

DOI :10.3880/j.issn.1006-7647.2009.04.010

核主成分分析的参数优化及其在水资源可持续利用综合评价中的应用

李 磊¹,金菊良²,梁忠民¹

(1. 河海大学水文水资源学院,江苏 南京 210098 ;2. 合肥工业大学土木与水利工程学院,安徽 合肥 230009)

摘要 针对核主成分方法应用时核函数参数确定困难,提出建立基于贡献率的优化问题,并用实码加速遗传算法对其实施优化求解,客观地确定了各参数的值,并将该核主成分方法应用于水资源可持续利用综合评价问题中,取得了良好的效果。

关键词 水资源可持续利用 综合评价 核主成分分析 参数优化 加速遗传算法

中图分类号 :TV213.9 **文献标识码** :A **文章编号** :1006-7647(2009)04-0036-03

Parameters optimization of kernel principal component analysis and its application in evaluation of water resources sustainable utilization//LI Lei¹, JIN Ju-liang², Liang Zhong-min¹ (1. College of Water Resources and Hydrology, Hohai University, Nanjing 210098, China ;2. College of Civil Engineering, Hefei University of Technology, Hefei 230009, China)

Abstract :According to the choice of kernel function parameters in the kernel principal component method, an optimization question based on the contribution rate is established. The real coded accelerating genetic algorithm is used to find an optimization solution. The values of the parameters are determined objectively. The kernel principal component method was used to evaluate and analyze water resources sustainable utilization, and a satisfactory result was obtained.

Key words :water resources sustainable utilization ; comprehensive evaluation ; kernel principal components analysis ; parameter optimization ; accelerating genetic algorithm

主成分分析(principal component analysis,PCA)法利用降维的思想,在损失很少信息的前提下把多个指标转化为几个综合指标,以其理论和实践的简洁性、所得结果的客观性等特点,广泛应用于经济、社会、科教等领域中众多对象的综合评价。该方法对于处理线性问题能取得很好的效果,但当指标之间为非线性相关时往往会出现降维效果不明显等问题,这是由其基本原理决定的^[1]。核思想和核函数是一种非常有效的设计非线性算法的数学手段,随着支持向量机的提出备受注目,近 10 多年来广泛应用于改进各种经典的线性算法,被称为线性算法的核化^[2]。核思想是将原非线性空间映射到一个高维空间,在高维的空间中设计或应用已有线性算法处理问题^[2-4]。Scholkopf 等^[5]于 1998 年提出的核主成分分析(Kernel-PCA)就是基于这种思想,在增加少量的计算复杂度情况下实现了一种非线性的主成分分析法。核主成分分析自提出以来逐步被引入了各个领域^[5-9],取得了一定的效果。但在应用核中,核参数的确定一直是个难题,而它直接影响到算法的效

果。关于核函数参数的选择已有不少研究^[10],但多是基于支持向量机的,对于核主成分分析的核参数选择仍以经验为主。本文针对核主成分分析法效率较高的特点,即当参数选择适当时第一主成分的贡献率已经可以满足需要,拟通过使第一主成分贡献率最大构造优化问题,用基于实码加速遗传算(real coding based accelerating genetic algorithm,RAGA)法^[11]对核主成分分析参数进行优化选择,实现客观的最优参数选择,提出了基于实码加速遗传算法的核主成分分析法(RAGA-KPCA 法),并应用于中国北方地区水资源可持续利用综合评价中。

1 Kernel-PCA 法的基本原理

1.1 传统的 PCA 法

设有 m 个评价对象、 n 个评价指标构成样本矩阵为 $X = [x_1, x_2, \cdots, x_m]$ 。假设评价对象已进行中心化,即 $\sum_{i=1}^m x_i = 0$,则样本的协方差矩阵为 $C_{n \times n} = \frac{1}{m} \sum_{j=1}^m x_j x_j'$ 。

