软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2022,33(12):4574−4589 [doi: 10.13328/j.cnki.jos.006515]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种基于各向异性高斯核核惩罚的 PCA 特征提取算法*

刘 俊[1], 李 威[1], 陈蜀宇[2], 徐光侠[1]

[1](重庆邮电大学 软件工程学院, 重庆 400065)

[2](重庆大学 大数据与软件学院, 重庆 401331)

通信作者: 刘俊, E-mail: junliu@cqupt.edu.cn

摘 要: 提出了一种基于各向异性高斯核核惩罚的主成分分析的特征提取算法. 该算法不同于传统的核主成分分析算法. 在非线性数据降维中, 传统的核主成分分析算法忽略了原始数据的无量纲化. 此外, 传统的核函数在各维度上主要由一个相同的核宽参数控制, 该方法无法准确反映各维度不同特征的重要性, 从而导致降维过程中准确率低下. 为了解决上述问题, 首先针对现原始数据的无量纲化问题, 提出了一种均值化算法, 使得原始数据的总方差贡献率有明显的提高. 其次, 引入了各向异性高斯核函数, 该核函数每个维度拥有不同的核宽参数, 各核宽参数能够准确地反映所在维度数据特征的重要性. 再次, 基于各向异性高斯核函数建立了核主成分分析的特征惩罚目标函数, 以便用较少的特征表示原始数据, 并反映每个主成分信息的重要性. 最后, 为了寻求最佳特征, 引入梯度下降算法来更新特征惩罚目标函数中的核宽度和控制特征提取算法的迭代过程. 为了验证所提出算法的有效性, 各算法在 UCI 公开数据集上和 KDDCUP99 数据集上进行了比较. 实验结果表明, 所提基于各向异性高斯核核惩罚的主成分分析的特征提取算法比传统的主成分分析算法在 9 种公开的 UCI 公开数据集上准确率平均提高了 4.49%. 在 KDDCUP99 数据集上, 所提基于各向异性高斯核核惩罚的主成分分析的特征提取算法比传统的主成分分析算法准确率提高了 8%.

关键词: 各向异性高斯核; 特征惩罚函数; 主成分分析; 梯度下降法

中图法分类号: TP18

中文引用格式: 刘俊, 李威, 陈蜀宇, 徐光侠. 一种基于各向异性高斯核核惩罚的PCA特征提取算法. 软件学报, 2022, 33(12): 4574–4589. http://www.jos.org.cn/1000-9825/6515.htm

英文引用格式: Liu J, Li W, Chen SY, Xu GX. PCA Feature Extraction Algorithm Based on Anisotropic Gaussian Kernel Penalty. Ruan Jian Xue Bao/Journal of Software, 2022, 33(12): 4574–4589 (in Chinese). http://www.jos.org.cn/1000-9825/6515.htm

PCA Feature Extraction Algorithm Based on Anisotropic Gaussian Kernel Penalty

LIU Jun[1], LI Wei[1], CHEN Shu-Yu[2], XU Guang-Xia[1]

[1](School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

[2](School of Big Data & Software Engineering, Chongqing University, Chongqing 401331, China)

Abstract: This study proposes a feature extraction algorithm based on the principal component analysis (PCA) of the anisotropic Gaussian kernel penalty which is different from the traditional kernel PCA algorithms. In the non-linear data dimensionality reduction, the nondimensionalization of raw data is ignored by the traditional kernel PCA algorithms. Meanwhile, the previous kernel function is mainly controlled by one identical kernel width parameter in each dimension, which cannot reflect the significance of different features in each dimension precisely, resulting in the low accuracy of dimensionality reduction process. To address the above issues, contraposing the current problem of nondimensionalization of raw data, an averaging algorithm is proposed in this study, which has shown sound performance in improving the variance contribution rate of the original data typically. Then, anisotropic Gaussian kernel function is

* 基金项目: 国家自然科学基金 (61772099, 61772098); 重庆市自然科学基金 (cstc2021jcyj-msxmX0530); 重庆市“三百”科技创新领军人才支持计划 (CSTCCXLJRC201917); 重庆市创新创业示范团队培育计划 (CSTC2017kjrc-cxcytd0063)
收稿时间: 2021-04-09; 修改时间: 2021-09-12; 采用时间: 2021-10-14; jos 在线出版时间: 2021-11-24

introduced owing each dimension has different kernel width parameters which can critically reflect the importance of the dimension data features. In addition, the feature penalty function of kernel PCA is formulated based on the anisotropic Gaussian kernel function to represent the raw data with fewer features and reflect the importance of each principal component information. Furthermore, the gradient descent method is introduced to update the kernel width of feature penalty function and control the iterative process of the feature extraction algorithm. To verify the effectiveness of the proposed algorithm, several algorithms are compared on UCI public data sets and KDDCUP99 data sets, respectively. The experimental results show that the feature extraction algorithm of the PCA based on the anisotropic Gaussian kernel penalty is 4.49% higher on average than the previous PCA algorithms on UCI public data sets. The feature extraction algorithm of the PCA based on the anisotropic Gaussian kernel penalty is 8% higher on average than the previous PCA algorithms on KDDCUP99 data sets.

Key words: anisotropic Gaussian kernel; feature penalty function; principal component analysis (PCA); gradient descent algorithm

1 引 言

在数据分析、数据挖掘、模式识别等研究领域中, 数据维数问题是被广大研究学者一直关注的经典问题之一. 降维是解决维数灾难的常用方法. 主要是通过算法大幅度降低数据的维数, 并保留数据中的大部分信息. 特征提取是降维的重要部分, 其主要通过寻找一个子集的可用特征建立一个良好的预测模型来解决降维问题. 在特征提取算法中, 主成分分析 (principal component analysis, PCA) 算法[1]是科研工作者一直研究的无监督数据降维算法之一, 它是降维方法中最常用的一种算法[2]. 主成分分析算法的思想在于很简单, 即减少数据集的维数, 用更少的特征尽可能多地保留原始数据的信息.

主成分分析功能强大且用途广泛, 应用于许多领域的经典统计技术. 它能够提供复杂的多变量概述的方法整理数据[3], 它也被称为 KL (Karhunen-Loeve) 变换[4]. 但是原始数据集在通过主成分分析进行降维之后得到的各个特征维度的含义具有一定的模糊性, 解释不清楚其数据的具体含义, 不如原始样本的解释性强, 而且降维会有一定数据的丢失[5], 方差小的非主成分信息也可能含有对样本差异的重要信息, 会对后续数据的处理产生一定的影响.

目前, 众多改进的 PCA 算法研究都主要集中在对特征提取、特征向量方向以及综合值计算的改进. 在对特征提取的改进中, 主要是对原始数据进行对数变换、平方根变换处理消除主成分之间的无量纲化影响. 在特征向量方向的确定上, 不同的方向会直接影响各个主成分的方差贡献率. 此外, 用熵值法改进传统的主成分评价方法能很好地消除用方差贡献率作为权值所带来的主观成分. 但在非线性主成分分析中, 常用的核函数在各维度上主要由同一个参数控制, 各个方向的参数都一样, 体现出一定的局限性.

因此选择由不同核宽参数控制每个维度的核函数一个重要的关注方面. 各向异性高斯核函数的核宽可以反映每个特征的重要程度[6], 许多研究学者通过各向异性高斯核代替传统的高斯核函数去控制不同方向的参数, 提取出重要的信息. Li 等人[7]为了实现亚像素级的精确定位和精确的形状估计, 提出了一种新的仿射不变斑点测量方法, 通过各向异性高斯核描述斑点的五参数形状, 并提出基于梯度搜索收敛的斑点识别用于自动去除低质量的斑点, 最后得到了较高的定位精度和形状估计精度. Zhao 等人[8]在图像分割目标识别的关键技术研究中, 提出了一种基于各向异性高斯核 (ANGK) 边缘检测和区域邻接图 (RAG) 合并算法的混合分割方法. 利用角度 ANGK 构造各向异性方向导数滤波器来检测原始图像的边缘轮廓. 与传统的基于边缘和区域的方法相比, 该方法具有更好的分割效果.

目前基于核函数的主成分分析研究大多数都是针对几种常见的核函数. 并且在各维度上的映射主要由一个核宽控制, 而各个方向的相同核宽参数不能具体反映出每个特征的重要程度. 各向异性高斯核每个方向的控制参数可以为不相同值, 因此可以从不同的方向反映数据特征的变换信息. 并且在数据局部结构特征不清晰的情况下, 通过给不同方向设置不同的参数, 各向异性高斯核已经被证明能够较好地提取各个方向的有用特征[9]. 因此近年来, 用各向异性高斯代替传统的核函数去解决非线性问题越来越多, 并取得了显著的效果.

