conjunto de datos de entrada que actualmente no son medoides se examinan uno por uno para ver si deberían ser medoides.

1. $t_j \in K_i$, but \exists another medoid t_m where $\text{dis}(t_j, t_m) \leq \text{dis}(t_j, t_h)$;
2. $t_j \in K_i$, but $\text{dis}(t_j, t_h) \leq \text{dis}(t_j, t_m) \forall$ other medoids t_m ;
3. $t_j \in K_m, \notin K_i$, and $\text{dis}(t_j, t_m) \leq \text{dis}(t_j, t_h)$; and
4. $t_j \in K_m, \notin K_i$, but $\text{dis}(t_j, t_h) \leq \text{dis}(t_j, t_m)$.

ALGORITMO DE ENERGÍA DE ENLACE (BEA)

Se ha utilizado en el área de diseño de bases de datos para determinar cómo agrupar los datos y cómo colocarlos físicamente en un disco.

Se puede utilizar para agrupar atributos según el uso y luego realizar el diseño lógico o físico en consecuencia. Con BEA, la afinidad (bond) entre los atributos de la base de datos se basa en el uso común.

Los pasos básicos de este algoritmo de agrupación son:

1. Crear una matriz de afinidad de atributos en la que cada entrada indique la afinidad entre los dos atributos asociados. Las entradas en la matriz de similitud se basan en la frecuencia de uso común de los pares de atributos.
2. La BEA luego convierte esta matriz de similitud en una matriz BOND en la que las entradas representan un tipo de vínculo vecino más cercano basado en la probabilidad de coacceso. El algoritmo BEA reorganiza filas o columnas para que los atributos similares aparezcan juntos en la matriz.
3. Finalmente, el diseñador dibuja cuadros alrededor de las regiones de la matriz con alta similitud.

Dos atributos A_i y A_j tienen una alta afinidad si se usan juntos con frecuencia en aplicaciones de bases de datos. En el corazón del algoritmo BEA se encuentra la medida de afinidad global.

Suponga que el esquema de una base de datos consta de n atributos $\{A_1, A_2, \dots, A_n\}$. La medida de afinidad global, AM , se define como

$$AM = \sum_{i=1}^n (\text{bond}(A_i, \overline{A_{i-1}}) + \text{bond}(A_i, \overline{A_{i+1}}))$$

	A_1	A_2	A_3	A_4	A_{n-1}	A_n
A_1						
A_2						
A_3						
A_4						
A_{n-1}						
A_n						

AGRUPACIÓN CON ALGORITMOS GENÉTICOS

Para determinar cómo realizar la agrupación con algoritmos genéticos, primero debemos determinar cómo representar cada grupo.

Un enfoque simple sería utilizar una representación de mapa de bits para cada posible agrupamiento.

Entonces, dada una base de datos con cuatro elementos, {A, B, C, D}, representaremos una solución para crear dos grupos como 1001 y 0110. Esto representa los dos grupos {A, D} y {B, C} .

ALGORITHM 5.9

Input:

$D = \{t_1, t_2, \dots, t_n\}$ //Set of elements

k //Number of desired clusters

Output:

K //Set of clusters

GA clustering algorithm:

randomly create an initial solution;

repeat

 use crossover to create a new solution;

until termination criteria is met;

