第三章　聚类分析

[教学目的与要求]

1. 了解适合用聚类分析解决的问题；
2. 理解对象之间的相似性如何测量的；
3. 区别不同的距离；
4. 区分不同的聚类方法及其相应的应用；
5. 理解如何选择类的个数；
6. 了解聚类分析的局限。

[教学重点和难点]

1. 理解对象之间的相似性如何测量的。
2. 理解如何选择类的个数。

[教学过程]

第一节　聚类分析方法

聚类分析是根据“物以类聚”的道理，对样品或指标进行分类的一种多元统计分析方法，它们讨论的对象是大量的样品，要求能合理地按各自的特性来进行合理的分类，没有任何模式可供参考或依循，即是在没有先验知识的情况下进行的。

基本思想是根据事物本身的特性研究个体分类的方法；聚类原则是同一类中的个体有较大的相似性，不同类中的个体差异很大。

基本程序：是根据一批样品的多个观测指标，具体地找出一些能够度量样品或指标之间相似程度的统计量，然后利用统计量将样品或指标进行归类。

具体进行聚类时，由于目的、要求不同，因而产生各种不同的聚类方法：

由小类合并到大类的方法

由大类分解为小类的方法

静态聚类法、动态聚类法

按样本聚类（Q）、按指标聚类（R）

在社会经济领域中存在着大量分类问题，如：

对我国31个省市自治区独立核算工业企业经济效益进行分析，一般不是逐省市自治区去分析，而较好地做法是选取能反映企业经济效益的代表性指标，如百元固定资产实现利税、资金利税、产值利税率等，根据这些指标对全国各省市自

治区进行分类，然后根据分类结果对企业经济效益进行综合评价，就易于得出科学的分析。

第二节　聚类统计量

概述

设有n个样本单位，每个样本测得p项指标（变量），原始资料阵为：

Q型聚类以距离作为统计量，R型聚类以相似系数作为统计量。

Q型聚类统计量（距离）

把n个样本点看成p维空间的n个点

1、绝对距离（Block距离）

$$d_{ij}(1)=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|$$

2、欧氏距离(Euclidean distance)

$$d_{ij}(2)=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2\right]^{1/2}$$

3、明考斯基距离(Minkowski)

$$d_{ij}=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^q\right]^{1/q}$$

4、兰氏距离

$$d_{ij}(L)=\sum_{k=1}^{p}\frac{\left|x_{ik}-x_{jk}\right|}{x_{ik}+x_{jk}}$$

5、马氏距离

6、切比雪夫距离(Chebychev)

R型聚类统计量

对两个指标之间的相似程度用相似系数来刻划，相似系数的绝对值越接近于1，表示指标间的关系越密切，绝对值越接近于0，表示指标间的关系越疏远。

1、夹角余弦

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

2、相关系数

$$r_{ij}=\frac{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)^2\sum_{k=1}^{n}(x_{kj}-\bar{x}_j)^2}}$$

3、同号率

$$C_{ij}=\frac{n_+-n_-}{n_++n_-}$$

第三节　无量纲化方法

所谓无量纲化处理，是将原始数据矩阵中每个元素按照某种特定的运算把它变成一个新值，且是数值的变化不依赖于原始数据中其它数据的新值。

1、极差正规化（规格化变换、阈值法）

$$x'_{ij}=\frac{x_{ij}-\min(x_{ij})}{\max(x_{ij})-\min(x_{ij})}$$

2、标准化变换

3、功效系数法

$$x'_i=\frac{x_i-x_{si}}{x_{hi}-x_{si}}\times 40+60$$

4、相对化变换

例：某年我国部分省市经济效益情况

指标	实际值				
	北京	天津	上海	江苏	广东
产品销售率（%）	96.01	95.72	98.42	93.43	95.16
资金利税率（%）	14.9	9.21	13.88	10.75	10.25
成本利润率（%）	9.51	3.35	7.55	3.99	5.03
劳动生产率（元/人）	14830	10004	15545	9708	14590
流动资金周转次数（次）	1.68	1.79	1.8	2.21	1.87
净资产率（%）	28.4	26.48	25.56	22.3	25.01

用以上几种方法对其无量纲化。

第四节　Q型系统聚类法

系统聚类法（层次聚类法）：在聚类分析的开始，每个样本自成一类；然后，按照某种方法度量所有样本之间的亲疏程度，并把最相似的样本首先聚成一小类；接下来，度量剩余的样本和小类间的亲疏程度，并将当前最接近的样本或小类再聚成一类；再接下来，再度量剩余的样本和小类间的亲疏程度，并将当前最接近的样本或小类再聚成一类；如此反复，直到所有样本聚成一类为止。

