Unit - 4

Clustering and Applications :-

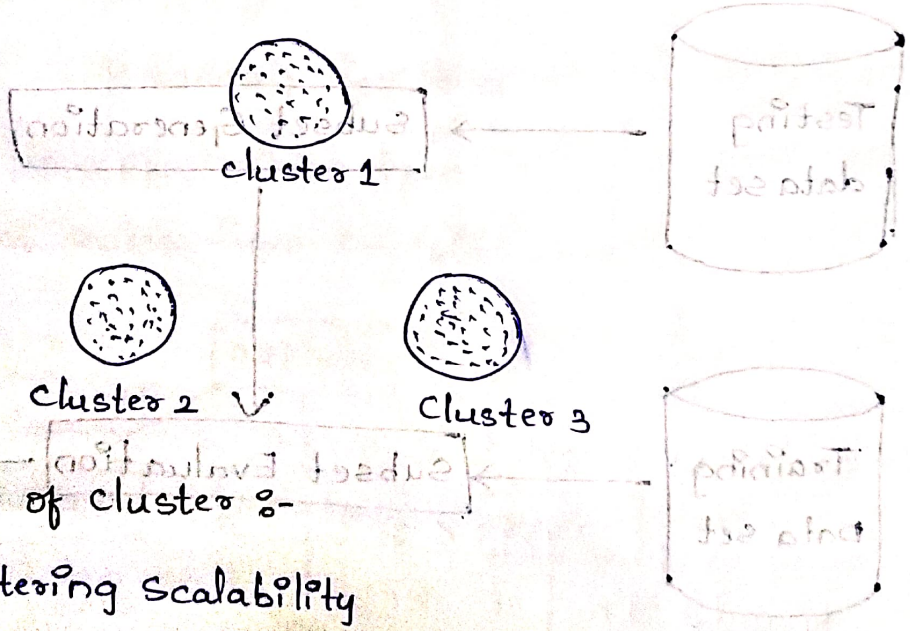
Cluster Analysis :-

Process of forming similar groups of objects together in form of a cluster.

The data items are clustered based on principles of maximising the intraclass similarities and Minimising the interclass similarities.

Ex:- In our class we are placing all the girls at one side and all the boys at one side. Girls are Girls cluster and boys are boys cluster.

- It is unsupervised Machine Learning algorithm.



Properties of cluster :-

- clustering scalability
- Algorithms usability with multiple types of data
- Dealing with Unstructured data
- Interoperatability.

Clustering methods :-

- Partioning methods
- Hierarchial Methods
- Densitity Based Methods
- Grid based Methods.

Types of Data in cluster Analysis :-

- clustering Analysis supports. Data structures.
- Data structures are of two types. They are:

1. Data Matrix
2. Dissimilarity Matrix.

1. Data Matrix :

Here, Data is represented as table (or) n by p matrix.

rows → real world entities (names)

Columns → Properties of entities.

2. Dissimilarity Matrix :

- It is represented as nxn matrix.
- It identifies dissimilarities between two objects.

$$\begin{bmatrix} 0 & & \\ d(2,1) & 0 & \\ \vdots & \vdots & \\ d(n,1) & d(n,2) & \cdots 0 \end{bmatrix}$$

Types of Data :

- There are Four types of data. They are:

1. Interval Scalled data
2. Binary Variables

3. Categorical Variables

4. Mixed Variables.

1. Interval scaled data :-

- It has continuous variables.

  Ex: 10-20, 20-30, --- etc.

- To Convert individual data into continuous variables.

  - do the ~~data~~ data standardisation before that.

- Data standardisation means removal of units.
- For standardisation data we should calculate mean absolute deviation.
- Then divide the data into intervals.

2. Binary variables :-

- It has only 2 states — 0, 1

  0 - Variables is absent
  1 - Variables is present.

- It has 2 subtypes.

