Clustering

Clustering is a process of partitioning a set of objects (data) into a set of meaningful sub-classes called **clusters**.

- Cluster: a collection of objects that are similar to each other and thus can be treated collectively as one group

Difference between Classification and Clustering

- Data classification: the name and number of classes are given in the training set
- Clustering: the name and number of classes are unknown

What Is Good Clustering?

A good clustering method will produce high quality clusters where:

- the ***intra****-class similarity* (that is within a cluster) is high.
- the ***inter****-class similarity* (that is between clusters) is low.

The quality of a clustering result also depends on the similarity measure used by the method.

The quality of a clustering result also depends on the definition and representation of cluster – different clustering algorithms may have different underlying notions of clusters.

Major Clustering Techniques

Partitioning algorithms: Construct various partitions and then evaluate them by some criterion.

Hierarchy algorithms: Create a hierarchical decomposition of the set of data (or objects) using some criterion. There is an agglomerative approach and a divisive approach.

Density-based: based on connectivity and density functions.

Grid-based: based on a multiple-level granularity structure.

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.

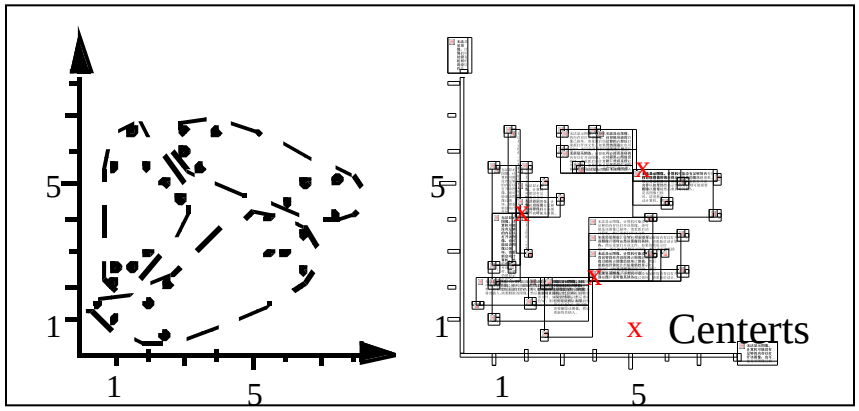
Partitioning Algorithms: Basic Concept

Partitioning method: Given a number ***k***, partition a database ***D*** of ***n*** objects into a set of ***k*** clusters so that a chosen objective function is minimized (e.g., sum of distances to the center of the clusters).

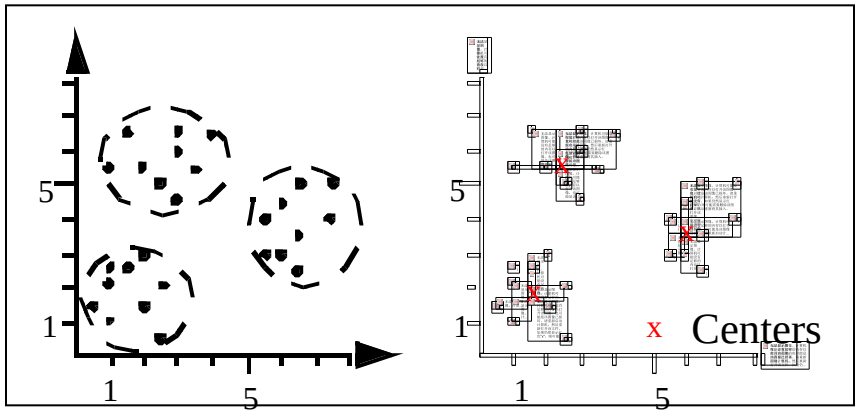
Global optimum: exhaustively enumerate all partitions – too expensive!

