CLUSTERING

Searching the data for a structure of natural grouping is an important exploratory technique. Grouping can provide information for assessing dimensionality, identifying outliers and suggests interesting hypothesis about relationships.

Cluster analysis is a primitive technique in that no assumptions are made regarding number of groups or group structure. Grouping is done on the basis of similarity or distances. The inputs required are similarity measures or data from which similarity can be computed. CLUSTER ANALYSIS IS AN UNSUPERVISED LEARNING TECHNIQUE WHICH AIMS AT GROUPING SET OF OBJECTS INTO CLUSTERS, SUCH THAT OBJECTS IN THE SAME CLUSTER SHOULD BE SIMILAR AS MUCH AS POSSIBLE WHERE AS OBJECTS IN ONE CLUSTER SHOULD BE DISSIMILAR AS MUCH AS POSSIBLE WITH THE OTHER CLUSTER.

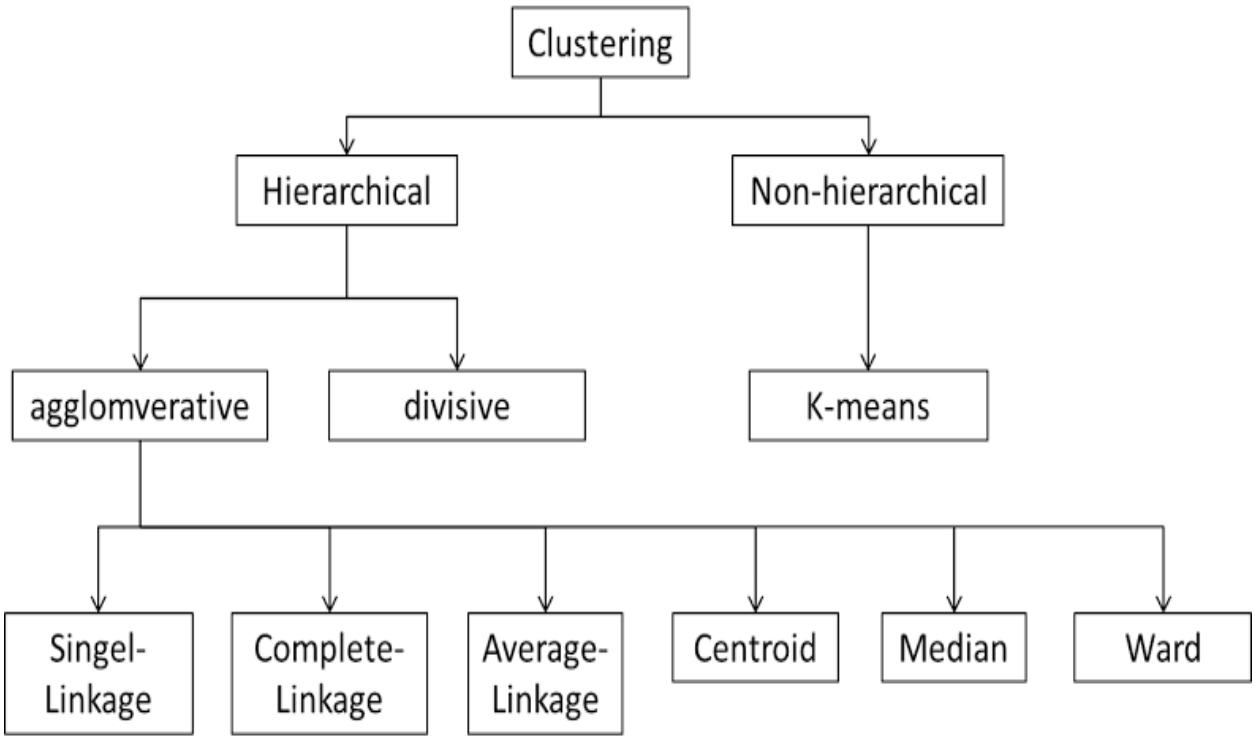

Figure-1

Figure-1 shows how clustering method can be classified.

Cluster classification.

1. Hierarchical clustering:

In hierarchical clustering method we proceed either by series of merges or by series of successive division. So it further classified into two groups.

1.1 Agglomerative hierarchical clustering:

The method in which clusters are formed by a series of merges is called as agglomerative hierarchical clustering. In this method initially there are as many cluster as objects. The most similar objects are first grouped and then initial groups are merged according to their similarities. Eventually the similarities decreases and all subgroups are fused into single cluster. Or else if we specify the number of clusters we have to form, until we get the number of specified clusters the grouping procedure continuous.

