第七章 聚类分析
Cluster Analysis

- §7.1 引言
- §7.2 距离和相似系数
- §7.3 系统聚类法
- §7.4 快速聚类法

§7.1 引言

- 聚类分析：将分类对象分成若干类，相似的归为同一类，不相似的归为不同的类。
- 聚类分析和判别分析有着不同的分类目的，彼此之间既有区别又有联系。
- 聚类分析分为Q型（分类对象为样品）和R型（分类对象为变量）两种。

按观测对象的“**相似**”程度分类

对样品分类

- 欧氏距离
- 马氏距离
- 兰氏距离

对变量分类

- 夹角余弦
- 相关系数

§7.2 距离和相似系数

一、距离

- 设$\boldsymbol{x}=(x_1,x_2,\cdots,x_p)'$ 和$\boldsymbol{y}=(y_1,y_2,\cdots,y_p)'$为两个样品，则所定义的距离一般应满足如下三个条件：
- (i)非负性：$d(\boldsymbol{x},\boldsymbol{y})\geq 0$，$d(\boldsymbol{x},\boldsymbol{y})=0$当且仅当$\boldsymbol{x}=\boldsymbol{y}$；
- (ii)对称性：$d(\boldsymbol{x},\boldsymbol{y})=d(\boldsymbol{y},\boldsymbol{x})$；
- (iii)三角不等式：$d(\boldsymbol{x},\boldsymbol{y})\leq d(\boldsymbol{x},\boldsymbol{z})+d(\boldsymbol{z},\boldsymbol{y})$。

常用的距离

- 1.明考夫斯基（Minkowski）距离
- 2.兰氏（Lance和Williams）距离
- 3.马氏距离

1.明考夫斯基距离

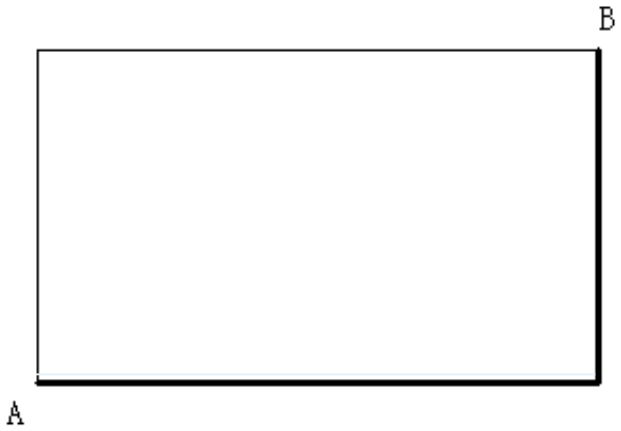
- 明考夫斯基距离（简称明氏距离）：

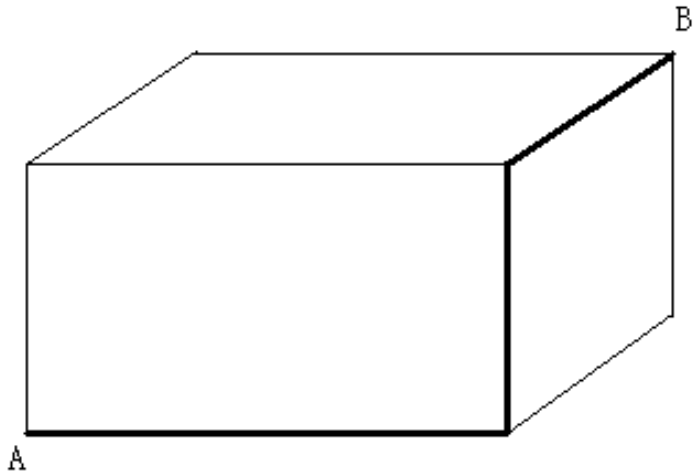
$$d(\boldsymbol{x},\boldsymbol{y})=\left[\sum_{i=1}^{p}|x_i-y_i|^q\right]^{1/q}$$ 这里$q\geq 1$。

- 明氏距离的三种特殊形式：

- (i)当$q=1$时，$d(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{p}|x_i-y_i|$，称为绝对值距离，常被形象地称作“城市街区”距离；
- (ii)当$q=2$时，$d(\boldsymbol{x},\boldsymbol{y})=\left[\sum_{i=1}^{p}|x_i-y_i|^2\right]^{1/2}=\sqrt{(\boldsymbol{x}-\boldsymbol{y})'(\boldsymbol{x}-\boldsymbol{y})}$，这是**欧氏距离，它是聚类分析中最常用的一个距离；**
- (iii)当$q=\infty$时，$d(\boldsymbol{x},\boldsymbol{y})=\max_{1\leq i\leq p}|x_i-y_i|$，称为切比雪夫距离。

绝对值距离图示





对各变量的数据作标准化处理

- 当各变量的单位不同或测量值范围相差很大时，应先对各变量的数据作标准化处理。最常用的标准化处理是，令

$$x_i^* = \frac{x_i - \bar{x}_i}{\sqrt{s_{ii}}}, \quad i = 1, 2, \cdots, p$$

其中$\bar{x}_i$ 和s_{ii}分别为x_i的样本均值和样本方差。

2.兰氏距离

- 当所有的数据皆为正时，可以定义$\boldsymbol{x}$与$\boldsymbol{y}$之间的兰氏距离为

$$d(\boldsymbol{x},\boldsymbol{y})=\sum_{i=1}^{p}\frac{|x_i-y_i|}{x_i+y_i}$$

- 该距离与各变量的单位无关，且适用于高度偏斜或含异常值的数据。

3.马氏距离

- $\boldsymbol{x}$和$\boldsymbol{y}$之间的马氏距离为

$$d(\boldsymbol{x},\boldsymbol{y})=\sqrt{(\boldsymbol{x}-\boldsymbol{y})'\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{y})}$$

  其中$\boldsymbol{S}$为样本协差阵。

- 聚类过程中的类一直变化着，$\boldsymbol{S}$一般难以确定，除非有关于不同类的先验知识。因此，在实际聚类分析中，马氏距离一般不是理想的距离。

定性变量（属性变量）的一种距离定义

- 例1　某高校举办一个培训班，从学员的资料中得到这样六个变量：

x_1：性别（男，女）

x_2：外语语种（英语，非英语）

x_3：专业（统计，非统计）

x_4：职业（教师，非教师）

x_5：居住处（校内，校外）

x_6：学位（硕士，学士）

- 现有两名学员：

$\boldsymbol{x}$=(男，英语，统计，非教师，校外，学士)′

$\boldsymbol{y}$=(女，英语，非统计，教师，校外，硕士)′

- 一般地，若记

m_1：配合的变量数

m_2：不配合的变量数

则它们之间的距离可定义为

$$d(\boldsymbol{x},\boldsymbol{y})=\frac{m_2}{m_1+m_2}$$

- 故按此定义，本例中$\boldsymbol{x}$ 与$\boldsymbol{y}$ 之间的距离为2/3

二、相似系数

- 变量之间的相似系数（或其绝对值）越大，认为变量之间的相似性程度就越高；反之，则越低。
- 聚类时，比较相似的变量倾向于归为一类，不太相似的变量归属不同的类。

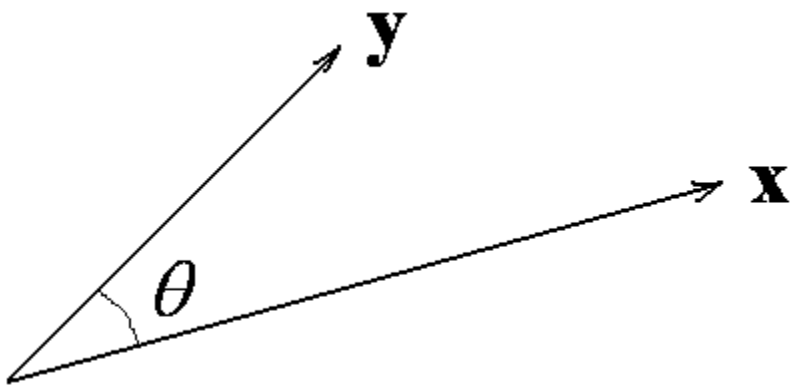
变量间相似系数一般应满足的条件

(1)$c_{ij}=\pm 1$，当且仅当$x_i=ax_j+b$，$a(\neq 0)$ 和b是常数；

(2)$|c_{ij}|\leq 1$，对一切i,j；

(3)$c_{ij}=c_{ji}$，对一切i,j。

两个向量的夹角余弦



$$\cos(\theta) = \frac{\boldsymbol{x}'\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|}$$

1.夹角余弦

- 变量x_i与x_j的夹角余弦定义为

$$c_{ij}(1)=\frac{\sum_{k=1}^{n}x_{ki}x_{kj}}{\left[\left(\sum_{k=1}^{n}x_{ki}^2\right)\left(\sum_{k=1}^{n}x_{kj}^2\right)\right]^{1/2}}$$

它是R^n中变量x_i的观测向量$(x_{1i},x_{2i},\cdots,x_{ni})'$与变量$x_j$的观测向量$(x_{1j},x_{2j},\cdots,x_{nj})'$之间夹角$\theta_{ij}$的余弦函数，即$c_{ij}(1)=\cos\theta_{ij}$。

2.相关系数

- 变量x_i与x_j的相关系数为

$$c_{ij}(2) = r_{ij} = \frac{\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_i\right)\left(x_{kj}-\bar{x}_j\right)}{\left\{\left[\sum_{k=1}^{n}\left(x_{ki}-\bar{x}_i\right)^2\right]\left[\sum_{k=1}^{n}\left(x_{kj}-\bar{x}_j\right)^2\right]\right\}^{1/2}}$$

- 如果变量x_i与x_j是已标准化了的，则它们间的夹角余弦就是相关系数。

3. 变量间的距离

(1) 利用相似系数来定义变量间的距离

令 $d_{ij}=1-|C_{ij}|$ 或 $d^2_{ij}=1-C^2_{ij}$ $(i,j=1,2,\ldots,m)$.

(2) 利用样本协差阵来定义距离

设样本协差阵$S=(s_{ij})>0$,变量X_i和X_j 间的距离可定义为

$$d_{ij}= s_{ii} + s_{jj} - 2 s_{ij}$$

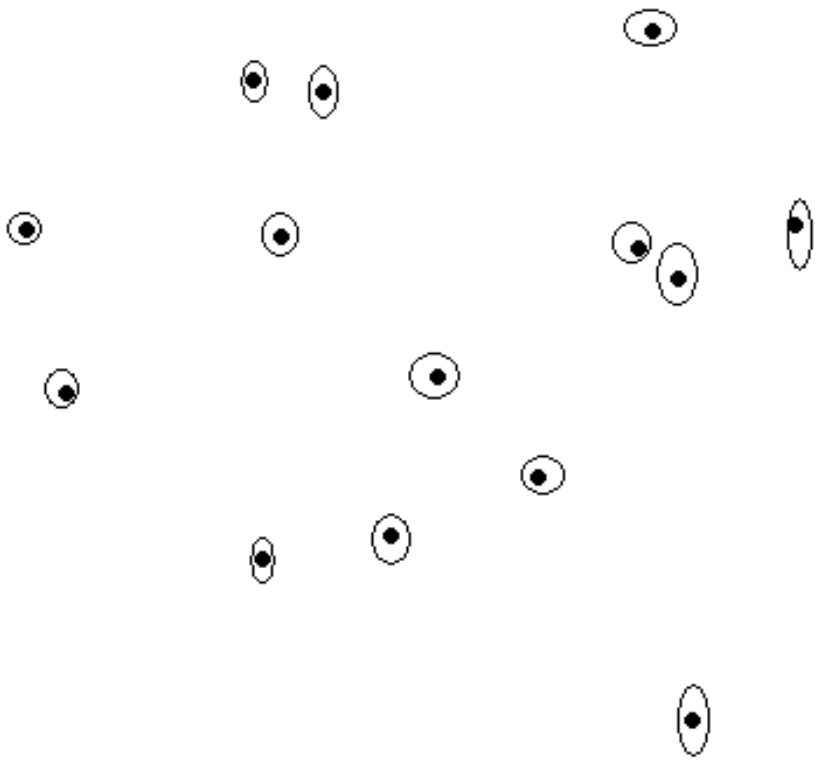
(3) 把变量X_i的n次观测值看成n维空间的点.在n维空间中类似可定义m个变量间的种种距离.

