计算机科学
COMPUTER SCIENCE

基于最大相关熵的 KPCA 异常检测方法

李其烨，邢红杰

引用本文

李其烨，邢红杰. 基于最大相关熵的 KPCA 异常检测方法[J]. 计算机科学, 2022, 49(8): 267-272.

LI Qi-ye, XING Hong-jie. KPCA Based Novelty Detection Method Using Maximum Correntropy Criterion[J]. Computer Science, 2022, 49(8): 267-272.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210700175

基于最大相关熵的 KPCA 异常检测方法

李其烨　邢红杰
河北大学数学与信息科学学院河北省机器学习与计算智能重点实验室　河北 保定 071002
(137170651@qq.com)

摘　要　异常检测是机器学习中一个重要的研究内容,目前已存在大量的异常检测方法。作为一种常用的核方法,核主成分分析(Kernel Principal Component Analysis,KPCA)已被成功地用于解决异常检测问题。然而,传统的 KPCA 异常检测方法对噪声非常敏感,若训练样本中存在噪声,则会降低 KPCA 异常检测方法的检测性能。为了提高 KPCA 异常检测方法的抗噪声能力,提出了一种基于最大相关熵(Maximum Correntropy Criterion,MCC)的 KPCA 异常检测方法。利用信息理论学习中的相关熵代替 KPCA 异常检测方法中基于 ℓ_2 范数的度量,通过调节相关熵函数中的宽度参数,可以有效抑制噪声带来的不利影响;利用半二次优化技术对所提方法的优化问题进行求解,仅需较少的迭代次数即可取得局部最优解。此外,给出了所提方法的算法描述,并分析了算法的计算复杂度。在 16 个 UCI 基准数据集上的实验结果表明,与其他 4 种相关方法相比,所提方法取得了更优的抗噪声能力和泛化性能。

关键词:核主成分分析;相关熵;半二次优化;异常检测;信息理论学习
中图法分类号　TP391.4

KPCA Based Novelty Detection Method Using Maximum Correntropy Criterion

LI Qi-ye and XING Hong-jie
Hebei Key Laboratory of Machine Learning and Computational Intelligence,College of Mathematics and Information Science,Hebei University,Baoding,Hebei 071002,China

Abstract　Novelty detection is an important research issue in the field of machine learning. Till now,there exist lots of novelty detection approaches. As a commonly used kernel method,kernel principal component analysis(KPCA)has been successfully applied to deal with the problem of novelty detection. However,the traditional KPCA based novelty detection method is very sensitive to noise. If there exist noise in the given training samples,the detection performance of KPCA based novelty detection method may be decreased. To enhance the anti-noise ability of KPCA based novelty detection method,a maximum correntropy criterion(MCC)based novelty detection method is proposed. Correntropy in information theoretic learning is utilized to substitute the ℓ_2-norm based measure in KPCA based novelty detection method. By adjusting the width parameter of the correntropy function,the adverse effect of noise can be alleviated. The half-quadratic optimization technique is used to solve the optimization problem of the proposed method. The local optimal solution can thus be obtained after a few iterations. Moreover,the algorithmic description of the proposed method is provided,and the computational complexity of the corresponding algorithm is analyzed. Experimental results on the 16 UCI benchmark data sets demonstrate that the proposed method obtains better anti-noise and generalization performance in comparison with the other four related approaches.

Keywords　Kernel principal component analysis,Correntropy,Half-quadratic optimization,Novelty detection,Information theoretic learning

1　引言

异常检测(Novelty Detection)是机器学习领域的常见任务,其实际应用价值在于当异常情况发生时,该异常能够被及时发现并解决,以减小损失。异常检测常常被看作机器学习中的单类分类问题[1],而单类分类器能够有效地处理类别极端不平衡问题。异常检测模型在训练阶段仅由正常数据构成训练集,在测试阶段待测样本则被分类为正常数据或异常数据。在实际应用中,用于解决异常检测问题的正常样本采样充分,而异常样本采样困难,如临床诊断[2]、桥面诊断[3]等。

到稿日期:2021-07-18　返修日期:2022-02-28
基金项目:国家自然科学基金(61672205);河北省自然科学基金(F2017201020);河北大学高层次人才科研启动项目(521100222002)
This work was supported by the National Natural Science Foundation of China(61672205), Natural Science Foundation of Hebei Province(F2017201020) and High-Level Talents Research Start-Up Project of Hebei University(521100222002).
通信作者:邢红杰(hjxing@hbu.edu.cn)

迄今为止，存在大量用于解决异常检测问题的单类分类方法，其中最常用的两种方法是单类支持向量机(One-class Support Vector Machine，OCSVM)[4]和支持向量数据描述(Support Vector Data Description，SVDD)[5]。在一定条件下，可以证明 OCSVM 等价于 SVDD[4,5]。此外，核主成分分析[6]也是一种常用的核方法，它将核计巧应用于主成分分析(Principal Component Analysis，PCA)[7]，首先将样本映射到高维的特征空间，然后在特征空间中使用 PCA 求取投影向量。因为具有更强的特征信息保留能力，KPCA 已被广泛地应用于去噪[8]、数据分析和压缩[9]、图像分析[10]等。

虽然 KPCA 具有获取非线性数据特征的能力，但若训练样本中存在噪声，则其性能会受到严重影响。为了提高 KPCA 的抗噪声能力，相关学者提出了以下 4 种解决方案：使用更为鲁棒的估计器计算 Gram 矩阵[11]，利用更为鲁棒的损失代替目标函数中基于 ℓ_2 范数的损失[12]，剔除训练样本中的噪声[13]，将鲁棒 PCA 用于 KPCA[14]。

为了使 KPCA 能够用于解决异常检测问题，Hoffmann[15]首次在高维特征空间中计算基于重构误差的异常度量，从而提出了一种基于 KPCA 的异常检测方法，实验结果表明它能够取得优于 OCSVM 和 Parzen 窗密度估计器[16]的检测性能，此外，该方法所构造的决策边界更为紧致，因此具有更强的泛化性能。如前文所述，由于 KPCA 的目标函数采用了基于 ℓ_2 范数的度量，若训练样本中存在噪声，则 KPCA 取得的结果将会较差。为了提高基于 KPCA 的异常检测方法的抗噪声能力，Xiao 等[17]利用 ℓ_1 范数代替传统 KPCA 中的 ℓ_2 范数，提出了基于 ℓ_1-KPCA 的异常检测方法。为了求取正交载荷向量(Loading Vector)，他们提出了包含顶层算法和底层算法的贪婪算法。实验结果表明，ℓ_1-KPCA 比传统 KPCA 具有更为鲁棒的抗噪声能力。Alzate 等[18]利用 Epsilon 非灵敏损失函数代替 KPCA 中基于 ℓ_2 范数的损失函数，提出了两种不同的算法，第一种算法以 KPCA 为出发点，对非线性方程组进行求解；第二种算法利用迭代加权过程求解了一系列广义特征值问题。然而，上述两个求解过程均非常复杂，而且每次仅能取得一个主成分。受鲁棒 PCA 的启发，Wang 等[19]对 KPCA 的目标函数进行了改进，并在目标函数中增加了基于 $\ell_{2,1}$ 范数的正则化项，与相关方法相比，所提方法具有更优的检测性能。

近年来，基于信息理论学习(Information Theoretic Learning，ITL)的机器学习方法逐渐引起广泛关注[20]。在 ITL 中，相关熵(Correntropy)是一种局部相似性度量[21]，当两个样本之间的距离较近时，它被看作是基于 ℓ_2 范数的度量，当两个样本之间的距离较远时，它被看作是基于 ℓ_1 范数的度量，当两个样本之间的距离非常远时，它被认为是基于 ℓ_0 范数的度量。因为它对非高斯噪声尤为鲁棒，所以 He 等[22]提出了基于最大相关熵的主成分分析，并利用半二次优化技术[23]对相应的优化问题进行求解。实验结果表明，所提方法具有比 PCA-ℓ_1[24]更优的抗噪声能力。

受基于 MCC 的 PCA 方法的启发，为了提高基于 KPCA 异常检测方法的鲁棒性，本文提出了基于 MCC 的 KPCA 异常检测方法。本文的贡献包括：

(1)利用相关熵代替 KPCA 目标函数中基于 ℓ_2 范数的度量，使得所提异常检测方法具有更为鲁棒的抗噪声能力；

(2)利用半二次优化技术对所提异常检测方法的优化问题进行迭代求解，使得在较少的迭代次数下即可求得局部最优解；

(3)描述了所提方法的算法实现，给出了整个算法的计算复杂度，并分析了算法的收敛性。

本文第 2 节简要回顾了相关熵、KPCA 及其在异常检测上的应用；第 3 节从数学模型和算法描述两方面详细阐述了所提的基于 MCC 的 KPCA 异常检测方法；第 4 节给出了所提方法与其相关方法在基准数据集上的对比实验结果，验证了所提方法的有效性；最后总结全文并展望未来。

2 相关知识

2.1 最大相关熵准则

在信息理论学习[20]中，相关熵被看作是一种广义相关函数[21]，它能够有效处理非高斯噪声。此外，相关熵通常被用作局部相似性度量。对于两个随机变量 A 和 B，它们之间的相关熵定义为：

$$V_\sigma(A,B)=E[k_\sigma(A-B)] \tag{1}$$

其中，$E[\cdot]$表示数学期望，$k_\sigma(\cdot)$是满足 Mercer 定理的核函数[25]，且 σ 为核函数的宽度参数。式(1)中的相关熵可以通过有限个样本$\{a_i,b_i\}_{i=1}^N$估计得到，即：

$$\hat{V}_\sigma(A,B)=\frac{1}{N}\sum_{i=1}^{N}k_\sigma(a_i-b_i) \tag{2}$$

通常情况下，核函数 $k_\sigma(\cdot)$取为高斯核函数，则式(2)可以表示为：

$$\hat{V}_\sigma(A,B)=\frac{1}{N}\sum_{i=1}^{N}G(a_i-b_i)=\frac{1}{N}\sum_{i=1}^{N}\exp\left(-\frac{(a_i-b_i)^2}{2\sigma^2}\right) \tag{3}$$

式(1)的最大化被称为最大相关熵准则[22]。与全局相似性度量准则(如均方误差)不同，MCC 关注的是局部相似性。由于相关熵对噪声和野点不敏感，当训练样本中存在脉冲噪声时，MCC 的性能会优于 MSE[20]。

2.2 KPCA

KPCA[6]首先通过非线性映射 $\phi:\Re^d\rightarrow F$ 将输入空间中的样本集 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n]\in\Re^{d\times n}$映射到高维特征空间 F 中，样本集在特征空间中的像为 $\boldsymbol{\Phi}=[\phi(\boldsymbol{x}_1),\phi(\boldsymbol{x}_2),\cdots,\phi(\boldsymbol{x}_n)]$。然后，对样本集的像去中心化。

$$\tilde{\phi}(\boldsymbol{x}_i)=\phi(\boldsymbol{x}_i)-\boldsymbol{\mu}=\phi(\boldsymbol{x}_i)-\frac{1}{n}\sum_{j=1}^{n}\phi(\boldsymbol{x}_j) \tag{4}$$

