# 基于核主成分 - 聚类分析的土壤质量评价方法

童新安,王云鹏　(洛阳理工学院数理教学部,河南洛阳 471023)

**摘要**　针对传统主成分分析处理非线性问题的不足,采用基于核主成分 - 聚类分析的评价方法对土壤质量进行评价并分类。先利用核主成分分析剔除原始数据中存在线性相关和信息重叠的指标,再利用得到的主成分代替原来的评价指标,对土壤质量进行聚类分析。结果表明,该方法准确客观,可为土壤评价提供参考依据。

**关键词**　核主成分聚类分析;土壤质量;评价方法
**中图分类号**　S11+9　　**文献标识码**　A　　**文章编号**　0517 - 6611(2009)10 - 04355 - 03

**Evaluating Method of Soil Quality Based on Kernel Principal Component-cluster Analysis**

**TONG Xin-an et al**　(Department of Mathematics and Physics, Luoyang Institute of Science and Technology, Luoyang, Henan 471023)

**Abstract**　Aiming at the deficiency of traditional principal component analysis solving nonlinear problems, the evaluating method based on kernel principal component - cluster analysis was taken to evaluate and classify the soil quality. Firstly, the kernel principal component analysis was taken to eliminate the indexes that had linear correlation and repeated information in original data, and then the original evaluation index was replaced by principal component, and according which, the cluster analysis on soil quality was conducted. The results showed that the method was accurate and objective, and it could provide the reference for evaluating soil quality.

**Key words**　Kernel principal component - cluster analysis; Soil quality; Evaluating method

土壤质量作为土壤肥力质量、环境质量和健康质量的综合量度,是土壤维持生产力、环境净化能力及保障动植物健康能力的集中体现[1]。土壤质量评价是土壤质量研究的基础和重要内容之一,土壤质量评价可为土壤的整治、规划和合理利用提供科学依据。但迄今为止,土壤质量评价尚没有统一标准[2]。土壤科学评价的基础是土壤分类[3],不同时期所拟定的土壤分类系统反映了该时期人们对土壤的认识水平和土壤科学本身的发展水平。现在越来越多的评价方法如灰色关联度法、模糊数学、多元统计分析、层次分析模型、系统评价模型、Fuzzy 聚类分析、投影寻踪、物元分析等被应用到土壤质量综合评价中[4-9]。笔者介绍了基于核主成分 - 聚类分析的土壤质量评价方法,该方法对土壤质量的评价结果符合客观实际,是一种科学合理的土壤质量评价方法。

## 1　核主成分 - 聚类分析方法

**1.1　核主成分分析(Kernel PCA)**　主成分分析(PCA)是一种经典的特征提取和降维方法[10]。但 PCA 是一种线性映射方法,降维后的矩阵是由线性映射生成的,忽略了数据之间高于两阶的相互关系,所以抽取的特征并不是最优的。为此,Scholkopf 等通过借鉴 SVM 的核技巧[11],将 PCA 方法推广到代表非线性领域的高维特征空间,提出了核主成分分析(KPCA)。KPCA 是一种能从数据样本中提取非线性特征的有效方法。

核主成分分析的基本思想是先通过一个非线性变换 $\Phi(x)$ 将输入数据 $X$ 映射到一个高维特征空间 $F$ 上,然后在特征空间 $F$ 上进行经典的线性主成分分析。

设有 $n$ 维的 $m$ 个样本集为 $x_1,x_2,\cdots,x_m$,其中 $x_k \in R^n$。首先,使用非线性映射 $\Phi$,将输入空间中的样本点 $x_1,x_2,\cdots,x_m$ 映射到高维特征空间 $F$ 中的 $\Phi(x_1),\Phi(x_2),\cdots,\Phi(x_m)$。假设 $\sum_{k=1}^{m}\Phi(x_k)=0$,则 *KPCA* 就可在高维特征空间 $F$ 中对协方差矩阵 $\bar{C}=\frac{1}{m}\sum_{j=1}^{m}\Phi(x_j)\cdot\Phi(x_j)^T$ 进行线性主成分分析。矩阵 $\bar{C}$ 的特征值 $\lambda(\lambda\geqslant 0)$ 和特征向量 $V(V\in F\backslash\{0\})$ 满足 $\lambda V=\bar{C}V$,两边均乘以 $\Phi(x_k)$,得到 $\lambda(\Phi(x_k)\cdot V)=(\Phi(x_k)\cdot\bar{C}V)$。对 $\lambda\neq 0$ 的所有特征向量 $V$,存在系数 $a_1,a_2,\cdots,a_m$,使得 $V=\sum_{k=1}^{m}a_k\Phi(x_k)$。引入 $m\times m$ 维矩阵 $K$,其元素定义为 $K_{ij}=K(x_i,x_j)=(\Phi(x_i)\cdot\Phi(x_j))$,可以得到 $m\lambda Ka=K^2a$,证明其等价于 $m\lambda a=Ka$,求解此式可得到特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_m$,而 $a_1,a_2,\cdots,a_m$ 则为其相应的特征向量。

