

Grassberger 熵随机森林在窃电行为检测的应用

阙华坤, 冯小峰, 刘盼龙, 郭文翀, 李健, 曾伟良, 范竞敏

引用本文

阙华坤, 冯小峰, 刘盼龙, 郭文翀, 李健, 曾伟良, 范竞敏. [Grassberger 熵随机森林在窃电行为检测的应用](#)[J]. 计算机科学, 2022, 49(6A): 790-794.

QUE Hua-kun, FENG Xiao-feng, LIU Pan-long, GUO Wen-chong, LI Jian, ZENG Wei-liang, FAN Jing-min.

[Application of Grassberger Entropy Random Forest to Power-stealing Behavior Detection](#)[J]. Computer Science, 2022, 49(6A): 790-794.

相似文章推荐 (请使用火狐或 IE 浏览器查看文章)

Similar articles recommended (Please use Firefox or IE to view the article)

[自适应的集成定序算法](#)

Adaptive Ensemble Ordering Algorithm

计算机科学, 2022, 49(6A): 242-246. <https://doi.org/10.11896/jsjcx.210200108>

[基于眼前节相干光断层扫描成像的核性白内障分类算法](#)

Classification Algorithm of Nuclear Cataract Based on Anterior Segment Coherence Tomography Image

计算机科学, 2022, 49(3): 204-210. <https://doi.org/10.11896/jsjcx.201100085>

[一种可用于分类型属性数据的多变量回归森林](#)

Multivariate Regression Forest for Categorical Attribute Data

计算机科学, 2022, 49(1): 108-114. <https://doi.org/10.11896/jsjcx.201200189>

[基于多头注意力机制的用户窃电行为检测](#)

Electricity Theft Detection Based on Multi-head Attention Mechanism

计算机科学, 2022, 49(1): 140-145. <https://doi.org/10.11896/jsjcx.210100177>

[基于随机森林的空域-频域联合特征全参考彩色图像质量评价方法](#)

Full Reference Color Image Quality Assessment Method Based on Spatial and Frequency Domain Joint Features with Random Forest

计算机科学, 2021, 48(8): 99-105. <https://doi.org/10.11896/jsjcx.200700106>

Grassberger 熵随机森林在窃电行为检测的应用

阙华坤¹ 冯小峰¹ 刘盼龙² 郭文舫¹ 李 健¹ 曾伟良² 范竞敏²

1 广东电网有限责任公司计量中心 广州 518049

2 广东工业大学自动化学院 广州 510006

(quehuakun@126.com)

摘 要 窃电行为严重危害电网安全,为了提高对窃电行为的检测效率,提出一种新型的基于 Grassberger 熵随机森林的电网用户窃电检测方法。首先,采用核主成分分析方法(Kernel Principal Component Analysis,KPCA)对用户的原始用电量的时间序列向量进行降维,提取用户的用电特征;接着,考虑到窃电样本和正常样本数量相差较大时,窃电检测的分类器训练效果较差,因此,采用数据欠采样方法建立多个数量平衡的样本子集,并采用改进的 Grassberger 熵随机森林(Random Forest,RF)算法计算信息增益,对各样本子集进行训练再集成,从而提高模型对窃电检测的准确度。以中国南方电网的专变用户窃电检测为案例,将各用户的电表采集电量数据作为模型输入,验证所提模型的窃电检测效果。

关键词: 窃电检测;随机森林;核主成分分析;Grassberger 熵

中图法分类号 F407.6

Application of Grassberger Entropy Random Forest to Power-stealing Behavior Detection

QUE Hua-kun¹, FENG Xiao-feng¹, LIU Pan-long², GUO Wen-chong¹, LI Jian¹, ZENG Wei-liang² and FAN Jing-min²

1 Metrology Center of Guangdong Power Grid Corporation, Guangzhou 518049, China

2 School of Automation, Guangdong University of Technology, Guangzhou 510006, China

Abstract Power stealing seriously endangers the grid security. In order to improve the efficiency of electricity theft detection, this paper proposes a novel method for electricity stealing detection based on Grassberger entropy random forest. First, KPCA is applied to reduce the dimensionality of the original power time series for extracting the user power consumption characteristics. Then, considering the unbalance of the number of theft samples and normal samples, the data under sampling method is used to establish multiple quantitatively balanced sample subsets. The random forest with improved Grassberger entropy is used to compute information gain, so as to improve the accuracy of the model in power theft detection. Finally, the electricity consumption dataset of China Southern Power Grid is used to verify the power stealing detection effect of the proposed model.

