CST466
DATA MINING

MODULE-3

Module - 3 (Advanced classification and Cluster analysis)

Classification- Introduction, Decision tree construction principle, Splitting indices -Information Gain, Gini indexDecision tree construction algorithms-ID3, Decision tree construction with presorting-SLIQ, Classification Accuracy-Precision, Recall.

Introduction to clustering-Clustering Paradigms, Partitioning Algorithm- PAM, Hierarchical Clustering-DBSCAN, Categorical Clustering-ROCK

Introduction to Clustering:

- *Consider an electronics shop datastore.*
- *Suppose there are five managers working there.*
- *The customer relationship director of the shop may be interested in organizing all the company's customers into five groups so that each group can be assigned to a different manager.*
- *Grouping should be done in such a way that the customers in each group are as similar as possible.*
- *Moreover, two given customers having very different business patterns should not be placed in the same group.*
- *The intention behind this business strategy may be to develop customer relationship campaigns that specifically target each group, based on common features shared by the customers per group.*

What kind of data mining techniques help to accomplish this task?

- *Here, the class label (or group ID) of each customer is unknown.*
- *We have to discover these groupings.*
- *Given a large number of customers and many attributes describing customer profiles, it can be very costly or even infeasible to have a human study the data and manually come up with a way to partition the customers into strategic groups. An appropriate tool must be used.*

CLUSTERING:

- Clustering is the process of **grouping a set of data objects into multiple groups or clusters** so that **objects within a cluster have high similarity**, **but are very dissimilar to objects in other clusters.**
- Cluster analysis is the process of partitioning a set of data objects into subsets.
  - Each subset is a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters.
  - The set of clusters resulting from a cluster analysis can be referred to as the clusters.
- Cluster analysis has been widely used in many applications such as business intelligence, image pattern recognition,outlier detection, web search, biology, and security.

CLUSTERING PARADIGMS:

- There are many clustering approaches.
- It is difficult to provide a crisp categorization of clustering methods because these categories may overlap so that a method may have features from several categories.
- In general, the major fundamental clustering methods can be classified into the following categories,

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

Partitioning methods:

- **Given a set of *n* objects, a partitioning method constructs *k* partitions of the data, where each partition represents a cluster and *k* ≤ *n*.**
  - ie, it divides the data into *k* groups such that each group must contain at least one object.
- The basic partitioning methods typically adopt exclusive cluster separation.
  - ie, each object must belong to exactly one group.
  - This requirement may be relaxed, for eg, in fuzzy partitioning techniques.
- Most partitioning methods are **distance-based.**
- Given *k*(the number of partitions to construct), a partitioning method creates an initial partitioning.
- It then uses an **iterative relocation technique** that attempts to improve the partitioning by moving objects from one group to another.
- The general criterion of a good partitioning is that objects in the same cluster are “close” or related to each other, whereas objects in different clusters are “far apart” or very different.
- There are various kinds of other criteria for judging the quality of partitions.

- Achieving global optimality in partitioning-based clustering is often computationally prohibitive, potentially requiring an exhaustive enumeration of all the possible partitions.
- Instead, most applications adopt popular heuristic methods, such as greedy approaches like the **k-means** and the **k-medoids(PAM) algorithms**, which progressively improve the clustering quality and approach a local optimum.
- These heuristic clustering methods **works well for finding spherical-shaped clusters** in small- to medium-size databases.
- To find clusters with complex shapes and for very large data sets, partitioning-based methods need to be extended.

Hierarchical methods:

- A hierarchical method **creates a hierarchical decomposition** of the given set of data objects.
- Hierarchical clustering methods can be **distance-based** or **density- and continuity-based.**
- Based on how the hierarchical decomposition is formed, a hierarchical method can be classified as;
  - **Agglomerative approaches.**
  - **Divisive approaches.**

Agglomerative approach:

- Also called the bottom-up approach.
- Starts with each object forming a separate group.
- It successively merges the objects or groups close to one another, until all the groups are merged into one (the topmost level of the hierarchy), or a termination condition holds.

