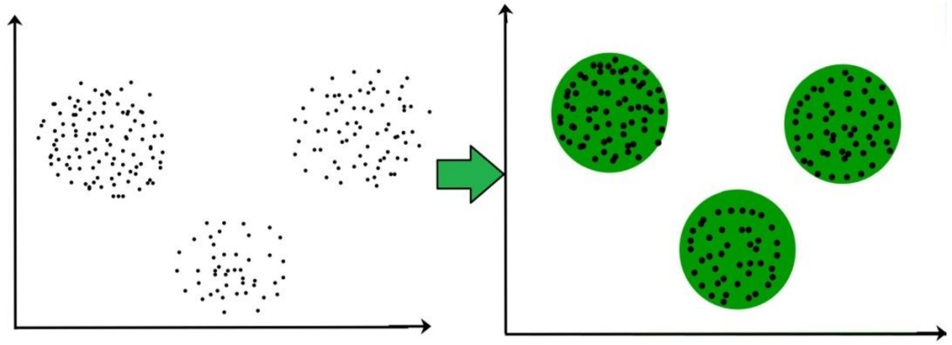
Clustering and Applications and Trends In Data Mining

- **Clustering:**
  1. Clustering is the process of making a group of abstract objects into classes of similar objects.
  2. Clustering is a Machine Learning technique that involves the grouping of data points.
  3. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.
  4. It is basically a collection of objects on the basis of similarity and dissimilarity between them.



3 clusters in the above picture.(It is not necessary for clusters to be a spherical it may be spherical as well as non spherical)

- **Cluster analysis:**
  1. Cluster analysis is a statistical classification technique in which a set of objects or points with similar characteristics are grouped together in clusters.
  2. Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.
  3. Cluster analysis is used to discover the hidden structures or relationships within data without having the need to explain or interpret what this relationship is.
  4. In essence, cluster analysis is only used to discover the structures found in data without explaining why those structures or relationships exist.
  5. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
  6. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.
- **Applications of Cluster Analysis**
  1. Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
  2. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
  3. In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.

4. Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
5. Clustering also helps in classifying documents on the web for information discovery.
6. Clustering is also used in outlier detection applications such as detection of credit card fraud.
7. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster

- **Clustering in Data Mining**
    1. **Scalability** − need highly scalable clustering algorithms to deal with large databases.
    2. **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
    3. **Discovery of clusters with attribute shape** − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
    4. **High dimensionality** − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
    5. **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
    6. **Interpretability** − The clustering results should be interpretable, comprehensible, and usable.

- **Clustering Methods**

    Clustering methods can be classified into the following categories −

    1. Partitioning Method: methods partition the objects into k clusters and each partition forms one cluster.
    2. Hierarchical Method: method creates a hierarchical decomposition of the given set of data objects.
    3. 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 neighborhood exceeds some threshold,
    4. Grid-Based Method: the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.
    5. Model-Based Method: model is hypothesized for each cluster to find the best fit of data for a given model.
    6. Constraint-based Method: clustering is performed by the incorporation of user or application-oriented constraints.
- **Need of cluster Evaluation:**

    In order to determine how well the clustering has performed there is need to evaluate cluster quality. We use number of metrics for **Evaluation of cluster**

    1. Ideal clustering is characterized by minimal intra cluster distance and maximal inter cluster distance.

2. ***Extrinsic Measures*:** which require ground truth labels? Examples are Adjusted Rand index, Fowlkes-Mallows scores, Mutual information based scores, Homogeneity, Completeness and V-measure.

3. ***Intrinsic Measures:*** that does not require ground truth labels. Some of the clustering performance measures are Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc.