作者简介 李磊(1981—),男,安徽巢湖人,博士研究生,从事水文不确定性及防灾减灾研究。E-mail :lileilily@tom.com

根据 $\lambda_i v_i = C v_i$ 可计算出第 i 主成分 $y_i = v_i' X$, 可得到 n 个新的综合评价指标, 由于 $D(y_i)$ 可以表达为式(1), 所以当 λ_i 按从大到小排列时, y_i 反映第 i 个方差最大方向, 其所反映的原指标信息即为第 i 多。

$$D(y_i) = D(v_i' X) = v_i' D(X) v_i = v_i' C v_i = \lambda_i v_i' v_i \tag{1}$$

式中 v_i 为标准化后的向量, 则 $D(y_i) = \lambda_i$ 。

按照累积方差贡献率大于或等于 85% 的准则, 取前 p 个主成分进行综合评价, 以各主成分贡献率

$$\omega_j = \lambda_j / \sum_{i=1}^n \lambda_i \text{ 为权系数, 则综合评价函数为 } f_i = \sum_{j=1}^p \omega_j v_j' x_i$$

若原指标间线性相关关系较强时, p 值可能较小, 则采用少数几个主成分即可概括原指标绝大部分信息, 指标得到降维。但当指标非线性相关时, p 值较大, 降维效果较差, 此时需要适合处理非线性的方法来进行降维。

1.2 Kernel-PCA 法^[5]

传统主成分分析只涉及向量内积, 故可引入核函数对原主成分方法进行改进以处理非线性问题。其目的是采用核思想对原指标进行升维, 在新的高维特征空间进行线性主成分分析。

如前, 评价对象向量 $x_i \in R^n (i = 1, 2, \dots, m)$, 存在非线性投影 $\Phi: R^n \rightarrow F, x \mapsto \Phi(x)$ 。设映射已中心化, 即 $\sum_{i=1}^n \Phi(x_i) = 0$, 则特征空间的协方差矩阵

$$C' = \frac{1}{m} \sum_{j=1}^m \Phi(x_j) \Phi(x_j)'$$

在特征空间中应用线性 PCA 求解 $\lambda v = C' v$, 得特征值 $\lambda \geq 0$ 和特征向量 $v \in F \setminus \{0\}$, 且 $\lambda (\Phi(x_k) \cdot v) = (\Phi(x_k) \cdot C' v) \quad (k = 1, 2, \dots, m)$ (2)

由于 v 属于由 $\Phi(x_1), \Phi(x_2), \dots, \Phi(x_m)$ 构成的生成子空间, 所以必存在参数 $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_m)$ 满足

$$v = \sum_{i=1}^n \alpha_i \Phi(x_i) \tag{3}$$

将式(2)、式(3)代入 $\lambda v = C' v$ 可得 $m \lambda \alpha = K \alpha$, 其中 $K = (k_{ij})_{m \times m} = (\Phi(x_i) \cdot \Phi(x_j))$ 。

由 $m \lambda \alpha = K \alpha$ 计算出 λ_k 及其对应的向量 $\alpha^k (k = 1, 2, \dots, m)$ 。

设在 $\lambda v = C' v$ 中对应于 λ_k 的特征向量为 v^k , 则对任意待评价对象 x 在 F 中的映射 $\Phi(x)$:

$$(v^k \cdot \Phi(x)) = \sum_{i=1}^m \alpha_i^k (\Phi(x_i) \cdot \Phi(x)) = \sum_{i=1}^m \alpha_i^k K(x_i, x) \tag{4}$$

式中 α_i^k 为 v^k 对应的参数向量 α^k 的第 i 个分量。

按累积方差贡献率大于或等于 85% 的标准, 取前 p 个主成分进行综合评价, 以各主成分贡献率为权系数, 则综合评价函数为

$$f = \sum_{k=1}^p \omega_k (v^k, \Phi(x)) = \sum_{k=1}^p \omega_k \sum_{i=1}^m \alpha_i^k K(x_i, x) \tag{5}$$

以上推导是在假设 $\sum_{i=1}^n \Phi(x_i) = 0$ 的情况下进行的, 当不满足时需进行中心化处理。

由 $\tilde{\Phi}(x_i) = \Phi(x_i) - \frac{1}{m} \sum_{i=1}^m \Phi(x_i)$ 可推出:

$$\begin{cases} \tilde{k}_{ij} = k_{ij} - \frac{1}{m} \sum_{s=1}^m k_{sj} - \frac{1}{m} \sum_{l=1}^m k_{il} + \frac{1}{m^2} \sum_{s=1}^m \sum_{l=1}^m k_{sl} \\ \tilde{K} = K - L_m K - K L_m + L_m K L_m \end{cases} \tag{6}$$