在特征空间中对去中心化后的像集 $\tilde{\boldsymbol{\Phi}}=[\tilde{\phi}(\boldsymbol{x}_1),\tilde{\phi}(\boldsymbol{x}_2),\cdots,\tilde{\phi}(\boldsymbol{x}_n)]$进行主成分分析，求得特征空间中的投影矩阵 $\boldsymbol{V}$，即求解式(5)所示的优化问题：

$$\begin{aligned}&\min\sum_{i=1}^{n}\|\phi(\boldsymbol{x}_i)-(\boldsymbol{\mu}+\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\tilde{\phi}(\boldsymbol{x}_i))\|^2\\&\text{s.t. }\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{I}\end{aligned} \tag{5}$$

其中，$\boldsymbol{I}$ 为单位矩阵。由文献[26]中的数学推导可知，上述优化问题等同于：

$$\begin{aligned}&\min_{\boldsymbol{V}}-Tr(\boldsymbol{V}^{\mathrm{T}}\tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\Phi}}^{\mathrm{T}}\boldsymbol{V})\\&\text{s.t. }\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{I}\end{aligned} \tag{6}$$

考虑到 $\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}^{\mathrm{T}}$ 是协方差矩阵，为解决上述优化问题，对式(6)使用拉格朗日乘子法可得：

$$\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}^{\mathrm{T}}\boldsymbol{v}=\lambda\boldsymbol{v} \tag{7}$$

可以看出，只需对协方差矩阵 $\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}^{\mathrm{T}}$ 进行特征值求解，得到的特征向量所构成的投影矩阵即为 KPCA 的解。然而，式(7)中的 $\boldsymbol{v}$ 不能直接求得，原因是映射函数 ϕ 的表达形式往往是未知的，故无法得到协方差矩阵 $\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}^{\mathrm{T}}$ 的具体值。根据文献[15]，协方差矩阵 $\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}^{\mathrm{T}}$ 的特征向量 $\boldsymbol{v}$ 可以表示为样本点的像 $\{\phi(\boldsymbol{x}_j)\}_{j=1}^{n}$ 的线性组合，即：

$$\boldsymbol{v}=\widetilde{\boldsymbol{\Phi}}\boldsymbol{\alpha} \tag{8}$$

其中，$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_n)^{\mathrm{T}}$ 为线性系数向量，可以由核矩阵 $\widetilde{\boldsymbol{K}}=\widetilde{\boldsymbol{\Phi}}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}$ 的特征值分解求得，$\widetilde{\boldsymbol{K}}$ 中的元素 $\widetilde{k}_{ij}=\widetilde{k}(\boldsymbol{x}_i,\boldsymbol{x}_j)=\widetilde{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_j)$ 由式(9)求得。

$$\widetilde{k}_{ij}=k_{ij}-\frac{1}{n}\sum_{r=1}^{n}k_{ir}-\frac{1}{n}\sum_{r=1}^{n}k_{rj}+\frac{1}{n^2}\sum_{r,s=1}^{n}k_{rs} \tag{9}$$

假设求得前 q 个特征值对应的特征向量为 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_q$，则由 q 个特征向量所组成的投影矩阵可表示为 $\boldsymbol{V}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_q]$。

将 KPCA 用作异常检测模型时，需要计算原数据点与投影后重构的数据点之间的重构误差，将其作为异常度量来进行异常检测。令 $f_l(\boldsymbol{z})$ 表示数据 $\boldsymbol{z}$ 在第 l 个投影向量上的投影。

$$\begin{aligned}f_l(\boldsymbol{z})&=(\widetilde{\phi}(\boldsymbol{z})\cdot\boldsymbol{v}_l)\\&=\{[\phi(\boldsymbol{z})-\frac{1}{n}\sum_{r=1}^{n}\phi(\boldsymbol{x}_r)]\cdot[\sum_{j=1}^{n}\alpha_j^l\phi(\boldsymbol{x}_j)-\\&\quad\frac{1}{n}\sum_{j=1}^{n}\sum_{r=1}^{n}\alpha_j^l\phi(\boldsymbol{x}_r)]\}\\&=\sum_{j=1}^{n}\alpha_j^l[k(\boldsymbol{z},\boldsymbol{x}_j)-\frac{1}{n}\sum_{r=1}^{n}k(\boldsymbol{x}_j,\boldsymbol{x}_r)-\frac{1}{n}\sum_{r=1}^{n}k(\boldsymbol{z},\boldsymbol{x}_r)+\\&\quad\frac{1}{n^2}\sum_{r,s=1}^{n}k(\boldsymbol{x}_r,\boldsymbol{x}_s)]\end{aligned} \tag{10}$$

由文献[15]可知 $(\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{z})\cdot\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{z}))=\sum_{l=1}^{q}f_l^2(\boldsymbol{z})$，数据点 $\boldsymbol{z}$ 的重构误差值可表示为：

$$\begin{aligned}RE(\boldsymbol{z})&=(\widetilde{\phi}(\boldsymbol{z})\cdot\widetilde{\varphi}(\boldsymbol{z}))-(\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{z})\cdot\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{z}))\\&=\widetilde{k}(\boldsymbol{z},\boldsymbol{z})-\sum_{l=1}^{q}f_l^2(\boldsymbol{z})\end{aligned} \tag{11}$$

