CLUSTERING

Definition :-

Clustering is the process of partitioning a set of data objects into subsets. Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. A cluster of data objects can be treated as one group.

- The process of partitioning data objects into subclasses is called as cluster.
- Clustering is also called as data segmentation, because it partitions large data sets into groups according to their similarity.

Applications :-

- Business Intelligence
- Image pattern recognition
- Web search
- Biology
- Security.

In marketing field clustering helps to find group of customers with similar behaviour from a given dataset customer record.

In biology classification of plants and animal according to their features.

In library clustering is very useful in book ordering

- clustering is sometimes called automatic classification.
- clustering is known as unsupervised learning because the class label information is not present.

Why?

clustering is very much important as it determines the intrinsic grouping among the unlabeled data present.

Requirements:-

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Requirements for domain knowledge to determine input parameters
- Ability to deal with noisy data.
- Incremental clustering and insensitivity to input order
- Capability of clustering high-dimentional data.
- Constraint based clustering
- Interpretability and usability.

→ we need highly scalable clustering algorithms to deal with large databases.

→ Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical and binary data.

→ clustering algorithm should be capable of detecting cluster of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.

→ High dimencionality :- The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.

→ Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to ~~to~~uch data and may lead to poor quality clusters.

Clustering Methods :-

Clustering methods can be classified into following categories.

- Partitioning method
- Hierarchical method
- Density-based method
- Grid based method

Partitioning method:-

Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of each data. Each partition will represent a cluster and $k \leq n$. It means that it classifies the data into k groups, which satisfy the following requirements

It conducts one-level partitioning on dataset. The basic partitioning methods typically adopt exclusive cluster seperation.

- Each group contains atleast one object
- Each object must belong to exactly one group.
- For a given number of partitions (say k) the partitioning method will create an initial partitioning
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

Hierarchical Methods:-

This method creates a hierarchical decomposition of the given set of data objects. There are two approaches

- Agglomerative approach (Bottom-up approach)
- Divisive approach (Top-down)

Agglomerative approach:-

This apprach is also known as bottom-up approach. In this we start with ~~all~~ each of the objects in the same cluster. In the continuous iteration forming seperate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach:-

This approach is also known as top-down approach. In this we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid i.e once a merging or splitting is done, it can never be undone.

Approaches to improve Quality of Hierarchical clustering:-

Two approaches

① Perform careful analysis of object linkages at each hierarchical partitioning.

② Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

Density-based Method:-

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold i.e for each data point within a given cluster, the radius of a given cluster has to contain atleast a minimum number of points.

Grid-based Method:-

In this method, a model is hypothesized for each cluster to find

In this the objects together form a grid. The object space is quantized to finite no. of cells that form a grid structure.

Advantages:-

- The major advantage is fast processing time.
- It is dependent only on the numbers of cells in each dimensions in the quantized space.

Method	General characteristics
Partitioning Method	- Find mutually exclusive clusters of spherical shape - Distance-based - May use mean (or) medoid to represent cluster center - Effective for small-to-medium size data sets
Hierarchical Method	- clustering is hierarchical decomposition (i.e multiple levels) - cannot correct erroneous merges or splits - May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	- can find arbitrarily shaped clusters - clusters are dense regions of objects in space that are seperated by low-density regions - cluster density - May filter out outliers.
Grid-based methods	- use a multiresolution grid data structure - fast processing time.

Partitioning Methods

① K-Means

② K- Medoids

Hierarchical Methods

① Diana

② Agnes, BIRCH, ROCK, CHAMELEON

Density - based

① DBSCAN

② OPTICS

③ DenClue

Grid based

① DB SCAN

② OPTICS

③ Denclue.

K-Means clustering algorithm:

K-Means performs division of objects into clusters. The term K is basically a number. If K=2, we have two clusters if we have K=3 then three clusters.

We have to produce K clusters.

objects $X = \{x_1, x_2, \ldots, x_m\}$

m objects

Each objects is described in terms of n features.

$$x_i = (x_{i1}, x_{i2}, x_{i3} \cdots x_{in})$$

we should get output as k clusters.

$$S = (S_1, S_2, \ldots, S_K)$$

S_i is represented by cluster center u_i;

Steps:-

① Take mean value

② Find the nearest number to mean and put it in the cluster.

③ Repeat ① & ② until we get same mean.