§7.3 系统聚类法（hierarchical clustering method）

系统聚类法（或层次聚类法）的基本思想：

开始时将每个样品各自作为一类，并规定样品之间的距离和类与类之间的距离，然后将距离最近的两类合并成一个新类，计算新类与其他类的距离；重复进行两个最近类的合并，每次减少一类，直至所有的样品合并为一类。

一开始每个样品各自作为一类



§7.3 系统聚类法

类间距离计算方法

（1）最短距离法（single）

（2）最长距离法（complete）

（3）中间距离法（median）

（4）类平均法（average）

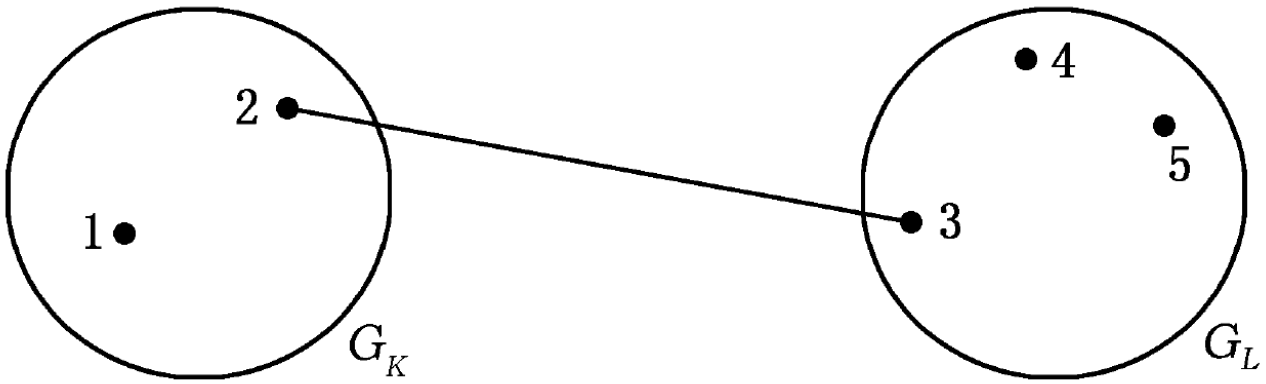
（5）重心法（centroid）

（6）离差平方和法（Ward）

一、最短距离法

- 定义类与类之间的距离为两类最近样品间的距离，即

$$D_{KL} = \min_{i \in G_K, j \in G_L} d_{ij}$$



最短距离法：$D_{KL}=d_{23}$

最短距离法的聚类步骤

(1) 规定样品之间的距离，计算n个样品的距离矩阵$\boldsymbol{D}_{(0)}$，它是一个对称矩阵。

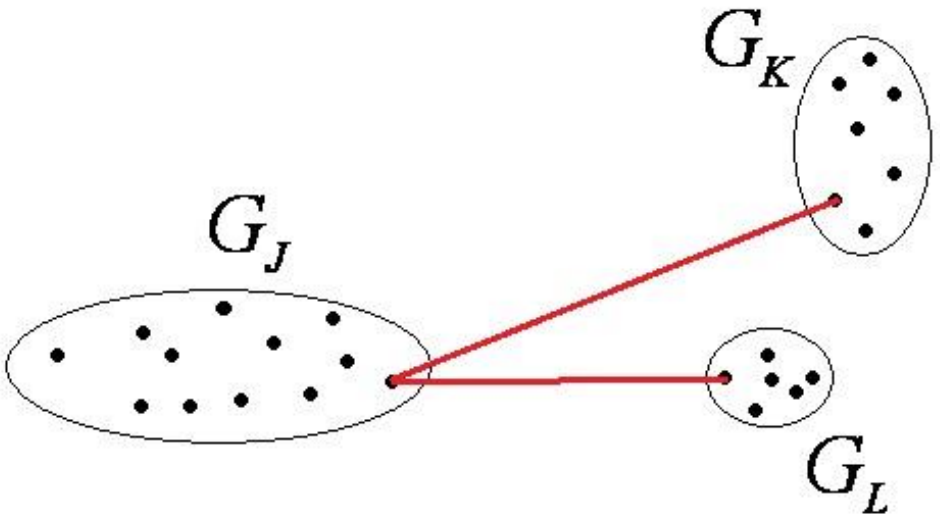
(2)选择$\boldsymbol{D}_{(0)}$中的最小元素，设为D_{KL}，则将G_K和G_L合并成一个新类，记为G_M，即$G_M = G_K \cup G_L$。

(3)计算新类G_M与任一类G_J之间距离的递推公式为

$$D_{MJ} = \min_{i\in G_M,\, j\in G_J} d_{ij} = \min\left\{ \min_{i\in G_K,\, j\in G_J} d_{ij},\ \min_{i\in G_L,\, j\in G_J} d_{ij} \right\}$$

$$= \min\left\{ D_{KJ}, D_{LJ} \right\}$$

递推公式的图示理解



最短距离法的聚类步骤（续）

在$\boldsymbol{D}_{(0)}$中，G_K和G_L所在的行和列合并成一个新行新列，对应G_M，该行列上的新距离值由上述递推公式求得，其余行列上的距离值不变，这样就得到新的距离矩阵，记作$\boldsymbol{D}_{(1)}$。

- (4)对$\boldsymbol{D}_{(1)}$重复上述对$\boldsymbol{D}_{(0)}$的两步得$\boldsymbol{D}_{(2)}$，如此下去直至所有元素合并成一类为止。

例2 设有五个样品，每个只测量了一个指标，分别是1，2，6，8，11，试用最短距离法将它们分类。

记$G_1=\{1\}$，$G_2=\{2\}$，$G_3=\{6\}$，$G_4=\{8\}$，$G_5=\{11\}$，样品间采用绝对值距离。

$\boldsymbol{D_{(0)}}$

	G_1	G_2	G_3	G_4	G_5
G_1	0				
G_2	1	0			
G_3	5	4	0		
G_4	7	6	2	0	
G_5	10	9	5	3	0

$D_{(1)}$

	G_6	G_3	G_4	G_5
G_6	0			
G_3	4	0		
G_4	6	2	0	
G_5	9	5	3	0

其中$G_6 = G_1 \cup G_2$

$D_{(2)}$

	G_6	G_7	G_5
G_6	0		
G_7	4	0	
G_5	9	3	0

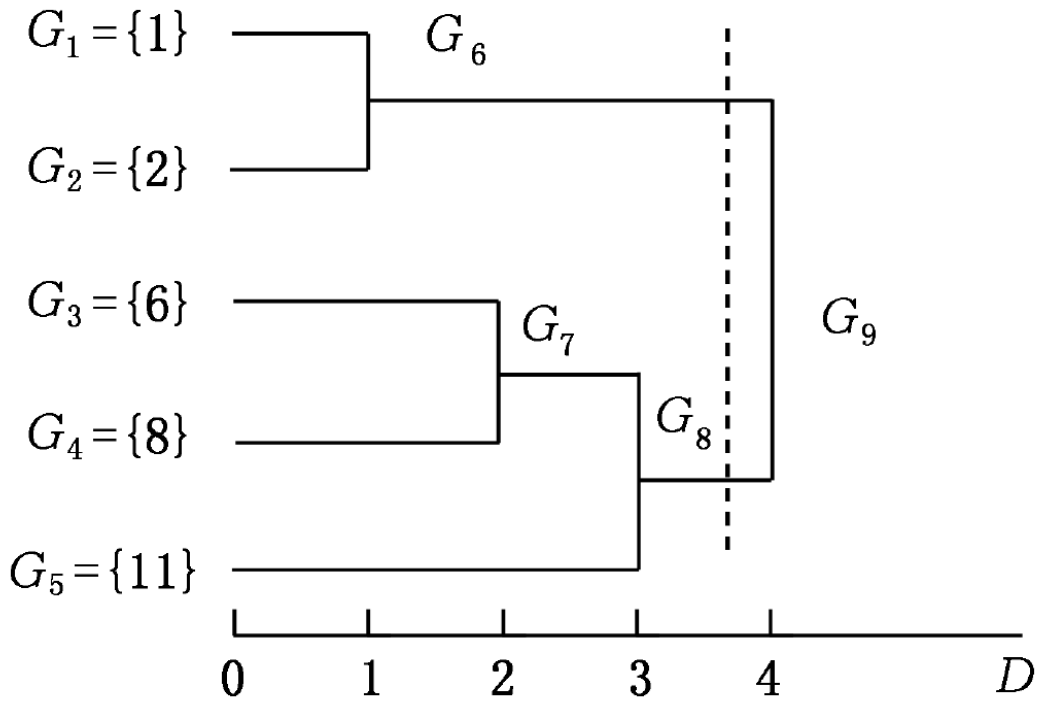
其中$G_7 = G_3 \cup G_4$

$D_{(3)}$

	G_6	G_8
G_6	0	
G_8	4	0

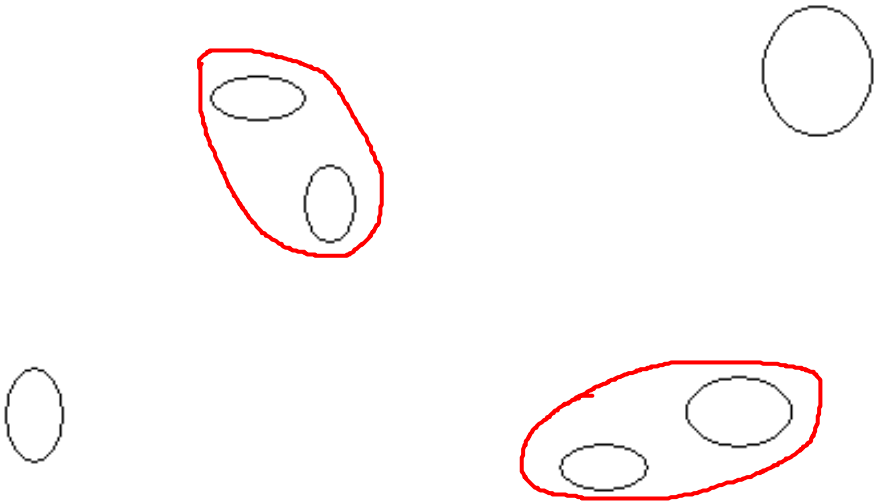
其中$G_6 = G_1 \cup G_2$

最短距离法树形图



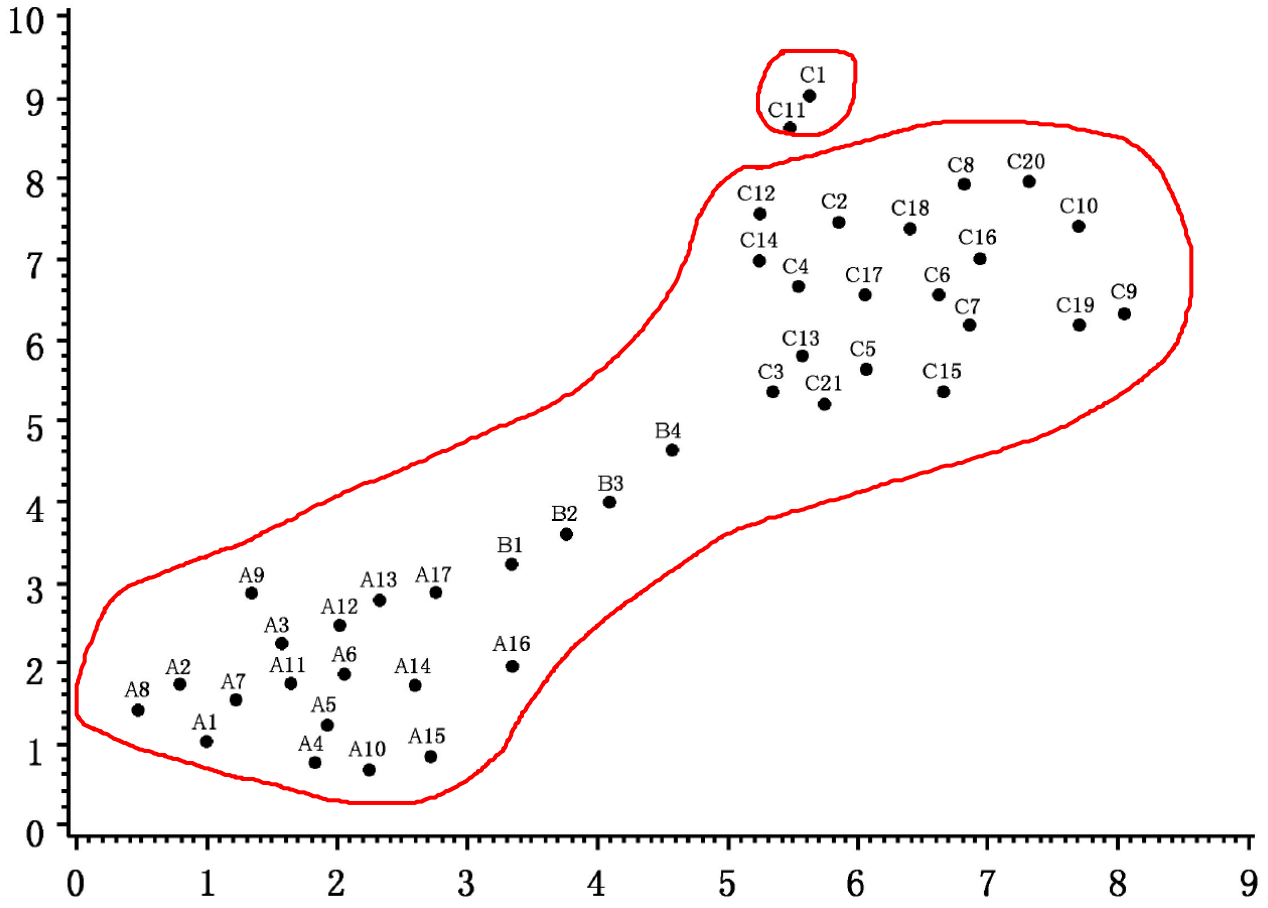
- 如果某一步$\boldsymbol{D}_{(m)}$中最小的元素不止一个，则称此现象为结，对应这些最小元素的类可以任选一对合并或同时合并。最短距离法最容易产生结，且有一种挑选长链状聚类的倾向，称为**链接**倾向。

结的图示：



- 最短距离法不适合对分离得很差的群体进行聚类。

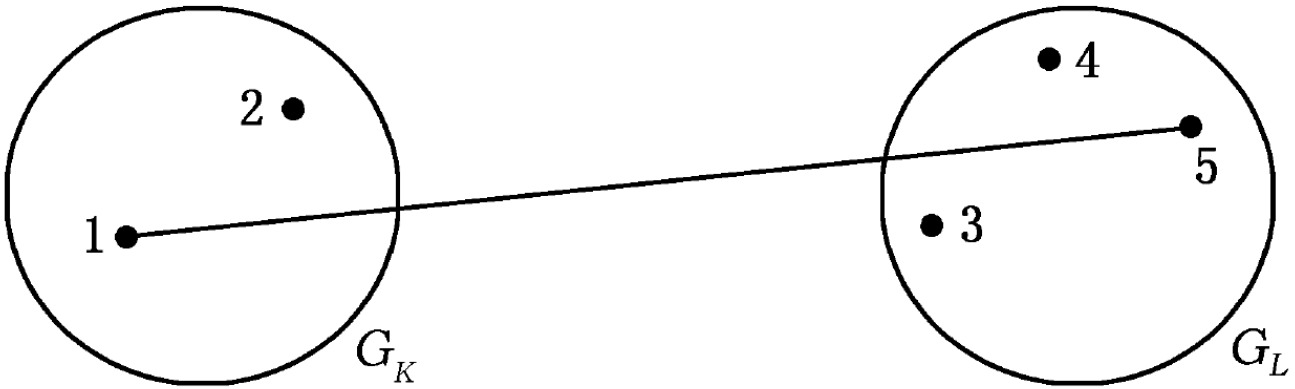
一个最短距离法产生链接的例子



二、最长距离法

- 类与类之间的距离定义为两类最远样品间的距离，即

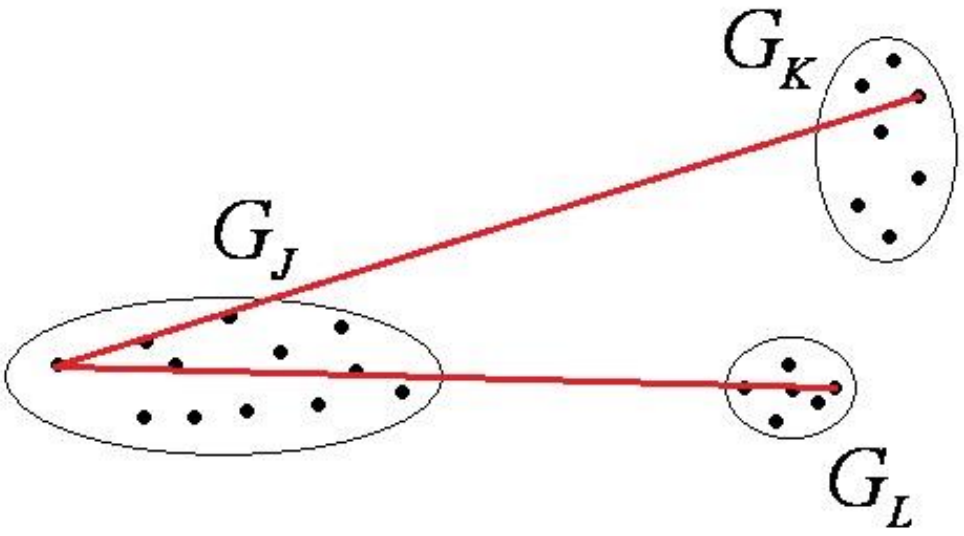
$$D_{KL} = \max_{i \in G_K, j \in G_L} d_{ij}$$



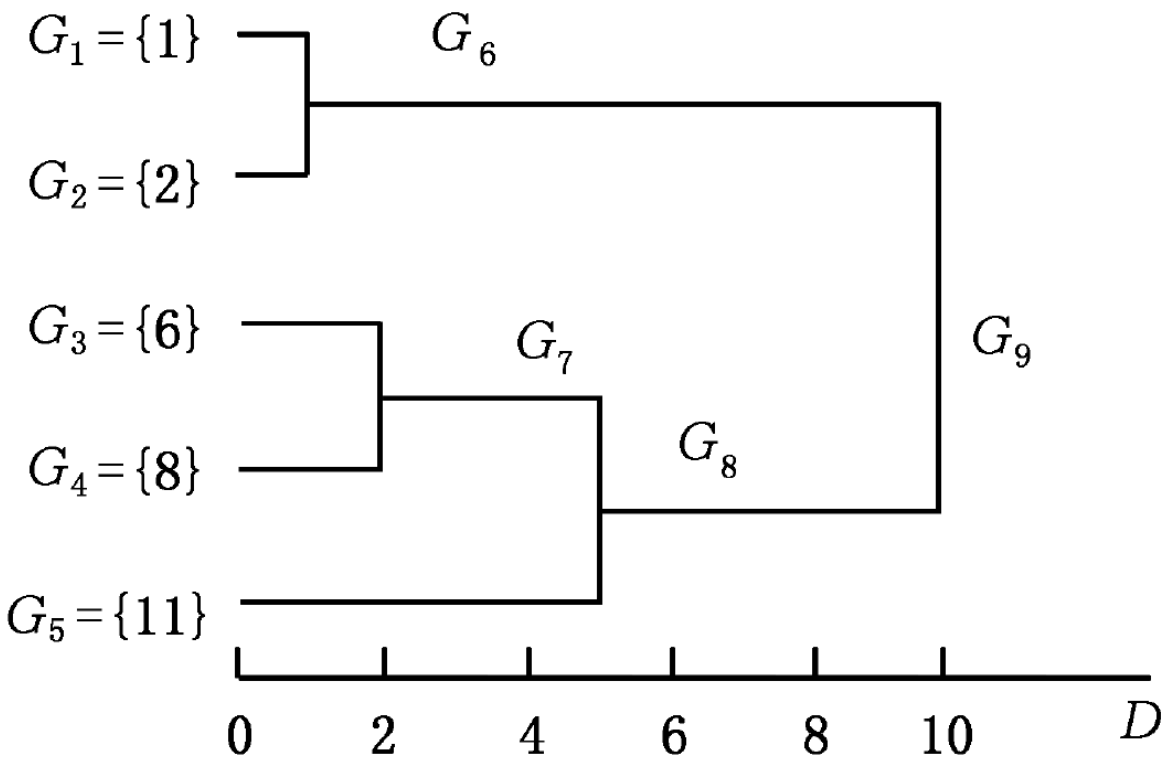
最长距离法：$D_{KL}=d_{15}$

- 最长距离法与最短距离法的并类步骤完全相同，只是类间距离的递推公式有所不同。
- 递推公式：

$$D_{MJ} = \max\{D_{KJ}, D_{LJ}\}$$



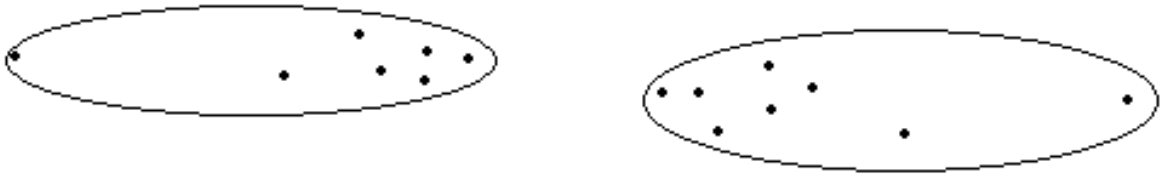
对例2采用最长距离法。



最长距离法树形图

异常值的影响

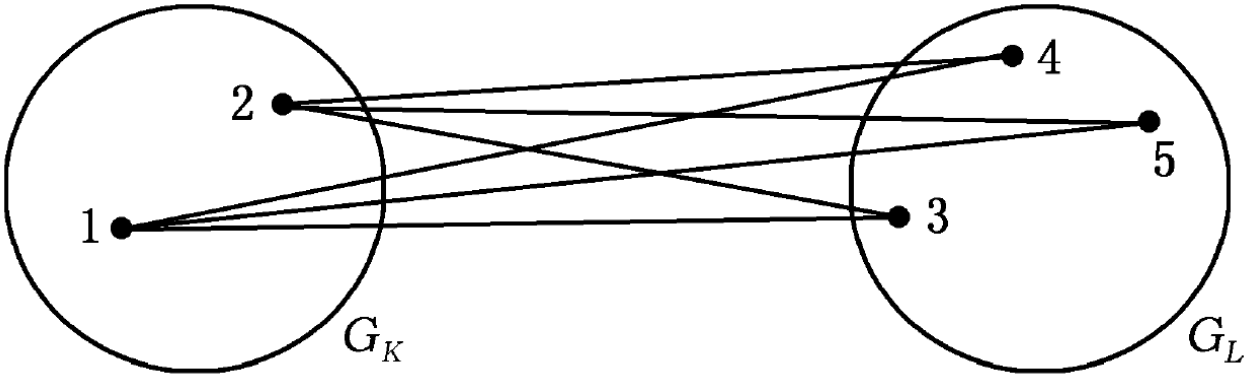
- 最长距离法容易被异常值严重地扭曲。



三、类平均法

- 类平均法较好地利用了所有样品之间的信息，在很多情况下它被认为是一种比较好的系统聚类法。
- 类G_K和G_L之间的距离定义为

$$D_{KL}=\frac{1}{n_K n_L}\sum_{i\in G_K,\,j\in G_L} d_{ij}$$

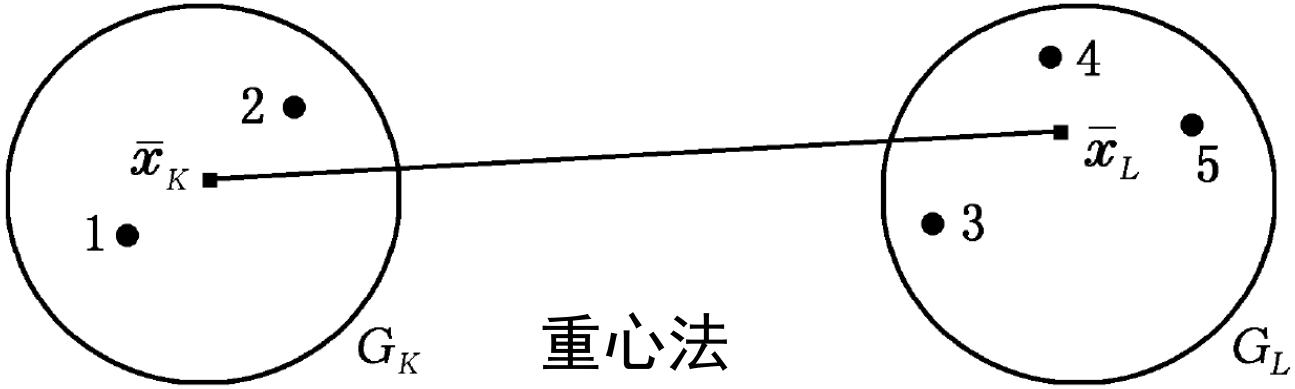


类平均法

四、重心法

- 设类G_K和G_L的重心（均值）分别为$\bar{\boldsymbol{x}}_K$和$\bar{\boldsymbol{x}}_L$，则G_K与G_L之间的平方距离定义为

$$D_{KL}^2 = d_{\bar{x}_K \bar{x}_L}^2 = \left(\bar{\boldsymbol{x}}_K - \bar{\boldsymbol{x}}_L\right)' \left(\bar{\boldsymbol{x}}_K - \bar{\boldsymbol{x}}_L\right)$$



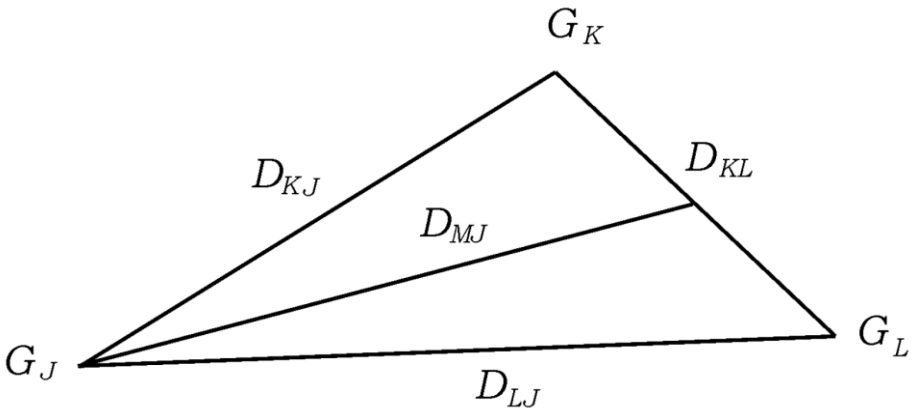
重心法

与其他系统聚类法相比，重心法在处理异常值方面更稳健，但是在别的方面一般不如类平均法或离差平方和法的效果好。

五、中间距离法

- 设$G_M = G_K \cup G_L$，对于任一类G_J，考虑由D_{KJ}，D_{LJ}和D_{KL}为边长组成的三角形，取D_{KL}边的中线作为D_{MJ}。D_{MJ}的计算公式为

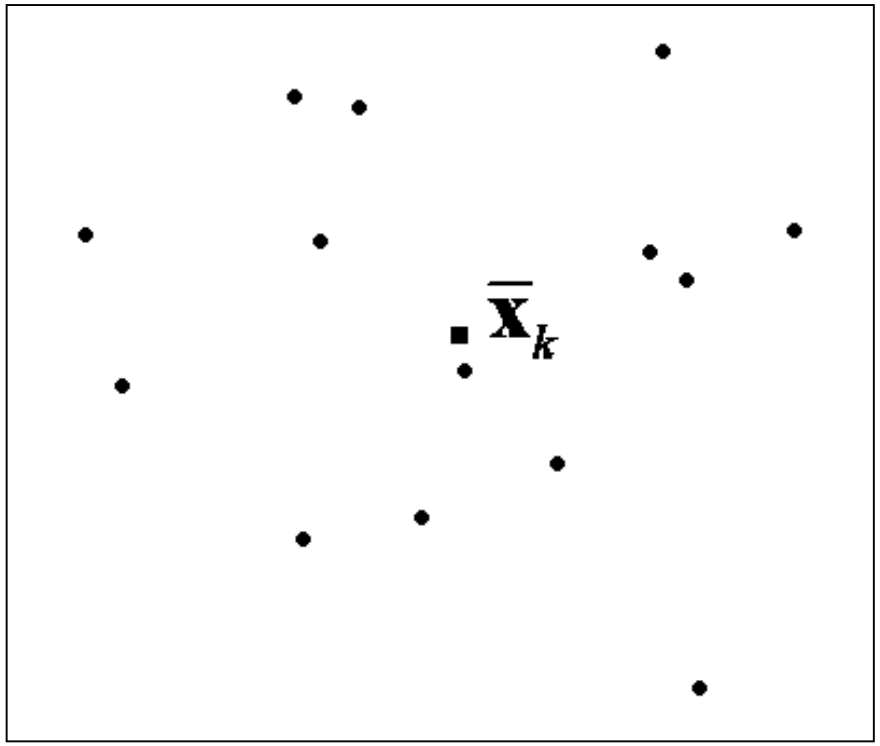
$$D_{MJ}^2 = \frac{1}{2}D_{KJ}^2 + \frac{1}{2}D_{LJ}^2 - \frac{1}{4}D_{KL}^2$$

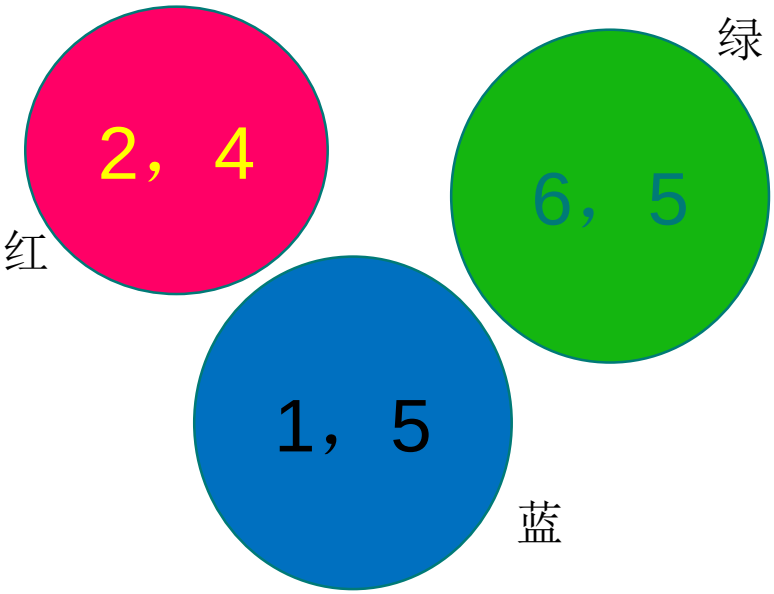


中间距离法的几何表示

六、离差平方和法(Ward方法)

- （类内）离差平方和：类中各样品到类重心（均值）的平方欧氏距离之和。
- 此方法并类时总是使得并类导致的类内离差平方和增量最小。首先,每个样品各成一类，每次类合并选择使离差平方和增加最小的两类进行合并。





$(2-3)^2+(4-3)^2=2$

$(6-5.5)^2+(5-5.5)^2=0.5$

$(1-3)^2+(5-3)^2=8$

- 红绿（2，4，6，5） 8.75
- 离差平方和增加8.75－2.5＝6.25
- 蓝绿（6，5，1，5）14.75
- 离差平方和增加14.75－8.5＝6.25
- 蓝红（2，4，1，5）10
- 离差平方和增加10－10＝0
- 故按该方法的连接应是蓝红首先连接。

- 类内离差平方和W_K是类G_K内各点到类重心点$\bar{\boldsymbol{x}}_k$的直线距离之平方和。设类G_K和G_L合并成新类G_M，则G_K, G_L和G_M的离差平方和分别是

$$W_K = \sum_{i \in G_K} \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_K\right)' \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_K\right)$$

$$W_L = \sum_{i \in G_L} \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_L\right)' \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_L\right)$$

$$W_M = \sum_{i \in G_M} \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_M\right)' \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_M\right)$$

对固定的类内样品数，它们反映了各自类内样品的分散程度。

- 定义G_K和G_L之间的平方距离为

$$D_{KL}^2 = W_M - W_K - W_L$$

- D_{KL}^2也可表达为

$$D_{KL}^2 = \frac{n_K n_L}{n_M}(\bar{x}_K - \bar{x}_L)'(\bar{x}_K - \bar{x}_L)$$

  - $\frac{n_K n_L}{n_M} = \frac{n_K n_L}{n_K + n_L} = \frac{1}{1/n_L + 1/n_K}$，当$n_K = n_L$时，$\frac{n_K n_L}{n_M} = \frac{n_K}{2}$

- 离差平方和法使得两个大的类倾向于有较大的距离，因而不易合并；相反，两个小的类却因倾向于有较小的距离而易于合并。这往往符合我们对聚类的实际要求。

R程序应用

- R中，系统聚类的函数为**hclust()**，**dist()**函数用来计算距离矩阵，**plot()**函数可以画出系统聚类的谱系图，**rect.hclust()**函数用来给定类的个数或给定阈值来确定聚类的情况。

- (1)**dist()**的使用方法：
  dist(x,method="euclidean",diag=F,upper=F,p=2)

其中，x为数据矩阵或数据框。 method为计算方法，包括：euclidean（欧氏距离）、maximum（切比雪夫距离）、manhattan（绝对值距离）、nberra（兰氏距离）、minkoeski（明氏距离）。diag为是否包含对角线元素。upper为是否需要上三角。p为明氏距离的幂次。

- (2)**hclust()**的使用方法：
  hclust(d,method="ward.D",….)

其中，d为距离矩阵。method为系统聚类方法：single（最短距离法）、complete（最长距离法，缺省）、average（类平均法）、median（中间距离法）、centroid（重心法）、ward.D（ward法）。

- (3)**plot()**的使用方法：**plot(x, labels = NULL, hang = 0.1,axes = TRUE, frame.plot = FALSE, ann = TRUE, main = "Cluster Dendrogram", sub = NULL, xlab = NULL, ylab = "Height", ...)**

其中，x是由hclust()函数生成的对象。hang是表明谱系图中各类所在的位置，当hang取负值时，谱系图中的类从底部画起。其他参数见帮助文档。

❖ (4)**rect.hclust()**的使用方法：

rect.hclust(tree, k = NULL, which = NULL, x = NULL, h = NULL, border = 2, cluster = NULL)

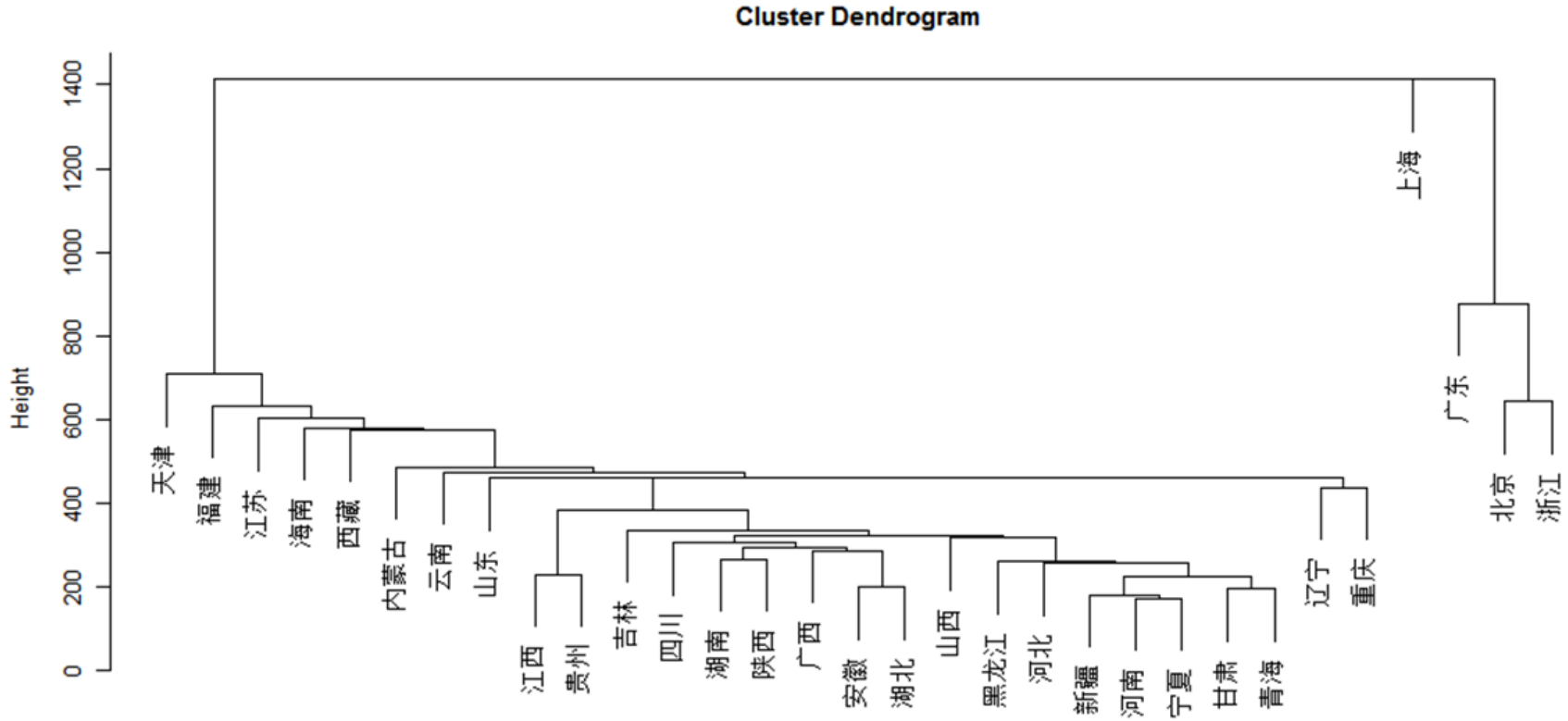
其中，tree是由hclust()生成的结构。k是类的个数。h是谱系图中的阈值，要求分成的各类的距离大于h。border是数或向量，表明矩形框的颜色。

课本例7.2 续例3.1，研究全国31个省、市、自治区2007年城镇居民生活消费的分布规律，根据调查资料做区域消费类型划分。

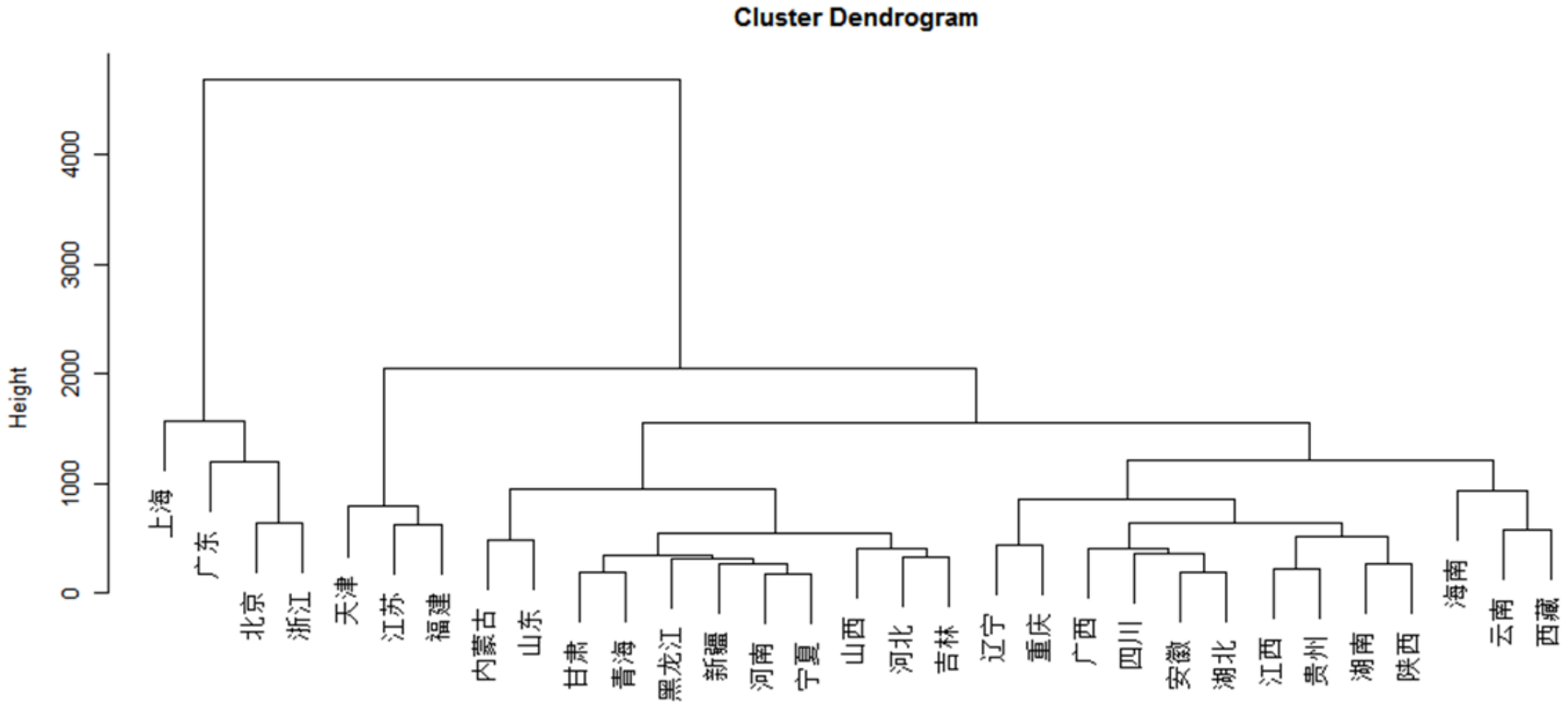
D=dist(X7.2);D

	北京	天津	河北	山西	内蒙古	辽宁	吉林
北京	0.0	1558.3	3081.6	3164.9	2800.4	2479.8	3030.1
天津	1558.3	0.0	1794.7	1930.4	1679.9	1105.2	1715.5
河北	3081.6	1794.7	0.0	342.8	607.0	818.0	335.0
山西	3164.9	1930.4	342.8	0.0	521.6	1004.6	409.6
内蒙古	2800.4	1679.9	607.0	521.6	0.0	886.5	508.9
辽宁	2479.8	1105.2	818.0	1004.6	886.5	0.0	746.4
吉林	3030.1	1715.5	335.0	409.6	508.9	746.4	0.0