AGRUPAMIENTO CON REDES NEURONALES

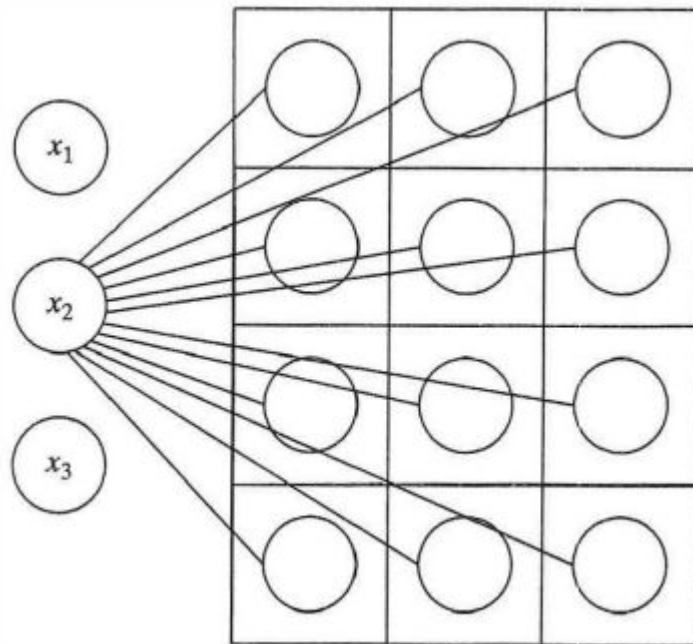
Las redes neuronales (NN) que utilizan el aprendizaje no supervisado intentan encontrar características en los datos que caracterizan la salida deseada. Buscan grupos de datos similares. Estos tipos de NN a menudo se denominan redes neuronales autoorganizadas.

Hay dos tipos básicos de aprendizaje no supervisado: no competitivo y competitivo.

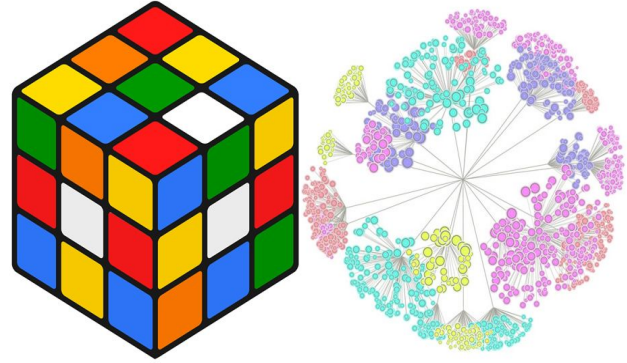
Con el aprendizaje no competitivo o aprendizaje de Hebbian, el peso entre dos nodos se cambia para que sea proporcional a ambos valores de salida. Es decir:

$$\Delta w_{ji} = \eta y_j y_i$$

Con el aprendizaje competitivo, los nodos pueden competir y el ganador se lo lleva todo.



5-6



CLUSTERING EN GRANDES BASES DE DATOS

CARACTERÍSTICAS DEL CLUSTERING EN GRANDES BASES DE DATOS

1. No se requerirá más de un escaneo de la base de datos.
2. Tener la capacidad de proporcionar el estado y la "mejor" respuesta hasta el momento durante la ejecución del algoritmo. En ocasiones, esto se conoce como la capacidad de estar en línea.
3. Ser suspendido, bloqueable y reanudable.
4. Ser capaz de actualizar los resultados de forma incremental a medida que se agregan o eliminan datos del base de datos.
5. Trabajar con memoria principal limitada.
6. Ser capaz de realizar diferentes técnicas para escanear la base de datos. Esto puede
7. incluir muestreo.
8. Procesa cada tupla solo una vez.

