

一种鲁棒的概率主成分分析方法

穆向阳^{1,2}, 张太镒¹, 周亚同³

(1. 西安交通大学电子与信息工程学院, 710049, 西安; 2. 西安石油大学电子工程学院, 710065, 西安;
3. 河北工业大学信息工程学院, 300401, 天津)

摘要: 针对传统主成分对实际样本的奇点不敏感的缺陷, 提出了一种鲁棒概率主成分分析(RPP-CA)方法. 首先引入连续的决策变量构造新能量函数, 将事先给定的硬门限改为自适应确定的软门限, 门限值由样本自动确定, 再计算概率主成分进行特征提取. 与线性主成分分析(LPCA)和概率主成分分析(PPCA)方法相比, RPPCA方法更为实用, 有效地减小了奇点的影响, 显示出比PPCA更强的稳健性, 也扩大了实用范围. 实验结果表明, RPPCA方法的分类准确率比LPCA方法平均提高了3.2%, 比PPCA方法平均提高了0.7%.

关键词: 主成分; 鲁棒; 概率主成分分析; 特征提取

中图分类号: TP18 **文献标志码:** A **文章编号:** 0253-987X(2008)10-1217-04

A Robust Probability Principle Component Analysis Method

MU Xiangyang^{1,2}, ZHANG Taiyi¹, ZHOU Yatong³

(1. School of Electronics and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China; 2. School of Electronic Engineering, Xi'an Shiyou University, Xi'an 710065, China; 3. School of Information Engineering, Hebei University of Technology, Tianjin 300401, China)

Abstract: In order to overcome the drawback that traditional principal component analysis fails to the outliers existing in the realistic data, a robust probability principal component analysis (RPP-CA) method is proposed. A continuous decision variable is introduced into the energy function, and the preset hard threshold is replaced by a soft adaptive threshold which is automatically determined by the data. The algorithm is then embedded in the procedure of PPCA's principal component feature extraction. Compared with PCA and probability principal component analysis (PP-CA), the proposed RPPCA can resist outlier well, is more robust than PPCA, and enlarges the real application area. The simulation results show that the algorithm improves 3.2% classification accuracy to LPCA, and 0.7% to PPCA on average.

Keywords: principle component; robustness; probability principal component analysis; feature extraction

线性主成分分析(LPCA)是一种经典的特征提取方法, 但分析结果缺乏样本分布的概率信息. 文献[1-2]提出的概率主成分分析(PPCA)方法赋予了LPCA概率含义, 可以看作是概率化的LPCA. 这两种方法均是通过对样本集的协方差矩阵进行特征值分解获得主成分, 若样本集内含奇点, 则不能得到协

方差矩阵的准确估计, 最终导致求出的主成分不准确. 所谓奇点, 直观上是指与样本集内绝大部分样本分布差异过大的极少量样本, 如测量或记录的偶尔出错、突发的随机噪音等都会产生奇点. 文献[3-4]利用神经网络方法检测奇点, 但检测准确度有待提高; 文献[5]利用自组织学习规则进行主成分分析,

但硬门槛 η 不易事先指定,从而导致对奇点可能形成误判.本文在分析 PPCA 方法的优缺点后,引入连续的决策变量来构建新的能量函数,并将事先给定的硬门限改为自适应确定软门限,然后将改进后的算法融入 PPCA 的主成分特征提取过程.仿真结果表明,与 LPCA 和 PPCA 方法相比,本文提出的 RPPCA 方法的分类准确率明显提高,有效地减小了奇点的影响,并显示出比 PPCA 更强的稳健性.

1 问题的提出与描述

假设有 N 个维数为 d 的学习样本 $x_1, x_2, \dots, x_N$, 其对应的总体为 $\mathbf{X}$. PPCA 首先假设 $\mathbf{X}$ 由 q ($q < d$) 维向量 $\mathbf{z}$ 按下式产生^[1-2]

$$\mathbf{X} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \quad (1)$$

式中: $\mathbf{z}$ 为潜在向量; $\mathbf{w}$ 为权矩阵; $\boldsymbol{\mu}$ 为总体 $\mathbf{X}$ 的均值; $\boldsymbol{\varepsilon}$ 为噪声,一般假设其服从方差为 σ 的高斯分布

$$P(\boldsymbol{\varepsilon}) = P(\mathbf{x} | \mathbf{z}) =$$

$$(2\pi\sigma^2)^{-d/2} \exp(-\|\mathbf{x} - \mathbf{w}\mathbf{z} - \boldsymbol{\mu}\|_2^2 / 2\sigma^2) \quad (2)$$

