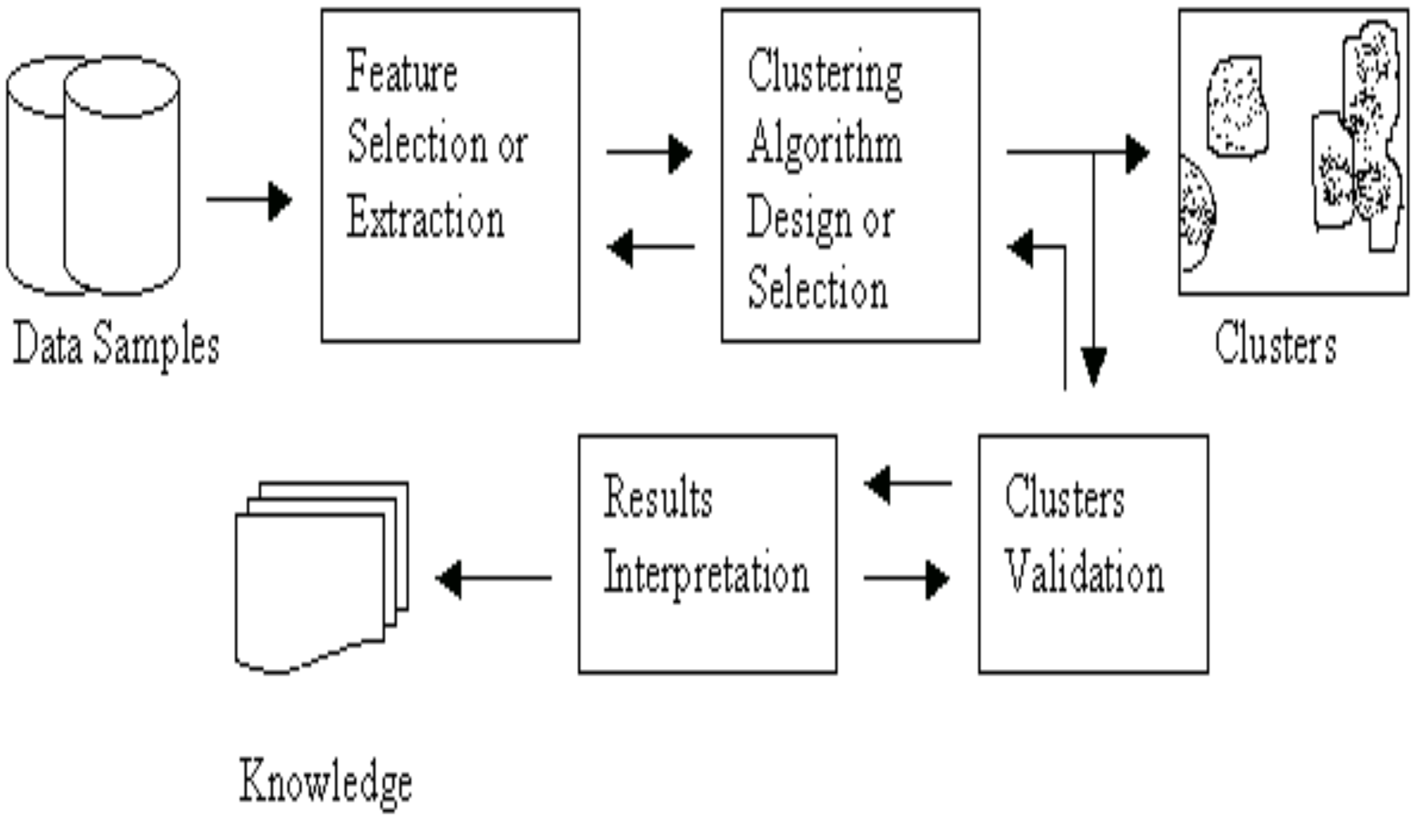
Cluster Analysis

- The aim of the clustering process is to discover overall distribution patterns and interesting correlations among the data attributes.

Steps of Clustering Process



TYPES OF CLUSTERING METHODS

- There are many clustering methods available and each of them may give a different grouping of data set. The choice of the particular method will depend on the type of output desired.