Mechanism for agglomerative hierarchical clustering as follows.

I. Given 'p' objects, start with 'p' clusters each having single element. Compute distance matrix $D_{(p\times p)}$.

II. Search $D_{(p\times p)}$ for the neariest pair of cluster. Let distance between most similar cluster U, V be d_{UV}.

III. Merge cluster U and V to form a new cluster (U,V). Update $D_{(p\times p)}$ by,

- Deleting row and column corresponding to U and V.
- Adding a row and column which gives distance between cluster (U,V) with remaining clusters.

IV. Repeat the step (II) until all clusters are merged into a single cluster.

Linkage methods:

Linkage methods describes the way in which grouping of observations is taking place. There are three linkage methods.

1.1.1 Single(Minimum) linkage method:

Mechanism for single linkage method is as follows.

i. Initially find smallest distance in $D_{(p\times p)}$ and merge the corresponding objects say U and V to form a new cluster (U,V).

ii. Compute the distance between cluster (U,V) and any other cluster say W by computing the distance $d_{(UV)W}$ as,

$$d_{(UV)W} = \text{minimum } (d_{UW}, d_{VW})$$

Where d_{UW} is the distance between cluster U and W.

d_{VW} the distance between cluster V and W.

iii. Continue step (i) and (ii) until the objects are merged into single cluster.

1.1.2 Complete(Maximum) linkage method:

Mechanism for complete linkage method is as follows.

i. Initially find smallest distance in $D_{(p\times p)}$ and merge the corresponding objects say U and V to form a new cluster (U,V).

ii. Compute the distance between cluster (U,V) and any other cluster say W by computing the distance $d_{(UV)W}$ as,

$d_{(UV)W}$ = maximum (d_{UW}, d_{VW})

Where d_{UW} is the distance between cluster U and W.

d_{VW} the distance between cluster V and W.

iii. Continue step (i) and (ii) until the objects are merged into single cluster.

1.1.3 Average linkage method:

Mechanism for average linkage method is as follows.

i. Initially find smallest distance in $D_{(p\times p)}$ and merge the corresponding objects say U and V to form a new cluster (U,V).

ii. Compute the distance between cluster (U,V) and any other cluster say W by computing the distance $d_{(UV)W}$ as,

$d_{(UV)W}$ = average (d_{UW}, d_{VW})

Where d_{UW} is the distance between cluster U and W.

d_{VW} the distance between cluster V and W.

iii. Continue step (i) and (ii) until the objects are merged into single cluster.

