빅데이터 분석 기술의 이해

이상철

Outline

- Unsupervised learning: clustering approaches
  - Partitioning methods
  - Hierarchical methods
  - Density-based methods
  - Grid-based methods

- LAB
  - Review of classification
  - Clustering methods

Introduction to Machine Learning

- **Supervised:** We are given input/output samples (X, y) which we relate with a function y = f(X). We would like to "learn" f, and evaluate it on new data. Types:
  - **Classification:** y is discrete (class labels).
  - **Regression:** y is continuous, e.g. linear regression.

- **Unsupervised:** Given only samples X of the data, we compute a function f such that y = f(X) is "simpler".
  - **Clustering:** y is discrete
  - y is continuous: **Matrix factorization, Kalman filtering, unsupervised neural networks**.

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)

Clustering for Data Understanding

- Information retrieval: document clustering
- 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
- Land use: Identification of areas of similar land use in an earth observation database
- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Economic Science: market research

Clustering as a Preprocessing Tool

- Summarization:
  - Preprocessing for regression, 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

What is Good Clustering?

- A good clustering method will produce high quality clusters
  - high intra-class similarity: cohesive within clusters
  - low inter-class similarity: distinctive between clusters
- The quality 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 hidden 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

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, CLARANS
- **Hierarchical approach:**
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Diana, Agnes, BIRCH, CAMELEON
- **Density-based approach:**
  - Based on connectivity and density functions
  - Typical methods: DBSACN, OPTICS, DenClue
- 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

Partitioning Algorithms: Basic Concept

- Partitioning method: Partitioning a database **D** of *n* objects into a set of *k* clusters, such that the sum of squared distances is minimized (where c_i is the centroid or medoid of cluster C_i)

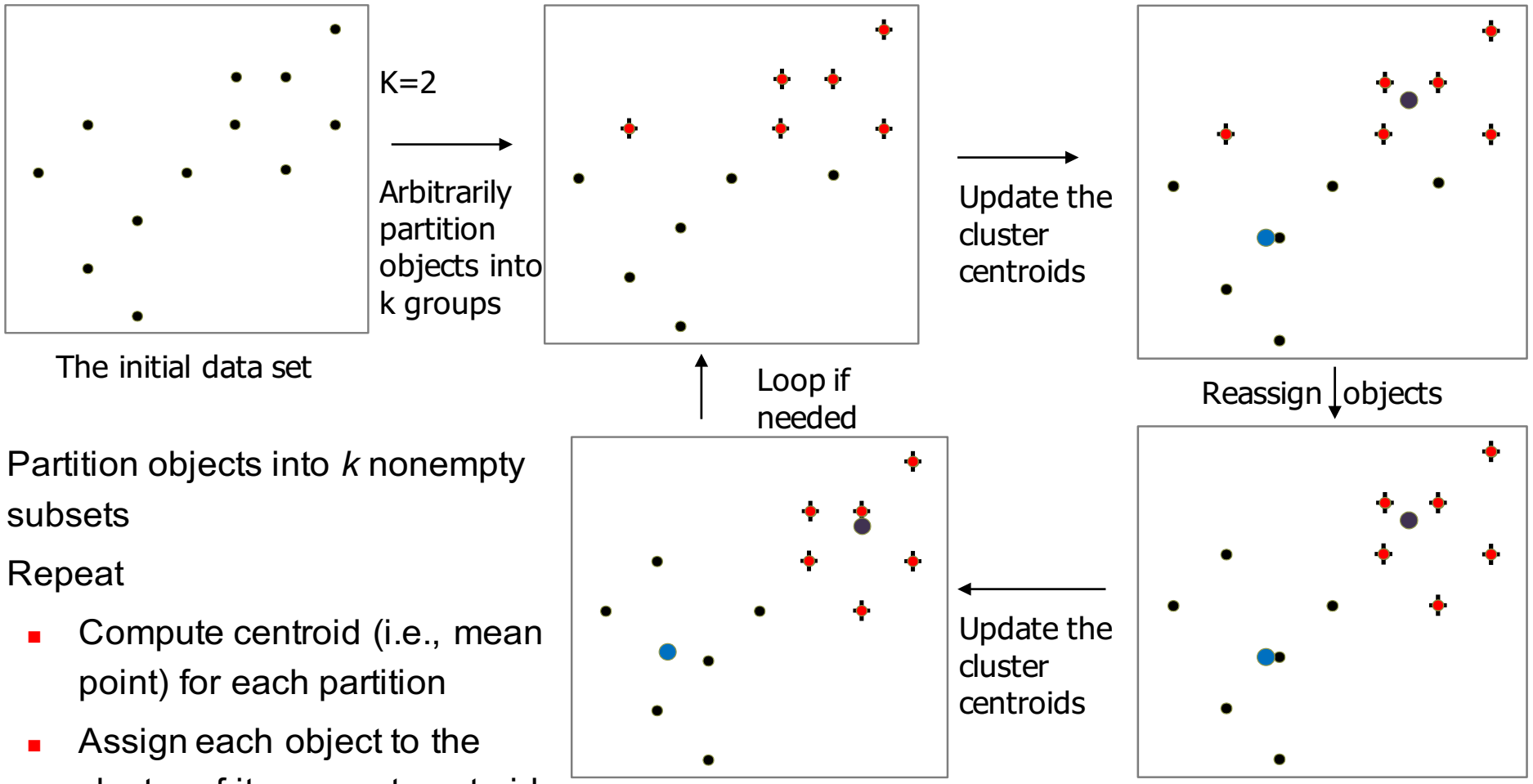
$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (p - c_i)^2$$

