

• 讲座 •

医学统计学的基本方法(七)

潘子昂 王石麟 李占魁

七 聚类分析

运用聚类分析的前提是存在有大的数据集,这些数据就表面来看,它们之间看不出有太多的相互联系,但是俗话说得好:“物以类聚”,所以这些实验数据作为事物特征的测量值也会有着必然的相互联系,可谓是以类而聚的。依据实验数据本身所具有的定性或定量的特征来对大量的数据进行分组归类,以了解数据集的内在结构,这种数据分析方法称为聚类分析。

聚类分析这种研究事物分类的方法,既可对样本,也可对所观察的指标进行分类。这种分析方法在许多科学领域,尤其是行为科学、生物分类学以及医学流行病学调查中普遍采用的一种分析方法。

1 聚类分析中基本思路

1.1 聚类分析目的

聚类分析的目的在于解决被分析数据(事物特征)的分组或归类,以了解样本或指标间内在的联系规律。

1.2 相似性的量度

要对数据进行分类,其根据是数据之间有无相似性,如相似就可以归为一类,不相似则归于不同类中。

怎样来认定数据之间有无相似性呢?这里存在一个相似性的量度问题。在聚类分析中,一般采用相关系数、距离或信息量来进行量度,其中常用的为相关系数和距离。

1.3 聚类分析的一般步骤

(1) 对实验数据进行整理,列出数据矩阵(数据集)。

(2) 由于对指标测量所用的物理量不同或指标测量值的数量级存在大的差异,在计算和

分析时会造成误差,故需对数据进行规范化处理,化成无量纲的数据。

(3) 计算样本或指标之间的相关系数或距离,并按相似性大小逐步进行归类。

(4) 重复计算和归类,直至将分析对象归为一类。

(5) 画出聚类图。

(6) 对实验和聚类分析结果作必要解释。

1.4 聚类分析方法

先把所要分析的 n 个对象看成 n 个类,再逐步按照之间的相似性大小进行合并归类,最终归为一类,这种方法叫系统聚类法或谱系聚类分析,其结果构成一个谱系图或聚类图。还有一种叫做分部聚类方法,它是将分析对象最终划分为 k 个类,使得同一类中彼此相似,且不同类能得到很好分离。由于 k 个类是同时产生的,所以所得的类是非谱系的,这种方法又叫非谱系聚类方法。

2 聚类分析中数学方法

2.1 数据矩阵(集)

用 n 个指标或特征值对 m 个待分析的样本进行描述,构成 $m \times n$ 的数据集即数据矩阵 $X_{m \times n}$,矩阵的元素为 x_{ij} (代表第 i 样品的第 j 个指标),数据矩阵表示为:

$$\begin{array}{c} \text{样本} \\ \begin{matrix} 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ m \end{matrix} \end{array} \begin{array}{c} \text{指标} \\ \left[\begin{array}{cccccc} 1 & 2 & \cdots & j & \cdots & n \\ x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1n} \\ x_{21} & & & \vdots & & \vdots \\ \vdots & & & & & \\ x_{i1} & & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & & & \vdots & & \vdots \\ x_{m1} & & \cdots & x_{mj} & \cdots & x_{mn} \end{array} \right] \end{array}$$

这一数据集也可表示为:

$$\begin{array}{c}
 \text{样本} \\
 \begin{array}{c} 1 \\ 2 \\ \vdots \\ j \\ \vdots \\ n \end{array}
 \begin{array}{c}
 \left[\begin{array}{cccccc}
 1 & 2 & \cdots & j & \cdots & m \\
 x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1m} \\
 x_{21} & & & \vdots & & \vdots \\
 \vdots & & & \vdots & & \vdots \\
 x_{j1} & & \cdots & x_{jj} & \cdots & x_{jm} \\
 \vdots & & & \vdots & & \vdots \\
 x_{n1} & & \cdots & x_{nj} & \cdots & x_{nm}
 \end{array} \right]
 \end{array}
 \end{array}
 \begin{array}{c}
 \text{指标} \\
 \end{array}$$

这一矩阵实际上是第一个矩阵的转置矩阵。

2.2 数据规范化处理

将数据矩阵中元素化为无量纲的数,在统计上称为Z变换。原来的数值减去平均值再除以标准差,组成矩阵Z。

$$Z_{ij} = \frac{x_{ij} - \bar{x}_i}{S_i}$$

其中

$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ij}$$

$$S_i = \sqrt{\frac{1}{n-1} \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2}$$

