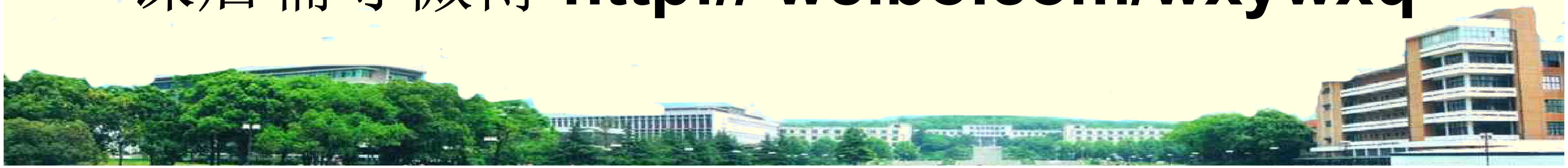
第3讲 分类与判别

汪晓银 教授

课后辅导微博 **http:// weibo.com/wxywxq**



3.1 模糊聚类分析

一、基本概念及定理

定义：设$R=(r_{ij})_{n\times n}$是n阶模糊方阵，I是n阶单位方阵，若R满足

（1）自反性：$I \le R (\Leftrightarrow r_{ij}=1)$;

（2）对称性：$R^T=R (\Leftrightarrow r_{ij}=r_{ji})$;

（3）传递性：$R^2 \le R (\Leftrightarrow \max\{(r_{ik}\wedge r_{kj}) \mid 1\le k\le n\} \le r_{ij})$;

则称R为模糊等价矩阵。

3.1 模糊聚类分析

自反性可推出：$R \le R^2$

与传递性：$R \ge R^2$ 结合，可得到：

模糊等价矩阵实际满足：$R = R^2$

传递性的理解：

若x_i与x_k有关系R，x_k与x_j有关系R，则x_i与x_j有关系R，这种关系可以理解为大于等于某个阈值 λ，在传递性下，$R \ge R^2$.

等价布尔矩阵是一种普通关系，在传递性条件下，是可以分类的，即$r_{ij}=1$，则x_i与x_j为一类。
我们要分类必须将模糊等价矩阵转化为等价布尔矩阵。所以引入 λ 截矩阵。

3.1 模糊聚类分析

定理：设$\boldsymbol{R}$是$\boldsymbol{n}$阶模糊等价矩阵，则$\forall \mathbf{0} \le \lambda < \mu \le \mathbf{1}, \boldsymbol{R}_{\mu}$所决定的分类中的每一个类是$\boldsymbol{R}_{\lambda}$所决定的分类中的某个子类。

该定理表明，当$\lambda < \mu$时，$\boldsymbol{R}_{\mu}$的分类是$\boldsymbol{R}_{\lambda}$分类的加细，当λ由 **1** 变到 **0** 时，$\boldsymbol{R}_{\lambda}$的分类由细变粗，形成一个动态的聚类图。

3.1 模糊聚类分析

例：设 $U=\{x_1,x_2,x_3,x_4,x_5\}$, 对于模糊等价矩阵

$$R=\begin{pmatrix} 1 & 0.4 & 0.8 & 0.5 & 0.5 \\ 0.4 & 1 & 0.4 & 0.4 & 0.4 \\ 0.8 & 0.4 & 1 & 0.5 & 0.5 \\ 0.5 & 0.4 & 0.5 & 1 & 0.6 \\ 0.5 & 0.4 & 0.5 & 0.6 & 1 \end{pmatrix}$$

当$\lambda=1$时，分类为$\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\}$;

当$\lambda=0.8$时，分类为$\{x_1,x_3\},\{x_2\},\{x_4\},\{x_5\}$;

当$\lambda=0.6$时，分类为$\{x_1,x_3\},\{x_2\},\{x_4,x_5\}$;

当$\lambda=0.5$时，分类为$\{x_1,x_3,x_4,x_5\},\{x_2\}$;

当$\lambda=0.4$时，分类为$\{x_1,x_2,x_3,x_4,x_5\}$.

3.1 模糊聚类分析

实际应用中建立一个模糊等价矩阵式不容易的，传递性不易满足。

定义：设$R=(r_{ij})_{n\times n}$是n阶模糊方阵，I是n阶单位方阵，若R满足

（1）自反性：$I \le R$；

（2）对称性：$R^T = R$；

则称R为模糊相似矩阵。

3.1 模糊聚类分析

定理：设R是n阶模糊相似矩阵，则存在一个最小的自然数$k(k \le n)$，使得R^k为模糊等价矩阵，且对一切大于k的自然数l，恒有$R^l = R^k$.

R^k称为R的传递闭包矩阵，记为$t(R)$.

$$R \to R^2 \to R^4 \to \cdots R^{2^i} = R^{2^{i+1}}, t(R) = R^{2^i}$$

3.1 模糊聚类分析

例：设有模糊相似矩阵 $R = \begin{pmatrix} 1 & 0.1 & 0.2 \\ 0.1 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{pmatrix}$

$$R \circ R = \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{pmatrix} = R^2$$

$$R^2 \circ R^2 = \begin{pmatrix} 1 & 0.2 & 0.2 \\ 0.2 & 1 & 0.3 \\ 0.2 & 0.3 & 1 \end{pmatrix} = R^2 = t(R).$$

3.1 模糊聚类分析

二、模糊聚类的一般步骤

1、建立数据矩阵

设论域$U=\{x_1,x_2,\cdots,x_n\}$为被分类对象，每个对象又由m个指标表示其性状：

$$x_i=\{x_{i1},x_{i2},\ldots,x_{im}\}\quad(i=1,2,\cdots,n)$$

则得到原始数据矩阵为$X=(x_{ij})_{n\times m}$.

