引用格式:刘建锋,梅智聪,刘梦琪,等.基于改进旋转森林算法的窃电检测研究[J].电力科学与技术学报,2024,39(1):93-104.
Citation:LIU Jianfeng,MEI Zhicong,LIU Mengqi,et al.Research on electricity theft detection based on improved rotation forest algorithm[J].Journal of Electric Power Science and Technology,2024,39(1):93-104.

基于改进旋转森林算法的窃电检测研究

刘建锋,梅智聪,刘梦琪,周　海,董倩雯
(上海电力大学电气工程学院,上海 200090)

摘　要:如何准确检测出用户侧窃电行为是长期存在于各供电公司一个难点,传统的窃电检测方案均存在一定的局限性。针对窃电检测领域正负类样本高度不平衡,且单一分类模型表现不佳的问题,提出一种基于改进旋转森林算法的窃电检测方法。旋转森林算法采用主成分分析(principal component analysis,PCA)进行特征提取,利用原始训练集的所有主成分训练每个基分类器。在经典的旋转森林算法基础上,使用改进合成少数类过采样(synthetic minority oversampling technique,SMOTE)算法平衡样本子集中的正负类样本;使用Bagging算法中的Bootstrap抽样对训练子集进一步抽样;按准确率对基分类器进行选择性集成等3个方面的改进。算例使用华东某地区实际用户数据,结果表明所提窃电检测方法对比单一分类模型和现有集成学习策略,在多项评价指标下均取得更好的效果。

关　键　词:窃电检测;集成学习;改进SMOTE算法;旋转森林;特征工程
DOI:10.19781/j.issn.1673-9140.2024.01.009　　**中图分类号**:TM731　　**文章编号**:1673-9140(2024)01-0093-12

Research on electricity theft detection based on improved rotation forest algorithm

LIU Jianfeng, MEI Zhicong, LIU Mengqi, ZHOU Hai, DONG Qianwen
(College of Electrical Engineering, Shanghai University of Electric Power, Shanghai 200090, China)

Abstract: Detecting user-side electricity theft accurately has long been a challenge for power supply companies, with traditional theft detection methods having certain limitations. Addressing the highly imbalanced positive and negative samples in the field of theft detection, and the poor performance of single classification models, this study proposes a theft detection method based on an improved Rotation Forest algorithm. The Rotation Forest algorithm uses Principal Component Analysis (PCA) for feature extraction, training each base classifier with all principal components of the original training set. Building upon the classical Rotation Forest algorithm, improvements are made in three aspects: balancing the positive and negative samples in the subset using the Synthetic Minority Oversampling Technique (SMOTE) algorithm, further sampling the training subset using Bootstrap sampling in the Bagging algorithm, and selectively integrating base classifiers based on accuracy. A case study using actual user data from a region in East China demonstrates that the proposed theft detection method achieves better results in multiple evaluation metrics compared to single classification models and existing ensemble learning strategies.

Key words: electricity theft detection; ensemble learning; improved SMOTE algorithm; rotation forest; feature engineering

用户的恶意窃电是造成电能非技术性损失(non-technical loss,NTL)的最主要因素[1],不仅给供电公司造成了巨额的经济损失,还给电力系统的运行安全带来了巨大的挑战。随着智能电表的普及和高级量测体系(advanced metering infrastructure,AMI)的建立,一方面窃电方式从物理手段逐

收稿日期:2022-12-21;**修回日期**:2023-04-27
基金项目:国家自然科学基金(61873159)
通信作者:梅智聪(1998—),男,硕士研究生,主要从事电力系统窃电检测方向的研究;E-mail:1762860273@qq.com

渐升级为信息手段[2]；另一方面15/30 min一次的用电信息采集为窃电检测带来了契机。

目前，窃电检测研究主要集中于5类方法：基于状态估计、基于博弈论、基于线损电量归因、基于聚类和基于分类的方法[3]。基于状态估计的方法通过判断用户的计量数据与配电网的潮流约束是否相矛盾展开窃电检测，该方法需要详细的配电网络拓扑结构和参数，在实际工程应用中存在较强的局限性[4-5]。基于博弈论的方法分析供电公司和窃电者之间的博弈[6]，对供电公司和用户双方的决策行为进行建模，但是忽视了用户多样的用电习惯，且尚处于理论阶段[3]。基于线损电量归因的方法意在分析用户用电情况与线损电量之间的关系[7-9]，判断对线损电量和波动影响显著的用户为窃电用户，但是该方法在用户数量较多时难以有效找到窃电用户与线损电量间的对应关系，且在华东某地供电公司的实际应用中发现其准确率较低。基于聚类的方法采用无监督学习对用电特征指标聚类成不同的组别，将结果远离聚类簇的用户判别为异常用户[10-11]，但是相当行业的用户用电行为并不连续，聚类方法的误报率通常远高于分类方法[12]。

基于分类方法的窃电检测采用有监督学习，根据选择的特征指标将用户分为正常用户和窃电用户两类[13-18]。对比基于聚类的窃电检测，前者使用有标签的样本进行训练，有利于提高检测精度，但由于窃电用户在实际中占比很小，存在数据类别高度不平衡的问题。多数机器学习算法在不平衡数据集中并不能取得良好的训练效果，仅以分类准确率(accuracy，ACC)为目标的算法其实际表现远差于指标表现[19-20]。文献[21]认为集成分类器在样本类不平衡中的条件下要优于单一分类器模型，实验还指出复杂的集成方法有时表现不如简单的集成方法。将标准的集成学习方法与数据层的不平衡处理技术相结合可以有效提升模型表现。文献[22]提出基于多异学习器融合Stacking集成学习的窃电检测，提高了分类器的多样性，增强了模型的泛化能力；文献[23]提出基于Bagging二次加权集成的孤立森林窃电检测算法，按照孤立类间相似度最低准则优选特征顺序建立模型，结果优于异质集成学习。上述研究在集成学习的基础上从分类器选择，特征优选等方面提升模型分类性能，但仍存在难以兼顾分类器的多样性和单个分类器准确性；模型复杂度过高调参困难；训练样本数量要求较多等缺陷。

旋转森林(rotation forest，RoF)算法[24]是Rodriguez等提出的一种集成学习方法，对比目前已经应用于窃电检测领域的集成学习算法如AdaBoost、Bagging、随机森林(random forest，RF)、Stacking等集成策略，RoF算法能够兼顾分类器的多样性和准确性，且在样本数量不大的条件下仍有出色的表现。本文将经典的旋转森林算法应用于窃电检测，并在此基础上提出基于改进旋转森林算法的窃电检测方法：在数据方面，将SMOTE算法和K-means聚类算法相结合，解决了数据高度不平衡的问题；在基分类器方面，从基分类器的多样性和基分类器的准确性2个角度提升了模型的窃电检测性能。另外，为了避免仅从用电量方面设计特征指标可能存在信息不足的缺陷，还增加了电压、电流和功率因数方面的特征指标。

