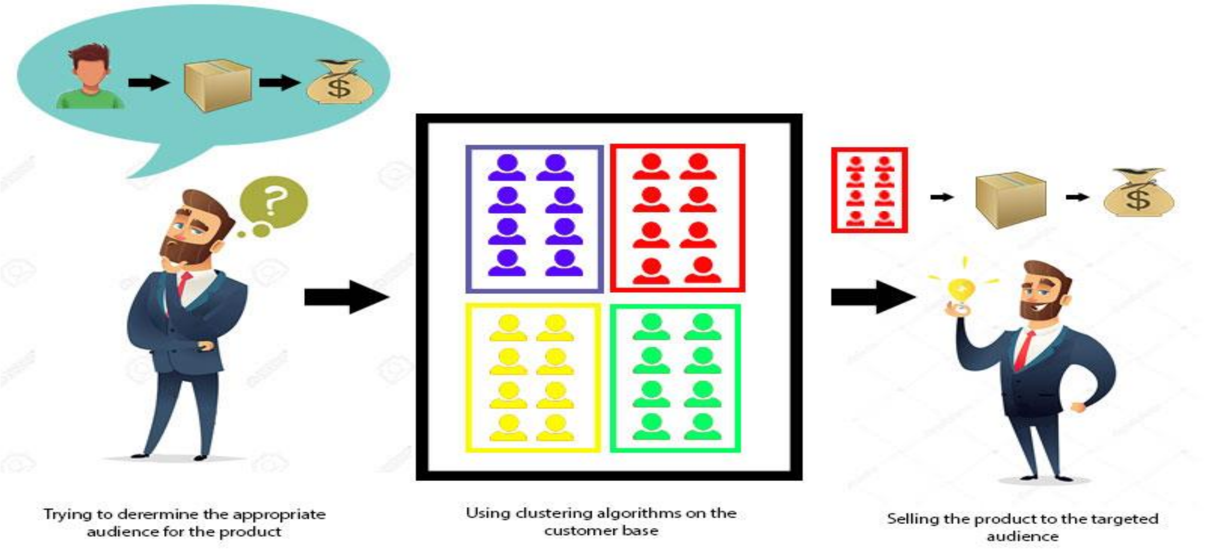
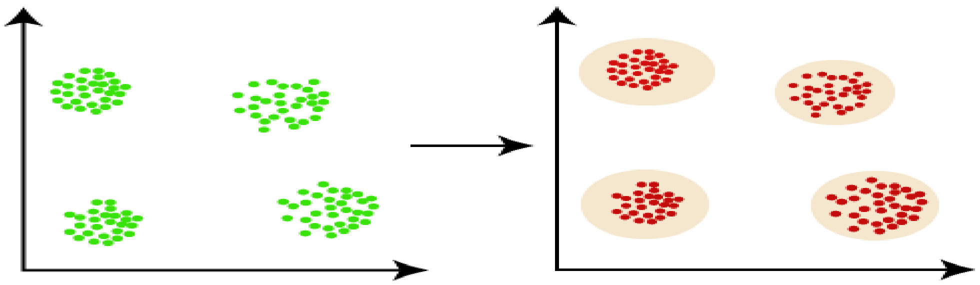
DATA MINING (COCSC16)
UNIT 5

Clustering in Data Mining

- Clustering is an unsupervised Machine Learning-based Algorithm that comprises a group of data points into clusters so that the objects belong to the same group.
- Clustering helps to splits data into several subsets. Each of these subsets contains data similar to each other, and these subsets are called clusters.
- Now that the data from our customer base is divided into clusters, we can make an informed decision about who we think is best suited for this product.


Trying to derermine the appropriate audience for the product
Using clustering algorithms on the customer base
Selling the product to the targeted audience

Let's understand this with an example, suppose we are a market manager, and we have a new tempting product to sell. We are sure that the product would bring enormous profit, as long as it is sold to the right people. So, how can we tell who is best suited for the product from our company's huge customer base?