Eg:-

$S = \{2, 3, 4, 10, 11, 12, 20, 25, 30\}$.

If $K = 2$

Randomly we take

$m_1 = 4$ $\quad$ $m_2 = 12$

By following step ①

$K_1 = \{2, 3, 4\}$ $\quad m_1 = \frac{2+3+4}{3} = 3$.

$K_2 = \{10, 11, 12, 20, 25, 30\}$ $\quad m_2 = \frac{108}{6} = 18$

$m_1 = 3$ $\quad$ $m_2 = 18$

again nearest values to 3 & 18

$K_1 = \{2, 3, 4, 10\}$ $\quad$ $K_2 = \{11, 12, 20, 25, 30\}$

$m_1 = \frac{19}{4} = 4.75$ $m_2 = 19.6$

$m_1 = 5$ $m_2 = 20$.

Again find out nearest cluters

$K_1 = \{2, 3, 4, 10, 11, 12\}$ $K_2 = \{25, 30\}$

$\frac{25+30}{2}$

$m_1 = 7$ $m_2 = 25$

$K_1 = \{2, 3, 4, 10, 11, 12\}$ $K_2 = \{20, 25, 30\}$ } same mean value

$m_1 = 7$ 1st group

$m_2 = 25$ 2nd group

K- Medoid algorithm :- (PAM - Partitioning Around Method)

→ Arrange values in increasing order and take middle values as medoid.

→ when your data set is $\{1, 2, 4\}$ then 2 is middle one

when it is $\{1, 2, 3, 4\}$ then take average of 2 & 3 i.e $\frac{2+3}{2} = 2.5$

→ In single dimensions it is ok to arrange points in increasing order. In multi dimensional ordering is complex. Definition to higher dimension we use medoid.

$S = \{x_1, x_2 \cdots, x_n\} \subseteq R^M$ ↳ m-dimensions.

↓ data set

d = euclidean distance

K-medoid is also known as partitioning around medoid. This was proposed by Kaufman and Rousseeuw. A medoid can be defined as point in a cluster.

Medoid (c_i)

object (p_i)

$$c = \sum_{c_i} \sum_{p_i \in c_i} |p_i - c_i|$$

Algorithm:-

1. Initialize:
   select k random points out of n data points as the medoids.
2. Associate each data point to the closest medoid by using any common distance metric methods.
3. While the cost decreases
   For each medoid m, for each data point o which is not a medoid
   (i) swap m and o, associate each data point to the closest medoid, recompute the cost.
   (ii) If the total cost is more than that in previous step undo the swap

Example:-

We use Manhattan distance i.e

$$(x_1, y_1) \& (x_2, y_2)$$
$$= |x_1 - x_2| + |y_1 - y_2|$$

S.No	x	y	C_1 Distance from (3,4)	C_2 Distance from (7,4)
1	2	6	$\|2-3\|+\|6-4\| = 3$	$\|2-7\|+\|6-4\| = 7$
2	3	4		
3	3	8	$\|3-3\|+\|8-4\| = 4$	$\|3-7\|+\|8-4\| = 8$
4	4	7	$\|4-3\|+\|7-4\| = 4$	$\|4-7\|+\|7-4\| = 6$
5	6	2	$\|6-3\|+\|2-4\| = 5$	$\|6-7\|+\|2-4\| = 3$
6	6	4	$\|6-3\|+\|4-4\| = 5$	$\|6-7\|+\|4-4\| = 1$
7	7	3	$\|7-3\|+\|3-4\| = 5$	$\|7-7\|+\|3-4\| = 1$
8	7	4		
9	8	5	$\|8-3\|+\|5-4\| = 6$	$\|8-7\|+\|5-4\| = 2$
10	7	6	$\|7-3\|+\|6-4\| = 6$	$\|7-7\|+\|6-4\| = 2$

Step 1:- We first select 2 medoids (3,4) & (7,4)

Step 2:- We calculate the distance between the rest of data points and both medoids.

Step 3:- We calculate the total cost involved in forming the cluster using these medoid.

Step 4:- We again choose some other medoids & repeat step 1 to step 2. If we don't get better cost we will stop.