VARIOUS METHODS

- Major clustering methods are classified into following categories:

  - Partitioning methods
  - Hierarchical methods
  - Density based methods
  - Grid based methods
  - Model based methods

Partitioning methods

- The partitioning methods generally result in a set of k clusters, where $k \leq n$ (no. of objects in database).
- Clusters are formed to optimize an objective partitioning criterion (called similarity function) so that objects within a cluster are similar, whereas objects of different clusters are dissimilar in terms of database attributes.
- Well known methods: k-Means, k-Medoids

K-Means Algorithm

- K-means is one of the simplest unsupervised learning algorithms
- The main idea is to define k centroids, one for each cluster.
- The next step is to consider each data belonging to a given data set and associate it to the nearest centroid.

K-Means Algorithm

- k initial clusters are formed.
- k new centroids are re-calculated and repeat the same process until the centroids do not move any more.

K-Means algorithm

- Unfortunately, there is no general theoretical solution to find the optimal number of clusters
- A simple approach is to compare the results of multiple runs with different k classes and choose the best one according to a given criterion

K-Means Algorithm

- Computational complexity is O(nkt), t is the number of iterations.
- Can be applied only when the mean of a cluster is defined
- Can’t apply in some applications, where categorical attributes are involved

K-Means Algorithm

- K, the number of clusters is predefined
- Only suitable for discovering clusters of spherical shapes
- It is sensitive to noise and outlier data points since a small number of such data can substantially influence the mean value

K-Modes method

- One variant of k-means is the k-modes method, which extends the k-means paradigm to cluster categorical data by replacing the means of clusters with modes.

K-prototypes method

- The k-means and k-modes methods can be integrated to cluster data with mixed numeric and categorical values, resulting in the k-prototypes method.
- A new dissimilarity measures to deal with categorical objects ($d(i, j) = (p – m) / p$, s.t. p = total no. of variables and m = no. of matches) and a frequency-based method to update mode of clusters are introduced to modify the original k-means algorithm.

Cluster Validation

Cluster Validation

- Many interesting algorithms applied to analyze very large datasets. Most algorithms don't provide any means for its validation and evaluation.
- So it is very difficult to conclude which are the best clusters and should be taken for analysis.

Cluster Validation

- whatever the intention of clustering may be, the number of clusters sought is always unknown (more or less).
- Some internal measures:

  compactness, connectedness, separation and combinations

Internal measures

A: Compactness B: Connectedness C: Spatial separation

Optimal Clusters

- There are several robust strategies for predicting optimal clusters:

(1) silhouette index

(2) Dunn's index, and

(3) Davies-Bouldin (DB) index.

(4) *Xie-Beni* (XB) index

(5) *I-index*

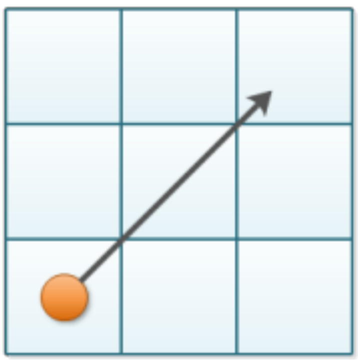
(6) *CS-index*

Distance

- Calculate a distance between 2 points **p** (x_1, y_1) and **q** (x_2, y_2) in XY-plane.
- Euclidean distance
- Chebyshev distance
- Manhattan distance
- Mahalanobis distance
- Minkowski distance

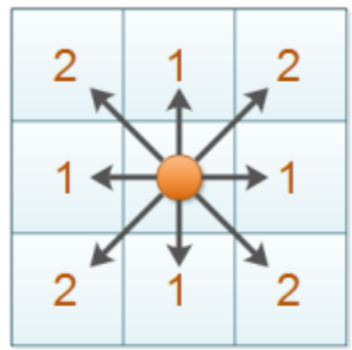
Distance

Euclidean Distance



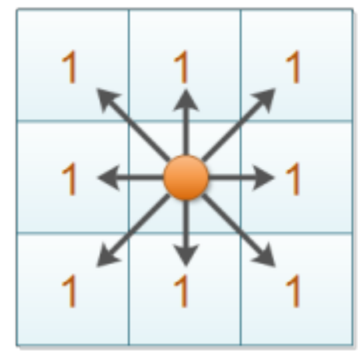
$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Manhattan Distance



$$|x_1 - x_2| + |y_1 - y_2|$$

Chebyshev Distance



$$\max(|x_1 - x_2|\,, |y_1 - y_2|)$$

Intercluster distance (δ)

- Let S and T are clusters formed using partition U. $d(x, y)$ is the distance between two objects x and y belonging to S and T respectively. $d(x, y)$ is calculated using well known metrics such as Euclidean, Manhattan and Chebychev. $|S|$ and $|T|$ are the number of objects in clusters S and T respectively.

