VIT UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)
VELLORE ▪ CHENNAI
www.vit.ac.in

Data Visualisation
CSE613

Visualizing Cluster Analysis

- Data visualization helps to understand the multivariate data structures.
- Hierarchical Cluster analysis can be accomplished by dendrograms.
- The results of Partitioning Cluster can be visualized by projecting the data into the 2-D space or parallel coordinates.
- Cluster membership can be represented by
  - Different Colors and Glyphs.
  - Diving the clusters into panels of trellis display.
- silhouette plots provide a popular tool for diagnosing the quality of a partition.
- Self-Organizing features maps (SOM) results can be easily visualized.

- Most popular plots for Visualization of Cluster Analysis Results are
  - Dendrograms
  - Convex cluster hulls
  - silhouettes
- **Hierarchical Cluster Analysis:**
  - There are many methods in hierarchical clusters.
  - Hierarchical clustering is a **most intuitive** approach for grouping data.
  - Because it works similar way to how a human can k-group from N-objects.
  - The idea is to build a binary tree of the data that successively merges similar groups of points.
  - Visualizing this tree provides a useful summary of the data.
  - Each level of the resulting tree is a segmentation of the data.
  - The algorithm results in a sequence of groupings.

- **Divisive Hierarchical Clustering:**
  - It starts with Complete Dataset.
  - Divides the dataset into two groups.
  - Each of these groups are then recursively divided into two subgroups.
  - This process takes places until each point forms a cluster of its own.
- **Agglomerative Hierarchical clustering**
  - It starts with N single clusters. i.e N total number of points in dataset
  - A Hierarchy of cluster is created by repeatedly joining two closest clusters.
  - This process takes places until the complete data set forms one cluster.

- Both the cases, we need to measure the distances d(x,y) between the d-dimensional data points x and y and between groups of points.
- **Euclidean distance:**
- **Manhattan distance:**
- Distances between groups of points are used to join clusters $\rightarrow$ links $\rightarrow$ referred as link methods
- Different types of linking methods:
  - Single - link method:
    - it is also called as single-link clustering or connectedness or minimum method $\rightarrow$ $l(A,B) = min_{a \in A, b \in B} d(a, b)$
    - The distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster.
    - If the data consist of similarities, we consider the similarity between one cluster and another cluster to be equal to the greatest similarity from any member of one cluster to any member of the other cluster.

- Complete - link Clustering:
  - it is also called diameter or maximum method.
  - he distance between one cluster and another cluster to be equal to the longest distance from any member of one cluster to any member of the other cluster.
  - $l(A,B) = max_{a \in A, b \in B} d(a, b)$
- Average - link Clustering:
  - The distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.
  - $l(A,B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$
  - A variation on average-link clustering is the UCLUS method of D'Andrade (1978) which uses the median distance.

- Complete - link Clustering:
  - it is also called diameter or maximum method.
  - he distance between one cluster and another cluster to be equal to the longest distance from any member of one cluster to any member of the other cluster.
  - l(A,B) = $max_{a \in A, b \in B} d(a, b)$
- Average - link Clustering:
  - The distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.
  - l(A,B) = $\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$
  - A variation on average-link clustering is the UCLUS method of D'Andrade (1978) which uses the median distance.

- **Dendrograms:**
  - The results from hierarchical clustering are typically presented as a dendrogram.
  - it is a tree $\rightarrow$ one-cluster solution
  - The root of the tree represents the one -cluster for complete data set
  - The leaves of the tree represents the single data points
  - The heights of the branches correspond to the distances between the clusters.
  - The layout of the a dendrogram is not unique, because at each branching point the top and bottom branches could be exchanged.
  - **N** data points $\rightarrow$ **N-1** branching points are needed.
  - Total number of dendrograms will be drawn for exactly the same clustering is 2^{N-1}.