Keywords Power stealing detection, Random Forest, Kernel principal component analysis, Grassberger entropy

以物理电网为基础,与计算机技术、控制技术、传感测量技术等现代先进技术高度集成而形成的智能电网(Smart Grid)正在高速建设当中。智能电网与互联网紧密相连,由于电力计量系统的复杂性以及互联网的连通性和开放性,不法分子利用漏洞进行窃电时有发生^[1]。

窃电行为会对电力系统造成非技术性损失(Non-technical Losses, NTL),不仅影响电网的正常生产运营,降低电力系统的稳定性和可靠性,甚至会严重损害国民经济。窃电的不正当操作会损害电力设备,造成停电事故,严重扰乱居民的正常生活和社会的稳定发展。用户的窃电手段多样,如欠流法窃电、欠压法窃电、移相法窃电、高频电源窃电等。常用的窃电检测方法是通过分析电网用户的电流、电压和相位等特征量来实判断窃电行为。

智能电网生成并储存的用户数据中,窃电用户数据只是

很小的一部分,更多的是正常使用的用户数据。因此,有窃电行为的用户用电数据与正常用电的用户数据构成了一个不平衡数据集。在基于用户数据分析的窃电检测研究中,由于数据集不平衡,分类算法在训练过程中会忽略窃电用户数据集中包含的信息,使得算法在检测时性能严重下降。因此,检测窃电行为时,需要对不平衡数据集进行处理,以提高算法对窃电行为的检测性能。

目前对窃电行为的检测大多是基于人工的现场检测,如现场检查设备或者硬件问题、比较抄表数据、检查线路情况等。人工现场检测的成本代价较高,同时考虑到前面所说的窃电行为带来的严重后果,开发一种准确而快速的窃电行为检测技术非常有必要。本文第 1 节主要回顾相关的文献;第 2 节主要介绍基于 KPCA 和改进 Grassberger 熵的随机森林电网用户窃电检测方法的相关原理;第 3 节结合算例得到

基金项目:中国南方电网有限责任公司科技项目(GDKJXM20185800);国家自然科学基金(61803100)

This work was supported by the Science and Technology Project of China Southern Power Grid Co. Ltd(GDKJXM20185800) and National Natural Science Foundation of China(61803100).

通信作者:曾伟良(weiliangzeng@gdut.edu.cn)

实验结果并进行分析;最后总结全文。

1 文献回顾

1.1 窃电检测研究现状

目前反窃电措施主要分为技术驱动和数据驱动。其中技术驱动是利用先进的封装技术或者设备电路硬件技术来防止窃电行为的发生,主要分为基于硬件和基于物理的防护;数据驱动措施主要是利用算法对采集到的电网节点信息和电力系统计量设备等得到的用户数据进行处理分析,以达到对窃电行为进行检测的目的。数据驱动的窃电检测方法目前主要有基于专家系统、节点状态、回归分析、机器学习的方法。本文主要是利用改进的随机森林算法对用户数据等进行处理分析,即基于数据驱动的机器学习窃电检测方法。

机器学习分为监督学习和无监督学习,其应用在窃电检测领域均有研究。Tian 等提出一种基于密度的聚类方法(Density-Based Spatial Clustering of Applications with Noise DBSCAN),通过对用户电力曲线进行聚类,以离群异常度为指标完成窃电检测^[1]。Wang 等改进了 DBSCAN 算法,完成了用户大数据流实时窃电检测^[2]。Fahim 等通过对用户电量数据采用模糊 C 均值聚类分析来进行窃电行为检测^[3]。Leandro 等采用最优路径森林(Optimal-Path Forest, OPF)算法进行窃电行为检测^[4]。Zanetti 等采用 K-mean 聚类算法提取用户用电特征,完成窃电检测^[5]。

