主成分集成评价方法的问题探析与模型拓展

王德青,李凯风,周　娇
(中国矿业大学 管理学院,江苏 徐州 221116)

摘　要:文章针对主成分综合评价主要环节的一般性问题展开讨论,给出可行的解决方案并进行了理论分析。在总结现有关于主成分聚类分析重要文献的基础上,通过构建客观赋权的加权主成分距离为聚类统计量,有效地解决了现有聚类模型不能处理指标共线性和重要性差异悬殊的问题。对比本文拓展的聚类模型与同类模型的分类效率发现,加权主成分聚类分析蕴含的客观合理性是其优势所在的根本原因。

关键词:数据挖掘;加权主成分聚类;主成分分析;集成评价
中图分类号:C81　　**文献标识码**:A　　**文章编号**:1002-6487(2015)01-0004-05

0　引言

评价问题在自然科学和社会科学的研究领域中普遍存在。广义上讲,大凡与选优或排序相关的问题都会涉及对所研究对象的评价,因此,没有科学的评价便没有正确的决策[1]。近年来,多指标评价问题的理论研究和实践应用取得了巨大进展,日趋复杂化、数学化、多学科化的评价方法层出不穷,如综合指数法、多元统计分析法、模糊评价法、灰色系统法、层次分析法、数据包络法、人工神经网络法、功效系数法等。从当前综合评价的研究现状看,绝大部分的研究成果还停留在"具体理论方法+实际案例应用"的初级阶段[1~3],理论方法研究同实际应用之间的衔接较为薄弱。从研究的具体内容看,评价问题研究大致分为两类,一类是针对不同的评价问题,如何构建科学的评价指标体系;另一类是各种评价模型的应用、修正和拓展,本文着眼于后者。事实上,随着评价层面的扩展和问题复杂程度的加深,综合评价研究的重心逐渐转变为方法论问题,因为在分析内在规律复杂的评价问题时,仅使用一种理论或方法往往难以取得良好效果。这时需要针对具体问题特点,将多个同类理论和方法的长处有机融合,集成能有效分析复杂问题规律的新方法,并从理论上完善处理过程。

在评价问题的方法研究和实际应用中,主成分分析无疑是提及频率较高的词汇。由于其强大的浓缩数据信息、简化数据结构功能,主成分分析逐渐成为一种独具特色的多指标评价技术。在这样的背景下,国内外学者针对主成分评价的相关问题进行了深入系统的分析研究,而《统计与决策》期刊无疑为这些研究成果的展示、交流和相互探讨提供了良好的平台。纵观《统计与决策》近年来刊登的关于主成分分析文章内容看,研究成果主要分为两类,第一类是主成分分析在评价问题中的直接应用,如:文献[4]基于国家统计局发布的客观数据,应用主成分方法深入比较分析了中国制造业与国际水平的差异,并依此给出提升中国制造业科技创新能力的政策建议;文献[5]基于2007~2009年19家上市公司的年度报告,依据主成分分析的结论提出了完善上市公司透明度和内部信息控制的改进建议。第二类是集成其它评价方法的优点对经典主成分分析进行拓展改进,如:文献[6]采用主成分分析和BP神经网络相结合的方法,通过对样本反复训练和仿真,准确地拟合出训练样本并对检验样本具有较高的预测精度;文献[7]综合运用主成分分析和聚类分析,构造了我国中小企业成长性评价模型,并以首批28家创业板上市企业作为样本进行实证检验,在此基础上提出了管理建议。综合来看,主成分分析在自然科学和社会科学领域的评价问题中均有广泛的应用,但与此同时应该注意到,直接将主成分分析方法应用于多指标评价存在不少问题[1],而现有主成分集成评价模型的科学性也未从理论上进行论证。主成分分析的实际应用中,如果忽略上述两个方面则会出现诸多问题,依此所得结论与建议的有效性有待商榷。

现有的研究成果对主成分分析的方法应用和进一步理论研究具有重要的借鉴意义,但包括主成分分析在内的每一评价方法都是有针对性的分析数据中的特定规律,如果忽略模型适用的前提和待分析对象的具体特点,直接套用现成模型或机械拼凑不同评价模型是否一定能提高分析结果的科学性,需要理论分析和实践应用的双重检验。基于以上认识,本文:(1)在机理分析的基础上,梳理主成分分析用于多指标综合评价存在的问题,并提出可行的解决方案以填补现有主成分评价方法实际应用中的漏洞;(2)针对文献[7]、[8]中主成分聚类模型的合理性提出质疑,在方法相容性研究的基础上对主成分聚类模型进行拓展,并以实例应用比较新方法与同类方法的优劣,依据实

基金项目:国家自然科学基金青年项目(710201139);国家社会科学基金资助项目(11BTJ001);全国统计科学研究计划重大项目(2012LD001)
作者简介:王德青(1983-),男,山东青岛人,博士,助理教授,研究方向:数据挖掘。

证结果论证新方法的有效性。

1 主成分评价模型与问题探析

1.1 主成分评价模型

主成分分析本质上是在不改变原始数据信息量的基础上，通过正交变换将线性相关的原始指标转化为相互独立的综合指标。假设 $X=(X_1,X_2,\cdots,X_p)$ 为 p 维指标向量，其协方差矩阵 $\mathrm{Cov}(X)=\sum_X$ 的特征根为 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$，则主成分分析的优化问题可表示为：

$$Max\ \mathrm{Var}(F_i)=Max\ \mathrm{Var}(u_i^{'}X)$$
$$s.t.\ \begin{cases} u_i^{'}u_i=1 \quad i=1,2,\cdots,p \\ \mathrm{Cov}(F_i,F_j)=0 \quad i\neq j \\ \mathrm{Var}(F_i)=\lambda_i \quad i=1,2,\cdots,p \\ \sum_{i=1}^{p}\mathrm{Var}(F_i)=\sum_{i=1}^{p}\mathrm{Var}(X_i) \end{cases} \tag{1}$$

解得(1)式最优解 $F_i=u_i^{'}X(i=1,2,\cdots,p)$ 即为 X 的主成分，定义 $\alpha_i=\lambda_i\Big/\sum_{k=1}^{p}\lambda_k$ 为主成分 F_i 的方差贡献率，并称 $\alpha=\sum_{i=1}^{m}\lambda_i\Big/\sum_{k=1}^{p}\lambda_k$ 为前m个主成分的累计方差贡献率。实际应用时，为了排除次要信息的干扰以达到数据简化的目的，通常按累积方差贡献率≥85%原则提取前m个主成分。令 $W=(\alpha_1,\alpha_2,\cdots,\alpha_m)^{'}$ 为主成分的权重向量，则主成分综合评价模型可表示为：

$$Z=\alpha_1F_1+\alpha_2F_2+\cdots\alpha_mF_m=W^{'}F \tag{2}$$

可以看出，主成分综合评价模型本质上是按重要程度(方差贡献率)客观加权各主成分得分，依据待分析对象的主成分综合得分进行排序选优。

1.2 存在的问题及系统分析

主成分综合评价的理论框架主要包括：(1)评价指标体系的构建；(2)原始指标数值的规则化；(3)确定权重进行指标合成。其中(1)由评价者主观设定，体现其价值判断；(2)和(3)由数学计算所得，反映了被评价对象的客观属性。以上三个方面紧密联系，任何一方面的处理不当都会导致评价结果的失真，本文针对上述三个过程中出现的一般性问题展开讨论。

问题1：是否可以忽略评价指标的相关性和重要性差异，选择尽量多的指标？

指标体系是多指标评价活动开展的基础，指标体系设计的好坏对任何多指标评价方法都至关重要。科学的指标体系应具有代表性、独立性和全面性，但实践应用中，误以为主成分分析能够区分不同指标重要程度的差异和将重复信息完全剔除的缘故，往往对评价指标不加选择，甚至有意选用较强相关性的指标描述待评价对象的特征。多指标的大样本无疑为评价研究提供了丰富信息，但也在一定程度上增加了问题分析的复杂程度。事实上，若指标体系中各指标之间的相关程度差异悬殊，则综合评价函数中的权重分配存在明显的集结倾向，这一点通过下例说明。