设 $\mathbf{z}$ 的先验为 $\mathbf{z} \sim N(0, \mathbf{I}_q)$, 将之与式(2)代入贝叶斯公式,可推导出 $\mathbf{z}$ 的后验概率

$$P(\mathbf{z} | \mathbf{x}) = (2\pi)^{-q/2} |\sigma^2 \mathbf{M}|^{-1/2} \cdot$$

$$\exp\left(-\frac{1}{2}(\mathbf{z} - \langle \mathbf{z} \rangle)^T (\sigma^2 \mathbf{M})^{-1} (\mathbf{z} - \langle \mathbf{z} \rangle)\right) \quad (3)$$

式中

$$\mathbf{M} = (\sigma^2 \mathbf{I}_q + \mathbf{w}^T \mathbf{w})^{-1} \quad (4)$$

$$\langle \mathbf{z} \rangle = \mathbf{M}^{-1} \mathbf{w}^T (\mathbf{x} - \boldsymbol{\mu}) \quad (5)$$

PPCA 是概率化的 LPCA,因此可以像 LPCA 一样提取主成分. PPCA 认为 $\langle \mathbf{z} \rangle$ 的 q 个分量就是其求出的 q 个主成分. $\mathbf{w}$ 与 σ 是上述两式中的未知参数,可用最大似然方法求出其估计值

$$\hat{\mathbf{w}} = \mathbf{U}_q (\boldsymbol{\Lambda}_q - \sigma^2 \mathbf{I}_q)^{1/2} \mathbf{R}_o \quad (6)$$

$$\hat{\sigma}^2 = \frac{1}{d-q} \sum_{j=q+1}^d \lambda_j \quad (7)$$

式中: λ_j 是通过矩阵 $\boldsymbol{\Sigma}$ 的特征值分解得到的,即

$$\boldsymbol{\Sigma} \mathbf{v}_j = \lambda_j \mathbf{v}_j \quad j = 1, 2, \dots, d \quad (8)$$

其中 $\boldsymbol{\Sigma}$ 是样本 $x_1, x_2, \dots, x_N$ 的协方差矩阵; λ_j 与 $\mathbf{v}_j$ 参与构建式(6)中的矩阵 $\mathbf{U}_q$ 与 $\boldsymbol{\Lambda}_q$, 即

$$\mathbf{U}_q = (\mathbf{v}_1, \dots, \mathbf{v}_q) \quad (9)$$

$$\boldsymbol{\Lambda}_q = \text{diag}(\lambda_1, \dots, \lambda_q) \quad (10)$$

无论是 LPCA 还是 PPCA,均通过式(8)获得 λ_j 与 $\mathbf{v}_j$,但式(8)中的 $\boldsymbol{\Sigma}$ 是基于样本集估计得到的,如果样本集内存在奇点,估计出的 $\boldsymbol{\Sigma}$ 将不准确,相应地会使 λ_j 与 $\mathbf{v}_j$ 产生误差.奇点产生是多方面的,如测量或记录的偶尔出错、突发的随机噪音等都会产

生奇点.文献[5]引入能量函数,选用迭代方式的学习规则并借助模拟退火法最小化能量函数获取主方向,所采用的能量函数为

$$E(V_i, \mathbf{v}) = \sum_{i=1}^N V_i e(\mathbf{x}_i, \mathbf{v}) + \eta \sum_{i=1}^N (1 - V_i) \quad (11)$$

式中: $e(\mathbf{x}_i, \mathbf{v}) = \left(\mathbf{x}_i^T \mathbf{x}_i - \frac{\mathbf{v}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \right)$; V_i 为二值函数; η 为事先设定的硬门槛值.每个样本 $\mathbf{x}_i$ 对应一个决策变量 V_i ,若 $e(\mathbf{x}_i, \mathbf{v}) \geq \eta$,则判定 $\mathbf{x}_i$ 为奇点,反之 $\mathbf{x}_i$ 为正常点.由此可以看出文献[5]算法有 2 个不足:①能量函数中既含有连续变量 $\mathbf{v}$ 又含有离散变量 V_i ,这增加了使其最小化的难度;②硬门槛 η 不易事先指定,若 η 不合适,判断一个样本是否为奇点时不可避免地会发生误判.

为了减小奇点对样本协方差矩阵和分类效果的影响,拓宽 PPCA 算法的应用范围,本文提出了一种新的鲁棒概率主成分分析方法.

2 鲁棒概率主成分分析方法的构建

2.1 鲁棒概率主成分分析方法的基本思想

基于以上原因,本文引入连续的决策变量 u_i 构建新能量函数,来降低函数最小化的难度,并将事先指定的硬门限改为自适应确定的软门限.首先将能量函数(式(11))改写成

$$E(u_i, \mathbf{v}) = \sum_{i=1}^N u_i^2 e(\mathbf{x}_i, \mathbf{v}) + \eta \sum_{i=1}^N (1 - u_i)^2 \quad (12)$$

式中: $u_i \in [0, 1]$ 是连续变量,用于度量 $\mathbf{x}_i$ 是否为奇点, u_i 值越小, $\mathbf{x}_i$ 是奇点的可能性越大.

