DATA MINING AND DATA WAREHOUSING

(20CS553)

By,
Anusha K S
Assistant Professor, Dept. of CSE
VVCE, Mysuru.

MODULE 4

Cluster Analysis, Overview, K-Means, Agglomerative Hierarchical Clustering. DBSCAN. Characteristics of data, clusters and clustering algorithms: Data Characteristics, Cluster Characteristics, General Characteristics of Clustering Algorithms. Which Clustering Algorithm?

SLT: Comparing K-means and DBSCAN

Textbook 2: Ch. 8.1-8.4, Ch. 9.1, 9.6

CLUSTER ANALYSIS

- Cluster Analysis groups data objects based only on information found in the data that describes the objects and their relationships

- The goal is that the objects within a group be similar to one another and different from the objects in other groups

- The greater the similarity within a groups and the greater the difference between groups the better or more distinct the clustering

- A clustering is a set of clusters

DIFFERENT TYPES OF CLUSTERING

- Partitional v/s Hierarchical
- Exclusive v/s overlapping v/s Fuzzy
- Complete v/s partial

DIFFERENT TYPES OF CLUSTERS

- Well separated
- Prototype – Based clusters (center-based)
- Graph – Based clusters (Contiguity –based)
- density – Based cluster
- Shared – property clusters (Conceptual)

K - MEANS

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- K-means clustering intends to partition “n” objects into k-clusters in which each object belongs to the cluster with the nearest mean
- Number of clusters, K, must be specified

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

BISECTING K-MEANS

The bisecting K-means algorithm is a straightforward extension of the basic K-means algorithm that is based on a simple idea: to obtain K clusters, split the set of all points into two clusters, select one of these clusters to split, and so on, until K clusters have been produced.

Algorithm 8.2 Bisecting K-means algorithm.

1: Initialize the list of clusters to contain the cluster consisting of all points.
2: **repeat**
3: Remove a cluster from the list of clusters.
4: {Perform several "trial" bisections of the chosen cluster.}
5: **for** $i = 1$ to *number of trials* **do**
6: Bisect the selected cluster using basic K-means.
7: **end for**
8: Select the two clusters from the bisection with the lowest total SSE.
9: Add these two clusters to the list of clusters.
10: **until** Until the list of clusters contains K clusters.

- Complexity of k-mean is **O(n * K * I * d)**

  n = number of points K = number of clusters
  I = number of iterations d = number of attributes

Strengths and Weaknesses of K-means (Limitations)

- K-means is simple and can be used for a wide variety of data types.
- It is also quite efficient, even though multiple runs are often performed.
- K-means is not suitable for all types of data.
- K-means has problems when clusters are of differing
- K-means has problems when the data contains outliers

PROBLEM

1. Consider the following data set **K = {2, 3, 4, 10, 11, 12, 20, 25, 30}** and number of clusters to be formed = 2 (i.e., **k=2**), random mean value **m1= 4 and m2=12.**