例1设四维综合评价问题的协方差矩阵如下所示(具体数据略)：

$$\begin{bmatrix} 1 & 0.988 & 0.86 & 0.081 \\ 0.988 & 1 & 0.856 & 0.146 \\ 0.86 & 0.856 & 1 & -0.007 \\ 0.081 & 0.146 & -0.007 & 1 \end{bmatrix}$$

其中 x_1,x_2,x_3 为高度线性相关的同类指标，x_4 为与 x_1,x_2,x_3 相关性较低的另一类指标。计算协方差矩阵的特征值分别为 $\lambda_1=2.814,\lambda_2=1.004,\lambda_3=0.173,\lambda_4=0.009$。前两个主成分的累积方差贡献率已达95.44%，故取前两个主成分并表示为：$F_1=0.586x_1+0.586x_2+0.5544x_3+0.0709x_4$，$F_2=-0.029x_1+0.0363x_2-0.135x_3+0.9901x_4$，则(2)式定义的综合评价函数为：$Z=0.405x_1+0.403x_2+0.356x_3+0.298x_4$。可以看出：$x_1,x_2,x_3$ 的系数之和明显比 x_4 的系数大，表明权重明显向相关性较高的指标倾斜，也即同类指标在综合评价函数中占据绝对的主导作用。但就单个指标的实际含义而言，x_4 可能是综合评价时应着重考虑的关键性指标，其权重应该占较大比重，而 x_1,x_2,x_3 仅是可有可无的辅助性指标，其权重应占较小份额。本例的权重分配不合理现象说明，传统的主成分综合评价函数不能区分原始指标重要程度的差异，反而可能会因为信息重叠而放大同类指标的重要性，歪曲被评价对象的相对地位造成评价结果的失真。因此，构建综合评价的指标体系时，应该在保证指标重要程度相差不大的前提下缩减同类指标的数目，节省指标收集成本的同时减少信息交叉与冗余。以下讨论均作此假定。

问题2：如何选择原始指标数据的无量纲化方法？

主成分实际应用中，为了消除分析结果因指标取值相差悬殊而剧烈变化的影响，往往对原始数据进行归一化[8]或标准化[11]的无量纲处理，但上述处理方法在消除量纲影响的同时也造成了信息损失。事实上，主成分分析过程利用的原始指标数据信息主要为两类：一是各指标的变异程度信息，由其方差(或变异系数)大小反映；二是指标之间相互影响程度的信息，由其相关系数(或协方差)体现。原始数据的协方差矩阵能完整刻画上述两类信息，但标准化后的指标数据方差均为1，从标准化数据的协方差矩阵提取主成分相当于仅从原始指标的相关系数矩阵中提取主成分，遗漏了原始指标变异程度的信息。

为了使主成分分析过程充分利用原始数据的全部信息，同时避免数据量纲的影响，可行的做法是：(1)当原始数据的变化范围相差不大时，宜直接从原始指标数据的协方差矩阵进行主成分分析[10]；(2)若各指标量纲差异悬殊，本文建议使用“均值化”方法进行无量纲处理，即各指标原始数据分别除以其相应的均值。数学上可以证明，均值化数据的协方差矩阵全面反映了原始数据的信息。首先，协方差矩阵的对角元素为各指标的变异系数，反映了原始指标的变异程度；其次，均值化处理未改变指标间的相互关系。综合考虑，相对归一化或标准化的无量纲化方法，均