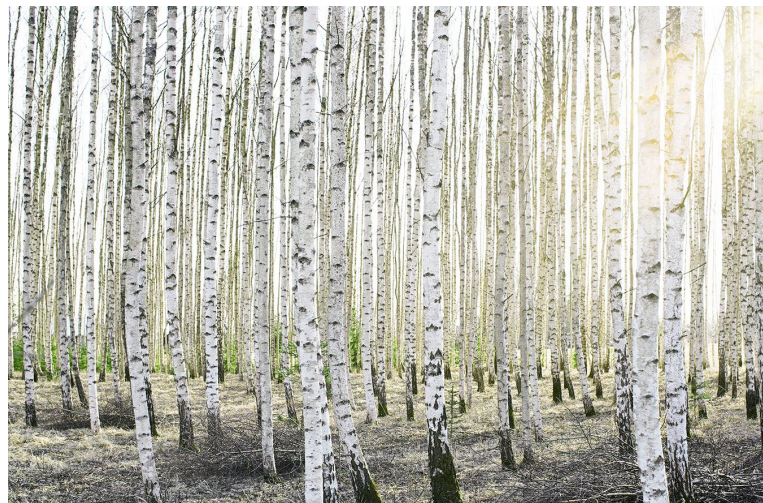


TÉCNICA DE ESCALAMIENTO BÁSICA

1. Leer un subconjunto de la base de datos en la memoria principal.
2. Aplicar la técnica de agrupación en clústeres a los datos en la memoria.
3. Combinar los resultados con los de muestras anteriores.
4. Los datos en memoria se dividen en tres tipos diferentes: aquellos elementos que siempre serán necesarios incluso cuando se traiga la siguiente muestra, las que pueden ser descartado con las actualizaciones apropiadas de los datos que se mantienen con el fin de responder a la problema y los que se guardarán en formato comprimido. Según el tipo, cada elemento de datos se guarda, elimina o comprime en la memoria.
5. Si no se cumplen los criterios de terminación, repita desde el paso 1.



BIRCH (BALANCED ITERATIVE REDUCING AND CLUSTERING USING HIRARCHIES)



La idea básica del algoritmo es que se construye un árbol que captura la información necesaria para realizar la agrupación. El agrupamiento es luego realizado en el mismo árbol, donde las etiquetas de los nodos en el árbol contienen la información necesaria para calcular los valores de distancia.

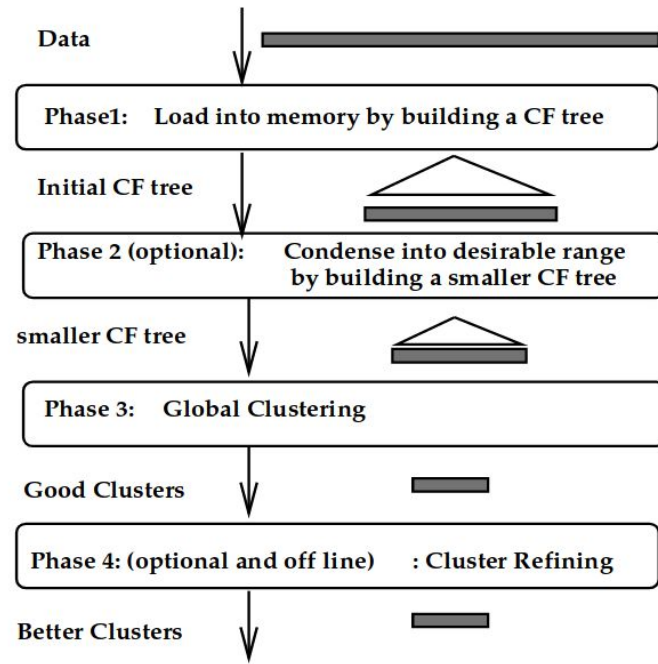
Una característica importante del algoritmo BIRCH es el uso del **Clustering Feature (CF)** que proporciona un resumen de la información alrededor de un grupo.

Por lo mencionado anteriormente podemos decir que BIRCH solo se aplica para datos numéricos.

CF (CLUSTERING FEATURE)