在实际问题中，不同的数据一般有不同的量纲，为了使有不同量纲的量能进行比较，需要将数据规格化，常用的方法有：

3.1 模糊聚类分析

（1）标准差标准化

对于第i个变量进行标准化，就是将x_{ij}换成x'_{ij}，即

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j} \quad (1 \le i \le n, 1 \le j \le m)$$

式中：$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, S_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}.$

3.1 模糊聚类分析

（2）极差正规化 $$x'_{ij}=\frac{x_{ij}-\min\{x_{ij}\}}{\max\{x_{ij}\}-\min\{x_{ij}\}}$$

（3）极差标准化 $$x'_{ij}=\frac{x_{ij}-\bar{x}_j}{\max\{x_{ij}\}-\min\{x_{ij}\}}$$

（4）最大值规格化 $$x'_{ij}=\frac{x_{ij}}{M_j}$$

其中： $M_j=\max(x_{1j},x_{2j},\cdots,x_{nj})$

3.1 模糊聚类分析

2、建立模糊相似矩阵

建立x_i与x_j相似程度$r_{ij}=R(x_i,x_j)$的方法主要有：

（1）相似系数法

①夹角余弦法 $$r_{ij}=\frac{\sum_{k=1}^{m}x_{ik}\cdot x_{jk}}{\sqrt{\sum_{k=1}^{m}x_{ik}^2}\cdot\sqrt{\sum_{k=1}^{m}x_{jk}^2}}$$

②相关系数法 $$r_{ij}=\frac{\sum_{k=1}^{m}\left|x_{ik}-\bar{x}_i\right|\left|x_{jk}-\bar{x}_j\right|}{\sqrt{\sum_{k=1}^{m}(x_{ik}-\bar{x}_i)^2}\cdot\sqrt{\sum_{k=1}^{m}(x_{jk}-\bar{x}_j)^2}}$$

3.1 模糊聚类分析

（2）距离法

一般地，取$r_{ij}=1-c(d(x_i,x_j))^{\alpha}$，其中$c,\alpha$为适当选取的参数，它使得$0\le r_{ij}\le 1$.采用的距离有：

①Hamming距离 $$d(x_i,x_j)=\sum_{k=1}^{m}\left|x_{ik}-x_{jk}\right|$$

②Euclid距离 $$d(x_i,x_j)=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}$$

③Chebyshev距离 $$d(x_i,x_j)=\max_{1\le k\le n}\left|x_{ik}-x_{jk}\right|$$

3.1 模糊聚类分析

（3）贴近度法

①最大最小法

$$r_{ij}=\frac{\sum_{k=1}^{m}(x_{ik}\wedge x_{jk})}{\sum_{k=1}^{m}(x_{ik}\vee x_{jk})}$$

②算术平均最小法

$$r_{ij}=\frac{\sum_{k=1}^{m}(x_{ik}\wedge x_{jk})}{\frac{1}{2}\sum_{k=1}^{m}(x_{ik}+x_{jk})}$$

③几何平均最小法

$$r_{ij}=\frac{\sum_{k=1}^{m}(x_{ik}\wedge x_{jk})}{\sum_{k=1}^{m}\sqrt{x_{ik}.x_{jk}}}$$

3.1 模糊聚类分析

3、聚类并画出动态聚类图

（**1**）模糊传递闭包法

步骤：

①求出模糊相似矩阵$\boldsymbol{R}$的传递闭包$\boldsymbol{t(R)}$；

②按λ由大到小进行聚类；

③画出动态聚类图。

3.1 模糊聚类分析

例：考虑某环保部门对该地区 **5** 个环境区域 $X=\{x_1,x_2,x_3,x_4,x_5\}$按污染情况进行分类。设每个区域包含空气、水分、土壤、作物 **4** 个要素，环境区域的污染情况由污染物在 **4** 个要素中的含量超过的程度来衡量。设这 **5** 个环境区域的污染数据为：

$\boldsymbol{x_1=(80,10,6,2), x_2=(50,1,6,4), x_3=(90,6,4,6),}$

$\boldsymbol{x_4=(40,5,7,3), x_5=(10,1,2,4).}$

试对 X 进行分类。

3.1 模糊聚类分析

解：由题设知特性指标矩阵为

$$X^* = \begin{pmatrix} 80 & 10 & 6 & 2 \\ 50 & 1 & 6 & 4 \\ 90 & 6 & 4 & 6 \\ 40 & 5 & 7 & 3 \\ 10 & 1 & 2 & 4 \end{pmatrix}$$

采用最大值规格化法将数据规格化为

$$X = \begin{pmatrix} 0.89 & 1 & 0.86 & 0.33 \\ 0.56 & 0.10 & 0.86 & 0.67 \\ 1 & 0.60 & 0.57 & 1 \\ 0.44 & 0.2 & 1 & 0.5 \\ 0.11 & 0.10 & 0.29 & 0.67 \end{pmatrix}$$

