Data Mining:

Concepts and Techniques

(3rd ed.)

Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign &

Simon Fraser University

Cluster Analysis: Basic Concepts and Methods

- Cluster Analysis: Basic Concepts

- Partitioning Methods

- Density-Based Methods

What is Cluster Analysis?

- Cluster: A collection of data objects
    - similar (or related) to one another within the same group
    - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis (or *clustering*, *data segmentation, …*)
    - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters
- Unsupervised learning: no predefined classes (i.e., *learning by observations* vs. learning by examples: supervised)
- Typical applications
    - As a stand-alone tool to get insight into data distribution
    - As a preprocessing step for other algorithms

Clustering for Data Understanding and Applications

- Costumer reviews clustering (hotel, movies, books, …)

- Information retrieval: document clustering

- Land use: Identification of areas of similar land use in an earth observation database

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs

- City-planning: Identifying groups of houses according to their house type, value, and geographical location

- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults

- Climate: understanding earth climate, find patterns of atmospheric and ocean

- Economic Science: market research

news.google.com

For you

Following

Saved searches

COVID-19

U.S.

World

Business

Technology

Entertainment

Sports

Science

Health

Headlines

More Headlines

> COVID-19 news: See the latest coverage of the coronavirus (COVID-19) ›

The US enters crucial weeks after a grim milestone of most deaths reported in one day

CNN · 47 minutes ago

- Coronavirus Task Force Briefing Live Stream Saturday: How to Watch White House's Latest COVID-19 Announcements

  Newsweek · 12 hours ago

- Trump puzzles allies by hedging on China criticism amid coronavirus outbreak

  Washington Examiner · 4 hours ago

- How Trump's attempts to win the daily news cycle feed a chaotic coronavirus response

  The Washington Post · Yesterday

- White House monitors big cities for signs of coronavirus outbreaks

  ▶ CBS This Morning · Yesterday

🔲 View Full Coverage ∧

Clustering as a Preprocessing Tool (Utility)

- Summarization:
    - Preprocessing for regression, PCA, classification, and association analysis
- Compression:
    - Image processing: vector quantization
- Finding K-nearest Neighbors
    - Localizing search to one or a small number of clusters
- Outlier detection
    - Outliers are often viewed as those "far away" from any cluster

Quality: What Is Good Clustering?

- A <u>good clustering</u> method will produce high quality clusters

  - high <u>intra-class</u> similarity: cohesive within clusters

  - low <u>inter-class</u> similarity: distinctive between clusters

- The <u>quality</u> of a clustering method depends on

  - the similarity measure used by the method

  - its implementation, and

  - Its ability to discover some or all of the <u>hidden</u> patterns

Measure the Quality of Clustering

- **Dissimilarity/Similarity metric**
    - Similarity is expressed in terms of a distance function, typically metric: $d(i, j)$
    - The definitions of distance functions are usually rather different for interval-scaled, boolean, categorical, ordinal ratio, and vector variables
    - Weights should be associated with different variables based on applications and data semantics
- Quality of clustering:
    - There is usually a separate "quality" function that measures the "goodness" of a cluster.
    - It is hard to define "similar enough" or "good enough"
        - The answer is typically highly subjective

Considerations for Cluster Analysis

- Partitioning criteria

  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)

- Separation of clusters

  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)

- Similarity measure

  - Distance-based (e.g., Euclidian, road network, vector) vs. connectivity-based (e.g., density or contiguity)

- Clustering space

  - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

Requirements and Challenges

- Scalability
    - Clustering all the data instead of only on samples
- Ability to deal with different types of attributes
    - Numerical, binary, categorical, ordinal, linked, and mixture of these
- Constraint-based clustering
    - User may give inputs on constraints
    - Use domain knowledge to determine input parameters
- Interpretability and usability
- Others
    - Discovery of clusters with arbitrary shape
    - Ability to deal with noisy data
    - Incremental clustering and insensitivity to input order
    - High dimensionality

