文章编号:1004-1478(2012)02-0105-04

全局集成聚类法的应用研究

曲双红, 汪远征, 徐雅静

(郑州轻工业学院 数学与信息科学系, 河南 郑州 450002)

摘要:针对用主成分分析做综合评价存在的不足,基于全局主成分分析、熵值法以及聚类分析,提出全局集成聚类分析方法.结合我国各地农村居民消费现金支出的数据,通过应用实例验证了该方法的可行性.

关键词:全局主成分分析;熵值法;集成聚类法

中图分类号:O212 **文献标志码**:A

Application research of the whole-integrated-clustering

QU Shuang-hong, WANG Yuan-zheng, XU Ya-jing

(*Dept. of Math. and Infor. Sci., Zhengzhou Univ. of Light Ind., Zhengzhou* 450002, *China*)

Abstract: Aiming at the problem that principal component analysis evaluation is insufficient, based on the whole principal component analysis, entropy method and cluster analysis, the method of global integration clustering was proposed. The cash data of rural residents living consumption tendency in China was analyzed. The feasibility of this method was proved by using example.

Key words: whole principal component analysis; entropy method; integrated clustering

0 引言

主成分分析是一种通过降维技术化繁为简的多指标评价技术.但在实际应用中,直接使用传统的主成分分析方法存在不少问题,为此很多学者提出了不同的改进方法[1].在现实生活中,随着时间的推移和数据的累积,形成了按时间顺序排列的平面数据表.若对不同的数据表分别进行主成分分析,则会产生不同的主超平面,那样就无法保证系统分析的统一性、整体性和可比性.传统主成分分析在求主成分综合得分时,如果第一主成分的方差贡献率不够高,常取各主成分的方差贡献率作为权重,这显然太主观.王学民[2]已证明了其错误所在.

本文拟在已有研究基础上,提出全局集成聚类法:用全局主成分分析[3]做综合评价时,先按照全局第一主成分得分得到第一次综合得分值,然后采用熵值法[4]客观赋权,得到第二次综合得分值,最后对2种评价结果进行一致性检验.如果通过一致性检验,就将2种评价结果进行集成,形成最终的评价结果;如果2种评价方法不具有一致性,则采用全局主成分聚类法进行评价.

1 预备知识

1.1 全局主成分分析

全局主成分分析[2]建立在经典主成分分析的基础上,通过对立体时序数据表进行主成分分析,

收稿日期:2011-09-26

基金项目:河南省科学技术研究项目(112300410156);河南省教育厅自然科学基金项目(2011A110022)

作者简介:曲双红(1973—),女,河南省偃师市人,郑州轻工业学院讲师,主要研究方向为应用数学.

可以得到一个统一的主成分子空间. 从全局来看,每张不同年份的数据表在该子空间的投影将在该子空间得到最佳的近似表达,从而可动态地描述数据表所反映的潜在信息. 根据叶双峰[5]的研究,我们先对原始数据作均值化处理,再由协方差阵出发做全局主成分分析,根据分析结果按全局第一主成分得分即可得到第一次综合得分.

1.2 熵值法

假设有 n 个待评对象,通过上述主成分分析提取了 m 个主成分,即将 p 个指标浓缩为 m 个,则构成一个具有 n 个样本、m 项指标的主成分矩阵,我们可以利用信息熵的工具,计算各指标的权重. 熵的概念产生于热力学,其定义为 $E = -\sum_{i=1}^{n} p_i \ln p_i$,其中 p_i 为第 i 种状态出现的概率. 某个指标的信息熵越小,表明该指标的变异程度越大,提供的信息量越大,在综合评价中的作用越大,权重也就越大. 反之,若某个指标的信息熵越大,其权重越小. 主成分矩阵 $\boldsymbol{F}_{ij}$ 不包含量纲,所以无需标准化,熵值法计算步骤如下:

1)计算第 j 项指标下第 i 个样本的比重

$$p_{ij} = \boldsymbol{F}_{ij} / \sum_{i=1}^{n} \boldsymbol{F}_{ij}$$

2)计算第 j 项指标的熵值

$$e_j = -\frac{1}{\ln n} \sum_{i=1}^{n} p_{ij} \ln p_{ij}$$

3)计算第 j 项指标的权重

$$w_j = (1 - e_j) / \sum_{j=1}^{m} (1 - e_j)$$

4)计算第 i 个样本的综合值

$$v_i = \sum_{j=1}^{m} w_j p_{ij}$$

在第 1 步中,为防止矩阵中有负值,可以将所有数据加上一个最小负数的绝对值,这样处理不会改变结果;然后按照平移后的矩阵进行第 2 到第 4 步的计算.