→ min_point

Minimum distance is considered

Total cost = 3 + 4 + 4 + 3 + 1 + 1 + 2 + 2

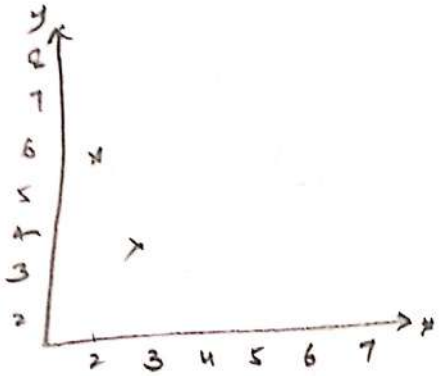
= 20

clusters with medoid {3, 4} are

{{3,4} {2,6} {4,7} {3,8}}

{7, 4} are

{{7,4} {6,2} {6,4} {7,3} {8,5} {7,6}}



Hierarchical Method:-

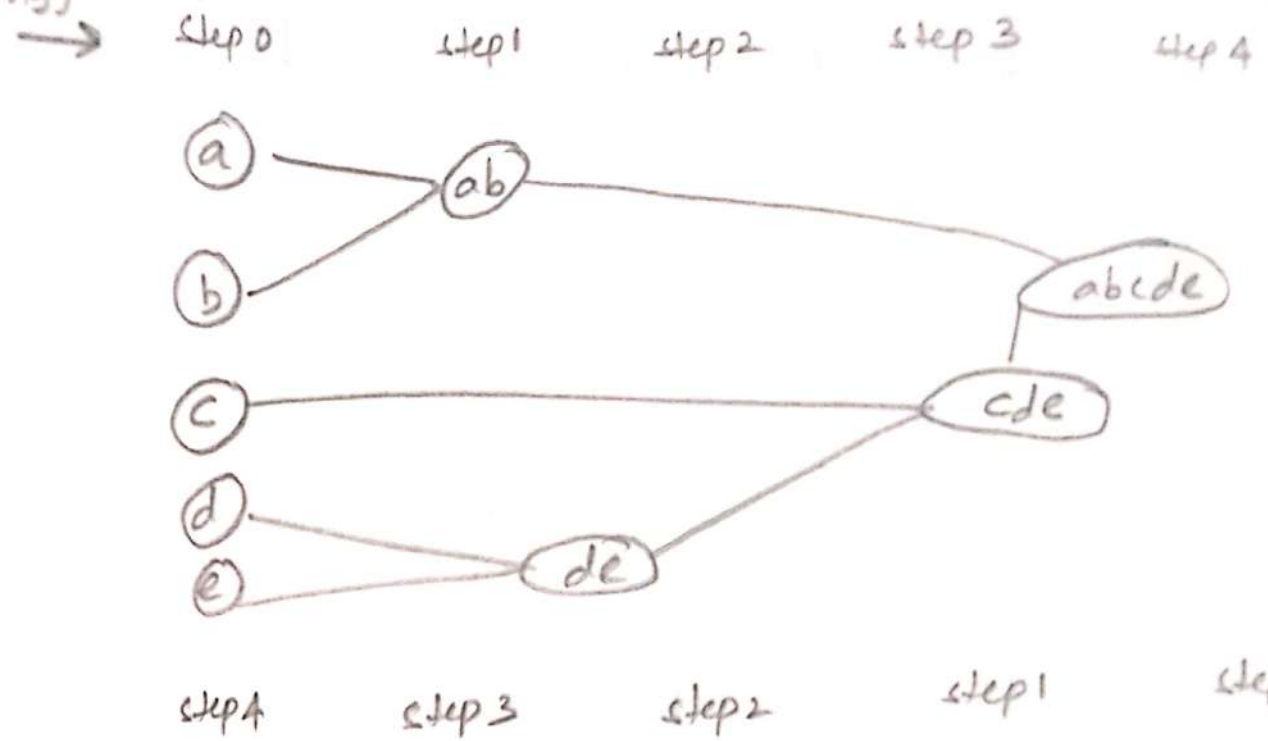
A hierarchical clustering method works by grouping data objects into a hierarchy or tree structure.

Representing data objects in the form of a hierarchy is useful for summarization and visualization.

eg:- Manager of human resources at All electronics may organize your employees into major groups such as executives, managers and staff.

senior officers — officers — trainees.