基于以上分析, 本文提出了一种基于各向异性高斯核核惩罚的主成分分析方法, 利用各向高斯核每个方向可以设置不同参数的特点, 反映出主成分信息每个特征的重要程度, 易于更好的大数据分析处理.

本文的贡献总结如下.

(1) 针对现原始数据的无量纲化问题, 提出了一种均值化算法以提高原始数据信息主成分的总方差贡献率.

(2) 针对现有核函数用相同核宽表示不同维度特征重要性而导致特征提取准确率低的问题, 提出了用各向异性高斯核的多维核宽向量表征不同维度特征重要性的方法.

(3) 基于各向异性高斯核, 建立了基于核惩罚函数的主成分分析特征提取目标函数. 目标函数通过 l_0 范数惩罚函数删除不重要的核宽向量. 为了获得特征提取算法合适的迭代次数和得到特征提取的最优解, 提出了一种梯度下降算法更新各向异性高斯核函数的核宽度和不重要特征的删除, 以尽可能少的特征去表示原始数据.

本文第 2 节介绍了本文的相关工作; 第 3 节简要介绍了 PCA 线性降维方法和 KPCA 非线性降维方法; 第 4 节说明了本文的动机; 第 5 节提出了一种基于各向异性高斯核核惩罚的主成分分析的特征提取算法; 第 6 节给出了 9 个公开数据集以及 KDDCUP99 数据集的实验结果并进行分析评价; 第 7 节进行了总结并展望未来工作.

2 相关工作

主成分分析是一个无监督学习问题, 它是基于方差去提取最有价值的信息[10]. 此外, 通过数据降维可以减轻维数灾难[11]和高维空间中其他不相关属性. 目前在对主成分分析算法的优化主要集中在特征提取的改进、特征向量方向的确定以及主成分综合值计算的改进方法上.

在特征提取的改进中, 主要是对需要进行主成分分析的原始数据集进行对数变换或者平方根变换等一些消除变量之间的无量纲化方法. 宋昱等人[12]对主成分分析在图像识别的研究中, 提出了一种对数变换的主成分分析算法, 将原始数据集进行对数变换处理, 进一步提升了传统的主成分分析算法处理图像识别领域中含有异常样本数据的性能, 能得到最高的识别精度和最低的重构误差. Tucker 等人[13]提出了一种弹性函数主成分回归算法, 该算法对时间测量中的误差, 函数未对齐等问题进行相位去除和功能校准达到对平方根斜率函数的改进, 大大提高了模型的预测正确性.

特征向量方向的正确选择会直接影响各个样本降维之后的总方差贡献率. Gu[14]提出了一种基于多特征和主成分分析的海上监视雷达小浮目标检测方法. 该方法主要分为 3 个阶段: 第 1 阶段是对雷达波中的特征进行选取组合成一个特征向量以达到最高的总方差贡献率. 第 2 阶段为对特征向量进行矩阵分解. 第 3 阶段为构造基于 PCA 的异常检测器. 第 1 阶段为第 2 和第 3 阶段的特征矩阵分解和异常检测器的建立提供最准确的信息, 该方法大大提高了模型目标检测的检测率. Bhandary 等人[15]为了能够快速准确地诊断出人类肺部异常以进行快速有效的治疗, 采用序列融合和基于主成分分析的特征选择来增强特征向量的选择, 提高了肺癌评估期间的分类准确性.

熵值法、主成分聚类法是从主成分综合值计算方面进行优化. 高光谱图像通常将土地覆盖类型的信息保存为一组连续的窄光谱波段, 为了有效进行分类, Uddin 等人[16]首先对主成分分析算法的目标函数进行优化, 然后对最终得到的主成分进行熵值法处理, 以避免由于根据方差贡献率作为权值来计算综合评价带来的主观成分, 最后提出了一种基于 Renyi 二次熵的特征选择和改进的主成分分析结合使用的算法, 提高了模型的分类性能.

在美国, 农民的生产水平与信贷约束有很大的联系. Griffin 等人[17]用基于主成分聚类的倾向得分匹配模型, 从可用的农业资源管理调查数据中得出农民的生产水平以确定其信贷约束.

本研究中, 对主成分分析算法的优化主要是在核函数的选择上即特征向量方向的确定, 选择了各向异性高斯核函数来代替传统的高斯核函数, 既展现了核函数线性不可分特点, 又体现了各个主成分特征的重要程度, 并且提高了核主成分分析算法的性能.

3 PCA 线性降维和 KPCA 非线性降维

3.1 PCA 线性降维

主成分分析是一个无监督学习问题, 它是一种常用的降维和特征提取方法, 通过将高维数据映射到方差最大的数轴上, 丢弃方差较小的数轴来达到降维目的. 它的主要思想是将一组 N 维向量数据降为 $K(0 < K < N)$ 维不相

关变量, 即主成分[18]. 具体目标是使原始数据变换到一个正交基上后各字段两两间协方差为0, 字段的方差则尽可能大. 并且使得原始数据在相互独立的方向上的投影能够尽可能分散, 其中, 尽可能分散就是为了能更多地保留原始信息, 当然寻找的方向相互独立就是为了避免保留下来的信息存在冗余.

将求最大方差的问题通过拉格朗日乘子法转化为求数据矩阵的特征值问题:

① 求最大方差:

$$\begin{cases} \max\ V^{\mathrm{T}}CV, \\ \text{s.t. } \|V\| = 1. \end{cases}$$

又由于$\|V\| = V^{\mathrm{T}}V$, 故s.t. $V^{\mathrm{T}}V = 1$即可.

② 转化为求特征值: 利用拉格朗日乘子法可以将上述问题转化为:

$$f(v,\lambda) = V^{\mathrm{T}}CV - \lambda\left(V^{\mathrm{T}}V - 1\right) \tag{1}$$

其中, $f(v,\lambda)$的平稳点和求最大方差问题等价:

$$\begin{cases} \dfrac{\partial f}{\partial v} = 2CV - 2\lambda V = 0 \\ \dfrac{\partial f}{\partial v} = V^{\mathrm{T}}V - 1 = 0 \end{cases} \tag{2}$$

公式(2)等价于: $\begin{cases} CV = \lambda V \\ \|V\| = 1 \end{cases}$. $CV = \lambda V$就是求数据矩阵的特征值和特征向量. 因此求出特征值和特征向量并从大到小进行排序, 选择与最大特征值对应的特征向量组成一个正交变换矩阵. 最后原始矩阵通过这个正交矩阵进行变换就得到了线性无关向量组成的矩阵, 即原始矩阵的主成分. Delchambre[19]也证明过PCA降维效果就是由协方差矩阵的特征值和相应的特征向量的大小确定的.

3.2 核PCA (KPCA)非线性降维

传统的主成分分析算法不能处理非线性的数据. 因此为了扩展对非线性数据的处理, 核PCA[20]被引入. 其算法思想是对于输入空间中的矩阵X, 通过核函数即非线性映射把X中的所有样本数据映射到一个高维度甚至是无穷纬度的特征空间F中, 再利用PCA算法对其在高维度F空间中的数据集进行降维. 与PCA算法相似之处在于, 二者都是通过数据矩阵的变换将其投影到新的低维空间中. 不同的是核PCA算法可实现数据的非线性降维, 用于处理线性不可分的数据集, 它其实是一个改进的PCA方法, 采用了非线性的核函数来提取主成分.

设$x_1, x_2, x_3, \ldots, x_n \in R$是要进行KPCA特征提取的$n$个高维数据, 那么降维后的主成分信息$t_i$就可以通过以下的方法来获取:

$$t_i = \frac{1}{\sqrt{\lambda_i}}\gamma_i^{\mathrm{T}}[k(x_1, x_{\mathrm{new}}), k(x_2, x_{\mathrm{new}}), \ldots, k(x_n, x_{\mathrm{new}})]^{\mathrm{T}},\ i = 1,2,\ldots,p \tag{3}$$

这里的列向量$\gamma_i(i = 1,2,\ldots,p;\ 0 < p < n)$是相对应的$p$个最大特征值$\left(\lambda_1 \geqslant \lambda_2, \ldots, \geqslant \lambda_p\right)$的正交特征向量, $k\left(x_i, x_j\right)$是内核函数实现在高维空间F中两个向量的内积运算, 不用知道核函数具体的表达式, 它可以直接得到低维数据映射到高维后的内积, 提供了一个从线性到非线性的连接. 它会根据你选择的核函数来生成内核矩阵, 并且每个核函数都有不同的表达式来计算自己的内核矩阵. 相比于PCA, KPCA不仅能够解决非线性结构的问题, 还能得到更高质量的主成分信息, 尽可能地去抽取原始指标包含的信息. 但是两者在降维之后的主成分信息具体的实际意义不明确, 需要根据实际情况选择不同的特征提取方法.