1.3 一致性检验

对于不同的评价方法,需要了解评价结果的一致性. 我们采用 KENDALL-W 协和系数法进行一致性检验[6]. 令

$$W = \frac{12 \sum_{i=1}^{n} R_i^2 - 3m^2 n(n+1)^2}{m^2 n(n^2 - 1)}$$

其中,m 为评价方法的数目,n 为评价对象的数目,R_i 为各评价对象的等级之和. 检验统计量 $\chi^2 = m(n-1)W$ 在大样本情况下近似服从 $\chi^2_\alpha(n-1)$. 当 $\chi^2 > \chi^2_\alpha(n-1)$ 时,认为 m 种评价方法的平均等级之间具有一致性;否则,不具有一致性.

1.4 集成聚类

使用全局第一主成分和熵值法进行评价时,会出现不同的结果,我们对排序结果进行一致性检验,若通过一致性检验,则对这 2 种方法得到的平均值进行集成,形成总的评价. 第 i 个评价对象的综合得分 $F_i = \alpha F_{i(1)} + \beta F_{i(2)}$. 其中 α, β 为权重,$\alpha + \beta = 1$;$F_{i(1)} + F_{i(2)}$ 分别为全局主成分法和熵值法得到的平均值. 当 2 种方法的评价结果通过一致性检验时,对各样本的集成得分做聚类分析,然后计算类中样品的均值得分来确定类间的排序,最后根据类中样品的集成得分,确定各类中的样品的排序,得到综合评价. 如果 2 种方法没有通过一致性检验,可以采用主成分聚类法[7]进行综合评价.

2 算例

笔者利用上述全局集成聚类法对我国各省、市、自治区农村居民生活消费现金支出进行综合评价. 本文的数据分别取自 2001—2009 年的《中国统计年鉴》. 从各地农村居民家庭平均每人生活消费现金支出(2000—2008 年的数据)中选择了反映农村居民生活消费的 8 项指标:食品支出(x_1)、衣着支出(x_2)、家庭设备用品及服务支出(x_3)、医疗保健支出(x_4)、交通和通信指出(x_5)、教育文化娱乐服务支出(x_6)、居住支出(x_7)、杂项商品和服务支出(x_8). 本文的所有结果均在 SAS 中运行得出.

2.1 综合评价结果

对我国 2000—2008 年 31 个省、市、自治区的农村居民生活消费现金支出进行全局主成分分析. 取 31 个样本点,8 项指标,9 年的 9 张数据表构成 31 × 8 × 9 维的时序立体数据表,对数据均值化处理后进行全局主成分分析. 从结果可以得到,各项指标间的相关系数普遍较大,特别是 x_1, x_3 与其他指标的相关系数均大于0.81,说明这 2 项指标与其他 6 项指标密切相关. 同时,第 1 主成分对方差贡献率已达 86.29%,我们先整理得到各地区按年度第 1 主成分得分(见表 1),然后求各地区的第 1 主成分得分的

历年平均值,并按平均值大小对各地区排序.

对以上样本再用熵值法(为提高准确度,提取前 2 个主成分,前 2 个主成分累计贡献率为 91.34%)做综合评价. 其中第 1 主成分的权重为 0.934 947,第 2 主成分的权重为 0.065 053,这与以贡献率为权重有很大差别,整理出表 2,求各地区的熵值法得分的历年平均值,然后进行排序.

对上述 2 种方法的评价结果进行一致性检验,结果表明,在显著性水平 0.005 下,这 2 种评价结果具有一致性. 对这 2 种评价结果进行集成,集成得分见表 3. 在评价函数 $F_i = \alpha F_{i(1)} + \beta F_{i(2)}$ 中,取 $\alpha = \beta = 0.5$[6],最终的评价结果见表 4(限于篇幅,这里只列出几种评价结果不同的地区以便比较).