Major Clustering Approaches (I)

- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-means, k-medoids
- Density-based approach:
  - Based on connectivity and density functions
  - Typical methods: DBSACN, OPTICS, DenClue
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Diana, Agnes, BIRCH, CAMELEON
- Grid-based approach:
  - based on a multiple-level granularity structure
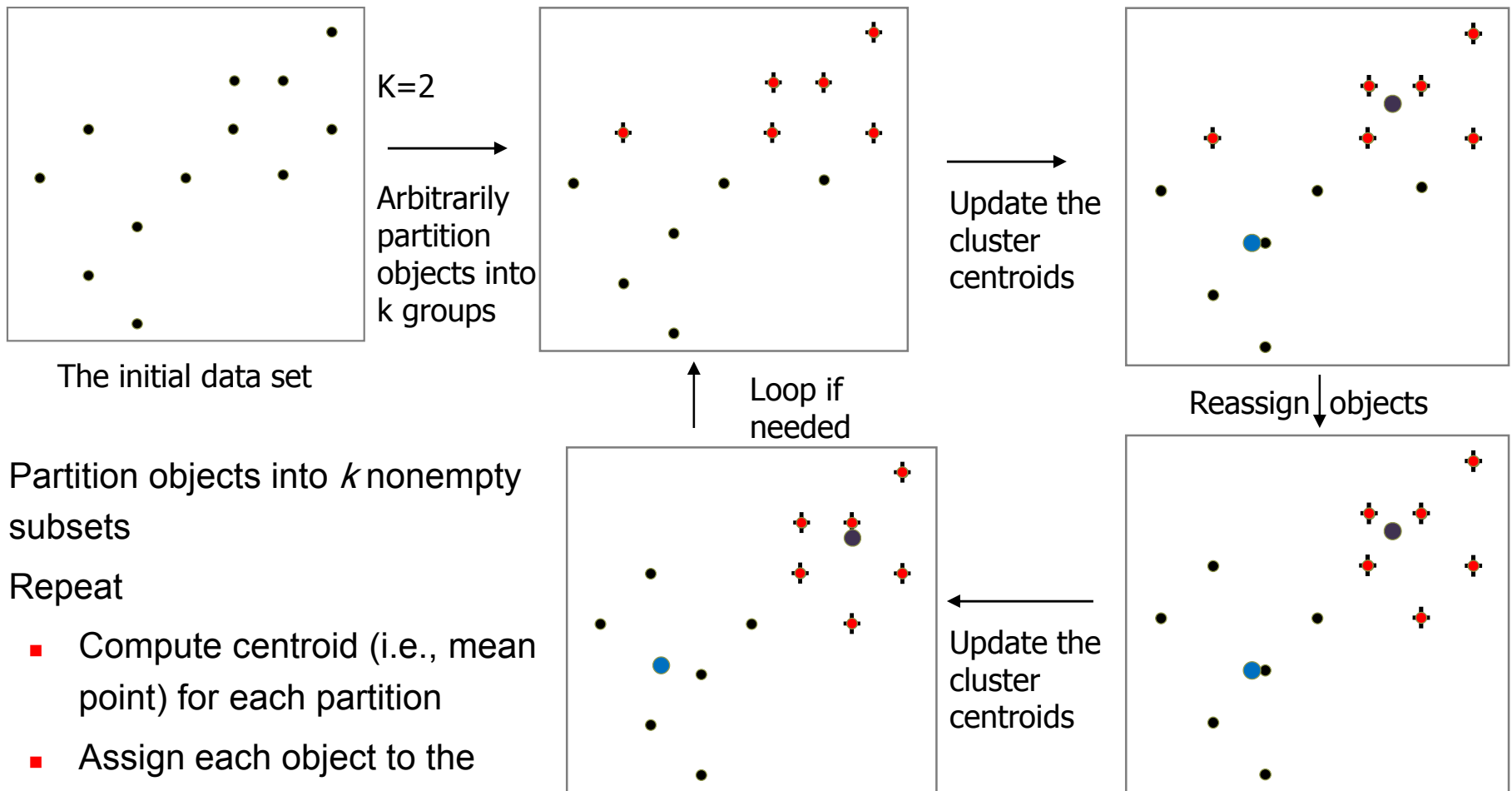 - Typical methods: STING, WaveCluster, CLIQUE

