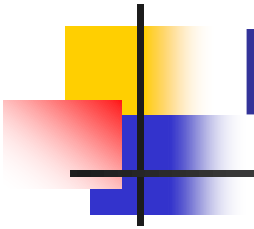
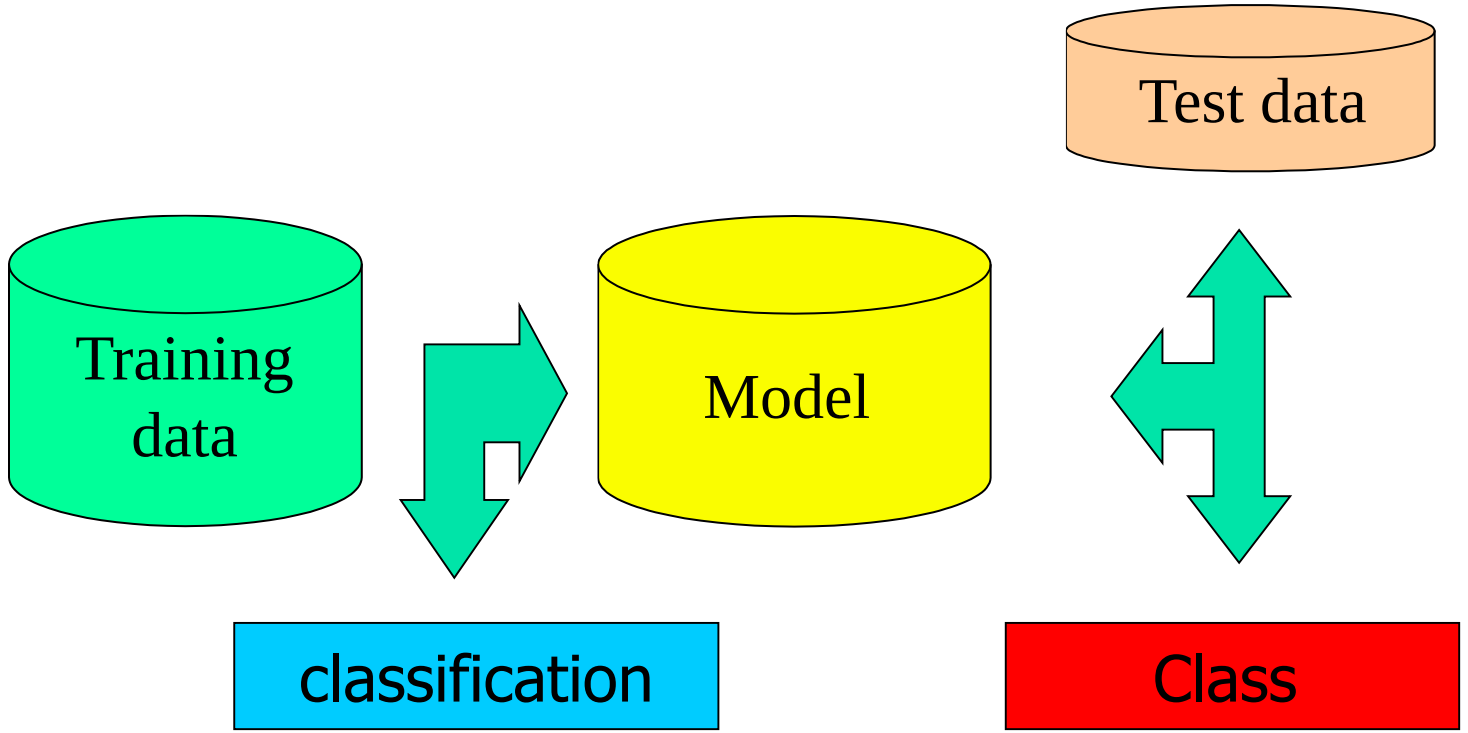
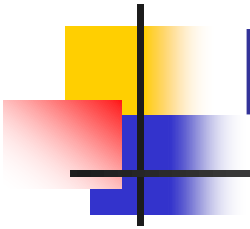
The Classification Problem in Data-Mining



- We are given a collection of records, these records call *the training set.*
  - Each record contains a set of *attributes* and the *class* that it belongs to.
- We are asked to find a *model* that describes the records of each class as a function of the values of their attributes.
- The goal is to use this model to classify new records for which we don't know the class in which they belong to.

