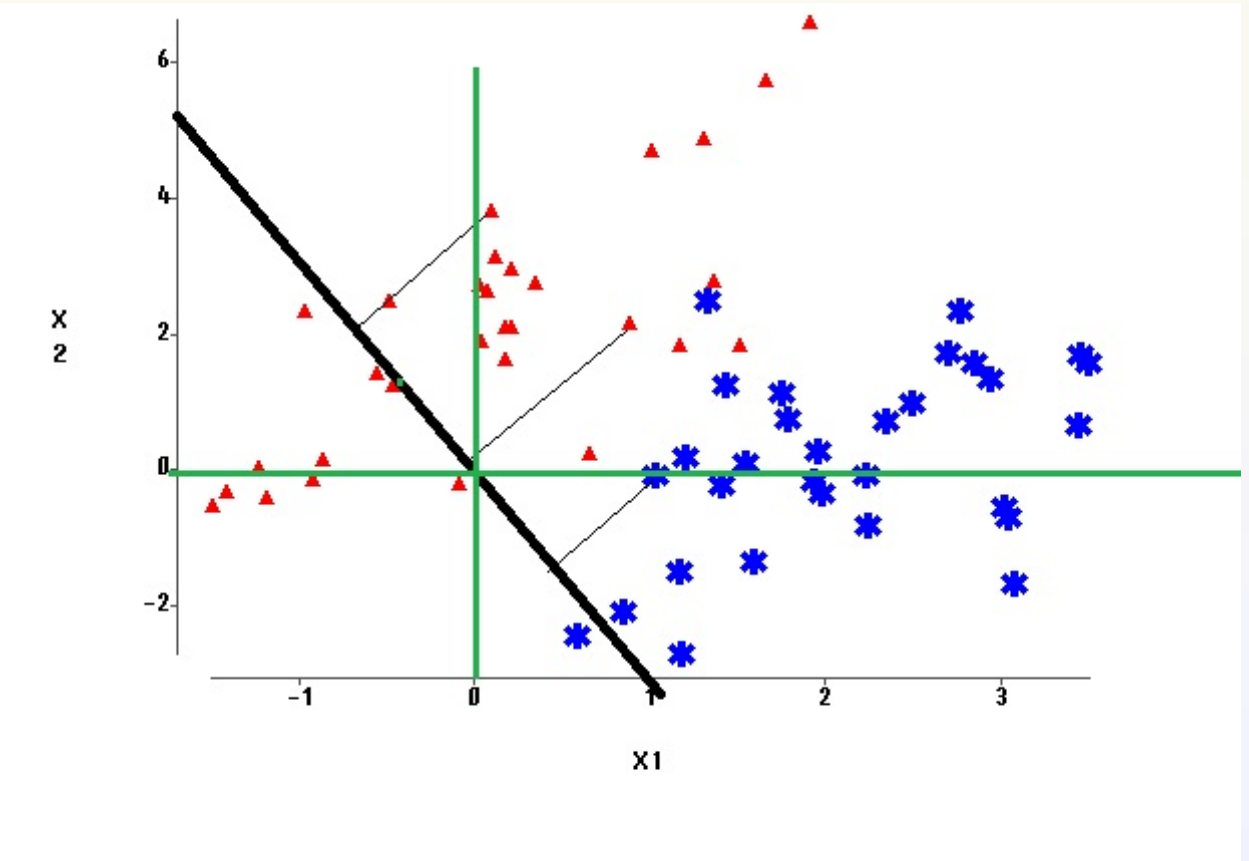
5.3 Fisher判别法

一、Fisher判别的基本思想

通过投影后，使组与组之间尽可能分开. 如下图当$m = 2, k = 2$时,寻找方向a, 使两组数据投影后在一维直线上尽可能区分开



利用线性投影和方差分析的思想.

设从m元总体$G_t(t=1,\cdots,k)$分别抽取样本如下:

$$X_{(i)}^{(t)}=(x_{i1}^{(t)},\cdots,x_{im}^{(t)}) \quad (t=1,\cdots,k;i=1,\cdots,n_t)$$

投影后为:

$$G_1: \quad a'X_{(1)}^{(1)},\cdots,a'X_{(n_1)}^{(1)}, \qquad \bar{X}^{(1)}=\frac{1}{n_1}\sum_{j=1}^{n_1}X_{(j)}^{(1)}$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$G_k: \quad a'X_{(1)}^{(k)},\cdots,a'X_{(n_k)}^{(k)}, \qquad \bar{X}^{(k)}=\frac{1}{n_k}\sum_{j=1}^{n_k}X_{(j)}^{(k)}$$

注：投影后可以看成是一维的分组样本，考虑样本总的平方和、组内平方和以及组间平方和。

投影后为一元数据,其组间平方和为

$$\begin{aligned} B_0 &= \sum_{t=1}^{k} n_t (a'\bar{X}^{(t)} - a'\bar{X})^2 \\ &= a'[\sum_{t=1}^{k} n_t (\bar{X}^{(t)} - \bar{X})(\bar{X}^{(t)} - \bar{X})']a \\ &= a'Ba \end{aligned}$$

其中B为组间离差阵

$$B = \sum_{t=1}^{k} n_t (\bar{X}^{(t)} - \bar{X})(\bar{X}^{(t)} - \bar{X})'$$

合并的组内平方和:

$$\begin{aligned} A_0 &= \sum_{t=1}^{k}\sum_{j=1}^{n_t}(a'X_{(j)}^{(t)} - a'\bar{X}^{(t)})^2 \\ &= a'Aa \end{aligned}$$

其中:

$$A = \sum_{t=1}^{k}\sum_{j=1}^{n_t}(X_{(j)}^{(t)} - \bar{X}^{(t)})(X_{(j)}^{(t)} - \bar{X}^{(t)})'$$

要使得投影后，k个总体（类）的均值差异尽可能的大. 即使得比值

$$\frac{a'Ba}{a'Aa} \overset{def}{=} \Delta(a)$$

尽可能的大.

更自然的定义方法: 考虑线性判别: $Y = a'X$ 满足:

$$\begin{aligned} \mu_{iY} &= E(Y|G_i) = a'E(X|G_i) = a'\mu_i, \\ \mathrm{Var}(Y|G_i) &= a'\mathrm{Cov}(X|G_i)a = a'\Sigma_i a. \end{aligned}$$

定义:

$$B_\mu = \sum_{i=1}^{k}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})', \ \bar{\mu} = \sum_{i=1}^{k}\mu_i, \ \Sigma = \sum_{i=1}^{k}\Sigma_i,$$

线性判别系数a 满足:

$$\widehat{a} = \mathrm{argmax}_a \left(\frac{\sum_{i=1}^{k}(\mu_{iY} - \bar{\mu})^2}{\sigma_Y^2} \right) = \mathrm{argmax}_a \left(\frac{a'B_\mu a}{a'\Sigma a} \right).$$

归纳为如下的规划问题：求$a \in R^m$使$a'Ba$ 在$a'Aa = 1$条件下达极大. 即

$$\begin{cases} \max_a \frac{a'Ba}{a'Aa}, \\ s.t. a'Aa = 1. \end{cases}$$

二、线性判别函数的求法

利用Lagranre方法，令$\varphi(a) = a'Ba - \lambda(a'Aa - 1)$

解方程组

$$\begin{cases} \frac{\partial \varphi}{\partial a} = 2(B - \lambda A)a = 0 \\ \frac{\partial \varphi}{\partial \lambda} = 1 - a'Aa = 0 \end{cases}$$

结论： 设$A^{-1}B$的非零特征值为$\lambda_1 \geq \cdots \geq \lambda_r > 0$，相应的满足约束条件的特征向量为$l_1, \cdots, l_r$,取$a = l_1$ 可使$\Delta(a)$达最大,且最大值为λ_1。

一般$\Delta(a)$称为判别效率. 由以上结论知，此时判别效率为λ_1。

附录：定理7.2

设B 是p 阶对称矩阵, λ_i 是B的第i大的特征值, l_i 是相应于λ_i的B 的标准化特征向量$(i = 1, \ldots, p)$ 为任一非零p 维向量, 那么有

$$\lambda_p \leq \frac{x'Bx}{x'x} \leq \lambda_1,$$

上式右边等号当$x = cl_1$时成立, 左边等号当$x = cl_p$时成立.

令$x = A^{\frac{1}{2}}a$, 则

$$\Delta(a) = \frac{x'Cx}{x'x}$$

其中$C = A^{-\frac{1}{2}}BA^{-\frac{1}{2}}$, 与$A^{-1}B$有相同的特征根.

定义5.3.1 设$A^{-1}B$的非零特征值为$0 < \lambda_1 \leq \cdots \leq \lambda_r$，相应的满足约束条件的特征向量为$l_1, \cdots, l_r$,称

$$P_1 = \lambda_1 / \sum_{i=1}^{r} \lambda_i$$

为线性判别函数$u_1(X) = l_1' X$的判别能力;称

$$P_{(l)} = (\lambda_1 + \cdots + \lambda_l) / \sum_{i=1}^{r} \lambda_i$$

为前l个$(l \leq r)$线性判别函数$u_1(X) = l_1' X, \cdots, u_l(X) = l_l' X$ 的累计判别能力.

三、Fisher判别准则

判别准则I(基本Fisher判别)

- 取最大特征根对应的特征向量l_1, 构作Fisher判别函数$l_1'X$, 将原来m维的总体简单化为一维总体的形式;
- 采用距离判别（或其它一维总体的判别方法）进行判别.

例5.3.1 若$k=2$,试求Fisher线性判别函数及其相应的判别效率.

解 两总体的组间离差阵B为

$$B=n_1(\bar{X}^{(1)}-\bar{X})(\bar{X}^{(1)}-\bar{X})'+n_2(\bar{X}^{(2)}-\bar{X})(\bar{X}^{(2)}-\bar{X})'$$

利用$\bar{X}=\frac{1}{n_1+n_2}(n_1\bar{X}^{(1)}+n_2\bar{X}^{(2)})$得

$$B=\frac{n_1n_2}{n_1+n_2}(\bar{X}^{(1)}-\bar{X}^{(2)})(\bar{X}^{(1)}-\bar{X}^{(2)})'.$$

$$B_0=a'Ba$$

合并组内离方差阵$A = A_1 + A_2$,其中

$$A_t = \sum_{i=1}^{n_t} (x_{(i)}^{(t)} - \bar{X}^{(t)})(x_{(i)}^{(t)} - \bar{X}^{(t)})' \quad t = 1, 2$$

$$A_0 = a' A a.$$

and

$$\Delta(a) = \frac{B_0}{A_0}.$$

由于B的秩等于1,故特征方程$|A^{-1}B - \lambda I| = 0$的非零特征根只有一个.

因为

$$A^{-1}B = \frac{n_1 n_2}{n_1 + n_2} A^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})(\bar{X}^{(1)} - \bar{X}^{(2)})'$$

有线性代数知:AB和BA的非零特征根相同.所以$A^{-1}B$的非零特征根等同于:

$$\frac{n_1 n_2}{n_1 + n_2}(\bar{X}^{(1)} - \bar{X}^{(2)})' A^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) = \frac{n_1 n_2}{n_1 + n_2} d^2,$$

其中

$$d^2 = (\bar{X}^{(1)} - \bar{X}^{(2)})' A^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})$$

因为$\frac{n_1 n_2}{n_1+n_2} d^2$是个数值,所以它就是所求的非零特征根$\lambda$.

记l为对应于λ的在条件$l'Al=1$的特征向量,满足$Bl=\lambda Al$,
所以取$l=\frac{1}{d}A^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$可满足上述条件.于是得Fisher线性判别函数为

$$u(X)=l'X=\frac{1}{d}X'A^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$$

相应的判别效率

$$\Delta(l)=\frac{n_1n_2}{n_1+n_2}(\bar{X}^{(1)}-\bar{X}^{(2)})'A^{-1}(\bar{X}^{(1)}-\bar{X}^{(2)})$$

序贯分析（sequential analysis）

- 序贯分析（sequential analysis）数理统计学的一个分支。其名称源出于美国统计学家瓦尔德在1947年发表的一本同名著作。
- 序贯最开始是指抽样时采用的一种方法。序贯抽样方案是指在抽样时，不事先规定总的抽样个数（观测或实验次数），而是先抽少量样本, 根据其结果，再决定停止抽样或继续抽样、抽多少，这样下去，直至决定停止抽样为止。
- 将序贯思想应用于判别分析，即得出Fisher判别的第二类判别方法。

判别准则II

设

$$u_{(ij)}^{(t)} = l_j' X_{(i)}^{(t)}, t = 1, \cdots, k, j = 1, \cdots, r, i = 1, \cdots, n_t$$

则k个m元样本在$l_1, l_2, \cdots, l_r$上投影后的均值分别为：

$$\bar{u}_j^{(t)} = \frac{1}{n} \sum_{i=1}^{n_t} u_{(ij)}^{(t)} = l_j' \bar{X}^{(t)}, j = 1, \cdots r, t = 1, \cdots, k.$$

投影后的方差分别为：

$$\hat{\sigma}_j^{(t)} = \frac{1}{n-1} \sum_{i=1}^{n_t} (u_{(ij)}^{(t)} - \bar{u}_j^{(t)})^2 = l_j' S_t l_j, j = 1, \cdots r, t = 1, \cdots, k.$$

序贯判别思想:

- 先取判别效率最大的线性判别函数$u_1(X) = l_1'X$,对样品X,计算它在l_1上投影:若存在唯一的i_1,使

$$\frac{|u_1(X) - \bar{u}_1^{(i_1)}|}{\hat{\sigma}_1^{(i_1)}} = \min_{t=1,\cdots,k} \frac{|u_1(X) - \bar{u}_1^{(t)}|}{\hat{\sigma}_1^{(t)}}$$

  时,判$X \in G_{i_1}$.

- 如果存在j个总体$G_{k_1}, \cdots, G_{k_j}$,使其与$u_1(X)$距离相等且为最小,记序号集$L = \{k_1, \cdots, k_j\}$, 则再取判别效率为λ_2(次大)的判别函数$u_2(X) = l_2'X$.

当存在唯一的i_2,使

$$\frac{|u_2(X)-\bar{u}_2^{(i_2)}|}{\hat{\sigma}_2^{(i_2)}}=\min_{t\in L}\frac{|u_2(X)-\bar{u}_2^{(t)}|}{\hat{\sigma}_2^{(t)}}$$

时,判$X\in G_{i_2}$.

- 如果第二个判别函数仍不能判别样品X 所属总体,则还可以取第三个线性判别函数,依此类推.

有了Fisher判别函数(即将多空间上的点投影到一维)后，可采用一维的判别方法给出判别准则. 下面以$k=2$为例推导出按距离准则判别样品归类的判别法. 设两总体均值为$\bar{X}^{(1)}$和$\bar{X}^{(2)}$,线性判别函数的数值为

$$\bar{\mu}^{(1)}=l'\bar{X}^{(1)},\bar{\mu}^{(2)}=l'\bar{X}^{(2)}$$

若投影后两总体方差相等,则阈值点

$$\bar{\mu}=\frac{1}{2}(l'\bar{X}^{(1)}+l'\bar{X}^{(2)})$$

若投影后两总体方差不等,则阈值点

$$\mu^{*}=\frac{\hat{\sigma}_2\bar{\mu}^{(1)}+\hat{\sigma}_1\bar{\mu}^{(2)}}{\hat{\sigma}_1+\hat{\sigma}_2}$$

其中$\hat{\sigma}_t$是投影后总体G_t的样本方差,且$\hat{\sigma}_t=l'S_tl$

判别准则　(不妨设$l'\bar{X}^{(1)} > l'\bar{X}^{(2)}$)

- 投影后方差相等

$$\begin{cases} X \in G_1 & u(X) > \bar{\mu} \\ X \in G_2 & u(X) < \bar{\mu} \\ \text{待判} & u(X) = \bar{\mu} \end{cases}$$

- 投影后方差不等

$$\begin{cases} X \in G_1 & u(X) > \mu^* \\ X \in G_2 & u(X) < \mu^* \\ \text{待判} & u(X) = \mu^* \end{cases}$$

判别准则III(主成分分析思想)

如果有r个非零特征根,相应的有r个线性判别函数$u_1(X), \cdots,$ $u_r(X)$, 这时, 相当于把原来m 个变量综合成r 个新变量.