Showing four clusters formed from the set of unlabeled data

- Clustering, falling under the category of **unsupervised machine learning**, is one of the problems that machine learning algorithms solve.
- Clustering only utilizes input data, to determine patterns, anomalies, or similarities in its input data.

A good clustering algorithm aims to obtain clusters whose:

- The intra-cluster similarities are high, it implies that the data present inside the cluster is similar to one another.
- The inter-cluster similarity is low, and it means each cluster holds data that is not similar to other data.

What is a Cluster?

- A cluster is a subset of similar objects
- A subset of objects such that the distance between any of the two objects in the cluster is less than the distance between any object in the cluster and any object that is not located inside it.
- A connected region of a multidimensional space with a comparatively high density of objects.

What is clustering in Data Mining?

- Clustering is the method of converting a group of abstract objects into classes of similar objects.
- Clustering is a method of partitioning a set of data or objects into a set of significant subclasses called clusters.
- It helps users to understand the structure or natural grouping in a data set and used either as a stand-alone instrument to get a better insight into data distribution or as a pre-processing step for other algorithms

Important points:

- Data objects of a cluster can be considered as one group.
- We first partition the information set into groups while doing cluster analysis. It is based on data similarities and then assigns the levels to the groups.
- The over-classification main advantage is that it is adaptable to modifications, and it helps single out important characteristics that differentiate between distinct groups.

Applications of cluster analysis in data mining:

- In many applications, clustering analysis is widely used, such as data analysis, market research, pattern recognition, and image processing.

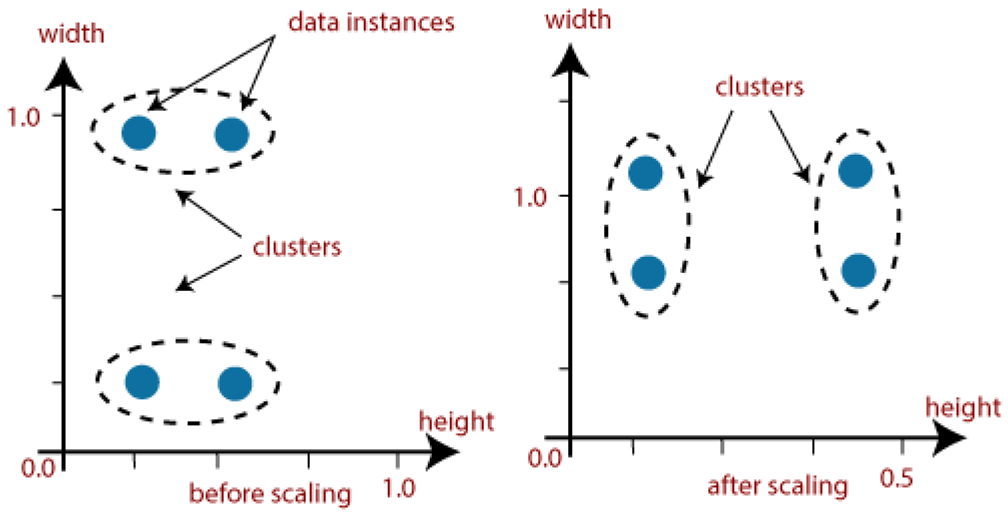
- It assists marketers to find different groups in their client base and based on the purchasing patterns. They can characterize their customer groups.
- It helps in allocating documents on the internet for data discovery.
- Clustering is also used in tracking applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to analyze the characteristics of each cluster.
- In terms of biology, it can be used to determine plant and animal taxonomies, categorization of genes with the same functionalities and gain insight into structure inherent to populations.
- It helps in the identification of areas of similar land that are used in an earth observation database and the identification of house groups in a city according to house type, value, and geographical location.

Why is clustering used in data mining?

- Clustering analysis has been an evolving problem in data mining due to its variety of applications. The advent of various data clustering tools in the last few years and their comprehensive use in a broad range of applications, including image processing, computational biology, mobile communication, medicine, and economics, must contribute to the popularity of these algorithms.
- The main issue with the data clustering algorithms is that it can't be standardized. The advanced algorithm may give the best results with one type of data set, but it may fail or perform poorly with other kinds of data set.
- Although many efforts have been made to standardize the algorithms that can perform well in all situations, no significant achievement has been achieved so far. Many clustering tools have been proposed so far.
- However, each algorithm has its advantages or disadvantages and can't work on all real situations.