值化法不仅简单易行，而且客观充分地保留了原始数据信息。

问题3：综合评价是仅用第一主成分还是取前几个主成分的加权综合得分？

主成分综合评价的多数文献以(2)式为模型[8,11~14]，其本质上是以方差贡献率客观加权各主成分，主成分个数的选取问题也是权重的分配问题[17]。需要说明的是累积方差贡献率表明前几个主成分共同反映原始数据信息能力的大小，但若以方差贡献率加权前几个主成分—不妨称作主成分综合得分，则其反映的原始数据信息并不能超过第一主成分反映的原始数据信息。事实上，由于各主成分之间相互正交，则有：

$$\mathrm{Var}(Z)=\sum_{i=1}^{m}\alpha_i^2\,\mathrm{Var}(F_i)=\sum_{i=1}^{m}\alpha_i^2\lambda_i\le\lambda_1\sum_{i=1}^{m}\frac{\lambda_i^2}{(\sum_j^p\lambda_j)^2}=\lambda_1\frac{\sum_1^m\lambda_i^2}{(\sum_j^p\lambda_j)^2}<\lambda_1=\mathrm{Var}(F_1) \tag{3}$$

可见，主成分综合得分虽然由几个主成分综合而成，并且各主成分反映原始数据信息能力累加起来超越第一主成分，但(2)式模型综合原始数据的信息并不能超越第一主成分反映的原始数据信息。因此，主成分分析用于多指标综合评价时，使用第一主成分优于使用(2)式的主成分综合得分模型。

2 主成分聚类分析存在的问题与模型拓展

2.1 主成分聚类分析及存在的问题

传统的Q型聚类分析多是基于样本之间距离的亲疏关系进行分类，聚类算法要求描述样本的指标重要性相同并且彼此独立[10,13]。如果对存有高度共线性的指标不加处理直接聚类，那么聚类统计量将同类指标重复计算，过于放大共线性指标的作用而淹没独立性指标的贡献[16]。考虑到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集成，即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—主成分聚类分析[6]。应该肯定的是，主成分聚类克服了传统聚类分析不能处理指标高度共线性的缺点，但是当各主成分的方差贡献率相差悬殊时，忽略不同主成分重要程度(方差贡献率)的差异，则势必会影响主成分聚类分析的准确性[13]。

为了体现不同主成分重要性的差异，文献[11]、[12]提出以(2)式主成分综合得分代替原始指标聚类分析—不妨称作主成分综合得分聚类。主成分综合得分将多维的主成分信息压缩到一维的主成分综合得分中，各主成分重要性的差异也通过方差贡献率的客观赋权得到体现。但由(3)式可知，主成分综合得分反映原始数据信息的能力并不能超越第一主成分的信息反映能力，所以这种看似合理的信息融合方法未必能提高聚类的效果，甚至可能会因为损失原始数据信息而降低聚类质量。

2.2 模型进一步拓展

指标之间的高度共线性影响和指标之间重要性的客观差异是限制经典聚类模型广泛应用的两个方面，对经典聚类模型的改进必须综合考虑以上两个缺点。聚类质量高低取决于所构建的聚类统计量合理与否，显然，如果某个指标的信息含量相对其它指标大，则聚类统计量中该指标所占的比重也应较大。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，本文拓展加权主成分聚类分析如下。

定义：$F_i,\lambda_i,\alpha_i(i=1,2,\ldots,m;m\le p)$ 定义同上，令 $\beta_k=\alpha_k\big/\sum_i^m\alpha_i$ $(k=1,2,\ldots,m)$ 为主成分 F_k 的距离权重，称

$$d_{ij}(q)=[\sum_{k=1}^{m}(\beta_k(F_{ik}-F_{jk}))^q]^{\frac{1}{q}}\qquad(m\le p) \tag{4}$$