1 改进旋转森林算法

1.1 旋转森林算法

RoF算法使用PCA对特征轴进行旋转，决策树算法对于特征轴的旋转较为敏感且能保证旋转后的精确度，因此RoF选取决策树作为基分类器。由此算法得名旋转森林。

1.1.1 主成分分析

RoF算法生成各基分类器的训练子集前，采用PCA对原训练集进行特征提取和特征降维。数据集$\boldsymbol{X}=[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N]^{\mathrm{T}}$包含$N$个样本，每个样本$\boldsymbol{X}_i$由$n$个特征组成$\boldsymbol{X}_i=[x_{i1},x_{i2},\cdots,x_{in}]$，对$\boldsymbol{X}_i$进行线性变换，得到的结果表示为$\boldsymbol{Y}_i=(y_{i1},y_{i2},\cdots,y_{in})$，过程如下：

$$\begin{cases} y_{i1}=l_{11}x_{i1}+l_{12}x_{i2}+\cdots+l_{1n}x_{in}=\boldsymbol{L}_1{}^{\mathrm{T}}\boldsymbol{X}_i \\ y_{i2}=l_{21}x_{i1}+l_{22}x_{i2}+\cdots+l_{2n}x_{in}=\boldsymbol{L}_2{}^{\mathrm{T}}\boldsymbol{X}_i \\ \vdots \\ y_{in}=l_{n1}x_{i1}+l_{n2}x_{i2}+\cdots+l_{nn}x_{in}=\boldsymbol{L}_n{}^{\mathrm{T}}\boldsymbol{X}_i \end{cases} \tag{1}$$

矩阵形式表示如下：

$$\boldsymbol{Y}_i=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{X}_i \tag{2}$$

其中，$\boldsymbol{L}=(\boldsymbol{L}_1,\boldsymbol{L}_2,\cdots,\boldsymbol{L}_n)$。

设一组变量$y_{i1},y_{i2},\cdots,y_{im}$，其中$m\leqslant n$，在该变量相互独立的前提下求矩阵$\boldsymbol{L}$，使得下式值达到最大：

$$D(y_{ij})=D(\boldsymbol{L}_j{}^{\mathrm{T}}\boldsymbol{X}_i)=\boldsymbol{L}_j{}^{\mathrm{T}}\sum\boldsymbol{L}_j,j=1,2,\cdots,m \tag{3}$$

满足条件的变量$y_{i1},y_{i2},\cdots,y_{im}$分别称为原始变量$x_{i1},x_{i2},\cdots,x_{in}$的第1主成分，第2主成分，…，第$m$

主成分。

在使用PCA时，RoF算法保留了所有的成分，防止某些有判断信息但方差较小的成分被忽略。

1.1.2　旋转森林的步骤

图1为RoF算法流程，其中L为基分类器的数量。首先根据特征维数n，将分类器$\boldsymbol{D}_i$中的训练集$\boldsymbol{X}_i$分为K个子集，为了提高分类器的多样性，这K个子集一般彼此不相交，每个子集包含$M=n/K$个特征，$\boldsymbol{X}_{i,j}$，$\left(i\in L, j\in K\right)$表示用来训练第$i$个基分类器的第$j$个特征子集。为避免训练出的多个分类器系数一样，旋转森林算法对$\boldsymbol{X}_{i,j}$进行Bootstrap采样，得到$\boldsymbol{X}'_{i,j}$。在$\boldsymbol{X}'_{i,j}$上运行PCA并按照特征顺序调整后得到旋转矩阵$\boldsymbol{R}_i^a$，由此可以构造出每个基分类器的训练子集$\boldsymbol{XR}_i^a$，最后通过投票得到最后的分类结果Y。

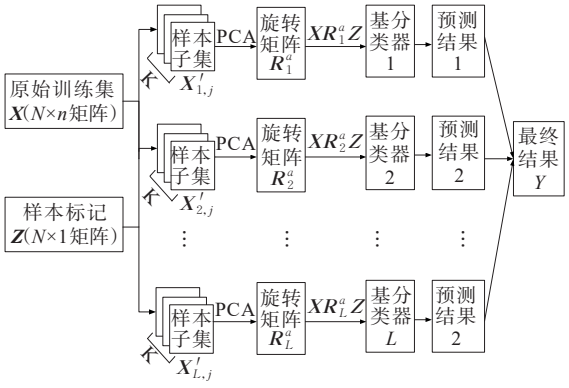

图1　RoF算法流程

Figure 1　Flowchart of RoF algorithm

1.2　改进旋转森林算法介绍

本文在传统RoF算法的基础上，从以下3个方面进行改进：

1）使用SMOTE算法解决窃电检测中的数据高度不平衡问题；

2）对PCA生成的训练子集$\boldsymbol{XR}_i^a$再次进行Bootstrap抽样以增加基分类器训练子集的差异性；

3）在验证阶段以ACC为标准对基分类器评价并筛选，剔除效果较差的基分类器。

1.2.1　窃电检测的数据不平衡

窃电检测是一种二分类问题，不平衡比（imbalanced ratio，IR）表示某一数据集的类分布情况，定义为正（多数）类样本数与负（少数）类样本数之比：

$$I_{\mathrm{R}}=\frac{N^+}{N^-} \tag{4}$$

当$I_{\mathrm{R}}>1$时，该数据集即为不平衡数据集；当$I_{\mathrm{R}}>9$时，该数据集类高度不平衡。相关研究证明I_{R}值越高时，负类样本占比越小，越难从负类样本获取二分类的有效信息，此时算法主要关注正类样本，因此学习难度也越大。对于一个类高度不平衡数据集，即使其负类样本全部被误分，得到的ACC值仍会超过0.9。

国家电网公司发布的数据集[25]包含42 372个电力用户接近3年的用电数据，并给出了正常和窃电用户的标签，其中正常用电用户数量为38 757户，而窃电用户数量为3 615户，其$I_{\mathrm{R}}\approx 10.72$。本文所用华东某地区电网数据集窃电用户数量占比同样很小，其$I_{\mathrm{R}}\approx 10.887$。在工程实践中，窃电检测数据集大多存在类高度不平衡的问题。

1.2.2　基于SMOTE-K的不平衡数据处理方法

SMOTE算法的思想是合成新的少数类样本，对每个少数类样本x_i，从它的最近邻少数类样本中随机选一个样本$\hat{x}_i$，然后在x_i、$\hat{x}_i$之间的连线上随机选一点作为新合成的少数类样本：

$$r_i=x_i+\mathrm{rand}(0,1)\cdot(\hat{x}_i-x_i) \tag{5}$$

SMOTE算法主要有下列3个关键参数：

1）最邻近样本个数T：寻找少数类样本最邻近的T个少数类样本；

2）距离决定因子r：当$r=2$时为欧式距离；

3）过采样倍率$p_{\mathrm{ercentage}}$：合成样本数量占原少数类样本数量的百分比。

然而传统的SMOTE方法也存在一定的局限性：当原始数据集中少数类样本存在噪声时，SMOTE过采样方法会合成新的噪声，进而放大噪声的影响，加大了原本数据特征的学习难度。另外，如果少数类样本散落在较多的多数类样本之间，使用SMOTE方法会造成分类边界模糊的问题，导致分类难度更大[26]。

针对传统SMOTE算法存在的缺陷，本文联合K-means聚类算法，使用一种基于SMOTE-K的不平衡数据处理方法。该方法的思路：首先使用K-means算法将原始数据集中的负类样本进行聚类，将原始数据集分为若干个数据集，计算得到每一个数据集的聚类簇心，再将簇心作为原始负类样本的中心，在此基础上使用SMOTE方法进行过采样，在簇类中心与本数据集的其他负类样本之间的连线上人工合成新的负类样本。

通过K-means聚类，簇类中心能够很好地代表本类数据的负类样本特征，且人工合成的新样本对

比使用传统SMOTE方法得到的样本噪声现象更小，更符合实际，并且能够有效解决传统SMOTE方法会模糊正负类样本间边界的问题。SMOTE-K方法得到的插值样本集$\boldsymbol{x}_{\text{new}}$的计算公式如下：

$$\boldsymbol{x}_{\text{new}}^{i}=\mu_i+\text{rand}(0,1)\cdot\left(y_i^j-\mu_i\right) \tag{6}$$