4 核函数选择的重要性

核函数的正确选择是核主成分分析算法的重要步骤, 直接决定了KPCA算法的非线性处理能力[21]. 常见的核函数有线性核函数、多项式核函数、径向基核函数、Sigmoid核函数等. 选择正确适合的核函数可以高效处理高维特征空间中计算量巨大、“维数灾难”等问题. 常见的几种核函数如表1所示.

表 1 4 个常见的核函数

英文名称	缩写	数学表达式
Linear kernel	Linear	$k(x,y)=x^{\mathrm{T}}y+c$
Polynomial kernel	Poly	$k(x,y)=\left(ax^{\mathrm{T}}y+c\right)^{d}$
Radial basis kernel	RBF	$k(x,y)=\exp\left(-\gamma\|x-y\|^2\right)$
Exponential kernel	EK	$k(x,y)=\exp\left(-\|x-y\|/2\sigma^2\right)$

通常在没有先验知识的情况下, 人们都会利用自己的主观经验去选择核函数, 具有很大的随意性. 并且常见的几种传统核函数在各维度上主要由同一个参数控制, 各个方向的参数都一样, 无法准确反应 KPCA 体现每个主成分的重要性, 有一定的局限性.

因此, 选择一个能够让不同的方向由不同的参数控制的核函数, 使其能够体现出每个特征的重要程度, 并提高核主成分分析算法的性能非常重要.

5 基于各向异性高斯核核惩罚的主成分分析方法

本文提出的基于各向异性高斯核核惩罚的 PCA (AP-KPCA) 算法主要由均值化算法、核函数的确定、特征惩罚函数的选择、改进的目标函数组成. 均值化算法主要用来处理原始样本, 改进样本数据的无量纲化. 核函数主要用各向异性高斯核替代了传统的高斯核函数, 利用多核宽参数来控制原始数据映射到高维空间的过程. 特征惩罚函数对降维过程中核参数进行惩罚, 以便提取重要的特征.

5.1 均值化处理数据集算法

AP-KPCA 的一个关键步骤是求主成分信息, 通常会对原始数据集进行标准化处理来消除变量量纲的影响, 但在消除量纲的同时, 也消除了各指标之间变异程度的差异信息. 事实上, 原始指标是包含两方面的信息. 一部分是由相关系数矩阵来体现的各指标之间的相关信息, 另一部分是由各指标的方差大小来反映的各指标变异程度的差异信息. 原始数据的标准化使各指标的方差都变成了 1, 消除了各指标之间变异程度上的差异信息. 因此从数据标准化之后得到的主成分, 不能准确反映原始数据的全部信息.

均值化后数据的协方差矩阵的对角元素是各指标的变异系数的平方, 它反映了各指标变异程度上的差异[22]. 因此, 均值化处理不会改变各指标间的相关系数, 并且协方差矩阵反映了相关系数矩阵的全部信息. 它不仅消除了原始指标量纲和数量级的影响, 还能包含原始数据的全部信息, 即用更少的特征包含更多的原始信息, 提高方差贡献率.

因此, 在原始样本数据处理方面, 本文提出的 AP-KPCA 采用均值化算法对 PCA 算法的原始样本数据进行改进. 假设有 n 个被评价的对象和 p 个指标, 那么原始样本数据中第 i 个对象的第 j 个评价指标可以定义为: X_{ij} ($i=1,2,\ldots,n$; $j=1,2,\ldots,p$); 每个评价指标的均值为 $\overline{x}_j$, $\overline{x}_j$ 定义为: $\overline{x}_j=X_{ij}/j$. 均值化定义为各个指标的均值除以它们对应的原始数据, 假设均值化定义为 Z_{ij}, 其可定义为公式 (4):

$$Z_{ij}=X_{ij}/\overline{x}_j \tag{4}$$

经过均值化后每个协方差矩阵 V 的元素为:

$$u_{ij}=\frac{1}{n-1}\sum_{i=1}^{n}(Z_{li}-Z_i)(Z_{lj}-Z_j) \tag{5}$$

根据上述表达可知, 在经过均值化后的各个指标的均值为 1, 因此有:

$$u_{ij}=\frac{1}{n-1}\sum_{i=1}^{n}(Z_{li}-1)(Z_{lj}-1)=\frac{1}{n-1}\sum_{l=1}^{n}\frac{(x_{li}-x_i)(x_{lj}-x_j)}{x_ix_j}=\frac{R_{ij}}{x_ix_j} \tag{6}$$

其中, R_{ij} 就为原始数据的协方差. 假设当 $i=j$ 时, 协方差矩阵就为 $u_{ij}=\left(\frac{\overline{R_{ii}}}{x_i}\right)^2$, $R_{ii}=\frac{1}{n}\sum_{l=1}^{n}(x_{li}-x_i)^2$, 所以说在经过均值化后数据的协方差矩阵 V 的对角元素是各个指标的变异系数 S_{ii}/X_i 的平方. 在均值化前, 各个指标之间的相互影响程度的相关系数 r'_{ij} 为: $r'_{ij}=\frac{R_{ij}}{\overline{R_{ii}}\,\overline{R_{jj}}}$, 均值化后的各个指标之间的相互影响程度的相关系数 $r_{ij}=\frac{u_{ij}}{\overline{u_{ii}}\,\overline{u_{jj}}}$, 将公式 (6) 代入可得:

$$r'_{ij}=\frac{R_{ij}}{x_ix_j}\Big/\frac{\overline{R_{ii}}}{x_i}\frac{\overline{R_{jj}}}{x_j}=\frac{R_{ij}}{\overline{R_{ii}}\,\overline{R_{jj}}}=r_{ij} \tag{7}$$

根据以上证明可知, 经过均值化后原始数据的协方差矩阵中不仅包含了由于标准化所带来的各个指标各指标之间变异程度, 并且也消除了指标量纲与数量级的影响.

5.2 各向异性高斯核函数

常见的核函数有高斯核、线性核、多项式核、西蒙核等, 本文选择的各向异性高斯核是传统高斯核的一种改进. 传统的高斯核函数将原始数据映射到高维空间的过程中主要由一个参数控制, 即每个方向维度的参数都一样, 不能反映出每个特征的重要性. 而各向异性高斯核可以对特征向量的每个维度设置不同的核参数, 通过设置不同方向的参数, 提取各个方向有用的特征信息[23]. 其定义为:

$$K(x_i,x_s,v)=\exp\left[-\sum\nolimits_{j=1}^{n}\frac{(x_{ij}-x_{sj})^2}{2\sigma_j^2}\right] \tag{8}$$

其中, x_i,x_s 为样本空间, n 为样本的特征维度, σ_j 为各向异性高斯核函数的核宽参数, 对应 n 维度样本特征, 具体为 $\sigma=[\sigma_1,\sigma_2,\sigma_3,\ldots,\sigma_n]$. 不同的核宽参数能反映样本数据中某个特征的重要程度[9], 即特征 n 的重要程度由 σ_n 来决定. 例如: σ_n 的值越大, 那么 $K(x_i,x_s,v)$ 值就越接近于 0, 贡献也接近于 0, 间接就说明特征 n 的重要程度较低. σ_n 的值越小, 那么 $K(x_i,x_s,v)$ 值就会很大, 贡献也就会很大, 间接就说明特征 n 的重要程度较高. 将 σ_n 值较大的对应的特征 n 删掉, 提取 σ_n 值较小的对应的特征, 这样就达到了特征提取的目的. 因此, 核宽向量 v[23]的定义为:

$$v=\left[\frac{1}{\sigma_1},\frac{1}{\sigma_2},\frac{1}{\sigma_3},\ldots,\frac{1}{\sigma_n}\right] \tag{9}$$

$$\begin{aligned}K(x_i,x_s,v)=&\exp\left[-\sum\nolimits_{j=1}^{n}\frac{(x_{ij}-x_{sj})^2}{2\sigma_j^2}\right]=\exp\left[-\sum\nolimits_{j=1}^{n}\frac{(x_{ij}-x_{sj})^2}{2\left(\frac{1}{v_j}\right)^2}\right]=\exp\left[-\frac{\sum_{j=1}^{n}\left[v_j(x_{ij}-x_{sj})\right]^2}{2}\right]\\=&\exp\left[-\frac{\sum_{j=1}^{n}\left[v_jx_{ij}-v_jx_{sj}\right]^2}{2}\right]=\exp\left[-\frac{(v_1x_{i1}-v_1x_{s1})^2+(v_2x_{i2}-v_2x_{s2})^2+\ldots+(v_nx_{in}-v_nx_{sn})^2}{2}\right]\\=&\exp\left[-\frac{\sum_{j=1}^{n}\left[v_jx_{ij}-v_jx_{sj}\right]^2}{2}\right]\end{aligned} \tag{10}$$

因此核函数变为:

$$K(x_i,x_s,v)=\exp\left[-\frac{v*x_i-v*x_s{}^2}{2}\right] \tag{11}$$

其中, $a*b=(a_1b_1,a_2b_2,\ldots,a_nb_n)$, v 也被称为各向异性高斯核的核宽向量.