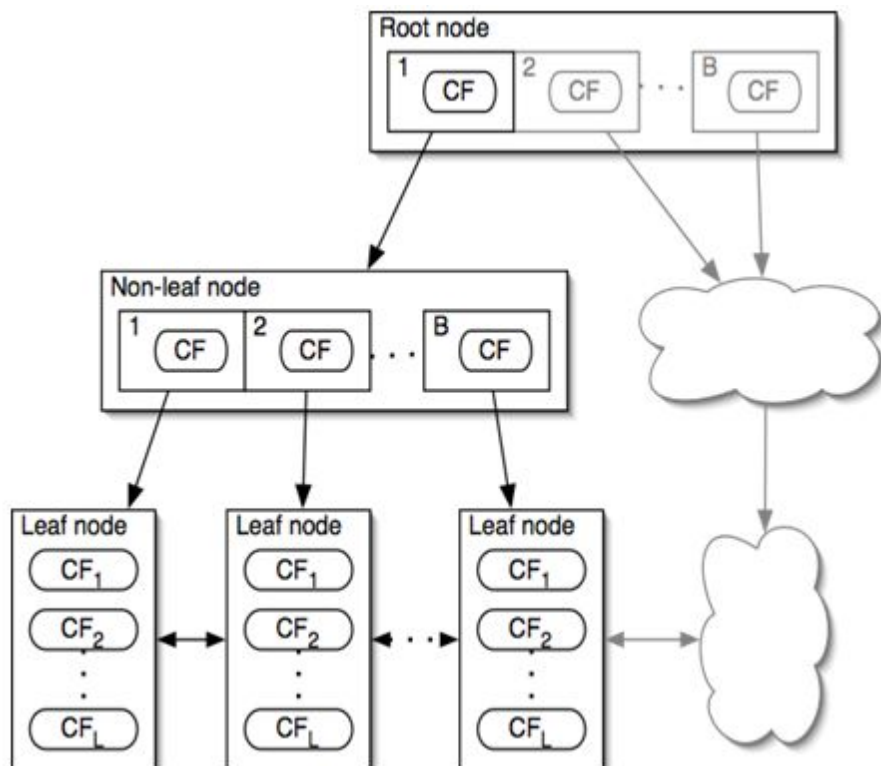
La función de agrupamiento llamado en ingles como Clustering Feature es un triple $CF = (N, LS, SS)$, donde N es el número de puntos en el cluster, LS es la suma de los puntos en el cluster y SS es la suma de los cuadrados de los puntos en dicho cluster.

Figure 2. BIRCH Overview



CLUSTERING FEATURE TREE

Es un árbol equilibrado con un factor de ramificación (máximo número de hijos que puede tener un nodo) B . Cada nodo interno contiene un triple CF para cada uno de sus hijos. Cada nodo hoja también representa un clúster y contiene un CF entrada para cada subclúster en él. Un subclúster en un nodo hoja debe tener un diámetro no mayor que un valor umbral determinado T .



ALGORITMO BIRCH

Input:

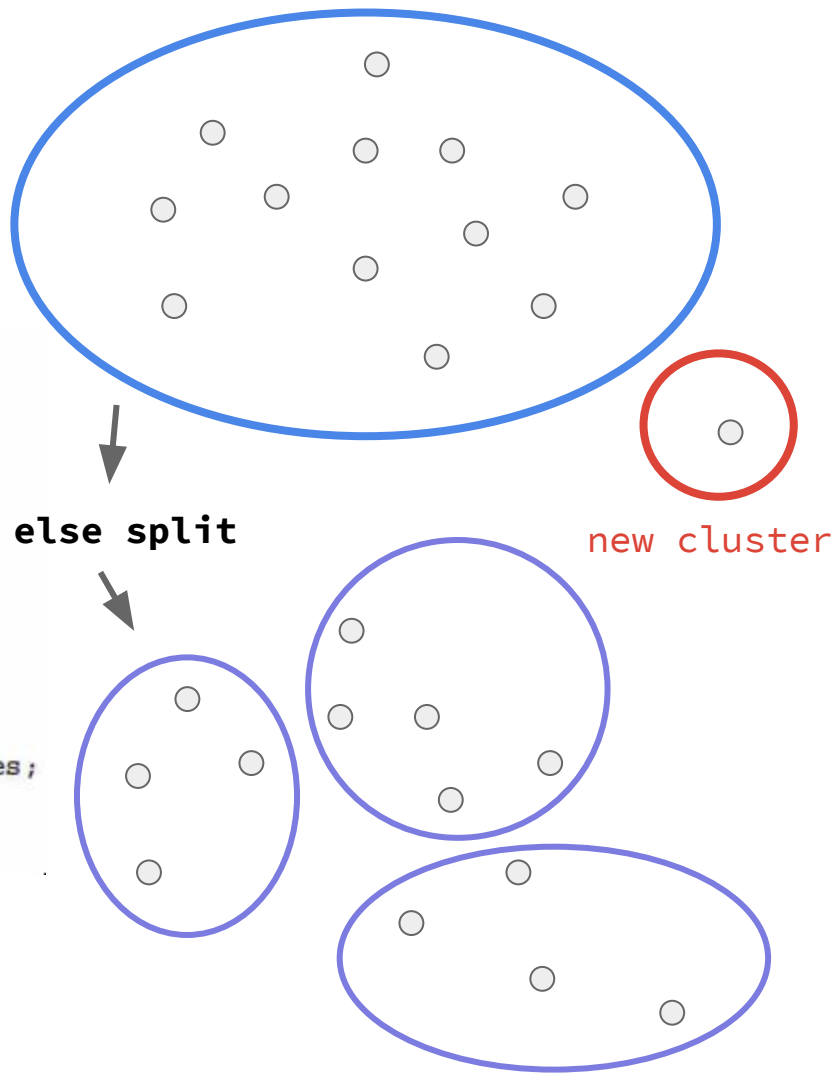
$D = \{t_1, t_2, \dots, t_n\}$ //Set of elements
 T // Threshold for CF tree construction