无监督学习虽然不需要分类标签和训练数据集,但是对参数设定依赖性较强,普适性较差。监督学习通过对已有的训练样本进行训练得到一个最优的检测模型,然后输入待检测数据完成检测任务。Muniz 等利用 5 个神经网络联合运算,对 2 万用户的用电数据进行训练^[6]。随后 Costa 等利用 12 个月的用户用电数据,并结合不同的特征值训练神经网络,检测精度在前文基础上有所提高^[7]。Lin 等首先利用思维进化算法(Mind Evolutionary Algorithm, MEA)优化 BP 神经网络,然后用优化的 BP 根据用户特征曲线与用电值的关系完成窃电检测^[8]。Spiri'c 根据用户类型、日期、位置、气候等采取单个决策树方法进行窃电行为检测^[9]。Yang 等首次采用深度森林分类算法从电量、电压、电流、功率因素等数据提取的特征对用户窃电行为进行检测,利用了深度森林超参数少、计算效率高的优点^[10]。Cai 等首先采用密集卷积神经网络(DenseNet)对大规模的用户用电数据集进行特征提取,然后根据得到的特征使用随机森林算法进行窃电行为检测,在融合模型时,使用网格搜索算法确定最优参数^[11]。

目前对于窃电检测的研究在实际应用中较少能达到理想状态,本文采用改进的 Grassberger 熵计算随机森林的信息增益,并利用集成思想集成各子分类器来提高检测准确率。

1.2 不平衡数据处理研究现状

智能电网获取到的用户用电数据中,正常用电用户数据数量远大于窃电用户数据数量,即构成了不平衡数据集。对于不平衡数据集,智能算法在训练过程中容易忽略窃电用户数据中的重要信息,若直接分类,算法倾向于将窃电用户分类为正常用户,从而导致算法的检测性能降低。

目前针对不平衡数据处理方法的研究主要有重采样^[12-13]、代价敏感学习^[14]、集成学习^[15]等。在数据层面,目前最成熟方法为重采样,重采样分为过采样和欠采样。Chen 等

提出了合成样本技术(Synthetic Minority Oversampling Technique, SMOTE)对少数样本进行过采样^[16],随后对 SMOTE 方法进行了改进^[12,17-18]。Del Rio 提出了一种 k -Nearest Neighbor(KNN)的欠采样算法^[19]。Zhang 等将欠采样和过采样结合使用来进行数据处理^[20]。Li 等针对软件缺陷预测样本数据不均衡问题,提出了基于代价敏感的软件缺陷预测方法^[14]。代价敏感学习虽然通过对正常用户和窃电用户分别设置不同的权重,使得分类算法更加重视窃电样本,但对窃电用户的惩罚需要考虑经济成本等多方面的因素,且权重的大小设置主观性较强。

2 基于 KPCA 和改进 Grassberger 熵随机森林的检测模型

2.1 基于 KPCA 的用电特征提取

原始电量数据的维度较高,若直接应用于分类器的训练,会导致训练效率低、计算复杂度高、训练效果差等问题。特征提取技术从原始数据集中提取最相关特征,可以在尽可能地压缩数据集、提取有用信息的同时,最大限度地保留重要信息,KPCA 算法因其较高的信息保留度、较快的计算效率等而在特征提取中的应用较好。

KPCA 是 PCA 的一种变体,通过将核方法推广到 PCA 中发展起来。核方法最初用于 SVM,后来被引入到许多具有点积术语的算法。考虑到 PCA 局限于线性可分特征的处理,KPCA 引入核方法将原始输入映射到高维特征空间,在高维特征空间中它是线性可分离的,因此具有可提取非线性可分特征的优越性,常用核函数如表 1 所列。映射公式如式(1)所示:

$$(x_i, x_j) \rightarrow F(x_i, x_j) = \Phi(x_i)^T \cdot \Phi(x_j) \quad (1)$$

其中, Φ 是映射函数。

表 1 常用核函数及其计算公式

Table 1 Common kernel functions and their calculation formulas

核函数	计算公式
P 阶多项式核	$F(x_i, x_j) = (\gamma \cdot x_i^T x_j + b)^p$
径向基核	$F(x_i, x_j) = \exp(-\gamma \ x_i - x_j\ ^2)$
多层感知器核	$F(x_i, x_j) = \tanh(\gamma \cdot x_i^T x_j + b)$