在实用中常取$l \leq r$,且满足

$$\sum_{i=1}^{l} \lambda_i / \sum_{i=1}^{r} \lambda_i \geq P_0 \quad (p_0\text{一般取}0.7)$$

这样m元总体的判别问题即化为l元总体的判别问题(降维), 而且l个新变量互不相关,可按距离判别准则来归类.

例5.3.2 对表5.2中胃癌检验的生化指标值用主成分判别法进行判别归类.

- 第一步：均值检验和协方差检验
- 第二步：加载adegenet(MASS, ade4)
- 第三步：读入数据
  - $D522 <- read.table("ex522.txt", header = F)$
  - $X <- D522[, 2:5]$
  - $Y <- D522[, 1]$

- 第四步：进行主成分判别
  - $re1 < -dapc(X, Y)$,# 会出现累积贡献率图，
  - Choose the number PCs to retain (>= 1): 1 #根据图选择主成分的个数
  - Choose the number discriminant functions to retain (>= 1): 1 # 选择用于主成分判别的主成分个数

- 第五步：回判
  - $predict(re1, X)$

```
$assign
 [1] 1 1 1 2 3 3 1 3 3 2 2 3 1 2 3
Levels: 1 2 3

$posterior
           1         2         3
 [1,] 0.5891914 0.2285853 0.1822233
 [2,] 0.6750288 0.1856064 0.1393648
 [3,] 0.4794356 0.2804507 0.2401138
 [4,] 0.3300522 0.3441481 0.3257997
 [5,] 0.1210917 0.4068540 0.4720543
 [6,] 0.1785776 0.3951101 0.4263123
 [7,] 0.5728086 0.2365553 0.1906362
 [8,] 0.1631301 0.3989081 0.4379618
 [9,] 0.2429939 0.3758935 0.3811126
[10,] 0.2755332 0.3646495 0.3598173
[11,] 0.3200217 0.3480557 0.3319226
[12,] 0.1944713 0.3908279 0.4147007
[13,] 0.3874317 0.3208019 0.2917665
[14,] 0.3048609 0.3538515 0.3412876
[15,] 0.1104645 0.4081241 0.4814114
```

- 采用一个主成分的判别效果并不好。
- 如果选用二个主成分去判别，出输出结果:

```
$assign
 [1] 1 1 1 1 1 2 3 2 3 2 3 2 2 3 3
Levels: 1 2 3
```

Figure: 主成分判别回判结果

- 采用二个主成分后，第一类判别正确，但第二第三类仍不理想。

R中还有一些用于判别分析的包：

例如：Mixture Discriminant Analysis, 包的名称：mda, 函数

- mda(formula, data, subclasses, sub.df, tot.df, dimension, eps, iter, weights, method, keep.fitted, trace, ...)
- fda(formula, data, weights, theta, dimension, eps, method, keep.fitted, ...)

R中还有一些用于判别分析的包：

- rda: rda provides classification for high dimensional data by means of shrunken centroid regularized discriminant analysis;
- class: it provides k-nearest neighbours by knn().
- knncat: it provides k-nearest neighbours for categorical variable.
- SensoMineR: it provides FDA() for factorial discriminant analysis.
- A number of packages provide for dimension reduction with the classification, such as klaR, superpc, gpls, hddplot, ROCR, predbayescor, etc.

判别效果的检验及各变量判别能力的检验

判别能力依赖:

- (1)样本是否来自不同的总体; 即: 进行总体均值是否相等的假设检验.
- (2)m 个判别指标区组能力; 即: 进行各变量判别能力的假设检验.

假设总体G_i 的分布为$N_m(\mu^{(t)}, \Sigma_t)$, $t = 1, 2, \cdots, k$, $X_{(i)}^{(t)}$, $t = 1, 2, \cdots, k$; $i = 1, 2, \cdots, n_t$, 为来自G_t 的m 元样本.

一、两总体判别效果的检验

$$H_0 : \mu^{(1)} = \mu^{(2)}$$

根据第三章的结论, 计算两总体之间的马氏距离

$$d^2(1,2) = (\overline{X}^{(1)} - \overline{X}^{(2)})' S^{-1} (\overline{X}^{(1)} - \overline{X}^{(2)})$$

构造F统计量

$$F = \frac{(n_1 + n_2 - m - 1) n_1 n_2}{m(n_1 + n_2)(n_1 + n_2 - 2)} d^2(1,2) \sim F(m, n_1 + n_2 - m - 1)$$

两个总体均值有显著性时，对两总体讨论判别问题是有意义的.
但当两个总体均值没有显著性差异时，如果盲目地应用判别分析的方法进行判别分析，则错判的机会很大，判别分析没有意义.

二、k个总体判别效果的检验

第一步：检验各总体均值是否有显著差异

$$H_0 : \mu^{(1)} = \mu^{(2)} = \cdots = \mu^{(k)}$$

根据第三章的结论, 利用似然比原则导出Λ 统计量

$$\Lambda = \frac{|A|}{|A+B|} = \frac{|A|}{|T|} \sim \Lambda(m, n-k, k-1)$$

第二步：如果上述检验显著，则进行两两配对检验，说明各总体之间的差异

$$H_0^{(ij)}:\mu^{(i)}=\mu^{(j)}$$

如果假设各总体的协方差阵相等，则采用马氏距离构造的F统计量

$$F_{ij}=\frac{(n-k-m+1)}{m(n-k)}\frac{n_in_j}{n_i+n_j}d^2(i,j)\sim F(m,n-k-m+1)$$

如果检验发现$H_0^{(ij)}$ 为真，则将第i 类和第j 类合并.

第三步：分类检验结束后，检验各变量判别能力

- 变量判别能力的度量
- 变量判别能力的检验

变量判别能力的度量

- 消去法求行列式的值

$$|A| = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{vmatrix} = a_{11} \begin{vmatrix} 1 & \frac{a_{12}}{a_{11}} & \cdots & \frac{a_{1m}}{a_{11}} \\ 0 & a_{22}^{(1)} & \cdots & a_{2m}^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{m2}^{(1)} & \cdots & a_{mm}^{(1)} \end{vmatrix}$$

$$= a_{11} \begin{vmatrix} a_{22}^{(1)} & \cdots & a_{2m}^{(1)} \\ \vdots & \vdots & \vdots \\ a_{m2}^{(1)} & \cdots & a_{mm}^{(1)} \end{vmatrix} = a_{11} a_{22}^{(1)} \cdots a_{mm}^{(m-1)}.$$

- 上述的消去法并不一定要从a_{11}开始，可以从$a_{i_1 i_1}$ 开始

$$|A| \quad = \quad = a_{i_1 i_1} a_{i_2 i_2}^{(1)} \cdots a_{i_m i_m}^{(m-1)}.$$

- 对$|T|$也可以进行类似的计算

$$|T| \quad = \quad = t_{i_1 i_1} t_{i_2 i_2}^{(1)} \cdots t_{i_m i_m}^{(m-1)}.$$

仍从统计量

$$\Lambda = \frac{|A|}{|A+B|} = \frac{|A|}{|T|}$$

开始

- 构造单个变量的判别能力统计量

(1) m个指标的判别能力统计量

$$\begin{aligned}\Lambda_{(m)} &= \frac{|A|}{|T|} = \frac{a_{11}a_{22}^{(1)}\cdots a_{mm}^{(m-1)}}{t_{11}t_{22}^{(1)}\cdots t_{mm}^{(m-1)}} \\ &= U_{(1,2,\cdots,m)}\end{aligned}$$

度量m个指标$X_1,\cdots,X_m$ 对k 个总体的判别能力，$\Lambda_{(m)}$越小，说明判别效果越好.

(2) 现考虑前$m-1$个变量的判别能力统计量

$$\begin{aligned}\Lambda_{(m-1)} &= \frac{|A_{m-1}|}{|T_{m-1}|} = \frac{a_{11}a_{22}^{(1)}\cdots a_{(m-1)(m-1)}^{(m-2)}}{t_{11}t_{22}^{(1)}\cdots t_{(m-1)(m-1)}^{(m-2)}} \\ &= U_{(1,2,\cdots,m-1)}\end{aligned}$$

(3)定义第m个指标的判别能力统计量记：

$$U_{m|1,2,\cdots,m-1}=\frac{U_{(1,2,\cdots,m)}}{U_{(1,2,\cdots,m-1)}}=\frac{a_{mm}^{(m-1)}}{t_{mm}^{(m-1)}}$$

显然，$U_{m|1,2,\cdots,m-1}$ 表示第m个指标X_m 对k 个总体的判别能力，$U_{m|1,2,\cdots,m-1}$ 越小，说明判别效果越好.

(4)定义第j个指标的判别能力统计量对于第j个变量的判别能力可以用类似的统计量来

$$U_{j|1,2,\cdots,j-1,j+1,\cdots,m-1}=\frac{U_{(1,2,\cdots,m)}}{U_{(1,2,\cdots,j-1,j+1,\cdots,m-1)}}$$

- 构造单个变量的判别能力检验

(1) 若已经r个变量$X_{i_1},\cdots,X_{i_r},r<m$ 的k 个总体的判别效果显著. 其判别能力的统计量为

$$U_{(i_1,i_2,\cdots,i_r)} = \frac{a_{i_1i_1}a^{(1)}_{i_2i_2}\cdots a^{(r-1)}_{(i_r)(i_r)}}{t_{i_1i_1}t^{(1)}_{i_2i_2}\cdots t^{(r-1)}_{i_ri_r}}$$

(2) 考虑增加新的变量$X_{i_{r+1}}$，相应的对k 个总体的判别能力会提高（至少保持原来的判别能力）. 统计量为

$$U_{(i_1,i_2,\cdots,i_r,i_{r+1})} = U_{(i_1,i_2,\cdots,i_r)}\cdot U_{i_{r+1}|(i_1,i_2,\cdots,i_r)}.$$

其中：

$$U_{i_{r+1}|(i_1,i_2,\cdots,i_r)}=\frac{a^r_{i_{r+1}i_{r+1}}}{t^{(r)}_{i_{r+1}i_{r+1}}}.$$

注意：统计量$U_{i_{r+1}|(i_1,i_2,\cdots,i_r)}$检验的是除去了$X_{i_1},\cdots,X_{i_r}$个变量后，新增加的变量$X_{i_{r+1}}$的判别能力. 即检验假设

$$H_0:\mu^{(1)}_{i_{r+1}|(i_1,\cdots,i_r)}=\mu^{(2)}_{i_{r+1}|(i_1,\cdots,i_r)}=\cdots=\mu^{(k)}_{i_{r+1}|(i_1,\cdots,i_r)}$$

- 判别能力统计量的分布

在假设总体为正态的前提下，可以证明

(1) $U_{(i_1,i_2,\cdots,i_r)} \sim \Lambda(r, n-k, k-1)$

(2) $U_{(i_1,i_2,\cdots,i_r,i_{r+1})} \sim \Lambda(r+1, n-k, k-1)$

(3) $U_{i_{r+1}|(i_1,i_2,\cdots,i_r)} \sim \Lambda(1, n-k-r, k-1)$.

利用Λ 统计量与F统计量的关系，对H_0 给出检验.

1. 向前法
2. 向后法
3. 逐步法

逐步法的主要思想：

逐个引入变量，每次把一个判别能力最强的变量引入判别式，每引入一个新变量，对判别式中的老变量逐个进行检验，如其判别能力因新变量的引入而变得不显著，应把它从判别式中剔除。

基本步骤

- 引入所有变量中判别能力最强的一个变量.
  (一元方差分析，分析每个变量的差别能力，用F大的为最有差别能力)
- 检验变量对k类判别是否显著; 如果显著则引入，如果不显著则无法建立判别式，即所有的变量对总体无法判别, 变量选择过程停止, 需要引入新的变量.
- 第一个变量引入后，再从其余的变量中选择一个判别能力再强的变量，检验第二个变量的判别能力，如果显著则引入，如果不显著，判别变量选择过程结束.

- 第二个变量引入后，对于引入的变量X_{i_1}, X_{i_2} 重新进行检验，从中先取一个判别能力最弱的变量进行检验，如果显著，则进入第三个变量的选择，如果不显著，则把入选的变量重新删除，然后进入第三个变量的选择。
- ……
- 如果现在有r个变量已经入选，在引入$r+1$个变量时，要先对r 个变量中判别能力最弱的一个作检验，显著则进一下个变量的选择，如果不显著则要删除变量，并对剩余的$r-1$个变量中最弱的一个作检验，
- 如果没有变量可以入选，也没有变量可以删除，则变量选择过程最后结束.

Stepwise Diagonal Discriminant Analysis: SDDA

Usage: sdda(X, y, priors, start = rep(FALSE, ncol(X)), never = rep(FALSE, ncol(X)),method="lda",...)

- **X**: Training data matrix - rows are observations, columns are variables.
- **y**: A factor of true class labels, or a numeric vector with values 1, 2, 3, ... G, where G is the number of classes.
- **priors**: Prior probabilities for the different classes, if left unspecified these default to equal probability to belong to each group
- ....

- $D511 <- read.table("ex511.txt", header = F)$
- $attach(D511)$
- $A <- as.factor(V5)$
- $MD511 <- as.matrix(D511[, 1:4])$
- $s1 <- sdda(MD511, A)$
- $s2 <- predict(s1, MD511)$
- $table(s2, A)$

第六章、聚类分析

November 27, 2018

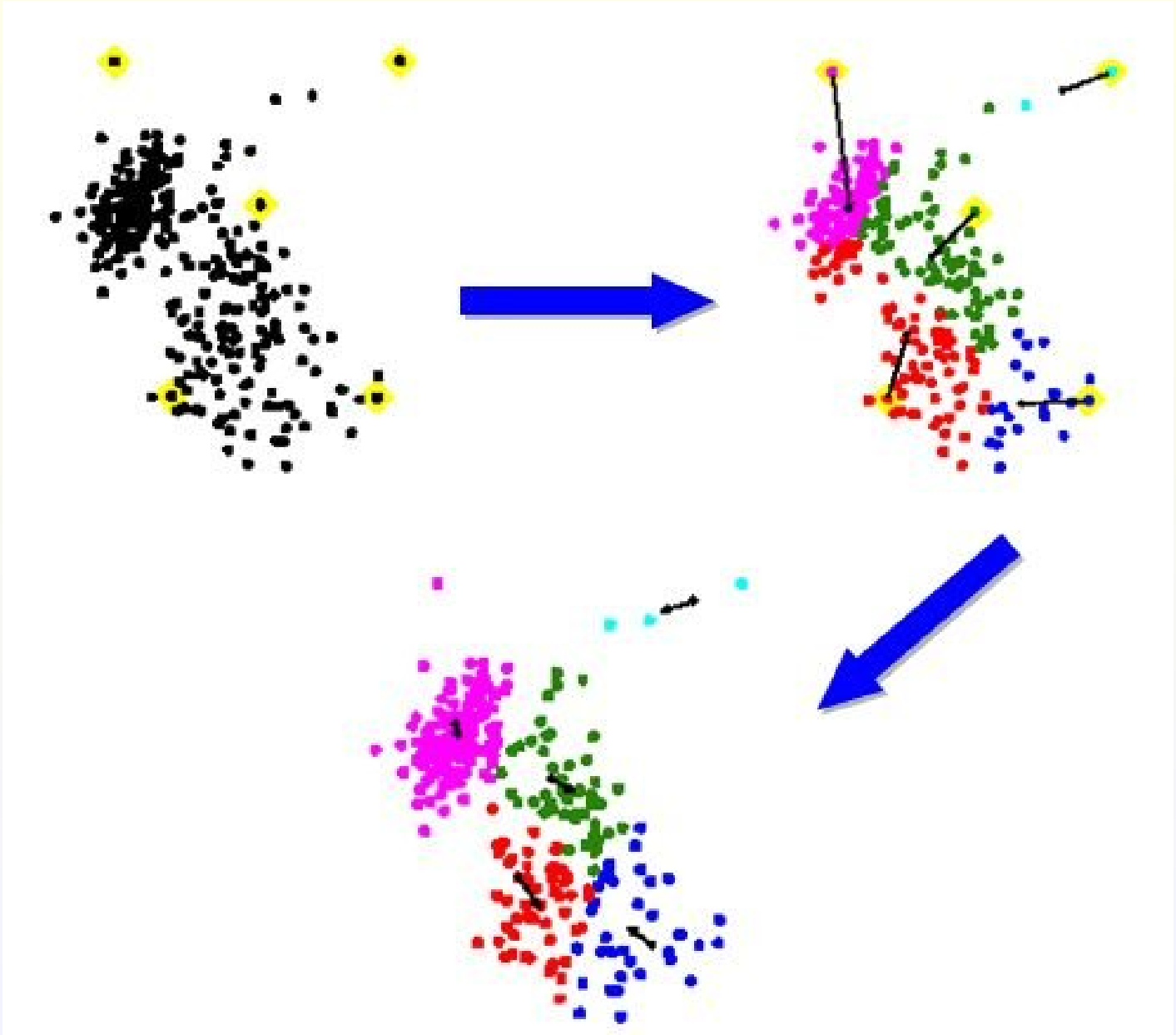
聚类分析又称群分析, 顾名思义是一种分类的多元统计分析方法。按照个体或样品(individuals, objects or subjects) 的特征将它们分类，使同一类别内的个体具有尽可能高的同质性(homogeneity)，而类别之间则应具有尽可能高的异质性(heterogeneity)。

例如

1. 对某城市大气污染的轻重分成几类区域;
2. 对某年级学生按各科的学习情况分为几种类型;
3. 在经济学中根据人均国民收入,人均工农业产值,人均消费水平等指标对世界上所有国家的经济发展状况进行分类等等.