其中 $(L_m)_{ij} = \frac{1}{m}$

在计算时需要对 v^k 进行规范化, 即

$$1 = (v^k \cdot v^k) = \sum_{i=1}^m \alpha_i^k \alpha_j^k (\Phi(x_i) \cdot \Phi(x_j)) =$$

$$\sum_{i,j=1}^m \alpha_i^k \alpha_j^k \tilde{k}_{ij} = (\alpha^k \cdot K \alpha^k) = \lambda_k (\alpha^k \cdot \alpha^k) \tag{7}$$

核函数是满足 Mercer 定理的非线性函数, 它可以实现某一个高维特征空间中两向量的内积运算。常用的核函数有:

多项式核函数

$$K(x, x_i) = [a(x, x_i) + b]^q \tag{8}$$

Gauss 核函数

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{9}$$

Sigmoid 核函数

$$K(x, x_i) = \tanh(v(x, x_i) + c) \tag{10}$$

从上面推导可以看出, 从 $C_{n \times n} \rightarrow C' \rightarrow K_{m \times m}$, KPCA 实现了从 n 维空间升高到一个高维特征空间 F , 在无需知道 F 维数的情况下转化在 m 维空间中计算内积。当 $m > n$ 时, 只增加少量的维数和计算量就实现了 PCA 的非线性化。

KPCA 评价的一般步骤如下: ①选择适当的参数代入选择的核函数, 得到核函数 $K(x, x_i)$; ②由评价样本计算 $K_{m \times m}$ 并进行中心化; ③由 $m \lambda \alpha = K \alpha$ 计算出 λ_k 和 α^k , 计算出对应的贡献率 $w_k (k = 1, 2, \dots, m)$, 并进行归一化; ④根据式(5)进行综合评价。

2 RAGA-KPCA 法

上述几种核函数都存在未知参数, 不同参数对 KPCA 法的效果有较大影响, 参数选取适当可以尽

可能使用少的主成分来概括数据本身信息,使降维效果更明显。一般当 KPCA 法的参数选择适当时,第一主成分的贡献率就已经超过 85%,然而目前这些参数的确定基本凭经验。实际上主成分的贡献率,也即核函数矩阵的特征值与参数之间具有一一对应的非线性关系,因此笔者以第一主成分贡献率为目标函数,以核函数的参数为优化变量建立优化问题来选择合适的核参数。显然该问题是一个复杂的非线性的优化问题,常规的方法很难求解。模拟生物优胜劣汰规则与群体内部染色体信息交换机制的 RAGA 法是一种通用的全局优化新方法,对这类问题可以简便有效求解^[11]。将 RAGA 与 KPCA 评价方法相结合,以 $f(i) = \max(w_1)$ 为适应度函数,多次迭代即可得到核函数的最优参数。将最优参数代入核函数,按 KPCA 法评价的一般步骤进行评价。

3 基于 RAGA-KPCA 法的北方地区水资源可持续利用综合评价

如何评价水资源可持续利用程度是目前国内外研究的前沿和热点之一,现有的可持续评价方法多是通过建立综合评价指标体系进行综合评价。文献[12]建立了北方地区水资源可持续利用综合评价指标体系,并对其采用常规的主成分分析法进行综合评价。文献[12]中提取了前 4 个主成分,累积贡献率为 88.82%,降维效果并不很明显。这里采用 RAGA-KPCA 法对其进行了评价,见表 1。

KPCA 法进行评价常使用多项式核函数 $K(x, x_i) = [a(x, x_i) + b]^q$ 。经过多次数据试验,参数 a ,

b, q 对应的搜索空间分别取 $[0,1], [4,7], [2,3]$,用 RAGA 法优化可以得到最优核函数参数 $a = 0.746\ 675, b = 5.639\ 153, q = 2.438\ 400$,第一主成分对应的贡献率 $w_1 = 0.960\ 264$,综合评价结果见表 1。