Flow Chart



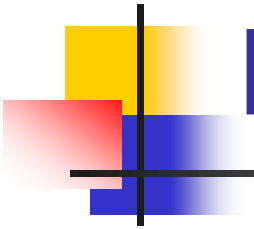


Record Structure

- Each record consists of a set of attributes
  - Categorical attributes
    - Each takes a value from a finite discrete set
      - Ex. attribute color takes a value form the set { red, green, blue }
  - Continuous attributes
    - Each takes a value from a continuous interval
      - Ex. attribute salary takes a value in the range [20,000, 15000]
- Each record contains a class label
  - Its characteristics are similar to those of the categorical attributes.

Sample Training Data

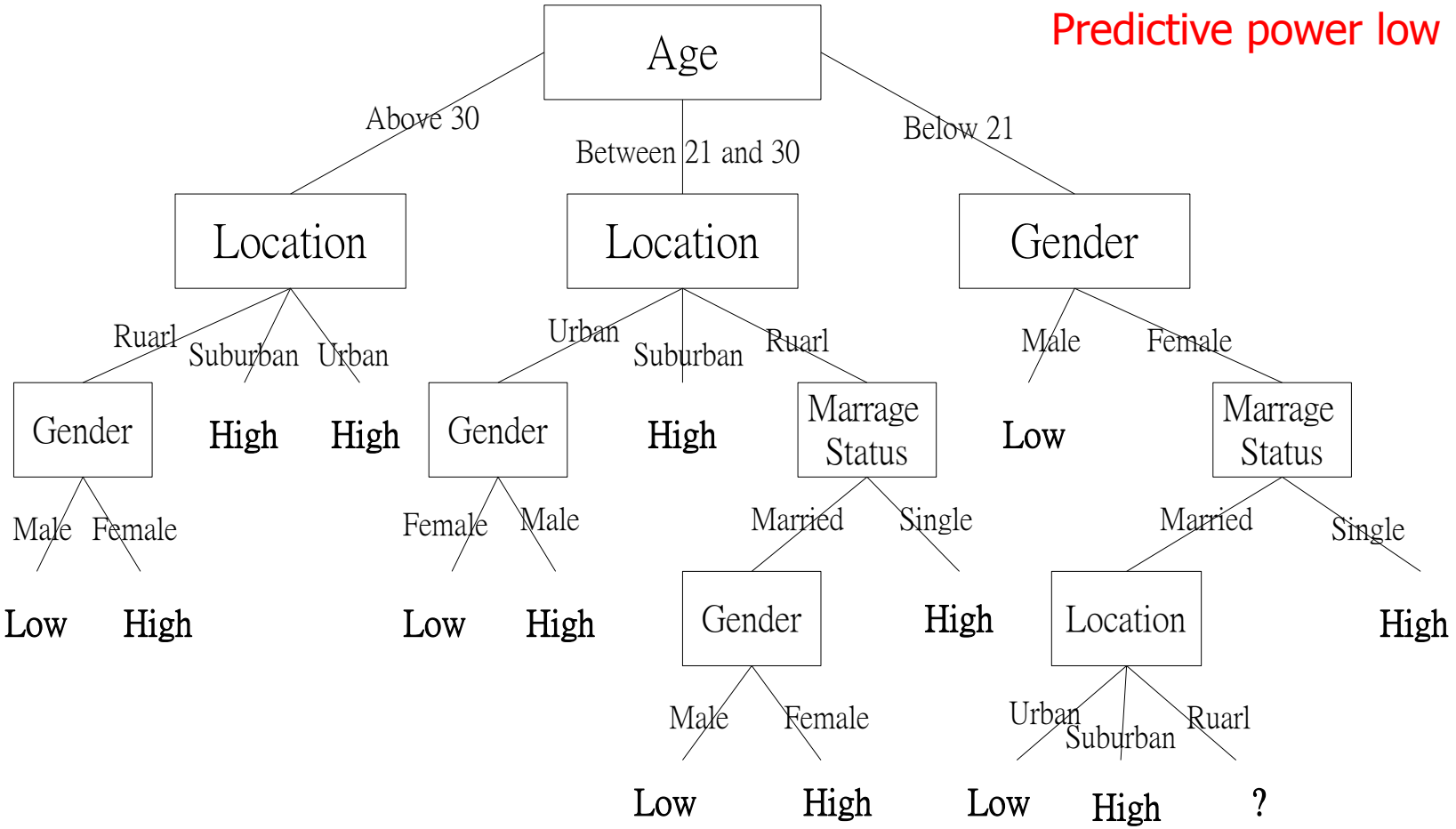
No	Attributes				Class
	Location	Age	Marriage status	Gender	Low
1	Urban	Below 21	Married	Female	Low
2	Urban	Below 21	Married	Male	Low
3	Suburban	Below 21	Married	Female	High
4	Rural	Between 21 and 30	Married	Female	High
5	Rural	Above 30	Single	Female	High
6	Rural	Above 30	Single	Male	Low
7	Suburban	Above 30	Single	Male	High
8	Urban	Between 21 and 30	Married	Female	Low
9	Urban	Above 30	Single	Female	High
10	Rural	Between 21 and 30	Single	Female	High
11	Urban	Between 21 and 30	Single	Male	High
12	Suburban	Between 21 and 30	Married	Male	High
13	Suburban	Below 21	Single	Female	High
14	Rural	Between 21 and 30	Married	Male	Low



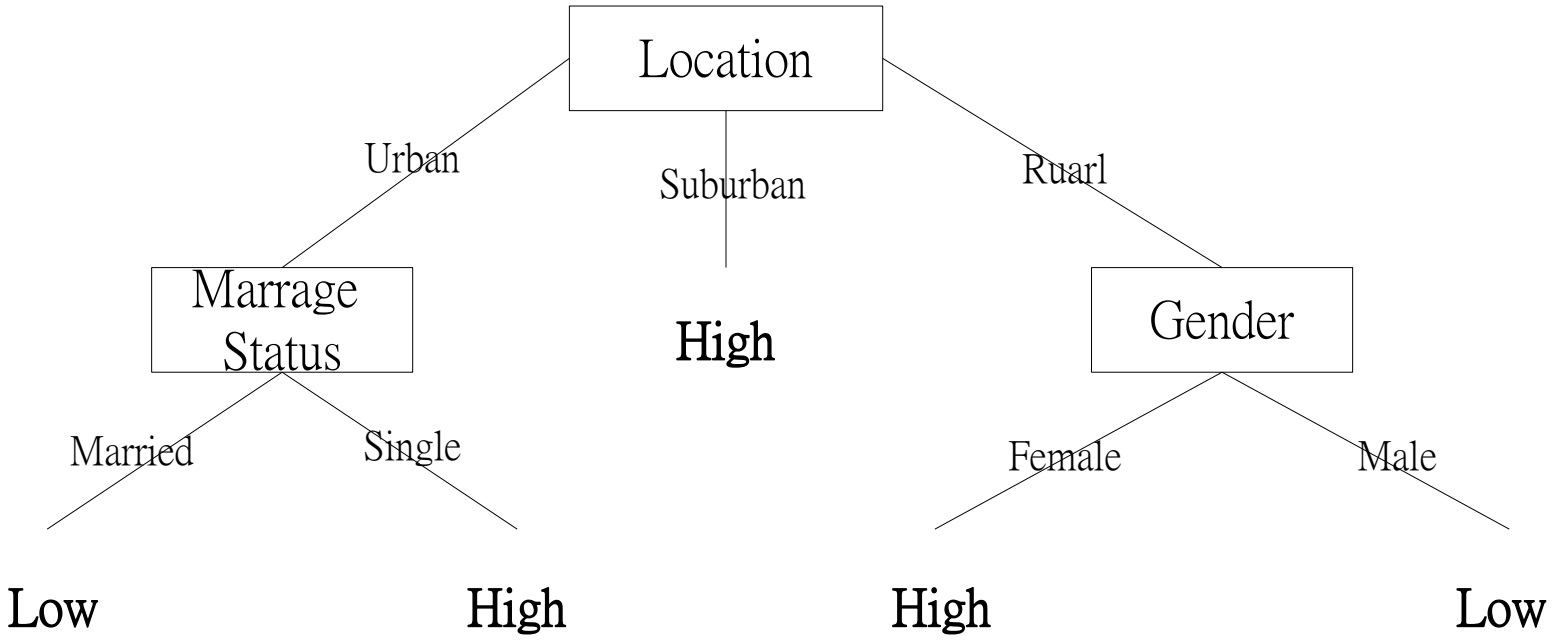
Decision tree

- A decision tree is a special type of classifier.
  - Each leaf carries a class name.
  - Each internal node specifies an attribute.
  - top-down irrevocable strategy.
  - divide-and-conquer.
- The primary problems remain finding good split-points.

A Complex Decision Tree



A Compact Decision Tree



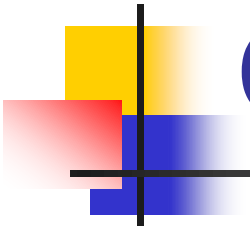
Its predictive power is often higher than that of a complex decision tree.

The Clustering Problem in Data-Mining

- In clustering analysis, there is no pre-classified data. Instead, clustering analysis is a process where by a set of objects is partitioned into several clusters in which all members in one cluster are similar to each other and different from the members of other clusters, according to some similarity metric.

Problem Formulation

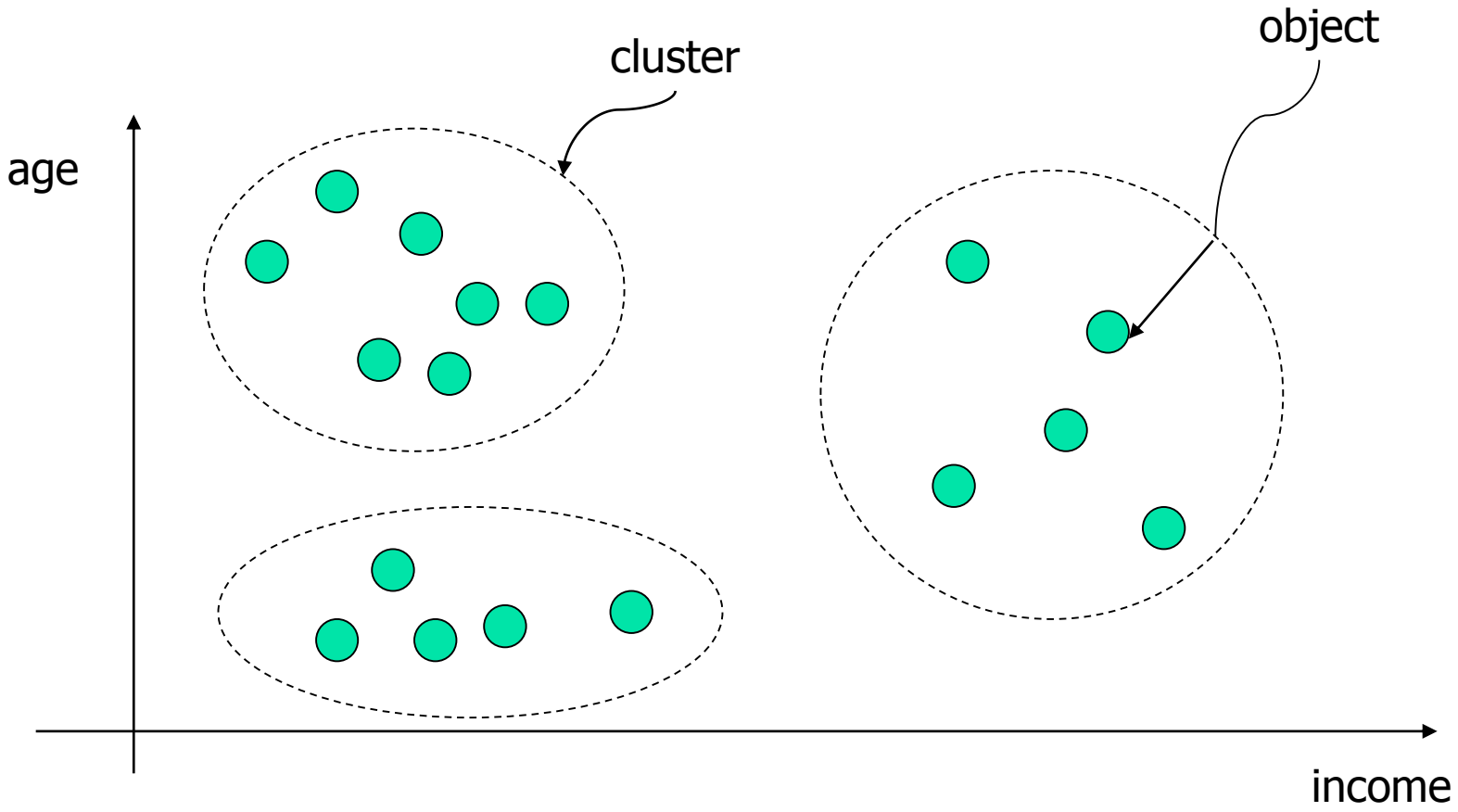
- Let O = $\{o_1, o_2, \ldots, o_n\}$ be a set of n objects and let C = $\{C_1, \ldots, C_k\}$ be a *partition* of O into subsets; i.e. $C_i \cap C_j = \emptyset$,i≠j and $\cup_k C_k$ = O. Each subset is called a cluster, and C is a clustering solution.
- Given an input, find a partition C of a set of objects O such that the resulting clusters are homogeneous and well separated.



Objective of Cluster Analysis

- This implies that objects belonging to the same cluster be as similar to each other as possible, while objects belonging to different clusters to be as dissimilar as possible.

Example



Dissimilarity Measures

- Manhattan Distance Measure
  - Suppose the data set contains p attributes. The Manhattan distance measure defines the dissimilarity d(i, j) between objects i and j as:

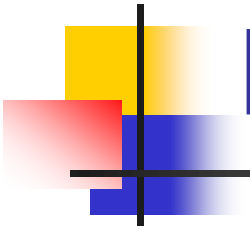
$$d(i, j) = \frac{|x_{i1} - x_{j1}|}{R_1} + \frac{|x_{i2} - x_{j2}|}{R_2} + \ldots + \frac{|x_{ip} - x_{jp}|}{R_p}$$

Example of Manhattan

Object	Age(20~70)	Income (20000~120000)
1	20	30000
2	30	50000

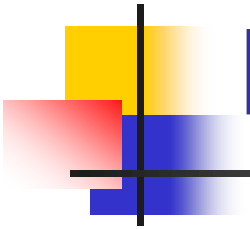
$$d(1,2) = \frac{|20-30|}{50} + \frac{|30000-50000|}{100000}$$

$$= \frac{1}{5} + \frac{1}{5} = 0.4$$



K-means

- Suppose that n objects described by the attribute vectors $\{x_1, x_2, \ldots, x_n\}$ be partitioned into k clusters, where k<n. Let m_i be the mean of the vectors in cluster i. That is, object y is in cluster i if the distance between y and m_i is the minimum.



K-means

- Algorithm of K-means
  - Randomly initialize the means m_1, m_2, …, m_k

  Repeat

  Use the means to classify all objects into clusters

  For i=1 to k

  Replace m_i with the mean of all objects in cluster i

  End-for

  Until there is no change in any mean

Example

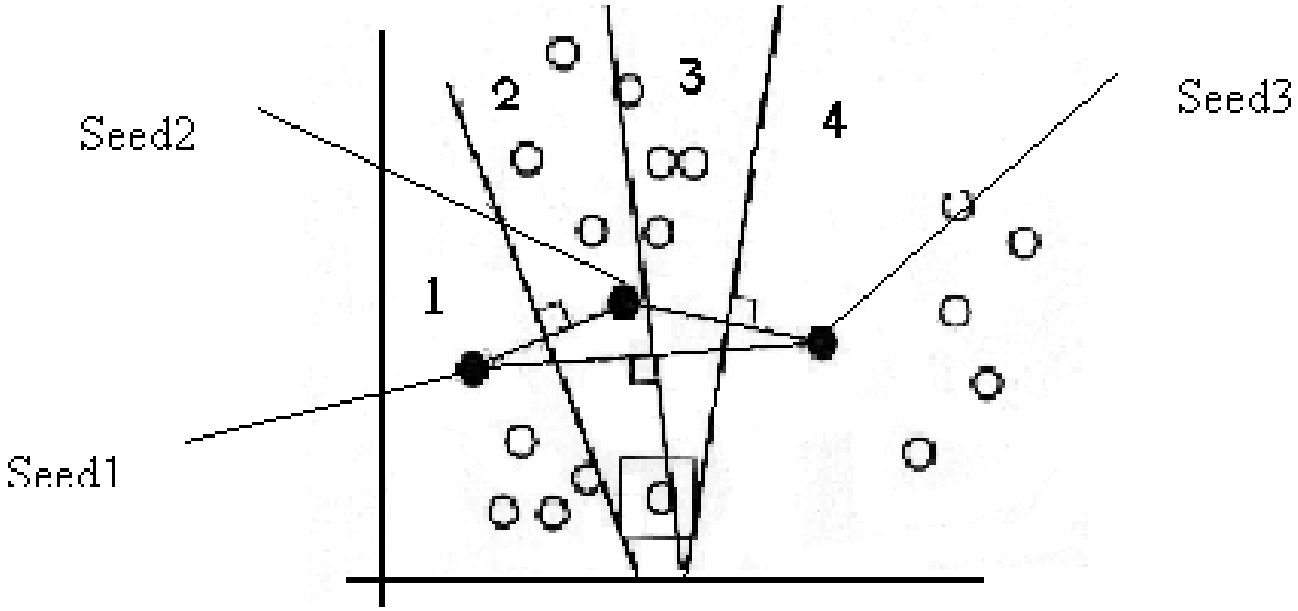
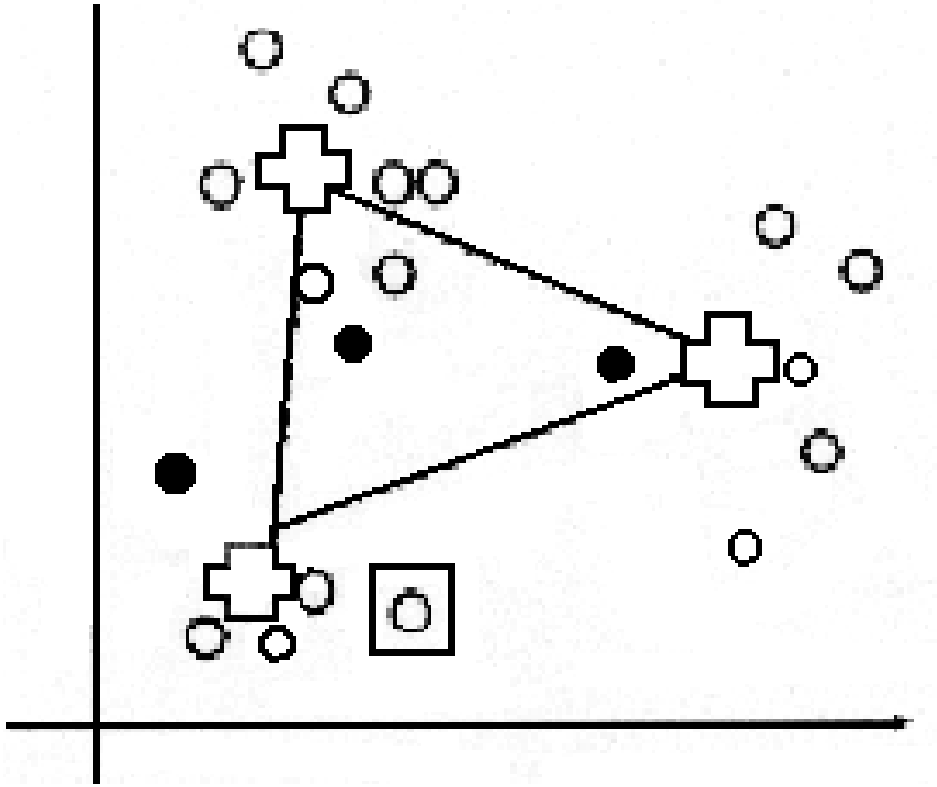