任意向量 $x$ 在特征空间主成分方向 $\Phi(x)$ 上的投影为 $V\cdot\Phi(x)=\sum_{k=1}^{m}a_i\Phi(x_k)\cdot\Phi(x)=\sum_{k=1}^{m}a_iK(x_i,x)$。由于 $K(x_i,x)$ 为核函数,核主成分分析只需在原空间作内积即可,而不需要映射 $\Phi(x)$ 的具体表现形式。

与 PCA 类似,比值 $\lambda_i/\sum_{k=1}^{m}\lambda_k$ 反映了 $V\cdot\Phi(x)$ 对整体方差的贡献,其重要分量对应的比值较大,故可用保留前 $s$ 个特征向量的方法使系统降维。主成分数量 $s$ 的选取一般根据 $(\sum_{k=1}^{s}\lambda_k/\sum_{k=1}^{m}\lambda_k)\geqslant E$ 来确定,阈值 $E$ 通常取 85%。

上述算法是在假设 $\sum_{k=1}^{m}\Phi(x_k)=0$ 的情况下得出的,但该假设一般情况下是不成立的,因此,通常先令 $\bar{\Phi}(x_i)=\Phi(x_i)-\frac{1}{m}\sum_{k=1}^{m}\Phi(x_k)$,对映射数据进行中心化,再用 $\bar{K}=K-AK-KA+AKA$ 代替原来的 $K$,进而求解特征值和特征向量,其中,$A$ 为系数为 $\frac{1}{m}$ 的 $m\times m$ 阶单位矩阵。

目前常用的核函数形式有[12-13]:(1)多项式核函数 $K(x,x_i)=[a(x\cdot x_i)+b]^d$;(2)高斯径向基函数(*RBF*)核函数 $K(x,x_i)=\exp\left(-\frac{\|x\cdot x_i\|^2}{2\sigma^2}\right)$;(3)多层感知器(*MLP*)核函数 $K(x,x_i)=\tanh[a(x,x_i)+b]$。

到目前为止,核函数的选择及相关参数的确定并无太多理论指导[13],一般靠试验和经验获取。

**1.2　层次聚类法(Hierarchical Clustering Method)**　层次聚类法(HCM)又称系统聚类法,分级聚类法,是将样品或变量按照其性质上的亲疏相似程度进行分类的一种多元统计

作者简介　童新安(1982 - ),男,湖北荆州人,硕士,助教,从事数据处理与数据挖掘研究。
收稿日期　2009-01-07

方法[13]。其思想是先将所有样本各自视为一类,然后计算类与类之间的相似性,将相似性最大的一对类别合并成一个新类,进而在新的类别划分下重复合并操作,直到满足停止条件。常用的类间距离定义方法有最短距离法、最长距离法、中间距离法、重心法、类平均法、离差平方和法等,而停止判断条件有:所有样本合并成一类;样本类别数达到规定的数目;所有类别的相似性测度大于给定的阈值。此方法的基本步骤是:①先将所有样本各自看成一类,计算样本与样本之间的距离和类与类之间的距离;②合并距离最近的两类为一个新类,新类的属性基于原类的属性构成。如果合并的新类中有一个公共类则这两个新类为一类;③把新类看成新的样品并提取其相应的属性值,重新计算类与类之间的距离;④判断是否达到聚类终止条件,如达到,则聚类停止;否则转第二步;⑤画出聚类表,确定聚类类别。

层次聚类法的优点是不必事先知道分类对象的分类结构就可以给出很好的分类结果,而且划分的每个子集中的点具有高度的内在相似性[14]。但该方法并不能得到各类别间优劣程度的综合评价结果。

**1.3 核主成分-聚类方法(KPCA-HCM)** 对大样本多指标系统进行聚类分析时,众多的指标变量往往导致聚类过程复杂,聚类结果不易分析,此时可先通过主成分分析对其进行降维[15-16]。然而,主成分分析尽管能够很方便地用较少的数据对多指标系统进行综合评价,但样本间的线性关系不太强时,可能需要较多数目的主成分,降维效果不明显,故此时可以用核主成分法对其进行改进。鉴于这两者的特点,可以将核主成分方法与聚类分析法结合起来,采取"核主成分-聚类"分析方法进行综合评价。

