文章编号:1000-5471(2012)03-0006-05

基于模糊点数据的核主成分分析①

纳艳萍, 魏立力

宁夏大学 数学计算机学院,银川 750021

摘要:在核主成分分析中,给每个训练数据赋予一个置信权重,将训练数据视为样本空间的模糊点,研究了基于模糊点数据的核主成分分析.数值模拟表明,该方法能够有效控制异常点对主成分的影响.同时,该方法也为数据先验信息的利用提供了一个可行的途径.

关 键 词:主成分分析;核主成分分析;模糊点数据

中图分类号:O212.4;TP391.4 **文献标志码**:A

主成分分析[1](PCA)是将多个相关指标转化成少数不相关指标的一种多元统计方法,常常用于数据压缩和特征提取[2].由于主成分分析是一种线性映射方法,在处理非线性问题时往往不能取得好的效果,所以,近年来把主成分分析推广到非线性情形[3-5]成了一个研究热点.核主成分分析[6-7](KPCA)就是非线性情形的方法之一.这种方法通过运用核函数,把输入空间映射到高维特征空间,在特征空间中进行主成分分析.核主成分分析在一系列应用中表现出较主成分分析更加优秀的性能,如:基于核的特征提取方法可有效解决特征提取中的非线性问题;将核主成分分析用于支持向量机可以降低分类面的复杂度、提高分类精度等.

然而,在实际问题中,由于主成分分析对异常数据非常敏感,常常导致错误的分析结果.为了克服这个缺点,许多学者先后提出了稳健主成分分析[8-9]和模糊主成分分析[10-12].稳健主成分分析主要有两种方法:第一种是建立稳健协方差矩阵来处理一般情况;第二种是投影寻踪方法,用来处理指标多于样本个数的情况.模糊主成分分析主要是对模糊数据集做主成分分析.文献[10]通过在模糊数的两个边界和中心做主成分分析得到 3 个模糊协方差矩阵和 3 个主成分公式;文献[11-12]从隶属度方面改进,而隶属度通常由数据到数据集的中心距离来确定.这些方法都是从客观角度出发,没有充分利用训练数据的先验信息.

本文从一个新的角度将模糊性引入到核主成分分析中.在经典的核主成分分析中,每个训练数据在构建主成分时的作用是相同的.然而,在许多实际应用中,训练数据的意义和作用是不同的,通常有些数据比其它数据更为重要.对于重要的数据我们应给予充分的重视,在构建主成分时应起更大的作用,而对于不可信数据(可能是异常数据),应限制其作用.受文献[13-14]的启发,本文给每个训练数据赋予一个置信权重,将训练数据视为样本空间的模糊点,在此基础上,重新研究了核主成分分析,为更有效地分析数据以及控制可能的异常点提供了一个可行的途径.

1 核主成分分析

给定一个中心化的数据集 $\boldsymbol{x}_k \in \mathbb{R}^N$, $k=1,\cdots,M$. 通过一个非线性函数 Φ 把数据从输入空间 $\mathbb{R}^N$ 映射到

① 收稿日期:2010-09-27
基金项目:国家自然科学基金资助项目(60663003);宁夏高等学校科学研究项目(2010);宁夏大学科研基金项目 E(ndzr0922).
作者简介:纳艳萍(1977-),女,宁夏永宁人,讲师,主要从事统计数据分析和模式识别研究.

特征空间 F，

$$\begin{aligned}\Phi:\ &\mathbb{R}^N \longrightarrow F\\ &\boldsymbol{x} \longrightarrow \boldsymbol{X}\end{aligned} \tag{1}$$

其中 F 可以有任意大的维数甚至可以有无穷维. 利用内积在 F 中做主成分分析就可以得到核主成分分析.