3.1 模糊聚类分析

用最大最小法构造模糊相似矩阵得到

$$R=\begin{pmatrix} 1 & 0.54 & 0.62 & 0.63 & 0.24 \\ 0.54 & 1 & 0.55 & 0.70 & 0.53 \\ 0.62 & 0.55 & 1 & 0.56 & 0.37 \\ 0.63 & 0.70 & 0.56 & 1 & 0.38 \\ 0.24 & 0.53 & 0.37 & 0.38 & 1 \end{pmatrix}$$

用平方法合成传递闭包

$$t(R)=R^4=\begin{pmatrix} 1 & 0.63 & 0.62 & 0.63 & 0.53 \\ 0.63 & 1 & 0.62 & 0.70 & 0.53 \\ 0.62 & 0.62 & 1 & 0.62 & 0.53 \\ 0.63 & 0.70 & 0.62 & 1 & 0.53 \\ 0.53 & 0.53 & 0.53 & 0.53 & 1 \end{pmatrix}$$

3.1 模糊聚类分析

将$t(R)$中的元素从大到小编排如下：

1>0.70>0.63>0.62>0.53

取 $\lambda=1$，得

$$t(R)_1=\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

X被分成 5 类：$\{x_1\},\{x_2\},\{x_3\},\{x_4\},\{x_5\}$.

3.1 模糊聚类分析

取$\lambda = 0.7$，得

$$t(R)_{0.7} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

X被分成 4 类：

$\{x_1\},\{x_3\},\{x_2,x_4\},\{x_5\}.$

取$\lambda = 0.63$，得

$$t(R)_{0.63} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

X被分成 3 类：

$\{x_1,x_2,x_4\},\{x_3\},\{x_5\}.$

3.1 模糊聚类分析

取$\lambda = 0.62$，得

$$t(R)_{0.62} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

X被分成 2 类：

$\{x_1, x_2, x_3, x_4\}, \{x_5\}$.

取$\lambda = 0.53$，得

$$t(R)_{0.53} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

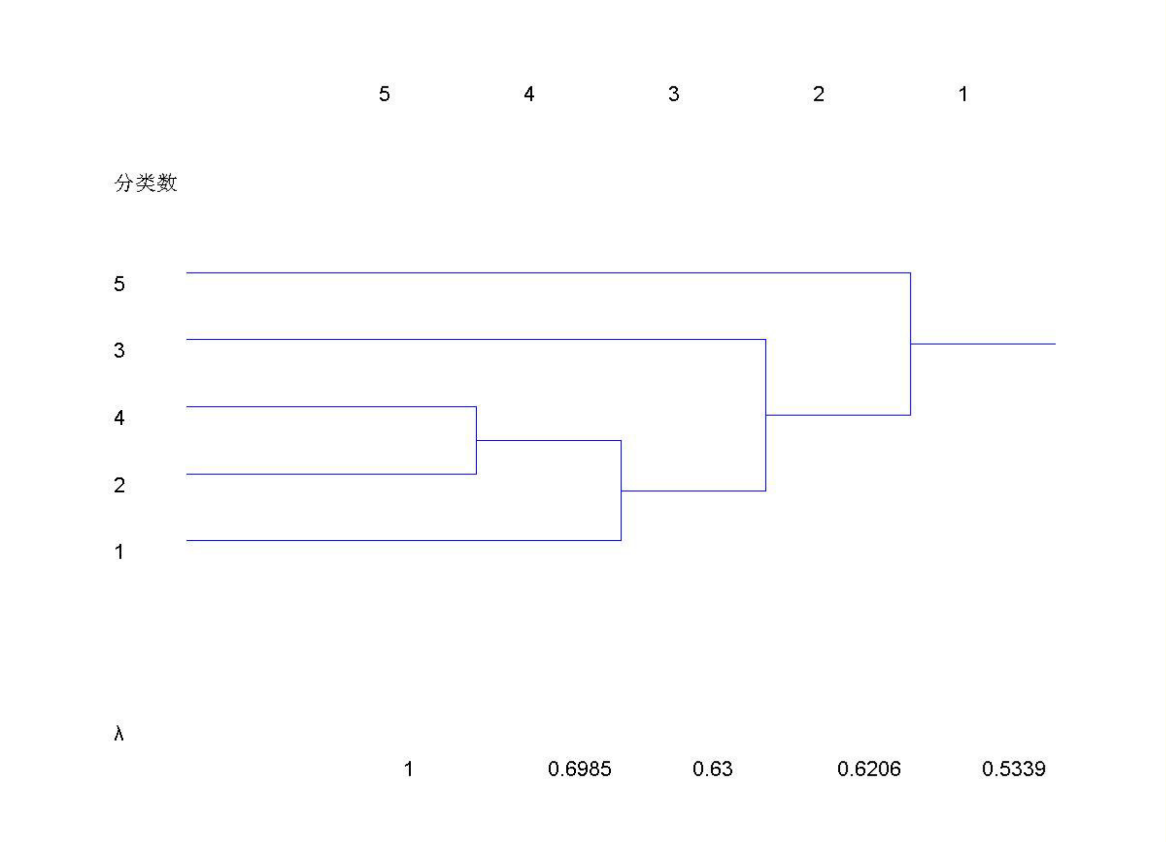
X被分成 1 类：

$\{x_1, x_2, x_3, x_4, x_5\}$.

