Chapter 7. Cluster Analysis

1. What is Cluster Analysis?
2. Types of Data in Cluster Analysis
3. A Categorization of Major Clustering Methods
4. Partitioning Methods
5. Hierarchical Methods
6. Density-Based Methods
7. Clustering High-Dimensional Data
8. Constraint-Based Clustering
9. Outlier Analysis
10. Summary

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
- Cluster analysis
  - Finding similarities between data according to the characteristics found in the data
  - Grouping similar data objects into clusters

What is Cluster Analysis?

- Unsupervised learning: no predefined classes
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms

Clustering: Rich Applications and Multidisciplinary Efforts

- Spatial Data Analysis
  - Detect spatial clusters or for other spatial mining tasks
- Economic Science (especially market research)
  - Identify customers whose behaviors are similar
- WWW
  - Cluster documents
  - Cluster Weblog data to discover groups of similar access patterns
- Image Processing & Pattern Recognition

Examples of Clustering Applications

- <u>Marketing:</u>
  - Help marketers discover distinct groups in their customer bases
  - Use this knowledge to develop targeted marketing programs
- <u>Land use:</u>
  - Identification of areas of similar land use in an earth observation database

Examples of Clustering Applications

- 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

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
  - 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 semantics
- Hard to define "similar enough" or "good enough"
  - The answer is typically highly subjective

Requirements of Clustering in Data Mining

- Ability to deal with different types of attributes
- Ability to handle dynamic data
- Discovery of clusters with an arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noises and outliers
- Insensitive to the order of input records
- High dimensionality
- Scalability
- Incorporation of user-specified constraints

Chapter 7. Cluster Analysis

1. What is Cluster Analysis?
2. Types of Data in Cluster Analysis
3. A Categorization of Major Clustering Methods
4. Partitioning Methods
5. Hierarchical Methods
6. Density-Based Methods
7. Clustering High-Dimensional Data
8. Constraint-Based Clustering
9. Outlier Analysis
10. Summary