Cluster Analysis

——Hierarchical 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

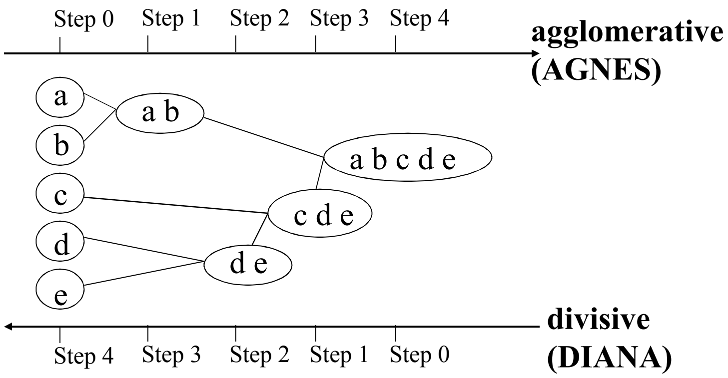
Cluster Analysis

- 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

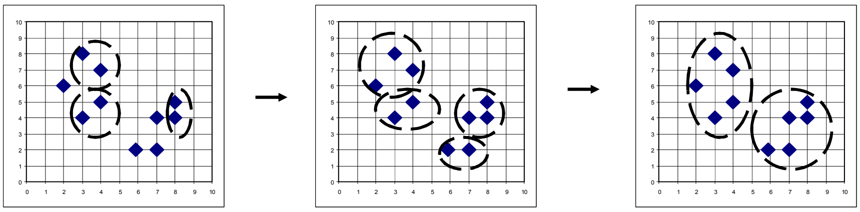
Hierarchical Clustering

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



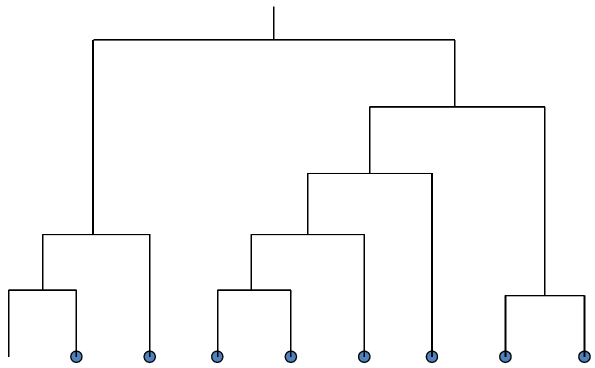
AGNES (Agglomerative Nesting)

- Introduced in Kaufmann and Rousseeuw (1990)
- Implemented in statistical analysis packages, e.g., Splus
- Use the Single-Link method and the dissimilarity matrix.
- Merge nodes that have the least dissimilarity
- Go on in a non-descending fashion
- Eventually all nodes belong to the same cluster



A Dendrogram Shows How the Clusters are Merged Hierarchically

- 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.



More on Hierarchical Clustering Methods

- Major weakness of agglomerative clustering methods
  - do not scale well: time complexity of at least $O(n^2)$, where n is the number of total objects
  - can never undo what was done previously
- Integration of hierarchical with distance-based clustering
  - BIRCH (1996): uses CF-tree and incrementally adjusts the quality of sub-clusters
  - CURE (1998): selects well-scattered points from the cluster and then shrinks them towards the center of the cluster by a specified fraction
  - CHAMELEON (1999): hierarchical clustering using dynamic modeling

BIRCH (1996)

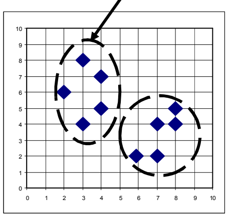
- **Birch: Balanced Iterative Reducing and Clustering using Hierarchies, by Zhang, Ramakrishnan, Livny (SIGMOD'96)**
- **Incrementally construct a CF (Clustering Feature) tree, a hierarchical data structure for multiphase clustering**
  - **Phase 1: scan DB to build an initial in-memory CF tree (a multi-level compression of the data that tries to preserve the inherent clustering structure of the data)**
  - **Phase 2: use an arbitrary clustering algorithm to cluster the leaf nodes of the CF-tree**
- ***Scales linearly*: finds a good clustering with a single scan and improves the quality with a few additional scans**
- ***Weakness:* handles only numeric data, and sensitive to the order of the data record.**