其中，$\boldsymbol{x}_{\text{new}}^{i}$为第$i$个簇类中心$\mu_i$人工合成的新负类样本集；$y_i^j$为第$i$个聚类簇群中，除簇类中心$\mu_i$之外的其他负类样本；rand(0,1)为0到1之间的一个随机数。

基于SMOTE-K的不平衡数据处理方法的流程如图2所示。

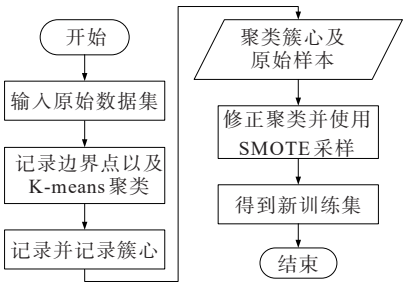

图2 SMOTE-K方法流程

Figure 2 Flowchart of SMOTE-K

1.2.3 Bootstrap二次抽样

Bootstrap抽样技术是Bagging集成学习算法的核心，利用Bootstrap抽样技术可以在样本数量不大的条件下构造出多个彼此不同的训练样本，而且决策树等对于所给训练集敏感的算法，使用Bootstrap抽样会更加有效。本文在传统旋转森林原有的步骤基础上对PCA生成的训练子集$\boldsymbol{XR}_i^a$再次进行Bootstrap抽样，使得基分类器能够获取差异性更大的训练子集，有助于进一步提高基分类器的多样性。

1.2.4 选择性集成

基分类器的数目不仅影响集成学习模型的计算速度，还会直接影响其分类精度。传统的集成学习算法通常是集成所有的基分类器并通过大规模集成的方式以提高泛化能力。目前已有研究表明[27]，优选部分基分类器来构建模型在性能上更优于集成所有基分类器构建的模型。因此通过选择性集成的方式，不仅可以提高模型计算速度、节约存储空间还能提高模型的精度。文献[24]表明，旋转森林算法使用10个基分类器时可以取得较好的成绩，因此原始基分类器数量设置为10，本文以ACC为评价指标对生成的基分类器进行筛选，筛除性能较差的2个基分类器。

1.3 改进旋转森林算法流程

1.3.1 训练阶段

输入：原始训练集$\boldsymbol{X}$，样本标签Z，基分类器数量L，每个基分类器的样本子集数量K，SMOTE算法中的最邻近样本个数T，过采样倍率$p_{\text{ercentage}}$，距离决定因子r，由于窃电检测为二分类问题，所有类标集合$\{z_1,z_2,\cdots,z_t\}$中$t=2$。

1）在得到，分类器$\boldsymbol{D}_i$中的特征子集$\boldsymbol{X}_{i,j}$后选择有放回的Bootstrap采样以75%的采样率得到不平衡样本子集$\boldsymbol{X}'_{i,j}$；

2）对$\boldsymbol{X}'_{i,j}$使用SMOTE算法，将其变为平衡状态，得到平衡后的样本子集表示为$\boldsymbol{X}''_{i,j}$；

3）对$\boldsymbol{X}''_{i,j}$进行PCA提取特征，得到系数矩阵$\boldsymbol{C}_{i,j}$来储存每个主成分系数$c_{i,j}^{(1)},c_{i,j}^{(2)},\cdots,c_{i,j}^{(M_j)}$，其中$M_j\leqslant M$；

4）生成稀疏的矩阵$\boldsymbol{R}_i$为

$$\boldsymbol{R}_i=\begin{bmatrix} c_{i,1}^{(1)},c_{i,1}^{(2)},\cdots,c_{i,1}^{(M_1)} & 0 & \cdots & 0 \\ 0 & c_{i,2}^{(1)},c_{i,2}^{(2)},\cdots,c_{i,2}^{(M_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_{i,K}^{(1)},c_{i,K}^{(2)},\cdots,c_{i,K}^{(M_K)} \end{bmatrix} \tag{7}$$

按照$\boldsymbol{X}$中的特征顺序重新排列矩阵$\boldsymbol{R}_i$的列，得到旋转矩阵$\boldsymbol{R}_i^a(N\times n)$，$\boldsymbol{D}_i$的训练子集为$\boldsymbol{X}'_i=\boldsymbol{XR}_i^a$；

5）对上一步中的训练子集$\boldsymbol{X}'_i$再次以75%的比例进行Bootstrap抽样，得到$\boldsymbol{D}_i$最终的训练子集$\boldsymbol{X}''_i$；

6）重复上述步骤，构造并训练L个决策树基分类器。

1.3.2 验证阶段

以ACC为标准，对L个基分类器进行评价，筛选出精度最好的L_{f}个分类器作为改进RoF的基分类器。

1.3.3 测试阶段

给定测试样本$\boldsymbol{x}_{\text{un}}$，基分类器$D_i$预测该样本类别为$z_t$，($t=1,2$)的概率表示为$d_{i,j}(\boldsymbol{x}_{\text{un}}\boldsymbol{R}_i^a)$，计算由$L_{\text{f}}$个分类器集成的改进RoF算法判断该样本类别为$z_t$的置信度表示如下：

$$\mu_t(\boldsymbol{x}_{un})=\frac{1}{L_{\text{f}}}\sum_{i=1}^{L_t}\boldsymbol{d}_{i,t}(\boldsymbol{x}_{\text{un}}\boldsymbol{R}_i^a),t=1,2 \tag{8}$$

最终判断测试样本x_{un}为置信度最高的类别。改进RoF算法的流程如图3所示。

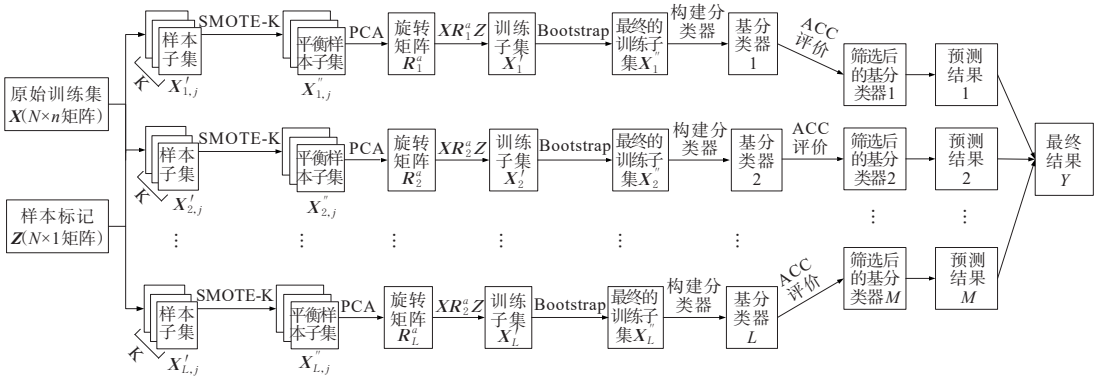

图 3　改进 RoF 算法流程

Figure 3　Flowchart of improved RoF algorithm

2　窃电行为分析及特征构建

2.1　窃电行为分析

用户窃电的核心目标是减少或篡改用电数据以减少需要缴纳的电费。结合分时电价标准，其目标可表示为

$$\sum_{i=1}^{t}P_i x_i' \leqslant \sum_{i=1}^{t}P_i x_i \tag{9}$$

式中，t为每天的电价总段数；P_i为第i时段的单位电价；x_i'为第i时段用户篡改后智能电表上记录的用电量；x_i为第i时段用户实际的用电量。

侵入式干扰计量装置的窃电方法包括：分压法、分流法、移相法和扩差法；非侵入式干扰计量装置的窃电法包括：绕越接线、数字编程和网络攻击等。根据现有的窃电手法，文献[28]总结了6种窃电模式，将窃电的电量变化概括为表1中的6类。

