文章编号: 1001- 2486(2008) 03- 0081- 05

基于核主成分分析的高校科技创新能力评价研究*

吕 蔚[1,2], 王新峰[2], 孙智信[2]
(1. 北京理工大学 管理与经济学院, 北京 100081; 2. 国防科技大学 信息系统与管理学院, 湖南 长沙 410073)

摘 要: 高校在国家创新体系中占有举足轻重的地位, 高校科技创新能力评价研究对提高高校科技创新能力具有积极的意义。基于核主成分分析的评价方法能够有效去除高校科技创新能力评价指标体系中的非线性相关信息, 从而取得较好的评价结果。以 15 所教育部直属高校科研统计数据为样本, 利用核主成分分析法进行了高校科技创新能力评价的实证分析, 并与主成分分析结果进行了比较, 结果表明, 核主成分分析能够取得更高的主特征值累积贡献率, 从而产生更为合理的评价结果。

关键词: 科技创新能力评价; 核主成分分析; 评价指标
中图分类号: C931 文献标识码: A

Application of the Kernel Principal Component Analysis Method to University S&T Innovation Capability Evaluation

LU Wei[1,2], WANG Xin-feng[2], SUN Zhi-xin[2]
(1. School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China;
2. College of Information System and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Universities play an important role in the national innovation system, and the research on the university S&T innovation capability evaluation is of great significance. Kernel principal component analysis (KPCA) method is proposed for the evaluation of university S&T innovation capability. This method, in comparison with Principal Component Analysis (PCA) method, can solve nonlinear correlation problem of evaluation index in the analysis of the S&T data released in 2002 from 15 universities directly subordinate to Chinese Ministry of Education. In particular, the result shows that contribution of the first and second principle components are more concentrated by KPCA than by PCA, and KPCA has better evaluation performance than PCA.

Key words: S&T innovation capability evaluation; kernel principal component analysis (KPCA); evaluation index

高校在国家科技创新体系中发挥着重要的作用, 是重要的技术创新源。客观、科学地评价高校科技创新能力, 对提出提升高校科技创新能力的对策, 优化高校科技资源配置, 提高高校科技创新能力, 具有非常重要的理论和现实意义。

国外院校的科技创新能力评价研究主要出现在院校评估与绩效评估中。我国对高校科研绩效和办学实力的排行性评价目前正在成为研究热点。王章豹等完成了"高校科研排行性评价科技创新能力评价指标设计"研究[1]; 中国科学评价研究中心的邱均平等利用美国出版的《基本科学指标》(ESI) 作为原始数据来源, 对世界大学和科研院所的科研竞争力进行了评价和排名[2]等。在评价方法上, 王章豹等采用线性加权求和法计算综合评价得分[1]; 敖慧等研究了多级模糊综合评价在高校科技创新能力评价中的应用[3]; 王晓红等应用改进的 E-DEA 模型, 给出各个大学科研绩效评估及排名[4]。虽然目前国内外对评价方法的研究可谓硕果累累, 但在能力评价实践中对新的评价方法的应用却非常有限。在评价实践中, 仍然是一些较为经典的评价方法起主导作用。

在高校科技创新能力评价中, 经常使用的一种方法是主成分分析法(PCA)。如张浩等曾以 15 所教育部直属高校为样本, 利用主成分分析法进行了高校科技创新能力评价的实证分析[5]。这是因为在高

* 收稿日期: 2007- 12- 11
基金项目: "十一五"全国教育科学规划军队重点课题资助项目(PLA061003)
作者简介: 吕蔚(1973—), 女, 助理研究员, 在职博士生。