Major Clustering Approaches (II)

- **Model-based**:
  - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
  - Typical methods: EM, SOM, COBWEB
- **Frequent pattern-based**:
  - Based on the analysis of frequent patterns
  - Typical methods: p-Cluster
- **User-guided or constraint-based**:
  - Clustering by considering user-specified or application-specific constraints
  - Typical methods: COD (obstacles), constrained clustering
- **Link-based clustering**:
  - Objects are often linked together in various ways
  - Massive links can be used to cluster objects: SimRank, LinkClus

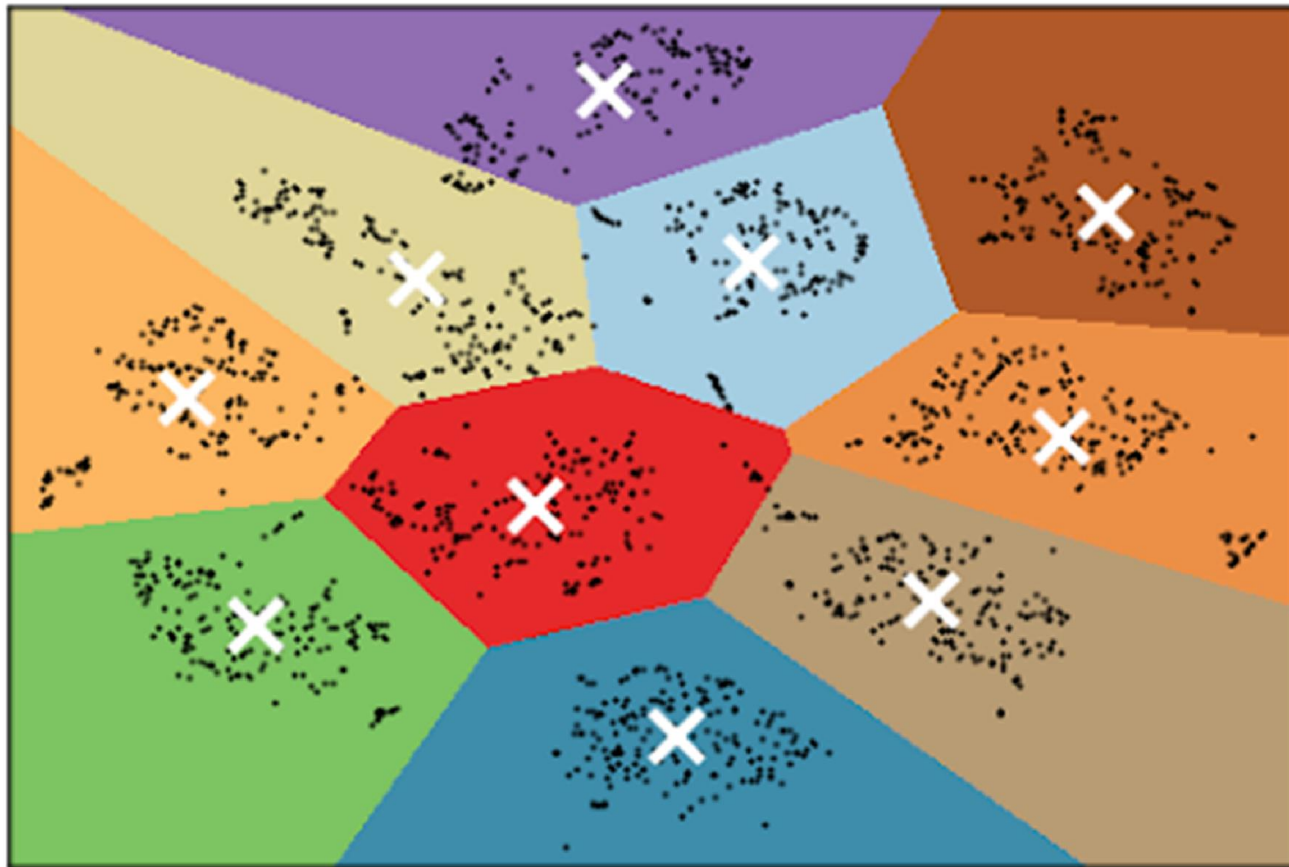# Cluster Analysis: Basic Concepts and Methods

