

Chapter 5 Cluster Analysis

- It is similar to classification where similar data are grouped together.
- Groups are not predefined as in classification i.e clustering is an unsupervised way of classification.
- It is also called data segmentation and mostly used for detection.
- Cluster is a collection of data objects in which the objects are similar to one another within the same cluster and dissimilar to the objects in other clusters.

Given a data base,

$D = \{x_1, x_2, \dots, x_m\}$, a distance measure $dist(x_i, x_j)$ defined among two objects x_i and x_j and an integer value K (the number of clusters). The clustering problem is to define a mapping $f: D \rightarrow \{1, 2, \dots, K\}$ where, each x_i is assigned to one cluster K_j , $1 \leq j \leq K$.

Techniques:

- (1) Hierarchical clustering
- (2) Partitioning clustering
- (3) Density based clustering
- (4) Grid based clustering
- (5) Model based clustering

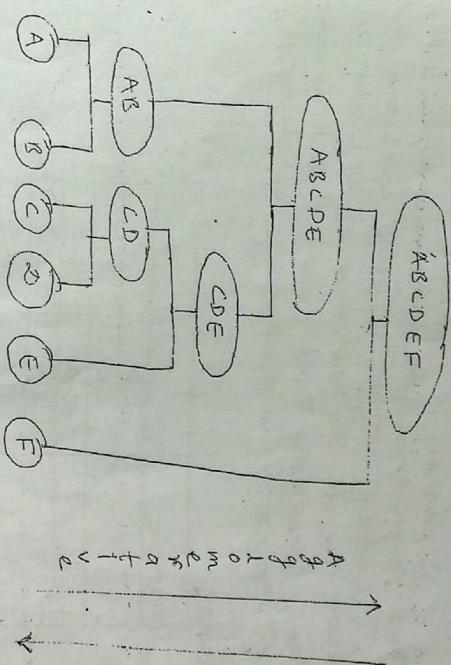
1. Hierarchical clustering:-

- A nested set of clusters is created with each level in the hierarchy that has the separate set of clusters.
- At the lowest level each item is in its own unique cluster.
- At the highest level, all items belong to the same cluster. Types of hierarchical clustering:-
- (a) Agglomerative (bottom up):-
 - starts from clustering individual point only with each cluster having only one record.
 - repeat merging the cluster until a certain number of clusters.
 - the merging is done on the basis of pair nearest to each other.
 - If the merging is continued, it terminates in hierarchy of clusters which ends into single cluster.
 - It is one of the more powerful approach for clustering.
- (b) Divisive (top down):-
 - starts from a cluster including all the data points, repeating the cluster until a certain number of clusters are generated.

Disadvantage:-

- Once the step is done, it cannot be undone.

Algorithms: BIRCH, ROCK



2. Partitioning Clustering (Iterative relocation method)

- Partition the database into a predefined number of cluster.
- It attempts to determine the k-partitions that optimize the certain criteria.
- construct a partition of database D of n -objects into a set of k -clusters such that we have the minimum sum of squared.
- Some of the common algorithms are
 - (i) K-means
 - (ii) K-Medoid
 - (iii) K-Medians

Iteration 2nd

K-Means Algorithm: K means algorithm choose k i.e. number of cluster to be determined from the data as the centers.

Step 1: choose k objects randomly from the data as the centers.

Step 2: choose k objects randomly from the remaining objects to the cluster where

the remaining objects to the cluster where centre is most close to using euclidean distance.

Step 3: Assign each of the remaining objects to the cluster using mean point.

Step 4: Compute the new centre of the cluster using mean point.

Step 5: Repeat step 3 and 4 until there is no change in cluster

center or no object change of cluster.

Example:

Instance	X	Y
1	1.0	1.5
2	1.0	4.5
3	2.0	1.5
4	2.0	3.5
5	3.0	2.5
6	3.0	6.0

(1) Let k = 2,

(a) centre for cluster 1, c_1 = instance 2 (1.0, 4.5)

centre for cluster 2, c_2 = instance 4 (2.0, 3.5)

(b) Now, distance of instance 1 to c_1 = $\sqrt{(1.0-1.0)^2 + (1.5-4.5)^2} = \sqrt{9}$

distance of instance 1 to c_2 = $\sqrt{(1.0-2.0)^2 + (1.5-3.5)^2} = \sqrt{5}$.