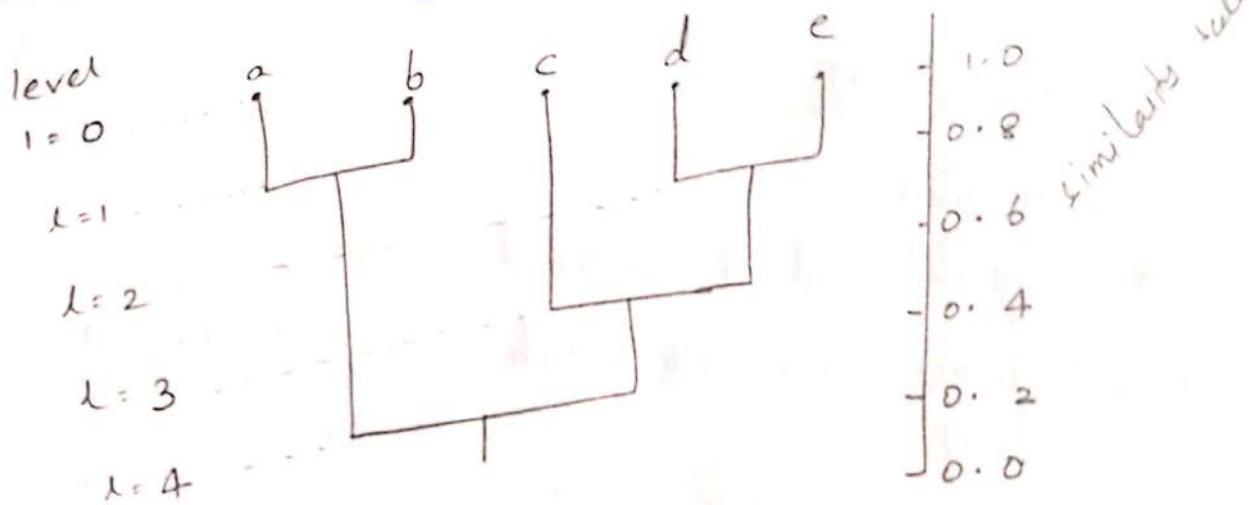
Agglomerative



Agglomerative and divisive hierarchical clustering on data objects {a, b, c, d, e}.

Dendrogram:

It is used to represent the process of hierarchical clustering.



At $l = 0$ five objects as singleton clusters.

At $l = 1$ a & b are grouped together.

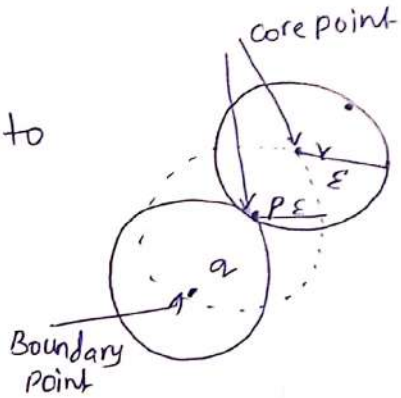
DBSCAN:

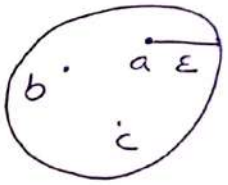
Density Based Spatial clustering of Applications with noise.

Eps.

Minimum points = 3.

→ Two points belong to same region



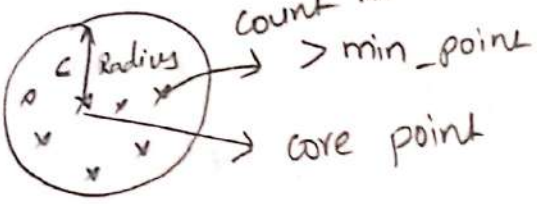


→ Density is lower.

Two parameters

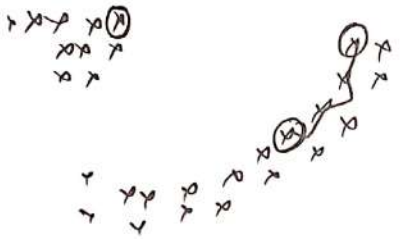
→ Min points : kind of threshold value

→ ε (epsilon) : Area over which you will perform count. (radius)



Core point is a point lies in high density.

A point is Density reachable by traversing only through core point. ②



Density connected

i & j are density connected if there exists core point K from which both of them are density reachable. then i & j are density connected.

→ i & j are in same cluster if and only if density connected

→ DBScan computation is signifant

→ All kinds of arbitraral clustering.

Optics

ordering data such points such that different points of

- Neighbourhood - ε
- Min - point
- core point
- Border point
- Noise point