- Cluster Analysis: Basic Concepts

- Partitioning Methods

- Density-Based Methods
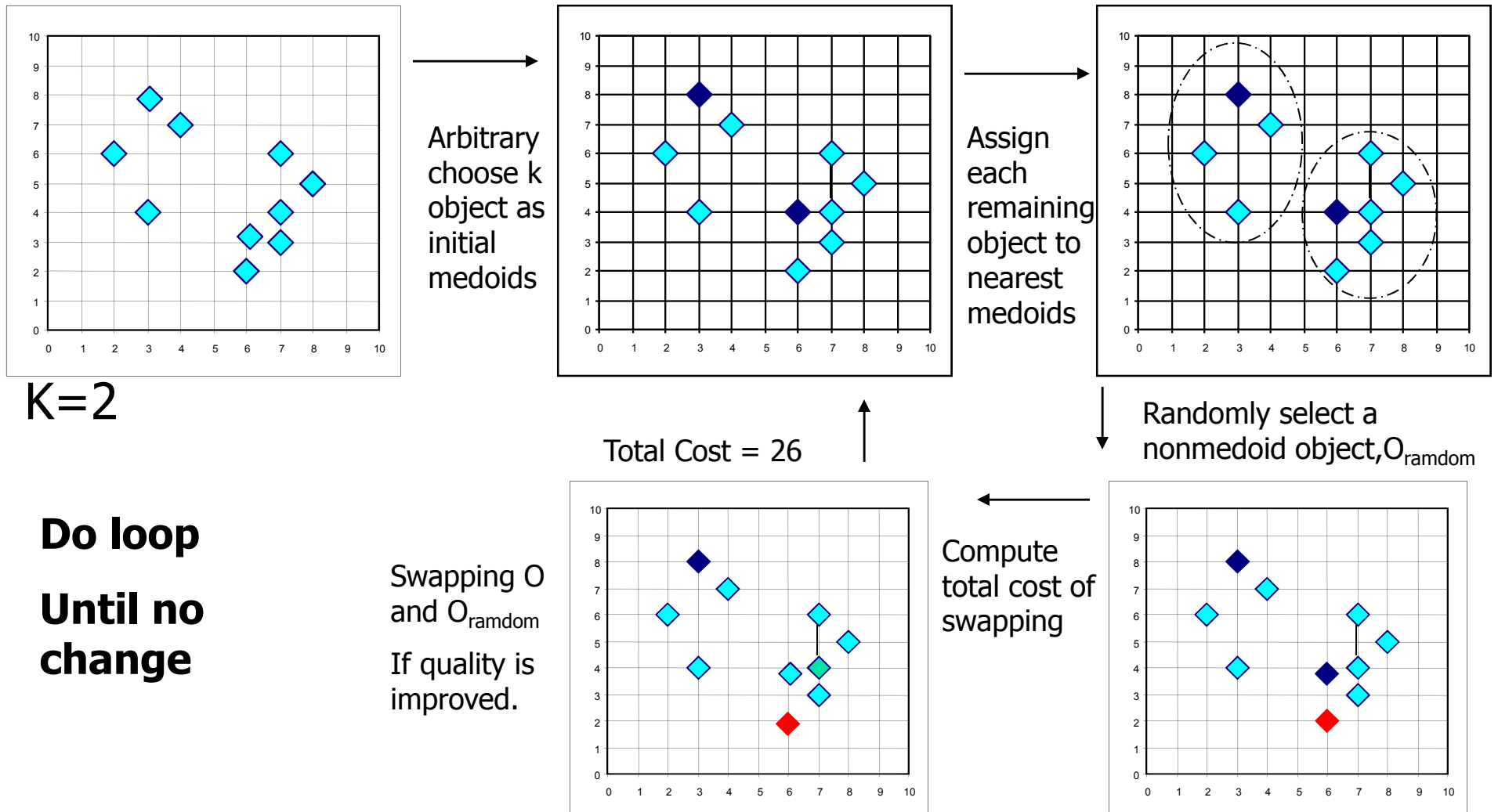
Review: An Example of *K-Means* Clustering