1. Scalability:

- Scalability in clustering implies that as we boost the amount of data objects, the time to perform clustering should approximately scale to the complexity order of the algorithm. For example, if we perform K- means clustering, we know it is O(n), where n is the number of objects in the data. If we raise the number of data objects 10 folds, then the time taken to cluster them should also approximately increase 10 times. It means there should be a linear relationship. If that is not the case, then there is some error with our implementation process.



Showing example where scalability may leads to wrong result

- *Data should be scalable if it is not scalable, then we can't get the appropriate result. The figure illustrates the graphical example where it may lead to the wrong result.*

2. Interpretability:

- The outcomes of clustering should be interpretable, comprehensible, and usable.

3. Discovery of clusters with attribute shape:

- The clustering algorithm should be able to find arbitrary shape clusters. They should not be limited to only distance measurements that tend to discover a spherical cluster of small sizes.

4. Ability to deal with different types of attributes:

- Algorithms should be capable of being applied to any data such as data based on intervals (numeric), binary data, and categorical data.

5. Ability to deal with noisy data:

- Databases contain data that is noisy, missing, or incorrect. Few algorithms are sensitive to such data and may result in poor quality clusters.

6. High dimensionality:

- The clustering tools should not only able to handle high dimensional data space but also the low-dimensional space.

Clustering Methods:
The clustering methods can be classified into the following categories:

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method:

It is used to make partitions on the data in order to form clusters. If "n" partitions are done on "p" objects of the database then each partition is represented by a cluster and n < p. The two conditions which need to be satisfied with this Partitioning Clustering Method are:

- One objective should only belong to only one group.
- There should be no group without even a single purpose.

In the partitioning method, there is one technique called iterative relocation, which means the object will be moved from one group to another to improve the partitioning

Hierarchical Method:

In this method, a hierarchical decomposition of the given set of data objects is created. We can classify hierarchical methods and will be able to know the purpose of classification on the basis of how the hierarchical decomposition is formed. There are two types of approaches for the creation of hierarchical decomposition, they are:

- **Agglomerative Approach:** The agglomerative approach is also known as the bottom-up approach. Initially, the given data is divided into which objects form separate groups. Thereafter it keeps on merging the objects or the groups that are close to one another which means that they exhibit similar properties. This merging process continues until the termination condition holds.

- **Divisive Approach:** The divisive approach is also known as the top-down approach. In this approach, we would start with the data objects that are in the same cluster. The group of individual clusters is divided into small clusters by continuous iteration. The iteration continues until the condition of termination is met or until each cluster contains one object.

Once the group is split or merged then it can never be undone as it is a rigid method and is not so flexible. The two approaches which can be used to improve the Hierarchical Clustering Quality in Data Mining are: –

- One should carefully analyze the linkages of the object at every partitioning of hierarchical clustering.

- One can use a hierarchical agglomerative algorithm for the integration of hierarchical agglomeration. In this approach, first, the objects are grouped into micro-clusters. After grouping data objects into microclusters, macro clustering is performed on the microcluster.

Density-Based Method:

The density-based method mainly focuses on density. In this method, the given cluster will keep on growing continuously 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 at least a minimum number of points.

Grid-Based Method:

In the Grid-Based method a grid is formed using the object together, the object space is quantized into a finite number of cells that form a grid structure. One of the major advantages of the grid-based method is fast processing time and it is dependent only on the number of cells in each dimension in the quantized space. The processing time for this method is much faster so it can save time.

Model-Based Method:

In the model-based method, all the clusters are hypothesized in order to find the data which is best suited for the model. The clustering of the density function is used to locate the clusters for a given model. It reflects the spatial distribution of data points and also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. Therefore, it yields robust clustering methods.