Output:

K //Set of clusters

BIRCH clustering algorithm:

```
for each  $t_i \in D$  do
    determine correct leaf node for  $t_i$  insertion;
    if threshold condition is not violated, then
        add  $t_i$  to cluster and update CF triples;
    else
        if room to insert  $t_i$ , then
            insert  $t_i$  as single cluster and update CF triples;
        else
            split leaf node and redistribute CF features;
```



DBSCAN (DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE)

El enfoque utilizado por DB SCAN (agrupación espacial de aplicaciones basada en densidad con ruido) consiste en crear clústeres con un tamaño y una densidad mínimos. La densidad se define como número mínimo de puntos dentro de una cierta distancia entre sí. Esto maneja el problema de valores atípicos asegurándose de que un valor atípico (o un pequeño conjunto de valores atípicos) no creará un grupo. Un parámetro de entrada, MinPts, indica el número mínimo de puntos en cualquier grupo. Además, para cada punto de un clúster debe haber otro punto en el clúster, cuya distancia desde él es menor que un valor de entrada de umbral, Eps. El barrio de Eps o vecindad de un punto es el conjunto de puntos dentro de una distancia de Eps. El deseado el número de conglomerados, k , no es una entrada, sino que está determinado por el algoritmo mismo.



DBSCAN

DBSCAN utiliza un nuevo concepto de densidad. la siguiente definición define directamente la densidad alcanzable. La primera parte de la definición asegura que el segundo punto está "lo suficientemente cerca" del primer punto. El segundo parte de la definición asegura que hay suficientes puntos centrales lo suficientemente cerca de cada otro. Estos puntos centrales forman la parte principal de un grupo en el sentido de que todos están cerca de mutuamente. Un punto directamente alcanzable por densidad debe dosificarse en uno de estos puntos centrales, pero no tiene por qué ser un punto central en sí mismo. En ese caso, se llama punto fronterizo. Se dice que un punto es de densidad alcanzable desde otro punto si hay una cadena de uno a otro que contiene solo puntos que son directamente alcanzables en densidad desde el punto anterior. Esta garantiza que cualquier clúster tendrá un conjunto central de puntos muy cercano a un gran número de otros puntos (puntos centrales) y luego algunos otros puntos (puntos fronterizos) que están suficientemente cerca de al menos un punto central.