5.3 特征惩罚函数

特征惩罚函数可以通过外部罚函数法、内部罚函数法等将目标函数由有约束问题转化为无约束优化问题. 其中 l_0 范数逼近具有非平滑的特性并且可以用来寻找最少最优的系数特征项, 已经被广泛应用在有向量参数的目标函数中, 在基于支持向量机和 K-means 的特征选择算法中均得到很好的验证[9,23]. 因此本文也采用 l_0 范数逼近的

方法应用在核主成分分析的特征提取过程中. 根据文献 [9,23], l_0 范数 $\|w\|_0$ 可以由一个凹函数近似表达为:

$$\|w\|_0 \approx e^{\mathrm{T}}(e-\exp(-\beta|w|)) \tag{12}$$

其中, $e=(1,\ldots,1)^{\mathrm{T}}$, $\beta\in R_+$, 并根据公式 (12) 提出了特征惩罚函数:

$$f(v)=e^{\mathrm{T}}(e-\exp(-\beta|w|))=\sum_{j=1}^{n}\left(1-\exp\left(-\beta v_j\right)\right) \tag{13}$$

这里的 v_j 描述的是各向异性高斯核中的核宽参数. β 为近似参数, 根据文献 [9,23] 的结论, β 的值设置为 5 时效果较好而且适应的范围最广, 因此本文也将 β 的值设置为 5.

5.4 目标函数的建立

在求解 PCA 运算过程中, 目标是选择更少的单位正交基, 使原始数据变换到这组基上后, 各字段两两间协方差为 0, 方差尽可能大. 而计算得到的协方差矩阵中对角线元素则是两两字段间的方差, 其他元素则是两两字段间的协方差, 将协方差矩阵进行对角化便可得到其特征值和特征向量. 因此它的求解目标公式 (14) 定义为:

$$\left(\sum_{i=1}^{m}Z_iZ_i^{\mathrm{T}}\right)W=\lambda W \tag{14}$$

其中, Z_i 是样本点 x_i 在高维空间中的像, W 为特征向量组成的矩阵, λ 为特征值, 简化得:

$$W=\sum_{i=1}^{m}Z_i\left(Z_i^{\mathrm{T}}W\right)/\lambda=\sum_{i=1}^{m}Z_i\alpha_i \tag{15}$$

然后假定 Z_i 是由原始属性空间中的样本点 x_i 通过非线性映射 ϕ 产生的, 那么将公式 (14) 和公式 (15) 改成如下:

$$\left(\sum_{i=1}^{m}\phi(x_i)\phi(x_i)^{\mathrm{T}}\right)W=\lambda W \tag{16}$$

$$W=\sum_{i=1}^{m}\phi(x_i)\alpha_i \tag{17}$$

引入各向异性高斯核函数:

$$K\left(x_i,x_j,v\right)=\exp\left[-\frac{v*x_i-v*x_j{}^2}{2}\right] \tag{18}$$

化简后得到 $KA=\lambda A$, 其中, K 为对应的核矩阵, $A=(\alpha_1;\alpha_2;\ldots;\alpha_m)$

最后得到样本 x_i 在投影后的第 j 维坐标 Z_j 为:

$$Z_j=W_j^{\mathrm{T}}\phi(x)=\sum_{i=1}^{m}\alpha_i^jK\left(x_i,x_j,v\right) \tag{19}$$

因此, 目标函数可以写为:

$$\min_v F(v)=\sum_{i=1}^{m}\alpha_i^jK\left(x_i,x_j,v\right) \tag{20}$$

目标函数通过 l_0 范数惩罚函数删除不重要的核宽向量 v 以及对应的特征向量, 并且所选择的特征尽可能地包含了原始特征信息. 因此将特征惩罚函数公式 (13) 引入公式 (20) 中建立 AP-KPCA 算法的最小化目标函数公式 (21):

$$\begin{cases}\min\limits_v F(v)=\sum\limits_{i=1}^{m}\alpha_i^jK\left(x_i,x_j,v\right)+\mu f(v)\\ v_i\geqslant 0,\ \forall i\in\{1,\ldots,N\}\end{cases} \tag{21}$$

其中, μ 是预定义的参数, 用于惩罚 $f(v)$.

基于各向异性高斯核的核惩罚的主成分分析算法的具体步骤如算法 1.

算法 1. 基于各向异性高斯核的核惩罚的主成分分析算法.

Input: x_i: Data sample, $x_1,\ldots,x_k$;

Output: Z_j: Data samples after dimensionality reduction, $Z_1,\ldots,Z_k$ $(k\ll i)$.

1. Begin

2. Define the eigenvector matrix: W; Eigenvalues: λ;

3. for $(i=0;i\leqslant m;i++)\{$
4. $\quad W=\phi(x_i)\alpha_i;$
5. $\}$
6. for $(i=0;i\leqslant m;i++)\{$
7. $\quad Z_j=\alpha_i^j K\left(x_i,x_j,v\right);$
8. $\}$
9. The feature vector with variance contribution rate > 85% is selected as the data sample after dimensionality reduction: $Z_1,\ldots,Z_k\ (k\ll i)$;
10. End

5.5 特征删除和核宽的更新

基于梯度下降的特征删除算法已经在 SVM 和 K-means[9,23,24]的算法中得到了验证, 因此, 本文也采用了梯度下降的算法对特征进行删除和更新核宽, 在迭代的过程中最小化核宽向量 $v=\left[\frac{1}{\sigma_1},\frac{1}{\sigma_2},\frac{1}{\sigma_3},\ldots,\frac{1}{\sigma_n}\right]$, 并在每次迭代中对不重要的特征信息 λ_j 进行删除以及核宽向量 v 的更新, 具体算法步骤如算法 2.

算法 2. 特征删除和核宽的更新.

1. initialization: $v=v_0e$
2. definition: $EndFlag=$ true; $t=0$
3. while $(EndFlag==$ true$)$ do
4. $\quad$ KPCA (ANGKS);
5. $\quad v^{t+1}=v^t-\gamma\nabla F(v^t)$;
6. $\quad$ for all $\left(\lambda_j^{t+1}<\in\right)$ do
7. $\quad\quad v_j^{t+1}=\lambda_j^{t+1}=0$;
8. $\quad\quad$ endfor
9. $\quad$ if $(v^{t+1}==v^t)$ then
10. $\quad\quad EndFlag=$ false;
11. $\quad\quad$ endif
12. $\quad t=t+1$;
13. endwhile

利用梯度下降算法对核宽向量进行更新, 当核宽向量集合 v 中对应特征值的方差贡献率小于预先定义的 ϵ 则将其删除 (对应算法 2 第 5–8 行); 当 t 时刻的核宽向量与 $t+1$ 时刻的核宽向量相近似的时候, 则整个算法结束 (对应算法 2 的第 9–11 行). 对于特征 j, 梯度下降函数为:

$$\nabla_j F(v)=\sum_{i,s=1}^{m} v_j\left(x_{i,j}-x_{s,j}\right)^2\alpha_i^j K(x_i,x_s,v)+\mu\beta\exp\left(-\beta v_j\right) \tag{22}$$

6 实验与分析

6.1 实验数据集

UCI 公开数据集[25,26]是一种标准的测试数据集并被广泛地应用于各种机器学习算法的测试中. 因此本文为了检测本文提出的 AP-KPCA 算法的有效性, 也选取了 UCI 公开数据集作为测试数据. 此外, 为了更有效地验证算法在实际场景中的有效性, 引入了网络攻击环境中常用的 KDDCUP99 数据集中进行验证. 在 UCI 公开数据集中, 选

择了 abalone、column、glass、iris、cmc、south、segment、waveform 和 wine 这 9 个大小不同的测试数据集, 这些测试数据集涉及医学、自然科学和物理等学科领域. 数据集的样本数目的范围较广, 样本数目较小的 iris 数据集为 150, 而样本数目最大的数据集 waveform 个数为 5 000. 而特征维度最小的是 iris 数据集, 维度为 3, 特征最大的维度是 waveform, 维度为 21. 从数据集的多样性、样本和维度的广泛围可以在一定程度上测试所提出算法的有效性. 各数据集的详细信息 (数据集名称、类别数目、样本数和特征个数) 如表 2 所示.