校科技创新能力评价中,一般选择的评价指标较多且指标间有一定的相关性,因此所得到的统计数据反映的信息有一定程度的重叠,增加了评价的复杂性。而PCA可利用几个不相关的主成分作为原来众多变量的线性组合,在保留了原始变量大部分信息的基础上,减少了计算量,综合评价时更简洁,因此在高校科技创新能力评价中得到了广泛应用。但是,PCA只能去除评价指标之间的线性相关信息,忽略了多个评价指标间的非线性相关性。而在人为设计的高校科技创新指标体系中,各个指标之间往往表现出非线性,这时采用PCA方法进行评价,就不能取得较好的结果。核主成分分析(KPCA)方法不仅特别适合于处理非线性相关问题,且能提供更多的特征信息。KPCA通过某种事先选择的非线性映射 Φ 将输入矢量 $\boldsymbol{x}$ 映射到一个高维特征空间 $\boldsymbol{F}$,从而使输入矢量具有更好的可分性,然后对高维空间中的映射数据做PCA,从而得到数据的非线性主成分。根据核学习理论[6],只要能选择满足Mercer条件的核函数 $\boldsymbol{K}=\Phi\cdot\Phi$,就可以在这个特征空间中进行运算,求得特征空间的主成分,实现评价对象的综合评价。

1 高校科技创新能力评价指标体系的选择

结合教育部对高校科技竞争力评价所用的评价指标,并参考国内许多专家学者提出的高校科技创新能力评价指标体系[1-5,7],选择确立了一种较为常用、较易被人接受的指标体系。

把高校科技创新能力分解为3个主要因素:科技创新资源投入能力、知识创新能力以及科技创新支撑能力。正是它们之间的相互影响导致整个高校科技创新能力螺旋式上升。在高校科技创新能力的3个分能力中,资源投入能力是指高校人力和财力投入科技创新资源的数量和质量,是完成科技创新的必要条件;知识创新能力是高校所有能产生新知识的科学研究活动,是增加整个人类知识财富的能力,在高校主要可体现为科学研究成果和科技成果转让情况两个方面;创新支撑能力通过高校学术研究环境的营造,为科技创新的进行奠定基础,它主要包括学术资源、国内外科技合作交流等方面的质量和数量。最后,对高校科技创新能力评价指标设置一级指标3个、二级指标6个,三级指标14个,如表1所示。

表1 高校科技创新能力评价指标体系

Tab.1 Universities S&T innovation capability evaluation index

目标层	准则层	影响因素	指标因子	单位	编号
高校科技创新能力	科技创新资源投入能力	人力资源投入	全校科研全年人员数	人	E_1
			副教授以上人员比例	%	E_2
		财力资源投入	全校科研经费年度筹集总额	千元	E_3
			科研经费中政府投入比例	%	E_4
	知识创新能力	研究与开发成果	年度承担科研项目总数	项	E_5
			获得国家及省部级科技奖励数	项	E_6
			申请国内外专利数	项	E_7
			已获授权国内外发明专利数	项	E_8
		科技成果转让情况	当年转让合同数	项	E_9
			当年科技成果转化实际收入	千元	E_{10}
	科技创新支撑能力	学术资源	年度研发全时人员人均项目经费	千元	E_{11}
			发表学术论文数	篇	E_{12}
		国内外科技合作与交流	年度举办国际学术会议次数	次	E_{13}
			年度派遣和接受进修访问学者人次	人次	E_{14}

2 基于核主成分分析的创新能力综合评价

设评价对象个数为 n,评价指标个数为 p,则 n 个对象的指标值组成样本数据矩阵为 y,$y_i=\{y_{i1},y_{i2},\cdots,y_{ij}\}(i=1,2,\cdots,n;j=1,2,\cdots,p)$。

为了排除数量级和量纲不同带来的影响,首先对原始数据进行标准化处理,

$$x_{ij}=\frac{(y_{ij}-m_j)}{\sigma_j}\quad(i=1,2,\cdots,n;j=1,2,\cdots,p) \tag{1}$$

其中, m_j, σ_j 分别为第j 个指标向量$(y_{1j}, y_{2j}, \cdots, y_{nj})$的均值和方差, 这样形成新的数据矩阵 $\boldsymbol{x}$, $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \cdots, x_{ij}\}$, $(i = 1, 2, \cdots, n; j = 1, 2, \cdots, p)$。

设变换 Φ实现了样本空间$\boldsymbol{R}^p$ 到特征空间$\boldsymbol{F}$ 的映射, 即样本数据 x_i 在F 空间的像为 $\Phi(x_i)$, 则映射数据的协方差矩阵为

$$\boldsymbol{C} = \frac{1}{n}\sum_{i=1}^{n}\Phi(\boldsymbol{x}_i)\,\Phi(\boldsymbol{x}_i)^{\mathrm{T}} \tag{2}$$

对 $\boldsymbol{C}$ 求特征 $\lambda(\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant 0)$ 和特征向量 $\boldsymbol{V} \in \boldsymbol{F}$, 则

$$\boldsymbol{C}\boldsymbol{V} = \lambda\boldsymbol{V} \tag{3}$$

然后进行核变换, 即对每个样本与该式求内积, 得

$$\lambda(\Phi(\boldsymbol{x}_i)\cdot\boldsymbol{V}) = \Phi(\boldsymbol{x}_i)\cdot\boldsymbol{C}\boldsymbol{V}, \quad i = 1, \cdots, n \tag{4}$$

特征向量矩阵 $\boldsymbol{V}$可以用 $\Phi(x_i)$ 表示为

$$\boldsymbol{V} = \Phi(\boldsymbol{x})\,\alpha = \sum_{j=1}^{N}\alpha_j\Phi(\boldsymbol{x}_j) \tag{5}$$

式中, $\Phi(\boldsymbol{x}) = (\Phi(\boldsymbol{x}_1), \cdots, \Phi(\boldsymbol{x}_n))$, $\alpha = (\alpha_1, \cdots, \alpha_n)^{\mathrm{T}}$, 则代入式(2) 有

$$\lambda\sum_{j=1}^{n}\alpha_j[\Phi(\boldsymbol{x}_k)\cdot\Phi(\boldsymbol{x}_j)] = \frac{1}{n}\cdot\sum_{j=1}^{n}\alpha_j\Phi(\boldsymbol{x}_k)\cdot\sum_{i=1}^{n}\Phi(\boldsymbol{x}_j)[\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x}_i)] \tag{6}$$

定义 $n \times n$ 矩阵$\boldsymbol{K}$, $K_{ij} = \Phi(x_i)\,\Phi(x_j)$, 注意到 $\boldsymbol{K}$ 是一个对称阵, 式(5) 可写为

$$n\lambda\alpha = \boldsymbol{K}\alpha \tag{7}$$

一般映射数据为非零均值的, 这时可以通过修正式(7) 得到

$$\lambda\alpha = \left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{I}_{n\times n}\right)\boldsymbol{K}\alpha \tag{8}$$

式中, $\boldsymbol{I}_n$ 为n 维大小的单位矩阵, $\boldsymbol{I}_{n\times n}$表示各元素为 1 的 $n \times n$ 维矩阵。对式(8) 求解, 获得要求的特征值 $\lambda(\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant 0)$ 和特征向量 $\boldsymbol{V}$。测试样本 $\boldsymbol{x}$ 在$\boldsymbol{F}$ 空间向量 $\boldsymbol{V}^k$ 的投影为

$$\boldsymbol{V}^k\cdot\Phi(\boldsymbol{x}) = \sum_{i=1}^{n}\alpha_i^k[\Phi(\boldsymbol{x}_i)\cdot\Phi(\boldsymbol{x})] \tag{9}$$

按照前 m 个非线性主成分的累积贡献率大于 85% 的原则选取非线性主成分, 确定前 m 个主成分, 则测试样本的综合评价函数为

$$\boldsymbol{J} = \sum_{i=1}^{m}\boldsymbol{V}^i\cdot\Phi(\boldsymbol{x}) = \sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_j^k[\Phi(\boldsymbol{x}_j)\cdot\Phi(\boldsymbol{x})] \tag{10}$$