3 基于最大相关熵的 KPCA 异常检测方法

本节将引入基于最大相关熵的 KPCA 异常检测方法，首先介绍其数学模型和相应的求解方法，然后描述其算法的实现过程。

3.1 数学模型

利用相关熵替换式(5)中基于 ℓ_2 范数的度量 $\|\cdot\|^2$，可得如下的优化问题：

$$\begin{aligned}\max J(\boldsymbol{\mu},\boldsymbol{V})&=\sum_{i=1}^{n}g(\phi(\boldsymbol{x}_i)-\boldsymbol{\mu}-\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_i))\\&=\sum_{i=1}^{n}\exp\left(-\frac{\|\phi(\boldsymbol{x}_i)-\boldsymbol{\mu}-\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_i)\|^2}{2\sigma^2}\right)\end{aligned} \tag{12}$$

由 $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{I}$，可得：

$$\begin{aligned}&\|\phi(\boldsymbol{x}_i)-\boldsymbol{\mu}-\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_i)\|\\&=\|\widetilde{\phi}(\boldsymbol{x}_i)-\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_i)\|\\&=\sqrt{\widetilde{k}(\boldsymbol{x}_i,\boldsymbol{x}_i)-\widetilde{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_i)}\end{aligned} \tag{13}$$

则优化问题(12)可以改写为：

$$\max J(\boldsymbol{\mu},\boldsymbol{V})=\sum_{i=1}^{n}\exp\left(-\frac{\widetilde{k}(\boldsymbol{x}_i,\boldsymbol{x}_i)-\widetilde{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_i)}{2\sigma^2}\right) \tag{14}$$

可以利用半二次优化技术[23]对优化问题(14)进行求解。根据凸共轭函数理论[27]，可知下述定理成立。

命题 1 对于 $g(x)=\exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)$，存在凸共轭函数 φ，使得：

$$g(x)=\max_{p'\in\Re^-}\left(p'\frac{\|x\|^2}{2\sigma^2}-\varphi(p')\right) \tag{15}$$

对于固定的 x，当 $p'=-g(x)$ 时，式(15)取得最大值。

根据命题 1，式(14)可改写为：

$$\begin{aligned}&\max J(\boldsymbol{\mu},\boldsymbol{V},p)=\\&\sum_{i=1}^{n}(p_i\left(\frac{\widetilde{k}(\boldsymbol{x}_i,\boldsymbol{x}_i)-\widetilde{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_i)}{2\sigma^2}\right)-\varphi(p_i))\end{aligned} \tag{16}$$

其中，向量 $\boldsymbol{p}=(p_1,p_2,\cdots,p_n)^{\mathrm{T}}$ 用于存储辅助变量。根据命题 1，优化问题(16)的最优解可以迭代求取，则有：

$$p_i^{t+1}=-\exp\left(-\frac{\widetilde{k}(\boldsymbol{x}_i,\boldsymbol{x}_i)-\widetilde{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{V}^t(\boldsymbol{V}^t)^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_i)}{2\sigma^2}\right) \tag{17}$$

$$\boldsymbol{\mu}^{t+1}=\frac{1}{(\sum_{i=1}^{n}p_i^{t+1})}\sum_{i=1}^{n}p_i^{t+1}\phi(\boldsymbol{x}_i) \tag{18}$$

和

$$\begin{aligned}&\boldsymbol{V}^{t+1}=\arg\max_{\boldsymbol{V}}Tr(\boldsymbol{V}^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}^{t+1}\boldsymbol{P}^{t+1}(\widetilde{\boldsymbol{\Phi}}^{t+1})^{\mathrm{T}}\boldsymbol{V})\\&\text{s.t. }\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}=\boldsymbol{I}\end{aligned} \tag{19}$$

然而，式(18)中映射函数 $\phi(\cdot)$ 的表达形式未知，因此对中心向量 $\boldsymbol{\mu}$ 的更新可以转化为对核矩阵 $\widetilde{\boldsymbol{K}}$ 的更新，即：

$$\begin{aligned}\widetilde{k}_{rs}^{t+1}&=(\phi(\boldsymbol{x}_r)-\boldsymbol{\mu}^{t+1})\cdot(\phi(\boldsymbol{x}_s)-\boldsymbol{\mu}^{t+1})\\&=k_{rs}-\frac{\sum_{i=1}^{n}p_i^{t+1}k_{ri}}{\sum_{i=1}^{n}p_i^{t+1}}-\frac{\sum_{i=1}^{n}p_i^{t+1}k_{is}}{\sum_{i=1}^{n}p_i^{t+1}}+\frac{\sum_{i,j=1}^{n}p_i^{t+1}p_j^{t+1}k_{ij}}{\sum_{i=1}^{n}p_i^{t+1}p_j^{t+1}}\end{aligned} \tag{20}$$

将式(8)代入式(19)，可得：

$$\begin{aligned}\boldsymbol{A}^{t+1}&=\arg\max_{\boldsymbol{A}}Tr(\boldsymbol{A}^{\mathrm{T}}(\widetilde{\boldsymbol{\Phi}}^{t+1})^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}^{t+1}P^{t+1}(\widetilde{\boldsymbol{\Phi}}^{t+1})^{\mathrm{T}}\widetilde{\boldsymbol{\Phi}}^{t+1}\boldsymbol{A})\\&=\arg\max_{\boldsymbol{A}}Tr(\boldsymbol{A}^{\mathrm{T}}\widetilde{\boldsymbol{K}}^{t+1}\boldsymbol{P}^{t+1}\widetilde{\boldsymbol{K}}^{t+1}\boldsymbol{A})\end{aligned} \tag{21}$$

其中，$\boldsymbol{A}=[\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_q]$，$\boldsymbol{P}^{t+1}$ 矩阵为对角矩阵，其主对角线上的元素为 $\boldsymbol{P}^{t+1}(i,i)=-p_i^{t+1}$。优化问题(21)的最优解 $\boldsymbol{A}^*$ 通过求解下面的特征值分解问题求得。

$$\boldsymbol{P}^{t+1}\widetilde{\boldsymbol{K}}^{t+1}\boldsymbol{A}=\boldsymbol{\Lambda}\boldsymbol{A} \tag{22}$$