3.1 模糊聚类分析

X=[80 10 6 2;50 1 6 4;90 6 4 6;40 5 7 3;10 1 2 4]

输出动态聚类图如下：



3.1 模糊聚类分析

最佳分类（最佳阈值λ）

方法：对每个阈值下的分类计算一个**F**值，取最大**F**值对应的分类作为最佳分类。

计算方式如下：

设某个阈值λ水平下，共分了r个类，第i类有n_i个对象。

$\bar{x}_k(i)$： 第**i**类中全体对象的第**k**个指标的均值；

类中指标均值向量： $\bar{x}(i)=(\bar{x}_1(i)\cdots\bar{x}_m(i))$

$\bar{x}(i)$： 全体对象的第**k**个指标的均值；

总指标均值向量： $\bar{x}=(\bar{x}_1\cdots\bar{x}_m)$

3.1 模糊聚类分析

模糊统计量

$$F=\frac{\sum_{i=1}^{r} n_i\left[M\left(\bar{x}(i),\bar{x}\right)\right]^2 \Big/ (r-1)}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}\left[M\left(x_j(i),\bar{x}(i)\right)\right]^2 \Big/ (n-r)}$$

其中**M**为向量间的欧氏距离

分子为类均值与总均值的差异，描述类与类间距离

分母为每个元素与类均值的差异，描述类内元素间距离

故**F**越大，类之间差异越大，从而分类越合理。

3.2 系统聚类

聚类分析又称群分析，它是研究分类问题的一种多元统计方法。所谓类，通俗地说，就是指相似元素的集合。那么要将相似元素聚为一类，通常选取元素的许多共同指标，然后通过分析元素的指标值来分辨元素间的差距，从而达到分类的目的。

聚类分析可以分为：**Q**型（样品分类）分类、**R**型（指标分类）分类。这里介绍的是**Q**型（样品分类）分类。

3.2 系统聚类

聚类分析前的预处理步骤：

1)确定聚类类型：对样品聚类称**Q**型聚类；

对变量聚类称**R**型聚类。

2)数据预处理

原因：实际应用所使用的样本资料中，由于不同的变量具有不同的计量单位（或量纲），并且具有不同的数量级，为了使具有不同计量单位和数量级的数据能够放在一起进行比较分析，通常都要对数据进行变换处理。

常用方法有：中心化变换；规格化变换（极差正规化）；标准化变换；对数变换等

3.2 系统聚类

3) 研究样品之间的关系。通常有两种方法：

相似系数。性质相近的相似系数的绝对值越接近于**1**，彼此不相关的相似系数的绝对值越接近于**0**。

常用相似系数有：夹角余弦；相关系数；指数相似系数；非参数方法灯

计算距离。将样品看作**P**维空间的一点，通过计算不同样品的距离，距离越接近的点归为一类，距离远的点归为不同类。

常用距离有：明科夫斯基距离；欧氏距离；绝对值距离；切比雪夫距离；兰氏距离；马氏距离。

4)计算距离矩阵或相似性系数矩阵**D**。

3.2 系统聚类

聚类分析的一般步骤(Q-型分类)

1) 每个样本独自成类，$G_i = \{X_i\} \quad i = 1,2,..n$

2) 由距离矩阵或相似性系数矩阵D，找到当前最小的D_{ij}, 并将类G_i、G_j合为一类得到一个新类$G_r=\{G_i、G_j\}$

3) 从新计算类间的距离，得到新的矩阵D。

4) 重复第2步直到全部合为一类。

3.2 系统聚类

进行聚类分析时，由于对类与类之间的距离的定义和理解不同，并类的过程中又会产生不同的聚类方法。常用的系统聚类方法有**8**种：最短距离法；最长距离法；中间距离法；重心法；类平均法；可变类平均法；可变法；离差平方和法。

3.2 系统聚类

例：从**21**个工厂中抽出同类产品，每个产品测两个指标，欲将各厂的质量情况进行分类。

工厂指标观测值

工厂	1	2	3	4	5	6	7	8	9	10	11
指标1	0	0	2	2	4	4	5	6	6	7	-4
指标2	6	5	5	3	4	3	1	2	1	0	3

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

3.2 系统聚类

```
data ex;input x1 x2 factory$@@;
cards;
/*数据省略*/
;
proc cluster
data=ex method=ward ccc pseudo outtree=tree;
id factory;
run;
proc tree data=tree horizontal;
id factory;
run;
```

3.2 系统聚类

ccc表示要计算半偏**R2**，**R2**和**ccc**立方聚类标准统计量，这三个统计量和下面的伪**F**和伪**t2**统计量，主要用于检验聚类的效果。当把数据从**G+1**类合并为**G**类时，半偏**R2**统计量说明了本次合并信息的损失程度，统计量大表明损失程度大。 **R2**统计量反映类内离差平方和的大小，统计量大表明类内离差平方和小。 **ccc**统计量的值大说明聚类的效果好。

Pseudo说明要计算伪**F**和伪**t2**统计量。一般认为，伪**F**统计量出现峰值时的所对应的分类是较佳的分类选择。当把数据从**G+1**类合并为**G**类时，伪**t2**统计量的值大，说明不应该合并这两类。

