Cluster Analysis

Part 3: Hierarchical and Density-based clustering

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
Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters $k$ as an input, but needs a termination condition.
- Two types: agglomerative and divisive.
AGNES (Agglomerative Nesting)

- Introduced by Kaufmann and Rousseeuw (1990)
- Use the single-link distance measurement
- Merge nodes that have the most similarity
- Iterate in a non-descending fashion
- Eventually *all nodes belong to the same cluster (?)*
**Dendrogram**: shows how the clusters are merged.

**Dendrogram**: a tree structure that is commonly used to represent the process of hierarchical clustering.
DIANA (Divisive Analysis)

- Introduced by Kaufmann and Rousseeuw (1990)
- Inverse order of AGNES
- Eventually each node forms a cluster on its own
Recent Hierarchical Clustering Methods

- Major weakness of agglomerative clustering methods
  - do not scale well: time complexity of at least $O(n^2)$, where $n$ is the number of total objects
  - can never undo what was done previously

- Integration of hierarchical with distance-based clustering
  - **BIRCH (1996)**: uses CF-tree and incrementally adjusts the quality of sub-clusters
  - **ROCK (1999)**: clustering categorical data by neighbor and link analysis
  - **CHAMELEON (1999)**: hierarchical clustering using dynamic modeling
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
Density-Based Clustering Methods

- Clustering based on density distribution of data points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One dataset scan
  - Need density parameters as termination condition

- Several interesting studies:
  - **DBSCAN**: Ester, et al. (KDD’96)
  - **DENCLUE**: Hinneburg & D. Keim (KDD’98)
  - **CLIQUE**: Agrawal, et al. (SIGMOD’98) (more grid-based)
Density-Based Clustering: Basic Concepts

- Two parameters:
  - **Eps**: Maximum radius of the neighbourhood
  - **MinPts**: Minimum number of points in an Eps-neighbourhood of that point

- \( N_{Eps}(p) \): \( \{q \text{ belongs to } D \mid \text{dist}(p,q) \leq Eps\} \)

- Directly density-reachable: A point \( p \) is directly density-reachable from a point \( q \) w.r.t. \( Eps, MinPts \) if
  - \( p \) belongs to \( N_{Eps}(q) \)
  - core point condition:
    \[ |N_{Eps}(q)| \geq MinPts \]

MinPts = 5
Eps = 1 cm
Density-Reachable and Density-Connected

- Density-reachable:
  - A point $p$ is density-reachable from a point $q$ w.r.t. $Eps$, $MinPts$ if there is a chain of points $p_1, ..., p_n$, $p_1 = q$, $p_n = p$ such that $p_{i+1}$ is directly density-reachable from $p_i$

- Density-connected
  - A point $p$ is density-connected to a point $q$ w.r.t. $Eps$, $MinPts$ if there is a point $o$ such that both, $p$ and $q$ are density-reachable from $o$ w.r.t. $Eps$ and $MinPts$
DBSCAN: Density Based Spatial Clustering of Applications with Noise

- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points.
- Discovers clusters of arbitrary shape in spatial databases with noise.

![Diagram showing DBSCAN concepts: Core points, border points, and outliers. Eps = 1cm, MinPts = 5.]
DBSCAN: The Algorithm

1. Arbitrary select a point \( p \)
2. Retrieve all points density-reachable from \( p \) w.r.t. \( Eps \) and \( MinPts \).
3. If \( p \) is a core point, a cluster is formed.
4. If \( p \) is a border point, no points are density-reachable from \( p \) and DBSCAN visits the next point of the database.
5. Continue the process until all of the points have been processed.
DBSCAN: Sensitive to Parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.
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
References (1)

- 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)

- R. Ng and J. Han. Efficient and effective clustering method for spatial data mining. VLDB'94.
References (3)

- 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.