其中，$\boldsymbol{\Lambda}$ 是对角矩阵，它的主对角线上的元素为 $\boldsymbol{P}^{t+1}\widetilde{\boldsymbol{K}}^{t+1}$ 的前 q 个最大特征值，这些特征值对应的特征向量组成最优解 $\boldsymbol{A}^*$。

3.2 学习算法

基于最大相关熵的 KPCA 异常检测方法的训练过程如

算法 1 所示。需要指出，算法 1 第 2 步中的 E 表示最大相关熵的值，即：

$$E=\sum_{i=1}^{n}\exp\left(-\frac{\widetilde{k}(\boldsymbol{x}_i,\boldsymbol{x}_i)-\widetilde{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}VV^{\mathrm{T}}\widetilde{\phi}(\boldsymbol{x}_i)}{2\sigma^2}\right)$$

$$=\sum_{i=1}^{n}\exp\left(-\frac{\widetilde{k}(\boldsymbol{x}_i,\boldsymbol{x}_i)-\sum_{l=1}^{q}f_l^2(\boldsymbol{x}_i)}{2\sigma^2}\right) \tag{23}$$

算法 1 中第 3 步更新辅助向量 $\boldsymbol{p}$ 的计算代价为 $O(n)$，其中 n 为数据集中的样本个数；第 4 步更新核矩阵 $\widetilde{\boldsymbol{K}}$ 的计算复杂度为 $O(n^2+2n)$；第 5 步更新特征矩阵 $\boldsymbol{A}$ 的计算复杂度为 $O(n^3)$。因此，整个算法过程的计算复杂度为 $O(m(n^3+n^2+3n))$，其中 m 表示迭代次数。

算法 1 基于最大相关熵的 KPCA 异常检测

输入：数据集 $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n]\in\Re^{d\times n}$，特征向量个数 q，停止阈值 ε

输出：投影矩阵 $\mathbf{A}^*=[\boldsymbol{\alpha}_1^*,\boldsymbol{\alpha}_2^*,\cdots,\boldsymbol{\alpha}_q^*]$

初始化：使用传统的 KPCA 求取 q 个特征向量，组成正交矩阵 $\mathbf{A}$

过程：

1. t=0；
2. while E<ε do
3. 利用式(17)更新辅助向量 $\mathbf{p}^{t+1}=(p_1^{t+1},p_2^{t+1},\cdots,p_n^{t+1})^{\mathrm{T}}$；
4. 利用式(20)更新核矩阵 $\widetilde{\mathrm{K}}^{t+1}$；
5. 利用式(21)更新特征矩阵 A^{t+1}；
6. t=t+1；
7. end while
8. $\mathbf{A}^*=\mathbf{A}^t$.

算法 1 的收敛性如命题 2 所述。

命题 2 算法 1 产生的序列 $J(\boldsymbol{\mu}^t,\boldsymbol{V}^t,\boldsymbol{p}^t)(t=1,2,\cdots)$ 收敛。

证明：由命题 1 和式(19)可知 $J(\boldsymbol{\mu}^t,\boldsymbol{V}^t,\boldsymbol{p}^t)\leqslant J(\boldsymbol{\mu}^t,\boldsymbol{V}^t,\boldsymbol{p}^{t+1})\leqslant J(\boldsymbol{\mu}^{t+1},\boldsymbol{V}^{t+1},\boldsymbol{p}^{t+1})$。因此 $J(\boldsymbol{\mu}^t,\boldsymbol{V}^t,\boldsymbol{p}^t)$ 为非递减函数。此外，由最大相关熵的有界性[28]可知 $J(\boldsymbol{\mu},\boldsymbol{V})$ 有界，而 $\max J(\boldsymbol{\mu},\boldsymbol{V})=\max J(\boldsymbol{\mu},\boldsymbol{V},\boldsymbol{p})$，则 $J(\boldsymbol{\mu},\boldsymbol{V},\boldsymbol{p})$ 亦有界。因此，$J(\boldsymbol{\mu}^t,\boldsymbol{V}^t,\boldsymbol{p}^t)(t=1,2,\cdots)$ 收敛。

最后，对于待测样本 $\boldsymbol{z}$，相应的判别式为：

$$ps(\boldsymbol{z})=-(\widetilde{k}(\boldsymbol{z},\boldsymbol{z})-\sum_{j=1}^{n}\widetilde{k}(\boldsymbol{z},\boldsymbol{x}_j)\boldsymbol{\alpha}_j\boldsymbol{\alpha}_j^{\mathrm{T}}\widetilde{k}(\boldsymbol{x}_j,\boldsymbol{z})) \tag{24}$$

设判别阈值为 ξ，如果 $ps(\boldsymbol{z})<\xi$，则将 $\boldsymbol{z}$ 判定为异常数据，否则将 $\boldsymbol{z}$ 判定为正常数据。

4 实验结果

为了验证所提出的基于最大相关熵的 KPCA 异常检测方法(简写为 KPCA-MCC)的有效性，将它与单类支持向量机(OCSVM)[4]、基于 KPCA 的异常检测方法(KPCA)[15]、基于局部多核学习的异常检测方法(KA-MKOCSVM)[29]和基于 L1-KPCA 的异常检测方法(KPCA-L1)[17]进行了比较。5 种方法中的核函数均选用高斯核函数，即：

$$K(\boldsymbol{x},\boldsymbol{y})=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2\sigma^2}\right) \tag{25}$$