Constraint-Based Method:

The constraint-based clustering method is performed by the incorporation of application or user-oriented constraints. A constraint refers to the user expectation or the properties of the desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. The user or the application requirement can specify constraints.

Hierarchical clustering in data mining

- Hierarchical clustering refers to an unsupervised learning procedure that determines successive clusters based on previously defined clusters.
- It works via grouping data into a tree of clusters. Hierarchical clustering stats by treating each data points as an individual cluster.
- The endpoint refers to a different set of clusters, where each cluster is different from the other cluster, and the objects within each cluster are the same as one another.

There are two types of hierarchical clustering

- Agglomerative Hierarchical Clustering
- Divisive Clustering

Agglomerative hierarchical clustering

- Agglomerative clustering is one of the most common types of hierarchical clustering used to group similar objects in clusters. Agglomerative clustering is also known as AGNES (Agglomerative Nesting).
- In agglomerative clustering, each data point act as an individual cluster and at each step, data objects are grouped in a bottom-up method. Initially, each data object is in its cluster.
- At each iteration, the clusters are combined with different clusters until one cluster is formed.

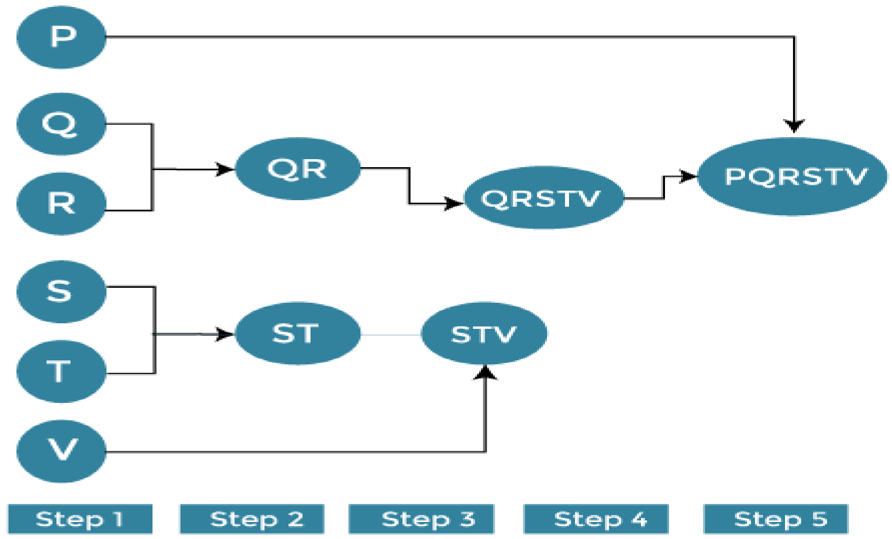
Agglomerative hierarchical clustering algorithm

1. Determine the similarity between individuals and all other clusters. (Find proximity matrix).
2. Consider each data point as an individual cluster.
3. Combine similar clusters.
4. Recalculate the proximity matrix for each cluster.
5. Repeat step 3 and step 4 until you get a single cluster.

Let’s understand this concept with the help of graphical representation using a dendrogram.

With the help of given demonstration, we can understand that how the actual algorithm work. Here no calculation has been done below all the proximity among the clusters are assumed.

Let's suppose we have six different data points P, Q, R, S, T, V.



Step 1:

Consider each alphabet (P, Q, R, S, T, V) as an individual cluster and find the distance between the individual cluster from all other clusters.

Step 2:

Now, merge the comparable clusters in a single cluster. Let's say cluster Q and Cluster R are similar to each other so that we can merge them in the second step. Finally, we get the clusters [(P), (QR), (ST), (V)]

Step 3:

Here, we recalculate the proximity as per the algorithm and combine the two closest clusters [(ST), (V)] together to form new clusters as [(P), (QR), (STV)]

Step 4:

Repeat the same process. The clusters STV and PQ are comparable and combined together to form a new cluster. Now we have [(P), (QQRSTV)].