2.2　集成聚类

在 SAS 中对表 3 数据用离差平方和法进行聚类分析,得到集成聚类分析树形图(见图 1). 聚类分析显示 31 个省、市、自治区可分为 6 类,按集成得分的平均值从高到低排序有:第 1 类:{上海};第 2 类:{北京,浙江};第 3 类:{天津,山东,辽宁,江苏,福建,广东};第 4 类:{河北,湖北,内蒙古,吉林,湖南,黑龙江};第 5 类:{山西,安徽,陕西,河南,江西,宁夏,广西,海南,重庆,四川,青海,新疆};第 6 类:{贵州,西藏,云南,甘肃}.

从分类及排序结果可以清楚地看到:由于上海、北京、浙江这几个地区的经济发达,农村消费水平最高;天津、江苏、福建、广东等次之;农村消费水

表 1　各省、市、自治区 2000—2008 年农村居民生活消费水平第一主成分得分

地区	Prin2000	Prin2001	Prin2002	Prin2003	Prin2004	Prin2005	Prin2006	Prin2007	Prin2008
北京	1.654 751	1.886 021	2.230 799	2.460 617	3.133 482	4.395 679	4.914 894	5.699 568	6.852 116
天津	-0.333 57	-0.435 6	-0.149 83	-0.067 02	0.257 404	0.871 997	1.346 465	1.483 378	1.811 766
…	…	…	…	…	…	…	…	…	…
新疆	-1.547 49	-1.414 63	-1.365 51	-1.307 83	-1.093 65	-0.679 09	-0.533 12	-0.175 94	0.095 283

表 2　各省、市、自治区 2000—2008 年农村居民生活消费水平熵值法得分

地区	2000 年	2001 年	2002 年	2003 年	2004 年	2005 年	2006 年	2007 年	2008 年
北京	1.621 122	1.842 905	2.174 124	2.361 155	2.962 957	4.074 834	4.552 976	5.282 161	6.354 337
天津	-0.314 33	-0.391 56	-0.123 49	-0.036 29	0.265 265	0.767 297	1.216 197	1.361 673	1.676 828
…	…	…	…	…	…	…	…	…	…
新疆	-1.447 92	-1.320 97	-1.283 3	-1.239 07	-1.043 41	-0.653 55	-0.519 03	-0.181 22	0.073 79

表 3　各省、市、自治区 2000—2008 年农村居民生活消费水平集成得分

地区	2000 年	2001 年	2002 年	2003 年	2004 年	2005 年	2006 年	2007 年	2008 年
北京	1.637 936	1.864 463	2.202 462	2.410 886	3.048 22	4.235 256	4.733 935	5.490 864	6.603 227
天津	-0.323 95	-0.413 58	-0.136 66	-0.051 65	0.261 334	0.819 647	1.281 331	1.422 525	1.744 297
…	…	…	…	…	…	…	…	…	…
新疆	-1.497 71	-1.367 8	-1.324 41	-1.273 45	-1.068 53	-0.666 32	-0.526 08	-0.178 58	0.084 537

表 4　各省、市、自治区 2000—2008 年农村居民生活消费水平评价排序

地区	Prin1 历年平均	Prin1 排序	熵值法历年平均值	熵值法排序	集成综合评价法	集成综合评价排序	集成聚类排序
…	…	…	…	…	…	…	…
安徽	-0.646 36	18	-0.608 93	19	-0.627 6	18	18
陕西	-0.650 34	19	-0.607 42	18	-0.628 9	19	19
…	…	…	…	…	…	…	…
青海	-0.901 65	25	-0.858 06	26	-0.879 9	26	26
海南	-0.910 61	26	-0.847 85	25	-0.879 2	25	25
…	…	…	…	…	…	…	…

平最低的地区是贵州、西藏、云南、甘肃. 表 4 显示安徽和陕西、青海和海南的排序发生了变化,考察这 4 个地区的各项指标发现,集成聚类排序更为合理.