DEFINITION 5.5. Given values Eps and $MinPts$, a point p is **directly density-reachable** from q if

- $dis(p, q) \leq Eps$
and
- $|\{r \mid dis(r, q) \leq Eps\}| \geq MinPts$

DBSCAN EXAMPLE

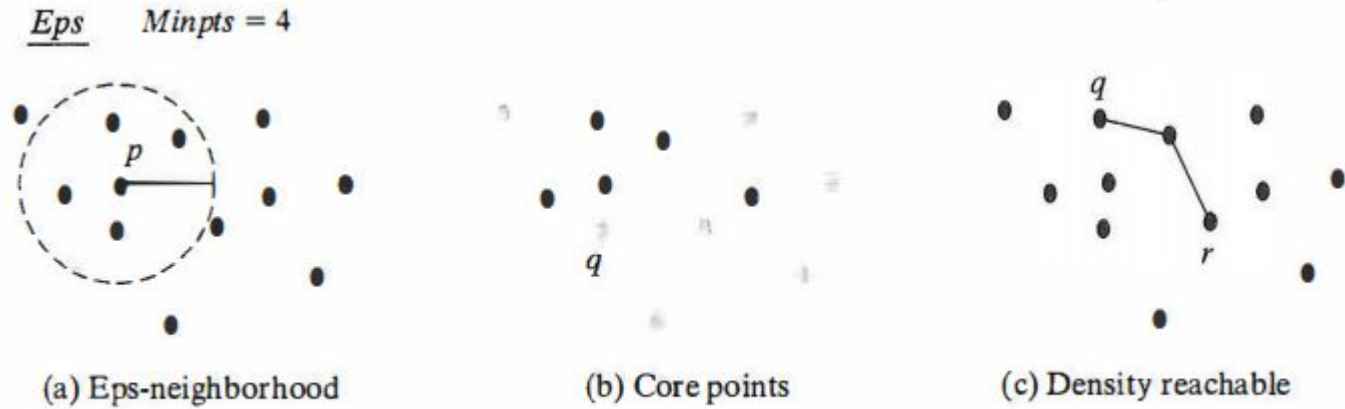


FIGURE 5.12: DBSCAN example.

DBSCAN ALGORITHM

ALGORITHM 5.11

Input:

$D = \{t_1, t_2, \dots, t_n\}$ //Set of elements
 $MinPts$ // Number of points in cluster
 Eps // Maximum distance for density measure

Output:

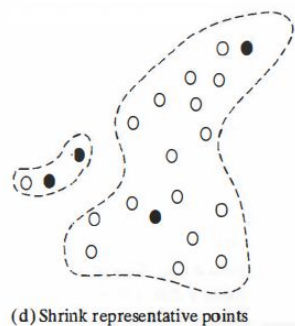
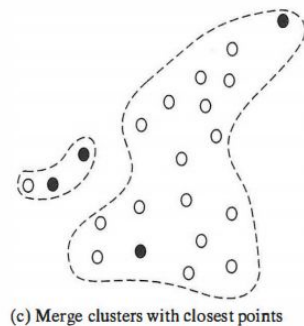
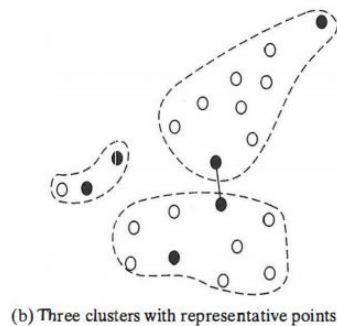
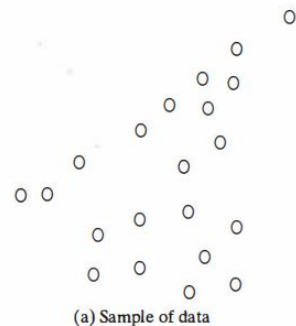
$K = \{K_1, K_2, \dots, K_k\}$ //Set of clusters