由上述推导过程可知, KPCA 实际上就是在特征空间中进行标准的主成分分析。当核函数选择采用线性核函数时, 它就是标准的PCA。因此, KPCA 保留了主成分分析的优点并具有处理非线性的能力。首先, 它在特征空间是一种正交变换, 所得非线性主成分是互不相关的。其次, 非线性主成分是按照特征值从大到小自动排序的, 因此, 前 m 个主成分携带了最多的数据结构变异信息。第三, 用前 m 个主成分代表样本所得的均方估计误差最小。

3　评价案例分析

选择 2002 年度 15 所教育部直属高校的统计数据[8] 作为样本进行科技创新能力的综合评价案例分析。样本数据见表 2。

对样本数据进行 KPCA, 并与 PCA 方法进行比较。KPCA 核函数选择多项式核函数, 相应的参数 $d = 3$, 即核函数为

$$K(\boldsymbol{x}, \boldsymbol{y}) = (1 + \boldsymbol{x}\cdot\boldsymbol{y})^3 \tag{11}$$

表2 15所教育部直属高校科技创新能力评价指标数据

Tab.2 S&T innovation capability evaluation data of university directly under ministry of education

编号	指标	E_1	E_2	E_3	E_4	E_5	E_6	E_7
1	武汉大学	1750	48.46	205 400	66.08	887	54	128
2	华中科技大学	1517	54.71	350 262	43.14	1404	37	116
3	中国地质大学	1016	50.39	101 044	44.87	456	18	11
4	武汉理工大学	992	56.45	139 459	25.73	652	16	15
5	华中农业大学	560	46.96	82 191	89.51	450	22	18
6	复旦大学	1224	49.84	218 542	51.94	1217	37	180
7	同济大学	1272	45.28	423 720	31.70	1580	32	38
8	上海交通大学	1283	63.76	628 062	52.52	1809	34	198
9	华东理工大学	323	49.54	119 898	44.45	377	11	61
10	南京大学	716	69.55	150 191	60.63	772	19	38
11	东南大学	1210	53.97	272 754	37.62	890	23	75
12	浙江大学	1476	59.76	693 988	31.77	3491	94	177
13	合肥工业大学	718	52.51	93 269	53.97	434	5	5
14	厦门大学	538	55.20	38 372	69.35	470	17	21
15	山东大学	915	65.79	95 226	83.84	770	74	41

编号	指标	E_8	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}
1	武汉大学	33	38	7000	53.25	3279	249	380
2	华中科技大学	60	75	4447	135.01	5891	141	558
3	中国地质大学	6	3	5000	47.93	1376	92	31
4	武汉理工大学	9	7	300	96.26	1424	53	14
5	华中农业大学	0	11	1046	52.36	828	34	104
6	复旦大学	35	3	18 540	74.70	2126	122	46
7	同济大学	22	65	6520	171.92	2789	208	922
8	上海交通大学	38	485	128 871	359.48	4003	183	52
9	华东理工大学	6	32	5809	171.54	1157	35	68
10	南京大学	13	4	3312	126.74	1979	208	234
11	东南大学	24	56	20 800	139.88	2223	161	93
12	浙江大学	45	132	29 726	238.18	6223	252	387
13	合肥工业大学	2	41	3120	56.86	932	10	29
14	厦门大学	0	4	2026	54.97	1360	145	41
15	山东大学	1	16	1810	53.72	2514	134	49

KPCA和PCA特征值、方差贡献率和方差累计贡献率如表3所示。

表3 KPCA和PCA特征值的贡献率(%)

Tab.3 Rate of KPCA and PCA eigenvalue

	No	特征值	贡献率(%)	累计贡献率(%)		No	特征值	贡献率(%)	累计贡献率(%)
	1	105.1349	53.64	53.64		1	85 935	56.68	56.68
	2	31.8809	16.27	69.91		2	51 840	34.19	90.88
	3	20.5948	10.51	80.41		3	3495	2.31	93.18
PCA	4	11.9110	6.08	86.49	KPCA	4	2690	1.77	94.95
	5	9.1076	4.65	91.14		5	2075	1.37	96.32
	6	7.0901	3.62	94.76		6	1667	1.10	97.42
	…	…	…	…		…	…	…	…