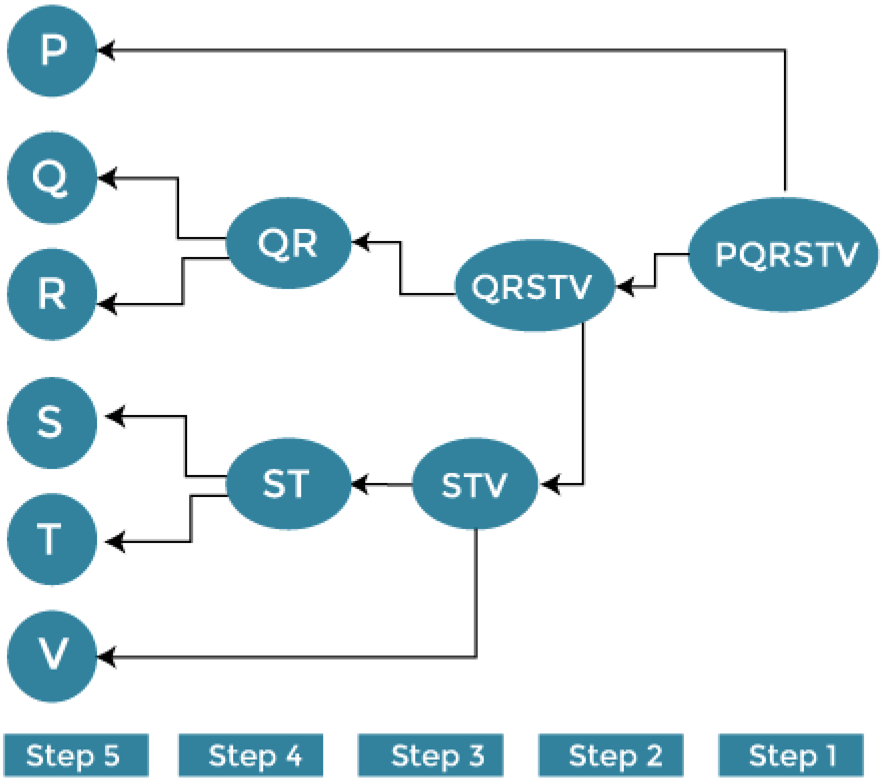
Step 5:

Finally, the remaining two clusters are merged together to form a single cluster [(PQRSTV)]

Divisive Hierarchical Clustering

- Divisive hierarchical clustering is exactly the opposite of Agglomerative Hierarchical clustering.
- In Divisive Hierarchical clustering, all the data points are considered an individual cluster, and in every iteration, the data points that are not similar are separated from the

cluster. The separated data points are treated as an individual cluster. Finally, we are left with N clusters.



Advantages of Hierarchical clustering

- It is simple to implement and gives the best output in some cases.
- It is easy and results in a hierarchy, a structure that contains more information.
- It does not need us to pre-specify the number of clusters.

Disadvantages of hierarchical clustering

- It breaks the large clusters.
- It is Difficult to handle different sized clusters and convex shapes.
- It is sensitive to noise and outliers.
- The algorithm can never be changed or deleted once it was done previously.

Density-based clustering in data mining

Density-based clustering refers to a method that is based on local cluster criterion, such as density connected points. In this tutorial, we will discuss density-based clustering with examples.

What is Density-based clustering?

Density-Based Clustering refers to one of the most popular unsupervised learning methodologies used in model building and machine learning algorithms. The data points in the region separated by two clusters of low point density are considered as noise. The surroundings with a radius ε of a given object are known as the ε neighborhood of the object. If the ε neighborhood of the object comprises at least a minimum number, MinPts of objects, then it is called a core object.

Density-Based Clustering - Background

There are two different parameters to calculate the density-based clustering

EPS: It is considered as the maximum radius of the neighborhood.

MinPts: MinPts refers to the minimum number of points in an Eps neighborhood of that point.