表 2 数据集

数据集名称	类别数目	样本数	维数 (原始特征个数)
abalone	3	4 177	9
column	2	310	6
glass	6	214	9
iris	3	150	4
cmc	3	1 473	9
south	2	1 000	20
segment	7	2 310	19
waveform	3	5 000	21
wine	3	178	13
KDDCUP99	2	2 000 000	41

6.2 数据集的预处理

用均值化方法对选取的 9 个 UCI 公开数据集以及 KDDCUP99 进行预处理, 为了证明所提出的均值化方法的有效性, 将其得到的方差贡献率与传统的 PCA 算法进行对比, 得到表 3.

表 3 方差贡献率对比

数据集名称	维数 (降维之后的)	传统的PCA算法	改进的PCA算法
abalone	2	0.947 792 06	0.960 330 89
column	3	0.866 910 33	0.979 667 76
glass	5	0.893 104 96	0.999 061 41
iris	2	0.958 009 75	0.978 840 42
cmc	6	0.850 183 34	0.990 605 49
south	13	0.858 834 66	0.925 521 02
segment	7	0.886 513 02	0.988 264 80
waveform	9	0.851 143 92	0.999 931 17
wine	7	0.893 367 94	0.947 657 88
KDDCUP99	13	0.908 179 23	0.941 579 03

从表 3 可以明显看出所选取的 10 个公开数据集在进行均值化处理之后, 明显提高其相应的方差贡献率, 用同样的特征维度表示出更多的原始信息, 并剔除了原始数据中不重要的冗余信息.

6.3 评价指标与评价方法

为了验证本文提出的 AP-KPCA 算法的有效性. 首先将 UCI 数据样本和 KDDCUP99 数据集采用 AP-KPCA 算法进行降维, 然后采用 SVM 算法对降维后的数据进行分类, 得到全局最优解 [27]. 最后采用谢娟英等人 [28]在研究基于基因表达数据进行疾病诊断时所选择的准确率 (ACC)、精确率 (AUC)、召回率 (recall)、$F1$-score 这 4 个经典的评价指标对分类的有效性进行验证.

本研究实验的评价标准都是通过 5 次实验取平均值作为最后的结果. 实验首先比较了未加入惩罚项的基于各向异性高斯核的主成分分析算法 (KPCA(ANGKS)) 与线性主成分分析方法 (PCA)、基于线性核函数的主成分分析方法 (KPCA(linear))、基于高斯核函数的主成分分析方法 (KPCA(rbf))、基于多项式核函数的主成分分析方法

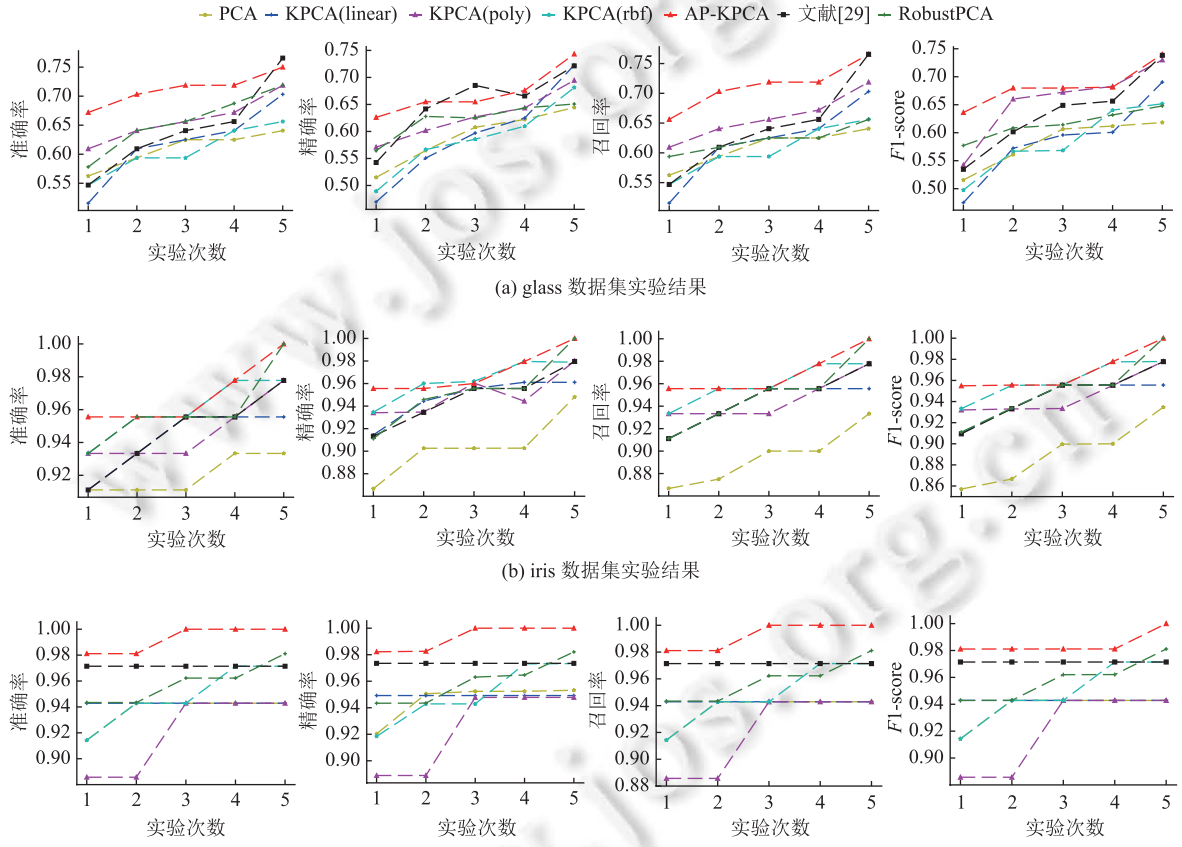
(KPCA(ploy))、文献 [29] 所提出的改进 LDA 的特征提取算法以及鲁棒性主成分分析方法 (RobustPCA) 分别对 10 个公开数据集进行降维, 再分别用 SVM 分类器[30] (统一惩罚因子 C 取 100, 核函数采用线性核函数, 其余参数均取默认值) 对降维之后的主成分信息作分类预测, 计算出准确率、精确率、召回率和 $F1$-score 并进行对比. 然后比较加入惩罚项的基于各向异性高斯核的主成分分析算法 (AP-KPCA) 与 KPCA(ANGKS) 算法的性能.

6.4 实验结果与分析

本节首先比较了本文提出的 AP-KPCA 算法、PCA 算法、KPCA(linear) 算法[31]、KPCA(rbf) 算法[32]、KPCA(ploy) 算法[33]、文献 [29] 提出的改进 LDA 的特征提取算法以及 RobustPCA 算法在表 2 数据集中的性能, 比较各特征提取的主成分信息对应分类器的各指标值, 然后对各算法在 5 次实验之后准确率平均值进行了比较, 最后对核主成分分析的核函数的惩罚项以及核参数进行了分析和讨论.

6.4.1 实验结果

本小节以 iris、glass 和 wine 数据集为例, 采用 10 折交叉验证方法划分训练集与测试集, 对比 AP-KPCA 算法与各算法采用 SVM 分类的实验结果. 图 1 分别是各算法在 iris、glass 和 wine 数据集的实验结果.



(a) glass 数据集实验结果

(b) iris 数据集实验结果

(c) wine 数据集实验结果

图 1 各算法在不同数据集下的 SVM 分类器指标值

从图 1 实验结果显示, 红色虚线代表的是本文所提出的 AP-KPCA 算法, 可以看出在本文所举例的 3 个数据集中, 进行 SVM 分类之后. 在准确率、精确率, 召回率和 $F1$-score 值这 4 个指标中, 都优越于其他对比的算法, 证明了 AP-KPCA 算法的有效性.

从 glass 数据集的实验结果显示, 本文所提出的 AP-KPCA 算法提取的主成分信息的 SVM 分类器的各指标值绝对地优于对比算法, 然后是 RobustPCA 算法, 接着是文献 [29] 所提出的改进 LDA 的降维算法和基于多项式核

的主成分分析算法, 线性主成分分析算法所提取的主成分信息的 SVM 分类器的性能最差.

