

80206085244011 (40/2)

# Pattern Recognition & Machine Learning : Clustering



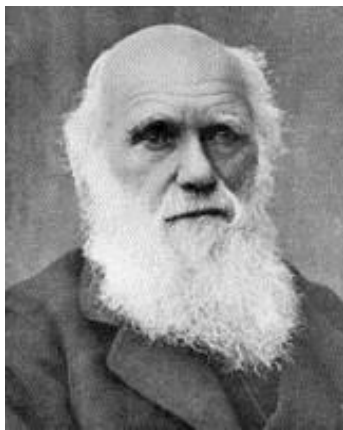
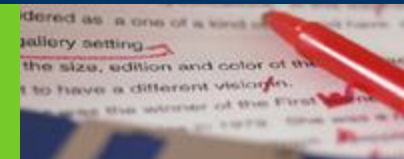
**Yong Wang**

**University of Chinese Academy of Sciences**

**2018.04.15**



# 聚类分析



大自然是一有机会就要说谎的。

—— Charles Robert Darwin

英国剑桥大学生物学家 (1809 – 1882)



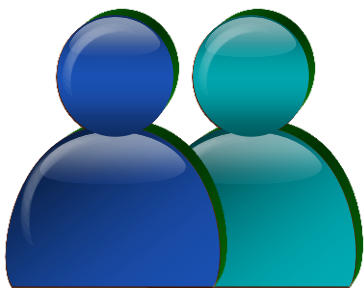
谬误无处不在，无孔不入。没有一种方法是万无一失的。

—— Charles Jules Henry Nicolle

法国细菌学家 (1866 – 1936)

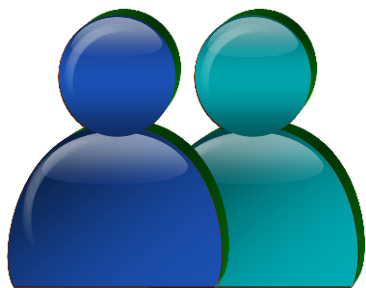
# 聚类分析

- 1 聚类分析基本概念
- 2 聚类分析中的数据类型
- 3 聚类分析的主要方法
- 4 离群点分析方法



# 聚类分析

- 1 聚类分析基本概念
- 2 聚类分析中的数据类型
- 3 聚类分析的主要方法
- 4 离群点分析方法



# 什么是聚类

## ■ 肤纹

皮肤纹理简称“肤纹”，包含指纹、掌纹、足纹等，是灵长目动物特有的、外露的生物学特征。

- 早在1000多年前，人们就认识到，每个人的肤纹与生俱来、终生不变。肤纹在个人是各不相同、终身稳定，在民族群体间也有很大差异，但在同一民族群体的肤纹却相对稳定。



- 经过30年不懈努力，我国上百家研究单位、千余名研究者共同参与的一项肤纹研究项目最新成果，证实了中华56个民族的肤纹特征有很强的民族杂合性、各少数民族互相间的肤纹基因有渊源且影响至今。应用**肤纹聚类分析统计法**，研究人员将中华56个民族梳理成南方和北方两大民族群，找到了民族肤纹的标志性群体，并明确了民族主支和支系的关系等。

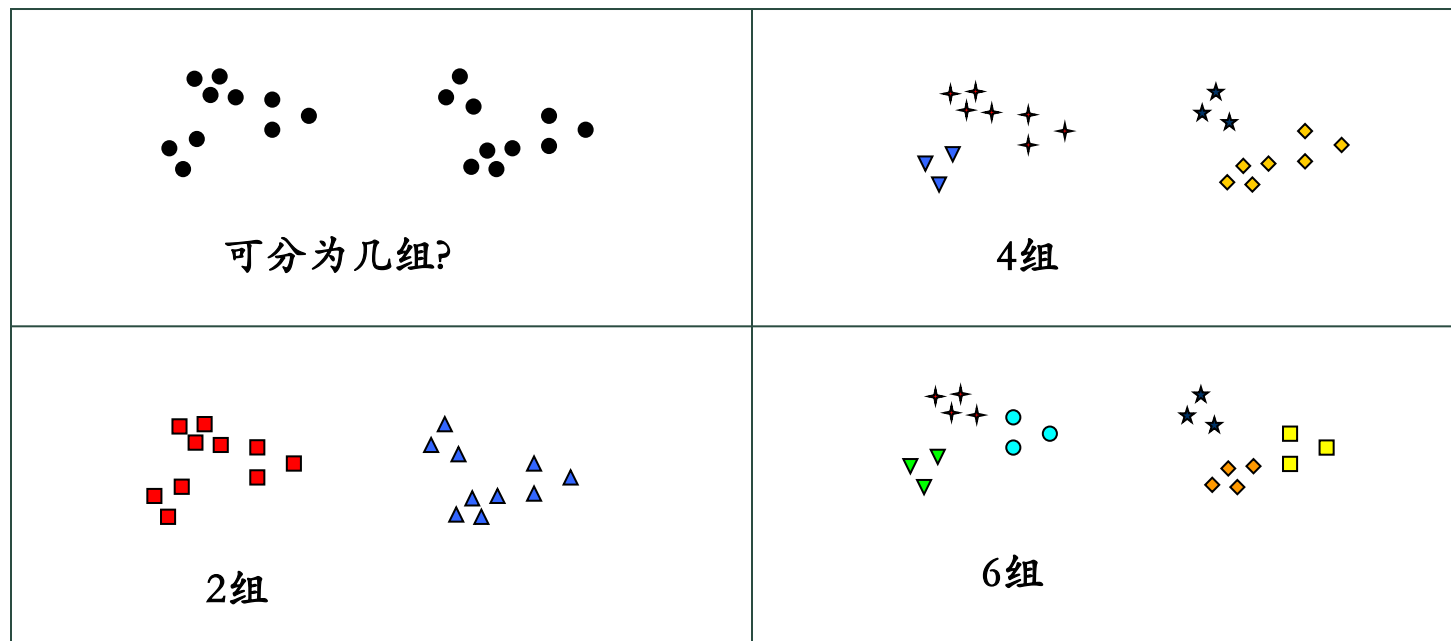


# 什么是聚类

- 将物理或抽象对象的集合划分成相似的对象类的过程称为聚类

—— — Jiawei Han *Data Mining: Concepts and Techniques*

- 聚类分析时不考虑数据类标号，而是通过聚类产生新类标号（自动分类）
- 对象根据最大化内部的相似性（similarity）、最小化类之间的相似性的原则进行聚类或分组（簇，cluster）



# 为什么聚类

## ■ 增进对事务现象的理解

- 商务智能    ➤ 生物学研究
- 模式识别    ➤ 空间数据分析
- 图像处理    ➤ Web文档分析
- .....

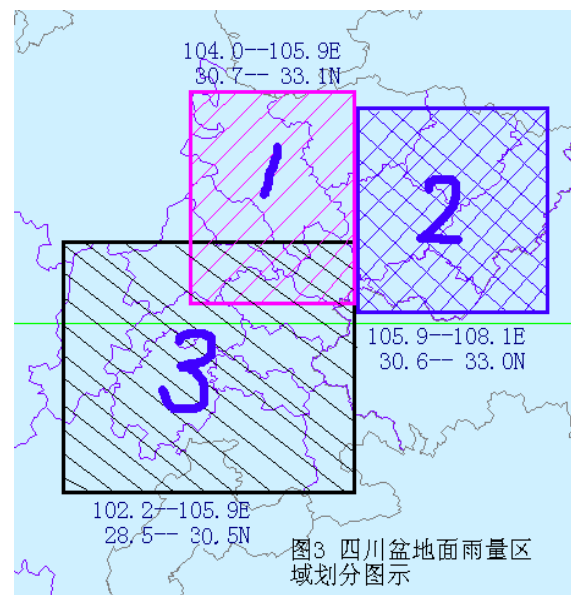
## ■ 对数据进行概括

降低大型数据库规模

## ■ 四川盆地降雨量分析

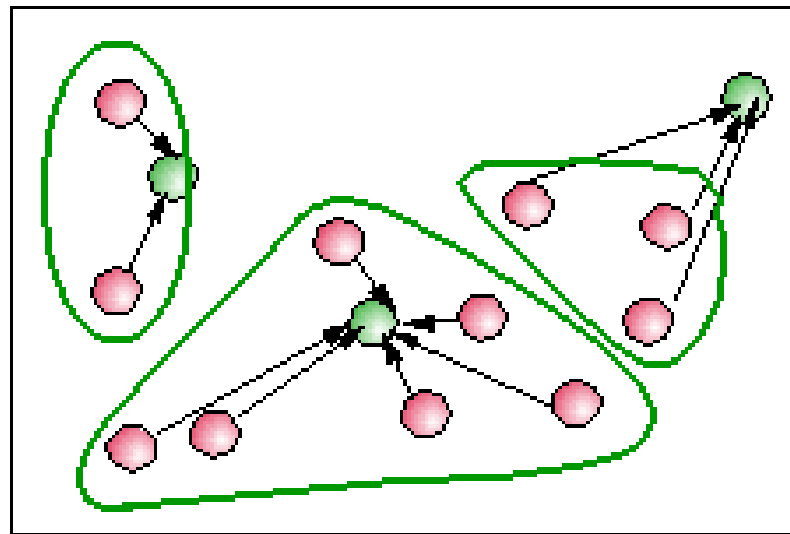


	Discovered Clusters	Industry Group
1	Applied-Matl-DOWN,Bay-Network-DOWN,3-COM-DOWN,Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN,DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN,Micron-Tech-DOWN,Texas-Inst-DOWN,Tellabs-Inc-DOWN,Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN,Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN,ADV-Micro-Device-DOWN,Andrew-Corp-DOWN,Computer-Assoc-DOWN,Circuit-City-DOWN,Compaq-DOWN,EMC-Corp-DOWN,Gen-Inst-DOWN,Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN,MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP,Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP,Schlumberger-UP	Oil-UP



# 数据挖掘对聚类的典型要求

- 可伸缩性 ( Scalability )
  - 可处理百万级数据对象
- 处理不同类型属性的能力
  - 数值、名词、序列、图、文档、混合
- 发现任意形状的聚类
  - 发现除球状簇之外的任意形状簇
- 对于决定输入参数的领域知识需求最小
  - 聚类结果对于输入参数十分敏感
- 处理带噪声数据的能力
- 聚类高维数据的能力
- 增量聚类和对输入记录的次序不敏感

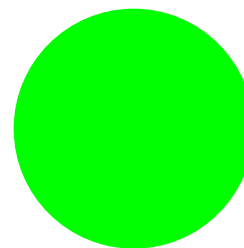
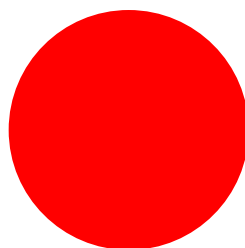
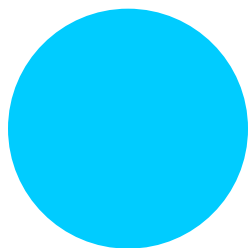


- 基于约束的聚类
  - 对每簇数据的类型和数量等施加约束
- 可解释性和可用性
  - 应用目标影响聚类特征和方法的选择

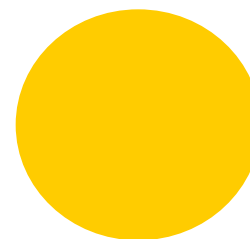
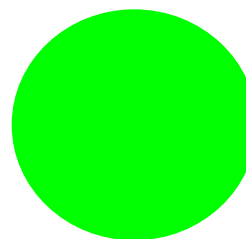
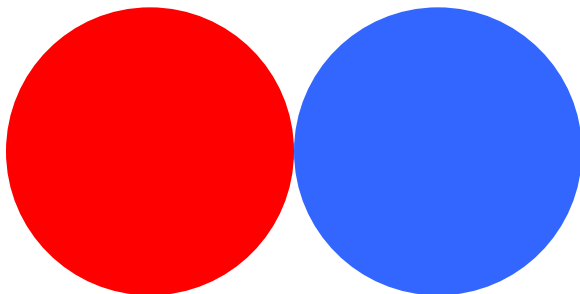


# 发现任意形状的聚类 (1)

## ■ 簇的类型 ( Types of Clusters )



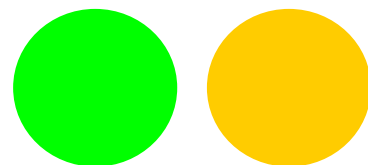
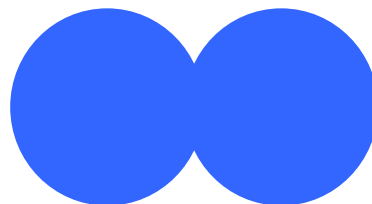
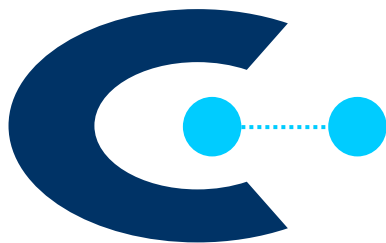
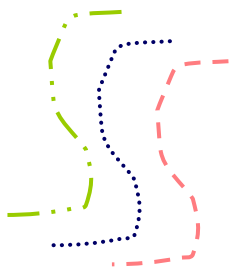
3 well-separated clusters



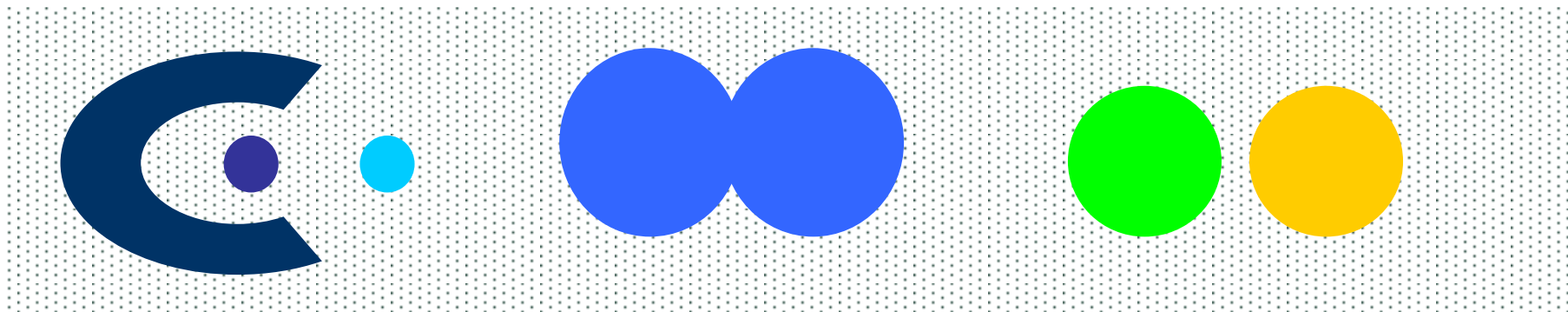
4 center-based clusters

# 发现任意形状的聚类 (2)

## ■ 簇的类型 ( Types of Clusters )



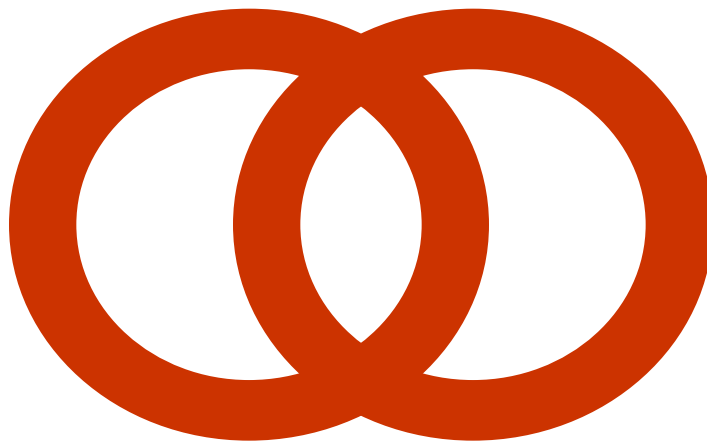
8 contiguous clusters



6 density-based clusters

# 发现任意形状的聚类 (3)

## ■ 簇的类型 ( Types of Clusters )



2 Overlapping Circles

- 根据目标函数定义的簇发现最小化或最大化目标函数的簇

# 什么不是聚类

- 分类

- 监督学习，有类标

- 简单分组

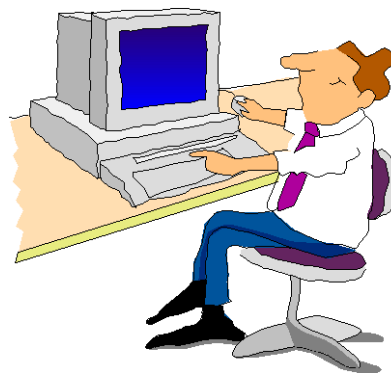
- 根据姓名的字母顺序将学生分为不同的学习兴趣小组

- 检索结果

- 由根据外部约束进行检索所得到的结果构成的集合

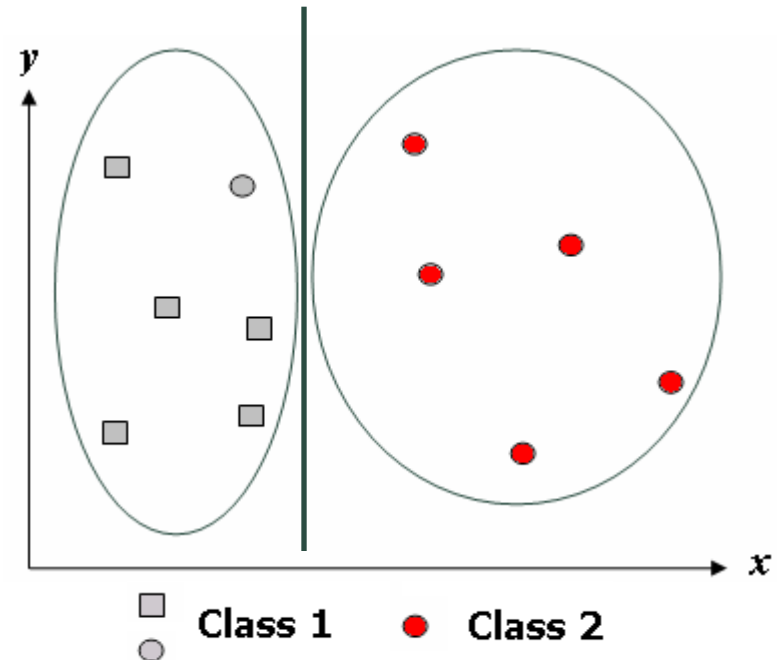
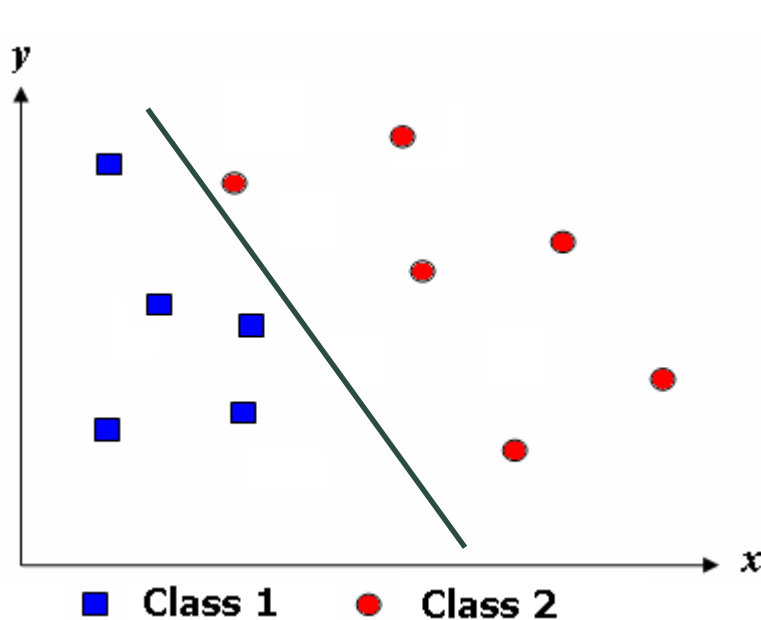
- 图分割

- 部分之间相互相关并且协作，但不完全一样

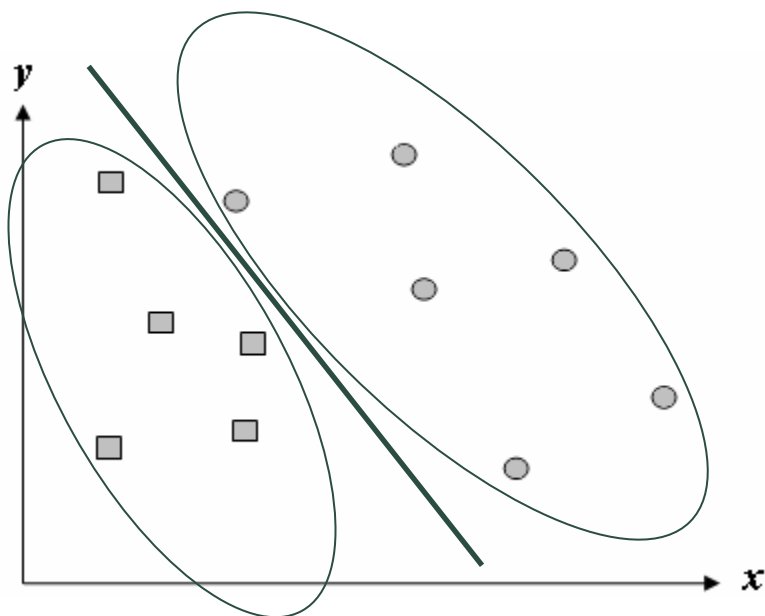


# 分类 VS. 聚类

- 分类需要高昂的代价收集和标记大量训练元组集或模式，以便分类法使用它们对每个组建模
- 聚类把数据集合划分成组（簇，cluster），然后给这些数量相对较少的组指定标号



# 聚类过程与结果的动态性



■ Class 1    ● Class 2

## ■ 与分类相比，聚类有如下一些特征

- （最优）分组数一般未知
- 可能没有关于聚类的任何先验知识
- 聚类结果是动态的

不同的相似性度量，不同的目的要求，产生的聚类结果是不一样的

## ■ 对顾客分组，促进童装销售

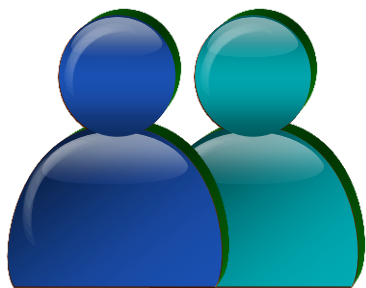
- 孩子数是最重要属性
- 年收入和受教育水平（相关）
- 年龄和婚姻状况（无关）

年收入/美元	年 龄	孩 子 数	婚姻状况	受教育程度
25 000	35	3	单身	高中
15 000	25	1	已婚	高中
20 000	40	0	单身	高中
30 000	20	0	离异	高中
20 000	25	3	离异	大专
70 000	60	0	已婚	大专
90 000	30	0	已婚	研究生
200 000	45	5	已婚	研究生
100 000	50	2	离异	大专



# 聚类分析

- 1 聚类分析基本概念
- 2 聚类分析中的数据类型
- 3 聚类分析的主要方法
- 4 离群点分析方法



# 聚类分析中的数据类型 (1)

## ■ 数据结构

### ■ 数据矩阵 (Data matrix)

对象—变量结构

用 $p$ 个变量表示 $n$ 个对象

**二模 (two-mode) 矩阵**

### ■ 相异度矩阵 (Dissimilarity matrix)

$d(i,j)$ 表示对象 $i$ 、 $j$ 之间的相异度

相异度非负，对象越接近，其值越小

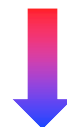
**单模 (one-mode) 矩阵**

### ■ 数据矩阵的行和列代表不同的实体

相异度矩阵的行和列代表相同的实体

**多数聚类算法都是对相异度矩阵运行**

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



$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

# 聚类分析中的数据类型 (2)

## ■ 数据类型



# 聚类分析中的数据类型 (3)

## ■ 区间标度变量 Interval-scaled variables

- 一种粗略线性标度的连续变量，例如重量、高度、经纬度、气温等
- 选用的度量单位将影响聚类分析的结果

### 度量标准化

#### (1) 计算均值绝对偏差 (mean absolute deviation)

$$s_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

$$m_f = \frac{1}{n}(x_{1f} + x_{2f} + \dots + x_{nf})$$

#### (2) 计算标准度量值 (standardized measurement, or z-score)

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

数据矩阵

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



$$x_{if} \Rightarrow z_{if}$$

- 均值绝对偏差比标准差对于离群点具有更好的鲁棒性？ 无差值平方
- 要根据问题决定是否和如何进行标准化 (例如给篮球运动员进行聚类)

# 聚类分析中的数据类型 (4)

## ■ 区间标度变量 Interval-scaled variables

### ■ 计算相异度

相异度矩阵

#### (1) 闵可夫斯基距离 ( *Minkowski distance* )

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

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{bmatrix}$$

#### (2) 曼哈顿 (或城市块) 距离 ( *Manhattan distance* )

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

#### (3) 欧几里得距离 ( *Euclidean distance* )

$$d(i, j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \dots + |x_{ip} - x_{jp}|^2)}$$

- $d(i, j) \geq 0$
- $d(i, i) = 0$
- $d(i, j) = d(j, i)$
- $d(i, j) \leq d(i, k) + d(k, j)$

#### (4) 加权的欧几里得距离 ( *Weighted Euclidean distance* )

$$d(i, j) = \sqrt{w_1 |x_{i1} - x_{j1}|^2 + w_2 |x_{i2} - x_{j2}|^2 + \dots + w_p |x_{ip} - x_{jp}|^2}$$

# 聚类分析中的数据类型 (5)

## ■ 二元变量 Binary variables

- 每个变量只有两种状态：0或1。0表示该变量不出现，1表示该变量出现，即  $x_{ij}=0$  或  $x_{ij}=1$
- 如果所有的二元变量具有相同的权重，则得到一个两行两列的相依表

数据矩阵

$x_{11}$	$\dots$	$x_{1f}$	$\dots$	$x_{1p}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_{i1}$	$\dots$	$x_{if}$	$\dots$	$x_{ip}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$x_{n1}$	$\dots$	$x_{nf}$	$\dots$	$x_{np}$

- $a$ 表示对象*i*和*j*值都为1的变量的数目
- $b$ 表示对象*i*值为1但对象*j*值为0的变量的数目
- $c$ 表示对象*i*值为0但对象*j*值为1的变量的数目
- $d$ 表示对象*i*和*j*值都为0的变量的数目

二元变量相依表

	Object <i>j</i>		<i>sum</i>
	1	0	
Object <i>i</i>	1	$a$ $b$	$a+b$
	0	$c$ $d$	$c+d$
<i>sum</i>	$a+c$	$b+d$	$p$



# 聚类分析中的数据类型 (6)

## ■ 二元变量 Binary variables

### ■ 计算相异度

#### (1) 对称二元相异度 ( *Dissimilarity of symmetric binary variables* )

一个二元变量是对称的，如果它的两个状态具有同等价值和相同的权重，例如，性别变量

$$d(i, j) = \frac{b+c}{a+b+c+d}$$

二元变量相依表

		Object <i>j</i>		
		1	0	<i>sum</i>
Object <i>i</i>	1	<i>a</i>	<i>b</i>	<i>a+b</i>
	0	<i>c</i>	<i>d</i>	<i>c+d</i>
<i>sum</i>		<i>a+c</i>	<i>b+d</i>	<i>p</i>

#### (2) 非对称二元相异度 ( *Dissimilarity of asymmetric binary variables* )

$$d(i, j) = \frac{b+c}{a+b+c}$$

通常将比较重要的输出结果，也是出现几率较小的结果编码为1

#### (3) 非对称二元相似度 ( *similarity of asymmetric binary variables* )

$$sim_{Jaccard}(i, j) = \frac{a}{a+b+c} = 1 - d(i, j)$$

# 聚类分析中的数据类型 (7)

## ■ 二元变量 Binary variables

例题

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

		<i>Marry</i>	
		1	0
<i>Jack</i>	1	$a=2$	$b=0$
	0	$c=1$	$d=3$

$$d(jack, mary) = \frac{0+1}{2+0+1} = 0.33$$

$$d(jack, jim) = \frac{1+1}{1+1+1} = 0.67$$

$$d(jim, mary) = \frac{1+2}{1+1+2} = 0.75$$

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	1	0	1	0	0	0
Mary	F	1	0	1	0	1	0
Jim	M	1	1	0	0	0	0

# 聚类分析中的数据类型 (8)

## ■ 分类 (标称) 变量 Categorical variables

- 二元变量的推广，每个变量可以有 $M$ 种状态，例如，颜色变量

对象标识符	test-1(分类的)	test-2(序列的)	test-3(比例标度的)
1	color-A	优秀	445
2	color-B	一般	22
3	color-C	好	164
4	color-A	优秀	1210

- 状态标识1~ $M$ 只是用来数据处理，并不代表任何特定的顺序

## ■ 计算相异度

$$d(i, j) = \frac{p - m}{p}$$

### (1) 简单计算

$m$ 是对象 $i$ 和 $j$ 取值相同状态的变量的数目， $p$ 是全部变量的数目

### (2) 复杂计算

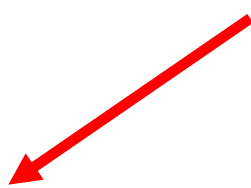
为 $M$ 个状态的每一个创建一个二元变量，计算其非对称二元变量相异度

# 聚类分析中的数据类型 (9)

## ■ 分类 (标称) 变量 Categorical variables

### 例题

对象标识符	test-1(分类的)	test-2(序列的)	test-3(比例标度的)
1	color-A	优秀	445
2	color-B	一般	22
3	color-C	好	164
4	color-A	优秀	1210


$$\begin{bmatrix} 0 \\ d(2,1) & 0 \\ d(3,1) & d(3,2) & 0 \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \xrightarrow{d(i,j) = \frac{p-m}{p}} \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

# 聚类分析中的数据类型 (10)

## ■ 序列变量 Ordinal variables

### ■ 离散序列变量 (discrete ordinal variable)

类似于分类变量，但 $M$ 个状态值是以有意义的序列排序，例如助理工程师、工程师、高级工程师等

数据矩阵

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

### ■ 连续序列变量 (continuous ordinal variable)

类似于一个刻度未知的连续数据的集合，值的相对次序是基本的，但其实际的大小并不重要，例如比赛名次

## ■ 计算相异度

(1) 将序列变量 $f$ 的 $M_f$ 个状态转化为一个秩评定，用相应的秩 $r_{if} \in \{1, \dots, M_f\}$ 代替 $x_{if}$

(2) 将每个变量的值域映射到 $[0.0, 1.0]$ 用 $z_{if}$ 代替 $r_{if}$

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

(3) 按照区间标度变量方式计算相异度

# 聚类分析中的数据类型 (11)

## 例题

对象标识符	test-1(分类的)	test-2(序列的)	test-3(比例标度的)
1	color-A	优秀	445
2	color-B	一般	22
3	color-C	好	164
4	color-A	优秀	1210

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \xrightarrow{d(i,j) = \sqrt{\sum_{k=1}^p |x_{ik} - x_{jk}|^2}} \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

对象标识符	test-1(分类的)	test-2(序列的)	test-3(比例标度的)
1	color-A	1	445
2	color-B	0	22
3	color-C	0.5	164
4	color-A	1	1210



# 聚类分析中的数据类型 (12)

## ■ 比例标度变量 Ratio-scaled variables

- 变量取值是在非线性的刻度上取正的度量值，例如指数刻度

$$x_{if} = Ae^{Bt} \quad \text{或} \quad x_{if} = Ae^{-Bt}$$

( $A$ 和 $B$ 是正的常数,  $t$ 通常表示时间)

数据矩阵

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

- 典型实例包括细菌群体的增长或放射性元素的衰变

## ■ 计算相异度

- (1) 采用与处理区间标度变量同样的方法来处理比例标度变量

会出现刻度扭曲，不宜采取！

- (2) 对比例标度变量进行对数变换，将变换后得到的变量当作区间标度变量

$$y_{if} = \log(x_{if})$$

- (3) 将比例变量看做连续的序列数据，将其秩作为区间标度变量来对待

# 聚类分析中的数据类型 (13)

## ■ 比例标度变量 Ratio-scaled variables

### 例题

对象标识符	test-1(分类的)	test-2(序列的)	test-3(比例标度的)
1	color-A	优秀	445
2	color-B	一般	22
3	color-C	好	164
4	color-A	优秀	1210

$$y_{if} = \log(x_{if})$$

对象标识符	test-1(分类的)	test-2(序列的)	test-3(比例标度的)
1	color-A	优秀	2.65
2	color-B	一般	1.34
3	color-C	好	2.21
4	color-A	优秀	3.08

$$\begin{bmatrix} 0 & & & \\ 1.31 & 0 & & \\ 0.44 & 0.87 & 0 & \\ 0.43 & 1.74 & 0.87 & 0 \end{bmatrix}$$

$$d(i, j) = \sqrt{\sum_{k=1}^p |x_{ik} - x_{jk}|^2}$$

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

# 聚类分析中的数据类型 (14)

## ■ 混合类型变量 Variables of mixed types

■ 如果  $f$  是区间标度变量：
$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$$
 $h$  遍取变量  $f$  的所有非缺失对象

■ 如果  $f$  是二元或分类变量：
$$d_{ij}^{(f)} = \begin{cases} 0 & x_{if} = x_{jf} \\ 1 & x_{if} \neq x_{jf} \end{cases}$$

■ 如果  $f$  是序列变量或比例标度变量：按单一变量类型的处理方法计算

数据矩阵

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

注意：基于区间的变量，其变量值被映射到区间  $[0.0, 1.0]$

(2) 将所有类型的变量一起处理，只进行一次聚类分析

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

$\delta_{ij}^{(f)}$  是指示项

如果  $x_{if}$  或  $x_{jf}$  缺失，或者  $x_{if} = x_{jf} = 0$ ，且变量  $f$  是非对称二元变量，则指示项等于 0；否则等于 1

# 聚类分析中的数据类型 (15)

## ■ 混合类型变量 Variables of mixed types

例题

对象标识符	test-1(分类的)	test-2(序列的)	test-3(比例标度的)
1	color-A	优秀	445
2	color-B	一般	22
3	color-C	好	164
4	color-A	优秀	1210

$$\begin{bmatrix} 0 & & & \\ \mathbf{1} & 0 & & \\ \mathbf{1} & \mathbf{1} & 0 & \\ \mathbf{0} & \mathbf{1} & \mathbf{1} & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ \mathbf{1} & 0 & & \\ \mathbf{0.5} & \mathbf{0.5} & 0 & \\ \mathbf{0} & \mathbf{1.0} & \mathbf{0.5} & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ \mathbf{1.31} & 0 & & \\ \mathbf{0.44} & \mathbf{0.87} & 0 & \\ \mathbf{0.43} & \mathbf{1.74} & \mathbf{0.87} & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ \mathbf{0.92} & 0 & & \\ \mathbf{0.58} & \mathbf{0.67} & 0 & \\ \mathbf{0.08} & \mathbf{1.00} & \mathbf{0.67} & 0 \end{bmatrix}$$

$$d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$$

除以 (3.08-1.34)=1.74

$$\begin{bmatrix} 0 & & & \\ \mathbf{0.75} & 0 & & \\ \mathbf{0.25} & \mathbf{0.50} & 0 & \\ \mathbf{0.25} & \mathbf{1.00} & \mathbf{0.50} & 0 \end{bmatrix}$$

# 聚类分析中的数据类型 (16)

## ■ 向量对象 Vector Objects

- 信息检索、文本文档聚类或生物分类中，所包含的有大量符号实体（如关键词和短语）的复杂对象

数据矩阵

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

## ■ 计算相异度

- 放弃传统的度量距离或属性匹配的计算方式

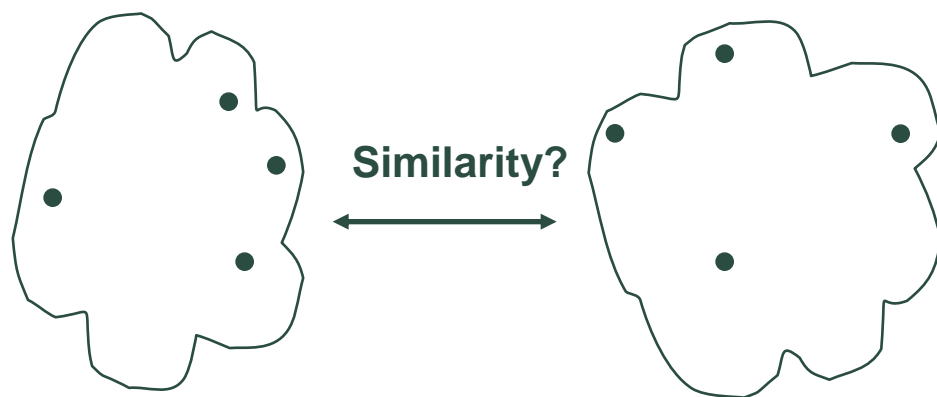
(1) 余弦度量 
$$s(\bar{X}, \bar{Y}) = \frac{\bar{X}^t \cdot \bar{Y}}{\|\bar{X}\| \|\bar{Y}\|}$$

(2) Tanimoto系数 
$$s(\bar{X}, \bar{Y}) = \frac{\bar{X}^t \cdot \bar{Y}}{\bar{X}^t \cdot \bar{X} + \bar{Y}^t \cdot \bar{Y} - \bar{X}^t \cdot \bar{Y}}$$

例题

$$\begin{aligned} \bar{X} &= (1, 1, 0, 0) \\ \bar{Y} &= (0, 1, 1, 0) \end{aligned} \quad \longrightarrow \quad s(\bar{X}, \bar{Y}) = \frac{\bar{X}^t \cdot \bar{Y}}{\|\bar{X}\| \|\bar{Y}\|} = \frac{(0 + 1 + 0 + 0)}{(\sqrt{2} \sqrt{2})} = 0.5$$

# 聚类分析中簇的距离 (1)



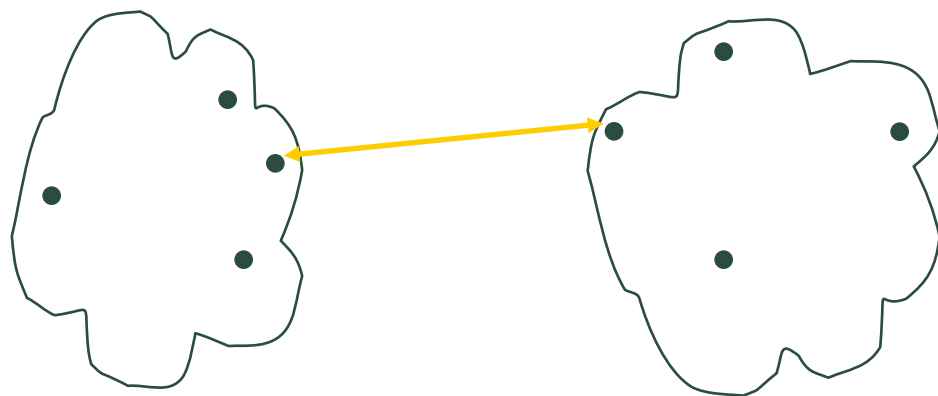
数据矩阵

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

- MIN Single link
- MAX Complete link
- Group Average
- Distance Between Centroid (质心)
- Distance Between Medoid (中心)



## 聚类分析中簇的距离 (2)



数据矩阵

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

- **MIN** Single link

- **MAX** Complete link

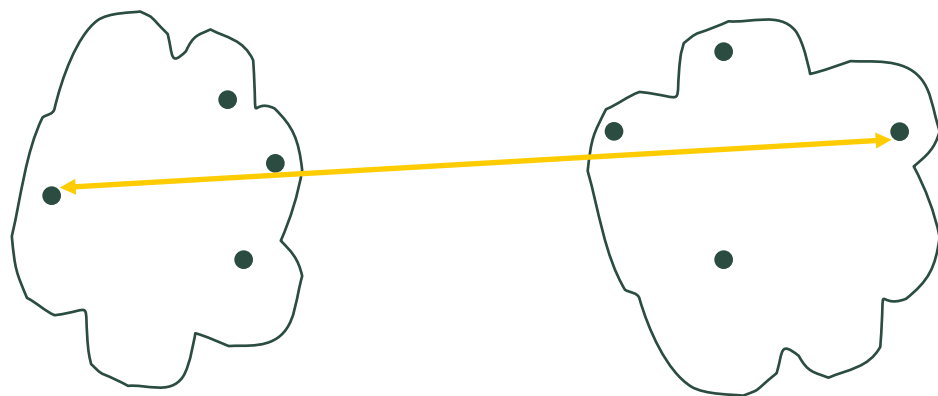
- **Group Average**

- **Distance Between Centroid (质心)**

- **Distance Between Medoid (中心)**

$$dis(K_i, K_j) = \min(x_{ip}, x_{jq})$$

# 聚类分析中簇的距离 (3)



数据矩阵

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

- MIN Single link

- MAX Complete link

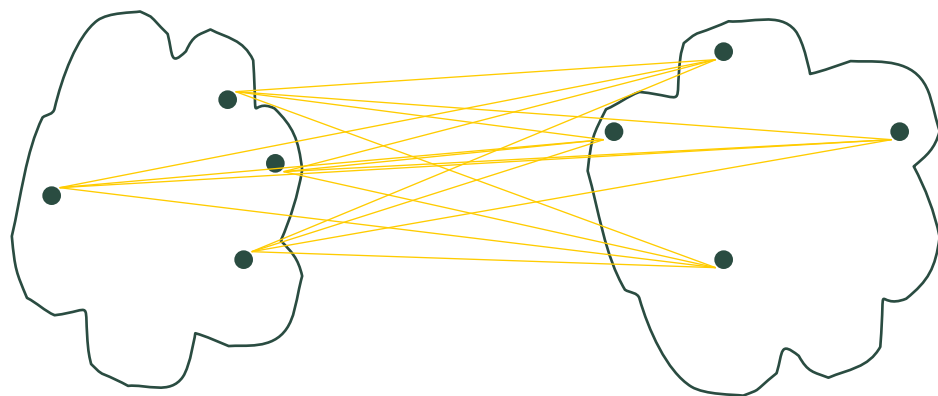
- Group Average

- Distance Between Centroid (质心)

- Distance Between Medoid (中心)

$$dis(K_i, K_j) = \max(x_{ip}, x_{jq})$$

# 聚类分析中簇的距离 (4)



数据矩阵

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

- MIN Single link

- MAX Complete link

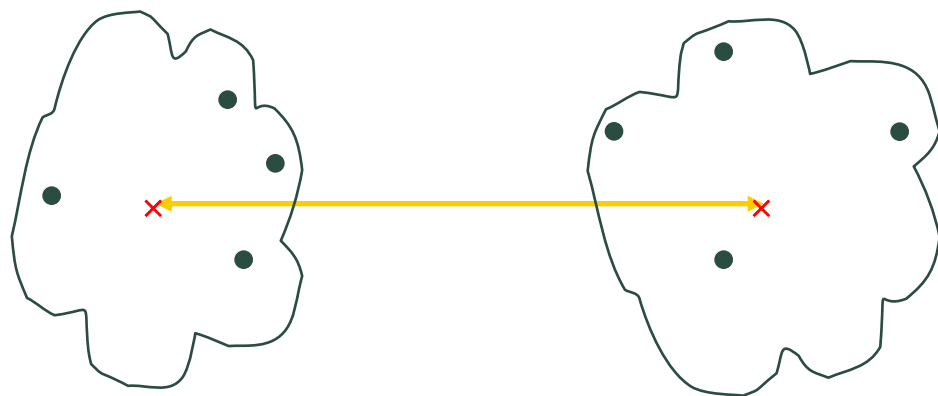
$$dis(K_i, K_j) = avg(x_{ip}, x_{jq})$$

- Group Average

- Distance Between Centroid (质心)

- Distance Between Medoid (中心)

# 聚类分析中簇的距离 (5)



数据矩阵

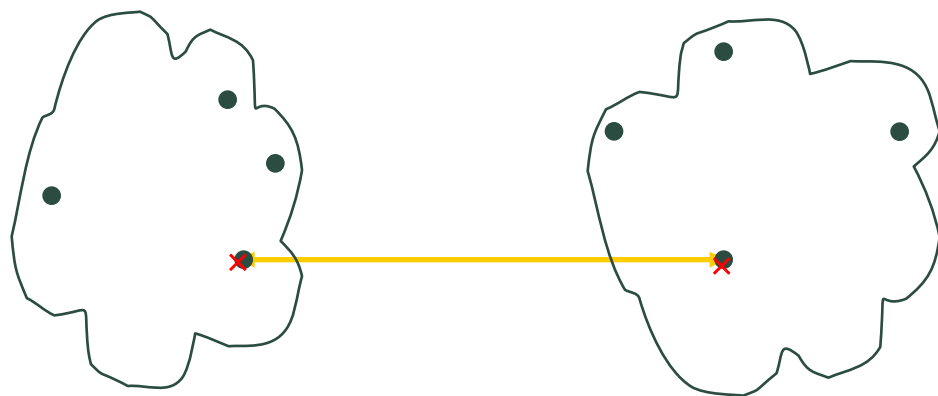
$$\begin{bmatrix} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{bmatrix}$$

- MIN Single link
- MAX Complete link
- Group Average
- Distance Between Centroid (质心)
- Distance Between Medoid (中心)

$$dis(K_i, K_j) = dis(C_i, C_j)$$

$$C_i = \frac{\sum_{k=1}^N x_{ki}}{N}$$

# 聚类分析中簇的距离 (6)



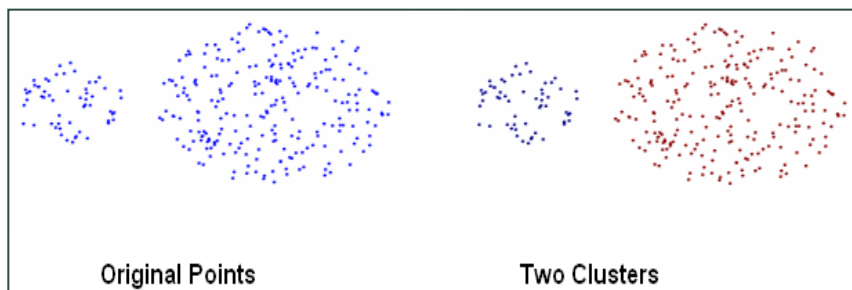
数据矩阵

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

- MIN Single link
- MAX Complete link
- Group Average
- Distance Between Centroid (质心)
- Distance Between Medoid (中心)

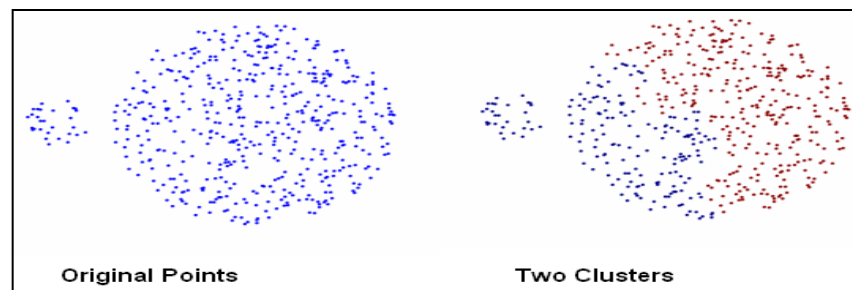
$$dis(K_i, K_j) = dis(M_i, M_j)$$

# 聚类分析中簇的距离 (7)



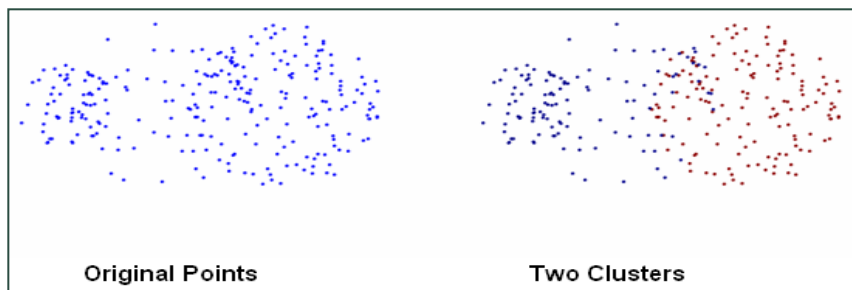
**MIN** Single link

**优点:** 可以发现球状簇以外的其它形状簇



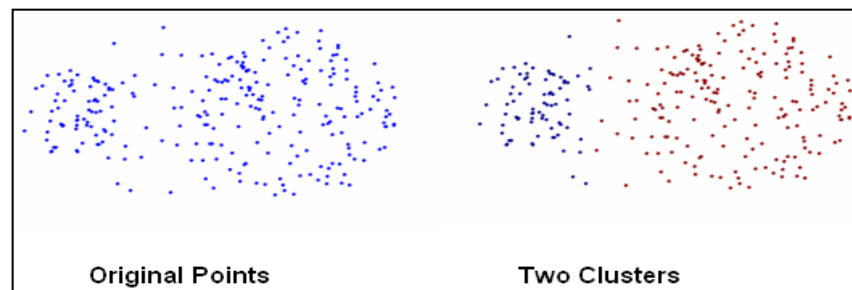
**MAX** Complete link

**缺点:** 容易分裂大的簇



**MIN** Single link

**缺点:** 对噪声和离群点敏感

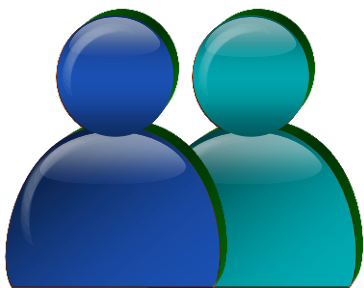


**MAX** Complete link

**优点:** 可以有效处理噪声和离群点干扰

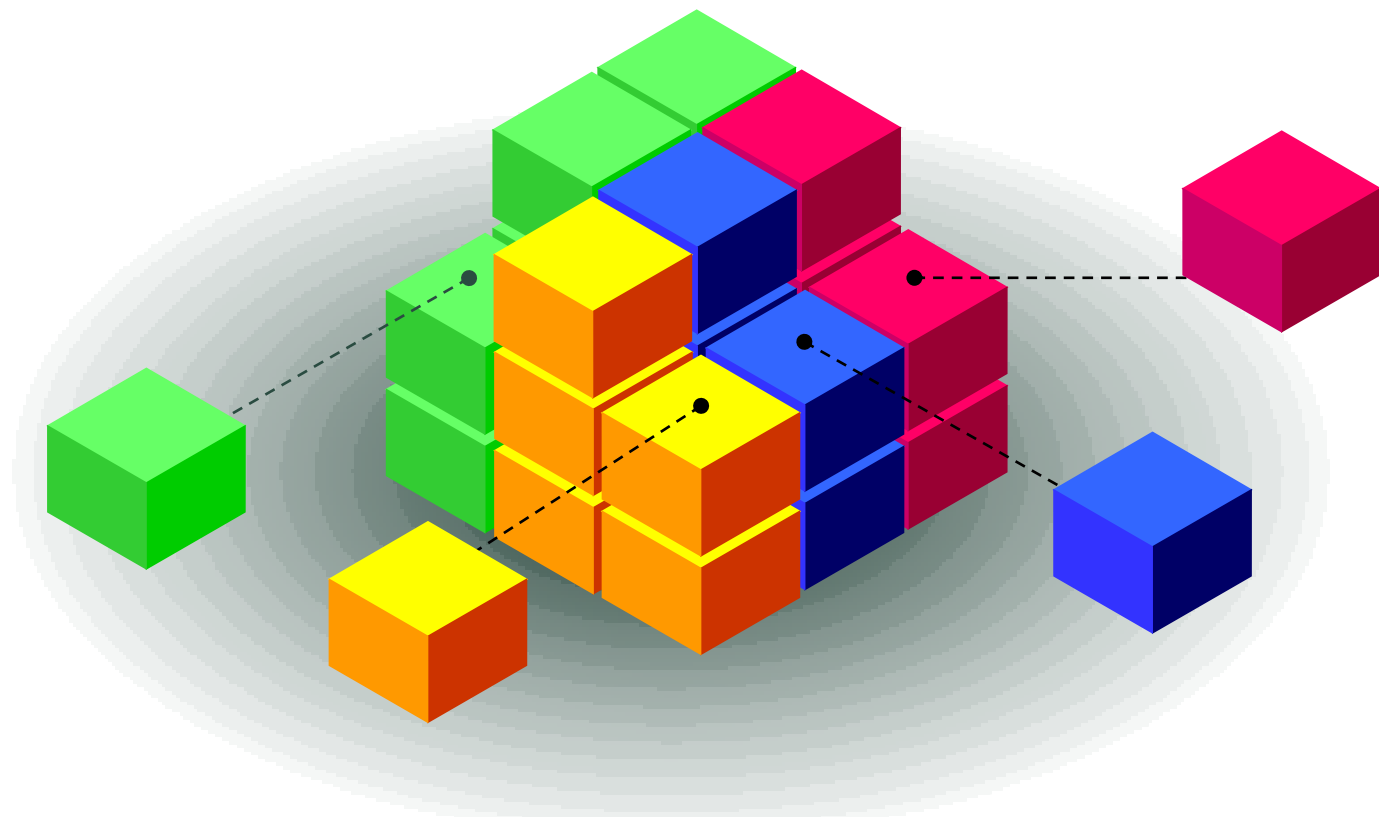
# 聚类分析

- 1 聚类分析基本概念
- 2 聚类分析中的数据类型
- 3 聚类分析的主要方法
- 4 离群点分析方法



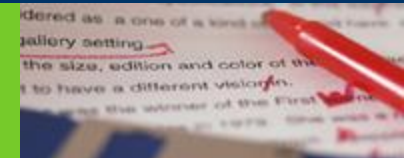


# 聚类分析的主要方法 (1)



聚类分析既是无监督学习又是观察式学习（不是示例式学习），其主要方法分为：划分方法、层次方法、基于密度的方法、基于网格的方法、基于模型的方法、高维数据的方法和基于约束的聚类等

# 聚类分析的主要方法 (2)



## ■ 划分方法 Partitioning methods

- 首先创建  $k$  个划分的初始集合，其中参数  $k$  是要构建的划分数目
- 然后采用**迭代重定位**技术，设法通过将对象从一个簇转移到另一个来改进划分的质量
- 典型的划分方法包括  $k$  均值， $k$  中心点和它们的改进算法

## ■ $k$ 均值法 $k$ -means algorithm

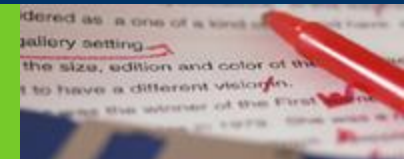
- 聚类结果是簇内的相似度高，而簇间的相似度低
- 簇的相似度是关于簇中对象的均值度量，可以看做是簇的质心 (Centroid) 或中心 (centre of gravity)
- 平方误差准则：

$$E = \sum_{i=1}^k \sum_{x \in C_i} |x - m_i|^2$$

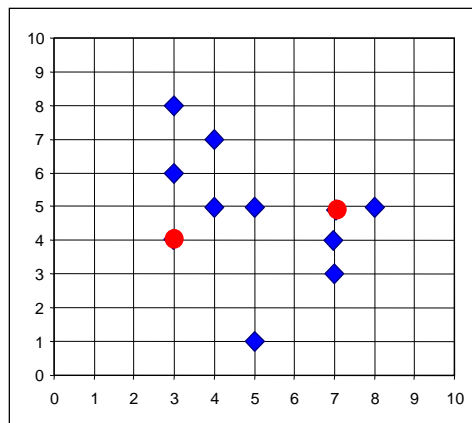
使生成的  $k$  个结果簇  
尽可能的紧凑和独立

$E$  是数据集中所有对象的平方误差和， $x$  表示对象， $m_i$  是簇  $C_i$  的均值

# 聚类分析的主要方法 (3)



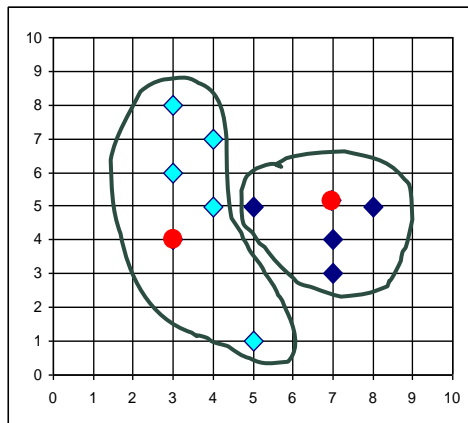
## 例题



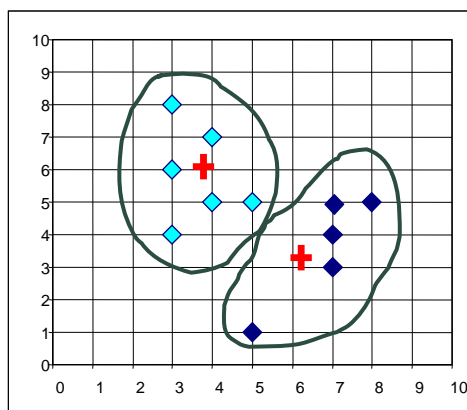
$K = 2$

任意选择  $k$  个对象当作初始簇中心

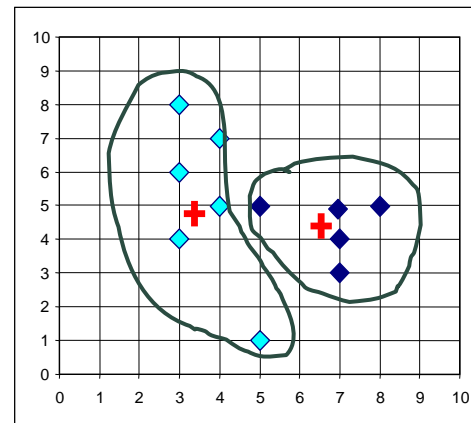
根据对象与簇中心的距离进行聚类



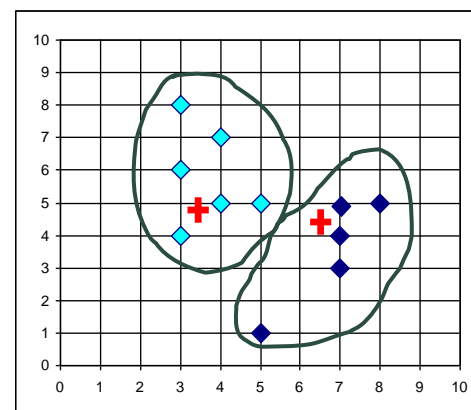
迭代重定位



更新簇中心

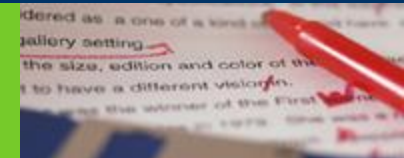


迭代重定位



更新簇中心

# 聚类分析的主要方法 (4)



## ■ $k$ 均值法 $k$ -means algorithm

算法： $k$  均值。用于划分的  $k$  均值算法，每个簇的中心用簇中对象的均值表示。

输入：

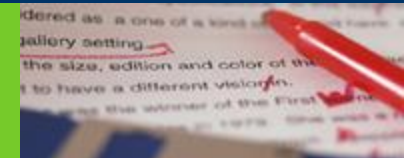
- $k$ ：簇的数目，
- $D$ ：包含  $n$  个对象的数据集。

输出： $k$  个簇的集合。

方法：

- (1) 从  $D$  中任意选择  $k$  个对象作为初始簇中心；
- (2) **repeat**
- (3)     根据簇中对象的均值，将每个对象（再）指派到最相似的簇；
- (4)     更新簇均值，即计算每个簇中对象的均值；
- (5) **until** 不再发生变换

# 聚类分析的主要方法 (5)

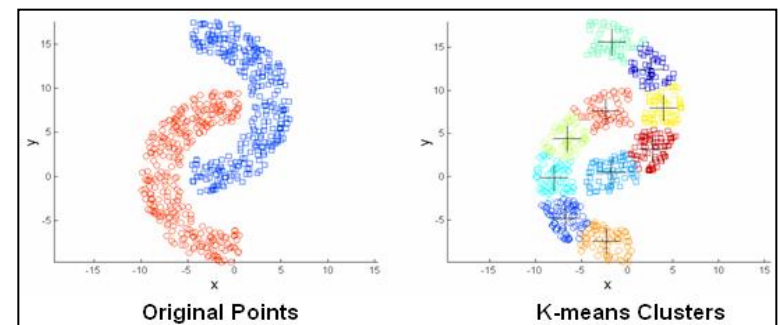
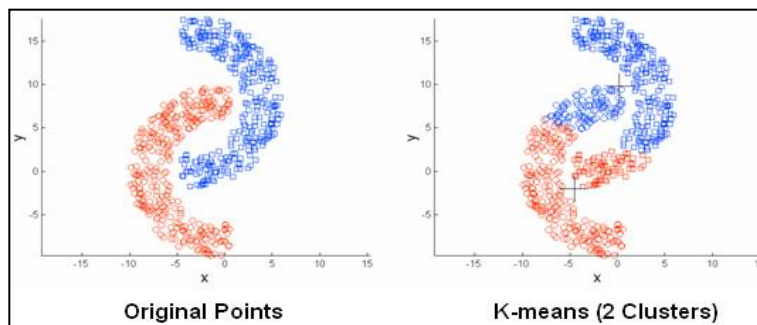
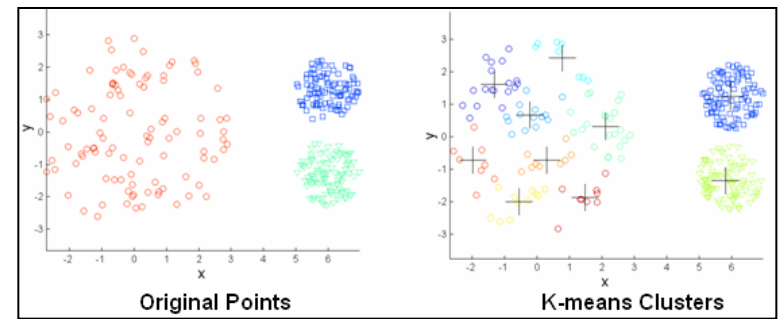
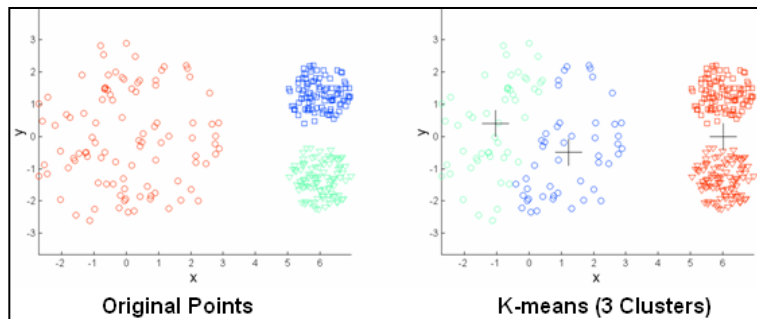
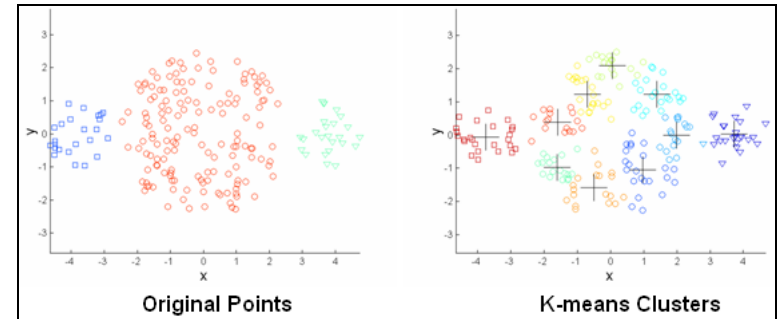
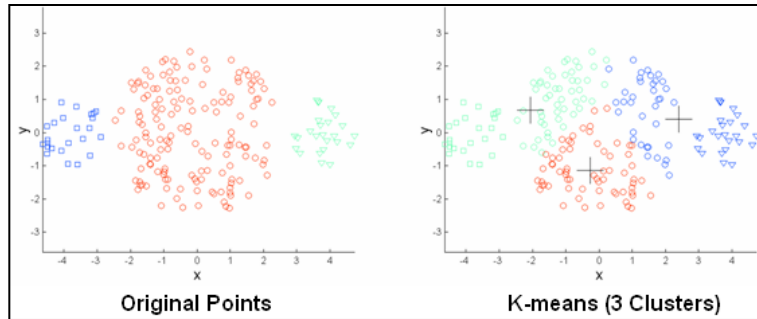


## ■ $k$ 均值法的总结

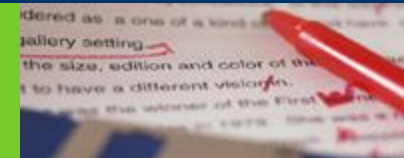
优点	<ul style="list-style-type: none"><li>■ 处理大数据集，该算法是相对可伸缩和有效率的 <math>O(nkt)</math> <math>n</math> 是对象的总数，<math>k</math> 是簇的个数，<math>t</math> 是迭代次数 <math>k \ll n, t \ll n</math></li></ul>
缺点	<ul style="list-style-type: none"><li>■ 当结果簇紧凑，且簇与簇之间明显分离时，它的效果较好</li><li>■ 该算法经常终止于局部最优解</li><li>■ 适用于簇均值有定义的情况，无法处理具有分类属性的数据</li><li>■ 对噪声和离群点数据敏感</li><li>■ 不适合于发现非凸形状的簇</li><li>■ 用户必须事先给出要生成的簇的数目 <math>k</math></li><li>■ 初始的簇中心的选择对聚类结果影响很大</li></ul>

# 聚类分析的主要方法 (6)

## 例题



# 聚类分析的主要方法 (7)



## ■ 其它划分方法:

### ■ $k$ 中心点法 $k$ -medoids algorithm

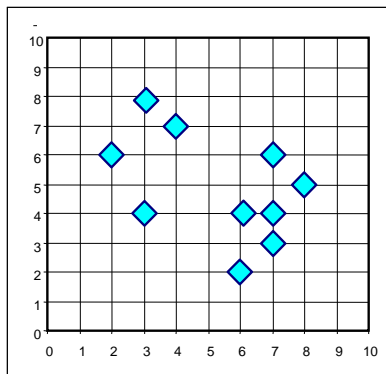
$k$ -means 方法对于脏数据很敏感, 改进的  $k$ -medoids 方法选取一个对象叫做 *mediod* 来代替  $k$  均值法中的中心的作用, 这样的 *medoid* 就标识了这个类

1. 任意选取  $k$  个对象作为 *medoids* ( $O_1, O_2, \dots, O_i, \dots, O_k$ ) ;
2. 将余下的对象分到各个类中去 (根据与 *medoid* 最相近的原则) ;
3. 对于每个类 ( $O_i$ ) 中, 顺序选取一个  $O_r$ , 计算用  $O_r$  代替  $O_i$  后的消耗— $E(O_r)$ , 选择  $E$  最小的那个  $O_r$  来代替  $O_i$ , 这样  $k$  个 *medoids* 就改变了, 下面就再转到2;
4. 这样循环直到  $k$  个 *medoids* 固定下来。

这种算法对于脏数据和异常数据不敏感, 但计算量显然要比  $k$  均值要大, 一般只适合小数据量  $O(k(n-k)^2)$

# 聚类分析的主要方法 (8)

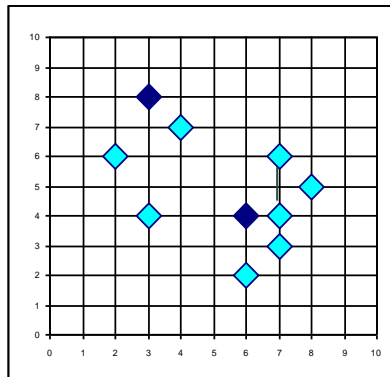
## 例题



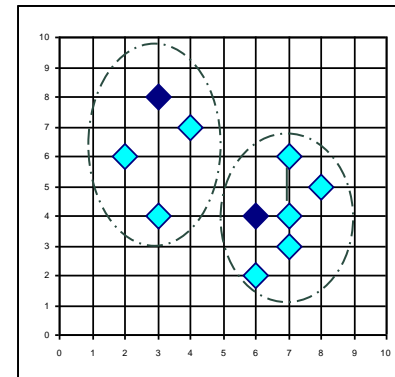
$k=2$

循环直到  
没有变化

随机选择  $k$   
个点作为初  
始 medoids

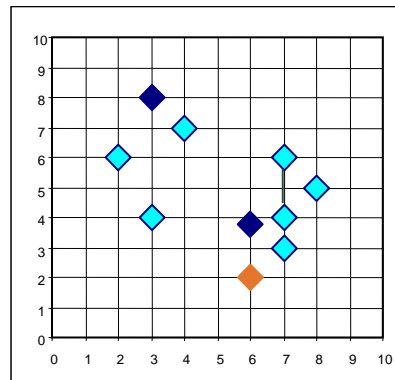


指定剩余点  
到离它最近  
的 medoids  
类中



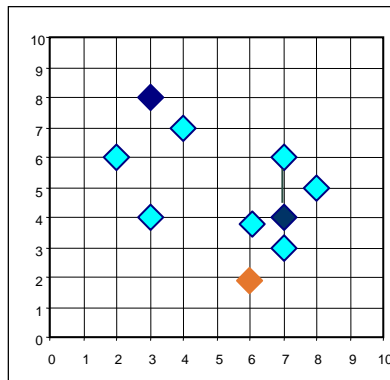
Total Cost = 20

顺序选择非  
medoids点  $O_{\text{random}}$



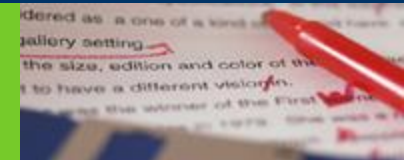
计算medoids  
交换的代价

如果交换  
medoids提升  
了聚类质量,  
那就用  
 $O_{\text{random}}$  代替  
 $O$





# 聚类分析的主要方法 (9)



## ■ 围绕中心点的划分 Partitioning Around Medoids

算法： $k$  中心点。PAM，一种基于中心点或中心对象进行划分的  $k$  中心点算法。

输入：

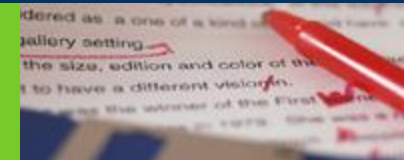
- $k$ ：簇的数目，
- $D$ ：包含  $n$  个对象的数据集。

输出： $k$  个簇的集合。

方法：

- (1) 从  $D$  中任意选择  $k$  个对象作为代表对象或种子；
- (2) **repeat**
- (3) 将每个剩余对象指派到最近的代表对象所代表的簇；
- (4) **随机** 选择一个非代表对象  $O_{random}$ ；
- (5) 计算用  $O_{random}$  交换代表对象  $O_j$  的总代价  $S$ ；
- (6) 如果  $S < 0$ ，就用  $O_{random}$  替换  $O_j$ ，形成新的  $k$  个代表对象的集合；
- (7) **until** 不再发生变换

# 聚类分析的主要方法 (10)



## ■ 其它划分方法:

### ■ Clara 算法 Clustering LARge Applications algorithm

*k-medoids* 算法不适合于大数据量的计算。*Clara* 算法的思想就是用实际数据的抽样来代替整个数据, 然后再在这些抽样的数据上利用 *k-medoids* 算法得到最佳的 *medoids*。*Clara* 算法能够处理大量的数据, 但其有效性取决于样本的大小  $O(ks^2 + k(n-k))$

### ■ Clarans 算法 Clustering LARge Application based upon RANdomized Search algorithm

*Clara* 算法的效率取决于采样的大小, 一般不太可能得到最佳的结果。在 *Clara* 算法的基础上, 又提出了 *Clarans* 算法, 它将抽样技术与 *k-medoids* 结合起来。与 *Clara* 算法在搜索开始时就抽取节点样本不同, *Clarans* 算法在搜索的每一步都动态地按照某种随机性抽取近邻的随机样本。它的聚类效果取决于所用的抽样方法  $O(n^2)$

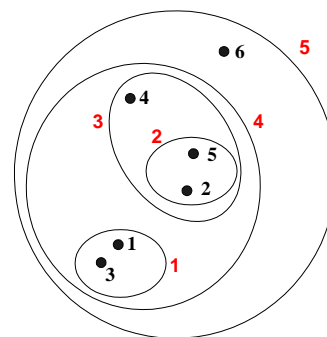
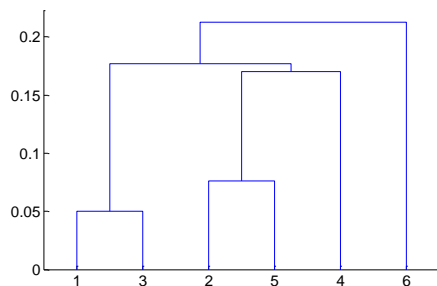
# 聚类分析的主要方法 (11)

## ■ 层次方法 Hierarchical methods

- 创建给定数据对象集的层次分解
- 根据层次分解的形成方式，该方法可以分为凝聚的（自底向上）或分裂的（自顶向下）

**凝聚层次聚类：**首先将每个对象作为其簇，然后合并这些原子簇为越来越大的簇，直到所有的对象都在一个簇中，或者某个终止条件满足

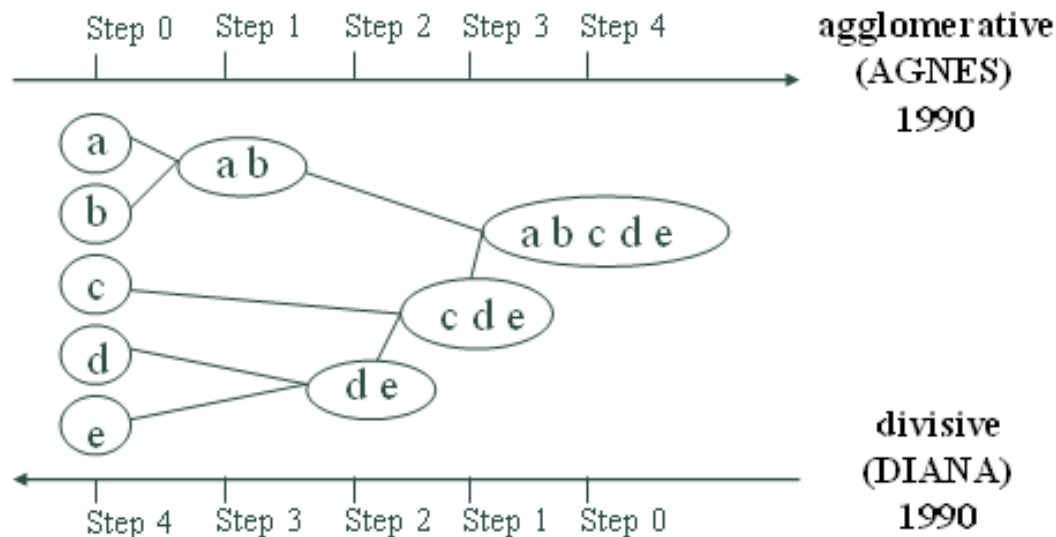
**分裂层次聚类：**首先将所有对象置于一个簇中，然后将它逐渐细分为越来越小的簇，直到每个对象自成一簇，或者达到某个终止条件



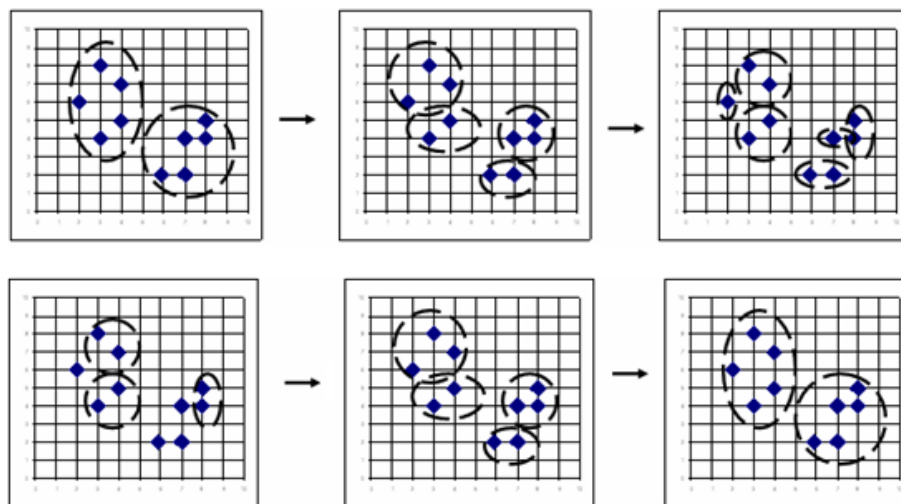
树状图 (*dendrogram*)

# 聚类分析的主要方法 (12)

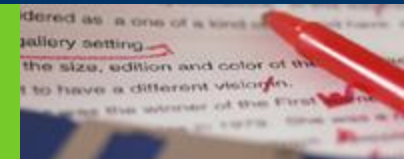
## 例题



- MIN [Single link](#)
- MAX [Complete link](#)
- Group Average
- Distance Between Centroid (质心)
- Distance Between Medoid (中心)

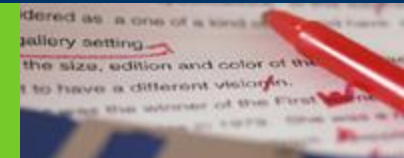


# 聚类分析的主要方法 (13)



方法	空间性质	单调性	对距离的要求	结果的要求	适用性	备注
MIN	压缩	单调		条形, S形	唯一	太压缩 不够灵敏
MAX	扩张	单调		椭圆形	不唯一	太扩张, 样本 大时易失真
中间距离	守恒	非单调	欧氏距离的平方	椭圆形	不唯一	
Centroid	守恒	非单调	欧氏距离的平方	椭圆形	不唯一	
Group Average	守恒	单调		椭圆形	不唯一	效果较好 较常用
变差平方和	扩张	单调	欧氏距离的平方	椭圆形	不唯一	效果较好 较常用

# 聚类分析的主要方法 (14)



## ■ 层次方法的优点:

- 不需要预先指定聚类的簇数 $k$ , 可以对树状图进行修剪获得需要的聚类数目
- 聚类结果对应更好的有意义的分类准则 (meaningful taxonomies)  
动物族谱、语系发展史等

对于纯粹的层次聚类方法, 一旦合并或分裂执行, 就不能修正

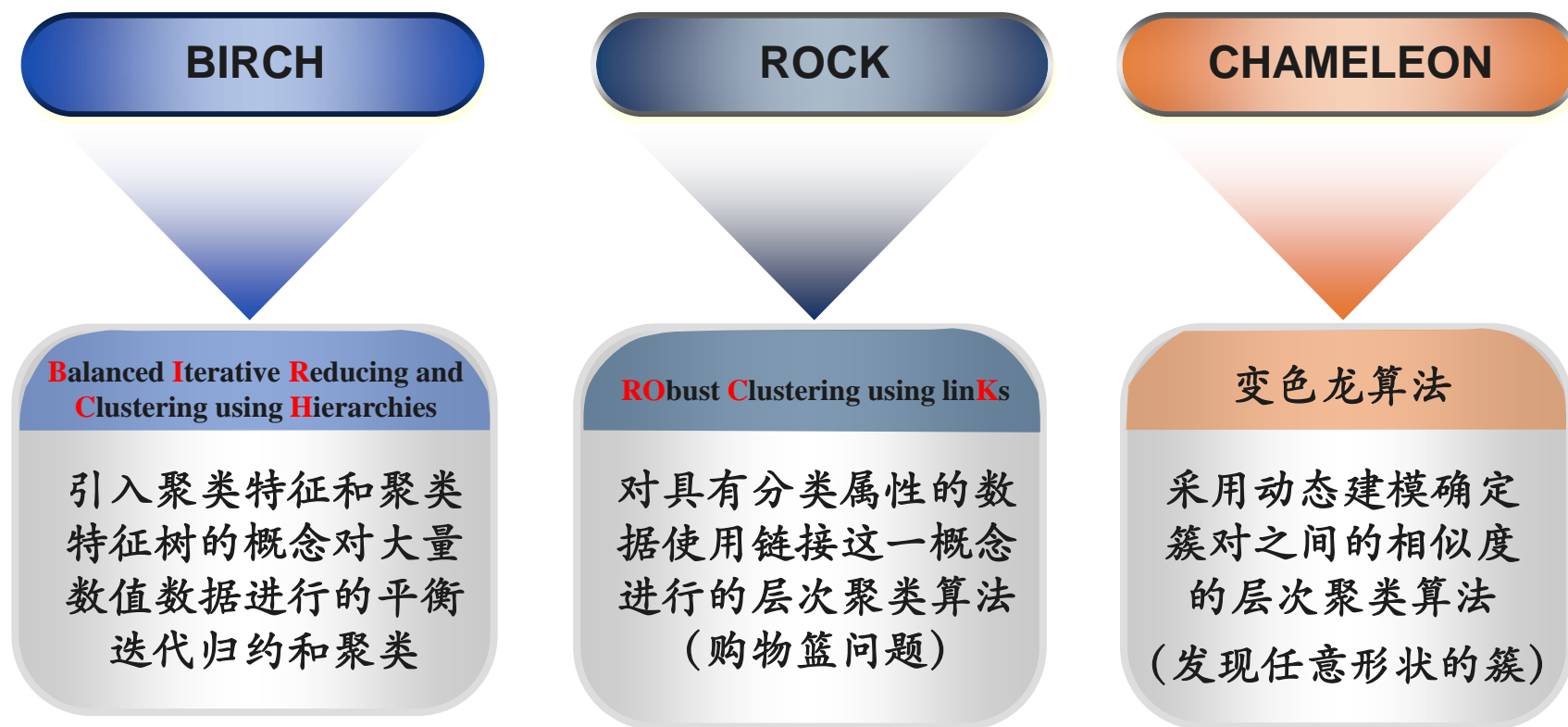
## ■ 层次方法的改进:

- 为了弥补合并或分裂的严格性, 凝聚层次方法的聚类质量可以通过以下方法改进:

分析每个层次划分中的对象链接, 或者首先执行微聚类 (把数据划分为“微簇”), 然后使用其它聚类技术对微簇聚类

# 聚类分析的主要方法 (15)

## ■ 其它层次方法:



## ■ 最近的研究强调凝聚层次聚类和迭代重定位方法的集成、概率层次聚类

*K.A. Heller and Z. Ghahramani. "Bayesian hierarchical clustering". ICML'05, pp.297-304, 2005*



# 聚类分析的主要方法 (16)

## ■ 基于密度的方法 Density-Based methods

- **指导思想**: 只要一个区域中的点的密度大过某个阈值, 就把它加到与之相近的聚类中去
- **根本区别**: 基于密度的方法不是基于各种各样的距离的, 而是基于密度的, 这样就能克服基于距离的算法只能发现“类圆形”的聚类的缺点, 并且可以过滤掉“噪声”和“孤立点”

代表算法

DBSCAN

根据邻域对象的密度

OPTICS

生成数据聚类结构的一个增广序

DENCLUE

根据某种密度函数



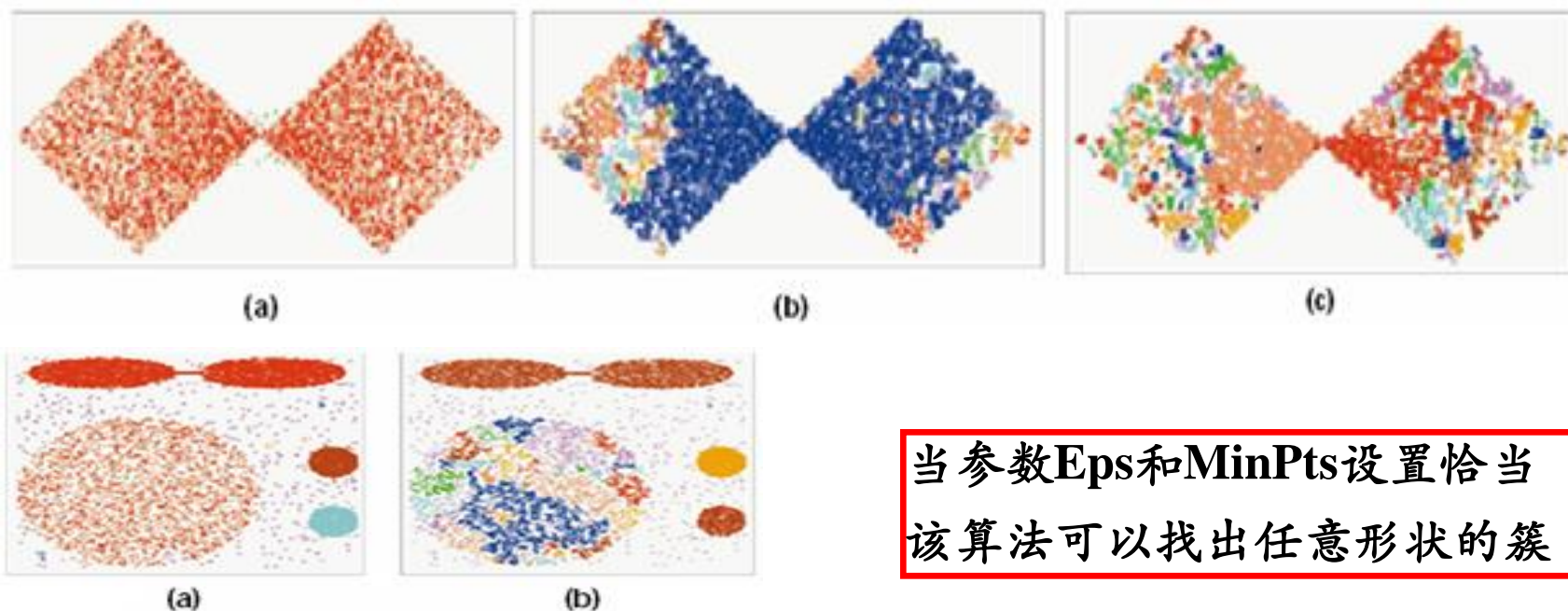
# 聚类分析的主要方法 (17)

## ■ DBSCAN: 基于高密度连通区域的基于密度的聚类算法

**Density-Based Spatial Clustering of Applications with Noise**

具有噪声的基于密度的聚类应用

- 将具有足够高密度的区域划分为簇，并在具有噪声的空间数据库中发现任意形状的簇，将簇定义为密度相连的点的最大集合



当参数Eps和MinPts设置恰当  
该算法可以找出任意形状的簇

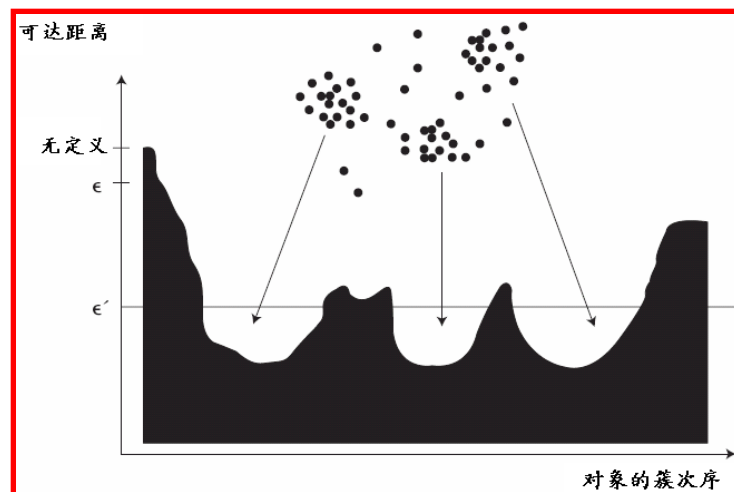
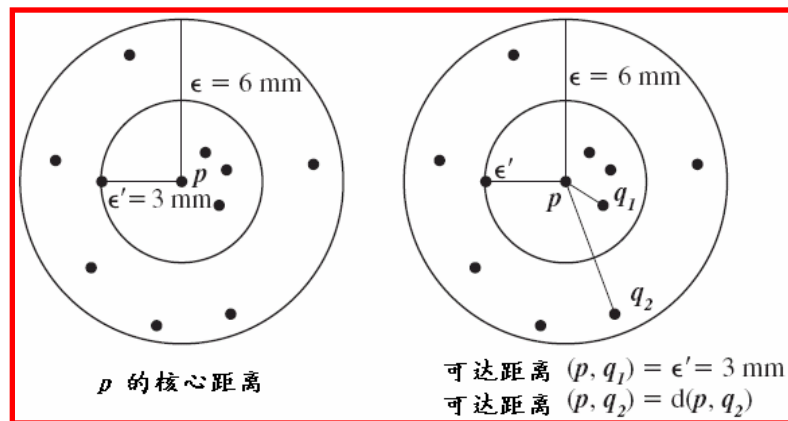
# 聚类分析的主要方法 (18)

## ■ OPTICS: 通过点排序识别聚类结构

### Ordering Points to Identify the Clustering Structure

#### 通过点排序识别聚类结构

- 并不显式地产生数据集聚类，而是为自动和交互的聚类分析计算一个增广的**簇排序** (cluster ordering)，它包含的信息等价于从一个广泛的参数设置所获得的基于密度的聚类
- 次序选择根据最小的  $\epsilon$  值密度可达的对象，以便较高密度（较低  $\epsilon$  值）的簇先完成



# 聚类分析的主要方法 (19)

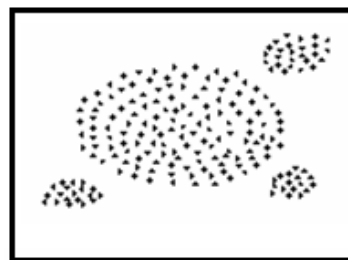
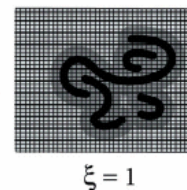
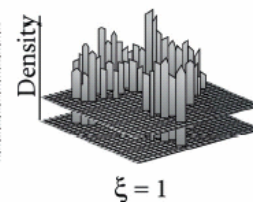
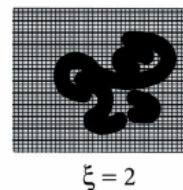
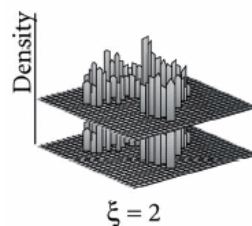
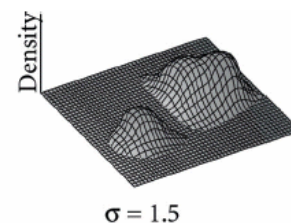
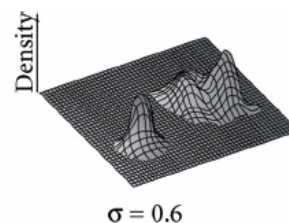
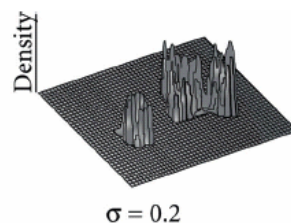
## ■ DENCLUE: 基于密度分布函数的聚类算法

- 坚实的数学基础，概括了各种聚类方法
- 适用于含有大量噪声的数据
- 聚类高维数据的任意形状簇
- 运算速度快
- 参数选择要求高

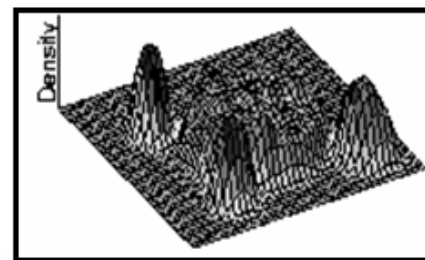
$$f_{\text{Gaussian}}(x, y) = e^{-\frac{d(x, y)^2}{2\sigma^2}}$$

$$f_{\text{Gaussian}}^D(x) = \sum_{i=1}^N e^{-\frac{d(x, x_i)^2}{2\sigma^2}}$$

$$\nabla f_{\text{Gaussian}}^D(x, x_i) = \sum_{i=1}^N (x_i - x) \cdot e^{-\frac{d(x, x_i)^2}{2\sigma^2}}$$



(a) 数据集



(b) 基于高斯影响的总密度函数

密度吸引点影响很多其它的点

# 聚类分析的主要方法 (20)

## ■ 基于网格的方法 Grid-Based methods

### 指导思想

- 使用多分辨率的网格数据结构
- 将对象空间量化为有限数目的单元，形成网格结构
- 对网格结构进行聚类

### 主要优点

- 处理速度快，其处理时间独立于数据对象的数目，仅依赖于量化空间中每一维的单元数目

### STING (**S**Tatistical **I**nformation **G**rid)

统计信息网格 将空间区域划分为多层次矩形单元，利用存储在网格单元中的统计信息进行聚类分析

### WaveCluster

利用小波变换聚类 在数据空间强加一个多维网络结构来汇总数据，采用小波变换来变换原特征空间，在变换后空间发现聚类区域

### CLIQUE (**C**lustering **I**n **Q**UEST)

维增长子空间聚类 聚类过程开始于单维的子空间，通过升维识别空间中的稀疏和拥挤区域（单元），把高维稠密单元的搜索限定在子空间稠密单元的交集上

# 聚类分析的主要方法 (21)

## ■ 聚类高维数据 Clustering High-Dimensional Data

- 当数据的维数增加时，通常只有少数的几维与某些簇相关，其它不相关维的数据可能会产生大量的噪声而屏蔽真实的簇
- 随着维数的增加，数据通常会变得更加稀疏，位于不同维的数据点可以认为是距离相等的，这就使得聚类分析中十分重要的距离度量失去意义

### 特征（属性）变换

主成分分析 (PCA)  
随机投影  
奇异值分解 (SVD)



适用于维数相关数据

### 特征（属性）选择

过滤 (filter)  
包装 (wrapper)



运算时间消耗大

### 子空间聚类

维增长子空间聚类(CLIQUE)  
维归约子空间聚类(PROCLUS)  
基于频繁模式的聚类(pCluster)



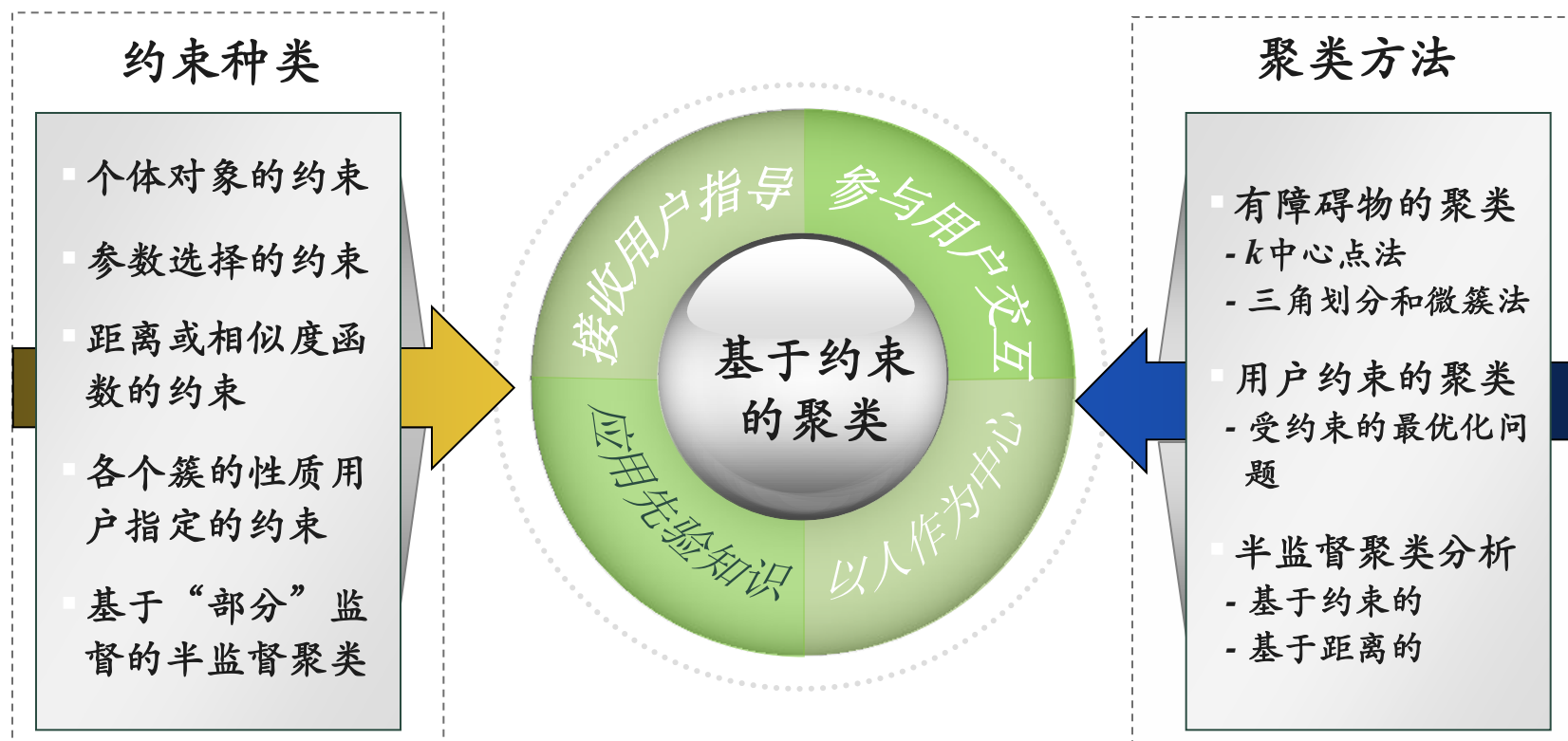
适用于高维聚类：无监督学习



# 聚类分析的主要方法 (22)

## ■ 基于约束的方法 Constraint-Based methods

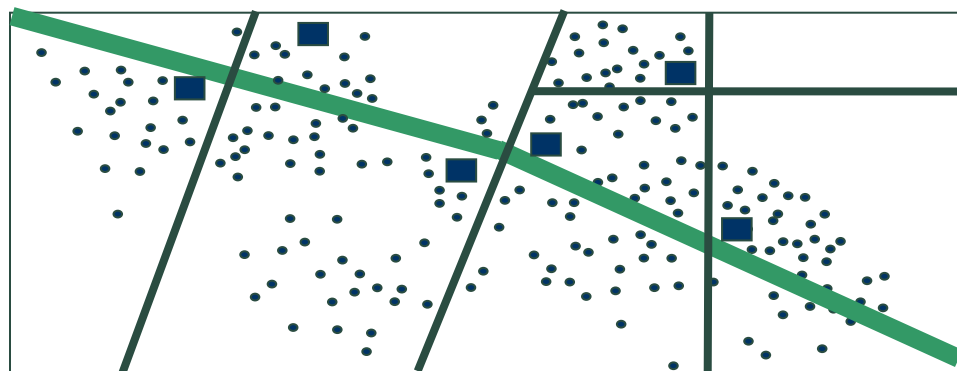
根据应用需求或者用户指定的约束（偏好）对对象进行分组



# 聚类分析的主要方法 (23)

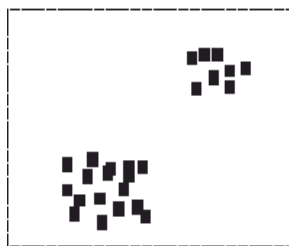
## 例题

- 在拥有河流、桥梁、公路、湖泊和山脉的城市中设置自动取款机的位置  
(包含障碍物对象的聚类)
- 包裹投递公司在某个城市确定  $k$  个服务站的位置  
(用户约束的聚类)

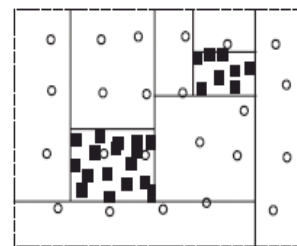


? 聚类是根据数量还是根据价值

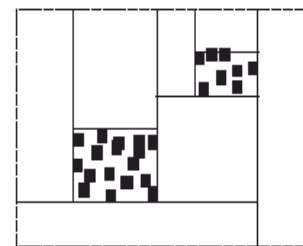
- CLTree (Clustering based on decision TREEs, 基于决策树的聚类)  
集成聚类与监督分类思想  
(基于约束的半监督聚类)



(a)



(b)



(c)

# 聚类分析的主要方法 (24)

## ■ 基于模型的方法 Model-Based methods

- **指导思想**: 假设数据都是根据潜在的混合概率分布生成, 为每个簇假设一个模型, 并找出数据与该模型的最佳拟合

### 基于模型的聚类方法

#### 统计方法

- 每个簇都可以用**参数**概率分布数学描述, 整个数据就是这些分布的混合
- 期望最大化算法 (EM算法)
- 贝叶斯聚类算法 AutoClass

#### 神经网络方法

- 自组织特征映射 (SOM)
- 用低维目标空间的点来表示高维源空间中的所有点

#### 概念聚类

- 首先聚类, 然后给出每组对象的**特征描述**其中每组对象代表一个概念或类

• COBWEB ..... → • CLASSIT



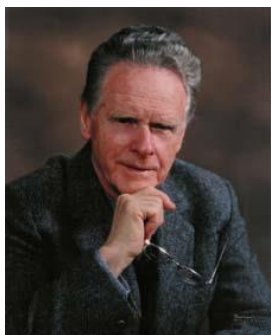
# 聚类分析的主要方法 (25)

## ■ 期望最大化算法 Expectation-Maximization algorithm

在已知部分相关变量的情况下，估计未知变量的迭代技术。

EM的算法流程如下：

- 初始化分布参数
  - 重复直到收敛：
- E步骤：估计未知参数的期望值，给出当前的参数估计
  - M步骤：重新估计分布参数，以使得数据的似然性最大，给出未知变量的期望估计



*Arthur Dempster, Nan Laird, and Donald Rubin.  
"Maximum likelihood from incomplete data via the  
EM algorithm". Journal of the Royal Statistical  
Society, Series B, 39(1):1–38, 1977*

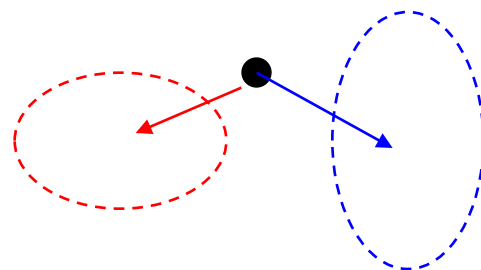
## ■ EM算法是学习模型中存在隐含变量时广泛使用的一种学习方法。

贝叶斯网络、径向基网络、聚类算法、马尔科夫模型、数据缺失等

# 聚类分析的主要方法 (26)

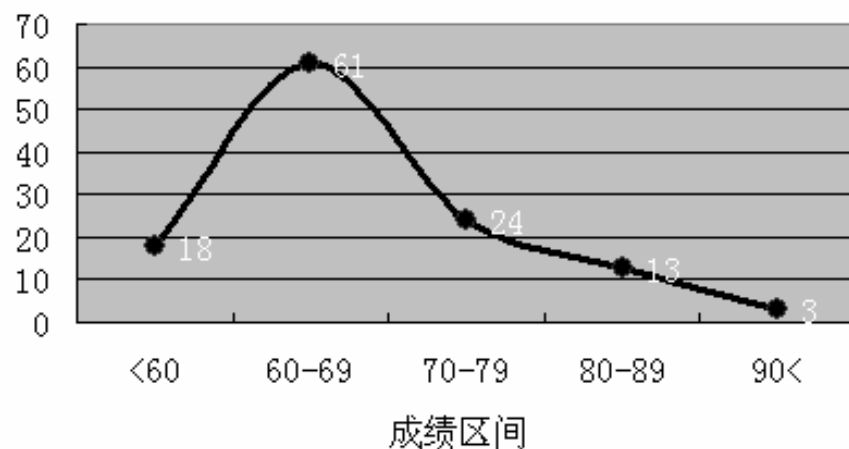
## ■ 期望最大化算法 Expectation-maximization algorithm

- 实例都以一定的可能性分属于每个聚类
- 聚类的目标是寻找给定数据的最有可能的聚类



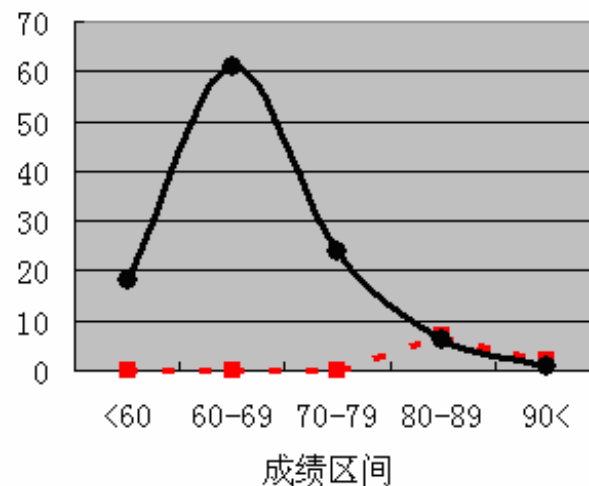
## ■ 有限混合 finite mixtures 统计模型

学生人数



2009高等工程数学成绩分布图

学生人数

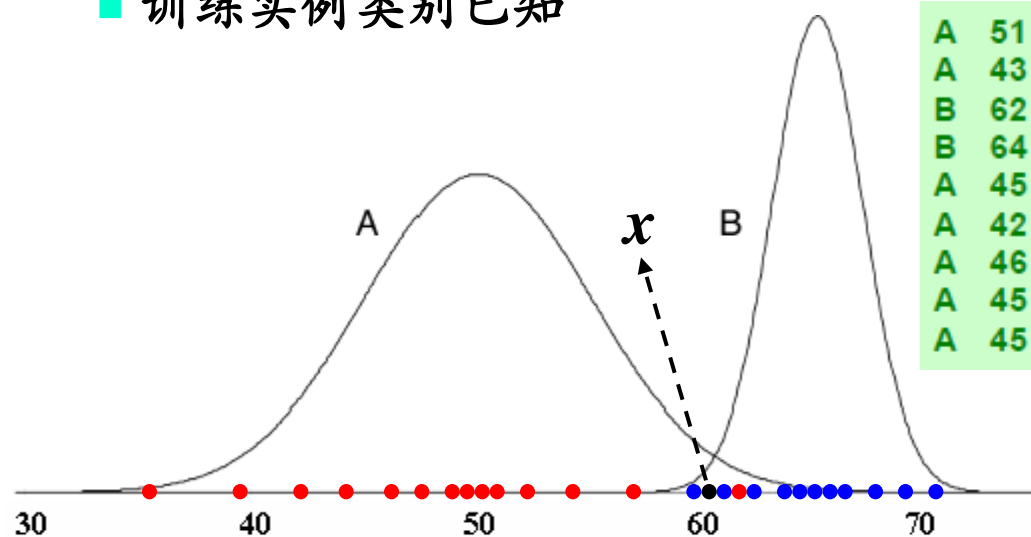


在职生与应届生成绩分布图

# 聚类分析的主要方法 (27)

## ■ 有限混合 finite mixtures 统计模型

### ■ 训练实例类别已知



$$\mu_A = 50, \sigma_A = 5, P_A = 0.6 \quad \mu_B = 65, \sigma_B = 2, P_B = 0.4$$

$$\Pr[A | x] = \frac{\Pr[x | A] \cdot \Pr[A]}{\Pr[x]} = \frac{f(x; \mu_A, \sigma_A) \cdot P_A}{\Pr[x]}$$

$$\Pr[B | x] = \frac{\Pr[x | B] \cdot \Pr[B]}{\Pr[x]} = \frac{f(x; \mu_B, \sigma_B) \cdot P_B}{\Pr[x]}$$

A	51	B	62	B	64	A	48	A	39	A	51
A	43	A	47	A	51	B	64	B	62	A	48
B	62	A	52	A	52	A	51	B	64	B	64
B	64	B	64	B	62	B	63	A	52	A	42
A	45	A	51	A	49	A	43	B	63	A	48
A	42	B	65	A	48	B	65	B	64	A	41
A	46	A	48	B	62	B	66	A	48		
A	45	A	49	A	43	B	65	B	64		
A	45	A	46	A	40	A	46	A	48		

$$P_A + P_B = 1$$

$$\mu = \frac{x_1 + x_2 + \dots + x_n}{n}$$

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_n - \mu)^2}{n}$$

严格地说:  $f(x; \mu_A, \sigma_A) \neq \Pr[x | A]$

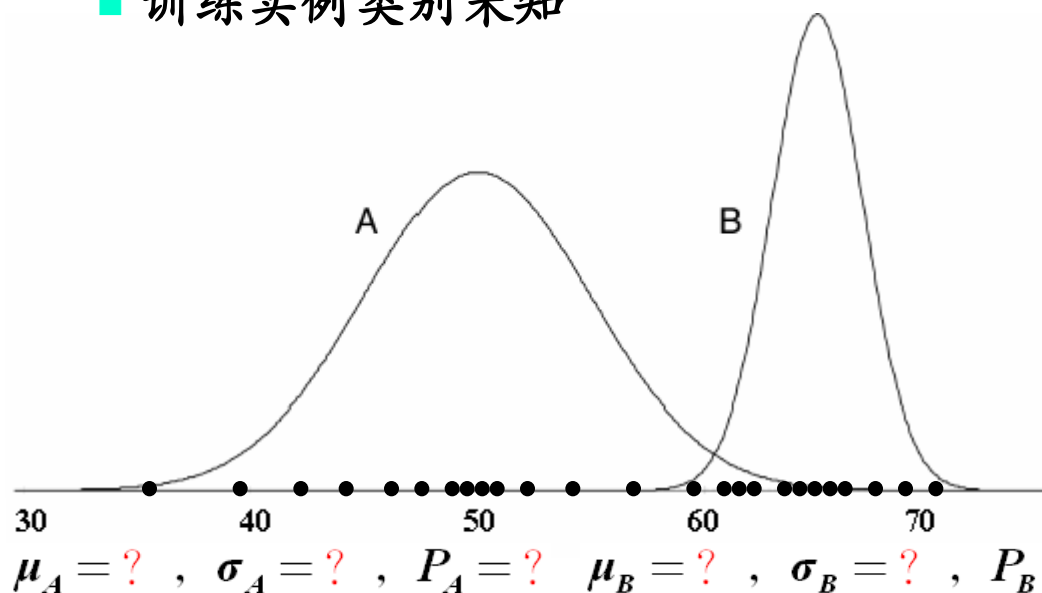
$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



# 聚类分析的主要方法 (28)

## ■ 期望最大化算法 Expectation-maximization algorithm

### ■ 训练实例类别未知



51	62	64	48	39	51
43	47	51	64	62	48
62	52	52	51	64	64
64	64	62	63	52	42
45	51	49	43	63	48
42	65	48	65	64	41
46	48	62	66	48	
45	49	43	65	64	
45	46	40	46	48	

$k$ 均值初始化聚类参数



$$\mu_A = \mu_{A0}, \sigma_A = \sigma_{A0}, P_A = P_{A0}$$

$$\mu_B = \mu_{B0}, \sigma_B = \sigma_{B0}, P_B = P_{B0}$$

$$\Pr[A | x] = \frac{\Pr[x | A] \cdot \Pr[A]}{\Pr[x]} = \frac{f(x; \mu_A, \sigma_A) \cdot P_A}{\Pr[x]}$$

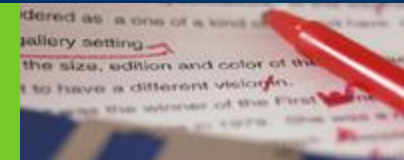
计算实例  $i$  属于聚类  $A$  的概率

$$\mu_{A1} = \frac{w_{A1}x_1 + w_{A2}x_2 + \dots + w_{An}x_n}{w_{A1} + w_{A2} + \dots + w_{An}}$$

$$\sigma_{A1} = \frac{w_{A1}(x_1 - \mu_{A1}) + \dots + w_{An}(x_n - \mu_{A1})}{w_{A1} + w_{A2} + \dots + w_{An}}$$

$$\Pr[A | x_i] = w_{Ai}$$

# 聚类分析的主要方法 (29)



## ■ 期望最大化算法 Expectation-maximization algorithm

数据集总体似然最大化 
$$\max \sum_i^n \log (P_A \cdot \Pr[x_i | A] + P_B \cdot \Pr[x_i | B])$$

