Data Mining - Cluster Analysis

What is Cluster?

Cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster.

What is Clustering?

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

Points to Remember

- A cluster of data objects can be treated as a one group.
- While doing the cluster analysis, we first partition the set of data into groups based on data similarity and then assign the label to the groups.
- The main advantage of Clustering over classification is that, It is adaptable to changes and help single out useful features that distinguished different groups.

Applications of Cluster Analysis

- Clustering Analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer basis. And they can characterize their customer groups based on purchasing patterns.
- In field of biology it can be used to derive plant and animal taxonomies, categorize genes with similar functionality and gain insight into structures inherent in populations.
- 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 house type, value, geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function Cluster Analysis serve as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

Requirements of Clustering in Data Mining

Here is the typical requirements of clustering in data mining:

- **Scalability** - We need highly scalable clustering algorithms to deal with large databases.
- **Ability to deal with different kind of attributes** - Algorithms should be capable to be applied on any kind of data such as interval based (numerical) data, categorical, binary data.
- **Discovery of clusters with attribute shape** - The clustering algorithm should be capable of detect cluster of arbitrary shape. The should not be bounded to only distance measures that tend to find spherical cluster of small size.

- **High dimensionality** - The clustering algorithm should not only be able to handle low- dimensional data but also the high dimensional space.
- **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.
- **Interpretability** - The clustering results should be interpretable, comprehensible and usable.

Clustering Methods

The clustering methods can be classified into following categories:

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

PARTITIONING METHOD

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

- Each group contain at least one object.
- Each object must belong to exactly one group.

Points to remember:

- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

HIERARCHICAL METHODS

This method create the hierarchical decomposition of the given set of data objects. We can classify Hierarchical method on basis of how the hierarchical decomposition is formed as follows:

- Agglomerative Approach
- Divisive Approach

AGGLOMERATIVE APPROACH

This approach is also known as 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.

DIVISIVE APPROACH

This approach is also known as 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.

Disadvantage

This method is rigid i.e. once merge or split is done, It can never be undone.

APPROACHES TO IMPROVE QUALITY OF HIERARCHICAL CLUSTERING

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

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into microclusters, and then performing macroclustering on the microclusters.

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 neighbourhood 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.

GRID-BASED METHOD

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

Advantage

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

MODEL-BASED METHODS

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

This method also serve a way of automatically determining number of clusters based on standard statistics , taking outlier or noise into account. It therefore yield robust clustering methods.

CONSTRAINT-BASED METHOD

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

Source:

http://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm