

# 第四章 统计模式识别中的 聚类方法

吉  
祥  
如意

中科大 自动化系 郑志刚  
2018.11





- 4.0 聚类分析基本概念
- 4.1 相似性准则（相似性度量）
- 4.2 聚类准则函数
- 4.3 两种简单的聚类算法
- 4.4 系统聚类
- 4.5 分解聚类
- 4.6 动态聚类
- 4.7 最小张树聚类

# 吉 祥 图

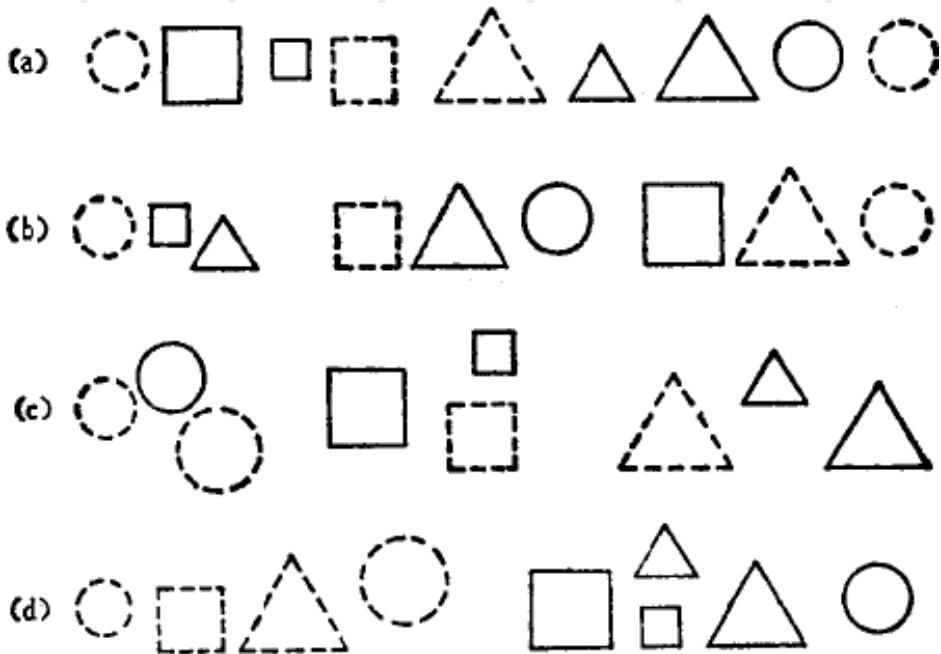
## 4.0 聚类分析基本概念

### ■ 分类与聚类的区别

➤ 分类：用已知类别的样本训练集来设计分类器（监督学习）

➤ 聚类（集群）：用事先不知样本的类别，而利用样本的先验知识来构造分类器（无监督学习）

聚类分析的关键问题：如何在聚类过程中自动地确定类型数目

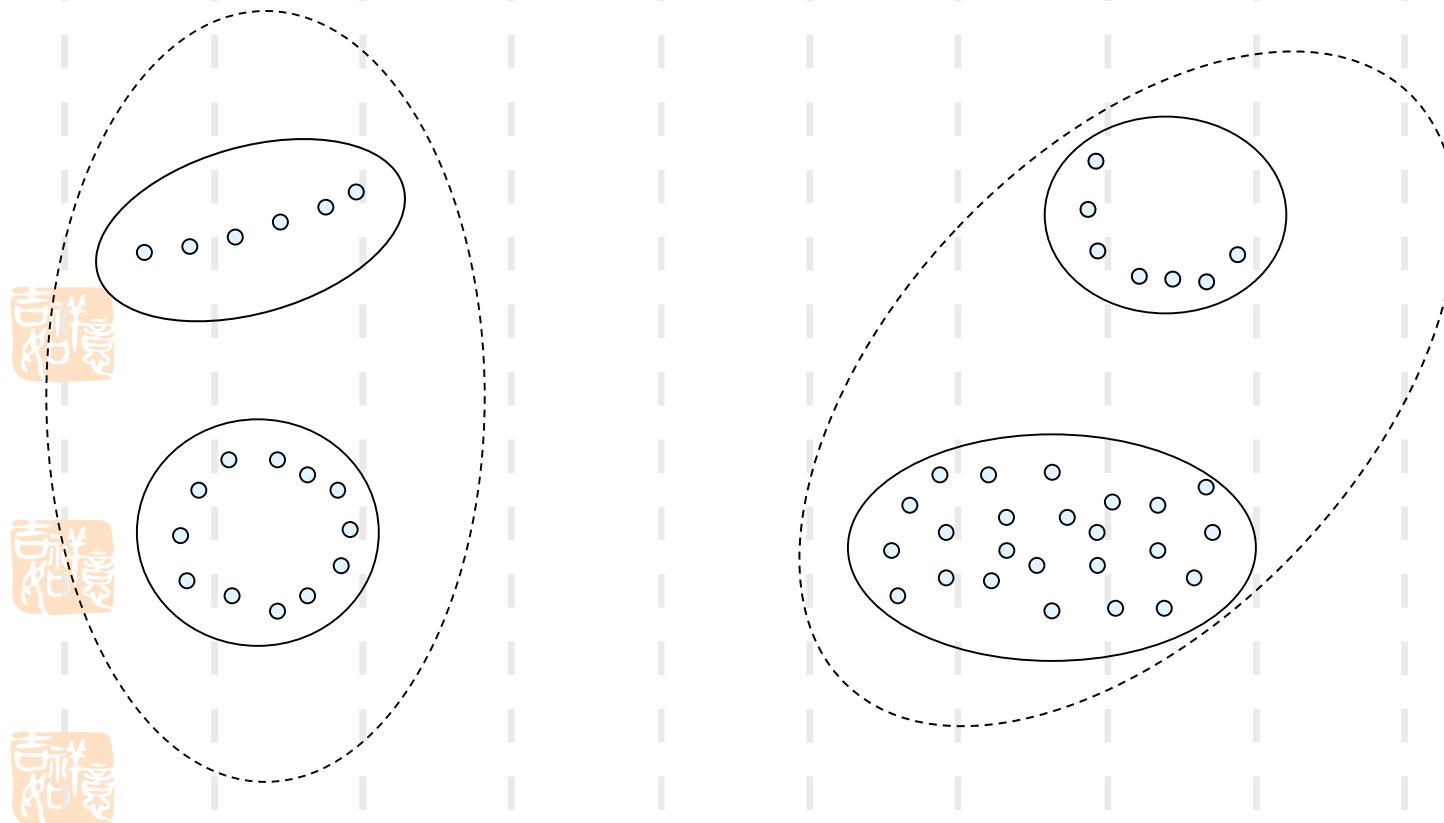


(a) 混合训练样本集； (b) 根据面积分成三类；  
(c) 根据外形分成三类； (d) 根据线型分成两类。

吉  
祥  
如  
意

## 4.0 聚类分析基本概念

距离测度不同,聚类结果也不同



数据的粗聚类是两类,细聚类为4类

吉祥如意

## 4.0 聚类分析基本概念

综上可见：

选择什么特征？

选择多少个特征？

选择什么样的量纲？

选择什么样的距离测度？

这些对分类结果都会产生极大影响。



# 聚类过程遵循的基本步骤

## 一、特征选择(feature selection)

尽可能多地包含任务关心的信息

## 二、近邻测度(proximity measure)

定量测定两特征如何“相似”或“不相似”

## 三、聚类准则 (clustering criterion)

以蕴涵在数据集中类的类型为基础

## 四、聚类算法 (clustering algorithm)

按近邻测度和聚类准则揭示数据集的聚类结构

## 五、结果验证 (validation of the results)

常用逼近检验验证聚类结果的正确性

## 六、结果判定 (interpretation of the results)

由专家用其他方法判定结果的正确性



## 4.1 相似性准则（相似性度量）

用于描述各模式之间特征的相似程度

- 距离测度
- 相似测度
- 匹配测度



# 吉祥如意

## 4.1 相似性准则（相似性度量）

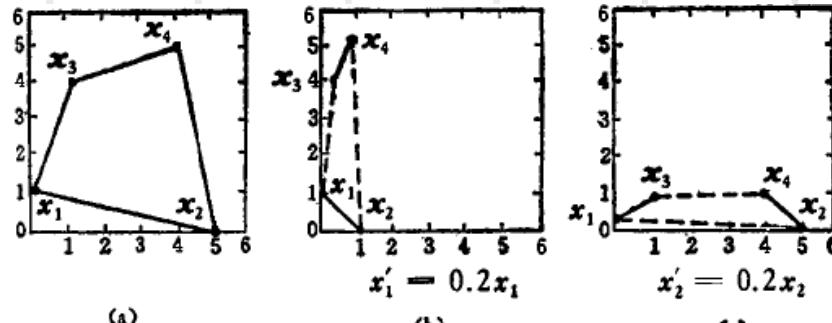
### 1. 距离相似性度量

(1) 欧氏距离

$$D_e(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^d |x_i - y_i|^2}$$

a、门限  $d_s$  的选择问题

b、模式特征坐标单位的选取也会强烈地影响聚类结果



特征坐标单位对聚类结果的影响

c、欧氏距离具有旋转不变的特性，但对于一般的线性变换不是不变的，此时要对数据进行标准化

# 在聚类分析中，常用的聚类要素的数据处理方法有 如下几种：

① 总和标准化。分别求出各聚类要素所对应的数据的总和，以各要素的数据除以该要素的数据的总和，即

$$x'_{ij} = \frac{x_{ij}}{\sum_{i=1}^m x_{ij}} \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, n)$$

这种标准化方法所得到的新数据满足

$$\sum_{i=1}^m x'_{ij} = 1 \quad (j = 1, 2, \dots, n)$$



## ② 标准差标准化，即

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, n)$$

由这种标准化方法所得到的新数据，各要素的平均值为0，标准差为1，即有

$$\bar{x}'_j = \frac{1}{m} \sum_{i=1}^m x'_{ij} = 0 \qquad s'_j = \sqrt{\frac{1}{m} \sum_{i=1}^m (x'_{ij} - \bar{x}'_j)^2} = 1$$





### ③ 极大值标准化，即

$$x'_{ij} = \frac{x_{ij}}{\max_i \{x_{ij}\}} \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, n)$$

经过这种标准化所得的新数据，各要素的极大值为1，其余各数值小于1。

### ④ 极差的标准化，即

$$x_{ij} = \frac{x_{ij} - \min_i \{x_{ij}\}}{\max_i \{x_{ij}\} - \min_i \{x_{ij}\}} \quad (i = 1, 2, \dots, m; j = 1, 2, \dots, n)$$

经过这种标准化所得的新数据，各要素的极大值为1，极小值为0，其余的数值均在0与1之间。



## 4.1 相似性准则（续）

d、还要注意模式样本测量值的选取，应该是有效反映类别属性特征（各类属性的代表应均衡）





## 4.1 相似性准则（续）

### （2）马氏（Mahalanobis）距离

定义：马氏距离的平方  $\gamma^2 = (x - \mu)^T \Sigma^{-1} (x - \mu)$

其中， $\mu$  为均值向量， $\Sigma$  为协方差矩阵

### （3）明氏（Minkowsky）距离

定义：明氏距离： $D_\lambda(x, y) = \left[ \sum_{i=1}^d |x_i - y_i|^\lambda \right]^{\frac{1}{\lambda}}$ ，  $\lambda > 0$

它是若干距离函数的通式：

$\lambda = 2$  时，等于欧氏距离；

$\lambda = 1$  时，称为“街坊”（city block）距离

## 例2.2.1

已知一个二维正态母体G的分布为  $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}\right)$

求点  $A: \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  和  $B: \begin{pmatrix} 1 \\ -1 \end{pmatrix}$  至均值点  $M: \vec{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$  的距离。

解：由题设，可得  $\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$   $\Sigma^{-1} = \frac{1}{0.19} \begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$

从而马氏距离

$$d_M^2(A, M) = (1 - 1) \Sigma^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0.2 / 0.19 \quad d_M^2(B, M) = (1 - (-1)) \Sigma^{-1} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = 3.8 / 0.19$$

它们之比达  $\sqrt{19}$  倍，若用欧氏距离，算得的距离值相同。

$$d_E^2(A, M) = 2 \quad d_E^2(B, M) = 2$$

由分布函数知，A、B两点的概率密度分别为

$$p(1, 1) = 0.2157 \quad p(1, -1) = 0.00001658$$



## 4.1 相似性准则（相似性度量）

### 二、相似测度

测度基础：以两矢量的方向是否相近作为考虑的基础，矢量长度并不重要。设

$$\vec{x} = (x_1, x_2, \dots, x_n), \vec{y} = (y_1, y_2, \dots, y_n)$$



#### 1. 角度相似系数(夹角余弦)

$$\cos(\vec{x}, \vec{y}) = \frac{\vec{x}'\vec{y}}{\|\vec{x}\|\|\vec{y}\|} = \frac{\vec{x}'\vec{y}}{[(\vec{x}'\vec{x})(\vec{y}'\vec{y})]^{1/2}}$$



注意：坐标系的旋转和尺度的缩放是不变的，但对一般的线形变换和坐标系的平移不具有不变性。



## 二、相似测度

### 2. 相关系数

它实际上是数据中心化后的矢量夹角余弦。



$$r(\vec{x}, \vec{y}) = \frac{(\vec{x} - \bar{\vec{x}})'(\vec{y} - \bar{\vec{y}})}{\left[ (\vec{x} - \bar{\vec{x}})'(\vec{x} - \bar{\vec{x}})(\vec{y} - \bar{\vec{y}})'(\vec{y} - \bar{\vec{y}}) \right]^{1/2}}$$



## ■ 4.1 相似性准则（相似性度量）

### 二、相似测度

#### 3. 指数相似系数

$$e(\vec{x}, \vec{y}) = \frac{1}{n} \sum_{i=1}^n \exp \left[ -\frac{3}{4} \frac{(x_i - y_i)^2}{\sigma_i^2} \right]$$

式中  $\sigma_i^2$  为相应分量的协方差， $n$  为矢量维数。  
它不受量纲变化的影响。



## ■ 4.1 相似性准则（相似性度量）

### 三、匹配测度

当特征只有两个状态（0，1）时，常用匹配测度。

0表示无此特征 1表示有此特征。故称之为二值特征。

对于给定的 $x$ 和 $y$ 中的某两个相应分量 $x_i$ 与 $y_j$

若 $x_i=1, y_j=1$ ，则称  $x_i$ 与 $y_j$ 是 (1-1) 匹配；

若 $x_i=1, y_j=0$ ，则称  $x_i$ 与 $y_j$ 是 (1-0) 匹配；

若 $x_i=0, y_j=1$ ，则称  $x_i$ 与 $y_j$ 是 (0-1) 匹配；

若 $x_i=0, y_j=0$ ，则称  $x_i$ 与 $y_j$ 是 (0-0) 匹配。





## ■ 4.1 相似性准则（相似性度量）

### 三、匹配测度

对于二值**n**维特征矢量可定义如下相似性测度

令  $a = \sum_i x_i y_i$  为  $\vec{x}$  与  $\vec{y}$  的 (1-1) 匹配的特征数目

$b = \sum_i y_i (1 - x_i)$  (0-1) 匹配的特征数目

$c = \sum_i x_i (1 - y_i)$  (1-0) 匹配的特征数目

$e = \sum_i (1 - x_i)(1 - y_i)$  (0-0) 匹配的特征数目





## ■ 4.1 相似性准则（相似性度量）

### 三、匹配测度

#### (1) Tanimoto 测度

$$s(\vec{x}, \vec{y}) = \frac{a}{a + b + c} = \frac{\vec{x}'\vec{y}}{\vec{x}'\vec{x} + \vec{y}'\vec{y} - \vec{x}'\vec{y}}$$





## ■ 4.1 相似性准则（相似性度量）

### 例4.1.2

设  $\vec{x} = (0, 1, 0, 1, 1, 0)'$        $\vec{y} = (0, 0, 1, 1, 0, 1)'$

则  $\vec{x}'\vec{x} = 3$  ,     $\vec{y}'\vec{y} = 3$  ,     $\vec{x}'\vec{y} = 1$

$$s(\vec{x}, \vec{y}) = \frac{1}{3+3-1} = \frac{1}{5}$$



可以看出，它等于共同具有的特征数目与分别具有的特征种类总数之比。这里只考虑(1-1)匹配而不考虑(0-0)匹配。



## ■ 4.1 相似性准则（相似性度量）

### 三、匹配测度

#### (2) Rao测度



$$s(\vec{x}, \vec{y}) = \frac{a}{a + b + c + e} = \frac{\vec{x}'\vec{y}}{n}$$



注: (1-1)匹配特征数目和所选用的特征数目之比。





## ■ 4.1 相似性准则（相似性度量）

### 三、匹配测度

#### (3) 简单匹配系数



$$m(\vec{x}, \vec{y}) = \frac{a + e}{n}$$



注：上式分子为**(1-1)**匹配特征数目与**(0-0)**匹配特征数目之和，分母为所考虑的特征数目。





## ■ 4.1 相似性准则（相似性度量）

### 三、匹配测度

#### (4) Dice系数

$$m(\vec{x}, \vec{y}) = \frac{a}{2a + b + c} = \frac{\vec{x}'\vec{y}}{\vec{x}'\vec{x} + \vec{y}'\vec{y}} = \frac{(1 - 1) \text{匹配个数}}{\text{俩矢量中1的总数}}$$



#### (5) Kulzinsky系数

$$m(\vec{x}, \vec{y}) = \frac{a}{b + c} = \frac{\vec{x}'\vec{y}}{\vec{x}'\vec{x} + \vec{y}'\vec{y} - 2\vec{x}'\vec{y}} = \frac{(1 - 1) \text{匹配个数}}{(0 - 1) + (1 - 0) \text{匹配个数}}$$





## 4.1 相似性准则（续）

- 样本相似性度量是聚类分析的基础，针对具体问题，选择适当的相似性度量是保证聚类质量的重要问题。但有了相似性度量还不够，还必须有适当的聚类准则函数。聚类准则函数对聚类质量也有重大影响。
  - 相似性度量 → 集合与集合的相似性。
  - 相似性准则 → 分类效果好坏的评价准则。



吉祥如意

## 4.2 聚类准则函数

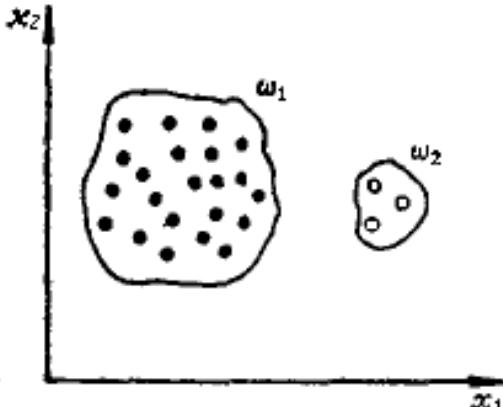
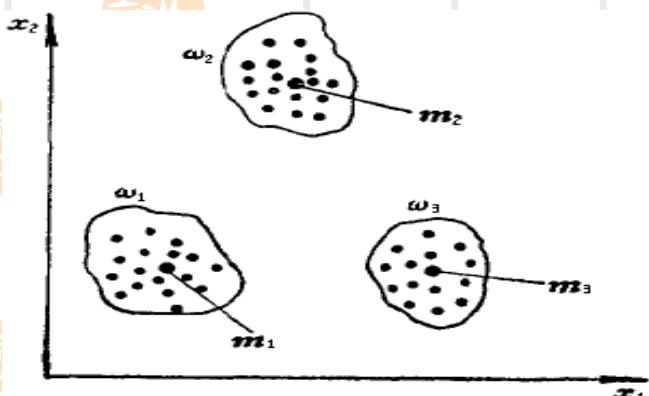
### 1. 误差平方和准则（最常用的）

$$J_c = \sum_{j=1}^c \sum_{k=1}^{n_j} \|x_k - m_j\|^2$$

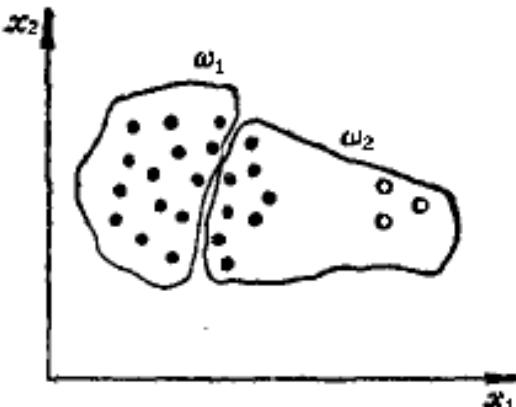
式中  $m_j$  为类型  $w_j$  中样本的均值:  $m_j = \frac{1}{n_j} \sum_{j=1}^{n_j} x_j$ ,  $j = 1, 2, \dots, c$ 。

$m_j$  是  $c$  个集合的中心，可以用来代表  $c$  个类型。

误差平方和准则适用于各类样本比较密集且样本数目悬殊不大的样本分布



(a) 正确分类



(b) 错误分类



例如：有5个样本，如下图所示  $x_1 \sim x_4 \in w_1$ ， $x_5 \in w_2$ 。

虚线为正确类型区分域，实线为采用误差平方和最小准则时的类别区分。

虚线划分时： $w_1 : X_1 = \{x_1, x_2, x_3, x_4\}$ ， $m_1 = \frac{1}{4} \sum_{k=1}^4 x_k = (0,0)^T$

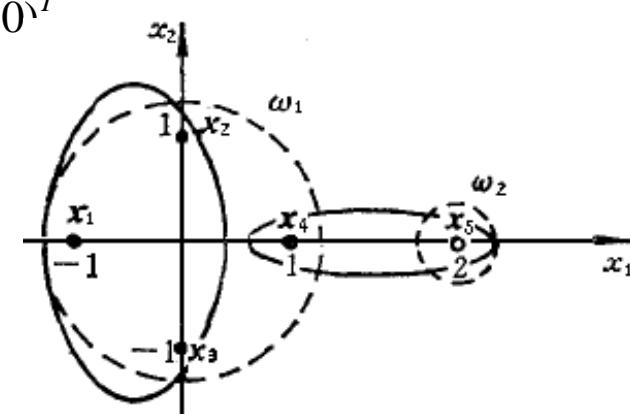
$w_2 : X_2 = \{x_5\}$ ， $m_2 = x_5 = (2,0)^T$

$$J_{c_1} = \sum_{x_k \in X_1} \|x_k - m_1\|^2 + \sum_{x_k \in X_2} \|x_k - m_2\|^2 = 4$$

实线划分时： $w_1 : X_1 = \{x_1, x_2, x_3\}$ ， $m_1 = \frac{1}{3} \sum_{k=1}^3 x_k = (-\frac{1}{3}, 0)^T$

$w_2 : X_2 = \{x_4, x_5\}$ ， $m_2 = \frac{1}{2} \sum_{x_k \in X_2} x_k = (1.5, 0)^T$

$$J_{c_2} = \sum_{x_k \in X_1} \|x_k - m_1\|^2 + \sum_{x_k \in X_2} \|x_k - m_2\|^2 = \frac{8}{3} + \frac{1}{2} = \frac{19}{6} = 3.17$$



$J_c$  准则错误聚类的示例

所以  $J_{c_1} > J_{c_2}$ ，如果按误差平方和准则聚类将得到错误结果。



## 4.2 聚类准则函数（续）

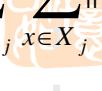
### 2. 加权平均平方距离和准则

$$J_l = \sum_{j=1}^c P_j \cdot S_j^*$$

式中： $S_j^*$  是类内样本间平均平方距离。

  $S_j^* = \frac{2}{n_j(n_j - 1)} \sum_{x \in X_j} \sum_{x' \in X_j} \|x - x'\|^2$ ，所有的样本之间距离的平均值。

  $X_j$  中的样本个数  $n_j$ ， $X_j$  中的样本两两组合共有  $\frac{n_j(n_j - 1)}{2}$  种。

  $\sum_{x \in X_j} \sum_{x' \in X_j} \|x - x'\|^2$  表示所有样本之间距离之和。



$P_j$  为  $w_j$  类的先验概率，可以用样本数目  $n_j$  和样本总数目  $n$  来估计。

$$P_j = \frac{n_j}{n}, \quad j=1,2,\dots,c \quad \text{因此: } J_l = \frac{1}{n} \sum_{j=1}^c n_j \cdot S_j^*$$

用  $J_l$  重新讨论误差平方和准则中所举例子。

虚线划分时:  $w_1 : X_1 = \{x_1, x_2, x_3, x_4\}$ ,  $S_1^* = \frac{1}{6}(4+2+2+2+2+4) = \frac{8}{3}$

$$w_2 : X_2 = \{x_5\}, \quad S_2^* = 0.$$

$$J_{l_1} = \frac{4}{5} \times \frac{8}{3} = 2.13$$

实线划分时:  $w_1 : X_1 = \{x_1, x_2, x_3\}$ ,  $S_1^* = \frac{1}{3}(4+2+2) = \frac{8}{3}$

$$w_2 : X_2 = \{x_4, x_5\}, \quad S_2^* = \frac{1}{1}(1) = 1.$$

$$J_{l_2} = \frac{3}{5} \times \frac{8}{3} + \frac{2}{5} \times 1 = 2$$



## 4.2 聚类准则函数（续）

### 3. 类间距离和准则

$$J_{b_1} = \sum_{j=1}^c (m_j - m)^T (m_j - m)$$

加权类间距离和:  $J_{b_2} = \sum_{j=1}^c P_j \cdot (m_j - m)^T (m_j - m)$


$$m_j = \frac{1}{n_j} \sum_{j=1}^{n_j} x_j \quad j = 1, 2, \dots, c$$



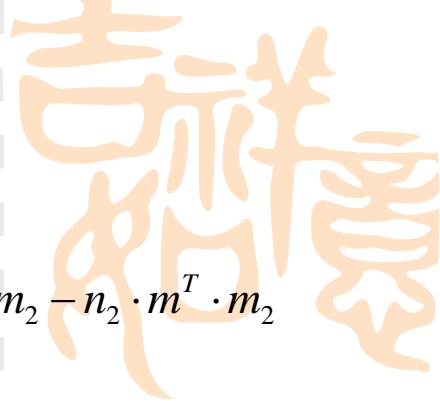
$m = \frac{1}{n} \sum_{k=1}^n x_k$   $P_j$  为  $w_j$  类型的先验概率, 可以用  $\frac{n_j}{n}$  来估计。



两类问题:  $w_1 / w_2$ , 类间距离常用


$$J_{b_3} = (m_1 - m_2)^T (m_1 - m_2)$$





两类问题的加权类间距离和：

$$J_{b_2} = \frac{1}{n} \sum_{j=1}^2 n_j \cdot (m_j - m)^T (m_j - m) = \frac{1}{n} (n_1 \cdot m_1^T \cdot m_1 - n_1 \cdot m^T \cdot m_1 + n_2 \cdot m_2^T \cdot m_2 - n_2 \cdot m^T \cdot m_2)$$

将  $n \cdot m = n_1 \cdot m_1 + n_2 \cdot m_2$  代入上式，有：

$$J_{b_2} = \frac{1}{n} \left[ \frac{n_1 n_2}{n} (m_1^T - m_2^T) m_1 + \frac{n_1 n_2}{n} (m_1^T - m_2^T) m_2 \right] = \frac{n_1 n_2}{n^2} [(m_1 - m_2)^T (m_1 - m_2)] = P_1 \cdot P_2 \cdot J_{b_3}$$

类间距离和准则描述不同类型之间的分离程度，所以  $J_b$  的值越

大，表示各类之间分离性好，聚类质量高。





## 4.2 聚类准则函数（续）

### 4. 散射矩阵

①类内散射矩阵  $S_w = \sum_{j=1}^c P_j \cdot S_j$

其中  $S_j$  为某一个类型的类内散射矩阵：  
 $S_j = \frac{1}{n_j} \sum (x_k^{(j)} - m_j)(x_k^{(j)} - m_j)^T$   
 $x_k^{(j)}$  表示  $w_j$  类型的第  $k$  个样本， $j = 1, 2, \dots, c$ 。

②类间散射矩阵  $S_b = \sum_{j=1}^c P_j \cdot (m_j - m)(m_j - m)^T$

③全部样本的总散射矩阵  $S_t = \frac{1}{n} \sum_{k=1}^n (X_k - m)(X_k - m)^T$

上述3个散射矩阵有如下关系：

$$S_t = S_w + S_b$$

吉祥如意

证明：

$$\begin{aligned} S_t &= \sum_{j=1}^c \frac{n_j}{n} \cdot \frac{1}{n_j} \sum_{k=1}^{n_j} (m_k^{(j)} - m)(m_k^{(j)} - m)^T \\ &= \sum_{j=1}^c P_j \left[ \frac{1}{n_j} \sum_{k=1}^{n_j} (m_k^{(j)} - m_j)(m_k^{(j)} - m_j)^T + (m_j - m)(m_j - m)^T \right] \\ &= \sum_{j=1}^c P_j \cdot S_j + \sum_{j=1}^c P_j \cdot (m_j - m)(m_j - m)^T \\ &= S_w + S_b \end{aligned}$$

可以定义如下的4个聚类准则：



$$J_1 = t_r(S_w^{-1}S_b)$$

$$J_2 = |S_w^{-1}S_b|$$



$$J_3 = t_r(S_w^{-1}S_t)$$

$$J_4 = |S_w^{-1}S_t|$$





考虑到矩阵的迹和行列式的旋转不变性，我们总可以找到一个正交矩阵  $U$ ，使  $U^T(S_w^{-1}S_b)U = A$ 。

$(S_w^{-1}S_b)$  是  $d \times d$  维的对称矩阵， $U$  是  $d \times d$  维正交归一化矩阵， $A$  是以特征值  $\lambda_i (i=1,2,\dots,d)$  为对角线的对角矩阵。则有：

$$J_1 = \sum_{i=1}^d \lambda_i \quad J_2 = \prod_{i=1}^d \lambda_i$$

又由于： $S_w^{-1}S_t = S_w^{-1}(S_w + S_b) = I + S_w^{-1}S_b$ ， $I$  为  $d \times d$  维单位矩阵。

$$U^T(S_w^{-1}S_t)U = U^T(I + S_w^{-1}S_b)U = I + A$$

所以： $J_3 = \sum_{i=1}^d (1 + \lambda_i)$   $J_4 = \prod_{i=1}^d (1 + \lambda_i)$



## 4.3 两种简单的聚类算法

### 1. 采用最近邻规则的聚类算法

假设已有混合样本集  $X = \{x_1, x_2, \dots, x_n\}$ ，按照最近邻原则进行聚类，

算法如下：

① 选取距离阈值  $T$ ，并且任取一个样本作为第一个聚合中心  $Z_1$ ，

如： $Z_1 = x_1$ 。

② 计算样本  $x_2$  到  $Z_1$  的距离  $D_{21}$ ：

若  $D_{21} \leq T$ ，则  $x_2 \in Z_1$ ，否则令  $x_2$  为第二个聚合中心， $Z_2 = x_2$ 。

设  $Z_2 = x_2$ ，计算  $x_3$  到  $Z_1$  和  $Z_2$  的距离  $D_{31}$  和  $D_{32}$ ，若  $D_{31} > T$  和  $D_{32} > T$ ，则建立第三个聚合中心  $Z_3$ 。否则把  $x_3$  归于最近邻的聚合中心。依此类推，直到把所有的  $n$  个样本都进行分类。

③ 按照某种聚类准则考察聚类结果，若不满意，则重新选取距离阈值  $T$ 、第一个聚合中心  $Z_1$ ，返回②，直到满意，算法结束。

## 4.3 两种简单的聚类算法（续）

### 2. 最大最小距离聚类算法

例：样本分布如图所示。

最大最小距离聚类算法步骤如下：

① 给定  $\theta$ ,  $0 < \theta < 1$ , 并且任取一个样本

作为第一个聚合中心,  $Z_1 = x_1$ 。

② 寻找新的集合中心:

计算其它所有样本到  $Z_1$  的距离  $D_{i1}$ :

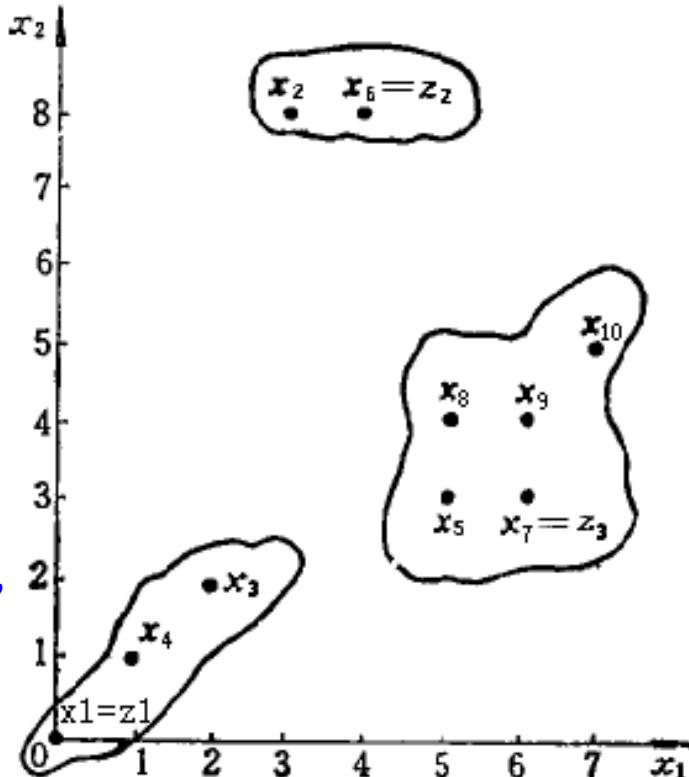
若  $D_{k1} = \max_i \{D_{i1}\}$ , 则取  $x_k$  为第二个聚合中心  $Z_2$ ,

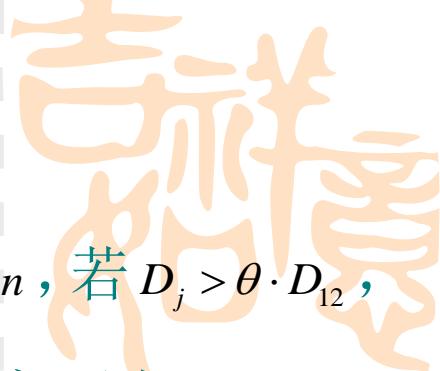
$Z_2 = x_6$

计算所有样本到  $Z_1$  和  $Z_2$  的距离  $D_{i1}$  和  $D_{i2}$ :

若  $D_l = \max \{\min(D_{i1}, D_{i2})\}$ ,  $i = 1, 2, \dots, n$ , 并且  $D_l > \theta \cdot D_{12}$ ,

$D_{12}$  为  $Z_1$  和  $Z_2$  间距离, 则取  $x_l$  为第三个集合中心  $Z_3$ ,  $Z_3 = x_7$ 。





【注意:  $D_{i1} = \|x_i - Z_1\| = \sqrt{\sum_{i=1}^d |x_i - z_1|^2}$ ,  $D_{i2} = \|x_i - Z_2\|$ 】

如果  $Z_3$  存在, 则计算  $D_j = \max\{\min(D_{i1}, D_{i2}, D_{i3})\}$ ,  $i = 1, 2, \dots, n$ , 若  $D_j > \theta \cdot D_{12}$ , 则建立第四个聚合中心。依次类推, 直到最大最小距离不大于  $\theta \cdot D_{12}$  时, 结束寻找聚合中心的计算。

本例中已知样本:  $x_1 = (0,0)^T$ ,  $x_2 = (3,8)^T$ ,  $x_3 = (2,2)^T$ ,  $x_4 = (1,1)^T$ ,  $x_5 = (5,3)^T$ ,  $x_6 = (4,8)^T$ ,  $x_7 = (6,3)^T$ ,  $x_8 = (5,4)^T$ ,  $x_9 = (6,4)^T$ ,  $x_{10} = (7,5)^T$ 。

样本	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$	$x_9$	$x_{10}$
到 $Z_1$ 的距离	0	$\sqrt{73}$	$\sqrt{8}$	$\sqrt{2}$	$\sqrt{34}$	$\sqrt{80}$	$\sqrt{45}$	$\sqrt{41}$	$\sqrt{52}$	$\sqrt{74}$
到 $Z_2$ 的距离	$\sqrt{80}$	$\sqrt{2}$	$\sqrt{40}$	$\sqrt{58}$	$\sqrt{26}$	0	$\sqrt{29}$	$\sqrt{17}$	$\sqrt{22}$	$\sqrt{18}$
$\min(D_{i1}, D_{i2})$	0	$\sqrt{2}$	$\sqrt{8}$	$\sqrt{2}$	$\sqrt{26}$	0	$\sqrt{29}$	$\sqrt{17}$	$\sqrt{22}$	$\sqrt{18}$



注意  $x_7$  所在第列， $\sqrt{29}$  在  $\min(D_{i1}, D_{i2})$  中为最大的，而且  
 $D_l = \sqrt{29} > \theta \cdot \sqrt{80}$ ，一般取  $\theta = \frac{1}{2}$ 。所以， $Z_3 = x_7$ 。

上图中只有三个集合中心， $Z_1 = x_1$ ， $Z_2 = x_6$ ， $Z_3 = x_7$ 。

③ 按最近邻原则把所有样本归属于距离最近的聚合中心，有：

$\{x_1, x_3, x_4\} \in Z_1$ ， $\{x_2, x_6\} \in Z_2$ ， $\{x_5, x_7, x_8, x_9, x_{10}\} \in Z_3$ 。

④ 按照某聚类准则考查聚类结果，若不满意，则重选  $\theta$ ，第一个聚合中心  $Z_1$ ，返回到②，直到满意，算法结束。



## 4.4 系统聚类

- 系统聚类：先把每个样本作为一类，然后根据它们间的相似性和相邻性聚合。
- 相似性、相邻性一般用距离表示
- (1) 两类间的距离



1、最短距离：两类中相距最近的两样品间的距离。

$$D_{pq} = \min_{\substack{x_i \in \omega_p \\ x_j \in \omega_q}} d_{ij}$$





- 2、最长距离：两类中相距最远的两个样本间的距离。

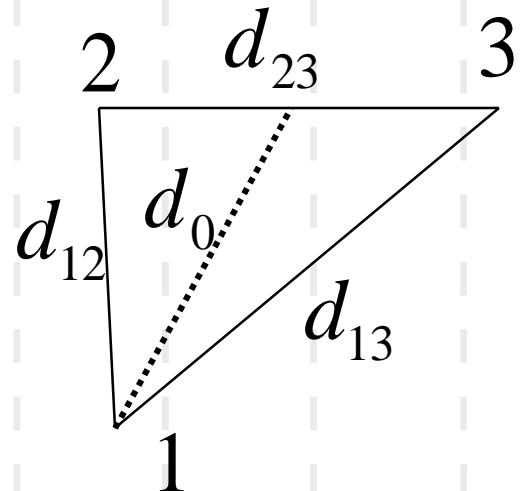
$$D_{pq} = \max_{\substack{x_i \in \omega_p \\ x_j \in \omega_q}} d_{ij}$$

- 3、中间距离：最短距离和最长距离都有片面性，因此有时用中间距离。设 $\omega_1$ 类和 $\omega_{23}$ 类间的最短距离为 $d_{12}$ ，最长距离为 $d_{13}$ ， $\omega_{23}$ 类的长度为 $d_{23}$ ，则中间距离为：

$$d_0^2 = \frac{1}{2} d_{12}^2 + \frac{1}{2} d_{13}^2 - \frac{1}{4} d_{23}^2$$



- 上式推广为一般情况：





$$d_0^2 = \frac{1}{2} d_{12}^2 + \frac{1}{2} d_{13}^2 + \beta d_{23}^2$$

其中  $\beta$  为参数,  $-\frac{1}{4} \leq \beta \leq 0$

- 4、重心距离: 均值间的距离
- 5、类平均距离: 两类中各个元素两两之间的距离平方相加后取平均值

$$D_{pq}^2 = \frac{1}{N_p N_q} \sum_{\substack{x_i \in \omega_p \\ x_j \in \omega_q}} d_{ij}^2$$

其中:  $N_p$ :  $\omega_p$  样本数,  $N_q$ :  $\omega_q$  样本数

$d_{ij}$  为  $\omega_p$  类点  $i$  与  $\omega_q$  类点  $j$  之间的距离

$$D_{kl}^2 = \alpha_p D_{kp}^2 + \alpha_q D_{kq}^2 + \beta D_{pq}^2 + \gamma |D_{kp}^2 - D_{kq}^2|$$

	$\alpha_p$	$\alpha_q$	$\beta$	$\gamma$
最近距离法	1/2	1/2	0	-1/2
最远距离法	1/2	1/2	0	1/2
中间距离法	1/2	1/2	-1/4	0
重心距离法	$\frac{n_p}{n_p + n_q}$	$\frac{n_q}{n_p + n_q}$	$-\alpha_p \alpha_q$	0
平均距离法	$\frac{n_p}{n_p + n_q}$	$\frac{n_q}{n_p + n_q}$	0	0
可变平均法	$(1-\beta) \frac{n_p}{n_p + n_q}$	$(1-\beta) \frac{n_q}{n_p + n_q}$	< 1	0
可变法	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	< 1	0
离差平方和法	$\frac{n_k + n_p}{n_k + n_l}$	$\frac{n_k + n_q}{n_k + n_l}$	$-\frac{n_k}{n_k + n_l}$	0

吉  
祥  
慶

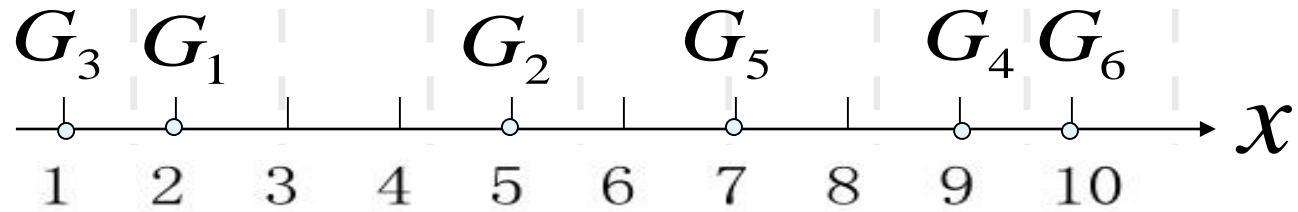
## (2) 系统聚类的算法

- 给定类型数目C
- 给定类间距离D
- 完成全部聚合过程，选取适当聚合级





- 例：如下图所示



- 设全部样本分为6类，  
作距离矩阵D(0)



吉祥如意

	$\omega_1$	$\omega_2$	$\omega_3$	$\omega_4$	$\omega_5$
$\omega_2$	9				
$\omega_3$	1	16			
$\omega_4$	49	16	64		
$\omega_5$	25	4	36	4	
$\omega_6$	64	25	81	1	9



吉  
祥  
慶

- 3、求最小元素:  $d_{31} = d_{64} = 1$
- 4、把  $\omega_1, \omega_3$  合并  $\omega_7 = (1, 3)$ 
  - $\omega_4, \omega_6$  合并  $\omega_8 = (4, 6)$
- 5、作距离矩阵 D(1)

	$\omega_7$	$\omega_2$	$\omega_8$
$\omega_2$	9		
$\omega_8$	49	16	
$\omega_5$	25	4	4

吉  
祥  
慶

吉  
祥  
慶

吉  
祥  
慶

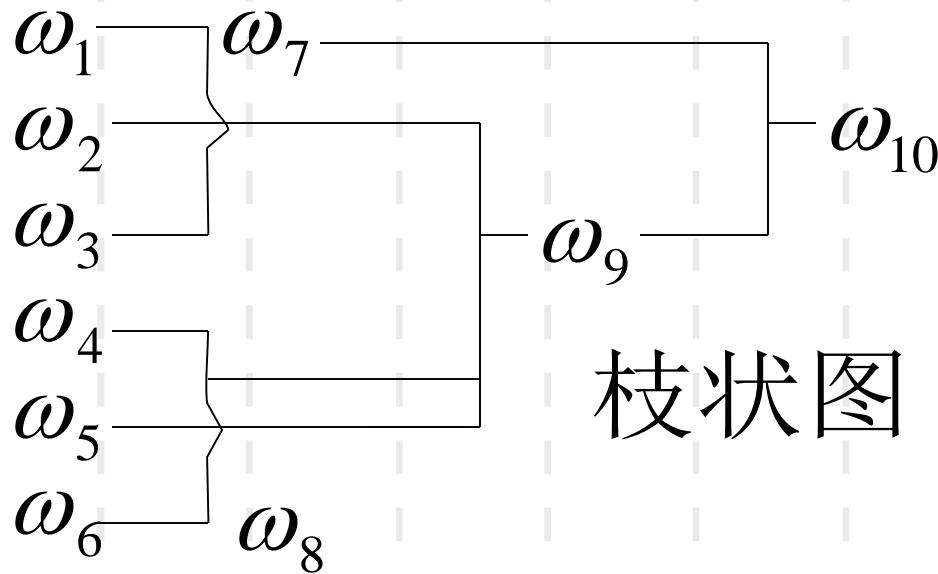
吉  
祥  
如  
意

- 6、若合并的类数没有达到要求，转3。  
否则停止。
- 3、求**最小**元素：

$$d_{52} = d_{58} = 4$$

- 4、 $\omega_8, \omega_5, \omega_2$ 合并,  $\omega_9 = (2, 5, 4, 6)$

吉  
祥  
如  
意



枝状图

吉  
祥  
如  
意

吉  
祥  
如  
意

吉  
祥  
如  
意

吉  
祥  
如  
意

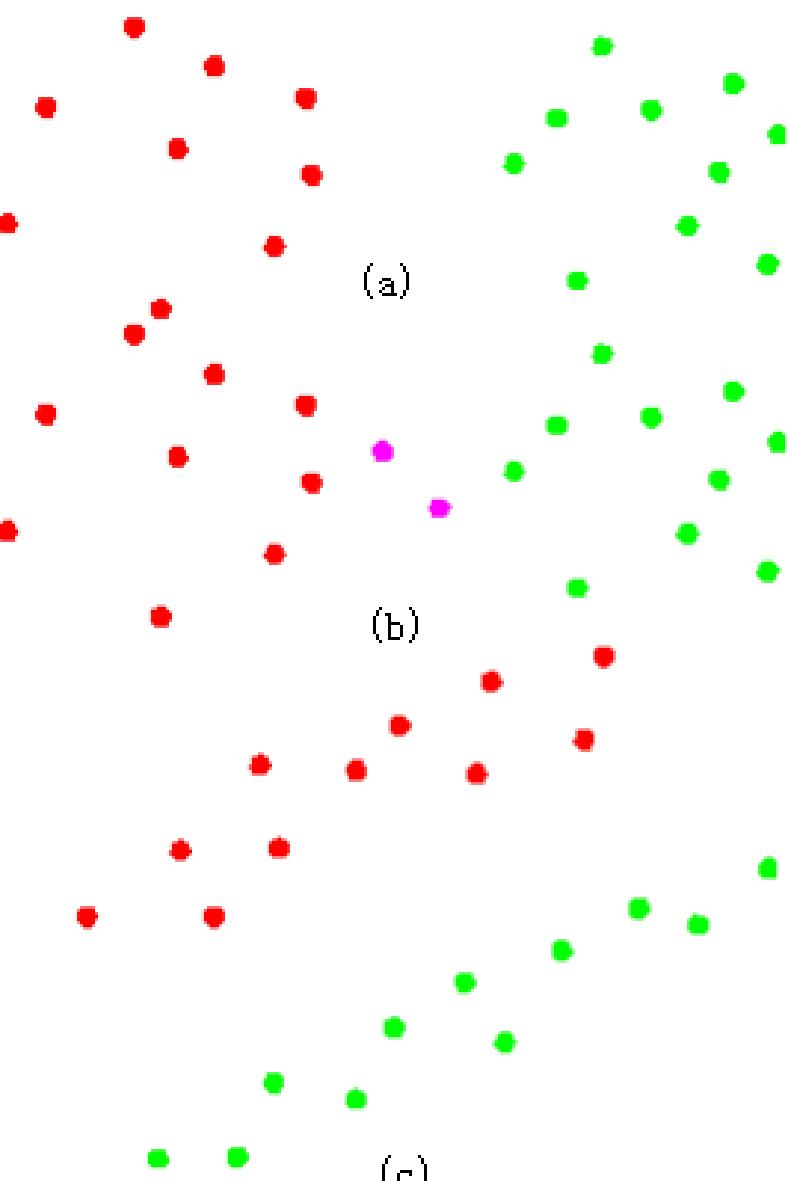
吉  
祥  
如  
意

吉祥如意

(a)

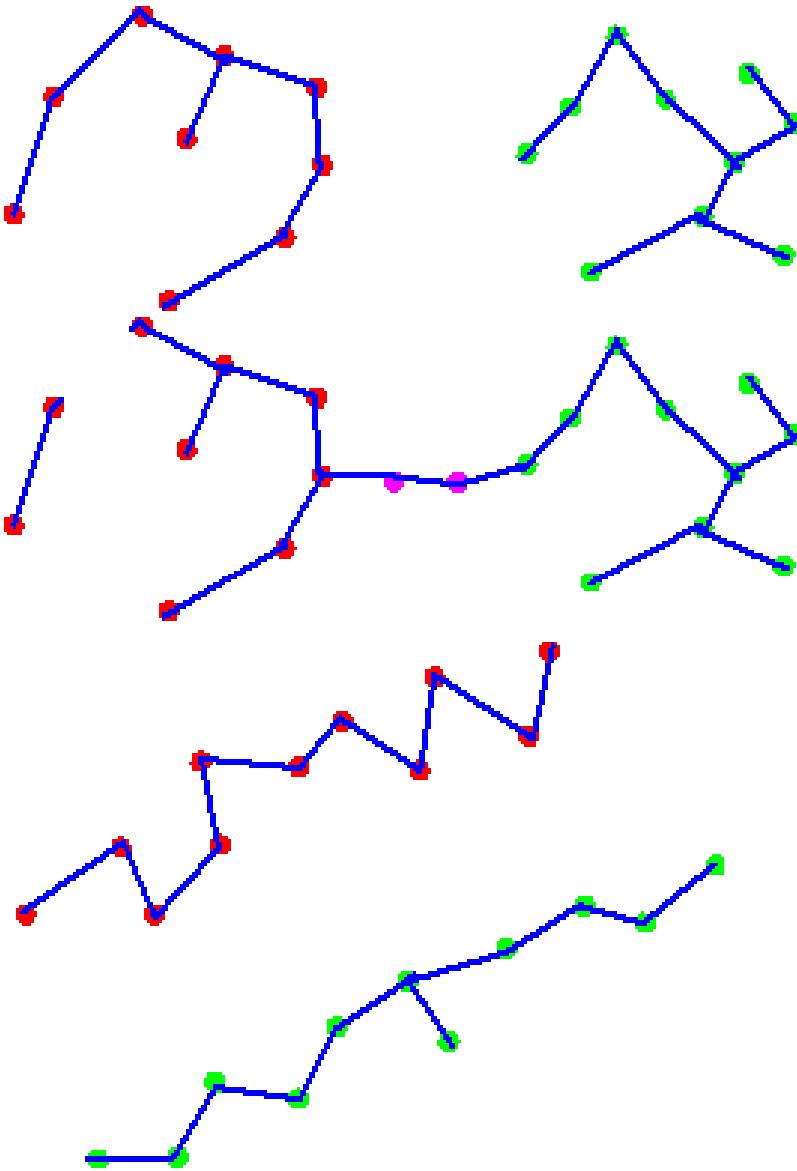
(b)

点集



吉祥如意

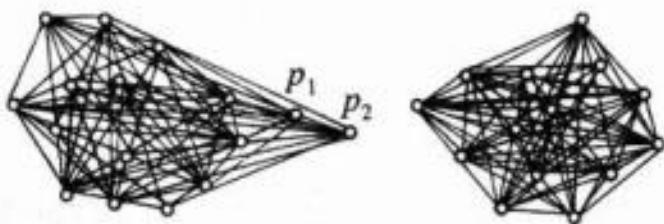
最短距离



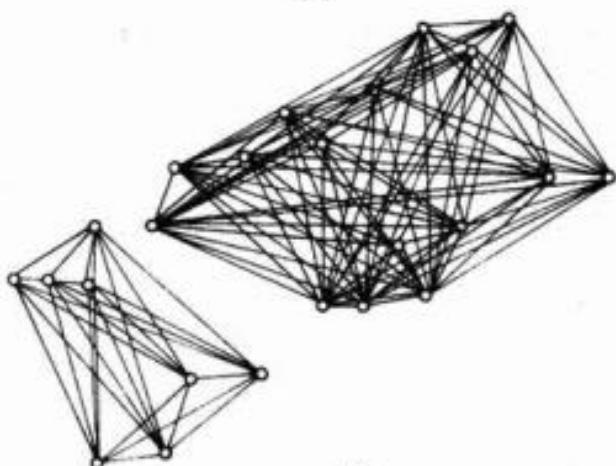
吉祥如意



(a)



(b)



(c)

最远距离





## 4.5 分解聚类

- 分解聚类：把全部样本作为一类，然后根据相似性、相邻性分解。
- 目标函数 两类均值方差

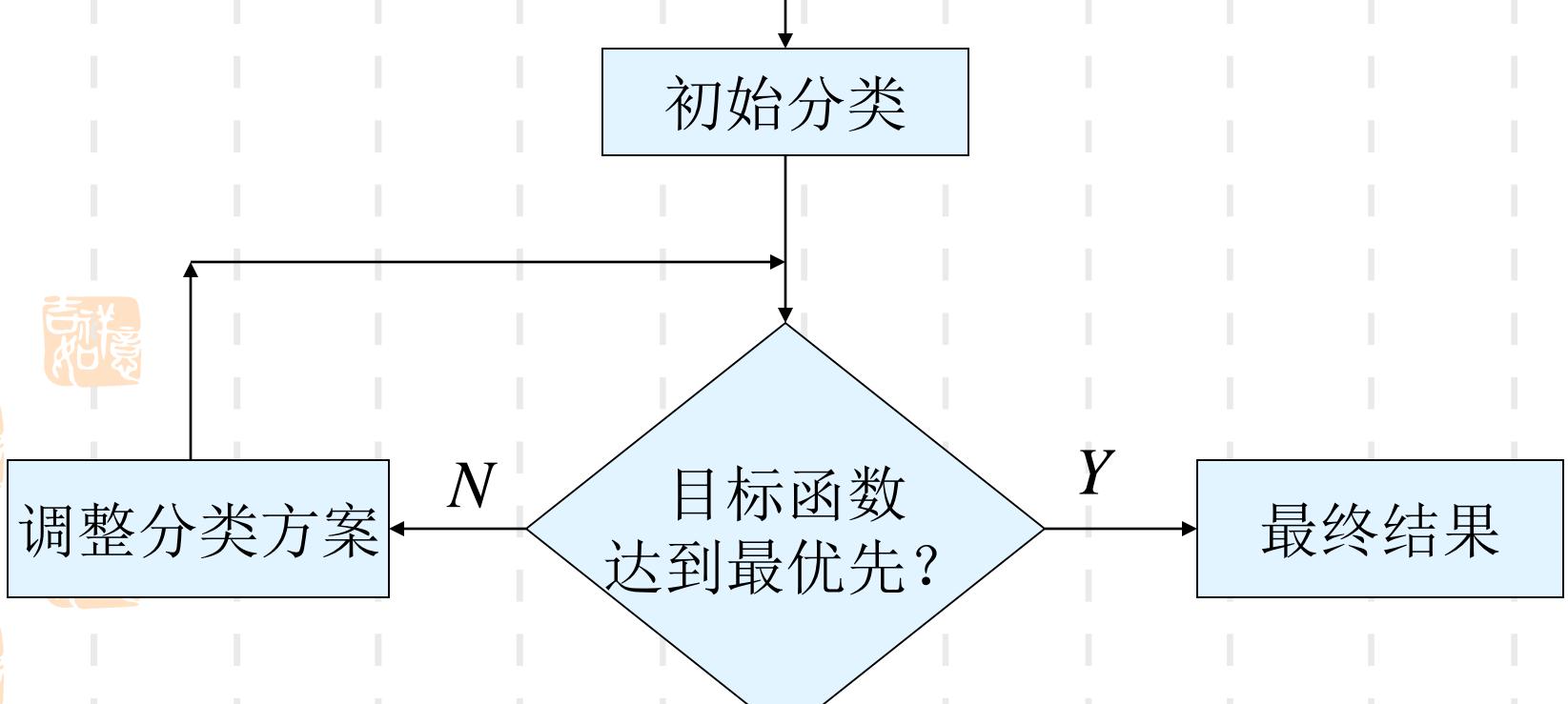
$$E = \frac{N_1 N_2}{N} (\bar{x}_1 - \bar{x}_2)^T (\bar{x}_1 - \bar{x}_2)$$

$N$ : 总样本数,  $N_1$  :  $\omega_1$ 类样本数

$N_2$ :  $\omega_2$ 类样本数,  $\bar{x}_1, \bar{x}_2$  : 两类均值

吉祥如意

# ❖ 分解聚类框图：





## ■ 对分算法：略

例：已知21个样本，每个样本取二个特征，原始资料矩阵如下表：

样本号	1	2	3	4	5	6	7	8	9	10
$x_1$	0	0	2	2	4	4	5	6	6	7
$x_2$	6	5	5	3	4	3	1	2	1	0

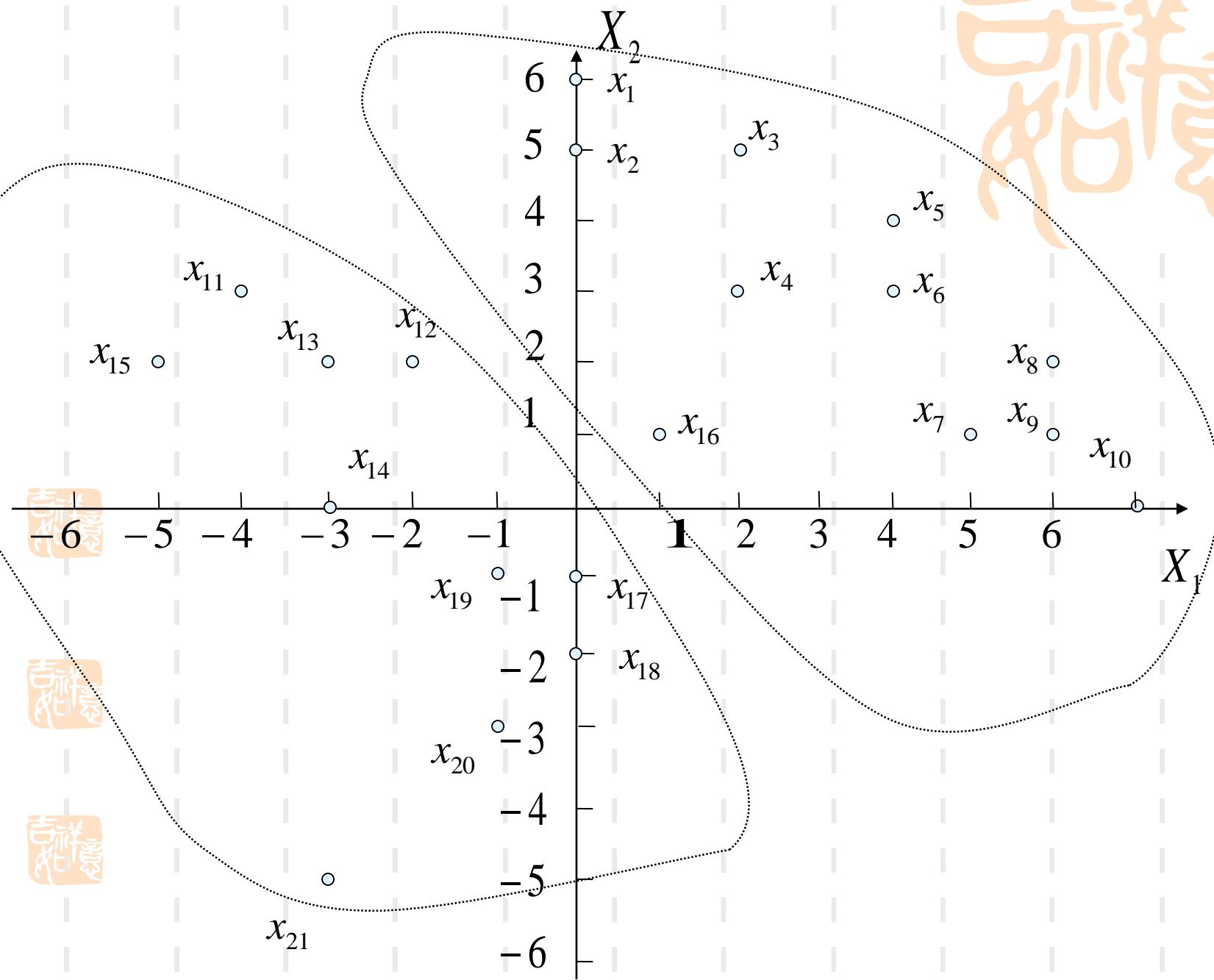
11	12	13	14	15	16	17	18	19	20	21
-4	-2	-3	-3	-5	1	0	0	-1	-1	-3
3	2	2	0	2	1	-1	-2	-1	-3	-5

吉祥如意

吉祥如意

吉祥如意

吉祥如意





解：第一次分类时计算所有样本，分别划到 $G_2$ 时的E值，找出最大的。

1、开始时， $G_1^{(0)} = (x_1, x_2, \dots, x_{21})$   $G_2^{(0)} = \text{空}$

$$\therefore \bar{x}_1^{(0)} = \begin{pmatrix} 0.714 \\ 1.333 \end{pmatrix} \quad \bar{x}_2^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad N_1^{(0)} = 21, N_2^{(0)} = 0$$

$$\therefore \text{目标函数 } E = \frac{N_1 N_2}{N} (\bar{x}_1 - \bar{x}_2)^T (\bar{x}_1 - \bar{x}_2) = 0$$





2、分别计算当  $x_1, x_2, \dots, x_{21}$  划入  $G_2$  时的E值  
把  $x_1$  划入  $G_2$  时有

$$\bar{x}_1^{(1)} = \bar{x}_1^{(0)} + \frac{\bar{x}_1^{(0)} - x_1}{N_1^{(0)} - 1}$$

$$= \begin{pmatrix} 0.714 \\ 1.333 \end{pmatrix} + \frac{\begin{pmatrix} 0.714 \\ 1.333 \end{pmatrix} - \begin{pmatrix} 0 \\ 6 \end{pmatrix}}{(21-1)} = \begin{pmatrix} 0.75 \\ 1.10 \end{pmatrix},$$

$$\bar{x}_2^{(1)} = \begin{pmatrix} 0 \\ 6 \end{pmatrix}$$

$$E = \frac{20 \times 1}{21} [0.75^2 + (1.10 - 6)^2] = 23.40$$



■ 然后再把  $x_2, x_3, \dots, x_{21}$  划入  $G_2$  时对应的  $E$  值，找出一个最大的  $E$  值。

把  $x_{21}$  划为  $G_2$  的  $E$  值最大。

$$\therefore G_1^{(1)} = (x_1, x_2, \dots, x_{20}), G_2^{(1)} = (x_{21})$$

$$\bar{x}_1 = \begin{pmatrix} 0.9 \\ 1.65 \end{pmatrix}, \bar{x}_2 = \begin{pmatrix} -3 \\ -5 \end{pmatrix}, N_1^{(1)} = 20, N_2^{(1)} = 1$$

$$E(1) = 56.6$$

再继续进行第二，第三次迭代...

计算出  $E(2), E(3), \dots$

吉  
祥  
慶  
喜

次数	$G_1 \rightarrow G_2$	E值
1	$x_{21}$	56.6
2	$x_{20}$	79.16
3	$x_{18}$	90.90
4	$x_{14}$	102.61
5	$x_{15}$	120.11
6	$x_{19}$	137.15
7	$x_{11}$	154.10
8	$x_{13}$	176.15
9	$x_{12}$	195.26
10	$x_{17}$	213.07
11	$x_{16}$	212.01





- 第10次迭代 $x_{17}$ 划入 $G_2$ 时，E最大。于是分成以下两类：
- $\therefore G_1 = (x_1, x_2, \dots, x_{10}, x_{16})$
- $G_2 = (x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{17}, x_{18}, x_{19}, x_{20}, x_{21})$
- 每次分类后要重新计算 $\bar{x}_1, \bar{x}_2$ 的值。可用以下递推公式：
$$x_1^{(k+1)} = \bar{x}_1^{(k)} + (\bar{x}_1^{(k)} - x_i) / (N_1^{(k)} - 1)$$
$$x_2^{(k+1)} = \bar{x}_2^{(k)} - (\bar{x}_2^{(k)} - x_i) / (N_2^{(k)} + 1)$$

$\bar{x}_1^{(k)}, \bar{x}_2^{(k)}$ 是第 $k$ 步对分时两类均值，  
 $x_1^{(k+1)}, x_2^{(k+1)}$ 是下一次对分时把 $x_i$ 从 $G_1^{(k)}$ 划到 $G_2^{(k)}$ 时的两类均值。

$N_1^{(k)}, N_2^{(k)}$ 为二类样品数

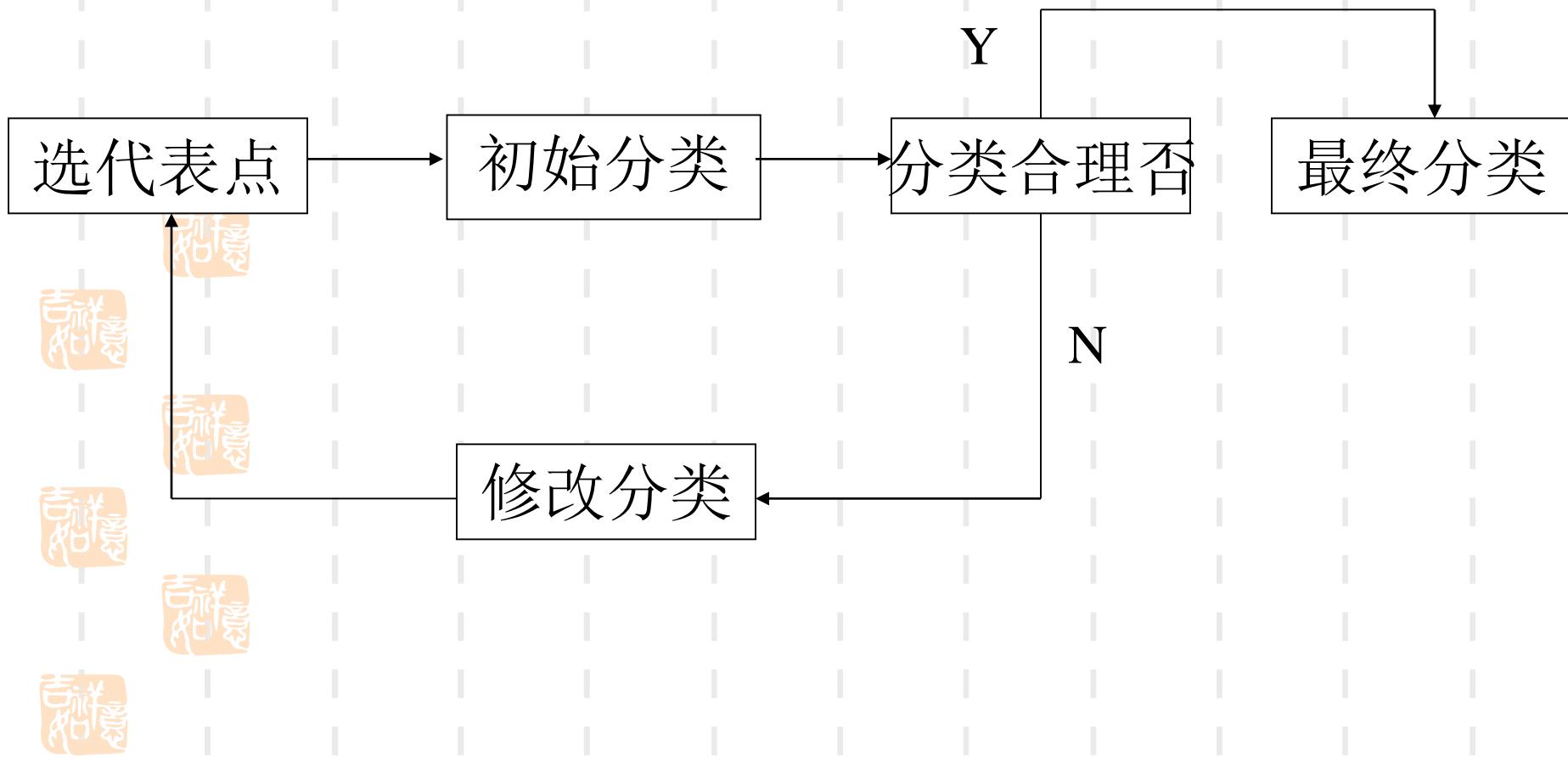
# 4.6 动态聚类——兼顾系统聚类和分解聚类

## 一、动态聚类的方法概要

- ① 先选定某种距离作为样本间的相似性的度量；
- ② 确定评价聚类结果的准则函数；
- ③ 给出某种初始分类，用迭代法找出使准则函数取极值的最好的聚类结果。

吉祥如意

# 动态聚类框图



## 二、代表点的选取方法：代表点就是初始分类的聚类中心数k

- ①凭经验选代表点，根据问题的性质、数据分布，从直观上看来较合理的代表点k；
- ②将全部样本随机分成k类，计算每类重心，把这些重心作为每类的代表点；

吉祥如意

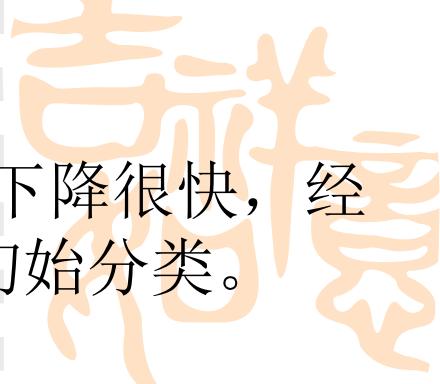
### ③ 按密度大小选代表点：

以每个样本作为球心，以 $d$ 为半径做球形；落在球内的样本数称为该点的密度，并按密度大小排序。首先选密度最大的作为第一个代表点，即第一个聚类中心。再考虑第二大密度点，若第二大密度点距第一代表点的距离大于 $d_1$ （人为规定的正数）则把第二大密度点作为第二代表点，否则不能作为代表点，这样按密度大小考察下去，所选代表点间的距离都大于 $d_1$ 。 $d_1$ 太小，代表点太多， $d_1$ 太大，代表点太小，一般选 $d_1=2d$ 。对代表点内的密度一般要求大于 $T$ 。 $T>0$ 为规定的一个正数。

### ④ 用前 $k$ 个样本点作为代表点。

### 三、初始分类和调整

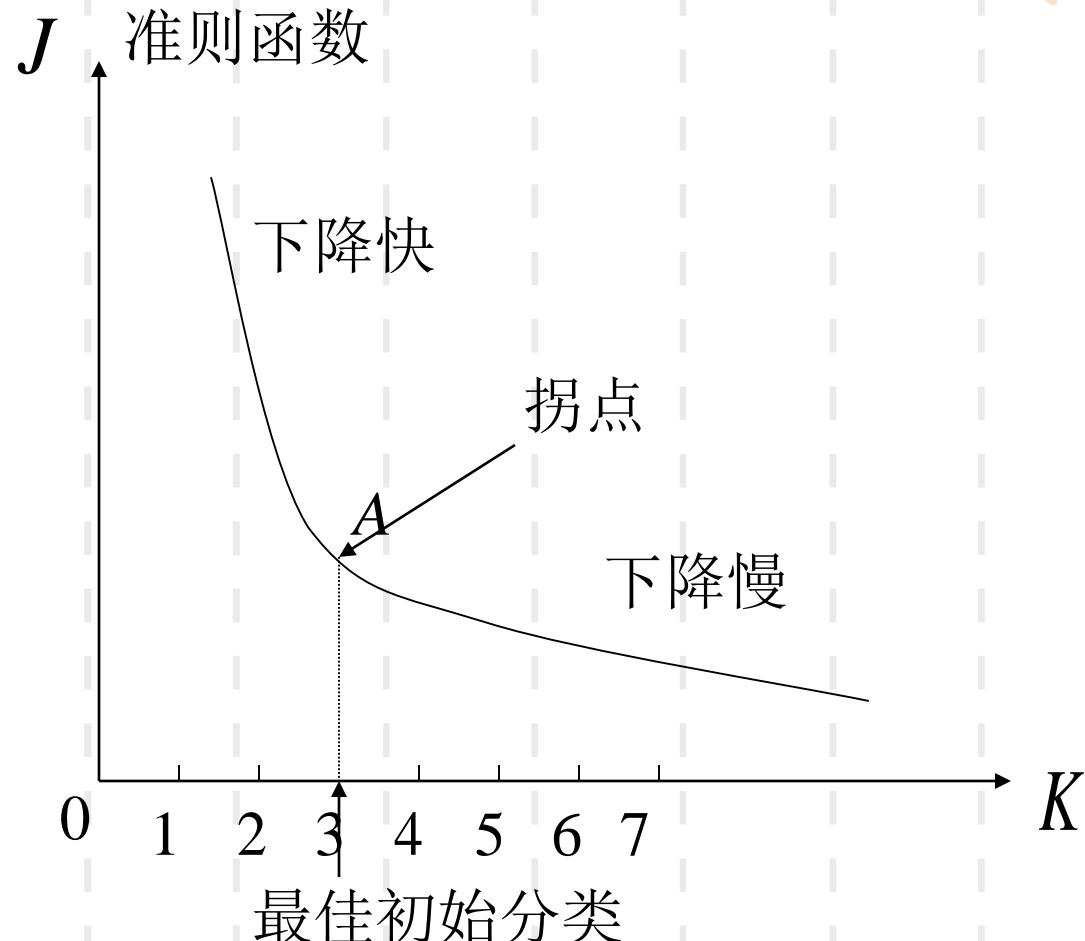
- ① 选一批代表点后，代表点就是聚类中心，计算其它样本到聚类中心的距离，把所有样本归于最近的聚类中心点，形成初始分类，再重新计算各聚类中心，称为成批处理法。
- ② 选一批代表点后，依次计算其它样本的归类，当计算完第一个样本时，把它归于最近的一类，形成新的分类。再计算新的聚类中心，再计算第二个样本到新的聚类中心的距离，对第二个样本归类。即每个样本的归类都改变一次聚类中心。此法称为逐个处理法。
- ③ 直接用样本进行初始分类，先规定距离 $d$ ，把第一个样本作为第一类的聚类中心，考察第二个样本，若第二个样本距第一个聚类中心距离小于 $d$ ，就把第二个样本归于第一类，否则第二个样本就成为第二类的聚类中心，再考虑其它样本，根据样本到聚类中心距离大于还是小于 $d$ ，决定分裂还是合并。



④

#### 最佳初始分类。

如图所示，随着初始分类 $k$ 的增大，准则函数下降很快，经过拐点A后，下降速度减慢。拐点A就是最佳初始分类。



# 四、c—均值与ISODATA聚类算法

## 1. C—均值聚类算法

聚类准则函数是误差平方和准则:  $J_c = \sum_{j=1}^c \sum_{k=1}^{n_j} \|x_k - m_j\|^2$

### (1) C—均值算法 (一)

① 给出  $n$  个混合样本, 令  $I=1$ , 表示迭代运算次数, 选取  $c$  个初始聚合中心  $Z_j(1)$ ,  $j=1,2,\dots,c$ ;

② 计算每个样本与聚合中心的距离  $D(x_k, Z_j(I))$ ,  $k=1,2,\dots,n$ ,

$j=1,2,\dots,c$ 。

若  $D(x_k, Z_i(I)) = \min_{j=1,2,\dots,c} \{D(x_k, Z_j(I)), k=1,2,\dots,n\}$ , 则  $x_k \in w_i$ 。

③ 计算  $c$  个新的集合中心:  $Z_j(I+1) = \frac{1}{n_j} \sum_{k=1}^{n_j} x_k^{(j)}$ ,  $j=1,2,\dots,c$ 。

④ 判断: 若  $Z_j(I+1) \neq Z_j(I)$ ,  $j=1,2,\dots,c$ , 则  $I=I+1$ , 返回②, 否则算法结束。

吉  
祥  
如  
意

## 算法特点：

- ① 每次迭代中都要考查每个样本的分类是否正确，若不正确，就要调整，在全部样本调整完之后，再修改聚合中心，进入下一次迭代。如果在某一个迭代运算中，所有的样本都被正确分类，则样本不会调整，聚合中心也不会有变化，也就是收敛了。
- ②  $c$ 个初始聚合中心的选择对聚类结果有较大影响。

吉祥如意

下面通过样本移动对  $J_c$  的影响来修改上述算法。

假定  $x_k^{(i)}$  由样本的子集  $X_i$  移入另一个子集  $X_j$ ，那么这次移动只影

响两个类型  $w_i$  和  $w_j$  的聚合中心  $Z_i$  和  $Z_j$ ，以及两类的类内误差平方和

$J_{c_i}$ 、 $J_{c_j}$ 。

移动后聚合中心：

$$Z_i(I+1) = \frac{1}{n_i - 1} [n_i \cdot Z_i(I) - x_k^{(i)}] = Z_i(I) + \frac{1}{n_i - 1} [Z_i(I) - x_k^{(i)}]$$

$$Z_j(I+1) = \frac{1}{n_j + 1} [n_j \cdot Z_j(I) + x_k^{(i)}] = Z_j(I) - \frac{1}{n_j + 1} [Z_j(I) - x_k^{(i)}]$$



因此：

$$J_{c_i}(I+1) = J_{c_i}(I) - \frac{n_i}{n_i - 1} \|x_k^{(i)} - Z_i(I)\|^2$$

$$J_{c_j}(I+1) = J_{c_j}(I) + \frac{n_j}{n_j + 1} \|x_k^{(j)} - Z_j(I)\|^2$$

如果： $\frac{n_j}{n_j + 1} \|x_k^{(j)} - Z_j(I)\|^2 < \frac{n_i}{n_i - 1} \|x_k^{(i)} - Z_i(I)\|^2$



那么， $J_c$  的值会减小为：

$$J_c(I+1) = J_c(I) - \left[ \frac{n_i}{n_i - 1} \|x_k^{(i)} - Z_i(I)\|^2 - \frac{n_j}{n_j + 1} \|x_k^{(j)} - Z_j(I)\|^2 \right]$$

根据上述分析，对C—均值算法作改进，变为C—均值算法(二)。





## (2) C—均值算法 (二)

- ① 给定  $n$  个混合样本，令  $I=1$  (迭代次数)，选取  $c$  个初始聚合中心  $Z_j(1)$ ， $j=1,2,\dots,c$ 。
- ② 计算每个样本与每个聚合中心的距离  $D(x_k, Z_j(I))$ ， $k=1,2,\dots,n$ ，  
 $j=1,2,\dots,c$ 。

若： $D(x_k, Z_i(1)) = \min_{j=1,2,\dots,c} \{D(x_k, Z_j(1)), k=1,2,\dots,n\}$ ，则  $x_k \in w_i$ 。

③ 令  $I=I+1=2$ ，计算新的聚合中心。 $Z_j(2)=\frac{1}{n_j} \sum_{k=1}^{n_j} x_k^{(j)}$ ， $j=1,2,\dots,c$

计算误差平方和  $J_c$  值： $J_c(2)=\sum_{j=1}^c \sum_{k=1}^{n_j} \|x_k^{(j)} - Z_i(2)\|^2$



④ 对每个聚合中的每个样本，计算：

$$\rho_{ii} = \frac{n_i}{n_i - 1} \|x_k^{(i)} - Z_i(I)\|^2, \quad i = 1, 2, \dots, c$$

$\rho_{ii}$  表示  $J_c$  减少的部分。

$$\rho_{ij} = \frac{n_j}{n_j + 1} \|x_k^{(i)} - Z_j(I)\|^2, \quad j = 1, 2, \dots, c, \quad j \neq i$$

$\rho_{ij}$  表示  $J_c$  增加的部分。

令：  $\rho_{il} = \min_{j \neq i} \{\rho_{ij}\}$ ， 若  $\rho_{il} < \rho_{ii}$ ， 则把样本  $x_k^{(i)}$  移到聚合中心  $w_l$  中，并修改聚合中心和  $J_c$  值。

$$Z_i(I+1) = Z_i(I) + \frac{1}{n_i - 1} [Z_i(I) - x_k^{(i)}]$$

$$Z_l(I+1) = Z_l(I) - \frac{1}{n_l + 1} [Z_l(I) - x_k^{(i)}]$$

$$J_c(I+1) = J_c(I) - (\rho_{ii} - \rho_{il})$$

⑤ 判断：若  $J_c(I+1) < J_c(I)$ ，则  $I = I + 1$ ，返回④。否则，算法结束。

例：现有混合样本集  $X$ ，共有样本20个，分布如右图所示，类型数目  $c=2$ 。试用C—均值算法进行聚类分析。

解：①  $c=2$ ，选2个集合中心： $Z_1(1)=x_1$ ， $Z_2(1)=x_2$ ，则：

$$Z_1(1)=(0,0)^T, \quad Z_2(1)=(1,0)^T, \quad \text{令 } I=1.$$

② 选用欧氏距离作为相似性度量，计算各样本到  $Z_1(1)$ 、 $Z_2(1)$  的距离，并把  $x_k$  归于最近的聚合范围内，有：

吉祥  $\|x_1 - Z_1(1)\| < \|x_1 - Z_2(1)\|$ ，所以  $x_1 \in Z_1(1)$

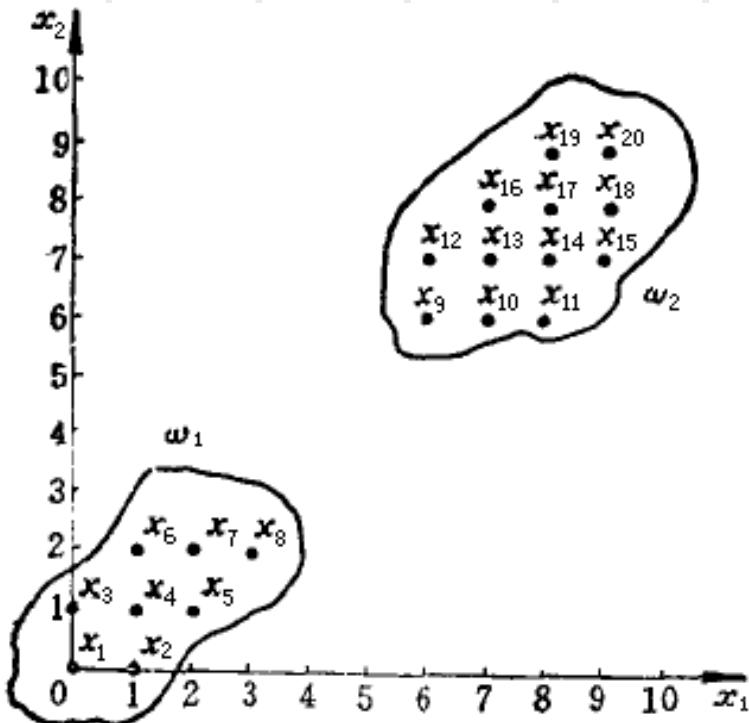
吉祥  $\|x_2 - Z_2(1)\| < \|x_1 - Z_1(1)\|$ ，所以  $x_2 \in Z_2(1)$

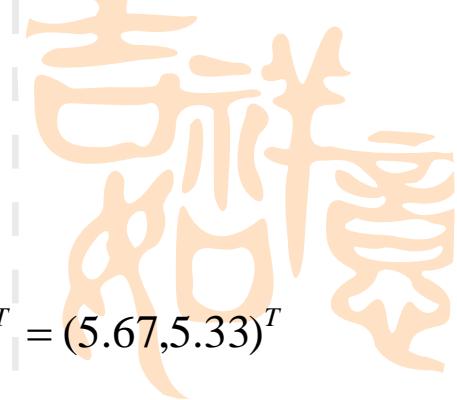
吉祥  $\|x_3 - Z_1(1)\| < \|x_3 - Z_2(1)\|$ ，所以  $x_3 \in Z_1(1)$

吉祥  
……

吉祥 可得到： $w_1 : X_1 = \{x_1, x_3\}, n_1 = 2$

吉祥  $w_2 : X_2 = \{x_2, x_4, x_5, \dots, x_{20}\}, n_2 = 18$





### ③ 计算新的聚合中心;

$$Z_1(2) = \frac{1}{2}(x_1 + x_3) = (0,0.5)^T$$

$$Z_2(2) = \frac{1}{18} \sum_{x \in X_2} x = \left( \frac{3+4+3+12+21+32+2}{18}, \frac{2+6+18+28+24+18}{18} \right)^T = (5.67, 5.33)^T$$

④ 判断  $Z_j(2) \neq Z_j(1)$  ,  $j = 1, 2$  。 令  $I = I + 1 = 2$  , 返回②。

② 计算各样本到  $Z_j(2)$  ,  $j = 1, 2$  的欧氏距离。有:

  $\|x_k - Z_1(2)\| < \|x_k - Z_2(2)\|, k = 1, 2, \dots, 8$        $\|x_k - Z_2(2)\| < \|x_k - Z_1(2)\|, k = 9, 10, \dots, 20$



得到新的聚合:  $w_1: X_1 = \{x_1, x_2, \dots, x_8\}, n_1 = 8$  ,  $w_2: X_2 = \{x_9, x_{10}, \dots, x_{20}\}, n_2 = 12$

### ③ 计算聚合中心:

  $Z_1(3) = \frac{1}{n_1} \sum_{x \in X_1} x = \frac{1}{8}(x_1 + x_2 + \dots + x_8) = \left( \frac{3+4+3}{8}, \frac{3+6}{8} \right)^T = (1.25, 1.13)^T$

  $Z_2(3) = \frac{1}{n_2} \sum_{x \in X_2} x = \frac{1}{12}(x_9 + x_{10} + \dots + x_{20}) = \left( \frac{12+21+32+27}{12}, \frac{18+28+24+18}{12} \right)^T = (7.67, 7.33)^T$



吉  
祥  
如  
意

④ 判断:  $Z_j(3) \neq Z_j(2)$ ,  $j = 1, 2$ 。令  $I = I + 1 = 2$ ,  
返回②。

② 聚类结果无变化。

③ 聚合中心无变化,  $Z_1(4) = Z_1(3)$ ,

$Z_2(4) = Z_2(3)$ 。

吉  
祥  
如  
意

④ 判断:  $Z_j(4) = Z_j(3)$ ,  $j = 1, 2$  算法结束。

聚类结果如上图所示。

吉  
祥  
如  
意

吉  
祥  
如  
意

吉  
祥  
如  
意

# 四、c—均值与ISODATA聚类算法（续）

## 2. ISODATA聚类算法

- ISODATA算法: Iterative Self—Organizing Data Analysis Techniques Algorithm, 迭代自组织的数据分析算法。
- ISODATA算法特点: 可以通过类的自动合并（两类合一）与分裂（一类分为二），得到较合理的类型数目 $c$ 。

具体算法步骤:

(1) 给定控制参数

$k$ : 预期的聚类中心数目。



$\theta_n$ : 每一聚类中最少的样本数目，如果少于此数就不能作为一个独立的聚类。

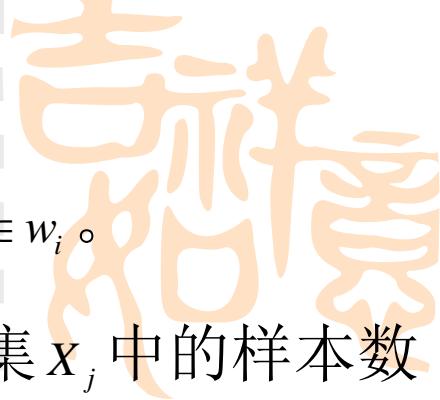
$\theta_s$ : 一个聚类域中样本距离分布的标准差（阈值）。

$\theta_c$ : 两个聚类中心之间的最小距离，如果小于此数，两个聚类合并。

$L$ : 每次迭代允许合并的最大聚类对数目。

$I$ : 允许的最多迭代次数。

给定 $n$ 个混合样本，令 $J=1$ （迭代次数），预选 $c$ 个起始聚合中心，  
 $Z_j(J)$ ， $j=1,2,\dots,c$ 。



(2) 计算每个样本与聚合中心距离:  $D(x_k, Z_j(J))$ 。

若:  $D(x_k, Z_j(J)) = \min_{j=1,2,\dots,c} \{D(x_k, Z_j(J)), k=1,2,\dots,n\}$ , 则:  $x_k \in w_i$ 。

把全部样本划分到 $c$ 个聚合中去, 且 $n_j$ 表示各子集 $X_j$ 中的样本数目。

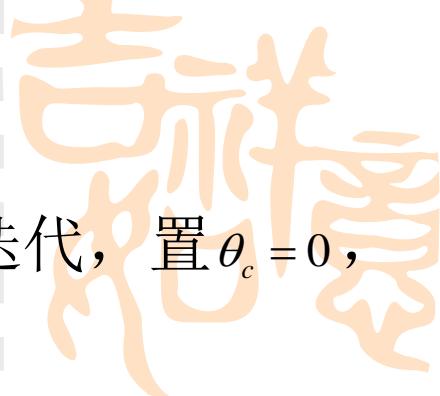
(3) 判断: 若  $n_j < \theta_n$ ,  $j=1,2,\dots,c$  则舍去子集  $X_j$ ,  $c=c-1$ , 返回②。

(4) 计算修改聚合中心:  $Z_j(J) = \frac{1}{n_j} \sum_{k=1}^{n_j} x_k^{(j)}$ ,  $j=1,2,\dots,c$ 。

(5) 计算类内距离平均值  $\bar{D}_j$ :  $\bar{D}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} D(x_k^{(j)}, Z_j(J))$ ,  $j=1,2,\dots,c$

(6) 计算类内总平均距离  $\bar{D}$  (全部样本对其相应聚类中心的总平均距离):  $\bar{D} = \frac{1}{n} \sum_{j=1}^c n_j \cdot \bar{D}_j$





(7) 判别分裂、合并及迭代运算等步骤。

(a) 如迭代运算次数已达 $I$ 次，即最后一次迭代，置 $\theta_c = 0$ ，

跳到(11)，运算结束。

(b) 如 $c \leq \frac{K}{2}$ ，即聚类中心的数目等于或不到规定值的一半，

则转(8)，将已有的聚类分裂。

(c) 如迭代运算的次数是偶数，或 $c \geq 2K$ ，则不进行分裂，跳

到(11)，若不符合上述两个条件，则进入(8)，进行分裂处理。

(8) 计算每个聚合的标准偏差向量： $\sigma_j = (\sigma_{j1}, \sigma_{j2}, \dots, \sigma_{jd})^T$ 。

每个分量为： $\sigma_{ji} = \sqrt{\frac{1}{n_j} \sum_{x \in X_j} (x_i - Z_{ji}(J))^2}$ ， $i = 1, 2, \dots, d$ ， $j = 1, 2, \dots, c$ 。

$x_i$  表示 $x$ 的第 $i$ 个分量， $Z_{ji}$  表示 $Z_j$ 的第 $i$ 个分量。 $d$  为维数。



(9) 求出每个聚合的最大标准偏差分量  $\sigma_{j \max}$  :

$$\sigma_{j \max} = \max_{i=1,2,\dots,d} \{\sigma_{ji}\}, \quad j=1,2,\dots,c.$$

(10) 考查  $\sigma_{j \max}$ ,  $j=1,2,\dots,c$  若有  $\sigma_{j \max} > \theta_s$ , 同时满足以下两条件之一,

(a)  $\bar{D}_j > \bar{D}$  及  $n_j > 2(\theta_n + 1)$ , (样本数目超过规定值一倍以上)。

(b)  $c \leq \frac{K}{2}$ 。

则把该集合分为两个新的聚合, 聚合中心分别为:

$$Z_j^+(J) = Z_j(J) + r_j, \quad Z_j^-(J) = Z_j(J) - r_j$$

其中:  $r_j = k\sigma_j$  或  $r_j = k[0,0,\dots,\sigma_{j \ max},0,\dots,0]^T$ ,  $0 < k \leq 1$ 。

令:  $c = c + 1$ ,  $J = J + 1$  返回(2)

其中,  $K$  的选择很重要, 应使  $x_j$  中的样本到  $Z_j^+(J)$  和  $Z_j^-(J)$  的距离不同, 但又使样本全部在这两个集合中。



(11) 计算两两聚合中心间的距离  $D_{ij}$  :

$$D_{ij} = D(Z_i(J), Z_j(J)) , \quad i = 1, 2, \dots, c-1 , \quad j = i+1, \dots, c .$$

(12) 比较  $D_{ij}$  与  $\theta_c$ ，并把小于  $\theta_c$  的  $D_{ij}$  按递增次序排队:

$$D_{i_1j_1} < D_{i_2j_2} < \dots < D_{i_Lj_L} , \quad L \text{ 为给定的合并参数。}$$

(13) 考查(12)中的不等式，对每一个  $D_{i_Lj_L}$ ，相应有两个聚类中心  $Z_{i_L}$  和  $Z_{j_L}$ ，假使在同一次迭代中，还没把  $Z_{i_L}$  和  $Z_{j_L}$  合并，则把两者合并，合并后中心为:

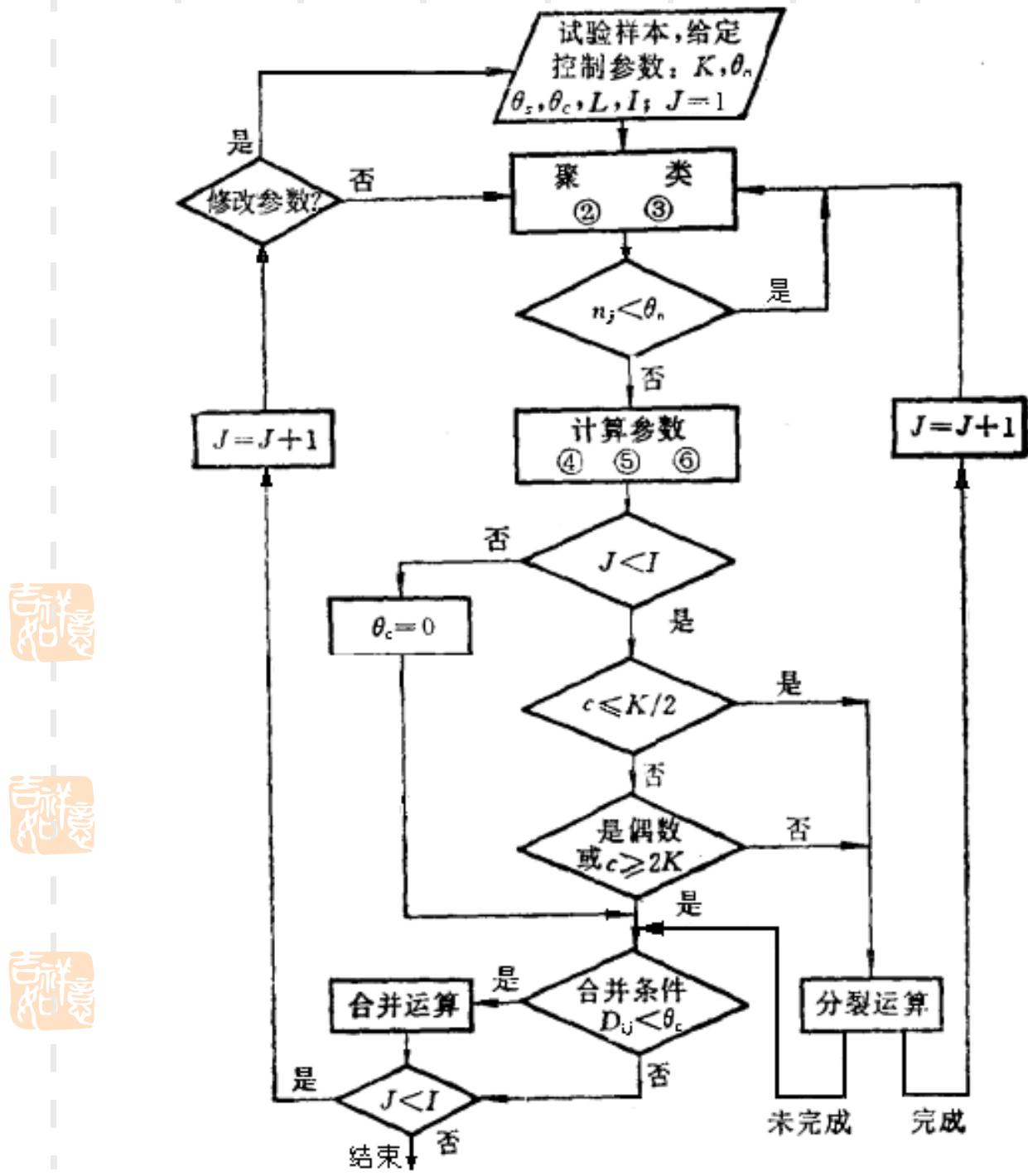
$$Z_L(J) = \frac{1}{n_{i_L} + n_{j_L}} [n_{i_L} \cdot Z_{i_L}(J) + n_{j_L} \cdot Z_{j_L}(J)] , \quad \text{令 } c = c + 1$$

