Data Mining
(EECS 6412)

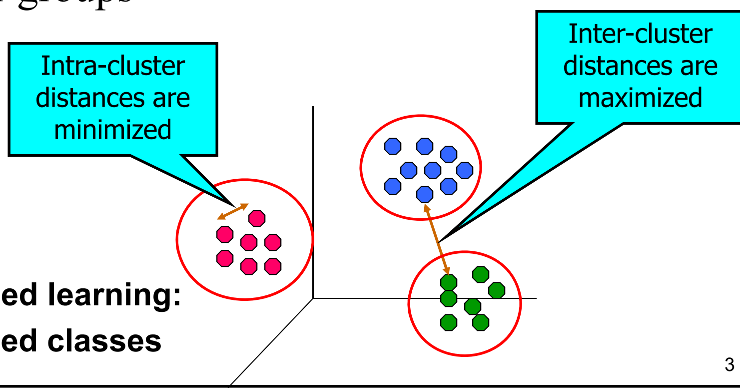
Clustering

Outline

- What is Clustering?
- Types of Data in Cluster Analysis and Similarity Measures
- Clustering Methods
  - K-means
  - K-medoids
  - Hierarchical clustering method
  - DBSCAN: a Density-based Algorithm
- Cluster Validity Measures

What Is Clustering?

- Group data into clusters so that the points in one group are similar to each other and are as different as possible from the points in other groups



Unsupervised learning: no predefined classes

3

Examples of Clustering Applications

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Image processing: Soil scientists filter trees from background
- Genomics: Group genes to predict possible functions of genes with unknown function
- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- WWW
  - Cluster web documents
  - Cluster web log data to discover groups of users

4

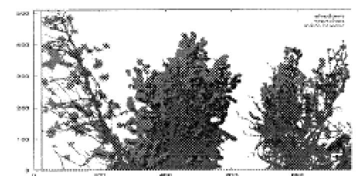
Example

- Filtering real images
  - Images of trees taken in near-infrared band (NIR) and visible wavelength (VIS)
  - 512x1024 pixels and each of them contains a pair of brightness values (NIR,VIS)

The images taken in NIR and VIS

1st step: 284 seconds

2nd step: 71 seconds



The sunlit leaves, branches and shadows

5

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
  - high intra-class similarity
  - low inter-class similarity
- The quality of a clustering method is measured by its ability to discover some or all of the hidden patterns.
- The quality of a clustering result depends on
  - the clustering method
  - the similarity measure used by the method.

6

Data Representation

- Typical Data matrix

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

- Similar to the table used in classification algorithm, but without the class attribute.

Similarity (or Dissimilarity) Measures

- For data containing only *numeric attributes*, a popular measure of distance between two data objects, *i* and *j*, is *Minkowski distance*:

$$d(i,j) = \sqrt[q]{(|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + ... + |x_{ip} - x_{jp}|^q)}$$

where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two *p*-dimensional data objects, and *q* is a positive integer.

- If $q = 1$, d is *Manhattan distance*

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}|$$

Similarity and Dissimilarity Between Objects (*Cont'd*)

- *If q = 2, d* is Euclidean distance:

$$d(i,j) = \sqrt{(x_{i_1} - x_{j_1})^2 + (x_{i_2} - x_{j_2})^2 + \ldots + (x_{i_p} - x_{j_p})^2}$$

- Properties of Minkowski distance
  - *d(i,j)* ≥ 0 (nonnegative)
  - *d(i,i)* = 0 (distance to itself is 0)
  - *d(i,j)* = *d(j,i)* (symmetric)
  - *d(i,j)* ≤ *d(i,k)* + *d(k,j)* (Triangular inequality)
- Also, one can use weighted distance (or other dissimilarity measures)

$$d(i,j) = \sqrt[q]{w_1 |x_{i_1} - x_{j_1}|^q + w_2 |x_{i_2} - x_{j_2}|^q + \ldots + w_p |x_{i_p} - x_{j_p}|^q)} \quad (q > 0)$$

9

Similarity (or Dissimilarity) Measures

- Another similarity measure for data with only numeric attributes is cosine similarity:

$$\text{Similarity}(\mathbf{x}, \mathbf{y}) = \cos(\theta) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

where θ is the angle between two examples **x** and **y** .

10

Similarity (or Dissimilarity) Measures

- What if data contain *both numeric and nominal attributes*?
- Can use the distance measure that we introduced in the kNN method:

$$\text{Distance } (x, y) = \sum_{i=1}^{m} dist\,(x_i, y_i)$$

where $(x_1, x_2, \ldots, x_m)$ and $(y_1, y_2, \ldots, y_m)$ are two instances;

$$dist(x_i, y_i) = \begin{cases} 0 & \text{if } x_i \text{ and } y_i \text{ are nominal and } x_i = y_i \\ 1 & \text{if } x_i \text{ and } y_i \text{ are nominal and } x_i \neq y_i \\ |\,norm(x_i) - norm(y_i)\,| & \text{if } x_i \text{ and } y_i \text{ are continuous} \end{cases}$$

11

Outline

- What is Clustering?
- Types of Data in Cluster Analysis and Similarity Measures
- *Clustering Methods*
  - K-means
  - K-medoids
  - Hierarchical clustering method
  - DBSCAN: a Density-based Algorithm
- Cluster Validity Measures

12

Major Clustering Approaches

- Partitioning algorithms: Construct various partitions and then evaluate them by some criterion
- Hierarchical algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion
- Density-based: based on connectivity and density functions
- Model-based: A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other

Partitioning Algorithms: Basic Concept

- Partition *n* objects into *k* clusters
  - Optimize the chosen partitioning criterion
- Global optimal: exhaustively enumerate all partitions
  - Infeasible in practice
- Heuristic methods: *k-means* and *k-medoids* algorithms
  - *k-means* (MacQueen'67): Each cluster is represented by the mean values of the objects in the cluster (as the cluster center)
  - *k-medoids* or PAM (Partition around medoids) (Kaufman & Rousseeuw'87): Each cluster is represented by one of the objects in the cluster

K-means

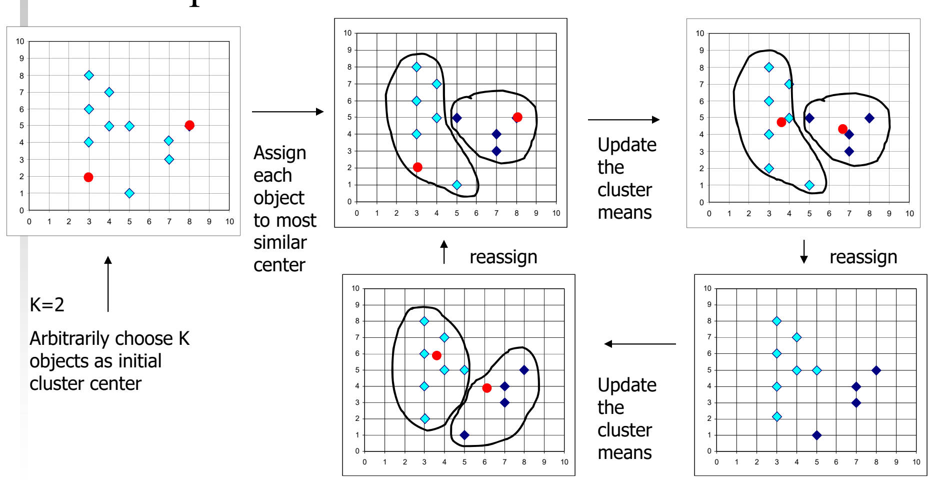
Group objects into *k* clusters:

- Arbitrarily choose *k* objects as the initial cluster centers
- Until no change, do
  - (Re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster
  - Update the cluster means, i.e., calculate the mean value of the objects for each cluster

15

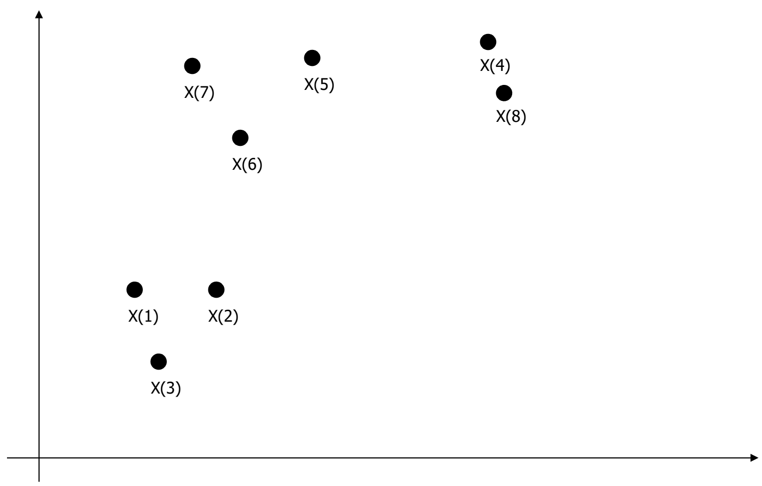
The *K-Means* Clustering Method

- Example



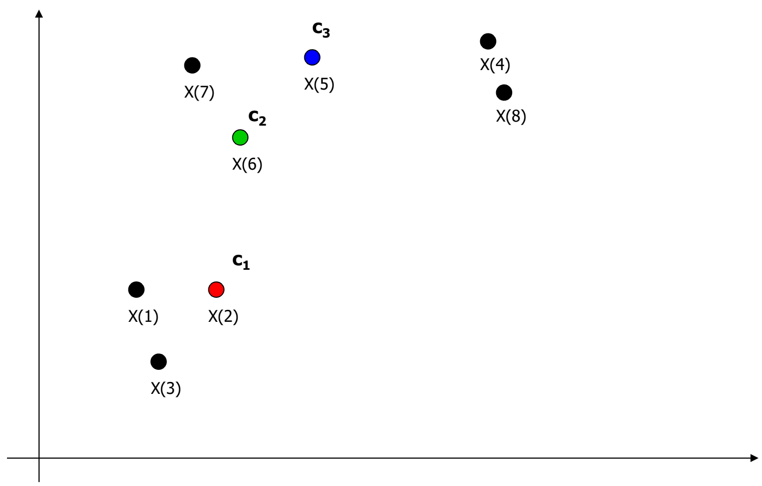
16

K-means example



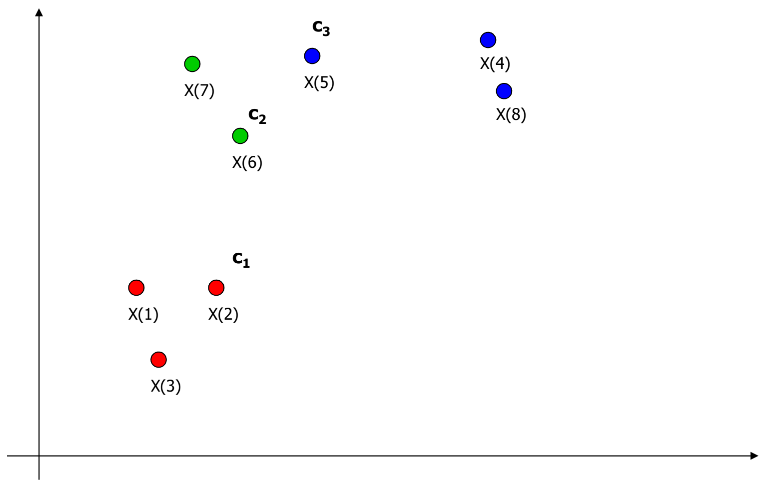
17

K-means example



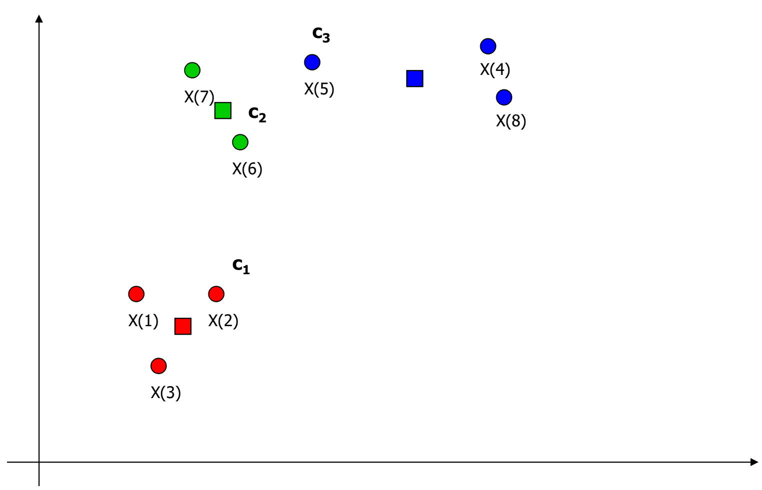
18

K-means example



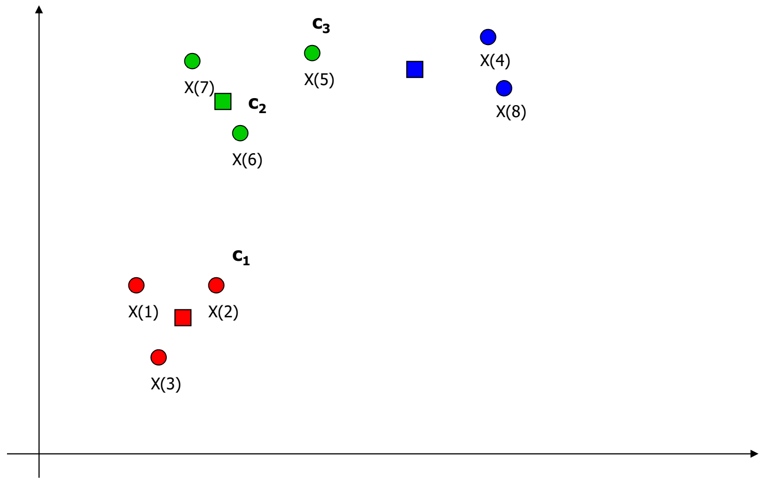
19

K-means example

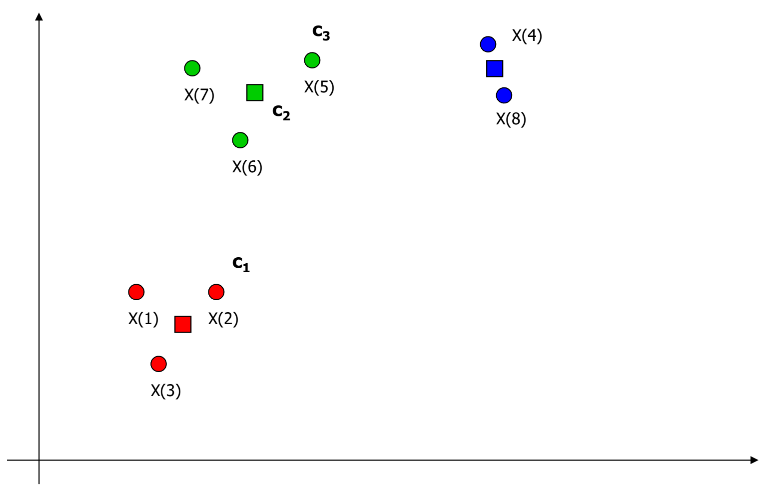


20

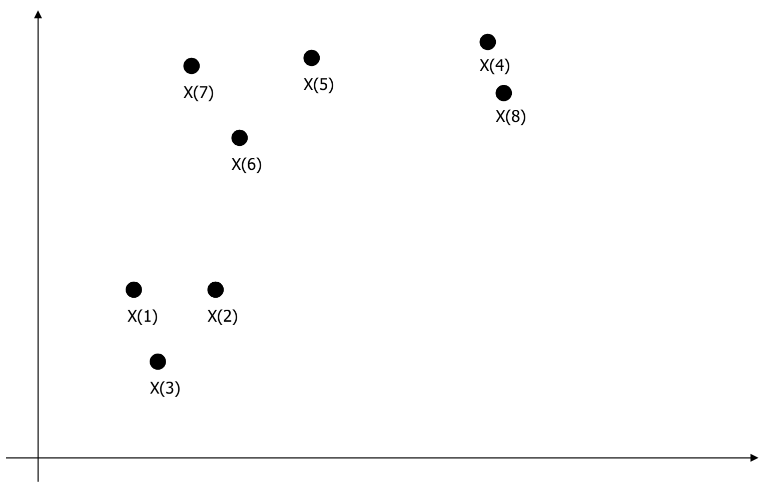
K-means example



K-means example

