Clustering

Clustering is the process of making a group of abstract objects into classes of similar objects.

Points

i. A cluster of data objects can be treated as one group.
ii. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
iii. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

Examples:

i. Given a collection of text, we need to organize them, according to the content similarities to create a topic hierarchy.
ii. Detecting distinct kinds of pattern in image data (Image processing). It's effective in biology research for identifying the underlying patterns.
iii. Identification of areas of similar land use in an earth observation database.
iv. Identifying groups of motor insurance policy holders with a high average claim cost.
v. Identifying groups of houses according to their house type, value, and geographical location Observed earth quake epicenters should be clustered along continent fault.

Applications of Cluster Analysis

i. Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
ii. Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
iii. In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
iv. Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
v. Clustering also helps in classifying documents on the web for information discovery.
vi. Clustering is also used in outlier detection applications such as detection of credit card fraud.
vii. As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Requirements of Clustering in Data Mining

i. **Scalability –** We need highly scalable clustering algorithms to deal with large databases.

ii. **Ability to deal with different kinds of attributes –** Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.

iii. **Discovery of clusters with attribute shape –** The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.

iv. **High dimensionality –** The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.

v. **Ability to deal with noisy data –** Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

vi. **Interpretability –** The clustering results should be interpretable, comprehensible, and usable.

Major Clustering Approaches / Methods

Clustering methods can be classified into the following categories −

1) **Partitioning Method**
2) **Hierarchical Method**
3) **Density-based Method**
4) **Grid-Based Method**
5) **Model-Based Method**
6) **Constraint-based Method**

1. Partitioning Method

Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements −

i. Each group contains at least one object.

ii. Each object must belong to exactly one group.

Points–

a) For a given number of partitions (say k), the partitioning method will create an initial partitioning.

b) Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

2. Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

1) Agglomerative Approach

2) Divisive Approach

1. Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

2. Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering

Here are the two approaches that are used to improve the quality of hierarchical clustering −

i. Perform careful analysis of object linkages at each hierarchical partitioning.

ii. Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

3. Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

4. Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

Advantages

i. The major advantage of this method is fast processing time.

ii. It is dependent only on the number of cells in each dimension in the quantized space.

5 .Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

6 .Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.