定义 1[6]　设 $\boldsymbol{x}$ 是一个测试点，$\Phi(\boldsymbol{x})$ 是它在 F 中的像，则称

$$(\boldsymbol{V}^k \cdot \Phi(\boldsymbol{x})) = \sum_{i=1}^{M} \alpha_i^k (\Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x})) \qquad k = 1, \cdots, p \tag{2}$$

为相应于 Φ 的核主成分. (2) 式中 $\boldsymbol{V}^k$ 是协方差矩阵 $\boldsymbol{C} = \frac{1}{M}\sum_{j=1}^{M} \Phi(\boldsymbol{x}_j)\Phi(\boldsymbol{x}_j)'$ 的第 k 个规范化的特征向量，$\boldsymbol{\alpha}^\kappa$ 是矩阵 $\boldsymbol{K}$ 的从大到小排列的第 k 个特征值对应的特征向量，α_i^k 是 $\boldsymbol{\alpha}^\kappa$ 的第 i 个分量，而 $M \times M$ 矩阵 $\boldsymbol{K}$ 的元素 $\boldsymbol{K}_{ij} := (\Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j))$.

2　基于模糊点数据的核主成分分析

受文献[13－14]的启发，本文给出如下定义.

定义 2[15]　设 $\{\boldsymbol{x}_i,\ i = 1, \cdots, n\}$ 是 p 维训练样本，给每个训练样本 $\boldsymbol{x}_i$ 赋予一个模糊隶属度 $s_i (0 < \sigma \leqslant s_i \leqslant 1)$，则称 $(\boldsymbol{x}_i,\ s_i)$ 是 R^p 中的模糊点，$\boldsymbol{x}_i$ 和 s_i 分别叫做支撑和高度，并称 $\{(\boldsymbol{x}_i,\ s_i),\ i = 1, \cdots, n\}$ 是模糊点数据集.

在实际应用中，s_i 可以认为是对应数据点对总体的重要度或置信度，σ 为其下界. s_i 的确定需依赖于具体的数据特征，特征不同其含义就不同：对于具有时间特性的数据，s_i 可以选择为 $\boldsymbol{x}_i$ 达到系统的时间 t_i 的函数；对于分组重复测量数据，s_i 可以表示为对应组测量数据的频率；对于分类数据，s_i 可以表示为类标示变量的函数；当然 s_i 也可以直接表示人们对每个数据置信度的先验信息. 总之，s_i 的确定并不困难，但或多或少的存在着一些主观色彩[15]. 为了便于研究，本文假设 s_i 是已知的.

设 $\{(\boldsymbol{x}_i,\ s_i),\ i = 1, \cdots, M\}$ 是模糊点数据集，并设一个非线性函数 Ψ 把输入空间 $\mathbb{R}^N$ 映射到特征空间 H

$$\begin{aligned}\Psi:\ &\mathbb{R}^N \longrightarrow H\\ &(\boldsymbol{x}_i,\ s_i) \longrightarrow \Psi(s_i\boldsymbol{x}_i) \qquad i = 1, \cdots, M\end{aligned}$$

设数据集 $\Psi(s\boldsymbol{x})$ 被中心化，$\sum_{k=1}^{M} \Psi(s_k\boldsymbol{x}_k) = 0$，在 H 中数据集的协方差矩阵为

$$\boldsymbol{G} = \frac{1}{M}\sum_{j=1}^{M} \Psi(s_j\boldsymbol{x}_j)\Psi(s_j\boldsymbol{x}_j)'$$

寻找特征值 $l \geqslant 0$ 和特征向量 $\boldsymbol{U} \in \boldsymbol{H}\backslash\{0\}$ 满足

$$l\boldsymbol{U} = \boldsymbol{G}\boldsymbol{U} \tag{3}$$

所有 $\boldsymbol{U}$ 的解都在由 $\Psi(s_1\boldsymbol{x}_1), \cdots, \Psi(s_M\boldsymbol{x}_M)$ 生成的线性空间中. 先考虑与(3) 式等价的方程

$$l(\Psi(s_k\boldsymbol{x}_k) \cdot \boldsymbol{U}) = (\Psi(s_k\boldsymbol{x}_k) \cdot \boldsymbol{G}\boldsymbol{U}),\ k = 1, \cdots, M \tag{4}$$