从 iris 数据集的实验结果显示, 各个算法所提取的主成分信息的分类性能指标虽然纵横交错, 在准确率、精确率分类性能指标中, 其他几种算法都有高于本文所提出的算法, 但本文所提出的 AP-KPCA 算法提取的主成分信息的分类性能指标都处于一个较高水平的位置, 线性主成分分析提取的主成分信息的分类性能最差.

从 wine 数据集的实验结果显示, 本文提出的 AP-KPCA 算法提取的主成分信息的分类性能优于其他几种算法提取的主成分信息的分类性能. 文献 [29] 提出的特征提取算法和 RobustPCA 算法各个分类指标都趋于稳定, 分类性能紧随其后, 基于线性核的主成分分析算法和基于高斯核的主成分分析算法提取的主成分信息的分类性能居中, 基于多项式核的主成分分析算法提取的主成分信息的分类性能最差.

综上所述, 本文所提出的 AP-KPCA 算法能提取出更优质的主成分信息, 而加入的惩罚项对各向异性高斯核函数进行惩罚能更好地剔除冗余信息, 证明 AP-KPCA 算法要由于其他对比算法.

6.4.2 各算法准确率性能比较

为了验证本文所提出算法的整体性能, 比较各算法在表 2 数据集中 5 次实验所提取的主成分信息对应分类器的各指标平均结果的最优值. 表 4 分别展示了各个算法在 9 个数据集和 KDDCUP99 数据集提取的主成分信息对应SVM 分类器的最优平均分类准确率. 表 5 分别展示了各个算法在 9 个数据集和 KDDCUP99 数据集提取的主成分信息对应SVM 分类器的平均方差. 其中加粗的表示各算法提取的主成分信息的分类最优准确率和方差. 具体如表 4–表 6 所示.

表 4 各算法特征提取准确率均值比较

数据集	*n*	AP-KPCA	KPCA(ANGKS)	PCA	KPCA(linear)	KPCA(poly)	KPCA(rbf)	文献[29]	RobustPCA
abalone	2	0.5371	0.5339	0.5382	0.5359	0.5219	0.5433	0.5508	**0.5565**
column	3	**0.8537**	0.8408	0.7526	0.7806	0.7634	0.7913	0.8279	0.8344
glass	5	**0.7125**	0.6875	0.6093	0.6187	0.6593	0.6062	0.64375	0.6562
iris	2	**0.9733**	0.9688	0.92	0.9422	0.9466	0.96	0.9466	0.96
cmc	6	0.4825	0.4784	0.4394	0.4616	0.4421	0.4562	**0.4943**	0.4820
south	13	0.7740	**0.7813**	0.7706	0.7586	0.758	0.7613	0.7666	0.7533
segment	7	0.8349	0.8308	0.8268	0.8317	0.7197	0.8513	0.9454	**0.9584**
waveform	9	**0.8758**	0.8736	0.8722	0.87	0.8710	0.8633	0.8665	0.8656
wine	7	0.9849	**0.9924**	0.9371	0.9428	0.92	0.9485	0.9714	0.9584
KDDCUP99	13	**0.9933**	0.9800	0.9133	0.9466	0.9466	0.9466	0.9371	0.9666

表 5 各算法特征提取性能准确率方差比较

数据集	*n*	AP-KPCA	KPCA(ANGKS)	PCA	KPCA(linear)	KPCA(poly)	KPCA(rbf)	文献[29]	RobustPCA
abalone	2	0.000300317	0.000207961	0.000105 92	0.000305922	0.000186623	0.000096 81	0.000142483	**0.000081528**
column	3	**0.000265927**	0.001237137	0.000982 77	0.001653372	0.004220141	0.001306 50	0.001329633	0.002289282
glass	5	**0.000805664**	0.003051758	0.000976 56	0.004589844	0.001635742	0.001879 88	0.006396484	0.002807617
iris	2	**0.000145679**	0.000395062	0.000148 14	0.000395062	0.000395062	0.000345 67	0.000641975	0.000592593
cmc	6	0.000288974	0.000239098	0.000574 34	0.000114664	0.000593888	0.000296 68	**0.000210818**	0.000708552
south	13	0.000753333	0.000352222	0.001185 55	0.000603333	0.000642222	**0.000258 88**	0.000672222	0.000416667
segment	7	0.000189069	0.000182614	0.000107 23	0.000870382	0.000429361	0.000207 18	0.000145133	**0.000021979**
waveform	9	**0.000149778**	0.000736889	0.000368 88	0.000480000	0.000463556	0.000333 33	0.000220889	0.000448000
wine	7	0.000427198	**0.000106842**	0.000163 26	0.000163265	0.000979592	0.000571 42	0.000163265	0.000249199
KDDCUP99	13	0.000222222	0.000333333	0.000333 33	0.000333333	0.000333333	0.000333 33	0.000163265	**0**

从表 4 中可以看出, 本文所提出的基于各向异性高斯核的主成分分析算法在 7/10 的数据集上表现更好, 其中加入惩罚项的 AP-KPCA 算法比未加入惩罚项的效果更优. 其中 *n* 表示的是数据集进行各个算法之后所提取的特征维数, 代表原始信息的特征维度. south 数据集从原来的 20 维降到 13 维, waveform 数据集从原来的 21 维降到 9 维, wine 数据集从原来的 13 维降到 7 维, AP-KPCA 算法都在该数据集中提取了很好的主成分信息, 展示出更好

的分类性能, 说明本文所提出的算法更适合于更高维的数据集的特征提取, 并且能达到一个很好的效果. 结合表 5 可以看出, 本文所提出的基于各向异性高斯核的主成分分析算法在 4/5 的数据集上求出的准确率均值越大方差就越小, 证明本文所提出的模型的有效性.

各算法特征选择的平均准确率如图 2 所示.

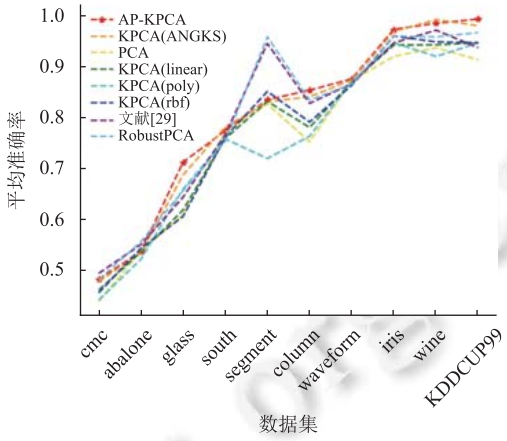


图 2 各算法特征提取性能比较

从图 2 中可以明显看出, 红色的线代表的本文所提出的 AP-KPCA 算法在不同数据集下的平均准确率, 大部分点红色的线都居于上位部分, 说明此方法的平均准确率在大部分的数据集下是要优于其他对比算法的.

综上所述, 本文提出的基于各向异性高斯核核惩罚的主成分分析的特征提取算法比传统的主成分分析算法在 9 种公开的 UCI 公开数据集上准确率平均提高了 4.49%. 在 KDDCUP99 数据集上, 本文提出的基于各向异性高斯核核惩罚的主成分分析的特征提取算法比传统的主成分分析算法准确率提高了 8%.

6.4.3 惩罚项影响及讨论

本小节主要验证了惩罚项对采用各向异性高斯核函数的 KPCA 算法的影响. 因此, 本节比较了加入惩罚项的 AP-KPCA 算法和未加入惩罚项的 KPCA 算法. 数据集采用表 2 中的数据, 评价方法仍采用 SVM 分类算法对使用 AP-KPCA 和 KPCA 算法降维后的数据进行分类, 再比较准确率 ACC、精确率 AUC、召回值 recall 和 $F1$-score 值, 通过 5 次实验的平均值来作为最终的结果.

图 2 是两种算法在 9 个数据集所选主成分对应 SVM 分类器的最优平均分类准确率 ACC、精确率 AUC、召回值 recall、$F1$-score 值的比较.

图 2 实验结果显示, 本文提出的 AP-KPCA 算法提取的特征子集的 SVM 分类器的各指标值在 9 个数据集下都优于未加入惩罚项的基于各向异性高斯核函数的主成分分析算法, 说明特征惩罚函数对惩罚核主成分分析算法的效果比较明显.

综合图 1 和图 2 的实验结果来看, 本文所提出的基于惩罚项各向异性高斯核的主成分分析算法能提取出类别区分能力很好的特征子集更优质的主成分信息, 优于其他对比算法.

6.4.4 核参数的影响及讨论

本文所提出的梯度下降法更新核宽向量的优点是能自动获得一个最优的特征子集, 根据特征子集中的核宽参数对应的准确率去进行核宽向量的更新, 并不断调整以达到一个最优的状态. 图 3 给出了 iris 数据集在不同核参数下所提取的主成分信息对应的 SVM 分类准确率, 改变核参数值为: $\{2^{-5}, 2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 2^{0}, 2^{1}, 2^{2}, 2^{3}, 2^{4}, 2^{5}\}$.