其中, F 是高维特征向量空间; γ 是控制映射缩放的核宽度参数。

KPCA 将原始输入映射到高维特征空间后,在高维特征空间中进行 PCA 过程。映射得到的特征空间中 x_i 满足中心化的条件,即:

$$\sum_{i=1}^n \Phi(x_i) = 0 \quad (2)$$

则特征空间中的协方差矩阵 σ 为:

$$\sigma = \frac{1}{n} \sum_{i=1}^n \Phi(x_i) \Phi(x_i)^T \quad (3)$$

可知 σ 的特征值 $\lambda \geq 0$, 特征向量 v 可由式(4)求得:

$$\sigma v = \lambda v \quad (4)$$

考虑到所有的特征向量 v 可由线性变换组成:

$$v = \sum_{i=1}^n \alpha_i \Phi(x_i) \quad (5)$$

其中, α 为线性变换系数。

则有:

$$(\Phi(x_j) \cdot \sigma v) = \lambda (\Phi(x_i) \cdot v) \quad (6)$$

显然满足:

$$n \lambda (\Phi(x_i) \cdot \Phi(x_j)) \alpha = (\Phi(x_i) \cdot \Phi(x_j))^2 \alpha \quad (7)$$

求解式(7)即可得到特征值 λ 和特征向量 $\mathbf{v}$ 。步骤如图1所示。

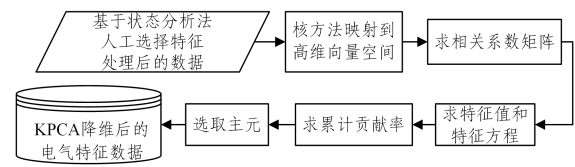


图1 KPCA提取用电特征步骤

Fig. 1 Steps of KPCA to extract electricity consumption characteristics

2.2 解决类别不平衡问题

由于窃电用户的数量一般远小于正常用户,因此窃电排查往往面临类别不平衡问题,若直接分类,算法倾向于将窃电用户分类为正常用户,排查结果很差。为了处理类别不平衡问题,目前的主要解决方法包括对多数样本的欠采样、对少数样本的过采样等。若类别不平衡比例过大,分类算法将不再适用,可以改用异常点识别算法进行分析。针对窃电排查问题,对正常用户样本进行欠采样,可以有效缩短训练时间,但同时会舍弃较多的未被采样的正常用户样本,导致分类器只能学习输入样本的局部特征,影响分类结果;过采样方法通过生成本来增加窃电用户样本的数量,从而平衡两类样本的比例,但窃电样本的生成方法不同,对分类结果有较大影响,且数量的增加会增加训练时间;代价敏感度学习通过对正常用户和窃电用户分别设置不同的权重,使得分类算法更加重视窃电样本,但对窃电用户的惩罚需要考虑经济成本等多方面的因素,且权重的大小设置主观性较强。

本文在将多数样本划分为与少数样本规模相等的子集的基础上,建立初始组合集合 $\mathbf{X}_{or}$,再对每个组合子集进行重采样得到 $\mathbf{X}_{re}$,将 $\mathbf{X}_{re}$ 输入到改进 Grassberger 熵随机森林中,最后集成每个随机森林的训练结果,具体步骤如下:

输入:设训练集有 N_r 个用户,其中正常用户样本集 $\mathbf{X}_p$,窃电用户样本集 $\mathbf{X}_q$, $|\mathbf{X}_q| < |\mathbf{X}_p|$ 。设共有 s 个组合子集,随机森林 H 中有 r_i 个基分类器($i=1,2,\dots,s$)。

步骤1 在 $\mathbf{X}_p$ 中随机不放回采样,依次得到 $\mathbf{X}_{p1}, \mathbf{X}_{p2}, \dots, \mathbf{X}_{ps}, |\mathbf{X}_{pi}| = |\mathbf{X}_q|$,且

$$(\mathbf{X}_{p1} + \mathbf{X}_{p2} + \dots + \mathbf{X}_{ps}) = \mathbf{X}_p \tag{8}$$