其次，设存在系数向量 $\boldsymbol{\beta}_i (i = 1, \cdots, M)$ 使得

$$\boldsymbol{U} = \sum_{i=1}^{M} \boldsymbol{\beta}_i \Psi(s_i\boldsymbol{x}_i) \tag{5}$$

综合(4) 和(5) 式得到

$$l\sum_{i=1}^{M} \boldsymbol{\beta}_i (\Psi(s_k\boldsymbol{x}_k) \cdot \Psi(s_k\boldsymbol{x}_k)) = \frac{1}{M}\sum_{i=1}^{M} \boldsymbol{\beta}_i (\Psi(s_k\boldsymbol{x}_k) \cdot \sum_{j=1}^{M} \Psi(s_j\boldsymbol{x}_j)(\Psi(s_j\boldsymbol{x}_j) \cdot \Psi(s_i\boldsymbol{x}_i))) \qquad k = 1, \cdots, M \tag{6}$$

定义一个 $M \times M$ 矩阵 $\widetilde{\boldsymbol{K}}$，$\widetilde{\boldsymbol{K}}$ 的第 $(i,\ j)$ 元为

$$\widetilde{\boldsymbol{K}}_{ij} := (\Psi(s_i\boldsymbol{x}_i) \cdot \Psi(s_j\boldsymbol{x}_j))$$

则(6) 式变为

$$Ml\widetilde{\boldsymbol{K}}\boldsymbol{\beta} = \widetilde{\boldsymbol{K}}^2\boldsymbol{\beta} \tag{7}$$

这里 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\ \cdots,\ \boldsymbol{\beta}_M)'$，(7) 式可以化为

$$Ml\boldsymbol{\beta}=\widetilde{\boldsymbol{K}}\boldsymbol{\beta}$$

容易证明，矩阵 $\widetilde{\boldsymbol{K}}$ 是一个非负定矩阵.

设 $l_1\geqslant l_2\geqslant\cdots\geqslant l_p$ 是 $\widetilde{\boldsymbol{K}}$ 的全部非零特征根，$\boldsymbol{\beta}^1,\cdots,\boldsymbol{\beta}^p$ 是相应的特征向量，当要求向量 $\boldsymbol{U}$ 标准化时，

$$(\boldsymbol{U}^k\cdot\boldsymbol{U}^k)=1\qquad k=1,\cdots,p \tag{8}$$

根据(5)式和(8)式得到

$$1=\sum_{i,\,j=1}^{M}\beta_i^k\beta_j^k(\Psi(s_i\boldsymbol{x}_i)\cdot\Psi(s_j\boldsymbol{x}_j))=\sum_{i,\,j=1}^{M}\beta_i^\kappa\beta_j^\kappa\widetilde{\boldsymbol{K}}_{ij}=(\boldsymbol{\beta}^\kappa\cdot\widetilde{\boldsymbol{K}}\boldsymbol{\beta}^\kappa)=l_k(\boldsymbol{\beta}^\kappa\cdot\boldsymbol{\beta}^\kappa)$$

定义 3 设 $\boldsymbol{x}$ 是测试点，$\Psi(\boldsymbol{x})$ 是它在特征空间 H 中的像，则基于模糊点数据的核主成分为

$$(\boldsymbol{U}^k\cdot\Psi(\boldsymbol{x}))=\sum_{i=1}^{M}\beta_i^k(\Psi(s_i\boldsymbol{x}_i)\cdot\Psi(\boldsymbol{x}))\ ,\ k=1,\cdots,p$$

显然，基于模糊点数据的核主成分也满足一般主成分的性质.

