第46卷 第24期 2022年12月25日
DOI：10.7500/AEPS20210731001
电力系统自动化
Automation of Electric Power Systems
Vol. 46 No. 24 Dec. 25，2022

# 基于多异学习器融合Stacking集成学习的窃电检测

游文霞，李清清，杨　楠，申　坤，李文武，吴泽黎
（三峡大学电气与新能源学院，湖北省宜昌市 443002）

**摘要：**针对窃电检测中用户用电数据类别不平衡、采用投票法作为结合策略的集成学习方法无法充分发挥多个不同学习器优势等问题，提出一种利用Stacking集成学习融合多异学习器的模型应用于窃电检测。首先，从影响电量计量的因素出发，根据常见的5种窃电方法模拟6种窃电行为模式；其次，采用合成少数类过采样技术(SMOTE)对不平衡的用电数据进行处理，并利用$K$折交叉验证法对平衡后的训练集进行划分以缓解因重复学习造成的过拟合；然后，使用评价指标和多样性度量优选模型的不同初级学习器和元学习器，构建融合不同学习器优势和差异的Stacking集成学习窃电检测模型；最后，算例对比分析结果表明所提窃电检测模型能有效解决用电数据类别不平衡，充分发挥不同学习器的优势，评价指标良好。

**关键词：**Stacking结合策略；集成学习；窃电检测；合成少数类过采样技术；$K$折交叉验证

## 0　引言

电力的传输和分配涉及技术损耗(technical loss，TL)和非技术损耗(non-technical loss，NTL)，而NTL中绝大多数损失与欺诈和能源盗窃有关[1-2]。窃电通过对用电数据进行恶意的攻击，给供电企业带来了巨大的经济损失[3]。随着供电公司对窃电检测重视程度的增加，传统通过诸如线路窃听或电表篡改之类的物理攻击的检测方法难以有效检测出窃电的行为[4]。同时，智能电表和用电信息采集系统的普及使得越来越多的研究者可以更有效地采集用户用电数据，这是利用机器学习进行窃电检测的基础[5]。

目前，应用于窃电检测的技术主要分为3种，即基于系统状态、基于博弈论和基于分类[6]。其中，基于系统状态的检测技术利用配电网状态估计与用户计量数据之间的矛盾进行窃电检测，但带来了附加的投资[7]；基于博弈论的检测技术根据窃电者和检测者的行为分析相应的博弈均衡，但难以确定参与人的效用水平[8]；基于分类的检测技术根据用户的电量以及用电曲线分布等特征采用数据驱动的方法进行窃电检测，目前已开展了广泛研究[9-17]。

对于窃电检测二分类问题，大部分都采用了单一学习方法[9-13]。而单一学习方法只能从单个角度观测用电数据，检测性能的提升空间有限。为了改善单一学习方法的局限性，近些年来在窃电检测中开展了集成学习方法研究。文献[14]采用日用电量为特征指标，提出一种基于稀疏随机森林(random forest，RF)的用电侧异常检测方法。文献[15]提出了一种采用决策树作为弱分类器的自适应提升(adaptive boosting，AdaBoost)树的窃电检测方法。文献[16]提出了一种特征工程的新框架，在该框架内应用梯度提升机(gradient boosting machine，GBM)算法进行窃电检测。文献[17]提出使用监督学习方法进行非技术损失检测，其中，极限梯度提升(eXtreme gradient boosting，XGBoost)树优于其他分类器。但是，这些集成学习方法一般采用投票法结合相同的学习器，不能体现出不同学习器的差异性。

上述研究为窃电检测领域提供了有效的解决方法，但依旧存在着以下不足：一是采用投票法作为结合策略的集成学习方法无法充分发挥不同学习器的优势；二是用户用电数据集中存在数据类别不平衡问题，导致分类结果出现偏倚。针对以上问题，本文提出一种利用元学习器融合多个不同初级学习器优势和差异的Stacking集成学习方法。首先，采用合成少数类过采样技术(synthetic minority oversampling technique，SMOTE)算法处理类别不平衡的用电数据，实现训练数据样本分布均衡；然后，利用评价指标和多样性度量优选融合的不同初级学习器和元学习器，并采用$K$折交叉验证的方法对训练集进行划分以减小过拟合；最后，使用爱尔兰

收稿日期：2021-07-31；修回日期：2022-01-22。
上网日期：2022-11-22。
国家自然科学基金资助项目(51607104)。

智能电表数据集验证模型的有效性。

## 1　相关理论介绍

### 1.1　SMOTE算法

用户用电数据集大多存在数据倾斜方面的问题，即窃电用户所占比例远低于正常用户。为达到少数类和多数类样本的平衡，提高检测窃电用户的性能，本文采用SMOTE算法进行无重复的新的少数类样本的生成[18]。