3.2 系统聚类

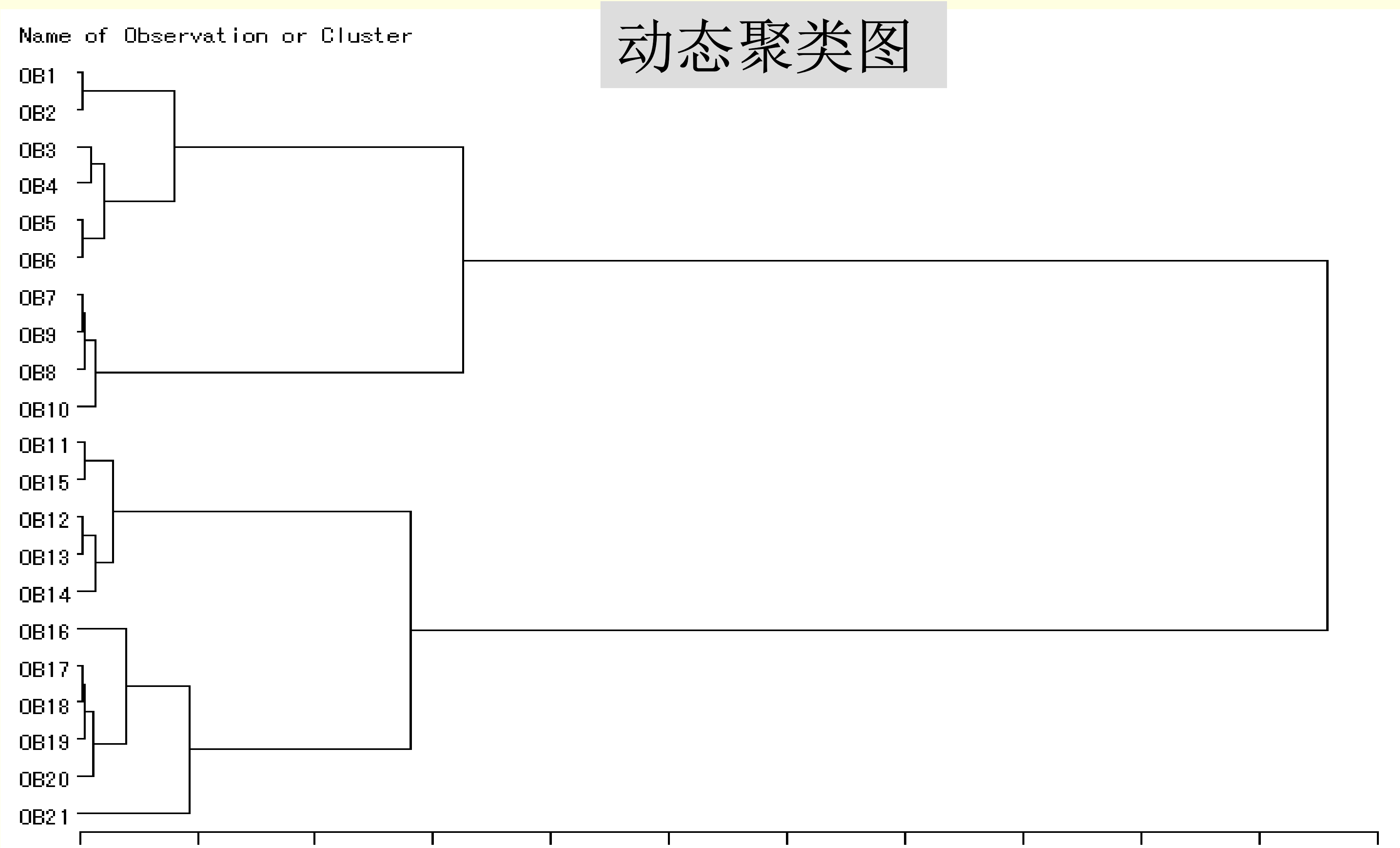
Cluster History

NCL	--Clusters Joined---		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Tie
20	f1	f2	2	0.0012	.999	.	.	42.2	.	T
19	f7	f9	2	0.0012	.998	.	.	44.4	.	T
18	f12	f13	2	0.0012	.996	.	.	47.0	.	T
17	f17	f18	2	0.0012	.995	.	.	49.9	.	T
16	f5	f6	2	0.0012	.994	.	.	53.1	.	
15	CL19	f8	3	0.0021	.992	.	.	51.1	1.7	T
14	CL17	f19	3	0.0021	.990	.	.	51.3	1.7	
13	f11	f15	2	0.0025	.987	.	.	51.1	.	
12	f3	f4	2	0.0050	.982	.	.	45.0	.	
11	CL14	f20	4	0.0060	.976	.	.	40.8	3.6	
10	CL15	f10	4	0.0067	.969	.	.	38.8	4.0	
9	CL18	f14	3	0.0071	.962	.	.	38.4	5.7	
8	CL12	CL16	4	0.0106	.952	.	.	36.7	3.4	
7	CL13	CL9	5	0.0141	.938	.	.	35.1	3.9	
6	f16	CL11	5	0.0196	.918	.	.	33.6	6.3	
5	CL20	CL8	6	0.0401	.878	.	.	28.8	8.9	
4	CL6	f21	6	0.0463	.832	.830	0.04	28.0	6.4	
3	CL7	CL4	11	0.1406	.691	.744	-1.2	20.1	12.6	
2	CL5	CL10	10	0.1623	.529	.538	-.13	21.3	19.0	
1	CL2	CL3	21	0.5288	.000	.000	0.00	.	21.3	