我们也可以对变量进行聚类—分类，但是更常见的还是对个体分类（样本聚类）。为了得到比较合理的分类，

- 首先要采用适当的指标来定量地描述研究对象（样本或变量，常用的是样本）之间的联系的紧密程度。常用的指标为“距离”和“ 相似系数”，假定研究对象均用所谓的“点”来表示。
- 然后根据紧密程度归类. 一般的规则是将“距离”较小的点或“相似系数”较大的点归为同一类，将“距离”较大的点或“相似系数”较小的点归为不同的类！（一般的相似系数就是相关系数了）



6.1 聚类分析的方法

聚类分析按其聚类方法分为:

1. 系统聚类法
2. 调优法(动态聚类法)
3. 最优分割法(有序样品聚类法)
4. 模糊聚类法：利用模糊集理论来处理分类问题。
5. 图论聚类法：利用图论中最小支撑树的概念来处理分类问题
6. 聚类预报法：

聚类分析按根据分类对象的不同分为R型和Q型,R型对变量进行分类,Q型对样品进行分类.

R型聚类分析的目的:

1. 相关性：了解变量间及变量组合间的亲疏关系
2. 分类：对变量进行分类
3. 降维：根据分类结果及它们之间的关系,在每一类中选择有代表的变量作为重要变量,利用少数几个重要变量进一步作分析计算,如进行回归分析等

Q型聚类分析的目的： 对样品进行分类.(注意分类与判别的区别)

6.2 距离与相似系数

为了对样品(或变量)进行分类,就必须研究它们之间的关系, 距离与相似系数是用来描述样品间亲疏关系的统计量.

- 定量变量：即连续变量，例如：长度、重量、产量、温度等.
- 定性变量：又分有序变量和名义变量. 有序变量是指虽然没有明确的数量关系，但有次序关系，如质量等级；名义变量的取值仅是分类的作用，没有序关系，例如性别和职业.

一、数据的变换方法

定量数据在进行聚类分析前要进行数据变换.

设有n个样品,每个样品有m个指标(变量),观测数据为$x_{ij}(i=1,\cdots,n;j=1,\cdots,m)$.

$$\text{均值}\bar{x}_j=\frac{1}{n}=\sum_{t=1}^{n}x_{tj}\quad(j=1,\cdots,m)$$

$$\text{标准差}s_j=\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(x_{tj}-\bar{x}_j)^2}\quad(j=1,\cdots,m)$$

$$\text{极差}R_j=\max_{t=1,\cdots,n}x_{tj}-\min_{t=1,\cdots,n}x_{tj}\quad(j=1,\cdots,m)$$

1.中心化变换

$$x_{ij}^{*}=x_{ij}-\bar{x}_{j}\quad(i=1,\cdots,n;j=1,\cdots,m)$$

变换后数据的均值为0,协方差阵不变.

2.标准化变换

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_{j}}{s_{j}}\quad(i=1,\cdots,n;j=1,\cdots,m)$$

标准化变换后,每个变量的样本均值为0,标准差为1,且标准化后的数据无量纲.

3.极差标准化变换

$$x_{ij}^{*}=\frac{x_{ij}-\bar{x}_{j}}{R_{j}}\quad(i=1,\cdots,n;j=1,\cdots,m)$$

变换后,每个变量的样本均值为0,极差为1,且变换后的数据无量纲.

4.极差正规化变换

$$x_{ij}^{*}=\frac{x_{ij}-\min_{t=1,\cdots,n} x_{tj}}{R_j} \quad (i=1,\cdots,n; j=1,\cdots,m)$$

变换后的数据$0 \leq x_{ij}^{*} \leq 1$;极差为1,无量纲.

5.对数变换

$$x_{ij}^{*}=\ln(x_{ij}) \quad (x_{ij}>0; i=1,\cdots,n; j=1,\cdots,m)$$

它将指数特征的数据结构变换为线性结构.

R中数据标准化的方法

1. $D641 <- read.table("ex641.txt", header = F)$# 读入数据
2. $D641A <- D641[, 2:7]$
3. 函数：scale
   - $scale(D641A, center = T, scale = T)$ #按列规范化
     说明: $z = (x - \bar{x})/\sigma$, $\sigma = \sqrt{\sum (x_i - \bar{x})^2/(n-1)}$.
4. 函数：sweep
   - $center <- sweep(D641A, 2, apply(D641A, 2, mean))$#按列中心化
   - $R <- apply(D641A, 2, max) - apply(D641A, 2, min)$#按列计算极差
   - $D641A1 <- sweep(center, 2, R, "/")$#按列极差标准化
   - $D641A1$

二、样品间的距离与相似系数

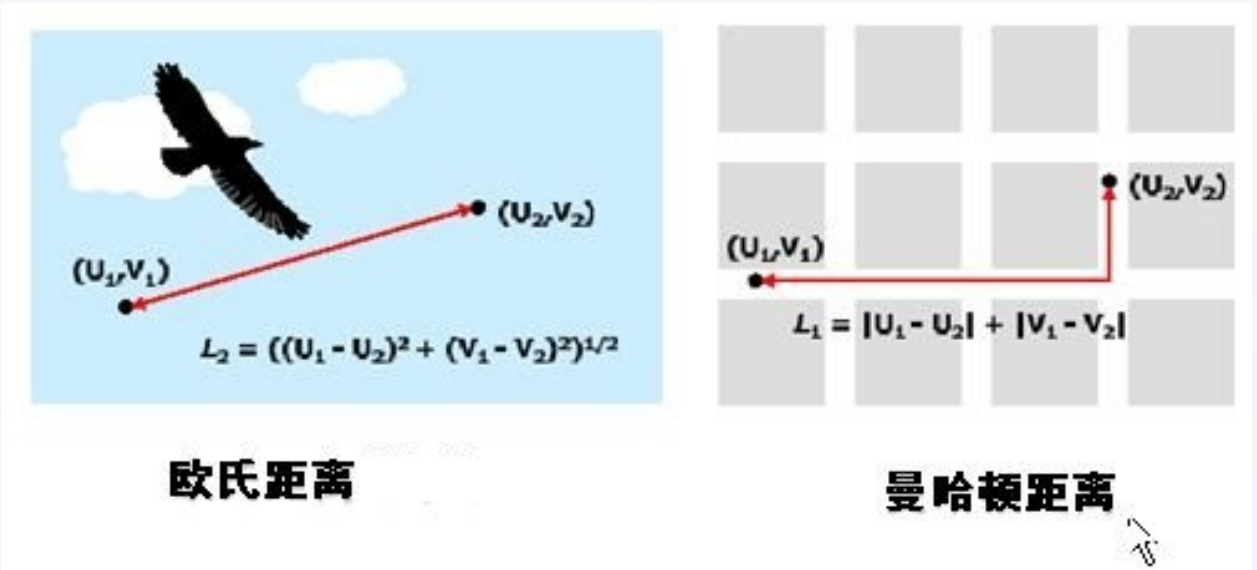
描述样品间的亲疏程度最常用的是距离.用d_{ij}表示样品$X_{(i)}$和$X_{(j)}$的距离,一般要求:

1. $d_{ij} \geq 0$ 对一切的i, j; $d_{ij} = 0$等价于$X_{(i)} = X_{(j)}$;
2. $d_{ij} = d_{ji}$ 对一切的i, j;
3. $d_{ij} \leq d_{ik} + d_{kj}$ 对一切的i, j, k(三角不等式)

1.闵科夫斯基(Minkowski)距离

$$d_{ij}(q)=\left[\sum_{t=1}^{m}|x_{it}-x_{jt}|^{q}\right]^{1/q}\quad(i,j=1\cdots,n)$$

- 当$q=1$时称为**绝对值距离(曼哈顿距离)**.
- 当$q=2$时称为**欧式距离**.
- 当q趋于∞时,称为**切比雪夫距离**.



- 欧式距离是最常用的距离,但该距离与量纲有关,没有考虑变量间的相关性,也没考虑各变量方差的不同,会造成方差大的变量在距离中的作用就会大.

$$d_{ij}(2) = \left[\sum_{t=1}^{m} |x_{it} - x_{jt}|^2\right]^{1/2}.$$

- 简单的处理就是对各变量加权,如以$1/s^2$作为权重可得出“**统计距离**”

$$d_{ij}^*(2) = \left[\sum_{t=1}^{m} \left(\frac{x_{it} - x_{jt}}{s_t}\right)^2\right]^{1/2}.$$

2.兰氏距离

$$d_{ij}(L)=\frac{1}{m}\sum_{t=1}^{m}\frac{|x_{it}-x_{jt}|}{x_{it}+x_{jt}}\quad(i,j=1\cdots,n)$$

- 兰氏距离是由Lance 和Williams 最早提出的.
- 这个量无量纲,克服了闵氏距离与各指标的量纲有关的缺点.
- 对较大的奇异值不敏感, 特别适合高度偏倚的数据.
- 但没有考虑到变量间的相关性.

3.马氏距离

$$d_{ij}(M) = (X_{(i)} - X_{(j)})' S^{-1} (X_{(i)} - X_{(j)})$$

表示样品$X_{(i)}$和$X_{(j)}$的马氏距离,S^{-1}表示样本协方差的逆矩阵.

- 可以排除变量之间相关性, 并且不受量纲的影响.
- 一般用同各个类的样本来计算各自的协方差阵.
- 样品间的马氏距离需要用到样本协方差阵来计算, 类的形成需要依赖马氏距离, 而样本间合理的马氏距离又依赖于分类,这就形成了一个恶性循环. 所以马氏距离不是理想的距离.

R软件中各种距离的计算：

函数dist(x, method=”？”, diag=F, upper=F, p=2)

1. method距离计算方法
2. diag=T输出对角线上的距离
3. upper=T输出上三角矩阵

method距离可选择：

- euclidean
- maximum
- manhattan (绝对值)
- minkowski, (p: the power of the Minkowski distance)
- binary, $d_{ij} = \frac{m_2}{m_1+m_2}$
- canberra (Lance 距离)

4.斜交空间距离

为使具有相关性变量的谱系结构不发生变化引进斜交空间距离

$$d_{ij}=\left[\frac{1}{m^2}\sum_{k=1}^{m}\sum_{l=1}^{m}(x_{ik}-x_{jk})(x_{il}-x_{jl})r_{kl}\right]^{1/2}\qquad(i,j=1\cdots,n)$$

r_{kl}为变量X_k和X_l之间的相关系数.

5.相似系数

参看小节三”变量间的相似系数和距离”中的定义.

6.定性变量样品间的距离或相似系数

Definition

- 项目：定性变量, 如性别为项目.
- 类目：定性变量的各种不同的取”值”, 如”男”和”女”是性别这个项目的类目.

设共有m个项目，第k个项目有r_k 个类目. 现共有n 个样本. 第i个样本的取值为

$$(\delta_i(k,1),\delta_i(k,2),\cdots,\delta_i(k,r_k)),\ \ i=1,2,\cdots,n,\ \ k=1,2,\cdots,m$$

其中

$$\delta_i(k,l)=\begin{cases}1, & \text{当第}i\text{个样品中第}k\text{个项目的定性数据为第}l\text{个类目时,}\\ 0, & \text{否则,}\end{cases}$$

称$\delta_i(k,l)$为第k 项目之l 类目在第i 个样品中的反应.

- 若$\delta_i(k,l)=\delta_j(k,l)=1$, 称第$i$个样本和第$j$个样本1-1配对;
- 若$\delta_i(k,l)=\delta_j(k,l)=0$, 称第$i$个样本和第$j$个样本0-0配对;
- 若$\delta_i(k,l)\neq\delta_j(k,l)$, 称第$i$个样本和第$j$个样本不配对.

记：

- m_0 为所有类目中0-0配对总数;
- m_1 为所有类目中1-1配对总数;
- m_2 为所有类目中不配对总数;

显然有$m_0 + m_1 + m_2 = p$总类目数，即$r_1 + r_2 + \cdots + r_m = p$.

类似于欧氏距离，可以定义

$$d_{ij}^2 = \sum_{k=1}^{m} \sum_{l=1}^{r_k} (\delta_i(k,l) - \delta_j(k,l))^2.$$

现有三个项目：性别，职称，年收入

序号	性别		职称				收入	
	男	女	初级	中级	副高	正高	$\leq$ 5万	>5 万
1	1	0	1	0	0	0	1	0
2	0	1	0	0	0	1	0	1

$$m_0 = 2,\ m_1 = 0,\ m_2 = 6.$$

$$d_{12}^2 = \sum_{k=1}^{m}\sum_{l=1}^{r_k}(\delta_i(k,l) - \delta_j(k,l))^2 = 6.$$

即为不配对的数。

样本间相似性的几种度量:

1. $(m_0 + m_1)/p$ 配对的总数在总类目中之比;
2. m_1/p 1-1配对的总数在总类目数中之比;
3. $m_1/(m_1 + m_2)$ 不考虑0-0 配对的情况;
4. $2(m_0 + m_1)/(p + m_0 + m_1)$ 对配对的数双倍加权;
5. $(m_0 + m_1)/(p + m_2)$
6. $2m_1/(2m_1 + m_2)$ 1-1配对数双倍加权, 不考虑0-0配对;
7. $m_1/(m_1 + 2m_2)$ 不考虑0-0配对，且对不配对数双倍加权;
8. m_1/m_2

三、变量间的距离与相似系数

通常采用变量间的距离与相似系数表示变量间的亲疏程度来对变量进行分类. 设C_{ij}表示变量X_i和X_j之间的相似系数,一般要求:

1. $C_{ij} = \pm 1 \leftrightarrow X_i = aX_j (a \neq 0)$
2. $|C_{ij}| \leq 1$,对一切i, j
3. $C_{ij} = C_{ji}$,对一切i, j

对于定量变量,通常采用的相似系数有夹角余弦和相关系数.

在R软件中没有现成的包用于夹角余弦的计算，但可以采用scale()，和sweep() 函数计算。

设变量X_i的n次观测值为$(x_{1i},\cdots,x_{ni})$

1.夹角余弦

$$C_{ij}(1)=\cos\alpha_{ij}=\frac{\sum_{t=1}^{n}x_{ti}x_{tj}}{\sqrt{\sum_{t=1}^{n}x_{ti}^2}\sqrt{\sum_{t=1}^{n}x_{tj}^2}}$$

2.相关系数

相关系数就是对数据标准化后的夹角余弦.

$$C_{ij}(2)=\frac{\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)(x_{tj}-\bar{x}_j)}{\sqrt{\sum_{t=1}^{n}(x_{ti}-\bar{x}_i)^2}\sqrt{\sum_{t=1}^{n}(x_{tj}-\bar{x}_j)^2}}$$

3.变量间的距离

1. 利用相似系数定义变量间的距离

$$d_{ij} = 1 - |C_{ij}| \text{或} d_{ij}^2 = 1 - C_{ij}^2$$

2. 利用样本协方差阵S来定义距离.