- **Difference between K Means and Hierarchical clustering**

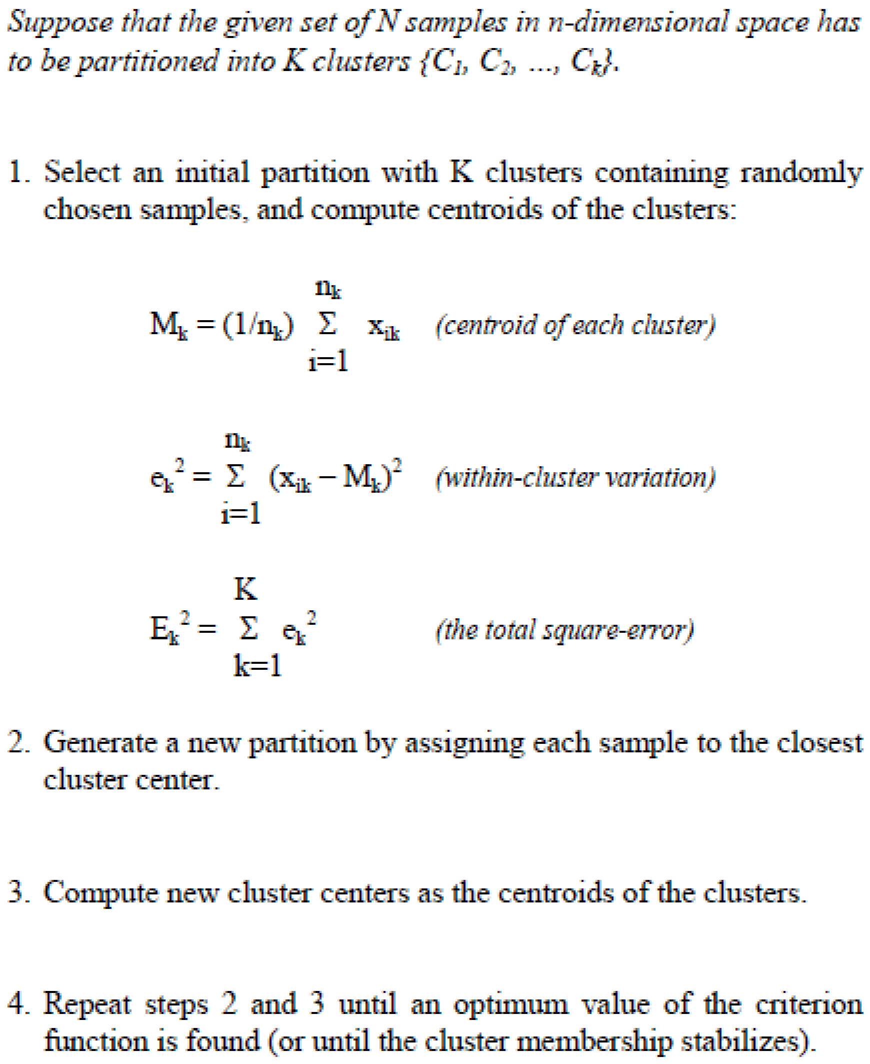
1. Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. $O(n^2)$.
2. In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
3. K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
4. K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

- **PARTITIONAL CLUSTERING**
    a. Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning.
    b. Such methods typically require that the number of clusters will be pre-set by the user.

1. K-means partition clustering algorithm:

a) The simplest and most commonly used algorithm, employing a squared error Criterion is the *K*-means algorithm.
b) One of the simplest unsupervised learning algorithms that solve the well known clustering problems.
c) The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori.
d) This algorithm partitions the data into K clusters ($C_1; C2;:::; CK$), represented by their centers or means.
e) The center of each cluster is calculated as the mean of all the instances belonging to that cluster.
f) The algorithm starts with an initial set of cluster centers, chosen at random or according to some heuristic procedure.
g) In each iteration, each instance is assigned to its nearest cluster center according to the Euclidean distance between the two.Then the cluster centers are re-calculated.

Algorithmic steps for k-means clustering



Suppose that the given set of N samples in n-dimensional space has to be partitioned into K clusters $\{C_1, C_2, \ldots, C_k\}$.

1. Select an initial partition with K clusters containing randomly chosen samples, and compute centroids of the clusters:

$$M_k = (1/n_k) \sum_{i=1}^{n_k} x_{ik} \quad \textit{(centroid of each cluster)}$$

$$e_k^2 = \sum_{i=1}^{n_k} (x_{ik} - M_k)^2 \quad \textit{(within-cluster variation)}$$

$$E_k^2 = \sum_{k=1}^{K} e_k^2 \quad \textit{(the total square-error)}$$

2. Generate a new partition by assigning each sample to the closest cluster center.

3. Compute new cluster centers as the centroids of the clusters.

4. Repeat steps 2 and 3 until an optimum value of the criterion function is found (or until the cluster membership stabilizes).