所谓"非线性主成分-聚类"方法,是指先对原始数据进行核主成分分析,再取其若干主成分进行聚类分析的一种新的综合评价方法。其实现步骤如下:①输入样本数据;②利用 KPCA 对样本数据进行核主成分分析,选取累计贡献率大于 85% 的特征值及其对应的特征向量,并计算主成分;③用选取的主成分替代原始数据进行聚类分析;④确定聚类类别,评价聚类结果。

## 2 土壤质量的核主成分-聚类评价过程

以福州海岛各类耕作土壤[17]为研究对象,以土属为评价单元,根据科学性、综合性、实用性和主导因素原则,选择土壤耕层厚度、有机质、全氮、全磷、全钾、有效磷、速效钾含量、pH 值和阳离子交换量(CEC)共 9 个指标作为土壤质量评价的参评因子[1,17],利用核主成分-聚类分析法对土壤质量进行定量评价。16 个土属各参评因子指标见表 1[17]。

**表 1 福州海岛土属的理化性质**

**Table 1 Physical and chemical properties of soil genera inislands of Fuzhou**

| 序号 Sequence | 土壤名称 Soil name | 耕层厚度 cm Topsoil thickness | 有机质 g/kg Organic matter | 全氮 g/kg Total nitrogen | 全磷 g/kg Total phosphorus | 全钾 g/kg Total potassium | 有效磷 mg/kg Available P | 速效钾 mg/kg Rapid available K | pH 值 pH value | 阳离子交换量 cmol(+)/kg CEC |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 赤土 Bare soil | 12 | 10.4 | 0.66 | 0.63 | 4.52 | 7.5 | 22.3 | 5.4 | 20.1 |
| 2 | 赤砂土 Bare sand soil | 10 | 7.3 | 0.44 | 0.44 | 4.56 | 11.5 | 19.5 | 6.1 | 8.44 |
| 3 | 红泥砂土 Red mud and sand soil | 12 | 9.7 | 0.53 | 8.82 | 30.8 | 26.0 | 43.0 | 6.1 | 14.00 |
| 4 | 黄泥田 Yellow mud field | 7 | 10.6 | 0.67 | 0.64 | 4.84 | 15.5 | 29.2 | 6.3 | 22.86 |
| 5 | 黄泥砂田 Yellow mud and sand field | 13 | 15.2 | 0.87 | 0.63 | 9.36 | 10.5 | 23.2 | 7.6 | 19.40 |
| 6 | 乌泥田 Black mud field | 16 | 24.6 | 1.46 | 1.39 | 29.6 | 17.5 | 90.1 | 5.1 | 42.07 |
| 7 | 灰泥田 Grey mud field | 11 | 36.6 | 2.61 | 3.44 | 33.2 | 30.0 | 453.0 | 7.3 | 24.25 |
| 8 | 灰砂田 Grey sand field | 10 | 14.5 | 0.70 | 1.53 | 31.8 | 31.5 | 72.6 | 5.5 | 8.62 |
| 9 | 烂泥田 Slush field | 16 | 22.2 | 1.42 | 1.24 | 33.4 | 13.0 | 96.1 | 6.2 | 13.52 |
| 10 | 盐斑田 Saline spot field | 12 | 14.1 | 1.17 | 1.68 | 6.04 | 28.0 | 408.0 | 8.2 | 33.54 |
| 11 | 埭田 Dike field | 12 | 28.2 | 1.91 | 7.67 | 29.4 | 14.5 | 99.6 | 6.1 | 30.98 |
| 12 | 潮砂土 Damp sand soil | 12 | 11.2 | 0.68 | 1.08 | 34.8 | 42.0 | 22.0 | 6.3 | 8.12 |
| 13 | 潮泥土 Damp mud soil | 17 | 18.4 | 0.69 | 1.30 | 32.2 | 26.5 | 194.0 | 5.0 | 10.46 |
| 14 | 旱砂土 Dry sand field | 13 | 2.9 | 0.013 | 0.31 | 16.8 | 6.0 | 26.1 | 6.2 | 2.63 |
| 15 | 润砂土 Damp sand soil | 10 | 5.8 | 0.028 | 1.15 | 18.2 | 20.5 | 15.3 | 6.7 | 3.18 |
| 16 | 砂埭土 Sand dike field | 21 | 3.3 | 0.27 | 0.34 | 17.9 | 6.5 | 57.6 | 6.5 | 8.52 |