步骤：

1、对数据进行变换处理，消除量纲

2、构造n个类，每个类只包含一个样本计算

3、n个样本两两间的距离$\{d_{ij}\}$

4、合并距离最近的两类为一新类

5、计算新类与当前各类的距离，重复（4）

6、画聚类图

7、决定类的个数和类

类与类间距离的确定

一、最短距离法

二、最长距离法

三、中间距离法

四、重心距离法

五、类平均法

六、离差平方和

最短距离法(Nearest Neighbor)

以当前某个样本与已经形成的小类中的各样本距离中的最小值作为当前样本与该小类之间的距离。

例 1：为了研究辽宁省 5 省区某年城镇居民生活消费的分布规律，根据调查资料做类型划分。

省份	x1	x2	x3	x4	x5	x6	x7	x8
辽宁	7.9	39.77	8.49	12.94	19.27	11.05	2.04	13.29
浙江	7.68	50.37	11.35	13.3	19.25	14.59	2.75	14.87
河南	9.42	27.93	8.2	8.14	16.17	9.42	1.55	9.76
甘肃	9.16	27.98	9.01	9.32	15.99	9.1	1.82	11.35
青海	10.06	28.64	10.52	10.05	16.18	8.39	1.96	10.81

G1={辽宁}，G2={浙江}，G3={河南}，G4={甘肃}，G5={青海}

$$d_{ij}=\left[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2\right]^{1/2}$$ =[(7.9-7.68)^2+(39.77-50.37)^2+(8.49-11.35)^2+(12.94-13.3)^2+(19.27-19.25)^2+(11.05-14.59)^2+(2.04-2.75)^2+(13.29-14.87)^2]$^{0.5}$=11.67

d_{13}=13.80　d_{14}=13.12　d_{15}=12.80　d_{23}=24.63　d_{24}=24.06　d_{25}=23.54　d_{34}=2.2　d_{35}=3.51　d_{45}=2.21

D_1=

	1	2	3	4	5
1	0				
2	11.67	0			
3	13.80	24.63	0		
4	13.12	24.06	2.20	0	

5　　12.80　　23.54　　3.51　　2.21　　0

河南与甘肃的距离最近，先将二者（3 和 4）合为一类 G6={G2，G4}

$d_{61}=d_{(3,4)1}=\min\{d_{13},d_{14}\}=13.12$　　$d_{62}=d_{(3,4)2}=\min\{d_{23},d_{24}\}=24.06$

$d_{65}=d_{(3,4)5}=\min\{d_{35},d_{45}\}=2.21$

$D_2=$

	6	1	2	5
6	0			
1	13.12	0		
2	24.06	11.67	0	
5	2.21	12.80	23.54	0

$d_{71}=d_{(3,4,5)1}=\min\{d_{13},d_{14},d_{15}\}=12.80$

$d_{72}=d_{(3,4,5)2}=\min\{d_{23},d_{24},d_{25}\}=23.54$

$D_3=$

	7	1	2
7	0		
1	12.80	0	
2	23.54	11.67	0

$d_{78}=\min\{d_{71},d_{72}\}=12.80$

$D_4=$

	7	8
7	0	
8	12.8	0

最长距离法(furthest neighbor)

以当前某个样本与已经形成的小类中的各样本距离中的最大值作为当前样本与该小类之间的距离。

例 2：对例 1 的数据以最长距离法聚类 $d_{13}=13.80$　$d_{14}=13.12$　$d_{15}=12.80$　$d_{23}=24.63$　$d_{24}=24.06$　$d_{25}=23.54$　$d_{34}=2.2$　$d_{35}=3.51$　$d_{45}=2.21$

	1	2	3	4	5
$D_1=$ 1	0				
2	11.67	0			
3	13.80	24.63	0		
4	13.12	24.06	2.20	0	
5	12.80	23.54	3.51	2.21	0

$d_{61}=d_{(3,4)1}=\max\{d_{13}, d_{14}\}=13.80$ $d_{62}=d_{(3,4)2}=\max\{d_{23}, d_{24}\}=24.63$