NEps (i) : { k belongs to D and dist (i,k) < = Eps}

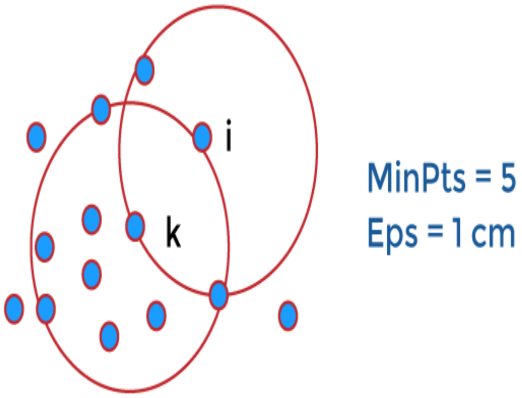
Directly density reachable:

A point i is considered as the directly density reachable from a point k with respect to Eps, MinPts if

i belongs to NEps(k)

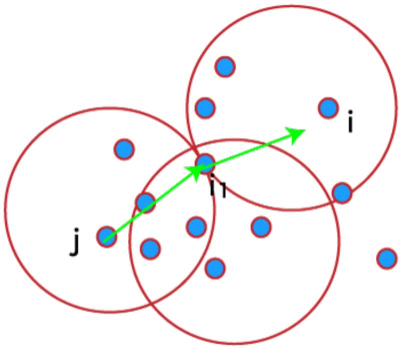
Core point condition:

NEps (k) >= MinPts



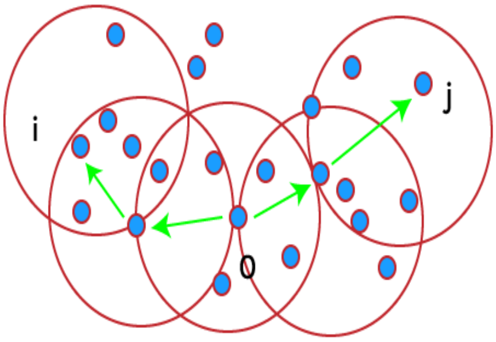
Density reachable:

A point denoted by i is a density reachable from a point j with respect to Eps, MinPts if there is a sequence chain of a point i1,…., in, i1 = j, pn = i such that i_i + 1 is directly density reachable from $i_{i.}$



Density connected:

A point i refers to density connected to a point j with respect to Eps, MinPts if there is a point o such that both i and j are considered as density reachable from o with respect to Eps and MinPts.



Working of Density-Based Clustering

Suppose a set of objects is denoted by D', we can say that an object I is directly density reachable form the object j only if it is located within the ε neighborhood of j, and j is a core object.

An object i is density reachable form the object j with respect to ε and MinPts in a given set of objects, D' only if there is a sequence of object chains point i1,…., in, i1 = j, pn = i such that i_i + 1 is directly density reachable from i_i with respect to ε and MinPts.

An object i is density connected object j with respect to ε and MinPts in a given set of objects, D' only if there is an object o belongs to D such that both point i and j are density reachable from o with respect to ε and MinPts.

Major Features of Density-Based Clustering

The primary features of Density-based clustering are given below.

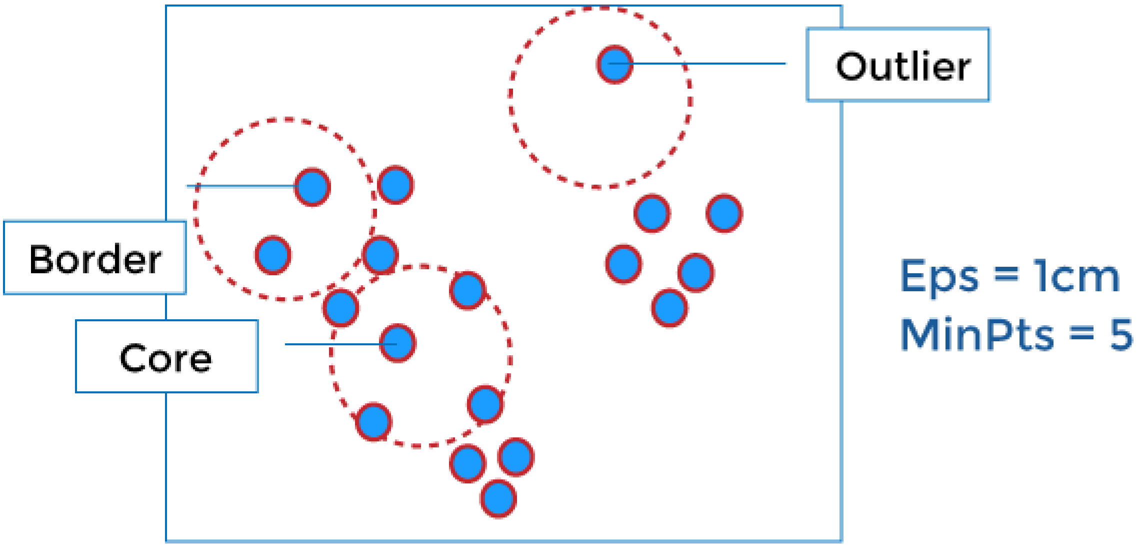
- It is a scan method.
- It requires density parameters as a termination condition.