  - Symmetric binary → states of variables can change
  - ~~Assymetric~~ Asymmetric binary → States of variables can't change

3. Categorical Variables :-

Data that can be divided into categories

- 2 types.

a) Nominal Variable :

It has no particular order to the categories

Ex: Gender

male Female

(can be in any order)

b) Ordinal Variable :-

It has particular internal order to the categories

Ex: Temperature

Low Medium High

(Should be in order)

4. Mixed Variables :-

Combination of different types of Variables.

Categorization of Major Clustering Methods :-

The clustering methods are categorized into Four types. They are :

1. Partioning Method
2. Hierarchial Method
3. Density-based Method
4. Grid-based Method.

1. Partioning Methods :-

- The data is divided into partitions.
- Partition represents a cluster.
- Each cluster should be less than (or) equal to size of total data items.

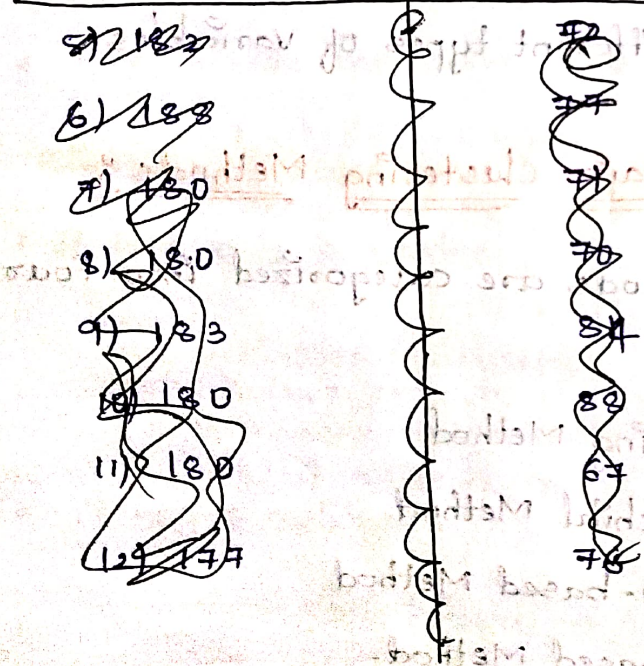
$k \leq n$

- Partition should ~~participate~~ satisfy ② Rules.
  - each partition should have atleast one object.
  - each object should be belongs to only one partition.

Example: K-Means algorithm.

- data is divided into clusters based on distance and centroid values.

Height (x)	Weight (Y).
1) 185	72
2) 170	56
3) 168	60
4) 174	68



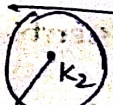
→ Divide the data into two clusters.

Each cluster should have centroid.

Cluster 1

(185, 72)

Cluster 2



(170, 56)

We can take any Order pair from example.

Now, remaining values should be divided based on Ecuclidian distance (ED)

$$ED = \sqrt{(x_0 - x_c)^2 + (y_0 - y_c)^2}$$

For ③, (168,60)

$$K_1 = \sqrt{(168-185)^2 + (60-72)^2} = 20.8$$

$$K_2 = \sqrt{(168-170)^2 + (60-56)^2} = 4.48$$

$K_2 < K_1$

It belongs to cluster 2.

Now, calculate the new centroid

(170,56) and (168,60)

$$\left(\frac{170+168}{2}, \frac{56+60}{2}\right) = (169,58)$$

$K_2 = (169,58)$

For ④, (174,68)

$$K_1 = \sqrt{(174-185)^2 + (68-72)^2} = 6.32$$

$$K_2 = \sqrt{(174-169)^2 + (68-58)^2} = 14.14$$

$K_1 < K_2$

∴ It belongs to cluster 1.

Calculate centroid for (185,72) and (174,68)

$$\left(\frac{185+174}{2}, \frac{72+68}{2}\right) = (182,70)$$

$K_1 = \{1,4\}$

$K_2 = \{2,3\}$

∴ The data is divided into two different clusters.