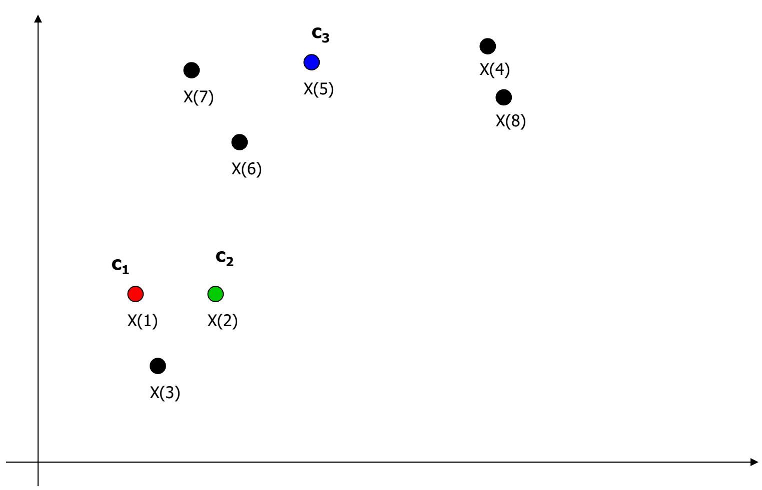


K-means example #2



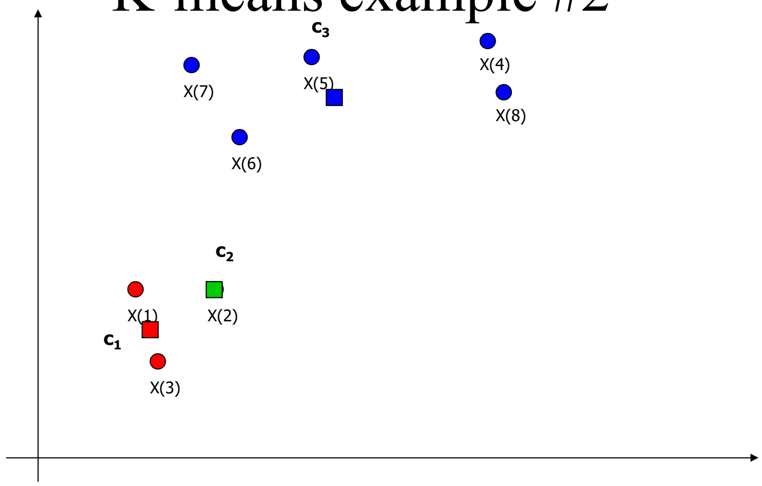
23

K-means example #2



24

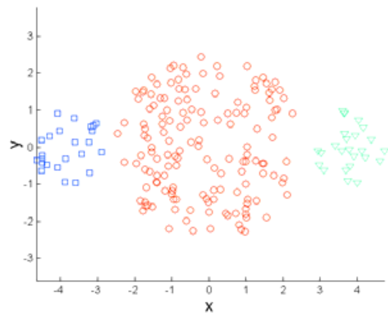
K-means example #2



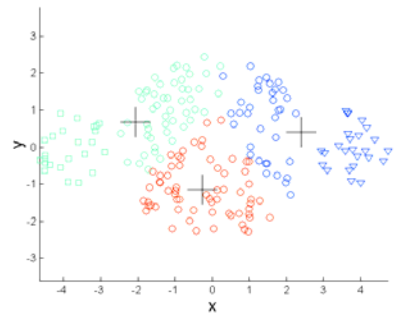
Comments on the *K-Means* Method

- Strength: *Relatively efficient*: $O(tkn)$, where n is # objects, k is # clusters, and t is # iterations. Normally, k, $t << n$.
  - Comparing: PAM: $O(k(n-k)^2)$, CLARA: $O(ks^2 + k(n-k))$
- Comment: Often terminates at a *local optimum*.
- Weakness
  - Sensitive to the initial clusters
  - Applicable only when *mean* is defined, then what about categorical data?
  - Need to specify *k,* the *number* of clusters, in advance
  - Very sensitive to noise and *outliers*
  - May have a problem when clusters have different sizes and are close to each other. example
  - Not suitable to discover clusters with *non-convex shapes* example