为样本 i,j 之间的加权主成分距离，其中 β_k 由原始指标体系提取，用以体现各主成分信息含量的客观差异。当 $\beta_i=\beta_j$ 特殊情形时，本文拓展模型也即现有文献的主成分聚类模型。不难证明，公式(4)满足距离定义的正定性、对称性和三角不等式。与经典聚类模型和文献[7]、[11]的研究成果相比，加权主成分聚类模型的优点在于：(1)在基本不损失原始数据信息的前提下，避免了聚类指标的共线性影响；(2)不同主成分的聚类效率差异由其对应的距离权重大小反映，并且权重来自原始数据信息，赋权标准客观合理。加权主成分聚类分析与现有同类聚类方法的核心区别，在于有机地融合了两种分类方法的长处，理论基础充分，有着同类复杂分类问题下的普遍适用性。实际应用时，通常只需提取方差贡献率较大的前几个主成分即可，但当评价对象的相似度较高，仅提取前几个主成分不能有效分类时，则需要提取全部的主成分参与聚类过程。

3 主成分聚类分析的实例

3.1 指标、数据与预处理结果

为了对比拓展的主成分聚类模型与现有聚类模型分类效率的优劣，本文采用文献[8]中上市公司创新能力评价的例子来实证说明。按科学性、简明性和可操作性原则，分三个层次确定八个指标反映上市公司的创新能力。综合考虑覆盖地点全面和经营业务全面原则，选取沪深两地30家上市公司为研究样本。对原始数据预处理后进行主成分分析的具体结果如表1所示。

表1　　主成分分析结果汇总

主成分	特征值	方差贡献率%	累计方差贡献率%	距离权重
F_1	2.623	32.782	32.782	0.364
F_2	1.943	24.291	57.073	0.269
F_3	1.128	14.102	71.174	0.156
F_4	0.996	12.446	83.620	0.138
F_5	0.526	6.572	90.193	0.073
F_6	0.442	5.526	95.719	——
F_7	0.342	4.281	100	——
F_8	–2.191E–16	–2.739E–15	100	——

由表1可知，前五个主成分的累积方差贡献率已达90.193%，基本包含了原始数据的全部信息，因此下文选取前五个主成分进行计算。应该注意到，五个主成分反映原始数据信息的能力相差悬殊，如第一主成分的方差贡献率为第五主成分的五倍之多。如果不加区别地将五个主成分代替原始指标直接聚类分析，则会过于放大第五主成分的重要性而削弱第一主成分的分类效率，抹煞了不同主成分聚类效率客观存在的差异。

3.2 分类结果对比

应用主成分聚类模型[8]、主成分综合得分聚类模型[11]、加权主成分聚类模型和第一主成分聚类模型分别对上市公司分类，需要说明的是，文献[9]中各主成分的取值范围相差悬殊，为了避免聚类结果因量纲不同而剧烈变化，聚类之前需要对主成分原始数据进行无量纲化处理。由于变量正交的欧氏距离具有明确的空间距离特征，借鉴文献[8]将上市公司分为四类的思路，本文以 $q=2$ 欧氏距离为聚类统计量，采用离差平方和法(Ward法)将上市公司分为四类，如表2所示。

表2 上市公司创新能力聚类分析结果

聚类模型	第一类	第二类	第三类	第四类
主成分聚类	浪潮软件 同方股份 天坛生物 四创电子 大唐电信 大恒科技 中国船舶 航天动力 长电科技 振华重工	海信电器 上海汽车 华仪电气	中国建筑	长征电子 莱钢股份 南钢股份 现代制药 抚顺特钢 迪康药业 凤凰光学 宇通客车 青岛啤酒 金枫酒业 光明乳业 哈药集团 凌钢股份 广州药业 华阳科技 吉恩镍业
加权主成分聚类	浪潮软件 天坛生物 振华重工 中国建筑	同方股份 四创电子 上海汽车 中国船舶 大唐电信 大恒科技 长电科技	海信电器 迪康药业 华仪电气	长征电子 莱钢股份 南钢股份 现代制药 抚顺特钢 华阳科技 凤凰光学 宇通客车 青岛啤酒 金枫酒业 光明乳业 哈药集团 凌钢股份 广州药业 航天动力 吉恩镍业
主成分综合得分聚类	天坛生物 中国建筑	浪潮软件 同方股份 上海汽车 迪康药业 华仪电气 振华重工	莱钢股份 南钢股份 抚顺特钢 凤凰光学 金枫酒业 华阳科技	海信电器 四创电子 长征电子 现代制药 宇通客车 中国船舶 大唐电信 青岛啤酒 光明乳业 哈药集团 大恒科技 吉恩镍业 凌钢股份 广州药业 航天动力 长电科技
第一主成分聚类	浪潮软件 天坛生物 中国建筑 振华重工	同方股份 四创电子 上海汽车 中国船舶 大唐电信 大恒科技 长电科技	海信电器 长征电子 莱钢股份 现代制药 抚顺特钢凤凰光学 青岛啤酒 华阳科技	南钢股份 宇通客车 迪康药业 金枫酒业 光明乳业 哈药集团 华仪电气 凌钢股份 广州药业 航天动力 吉恩镍业

