Data Mining

5. Cluster Analysis

5.1 Introduction

Fall 2009

Instructor: Dr. Masoud Yaghini

Outline

- What is Cluster Analysis?
- Active Themes of Research
- Measure the Quality of Clustering
- A Categorization of Major Clustering Methods
- References

What is Cluster Analysis?

What is Cluster Analysis?

- Cluster: a collection of data objects
  - Similar to one another within the same cluster
  - Dissimilar to the objects in other clusters
- Clustering / Cluster analysis
  - the process of grouping a set of objects into classes or clusters
  - the objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters.
  - also called data segmentation

Clustering

- Dissimilarities are assessed based on the **attribute values** describing the objects.
- Clustering has its roots in many areas, including
  - **Machine learning**: clustering is an example of **unsupervised learning** (no predefined classes).
  - **Data mining**: efforts have focused on finding methods for **efficient** and **effective** cluster analysis in large databases.
  - **Statistics**: focusing mainly on *distance-based cluster analysis.*

Typical applications

- Typical applications
  - **As a stand-alone tool**
    - to get insight into data distribution
    - to observe the characteristics of each cluster
    - to focus on a particular set of clusters for further analysis
  - **As a preprocessing step for other algorithms**
    - e.g. preprocessing for classification

Clustering: Rich Applications

- Pattern Recognition
- Spatial Data Analysis
  - Create thematic maps in GIS by clustering feature spaces
  - Detect spatial clusters or for other spatial mining tasks
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

Clustering: Rich Applications

- **Marketing:**
  - Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Land use:**
  - Identification of areas of similar land use in an earth observation database
- **Insurance:**
  - Identifying groups of motor insurance policy holders with a high average claim cost
- **City-planning:**
  - Identifying groups of houses according to their house type, value, and geographical location

Clustering: Rich Applications

- **Biology:**
  - It can be used to derive plant and animal taxonomies, categorize genes with similar functionality, and gain insight into structures inherent in populations.
- **World Wide Web (WWW):**
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns
- **Pattern Recognition**
- **Image Processing**

Active Themes of Research

Active Themes of Research

- Scalability
  - Many clustering algorithms work well only on small data sets
  - Clustering algorithms that can work on large data sets are needed.
- Ability to deal with different types of attributes
  - Many algorithms are designed to cluster interval-based (**numerical**) data.
  - Clustering algorithms that can work on other types of data, such as **binary**, **categorical** (**nominal**), and **ordinal data**, or **mixtures** of these data types are needed.

Active Themes of Research

- Minimal requirements for domain knowledge to determine input parameters
  - Many clustering algorithms require users to input certain parameters in cluster analysis (such as the number of desired clusters). The clustering results can be quite sensitive to input parameters.
  - Clustering algorithms with minimal input parameters are needed.
- Able to deal with noise and outliers
  - Some clustering algorithms are sensitive to data contain outliers or missing, unknown, or erroneous data. and may lead to clusters of poor quality.

Active Themes of Research

- Incremental clustering
  - Some clustering algorithms cannot incorporate newly inserted data (i.e., database updates) into existing clustering structures and, instead, must determine a new clustering from scratch.
  - It is important to develop incremental clustering algorithms
- Insensitive to order of input records
  - Some clustering algorithms are sensitive to the order of input data. That is, given a set of data objects, such an algorithm may return dramatically different clustering depending on the order of presentation of the input objects.
  - It is important to develop algorithms that are insensitive to the order of input.

Active Themes of Research

- High dimensionality
  - Many clustering algorithms are good at handling low-dimensional data, involving only two to three attributes.
  - Finding clusters of data objects in high-dimensional space is challenging
- Constraint-based clustering
  - Suppose that your job is to choose the locations for a given number of new automatic banking machines (ATMs) in a city. you may cluster households while considering constraints such as the city's rivers and highway networks, and the type and number of customers per cluster.
  - A challenging task is to find groups of data with good clustering behavior that satisfy specified constraints.

Measure the Quality of Clustering

Quality: What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering result depends on both the similarity measure used by the method and its implementation
- The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

Measure the Quality of Clustering

- Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, typically metric: d(i, j)
- There is a separate "quality" function that measures the "goodness" of a cluster.
- The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables.
- Weights should be associated with different variables based on applications and data.
- It is hard to define "similar enough" or "good enough"
  - the answer is typically highly subjective.

A Categorization of Major Clustering Methods

A Categorization of Major Clustering Methods

- The major clustering methods can be classified into the following categories:
  - Partitioning methods
  - Hierarchical methods
  - Density-based methods
- Some 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.

Partitioning Approach

- Given a database of n objects, a partitioning method constructs k partitions of the data, where each partition represents a cluster and $k \leq n$.
- It satisfis the following requirements:
  - (1) each group must contain at least one object, and
  - (2) each object must belong to exactly one group.
- Notice that the second requirement can be relaxed in some fuzzy partitioning techniques.
- The general criterion of a good partitioning is that objects in the same cluster are “close” or related to each other, whereas objects of different clusters are “far apart” or very different.

Partitioning Approach

- To achieve global optimality in partitioning-based clustering would require the exhaustive enumeration of all of the possible partitions.
- Popular heuristic methods:
  - (1) **k-means algorithm**: where each cluster is represented by the mean value of the objects in the cluster, and
  - (2) **k-medoids algorithm**: where each cluster is represented by one of the objects located near the center of the cluster.
- These heuristic clustering methods work well for finding spherical-shaped clusters in small to medium-sized databases.
- To find clusters with complex shapes and for clustering very large data sets, partitioning-based methods need to be extended.

Hierarchical Methods

- Create a hierarchical decomposition of the set of objects
- A hierarchical method can be classified as:
  - Agglomerative (bottom-up) approach
  - Divisive (top-down) approach
- Typical methods: Diana, Agnes, BIRCH, ROCK, CAMELEON

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 at discovering clusters of arbitrary shapes.
- Density-based methods continue growing the given cluster as long as the density (number of objects or data points) in the "neighborhood" exceeds some threshold
  - For each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points.

Density-based Methods

- Density-based approach are based on connectivity and density functions
- Typical methods: DBSACN, OPTICS, DenClue

The Choice of Clustering Algorithm

- The choice of clustering algorithm depends both on
  - the type of data available and
  - the particular purpose of the application.
- If cluster analysis is used as a descriptive or exploratory tool, it is possible to try several algorithms on the same data to see what the data may disclose.

References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 7)

The end