Heuristic methods based on iterative refinement of an initial partition

Bad Clustering



Optimal Clustering

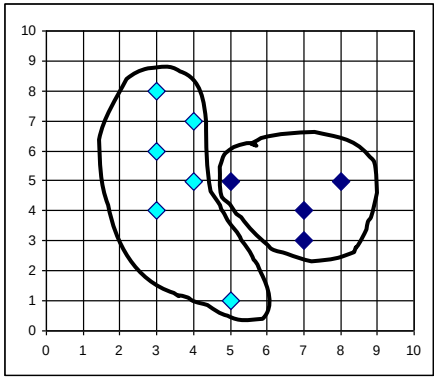
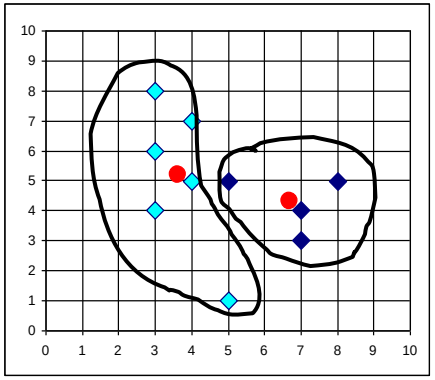
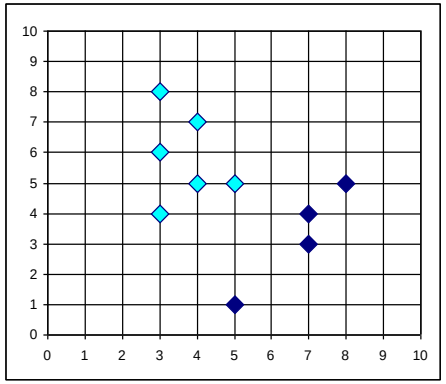
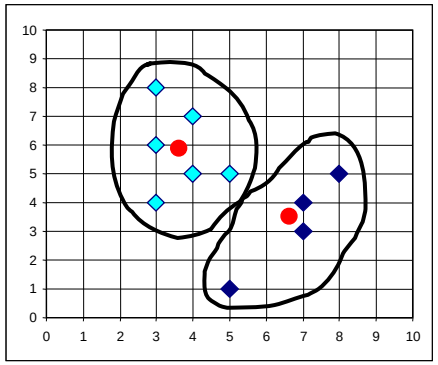


- Distance function
  - To measure similarity between two data objects
    - It is application dependent
  - For example: a distance between two points in a two dimensional plane
- Summarized representation of clusters
  - Mean (center): the average value
  - Radius: the average distance between all the objects in the cluster and the center

The K-Means Clustering Method

Given *k*, the *k-means* algorithm is implemented in 4 steps:

1. Partition objects into *k* nonempty subsets
2. Compute centers of the clusters of the current partition. The center a cluster for the *k*-means algorithm is the mean point of all points in the cluster.
3. Assign each object to the cluster with the nearest center.
4. Go back to Step 2, stop when no more new assignment.