- It is used to manage noise in data clusters.
- Density-based clustering is used to identify clusters of arbitrary size.

Density-Based Clustering Methods

DBSCAN

DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. It depends on a density-based notion of cluster. It also identifies clusters of arbitrary size in the spatial database with outliers.



OPTICS

OPTICS stands for Ordering Points to Identify the Clustering Structure. It gives a significant order of database with respect to its density-based clustering structure. The order of the cluster comprises information equivalent to the density-based clustering related to a long range of parameter settings. OPTICS methods are beneficial for both automatic and interactive cluster analysis, including determining an intrinsic clustering structure.

DENCLUE

Density-based clustering by Hinnebirg and Kiem. It enables a compact mathematical description of arbitrarily shaped clusters in high dimension state of data, and it is good for data sets with a huge amount of noise.

Partitioning Method (K-Mean) in Data Mining

Partitioning Method:

- This clustering method classifies the information into multiple groups based on the characteristics and similarity of the data.
- It’s the data analysts to specify the number of clusters that has to be generated for the clustering methods. In the partitioning method when database(D) that contains multiple(N) objects then the partitioning method constructs user-specified(K) partitions of the data in which each partition represents a cluster and a particular region.
- There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Medoids), CLARA algorithm (Clustering Large Applications) etc.

- **K-Mean (A centroid based Technique):** The K means algorithm takes the input parameter K from the user and partitions the dataset containing N objects into K clusters so that resulting similarity among the data objects inside the group (intracluster) is high but the similarity of data objects with the data objects from outside the cluster is low (intercluster). The similarity of the cluster is determined with respect to the mean value of the cluster. It is a type of square error algorithm. At the start randomly k objects from the dataset are chosen in which each of the objects represents a cluster mean(centre). For the rest of the data objects, they are assigned to the nearest cluster based on their distance from the cluster mean. The new mean of each of the cluster is then calculated with the added data objects.

Algorithm: K mean:
Input:
K: The number of clusters in which the dataset has to be divided
D: A dataset containing N number of objects

Output:
A dataset of K clusters
Method:

1. Randomly assign K objects from the dataset(D) as cluster centres(C)
2. (Re) Assign each object to which object is most similar based upon mean values.
3. Update Cluster means, i.e., Recalculate the mean of each cluster with the updated values.
4. Repeat Step 2 until no change occurs.

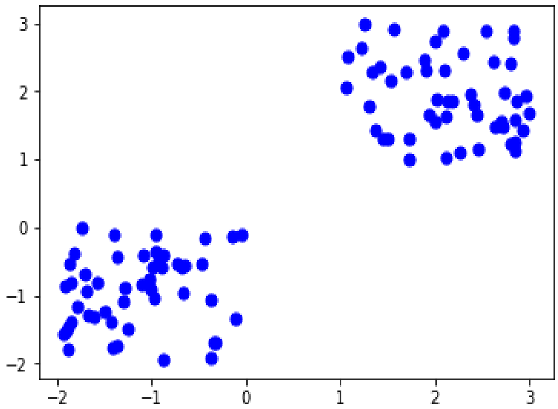


Figure – K-mean Clustering

Flowchart:

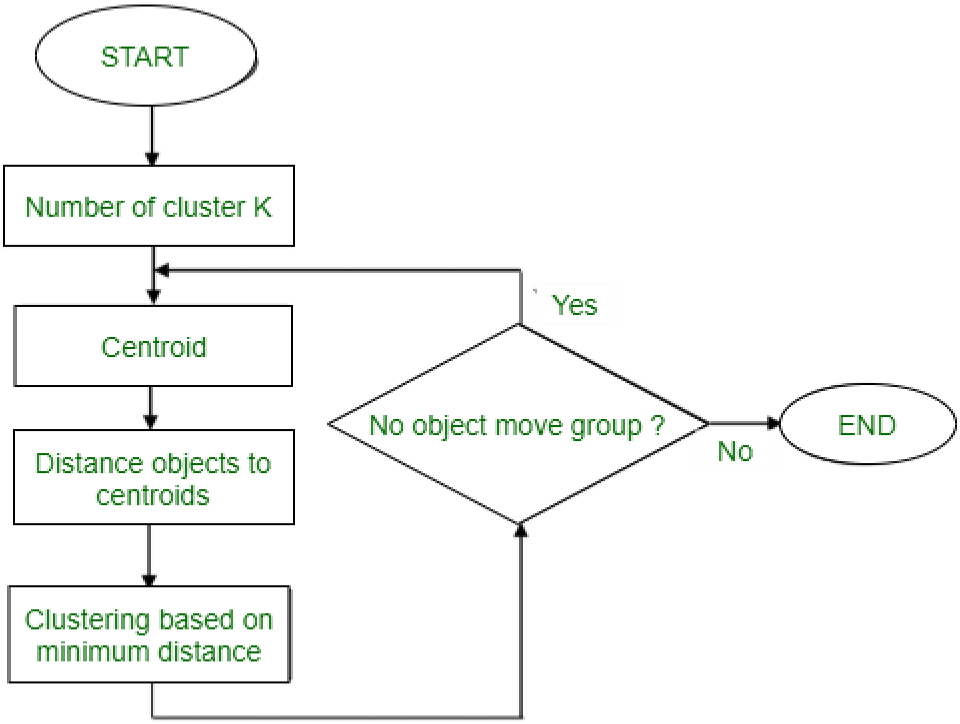

START
Number of cluster K
Centroid
Distance objects to centroids
Clustering based on minimum distance
No object move group ?
Yes
No
END

Figure – K-mean Clustering

Example: Suppose we want to group the visitors to a website using just their age as follows:
16, 16, 17, 20, 20, 21, 21, 22, 23, 29, 36, 41, 42, 43, 44, 45, 61, 62, 66

Initial Cluster:
K=2

Centroid(C1) = 16 [16]

Centroid(C2) = 22 [22]

Note: These two points are chosen randomly from the dataset. **Iteration-1:**
C1 = 16.33 [16, 16, 17]

C2 = 37.25 [20, 20, 21, 21, 22, 23, 29, 36, 41, 42, 43, 44, 45, 61, 62, 66]

Iteration-2:
C1 = 19.55 [16, 16, 17, 20, 20, 21, 21, 22, 23]

C2 = 46.90 [29, 36, 41, 42, 43, 44, 45, 61, 62, 66]

Iteration-3:
C1 = 20.50 [16, 16, 17, 20, 20, 21, 21, 22, 23, 29]

C2 = 48.89 [36, 41, 42, 43, 44, 45, 61, 62, 66]

Iteration-4:
C1 = 20.50 [16, 16, 17, 20, 20, 21, 21, 22, 23, 29]

C2 = 48.89 [36, 41, 42, 43, 44, 45, 61, 62, 66]

No change Between Iteration 3 and 4, so we stop. Therefore, we get the clusters **(16-29)** and **(36-66)** as 2 clusters we get using K Mean Algorithm.