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

3. 利用小节二”样品间的距离和相似系数”中介绍的方法类似定义变量间的距离.

4.定性变量间的相似系数

列联表

变量	t_1	t_2	$\cdots$	t_q	求行和
r_1	n_{11}	n_{12}	$\cdots$	n_{1q}	$n_{1+}=\sum_l n_{1l}$
r_2	n_{21}	n_{22}	$\cdots$	n_{2q}	$n_{2+}=\sum_l n_{2l}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r_p	n_{p1}	n_{p2}	$\cdots$	n_{pq}	$n_{p+}=\sum_l n_{pl}$
求列和	n_{+1}	n_{+2}	$\cdots$	n_{+q}	$n_{++}=\sum_i\sum_j n_{ij}$

利用列联表对两个属性变量的独立性检验中，经常要用到χ^2 统计量

$$\chi_{ij}^2 = n_{++}\left(\sum_{k=1}^{p}\sum_{l=1}^{q}\frac{n_{kl}^2}{n_{k+}n_{+l}} - 1\right),$$

建立在χ^2 统计量基础上的相似系数有：

- 联列系数
- 连关系数(有三种)
- 点相关系数
- 四分相关系数
- 夹角余弦

联列系数

$$C_{ij}(3)=\sqrt{\chi_{ij}^2/\left(\chi_{ij}^2+n\right)}$$

连关系数

$$\begin{aligned}
C_{ij}(4) &= \sqrt{\frac{\chi_{ij}^2}{n\cdot\max(p-1,q-1)}};\\
C_{ij}(5) &= \sqrt{\frac{\chi_{ij}^2}{n\cdot\min(p-1,q-1)}};\\
C_{ij}(6) &= \sqrt{\frac{\chi_{ij}^2}{n\cdot(p-1,q-1)}}.
\end{aligned}$$

点相关系数，四分相关系数，夹角余弦见书p227.

附录：关于多维标度问题:

- 多维标度法是一种将多维空间的研究对象(样本或变量)简化到低维空间进行定位、分析和归类，同时又保留对象间原始关系的数据分析方法。
- 多维标度法是一类多元统计分析方法的总称，包含各种各样的模型和手段，其目的是通过各种途径把高维的研究对象转化成低维情形进行研究.

- 多维标度法是以研究对象之间某种亲近关系为依据(如距离、相似系数，亲疏程度的分类情况等)，
- 合理地将研究对象(样品或变量)在低维空间中给出标度或位置，
- 全面而又直观地再现原始各研究对象之间的关系，同时在此基础上也可按对象点之间距离的远近实现对样品的分类
- 多维标度法能弥补聚类分析的不足之处，因为聚类分析将相似的样品归类，最后得到一个反映样品亲疏关系的谱系图。

Example

为了分析亚洲国家和地区的经济发展、文教、卫生水平。现采用48个国家和地区的如下的指标：

- 城市人口的比例
- 女（男）性的平均寿命
- 会阅读的人的比例（女（男）性）
- 人均GDP
- 死亡率

并收集相应的数据资料。可以通过多维标度法，将每个国家（地区）在直角座标系上直观呈现国家（地区）之间的关系。

具体的做法是：

- 第一步：利用数据给出相似矩阵；
- 第二步：给出一个欧氏距离矩阵，使这个距离矩阵与原来的相似矩阵在某种意义下非常接近；
- 第三步：将每个国家（地区）在图中标出，分析他们之间的亲疏关系。

聚类分析与多维标度法之间的关系：

- 聚类分析是分析样品之间亲疏关系简便易行的方法。聚类分析的缺点是将一些高维的样品强行纳入一个一维的谱系分类中，常常使原始样品之间的关系简单化，甚至有时失真。
- 多维标度法是将几个高维研究对象，在近似的意义下，从高维约简到一个较低维的空间内，并且寻求一个最佳的空间维数和空间位置如2维或3维), 较好保持各研究对象数据的原始关系。

- stats包的cmdscale()函数执行传统的多维尺度分析（multidimensional scaling，MDS）（主坐标分析Principal Coordinates Analysis）.
- MASS包的sammon()和isoMDS()函数分别执行Sammon和Kruskal非度量多维尺度分析.
- vegan包提供非度量多维尺度分析的包装(wrappers)和后处理程序.

6.3 系统聚类法

一、系统聚类法的基本思想和基本步骤

系统聚类法的基本思想:

- 首先定义样品间距离和类与类间的距离.
- 初始将n个样品看成n个类,按最小距离准则并类,
- 每次缩小一类重新计算类间距离,直至所有样品并称一类.

系统聚类法基本步骤:

1. 数据变换(如标准化变换等),定义样品间距离(如欧式距离),定义类间的距离(如最短距离法等).
2. 计算n个样品两两之间的距离,得到样品间的距离矩阵$D^{(0)}$.
3. 第一步将每个样品自构成一类,此时类间距离等于样品间的距离(即$D^{(1)} = D^{(0)}$).
4. 按某种距离原则（如：最小距离原则，详细见类间的距离计算见后面）把类间距离最小的两类并成一新类,此时类的总数少了一个,计算新类与其他类的距离,得到类间距离$D^{(2)}$.
5. 重复步骤4,可得到$D^{(3)}, D^{(4)} \cdots$,直到类的总数为1.
6. 画谱系聚类图.
7. 决定分类个数及各类的成员.

例6.3.1

设有5个样品,分别对每个产品测得一项指标X,其值如下:1,2,4.5,6,8.对这5个产品按质量指标进行分类.

解 计算$D^{(1)}$(采用**欧氏距离** 计算)

	$X_{(1)}$	$X_{(2)}$	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$
$X_{(1)}$	0	1	3.5	5.0	7.0
$X_{(2)}$		0	2.5	4.0	6.0
$X_{(3)}$			0.0	1.5	3.5
$X_{(4)}$				0.0	2.0
$X_{(5)}$					0.0

所以可知先合并$X_{(1)}$和$X_{(2)}$为一个新类.

计算$D^{(2)},D^{(3)},D^{(4)}$(类间**距离按最短距离** 法计算)

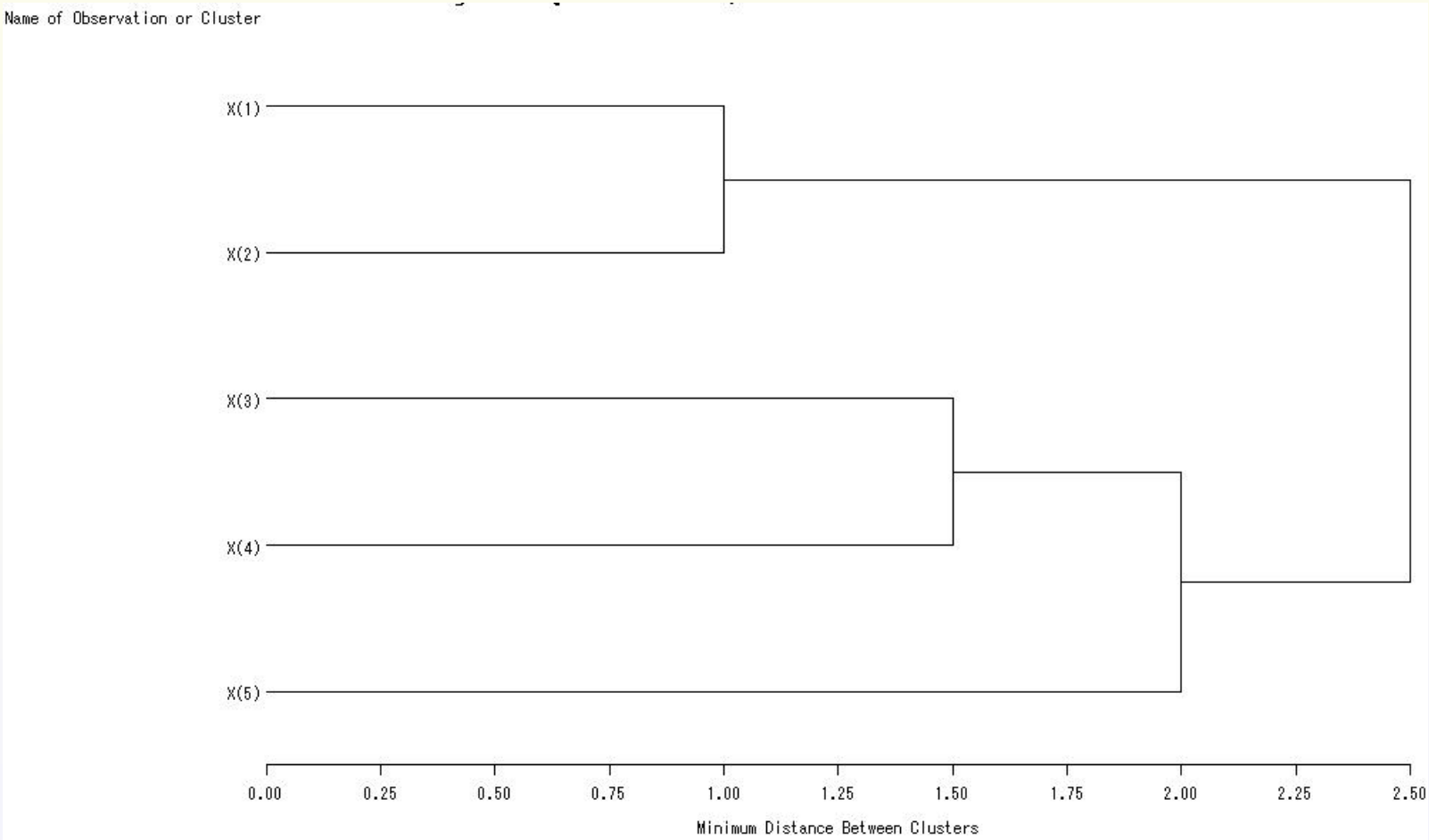
	$X_{(3)}$	$X_{(4)}$	$X_{(5)}$	$CL4$
$X_{(3)}$	0	1.5	3.5	2.5
$X_{(4)}$		0.0	2.0	4.0
$X_{(5)}$			0.0	6.0
$CL4$				0.0

	$X_{(5)}$	$CL4$	$CL3$
$X_{(5)}$	0	6	2.0
$CL4$		0	2.5
$CL3$			0.0

	$CL4$	$CL2$
$CL4$	0	2.5
$CL2$		0.0

	$CL1$
$CL1$	0

谱系聚类图:


Name of Observation or Cluster
X(1)
X(2)
X(3)
X(4)
X(5)
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
Minimum Distance Between Clusters

确定分类个数及各类成员:

- 若分为两类,则$G_1^{(2)} = \{X_{(1)}, X_{(2)}\}$,$G_2^{(2)} = \{X_{(5)}, X_{(4)}, X_{(3)}\}$
- 若分为三类,则$G_1^{(3)} = \{X_{(1)}, X_{(2)}\}$,$G_2^{(3)} = \{X_{(5)}\}$,
  $G_3^{(3)} = \{X_{(4)}, X_{(3)}\}$
- 若分为四类,则$G_1^{(4)} = \{X_{(1)}, X_{(2)}\}$,$G_2^{(4)} = \{X_{(5)}\}$,
  $G_3^{(4)} = \{X_{(3)}\}$, $G_4^{(4)} = \{X_{(4)}\}$

二、系统聚类分析的方法

从上面的例子可以看出，除d_{ij}表示样品$X_{(i)}$和$X_{(j)}$之间的距离，我们还需要定义类与类之间的距离。下面是几种定义类与类之间距离的方法。

1. 最短距离法(SINgle linkage) 类与类之间的距离定义为两类中相距最近的样品之间的距离,即类G_p和G_q之间的距离.

$$D_{pq} = \min_{i \in G_p, j \in G_q} d_{ij}$$

当类G_p和G_q合并成G_r后和其他类G_k之间的距离的递推公式:

$$D_{rk} = \min\{D_{pk}, D_{qk}\}$$

2. 最长距离法(COMplete method) 类与类之间的距离定义为两类中相距最远的样品之间的距离,即类G_p和G_q之间的距离.

$$D_{pq} = \max_{i \in G_p, j \in G_q} d_{ij}$$

当类G_p和G_q合并成G_r后和其他类G_k之间的距离的递推公式:

$$D_{rk} = \max\{D_{pk}, D_{qk}\}$$

3. 中间距离法(MEDian method)

这种方法介于最短距离法和最长距离法之间.其递推公式为:

$$D_{rk}^2 = \frac{1}{2}(D_{pk}^2 + D_{qk}^2) + \beta D_{pq}^2 \quad (-1/4 \leq \beta \leq 0)$$

可以知道，当$\beta = 1/4$时，D_{rk} 就是以D_{qk}, D_{pk}, D_{pq}为边的三角形的中D_{pq}边上的中线.

4. 重心法(CENtroid method)

将两类间的距离定义为两类重心间的距离，这种聚类方法称为重心法. 每一类的重心就是属于该类样本的均值.

设某一步骤将G_p和G_q合并成G_r后,它们所包含的样品个数分别为n_p, n_q和n_r,各类的重心分别为$\bar{X}^{(p)}, \bar{X}^{(q)}$和$\bar{X}^{(r)}$显然有:

$$\bar{X}^{(r)} = \frac{1}{n_r}(n_p\bar{X}^{(p)} + n_q\bar{X}^{(q)})$$

它与新类G_r的距离是

$$D_{rk} = d(\overline{X}^{(r)}, \overline{X}^{(k)}).$$

如果样品间的距离为欧氏距离，当类G_p和G_q合并成G_r后和其他类G_k之间的距离的递推公式:

$$D_{rk}^2 = \frac{n_p}{n_r}D_{pk}^2 + \frac{n_q}{n_r}D_{qk}^2 - \frac{n_p}{n_r}\frac{n_q}{n_r}D_{pq}^2$$

5. 类平均法(AVErage linkage)

用两类样品两两之间的平方距离平均作为类间距离.

$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{i \in G_p, j \in G_q} d_{ij}^2$$

递推公式:

$$D_{rk}^2 = \frac{n_p}{n_r} D_{pk}^2 + \frac{n_q}{n_r} D_{qk}^2$$

类平均法是一种使用比较广泛、聚类效果较好的方法.

6. 可变类平均法(FLExible-beta method)

类平均法的递推公式中没有考虑G_p和G_q之间的距离D_{pq}的影响,可变类平均法的递推公式为

$$D_{rk}^2 = (1-\beta)\left[\frac{n_p}{n_r}D_{pk}^2 + \frac{n_q}{n_r}D_{qk}^2\right] + \beta D_{pq}^2$$

- 在实用中β常取负值,如$\beta = -1/4$. 一般$\beta < 1$.
- $\beta = 0$,类平均法。
- $n_p = n_q$, 可变法。
- 可变类平均法是由类平均法和中间距离法适当推广得到的.

7. 可变法及McQuitty相似分析法(MCQ)

可变类平均法中$n_p = n_q$的形式.

$$D_{rk}^2 = \frac{(1-\beta)}{2}[D_{pk}^2 + D_{qk}^2] + \beta D_{pq}^2$$

当$\beta = 0$时，

$$D_{rk}^2 = [D_{pk}^2 + D_{qk}^2]/2,$$

称为McQuitty相似分析法。

8. 离差平方和法(WARD)

离差平方和法(WARD)是Ward(1936年)提出的，也称为Ward法.

其基本思想:

- 先将n个样品各自归一类,此时总离差平方和$W = 0$,
- 每次并类的原则为使W增加最小的两类并类,
- 直至并为一类.

假定已将n个样品分为k类，记为$G_1, G_2, \cdots, G_k$, n_t 表示G_t 类的样品个数，$\overline{X}^{(t)}$ 表示G_t 的重心，$X(i)^{(t)}$ 表示G_t中第i 个样品$(i = 1, 2, \cdots, n)$, G_t中样品的离差平方和为

$$W_t = \sum_{i=1}^{n_t} (X_{(i)}^{(t)} - \bar{X}^{(t)})'(X_{(i)}^{(t)} - \bar{X}^{(t)})$$

k个类总离差平方和

$$W = \sum_{t=1}^{k} W_t$$

WARD法把某两类合并后增加的离差平方和看成类间的平方距离,即

$$D_{pq}^2 = W_r - (W_p + W_q)$$

表示类G_p和G_q之间的平方距离.