由于原始数据指标较多,为确定指标间的线性关系,分别采用 PCA 和 KPCA 法对原始数据进行降维处理,提取特征向量,其中,KPCA 方法中选取高斯径向基函数(RBF)核函数,参数取 $2\sigma^2 = 10^6$,计算结果如下:

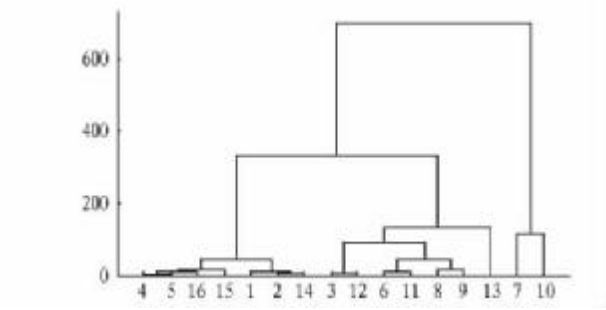
**表 2 PCA 法和 KPCA 法提取的特征向量的比较**

**Table 2 Comparisons of feature extraction between PCA and KPCA methods**

| 序号 Sequence | PCA 方法 Method | | | KPCA 方法 Method | | |
|---|---|---|---|---|---|---|
| | 特征值 Eigenvalue | 方差贡献率 Variance contribution | 方差累积贡献率 Variance cumulative contribution | 特征值 Eigenvalue | 方差贡献率 Variance contribution | 方差累积贡献率 Variance cumulative contribution |
| 1 | 3.570 939 | 0.396 771 | 0.396 771 | 0.510 16 | 0.962 346 | 0.962 346 |
| 2 | 1.725 64 | 0.191 738 | 0.588 509 | 0.009 776 | 0.018 441 | 0.980 787 |
| 3 | 1.311 231 | 0.145 692 | 0.734 201 | 0.004 856 | 0.009 161 | 0.989 947 |
| 4 | 1.009 397 | 0.112 155 | 0.846 356 | 0.003 037 | 0.005 729 | 0.995 676 |
| 5 | 0.689 336 | 0.076 593 | 0.922 949 | 0.001 512 | 0.002 852 | 0.998 528 |
| 6 | 0.388 126 | 0.043 125 | 0.966 074 | 0.000 447 | 0.000 844 | 0.999 372 |
| 7 | 0.243 66 | 0.027 073 | 0.993 148 | 0.000 186 | 0.000 35 | 0.999 722 |
| 8 | 0.043 747 | 0.004 861 | 0.998 009 | 9.05E-05 | 0.000 171 | 0.999 893 |
| 9 | 0.017 922 | 0.001 991 | 1 | 5.68E-05 | 0.000 107 | 1 |

由表 2 可知,PCA 方法降维后,第一特征值贡献率仅为 39.7%,若取阈值为 85%,需选取 5 个主成分,降维效果不明显,说明原始数据间线性关系不强;而采用 KPCA 方法降维后,第一特征值的贡献率达到 96.2%,大于阈值。故可仅选第一主成分代替原始数据。

分别对 PCA 法后的前 5 个主成分和 KPCA 法降维后的第一主成分进行层次聚类,类间距离用离差平方和法计算,聚类结果如:




注:数字代表每类土壤的序号。下同。

Note: The number represents the sequence of each type of soil. The same as follows.

**图 1　PCA 法降队后前 5 个主成分的聚类结果**

**Fig. 1　The clustering result of former five main components with PCA method**

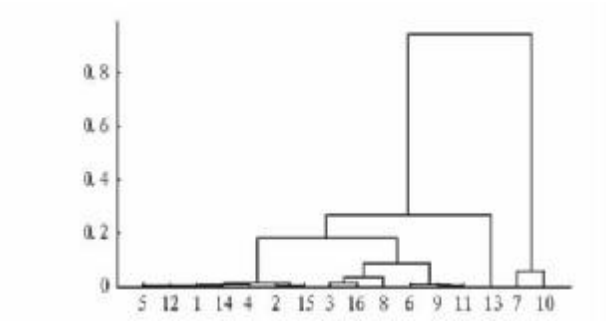



**图 2　KPCA 法降维后第一个主成分的聚类结果**

**Fig. 2　The clustering result of first main components with KPCA method**

由图 1 和图 2 可知,若将土壤划分成 4 个类别,PCA 和 KPCA 2 种聚类结果基本相同,验证了核主成分－聚类方法的有效性。图 2 所示分类结果列于表 3。