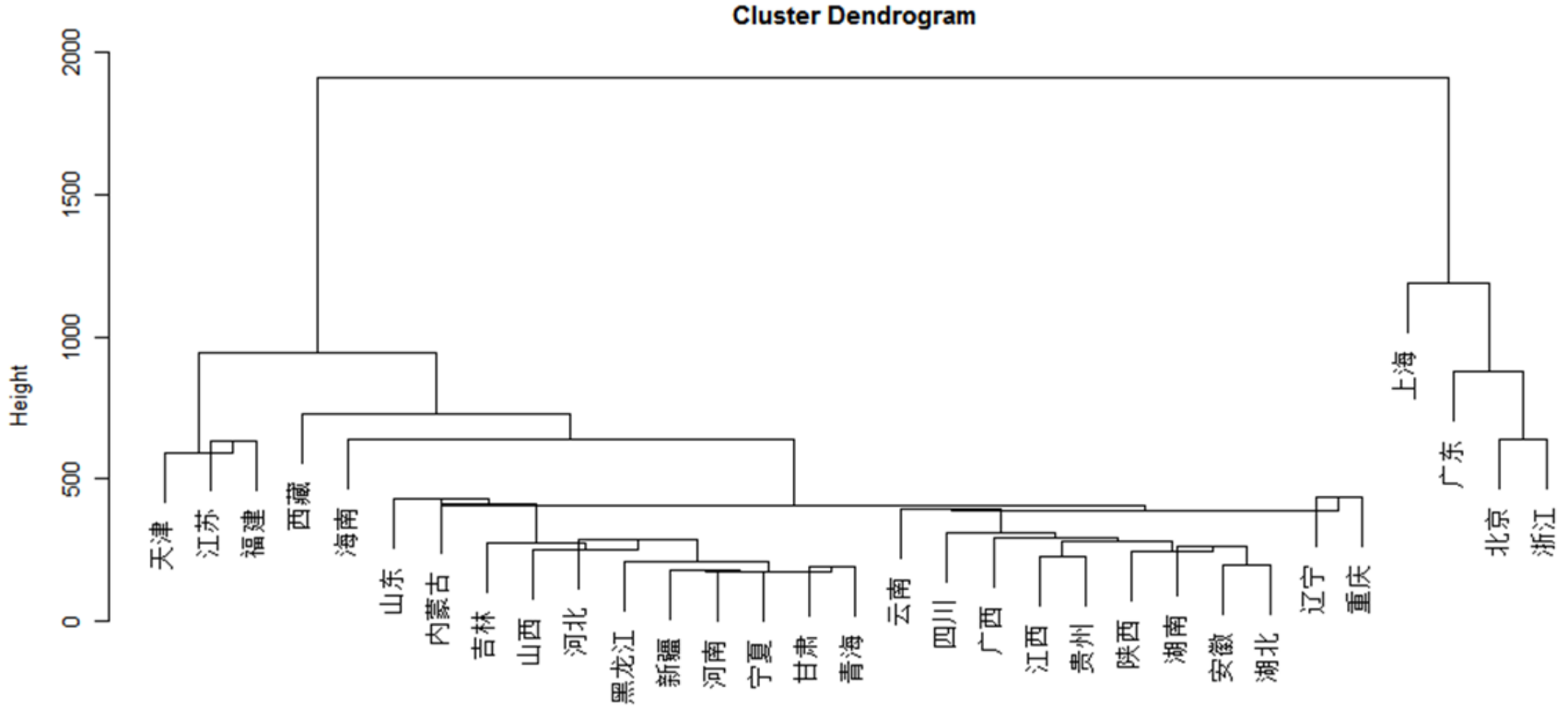
```
plot(hclust(D,'single'))    #最短距离法
```



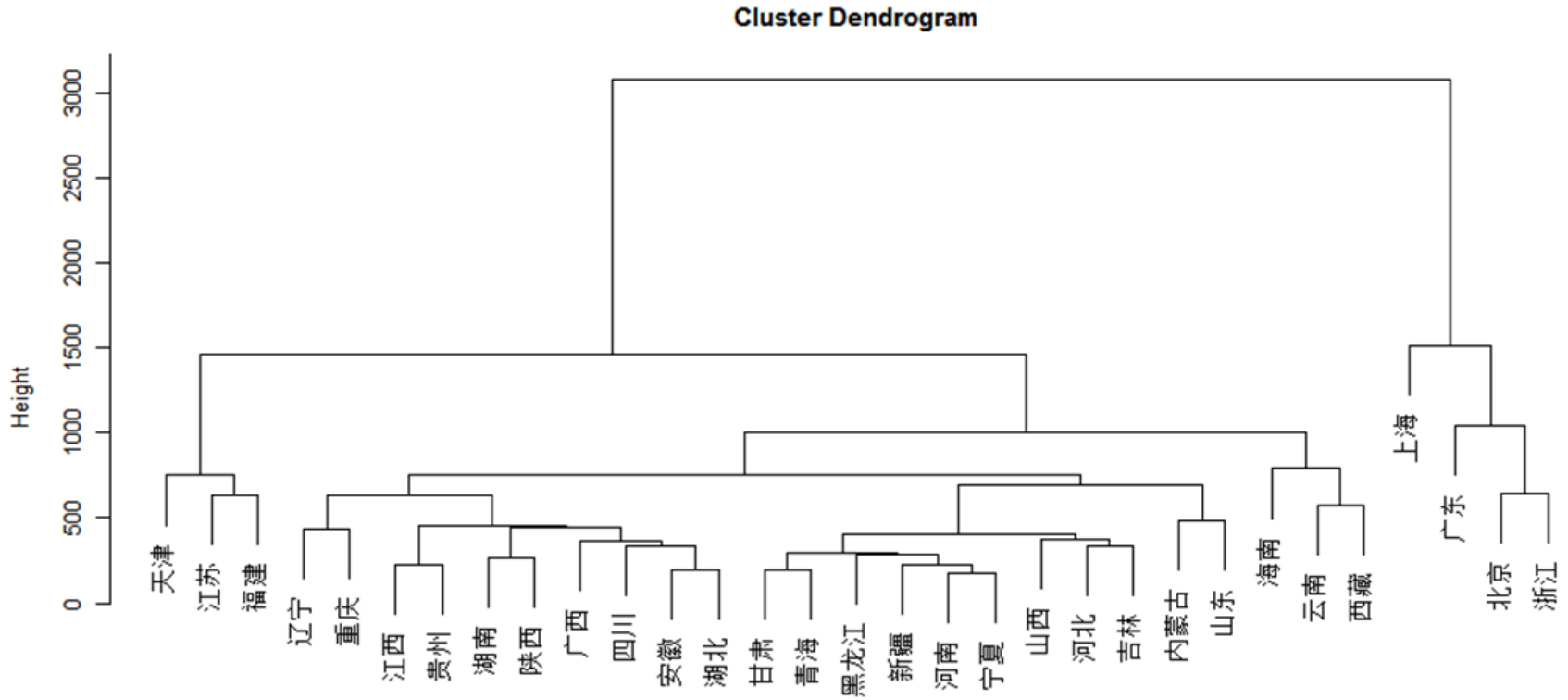
```
plot(hclust(D,'complete'))    #最长距离法
```



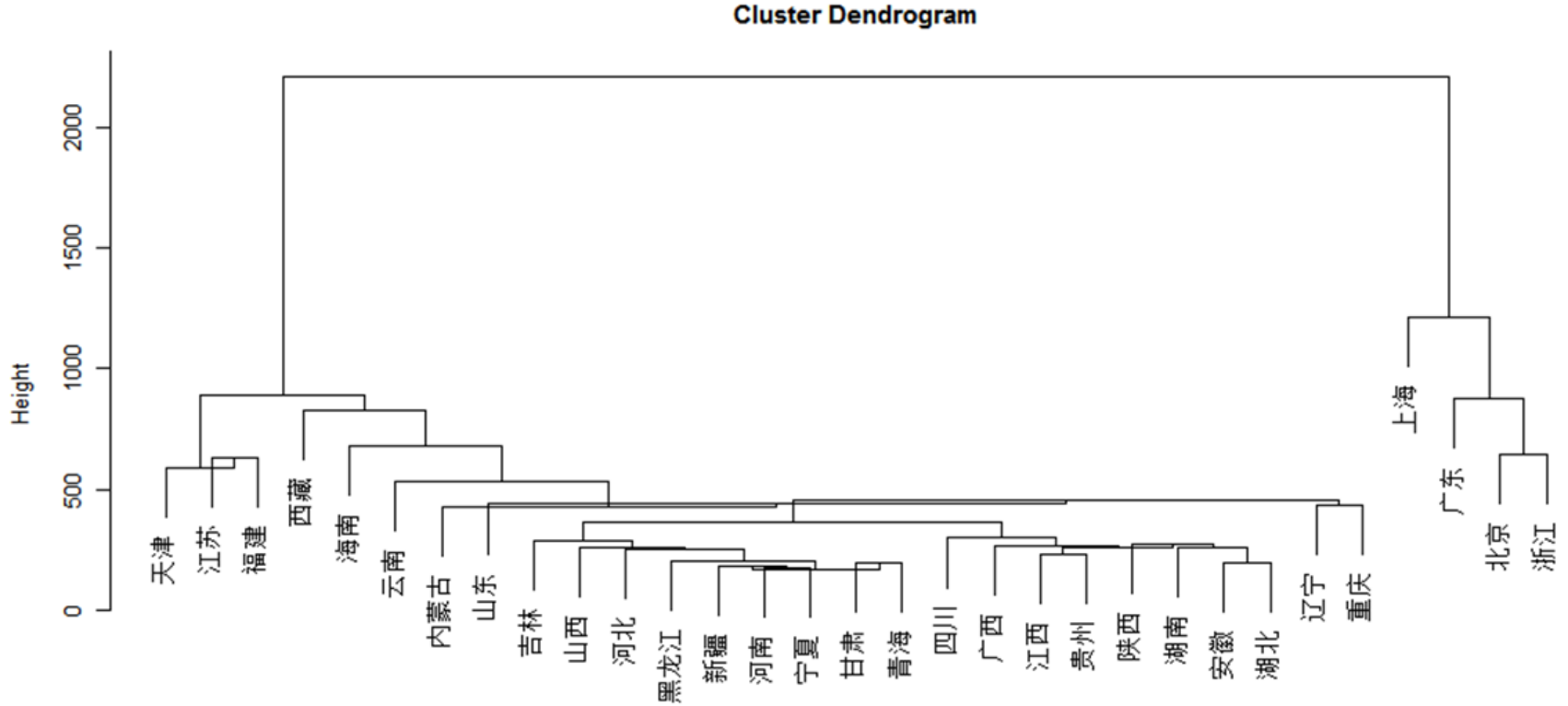
```
plot(hclust(D,'median'))      #中间距离法
```



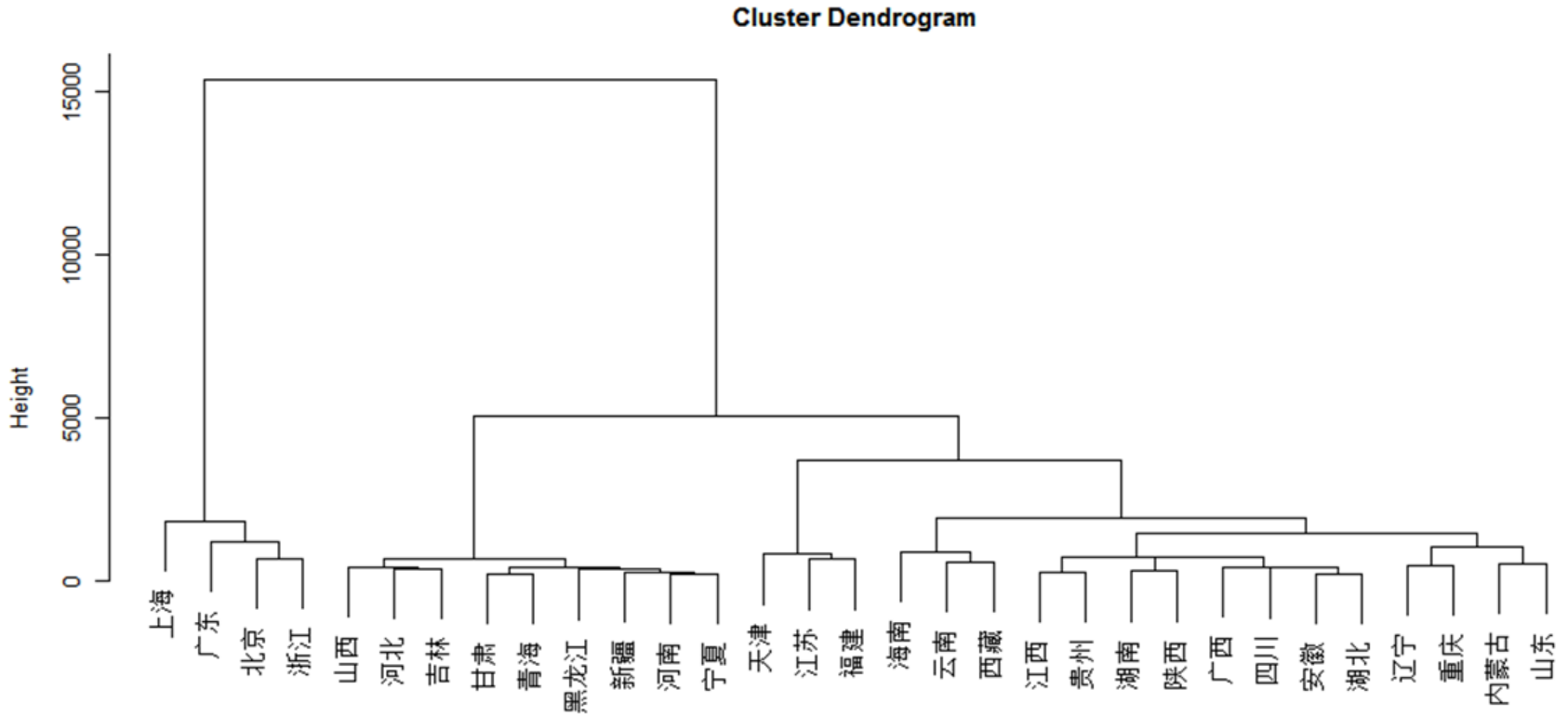
```
plot(hclust(D,'average'))     #类平均法
```