(14) 若  $J < I$ ，则  $J = J + 1$ ，如果修改给定参数则返回(1)，不修改参数返回(2)，否则  $J = I$ ，算法结束。



注意：(8)~(10)步为分裂，(11)~(13)为合并。

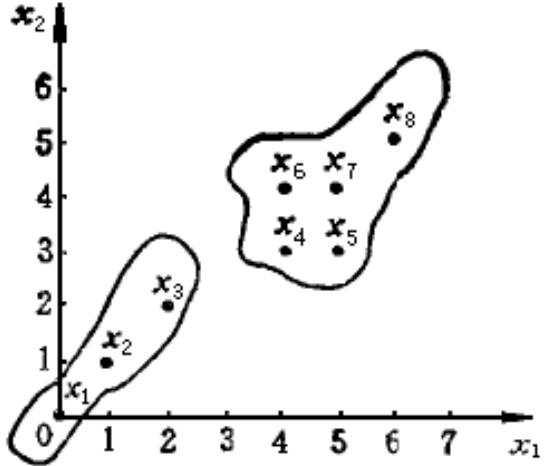
吉祥如意



吉  
祥  
如  
意

例：有一混合样本集，如下图所示，试用ISODATA进行聚类分析。

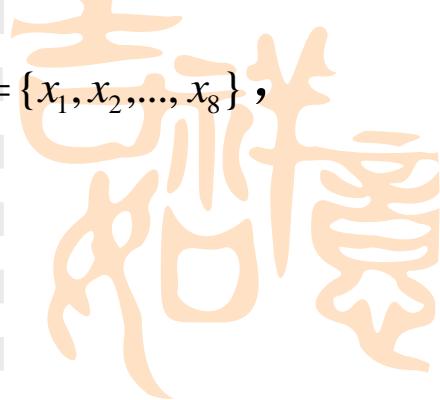
解：如下图所示，样本数目  $n=8$ ，取类型数目初始值  $c=1$ ，执行 ISODATA 算法：



(1) 给定参数（可以通过迭代过程修正这些参数）：

$$K = 2, \theta_n = 2, \theta_s = 1, \theta_c = 4, L = 0, I = 4$$

预选  $x_1$  为聚合中心，即： $Z_1 = (0,0)^T$ 。令  $J = 1$ ，迭代次数。



(2) 聚类: 因只有一个聚合中心  $Z_1 = (0,0)^T$ , 故  $w_1: X_1 = \{x_1, x_2, \dots, x_8\}$ ,  $n_1 = 8$ 。

(3) 因  $n_1 = 8 > \theta_n$ , 没有子集抛弃。

(4) 计算新聚合中心:

$$Z_1 = \frac{1}{8} \sum_{x \in X_1} x = \left( \frac{1+2+8+10+6}{8}, \frac{1+2+6+8+5}{8} \right) = (3.38, 2.75)^T$$

(5) 计算类内平均距离:

$$\overline{D}_1 = \frac{1}{n_1} \sum_{x \in X_1} \|x - Z_1\|$$

$$= \frac{1}{8} \left[ \sqrt{\left(\frac{27}{8}\right)^2 + \left(\frac{22}{8}\right)^2} + \sqrt{\left(\frac{19}{8}\right)^2 + \left(\frac{14}{8}\right)^2} + \sqrt{\left(\frac{11}{8}\right)^2 + \left(\frac{6}{8}\right)^2} + \sqrt{\left(\frac{5}{8}\right)^2 + \left(\frac{2}{8}\right)^2} + \right.$$

$$\left. \sqrt{\left(\frac{13}{8}\right)^2 + \left(\frac{2}{8}\right)^2} + \sqrt{\left(\frac{5}{8}\right)^2 + \left(\frac{10}{8}\right)^2} + \sqrt{\left(\frac{13}{8}\right)^2 + \left(\frac{10}{8}\right)^2} + \sqrt{\left(\frac{21}{8}\right)^2 + \left(\frac{18}{8}\right)^2} \right]$$

$$= 2.26$$



(6) 计算类内总平均距离:  $\bar{D} = \bar{D}_1 = 2.26$ 。

(7) 不是最后一次迭代, 且  $c = \frac{k}{2}$  转(8)

(8) 计算聚合  $x_1$  中的标准偏差  $\sigma_1$ :

$$\sigma_1 = (\sigma_{11}, \sigma_{12})^T$$

$$\sigma_{11} = \sqrt{\frac{1}{8} \sum_{x \in X_j} (x_1 - Z_{ji}(J))^2}$$

$$= \sqrt{\frac{1}{8} [(0 - \frac{27}{8})^2 + (1 - \frac{27}{8})^2 + (2 - \frac{27}{8})^2 + (4 - \frac{27}{8})^2 + (5 - \frac{27}{8})^2 + (4 - \frac{27}{8})^2 + (5 - \frac{27}{8})^2 + (6 - \frac{27}{8})^2]}$$

$$= \sqrt{3.98} = 1.99$$

$$\sigma_{12} = \sqrt{\frac{1}{8} [(\frac{22}{8})^2 + (\frac{14}{8})^2 + (\frac{6}{8})^2 + (\frac{2}{8})^2 + (\frac{22}{8})^2 + (\frac{10}{8})^2 + (\frac{10}{8})^2 + (\frac{18}{8})^2]} = 1.56$$

$$\sigma_1 = (1.99, 1.56)^T$$

(9)  $\sigma_1$  中的最大偏差分量为  $\sigma_{11} = 1.99$ , 即  $\sigma_{1\max} = 1.99$ 。

(10) 因为  $\sigma_{1\max} > \theta_s$ ，且  $c = \frac{K}{2}$ 。所以把聚合分裂成两个子集， $k = 0.5$ ，

则： $r_1 = (1,0)^T$ ，故新的聚合中心分别为：

$$Z_1^+ = (4.38, 2.75)^T, \quad Z_1^- = (2.38, 2.75)^T$$

为方便起见， $Z_1^+$  和  $Z_1^-$  改写为  $Z_1$  和  $Z_2$ ，令  $c = c + 1$ ， $J = J + 1 = 2$ ，返回(2)。

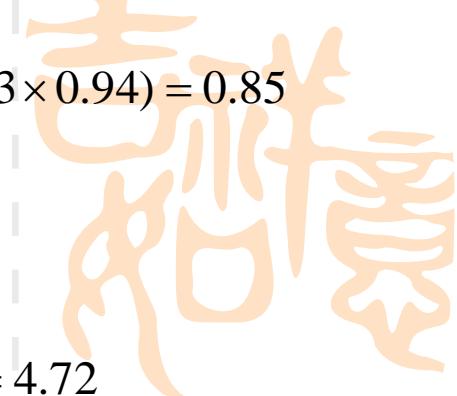
(2) 重新聚类： $w_1 : X_1 = \{x_4, x_5, x_6, x_7, x_8\}, n_1 = 5$        $w_2 : X_2 = \{x_1, x_2, x_3\}, n_2 = 3$

(3) 因为  $n_1 > \theta_n$ ， $n_2 > \theta_n$  没有子集抛弃。

(4) 重新计算聚合中心： $Z_1 = \frac{1}{n_1} \sum_{x \in X_1} x = (4.8, 3.8)^T$        $Z_2 = \frac{1}{n_2} \sum_{x \in X_2} x = (1.06, 1)^T$

(5) 计算类内平均距离：

$$\overline{D}_1 = \frac{1}{n_1} \sum_{x \in X_1} \|x - Z_1\| = 0.8 \quad \overline{D}_2 = \frac{1}{n_2} \sum_{x \in X_2} \|x - Z_2\| = 0.94$$



(6) 计算类内总平均距离:  $\bar{D} = \frac{1}{n} \sum_{j=1}^2 n_j \cdot \bar{D}_j = \frac{1}{8} (5 \times 0.8 + 3 \times 0.94) = 0.85$

(7) 因是偶次迭代, 所以转向(11)。

(11) 计算两个聚合中心之间的距离:  $D_{12} = \|Z_1 - Z_2\| = 4.72$

(12) 判断  $D_{12} > \theta_c$ 。

(13) 聚合中心不合并。

(14) 因为不是最后一次迭代, 令  $J = J + 1 = 3$ , 考虑是否修改参数。

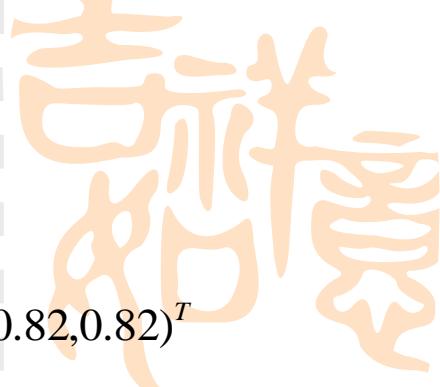
考虑:

(a) 已获得合理的聚合数目。

(b) 两聚合中心间距离大于类内总平均距离。

(c) 每个聚合内部有足够的比例的样本数目。

不必修改控制参数, 返回到(2)。



(2)步~(6)步与上次迭代相同。

(7) 所列情况均不满足，继续执行。

(8) 计算两个聚合的标准偏差。  $\sigma_1 = (0.75, 0.75)^T$ ,  $\sigma_2 = (0.82, 0.82)^T$

(9)  $\sigma_{1\max} = 0.75, \sigma_{2\max} = 0.82$

(10) 因为  $c = \frac{K}{2}$ ，且  $n_1$  和  $n_2$  均小于  $2(\theta_n + 1)$ ，分裂条件不满足。执行(11)。

(11)步~(13)步与前一次迭代结果相同。

(14) 因为  $J < I$ ，令  $J = J + 1 = 4$ ，无显著变化，返回(2)。



(2)步~(6)步与前一次迭代相同。

(7) 因为  $J = I$ ，是最后一次迭代，所以令  $\theta_c = 0$ ，转向(11)。



(11)步~(12)步与前一次迭代相同。

(13) 无合并发生。



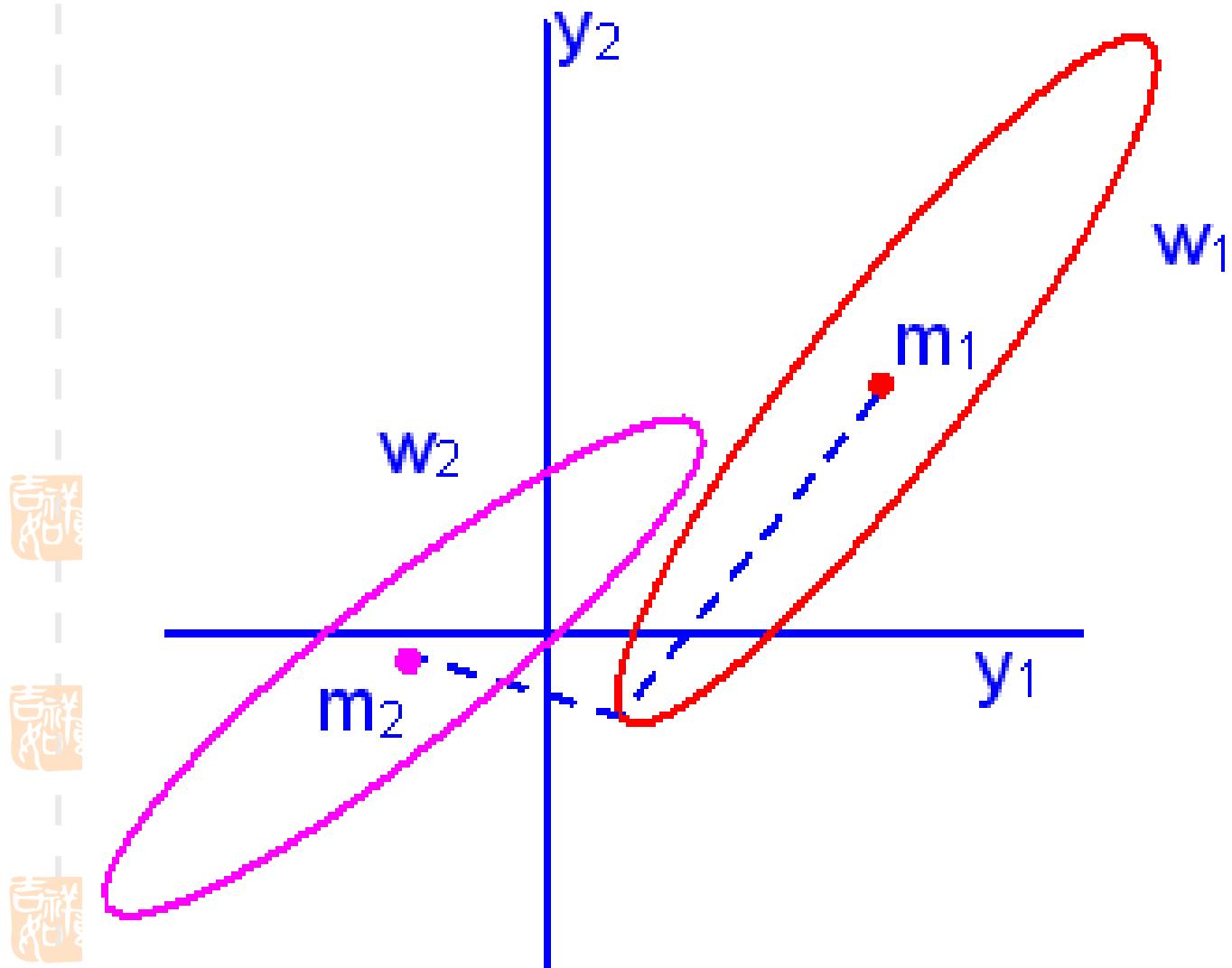
(14) 因  $J = I$ ，聚类过程结束。结果如上图。

吉  
祥  
慶

- ISODATA算法中，起始聚合中心的选取对聚类过程和结果都有较大影响，如果选择的好，则算法收敛快，聚类质量高。
- 注意：ISODATA与C—均值算法的异同点：
  - ① 都是动态聚类算法。
  - ② C—均值简单，ISODATA复杂。
  - ③ C—均值中，类型数目固定，ISODATA中，类型数目可变。



吉祥



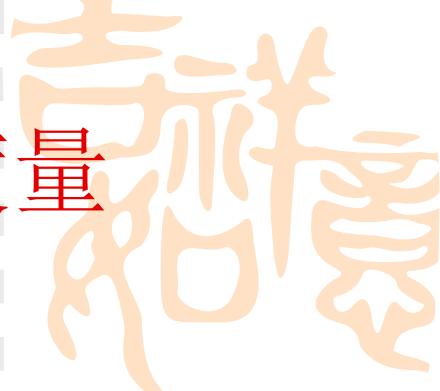
吉祥

吉祥

吉祥

吉祥

吉祥



## 五、基于样本和核相似性度量

- 基于样本和核相似性度量的聚类算法
  - 采用一个“核”来代表一个类
  - 核 $K_j$ 可以是一个函数，一个点集，等等
  - 样本和核之间相似性的度量  $\Delta(y, K_j)$
  - 准则函数，最小化的目标

$$J = \sum_{j=1}^c \sum_{y \in \Gamma_j} \Delta(y, K_j)$$





## ■ 样本和核相似性度量的聚类算法（续）

➤ 算法步骤类似于C-均值

1. 选择初始划分并计算初始核

2. 重新分配各个样本



$y \in \Gamma_j$  如果  $\Delta(y, K_j) = \min_k \Delta(y, K_k)$

3. 修正核，并重复2-3直至收敛



➤ C-均值算法=以类均值为核，欧氏距离作为样本和核之间的相似性度量





- 样本和核相似性度量的聚类算法（续）
  - 算法收敛的充分条件：准则函数  $J$  满足  
如果  $J(\Gamma, \tilde{K}) \leq J(\Gamma, K)$ , 那么  $J(\tilde{\Gamma}, \tilde{K}) \leq J(\Gamma, \tilde{K})$



- $\Gamma, K$  修正之前的分类和对应的核
  - $\tilde{\Gamma}, \tilde{K}$  修正之后的分类和对应的





- 样本和核相似性度量的聚类算法（续）
  - 正态核函数，适用于各类为正态分布

$$K_j(y, V_j) = \frac{1}{(2\pi)^{d/2} |\hat{\Sigma}_j|^{1/2}} \exp\left\{-\frac{1}{2}(y - m_i)^T \hat{\Sigma}_j^{-1} (y - m_i)\right\}$$

➤ 参数集  $V_j = (m_j, \hat{\Sigma}_j)$  从各类样本中估计

➤ 相似性度量

$$\Delta(y, K_j) = \frac{1}{2}(y - m_i)^T \hat{\Sigma}_j^{-1} (y - m_i) + \frac{1}{2} \log |\hat{\Sigma}_j|$$





- 样本和核相似性度量的聚类算法（续）
  - 主轴核函数，适用于各类样本集中在各自的主轴方向上的子空间里的情况

$$K(y, V_j) = U_j^T y$$

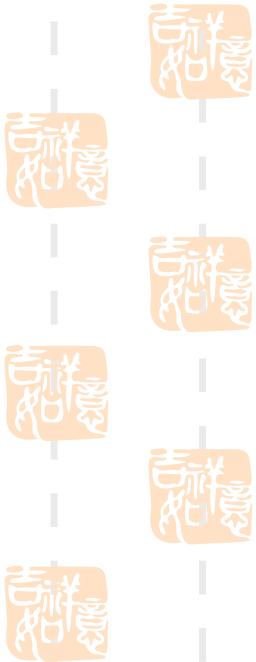
$U_j = (u_1, u_2, \dots, u_{d_j})$  是第  $j$  类样本协方差阵的前  $d_j$  个最大特征值对应的特征向量系统

$$\Delta(y, K_j) = [(y - m_j) - U_j U_j^T (y - m_j)]^T \square$$
$$[(y - m_j) - U_j U_j^T (y - m_j)]$$

