

## 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 outlier detection.
- Cluster is a collection of data objects in which the objects are all do one another within the same cluster and dissimilar to the objects belonging to other clusters.

Given a database,

$D = \{t_1, t_2, \dots, t_n\}$ , a distance measure  $\text{dist}(t_i, t_j)$  defined among two objects  $t_i$  and  $t_j$  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  $t_i$  is assigned to one cluster  $K_j$ ,  $1 \leq j \leq K$ .

Techniques:

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

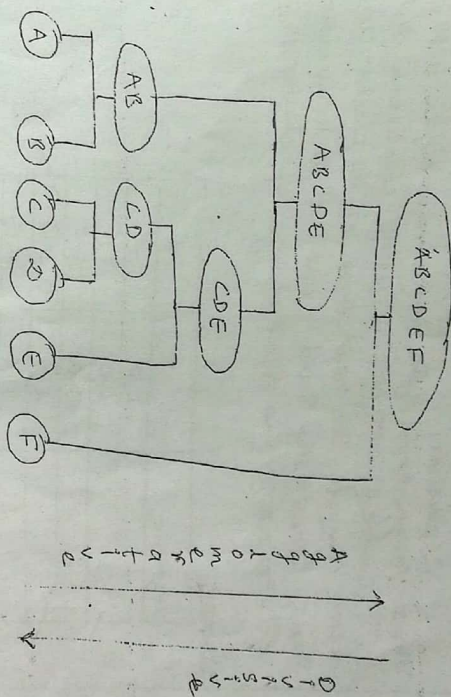
### 1. Hierarchical clustering:

- 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.
- (a) Agglomerative (bottom up):
  - start 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.
  - As the merging is continued, it terminates in hierarchy of clusters which ends in a single cluster.
  - It is one of the more powerful approach for clustering.

### (b) Divisive (top down):

- start from a cluster including all the data points, repeatedly the clusters until a certain number of clusters are generated.

- Splitting is done on the basis of certain optimization criteria.
- If the splitting is continued, it terminates in the hierarchy in cluster whose each item is identified as single cluster.



### 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 clusters.
- It attempts to determine the K-partitions that optimize the certain criteria.
- Construct a partition of database 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-Med
  - (iii) K-Median

23

K-Means Algorithm:  
 choose K i.e. number of clusters to be determined

step 1: choose K i.e. number of clusters to be determined  
 step 2: choose K objects randomly from the data as the centres

K clusters

step 3: Assign each of the remaining objects to the cluster whose centre is most close to using euclidean distance.

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 centre 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	5.0	6.0

Let  $K=2$ ,

(2) centre for cluster  $C_1$  = instance 2 (1.0, 4.5)  
 centre for cluster  $C_2$  = instance 4 (2.0, 3.5)

(3) Also, 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  $C_2 < \text{dist}(1, C_1)$ , 1 lies in  $C_2$

$\text{dist}(1, C_2) < \text{dist}(1, C_1)$ , 1 lies in  $C_2$

$\text{dist}(1, C_1) = \sqrt{(1.0-1.0)^2 + (1.5-4.5)^2} = \sqrt{9}$

$\text{dist}(1, C_2) = \sqrt{(1.0-2.0)^2 + (1.5-3.5)^2} = \sqrt{5}$ ,  $\therefore$  1 lies in  $C_2$

$\text{dist}(5, C_1) = \sqrt{(3.0-1.0)^2 + (2.5-4.5)^2} = \sqrt{8}$

$\text{dist}(5, C_2) = \sqrt{(3.0-2.0)^2 + (2.5-3.5)^2} = \sqrt{2}$ ,  $\therefore$  5 lies in  $C_2$

$\text{dist}(6, C_1) = \sqrt{(5.0-1.0)^2 + (6.0-4.5)^2} = \sqrt{18.25}$

$\text{dist}(6, C_2) = \sqrt{(5.0-2.0)^2 + (6.0-3.5)^2} = \sqrt{15.25}$ ,  $\therefore$  6 lies in  $C_2$

$C_1 = \{2, 3\}$ ,  $C_2 = \{1, 3, 4, 5, 6\}$

4) New mean centre,  $C_1 = (1.0, 4.5)$

$C_2 = \frac{1+2+3+5}{5}, \frac{1.5+1.5+3.5+2.5+6}{5}$   
 $= (1.0, 4.2)$

Iteration 2:

$\text{dist}(1, C_1) = \sqrt{9}$  ( $\because C_1$  is not changed)

$\text{dist}(1, C_2) = \sqrt{(1.0-2.0)^2 + (1.5-3.5)^2} = \sqrt{5}$ ,  $\therefore$  1 lies in  $C_2$

$\text{dist}(3, C_1) = \sqrt{10}$

$\text{dist}(3, C_2) = \sqrt{(3.0-2.0)^2 + (3.5-3.5)^2} = \sqrt{1.0}$ ,  $\therefore$  3 lies in  $C_2$

$\text{dist}(5, C_1) = \sqrt{8}$

$\text{dist}(5, C_2) = \sqrt{(3.0-2.0)^2 + (2.5-3.5)^2} = \sqrt{2.0}$ ,  $\therefore$  5 lies in  $C_2$

$\text{dist}(4, C_1) = \sqrt{(2.0-1.0)^2 + (3.5-4.5)^2} = \sqrt{2.0}$

$\text{dist}(4, C_2) = \sqrt{(2.0-2.0)^2 + (3.5-3.5)^2} = \sqrt{0.0}$ ,  $\therefore$  4 lies in  $C_2$

$\text{dist}(6, C_1) = \sqrt{18.25}$

$\text{dist}(6, C_2) = \sqrt{(5.0-2.0)^2 + (6.0-3.5)^2} = \sqrt{14.75}$ ,  $\therefore$  6 lies in  $C_2$

$\text{dist}(1, C_2) = \sqrt{(1.0-2.0)^2 + (1.5-3.5)^2} = \sqrt{5}$ ,  $\therefore$  1 lies in  $C_2$

Conclusion: Since there is no change in objects of clusters, so the final two clusters are  $C_1 = \{2, 3\}$ ,  $C_2 = \{1, 3, 4, 5, 6\}$

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 non-uniform distributed data

3. Density Based clustering:

→ DBSCAN (Density Based Spatial clustering of applications with Noise)

→ It groups its region with sufficiently high density into clusters

→ It groups clusters of arbitrary shape in spatial database with no

→ The neighbourhood of the object

→ The  $\epsilon$ -neighbourhood of the object contains atleast a minimum

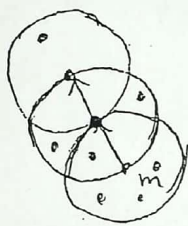
→ The  $\epsilon$ -neighbourhood of the object is called core object

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

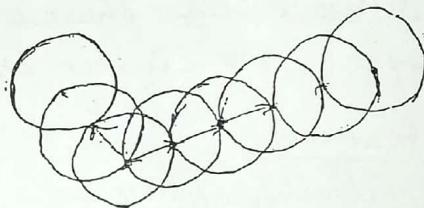
from object q, within  $\epsilon$ -neighbourhood of p and q is core object

- 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  $p_{i+1}$  is directly density reachable from  $p_i$  with respect to its radius and minimum point.
- An object 'p' is density connected to object 'q' with respect to its 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:



C1



C2

Algorithm:-

Step 1: DBSCAN searches for cluster by checking the  $\epsilon$ -neighbourhood of each point in data base.

Step 2: If  $\epsilon$ -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.