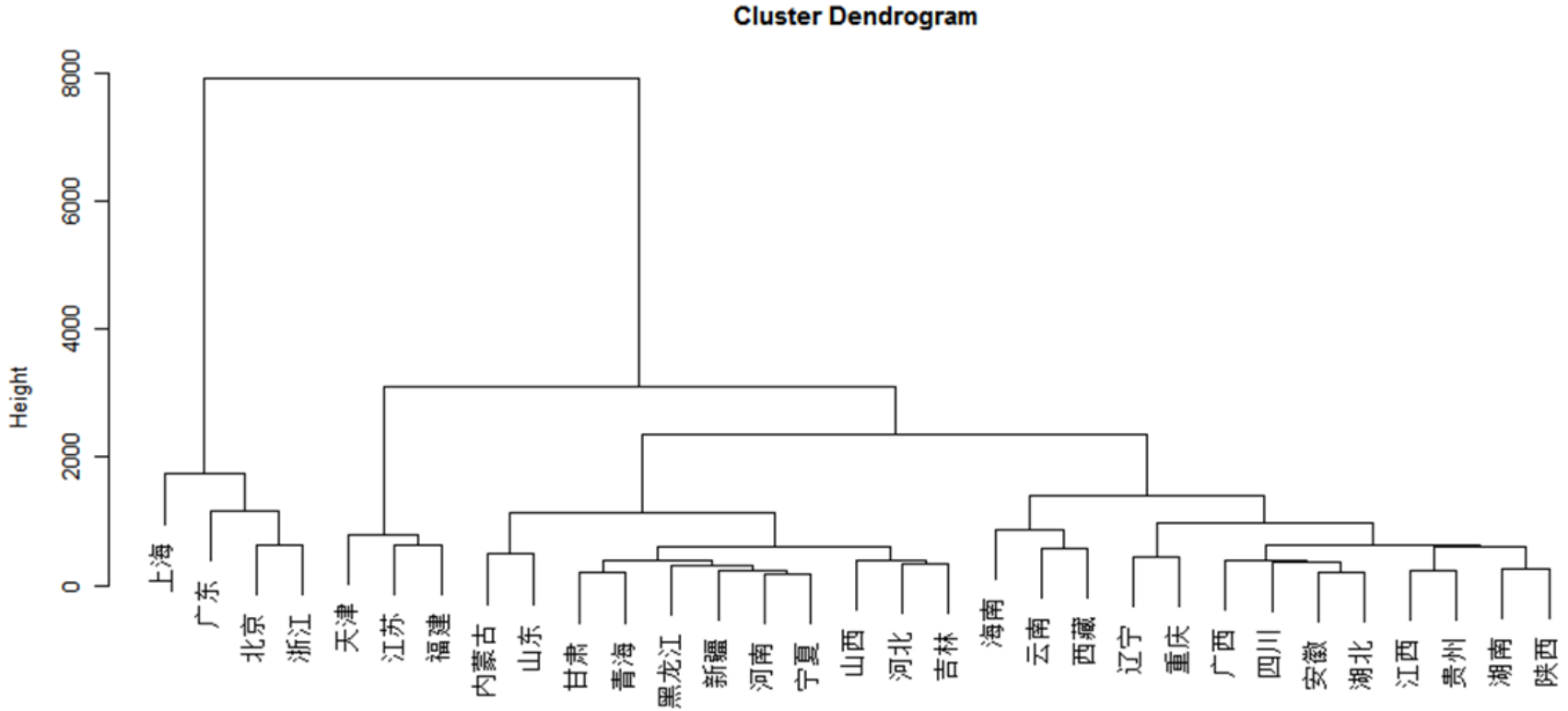
```
plot(hclust(D,'centroid'))      #重心法
```



```
plot(hclust(D,'ward.D'))      #ward.D 法
```

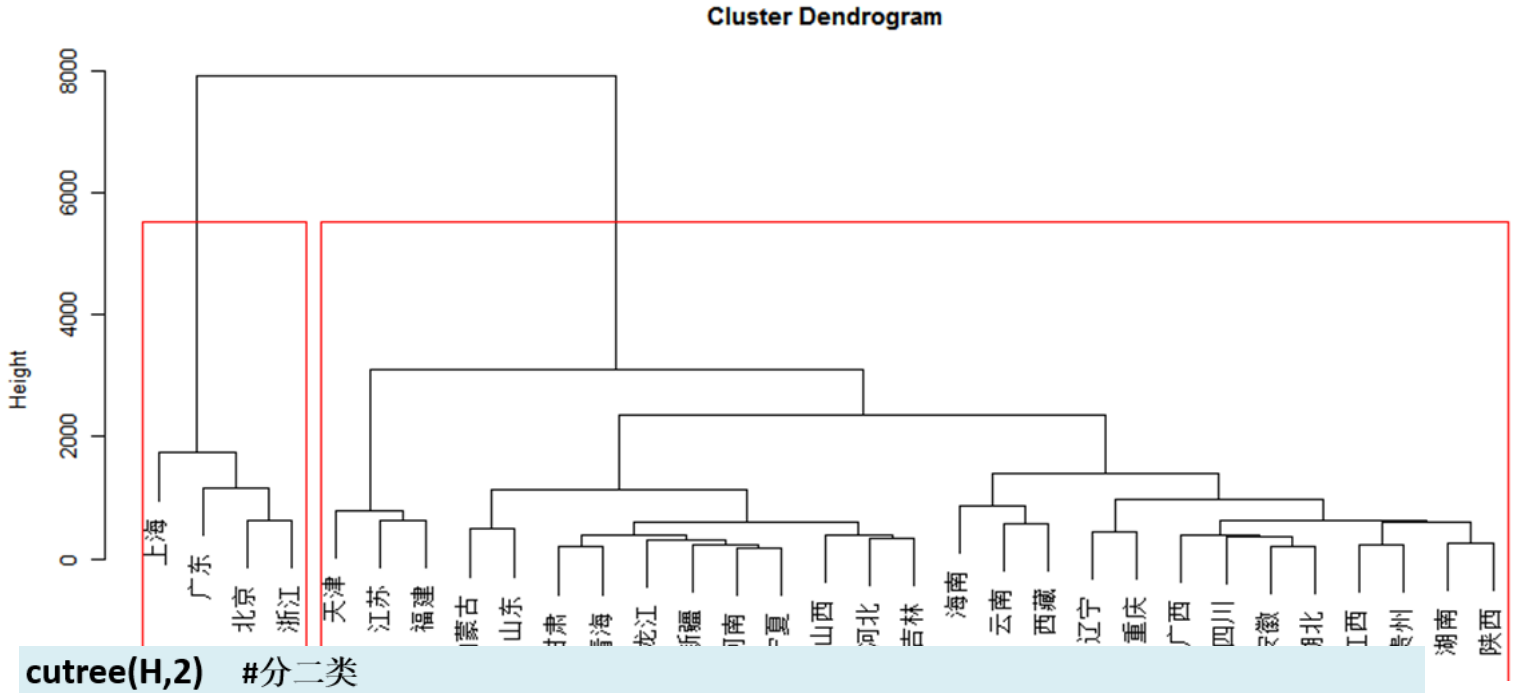


```
plot(hclust(D,'ward.D2'))      #ward.D2 法
```



```
H=hclust(D,'ward.D2')
```

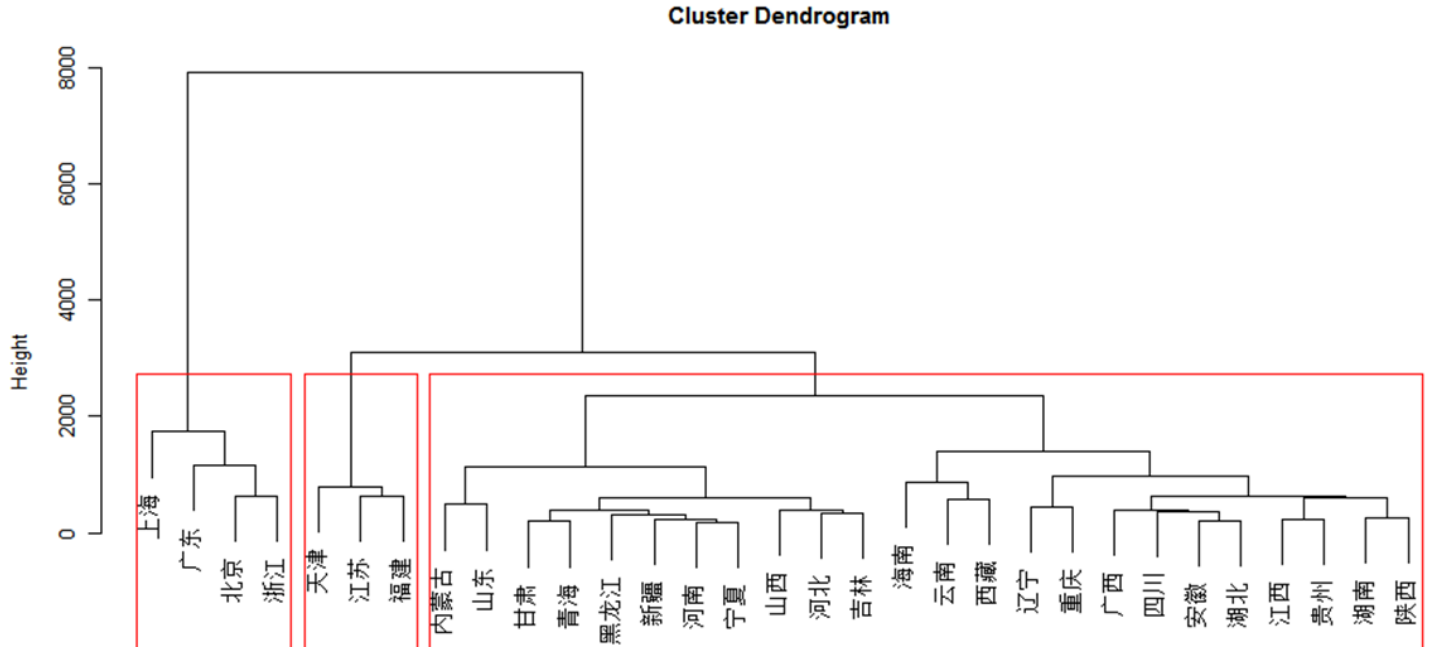
```
plot(H); rect.hclust(H,2)      #加二类框
```




```
cutree(H,2)    #分二类
```

北京	天津	河北	山西	内蒙古	辽宁	吉林	黑龙江	上海	江苏	浙江
1	2	2	2	2	2	2	2	1	2	1
安徽	福建	江西	山东	河南	湖北	湖南	广东	广西	海南	重庆
2	2	2	2	2	2	2	1	2	2	2
四川	贵州	云南	西藏	陕西	甘肃	青海	宁夏	新疆		
2	2	2	2	2	2	2	2	2		