2. Hierarchial Methods :-

- It groups the data into tree of clusters.
- The structure is called dendrogram.
- Dendrogram have sequences of all merges and splits.
- Hierarchial method has two sub methods

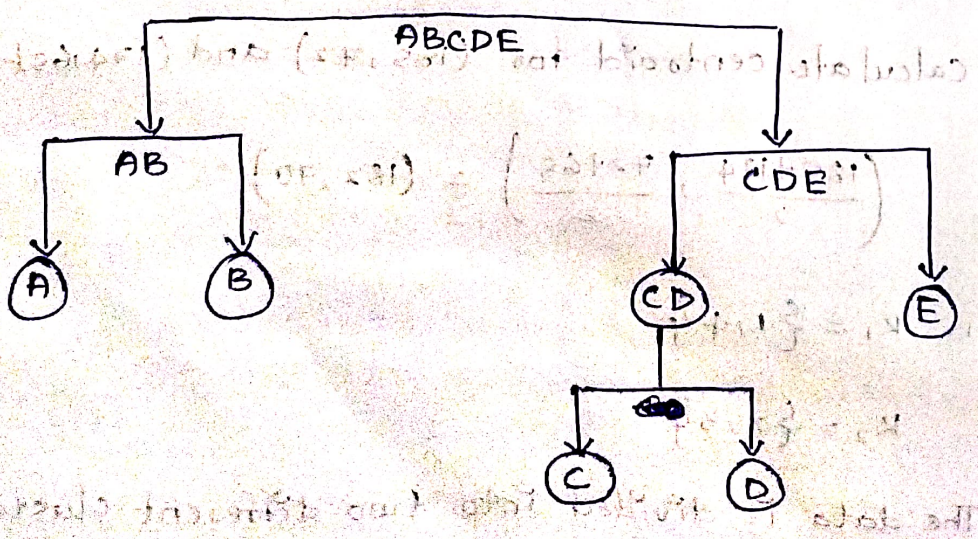
    1. Agglomerative Method
    2. Divisive Method.

1. Agglomerative Method :-

- It is bottom-up method.

Steps Involved :

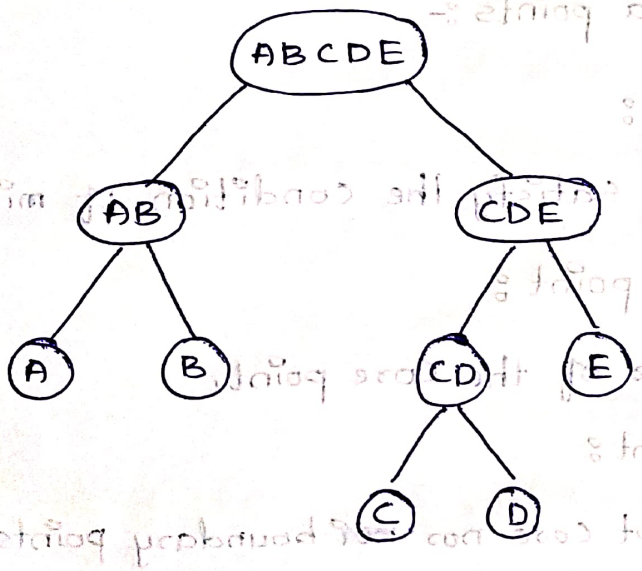
1) Calculate the similarity of one cluster with respect to all other clusters.
2) Consider every data point as Individual data.
3) Merge the clusters with highest similarity.
4) Recalculate similarity for each cluster.
5) Repeat ③ and ④ until the single cluster is obtained.



- There are 3 modes
  - single linkage - min
  - Complete linkage - max
  - Average linkage - Avg.

2. Divisive Method :-

- It is top-bottom method.
- We take all the data items into single cluster and in iterations, we split the data.
- At the end, we get N clusters.