$d_{65}=d_{(3,4)5}=\max\{d_{35}, d_{45}\}=3.51$

	6	1	2	5
6	0			
$D_2=$ 1	13.80	0		
2	24.63	11.67	0	
5	3.51	12.80	23.54	0

$d_{71}=d_{(3,4,5)1}=\max\{d_{13}, d_{14}, d_{15}\}=13.80$

$d_{72}=d_{(3,4,5)2}=\max\{d_{23}, d_{24}, d_{25}\}=24.63$

	7	1	2
$D_3=$ 7	0		
1	13.80	0	
2	24.63	11.67	0

$d_{78}=\max\{d_{71}, d_{72}\}=24.63$

	7	8
$D_4=$ 7	0	
8	24.63	0

中位数法（Median clustering）

用两位类的中位数间的距离作为两类的距离

$$D_{tr}^2=\frac{1}{2}D_{lt}^2+\frac{1}{2}D_{tm}^2-\frac{1}{4}D_{lm}^2$$

重心法

用两类的重心间的距离作为两类的距离

$$D_{tr}^2=\frac{1}{2}D_{lt}^2+\frac{1}{2}D_{tm}^2-\frac{1}{4}D_{lm}^2$$

$$n_l+n_m=n_r$$

组间平均链锁法（Between-groups linkage）

定义两个小类之间的距离为所有样本对间的平均距离。

利用了所有样本对距离的信息。

$$D_{tr}^2=\frac{1}{n_l n_r}\sum_{i\in G_t}\sum_{i\in G_r}d_{ij}^2$$

$$=\frac{1}{n_l n_r}\sum_{i\in G_t}\sum_{i\in G_l}d_{ij}^2+\frac{1}{n_l n_r}\sum_{i\in G_t}\sum_{i\in G_m}d_{ij}^2$$

$$=\frac{n_l}{n_r}\frac{1}{n_t n_l}\sum_{i\in G_t}\sum_{i\in G_l}d_{ij}^2+\frac{n_m}{n_r}\frac{1}{n_t n_m}\sum_{i\in G_t}\sum_{i\in G_m}d_{ij}^2$$

$$=\frac{n_l}{n_r}D_{tl}^2+\frac{n_m}{n_r}D_{tm}^2$$

组内平均链锁法（Within-groups linkage）

对所有样本对的距离求平均值，包括小类之间的样本对、小类内的样本对

离差平方和法（ Ward’s method word）

使小类内各样本的**欧氏距离**总平方和增加最小的两小类合并为一类。

$$S_t=\sum_{i=1}^{n_t}(X_i(t)-\overline{X}(t))(X_i(t)-\overline{X}(t))'$$

$$S=\sum_{t=1}^{q}S_t=\sum_{t=1}^{q}\sum_{i=1}^{n_t}(X_i(t)-\overline{X}(t))(X_i(t)-\overline{X}(t))'$$

将q固定时，要选择使S达到极小的分类，一切可能的分法有：

$$R(n,q)=\frac{1}{q!}\sum_{i=0}^{q}(-1)^{q-1}C_q^i i^n$$

Ward 寻找到一个局部最优解的方法。

先将n个样本各成一类，然后每次缩小一类，每缩小一类离差平方和就要增大，选择使离差平方和S增加最小的两类合并，直至所有样本归为一类为止。

$$D_{tr}^2=\frac{n_t+n_l}{n_r+n_t}D_{tl}^2+\frac{n_t+n_m}{n_r+n_t}D_{tm}^2-\frac{n_t}{n_r+n_t}D_{lm}^2$$

第五节　R型系统聚类法

一、最小系数法

二、最大系数法

三、中间系数法

对变量聚类，是一种降维的方法，用于在变量众多时寻找有代表性的变量，以便当用少量、有代表性的变量代替大变量时损失信息很少。

第六节　快速聚类

如果选择了N个数值型变量参与聚类分析，最后要求聚类数K，那么可以由系统首先选择K个观测量作为聚类的种子，也称初始类中心、凝聚点，按照距这几个类中心的距离最小原则把观测量分到各类中心所在的类中去，形成第一次迭代形成的K类。根据组成每一类的观测量计算各变量均值，每一类中的n个均值在N维空间中又形成K个点，这就是第二次迭代的类中心，按照这种方法依次迭代下去直到分类比较合理为止。

凝聚点的选择

1、经验选择

2、对样本人为或随机分类，以每类的重心作为凝聚点

3、最小最大距离法。如果欲将n个样本点分为q类，先选取距离最大的两点x_{i1}，x_{i2}为前两个凝聚点，然后选取第3个凝聚点x_{i3}，由于其余所有点与前两个凝聚点都有最短距离，在全部最短距离中选择最长距离，这个距离的两端一个是x_{i1}或x_{i2}，而另一个就是我们要选择的x_{i3}.

4、密度法

初始分类

1、人为地分类

2、选择凝聚点后，将与其最近的凝聚点归并

3、选择凝聚点后，每个凝聚点自成一类，将样本依次归入其距离最近的凝聚点那一类，并立即计算该类的重心，以代替原来的凝聚点，再计算下一个样本的归类。

4、先对样本数据标准化，然后计算统计量

快速聚类步骤

1、选择分析变量

2、指定聚类数目

3、选择k个样本作为凝聚点

4、按照距初始类中心最小的原则将各观察量分到聚类中心所在的类中去，形成第一步迭代的k类

5、计算每类中所有变量的均值，作为第二次迭代的中心

6、重复3、4步，直至指定的迭代次数或达到终止的条件