2.3 相似性矩阵

在相关系数和距离之中选取一种作为相似

性量度,来计算样本或指标的两两之间的相似性值,构成对称的相似性矩阵。

相关系数的计算可按如下公式进行,

$$r = \frac{l_{ij}}{\sqrt{l_{ii}l_{jj}}}$$

$$\text{其中 } l_{ij} = \sum_{k=1}^n (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$$

$$l_{ii} = \sum_{k=1}^n (x_{ik} - \bar{x}_i)^2$$

$$l_{jj} = \sum_{k=1}^n (x_{jk} - \bar{x}_j)^2$$

相关系数越大,相似性就越大,可归为一类。

距离的定义有多种,用得最多的是欧氏距离,即平方根距离,对象*i*、*j*之间的距离计算公式如下:

$$d_{ij} = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}$$

d_{ij} 越小,性质就越接近。

3 应用举例

3.1 数据矩阵

下面的数据是用14项不同指标对10个样本进行测量所得的结果,见表1。

表1 实验数据

指标	样 本										$\bar{x}$	s
	1	2	3	4	5	6	7	8	9	10		
1	5.14	5.34	4.51	1.88	4.27	6.55	8.57	6.58	6.48	8.58	5.79	2.03
2	2.35	2.52	2.21	1.16	0.99	0.93	0.54	0.77	0.56	3.43	1.55	1.00
3	2.53	2.34	2.48	0.88	1.69	2.82	3.29	2.62	2.85	3.71	2.52	0.79
4	0.50	0.50	0.53	0.80	1.05	0.43	0.62	0.84	0.63	2.89	0.68	0.20
5	3.02	5.28	3.69	3.62	3.16	2.69	2.93	3.24	2.70	2.58	3.31	0.80
6	2.23	1.84	1.59	0.83	2.44	4.01	4.53	2.85	2.93	3.39	2.66	1.12
7	2.86	2.26	2.48	1.22	2.63	3.87	3.71	2.77	3.25	3.86	2.89	0.83
8	9.54	9.02	6.31	9.71	10.30	8.06	6.64	8.38	7.59	8.45	8.40	1.30
9	17.53	17.55	21.32	21.20	18.41	15.44	14.12	18.45	15.40	14.11	17.35	2.61
10	10.32	11.25	13.06	14.90	9.29	10.53	13.75	11.30	14.03	9.37	11.78	2.01
11	2.13	1.92	1.86	1.08	2.20	3.84	3.71	2.10	2.14	2.73	2.37	0.84
12	19.22	17.50	13.03	18.61	19.14	16.82	17.12	16.75	17.36	16.41	17.20	1.77
13	20.01	19.68	24.46	21.57	22.34	17.07	15.31	20.02	20.20	19.27	19.99	2.56
14	1.49	1.76	1.23	1.65	1.16	1.27	0.98	1.13	1.51	1.29	1.35	0.25

现在需要对14项指标进行聚类分析。

3.2 数据规范化

由于指标间数值差别较大,故必须对数据

进行规范化处理,即进行Z变换。数字大小并不完全说明此项指标重要性。实验数据规范化值见表2。

表 2 实验数据的规范化值

指标	样 本									
	1	2	3	4	5	6	7	8	9	10
1	-0.32	-0.22	-0.63	-1.93	-0.75	0.37	1.37	0.39	0.34	1.37
2	0.80	0.97	0.66	-0.39	-0.56	-0.62	-1.01	-0.78	-0.98	1.88
3	0.01	-0.23	-0.05	-2.08	-1.05	0.38	0.97	0.13	0.42	1.51
4	-0.90	-0.90	-0.75	0.59	1.85	-1.25	-0.29	0.79	-0.24	1.05
5	-0.36	2.45	0.47	0.67	-0.19	-0.79	-0.47	-0.09	-0.76	-1.91
6	-0.39	-0.73	-0.96	-1.63	-0.20	1.21	1.67	0.17	0.24	0.45
7	-0.04	-0.16	-0.49	-2.01	-0.31	1.18	0.99	-0.14	0.43	1.17
8	0.88	0.48	-1.61	1.01	1.46	-0.26	-1.35	-0.02	-0.62	0.04
9	0.07	0.08	1.52	1.48	0.41	-0.73	-1.24	0.42	-0.75	-1.24
10	-0.73	-0.36	0.64	1.55	-1.24	-0.62	0.98	-0.24	1.12	-1.20
11	-0.29	-0.54	-0.61	-1.54	-0.20	1.75	1.60	-0.32	-0.27	0.43
12	1.14	0.17	-2.36	0.80	1.10	-0.21	-0.05	-0.25	0.09	-0.45
13	0.01	-0.12	1.75	0.62	0.92	-1.14	1.83	0.01	0.08	-0.28
14	0.56	1.84	-0.48	1.20	-0.76	-0.32	-1.48	-0.88	0.64	-0.24