表 1　不同窃电模式电量变化

Table 1　Electricity changes with different electricity theft modes

类型	模式
1	$x_i'=\alpha x_i, 0<\alpha<1$
2	$x_i'=\begin{cases} x_i, & x_i \leqslant \delta \\ \delta, & x_i > \delta \end{cases}$
3	$x_i'=\begin{cases} 0, & i_1 < i < i_2 \\ x_i, & \text{其他} \end{cases}$
4	$x_i'=\max\{x_i-\beta, 0\}$
5	$x_i'=x_{t-i}$
6	$x_i'=\text{mean}(\tilde{x})$

第1种模式将t时段的用电数据以一个固定的比例α缩减；第2种模式设置了一个阈值δ，当用电数据高于此阈值时就将用电数据固定为δ；第3种模式将时间段(t_1, t_2)内的用电数据记为0；第4种模式同样设置了一个阈值β，将用电数据与此阈值做差，取差值和0两者的最大值；第5种模式利用不同时段电价不同，对用电曲线进行移峰以减少电费；第6种模式中同样利用单位电价的不同，将1天中每个时段的用电记录记为1天的平均值。

2.2　窃电检测特征构建

本文选择4 612个用户，时长为2年的用电数据。数据的采样间隔为15 min，数据中包含用户ID、采样时间点、用电量、三相电压/电流、功率因数等信息。

由于用户进行窃电时在用电数据上最直接的体现就是电量变化，再结合前文分析的不同窃电模式下的电量变化，首先围绕用电量进行特征指标设计。用户每天的用电量数据$(x_1, x_2, \cdots, x_{96})$直接反应其用电行为，该数据作为基础特征。此外，本文分别将1周、1个月、1个季度3种周期作为切分粒度，使用滑动窗口技术提取用户的时序特征，包括峰时耗电量、日最大负荷、日最小负荷、日平均负荷，标准差[29]。

相似度特征对于用户的异常用电行为较为敏感，因此本文采用皮尔逊相关系数(pearson correlation coefficient，PCC)提取了用户用电量的波动特征和周期特征。波动特征描述用户当天用电量曲线与前6天用电量曲线之间的相似程度，表达式为

$$\rho_{\text{d}}=\frac{\text{cov}(X_{\text{d}}, X_{\text{d-}i})}{\sigma_{X_{\text{d}}}\sigma_{X_{\text{d-}i}}}=\frac{E(X_{\text{d}}X_{\text{d-}i})-E(X_{\text{d}})E(X_{\text{d-}i})}{\sqrt{E(X_{\text{d}}^2)-E^2(X_{\text{d}})}\sqrt{E(X_{\text{d-}i}^2)-E^2(X_{\text{d-}i})}} \tag{10}$$

式中，X_{d}为当天的96点用电量曲线；$X_{\text{d-}i}$为i天前的用电量曲线，$i\in(1, 2, \cdots, 6)$；ρ_{d}为X_{d}和$X_{\text{d-}i}$之间的PCC系数。

周期特征描述4周之间的用电量曲线相似程度，表达式为

$$\rho_{\mathrm{w}}=\frac{\mathrm{cov}(X_{\mathrm{w}},X_{\mathrm{w}\text{-}i})}{\sigma_{X_{\mathrm{w}}}\sigma_{X_{\mathrm{w}\text{-}i}}}=\frac{E(X_{\mathrm{w}}X_{\mathrm{w}\text{-}i})-E(X_{\mathrm{w}})E(X_{\mathrm{w}\text{-}i})}{\sqrt{E(X_{\mathrm{w}}^2)-E^2(X_{\mathrm{w}})}\sqrt{E(X_{\mathrm{w}\text{-}i}^2)-E^2(X_{\mathrm{w}\text{-}i})}} \tag{11}$$

式中，X_{w}为1周用电量曲线；$X_{\mathrm{w}\text{-}i}$为i周前的用电量曲线，$i\in(1,2,3)$；ρ_{w}为X_{w}和$X_{\mathrm{w}\text{-}i}$之间的PCC系数。

图4为一个正常用户1周用电量的PCC矩阵，矩阵描述了该1周内用电行为的相似程度，相似程度越高，相关系数值就越接近于1，可见1个正常用户每天用电相似度较高。

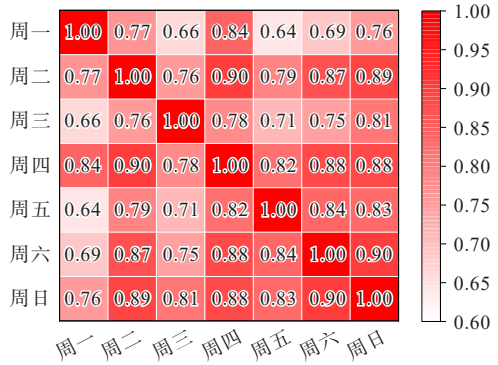

图4 正常用户1周用电量PCC矩阵

Figure 4 Seven-day PCC matrix of electricity consumption for normal user

图5为某窃电用户从周四开始有窃电行为的1周用电量的PCC矩阵。可以看出，该用户本周内前3天之间的相关系数较高，而后4天与前3天的相关系数较低，甚至有负相关的情况，与正常用户相比区别显著。而该用户后4天之间的相关系数又较高，这是该用户后4天采用了相同窃电模式所造成的。

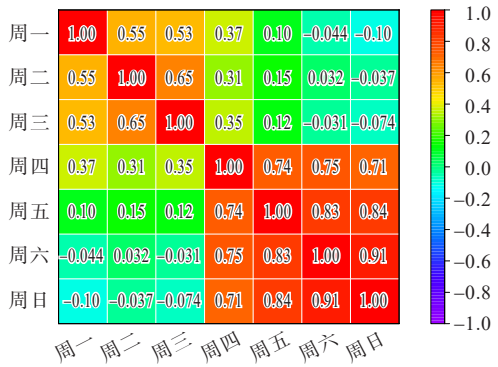

图5 窃电用户一周用电量PCC矩阵

Figure 5 Seven-day PCC matrix of electricity consumption for user with electricity theft

图6为1个正常用户4周的用电量PCC矩阵，可以看出正常用户周用电量数据之间存在很强的相关性，大部分的PCC值都超过了0.8，可见用户用电存在周期特征。

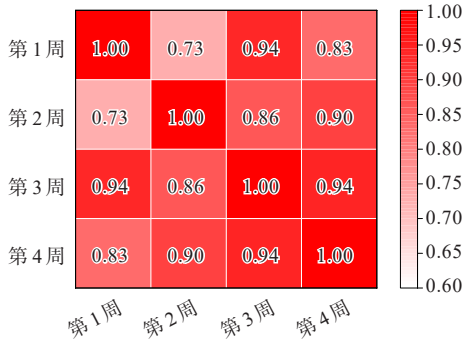

图6 正常用户4周用电量PCC矩阵

Figure 6 Four-week PCC matrix of electricity consumption for normal user

图7为某窃电用户4周的用电量PCC矩阵，该用户第2周开始有窃电行为，反映到此矩阵中则是周用电数据相关性较低，所有的PCC值均小于0.5，甚至出现很多负值，与正常用户相比区别显著。