Divisive approach:

- Also called top-down approach.
- Starts with all the objects in the same cluster.
- In each successive iteration, a cluster is split into smaller clusters, until each object is in one cluster, or a termination condition holds.

Drawback:

- Once a step (merge or split) is done, it can never be undone.
- This rigidity is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of different choices.
- Cannot correct erroneous decisions.

Density-based methods:

- Most partitioning methods cluster objects based on the distance between objects.
- Such methods can find only spherical-shaped clusters and encounter difficulty in discovering clusters of arbitrary shapes.
- Other clustering methods have been developed **based on the notion of density.**
- Their general idea is to **continue growing a given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold.**
    - For example, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points.
    - Such a method can be used to filter out noise or outliers and discover clusters of arbitrary shape.
- Density-based methods can **divide a set of objects into multiple exclusive clusters**, or a hierarchy of clusters.
- Typically, density-based methods consider exclusive clusters only, and do not consider fuzzy clusters.
- Eg, of density based clustering: **DBSCAN**, OPTICS, DENCLUE

Grid-based methods:

- Grid-based methods quantize the object space into a finite number of cells that form a grid structure.
- All the clustering operations are performed on the grid structure (i.e., on the quantized space).
- The main advantage of this approach is its **fast processing time**, which is typically independent of the number of data objects and dependent only on the number of cells in each dimension in the quantized space.
- Using grids is often an efficient approach to many spatial data mining problems,including clustering.
- Therefore, grid-based methods can be integrated with other clustering methods such as density-based methods and hierarchical methods.

Note:

- *Most of the clustering algorithms integrate the ideas of several clustering methods, so that it is sometimes difficult to classify a given algorithm as uniquely belonging to only one clustering method category.*
- *Furthermore, some applications may have clustering criteria that require the integration of several clustering techniques.*

DISTANCE MEASURES IN CLUSTER ANALYSIS:

- In cluster analysis, the dissimilarity(or similarity) between the objects described by interval scaled variables is computed based on the distance between each pair of objects.
- Following are the common distance measures used;
  - **Euclidean distance**
  - **Manhattan distance**
  - **Minkowski distance**
- Consider two n-dimensional data objects $i = (x_{i1}, x_{i2}, \ldots, x_{in})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jn})$
- **Euclidean distance** between i and j is defined as;

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2},$$

- **Manhattan (or city block) distance**, defined as;

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|.$$

- **Minkowski distance** is a generalization of both Euclidean distance and Manhattan distance and is defined as;

$$d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p}$$

,where p is a positive integer.

Note:

- *Both the Euclidean distance and Manhattan distance satisfy the following mathematic requirements of a distance function:*

1. $d(i, j) \geq 0$: Distance is a nonnegative number.
2. $d(i, i) = 0$: The distance of an object to itself is 0.
3. $d(i, j) = d(j, i)$: Distance is a symmetric function.
4. $d(i, j) \leq d(i, h) + d(h, j)$: Going directly from object i to object j in space is no more than making a detour over any other object h (*triangular inequality*).

Sample Questions:

1. Let x1 = (1, 2) and x2 = (3, 5) represent two objects. Calculate the euclidean and manhattan distance.
2. Given 5 dimensional samples A (1,0,2,5,3) and B (2,1,0,3,-1). Find the euclidean, manhattan and minkowski distance (given p = 3, for minkowski distance).

Eg:

- Let x_1 = (1,2) and x_2=(3,5) represent two objects.
- Find the euclidean and manhattan distance between x_1 and x_2.

Solution:

- Euclidian distance = $((1-3)^2+(2-5)^2)^{1/2}$

  = 3.61

- Manhattan distance = $|1-3|+|2-5| = 5$

PARTITIONING METHODS

- Given **D**, a data set of ***n* objects**, and ***k***, the **number of clusters to form.**
- A partitioning algorithm organizes the objects into k partitions (k ≤ n), where each partition represents a cluster.
- The clusters are formed to optimize an objective partitioning criterion, such as a dissimilarity function based on distance, so that the objects within a cluster are "similar," whereas the objects of different clusters are "dissimilar" in terms of the data set attributes.
- The most well-known and commonly used partitioning methods are;
  - **k-means**
  - **k-medoids**