从表3可以看出,采用PCA的前4个特征值累积贡献率分别为(53.64, 69.91, 80.41, 86.49),而采用KPCA方法前4个特征值累积贡献率分别为(56.68, 90.88, 93.18, 94.95),获得了比PCA更好的降维效果。

KPCA 和 PCA 一样, 选取主成分的数目也是一个值得研究的问题。适当的主元数目不仅要简化计算, 还要保证可以对工作过程进行充分描述。主元太少, 则所得主元特征不能最大可能地反映工作过程的变化; 主元太多时又会加大噪声干扰。在实际应用中, 一般按照前 m 个非线性主成分的累积贡献率大于 85% 的原则选取。由表 3 知, PCA 需要选择前 4 个主成分进行评价, 而 KPCA 只需选择前 2 个非线性主成分代表原来的 14 个指标综合评价高校的科技创新能力。由公式(10) 计算样本的综合评价值。由 KPCA 求出各所高校科技创新能力的综合得分以及排序如表 4 所示。

表 4　高校科技创新能力综合排名

Tab. 4　General rank of university S&T innovation capability

编号	高校名称	KPCA	
		得分 J	排名
1	武汉大学	- 2329	5
2	华中科技大学	- 1367	3
3	中国地质大学	- 3059	12
4	武汉理工大学	- 2691	10
5	华中农业大学	- 4465	15
6	复旦大学	- 2400	7
7	同济大学	- 2244	4
8	上海交通大学	20 022	1
9	华东理工大学	- 3027	11
10	南京大学	- 2443	8
11	东南大学	- 2397	6
12	浙江大学	16 253	2
13	合肥工业大学	- 3954	14
14	厦门大学	- 3381	13
15	山东大学	- 2519	9

4　结 论

高校科技创新能力评价属于高校科研管理的范畴, 涉及评价因素很多, 这些因素存在着不确定的非线性相关关系。本文提出了基于核主成分分析的综合评价方法, 并应用于 2002 年 15 所教育部直属高校科研统计数据, 从结果来看, 由于 KPCA 提取的主特征值累积贡献率优于 PCA 提取的主特征值贡献率, 只需选择前 2 个非线性主成分进行综合评价, 因而获得了更好的降维效果, 提高了评价的科学性与准确性。

参 考 文 献:

[1] 王章豹, 徐枞巍, 等. 高校科研排行性评价与科技创新能力评价指标设计[J]. 合肥工业大学学报(社会科学版), 2005, 19(1): 1- 8.

[2] 邱均平. 美国《科学引文索引》与科研绩效评价[J]. 科研管理, 2003(4): 22- 27.

[3] 敖慧. 高校科技创新能力的多级模糊综合评价[J]. 武汉理工大学学报(信息与管理工程版), 2004, 6(6): 169- 171.

[4] 王晓红. 基于投入产出分析的科研绩效评价理论模型及方法研究[D]. 哈尔滨: 哈尔滨工业大学论文, 2004.

[5] 张浩, 冯林. 主成分分析法在高校科技创新能力评价中的应用[J]. 武汉理工大学学报(信息与管理工程版), 2004, 26(6): 157- 161.

[6] Scholkopf B, Smola A, M ller K R. Kernel Principal Component Analysis[C]// Proceedings of ICANN, LNCS, Springer, 1997: 583- 589.

[7] 何晋秋, 孙志军. 创新能力及高校科技创新能力[C]// 教育部科技委管理科学部"高校科技管理与创新能力建设论坛"论文集, 南昌, 2006.

[8] 中华人民共和国教育部科学技术司. 2002 年高等学校科技统计资料汇编[M]. 北京: 高等教育出版社, 2002.