7

Clustering Feature Vector

- **Clustering Feature: *CF = (N, LS, SS)***
- ***N*: Number of data points**
- *LS: ∑Ni= ∑ Xi*
- *SS: ∑Ni= ∑ Xi^2*

CF = (5, (16,30),(54,190))



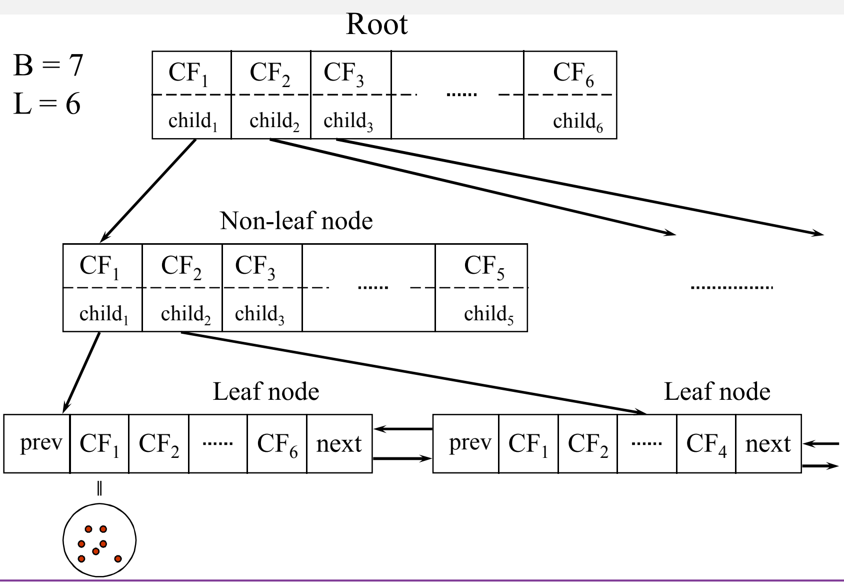
(3,4)
(2,6)
(4,5)
(4,7)
(3,8)

8

CF-Tree in BIRCH

- **Clustering feature:**
  - **summary of the statistics for a given subcluster: the 0-th, 1st and 2nd moments of the subcluster from the statistical point of view.**
  - **registers crucial measurements for computing cluster and utilizes storage efficiently**
- **A CF tree is a height-balanced tree that stores the clustering features for a hierarchical clustering**
  - **A nonleaf node in a tree has descendants or "children"**
  - **The nonleaf nodes store sums of the CFs of their children**
- **A CF tree has two parameters**
  - **Branching factor: specify the maximum number of children.**
  - **Threshold: max diameter of sub-clusters stored at the leaf nodes**

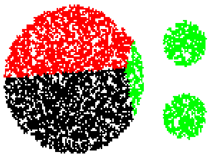
CF Tree



CURE (Clustering Using REpresentatives)

- CURE: proposed by Guha, Rastogi & Shim, 1998
  - Stops the creation of a cluster hierarchy if a level consists of *k* clusters
  - Uses multiple representative points to evaluate the distance between clusters, adjusts well to arbitrary shaped clusters and avoids single-link effect

(a)

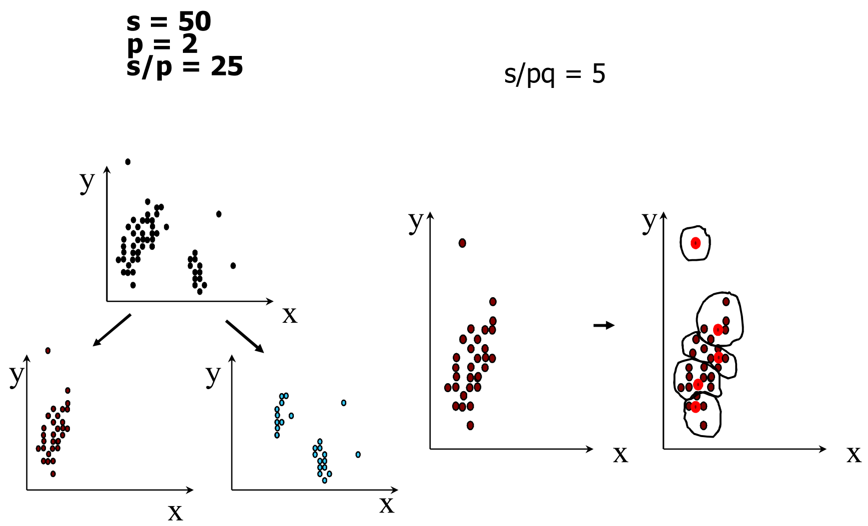


(b)