- Given *k*, find a partition of *k* clusters that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k*-means and *k*-medoids algorithms
  - *k*-means (MacQueen'67, Lloyd'57/'82): Each cluster is represented by the center of the cluster
  - *k*-medoids or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

The *k*-Means Clustering Method

- Given *k*, the *k*-means algorithm is implemented in four steps:
  - Partition objects into *k* non-empty subsets
  - Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e., mean point, of the cluster)
  - Assign each object to the cluster with the nearest seed point
  - Go back to Step 2, stop when the assignment does not change

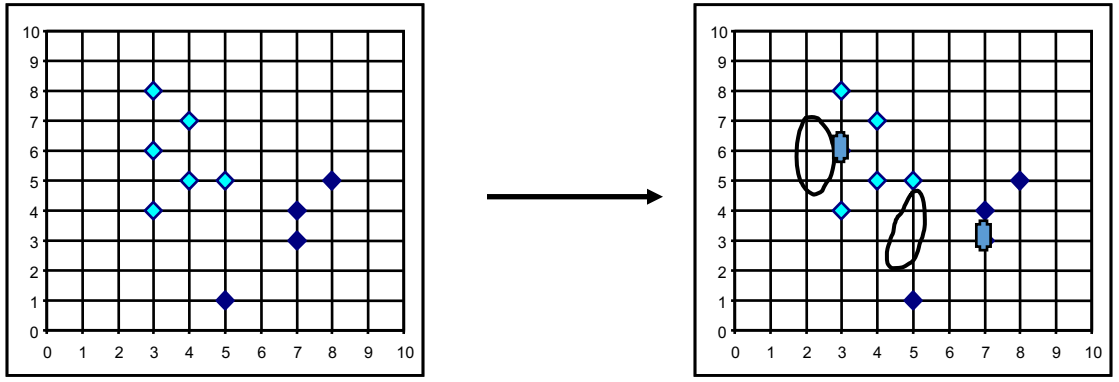
An Example of *k*-Means Clustering



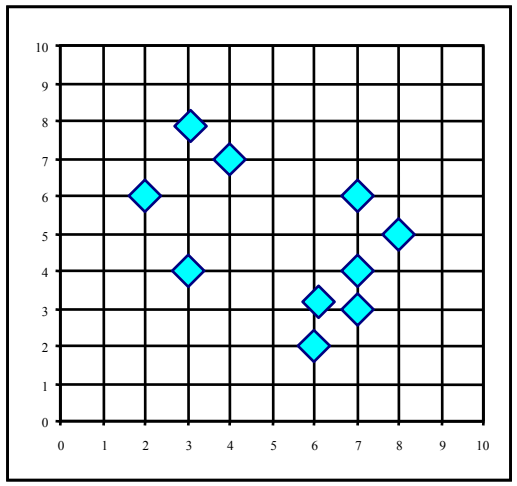
- Partition objects into *k* nonempty subsets
- Repeat
  - Compute centroid (i.e., mean point) for each partition
  - Assign each object to the cluster of its nearest centroid
- Until no change

What Is the Problem of the *k*-Means Method?

