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

P.2

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

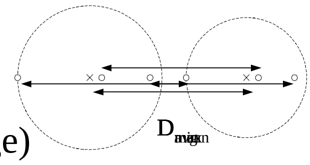
<u>Drawbacks</u>

1. Local optima
2. The number of clusters
3. Only convex clusters

P.3

Problems (2/2)

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



1. Termination condition
2. Not scale well

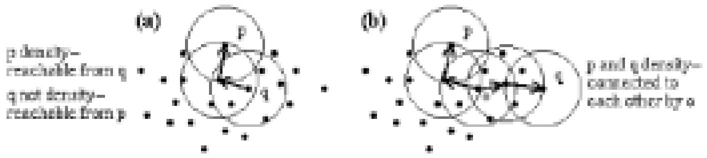
- **Density-based Methods**
  - – The previous work
    - Partition data set into cells
    - Get the histogram based on the cells
    - Identify cluster centers and boundaries
  - – DBSCAN, DBCLASD, DENCLUE, OPICS, etc.

1. Space & Run-time
2. Cell size

P.4

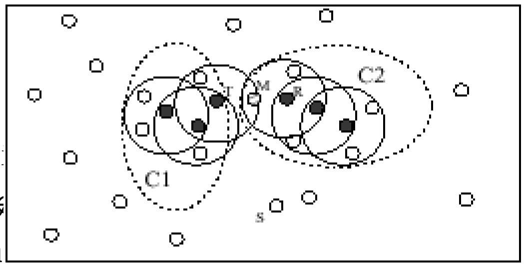
Solutions (1/3)

- **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)
    - ∃ k nearest neighbors in its ε-neighborhood



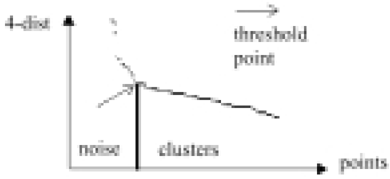
Solutions (2/3)

- **Definitions**
  - Cluster C (k,ε)
    - ∀p, q: p∈C if q∈C
    - ∀p, q∈C: p is dens
  - Noise: {p∈D|∀i: p
- **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 (ε) in the thinnest cluster
  - Interactive approach
    - The percentage of noise
    - The first valley of the graph



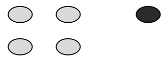
Experimental Results

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

number of points	1252	2503	3910	5213	6256
DBSCAN	3.1	6.7	11.3	16.0	17.8
CLAR-ANS	758	3026	6845	11745	18029
number of points	7820	8937	10426	12512	
DBSCAN	24.5	28.2	32.7	41.7	
CLAR-ANS	29826	39265	60540	80638	

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