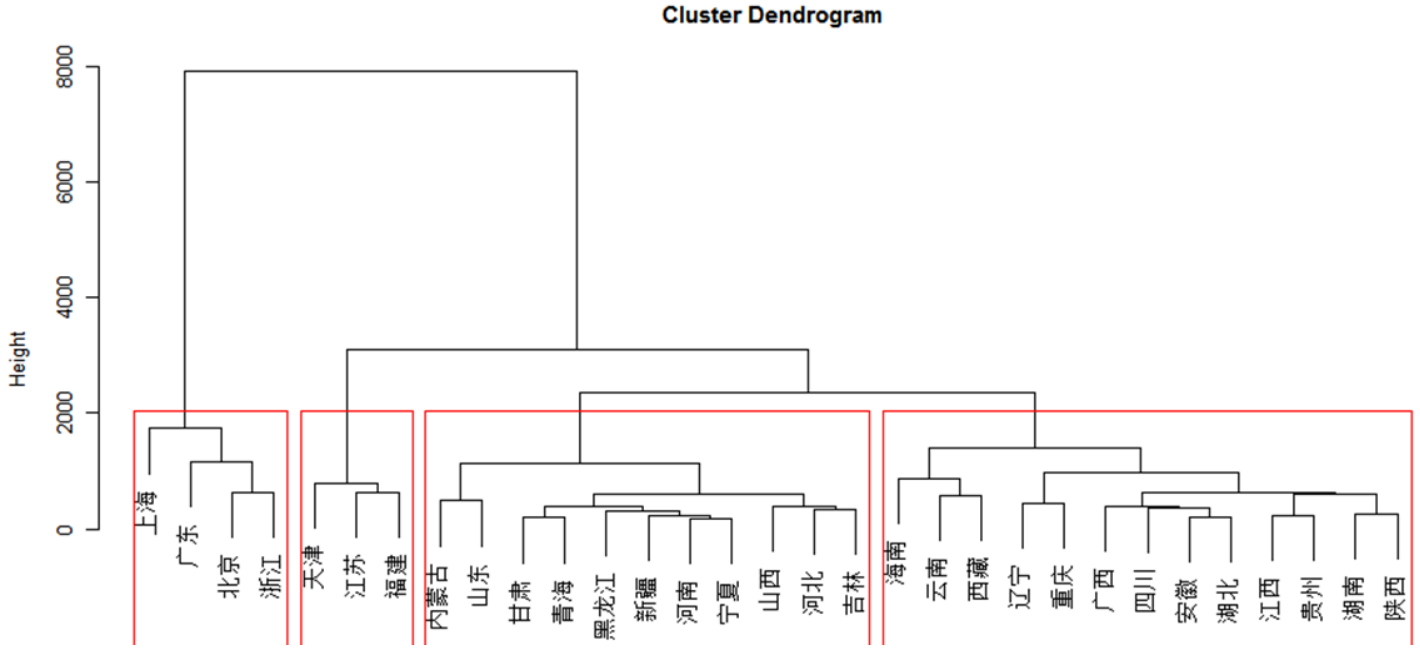
plot(H); rect.hclust(H,3) #加三类框



cutree(H,3) #分三类

北京	天津	河北	山西	内蒙古	辽宁	吉林	黑龙江	上海	江苏	浙江
1	2	3	3	3	3	3	3	1	2	1
安徽	福建	江西	山东	河南	湖北	湖南	广东	广西	海南	重庆
3	2	3	3	3	3	3	1	3	3	3
四川	贵州	云南	西藏	陕西	甘肃	青海	宁夏	新疆		
3	3	3	3	3	3	3	3	3		

plot(H); rect.hclust(H,4) #加四类框



cutree(H,4) #分四类

北京	天津	河北	山西	内蒙古	辽宁	吉林	黑龙江	上海	江苏	浙江
1	2	3	3	3	4	3	3	1	2	1
安徽	福建	江西	山东	河南	湖北	湖南	广东	广西	海南	重庆
4	2	4	3	3	4	4	1	4	4	4
四川	贵州	云南	西藏	陕西	甘肃	青海	宁夏	新疆		
4	4	4	4	4	3	3	3	3		

表 7-4　按类整理聚类图结果

分类	第一类	第二类		
分二类	北京、上海浙江、广东	天津、江苏、山东、河北、山西、内蒙古、辽宁、吉林、黑龙江、安徽、福建、江西、河南、湖北、湖南、广西、海南、重庆、四川、贵州、云南、西藏、陕西、甘肃、青海、宁夏、新疆		
分三类	第一类	第二类	第三类	
	北京、上海浙江、广东	天津、江苏、福建	山东、河北、山西、内蒙古、辽宁、吉林、黑龙江、安徽、江西、河南、湖北、湖南、广西、海南、重庆、四川、贵州、云南、西藏、陕西、甘肃、青海、宁夏、新疆	
分四类	第一类	第二类	第三类	第四类
	北京、上海浙江、广东	天津、江苏、福建	河北、山西、内蒙古、吉林、黑龙江、山东、河南、甘肃、青海、宁夏、新疆	辽宁、安徽、江西、湖北、湖南、广西、海南、重庆、四川、贵州、云南、西藏、陕西

七、使用图形作聚类及对聚类效果的评估

- 1.使用图形作直观的聚类
- 2.使用图形对聚类效果的评估

1.使用图形作直观的聚类

- 当p=2时，可以直接在散点图上进行主观的聚类，其效果未必逊于、甚至好于正规的聚类方法，特别是在寻找“自然的”类和符合我们实际需要的类方面。
- 当p=3时，我们可使用统计软件产生三维旋转图，通过三维旋转从各个角度来观测散点图，作直观的聚类。但由于其视觉效果及易操作性远不如平面散点图，故实践中很少采用。
- 当$p\geq3$时，有时我们可采用主成分分析或因子分析的技术将维数降至2（或3）维，然后再生成散点图（或旋转图），从直觉上进行主观的聚类。

寻找“自然的”类

2.使用图形对聚类效果的评估

- 经聚类分析已将类分好之后，常常希望从统计的角度看一下聚类的效果：不同类之间是否分离得较好，同一类内的样品（或变量）是否彼此相似。
- 通常可通过构造图形作直观的观测，所使用的图形有如下两种：
- (1)将p维数据画于平面图上，方法有平行（坐标）图、星形图、切尔诺夫脸谱图、星座图和安德鲁曲线图等；
- (2)使用费希尔判别的降维方法，将p维数据降至2（或3）维再构造散点图（或旋转图）。
- 如果方法(2)能够成功，则往往更值得推荐，尤其在样品数很大的场合下。

八、对变量的聚类

- 最短距离法、最长距离法和类平均法都属于连接方法，它们既可以用于样品的聚类，也能够用于变量的聚类。不过并非所有的系统聚类方法都适用于对变量的聚类。

例4 对305名女中学生测量八个体型指标：

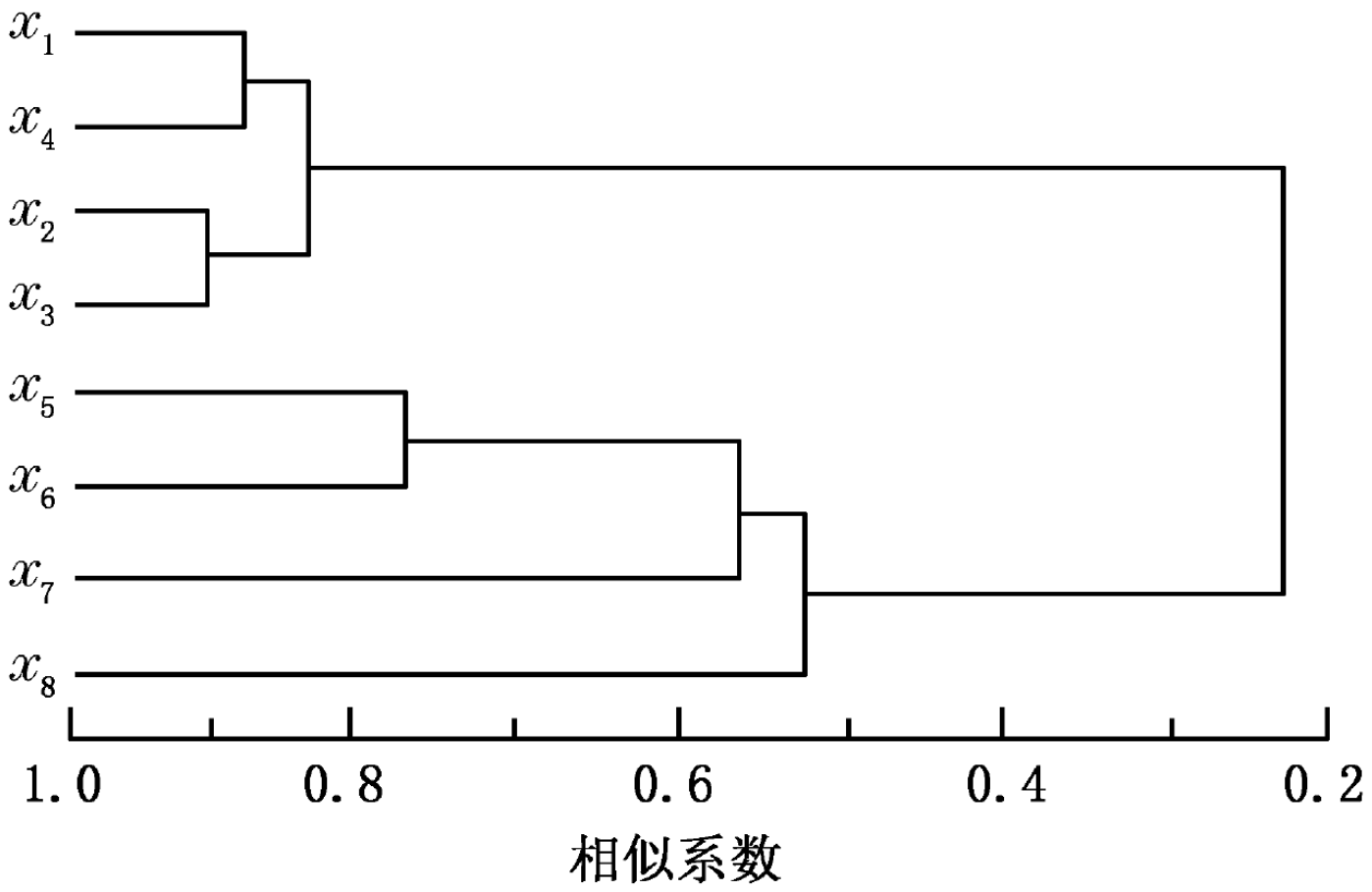
x_1：身高　x_2：手臂长　x_3：上肢长　x_4：下肢长

x_5：体重　x_6：颈围　x_7：胸围　x_8：胸宽

各对变量之间的相关系数

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	1.000							
x_2	0.846	1.000						
x_3	0.805	0.881	1.000					
x_4	0.859	0.826	0.801	1.000				
x_5	0.473	0.376	0.380	0.436	1.000			
x_6	0.398	0.326	0.319	0.329	0.762	1.000		
x_7	0.301	0.277	0.237	0.327	0.730	0.583	1.000	
x_8	0.382	0.415	0.345	0.365	0.629	0.577	0.539	1.000

- 单从该相关矩阵就可直观地判断出聚成两类：$\{x_1,x_2,x_3,x_4\}$和$\{x_5,x_6,x_7,x_8\}$，这两类的特征明显，其类内变量分别都是身材方面的“纵向”指标和“横向”指标。
- 分别用最短距离法、最长距离法和(类平均法对变量进行聚类，这三种方法的类与类之间的相似系数分别定义为两类变量间的最大、最小和平均相关系数，每次聚类时合并两个相似系数最大的类。
- 从图可见，聚成两类：$\{x_1,x_2,x_3,x_4\}$和$\{x_5,x_6,x_7,x_8\}$。
- 最短距离法和类平均法也都有与此相同的聚成两类的结果。



八个体型变量的最长距离法树形图

九、类的个数

- 如果能够分成若干很分开的类，则类的个数就比较容易确定；反之，如果无论怎样分都很难分成明显分开的若干类，则类个数的确定可能就比较困难了。
- 确定类个数的常用方法有：

  1.给定一个阈值T。

  2.观测样品的散点图。

1.给定一个阈值T

- 通过观测树形图，给出一个你认为合适的阈值T，要求类与类之间的距离要大于T，有些样品可能会因此而归不了类或只能自成一类。这种方法有较强的主观性，这是它的不足之处。

2.观测样品的散点图

- 如果样品只有两个（或三个）变量，则可通过观测数据的散点图（或旋转图）来主观确定类的个数。
- 如果变量个数超过三个，则可对每一可能考虑的聚类结果，将所有样品的前两个（或三个）费希尔判别函数得分制作成散点图（或旋转图），目测类之间是否分离得较好。
- 该图既能帮助我们评估聚类效果的好坏，也能帮助我们判断所定的类数目是否恰当。