利用W_r的定义,

$$\begin{aligned} W_r &= \sum_{t=1}^{n_r} (X_{(t)}^{(r)} - \bar{X}^{(r)})'(X_{(t)}^{(r)} - \bar{X}^{(r)}) \\ &= \sum_{i=1}^{n_p} (X_{(i)}^{(p)} - \bar{X}^{(r)})'(X_{(i)}^{(p)} - \bar{X}^{(r)}) \\ &+ \sum_{i=1}^{n_q} (X_{(i)}^{(q)} - \bar{X}^{(r)})'(X_{(i)}^{(q)} - \bar{X}^{(r)}) \end{aligned}$$

经整理可得

$$D_{pq}^2 = \frac{n_p n_q}{n_r}(\bar{X}^{(p)} - \bar{X}^{(q)})'(\bar{X}^{(p)} - \bar{X}^{(q)})$$

若采用欧氏距离,上式改为:

$$D_{pq}^2 = \frac{n_p n_q}{n_r} d_{pq}^2$$

则有递推公式:

$$D_{rk}^2 = \frac{n_p + n_k}{n_r + n_k} D_{pk}^2 + \frac{n_q + n_k}{n_r + n_k} D_{qk}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2$$

离差平方和法应用广泛,效果较好,但要求样品间距离为欧式距离.

还有其它的一些系统聚类法：

- 最大似然谱系聚类法(EML)
- 密度估计法(DEN)
- 两阶段密度估计法(TWO)

三、系统聚类方法的统一

当类G_p和G_q合并成G_r后和其他类G_k之间的距离的递推公式:

$$D_{rk}^2 = \alpha_p D_{pk}^2 + \alpha_q D_{qk}^2 + \beta D_{pq}^2 + \gamma |D_{pk}^2 - D_{qk}^2|$$

其中α_p, α_q, β 和γ 是参数，不同的系统聚类方法有不同的取值。见书本P237表6.6.

CRAN的Cluster任务列表全面的综述了R实现的聚类方法。

- stats里提供等级聚类hclust()和k-均值聚类kmeans().
- cluster包里有大量的聚类和可视化技术
- clv包里则有一些聚类确认程序.
- e1071包的classAgreement()可计算Rand index比较两种分类结果.
- Trimmed k-means聚类分析可由trimcluster包实现.

- 聚类融合方法（Cluster Ensembles）由clue包实现，clusterSim包能帮助选择最佳的聚类
- hybridHclust包提供一些混合聚类方法.
- energy包里有基于E统计量的距离测度函数edist()和等级聚类方法hclust.energy().
- LLAhclust包提供基于似然（likelihood linkage）方法的聚类，也有评定聚类结果的指标.
- fpc包里有基于Mahalanobis距离的聚类.

- clustvarsel包有多种基于模型的聚类.
- 模糊聚类（fuzzy clustering）可在cluster包和hopach包里实现。
- Kohonen包提供用于高维谱（spectra）或模式（pattern）的有监督和无监督的SOM算法.
- clusterGeneration包帮助模拟聚类.
- RAN的Environmetrics任务列表里也有相关的聚类算法的综述.
- mclust包实现了基于模型的聚类，MFDA包实现了功能数据的基于模型的聚类.

R软件中系统聚类法

hclust()函数提供了系统聚类的计算，格式为

hclust(d, method="?", members=NULL) 其中:

- d 是由"dist" 构成的距离,
- members gives the number of observations per cluster.
- method是系统聚类的方法（缺省是最长距离法）

method 参数为

- single 最短距离法
- complete 最长距离法
- median 中间距离法
- mcquitty Mcquitty相似法
- average 类平均法
- centroid 重心法
- ward 离差平方和法

plot()函数可画出系统聚类的树形图(dendrogram谱系图), 格式为

plot(x, hang=0.1, xlab=NULL, ylab="height")

- x 是由hclust()函数生成的对象，
- hang=1, 表示标志挂线的长度，如果hang=-1，则一定是从0开始.
- xlab, ylab 分别为横轴和纵轴的标志.

Example

设有5个产品，分别对每个产品测得一项质量指标X，其值如下：1, 2, 4.5, 6, 8. 试对这5个产品按质量指标进行分类.

(1) 输入数据，生成距离结构

$x <- c(1, 2, 4.5, 6, 8);$

$d <- dist(x).$

(2)生成系统聚类

hc1 <- hclust(d, "single")

hc2 <- hclust(d, "complete")

hc3 <- hclust(d, "median")

hc4 <- hclust(d, "average")

绘出所有树形结构图，并以2×2 的形式绘在一张图上.

$Opar < -par(mfrow = c(2, 2))$

其中: par这个命令是用来对图形进行设置或排队. mfcol, mfrow 是将图形设置成行排或列排.

plot(hc1, hang=-1)

plot(hc2, hang=-1)

plot(hc3, hang=-1)

plot(hc4, hang=-1)

6.4 系统聚类法的性质及类的确定

一、系统聚类法的简单性质

系统聚类法有两个简单性质:

(1)单调性:设D_k表示系统聚类法中第k次并类时的距离.一个系统聚类法若能保证$\{D_k\}$单调上升,则称它具有**单调性**.

可以证明,最短距离法、最长距离法、类平均法、可变类平均法,以及离差平方和法都具有单调性,只有重心法和中间距离法不具有单调性.

(2)空间的浓缩与扩张:比较最短距离法和最长距离法的并类过程,可以看出:

$$D_{ij}(\text{短}) \leq D_{ij}(\text{长})$$

这种性质称**最长距离法比最短距离法扩张**,或称**最短距离法比最长距离法浓缩**.

对于系统聚类的各种方法,有如下结论:

1. 类平均法比最短距离法扩张,比最长距离法浓缩
2. 类平均法比重心法扩张,比离差平方和法浓缩
3. 太浓缩的方法不太灵敏,太扩张的方法容易失真
4. 类平均法比较适中,有具有单调性,应用较广泛

二、类的定义及特征

1.类的几种定义

Definition

设阈值T是给定的正数,若集合G中任意两个元素的距离d_{ij}都满足:

$$d_{ij} \leq T \quad (i, j \in G)$$

则称G对于阈值T组成一个类.

Definition

阈值T是给定的正数,若集合G中每个$i \in G$都满足:

$$\frac{1}{n-1}\sum_{j\in G} d_{ij} \leq T$$

其中n是集合G中的元素的个数,则称G对于阈值T组成一个类

Definition

设T和$H(H>T)$是两个给定的正数,若集合G中两两元素的距离的平均满足:

$$\frac{1}{n(n-1)}\sum_{i\in G}\sum_{j\in G}d_{ij}\leq T,\quad d_{ij}\leq H\quad (i,j\in G)$$

Definition

设T是给定的正数,若对集合G中任一个$i \in G$,一定存在一个$j \in G$,使得这两个元素的距离d_{ij}满足:

$$d_{ij} \leq T \quad (i, j \in G)$$

则称G对于阈值T组成一个类.

Definition

设阈值T是给定的正数,将集合G任意分为两类:G_1和G_2,这两类之间的距离满足:

$$D(G_1, G_2) \leq T$$

则称G对于阈值T组成一个类.

2.类的特征

设类G包含的样品记为$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$,其中$X_{(t)}, t = 1, \cdots, n$为$m$ 元总体的样本.可以从不同角度来刻画G 的特征,常用的特征有以下三种:

(1)均值(或称重心):$\bar{X}_G = \frac{1}{n}\sum_{t=1}^{n} X_{(t)}$,

(2)样本离差阵A_G及样本协方差阵S_G:

$$A_G = \sum_{t=1}^{n}(X_{(t)} - \bar{X}_G)(X_{(t)} - \bar{X}_G)', \quad S_G = \frac{1}{n-1}A_G,$$

(3)类的直径D_G:常用的直径有

$$D_G = \sum_{t=1}^{n}(X_{(t)} - \bar{X}_G)'(X_{(t)} - \bar{X}_G) = \mathrm{tr}(A_G),$$

和$D_G = \max_{i,j\in G} d_{ij}$.

三、类个数的确定

1.由适当的阈值确定

选定某种聚类方法，按系统聚类的步骤并类后，得到一张谱系聚类图，规定一个临界相似性尺度，用以分割谱系图而得到样品（或变量）的分类。

2.根据数据点的散布图直观地确定类的个数

- 如果考察的指标只有两个，则可以通过数据点的散布图直观地确定类的个数.
- 如果有三个变量，可以绘制三维散布图并通过旋转三维坐标轴由数据点的分布来确定应分几个类.
- 当考察的指标在三个以上时，可以由这些指标综合出两个或三个综合变量（主成份分析法）后再绘制数据点在综合变量上的散布图，从而直观地确定分类个数。

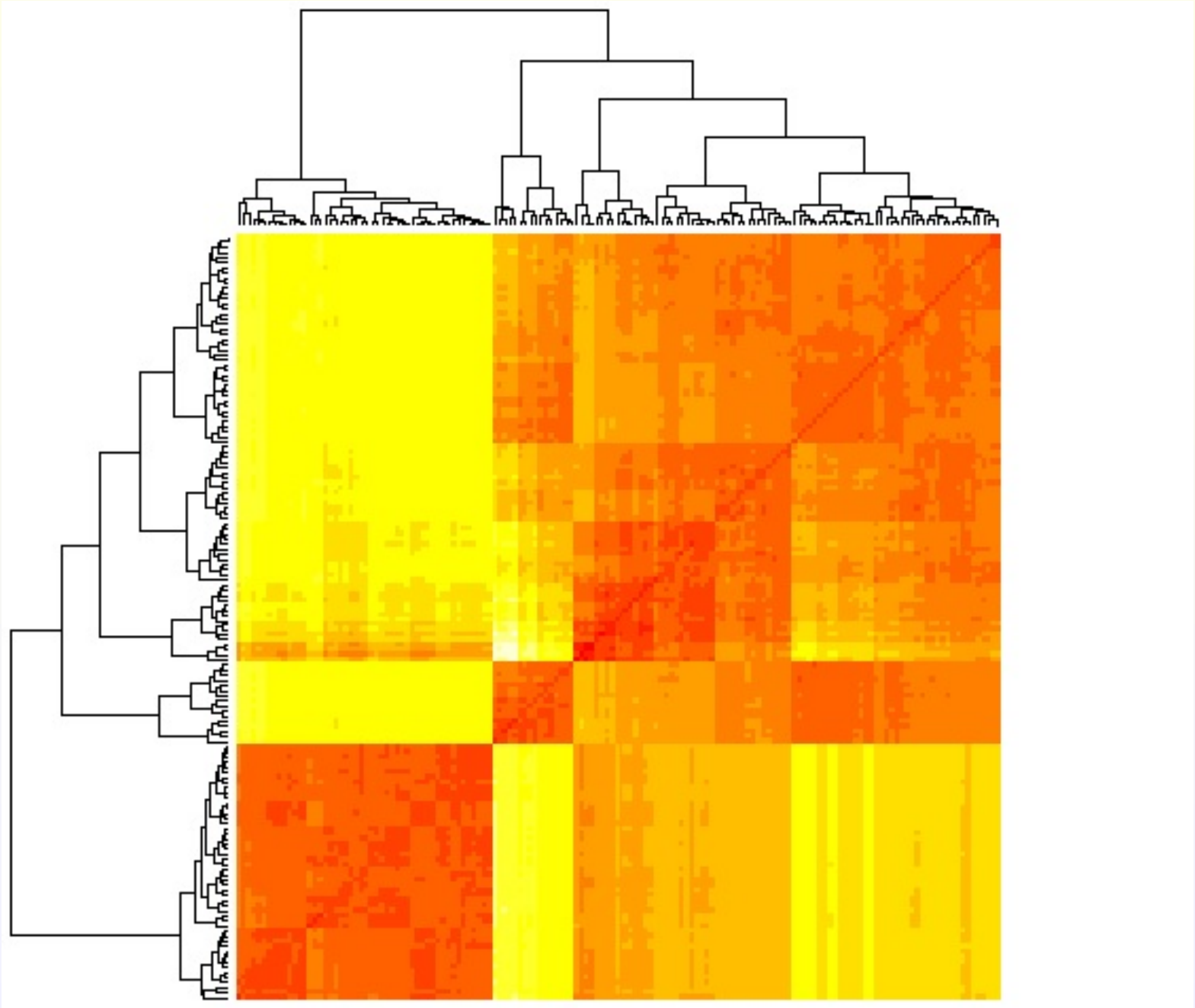
采用矩阵热图来发现类：

$data <- iris[,-5]$

$dist.e <- dist(data, method = "euclidean")$

$heatmap(as.matrix(dist.e), labRow = F, labCol = F)$

利用矩阵绘制热图，从图中可以看到颜色越深表示样本间距离越近。



从图中确定聚类的类的个数，然后进行聚类分析

$model1 = hclust(dist.e, method = "ward.D")$

$re < -cutree(model1, k = 3)$

采用多维标定，用第一和第二主成份，表示原本分类

$mds = cmdscale(dist.e, k = 2, eig = T)$

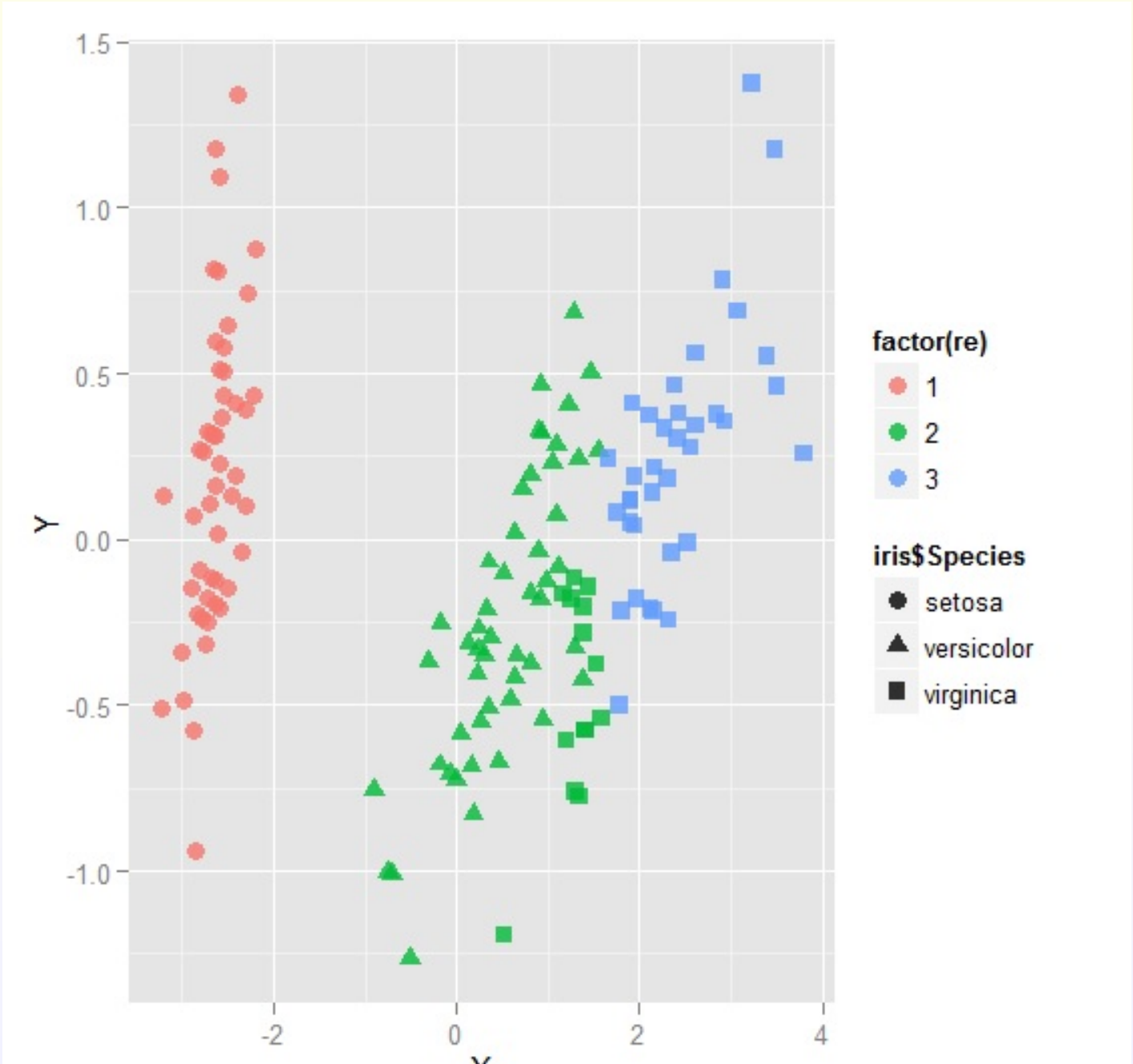
$X < -mds\$points[, 1]$

$Y < -mds\$points[, 2]$

加载ggplot2模块，用图象来表示原本分类和现在聚类分类后的差异

$p = ggplot(data.frame(X, Y), aes(X, Y))$

$p + geom_point(size = 3, alpha = 0.8, aes(colour = factor(re), shape = iris\$Species))$


1.5
1.0
0.5
0.0
-0.5
-1.0
Y
-2
0
2
4
X
factor(re)
1
2
3
iris$Species
setosa
versicolor
virginica

3.根据统计量确定分类个数

- R^2统计量:假定已将n个样品分成k类,记为$G_1,\cdots,G_k$, n_t表示G_t类的样品个数, $\bar{X}^{(t)}$表示G_t的重心.$\bar{X}_{(i)}^{(t)}$表示G_t中第i个样品$\bar{X}$表示所有样品的重心.

