主成分分析应用(II)——主成分聚类分析

胡良平[1,2*]

(1. 军事科学院研究生院,北京 100850;
2. 世界中医药学会联合会临床科研统计学专业委员会,北京 100029
*通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是介绍主成分聚类分析的概念、作用以及用软件实现计算的方法。无序样品聚类分析方法很多,本文仅介绍基于主成分变量的无序样品聚类分析方法。由于缺乏评价聚类效果的金标准,本文所介绍的聚类方法的实用价值有待进一步研究。

【关键词】 无序样品;样品聚类分析;主成分变量;主成分聚类分析
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.02.009

Application of the principal components analysis(II) —— the principal components cluster analysis

Hu Liangping[1,2*]

(1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
** Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【**Abstract**】 The aim of this paper was to introduce the concepts, functions and the calculation methods by using the statistical software of the principal components cluster analysis. Although there are many cluster analysis methods for the cluster analysis of no ordered samples, only the cluster method based on the principal components variables was introduced in this paper. Due to the lack of the gold standard for assessing the cluster effect, it needed further research on how much applied value of this method introduced in this paper.

【**Keywords**】 No ordered sample; Sample cluster analysis; Principal components variable; Principal components cluster analysis

1 概 述

1.1 基本概念

本期《基于 SAS 与 R 软件的主成分分析》一文介绍了“主成分分析方法”。此法不仅可以借用于多重线性回归分析(见本期《主成分分析应用(Ⅰ)——主成分回归分析》)之中,还可以借用于无序样品聚类分析之中。

1.2 何为主成分聚类分析

主成分聚类分析(the principal components Cluster analysis)是对拟用于无序样品聚类分析的定量变量先进行主成分分析,产生主成分变量,然后再基于这些主成分变量(注意:不是原变量)进行无序样品聚类分析。

1.3 主成分聚类分析的作用

通过主成分分析,将原先可能具有一定相关性的定量指标转变为相互独立的变量,期望能够更好地呈现不同样品之间的相对距离,从而更好地对无序样品实现精准聚类。

1.4 适合进行主成分聚类分析的数据结构

【例 1】沿用本期《基于 SAS 与 R 软件的主成分分析》一文中的“例 1 和表 1”[1],此处从略。此资料属于“单组设计多元定量资料”,假定资料具有“同质性”。

1.5 无序样品聚类分析的种类

1.5.1 概述

对于具有同质性的单组设计多元定量资料,若分析目的是希望将全部样品或个体按其“亲疏关系”聚成不同的类,被聚在同一类中的样品或个体被认为是“最接近的”或“最相似的”。这件事本身应属于“无序样品聚类分析问题”,有时也被称为“综合评价问题”。为了实现这一分析目的,可以基于多种不同的统计思想或思路来构造分析方法,通常有如下两大类:基于“距离”的聚类分析法和基于“综合评价指标”的聚类分析法。

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

1.5.2 基于“距离”的聚类分析法

所谓基于样品间“距离”大小来构造无序样品聚类分析法，就是把每个样品视为空间中的一个“点”，计算出任何两点之间的距离，再根据距离数值的大小，将距离最小且相邻的那些点聚在同一类中。此类分析方法可以进一步划分为两类：经典统计学中的无序样品聚类分析法和机器学习统计学中的无序样品聚类分析法。前者又可细分为 K－means聚类法、PAM 聚类法、层次聚类法和 EM 聚类法等[2-3]；后者相对较少，通常叫做自组织映射神经网络分析法，简称 SOM 方法[2-3]。

1.5.3 基于“综合评价指标”的聚类分析法

所谓基于“综合评价指标”的聚类分析法就是基于多项原始定量指标计算出一个“综合评价指标”来，计算出每个样品在综合评价指标上的取值，再按由小到大或由大到小进行排序（使无序样品变成了有序样品），进而基于某种规则将全部“有序样品”分为所需要的几档或几组。

1.5.4 以上两大类聚类分析方法的适用场合

当大部分或全部定量变量的取值大小在专业上既不是“高优指标（即指标的取值越大越好，例如疾病的治愈率）”，也不是“低优指标（即指标的取值越小越好，例如疾病的死亡率和复发率）”时，采用“基于距离的聚类分析方法”为宜；反之，采用“基于综合评价指标的聚类分析方法”为宜。

因篇幅所限，本文仅介绍“基于主成分变量和距离实现无序样品聚类分析”的方法。

2 主成分聚类分析的实现

2.1 基于主成分变量进行无序样品聚类

将表 1 中的 23 行 9 列数据按文本格式存储在“F:\CCC”文件夹中，命名为“23 种肿瘤类期刊文献计量学指标资料.txt”；设所需要的 SAS 程序名为“基于肿瘤类期刊文献计量学指标进行主成分聚类分析.SAS”：

```
data a1;
  infile 'F:\CCC\23 种肿瘤类期刊文献计量学指标资料.txt';
  input name $20. x1 - x8;
run;
proc princomp data = a1 out = b1 prefix = z;
  var x1 - x8;
run;
data a2;
  set b1;
  id = _n_;
ods graphics on;
proc cluster data = a2 method = war std simple ccc pseudo outtree = cluster;
  var z1 - z8;
  id id;
  copy name;
run;
proc tree horizontal;
  id name;
run;
ods graphics off;
```

【SAS 主要输出结果】

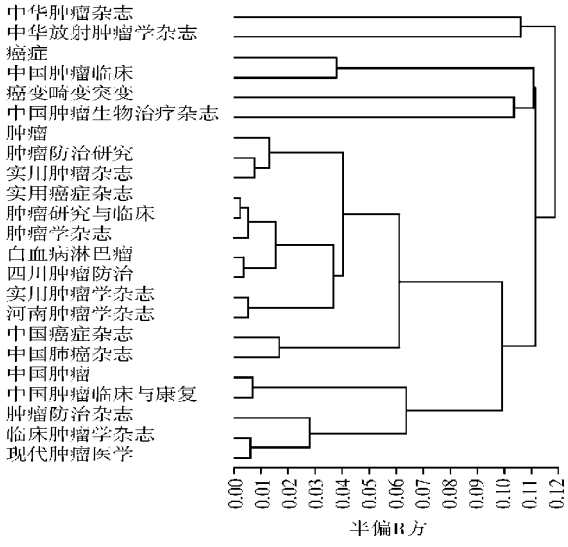

图 1 以树状图形式呈现的无序样品聚类结果 1

由图 1 可看出，若希望分成两类，则自上而下为：“中华肿瘤杂志”到“中国肿瘤生物治疗杂志”算作第一类，“肿瘤”到“现代肿瘤医学”算作第二类；若希望分成三类，则自上而下为：“中华肿瘤杂志”到“中国肿瘤生物治疗杂志”算作第一类，“肿瘤”到“肿瘤肺癌杂志”算作第二类，“中国肿瘤”到“现代肿瘤医学”算作第三类。

2.2 基于原变量进行无序样品聚类

所需要的 SAS 程序如下：

```
data a1;
  infile 'F:\CCC\23 种肿瘤类期刊文献计量学指标资料.txt';
```

```
    input name $20. x1 - x8;
    id = _n_;
run;
ods graphics on;
proc cluster data = a1 method = war std simple ccc
pseudo outtree = cluster;
    var x1 - x8;
    id id;
    copy name;
run;
proc tree horizontal;
    id name;
run;
ods graphics off;
```

【SAS 主要输出结果】

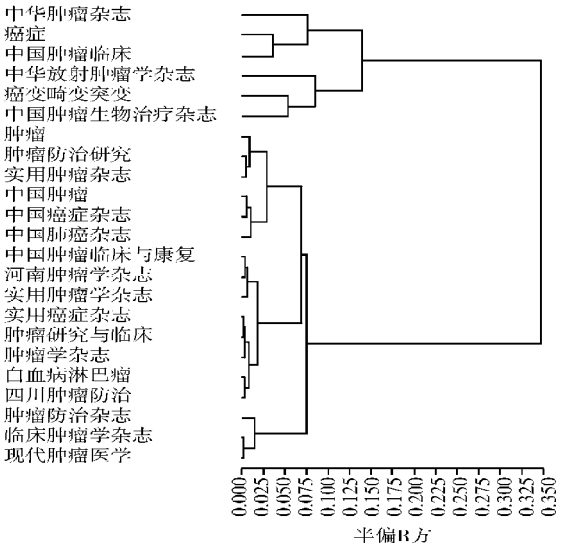

图 2 以树状图形式呈现的无序样品聚类结果 2

由图 2 可看出,若希望分成两类,则自上而下为:“中华肿瘤杂志”到“中国肿瘤生物治疗杂志”算作第一类,“肿瘤”到“现代肿瘤医学”算作第二类;若希望分成三类,则自上而下为:“中华肿瘤杂志”到“中国肿瘤生物治疗杂志”算作第一类,“肿瘤”到“四川肿瘤杂志”算作第二类,“肿瘤防治杂志”到“现代肿瘤医学”算作第三类。

以上两种聚类结果略有不同,究竟哪一个聚类结果更合理,目前尚无金标准。事实上,还有很多种聚类方法,例如可以采用《基于标准化变换的求和法:一种新的样品聚类分析方法》一文提及的方法以及其他方法,如秩和比法、Topsis 法等[4-5]。

参考文献

[1] 胡良平. 面向问题的统计学——(3)试验设计与多元统计分析[M]. 北京: 人民卫生出版社, 2012: 19 - 39.

[2] 薛薇. R 语言数据挖掘方法及应用[M]. 北京: 电子工业出版社, 2016: 226 - 281.

[3] 郑捷. 机器学习——算法原理与编程实践[M]. 北京: 电子工业出版社, 2015: 135 - 143, 208 - 213.

[4] 郭春雪, 沈宁, 胡良平. 基于标准化变换的求和法: 一种新的样品聚类分析方法[J]. 四川精神卫生, 2017, 30(3): 211 - 216.

[5] 胡良平, 黄国平. 医学科研设计方法与关键技术[M]. 成都: 四川大学出版社, 2017: 349 - 360.

(收稿日期:2018 - 04 - 02)

(本文编辑:陈 霞)