圖6-2

Example



Example

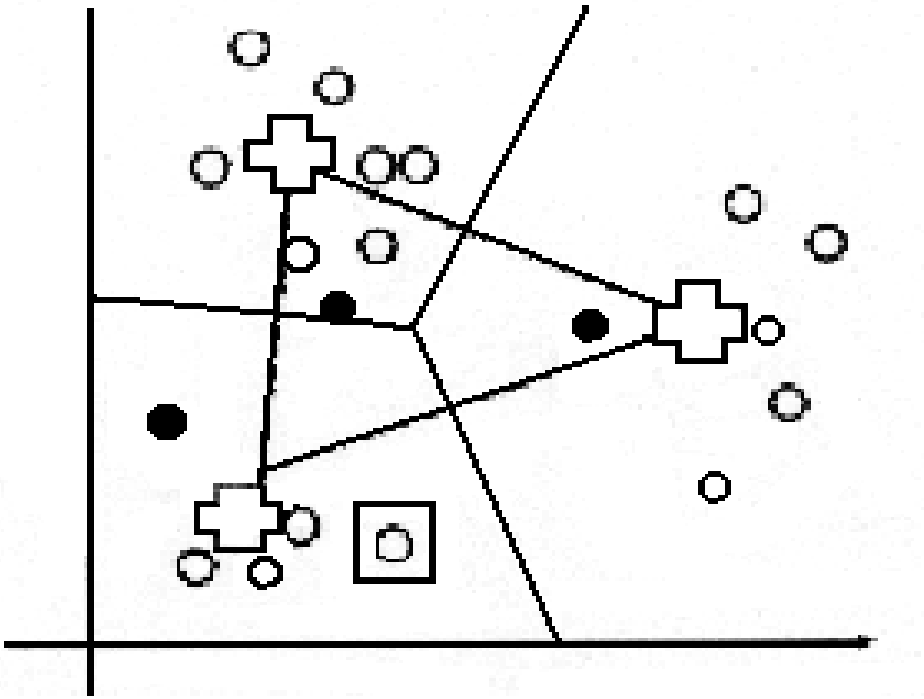


圖6-4