K=2

Arbitrarily partition objects into k groups

The initial data set

Update the cluster centroids

Loop if needed

Reassign objects

Update the cluster centroids

- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until Convergence

14

Review: K-Means Linear Separability

A Typical K-Medoids Algorithm (PAM)

Total Cost = 20

K=2

Do loop

Until no change

Arbitrary choose k object as initial medoids

Assign each remaining object to nearest medoids

Randomly select a nonmedoid object,O_{ramdom}

Total Cost = 26

Swapping O and O_{ramdom}

If quality is improved.

Compute total cost of swapping

PAM Clustering: Finding the Best Cluster Center

- Case 1: p currently belongs to o_j. If o_j is replaced by o_{random} as a representative object and p is the closest to one of the other representative object o_i, then p is reassigned to o_i
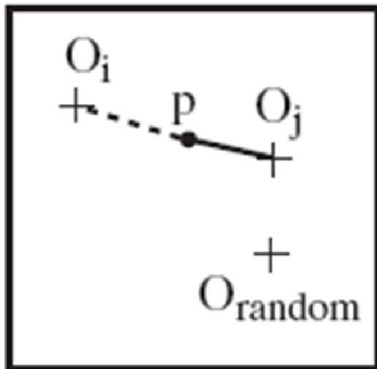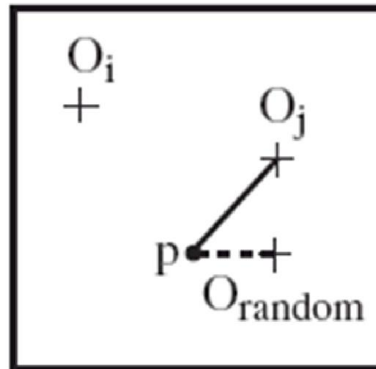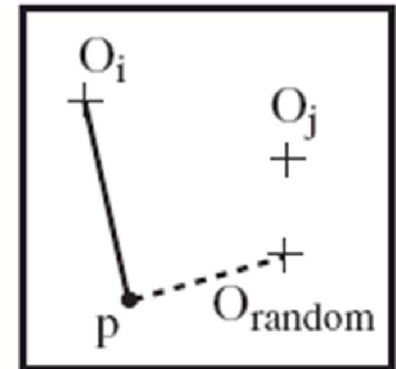

1. Reassigned to O_i 2. Reassigned to O_{random} 3. No change   4. Reassigned to O_{random}
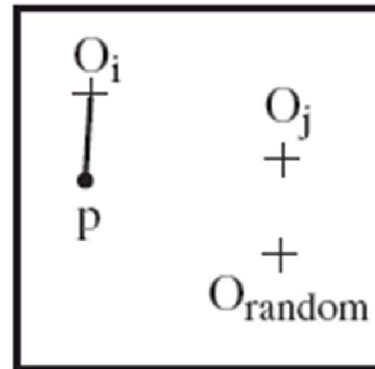
- • data object
- + cluster center
- — before swapping
- --- after swapping

What Is the Problem with PAM?

- Pam is more robust than k-means in the presence of noise and outliers because a medoid is less influenced by outliers or other extreme values than a mean

- Pam works efficiently for small data sets but does not **scale well** for large data sets.