- The *k*-means algorithm is sensitive to outliers !
  - Since an object with an extremely large value may substantially distort the distribution of the data
- *k*-Medoids: Instead of taking the **mean** value of the object in a cluster as a reference point, **medoids** can be used, which is the most **centrally located** object in a cluster

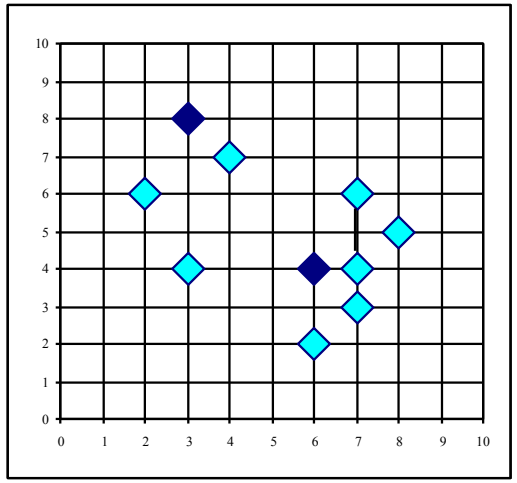


PAM: A Typical *k*-Medoids Algorithm

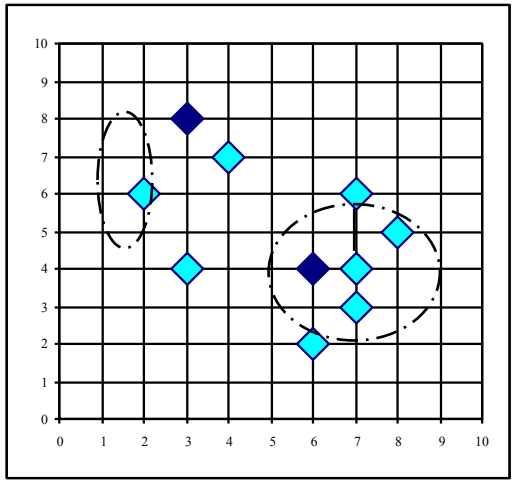


K=2

Arbitrary choose *k* object as initial medoids

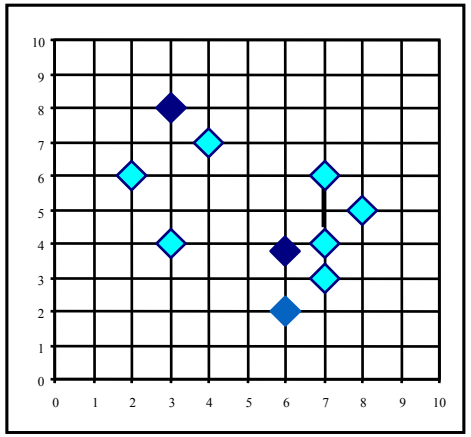


Assign each remaining object to nearest medoids

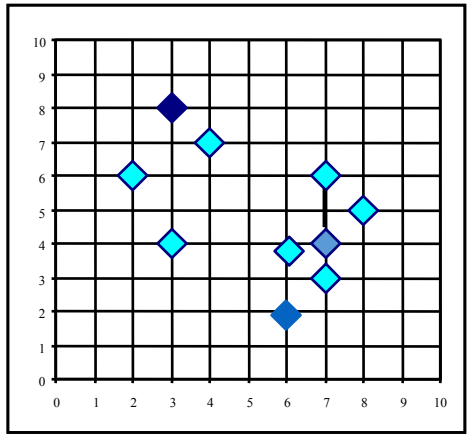


Total Cost = 20

Randomly select a nonmedoid object,O_{ramdom}



Compute total cost of swapping



Total Cost = 26

Swapping O and O_{ramdom}

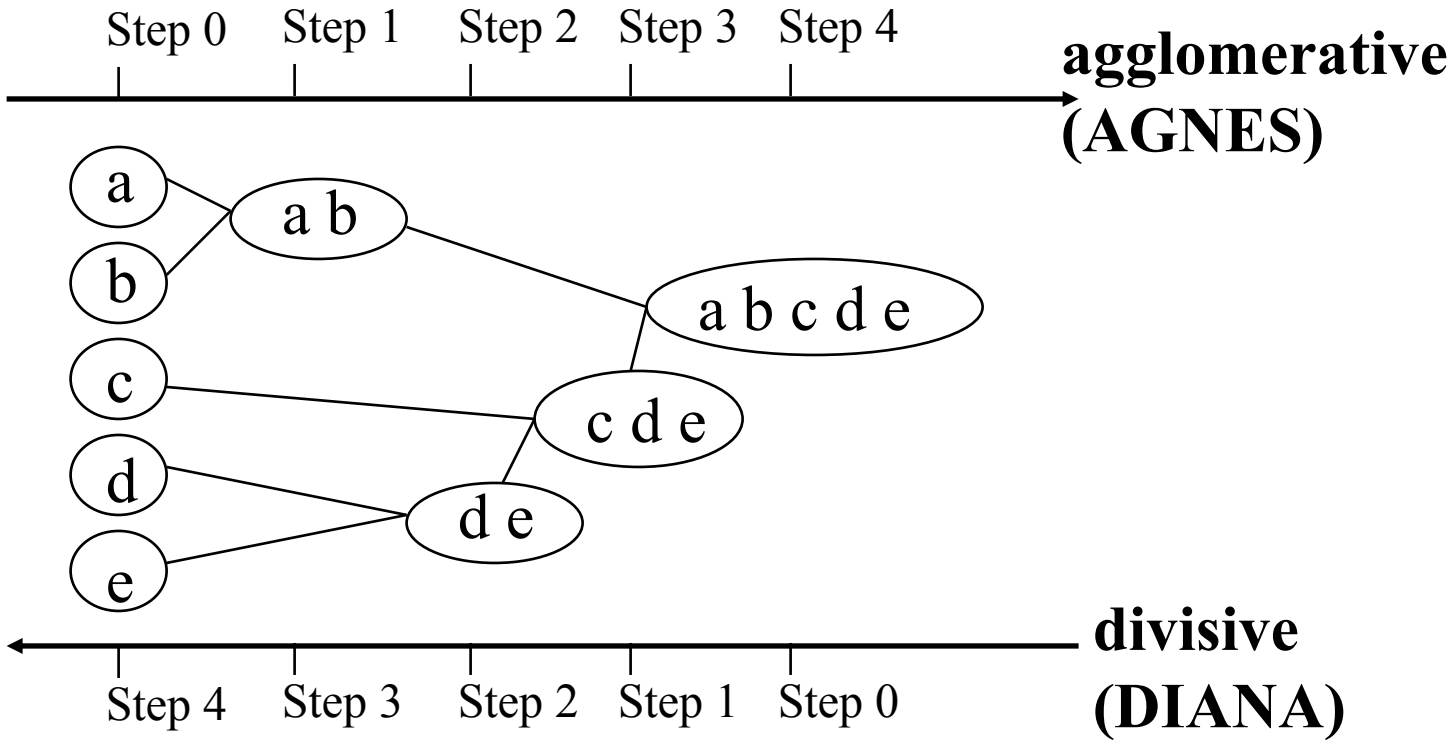
If quality is improved.

