Cluster Analysis

Part 4: Outlier Analysis

Hui Yang
Department of Computer Science
San Francisco State University
http://cose-stor.sfsu.edu/~huiyang
Cluster Analysis: Main Topics

- What is Cluster Analysis?
- Distance and Data Types
- A Categorization of Major Clustering Methods
  - Partitioning methods
  - Hierarchical methods
  - Density-Based methods
- Outlier analysis
- Summary
What Is Outlier Discovery?

- **Definition of outliers**
  - The set of objects are considerably dissimilar from the remainder of the data
  - Example: Sports: Michael Jordon, Ming Yao, ...

- **Outlier discovery**: Define and find outliers in large data sets

- **Applications**:
  - Credit card fraud detection
  - Telecom fraud detection
  - Customer segmentation
  - Medical analysis
Outlier Discovery: Statistical Approaches

**Assume a model underlying distribution that generates data set (e.g. normal distribution)**
- Use discordancy tests depending on
  - data distribution
  - distribution parameter (e.g., mean, variance)
  - number of expected outliers

**Drawbacks**
- most tests are for single attribute
- In many cases, data distribution may not be known
Outlier Discovery: Distance-Based Approach

- Overcome the main limitations exhibited by statistical methods
  - multi-dimensional analysis without knowing data distribution

- Distance-based outlier:
  - A $DB(p, d)$-outlier is an object $O$ in a dataset $T$ such that at least a fraction $p$ of the objects in $T$ lies at a distance greater than $d$ from $O$

- Algorithms for mining distance-based outliers
  - Index-based algorithm
  - Nested-loop algorithm
  - Cell-based algorithm
Density-Based Local Outlier Detection

- Main limitation of distance-based outlier detection
  - It is based on global distance distribution
  - It encounters difficulties to identify outliers if data is not uniformly distributed

- Example:
  - \( C_1 \) contains 400 loosely distributed points, \( C_2 \) has 100 tightly condensed points, 2 outlier points \( o_1, o_2 \)
  - Distance-based method cannot identify \( o_2 \) as an outlier

- Solution: introduce the concept of local outliers

- Local outlier factor (LOF)
  - Assume outlier is not crisp
  - Each point has a LOF
Outlier Discovery: Deviation-Based Approach

- Identifies outliers by examining the main characteristics of objects in a group
  - Objects that “deviate” from this description are considered outliers

**Techniques**

- Sequential exception technique
  - simulates the way in which humans can distinguish unusual objects from among a series of supposedly like objects
- OLAP data cube technique
  - uses data cubes to identify regions of anomalies in large multidimensional data
Summary

- Cluster analysis groups objects based on their similarity and has wide applications.
- Measure of similarity can be computed for various types of data.
- Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods.
- Outlier detection and analysis are very useful for fraud detection, etc. and can be performed by statistical, distance-based or deviation-based approaches.
- There are still many research issues in cluster analysis.
Problems and Challenges

- **Considerable progress** has been made in scalable clustering methods
  - Partitioning: k-means, k-medoids, CLARANS
  - Hierarchical: BIRCH, ROCK, CHAMELEON
  - Density-based: DBSCAN, OPTICS, DenClue
  - Grid-based: STING, WaveCluster, CLIQUE
  - Model-based: EM, Cobweb, SOM
  - Frequent pattern-based: pCluster
  - Constraint-based: COD, constrained-clustering

- Current clustering techniques do not **address** all the requirements adequately, still an active area of research
Requirements of Clustering in Data Mining

- Scalability
- Ability to deal with different types of attributes
- Ability to handle dynamic data
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality
- Incorporation of user-specified constraints
- Interpretability and usability
R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD'98


Beil F., Ester M., Xu X.: "Frequent Term-Based Text Clustering", KDD'02


D. Gibson, J. Kleinberg, and P. Raghavan. Clustering categorical data: An approach based on dynamic systems. VLDB’98.
References (2)

- V. Ganti, J. Gehrke, R. Ramakrishnan. CACTUS Clustering Categorical Data Using Summaries. KDD'99.
- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94.
L. Parsons, E. Haque and H. Liu, *Subspace Clustering for High Dimensional Data: A Review*, SIGKDD Explorations, 6(1), June 2004

E. Schikuta. Grid clustering: An efficient hierarchical clustering method for very large data sets. Proc. 1996 Int. Conf. on Pattern Recognition,

G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A multi-resolution clustering approach for very large spatial databases. VLDB’98.

A. K. H. Tung, J. Han, L. V. S. Lakshmanan, and R. T. Ng. *Constraint-Based Clustering in Large Databases*, ICDT’01.

A. K. H. Tung, J. Hou, and J. Han. *Spatial Clustering in the Presence of Obstacles*, ICDE’01


T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. SIGMOD’96.