  - $O(k(n-k)^2)$ for each iteration

    where n is # of data,k is # of clusters

➔ Sampling-based method

CLARA(Clustering LARge Applications)

CLARA (Clustering Large Applications) (1990)

- *CLARA* (Kaufmann and Rousseeuw in 1990)

    - Built in statistical analysis packages, such as SPlus

    - It draws *multiple samples* of the data set, applies *PAM* on each sample, and gives the best clustering as the output

- <u>Strength</u>: deals with larger data sets than *PAM*

- <u>Weakness:</u>

    - Efficiency depends on the sample size

    - A good clustering based on samples will not necessarily represent a good clustering of the whole data set if the sample is biased

CLARANS ("Randomized" CLARA) *(1994)*

- *CLARANS* (A Clustering Algorithm based on Randomized Search) (Ng and Han'94)
  - Draws sample of neighbors dynamically
  - The clustering process can be presented as searching a graph where every node is a potential solution, that is, a set of k medoids
  - If the local optimum is found, *it* starts with new randomly selected node in search for a new local optimum
- Advantages: More efficient and scalable than both *PAM* and *CLARA*
- Further improvement: Focusing techniques and spatial access structures (Ester et al.'95)

Cluster Analysis: Basic Concepts and Methods

- Cluster Analysis: Basic Concepts

- Partitioning Methods

- Density-Based Methods

Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - No need to specify K (number of clusters)
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - <u>DBSCAN</u>: Ester, et al. (KDD'96)
  - <u>OPTICS</u>: Ankerst, et al (SIGMOD'99).
  - <u>DENCLUE</u>: Hinneburg & D. Keim (KDD'98)
  - <u>CLIQUE</u>: 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 belongs to D | dist(p,q) ≤ 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:
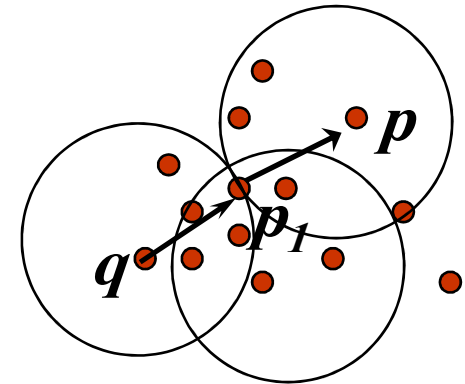
    $$|N_{Eps}(q)| \geq MinPts$$
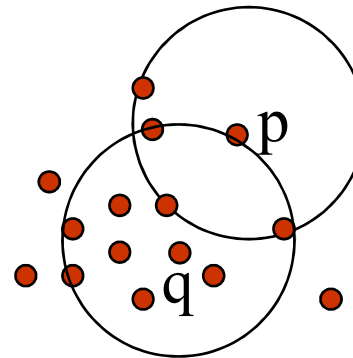
MinPts = 5

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*
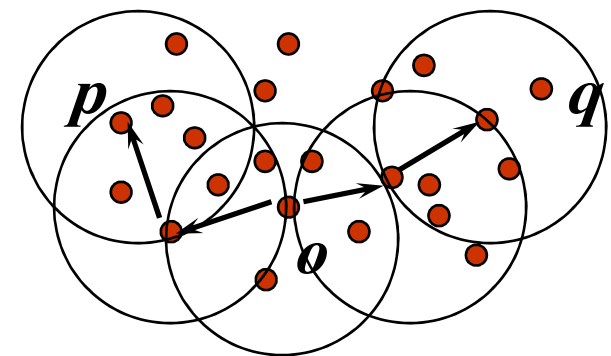
DBSCAN: Density-Based Spatial Clustering of Applications with Noise