步骤2 建立初始组合集合: $\mathbf{X}_{or} = \{(\mathbf{X}_{p1} + \mathbf{X}_q), (\mathbf{X}_{p2} + \mathbf{X}_q), \dots, (\mathbf{X}_{ps} + \mathbf{X}_q)\}$ 。

步骤3 用自助采样法对 $\mathbf{X}_{or}$ 中的每一个初始子集进行采样,最终得到重采样后的集合 $\mathbf{X}_{re}$ 。

步骤4 将重采样后的 s 个子集作为各个随机森林的输入,第 i 个随机森林的训练结果为:

$$H_i = \text{sgn}(\sum_{j=1}^{r_i} h_{i,j}(x)) \tag{9}$$

其中, $h_{i,j}(x)$ 为第 i 个随机森林中第 j 个决策树的训练结果。

输出:集成各个训练后的随机森林,得到分类器 H :

$$H = \text{sgn}(\sum_{i=1}^s \alpha_i H_i - \theta) \tag{10}$$

其中, α_i 为各个随机森林的集成权重, θ 为分类阈值。

这种方法将原来的训练集划分为 s 个欠采样和重采样的组合子集,将其作为输入对 s 个随机森林分类器进行训练。欠采样环节建立的初始组合集合 $\mathbf{X}_{or}$,避免了多数样本的特征丢失,而在此基础上的重采样则避免了各个随机森林采用同样的少数样本所导致的过拟合问题。在时间复杂度上,由于

s 个随机森林可以并行训练,欠采样和重采样所用时间相较于随机森林的训练时间可以忽略,所以本方法和单个随机森林的训练时间复杂度基本相同;在分类结果上,集成算法的分类效果主要和各子集的分类准确度以及子集间的多样性有关,因为随机森林算法的适应性强,各子分类器的分类准确度较高,同时重采样保证了各输入子集的多样性,因此对各子分类器集成后,可以得到较好的分类结果和窃电置信度。

2.3 改进 Grassberger 熵的随机森林

随机森林采用集成学习的思想,将多个子分类器相结合,从而获得比单一分类器更优越的性能,其性能与子分类器的准确性和输入训练子集的多样性密切相关,前者保证了集成分类器的准确性,后者保证了集成分类器的泛化能力。

在处理输入集时,首先,随机森林采用自助采样法,有放回地随机选择样本,建立 m 个训练子集,作为各子分类器的输入样本集(一般选择决策树作为子分类器);然后在决策树的节点属性中随机选择一个包含 k 个属性的子集,在子集中选择最佳属性,参数 k 的引入增加了属性选择的随机性,从而增加了子分类器之间的多样性,算法的泛化能力得以提升;最后通过投票法,结合各子分类器所占的权重,输出集成后的分类结果及置信度。

对于每个随机森林分类器中的分裂节点,最关键的点是搜寻信息增益最大的分裂属性 θ_j ,信息增益表达式为:

$$I(D_j, \theta) = H(D_j) - \sum_{k \in \{L, R\}} \frac{|D_j^k|}{|D_j|} H(D_j^k) \tag{11}$$

其中, D_j 为第 j 个节点的样本数目; R, L 为树的右、左孩子节点; $H(D_j)$ 为第 j 个节点的 Grassberger 熵,其表达式如下:

$$H(D_j) = \ln N - \frac{1}{N} \sum_{k=1}^K c_k G(c_k) \tag{12}$$

其中, c_k 为某个节点样本中第 k 类样本的数目, $G(c_k)$ 为改进的 Grassberger 熵函数。利用改进的 Grassberger 熵计算随机森林的信息增益,能提高检测的准确率^[21]。改进的 Grassberger 熵的函数 $G(c_k)$ 为:

$$G(c_k) = \psi(c_k) + \frac{1}{2} (-1)^{c_k} \left[\psi\left(\frac{c_k}{2} + 1\right) + \psi\left(\frac{c_k + 1}{2} + 1\right) \right] \tag{13}$$

其中, $\psi(c_k)$ 为 Gamma 函数对数的导数。

3 算例分析

为了验证本文模型的性能,相关环境如下:Windows10 专业版,CPU 为 AMD Ryzen 7 3750H;内存为 8GB;编程语言:Python,版本为 3.8;基于 Pytorch 的 scikit-learn 机器学习框架进行实验编程。

