聚类分析

《美国数学建模竞赛》

完整课程请长按下方二维码

聚类分析

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。

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) 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。

聚类分析的一般步骤(Q-型分类)

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

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。

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。

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	

```
data ex;input x1 x2 factory$@@; /*$：表示字符型变量*/
cards;
/*数据省略*/
;
proc cluster   /*系统聚类*/
data=ex method=ward ccc pseudo outtree=tree;
id factory;
run;
proc tree data=tree horizontal;    /*水平树*/
id factory;   /*工厂为样本*/
run;
```

ccc表示要计算半偏R^2，R^2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t^2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R^2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。 R^2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。

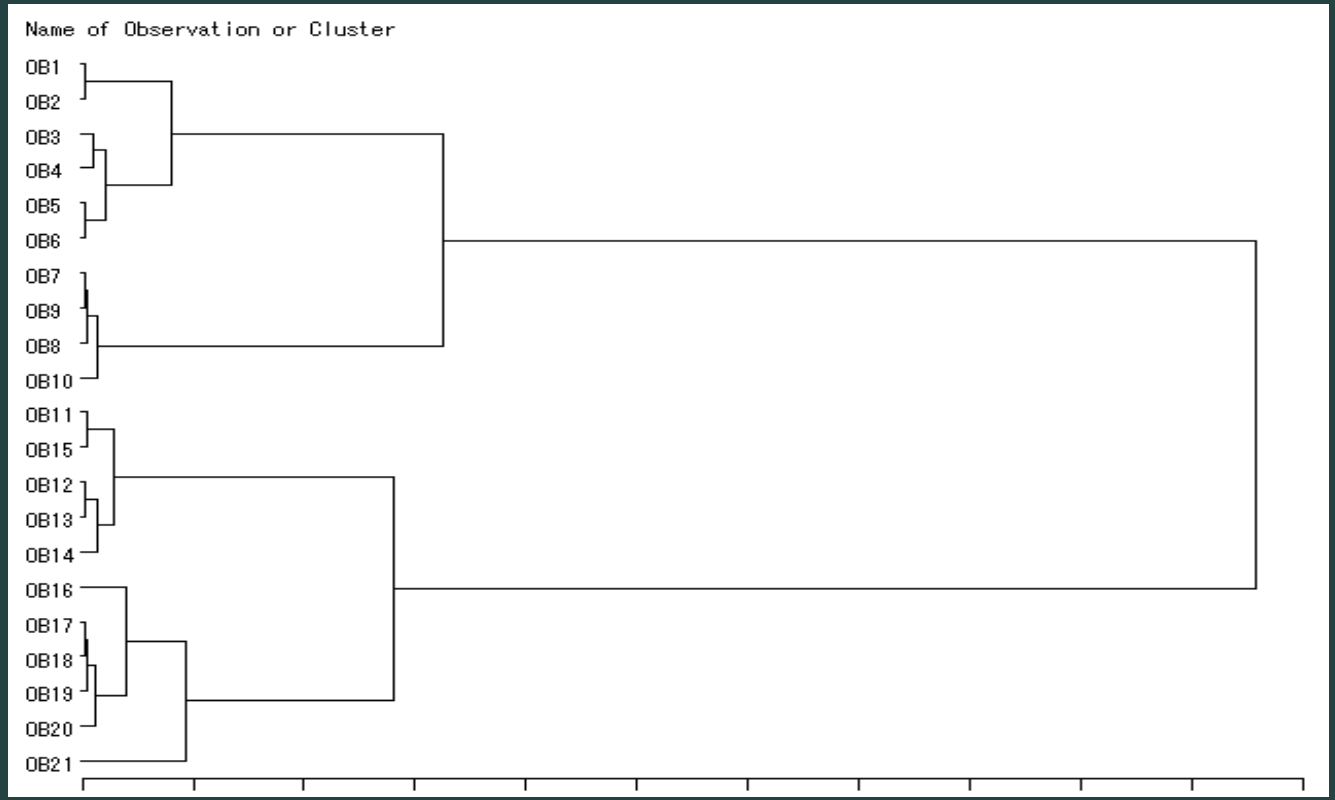
Pseudo说明要计算伪F和伪t^2统计量。一般认为，伪F统计量出现峰值时的所对应的分类是较佳的分类选择。当把数据从G+1类合并为G类时，伪t^2统计量的值大，说明不应该合并这两类。

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	

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表示类与类间最短规格化最短距离，RSQ表示R^2统计量，ERSQ表示半偏R^2统计量，CCC统计量值。PSF为伪F统计量，PST2为伪t^2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t^2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。

动态聚类图



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。