- 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

Outlier

Border

Core

Eps = 1cm
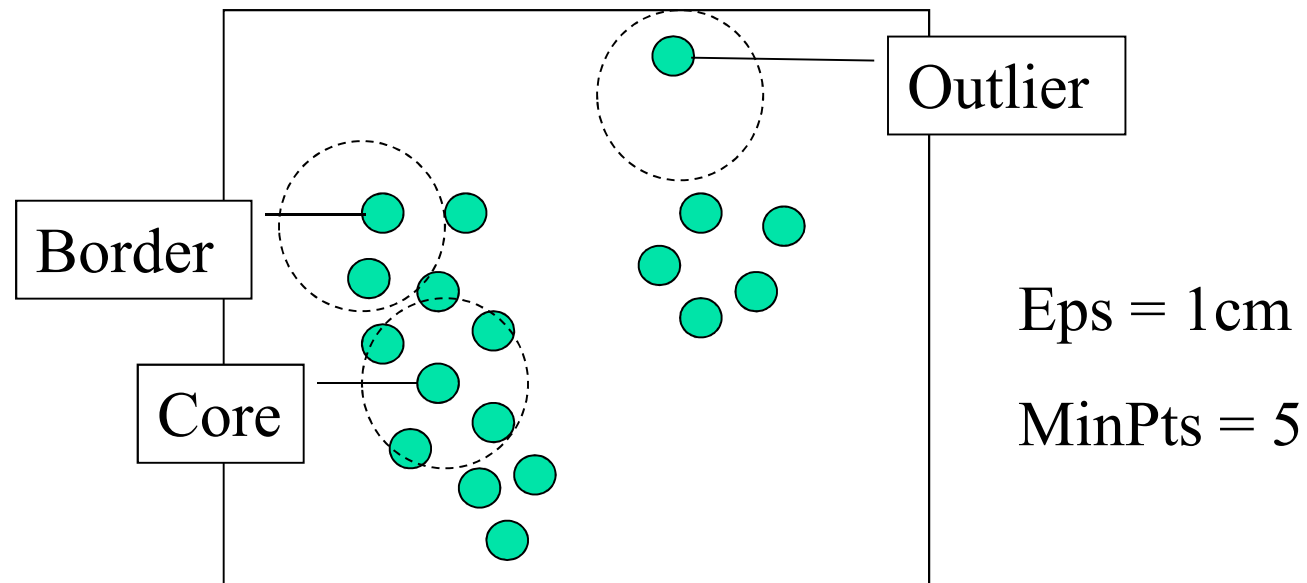
MinPts = 5

DBSCAN: The Algorithm

- Arbitrary select a point p

- Retrieve all points density-reachable from p w.r.t. *Eps* and *MinPts*

- If p is a core point, a cluster is formed

- If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database

- Continue the process until all of the points have been processed

DBSCAN: Sensitive to Parameters

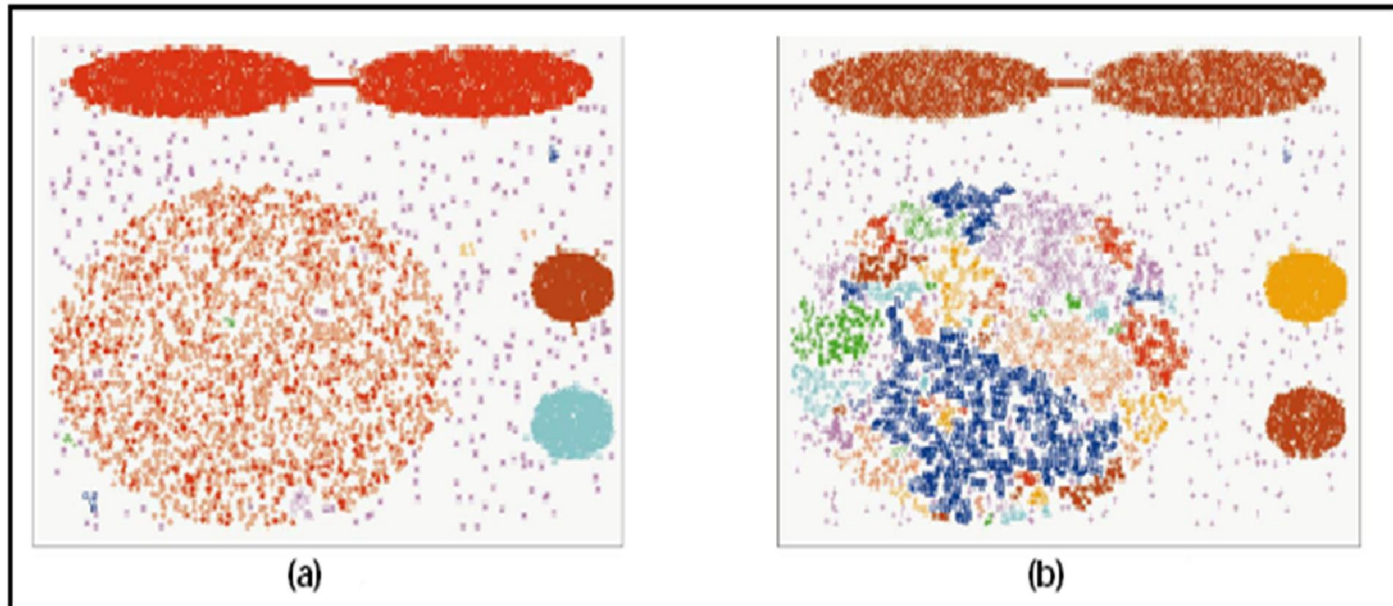Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.
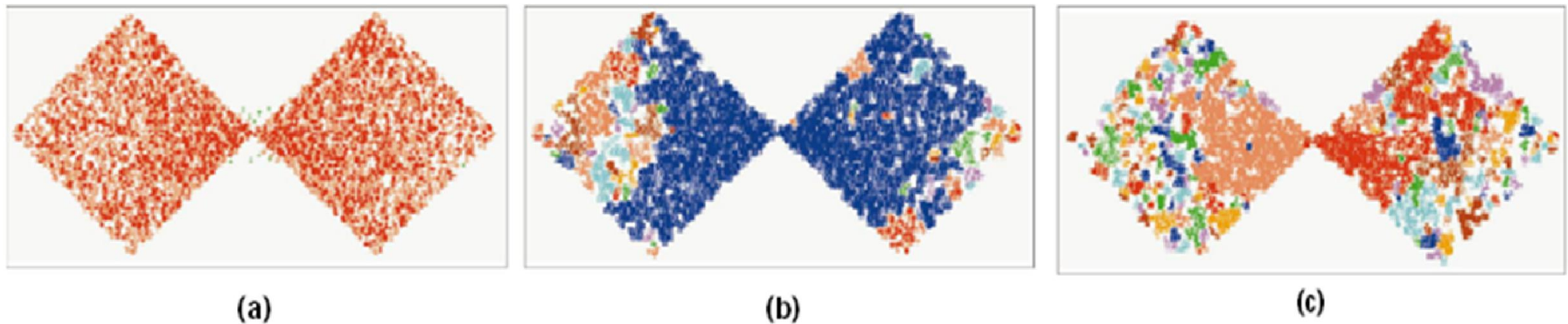
Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.

27

OPTICS: A Cluster-Ordering Method (1999)

- OPTICS: Ordering Points To Identify the Clustering Structure
    - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
    - Produces a special order of the database wrt its density-based clustering structure
    - This cluster-ordering contains info equiv to the density-based clusterings corresponding to a broad range of parameter settings
    - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
    - Can be represented graphically or using visualization techniques

OPTICS: Some Extension from DBSCAN

- Index-based:
  - k = number of dimensions
  - N = 20
  - p = 75%
  - M = N(1-p) = 5
  - Complexity: O(*NlogN*)
- Core Distance:
  - min eps s.t. point is core
- Reachability Distance

Max (core-distance (o), d (o, p))

r(p1, o) = 2.8cm. r(p2,o) = 4cm

p1

o

p2

D

MinPts = 5

o

ε = 3 cm