由表 1 可知,文献[12]中提取前 4 个主成分才使累积贡献率达到 88.82%,损失了很多信息,而且在综合时采用以贡献率为权重进行加权,以特征值对应的特征向量进行综合,而对于同一特征值所对应的特征向量不唯一,这些都有可能导致误差。而 RAGA-KPCA 法第一主成分贡献率为 96.03%,降维效果非常明显,直接以第一主成分进行综合,损失信息量和误差要小得多。从表 1 最后 2 列评价结果可以看出排名前 6 位和后 5 位的省(市、自治区)顺序没有发生变化,中间几个省(市、自治区)的排名略有差别,由于本文方法采用的是第一主成分,结果可信度更高。评价结果表明 新疆、宁夏、内蒙古、甘肃等省(市、自治区)可持续发展程度较好,而天津、北京、山西、山东等省(市、自治区)则相对较差、排名较后。这是由于前者人均水资源量较丰富以及经济发展水平不高造成的。

4 结 论

- a. RAGA 法对 KPCA 法中核函数参数进行确定,克服了原方法确定参数时依赖个人经验的缺点,可以精确地得到最优的参数。
- b. KPCA 法作为 PCA 法在非线性问题的一种推广,笔者将其应用于北方地区水资源可持续利用综合评价中,结果显示它的降维效果非常明显,克服了原有主成分存在的不足,且结果比 PCA 法可信度更高。

表 1 水资源可持续利用综合评价样本^[12]和评价结果

省(市、自治区)	I_1 / 亿 m^3	I_2 / 亿 mm	I_3 / 亿 m^3	I_4 / 亿 m^3	I_5 / 亿 m^3	I_6 / ($m^3 \cdot 人^{-1}$)	I_7 / (亿 $m^3 \cdot 万元^{-1}$)	I_8 / 万 hm^2	I_9	综合排名	
										PCA ^[12]	KPCA
北京	43.3	656	38.03	43.21	343	22.29	267	31.63	1	16	16
天津	22.5	596	10.54	23.36	246	13.00	212	38.07	4	17	17
河北	311.7	601	51.32	207.13	319	159.46	600	424.81	2	9	12
山西	152.8	576	8.60	57.66	184	40.17	437	108.60	4	15	15
内蒙古	638.9	296	23.35	152.60	661	98.24	1 550	170.69	3	3	3
辽宁	416.0	704	49.84	143.40	348	79.26	454	122.80	2	13	13
吉林	366.9	552	171.34	101.47	389	54.66	759	70.09	1	7	9
黑龙江	698.0	490	30.88	268.04	719	160.82	1 116	150.31	2	6	6
江苏	428.3	1 059	57.28	289.25	547	190.06	648	327.17	2	8	8
安徽	991.9	1 415	24.82	192.32	317	116.27	822	302.47	1	12	7
山东	271.1	661	41.85	261.21	294	116.08	431	453.06	1	14	14
河南	525.3	865	39.41	241.85	264	133.78	657	422.47	1	11	10
陕西	343.1	629	28.43	81.72	230	56.38	694	128.74	2	10	11
甘肃	191.9	270	40.33	120.91	490	70.17	1 693	89.00	3	4	4
青海	562.0	247	85.51	27.08	555	17.22	1 475	20.50	2	5	5
宁夏	13.3	325	0.59	91.85	1 728	32.99	4 649	32.41	1	2	2
新疆	893.4	212	11.33	437.16	2 588	324.87	4 793	291.66	1	1	1

注： I_1 表示地区的水资源总量； I_2 表示年降雨量； I_3 表示大中型水库蓄水量； I_4 表示供水总量； I_5 表示耗水总量； I_6 表示人均用水总量； I_7 表示单位 GDP(国内生产总值)用水量； I_8 表示实际灌溉面积； I_9 表示水质等级。

(下转第 66 页)

床纵剖面应及早认真研究,进行科学的论证。

参考文献：

[1] 毛世民. 淮河特点:七大江河特征的比较[C]//水利部淮河水利委员会. 淮河研究会第四届学术研讨会论文集. 北京:中国水利电力出版社,2004:474-477.
[2] 张义丰,李良义,钮仲勋. 淮河地理研究[M]. 北京:测绘出版社,1993:150-169.
[3] 朱松泉,窦鸿身. 洪泽湖[M]. 合肥:中国科学技术大学

出版社,1993.