因为 u_i 是连续变量,对式(12)求偏导可得

$$u_i = \frac{1}{1 + e(\mathbf{x}_i)/\eta} \quad (13)$$

然后,将式(13)代入到式(12)得到

$$E(u_i, \mathbf{v}) = \sum_{i=1}^n \left(\frac{1}{1 + e(\mathbf{x}_i)/\eta} \right) e(\mathbf{x}_i) \quad (14)$$

通过最小化式(14)中的 $E(u_i, \mathbf{v})$,同样可以得到 $\mathbf{v}$,其中 η 变成了软门限,其值为

$$\eta = \sum_{i=1}^N e(\mathbf{x}_i, \mathbf{v}) / N \quad (15)$$

由式(15)可知, η 是根据样本自动确定的,这样计算更为简单方便.

通过以上算法可获得鲁棒的 $\mathbf{v}$,另外从学习规则也可获得与 $\mathbf{v}$ 对应的 λ .将上述算法融入到 PPCA 的主成分提取过程中,最终得到鲁棒概率主成分分

析(RPPCA)方法.

2.2 RPPCA 方法步骤

RPPCA 方法的实现步骤如下:

(1)构建新的能量函数(式(12)),并对其求偏导,得到 u_i ,再将 u_i 代入得到式(14),并最小化式(14)求得 ν 和对应的 λ ;

(2)将已求出的稳健的 ν 与 λ 分别代入式(9)、式(10),可得 U_q 与 Λ_q 的稳健估计 U'_q 与 Λ'_q ;

(3)再将 U'_q 与 Λ'_q 分别代入式(6)、式(7),获得 w 与 σ 的稳健估计 $\hat{w}'$ 与 $\hat{\sigma}'$;

(4)最后将 $\hat{w}'$ 与 $\hat{\sigma}'$ 分别代入式(4)、式(5)以便求出 $\langle z' \rangle$, $\langle z' \rangle$ 是 $\langle z \rangle$ 的稳健估计,它的 q 个分量就是 RPPCA 求出的 q 个主成分.

3 数值实验

实验是在 MATLAB 6.5 环境下、主机 CPU CELERON 2.8 GHz、内存 256 MB、Windows XP 操作系统上进行的.

实验 1 本实验欲解决 DNA 序列分类问题. DNA 序列由 A、T、C、G 4 个字符组成,并具有一些规律性结构.现给出 200 个学习序列,其中前 100 个为一类(记为+1类),后 100 个为另一类(记为-1类);另给 162 个测试序列,它们的类别需要判断.为了简便起见,所给序列均等长.下面列出了部分学习与测试序列.

(1)第 1 类学习样本:DNA 序列+1 类

aggcacggaaaaacgggaataacggaggaggacttg-
gcacggcattacacggaggacgaggtaaaggaggctgtcta-
cgcccggaagtgaag
.....

(2)第 2 类学习样本:DNA 序列-1 类

gtagatttaacgttttttatggaatttatggaattataaattta-
aaaatttatatttttaggtaagtaacgaacgtttttattacttttt-
aaaattaaatattattat.....

(3)测试样本:DNA 序列

tttagctcagtcagctagctagtttacaatttcgacaccagttt-
cgcaccatcttaaatttcgatccgtaccgtaatttagcttagatttg-
gatttaaaggatttaga
.....

上述 DNA 序列较长,必须先进行特征提取,再用分类器分类.在特征提取前先将 DNA 序列中的字符转化为数字,以便于提取.本实验设计了 3 种方案,每种方案均提取 8 个特征,即 $q=8$.选用的分类器为支持向量机(SVM)^[9],并采用高斯核函数.

SVM 的正则化因子 $C=\infty$,核参数 σ 由交叉验证法自适应确定.表 1 列出了 3 种方案的分类准确率和计算时间.

表 1 DNA 序列的 3 种特征提取及分类方案比较

参数	LPCA-SVM	PPCA-SVM	RPPCA-SVM
分类准确率/%	96.3	98.6	99.1
计算时间/s	228.6	250.3	298.1

从表 1 可以看出,LPCA 提取的特征并不利于分类,PPCA 提取的特征有利于分类,但因为不稳健,所以分类准确率没有本文提出的 RPPCA 高. RPPCA 提取的特征有利于分类且稳健. RPPCA 之所以稳健是因为去除了学习样本中奇点的影响.不过由于 DNA 序列较长,因此奇点并不直观,在实验 2 中奇点会有更直观显示.