3.2 系统聚类

Cluster History表示聚类的具体过程，**NCL**表示当前系统存在类的总个数，**Clusters Joined**表示当前加入的编号，例如**NCL**等于**20**时，是类**1**，**2**聚为一类，**FREQ**表示新类的元素个数。**SPRSQ**表示类与类间规格化最短距离，**RSQ**表示**R2**统计量，**ERSQ**表示半偏**R2**统计量，**CCC**统计量值。**PSF**为伪**F**统计量，**PST2**为伪**t2**统计量。**Tie**表示"节"，是指当前类间最小距离不止一个的时候，此时可以任意选择一对最短距离进行聚类，在计算其他类与新类的距离。从**CCC**统计量的结果可以看出，最大值对应的类数为**4**。从四类合并为三类时，伪**t2**统计量显著的增加，伪**F**统计量下降显著，综合各方面的结果，因此分**4**类最为合适。

3.2 系统聚类



Name of Observation or Cluster
动态聚类图
OB1
OB2
OB3
OB4
OB5
OB6
OB7
OB9
OB8
OB10
OB11
OB15
OB12
OB13
OB14
OB16
OB17
OB18
OB19
OB20
OB21

3.2 系统聚类

综合以上分析，可以得到结果，将工厂分为**4**类，分别为

第**1**类：**f1,f2,f3,f4,f5,f6**；

第**2**类：**f7,f8,f9,f10**

第**3**类：**f11,f12,f13,f14,f15**；

第**4**类：**f16,f17,f18,f19,f20,f21**。

3.3 模糊模式识别

模式识别的本质特征：一是事先已知若干标准模式，称为标准模式库；二是有待识别的对象。

所谓模糊模式识别，是指在模式识别中，模式是模糊的，标准模式库中提供的模式是模糊的。或有待识别的对象是模糊的。

3.3 模糊模式识别

一、最大隶属原则

最大隶属原则Ⅰ：

设$A_1, A_2, \cdots, A_m$为给定的论域 U 上的 m 个模糊模式，$x_0 \in U$为一个待识别对象，若

$$A_i(x_0) = \max\{A_1(x_0), A_2(x_0), \cdots, A_m(x_0)\},$$

则认为x_0优先归属于模糊模式A_i。

3.3 模糊模式识别

最大隶属原则 II：

设 A 为给定论域 U 上的一个模糊模式，$x_1, x_2, \cdots, x_n$ 为 U 中的n个待识别对象，若

$$A(x_i) = \max\{A(x_1), A(x_2), \cdots, A(x_n)\},$$

则认为模糊模式 A 应优先录取x_i。

3.3 模糊模式识别

例：已知年轻人的模糊集隶属函数为

$$A_1(x)=\begin{cases}1, & x\le 25\\ [1+(\frac{x-25}{5})^2]^{-1}, & 25<x\le 100\end{cases}$$

老年人的模糊集的隶属函数为

$$A_2(x)=\begin{cases}0, & x\le 50\\ [1+(\frac{x-50}{5})^{-2}]^{-1}, & 50<x\le 100\end{cases}$$

现有某人55岁，问他相对来讲是老年还是年轻？

解：$A_1(55)=\frac{1}{37}, A_2(55)=\frac{1}{2}.$

$A_2(55)=\max\{A_1(55), A_2(55)\}$

按最大隶属原则，该人属于老年。

3.3 模糊模式识别

例：今考虑三角形的识别问题。设 U 为所有待识别的三角形所构成的集合，由于每一个三角形完全由其三个内角所确定，故可以三角形的三个内角α,β,γ作为特性指标。于是，论域 U 可记为

$$U=\{x=(\alpha,\beta,\gamma)\mid\alpha\geq\beta\geq\gamma\geq 0,\alpha+\beta+\gamma=180^{\circ}\}。$$

设 A 为 U 上的一个近似等腰三角形，其隶属函数为

$$A(x)=A(\alpha,\beta,\gamma)=[1-\frac{1}{60}\min(\alpha-\beta,\beta-\gamma)]^{2}.$$

3.3 模糊模式识别

给定 4 个三角形 $x_1=(93,50,37), x_2=(100,45,35)$, $x_3=(125,38,17), x_4=(80,56,44)$，试用最大隶属原则识别这 4 个三角形中哪个优先归属于近似等腰三角形 A。

解：经计算得

$$A(x_1)\approx 0.614, A(x_2)\approx 0.694$$

$$A(x_3)\approx 0.423, A(x_4)\approx 0.64,$$

$$A(x_2)=\max\{A(x_1),A(x_2),A(x_3),A(x_4)\}$$

按最大隶属原则，x_2应优先归属于 A。

3.3 模糊模式识别

阈值原则:

设$A_1, A_2, \cdots, A_m$为给定论域U上的m个模糊模式，规定一个阈值$\lambda \in [0,1]$, $x_0 \in U$为一个待识别对象。

（1）如果$\max\{A_1(x_0), A_2(x_0), \cdots, A_m(x_0)\} < \lambda$,
则作“拒绝识别”的判决，这时应查找原因，再作分析。

（2）如果$\max\{A_1(x_0), A_2(x_0), \cdots, A_m(x_0)\} \geq \lambda$,
并且有k个模糊模式$A_{i_1}(x_0), A_{i_2}(x_0), \cdots, A_{i_k}(x_0)$大于或等于$\lambda$, 则认为识别可行, 并将$x_0$划归于$A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}$.

