**CLUSTER ANALYSIS**

**Definition:** Clustering is a technique which groups or makes clusters of data only based on the information given by data which describe the objects and its relationship. In clusters the objects within the clusters will be similar and will have same properties to one another and will be different from the objects of other clusters, i.e. group of statistical methods that are used for identifying groups ("clusters") of similar items in multidimensional space.

The greater the similarity between intra cluster and greater the difference between inter clusters, better and distinct will be clustering.

Merging

Similarity computation

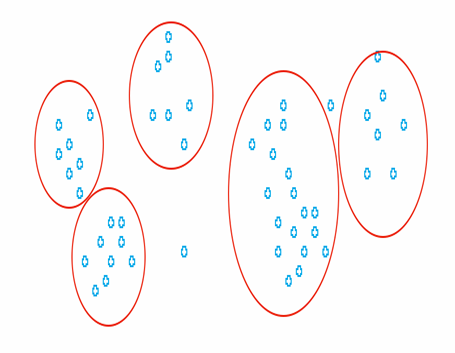
Feature Extraction

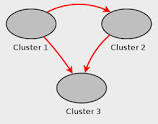
Objects

Object

Clusters

object





Here in cluster 1,cluster 2 and cluster 3 properties of the objects will be same within the clusters but different from other clusters.

**Types of Clustering:**

1. **Hierarchical Clustering.**
2. **Non- Hierarchical Clustering.**
3. **Fuzzy Clustering.**
4. **Hierarchical**: Hierarchical clustering creates a "tree" of clusters, with, at the bottom level, every item in its own cluster, and at the top level all items in one cluster.

It can be done with two ways:

1. **Bottom Up**: In this method n individual clusters will start from bottom and progressively merge as one move up the hierarchy and finally make a one cluster of all n individual clusters.
2. **Top Down:**In this method one cluster will start from up and progressively split as one moves down the hierarchy and finally make n individual clusters from one cluster.

Hierarchicalmethod is based on many Algorithms :

1. **Single Linkage**: In this method Clusters are merged with the minimum distance between the nearest two points in each cluster.
2. **Complete Linkage**: In this method Clusters are merged with the minimum distance between the farthest points in each cluster.
3. **Average Linkage**: In this method the minimum value of the mean distance between all pairs in the two clusters.

**Disadvantage**: Due to their computational complexity and memory requirements, hierarchical methods do not scale well to very large datasets.

1. **Non-Hierarchical**: It is also called as Partitional Clustering. In this all algorithms produces single partitioning of the dataset into different clusters.

The main algorithm of this method is:

**K-Means**: In this method we have to mention the number of clusters in advance based on our requirement. In this geometric center(Centroid) of the clusters computed first, then each item placed into the nearest cluster based on similarity between the items and each of the cluster centroid. After one follow the process centroid will be recalculated and it will change the assignments of clusters, then the process will be repeated until no more items change cluster, i.e clusters are stable.

**Advantages:**

1. K-Means clustering are very effective for large dataset with good speed compare to other methods.
2. By default always we use K-Means clustering because it’s easy to use.

**Disadvantages:**

1. In K-Means we have to give cluster number but it is not effective or obvious that how many clusters should be made for good result.
2. K-Means method is not effective if outliers are present.
3. It is not giving correct result if size, shapes of clusters are different and cluster is empty.
4. It cannot deal with overlapping of local minima of clusters.

**c)Fuzzy Clustering**: Till now we have seen that objects in two clusters will be totally different from each other but some time it happens that objects have two different types of properties and it belongs to two or three clusters. It will have partial properties of all three clusters but according to the more properties it will have membership degrees like 0,1,2 etc.

**Advantages:**

1. This method is good for large dataset.
2. It describes properly about local minima of clusters and deal effectively.
3. We can use this method to design global solution.

**Disadvantages:**

1. This method takes more time than other methods which is not advisable if we want to see result second wise or minute wise.
2. It does not recognize natural clusters.
3. We have to define number of cluster in advance.

**How to use clustering in dataset:**

**Method:**

1. If dataset is not large enough then Hierarchical Clustering Method is the best method for finding the clusters and results will be with less error.
2. If dataset is large then will go for K-Means Clustering Method and we will define number of clusters in advance.
3. We will see whether clusters are overlapping or not by seeing the range of the clusters.
4. If clusters are overlapping then we will apply Fuzzy Clustering Method.

**For Ruchi Soya how we have used Clustering**

**Objective**:In Ruchi Soya we have to calculate Projected Excessive Conversion Cost for 1 hour,2 hour and 24 hour.

We have used following steps:

1. Projected Excessive Conversion Cost data is in interval wise like when its running 1 interval and when its not running all values are zeros and it is 2ndinterval.All observation is divided into two intervals.
2. First we have tried to make different models for different intervals and then want to see overall result in running time, But it was not giving proper result and models were failing after few points(20 etc).
3. Then we have divided the full dataset into 5 different clusters using K-Means method. Every cluster was having different property from other clusters and then we built different models for every clusters and then we saw that result improved compare to earlier.

**Codes For Three Methods in R:**

1. **Hierarchical Method:**

d <- dist(X, method = "euclidean")   # find distance matrix of the data.  
 hc <- hclust(d)  # In hierarchical it takes Complete Linkage Method for clustering.

Tree<- cutree(hc, k = 10) # Tree for 10 clusters.

1. **K-Means Method:**

cluster<- kmeans (X, number of clusters, iter.max = 10) # By default number of iteration

is 10.

1. **Fuzzy Methods:**

Fuzzy<-fanny(x, k, memb.exp = 2) # by default it takes Euclidean, default membership exponent is 2.