3.1 数据集设置及评价指标

本文的数据来源为中国南方电网广东省某地区 2019 年 1 月 1 日至 2019 年 3 月 31 日 1052 个用户的电力数据,窃电用户有 61 户。其中数据信息包括了三相电压电流、功率因素、有功功率等,且用户分为工业用电用户、商用用电用户、居民用电用户 3 类。利用获取的数据带入到本文提出的基于 KPCA 和改进 Grassberge 熵的随机森林模型,以下简称 KP-CA-IGERF,利用得到结果进行分析验证。

为从多方面评价窃电算法的有效性,本文基于二分类的混淆矩阵(TP,FP,TN,FN),并将窃电样本作为正类,同时通过对比其他算法的准确度(ACC)以及 ROC 曲线下的面积

(AUC)来说明本文所提模型的有效性。本文还比较分析了不同窃电比以及 KPCA 提取维度后维数对 ACC、程序运行时间和 AUC 的影响。

3.2 准确率与 AUC 比较分析

基于以上数据模型输入,对比不同算法模型的准确率与 AUC 来说明本文算法模型的有效性。准确率(ACC)表示正确分类的样本数与样本总数的比率,值越高,表示检测算法的性能越好。ROC 曲线最开始在信号处理方面使用较广,最近被广泛应用于机器学习以评估分类算法的性能。ROC 曲线以 TPR 为纵坐标,以 FPR 为横坐标,通过改变算法的阈值来得到一条描述曲线,绘制出来的 ROC 曲线越靠近点(0,1),表示算法性能越好。假如 ROC 曲线并没有相交,则最靠近左上角的曲线代表的性能最好。实际曲线情况很复杂,因此引入 AUC 值来评估 ROC 曲线中算法的性能。AUC 值一般在 0.5~1 之间,越接近 1 表示性能越好。

由表 2 可知,本文提出的 KPCA-IGERF 模型准确率在测试集和训练集中的效果都优于目前常用的算法模型,同时也证明了相比未改进的随机森林算法模型 KPCA-RF,基于改进 Grassberger 熵的 KPCA-IGERF 模型的准确率高于未改进的 KPCA-RF 模型。

表 2 不同算法模型准确率对比
Table 2 Comparison of accuracy of different algorithm models

算法模型	训练集	测试集
KPCA-IGERF	99.61	98.64
KPCA-RF	97.06	95.24
WDNet	96.52	93.21
SVM	86.72	81.55

从表 3 和图 2 可知,本文提出的 KPCA-IGERF 模型在测试集上的性能较好,AUC 达到 0.99,相比未改进的 Grassberger 熵的 KPCA-RF 模型具有更好的泛化能力,能对新用户是否有窃电行为具有较好的适用性。

表 3 不同算法模型在测试集上的 AUC 对比
Table 3 AUC comparison of different algorithm models on test set

算法模型	AUC
KPCA-IGERF	0.99
KPCA-RF	0.97
WDNet	0.93
SVM	0.84

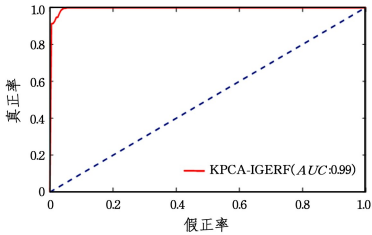


图 2 ROC 曲线
Fig. 2 ROC curve

3.3 窃电占比与特征向量维数影响下的性能指标

为了进一步分析该模型的性能,本文通过改变数据集中不同窃电比例以及利用 KPCA 降维特征向量时不同的维数值来分析本文提出的 KPCA-IGERF 的性能。不同窃电比例数据集设置如表 4 所列。

表 4 数据集设置
Table 4 Data set settings

	5%		10%		15%		20%	
	窃电	正常	窃电	正常	窃电	正常	窃电	正常
总样本	20	380	50	450	45	255	50	200
训练集	10	190	30	270	25	145	25	100
测试集	10	190	20	180	20	110	25	100

将表 4 中的数据集导入 KPCA-IGERF 模型中进行训练,得到该模型在不同窃电比例下的窃电检测评价,如表 5 所列。

表 5 不同窃电比例下的窃电检测评价
Table 5 Evaluation of electricity theft detection with different ratios of electricity theft