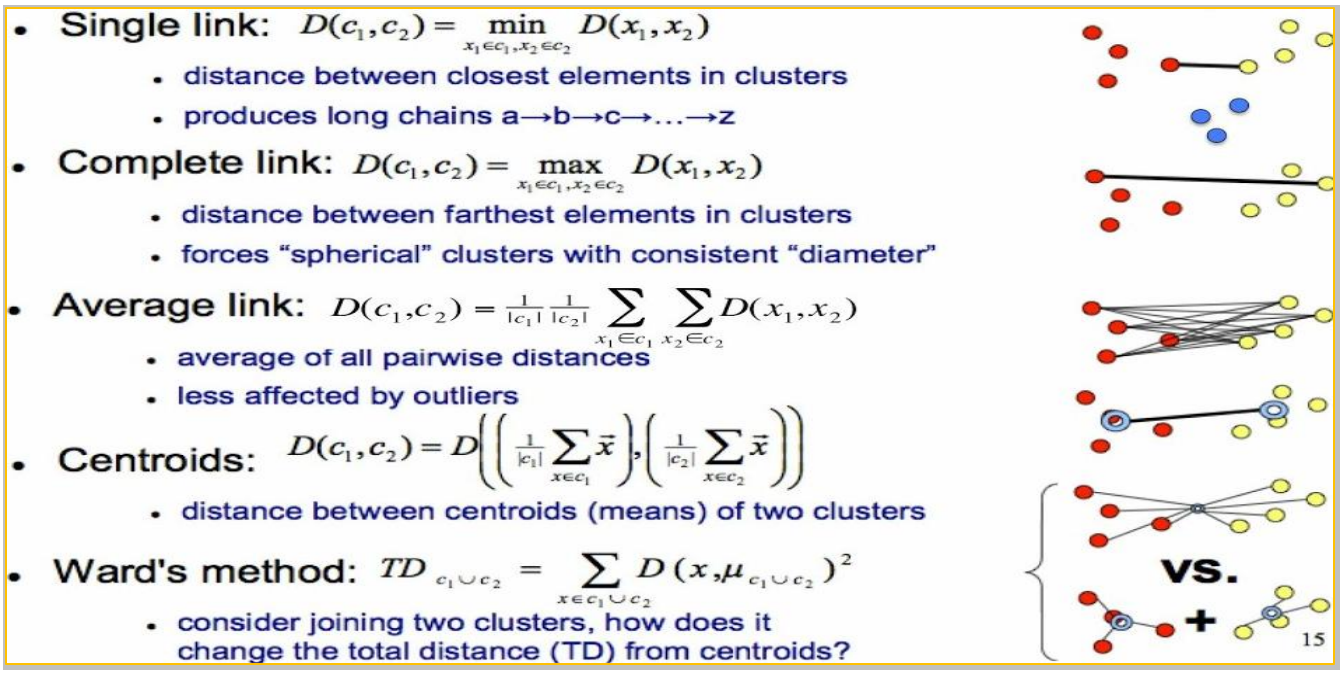

Figure-2

Figure-2 shows the Linkage methods formula with picture.

1.2 Divisive hierarchical clustering:

In divisive hierarchical clustering method clusters are formed by series of successive division. Thus initially single group of objects is divided into 2 subgroups such that object in one group is far away from object in other group. These subgroups are further divided into dissimilar subgroups. This procedure continuous until there are as many clusters as objects.

The results of both agglomerative and divisive clustering can be represented in the form of two dimensional diagram and is known as DENDROGRAM.