A Limitation of K-means: Differing Sizes



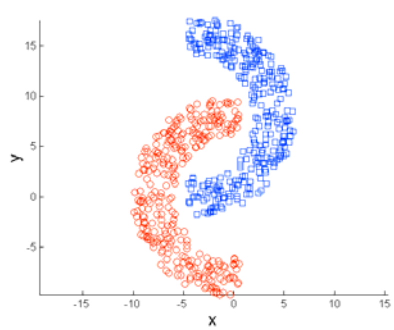
Original Points



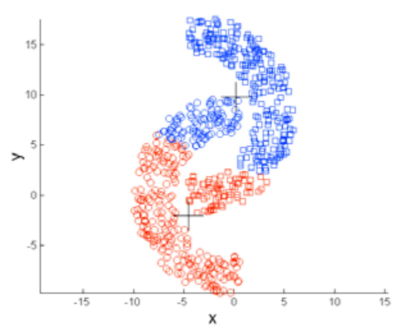
K-means (3 Clusters)

27

A Limitation of K-means: Non-globular (Non- Convex) Shapes



Original Points

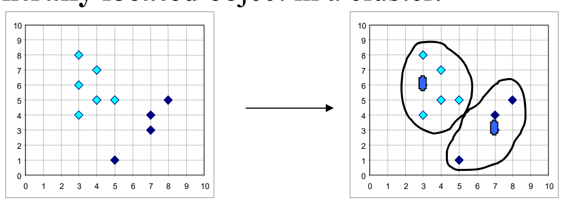


K-means (2 Clusters)

28

A Problem of 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 objects in a cluster as a reference point, **medoids** can be used, which is the **most centrally located** object in a cluster.



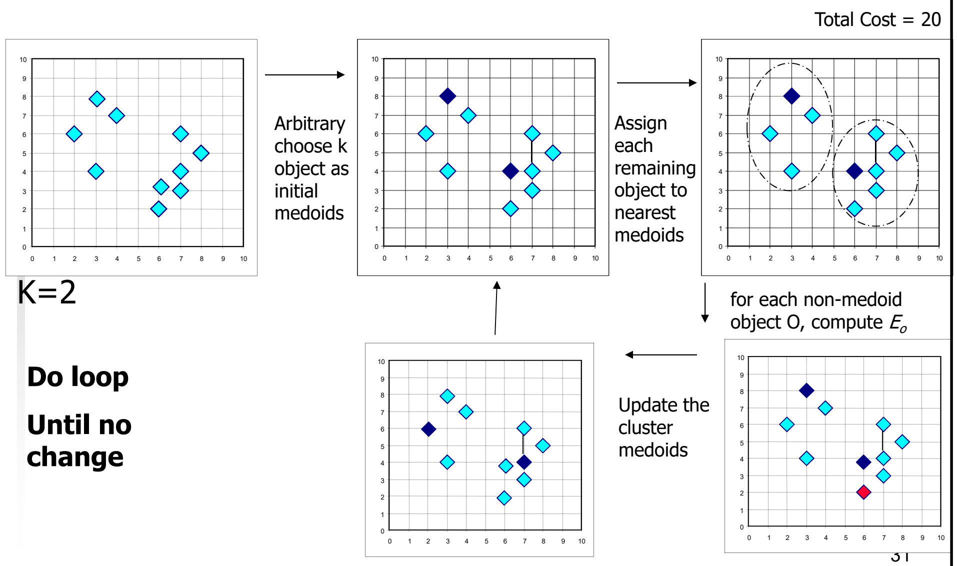
29

PAM: A K-medoids Method

- PAM: Partitioning Around Medoids (1987)
- Arbitrarily choose *k* objects as the initial medoids
- Until no change, do
  - (Re)assign each object to the cluster whose medoid is the nearest
  - For each pair of a non-medoid object *o'* and a medoid *o*, compute the total cost, *S*, of swapping medoid *o* with *o'*
  - If the lowest cost $S_{lowest} < 0$ then swap *o* with o_{lowest}' to form the new set of *k* medoids

30

Typical k-medoids algorithm (PAM)



31

How to Choose the new Medoid for a Cluster

- Use the squared-error criterion

$$E_o = \sum_{p \in C_i} d(p,o)^2$$

where o is a candidate medoid for the ith cluster C_i and $d(p, o)$ is the distance between o and point p in C_i.