DBSCAN algorithm:

```
k = 0; // Initially there are no clusters.
for i = 1 to n do
    if  $t_i$  is not in a cluster, then
         $X = \{t_j \mid t_j \text{ is density-reachable from } t_i\}$ ;
        if  $X$  is a valid cluster, then
            k = k + 1;
             $K_k = X$ ;
```

CURE (CLUSTERING USING REPRESENTATIVES)

El enfoque básico utilizado por CURE se muestra en la siguiente figura. El primer paso muestra una muestra de los datos. Existe un conjunto de grupos con sus puntos representativos en cada paso. en el procesamiento. En la figura hay tres grupos, cada uno con dos representativos puntos. Los puntos representativos se muestran como círculos oscuros. Como se discutió en el párrafos siguientes, estos puntos representativos se eligen para estar lejos unos de otros como así como de la media del conglomerado. En la parte (c), dos de los grupos se fusionan y dos se eligen nuevos puntos representativos. Finalmente, en la parte (d), estos puntos se reducen hacia la media del conglomerado. Observe que si se hubiera elegido un centroide representativo para los grupos, el grupo más pequeño se habría fusionado con el grupo inferior en su lugar de con el grupo superior.



PASOS BASICOS CURE

1. Obtenga una muestra de la base de datos.
2. Divida la muestra en p particiones de tamaño n / p . Esto se hace para acelerar el algoritmo porque la agrupación en clústeres se realiza primero en cada partición.
3. Agrupe parcialmente los puntos en cada partición utilizando el algoritmo jerárquico. Esto proporciona una primera suposición de lo que deberían ser los clústeres. los número de conglomerados es n / pq para alguna constante q .
4. Elimine los valores atípicos. Los valores atípicos se eliminan mediante el uso de dos técnicas diferentes. La primera técnica elimina los racimos que crecen muy lentamente. Cuando el número de clusters está por debajo de un umbral, se eliminan los clústeres con solo uno o dos elementos. Es posible que los valores atípicos cercanos sean parte de la muestra y no se identificarían por la primera técnica de eliminación de valores atípicos. La segunda técnica elimina los grupos muy pequeños hacia el final de la fase de agrupamiento.
5. Agrupe completamente todos los datos en la muestra usando Algoritmo. Aquí, para asegurar procesamiento en la memoria principal, la entrada incluye solo los representantes del clúster de los clústeres encontrados para cada partición durante el paso de agrupamiento parcial.
6. Agrupe toda la base de datos en el disco utilizando puntos c para representar cada grupo. Un articulo en la base de datos se coloca en el clúster que tiene el punto representativo más cercano a eso. Estos conjuntos de puntos representativos son lo suficientemente pequeños como para caber en la memoria principal, por lo que cada uno de los n puntos debe compararse con ck puntos representativos.

ALGORITMO CURE

Input:

$D = \{t_1, t_2, \dots, t_n\}$ //Set of elements

k // Desired number of clusters

Output:

Q //Heap containing one entry for each cluster

CURE algorithm:

$T = \text{build}(D);$

$Q = \text{heapify}(D);$ // Initially build heap with one entry per item;

