第 29 卷第 11 期
2012 年 11 月
统计研究
Statistical Research
Vol. 29, No. 11
Nov. 2012

主成分聚类分析有效性的思考*

王德青　朱建平　谢邦昌

内容提要：本文针对经典聚类分析和普通主成分聚类分析极端情形下的失效问题展开讨论，通过定义客观赋权的主成分距离为分类统计量，并以实证检验取得良好效果为依据，有效地解决了主成分聚类分析在极端情形下所不能揭示的问题。

关键词：分类；主成分聚类分析；极端情形
中图分类号：O212　　文献标识码：A　　文章编号：1002－4565(2012)11－0084－04

Remarks on the Efficiency of Principal Component Cluster Analysis

Wang Deqing　Zhu Jianping　Xie Bangchang

Abstract: A discussion about the invalidations of canonical cluster analysis and ordinary principal component cluster analysis under extreme situations is presented in this paper, by defining an objective weighted principal component distance as classification statistic, this paper solves the problems effectively which can't be revealed by principal component cluster analysis under extreme situations, and the empirical test proves its validation.

Key words: Classification; Principal component cluster analysis; Extreme situations

一、引言

面对规模庞大、复杂难辨的数据海洋，如何挖掘出隐含在其中的有用信息，清晰地展示系统结构是数据挖掘的热点问题。由于同一类事物之间具有更多的近似特性，分门别类地研究要远比在复杂多变的集合中更清晰明了，因此统计分类技术已成为数据挖掘最为常用的方法之一。然而，随着人们认识世界的不断深入和数据存储技术的飞速发展，基于传统统计技术建立的聚类分析假设条件较多，实际应用中面临诸多的局限[3,8]，于是许多学者关注经典聚类分析方法的改进研究。何跃等(2007)在分析经典聚类判别分析方法实质的基础上，提出了一种既可以对已有的聚类结果进行交叉验证，又可以对原始数据进行探索性分析的聚类判别分析框架；殷瑞飞、朱建平(2008)基于Q型因子分析的思想，结合对应分析在卡方距离框架下建立了一种新的大型数据库聚类方法，解决了Q型因子分析算法效率方面的缺陷；王惠文等(2009)以函数型数据为分析对象，定义了函数型数据的主成分距离，提出了一种低维空间上的函数型数据聚类分析的新方法；张兵等(2011)以因子分析和聚类分析为分析工具，探讨了"金砖国家"通货膨胀周期波动的影响因素及其治理政策。综观近年来关于分类方法研究的文献发现，将其他经典统计理论的优点科学合理地融合到聚类分析中，拓宽经典聚类分析的应用领域是延续经典聚类分析强大生命力的重要方向。

本文针对文献[7]－[10]提出的主成分聚类分析在极端情形下的失效问题提出了一种改进的分类方法——基于方差贡献率的加权主成分聚类分析。该方法通过定义客观赋权的主成分距离为分类统计量，并以实证检验取得良好效果为依据，有效地解决了一般主成分聚类分析在极端情形下所不能揭示的问题。

二、一般主成分聚类分析及其不足

应用传统的经典聚类分析解决实际分类问题时，通常是定性分析指标之间的关系，力图在筛选指标过程中达到增加指标独立性的目的。但定性筛选指标有较强的主观性，而且会损失部分重要信息。

* 本文获得国家社会科学基金项目"金融高频数据挖掘方法及应用研究"(11BTJ001)资助

由于主成分分析能在基本不损失原始指标信息的基础上,提取出彼此信息不重叠的主成分,因此可以将主成分分析与聚类分析有机集成。先对原始指标体系进行主成分分析,然后将主成分代替原始指标进行聚类,即一般主成分聚类分析[7-10]。

主成分分析克服了原始指标之间的共线性影响,保留了原始指标体系的大部分主要信息。但是当各主成分的方差贡献率相差较大时,忽略不同主成分聚类重要程度的差异,则必然会影响主成分聚类分析的准确性。因此,科学地构建分类统计量以体现主成分重要性的差异是聚类分析进一步改进的重要研究内容。

三、加权主成分聚类分析

(一)理论说明及具体步骤

借鉴主成分聚类分析的思想,考虑到主成分体现原始指标信息含量的差异性,本文定义加权主成分距离为分类统计量,通过赋予各主成分相应的客观权重体现其重要程度的不同。

定义:设 $F_1,F_2,\cdots,F_m(m\leqslant p)$ 为由 p 维指标向量 $X=(X_1,X_2,\cdots,X_p)^T$ 提取的主成分,记 $\alpha_i(i=1,2,\cdots,p)$ 为主成分 F_i 的方差贡献率。令 $\beta_k=\alpha_k/\sum_i^p\alpha_i(i=1,2,\cdots,p)$ 为主成分 F_k 的距离权重。定义样本 i,j 之间的加权主成分距离为:

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

通过主成分分析的特征提取,加权主成分聚类分析既剔除了原始指标共线性的重叠信息又体现了各主成分包含原始指标信息含量的差异。具体步骤如下:

(1)输入样本观测值,依据指标量纲的差异程度决定数据标准化的必要性;

(2)计算指标的相关系数矩阵及 KMO 值,判断进行主成分分析的可行性,如果指标之间存有高度的共线性,则进入(3);

(3)依据相关系数矩阵或协方差矩阵进行主成分分析,计算主成分的方差贡献率和距离权重,并结合因子载荷矩阵对主成分命名;

(4)将主成分代替原始指标,以本文定义的加权距离为分类统计量进行聚类分析,结合实际情况确定待分类样本最终的类别。

实际应用中,为了达到数据简化的目的,通常按累计方差贡献率≥85% 的原则提取前 m 个主成分。但当样品的相似度较高,仅提取前 m 个主成分不能有效区分样本时,则需要提取全部的主成分参与聚类过程。

(二)实证分析

为验证加权主成分聚类分析的有效性,本文以国际上常用的鸢尾花数据为标准测试数据[1]。描述鸢尾花属性的四个指标分别为:萼片长度、萼片宽度、花瓣长度、花瓣宽度。由于已知三种鸢尾的所属类别,本文分别用不同的聚类方法对样本进行分类,将不同聚类模型的分类结果与已知的分类结果对比,以错分率为标准判断不同聚类方法的优劣。为了直观地分析四个属性指标对鸢尾花类别区分程度的不同,绘制四个属性指标的散点图矩阵如图 1 所示。

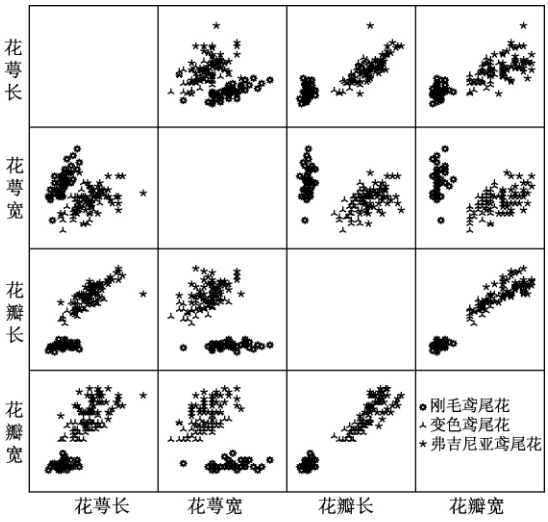

图 1　鸢尾花属性指标相关关系矩阵图

由图 1 可以看出,鸢尾花数据点在萼片长度和萼片宽度坐标上的分布密集,类与类之间的界限不明显,说明萼片属性指标较难正确区分三个类。而花瓣长度和花瓣宽度坐标上数据点的分散性比较大,类与类之间的界限明显,说明花瓣属性指标能够相对更有效地区分三个品种。如果忽略两类属性指标分类效率的差异,等权地用原始属性指标进行聚类,则相当于视四个原始指标同等重要。一方面显现不出花瓣属性指标有利于提高聚类质量的突出作用,另一方面会使不利于聚类结果的萼片属性指标削弱花瓣属性指标的分类效率。

应用本文提出的加权主成分聚类分析对鸢尾花进行分类,计算 KMO 值为 0.599,可知原始指标之

间存有高度的共线性，适宜做主成分分析。按累计方差贡献率≥85% 原则提取前两个主成分，并结合因子载荷矩阵确定主成分的实际含义，结果如表 1 所示。

表 1 的分析结果显示，第一主成分的信息含量是第二主成分信息含量的近 20 倍，说明两个主成分的重要程度相差悬殊。第一主成分的载荷主要集中在花瓣属性方面，第二主成分载荷主要集中在萼片宽度上，因此分别定义 F1、F2 为花瓣属性因子和萼片属性因子。为增强三种聚类模型的对比效果，本文统一采用 $q=2$ 的欧式距离为分类统计量，对比结果如表 2 所示。

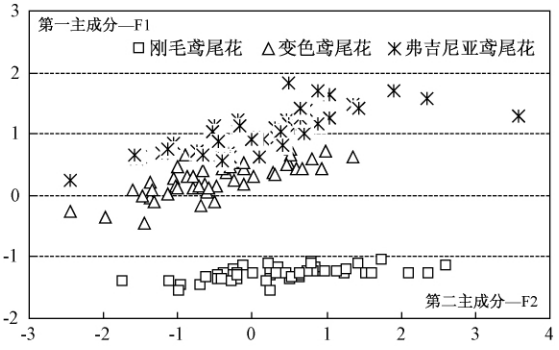

图 2 鸢尾花第一、第二主成分散点图

综观图 1、图 2 发现，刚毛鸢尾花与其他两种鸢尾花的原始指标值和主成分得分均相差较大，表现在散点图上即品种之间的界限明显，因此三种聚类模型都能成功地将刚毛鸢尾花与其他两种鸢尾花分开。但变色鸢尾花和弗吉尼亚鸢尾花之间的属性特征区别不大，不同聚类模型的分类结果差异较大。以错分率为标准，三种聚类分析的优劣次序依次是加权主成分聚类分析、经典聚类分析和一般主成分聚类分析。尤为引起注意的是，一般主成分聚类分析在鸢尾花的分类研究中效果不理想，错分率明显高于其他两种聚类方法，甚至出现将刚毛鸢尾花也错误分类的极端现象。结合表 1 的计算结果，究其原因在于本文提取的两个主成分信息含量相差近 20 倍。等权地将两个主成分代替原始指标聚类分析，则会过于放大第二主成分的重要性从而削弱第一主成分的分类效率，抹煞了各主成分重要性客观存在的悬殊差异，导致错误的分类结果。

四、结论

在本文的讨论中，一般主成分聚类分析的分类效果明显劣于其他两种聚类方法，这说明一般主成分聚类分析并不必然地优于经典聚类分析。事实上，指标之间的共线性影响和指标之间重要性的客观差异是经典聚类模型并存的两个缺点，对经典聚类模型的改进必须综合考虑以上两个缺点。

相对经典聚类分析，一般主成分聚类模型克服了指标之间的共线性影响，当主成分的信息含量相差不大时，一般主成分聚类模型会提高聚类的准确度。但当主成分的重要性相差悬殊时，一般主成分聚类分析是否一定优于经典聚类分析有待商榷。与

表 1 主成分分析结果

主成分	特征根 λ_i	方差贡献 α_i	权重 β_i	因子载荷				因子命名
				X1	X2	X3	X4	
F1	426.691	91.891	0.94	0.886	-0.398	0.997	0.968	花瓣属性因子
F2	25.839	5.565	0.06	0.417	0.779	-0.061	-0.047	萼片属性因子

表 2 三种聚类分析结果的比较

聚类方法	经典聚类分析			一般主成分聚类分析			加权主成分聚类分析		
类编号	植物品种			植物品种			植物品种		
	刚毛	变色	弗吉尼亚	刚毛	变色	弗吉尼亚	刚毛	变色	弗吉尼亚
1	50	0	0	49	0	0	50	0	0
2	0	35	1	1	36	19	0	43	3
3	0	15	49	0	14	31	0	7	47
合计	50	50	50	50	50	50	50	50	50
错分率	10.666%			22.7%			6.666%		

以上两种聚类方法相比,加权主成分聚类分析同时解决了经典聚类分析和一般主成分聚类分析存在的问题,聚类效果明显提高。但当原始指标的共线性较弱不满足主成分聚类分析的条件时,加权主成分聚类模型则会失效。

参考文献

[1] R. A. Fisher, The use of multiple measurement in taxonomic problems [J]. Annals of Eugenics,1936(7):179 - 188.
[2] A. K. JAIN,M. N. MURTY,P. J. FLYNN. Data Clustering: A Review [J]. ACM 31. 3(Sep. 1999):264 - 323.
[3] 王进. 聚类分析中的距离与变量选择 [J]. 山西财经大学学报,2007(29):36 - 43.
[4] 何跃,杨磊,徐玖平. 一种新的聚类判别分析框架及其实证研究 [J]. 计算机应用研究,2007(24):32 - 40.
[5] 殷瑞飞,朱建平. 数据挖掘中一种新的聚类方法——基于对应分析与因子旋转 [J]. 统计研究,2008(25):93 - 97.
[6] 张兵,李翠莲. "金砖国家"通货膨胀周期的协动性 [J]. 经济研究,2011(9):29 - 40.
[7] 王劼,黄可飞,王惠文等. 一种函数型数据聚类分析方法 [J]. 数理统计与管理,2009(28):839 - 844.
[8] 王德青. 主成分聚类分析在矿井安全评价应用中的思考 [J]. 中国矿业,2011(20):51 - 57.
[9] 庞丽、李显君. 我国汽车产业竞争力区域差异的实证研究 [J]. 数理统计与管理,2011(30):951 - 959.
[10] 岳斯玮. 基于主成分聚类分析对区域教育综合发展水平评价 [J]. 西南民族大学学报,2012(38):37 - 43.

作者简介

王德青,男,1983 年生,山东青岛人,厦门大学经济学院统计系、厦门大学数据挖掘研究中心博士生。研究方向为经济统计、数据挖掘、计量经济学。

朱建平,男,1962 年生,厦门大学经济学院教授、博士生导师、统计系主任、厦门大学数据挖掘研究中心主任。研究方向为数理统计、数据挖掘、计量经济学。

谢邦昌,男,1962 年生,台湾辅仁大学统计资讯学系教授、厦门大学经济学院统计系讲座教授、博士生导师。研究方向为数据挖掘与商业智勇。

(责任编辑:程　晞)

《统计研究》"中图分类号"要求

《统计研究》中图分类号可参考下表,并与以下的文献标识码列在一行用五号宋体标示。

《统计研究》主要栏目中图分类号简明对照表

主栏目	分栏目	分类号
统计工作的改革与发展		C829. 2
	法律法规	C829. 21
	统计方法制度	C829. 22
	统计管理体制	C829. 23
	统计资料管理,统计信息化建设,统计数据库	C816
国外统计工作		C829. 1
经济统计学		F222
国民经济核算		F222. 33
统计方法的应用与创新		C81
	统计调查、抽样与抽样分布	C811
	概率论	0211
	数理统计方法(如非参数统计、参数估计、假设检验、时间数例、贝叶斯统计、相关分析与回归分析)	0212
	统计指数	C813
统计实证分析		C812
	统计模型的应用	F222. 3
统计史		C829. 29
统计教育		C829. 29