实验 2 行人检测是目前智能车辆和计算机视觉研究领域备受关注的问题.常见的行人检测方法包括基于形状^[6]或基于神经网络^[7]的方法等.行人检测中的一个关键环节是判定待检图像是否为行人,这是一个二值分类问题.为此,首先从行人库与非行人库中各自挑出一些图像作为学习样本.本实验采用 CBCL 行人库^[8],图 1 为该库中的一些行人图像,每幅图的大小均为 128×64 像素的彩色图像.本实验采用的非行人库是作者自己构建的.可以看出,行人库中存在明显的奇点,比如第 2 排左起第 6 幅图不是一个行人,因此属于一个奇点.



图 1 行人库中的部分图像

本文设计了如表 2 所示的 3 种特征提取与分类方案.表中样本组 1 代表 100 个行人样本与 100 个非行人样本,样本组 2 代表测试样本始终保持 300

不变,方案中的特征提取均指提取 9 个特征,即 $q=9$. 选用的分类器仍为支持向量机^[9],并采用高斯核函数. 分类结果如表 2 所示,从表 2 可以看出,RPP-CA 提取的特征有利于分类而且稳健.

表 2 行人检测的 3 种特征提取及分类方案比较

分类方案	分类准确率/%	
	样本组 1	样本组 2
LPCA-SVM	88.9	90.3
PPCA-SVM	91.4	92.9
RPPCA-SVM	92.1	94.0

4 结 论

本文提出了一种 RPPCA 方法,与 LPCA 和 PPCA 方法相比,分类准确率有了明显提高,有效地减少了对奇点的误判,既增强了 PPCA 的稳健性,也拓宽了 PPCA 的适用范围. 在实验中,我们也注意到与 LPCA 和 PPCA 相比,本文提出的 RPPCA 计算时间稍长,因此下一步的研究方向是降低算法的复杂度和缩短计算时间. 另外,在实验中没有考虑特征选择环节,如果把 RPPCA 特征提取与特征选择结合起来,在特征提取后即计算各个特征在参与分类时的重要程度,并对特征按重要程度排序,只让重要的特征参与分类,从而提高分类算法的准确度与运算速度.

参考文献:

- [1] TIPPING M E, BISHOP C M. Probabilistic principal component analysis [J]. J Roy Statist Soc, Series B, 1999, 61(3): 611-622.
- [2] BISHOP C M. Latent variable models, learning in graphical models [M]. Jordan M I. Cambridge, MA, USA: MIT Press, 1999: 371-403.
- [3] ZHAO Weixiang, CHEN Dezhaoh, HU Shangxu. Detection of outlier and a robust BP algorithm against outlier [J]. Computers & Chemical Engineering, 2004, 28(8): 1403-1408.
- [4] BULLEN R J, CORNFORD D, NABNEY I T. Outlier detection in scatterometer data: neural network approaches [J]. Neural Networks, 2003, 16(3): 419-426.
- [5] XU Lei, YUILLIE A L. Robust principal component analysis by self-organizing rules based on statistical physics approach [J]. IEEE Trans on Neural Networks, 1995, 6(1): 131-143.
- [6] GAVRILA D M, GIEBEL J. Shape-based pedestrian detection and tracking [C]// Proceedings of IEEE Intelligent Vehicle Symposium. Piscataway, NJ, USA: IEEE Press, 2002: 8-14.
- [7] Zhao L, Thorpe C E. Stereo and neural network-based pedestrian detection [J]. IEEE Trans on Intelligent Transportation Systems, 2000, 1(3): 148-154.
- [8] The MIT-CBCL Face Recognition Database [EB/OL]. (2003-01-08) [2007-11-16]. <http://cbcl.mit.edu/software-datasets/heisele/facerecognition-database.html>.
- [9] CRISTIANINI N, SHAWE-TAYLOR J. An introduction to support vector machines and other kernel-based learning methods [M]. Cambridge, England: Cambridge University Press, 2000.

(编辑 刘杨)

徐宗本教授荣获 CSIAM 苏步青应用数学奖

2008 年 8 月 27 日,在郑州召开的中国工业与应用数学学会第五届代表大会暨第二届 CSIAM 苏步青应用数学奖颁奖大会上,西安交通大学理学院徐宗本教授荣获 CSIAM 苏步青应用数学奖,这也是西安交通大学首次获此殊荣.

据 CSIAM 苏步青应用数学奖奖励委员会公告,徐宗本教授由于“长期致力于数学与信息技术的结合与交叉研究,不仅在 Banach 空间几何与非线性算子理论等数学领域有重要贡献,而且在智能信息处理,特别是机器学习这一信息领域有系统而重要的贡献,多项研究成果处于国际前沿,得到国际学术界广泛认可与研发部门的实质性应用”而获奖. 据悉,本届大会与徐宗本教授同时获奖的还有中科院数学与系统科学研究院张关泉研究员.

(来源:西安交大科技在线 <http://www.xjtust.com>)