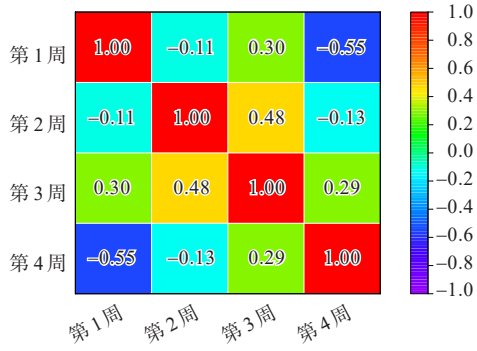

图7 窃电用户4周用电量PCC矩阵

Figure 7 Four-week PCC matrix of electricity consumption for user with electricity theft

由于不同类型、不同行业的用户用电习惯繁杂，在实际中，导致用户用电量突变的原因是多种多样的，例如工厂生产以订单为准，用电量本身就不平稳，因此仅围绕用电量设计特征指标容易出现误检的情况。为提升模型性能，降低误报率，本文额外增加了电流三相不平衡、电压三相不平衡、功率因数越限和功率因数突变作为特征指标。

3 基于改进旋转森林算法的窃电检测模型

3.1 数据预处理

原始数据集中包含缺失值，且在某些数据后带

有表记异常的标签将其视为错误值，造成这种情况的原因有：智能电表故障、测量数据传输不可靠、计划外的系统维护和存储问题等[30]。因此，需要对数据进行预处理。

利用插值方法恢复缺失值：

$$f(x_i)=\begin{cases}\dfrac{x_{i-1}+x_{i+1}}{2}, & x_i\in \mathrm{NaN}, x_{i-1}, x_{i+1}\notin \mathrm{NaN}\\ 0, & x_i\in \mathrm{NaN}, x_{i-1}\ \mathrm{or}\ x_{i+1}\in \mathrm{NaN}\\ x_i, & x_i\notin \mathrm{NaN}\end{cases} \tag{12}$$

式中，x_i为一个序列中第i时刻的数值，若x_i为空或非数字字符，则将其表示为NaN。

对于错误值，采用“Three-sigma”法则进行恢复，其表达式如下：

$$f(x_i)=\begin{cases}\mathrm{avg}(\boldsymbol{X})+2\cdot\mathrm{std}(\boldsymbol{X}), & x_i>\mathrm{avg}(\boldsymbol{X})+2\cdot\mathrm{std}(\boldsymbol{X})\\ x_i, & \mathrm{otherwise}\end{cases} \tag{13}$$

式中，$\boldsymbol{X}$为每日96个x_i数据组成的向量；avg(X)为$\boldsymbol{X}$的平均值；std(X)为$\boldsymbol{X}$的标准差。

3.2　基于改进旋转森林算法的窃电检测流程

将电量、电压、电流和功率因数数据进行数据预处理后构建特征。数据样本按照8∶2的比例划分成训练集和测试集，其中训练集利用K折交叉验证划分训练集和验证集，基于改进RoF算法的窃电检测流程如图8所示。

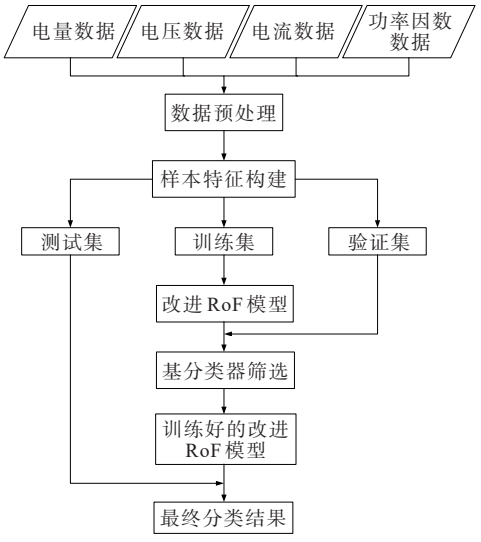

图8　改进RoF算法窃电检测流程

Figure 8　Flowchart for electricity theft detection based on improved RoF algorithm

3.3　多样性度量

分类器的多样性是衡量1个集成学习模型的重要指标[31]，分类模型想要在集成后提升性能，需要有多样化的基分类器。若每个分类器完全相同，集成再多的分类器显然也不能提升1个模型的性能。本文选取成对多样性度量方法：Q统计对改进RoF模型选择性集成前后的分类器多样性进行度量。

假设1个集成算法中包含A个分类器，D_i和D_j为2个不同的分类器($i,j=1,2,\cdots,A$; $i\neq j$)，对M个分类样本，将D_i和D_j分类正确与否情况列成如表2所示，样本总数M分成D_i和D_j都正确分类的M^{11}；D_i和D_j都错误分类的M^{00}；D_i正确分类，D_j错误分类的M^{10}；D_i错误分类，D_j正确分类的M^{01} 4类。

表2　2个分类器的预测结果统计

Table 2　Prediction statistics for 2 classifiers

分类器状态	D_j正确	D_j错误
D_i正确	M^{11}	M^{10}
D_i错误	M^{01}	M^{00}

2个分类器之间的Q统计值为

$$Q_{ij}=\frac{M^{11}M^{00}-M^{10}M^{01}}{M^{11}M^{00}+M^{10}M^{01}} \tag{14}$$

若D_i和D_j对每个样本都是同时分类正确或分类错误，则$M^{10}=M^{01}=0$，此时$Q_{ij}=1$，D_i和D_j之间的多样性程度最低；若D_i和D_j对每个样本分类结果都相反，则$M^{11}=M^{00}=0$，此时$Q_{ij}=-1$，D_i和D_j之间的多样性程度最高。若D_i和D_j是2个独立统计的分类器，则Q_{ij}的期望值为0。对于多分类器$D_1,D_2,\cdots,D_A$，Q统计平均值Q_{av}计算为

$$Q_{\mathrm{av}}=\frac{2}{A(A-1)}\sum_{i=1}^{A-1}\sum_{j=i+1}^{A}Q_{ij} \tag{15}$$

3.4　评价指标

评价指标是衡量分类结果的关键因数，混淆矩阵表达了二元分类问题各种分类结果，如表3所示。其中T_{P}和T_{N}分别表示被正确划分的正常用户和异常用户的数量，而F_{P}和F_{N}分别表示被错误划分的正常用户和异常用户的数量。

表3　混淆矩阵

Table 3　Confusion matrix

用户	检测为异常	检测为正常
实际异常	T_{P}	F_{N}
实际正常	F_{P}	T_{N}

选择ACC、受试者工作特征曲线下面积(area under the ROC,AUC)和误报率(false positive rate,FPR)3个指标作为模型的评价指标。

1) ACC为检测结果和实际结果一致的用户与所有用户之比,表示模型的分类准确度:

$$A_{CC}=\frac{T_P+T_N}{T_P+F_N+F_P+T_N} \tag{16}$$

2) 受试者工作特征(receiver operating characteristic,ROC)曲线是不平衡分类中广泛应用的综合指标,描述了命中率(true positive rate,TPR)和FPR 2个基础指标之间的关系,如图9所示。