是样本到主轴子空间的距离

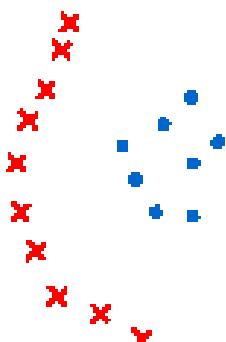
吉祥如意

- 样本和核相似性度量的聚类算法（续）
  - 可以拟合各种形状的分布
  - 需要预先知道数据的分布形状

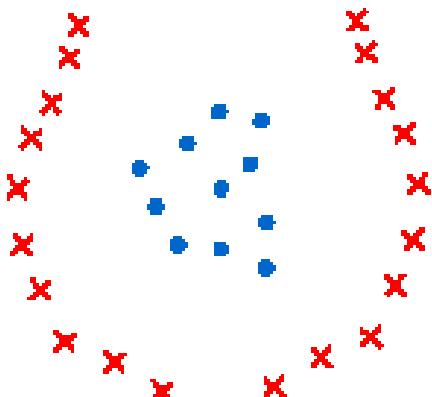


吉祥如意

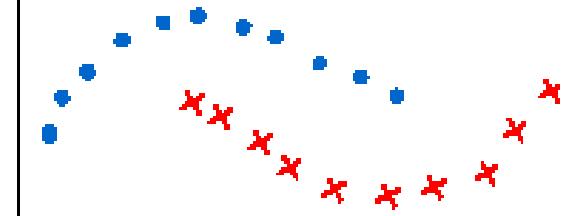
(a)



(b)



(c)





## 六、近邻函数准则

- 近邻函数准则算法

  - 近邻函数：样本间相似性的度量

  - 如果 $y_i$ 是 $y_j$ 的第 $I$ 个近邻， $y_j$ 是 $y_i$ 的第 $K$ 个近邻



$$a_{ij} = I + K - 2, i \neq j$$

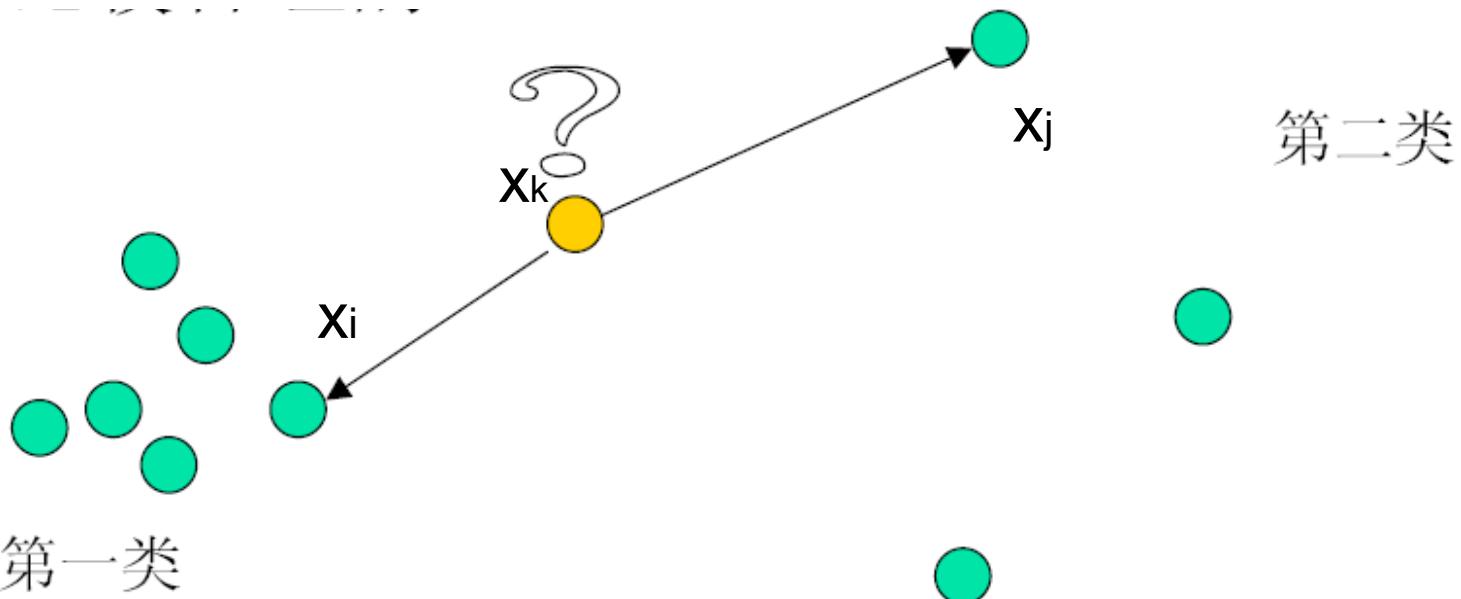
  - 近邻函数使得密度相近的点容易聚成一类





## ■ 近邻函数准则算法（续）

- 纯粹按照欧氏距离，则黄色点归为第一类
- 按照近邻函数值，黄色点归为第二类。这是比较合理的





## ■ 近邻函数准则算法（续）

- 同一类中的点之间存在“连接”。连接损失就定义为两点之间的近邻函数 $\alpha_{ij}$ 。
- 一个点和其自身的连接损失定义为 $\alpha_{ii}=2N$ ，以惩罚只有一个点的聚类
- 不同类的点不存在连接，连接损失为 $\alpha_{ij}=0$
- 总类内损失

$$L_{within} = \sum_{i=1}^N \sum_{j=1}^N \alpha_{ij}$$





- 近邻函数准则算法（续）

- 第*i*类和第*j*类之间的最小近邻函数值定义为

$$\gamma_{ij} = \min_{y_k \in \Gamma_i, y_l \in \Gamma_j} (\alpha_{kl})$$



- 第*i*类内最大连接损失记为  $\alpha_{i\max}$



- 第*i*类与第*j*类之间的连接损失定义为  $\beta_{ij}$ ,





## ■ 近邻函数准则算法（续）

$$\beta_{ij} = \begin{cases} -[(\gamma_{ij} - \alpha_{i\max}) + (\gamma_{ij} - \alpha_{j\max})], & \text{if } \gamma_{ij} > \alpha_{i\max}, \gamma_{ij} > \alpha_{j\max} \\ \gamma_{ij} + \alpha_{i\max}, & \text{if } \gamma_{ij} \leq \alpha_{i\max}, \gamma_{ij} > \alpha_{j\max} \\ \gamma_{ij} + \alpha_{j\max}, & \text{if } \gamma_{ij} > \alpha_{i\max}, \gamma_{ij} \leq \alpha_{j\max} \\ \gamma_{ij} + \alpha_{i\max} + \alpha_{j\max}, & \text{if } \gamma_{ij} \leq \alpha_{i\max}, \gamma_{ij} \leq \alpha_{j\max} \end{cases}$$

➤  $\beta_{ij}$  的设计目标是：如果两类间的最小近邻值大于任何一方的类内的最大连接损失时，损失代价就是正的，从而应该考虑把这两类合并





- 近邻函数准则算法（续）
  - 总类间损失


$$L_{between} = \sum_{i \neq j} \beta_{ij}$$

- 准则函数


$$J = L_{within} + L_{between}$$





## ■ 近邻函数准则算法（续）

### ➤ 算法步骤

- 1.计算距离矩阵  $\Delta_{ij} = \Delta(y_i, y_j)$
- 2.用距离矩阵计算近邻矩阵  $M_{ij}$ ,  $M_{ij}$ 表示  $y_j$  是  $y_i$  的第几个近邻
- 3.计算近邻函数矩阵  $L_{ij} = M_{ij} + M_{ji} - 2I$ ,  $L_{ii} = 2N$
- 4.在  $L$  中, 每个点与其最近邻连接, 形成初始的划分
- 5.对每两个类计算  $\gamma_{ij}$  和  $\alpha_{imax}$ ,  $\alpha_{jmax}$ , 只要  $\gamma_{ij}$  小于  $\alpha_{imax}$ 、 $\alpha_{jmax}$  中的任何一个, 就合并两类(建立连接)。重复至没有新的连接发生为止



吉  
祥  
慶

■ 例：如图所示样本，使用近邻函数准则聚类。

解：

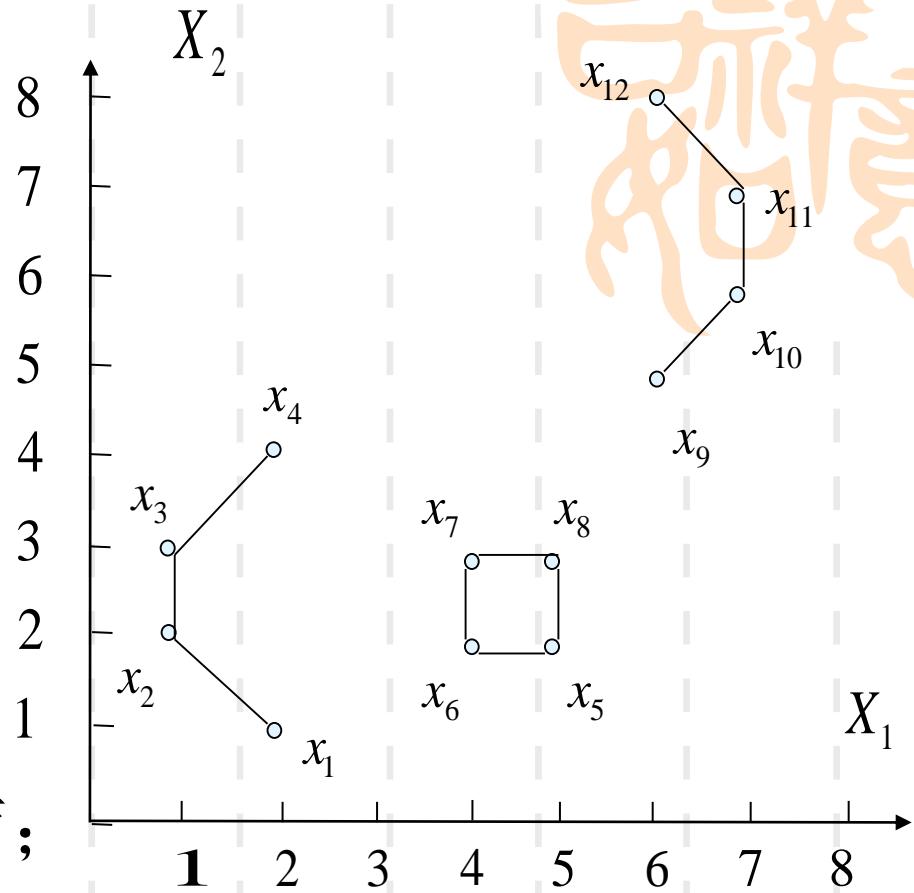


吉  
祥  
慶

- ① 计算D矩阵，结果从简；
- ② 计算M矩阵，结果从简；
- ③ 计算L矩阵，结果见表



吉  
祥  
慶



L	1	2	3	4	5	6	7	8	9	10	11	12
1	24	1	3	8	4	8	6	9	16	17	17	19
2	1	24	0	3	8	11	11	14	17	17	18	18
3	3	0	24	1	10	13	9	12	15	16	18	17
4	8	3	1	24	7	10	6	10	14	16	15	11
5	4	8	10	7	24	0	0	4	13	14	16	17
6	8	11	13	10	0	24	4	0	8	18	14	16
7	6	11	9	6	0	4	24	0	6	12	13	15
8	9	14	12	10	4	0	0	24	4	8	12	14
9	16	17	15	14	13	8	6	4	24	1	3	6
10	17	17	16	16	14	18	12	8	1	24	0	3
11	17	18	18	15	16	14	13	12	3	0	24	1
12	19	18	17	11	17	16	15	15	6	3	1	24

吉祥如意

初始聚合:

$$\omega_1 = X_1 = \{x_1, x_2, x_3, x_4\}, \quad n_1 = 4$$

$$\omega_2 = X_2 = \{x_5, x_6, x_7, x_8\}, \quad n_2 = 4$$

$$\omega_3 = X_3 = \{x_9, x_{10}, x_{11}, x_{12}\}, \quad n_3 = 4$$

最近邻函数值:

$$\omega_1 \text{ 和 } \omega_2: \gamma_{12} = 4$$

$$\omega_1 \text{ 和 } \omega_3: \gamma_{13} = 11$$

$$\omega_2 \text{ 和 } \omega_3: \gamma_{23} = 4$$

类内最大连接损失:

$$\alpha_{1\max} = 1$$

$$\alpha_{2\max} = 0$$

$$\alpha_{3\max} = 1$$



## 4.7 最小张树聚类

术语：

- ①线段：两个模式样本点的连线
- ②路径：连接两点的线段序列
- ③回路：闭合路径
- ④连通图形：任两点之间有一条或一条以上的路径者。（即各点之间是连结起来的，但不一定直接相连。）
- ⑤树图：没有回路的连通图形（单线顺序相连，不闭合，不返回）

- ⑥张树图：包含模式样本集合中每一点的树图（即连结每一个模式样本点且没有重复的连通图）
- ⑦线段权重（线的重要性），（可取点间距离）整个树图的权重为树图中各线段权重之和。
- ⑧最小张树图：权重最小的张树图（若以距离作权重，则各模式样本点以最小距离连结每一样本点，且无重复），沿着该路线，连结相邻每点间的距离总和  
 $\min$
- 一条路，经过最多点，代价最小。
- ⑨主直径：在最小张树图中走过最多模式样本点的那条路径

吉  
祥  
慶

- 最小张树图构成方法：

- ①计算距离表
- ②所有距离按从小到大顺序排列
- ③按距离从小到大顺序连结点对。

规则：  $\begin{cases} A & \text{先连距离小的点对，再连距离稍大的距离} \\ B & \text{所有点不能构成回路（构成时不连）} \end{cases}$

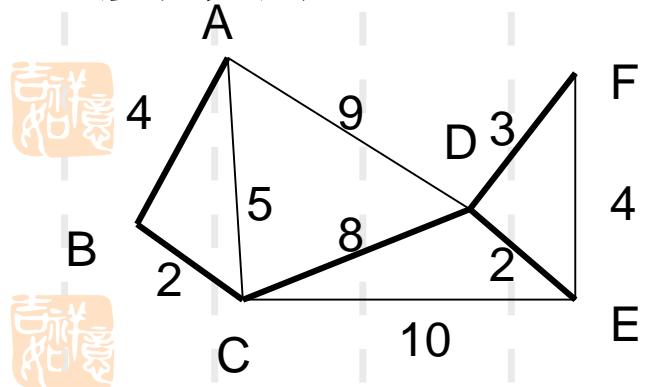


- 最小生成树算法，找到图的最小生成树，然后把树上最长的边去掉形成两个类。在两个子图中再去掉最长边，...，依次进行下去



吉  
祥  
圖

- 例：递增排序： BC、DE、DF、EF、AB、AC、CD、AD、CF
- 连接顺序： BC AB CD DE DF





- 缺点：对密集点集之间干扰敏感
- 改进：引入树的直径和点深度
  - 点的深度：与该点连接的最长分支的长度
- 算法步骤如下：
  - ①给定混合样本集，生成最小张树(可按距离给权值)；
  - ②确定最小张树上的直径和计算直径上各点深度；
  - ③绘制直径上各点的深度图，找出局部最小值；
  - ④去掉局部最小值的点，获得分离的类型聚合。

吉祥如意



(a) 混合样本集



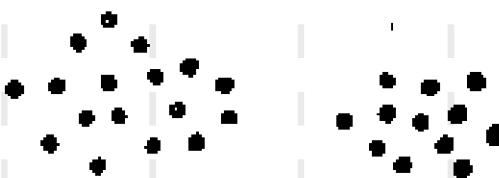
(b) 最小张树



(c) 直径上各点深度



(d) 深度图



(e) 分离后的聚类



吉祥如意

# 小结

- 聚类是一个面向具体数据，具体应用的任务
- 针对不同的数据，不同的目标选择不同的聚类算法
- 算法对数据尺度(**scale**)敏感的问题
- 采取什么样的相似性度量？数据的各个特征是可比较的吗？
- 除了预先指定，怎么确定分类的数目？