从图 4 中, 可以明显看出核参数在 2^{-2} 和 2^{2} 之间时, 基于各向异性高斯核核惩罚的主成分分析算法性能趋于稳定状态, 而对于更高或者更低的值, 其性能较差. 所以在选取核参数时, 选取 2^{-2} 和 2^{2} 之间的值能得到一个很好的效果.

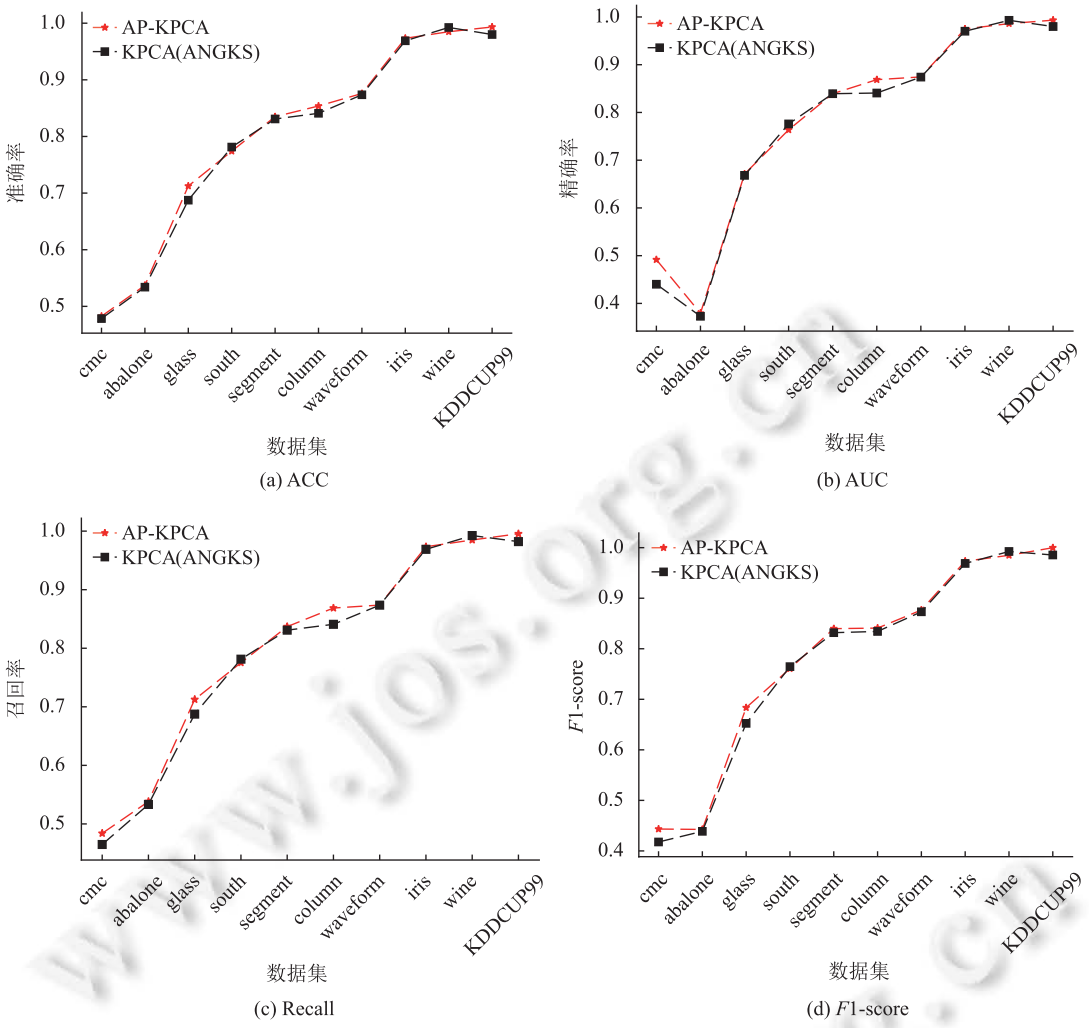


图 3　有无惩罚项的 KPCA(ANGKS) 算法在 9 种数据集下的 SVM 分类器指标值

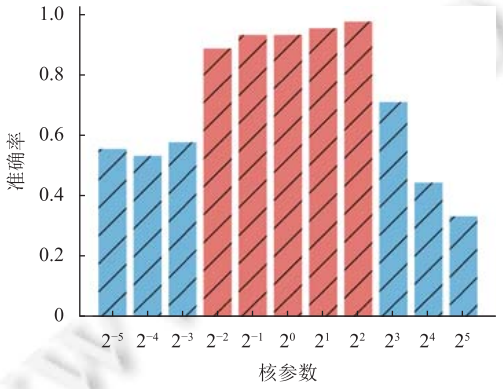


图 4　iris 数据集特征提取性能

6.4.5　总方差贡献率阈值影响及讨论

本文在进行思考数据降维到多少维度时, 选择了总方差贡献率大于 85% 的特征向量作为降维后的数据样本.

但是选取不同的阈值就会代表不同的原始数据的信息, 并且会影响后续的分类实验. 本文以 south 数据集为例, 讨论在阈值为 75%、80%、85%、90% 左右时进行 SVM 分类的准确率, 具体如表 6 所示.

从表 6 可以明显看出, 总方差贡献率越大, 降维之后的维数也越大, 平均准确率也相对应升高. 但是在选取较高的总方差贡献率之后, 维数也随之增大, 就脱离了本文所提出的初衷用较少的主成分去表征尽可能多的原始数据信息. 因此, 为了避免由于多变量带来的冗余性以及查阅相关文献, 特选取适中的总方差贡献率 85%.

表 6　south 数据集方差贡献率不同阈值性能分析

总方差贡献率 (%)	维数 (降维之后的)	平均准确率
75	7	0.7241
80	9	0.7531
85	11	0.7740
90	13	0.7804

7 总　结

本文提出了一种基于各向异性高斯核惩罚的主成分分析算法. 通过改变每个方向的控制参数, 从不同的方向反映数据特征的变换信息, 并在特征提取的过程中加入了特征惩罚函数, 同时引入梯度下降算法选择用更少的特征代表更多的原始特征信息, 为了验证所提方法的性能进行了一系列实验, 将本文提出的算法与其他几种常见的核函数以及文献 [29] 进行对比, 实验证明各向异性高斯核要优越于传统的高斯核函数及其他常见的几种核函数. 因此, 相信所提出的各向异性高斯核函数还可以应用在各个领域. 此外本文还提出了用均值法去改进传统的 PCA 算法, 实验证明了经过均值化处理数据, 可用更少的主成分去提取更多的原始信息, 提高了其方差贡献率.

所有实验表明, 所提出的各向异性高斯核函数的主成分分析算法是有效的, 但是仍有许多改进的余地. 进一步的工作可以选择不同的降维算法替代主成分分析算法, 比如线性判别分析、独立成分分析等.

References:

[1] Jeyanthi R, Jeyanthi R, Sireesha N, Srinivasan M, Devanathan S, Data reconciliation using MA-PCA and EWMA-PCA for large dimensional data. Journal of Intelligent & Fuzzy Systems, 2021, 41(5): 5731–5736. [doi: 10.3233/JIFS-189892]

[2] van Luong H, Deligiannis N, Seiler J, Forchhammer S, Kaup A. Compressive online robust principal component analysis via n-ℓ_1 minimization. IEEE Trans. on Image Processing, 2018, 27(9): 4314–4329. [doi: 10.1109/TIP.2018.2831915]

[3] Chu Z, Yu J, Hamdulla A. LPG-model: A novel model for throughput prediction in stream processing, using a light gradient boosting machine, incremental principal component analysis, and deep gated recurrent unit network. Information Sciences, 2020, 535: 107–129. [doi: 10.1016/j.ins.2020.05.042]

[4] Esmaeili M, Ahmadi M, Kazemi A. Kernel-based two-dimensional principal component analysis applied for parameterization in history matching. Journal of Petroleum Science and Engineering, 2020, 191: 107134. [doi: 10.1016/j.petrol.2020.107134]

[5] Reddy GT, Reddy MPK, Lakshmanna K, Kaluri R, Rajput DS, Srivastava G, Baker T. Analysis of dimensionality reduction techniques on big data. IEEE Access, 2020, 8: 54776–54788. [doi: 10.1109/ACCESS.2020.2980942]