下面介绍3种常用的核函数：

1) 多项式核函数：$\boldsymbol{k}(\boldsymbol{x},\boldsymbol{y})=((\boldsymbol{x}\cdot\boldsymbol{y})+1)^d$.

2) 高斯径向基核函数：$\boldsymbol{k}(\boldsymbol{x},\boldsymbol{y})=\exp\left\{-\dfrac{\boldsymbol{x}-\boldsymbol{y}^2}{2\sigma^2}\right\}$.

3) 多层感知器核函数：$\boldsymbol{k}(\boldsymbol{x},\boldsymbol{y})=\tanh(\boldsymbol{k}(\boldsymbol{x}\cdot\boldsymbol{y})+\boldsymbol{c})$.

3 数值模拟

为了说明基于模糊点数据的核主成分分析能有效地控制异常点对主成分的影响，下面用文献[6]中的例子进行数值模拟，并分别用高斯核和多项式核做核主成分分析.

例 1 给出一个二维数据点集，用高斯核 $\boldsymbol{k}(\boldsymbol{x},\boldsymbol{y})=\exp\left\{-\dfrac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right\}$ 作基于模糊点数据的核主成分分析.

二维数据点集是这样产生的：x 的取值服从[−1，1]上的均匀分布，y 由公式 $y_i=x_i^2+\xi$ 产生，其中 ξ 是均值为0标准差为0.2的噪声.

图1是取 $\sigma^2=0.5$ 的高斯核主成分图(经过数次模拟，$\sigma^2=0.5$ 拟合较好). 其中图1(a)是对原始数据做核主成分分析得到的第一主成分图(曲线所示)，从图中看出，第一主成分能较好地反映数据的分布. 图1(b)是加入5个异常点(用'o'表示)后做核主成分分析得到的第一主成分图，从图中看出，异常点影响了第一主成分的方向，成了一条封闭曲线. 图1(c)是给原始数据点分配0.9的置信度，给异常点分配0.1的置信度后，用基于模糊点数据的核主成分分析做出的第一主成分图，从图中看出，用基于模糊点数据的核主成分分析做出的第一主成分，方向更接近图1(a)的第一主成分方向.

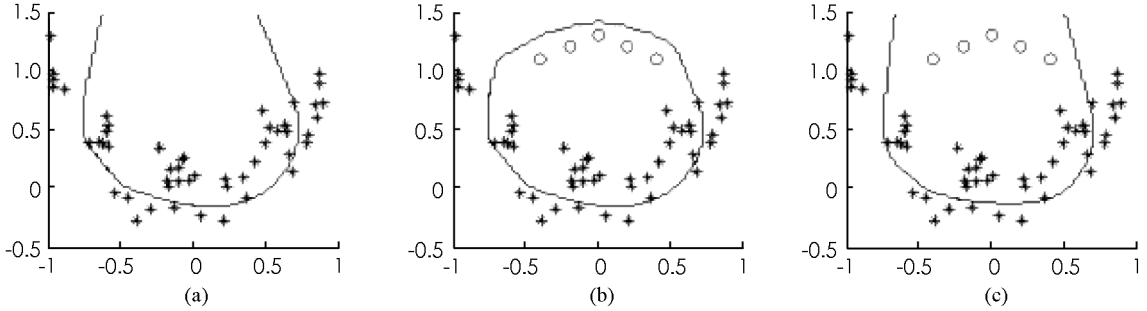

图1 高斯核主成分图

例 2 采用例1中的点集，用基于模糊点数据的多项式核作核主成分分析.

图2是对原始数据做多项式核主成分分析. 图2(a)，(b)分别是用一次多项式核做出的第一和第二主成分(直线所示)；图2(c)，(d)分别是用二次多项式核做出的第一和第二主成分(曲线所示).