Intercluster distance (δ)

(1) The single linkage distance is the closest distance between two objects belonging to two different clusters defined as:-.

–

$$\delta_1(S,T) = \min \left\{ \begin{matrix} d(x,y) \\ x \in S, y \in T \end{matrix} \right\}$$

Intercluster distance (δ)

(2) The complete linkage distance represents the distance between the most remote objects belonging to two different clusters, given as:-

$$\delta_2(S,T) = \max_{x \in S,\, y \in T} \left\{ d(x,y) \right\}$$

Intercluster distance (δ)

(3) The average linkage distance defines the average distance between all the objects belonging to two different clusters, described as:-

$$\delta_3(S,T) = \frac{1}{|S|\,|T|} \sum_{\substack{x \in S \\ y \in T}} d(x,y)$$

Intercluster distance (δ)

(4) The centroid linkage distance reflects the distance between the centers *vs* and *vt* of two clusters *S* and *T* respectively, presented below:-

where,

$$\delta_4(S, T) = d(v_s, v_t)$$

$$v_s = \frac{1}{|S|}\sum_{x \in S} x, \quad v_t = \frac{1}{|T|}\sum_{y \in T} y$$

Intercluster distance (δ)

(5) The average of centroids linkage represents the distance between the center of a cluster and all the objects belonging to a different cluster, explained as:-

$$\delta_5(S,T) = \frac{1}{|S|+|T|}\left\{\sum_{x \in S} d(x, v_t) + \sum_{y \in T} d(y, v_s)\right\}$$

Intracluster distance (Δ)

There are basically three types of intracluster distances:

(1) The complete diameter distance is the distance between the most remote objects belonging to the same cluster, as given below:-

$$\Delta_1(S) = \max_{x,\, y \in S} \{d(x, y)\}$$

Intracluster distance (Δ)

(2) The average diameter distance represents the average distance between all the objects belonging to the same cluster, as defined below:-

$$\Delta_2(S) = \frac{1}{|S| \cdot (|S| - 1)} \sum_{\substack{x, y \in S \\ x \neq y}} \{d(x, y)\}$$

Intracluster distance (Δ)

(3) The centroid diameter distance defines the double average distance between all of the objects and the cluster's center, as illustrated below:-

$$\Delta_3(S) = 2\left\{ \frac{\sum_{x \in S} d(x, \overline{v})}{|S|} \right\}$$

where

$$\overline{v} = \frac{1}{|S|} \sum_{x \in S} x$$

Dunn's Index

- the Dunn's validation index, DIndex, is defined as:

$$\text{Dindex}(U) = \min_{1 \le i \le c} \left\{ \min_{1 \le j \le c, j \ne i} \left\{ \frac{\delta(X_i, X_j)}{\max_{1 \le k \le c} \{\Delta(X_k)\}} \right\} \right\}$$

- $\delta(X_i, X_j)$ is the intercluster distance i.e. the distance between cluster X_i and X_j and $\Delta(X_k)$ is the intracluster distance of cluster X_k i.e. distance within the cluster X_k.

Dunn's Index

- The goal is to maximize the intercluster distances and minimizing the intracluster distances.
- The large values of Dindex corresponds to good quality cluster.
- Thus the number of clusters that maximizes DIndex is taken as the optimal number of clusters k

Davies-Bouldin index DBIndex

- Davies-Bouldin Index finds the set of clusters that are compact and well separated.
- The Davies-Bouldin index DBIndex is defined as:

$$\mathrm{DBIndex}(U) = \frac{1}{K}\sum_{i=1}^{k} \max_{i \neq j} \left\{ \frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)} \right\}$$

- Small values of DBIndex (U) represents the good quality clusters k.