Cure: The Algorithm

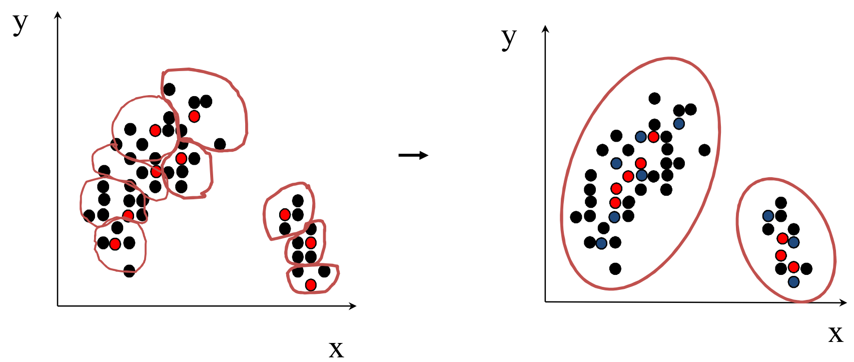
- Draw random sample *s*.
- Partition sample to *p* partitions with size *s/p*
- Partially cluster partitions into *s/pq* clusters
- Eliminate outliers
  - By random sampling
  - If a cluster grows too slow, eliminate it.
- Cluster partial clusters.
- Label data in disk

Data Partitioning and Clustering



Cure: Shrinking Representative Points

- Shrink the multiple representative points towards the gravity center by a fraction of α.
- Multiple representatives capture the shape of the cluster

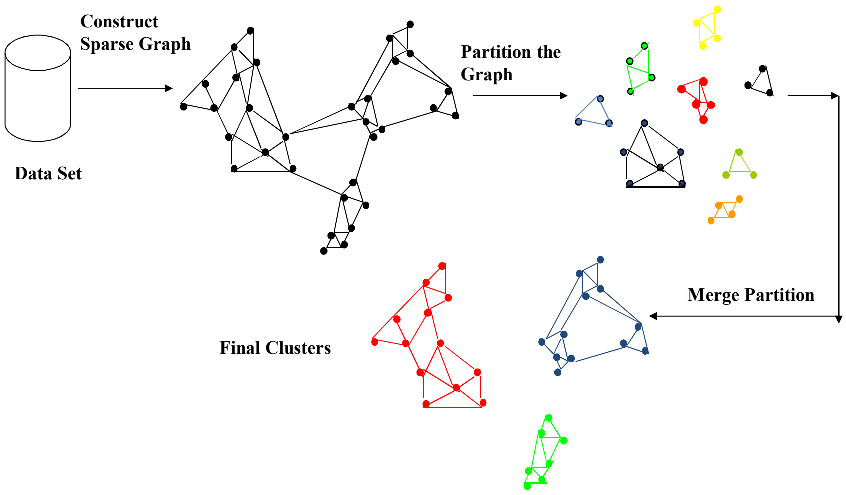


CHAMELEON (Hierarchical clustering using dynamic modeling)

- CHAMELEON: by G. Karypis, E.H. Han, and V. Kumar'99
- Measures the similarity based on a dynamic model
  - Two clusters are merged only if the *interconnectivity* and *closeness (proximity)* between two clusters are high *relative to* the internal interconnectivity of the clusters and closeness of items within the clusters
  - Cure ignores information about interconnectivity of the objects, Rock ignores information about the closeness of two clusters
- A two-phase algorithm
  1. Use a graph partitioning algorithm: cluster objects into a large number of relatively small sub-clusters
  2. Use an agglomerative hierarchical clustering algorithm: find the genuine clusters by repeatedly combining these sub-clusters

15

Overall Framework of CHAMELEON



16

Thanks!