Cluster Analysis

——Partitioning Methods——

徐华
清华大学 计算机系 智能技术与系统国家重点实验室
xuhua@tsinghua.edu.cn

1

Cluster Analysis

- What is Cluster Analysis?
- Types of Data in Cluster Analysis
- A Categorization of Major Clustering Methods
- Partitioning Methods
- Hierarchical Methods
- Density-Based Methods
- Grid-Based Methods
- Model-Based Clustering Methods
- Outlier Analysis
- Summary

2

Partitioning Algorithms: Basic Concept

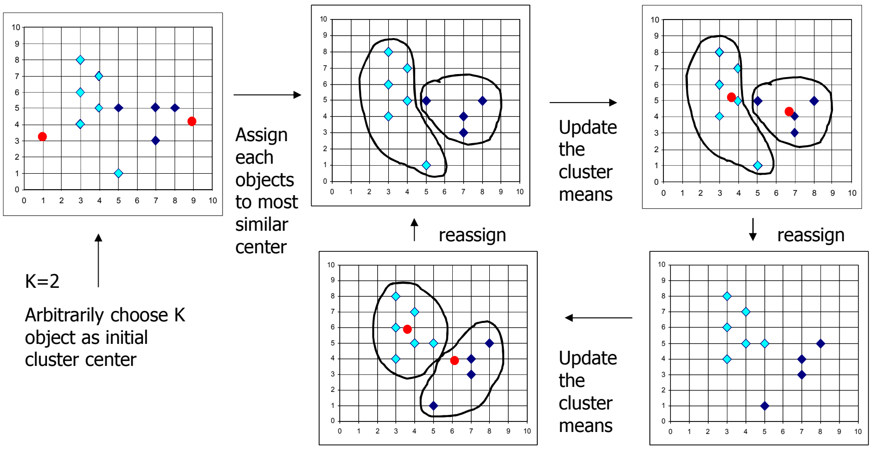
- Partitioning method: Construct a partition of a database *D* of *n* objects into a set of *k* clusters
- Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion
  - Global optimal: exhaustively enumerate all partitions
  - Heuristic methods: *k-means* and *k-medoids* (K-中心点)algorithms
  - *k-means* (MacQueen'67): 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 into *k* nonempty subsets
  - Compute seed *k*, the *k-means* algorithm is implemented in four steps:
  - Partition objects points as the centroids of the clusters of the current partition (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 no more new assignment

The K-Means Clustering Method

- **Example**

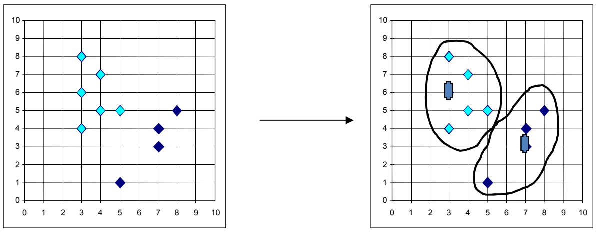


Comments on the K-Means Method

- **<u>Strength:</u>** ***Relatively efficient*****:** ***O*(*tkn*), where *n* is # objects, *k* is # clusters, and *t* is # iterations(迭代). Normally, *k*, *t* << *n*.**
  - **Comparing: PAM: O(k(n-k)^2), CLARA: O(ks^2 + k(n-k))**
- **<u>Comment:</u> Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* （模拟退火）and *genetic algorithms*（遗传算法）**
- **<u>Weakness</u>**
  - **Applicable only when *mean* is defined, then what about categorical data?**
  - **Need to specify *k*, the *number* of clusters, in advance**
  - **Unable to handle noisy data and *outliers***
  - **Not suitable to discover clusters with *non-convex shapes***

The K-Medoids Clustering 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.