Advantages

1. Fast, robust and easier to understand.
2. Gives best result when data set are distinct or well separated from each other.

Disadvantages

1. The learning algorithm requires apriori specification of the number of cluster centers.
2. The use of Exclusive Assignment - If there are two highly overlapping data then k-means will not be able to resolve that there are two clusters.
3. Applicable only when mean is defined i.e. fails for categorical data.
4. Unable to handle noisy data and outliers.

- **Hierarchical Methods**

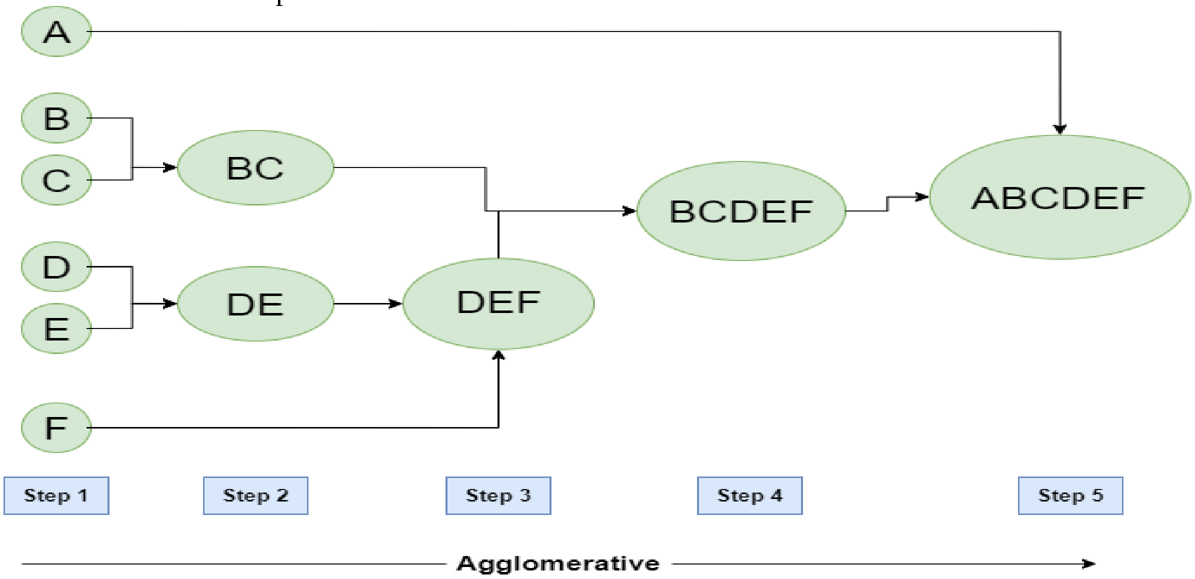
1. A **Hierarchical clustering** method works via grouping data into a tree of clusters.
2. Hierarchical clustering begins by treating every data points as a separate cluster. Then, it repeatedly executes the subsequent steps:

   a) Identify the 2 clusters which can be closest together, and
   b) Merge the 2 maximum comparable clusters. We need to continue these steps until all the clusters are merged together.

3. In Hierarchical Clustering, the aim is to produce a hierarchical series of nested clusters. A diagram called Dendrogram
4. The basic method to generate hierarchical clustering are:

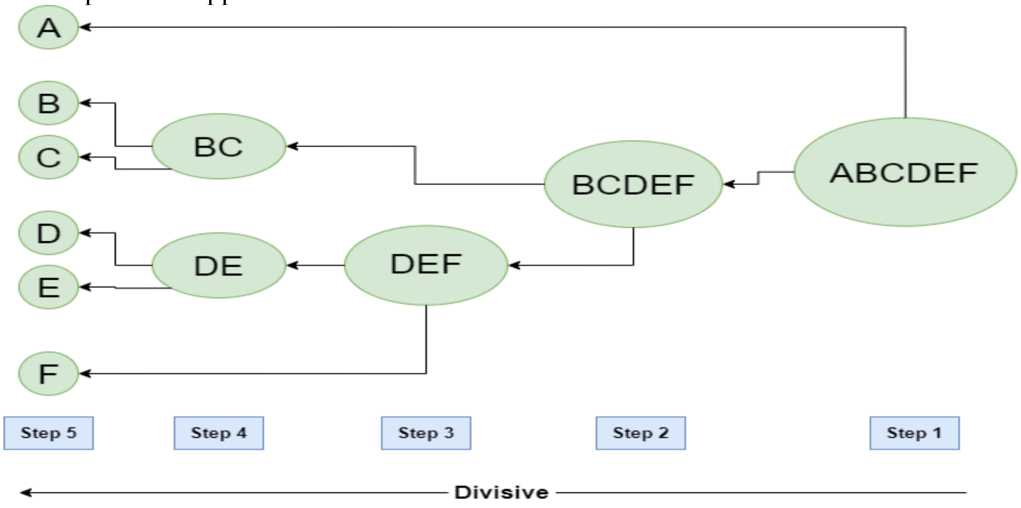
   a) **Agglomerative hierarchical clustering**—each object initially represents a cluster of its own. Then clusters are successively merged until the desired cluster structure is obtained.