3.3 相似性矩阵

是一个对称矩阵,即 $x_{ij}=x_{ji}$,故只列出一半数值。

选取指标间的相关系数作为相似性值,计算并取其绝对值,得到相似性矩阵,见表 3。这

表 3 相似性矩阵

1	2	3	4	5	6	7	8	9	10	11	12	13	13	14
1	1													
2	0.07	1												
3	0.96	0.25	1											
4	0.09	0.07	0.23	1										
5	0.48	0.26	0.46	0.21	1									
6	0.88	0.27	0.77	0.09	0.62	1								
7	0.92	0.00	0.90	0.16	0.68	0.92	1							
8	0.50	0.12	0.58	0.46	0.19	0.42	0.44	1						
9	0.88	0.01	0.80	0.09	0.50	0.87	0.87	0.22	1					
10	0.27	0.46	0.28	0.24	0.16	0.23	0.39	0.44	0.27	1				
11	0.76	0.20	0.69	0.27	0.52	0.94	0.88	0.42	0.78	0.27	1			
12	0.23	0.22	0.39	0.29	0.03	0.05	0.19	0.81	0.14	0.19	0.11	1		
13	0.69	0.25	0.54	0.27	0.30	0.81	0.64	0.20	0.81	0.03	0.79	0.24	1	
14	0.52	0.33	0.43	0.28	0.59	0.65	0.56	0.43	0.27	0.18	0.60	0.31	0.25	1

3.4 逐步归类

找出最大的相关系数,将与相关系数有关的两个指标归为一类。从表 3 中不难看出,最大的相关系数为 $r_{1,3}=0.96$,所以把指标 1 与 3 归为一类,记作 1.3。其它指标与这新的指标间的相关系数,不应小于归类前各自之间的相关系数,故取它们之间的最小值。例如指标 4 与 1.3

之间的相关系数为:

$$r_{1.3,4}=\min(r_{1,4},r_{3,4})=\min(0.09,0.23)=0.09$$

按比法列出第一次归类后的相似矩阵,见表 4。

在一次归类后的相似矩阵基础上,再找出最大的相关系数,进行二次归类,并列出二次归类后的相似矩阵,以此类推,直至把 14 项指数归为一类。

表 4 一次归类后相似性矩阵

1.3	2	4	5	6	7	8	9	10	11	12	13	13	14
1.3	1												
2	0.07	1											
4	0.09	0.07	1										
5	0.46	0.26	0.21	1									
6	0.77	0.27	0.09	0.62	1								
7	0.90	0.00	0.16	0.68	0.92	1							
8	0.50	0.12	0.46	0.19	0.42	0.44	1						
9	0.80	0.01	0.09	0.50	0.87	0.87	0.22	1					
10	0.27	0.46	0.24	0.16	0.23	0.39	0.44	0.27	1				
11	0.69	0.20	0.27	0.52	0.44	0.88	0.42	0.78	0.27	1			
12	0.23	0.22	0.29	0.03	0.05	0.19	0.81	0.14	0.19	0.11	1		
13	0.54	0.25	0.27	0.30	0.81	0.64	0.20	0.81	0.03	0.79	0.24	1	
14	0.43	0.31	0.28	0.59	0.65	0.56	0.43	0.27	0.18	0.60	0.31	0.25	1

3.5 聚类图

聚类分析的最终结果可以聚类图表示(图1)。图中可看到逐次归类结果。

当用不同大小相关系数去截割聚类图,就得出不同结果。取相关系数为0.7时,可以看到14项指标分成8个亚群。当取0.5时,则分为6个亚群。同一亚群中,它们之间的性质相近。

参 考 文 献

- 1 D.L. 马萨特, L. 考夫曼著, 刘昆元译, 聚类分析法 解析分析化学数据, 第一版, 北京: 化学工业出版社, 1990年1月。
- 2 金丕焕主编, 医用统计方法, 第一版, 上海: 上海医科大学出版社, 1993, 2月: 290~298。

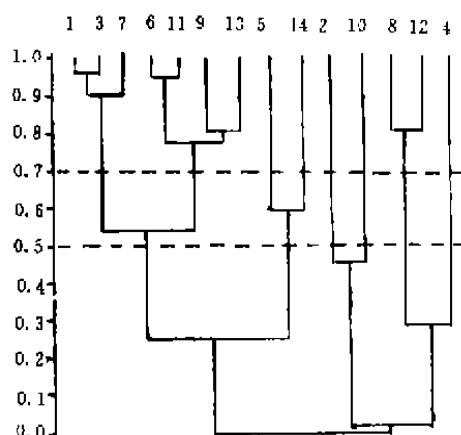


图 1 聚类图

第一届全国老年高血压研讨会征文征集

由中国老年学学会老年医学委员会和中国骨质疏松杂志社主办的第一届全国老年高血压研讨会决定在1997年6月下旬在黄山市举行。征文内容:有关老年高血压的基础研究及其并发症诊断治疗和进展,治疗的特殊性,药物配伍,合并症和并发症的防治;中西医结合及心理治疗、护理经验;国内外抗高血压药物研制情况;流行病学调查及人群防治等相关文章及产品。征文要求:按论文格式书写,全文不超过3200字和摘要400字各一份;来稿是未公开发表过的,要求打印并加盖单位公章。论文将由国家出版社正式出版。

截稿日期:1997年3月30日(邮戳为准)。

征稿日期:100071北京市丰台东大街5号-400 刘维平收。

欢迎厂家参展! 联系电话:(010)64976421 或(010)64219966 转 313 中国老年学学会老年医学委员会, 金信爱大夫

中国老年学学会老年医学委员会
中国骨质疏松杂志社
1997年2月18日