Do loop

Until no change

Hierarchical Clustering

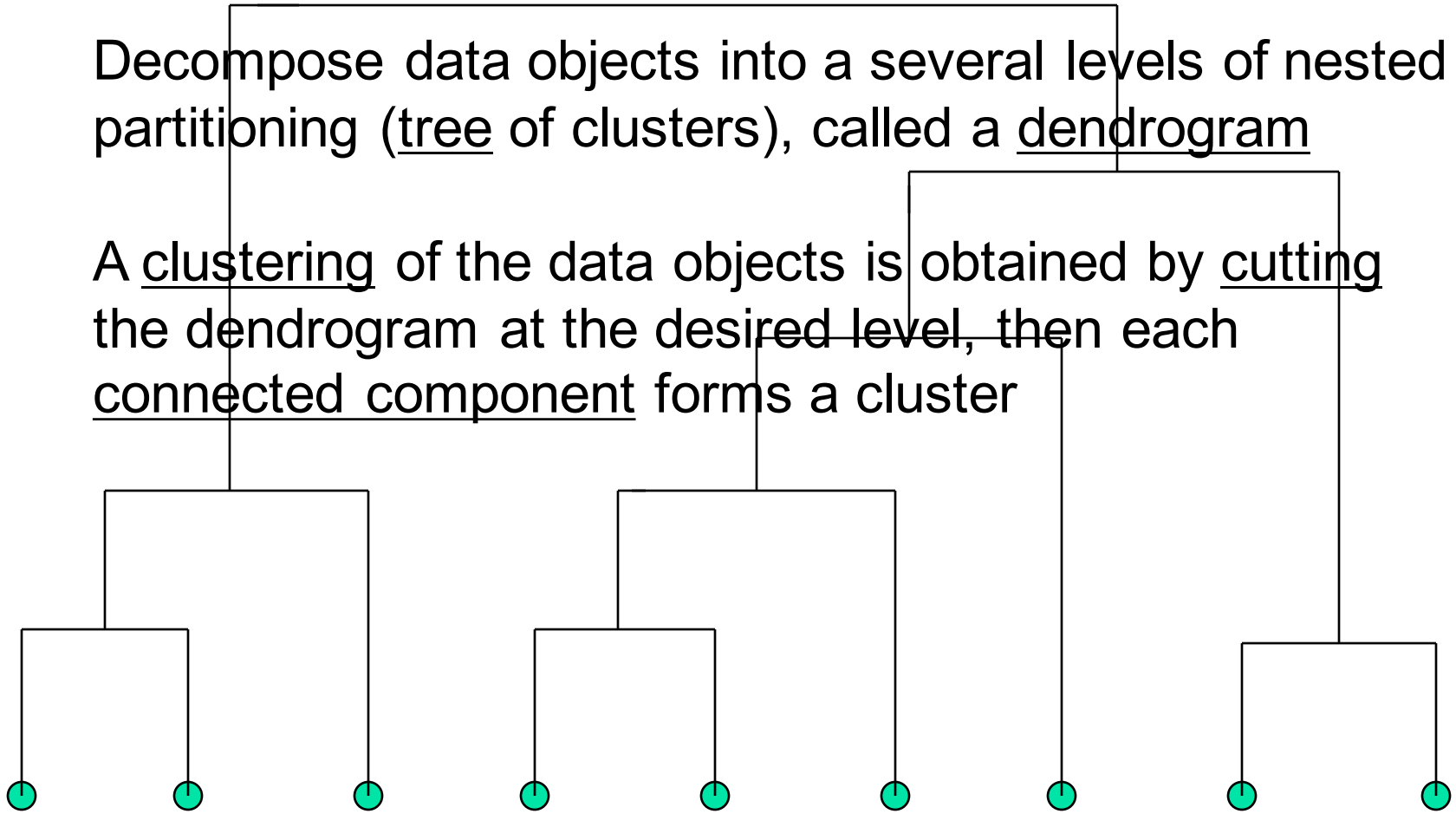
- Use distance matrix as clustering criteria. This method does not require the number of clusters *k* as an input, but needs a termination condition



Dendrogram: Shows How Clusters are Merged

Decompose data objects into a several levels of nested partitioning (tree of clusters), called a dendrogram

A clustering of the data objects is obtained by cutting the dendrogram at the desired level, then each connected component forms a cluster



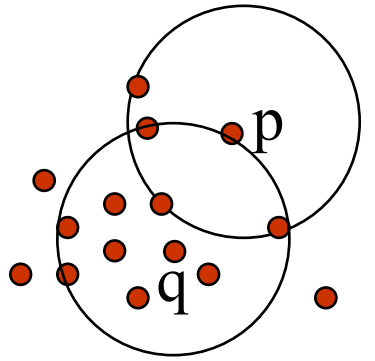
Density-Based Clustering Methods

- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
  - Discover clusters of arbitrary shape
  - Handle noise
  - One scan
  - Need density parameters as termination condition
- Several interesting studies:
  - DBSCAN: Ester, et al. (KDD'96)
  - OPTICS: Ankerst, et al (SIGMOD'99).
  - 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 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:

    $|N_{Eps}(q)| \geq MinPts$

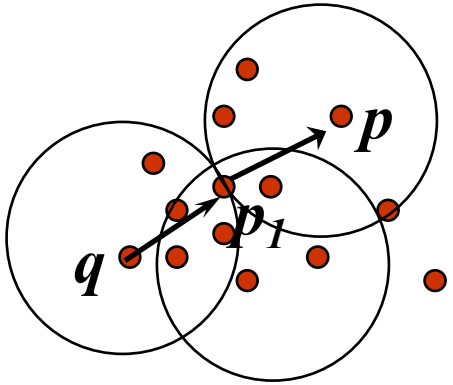


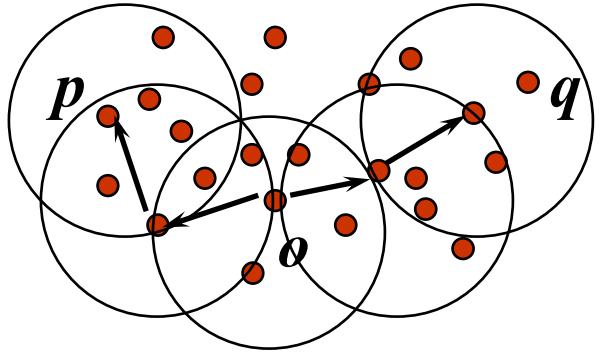
MinPts = 5

Eps = 1 cm

Density-Reachable and Density-Connected

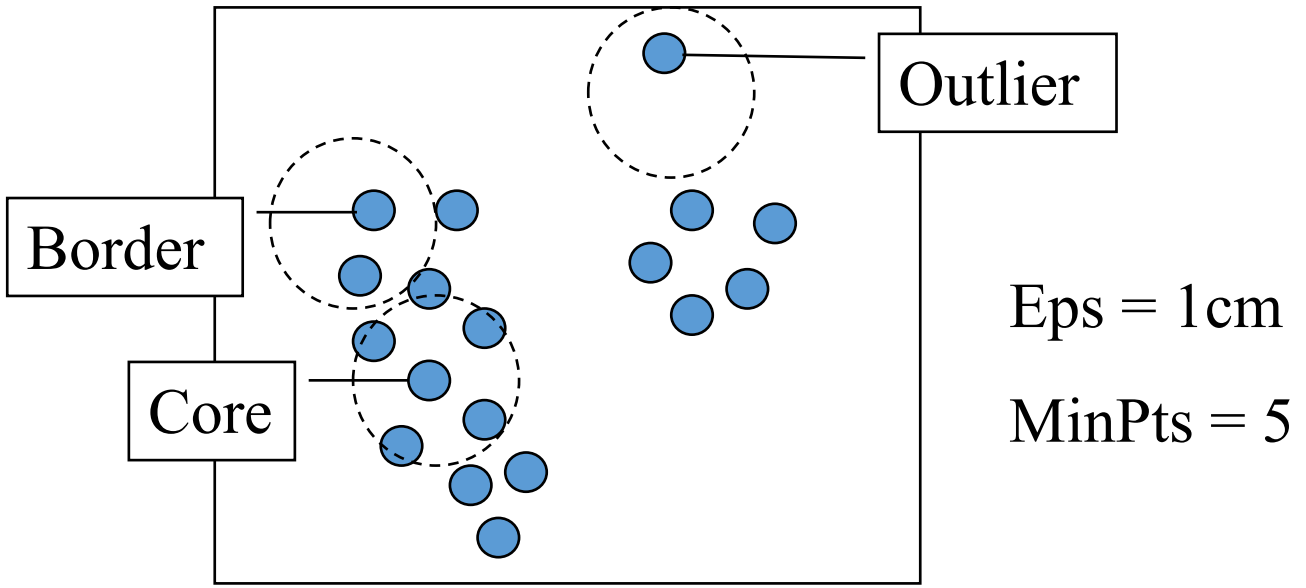
- Density-reachable:
  - A point p is density-reachable from a point q w.r.t. Eps, MinPts if there is a chain of points $p_1, \ldots, p_n, p_1 = q, p_n = p$ such that p_{i+1} is directly density-reachable from p_i
- Density-connected
  - A point p is density-connected to a point q w.r.t. Eps, MinPts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