所选用的 16 个基准数据集均取自 UCI 机器学习数据库[30]。由于所选取的数据集均被用于两类分类，为了使它们适用于单类分类，在每个数据集中，将某一类样本用作正常数据，将另一类样本用作异常数据。从正常数据中随机选取 70%作为训练集，并从异常数据中随机选取 5%作为噪声加入到训练集中，将剩余的所有正常数据和异常数据作为测试集。此外，将几何均值(g-mean)[31]用作性能度量，其表达式如下：

$$g=\sqrt{a^+\times a^-} \tag{26}$$

其中，a^+ 表示正常数据判别为正常的比例，a^- 表示异常数据判别为异常的比例。

为了使 KPCA-MCC 取得较好的分类性能，其核函数宽度参数 σ_1 和相关熵的宽度参数 σ_2 的最优取值是利用网格搜索策略分别在范围{0.01,0.05,0.1,0.15,…,8}和{2^{-4},2^{-3},…,1,2,3,4}内穷举选取的，特征向量个数 q 在范围{2,3,…,22}内穷举选取，其他方法的最优参数也采用相同的方法进行选取。

为了减小训练集随机选取对实验结果的影响，将上述实验在每个数据集上分别重复 20 次，将平均 g-mean 值作为最终的结果。训练和测试结果以及 20 次实验结果的标准差分别如表 1 和表 2 所列。

表 1 5 种不同方法在 UCI 基准数据集上的训练结果

Table 1 Training results of five different methods on UCI benchmark data sets

Data sets	OCSVM	KPCA	KPCA-L1	KA-MKOCSVM	KPCA-MCC
Breast cancer	0.7198±0.0289	0.7273±0.0116	0.7825±0.0275	0.5250±0.0062	0.8567±0.0575
Cancer	0.7915±0.0398	0.7120±0.0863	0.7568±0.0168	0.5767±0.0046	0.8170±0.0544
Cleverland heart	0.7280±0.0375	0.8011±0.0388	0.8298±0.0673	0.6689±0.0075	0.8741±0.0329
Diabetis	0.7325±0.0762	0.7875±0.0707	0.7826±0.0188	0.6766±0.0068	0.7258±0.0541
German	0.7454±0.0828	0.7639±0.0766	0.7591±0.0695	0.5125±0.0076	0.7125±0.0581
Glass	0.8347±0.0536	0.8853±0.0487	0.8727±0.0531	0.5785±0.0154	0.8974±0.0725
Heart	0.7946±0.0605	0.8360±0.0658	0.8490±0.0349	0.7579±0.0095	0.8865±0.0694
Hepatitis	0.7172±0.0454	0.7314±0.0621	0.7492±0.0524	0.7207±0.0076	0.7859±0.0215
Housing	0.8569±0.0453	0.8307±0.0597	0.8544±0.0907	0.8579±0.0217	0.8925±0.0533
Liver	0.7453±0.0652	0.7531±0.0347	0.8166±0.0811	0.7855±0.0435	0.8197±0.0894
Parkinsons	0.8265±0.0873	0.8136±0.0261	0.8655±0.0818	0.8299±0.0102	0.9169±0.0268
Pima	0.7485±0.0219	0.7542±0.0745	0.8127±0.0502	0.7509±0.0072	0.8397±0.0669
Sonar	0.8069±0.0352	0.8430±0.0316	0.8218±0.0219	0.7685±0.0299	0.8652±0.0714
Thyroid	0.8318±0.0708	0.8463±0.0822	0.8590±0.0591	0.8138±0.0166	0.8965±0.0343
Wdbc	0.8218±0.0536	0.8252±0.0726	0.8324±0.0413	0.6886±0.0398	0.8712±0.0359
Wholesale customers	0.8559±0.0645	0.8219±0.0574	0.8793±0.0413	0.7333±0.0192	0.8941±0.0289

注：平均 g-mean±标准差

表 2 5 种不同方法在 UCI 基准数据集上的测试结果

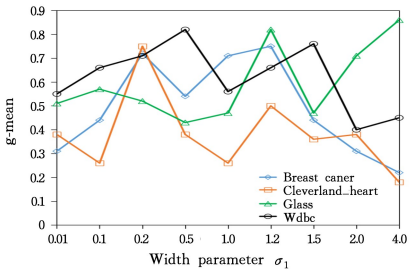
Table 2 Testing results of five different methods on UCI benchmark data sets

Data sets	OCSVM	KPCA	KPCA-L1	KA-MKOCSVM	KPCA-MCC
Breast cancer	0.6501±0.0675	0.6263±0.0326	0.7091±0.0168	0.5034±0.0066	**0.7731±0.0557**
Cancer	0.7226±0.0571	0.6140±0.0296	0.6809±0.0218	0.5342±0.0019	**0.7935±0.0681**
Cleverland heart	0.6292±0.0285	0.7022±0.0183	0.7158±0.0170	0.6192±0.0085	**0.7653±0.0186**
Diabetis	0.5419±0.0386	0.6773±0.0401	**0.6914±0.0522**	0.5811±0.0125	0.6834±0.0167
German	0.5748±0.0125	**0.6445±0.0437**	0.6051±0.0177	0.4644±0.0093	0.6169±0.0335
Glass	0.7397±0.0545	0.8803±0.0109	0.8560±0.0115	0.5118±0.0078	**0.8867±0.0225**
Heart	0.6833±0.0786	0.7126±0.0428	0.7295±0.0586	0.7043±0.0135	**0.7516±0.0523**
Hepatitis	0.5298±0.0217	0.5825±0.0153	0.5936±0.0428	0.5821±0.0217	**0.6051±0.0782**
Housing	0.6493±0.0461	0.5416±0.0255	0.6529±0.0112	0.6526±0.0137	**0.6593±0.0557**
Liver	0.5796±0.0175	0.6441±0.0328	0.7034±0.0339	**0.8426±0.0262**	0.7599±0.0905
Parkinsons	0.7719±0.0772	0.8072±0.0598	0.8225±0.0407	0.8102±0.0309	**0.8918±0.0512**
Pima	0.6682±0.0353	0.6850±0.0268	0.7696±0.0327	0.6980±0.0069	**0.7845±0.0258**
Sonar	0.6639±0.0410	0.7012±0.0442	0.7218±0.0524	0.5033±0.0056	**0.7370±0.0476**
Thyroid	0.7267±0.0336	0.7167±0.0142	0.7576±0.0305	0.7309±0.0282	**0.7651±0.0494**
Wdbc	0.7602±0.0413	0.7911±0.0987	0.8139±0.0781	0.5421±0.0099	**0.8255±0.0309**
Wholesale customers	0.7452±0.0591	0.7926±0.0537	0.8185±0.0156	0.7102±0.0138	**0.8309±0.0758**