3.3 模糊模式识别

二、择近原则

1、贴近度

$\sigma(A,B)$ 表示两个模糊集A，B之间的贴近程度。

（1）格贴近度

$$\sigma_0(A,B)=\frac{1}{2}[A\circ B+(1-A\odot B],$$

其中：$A\circ B=\max\{A(x)\wedge B(x)\}$

表示两个模糊集 A，B 的内积；

$$A\odot B=\min\{A(x)\vee B(x)\}$$

表示两个模糊集 A，B 的外积。

3.3 模糊模式识别

例：设论域$U=\{x_1,x_2,x_3,x_4\}$上的三个模式为
$A=(0.9,0.1,0.6,0.3), B=(0,0.3,0.4,0.8),$
$C=(0.1,0.6,0.3,0.4)$,判别 A 和 B 中哪个与 C 最贴近。

解：$A\circ C=0.1\vee 0.1\vee 0.3\vee 0.3=0.3$

$$A\odot C=0.9\wedge 0.6\wedge 0.6\wedge 0.4=0.4$$

$$B\circ C=0\vee 0.1\vee 0.3\vee 0.4=0.4$$

$$B\odot C=0.1\wedge 0.6\wedge 0.4\wedge 0.8=0.1$$

$$\sigma_0(A,C)=\frac{1}{2}[0.3+(1-0.4)]=0.45$$

$$\sigma_0(B,C)=\frac{1}{2}[0.4+(1-0.1)]=0.65$$

故B比A更贴近于 C.

3.3 模糊模式识别

定义 (公理化定义)若　(A, B)满足

① $\sigma(A, A) = 1$;

② $\sigma(A, B) = \sigma(B, A)$;

③ 若有 $A \leqslant B \leqslant C$, 则 $\sigma(A, C) \leqslant \sigma(A, B) \wedge \sigma(B, C)$.

则称 $\sigma(A, B)$为 A 与 B 的贴近度.

3.3 模糊模式识别

（2）最小最大贴近度

$$\sigma_1(A,B)=\frac{\sum_{k=1}^{n}(A(x_k)\wedge B(x_k))}{\sum_{k=1}^{n}(A(x_k)\vee B(x_k))}$$

（3）最小平均贴近度

$$\sigma_2(A,B)=\frac{2\sum_{k=1}^{n}(A(x_k)\wedge B(x_k))}{\sum_{k=1}^{n}(A(x_k)+B(x_k))}$$

3.3 模糊模式识别

（4）海明贴近度

$$\sigma_3(A,B)=1-\frac{1}{n}\sum_{k=1}^{n}\left|A(x_k)-B(x_k)\right|$$

（5）欧几里得贴近度

$$\sigma_4=1-\frac{1}{\sqrt{n}}\left[\sum_{k=1}^{n}(A(x_k)-B(x_k))^2\right]^{\frac{1}{2}}$$

3.3 模糊模式识别

2、择近原则

设$A_1, A_2, \cdots, A_n$为论域 U 上的 n 个模糊模式，B 为 U 上的一个待识别对象，若

$$\sigma(B, A_i) = \max\{\sigma(B, A_1), \sigma(B, A_2), \cdots, \sigma(B, A_n)\},$$

则认为 B 应归属于模式A_i。

3.3 模糊模式识别

例：设$U=\{u_1,u_2,\cdots,u_6\}$，$A_1,A_2,\cdots,A_6$为 U 上的 6 个模糊模式，且

$$A_1=(1,0.8,0.5,0.4,0,0.1), A_2=(0.5,0.1,0.8,1,0.6,0)$$
$$A_3=(0,1,0.2,0.7,0.5,0.8), A_4=(0.4,0,1,0.9,0.6,0.5)$$
$$A_5=(0.8,0.2,0,0.5,1,0.7), A_6=(0.5,0.7,0.8,0,0.5,1)$$

现给定一个待识别对象$B=(0.7,0.2,0.1,0.4,1,0.8)$，试判断 B 应归属哪一个模式比较合理。

3.3 模糊模式识别

解：采用最大最小贴近度公式计算 B 与 A_i 的贴近度如下：

$$\sigma(B, A_1) = 0.3333, \sigma(B, A_2) = 0.3778$$

$$\sigma(B, A_3) = 0.4545, \sigma(B, A_4) = 0.4348$$

$$\sigma(B, A_5) = 0.8824, \sigma(B, A_6) = 0.4565$$

由于 $\sigma(B, A_5) = \max\{\sigma(B, A_i) \mid i = 1, 2, \cdots, 6\}$

故由择近原则，B 应归属于模式 A_5.

3.4 判别分析

判别分析方法最初应用于考古学, 例如要根据挖掘出来的人头盖骨的各种指标来判别其性别年龄等. 近年来, 在生物学分类, 医疗诊断, 地质找矿, 石油钻探, 天气预报等许多领域, 判别分析方法已经成为一种有效的统计推断方法。

判别分析是一种在一些已知研究对象用某种方法已经分成若干类的情况下，确定新的样品的观测数据属于哪一类的统计分析方法。

肝病的判别

地震的判别

3.4 判别分析

为了能识别待判断的对象$\boldsymbol{x} = (x_1, x_2,\ldots, x_m)^T$是属于已知类$A_1, A_2,\ldots, A_r$中的哪一类？

事先必须要有一个一般规则, 一旦知道了x的值, 便能根据这个规则立即作出判断, 称这样的一个规则为**判别规则**(用于衡量待判对象与各已知类别接近程度的方法准则)。

判别规则往往通过的某个函数来表达, 我们把它称为**判别函数**, 记作$W(i; x)$.