**表 3　土壤质量的核主成分－聚类结果**

**Table 3　The clustering result of soil quality by using KPCA－HCM**

| 类别 Category | 土壤名称 Soil name |
|---|---|
| 1 | 红泥砂土、砂堺土、灰砂田、乌泥田、烂泥田、堺田 |
| 2 | 黄泥砂田、潮砂土、赤土、旱砂土、赤沙土、润砂土 |
| 3 | 潮泥土 |
| 4 | 灰泥田、盐斑田 |

通过对原始数据的分析,第一类土壤耕层较厚,养分充足,保肥能力强,属于优质土壤;第二类土壤各项指标与第一类均有一定差距,土壤耕层相对较薄,各养分含量相对较低,为比较贫瘠的劣质土壤;第四类土壤的速效钾含量显著高于其他土壤,且其他土属大多呈酸性而该类别土壤却呈偏碱性,讨论土壤养分含量时将其作为一类是合理的,该类土壤属于比较“特殊”的土属;第三类土壤只有一个土属——潮泥土,PCA 和 KPCA 2 种方法都将其单独作为一类,通过原始数据可以看出,该土壤速效钾含量较高,但 pH 值却明显偏低,与第四类土壤不同。但除 pH 值偏小外,该土壤各养分含量均高于第一类土壤均值,而部分元素含量又低于第二类土壤均值,故应将其与第一、二类区别开。

## 3　结论与讨论

对土壤质量进行测度评价,为土壤的整治、规划和合理利用提供科学依据,是土壤研究和农业生产面临的重要课题。该研究通过对 PCA 方法的不足进行改进,融合层次聚类法的优点,建立了基于核主成分－聚类分析的土壤质量综合评价模型,并运用该模型对耕作土壤质量进行实证分析。结果表明,该评价模型对土壤质量的评价总体来说是合理有效的。

但不可否认的是,尽管通过实例分析验证了该方法的可行性,但该方法对土壤的分类结果与解决实际问题还有一定距离。因此对该评价模型进行更加科学、可行的改进和完善,仍是今后工作的重要内容。

## 参考文献

[1] 刘世梁,傅伯杰,刘国华,等. 我国土壤质量及其评价研究的进展[J]. 土壤学报,2006,37(1):137－143.
[2] 郑昭佩,刘作新. 土壤质量及其评价[J]. 应用生态学报,2003,14(1):131－134.
[3] 蔡暄,徐惠,吴群. 土壤质量聚类分析——以封丘县为例[J]. 安徽农业科学,2008,36(25):10998－10999,11001.
[4] 侯文广,江聪世,熊庆文,等. 基于 GIS 的土壤质量评价研究[J]. 武汉大学学报,2003,28(1):60－64.
[5] 张庆利,史学正,潘贤章,等. 江苏省金坛市土壤肥力的时空变化特征[J]. 土壤学报,2004,41(2):315－319.
[6] 张华,张甘霖,漆智平,等. 热带地区农场尺度土壤质量现状的系统评价[J]. 土壤学报,2003,40(2):186－193.
[7] 万存绪,张效勇. 模糊数学在土壤质量评价中的应用[J]. 应用科学学报,1991,9(4):359－365.
[8] 付强,金菊良,门宝辉,等. 基于 RAGA 的 PPE 模型在土壤质量等级评价中的应用研究[J]. 水土保持通报,2002,22(5):51－54.
[9] 潘峰,梁川,付强. 基于层次分析法的物元模型在土壤质量评价中的应用[J]. 农业现代化研究,2002,23(2):93－97.
[10] 梅长林,周家良. 实用统计方法[M]. 北京:科学出版社,2002.
[11] SCHOLKOPF B, SMOLA A, MULLER K R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem[J]. Neural Computation, 1998, 10(5):1299－1319.
[12] 郭辉,刘贺平,王玲. 最小二乘支持向量机参数选择方法及其应用研究[J]. 系统仿真学报,2006,18(7):2033－2036.
[13] 肖建华. 智能模式识别方法[M]. 广州:华南理工大学出版社,2006.
[14] 王和勇,姚正安,李磊. 基于聚类的核主成分分析在特征提取中的应用[J]. 计算机科学,2005,32(4):64－66.
[15] 鲍艳,胡振琪,柏玉,等. 主成分聚类分析在土地利用生态安全评价中的应用[J]. 农业工程学报,2006,22(8):87－90.
[16] 童新安,许超. 基于非线性主成分和聚类分析的综合评价方法[J]. 统计与信息论坛,2008,(02):37－41,46.
[17] 陈松林. 福州海岛耕作土壤质量的定量评价[J]. 福建师范大学学报,1996,12(3):84－88.