窃电比例	准确率	精确率	召回率	F1 值	AUC
5%	0.9843	0.9626	0.9989	0.9685	0.9938
10%	0.9882	0.9549	0.9992	0.9642	0.9960
15%	0.9864	0.9633	0.9972	0.9759	0.9946
20%	0.9885	0.9721	0.9982	0.9814	0.9973

从表 5 可知,不同窃电比例下,本文提出的 KPCA-IGERF 模型的准确率、精确率、召回率、F1 值、AUC 性能指标都较好。

同时本文还对比分析了在不同窃电比例下,窃电检测效果随降维维度的变化关系,如图 3—图 6 所示。

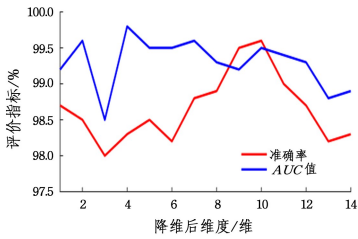


图 3 5%窃电比例下特征向量维度对检测效果的影响
Fig. 3 Influence of feature vector dimension on detection effect with theft ratio of 5%

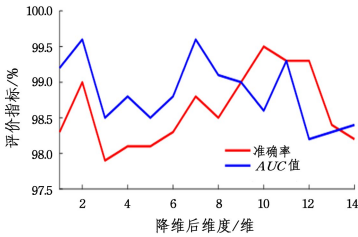


图 4 10%窃电比例下特征向量维度对检测效果的影响
Fig. 4 Influence of feature vector dimension on detection effect with theft ratio of 10%

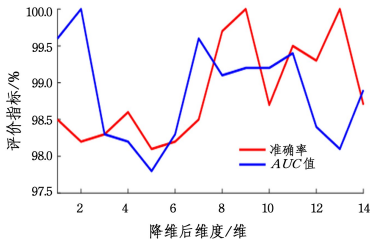


图 5 15%窃电比例下特征向量维度对检测效果的影响
Fig. 5 Influence of feature vector dimension on detection effect with theft ratio of 15%

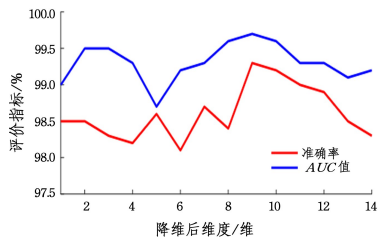


图 6 20%窃电比例下特征向量维度对检测效果的影响
Fig. 6 Influence of feature vector dimension on detection effect with theft ratio of 20%

从图 3—图 6 可以看出,即使在不同窃电比例下,改变 KPCA 对特征向量降维后的维数,该算法模型的准确率和 AUC 都较高,进一步说明了本文提出的模型具有较好的泛化能力,对检测新用户是否有窃电行为具有较好的适用性。

结束语 为了提高窃电检测技术检测性能,本文提出了一种新型的基于核主成分分析(KPCA)和改进 Grassberger 熵随机森林的电网用户窃电检测方法。采用 KPCA 对用户的电量向量进行降维,提取用户的用电特征;接着,考虑到窃电样本和正常样本数量相差较大时,窃电检测的分类器训练效果较差,因此,采用数据欠采样方法建立多个数量平衡的样本子集,并采用改进的 Grassberger 熵计算随机森林算法中的信息增益,然后对各样本子集进行训练。结果表明,该算法模型在窃电检测领域具有较高的准确率,AUC 也接近 1,具有较好的泛化能力,且其他性能指标也都较好。同时本文还分析了不同窃电比例以及 KPCA 对特征向量降维后的维数对算法模型的影响,结果显示,不同窃电比例以及 KPCA 对特征向量降维后的维数下,算法模型对窃电的检测性能均表现较好。

参 考 文 献

[1] TIAN L, XIANG M. Abnormal Power Consumption Analysis Based on Density-based Spatial Clustering of Applications with Noise in Power Systems[J]. Automation of Electric Power Systems, 2017, 41(5): 64-70.

[2] WANG G L, ZHOU G L, ZHAO H S, et al. Fast Clustering and Anomaly Detection Technique for Large-scale Power Data Stream[J]. Automation of Electric Power Systems, 2016, 40(24): 27-33.