repeat

$u = \min(Q);$

$\text{delete}(Q, u.\text{close});$

$w = \text{merge}(u, v);$

$\text{delete}(T, u);$

$\text{delete}(T, v);$

$\text{insert}(T, w);$

for each $x \in Q$ do

$x.\text{close} = \text{find closest cluster to } x;$

if x is closest to w , then

$w.\text{close} = x;$

$\text{insert}(Q, w);$

until number of nodes in Q is $k;$

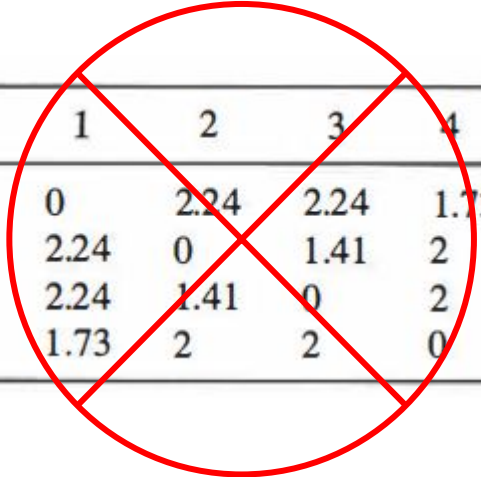
CLUSTERING CON ATRIBUTOS CATEGÓRICOS

Considere un sistema de recuperación de información donde los documentos pueden contener palabras clave {libro, agua, sol, arena, nadar, leer}. Suponga que hay cuatro documentos, donde el primero contiene la palabra {libro}, el segundo contiene {agua, sol, arena, nadar}, el tercero contiene {agua, sol, nadar, leer}, y el cuarto contiene {leer, arena}. Podemos representar los cuatro libros utilizando los siguientes puntos booleanos: (1, 0, 0, 0, 0, 0), (0, 1, 1, 1, 1, 0), (0, 1, 1, 0, 1, 1), (0, 0, 0, 1, 0, 1). Podemos usar la distancia euclidiana para desarrollar la siguiente matriz de adyacencia de distancias:

	1	2	3	4
1	0	2.24	2.24	1.73
2	2.24	0	1.41	2
3	2.24	1.41	0	2
4	1.73	2	2	0

CLUSTERING CON ATRIBUTOS CATEGÓRICOS

$k = 2 \rightarrow$ **clusters:** $\{\{1, 4\}, \{2, 3\}\}$.



	1	2	3	4
1	0	2.24	2.24	1.73
2	2.24	0	1.41	2
3	2.24	1.41	0	2
4	1.73	2	2	0

THE ROCK (ROBUST CLUSTERING USING LINKS)

El algoritmo de agrupación de clústeres ROCK (Robust Clustering using links) está dirigido a ambos datos booleanos y datos categóricos. Se basa un enfoque novedoso para identificar similitudes en el número de enlaces entre elementos. Se dice que un par de elementos son vecinos si su la similitud supera algún umbral. Esto no necesita definirse en función de una métrica precisa, pero podría utilizarse un enfoque más intuitivo utilizando expertos en el dominio. El objetivo del algoritmo de agrupación en clústeres es agrupar los puntos que tienen más enlaces. El algoritmo es un algoritmo aglomerativo jerárquico que utiliza el número de enlaces como medida de similitud en lugar de una medida basada en la distancia. En lugar de utilizar una distancia euclidiana. Una medida de similitud propuesta basada en el coeficiente de Jaccard se define como:

$$\text{sim}(t_i, t_j) = \frac{|t_i \cap t_j|}{|t_i \cup t_j|}$$

Ejemplo:

t1 = {1,0,0,0,0,0}

t2 = {0,1,1,1,1,0}

t3 = {0,1,1,0,1,1}

sim(t1,t2) = 0/5 = 0

sim(t2,t3) = 3/5 = 0.6

THE ROCK (ROBUST CLUSTERING USING LINKS)

Supongamos que decimos que el umbral para un vecino es 0.6, entonces tenemos los siguientes:

vecinos: $\{(2, 3), (2, 4), (3, 4)\}$.

	1	2	3	4
1	1	0	0	0
2	0	1	0.6	0.2
3	0	0.6	1	0.2
4	0	0.2	0.2	1

THE ROCK (ROBUST CLUSTERING USING LINKS)

La siguiente tabla muestra el número de enlaces (vecinos comunes entre puntos) asumiendo que el umbral para un vecino es 0.6:

En este caso, tenemos los siguientes grupos: $\{\{1\}, \{2, 3, 4\}\}$. Comparando esto con el conjunto de agrupamiento encontrado con una distancia euclidiana tradicional, vemos que un conjunto "mejor" de clusters se ha creado.

	1	2	3	4
1	1	0	0	0
2	0	3	3	3
3	0	3	3	3
4	0	3	3	3