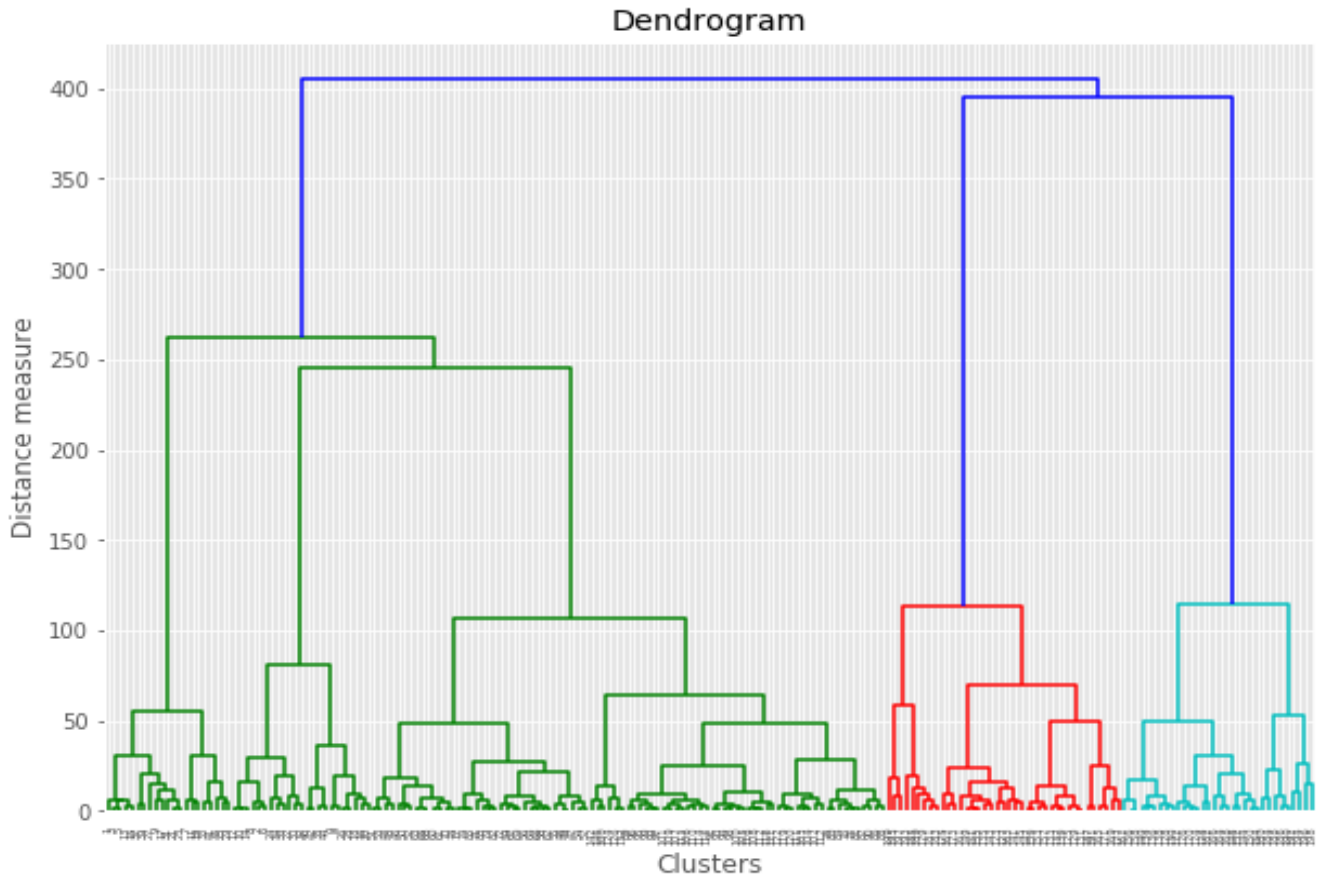

Figure-3

Figure-3 is an example of dendrogram.

2. Non-hierarchical clustering:

Non-hierarchical techniques are designed to group the items rather than variables into a collection of k clusters. The number of clusters may be either pre-specified or to be determined as a part of clustering procedure. Non-hierarchical clustering can be applied to large dataset than hierarchical clustering.

2.1 K-Means clustering:

Mac Queen suggested that the K-Means algorithm which assigns each observation to the cluster having nearest centroide(mean).

The algorithm has the following steps.

i. Partition the items into k initial clusters
ii. Proceed through list of items. Assign an item to the cluster whose centroid is nearest(distance usually calculated using Euclidean distance).
iii. Re-calculate the centroids for the cluster receiving the new item for the cluster loosing an item.
iv. Repeat step (ii) and (iii) until no more reassignments are takes place.

Note: K-Means clustering is depends on initially selected clusters.

Distance or similarity measures:

1. EUCLIDEAN DISTANCE:

Euclidean distance between two p-dimensional observation say,

$X^1 = (X_1, X_2, X_3, \dots X_p)$ and $Y^1 = (Y_1, Y_2, Y_3, \dots Y_p)$

Can be calculated as,

d (X,Y) = $\sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + \cdots + (X_p - Y_p)^2}$

$= \sqrt{(X - Y)^1 (X - Y)}$

2. MANHATTAN DISTANCE:

Manhattan distance between two p-dimensional observation say,

$X^1 = (X_1, X_2, X_3, \dots X_p)$ and $Y^1 = (Y_1, Y_2, Y_3, \dots Y_p)$

Can be calculated as,

d (X,Y) = $\sum_{i=1}^{p} |X_i - Y_i|$

This distance is less affected by outliers compared to Euclidean distance.

3. MINKOUSKI DISTANCE:

Minkouski distance between two p-dimensional observation say,

$X^1 = (X_1, X_2, X_3, \dots X_p)$ and $Y^1 = (Y_1, Y_2, Y_3, \dots Y_p)$

Can be calculated as,

d (X,Y) = $[\sum_{i=1}^{p}|X_i - Y_i|]^{1/m}$

Here if m=1, d(X,Y) measures the Manhattan distance

And if m=2, d(X,Y) measures the Euclidean distance

4. MAHANALOBIES DISTANCE:

Mahanalobies distance between two p-dimensional observation say,

$X^1 = (X_1, X_2, X_3, \ldots X_p)$ and $Y^1 = (Y_1, Y_2, Y_3, \ldots Y_p)$

Can be calculated as,

d (X,Y) = $=\sqrt{(X-Y)^1 S^{-1}(X-Y)}$

Where S is sample variance covariance matrix.

S = $\left(\frac{S_1 + S_2}{n_1 + n_2}\right)$

$$S_1 = \sum_{i=1}^{n_1}(X_i - \bar{X})\,(X_i - \bar{X})^1$$

$$S_2 = \sum_{i=1}^{n_2}(Y_i - \bar{Y})\,(Y_i - \bar{Y})^1$$

$$\bar{X} = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$$

$$\bar{Y} = \frac{\sum_{i=1}^{n_2} Y_i}{n_2}$$

However without prior knowledge distinct group, their sample quantities cannot be computed. For this reason Euclidean distance is often prefer in clustering.