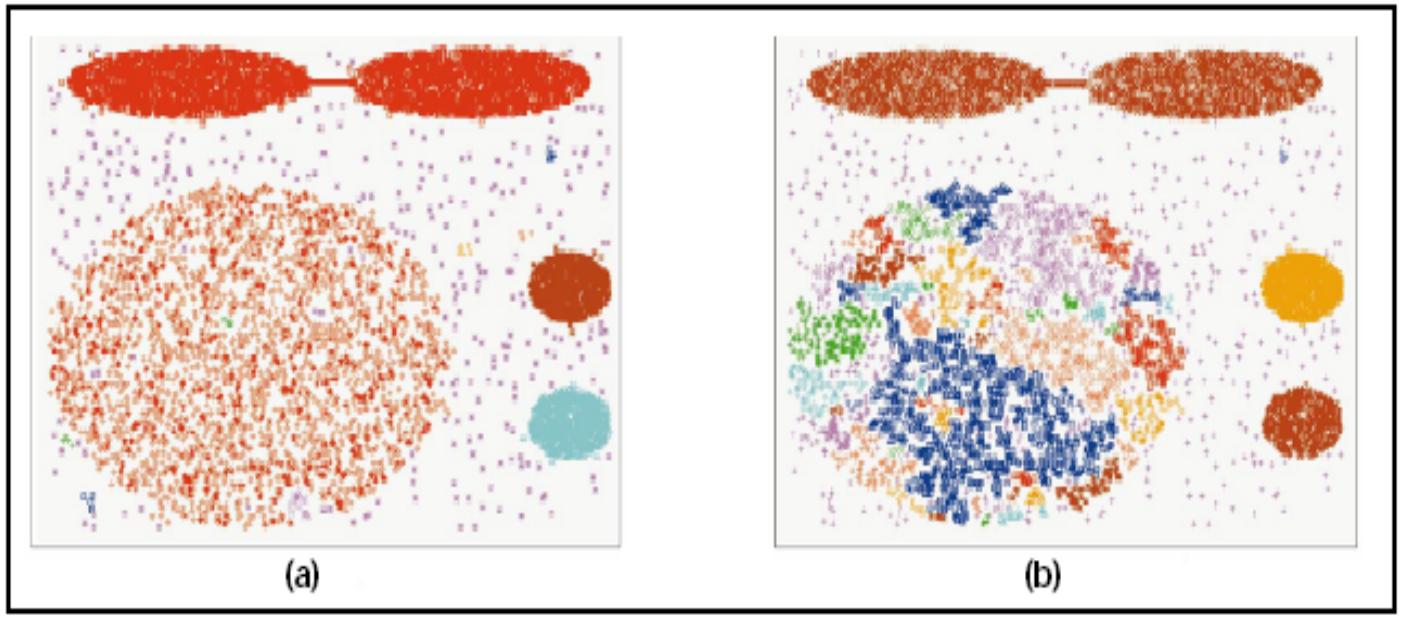
DBSCAN: Density-Based Spatial Clustering of Applications with Noise

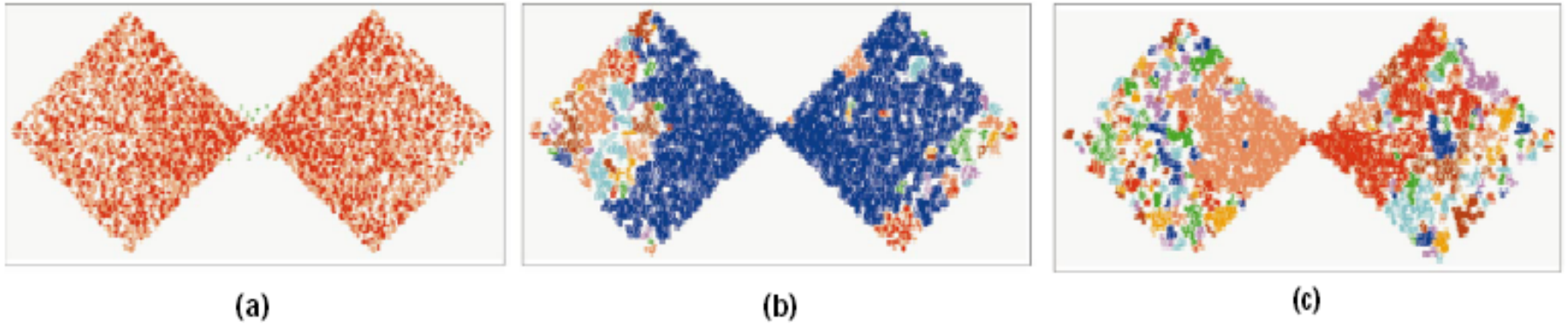
- Relies on a density-based notion of cluster: A cluster is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise



DBSCAN: Sensitive to Parameters

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





Summaries

- Unsupervised learning: clustering approaches
  - Partitioning methods: *k*-means, *k*-medoid
  - Hierarchical methods: AGNES, DIANA
  - Density-based methods: DBSCAN

- LAB
  - Review of classification
  - Clustering methods