  则G_t类中n_t个样品的离差平方和(组内)为

$$W_t=\sum_{i=1}^{n_t}(\bar{X}_{(i)}^{(t)}-\bar{X}^{(t)})'(\bar{X}_{(i)}^{(t)}-\bar{X}^{(t)})$$

  所有样品的总离差平方和为

$$T=\sum_{t=1}^{k}\sum_{i=1}^{n_t}(\bar{X}_{(i)}^{(t)}-\bar{X})'(\bar{X}_{(i)}^{(t)}-\bar{X})$$

T可分解为:

$$\begin{aligned} T &= \sum_{t=1}^{k}\sum_{i=1}^{n_t}(\bar{X}_{(i)}^{(t)}-\bar{X})'(\bar{X}_{(i)}^{(t)}-\bar{X}) = (\cdots) \\ &= \sum_{t=1}^{k} W_t + \sum_{t=1}^{k} n_t(\bar{X}^{(t)}-\bar{X})'(\bar{X}^{(t)}-\bar{X}) \\ &= P_k(\text{组内}) + B_k(\text{组间}) \end{aligned} \tag{1}$$

令

$$R_k^2 = \frac{B_k}{T} = 1 - \frac{P_k}{T}$$

R_k^2越大,即$\frac{B_k}{T}$越大,表示类间偏差平方和在总离差平方和占的比例越大,说明k各类能够很好的分开. 但R_k^2随着k的减少而变小,所以应该看R_k^2的变化来确定分类个数.

(2)半偏R_k^2统计量

$$半偏R_k^2 = B_{KL}^2/T = R_{k+1}^2 - R_k^2$$

其中$B_{KL}^2 = W_M - (W_K + W_L)$,

1. 半偏R_k^2统计量表示合并类G_K和G_L为新类G_M后类内离差平方和的增值.
2. 该统计量用于评价合并G_K 和G_L 类后的效果.
3. 根据公式，根据R_k^2的变化就可得到半偏R_k^2的值.
4. 用于$k+1$个类合并为k个类的效果，如果半偏R_k^2统计量的值大，说明$k+1$个类的效果好.

(3)伪F统计量

$$伪F_k = \frac{(T - P_k)/(k-1)}{P_k/(n-k)} = \frac{B_k}{P_k}\frac{n-k}{k-1}$$

1. 该统计量用于评价分为k个类的聚类效果,伪F_k值越大表示这n个样品可显著的分为k类.
2. 伪F_k 统计量可以作为确定类个数的有用指标, 但并不具有像F 统计量的分布.

(4)伪t^2统计量:

$$伪t^2 = \frac{B_{KL}^2}{(W_K + W_L)/(n_k + n_L) - 2}$$

1. 伪t^2统计量评价此步骤合并类G_K和G_L的效果, 即用于$k+1$个类合并为k个类的效果.伪t^2统计量值大,说明$k+1$个类的效果好.
2. 仅为分类指标，但并不具有t^2 统计量的分布.

4.根据谱系图确定分类个数的准则

Bemirmen(1972) 提出了应根据研究的目的来确定适当的分类方法, 并提出了一些根据谱系图来分析的准则：

准则A 各类重心之间的距离必须很大

准则B 确定的类中,各类所包含的元素不要很多

准则C 类的个数必须符合实用目的

准则D 若采用几种不同的聚类方法处理,则在各自的聚类图中应发现相同的类

例6.4.1

表6.7是我国16个地区农民在1982年支出情况的抽样调查数据的汇总,每个地区都调查了反映每人平均生活消费支出情况的六个指标.试对16个地区进行分类.

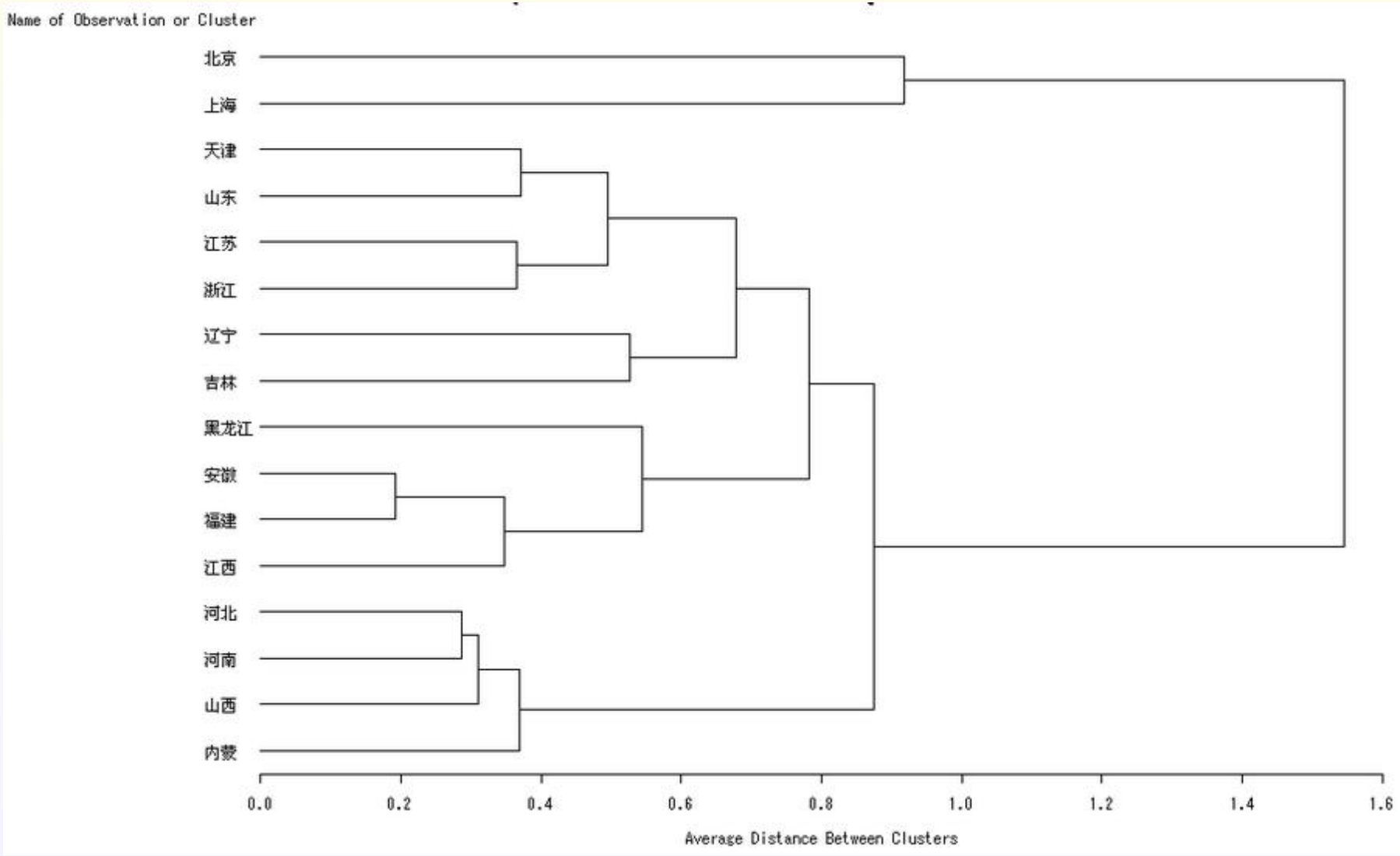
表 6.7　16 个地区农民生活水平的调查数据（单位：元）

地 区	食 品 (X1)	衣 着 (X2)	燃 料 (X3)	住 房 (X4)	生活用品及其他(X5)	文化生服务支出(X6)
北 京	190.33	43.77	9.73	60.54	49.01	9.04
天 津	135.20	36.40	10.47	44.16	36.49	3.94
河 北	95.21	22.83	9.30	22.44	22.81	2.80
山 西	104.78	25.11	6.40	9.89	18.17	3.25
内 蒙	128.41	27.63	8.94	12.58	23.99	3.27
辽 宁	145.68	32.83	17.79	27.29	39.09	3.47
吉 林	159.37	33.38	18.37	11.81	25.29	5.22
黑龙江	116.22	29.57	13.24	13.76	21.75	6.04
上 海	221.11	38.64	12.53	115.65	50.82	5.89
江 苏	144.98	29.12	11.67	42.60	27.30	5.74
浙 江	169.92	32.75	12.72	47.12	34.35	5.00
安 徽	153.11	23.09	15.62	23.54	18.18	6.39
福 建	144.92	21.26	16.96	19.52	21.75	6.73
江 西	140.54	21.50	17.64	19.19	15.97	4.94
山 东	115.84	30.26	12.20	33.61	33.77	3.85
河 南	101.18	23.26	8.46	20.20	20.50	4.30

类	2类	3类	4类	5类	6类	7类
R2统计量	0.593	0.745	0.870	0.911	0.935	0.951
伪F统计量	20.413	41.000	93.938	143.789	201.855	272.875

类的变化	$2 \to 1$	$3 \to 2$	$4 \to 3$	$5 \to 4$	$6 \to 5$	$7 \to 6$
偏R2统计量	0.593	0.153	0.125	0.040	0.024	0.016
伪t2统计量	20.413	14.739	7.775	7.461	4.473	–

综合认为采用类平均法分为5类合适.


Name of Observation or Cluster
北京
上海
天津
山东
江苏
浙江
辽宁
吉林
黑龙江
安徽
福建
江西
河北
河南
山西
内蒙
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
Average Distance Between Clusters

例6.41用系统聚类的方法进行聚类（样本聚类）

(1)输入数据

$D641 <- read.table("ex641.txt", header = F),$

$D641A <- D641[, 2:7].$

(2)计算距离

$d <- dist(D641A, "minkowski", p = 2)$

(3)聚类分析

$hc1 <- hclust(d, "average")$

$hc2 <- hclust(d, "complete")$

$hc3 <- hclust(d, "ward")$

$hc4 <- hclust(d, "centroid")$

(4)画谱系图

$$opar < -par(mfrow = c(2,1), mar = c(5.2,4,0,0))$$

$$plclust(hc1, hang = -1);$$

$$re1 < -rect.hclust(hc1, k = 5, border = "red")$$

$$plclust(hc2, hang = -1);$$

$$re2 < -rect.hclust(hc2, k = 5, border = "red")$$

$$par(opar)$$

rect.hclust()与确定类的个数聚类分析函数，它本质是由给定类的个数或给定阀值来确定聚类的情况格式为：

rect.hclust(tree, k, h, border=”red”)

其中：

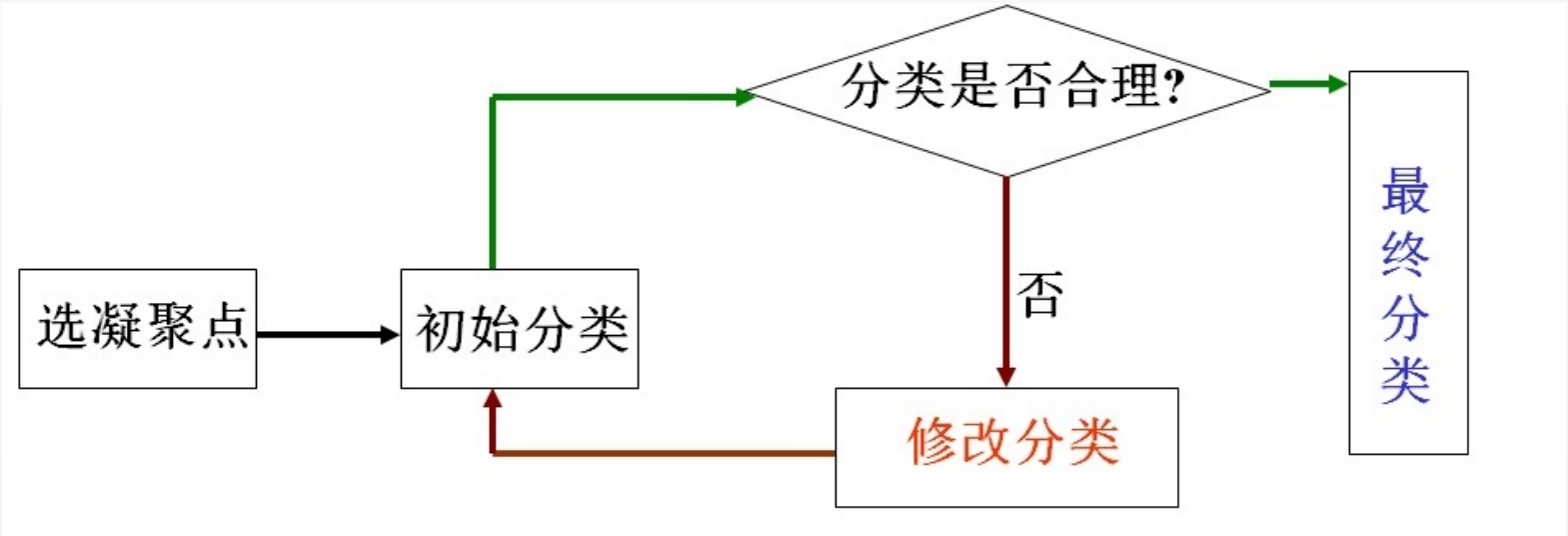
- tree是由hclust 生成的对象；
- k是类的个数；
- h 是谱系图中的阈值，要求分布的各类的距离大于h；
- border 是数或向量，标明矩形框的颜色

6.5 动态聚类法

系统聚类法的缺点:当样本较大时,计算量太大.

动态聚类法(又称逐步聚类法)的基本思想是,开始先粗略分类,然后按照某种最优原则修改不合理的分类,直至分类合理为止.用于大样本的样品间聚类.

动态聚类法的聚类过程可用下面框图描述



动态聚类法的优点:计算量小,方法简单,适用于大样本的Q型聚类分析.

动态聚类法的缺点:样品的最终聚类在某种程度上依赖于最初的划分，或种子点的选择。

改进方法：为了检验聚类的稳定性，可用一个新的初始分类重新检验整个聚类算法。如最终分类与原来一样，则不必再行计算；否则，须另行考虑聚类算法。

一、选择凝聚点和确定初始分类

凝聚点就是一批有代表性的点，是欲形成类的中心的点。凝聚点的选择直接决定初始分类，对分类结果也有很大的影响。

选择凝聚点的方法:

- **人为选择:**

1. 根据经验，预先确定分类个数和初始分类，然后从每一类中选择有代表性的样品作为凝聚点.
2. 将所有样品人为地分为k类，计算每一类的重心，并将这些重心作为凝聚点.
3. 人为选择一正数d,首先把所有样品的均值作为凝聚点,然后依次考察每个样品,若某样品与已选定的凝聚点的距离均大于d,该样品作为新凝聚点,否则考察下一个样品.