[4] 毛世民,金正越. 淮河干流入湖河口段的洪水水面比降[J]. 治淮,1999(12):36-38
[5] 向茂森. 论淮河流域水资源配置[C]//水利部淮河水利委员会. 淮河研究会第四届学术研讨会论文集. 北京:中国水利水电出版社,2004:166-171.
[6] 国家防汛抗旱总指挥部办公室,水利部南京水文水资源研究所. 中国水旱灾害[M]. 北京:中国水利水电出版社,1997:519-531.

(收稿日期:2009-02-23 编辑:方宇彤)

(上接第 38 页)

参考文献：

[1] 陈述云,张崇甫. 对多指标综合评价的主成分分析方法的改进[J]. 统计研究,1995(1):35-39.
[2] 许建华,张学工. 经典线性算法的非线性核形式[J]. 控制与决策,2006,21(1):1-6.
[3] VLADIMIRN V. 统计学习理论的本质[M]. 张学工,译. 北京:清华大学出版社,2000.
[4] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):32-42.
[5] SCHOLKOPF B, SMOLA A, MULLER K R. Nonlinear component analysis as a kernel eigenvalue problem [J]. Neural Computation,1998,10(5):1299-1319.
[6] 朱甫芹. 基于 KPCA 的城镇化水平综合评价[J]. 决策参

考,2004(1):40-41.

[7] 孔锐,施择生,郭立,等. 利用组合核函数提高核主分量分析的性能[J]. 中国图像图形学报,2004,9(1):40-45.
[8] 王岩,郑垂勇. 核函数主成分法和系统聚类法[J]. 水利经济,2006,24(3):10-14.
[9] 印兴耀,孔国英,张广智. 基于核主成分分析的地震属性优化方法及应用[J]. 石油地球物理勘探,2008,43(2):179-183.
[10] 刘向东,骆斌,陈兆乾. 支持向量机最优模型选择的研究[J]. 计算机研究与发展,2005,42(4):576-581.
[11] 金菊良,丁晶. 水资源系统工程[M]. 成都:四川科学技术出版社,2002.
[12] 门宝辉,赵燮京,梁川. 我国北方地区水资源可持续利用状况评价[J]. 南水北调与水利科技,2003,4(1):24-27.

(收稿日期:2008-10-21 编辑:方宇彤)

· 简讯 ·

第 13 届世界湖泊大会将在武汉市召开

第 13 届世界湖泊大会将于 2009 年 11 月 1~5 日在湖北省武汉市召开。大会主办单位为中国环境科学学会,中国环境科学研究院,武汉市人民政府,国际湖泊环境委员会;承办单位为中国环境科学学会,武汉市人民政府。会议议题如下:

a. 全球变暖对湖泊环境影响与新问题及新机制研究 ①21 世纪湖泊保护的难点、焦点与新问题;②全球变暖对湖泊环境与生态系统的影响;③跨界湖泊与河流的环境保护;④新型污染物与水污染机理;⑤湖泊保护与防治资金筹措机制;⑥湖泊流域综合管理的新机制和新体制;⑦其他新问题。

b. 湖泊流域水环境管理战略、政策及法规 ①湖泊水环境管理战略与理念;②湖泊流域水环境管理的政策和法律法规研究。

c. 湖泊流域水污染机理、污染监测及污染控制管理和技术对策 ①湖泊及流域水污染机理与控制基础研究;②湖泊流域城镇污染与农业面源污染防治技术;③水环境监测、预警与分析测试方法、技术及设备;④湖泊富营养化控制理念与新技术;⑤湖泊水华暴发机理与控制技术研究;⑥咸水湖泊水污染防治与咸化过程研究。

d. 湖泊流域综合管理及湖泊生态保护 ①湖泊流域水资源与可持续发展;②湖泊水量调控、水质改善与水源地保障技术研究;③湖泊萎缩与水量变化及生态退化;④湖泊生态系统健康与湿地保护;⑤湖泊流域人类活动影响及其综合管理。

e. 湖泊可持续利用的文化及公众参与 ①湖泊管理中的公众参与;②湖泊流域居民与传统文化及生态文明。

f. 中国湖泊污染控制研究 ①中国湖泊流域水污染问题及防治对策;②中国三湖(太湖、巢湖和滇池)水污染防治与蓝藻控制;③武汉湖泊的保护与管理。

会议查询网址: <http://www.chinalakes.org>,会议联系 E-mail: info@chinalakes.org

(本刊编辑部供稿)