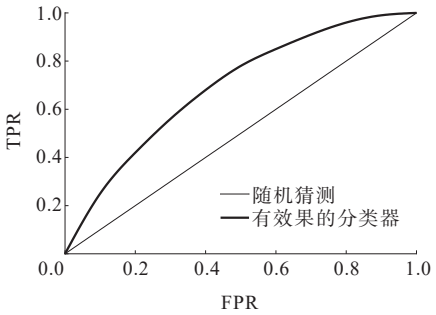

图9 ROC曲线示例

Figure 9 Example graph of ROC curves

命中率T_{PR}表示负类样本被正确分类的比例,误报率F_{PR}表示正类样本被错误分类的比例,命中率和误报率的表达式分别为

$$T_{PR}=\frac{T_P}{T_P+F_N} \tag{17}$$

$$F_{PR}=\frac{F_P}{F_P+T_N} \tag{18}$$

由图9可知,ROC曲线越凸,越靠近(0,1)则表示分类器效果越好,泛化能力越强。为避免2个分类模型的ROC曲线交叉,无法直观判断二者的性能,选择ROC曲线下面积A_{UC}来定量表示模型的性能优劣,当$A_{UC}=0.5$表示随机猜测,而$A_{UC}=1$代表1个完美分类器,因此A_{UC}的取值范围是[0.5,1],其值越接近1表示该分类模型的性能越好,泛化能力越强。

3) 实际工程中将正常用户误检为窃电用户的后果远甚漏检少量窃电用户,因此选择误报率(F_{PR})作为模型的评价指标,其含义为将实际正常却被误检为异常的用户与实际正常的用户总数之比。

4 算例分析

4.1 数据集和参数设置

本文选用的原始数据来自华东某地区,包含4 612个电力用户从2020年3月1日至2022年2月28日的数据,其中正常用电用户数量为4 224户,窃电用户数量为388户,窃电用户标的签记为1,正常用户的标签记为0。该数据集的数据采样间隔为15 min,1天总共96个数据点,在数据集中还包括用户ID、采样时间点、用电量、三相电压/电流、功率因数等信息。实验平台为数据挖掘软件Weka,实验中用于对比的各算法的迭代次数为20次,Adaboost、Bagging、RoF和改进RoF算法中的基分类器均采用C4.5算法,RF算法则直接使用自带的CART算法。Adaboost算法的基分类器个数为10,Bagging和RF算法的基分类器个数为40,其余参数大部分按照Weka软件中的推荐或者默认设置。

改进RoF算法的关键参数设置如表4所示。

表4 改进RoF算法关键参数设置

Table 4 Key parameters for improved RoF algorithm

算法	关键参数
C4.5	剪枝=Ture,置信度=0.25
SMOTE	最邻近样本个数T=2,过采样倍率percentage=平衡,距离决定因子=2
改进RoF	原始基分类器数量=10,迭代次数=20,交叉验证倍率=8

4.2 基分类器筛选

以总体分类准确率为评价指标对原始的10个基分类器进行排序、筛选,删除总体准确率较差的基分类器有助于提高最终模型的分类精度。将10个原始基分类器分别命名为b_0、b_1、…、b_9。

8折交叉验证将训练集均分8份,对每份样本子集从1到8进行编号,按照编号依次选择其中1份作为验证集,剩下7份作为训练集,记录每轮每个基分类器的ACC,最终计算每个基分类器的总体分类准确率,结果如图10所示。其中分类器b_2和分类器b_5的ACC指标评价较差,因此删除分类器b_2和b_5,选择性集成之后模型最终有8个基分类器。

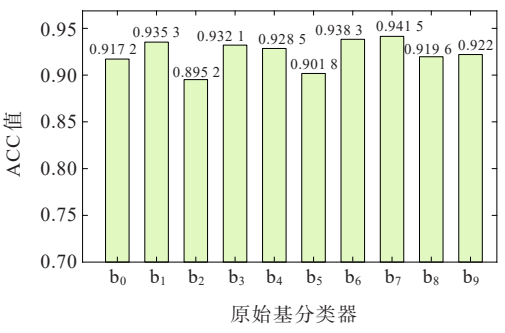

图10 10个原始基分类器的ACC评价

Figure 10 ACC of 10 raw base classifiers

4.3　选择性集成前后模型的Q统计值分析

表5记录了基分类器和整个模型在选择性集成前后Q统计平均值及其变化情况。

表5　选择性集成前后Q统计平均值

Table 5　Average of Q statistics before and after selective ensemble

分类器	Q统计平均值			
	选择性集成前		选择性集成后	
	基分类器	模型	基分类器	模型
b_0	0.834 8	0.901 1	0.797 4	0.877 7
b_1	0.916 5		0.884 2	
b_2	0.841 3		—	
b_3	0.866 7		0.841 3	
b_4	0.912 4		0.896 2	
b_5	0.923 2		—	
b_6	0.941 6		0.917 8	
b_7	0.898 9		0.868 0	
b_8	0.933 5		0.903 8	
b_9	0.942 1		0.912 7	

由表5可知，选择性集成前模型总体的Q统计平均值为0.901 1，剔除基分类器b_2和b_5之后剩余8个基分类器的Q统计平均值均有所下降，且最终模型总体的Q统计平均值降为0.877 7，即选择性集成不仅能提高模型的分类精度，还增加了分类器的多样性，增强了模型的泛化能力。

4.4　非用电量特征指标验证

本文在构建特征时设计了电流三相不平衡、电压三相不平衡、功率因数越限和功率因数突变作为非用电量特征指标意在降低窃电检测时的误报率。本节内容将对比输入特征增加上述非用电量特征指标前后各算法的FPR表现情况，实验在使用SMOTE-K算法将数据平衡的条件下进行。

图11统计了输入特征时增加电流三相不平衡、电压三相不平衡、功率因数越限和功率因数突变这几个非用电量特征前后各检测模型的FPR值，可以看出，各模型在增加非用电量特征之后误报率都有不同程度的降低，其中改进RoF增加非用电量特征之后的误报率最低，为5.59%，对比为增加非用电量特征时，其误报率降低了31.91%。在本文后续的实验中，输入特征均选择添加上述非用电量特征。

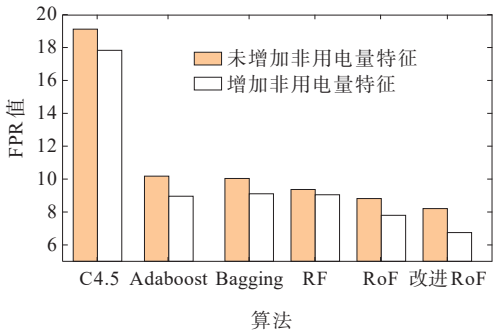

图11　非用电量特征对各算法FPR值的影响

Figure 11　The influence of non-power consumption features on FPR of each algorithm

4.5　各算法在不同过采样倍率下的表现及分析

本文所用SMOTE-K方法在数据层面解决了数据不平衡的问题，并且可以与各种算法相结合。为了验证本文所提基于改进RoF算法的窃电检测方法的有效性，本节实验对比了单分类算法C4.5，集成分类算法：Adaboost、Bagging、RF、RoF以及本文提出的改进RoF算法这几种集成学习算法在不同SMOTE-K过采样倍率条件下的AUC、ACC和FPR指标表现。其中，过采样倍率有4种状态，包括：$p_{ercentage}$＝初始值（未使用SMOTE-K算法增加窃电样本数量）、$p_{ercentage}$＝100（窃电样本的数量增大1倍）、$p_{ercentage}$＝500（窃电样本的数量增大5倍）、$p_{ercentage}$＝平衡（使用SMOTE-K算法后正常样本和窃电样本的数量基本平衡）。不同过采样倍率下各算法AUC值如图12所示。