- 密度法选择凝聚点: 以某个正数d为半径，以每个样品为球心，落在这个球内的样品数(不包括作为球心的样品)就叫做这个样品的密度。计算所有样品点的密度后，

1. 首先选择密度最大的样品作为第一凝聚点。
2. 人为地确定一个正数D(一般$D > d$，常取$D = 2d$).
3. 选出次大密度的样品点，若它与第一个凝聚点的距离大于D，则将其作为第二个凝聚点；否则舍去这点。
4. 选密度次于它的样品, 如果其与第一个凝聚点的距离大于D, 则将其作为第二个凝聚点；否则舍去这点。
5. 按密度大小依次考查，直至全部样品考查完毕为止.

- 随机选择: 如果对样品的性质一无所知，可采用随机数表来选择，打算分几类就选取几个凝聚点，或者就用前k个样品作为凝聚点。这个方法一般不用。

确定初始分类的方法有:

(1) 凭经验人为分类

(2) 选择凝聚点后,每个样品按与其距离最近的凝聚点归类.

(3) 选择一批凝聚点,每个凝聚点自成一类,将样品依次归入与其最近的凝聚点,并重新计算该类重心并作为新凝聚点,再考虑下个样品归类,直至所有样品都归类为止.

(4) 先将数据标准化处理,仍用x_{ij}表示标准化后第i个样品第j个指标.令

$$x_{i\cdot} - \min x_{i\cdot} = \sum_{j=1}^{m} x_{ij}, \quad R = \max_{1\le x\le n} x_{i\cdot} - \min_{1\le x\le n} x_{i\cdot}$$

对每个样品$X_{(i)}$计算

$$\frac{(k-1)(x_{i\cdot} - \min x_{i\cdot})}{R} + 1$$

假设与这个数最接近的整数l,则将样品$X_{(i)}$归入第l类.

(5) 用某种聚类方法得到一个分类,将其作为初始分类.当样本量大时,有时只用部分样品按某种聚类方法进行分类,如用每类重心作为凝聚点, 再用(2)或(3)的方法对全部样品归类后作为初始分类.

二、逐步聚类法

1.按批修改法

(1)按批修改的步骤:

1. 选择一批凝聚点(人为指定),并选定所用距离定义;
2. 将所有样品按其与最近的凝聚点归类;
3. 计算每一类的重心,将重心作为新的凝聚点,转到步骤2,如果某一步骤所有的新凝聚点与前一次的凝聚点重合,则过程终止.

(2)分类函数: 用t_i表示样品$X_{(i)}$所属类的标号.如$t_1 = 2$表示第一个样品属于第2类.

样品$X_{(i)}$到类G_j的距离D_{ij}定义为

$$D_{ij} = (X_{(i)} - \bar{X}^{(j)})'(X_{(i)} - \bar{X}^{(j)})$$

则分类函数定义为:

$$l(G_1, \cdots, G_k) = \sum_{i=1}^{n} D_{it_i}^2.$$

上式定义的分类函数实质上就是系统聚类分析中的离差平方和.

按批修改法的原则就是使这个分类函数逐渐减小,直到不能减小为止.

例6.5.1 已知5个样品的观测值:1,4,5,7,11.试用按批修改的动态聚类法对5个样品进行归类.

解 (1)选凝聚点:用密度法,取$d=2, D=4$;采用欧氏距离.
得到各点的密度分别为:0,1,2,1,0.
所以第一个凝聚点选$X_{(3)}$,第二第三凝聚点分别为$X_{(1)}$,$X_{(5)}$
(2)初始分类:

$$G_1^{(0)}=\{X_{(3)},X_{(2)},X_{(4)}\} \quad G_2^{(0)}=\{X_{(1)}\} \quad G_3^{(0)}=\{X_{(5)}\}$$

(3)修改分类:计算各类重心分别为$5\frac{1}{3}$,1,11.把他们作为新凝聚点重新归类,

$$G_1^{(1)}=\{X_{(3)},X_{(2)},X_{(4)}\}\quad G_2^{(1)}=\{X_{(1)}\}\quad G_3^{(1)}=\{X_{(5)}\}$$

归类后发现结果和原来相同,因此过程终止.最终分类就是$G_i^{(1)}$.

附注1 按批修改法的优点是计算量小,速度快,但结果依赖于凝聚点的选择.

附注2 有时并不要求过程收敛，只是人为规定这个修改过程重复若干次就行了.

附注3 在按批修改法中,有人将步骤3改为:计算每一类重心,取老凝聚点与重心连线的对称点作为新凝聚点转到步骤2.如果某一步新老凝聚点重合,则过程终止.这样做某些场合会有好的结果.

2.逐步修改法

按批修改法是当样品全部归类后才改变凝聚点.而**逐个修改法(也称K-均值法)**是对每个样品进行分类后,同时改变凝聚点.

比较常见的一种逐个修改法的聚类步骤:

(1) 规定样品间的距离,人为定出三个数:K**(分类数)**,C**(类间距离的最小值)**,R**(类内距离最大值)**;取前K个样品点作为凝聚点.

(2)计算这K个凝聚点两两之间的距离,如最小的距离小于C,将这两个凝聚点合并,并把这两个点的重心作为新的凝聚点,再重复步骤2,直至所有凝聚点之间的距离都$\geq C$为止.

(3) 将剩下的$n-K$个样品逐个归类,对每一个样品,计算该样品与所有凝聚点的距离,如最小距离$>R$,则该样品作为新凝聚点;如最小距离$\leq R$,则将该样品归入与它距离最小的凝聚点所在的类,并计算这类的重心并作为新的凝聚点.如凝聚点之间的距离都$\geq C$,则考虑下一个样品,否则用步骤2进行合并后再考虑下个样品,直至所有样品都归类

(4)采用步骤(3)的所有样品归类后的凝聚点，将样品从头至尾再逐个按步骤3进行归类,不同之处是:某个样品归类后,如分类与原来一致,则重心不必计算;如分类与原来不同,则涉及到的两类重心要重新计算.

如果新的分类与上一次相同,则聚类过程结束,否则重复步骤4.

例6.5.2 对例6.5.1的5个样品用逐个修改法进行聚类.

解 (1)用欧式距离,取$K=3, C=2, R=3$.取$X_{(1)}, X_{(2)}, X_{(3)}$作为凝聚点.

(2)计算凝聚点之间的距离,$d(1,2)=3, d(1,3)=4, d(2,3)=1<C=2$,将$X_{(2)}, X_{(3)}$合并为新类,重心为4.5,它与$X_{(1)}$的距离为$d(1,2^*)=3.5>C$.所以凝聚点有两个:1和4.5.

(3)考虑样品$X_{(4)}$,它与两个凝聚点的距离分别为$d(G_1, X_{(4)}) = 6$和$d(G_2, X_{(4)}) = 2.5(< R = 3)$,不能作为新凝聚点,应归入$G_2$,再考虑$X_{(5)}$,

因$d(G_1, X_{(5)}) = 10$和$d(G_2, X_{(5)}) = 17/3$最小值大于3,所以作为新凝聚点.

至此得到3类:$G_1 = \{X_{(1)}\}$,$G_2 = \{X_{(2)}, X_{(3)}, X_{(4)}\}$,$G_3 = \{X_{(5)}\}$

(4)将样品从头至尾按(3)进行归类,聚类结果同上.故过程结束,得到最终分类如上.

附注1 逐个修改法的最终分类与样品的考虑顺序有关，例如上题中如果按$X_{(5)} \Rightarrow X_{(1)}$ 的次序考虑分类，尽管所选择的K, C, R不变，则分类的结果完全不一样。

附注2 逐个修改法的最终分类与三个参数有关，因此在计算过程中最好让三个参数适当变化，最后根据实际问题的要求取舍聚类结果。

R语言中用于动态聚类的函数是kmeans()函数

- K 均值方法采用逐个修必方法，最早是MacQueen在1967年提出来的.
- 函数的格式为
  Kmeans(x, centers, iter.max=10, nstart=1,
  algorithm=c("Hartigan-Wong", "Lloyd", "Forgy",
  "MacQueen"))
  - x是由数据构成的矩阵或数据表单;
  - centers是聚类的个数或者是初始类的中心;
  - iter.max 为最大迭代次数(缺省为10);
  - nstart 是随机集合的个数
  - algorithm 为动态聚类的算法(缺省为Hartigam-Wong)

逐步聚类法

- 计算距离：$d <- dist(D641A, "minkowski", p = 2)$.
- 动态聚

  类：$km <- kmeans(d, 5, algorithm = "MacQueen")$
- 查看聚类结果：km

例6.5.3 试用R软件中kmeans(快速聚类)过程对16个地区农民生活水平的调查数据进行分类.得到输出结果为:

Cluster means:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	72.03562	37.15761	51.49619	42.98211	20.44668	25.77283	19.49910	30.84610	128.90169	29.42041	43.16667	14.26530	11.87911	13.16659	38.08154	45.510812
2	15.47807	47.36593	94.11984	91.68695	69.68590	45.26871	51.17506	78.76929	75.40502	41.03906	15.47807	49.65133	56.14565	61.34780	68.90071	89.955563
3	63.49706	112.73757	159.97941	161.28074	141.69522	117.05575	123.72783	149.36981	0.00000	108.52033	87.31299	120.10505	127.33166	131.64867	134.82408	157.0
4	63.86026	15.05978	46.44784	46.58815	31.57082	18.48500	38.14363	35.31583	118.28443	16.87456	37.42694	31.73775	29.69162	31.89667	21.48491	43.095207
5	102.75700	46.56484	12.03771	10.92143	25.54472	48.42295	57.02263	14.46146	156.92329	49.68781	74.50881	50.32871	42.24936	38.39859	26.77143	9.215736

Clustering vector:

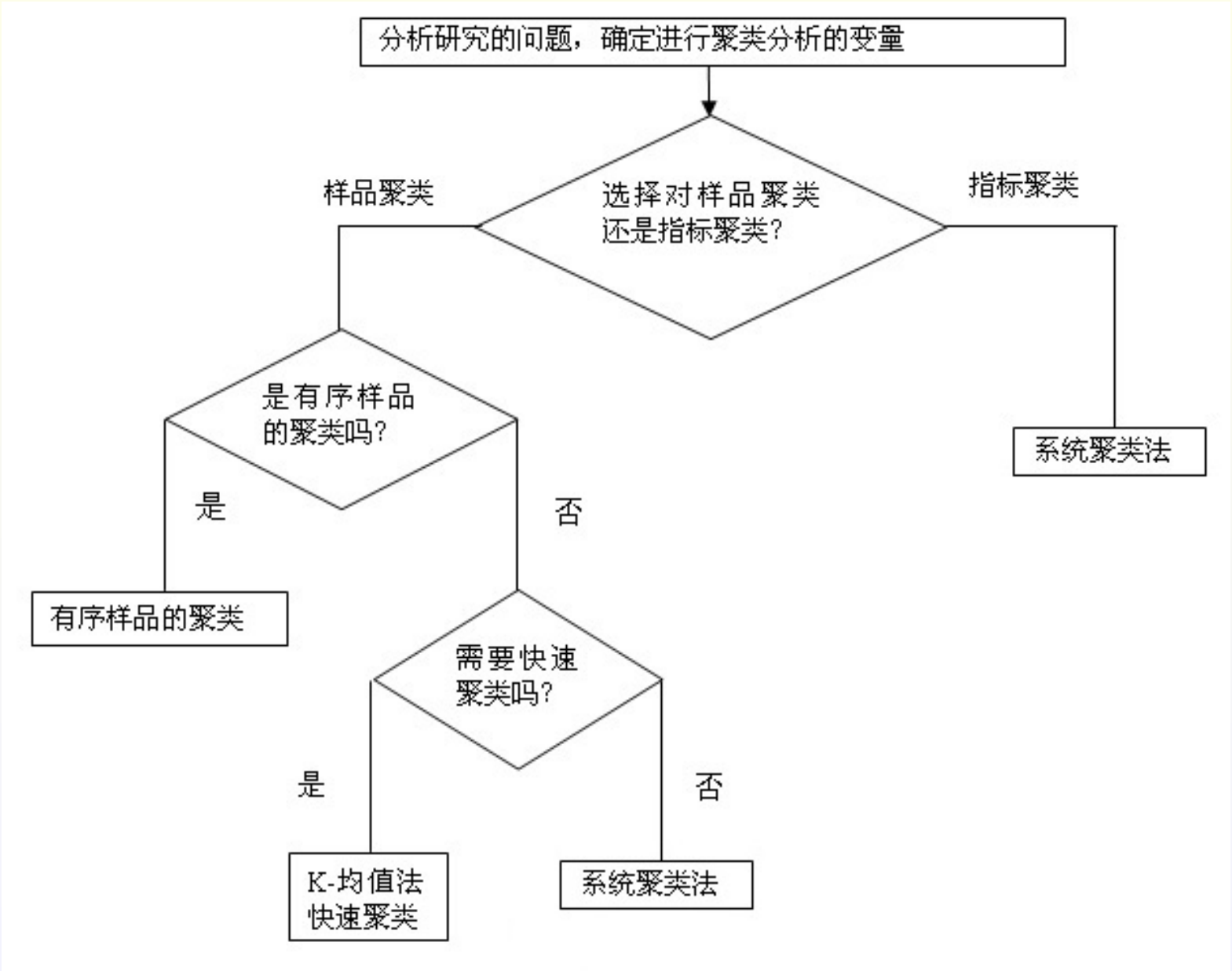
```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16
 2  4  5  5  1  4  1  5  3  4  2  1  1  1  4  5
```

Within cluster sum of squares by cluster:

```
[1] 6024.740 6331.728    0.000 5678.403 2911.569
 (between_SS / total_SS =  91.5 %)
```

Available components:

```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"    "size"
```


分析研究的问题，确定进行聚类分析的变量
选择对样品聚类还是指标聚类?
样品聚类
指标聚类
是有序样品的聚类吗?
是
否
系统聚类法
有序样品的聚类
需要快速聚类吗?
是
否
K-均值法快速聚类
系统聚类法

6.6 有序样品聚类法(最优分割法)

有序样品聚类法(最优分割法)

有些实际问题中,要求样品分类时不能打乱次序.例如在油田勘探中,需要通过岩心了解地层的结构,故而要求对地层的不同结构进行分类.这时岩心所在的位置(即样品次序)在分类时不能打乱.

如果用$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$表示$n$个有序的样品,则每一类必须是这样的形式:$X_{(i)}, X_{(i+1)}, \cdots, X_{(i+k)}$,其中$1 \leq i \leq n, k \geq 0$,即同一类样品必须相互邻接.这种分类问题称为**有序样品聚类法(最优分割法)**

有序样品聚类法的基本思想:

一开始将所有样品归为一类,然后分成二类,三类等等,直到分为n类,并且定义分类的损失函数(类似与系统聚类的Ward法),要求分类后产生的离差平方和达到最小.

一、最优分割法的聚类步骤

设有序样品依次是$X_{(1)}, X_{(2)}, \cdots, X_{(n)}$

1.定义类的直径

设某一类G包含样品$X_{(i)}, X_{(i+1)}, \cdots, X_{(j)}$,记为$G = \{i, i+1, \cdots, j\}$

常用的直径有:

$$D(i,j) = \sum_{t=i}^{j} (X_{(t)} - \bar{X}_G)'(X_{(t)} - \bar{X}_G)$$

即类G的离差平方和.

当$m = 1$时,也可定义直径:

$$D(i,j) = \sum_{t=i}^{j} |X_{(t)} - \tilde{X}_G| \qquad \tilde{X}_G\text{是这类数据的中位数.}$$

2.定义分类损失函数

用$b(n,k)$表示将n个有序样品分为k类的某种分法.常记分法$b(n.k)$为:

$$\begin{aligned} G_1 &= \{i_1, i_1+1, \cdots, i_2-1\} \\ G_2 &= \{i_2, i_2+1, \cdots, i_3-1\} \\ \cdots &\quad \cdots\cdots\cdots\cdots \\ G_k &= \{i_k, i_k+1, \cdots, n\} \end{aligned}$$

其中分点为$1 = i_1 < i_2 < \cdots < i_k < n$

定义上述损失函数为:

$$L[b(n,k)] = \sum_{t=1}^{k} D(i_t, i_{t+1} - 1)$$

即表示各类的离差平方和.