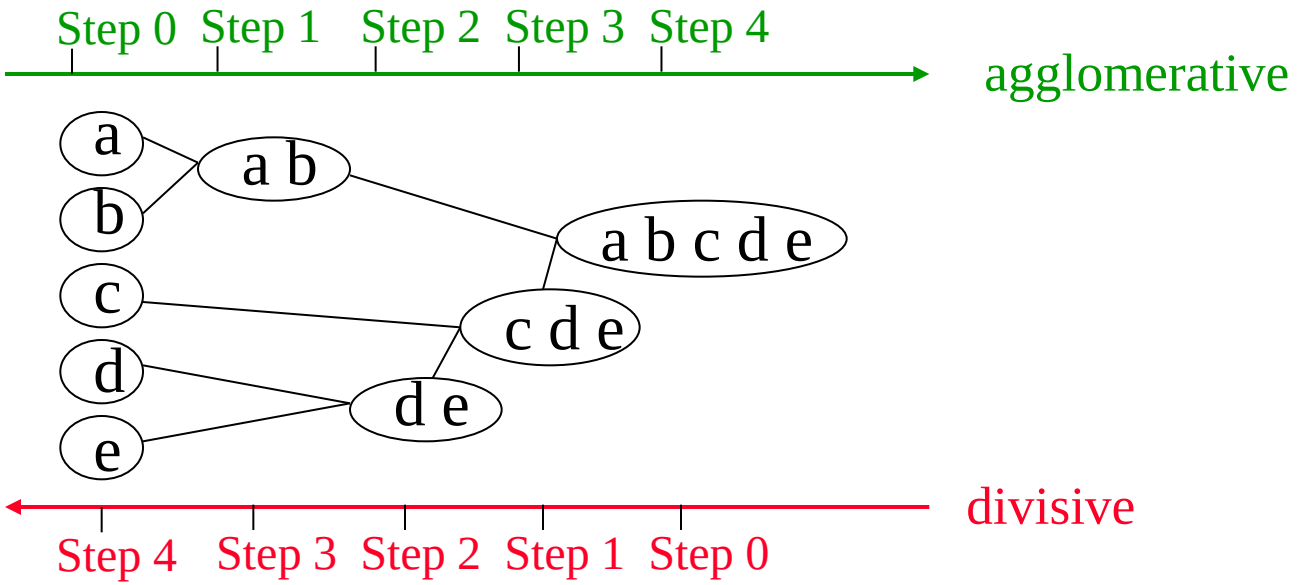





Hierarchical Clustering

Hierarchical decomposition of the data set (with respect to a given similarity measure) into a set of nested clusters

Result represented by a so called dendrogram

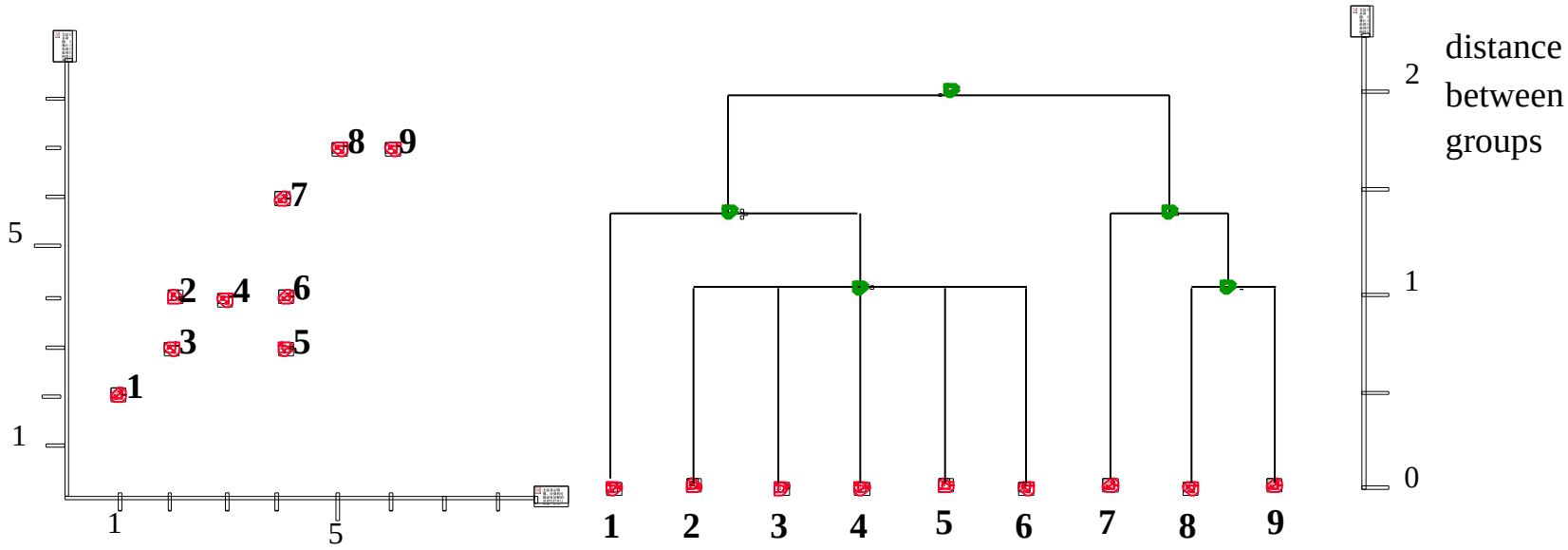
- Nodes in the dendrogram represent possible clusters
- can be constructed bottom-up (agglomerative approach) or top down (divisive approach)



Hierarchical Clustering: Example

Interpretation of the dendrogram

- The root represents the whole data set
- A leaf represents a single objects in the data set
- An internal node represent the union of all objects in its sub-tree
- The height of an internal node represents the distance/similarity between its two child nodes



Agglomerative Hierarchical Clustering

Single-Link Method and Variants:

start by placing each object in its own cluster.

keep merging "closest pairs" (most similar pairs) of clusters into larger clusters

until all objects are in a single cluster.

Most hierarchical methods belong to this category. They differ mainly in their definition of between-cluster similarity.

Single-Link: similarity is defined as the similarity between the "closest" (i.e., most similar) pair of objects.