§7.4 快速聚类法

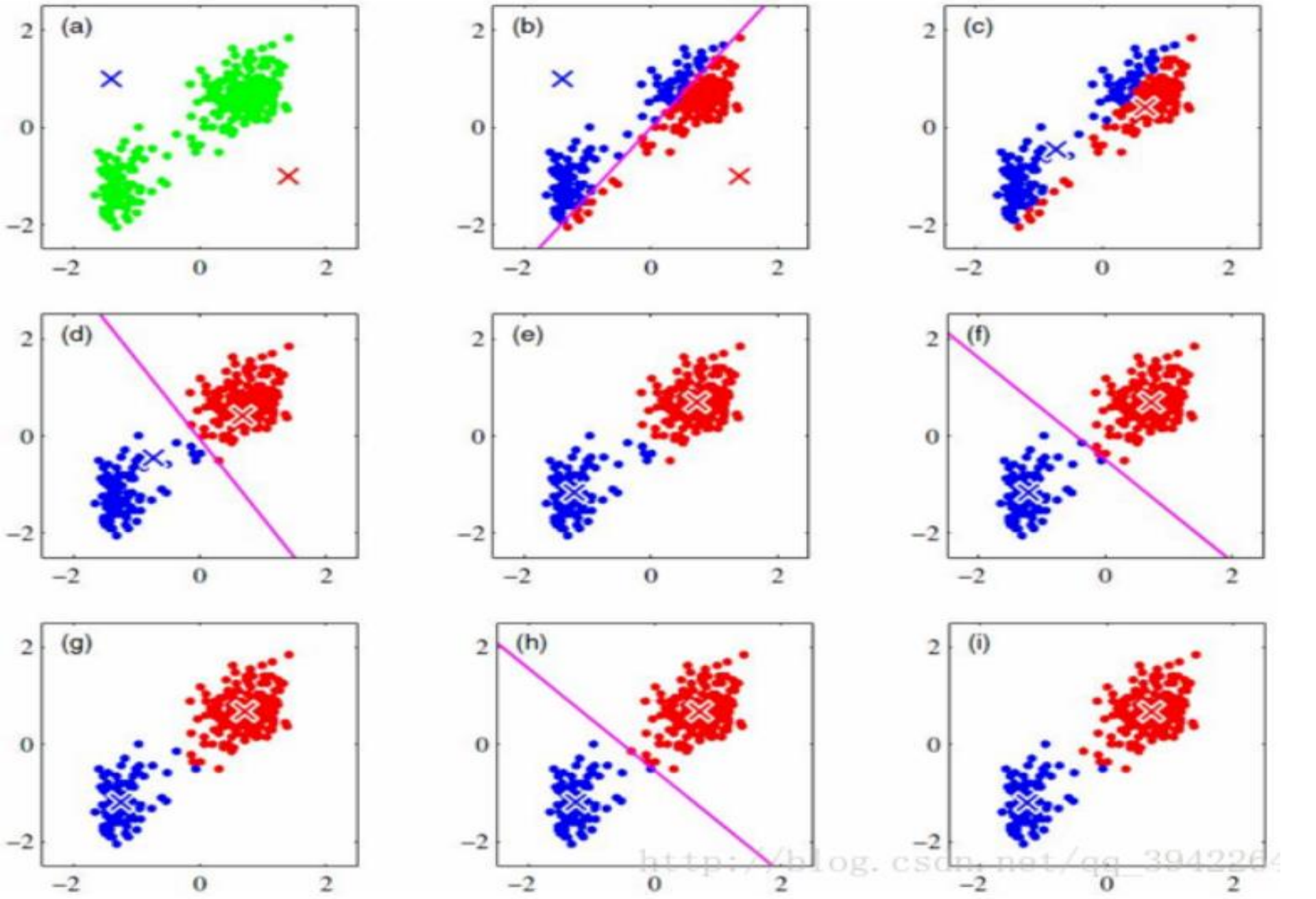
- 在系统聚类法中，对于那些先前已被“错误”分类的样品不再提供重新分类的机会，**k均值聚类法**是一种常用的**快速聚类法**，它允许样品从一个类移动到另一个类中。
- 快速聚类法的计算量要比建立在距离矩阵基础上的系统聚类法小得多。因此，使用快速聚类法计算机所能承受的样品数目n要远远超过使用系统聚类法所能承受的n。

*k*均值法(***k-means method***)的基本步骤

(1)选择*k*个样品作为初始凝聚点，或者将所有样品分成*k*个初始类，然后将这*k*个类的重心（均值）作为初始凝聚点。

(2)对所有的样品逐个归类，将每个样品归入凝聚点离它最近的那个类（通常采用欧氏距离），该类的凝聚点更新为这一类目前的均值，直至所有样品都归了类。

(3)重复步骤(2)，直至所有的样品都不能再分配为止。


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)

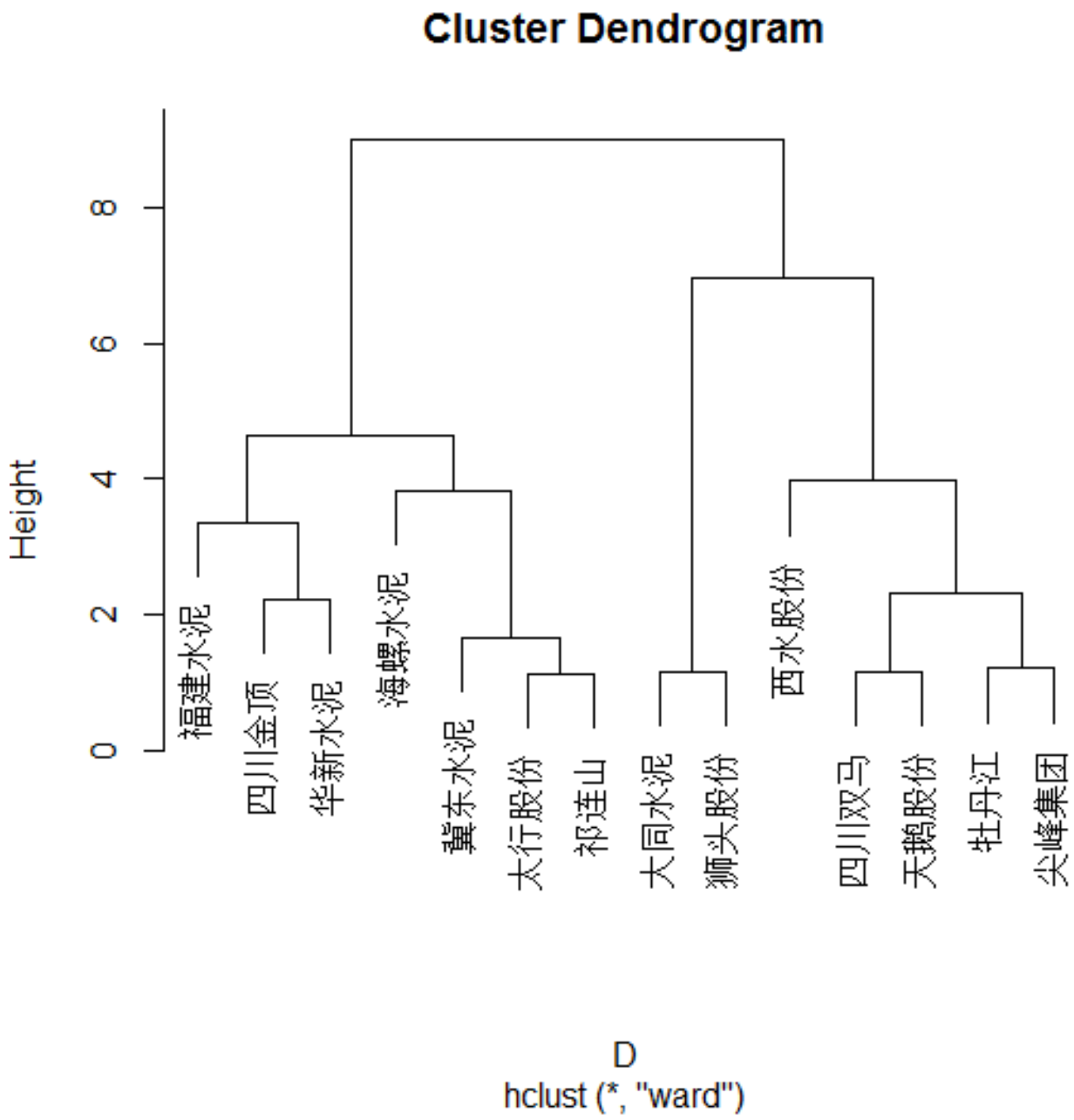
- 由于k均值法对凝聚点的初始选择有一定敏感性，故再试一下其他初始的凝聚点也许是个不错的想法。如果不同初始凝聚点的选择产生明显不同的最终聚类结果，或者迭代的收敛是极缓慢的，那么可能表明没有自然的类可以形成。
- k均值法有时也可用来改进系统聚类的结果，例如，先用类平均法聚类，然后将其各类的重心作为k均值法的初始凝聚点重新聚类，这可使得系统聚类时错分的样品能有机会获得重新的分类。不过，k均值法能否有效地改善系统聚类，我们不能一概而论，还应视聚类的最终结果而定。

案例1 对以下股票进行分类

x1：主营业务利润率 x2：销售毛利率 x3：速动比率
x4：资产负债率 x5：主营业务收入增长率 x6：营业利润增长率

	x1	x2	x3	x4	x5	x6
冀东水泥	33.8	34.75	0.67	59.77	15.49	16.35
大同水泥	27.54	28.04	2.36	35.29	-20.96	-46.45
四川双马	22.86	23.47	0.61	42.83	5.48	-49.22
牡丹江	19.05	19.95	1	48.51	-12.32	-65.99
西水股份	20.84	21.17	1.08	48.45	65.09	54.81
狮头股份	28.14	28.84	2.51	24.52	-6.43	-15.94
太行股份	30.45	31.13	1.02	46.14	6.57	-16.59
海螺水泥	36.29	36.96	0.27	58.31	70.85	117.59
尖峰集团	16.94	17.26	0.61	52.04	9.03	-94.05
四川金顶	28.74	29.4	0.6	65.46	-33.97	-55.02
祁连山	33.31	34.3	1.17	45.8	12.18	39.46
华新水泥	25.08	26.12	0.64	69.35	22.38	-10.2
福建水泥	34.51	35.44	0.38	61.61	23.91	-163.99
天鹅股份	25.52	26.73	1.1	47.02	-4.51	-68.79

complete
euclidean
0
50
100
200
300
福建水泥
尖峰集团
四川金顶
大同水泥
四川双马
牡丹江
天鹅股份
冀东水泥
祁连山
华新水泥
狮头股份
太行股份
西水股份
海螺水泥
D
hclust (*, "complete")



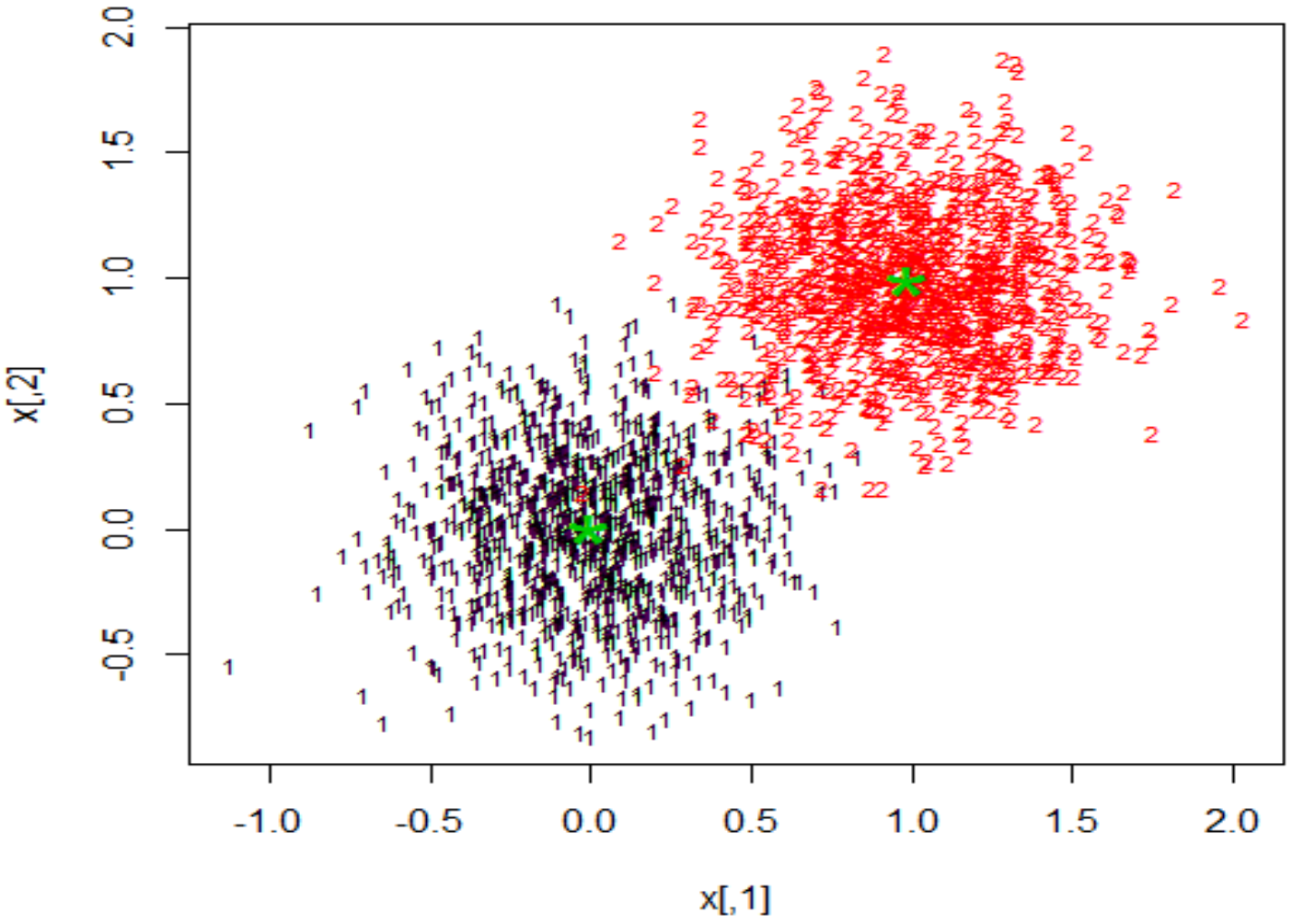
从分类图我们可以分三类：

第一类：福建水泥、四川金顶、华新水泥、海螺水泥、冀东水泥、太行股份、祁连山；

第二类：大同水泥、狮头股份；

第三类：西水股份、四川双马、天鹅股份、牡丹江、尖峰集团。

案例2 用服从正态分布的随机数模拟k均值分类



```
x=matrix(rnorm(1000,mean=0,sd=0.3),ncol=10)
 #均值为0，标准差为0.3的100*10的正态随机矩阵

y=matrix(rnorm(1000,mean=1,sd=0.3),ncol=10)
#均值为1，标准差为0.3的100*10的正态随机矩阵

z=rbind(x,y)
#按行合并为200*10的矩阵

cl<-kmeans(z,2)
#用k均值法分类
```

模拟任务：

(1)模拟10000个均值为0，标准差为0.3的正态分布随机数，再把这些随机数转化为10个变量、1000个观测的矩阵；

(2)然后再用同样的方法模拟10000个均值为1、标准差为0.3的正态分布数，再转化为10个变量、1000个观测的矩阵；

(3)然后把这两个矩阵合并成10个变量、2000个样品的数据矩阵；

(4)最后用k均值法将其聚类成两类，观察效果。