      i. It is a bottom-up method



   b) **Divisive hierarchical clustering** — All objects initially belong to one cluster. Then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters. This process continues until the desired cluster structure is obtained.

      i. It is Top Down Approach



- **Density based clustering algorithm**
    1. Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density.
    2. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm.
    3. It uses the concept of density reachability and density connectivity.
        - i. **Density Reachability -** A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has sufficient number of points in its neighbors which are within distance ε.
        - ii. **Density Connectivity -** A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbors and both the points "p" and "q" are within the ε distance.
        - iii. This is chaining process. So, if "q" is neighbor of "r", "r" is neighbor of "s", "s" is neighbor of "t" which in turn is neighbor of "p" implies that "q" is neighbor of "p".

Algorithmic steps for DBSCAN clustering

Let X = {x_1, x_2, x_3, ..., x_n} be the set of data points. DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a cluster (minPts).

i. Start with an arbitrary starting point that has not been visited.

ii. Extract the neighborhood of this point using ε (All points which are within the ε distance are neighborhood).

iii. If there is sufficient neighborhood around this point then clustering process starts and point is marked as visited else this point is labeled as noise (Later this point can become the part of the cluster).

iv. If a point is found to be a part of the cluster then its ε neighborhood is also the part of the cluster and the above procedure from step 2 is repeated for all ε neighborhood points. This is repeated until all points in the cluster is determined.

v. A new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

vi. This process continues until all points are marked as visited.

Advantages

1) Does not require a-priori specification of number of clusters.
2) Able to identify noise data while clustering.
3) DBSCAN algorithm is able to find arbitrarily size and arbitrarily shaped clusters.

Disadvantages

1) DBSCAN algorithm fails in case of varying density clusters.
2) Fails in case of neck type of dataset.
3) Does not work well in case of high dimensional data.

- **Grid-based Method**
  1. In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.
  2. The main advantage of the approach is its fast processing time.
  3. The grid-based clustering approach differs from the conventional clustering algorithms in that it is concerned not with the data points but with the value space that surrounds the data points.
  4. The computational complexity of most clustering algorithms is at least linearly proportional to the size of the data set.
  5. The great advantage of grid-based clustering is its significant reduction of the computational complexity, especially for clustering very large data sets.
  6. typical grid-based clustering algorithm consists of the following five basic steps:
     - i. Creating the grid structure, i.e., partitioning the data space into a finite number of cells.
     - ii. Calculating the cell density for each cell.
     - iii. Sorting of the cells according to their densities.
     - iv. Identifying cluster centers.
     - v. Traversal of neighbor cells.

Advantages

- i. The major advantage of this method is fast processing time.
- ii. It is dependent only on the number of cells in each dimension in the quantized space.

- **Model-based methods**

1. In this method, a model is hypothesized for each cluster to find the best fit of data for a given model.
2. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.
3. This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.
4. Method uses certain models for clusters and tries to optimize the fit between the data and the models.
5. In the model-based clustering approach, the data are viewed as coming from a mixture of probability distributions, each of which represents a different cluster.

- **Data Mining Applications.**

1. **Future Healthcare**: Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs.

2. **Market Basket Analysis**: Market basket analysis is a modeling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behavior of a buyer.

3. **Education:** There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments.

4. **Manufacturing Engineering**: Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process.

5. **Fraud Detection:** Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information.