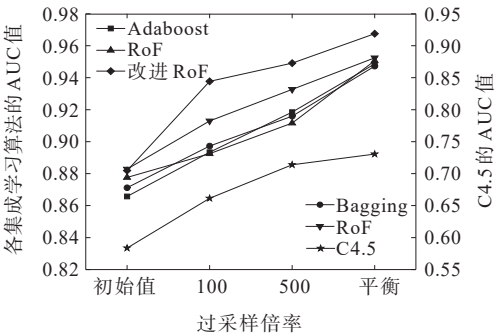

图12　不同过采样倍率下各算法的AUC值

Figure 12　AUC for each algorithm at different percentage

由图12可知，本文所提的基于改进RoF算法的窃电检测方法在AUC指标下能够取得最好的成绩。另外，实验中每一种算法随着过采样倍率的增加，AUC值都有所提升，证明使用SMOTE-K算法平衡数据能够有效提升分类模型的性能。当正常用电用户和窃电用户的数量基本平衡时，改进RoF算法取得了表7中所有情况下的最好成绩，AUC值为

0.967 5。对比单分类C4.5算法的最好成绩，改进RoF算法的AUC值提高了32.43%。传统的集成学习算法中表现最好的是传统RoF算法，其AUC值最好成绩为0.952 4，而改进RoF算法的AUC值在此基础上提高了1.6%。对比使用SMOTE-K算法平衡数据前后，改进RoF算法的AUC值提升了9.7%。

不同过采样倍率下各算法的ACC值如图13所示。可知，本文所提的改进RoF算法在4个过采样状态下都取得了ACC值的最好成绩，其ACC平均值为0.984 9。本文所提算法的ACC平均值比单分类算法C4.5提高了7.6%，比排名第2的传统RoF算法提高了0.77%。另外，每个分类模型的ACC值在SMOTE-K算法的不同过采样倍率下变化不大。

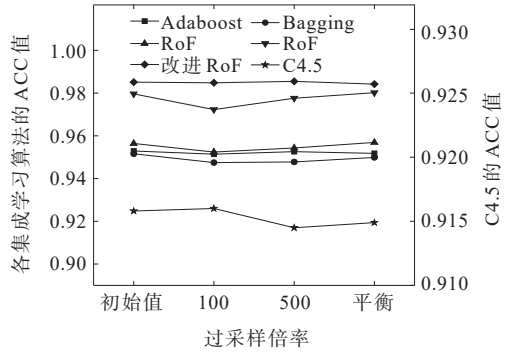

图13 不同过采样倍率下各算法的ACC值

Figure 13 ACC for each algorithm at different percentage

不同过采样倍率下各算法的FPR值如图14所示。可知，随着过采样倍率的增加，各算法的FPR值都有不同程度的降低，即使用SMOTE-K算法处理不平衡数据可以减少把正常用户误判为窃电用户的情况发生。对比各个算法，改进RoF算法在不同过采样倍率下的FPR指标均取得了最好成绩。在使用SMOTE-K算法平衡数据之后，改进RoF算法的FPR值仅为5.59%，比C4.5算法的FPR值下降了55.88%。排名第2的成绩由RoF算法在正负类样本平衡的条件下取得，而改进RoF算法的FPR指标在此基础上降低了5.41%。对比使用SMOTE-K算法平衡数据前后，改进RoF算法的FPR值降低了1.23%。

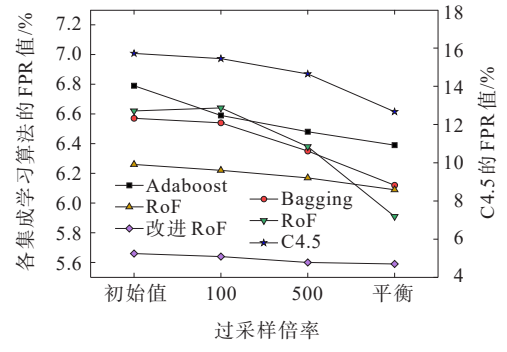

图14 不同过采样倍率下各算法的FPR值

Figure 14 FPR for each algorithm at different percentage

5 结语

本文提出了基于改进旋转森林算法的窃电检测方法，分别从用电量、电压、电流、功率因数4个方面提取出多个特征，在旋转森林算法的基础上使用SMOTE-K算法在克服传统过采样方法副作用的前提下解决了数据不平衡的问题，对训练子集再次使用Bootstrap抽样提高了基分类器的多样性，用选择性集成的方法提高了分类器的准确性。在某地区实际数据对该方法进行验证并对比单一分类算法和传统的集成学习算法，结果表明该方法在AUC、ACC和FPR指标上都具有优越性，在数据平衡时AUC值可以达到96.75%，ACC值可以达到98.42%，FPR值仅为5.59%，能够为供电公司精准找出窃电用户提供有力支持。

参考文献：

[1] 杨志东，丁建武，陈广久，等.基于LightGBM和LSTM模型的电力大数据异常用电检测方法研究[J/OL].电测与仪表：1-7[2024-04-03]. http://kns. cnki. net/kcms/detail/23.1202.TH.20220713.1958.004.html.
YANG Zhidong, DING Jianwu, CHEN Guangjiu, et al. Research on abnormal electricity consumption detection method for power big data based on lightGBM and LSTM models[J]. Electrical Measurement & Instrumentation: 1-7[2023-04-03]. http://kns. cnki. net/kcms/detail/23.1202.TH.20220713.1958.004.html.

[2] 俞林刚，李铭，伍栋文，等.配电网拓扑参数未知场景下中压用户窃电检测方法[J].电网与清洁能源，2023，39(9)：91-100.
YU Lingang, LI Ming, WU Dongwen, et al. An electricity theft detection method of medium voltage users in the case of unknown for distribution network topology parameters[J]. Power System and Clean Energy，2023，39(9)：91-100.

[3] 李双伟，付慧，史明明，等.基于堆叠稀疏自编码器和深度森林的窃电检测模型[J].供用电，2023，40(5)：77-83+99.
LI Shuangwei, FU Hui, SHI Mingming, et al. Electricity theft detection model based on stacked sparse auto encoder and deep forest[J]. Distribution & Utilization, 2023，40(5)：77-83+99.

[4] 徐明杰,赵健,王小宇,等.基于多任务联合模型的居民用电模式分类方法[J].电工技术学报,2022,37(21):5490-5502.
XU Mingjie,ZHAO Jian,WANG Xiaoyu,et al.Residential electricity consumption pattern classification method based on multi-task joint model[J].Transactions of China Electrotechnical Society,2022,37(21):5490-5502.

[5] 鄢仁武,郑杨,林志雄,等.基于改进ECC的含电动汽车用户负荷信息保护方案[J].中国电力,2023,56(1):150-157.
YAN Renwu, ZHENG Yang, LIN Zhixiong, et al. ECC based load information protection scheme for electric vehicle users[J].Electric Power,2023,56(1):150-157.

[6] 张衡,张沈习,程浩忠,等.Stackelberg博弈在电力市场中的应用研究综述[J].电工技术学报,2022,37(13):3250-3262.
ZHANG Heng,ZHANG Shenxi,CHENG Haozhong,et al. A state-of-the-art review on Stackelberg game and its applications in power market[J]. Transactions of China Electrotechnical,2022,37 (13):3250-3262.

[7] 殷涛,薛阳,杨艺宁,等.基于向量自回归模型的高损线路窃电检测[J].中国电机工程学报,2022,42(3):1015-1024.
YIN Tao,XUE Yang,YANG Yining,et al.Electricity theft detection of high-loss line with vector autoregression[J]. Proceedings of the CSEE,2022,42(3):1015-1024.