由于没有预先定义的类别标准来表明数据集中哪种期望关系是有效的，评判聚类模型的有效性必须定量分析和定性分析综合考虑。“可解释性”是评判模型分类质量的重要依据，聚类模型的优劣首先表现在能否对聚类结果做出合理的解释。表2的分类结果显示，四种聚类模型基本都能将浪潮软件、振华重工、天坛生物与其它上市公司分开，原因在于上述三个上市公司的创新综合评价值位居前三，其各主成分取值也远远领先其它上市公司，类别界限明显。但其余27个上市公司的各主成分取值相差不大，聚类空间狭窄而使得四种聚类模型的分类结果存在较大差异。特别的是，第一主成分聚类与加权主成分聚类的前两类结果完全一致，究其原因在于这两种聚类模型的聚类统计量中第一主成分信息比重占主导作用，在此极端情况下，加权主成分聚类与第一主成分聚类存有相似度较高的聚类结果。尤为注意的是，主成分聚类模型将中国建筑归类为创新水平落后于海信电器等上市公司的第三层次，但结合文献[9]中的原始数据发现，中国建筑的综合创新排名及多数主成分取值都领先海信电器等上市公司，所以将中国建筑划分为落后于海信电器的层级难以解释。出现上述极端聚类结果的原因在于，主成分聚类模型未曾区分不同主成分聚类效率的差异，使得分类质量下降。

为了对比各种聚类模型分类效率的优劣，本文方差分析的结果如表3所示。

表3 聚类模型方差分析对比

聚类模型	主成分	类间总离差	类内总离差	F—值
主成分聚类	F1	151.078	71.035	18.432***
	F2	2.904	1.709	14.729***
	F3	0.505	1.003	3.931**
	F4	0.575	1.069	4.659*
	F5	0.149	0.138	9.348***
加权主成分聚类	F1	188.619	33.494	48.806***
	F2	3.409	1.204	24.529***
	F3	0.541	1.076	4.85**
	F4	0.405	1.239	2.834*
	F5	0.071	0.217	2.837*
主成分综合得分聚类	F1	116.343	105.770	9.533***
	F2	1.394	3.219	3.753**
	F3	0.373	1.244	2.599*
	F4	0.548	1.095	4.339**
	F5	0.023	0.264	0.767
第一主成分聚类	F1	212.798	9.315	197.978***
	F2	0.145	4.468	0.281
	F3	0.207	1.410	1.272
	F4	0.344	1.300	2.292*
	F5	0.054	0.233	2.015

注：(1)、表中的F值为经自由度调整之后的组间方差与组内方差之比，组间离差、组内离差和总离差的自由度分别为3、26、29，F值越大，分类效果越好；(2)、***、**、*分别表示在1%、5%和10%的水平上差异显著。

表3中的对比结果显示，若以F值大小为标准：(1) 在信息含量最大的前三个主成分上，加权主成分聚类的分类质量明显优于主成分聚类；在信息含量较小的后两个主成分上，加权主成分聚类的分类质量稍逊主成分聚类。出现上述检验结果的原因在于，主成分聚类等同对待参与聚类的各个主成分，其算法目的仅是使各个主成分平衡地均能通过显著性检验即可，而加权主成分聚类则是按重要程度的差异，依次使用各主成分划分不同样品的所属类别，因此在信息含量较小的主成分上会出现分类效率劣于主成分聚类模型的情形，但加权主成分聚类模型的分类质量总