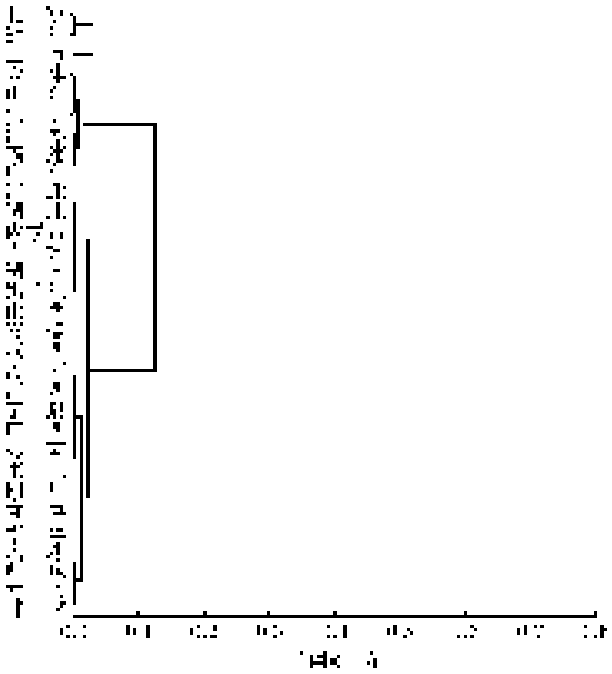

图 1 集成聚类分析树形图

2.3 各类地区现金消费趋势和消费水平分析

基于上述结果,笔者对各类地区这 6 年来农村居民现金消费趋势和消费水平进行了分析. 首先按年份取每类地区的集成得分平均值,然后在 SAS 中作趋势图(见图 2). 从图 2 不难发现,各类地区农村居民的现金消费支出都呈上升趋势,第 1 类和第 2 类地区发展较快,这说明采用本文方法所得数据是符合实际的,也进一步说明了该方法的可行性. 再将表 3 中的数据按行平均,得到各地区历年消费水平综合指标的平均值,排序并作图(见图 3). 从图 3 可以看出,上海、北京、浙江等经济发达地区农村居民消费水平明显高于其他地区,西藏、贵州、甘肃等经济落后地区农村居民消费水平比较低,这与前面所得结果是吻合的.

3 结语

本文将全局主成分分析和熵值法的评价结果进行全局集成聚类,有效结合 2 种方法的优点,动

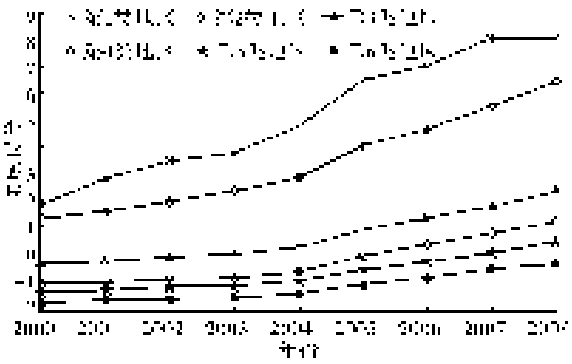

图 2 各类地区现金消费趋势图

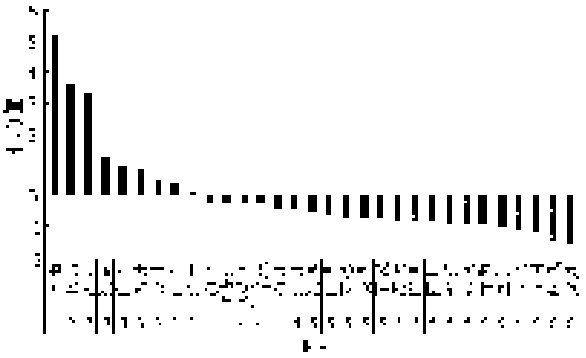

图 3 消费水平综合指标的平均值排序

态地描述立体时序数据表的潜在信息,并结合实例说明了该方法在做综合评价时的客观性和合理性. 由于多指标问题存在普遍性,针对不同多指标问题选取不同的综合评价方法,以及同一问题选取最好的综合评价方法将是下一步的研究方向.

参考文献:

[1] 曲双红,李华,李刚. 基于主成分分析的几种常用改进方法[J]. 统计与决策,2011(5):155.

[2] 王学民. 对主成分分析中综合得分的质疑[J]. 统计与决策,2007(8):31.

[3] 乔峰,姚俭. 时序全局主成分分析在经济发展动态描绘中的应用[J]. 数据统计与管理,2003,22(2):1.

[4] 徐永智,华惠川. 对主成分分析三点不足的改进[J]. 科技管理研究,2009(6):128.

[5] 叶双峰. 关于主成分分析做综合评价的改进[J]. 数理统计与管理,2001,20(2):52.

[6] 孙刘平,钱吴永. 基于主成分分析法的综合评价方法的改进[J]. 数学的实践与认识,2009,39(18):15.

[7] 张虎. 主成分聚类分析法的案例教学法[J]. 统计与决策,2007(20):163.