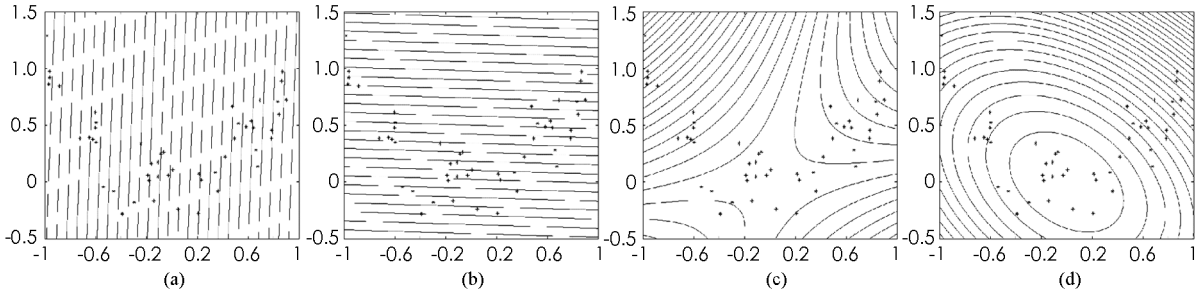

图 2　原始数据的多项式核主成分图

图 3 是给原始数据点集加入 5 个异常点(用'o'表示)后，用多项式核做出的主成分图. 图 3(a),(b)分别是用一次多项式核做出的第一和第二主成分(直线所示)；图 3(c),(d)分别是用二次多项式核做出的第一和第二主成分(曲线所示). 从图 3 中看出，加入异常点后，一次多项式核做出的第一和第二主成分方向都有所偏离，二次多项式核的第二主成分方向变化较大.

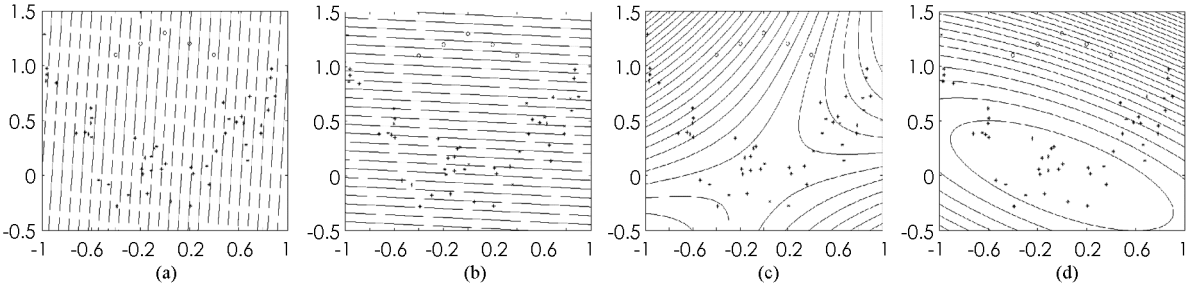

图 3　含有异常点的多项式核主成分图

图 4 是给原始数据点分配 0.9 的置信度，异常点分配 0.1 的置信度后，用基于模糊点数据的核主成分分析做出的主成分图. 顺序如上，从图 4 中看出，除二次多项式核的第一主成分外，其余主成分方向基本恢复到了图 2 的方向.