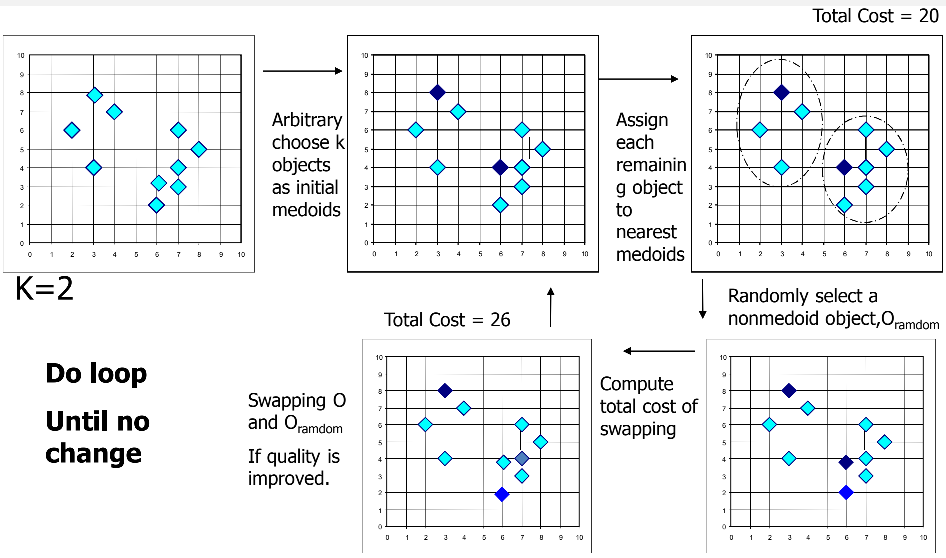
7

The K-Medoids Clustering Method (K中心聚类)

- Find *representative* objects, called medoids, in clusters
- *PAM* (Partitioning Around Medoids, 1987)
  - starts from an initial set of medoids and iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering
  - *PAM* works effectively for small data sets, but does not scale well for large data sets
- *CLARA* (Kaufmann & Rousseeuw, 1990)
- *CLARANS* (Ng & Han, 1994): Randomized sampling
- Focusing + spatial data structure (Ester et al., 1995)

8

Typical k-medoids algorithm (PAM)



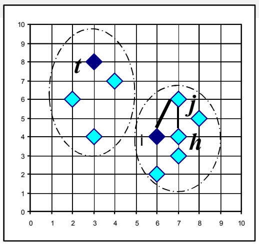
9

PAM (Partitioning Around Medoids) (1987)

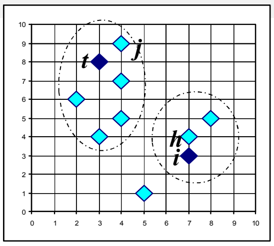
- **PAM (Kaufman and Rousseeuw, 1987), built in Splus**
- **Use real object to represent the cluster**
  1. Select *k* representative objects arbitrarily
  2. For each pair of non-selected object *h* and selected object *i*, calculate the total swapping cost TC_{ih}
  3. For each pair of *i* and *h*,
     a. If $TC_{ih} < 0$, *i* is replaced by *h*
     b. Then assign each non-selected object to the most similar representative object
  4. repeat steps 2-3 until there is no change

10

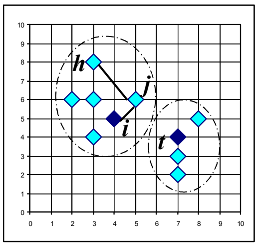
PAM Clustering: Total swapping cost $TC_{ih}=\sum_j C_{jih}$



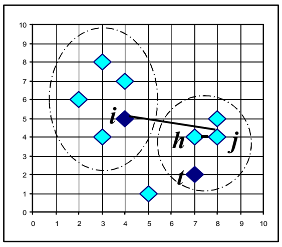
$C_{jih} = d(j, h) - d(j, i)$



$C_{jih} = 0$



$C_{jih} = d(j, t) - d(j, i)$



$C_{jih} = d(j, h) - d(j, t)$

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) and CLARANS

CLARA (Clustering Large Applications) (1990)

- *CLARA* (Kaufmann and Rousseeuw in 1990)
  - Built in statistical analysis packages
- 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

13

Thanks !

14