因此要寻找一种分法$b(n,k)$,使损失函数L最小.记$P(n,k)$是L达到极小的分类法.

3.$L[b(n,k)]$的递推公式

$$
\begin{aligned}
L[P(n,2)] &= \min_{2\le j\le n}\{D(1,j-1)+D(j,n)\};\\
L[P(n,3)] &= \min_{3\le j\le n}\{L[P(j-1,2)]+D(j,n)\};\\
&\cdots\cdots\\
L[P(n,k)] &= \min_{k\le j\le n}\{L[P(j-1,k-1)]+D(j,n)\}.
\end{aligned}
$$

由第二个公式可以看出,若要将n个样品分为k类的最优分割,应建立在将$j-1$个样品分为$k-1$ 类的最优分割基础上(这里$j=2,3,\cdots,n$)

4.最优解的求法

假设将n个有序样本分成k类，k已知, 求最优的分类法$P(n,k)$,使其在损失函数意义下最小,求法如下:

- 找分点j_k,使递推公式达到极小,即

$$L[P(n,k)] = L[P(j_k-1,k-1)] + D(j_k,n)$$

于是得第k类$G_k = \{j_k, \cdots, n\}$.

- 找j_{k-1}使得

$$L[P(j_k-1,k-1)] = L[P(j_{k-1},k-2)] + D(j_{k-1},j_k-1)$$

得第$k-1$类.

- 重复上述步骤.

总之,求最优解必须计算

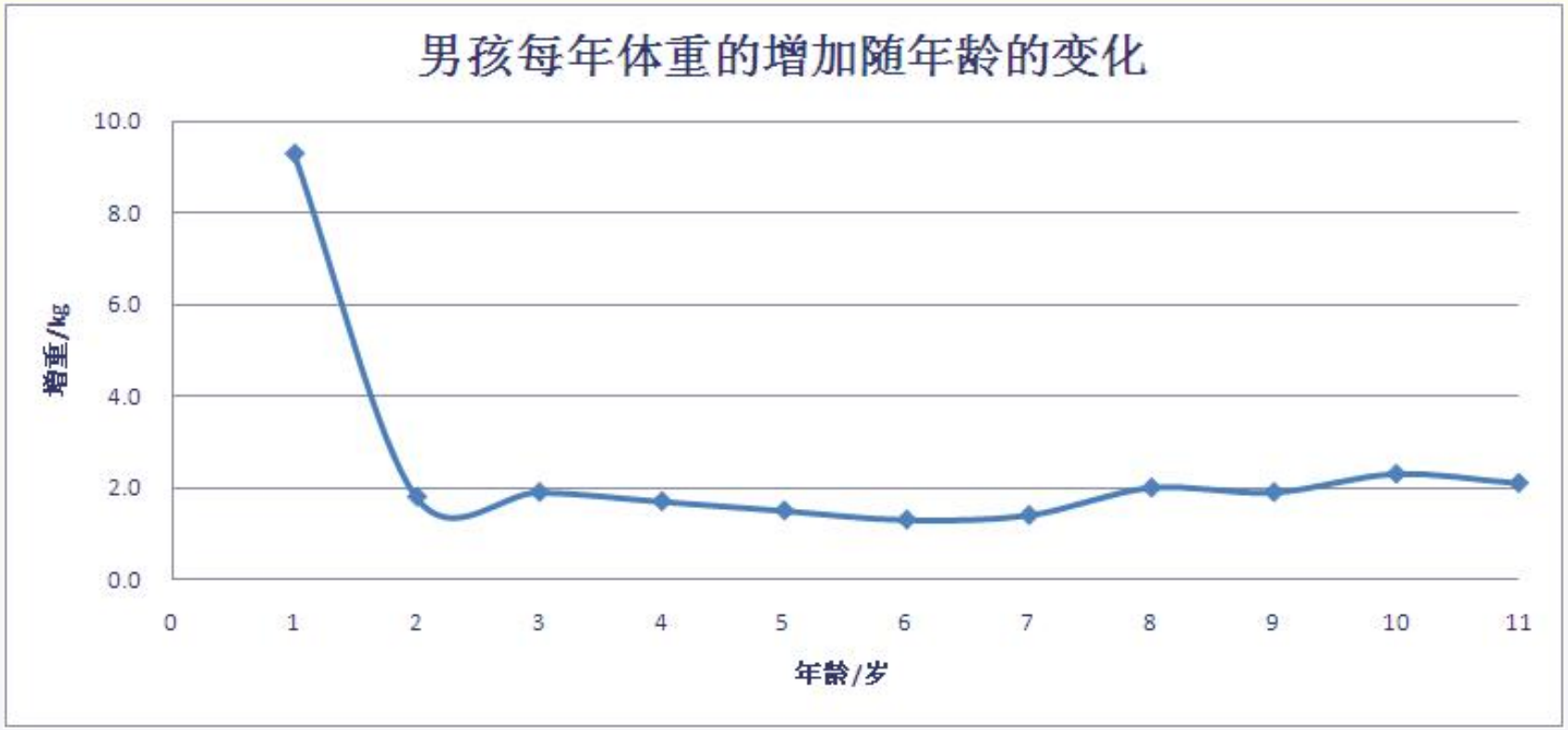
$$\{D(i,j); 1 \le i < j \le n\} \quad 和 \quad \{L[P(i,j)]; 1 \le i \le n, 1 \le j \le n\}$$

二、应用举例

Example

为了看了解儿童的生长发育规律，今统计了男孩出生到11岁每年平均增长的重量如下表所示。试问男孩发育可分为几个阶段？

年龄/岁	1	2	3	4	5	6	7	8	9	10	11
增重/kg	9.3	1.8	1.9	1.7	1.5	1.3	1.4	2.0	1.9	2.3	2.1


男孩每年体重的增加随年龄的变化
增重/kg
10.0
8.0
6.0
4.0
2.0
0.0
0
1
2
3
4
5
6
7
8
9
10
11
年龄/岁

第一步：计算直径$D(i,j)$

第一个元素：$D(1,2)$ 即计算(9.3,1.8)这个组样品之间的距离，即

$$(9.3-5.55)*(9.3-5.55)+(1.8-5.55)*(1.8-5.55)=28.125$$

	1	2	3	4	5	6	7	8	9	10
2	28.125	0.000	0.005	0.020	0.087	0.232	0.280	0.417	0.469	0.802
3	37.007	0.005	0.000	0.020	0.080	0.200	0.232	0.393	0.454	0.800
4	42.207	0.020	0.020	0.000	0.020	0.080	0.087	0.308	0.393	0.774
5	45.992	0.087	0.080	0.020	0.000	0.020	0.020	0.290	0.388	0.773
6	49.128	0.232	0.200	0.080	0.020	0.000	0.005	0.287	0.370	0.708
7	51.100	0.280	0.232	0.087	0.020	0.005	0.000	0.180	0.207	0.420
8	51.529	0.417	0.393	0.308	0.290	0.287	0.180	0.000	0.005	0.087
9	51.980	0.469	0.454	0.393	0.388	0.370	0.207	0.005	0.000	0.080
10	52.029	0.802	0.800	0.774	0.773	0.708	0.420	0.087	0.080	0.000
11	52.182	0.909	0.909	0.895	0.889	0.793	0.452	0.087	0.080	0.020

第二步：利用递推公式计算最小损失函数$L[P(l,k)]$

最小分类损失函数 $L[P(l,k)]$

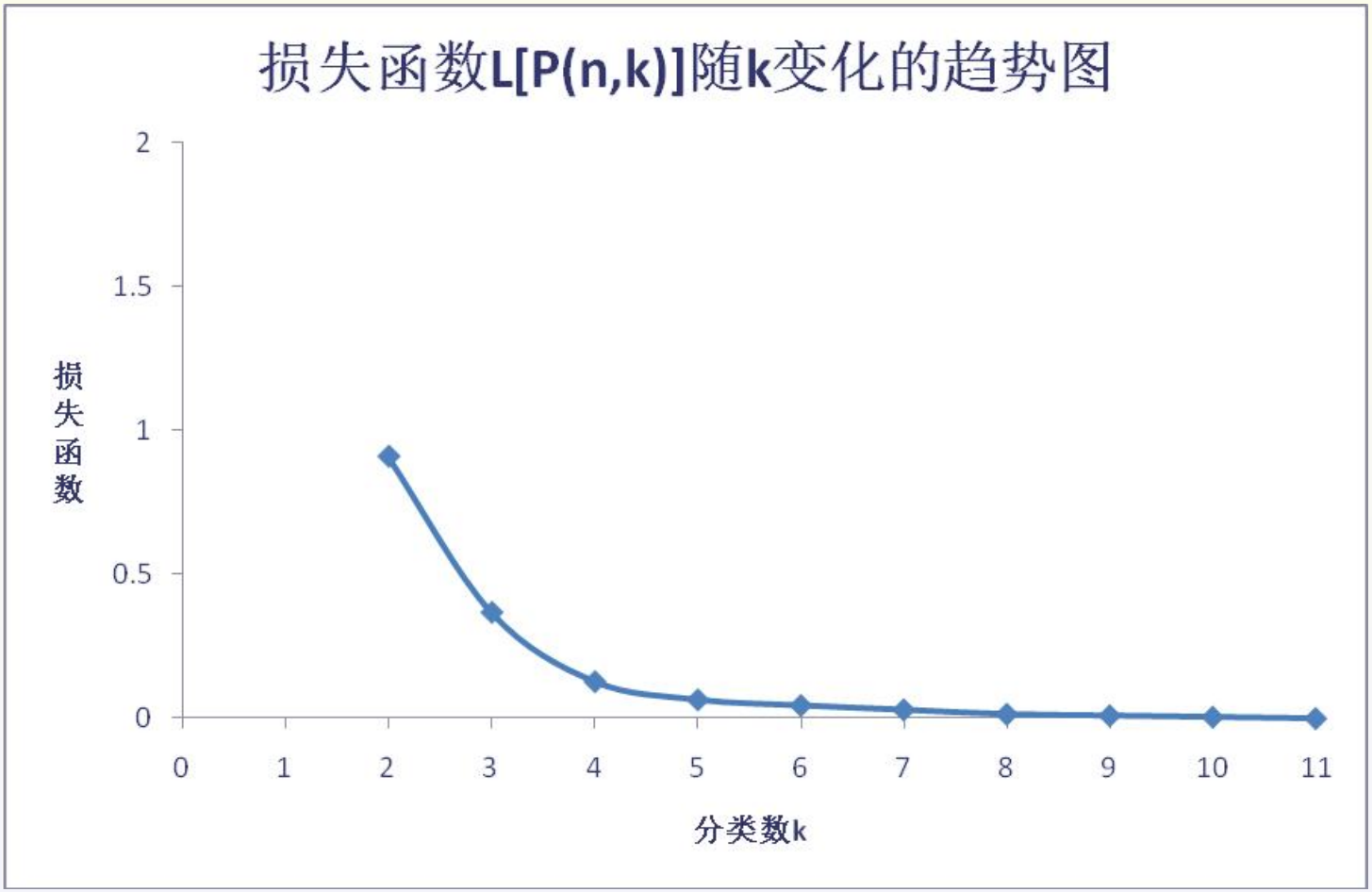
K / l	2	3	4	5	6	7	8	9	10
3	0.005(2)								
4	0.02(2)	0.005(4)							
5	0.088(2)	0.020(5)	0.005(5)						
6	0.232(2)	0.040(5)	0.02(6)	0.005(6)					
7	0.280(2)	0.040(5)	0.025(6)	0.010(6)	0.005(6)				
8	0.417(2)	0.280(8)	0.040(8)	0.025(8)	0.010(8)	0.005(8)			
9	0.469(2)	0.285(8)	0.045(8)	0.030(8)	0.015(8)	0.010(8)	0.005(8)		
10	0.802(2)	0.367(8)	0.127(8)	0.045(10)	0.030(10)	0.015(10)	0.010(10)	0.005(10)	
11	0.909(2)	0.368(8)	0.128(8)	0.065(10)	0.045(11)	0.030(11)	0.015(11)	0.010(11)	0.005(11)

由最小损失函数$L[P(l,k)]$ 就可以求得分为k类的最优分类$(k=1,2,\cdots,11)$.

第三步：决定分类个数，如果k已经，假设分3类。

- 查看最小分类损失函数，k=3列，最后11个元素分三类最后一个分划应是8;
- 再查看前1-7个数的分类，分二类的最优分划是2;
- 结果是$1||2,3,4,5,6,7,||8,9,10,11$

第四步：如何决定分类的个数k



由上图可知曲线在k=3,4处拐弯,所以分为3类或4类较好.

三、分类个数的确定

分类数k的确定对许多问题都很重要,上面给出的方法是通过$L[P(n,k)]$对k作图,在曲线拐弯处来确定k.当曲线拐弯的很平缓时,可以选取的k很多,这时需要其他方法来确定,比如均方比和特征根法.

6.7 变量聚类方法

聚类分析根据分类对象分为Q型(对样品)和R型(对变量)

变量聚类就是研究变量间的相似关系,按照变量的相关关系把他们聚合为若干类,然后观察和说明影响系统特性的主要特征.

一、变量聚类的系统聚类法

(1)可以把观测数据阵中的样品和变量的地位调换一下,也就是把数据阵做一转置,然后仍使用前面所述的聚类函数对变量进行聚类. 但这类方法没有考虑到变量的相关性。

(2)计算变量间的相关系数矩阵R,然后按6.2节中的”变量间的相似系数和距离”中的定义把相关矩阵转化为距离矩阵.然后把此矩阵作为输入数据,使用聚类函数，如hclust等对变量进行聚类.

变量聚类

例6.41用系统聚类的方法进行聚类（变量聚类）

1. 输入数据
   - $D641 <- read.table("ex641.txt", header = F)$
   - $D641A <- D641[, 2:7]$
2. 计算相关矩阵
   - $x <- cor(D641A)$

3 系统聚类分析

- as.dist()的作用是将普通矩阵转化为聚类分析用的距离结构
- $d <- as.dist(x);$
- $hc <- hclust(d, "average")$
- $dend <- as.dendrogram(hc)$

4 画谱系图

- 直接打$plog(hc)$
- $de <- dendrapply(dend, addE)$
- $plot(de, nodePar = nP)$

二、聚类分析在客户细分中的应用

目标：消费同一种类的商品或服务时，不同的客户有不同的消费特点，通过研究这些特点，企业可以制定出不同的营销组合，从而获取最大的消费者剩余，这就是客户细分的主要目的。

常用的客户分类方法主要有三类：

- 经验描述法，由决策者根据经验对客户进行类别划分；
- 传统统计法，根据客户属性特征的简单统计来划分客户类别；
- 非传统统计方法，即基于人工智能技术的非数值方法。

聚类分析法兼有后两类方法的特点，能够有效完成客户细分的过程。例如，

- 客户的购买动机一般由需要、认知、学习等内因和文化、社会、家庭、小群体、参考群体等外因共同决定。要按购买动机的不同来划分客户时，可以把前述因素作为分析变量，并将所有目标客户每一个分析变量的指标值量化出来，再运用聚类分析法进行分类。
- 在指标值量化时如果遇到一些定性的指标值，可以用一些定性数据定量化的方法加以转化，如模糊评价法. 除此之外，可以将客户满意度水平和重复购买机会大小作为属性进行分类；还可以在区分客户之间差异性的问题上纳入一套新的分类法.

- 将客户的差异性变量划分为五类：产品利益、客户之间的相互作用力、选择障碍、议价能力和收益率，依据这些分析变量聚类得到的归类，可以为企业制定营销决策提供有益参考。

以上分析的共同点在于都是依据多个变量进行分类，这正好符合聚类分析法解决问题的特点；不同点在于从不同的角度寻求分析变量，为某一方面的决策提供参考，这正是聚类分析法在客户细分问题中运用范围广的体现。