[3] FAHIM M, SILLITTI A. Analyzing Load Profiles of Energy Consumption to Infer Household Characteristics Using Smart Meters[J]. Energies, 2019, 12: 169-173.

[4] LEANDRO A P J, CAIO C O R, RODRIGUES D, et al. Unsupervised non-technical losses identification through optimum-path forest[J]. Electric Power Systems Research, 2016, 140: 413-423.

[5] ZANETTI, M, JAMHOUR E, PELLEZZI M, et al. A tunable fraud detection system for advanced metering infrastructure using short-lived patterns[J]. IEEE Transactions on Smart grid, 2019, 10(1): 830-840.

[6] MUNIZ C, FIGUEIREDO K, VELLASCO M, et al. Irregularity detection on low tension electric installations by neural network ensembles[C]// 2009 International Joint Conference on Neural Networks. IEEE, 2016: 2176-2182.

[7] COSTA B C, LA ALBERTO B, PORTELA A, et al. Fraud detection in electric power distribution networks using an ann-based knowledge-discovery process[J]. International Journal of Artificial Intelligence & Applications, 2019, 4(6): 17.

[8] LIN J N, CHENG Z H, LIN B X. Study on identification method of stolen electricity based on MEA-BP[J]. Electronic Design Engineering, 2021, 29(11): 175-180.

[9] SPIRIĆ J V, STANKOVIĆ S S, BDOČIĆ M, et al. Using the rough set theory to detect fraud committed by electricity customers[J]. International Journal of Electrical Power & Energy Systems, 2014, 62(1): 727-734.

[10] YANG X L, TAO X F, XIONG X, et al. Detection Method for Electricity Theft Based on Deep Forest Algorithm[J]. Smart Power, 2019, 47(10): 85-92.

[11] CAI J H, WANG K, DONG K, et al. Power user stealing detection based on DenseNet and random forest[J]. Journal of Computer Applications, 2021, 41(S1): 75-80.

[12] CIESLAK D A, CHAWLA N V, STRIEGEL A. Combating imbalance in network intrusion datasets[C]// IEEE International Conference on Granular Computing. 2006: 732-737.

[13] ZHAO Z X, WANG G L, LI X D. An Improved SVM Based Under-Sampling Method for Classifying Imbalanced Data[J]. Acta Scientiarum Naturalium Universitatis Sunyatseni, 2012, 51(6): 10-16.

[14] YANG J, YAN X F, ZHANG D P. Cost-sensitive Software Defect Prediction Method Based on Boosting[J]. Computer Science, 2017, 44(8): 176-180.

[15] LIU X Y, WU J, ZHOU Z H. Exploratory under sampling for class-imbalance learning[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2009, 39(2): 539-550.

[16] CHEN S Z, ZHU J P, YOU T G. Study on Unbalanced Customer Loss Based on SMOTERF Algorithm[J]. Journal of Mathematics in Practice and Theory, 2019, 1(9): 204-210.

[17] BARUA S, ISLAM M M, YAO X, et al. MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(2): 405-425.

[18] LIU X Y, WANG S T, ZHANG M L. Transfer synthetic oversampling for class-imbalance learning with limited minority class data[J]. Frontiers of Computer Science, 2019, 13(5): 406-415.

[19] DEL RÍO S, LÓPEZ V, BENÍ-TEZ J M, et al. On the use of MapReduce for imbalanced big data using Random Forest[J]. Information Sciences, 2014, 12(1): 235-239.

[20] ZHANG M, HU X H, WU J X. Imbalanced Data Processing Algorithm Based on Mixed Sampling[J]. Computer Engineering and Applications, 2019, 55(17): 68-75.

[21] MA J J, PAN Q, LIANG Y, et al. Object Detection Based on Improved Grassberger Entropy Random Forest Classifier[J]. Chinese Journal of Lasers, 2019, 46(7): 238-246.



QUE Hua-kun, born in 1986, senior engineer. His main research interests include metering automation and charging strategy.



ZENG Wei-liang, born in 1986, Ph.D, associate professor. His main research interests include routing problem in complex network, traffic simulation and big data visualization for smart city.