- Choose point o in C_i that minimizes E_o

32

Pros and Cons of PAM

- PAM is more robust than k-means in the presence of noise and outliers
  - Medoids are less influenced by outliers
- PAM is efficient for small data sets but does not scale well for large data sets
  - $O(k(n-k)^2)$ for each iteration
- Sampling based method: CLARA

Outline

- What is Clustering?
- Types of Data in Cluster Analysis and Similarity Measures
- Clustering Methods
  - K-means
  - K-medoids
  - <u>Hierarchical clustering method</u>
  - DBSCAN: a Density-based Algorithm
- Cluster Validity Measures

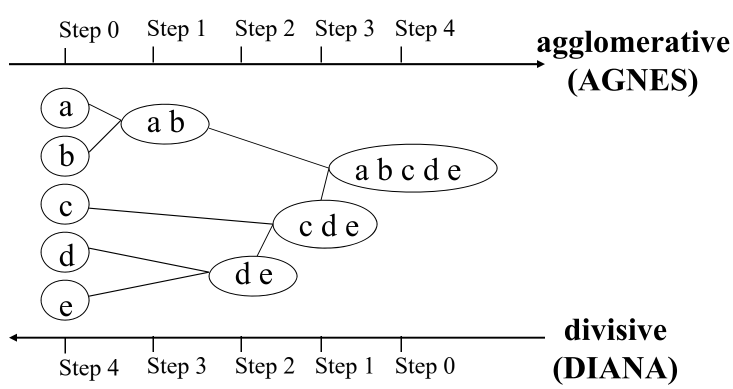
Hierarchical Clustering

- Iteratively merge or split clusters to form a tree of clusters
- Two types
  - Agglomerative (bottom-up): merge clusters iteratively
    - Start by placing each object in its own cluster
    - Merge these small clusters into larger and larger clusters
    - until all objects are in a single cluster
    - Most hierarchical methods belong to this category. They differ only in their definition of between-cluster similarity
  - Divisive (top-down): split a cluster iteratively
    - Start with all objects in one cluster and subdivide them into smaller pieces

35

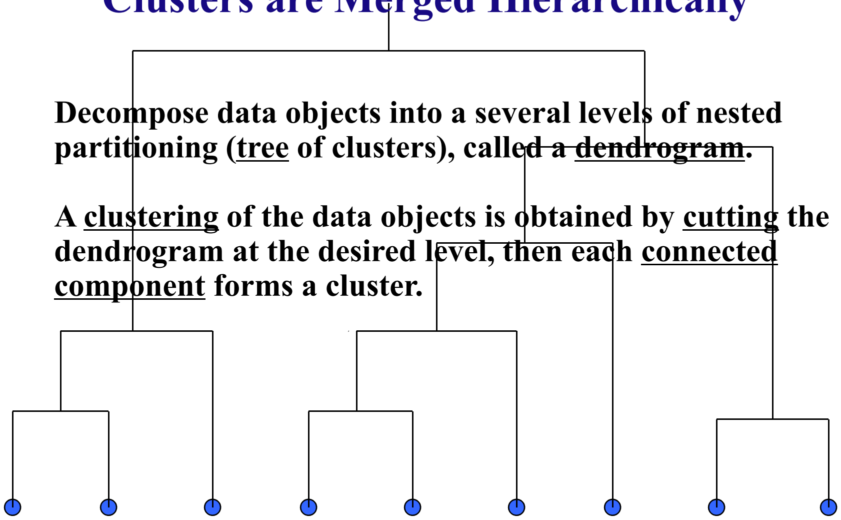
Hierarchical Clustering

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



36

A *Dendrogram* Shows How the Clusters are Merged Hierarchically



Decompose data objects into a several levels of nested partitioning (<u>tree</u> of clusters), called a <u>dendrogram</u>.

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

37

Inter-cluster Distances in Hierarchical Clustering

- Three widely used ways to define the distance between two separate clusters:
  - Single linkage method (nearest neighbor):

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} \{d(x, y)\}$$

  - Complete linkage method (furthest neighbor):

$$d(C_i, C_j) = \max_{x \in C_i, y \in C_j} \{d(x, y)\}$$

  - Average linkage method (unweighted pair-group average):

$$d(C_i, C_j) = \operatorname{avg}_{x \in C_i, y \in C_j} \{d(x, y)\}$$

38

Strengths and Limitations of Hierarchical Methods

- Conceptually simple
- Theoretical properties are well understood
- When clusters are merged/split, the decision is permanent
  - The number of different alternatives that need to be examined is reduced.
  - Erroneous decision are impossible to correct later
- Do not scale well:
  - Time complexity for agglomerative clustering is at least $O(n^2)$, where n is the number of total objects

39