* Since distance of instance 1 to c_2 is less than to c_1 ,

$$dist(1, c_2) < dist(1, c_1), \therefore 1 \in c_2$$

$$dist(3, c_1) = \sqrt{(2-1)^2 + (1.5-4.5)^2} = \sqrt{10}$$

$$dist(3, c_2) = \sqrt{(2-2)^2 + (1.5-3.5)^2} = \sqrt{2}, \therefore 3 \in c_2$$

$$dist(5, c_1) = \sqrt{(3-1)^2 + (2.5-4.5)^2} = \sqrt{8}$$

$$dist(5, c_2) = \sqrt{(3-2)^2 + (2.5-3.5)^2} = \sqrt{2}, \therefore 5 \in c_2$$

$$dist(6, c_1) = \sqrt{(3-1)^2 + (6-4.5)^2} = \sqrt{18.25}$$

$$dist(6, c_2) = \sqrt{(3-2)^2 + (6-3.5)^2} = \sqrt{14.75}, \therefore 6 \in c_2$$

$$dist(4, c_1) = \sqrt{(2-1)^2 + (3.5-4.5)^2} = \sqrt{10}$$

$$dist(4, c_2) = \sqrt{(2-2)^2 + (3.5-3.5)^2} = \sqrt{0.25}, \therefore 4 \in c_2$$

$$\text{New mean centre, } c_1 = (1.0, 4.5)$$

$$c_2 = \frac{1+2+2+3.5}{5}, \frac{1.5+1.5+3.5+2.5+6}{5}$$

$$\begin{aligned} dist(1, c_1) &= \sqrt{9} \quad (\because c_1 \text{ is not changed}) \\ dist(1, c_2) &= \sqrt{(2-1)^2 + (3-1)^2} = \sqrt{4.0}, \therefore 1 \in c_2 \\ dist(3, c_1) &= \sqrt{5} \\ dist(3, c_2) &= \sqrt{(2-2)^2 + (3-3)^2} = \sqrt{0.36+2.25} = \sqrt{2.61}, \therefore 3 \in c_2 \\ dist(5, c_1) &= \sqrt{10} \\ dist(5, c_2) &= \sqrt{(3-2)^2 + (3-3)^2} = \sqrt{0.16+0.25} = \sqrt{0.41}, \therefore 5 \in c_2 \\ dist(6, c_1) &= \sqrt{18.25} \\ dist(6, c_2) &= \sqrt{(3-2)^2 + (6-3)^2} = \sqrt{14.75}, \therefore 6 \in c_2 \end{aligned}$$

Conclusion: since there is no change in objects of clusters: so the final two cluster are $g = \{2, 3\}, c_2 = \{1, 3, 4, 5, 6\}$

Advantages:

→ Relatively efficient

→ Simple Implementation

Disadvantages:

→ Need to specify the number of clusters in advance

→ Unable to handle noisy data and outliers

→ Complexity increases with increase in size.

→ cannot handle categorical data

→ Not efficient for highly nonuniform distributed data.

3: Density Based clustering:

⇒ DBSCAN (Density based spatial clustering of applications with noise)

→ It grows its region with sufficiently high density into clusters of arbitrary shape in spatial database with n

discovered clusters or arbitrary shape in spatial database with n

→ The neighbourhood of the object contains atleast a minimum

number of points then the object is called core object.

→ The ϵ -neighbourhood of the object contains atleast a minimum

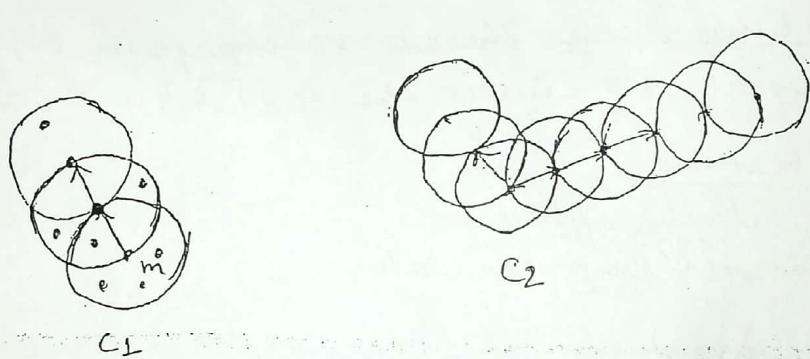
number of points then the object is called density reach

→ Given a set of objects D, and object p is directly density reach

from object q within ϵ 's neighbourhood of q and q is core object

- An object 'p' is density reachable from object 'q' with its radius and minimum points in a set of objects D, if there is a chain of objects $P_1 P_2, \dots, P_n$ where $P_1 = q$ and $P_n = p$ such that P_i is directly density reachable from P_{i+1} with respect to its radius and minimum point.
- An object 'p' is density connected to object 'q' with respect to radius and minimum points in a set of objects D. If there is an object 'o' belongs to D such that both p and q are density reachable from o.
- Density reachability is the transitive closure of direct density reachability and the relationship is asymmetric.
- Only core objects are mutually density reachable and symmetric in relation.

Eg:



Algorithm:-

- Step 1: DBSCAN searches for cluster by checking the ϵ -neighbourhood of each point in data base.
- Step 2: If ϵ -neighbourhood of a point (p) contains more than minimum point a new cluster with point 'p' as a core object is created
- Step 3: Iteratively collects directly density reachable objects from core objects, which may involve the merging of few density reachable clusters and points.
- Step 4: Terminates when no new point or cluster can be added to any cluster.

5.5 Evaluation, Scalability, comparison.

Evaluation:-

- Intrinsic
- Extrinsic.