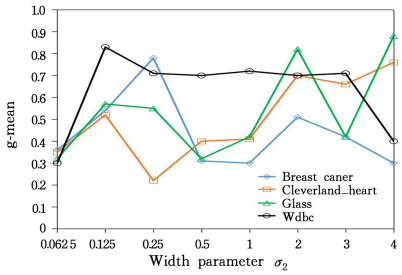
注：平均 g-mean±标准差

由表 1 中的训练结果可以看出，除了 Diabetis 和 German，所提方法在其余 14 个数据集上都取得了较高的平均 g-mean 值，这表明与其他 4 种方法相比，所提方法在这些数据集上得到了更为充分的训练。由表 2 中的结果可以看出，除了 Diabtis，German 和 Liver，所提方法在其余 13 个数据集上均取得了比其他 4 种方法更高的平均 g-mean 值，即所提 KPCA-MCC 具有优于其他 4 种相关方法的泛化性能。因此，利用最大相关熵代替基于 ℓ_2 范数的度量，可以有效提高 KPCA 异常检测方法的抗噪声能力。

此外，为了观察核函数宽度参数 σ_1 和相关熵宽度参数 σ_2 的不同取值对 KPCA-MCC 性能的影响，在 Breast can-cer，Cleverland heart，Glass 和 Wdbc 这 4 个基准数据集上对 KPCA-MCC 在不同参数设置下的平均 g-mean 值进行了探讨。图 1(a)给出了当 $\sigma_2=2$ 时，KPCA-MCC 的性能随 σ_1 在范围 $\{0.01, 0.1, 0.2, 1, 1.2, \cdots, 4\}$ 内不同取值下的变化情况；图 1(b)给出了当 $\sigma_1=0.2$ 时，KPCA-MCC 的性能随 σ_2 在范围 $\{2^{-4}, 2^{-3}, \cdots, 1, 2, 3, 4\}$ 内不同取值下的变化情况。



(a)随宽度参数 σ_1 的不同取值得到的不同分类性能



(b)随宽度参数 σ_2 的不同取值得到的不同分类性能

图 1 KPCA-MCC 在 4 个基准数据集上随宽度参数的不同取值得到的分类性能

Fig. 1 Classification performances of KPCA-MCC with different values of width parameters on four benchmark data sets

由图 1(a)可以发现，在数据集 Breast cancer，Cleverland heart，Glass 和 Wdbc 上，KPCA-MCC 的平均 g-mean 值在 σ_1 分别取 1.2，0.2，4 和 0.5 时达到最大，在最大值附近 KPCA-MCC 的性能下降很快。此外，从 4 条曲线的形状来看，平均 g-mean 值的波动都较大，由图 1(b)也可以观察到类似的结果。

因此，所提 KPCA-MCC 的分类性能取决于参数的设置，若参数设置得当，则 KPCA-MCC 能够取得较高的分类性能；若设置不当，则 KPCA-MCC 的分类性能会变差。

结束语 将传统的 KPCA 用于异常检测时，若训练样本中存在噪声，则 KPCA 的性能将会受到较为严重的影响。为了提高 KPCA 的抗噪声能力，利用相关熵代替其目标函数中基于 ℓ_2 范数的度量。此外，利用半二次优化技术对所提方法的优化问题进行迭代求解。通过与相关方法在 UCI 基准数据集上进行比较，验证了所提方法的抗噪声能力和泛化性能。

如实验结果所示，所提方法的分类性能严重依赖于核函数宽度参数和相关熵宽度参数的设置。为了使所提方法取得较优的性能，采用了穷举法对它们的参数值进行选取，因此非常耗时。迄今为止，针对基于高斯核函数的单类分类器，存在一些核函数宽度参数的选取方法[32-34]，这些方法都需要预先定义一个宽度参数候选集，并为候选集中的每一个宽度参数计算相应的目标函数值，最终基于最大(或最小)目标函数值，确定最优的宽度参数。尽管这些方法解决了高斯核函数的宽度参数选取的部分问题，但是仍存在不足，如宽度参数候选集难以确定、计算复杂度非常大等。针对相关熵的宽度参数选取，存在一些启发式方法，如 Silverman 规则[35]、K 近邻法[36]等。然而，当训练集中存在噪声时，这些启发式方法往往不能取得较优的效果。因此，在未来的工作中，针对核函数宽度参数和相关熵宽度参数的选取，将会尝试使用其他的参数选取方法，如智能优化算法。

参 考 文 献

[1] TAX D M J. One-class classification: concept learning in the absence of counter examples[D]. Delft: Delf University of Technology, 2001.

