A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu

KDD 1996 (pp: 226-231)

Presented by: Yi-Hung Wu
Date: 2002/8/21

Preliminaries

- **Statistic and Machine Learning Approaches**
  - Self-organized map, Neural gas, etc.

- **Model-based Approaches**
  - Partitioning, Hierarchical, Density-based

- **Approaches to Improving the Efficiency**
  - Grid-based, Multidimensional indexing

- **Approaches to Improving the Effectiveness**
  - Subspace projection, Outlier analysis, Constraint-based, Categorical data clustering
Problems (1/2)

• **Requirements for Clustering Algorithms**
  – Minimal priori-knowledge to determine parameters
  – Discovery of clusters with arbitrary shapes
  – Good efficiency on large databases

• **Partitioning Methods**
  – K-means, K-medoids, CLARANS, etc.
  – Iteration relocation
    • Find k representatives
    • Get the voronoi diagram/cells

<table>
<thead>
<tr>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Local optima</td>
</tr>
<tr>
<td>2. The number of clusters</td>
</tr>
<tr>
<td>3. Only convex clusters</td>
</tr>
</tbody>
</table>

Problems (2/2)

• **Hierarchical Methods**
  – Top-down: divisive (split)
  – Bottom-up: agglomerative (merge)

• **Density-based Methods**
  – The previous work
    • Partition data set into cells
    • Get the histogram based on the cells
    • Identify cluster centers and boundaries
  – DBSCAN, DBCLASD, DENCLUE, OPICS, etc.

| 1. Space & Run-time |
| 2. Cell size |
Solutions (1/3)

- **Motivation**
  - Use density to distinguish clusters from noises

- **Key Idea**
  - For each point of a cluster, the neighborhood of a given radius (Eps) has to contain at least a minimum number of points (MinPts)
    - $\exists k$ nearest neighbors in its $\varepsilon$-neighborhood

Solutions (2/3)

- **Definitions**
  - Cluster $C (k, \varepsilon)$
    - $\forall p, q: p \in C \text{ if } q \in C$
    - $\forall p, q \in C: p$ is density-reachable from $q$
  - Noise: $\{ p \in D | \forall i: p \notin C_i \}$

- **DBSCAN Algorithm**
  - Criteria
    - Every point belongs to at most one cluster
    - Two core points belong to one cluster if they are density connected
    - The remaining border points are noises
Solutions (3/3)

- **Parameter Determination**
  - K-dist: the distance from the k-th nearest neighbor
  - Sorted k-dist graph
  - Threshold point
    - The maximal k-dist value ($\epsilon$) in the thinnest cluster
  - Interactive approach
    - The percentage of noise
    - The first valley of the graph

Experimental Results

- **Performance Evaluation**
  - Accuracy
    - DBSCAN vs. CLARANS
  - Efficiency
    - SQUOIA 2000 benchmark
Conclusion Remarks

• Contribution
  – A density-based notion of clusters
  – Discover clusters of arbitrary shape
  – Only one parameter (k) is required
• Advantages of Density-based Methods
  – Identify unusual data objects (noise)
    • Distance-based outlier analysis: DB(p,D)-outlier, D^k_n outlier
    • Density-based outlier analysis: local outlier, top-n outlier
  – Generate natural clusters (arbitrary shape)

Paper Scoring

• Scores {bad, marginal, good, excellent}
  – Originality: excellent
  – Technical Depth: good
  – Impact/Practicability: excellent
  – Readability: good
  – Overall: good