3. Density-Based Methods :-

- Data objects are clustered based on density.
- Density means mass/volume.

Example : DBSCAN

(Density Based Spatial clustering of Applications with Noise).

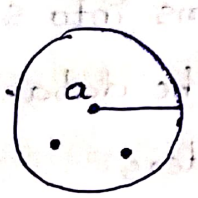
- It has a inputs
  - → ϵ - Epslon
  - → minimum points.

Where,

ϵ - radius of circle formed with data objects as centre.

min points - minimum no. of data points inside the circle

Ex: min pts = 3



Types of data points :-

1. Core point :

It should satisfy the condition of min points.

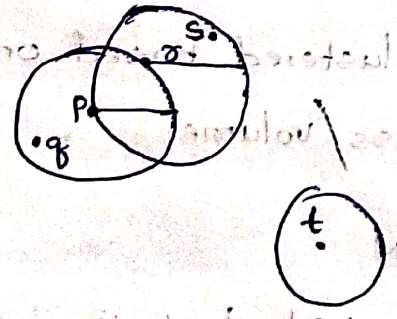
2. Boundary point :

neighbour of the core point.

3. Noise point :

It is not core nor boundary points.

Ex:



p, r → core

q, s → boundary

t → noise.

4. Grid-Based Methods :-

- It uses a multi resolution of grid data structure.
- It divides the objects into finite no. of cells that form a grid like structure.
  - then density is calculated for these cells
  - Sort the cells according to density.
  - Identify cluster centers
  - Update neighbour cells.
- Grid Based clustering is Quick processing time.

Example: STING

(Statistical Information Grid clustering Algorithm).

- Spatial data divided into rectangular cells at different levels of resolution, these cells form a tree structure.
- cells at higher levels contains small cells compared to lower levels
- clustering done based on parameters
  ↓
  mean, count, min, max.
- calculations of these parameters should start from root and go down till bottom layer.

Outlier Analysis :-

- Among the data items in Data Base, there may be some items which do not follow the general behavio -ur of data.
- Those data items are called outliers.
- Analysis of outlier is known as Outlier Analysis.

Outlier Detection :-

The process of identifying outliers and subsequently removing them.

There are two methods. They are :

1. Statistical Approach.
2. Proximity Approach.

1. Statistical Approach :

- It is based on probability of data points.
- Lowest probability is considered as Outlier.
  - parametric method.
  - Non Parametric method.

2. Proximility Approach :

- It is based on location of data points.
  - Density based Approach
  - Distance based Approach
  - Grid based Approach
  - Deviation based Approach.

Types of Outliers :-

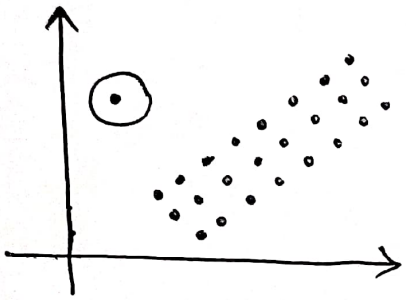
There are three types of Outlier. They are :

1. Global / point Outlier.
2. Collective outlier
3. Conditional Outlier.

1. Global / point Outlier :

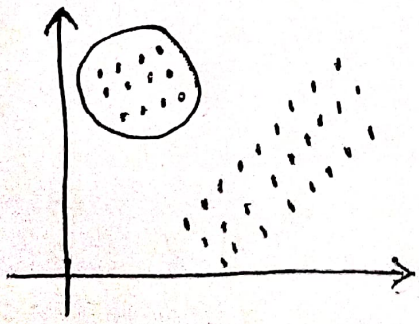
When a single data objects deviates from the rest of data.

points → Global / point outlier.



2. Collective Outlier :

When a group of data objects deviates from the rest of data.



3. conditional outlier :

Data objects deviates from others because of the specific condition.