[6] Liu J, Tang SL, Xu GX, Ma C, Lin MW. A novel configuration tuning method based on feature selection for hadoop MapReduce. IEEE Access, 2020, 8: 63862–63871. [doi: 10.1109/ACCESS.2020.2984778]

[7] Li O, Shui PL. Subpixel blob localization and shape estimation by gradient search in parameter space of anisotropic Gaussian kernels. Signal Processing, 2020, 171: 107495. [doi: 10.1016/j.sigpro.2020.107495]

[8] Zhao Z, Li B, Kang XQ, Chen L, Wei X, Xin MT. Hybrid image segmentation method based on anisotropic Gaussian kernels and adjacent graph region merging. Review of Scientific Instruments, 2020, 91(1): 015104. [doi: 10.1063/1.5095557]

[9] Maldonado S, Carrizosa E, Weber R. Kernel penalized K-means: A feature selection method based on kernel K-means. Information Sciences, 2015, 322: 150–160. [doi: 10.1016/j.ins.2015.06.008]

[10] Kouadri A, Hajji M, Harkat MF, Abodayeh K, Mansouri M, Nounou H, Nounou M. Hidden Markov model based principal component analysis for intelligent fault diagnosis of wind energy converter systems. Renewable Energy, 2020, 150: 598–606. [doi: 10.1016/j.renene.2020.01.010]

[11] Fernández-Martínez JL, Fernández-Muñiz Z. The curse of dimensionality in inverse problems. Journal of Computational and Applied Mathematics, 2020, 369: 112571. [doi: 10.1016/j.cam.2019.112571]

[12] Song Y, Sun WY, Chen CS. Logarithm transformation based principal component analysis for image recognition. Journal of Xi'an Jiaotong University, 2021, 55(1): 33–42 (in Chinese with English abstract). [doi: 10.7652/xjtuxb202101005]

[13] Tucker JD, Lewis JR, Srivastava A. Elastic functional principal component regression. Statistical Analysis and Data Mining, 2019, 12(2): 101–115. [doi: 10.1002/sam.11399]

[14] Gu TC. Detection of small floating targets on the sea surface based on multi-features and principal component analysis. IEEE Geoscience and Remote Sensing Letters, 2020, 17(5): 809–813. [doi: 10.1109/LGRS.2019.2935262]

[15] Bhandary A, Prabhu GA, Rajinikanth V, Thanaraj KP, Satapathy SC, Robbins DE, Shasky C, Zhang YD, Tavares JMRS, Raja NSM. Deep-learning framework to detect lung abnormality—A study with chest X-ray and lung CT scan images. Pattern Recognition Letters, 2020, 129: 271–278. [doi: 10.1016/j.patrec.2019.11.013]

[16] Uddin P, Mamun A, Afjal MI, Hossain A. Information-theoretic feature selection with segmentation-based folded principal component analysis (PCA) for hyperspectral image classification. Int'l Journal of Remote Sensing, 2021, 42(1): 286–321. [doi: 10.1080/01431161.2020.1807650]

[17] Griffin B, Hartarska V, Nadolnyak D. Credit constraints and beginning farmers' production in the US: Evidence from propensity score matching with principal component clustering. Sustainability, 2020, 12(14): 5537. [doi: 10.3390/su12145537]

[18] Ren L, Sun DS. Application of principal component based RBF kernel SVM in financial forecasting. Jiangsu Commercial Forum, 2019, (1): 102–104 (in Chinese with English abstract). [doi: 10.3969/j.issn.1009-0061.2019.01.027]

[19] Delchambre L. Weighted principal component analysis: A weighted covariance eigendecomposition approach. Monthly Notices of the Royal Astronomical Society, 2015, 446(4): 3545–3555. [doi: 10.1093/mnras/stu2219]

[20] Schölkopf B, Smola A, Müller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 1998, 10(5): 1299–1319. [doi: 10.1162/089976698300017467]

[21] John H, Christopher R, Mahmood R, Adaptive classification using incremental linearized kernel embedding. IEEE Trans. on Signal Processing, 2022, 70: 1764–1774. [doi: 10.1109/TSP.2022.3162407]

[22] Lee K, Lee CH, Kwak MS, Jang EJ. Analysis of multivariate longitudinal data using ARMA Cholesky and hypersphere decompositions. Computational Statistics & Data Analysis, 2021, 156: 107144. [doi: 10.1016/j.csda.2020.107144]

[23] Maldonado S, Weber R, Basak J. Simultaneous feature selection and classification using kernel-penalized support vector machines. Information Sciences, 2011, 181(1): 115–128. [doi: 10.1016/j.ins.2010.08.047]

[24] Liu J. Research on key technologies of performance tuning of jobs in distributed data processing system [Ph.D. Thesis]. Chongqing: Chongqing University, 2016 (in Chinese with English abstract).

[25] Asuncion A, Newman DJ. UCI machine learning repository. 2007. https://archive.ics.uci.edu/ml/index.php

[26] Amarnath B, Balamurugan SAA. Review on feature selection techniques and its impact for effective data classification using UCI machine learning repository dataset. Journal of Engineering Science and Technology, 2016, 11(11): 1639–1646.

[27] Zhang QS. Research of multiple kernel one-class support vector machine [MS. Thesis]. Beijing: Beijing University of Civil Engineering and Architecture, 2020 (in Chinese with English abstract). [doi: 10.26943/d.cnki.gbjzc.2020.000068]

[28] Xie JY, Ding LJ, Wang MZ. Spectral clustering based unsupervised feature selection algorithms. Ruan Jian Xue Bao/Journal of Software, 2020, 31(4): 1009–1024 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5927.htm [doi: 10.13328/j.cnki.jos.005927]

[29] Chen YR, Tao X, Xiong CC, Yang JC. An improved method of two stage linear discriminant analysis. KSII Trans. on Internet and Information Systems, 2018, 12(3): 1243–1263. [doi: 10.3837/tiis.2018.03.015]

[30] Wang Y, Yu WK, Fang ZC. Multiple kernel-based SVM classification of hyperspectral images by combining spectral, spatial, and semantic information. Remote Sensing, 2020, 12(1): 120. [doi: 10.3390/rs12010120]

[31] Si W, Qiao YL, Liu Z, Jin GW, Liu YF, Xue XY, Zhou H, Liu YM, Shen AJ, Liang XM. Combination of multi-model statistical analysis and quantitative fingerprinting in quality evaluation of Shuang-huang-lian oral liquid. Analytical and Bioanalytical Chemistry, 2020, 412(29): 8223. [doi: 10.1007/s00216-020-02937-6]

[32] Zhou T, Peng YB. Kernel principal component analysis-based Gaussian process regression modelling for high-dimensional reliability analysis. Computers & Structures, 2020, 241: 106358. [doi: 10.1016/j.compstruc.2020.106358]

[33] İkizoğlu S, Heydarov S. Accuracy comparison of dimensionality reduction techniques to determine significant features from IMU sensor-based data to diagnose vestibular system disorders. Biomedical Signal Processing and Control, 2020, 61: 101963. [doi: 10.1016/j.bspc.2020.101963]

附中文参考文献:

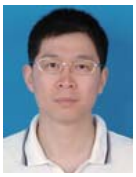
[12] 宋昱, 孙文赟, 陈昌盛. 对数变换主成分分析的图像识别. 西安交通大学学报, 2021, 55(1): 33–42. [doi: 10.7652/xjtuxb202101005]

[18] 任靓, 孙德山. 基于主成分的RBF核SVM在财务预测领域的应用. 江苏商论, 2019, (1): 102–104. [doi: 10.3969/j.issn.1009-0061.2019.01.027]

[24] 刘俊. 分布式数据处理系统中作业性能优化关键技术研究 [博士学位论文]. 重庆: 重庆大学, 2016.

[27] 张庆朔. 多核一类支持向量机方法研究 [硕士学位论文]. 北京: 北京建筑大学, 2020. [doi: 10.26943/d.cnki.gbjzc.2020.000068]

[28] 谢娟英, 丁丽娟, 王明钊. 基于谱聚类的无监督特征选择算法. 软件学报, 2020, 31(4): 1009–1024. http://www.jos.org.cn/1000-9825/5927.htm [doi: 10.13328/j.cnki.jos.005927]



刘俊(1978－), 男, 博士, 讲师, CCF 专业会员, 主要研究领域为信息安全, 人工智能, 高性能存储优化, 大规模机器学习, 大数据分析与处理, 区块链.

陈蜀宇(1963－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为云计算, 可信计算, 行业大数据.

李威(1997－), 男, 硕士生, 主要研究领域为机器学习, Spark 平台性能调优, 大数据分析与处理.

徐光侠(1974－), 女, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为区块链, 大数据安全与智能分析, 网络安全与管控.