循环迭代直到对数似然 ( log-likelihood ) 的增值可忽略不计

### ■ 几点说明:

- “坏”的参数初始值设置可以导致EM算法陷入一些局部最优点

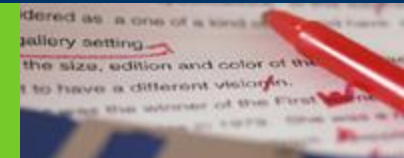
为了能有机会得到全局最大值，可使用不同的初始猜测参数值重复几次

- EM算法的收敛速度比较慢

降低数据规模，简化模型参数

- 只有在不存在直接解决的算法的情况下，才考虑使用EM算法，因为它并不是解决限制条件下优化问题的高效方法

# 聚类分析的主要方法 (30)

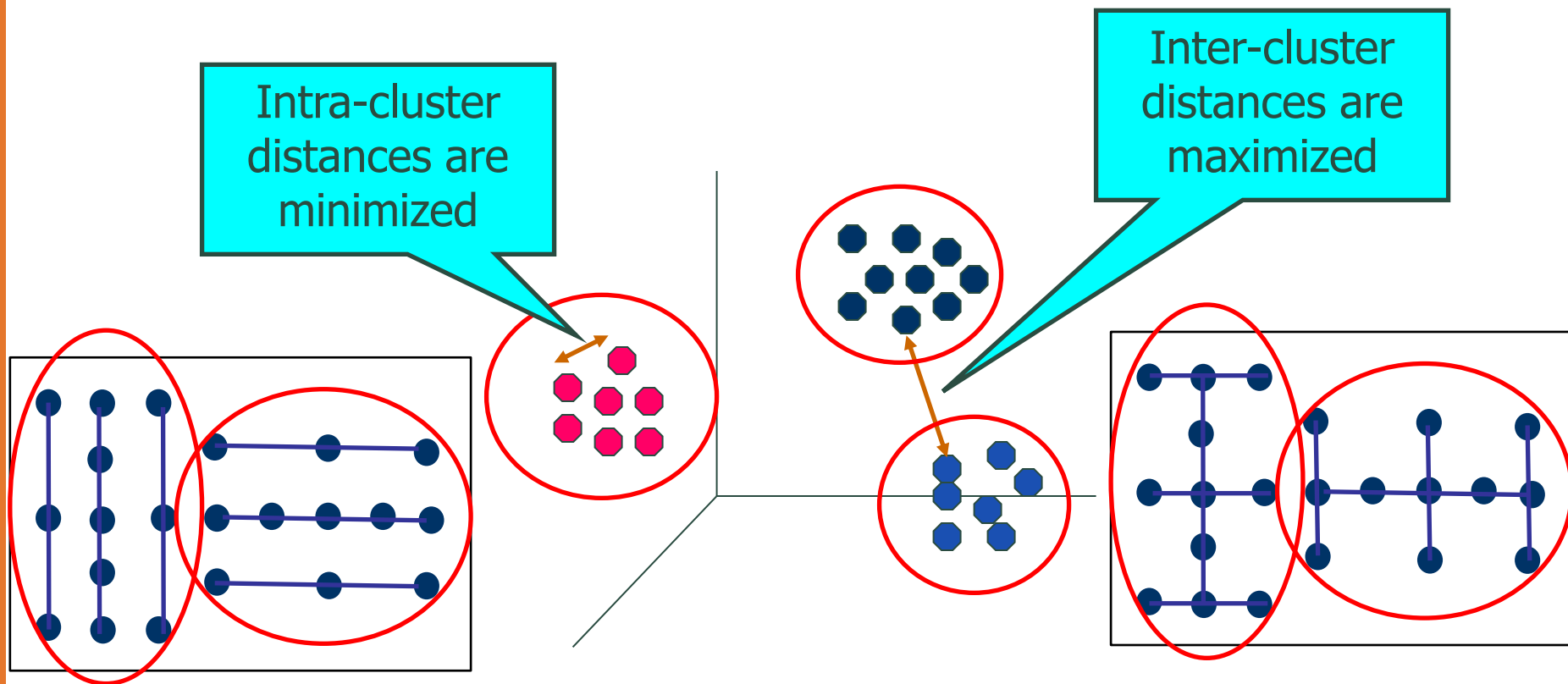


## ■ 各种典型聚类算法性能比较

	算法效率	可发现的簇形状	异常数据的敏感性	输入数据顺序敏感性
K-means	较高	凸形或球形	很敏感	不敏感
K-medoids	一般	凸形或球形	不敏感	不敏感
BIRCH	高	凸形或球形	不敏感	一般
CURE	较高	任意	不敏感	一般
DBSCAN	一般	任意	敏感	敏感
STING	高	任意	一般	不敏感
CLIQUE	较低	凸形或球形	一般	不敏感

# 如何评估一个好的聚类 (1)

- 同一簇中的对象之间具有很高的相似度，而不同簇之间的对象高度相异



- 聚类效果与不同聚类方法中所使用的相似度有关
- **聚类评估** 估计在数据集上进行聚类的可行性和由聚类方法产生的结果的质量

# 如何评估一个好的聚类 (2)

## ■ 聚类评估的主要任务

### ■ 估计聚类趋势

数据集上的聚类分析是有意义的，仅当数据中存在非随机结构

### 霍普金斯统计量 (Hopkins Statistic)

(1) 均匀地从  $D$  的空间中抽取  $n$  个点

$$p_1, p_2, \dots, p_n$$

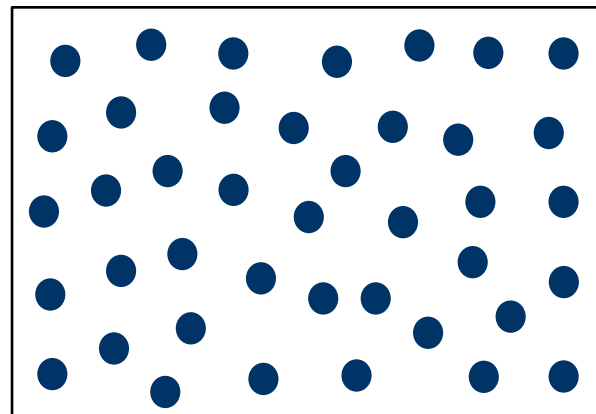
(2) 均匀地从  $D$  的空间中抽取  $n$  个点

$$q_1, q_2, \dots, q_n$$

(3) 计算霍普金斯统计量  $H$

$$H = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i}$$

$H \approx 0.5$   $\rightarrow$   $D$  均匀分布  
 $H \rightarrow 0$   $\rightarrow$   $D$  高度倾斜





# 如何评估一个好的聚类 (3)

## ■ 聚类评估的主要任务

聚类方法的有效性也根据其发现隐含模式的能力来衡量

### ■ 确定簇数

经验方法：对于  $n$  个点的数据集，设置簇数  $p$  大约为  $\sqrt{\frac{n}{2}}$

在期望情况下，每个簇大约有  $\sqrt{2n}$  个点

肘方法 (elbow method)：增加簇数有助于降低每个簇的簇内方差之和  
使用簇内方差和关于簇数的曲线的拐点

交叉验证：使用检验集中的所有点与它们的最近形心之间的距离的平方和来度量聚类模型拟合检验集的程度

### ■ 测定聚类质量 是否存在由专家构建的一种理想状态的聚类 (基准)？

外在方法

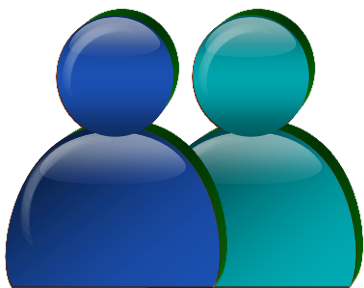
*E. Amigo et. al. "A comparison of extrinsic clustering evaluation metrics based on formal constraints". Information Retrieval, Vol.12(4):461–486, 2009*

内在方法

*L. Kaufman et. al. "Finding Groups in Data: An Introduction to Cluster Analysis". John Wiley & Sons, 1990*

# 聚类分析

- 1 聚类分析基本概念
- 2 聚类分析中的数据类型
- 3 聚类分析的主要方法
- 4 离群点分析方法

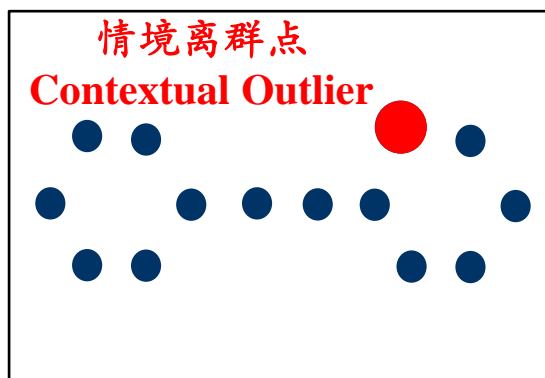


# 什么是离群点

- 与数据的一般行为或模型不一致的数据对象被称为离群点 (Outlier)

Data object that does not comply with the general behavior of the data

—— Jiawei Han *Data Mining: Concepts and Techniques*



- **全局离群点**是显著地偏离数据集中其他对象的一个对象
  - 交易审计系统中, 不遵守常规的交易
- **情境离群点**是在关于对象的特定情境下显著地偏离其他对象的一个对象
  - 今天的温度为28 °C, 这是一个异常吗?
- **集体离群点**是数据对象的一个子集作为整体显著地偏离整个数据集
  - 短期内, 相同股票在一小群当事人之间的大量交易

# 离群点检测的挑战

- 离群点挖掘的核心问题是离群点的定义（噪声还是离群点）和如何发现离群点（统计、距离、密度、偏差）
  - 噪声是被观测变量的随机误差或方差，噪声在数据分析中一般是无趣的
    - 一位顾客可能会产生“噪声交易”：比通常多购买了一件产品
  - 离群点的产生机制不同于其他数据的产生机制，因此是有趣的
    - 信用卡一次购买量比卡主通常购买量大得多，且该购买远离卡主的居住地
- 离群点检测的挑战
  - 正常对象和离群点的有效建模
  - 针对应用的离群点检测
  - 在离群点检测中处理噪声
  - 离群点的可理解性
- 数据可视化对于检测有很多分类属性或高维数据中的离群点效果很差

全局离群点检测最简单

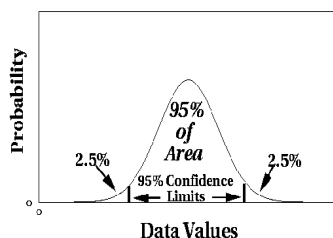
情境离群点检测需要背景知识来确定情境属性和情境

集体离群点检测需要背景知识来对对象之间的联系建模，以便找出离群点的组群

# 发现离群点的方法

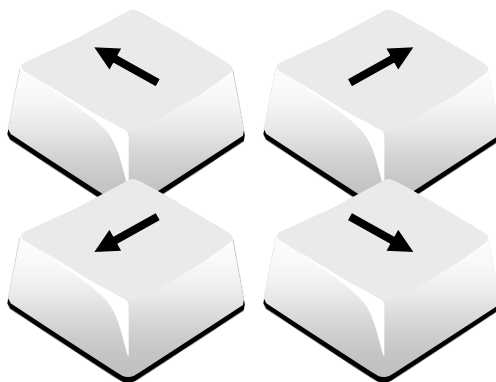
## 基于统计分布的离群点检测

Statistical Approach



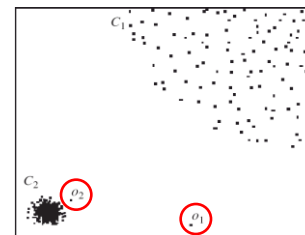
## 基于距离的离群点检测

Distance-Based Approach



## 基于密度的局部离群点检测

Density-Based Local Outlier Detection



## 基于偏差的离群点检测

Deviation-Based Approach

### ■ 基于统计分布的离群点检测

- 根据不同的检验统计量应用不和谐检验

### ■ 基于距离的离群点检测

- 定义离群点是没有“足够多”近邻的对象，其中近邻基于到给定对象的距离定义

### ■ 基于密度的局部离群点检测

- 局部离群点相对于它的局部邻域，特别是关于邻域密度，是远离的

### ■ 基于偏差的离群点检测

- 通过检查一组对象的主要特征来识别离群点

- 大多数检验是针对单个属性的

- 数据分布往往是未知的

- 距离参数设置涉及试凑

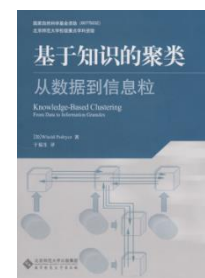
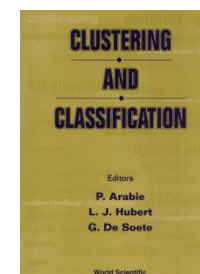
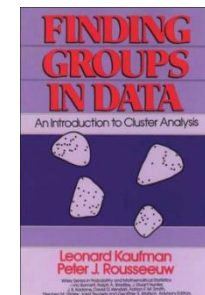
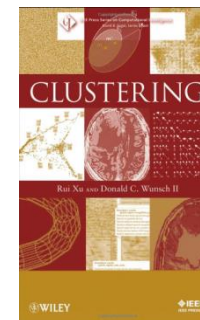
- 适用于非均匀分布数据

- 顺序异常技术：子集光滑因子

- OLAP数据立方体技术：下钻

# 参考文献

- **Clustering**. Wiley-IEEE Press, 2009.  
作者: Rui Xu, Don Wunsch
- **Finding Groups in Data: an Introduction to Cluster Analysis**. John Wiley & Sons, 1990.  
作者: Leonard Kaufman, Peter J. Rousseeuw
- **Clustering and Classification**. World Scientific, 1996.  
作者: Phipps Arabie, Lawrence J. Hubert, Geert De Soete
- **基于知识的聚类——从数据到信息粒**. 北京师范大学出版社, 2008  
作者: Witold Pedrycz 翻译: 于福生
- **模糊聚类算法及应用**. 国防工业出版社, 2011  
作者: 曲福恒 等
- **谱聚类集成算法研究**. 天津大学出版社, 2011  
作者: 贾建华





# 参考文献

- A.K. Jain, M.N. Murty, P.J. Flynn. **Data Clustering: A survey**. ACM Computing Surveys, Vol. 31(3): 264-323, 1999  
<http://eprints.iisc.ernet.in/273/01/p264-jain.pdf>
- L. Parsons, E. Haque, H. Liu. **Subspace clustering for high dimensional data: A review**. SIGKDD Explorations, Vol. 6(1): 90-105, 2004  
<http://www.sigkdd.org/explorations/issues/6-1-2004-06/parsons.pdf>
- A.K. Jain. **Data clustering: 50 years beyond K-means**. Pattern Recognition Letters, Vol. 31(8): 651-666, 2010  
[http://biometrics.cse.msu.edu/Publications/GeneralPRIP/JainDataClustering\\_PRL09.pdf](http://biometrics.cse.msu.edu/Publications/GeneralPRIP/JainDataClustering_PRL09.pdf)
- Algergawy Alsayed, Mesiti Marco, Nayak Richi, Saake Gunter. **XML data clustering : an overview**. ACM Computing Surveys, Vol. 43(4): No. 25, 2011  
<http://eprints.iisc.ernet.in/273/01/p264-jain.pdf>
- V. Chandola, A. Banerjee, V. Kumar. **Anomaly detection: A survey**. ACM Computing Surveys, Vol. 41(3): No. 15, 2009  
<http://rp-www.cs.usyd.edu.au/~comp4044/survey/anomaly.pdf>

# 聚类分析

- 1 聚类分析基本概念
- 2 聚类分析中的数据类型
- 3 聚类分析的主要方法
- 4 离群点分析方法



在观察的领域中，机遇只偏爱那种有准备的头脑。

—— Louis Pasteur

法国微生物学家和化学家 (1822 – 1895)





*To be continued .....*



[wangyong@ucas.ac.cn](mailto:wangyong@ucas.ac.cn)

<http://people.ucas.ac.cn/~wangyong>