常用的方法有：距离判别法、**Fisher**判别法、贝叶斯判别法、逐步判别法。这里仅介绍后两种。

3.4 判别分析

Bayes判别法

Bayes判别法的基本思想：总是假设对所研究的对象已有一定的认识，计算新给样品属于各总体的条件概率$P(G_i|x_0)$, $(i=1,\ldots k)$,比较这个概率的大小，然后将新样品判归为来自概率最大的总体。

3.4 判别分析

设有总体 $G_i(i=1,2,\cdots,k)$, G_i 具有概率密度函数 $f_i(x)$。并且根据以往的统计分析，知道 G_i 出现的概率为 q_i 。即当样本 x_0 发生时，求他属于某类的概率。由贝叶斯公式计算后验概率，有：

$$P(G_i \mid x_0)=\frac{q_i f_i(x_0)}{\Sigma q_j f_j(x_0)}$$

判别规则 $$P(G_h \mid x_0)=\max_{1\le i\le k} P(G_i \mid x_0)$$

则 x_0判给 G_h。

3.4 判别分析

Bayes判别法的一般步骤：

1.计算各类中变量的均值 $\bar{x}_j$ 及均值向量 $x_h(h=1,2,..k)$，各变量的总均值 $x_j(j=1,2..p)$及均值向量 x;

2.计算类内协方差矩阵S及其逆矩阵S^{-1}；

3.计算Bayes判别函数中，各个变量的系数及常数项并写出判别函数;

4.计算类内协方差矩阵W及总各协方差矩阵T作多个变量的全体判别效果的检验;

3.各个变量的判别能力的检验;

6.判别新样本应属于的类别。

3.4 判别分析

例题：人文发展指数是联合国开发计划署于1990年5月发表的一份<<人类发展报告>>中公布的数据如下，试通过已知的样品建立判别函数,误判率是多少?并判断待判的归类.

3.4 判别分析

类别	国家	寿命(X1)	成人识字率%(X2)	调整后GDP(X3)
1	美国	76	99	5374
1	日本	79.5	99	5359
1	瑞士	78	99	5372
1	阿根廷	73.2	93.9	5242
1	阿联酋	73.8	77.7	5370
2	保加利亚	71.2	93	4250
2	古巴	73.3	94.9	3412
2	巴拉圭	70	91.2	3390
2	格鲁吉亚	72.8	99	2300
2	南非	62.9	80.6	3799

待判样品:				
	中国	68.5	79.3	1950
	罗马尼亚	69.9	96.9	2840
	希腊	77.6	93.8	5233
	哥伦比亚	69.3	90.3	5159

3.4 判别分析

```
data ex;input g  x1-x3 @@;
cards;
1 76  99    5374  1  79.5   99    5359  1  78  99     5372   1  73.2   93.9    5242  1
73.8   77.7   53702    71.2    93   4250  2   73.3  94.9     3412 2   70    91.2   3390 2
72.8   99     2300 2    62.9   80.6    3799
;
data ex1; input x1-x3 @@;
cards;
68.5   79.3    1950
69.9   96.9    2840
77.6   93.8    5233
69.3   90.3    5159
;
proc discrim data=ex testdata=ex1
anova manova simple list testout=ex2;
class g; proc print data=ex2;run;
```

3.4 判别分析

Proc Discrim后的常用选择项有：

（1）**Data**=数据集名，指定输入数据集名，若缺省则指定最新建立的数据集。

（2）**Testdata**=数据集名，指定待作出判别的数据集名，其中的变量名须上**Data**数据集中的变量名一致。

（3）**Testout**=数据集名，指定输出数据集，输出**Testdata**数据集中所有观测值以及每个观测值的后验概率和判别后的类别。

（4）**List**，指定打印每个观测值的回代结果。

（5）**Anova**，指定输出各类均值检验的一元统计量。

（6）**Manova**，指定输出各类均值检验的多元统计量。

（7）**Simple**，指定打印总体和组内的简单统计量。

3.4 判别分析

Linear Discriminant Function for g

Variable	1	2
Constant	-323.21568	-236.03823
x1	5.79107	5.14034
x2	0.26498	0.25167
x3	0.03407	0.02533

因此**Bayes**判别函数为

y1=-323.21568+3.79107x1+0.26498x2+0.03407x3

y2=-236.03823+3.14034x1+0.25167x2+0.02533x3

3.4 判别分析

Error Count Estimates for g

	1	2	Total
Rate	0.0000	0.0000	0.0000
Priors	0.5000	0.5000	

从上面运行结果得知，两类的误判率均为**0**

The SAS System

Obs	x1	x2	x3	_1	_2	_INTO_
1	68.5	79.3	1950	0.00000	1.00000	2
2	69.9	96.9	2840	0.00000	1.00000	2
3	77.6	93.8	5233	0.99997	0.00003	1
4	69.3	90.3	5159	0.98524	0.01476	1

因而得知中国与罗马尼亚归入第二类，希腊与哥伦比亚归入第一类。