[2] PENNY K I, JOLLIFFE I T. A comparison of multivariate outlier detection methods for clinical laboratory safety data[J]. The

Statistician, 2001, 50(3): 295-307.
[3] OH C K, SOHN H, BAE I H. Statistical novelty detection within the Yeongjong suspension bridge under environmental and operational variations[J]. Smart Materials and Structures, 2009, 18(12): 5022-5029.
[4] SCHÖLKOPF B, WILLIAMSON R C, SMOLA A J. Support vector method for novelty detection[C]// Advances in Neural Information Processing Systems. 2000: 582-588.
[5] TAX D M J, DUIN R P W. Support vector data description[J]. Machine Learning, 2004, 54(1): 45-66.
[6] SCHÖLKOPF B, SMOLA A, MÜLLER K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural Computation, 1998, 10(5): 1299-1319.
[7] JOLLIFFE I T. Principal Component Analysis [M]. Berlin: Springer-Verlag, 2005.
[8] TEIXEIRA A R, TOMÉ A M, STADLTHANNER K, et al. KPCA denosing and the pre-image problem revisited[J]. Digital Signal Processing, 2008, 18(4): 568-580.
[9] LIAN H. On feature selection with principal component analysis for one-class SVM [J]. Pattern Recognition Letters, 2012, 33(9): 1027-1031.
[10] HILL J, CORONA E, AO J, et al. Information Theoretic Clustering for Medical Image Segmentation[M]. Berlin: Springer-Verlag, 2014.
[11] DEBRUYNE M, VERDONCK T. Robust kernel principal component and classification[J]. Advances in Data Analysis and Classification, 2010, 4(2): 151-167.
[12] KIM C, KLABIAN D. A simple and fast algorithm for L1-norm kernel PCA[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(8): 1842-1855.
[13] DUAN X, QI P, TIAN Z. Registration for variform object of remote-sensing image using improved robust weighted kernel principal component analysis[J]. Journal of The Indian Society of Remote Sensing, 2016, 44(5): 675-686.
[14] FAN J, CHOW T W S. Exactly robust kernel principal component analysis[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(3): 749-761.
[15] HOFFMANN H. Kernel PCA for novelty detection[J]. Pattern Recognition, 2007, 40(3): 863-874.
[16] DUDA R O, HART P E, STORK D G. Pattern Classification. 2nd Ed. [M]. New York: Wiley Press, 2001.
[17] XIAO Y, WANG H, XU W, et al. L1 norm based KPCA for novelty detection[J]. Pattern Recognition, 2013, 46(1): 389-396.
[18] ALZATE C, SUYKES J. Kernel component analysis using an epsilon-insensitive robust loss function[J]. IEEE Transactions on Neural Networks, 2008, 19(9): 1583-1598.
[19] WANG D, TANAKA T. Robust kernel principal component analysis with $l_{2,1}$-regularized loss minimization[J]. IEEE Access, 2020, 8(81): 864-875.
[20] PRINCIPE J C. Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives[M]. New York: Springer, 2010.
[21] LIU W, POKHAREL P P, PRINCIPE J C. Correntropy: properties and applications in non-Gaussian signal processing[J]. IEEE Transactions on Signal Processing, 2007, 55(11): 5286-5298.
[22] HE R, HU B, ZHENG W, et al. Robust principal component analysis based on maximum correntropy criterion [J]. IEEE Transactions on Image Processing, 2011, 20(6): 1485-1494.
[23] YUAN X, HU B. Robust feature extraction via information theoretic learning [C] // International Conference on Machine Learning, Montreal. 2009: 1193-1200.
[24] KWAK N. Principal component analysis based on L1-norm maximization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(9): 1672-1680.
[25] VAPNIK V N. The Nature of Statistical Learning Theory[M]. New York: Springer, 2000.
[26] ZHOU Z. Machine Learning[M]. Beijing: Tsinghua University Press, 2016.
[27] GÜLER O. Convex Analysis[M]. New York: Springer, 2010.
[28] SUN Q, ZHANG H, WANG X, et al. Sparsity constrained recursive generalized maximum correntropy criterion with variable center algorithm[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2020, 67(12): 3517-3521.
[29] GAUTAM C, BALAJI R, SUDHARSAN K, et al. Localized multiple kernel learning for anomaly detection: one-class classification[J]. Knowledge Based Systems, 2019, 165: 241-252.
[30] LICHMAN M. UCI Machine Learning Repository[EB/OL]. University of California, Irvine, School of Information and Computer Sciences, 2019. http://archive.ics.uci.edu/ml.
[31] WU M, YE J. A small sphere and large margin approach for novelty detection using training data with outliers [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(11): 2088-2092.
[32] DENG H, XU R. Model selection for anomaly detection in wireless ad hoc networks[C]// 2007 IEEE Symposium on Computational Intelligence and Data Mining. 2007: 540-546.
[33] WANG S, YU J, LAPIRA E, et al. A modified support vector data description based novelty detection approach for machinery components[J]. Applied Soft Computing, 2013, 13(2): 1193-1205.
[34] XIAO Y, WANG H, XU W. Parameter selection of Gaussian kernel for one-class SVM[J]. IEEE Transactions on Cybernetics, 2015, 45: 941-953.
[35] SILVERMAN B W. Density Estimation for Statistics and Data Analysis[M]. London: Chapman and Hall, 1986.
[36] LI Y, WANG Y, WANG Y, et al. Quantum clustering using kernel entropy component analysis[J]. Neurocomputing, 2016, 202: 36-48.

LI Qi-ye, born in 1995, postgraduate. His main research interests include novelty detection and kernel methods.

XING Hong-jie, born in 1976, Ph.D, professor, master supervisor. His main research interests include kernel methods, neural networks, novelty detection and ensemble learning.

(责任编辑:喻藜)