体检验效果显著。(2)主成分综合得分聚类与第一主成分聚类相比,在信息含量最大的第一主成分上,第一主成分聚类模型的分类效率明显优于主成分综合得分聚类模型,但由于第一主成分聚类模型未曾利用其它主成分的信息,其分类结果的其它主成分检验不显著。与之对比的是,尽管主成分综合得分平衡地使前四个主成分的分类显著性均通过检验,但其分类结果的方差显著性水平(F值)并不明显,也即类间离差相对类内离差并不显著。(3)比较而言,加权主成分聚类的分类结果不仅使得各主成分的检验效果显著,而且信息含量最大的前三个主成分F值相对其它模型的F值更大,说明加权主成分聚类模型较现有的三种主成分聚类模型分类效果明显提高。

综合加权主成分聚类模型的拓展机理及四种主成分聚类模型的分类质量对比结果可以发现,本文拓展的加权主成分聚类模型有机集成了多个分类理论和方法的长处,并且每一步都有充分的理论保证其必要性、合理性,有着同类分类评价问题下的普遍适用性。因此,新模型对样本的区分度更高,分类结果的"可解释性"更强,能够有效解决现有主成分聚类模型在极端情形下所不能解释的问题。

4 结束语

评价模型的层出不穷为评价问题的理论研究和实际应用提供了广阔的方法论选择空间,但是如果对经典评价方法的理论基础、适用性前提以及存在的缺陷缺乏深入理解,盲目地追求模型改进则可能陷入评价方法研究的误区。本文研究的内容主要是对主成分综合评价环节的一般性问题进行了简要梳理探讨,并在现有研究成果的基础上对主成分聚类模型进行了拓展。实证分析表明,改进的加权主成分聚类相对原始的主成分聚类分类效果更佳,评价结论的可靠性更高。

参考文献:

[1]李靖华,郭耀煌.主成分分析用于多指标评价的方法研究[J].管理工程学报,2002,16(1).

[2]王宗军.综合评价的方法、问题及其研究趋势[J].管理科学学报,1998,1(1).

[3]陈国宏,李美娟. 基于方法集的综合评价方法集化研究[J].中国管理科学,2004,12(1).

[4]吕红,王芳.中国制造业科技创新能力的国际比较 [J].统计与决策,2010,(18).

[5]陈宏明,史亚男.上市公司内部控制信息披露的影响因素研究[J].统计与决策,2011, (8).

[6]匡后权,吉松涛. 基于主成分BP神经网络的西部服务业产值预测[J]. 统计与决策,2011,(11).

[7]刘倩.基于主成分聚类分析的中小企业成长性研究[J]. 统计与决策,2011,(16).

[8]胡彦蓉,吴冲,刘洪久,李梅.基于主成分集成方法的上市公司创新能力评价研究[J].运筹与管理,2012,21(5).

[9]李巍巍,吴冲.上市公司财务绩效的改进集成评价研究[J].运筹与管理,2012,21(1).

[10]何晓群. 多元统计分析(第三版)[M].北京:中国人民大学出版社,2012.

[11]陆根尧,盛龙,唐辰华.中国产业生态化水平的静态与动态分析—基于省级数据的实证研究[J].中国工业经济, 2012, (3).

[12]袁建新,刘幸赟.技术引进促进经济增长作用省际差异性影响因素分析[J].中国工业经济, 2010, (5).

[13]王德青,朱建平,谢邦昌.主成分聚类分析有效性的思考[J].统计研究, 2012,29(11).

[14]徐雅静,汪远征.主成分分析应用方法的改进[J].数学的实践与认识,2006,(6).

[15]孟生旺.用主成分分析法进行多指标综合评价应注意的问题[J].统计研究, 1992,29(4).

[16]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,(5).

(责任编辑/亦 民)