Variants of k-medoid
  - **PAM** (Partitioning around medoids)
  - CLARA (Clustering Large Applications)
  - CLARANS (Clustering Large Applications based upon Randomized Search)

PAM:

Algorithm: *k*-medoids. PAM, a *k*-medoids algorithm for partitioning based on medoid or central objects.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects in *D* as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, o_{random};
(5) compute the total cost, *S*, of swapping representative object, o_j, with o_{random};
(6) **if** $S < 0$ **then** swap o_j with o_{random} to form the new set of *k* representative objects;
(7) **until** no change;

- PAM was one of the first k-medoids algorithms introduced.
- It attempts **to determine *k* partitions for *n* objects**.
- Initially, we'll select *k* representative objects randomly.
- After that, the algorithm repeatedly tries to make a better choice of cluster representatives.
- All of the possible pairs of objects are analyzed, where one object in each pair is considered a representative object and the other is not.
- The quality of the resulting clustering is calculated for each such combination.
- An object, o_j, is replaced with the object causing the greatest reduction in error.
- The set of best objects for each cluster in one iteration forms the representative objects for the next iteration.
- The final set of representative objects are the respective medoids of the clusters.

Drawback of PAM method:

- For large values of n and k, k-medoids method becomes very costly.
  - ie, for small data sets, PAM works well.
  - But, PAM does not scale well for large data sets.
- For dealing with large data sets, a sampling-based method called CLARA (Clustering LARge Applications) can be used.
- To improve the scalability and quality of CLARA, another k-medoid algorithm called CLARANS (Clustering Large Applications based upon Randomized Search) is used.
- CLARANS combines sampling technique with PAM.

k-medoid (PAM)

Algorithm:

Given the value of k and unlabelled data:

1. Choose k number of random points from the data and assign these k points to k number of clusters. These are the initial medoids.
2. For all the remaining data points, calculate the distance from each medoid and assign it to the cluster with the nearest medoid.
3. Calculate the total cost (Sum of all the distances from all the data points to the medoids)
4. Select a random point as the new medoid and swap it with the previous medoid. Repeat 2 and 3 steps.
5. If the total cost of the new medoid is less than that of the previous medoid, make the new medoid permanent and repeat step 4.
6. If the total cost of the new medoid is greater than the cost of the previous medoid, undo the swap and repeat step 4.
7. The Repetitions have to continue until no change is encountered with new medoids to classify data points.

Operating steps:

- The initial representative objects (medoids) are chosen arbitrarily.
- The iterative process of replacing representative objects by non-representative objects continues as long as the quality of the resulting clustering is improved.
- This quality is estimated using a cost function that measures the average dissimilarity between an object and the representative object of its cluster.
    - To determine whether a non-representative object, o_{random}, is a good replacement for a current representative object, o_j, the following four cases are examined for each of the non-representative objects,p.

Case 1: p currently belongs to representative object, o_j. If o_j is replaced by o_{random} as a representative object and p is closest to one of the other representative objects, o_i, $i \neq j$, then p is reassigned to o_i.

Case 2: p currently belongs to representative object, o_j. If o_j is replaced by o_{random} as a representative object and p is closest to o_{random}, then p is reassigned to o_{random}.

Case 3: p currently belongs to representative object, o_i, $i \neq j$. If o_j is replaced by o_{random} as a representative object and p is still closest to o_i, then the assignment does not change.

Case 4: p currently belongs to representative object, o_i, $i \neq j$. If o_j is replaced by o_{random} as a representative object and p is closest to o_{random}, then p is reassigned to o_{random}.

- Each time a reassignment occurs, a difference in absolute error, E, is contributed to the cost function.
    - Therefore, the cost function calculates the difference in absolute-error value if a current representative object is replaced by a nonrepresentative object.
- The total cost of swapping is the sum of costs incurred by all non-representative objects.
- If the total cost is negative, then o_j is replaced or swapped with o_{random} since the actual absolute error E would be reduced.
- If the total cost is positive, the current representative object, o_j, is considered acceptable, and nothing is changed in the iteration.