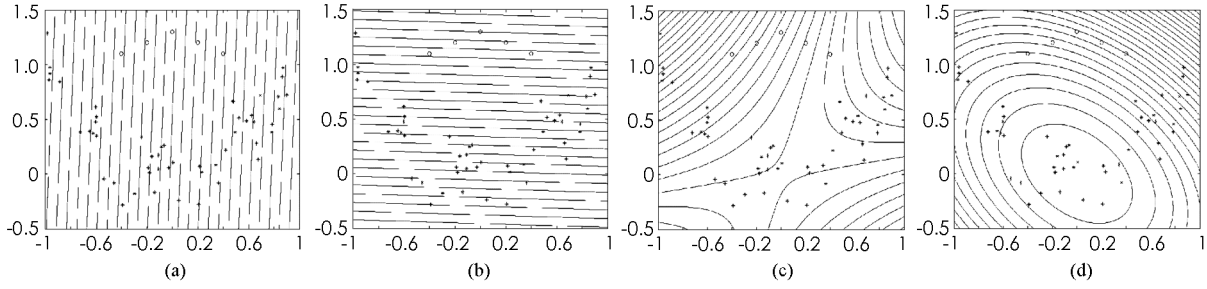

图 4　基于模糊点数据的多项式核主成分图

4　结　　论

当样本的置信度全部取 1 时，基于模糊点数据的核主成分分析方法就是一般的核主成分分析方法. 所以，基于模糊点数据的核主成分分析是一般核主成分分析的拓展.

本文提供的方法能够有效地利用样本的先验信息(模糊点)，从而能有效地控制异常点对主成分的影响.

参考文献：

[1] JOLLIFFE I T. Principal Component Analysis [M]. Berlin: Springer, 2002.

[2] 熊　明，王汝言，唐　琳. 基于局部线性嵌入与主成分分析的人脸识别方法 [J]. 重庆邮电大学学报：自然科学版，2009，21(1)：92－94.

[3] GNANADESIKAN R. Methods for Statistical Data Analysis of Multivariate Observations [M]. New York: John Wiley, 1977: 53－62.
[4] OJA E. A Simplified Neuron Model As a Principal Component Analysis [J]. Math Biology, 1982, 15(3): 267－273.
[5] 张军平，王 钰. 主曲线研究综述 [J]. 计算机学报，2003，26(2)：129－146.
[6] SCHOLKOPF B, SMOLA A, MULLER K R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem [J]. Neural Computation, 1998, 10(5): 1299－1319.
[7] JOHN S T, NELLO C. 模式分析的核方法 [M]. 赵玲玲，译. 北京：机械工业出版社，2006：20－104.
[8] YANG Tai-Ning, WANG Sheng-De. Robust Algorithm for Principal Component Analysis [J]. Pattern Recognition Letters, 1999(20): 927－933.
[9] BOLTON R J, HAND D J, WEBB A R. Projection Techniques for Nonlinear Principal Component Analysis [J]. Statistics and Computing, 2003, 13(3): 267－276.
[10] YABUUCHI Y, WATADA J. Fuzzy Principal Component Analysis for Fuzzy Data [C]//Proceedings of the Sixth IEEE International Conference on Fuzzy Systems. New York: IEEE Press, 1997: 1127－1130.
[11] 林和平，杨 晨. 模糊主成分分析方法的研究与分析 [J]. 航空计算技术，2006，36(6)：16－20.
[12] WU Xiao-hong , ZHOU Jian-jiang. Fuzzy Principal Component Analysis and Its Kernel-Base Model [J]. Journal of Electronics, 2007, 24(6): 772－775.
[13] WEI Li-li, LONG Wei-jiang, ZHANG Weng-xiu. 基于模糊支持向量机的数据域描述 [J]. 计算机科学，2004，31(1)：108－109.
[14] LIN Chun-fu, WANG Sheng-de. Fuzzy Support Vector Machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 464－471.
[15] 纳艳萍，魏立力. 基于模糊点数据的主成分分析 [J]. 甘肃联合大学学报，2009，23(5)：5－8.

Kernel Principal Component Analysis Based on Fuzzy Points Data

NA Yan-ping, WEI Li-li

School of Mathematics & Computer Science, Ningxia University, Yinchuan 750021, China

Abstract: Applying a confidence weight to each training data and considering the training data as fuzzy points in the sample space, this paper studies kernel principal component analysis (KPCA) based on fuzzy points data. Numerical simulation shows that this method can effectively control possible outliers and provides a feasible approach for the utilization of prior information.

Key words: principal component analysis; kernel principal component analysis; fuzzy points data

责任编辑 张 枸