[8] 陈光宇,徐嘉杰,卢兆军,等.基于相关性度量算法的台区线损异常判断及精准定位[J].电力工程技术,2022,41(4):67-74.
CHEN Guangyu, XU Jiajie, LU Zhaojun, et al. Judgment and precise location of abnormal line loss in station area based on correlation measurement algorithm[J]. Electric Power Engineering Technology,2022,41(4):67-74.

[9] 薛阳,张蓬鹤,杨艺宁,等.基于线损协方差分析的群体性固定比例窃电行为检测方法[J].电力系统自动化,2022,46(13):112-120.
XUE Yang,ZHANG Penghe,Yang Yining,et al.Detection method for group fixed-ratio electricity theft based on covariance analysis of line loss[J]. Automation of Electric Power Systems,2022,46(13):112-120.

[10] 武超飞,孙冲,刘厦,等.基于改进FCM聚类的窃电行为检测[J].电力科学与技术学报,2021,36(6):164-170.
WU Chaofei, SUN Chong, LIU Sha, et al. Detection of stealing electricity energy based on improved fuzzy C-means clustering[J]. Journal of Electric Power Science and Technology,2021,36(6):164-170.

[11] 邹念,魏梅芳,苏盛,等.基于水电关联信息的零电量低压用户窃电检测[J].中国电力,2023,56(12):206-216.
ZOU Nian, WEI Meifang, SU Sheng, et al. Detection of electricity theft by low voltage users with zero power consumption based on water-electricity correlation information[J].Electric Power,2023,56(12):206-216.

[12] 杨艺宁,薛阳,徐英辉,等.基于行业特性的路灯用电异常检测方法[J].电力科学与技术学报,2021,36(3):165-173.
YANG Yining, XUE Yang, XU Yinghui, et al. Sector electricity consumption behavior features based abnormal electricity consumption detection method for street lamps[J]. Journal of Electric Power Science and Technology,2021,36(3):165-173.

[13] ABDULRAHMAN T,MUHAMMAD I,USMAN Z,et al. Robust electricity theft detection against data poisoning attacks in smart grids[J]. IEEE Transactions on Smart Grid,2021,12(3):2675-2684.

[14] 伍栋文,于艾清,俞林刚,等.基于ICS-K-means聚类算法和WNN的有源低压台区线损估算方法[J].智慧电力,2022,50(4):8-14.
WU Dongwen, YU Aiqing, YU Lingang, et al. Line loss estimation method based on ICS-K-means clustering algorithm and WNN for transformer district with DGs[J]. Smart Power,2022,50(4):8-14.

[15] 李景歌,荣娜,陈庆超.基于生成对抗网络的分布式光伏窃电数据增强方法[J].电力科学与技术学报,2022,37(5):181-190.
LI Jingge, RONG Na, CHEN Qingchao. A data augmentation method for distributed photovoltaic electricity theft using generative adversarial network[J]. Journal of Electric Power Science and Technology,2022,37(5):181-190.

[16] SINGH S K,BOSE R,JOSHI A.Entropy-based electricity theft detection in AMI network[J]. IET Cyber-Physical Systems:Theory & Applications,2018,3(2):99-105.

[17] BUZAU M M,TEJEDOR-AGUILERA J,CRUZROMERO P, et al. Detection of non-technical losses using smart meter data and supervised learning[J].IEEE Transactions on Smart Grid,2019,10(3):2661-2670.

[18] MUHAMMAD I, SHAABAN M F, MAHESH N, et al. Deep learning detection of electricity theft cyber-attacks in renewable distributed generation[J]. IEEE Transactions on Smart Grid,2020,11(4):3428-3437.

[19] 刘康,刘鑫,张蓬鹤,等.基于负荷尖峰特征LSTM自编码器的窃电识别方法[J].电力系统自动化,2023,47(2):96-104.
LIU Kang,LIU Xin,ZHANG Penghe,et al.Identification method of electricity theft based on long short-term memory autoencoder with load peak features[J]. Automation of Electric Power Systems, 2023,47(2):96-104.

[20] 程超鹏,彭显刚,曾勇斌,等.相异模型下Stacking集成结构的异常用电用户识别方法[J].电网技术,2021,45(12):4828-4836.
CHENG Chaopeng, PENG Xiangang, ZENG Yongbin, et

al. An abnormal power user recognition method for stacking integrated structures with different models[J]. Power System Technology,2021,45(12):4828-4836.

[21] 刘钊瑞,高云鹏,郭建波,等.基于深度自编码器高斯混合模型的窃电行为检测[J].电力系统保护与控制,2022,50(18):92-102.
LIU Zhaorui,GAO Yunpeng,GUO Jianbo,et al.Abnormal detection of electricity theft using a deep auto-encoder Gaussian mixture model[J].Power System Protection and Control,2022,50(18):92-102.

[22] 游文霞,李清清,杨楠,等.基于多异学习器融合Stacking集成学习的窃电检测[J].电力系统自动化,2022,46(24):178-186.
YOU Wenxia, LI Qingqing, YANG Nan, et al. Electricity theft detection based on multiple different learners fusion by stacking ensemble learning[J]. Automation of Electric Power Systems,2022,46(24):178-186.

[23] 李国成,陆俊,王赟,等.基于Bagging二次加权集成的孤立森林窃电检测算法[J].电力系统自动化,2022,46(2):92-100.
LI Guocheng, LU Jun, WANG Yun. Electricity theft detection algorithm of isolation forest based on bagging secondary weighted ensemble[J]. Automation of Electric Power Systems,2022,46(2):92-100.

[24] RODRIGUEZ J J, KUNCHEVA L I, ALONSO C J. Rotation forest: A new classifier ensemble method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2006,28(10):1619-1630.

[25] YAO D,WEN M,LIANG X,et al.Energy theft detection with energy privacy preservation in the smart grid[J]. IEEE Internet of Things Journal,2019,6(5):7659-7669.

[26] 夏云舒,王勇,周林,等.基于改进生成对抗网络的虚假数据注入攻击检测方法[J].电力建设,2022,43(3):58-65.
XIA Yunshu, WANG Yong, ZHOU Lin, et al. False data injection attack detection method based on improved generative adversarial network[J]. Electric Power Construction,2022,43(3):58-65.

[27] MART NEZ-MU OZ G, HERN NDEZ-LOBATO D, SU REZ A. An analysis of ensemble pruning techniques based on ordered aggregation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 245-259.

[28] MARCELO Z,EDGARD J,MARCELO P,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.

[29] 潘骏,夏祥武,李梁,等.基于关联潮流感知与高斯混合模型的异常用电检测[J].电力建设,2023,44(11):138-148.
PAN Jun,XIA Xiangwu,LI Liang,et al.Abnormal power consumption detection based on associative power flow sensing and Gaussian mixture model[J]. Electric Power Construction,2023,44(11):138-148.

[30] ZHENG Z, YANG Y, NIU X, et al. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids[J]. IEEE Transactions on Industrial Informatics,2018,14(4):1606-1615.

[31] 徐耀松,李佳旺,段彦强.基于相似度机制AdaBoost-DBN的变压器故障层级诊断[J].高压电器,2023,59(6):154-164.
XU Yaosong, LI Jiawang, DUAN Yanqiang. Fault Hierarchical Diagnosis of Transformer Based on AdaBoost-DBN Similarity Mechanism[J]. High Voltage Apparatus,2023,59(6):154-164.