供电企业进行窃电检测的目的主要是识别窃电用户，采用SMOTE算法可以增加窃电用户的数量，使正常用户和窃电用户的比例为1∶1。

### 1.2　Stacking集成学习

#### 1.2.1　集成学习的结合策略

集成学习的思想就是利用多个学习器来解决某一问题，使用不同的学习器和不同的结合策略会产生不同的集成学习方法[19]。其中，结合策略是集成学习中最为关键的部分。

针对分类问题，常用的结合策略有投票法和学习法。其中，投票法又包括多数投票法和加权投票法，而它们仅是简单地对学习器的预测结果进行逻辑加工，通过某种特定的方式为学习器寻求权重，并未有效利用数据空间。因此，一种更为强大的结合策略是学习法，即通过另一个学习器进行结合[20]。

#### 1.2.2　Stacking结合策略

Stacking是学习法的典型代表，可利用某一学习器来集成不同学习器的分类结果，其中，不同学习器的多样性通过学习器的差异性来保证。它是一种有层次的集成学习，其层数可自由设置，但从各个领域的研究和应用来看，一般两层结构的Stacking既能强化学习效果又不至于造成模型过于复杂[21-22]。因此，本文以两层的Stacking集成学习为例进行说明。

Stacking第1层中的学习器称为初级学习器，第2层中的学习器称为元学习器。其基本思想是：首先，根据合适的比例，将原始的数据集依次划分为训练集、验证集和测试集；然后，在平衡训练集$D$上，采用$K$折交叉验证法训练不同的初级学习器，将它们的分类结果输入元学习器，而$D$的初始标记作为元学习器的标记，结合起来形成新的训练集来训练元学习器；最后，由元学习器输出最终的分类结果，如图1所示。

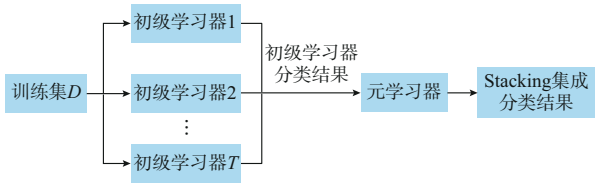


**图1　Stacking结合策略示意图**

**Fig. 1　Schematic diagram of Stacking combination strategy**

## 2　Stacking结合策略下融合多异学习器的窃电检测

用电用户包括正常用户和窃电用户。基于分类的窃电检测机理是利用机器学习中分类问题的相关方法来学习用户历史用电数据中蕴藏的规律，并以该规律来拟合大量未知的用电数据。

用户用电行为和用来判别异常的特征指标项是窃电检测的核心问题。在正常情况下，用户用电所形成的用电曲线分布具有较强的相似性。但是，在实际系统中，居民的用电行为更为多样化，由于住户旅游度假、改换工作、房屋更换租客等情况都可能导致用电行为习惯的突变。因此，本文在分析用户用电行为的基础上，从居民用户一天的用电量特征中提取出最大值、最小值、平均值和标准差这4个综合特征用于辅助模型进行窃电行为的判别[23-24]。

### 2.1　窃电模式分析

窃电是指非法使用电能的行为。窃电用户会通过破坏智能电表来发起窃电攻击，使电量减少或不计[25]。在用户计量准确的情况下，用户电量主要与电压、电流、功率因数和用电时间有关，窃电用户可以根据这4个影响电量的因素来进行窃电。窃电方法通常可分为5种：欠压法窃电、欠流法窃电、移相法窃电、扩差法窃电以及无表法窃电[26]。欠压法窃电通过减小电表电压线圈上两端的电压而使电量减少；欠流法窃电以减小电表电流线圈上的电流来进行窃电；移相法窃电通过改变电压与电流之间正常的相位使有功功率减少从而实现窃电；扩差法窃电改变电表内的构造使电表的误差发生变化从而使电量少记；无表法窃电绕开电表直接从供电企业的公共线路上接线来实现窃电[27]。根据现有窃电的方法[28]，可模拟为以下6种窃电模式。

1）按照固定比例$\alpha$减小$t$时段的电量$w_t$，可通过欠压法、欠流法和扩差法来实现。函数表达式为：

$$h_1(w_t)=\alpha w_t \qquad \alpha\in(0.2,0.8) \tag{1}$$

2）按照随机阈值$\gamma$削减电量$w_t$，高于阈值$\gamma$的电量固定为$\gamma$，可通过扩差法实现。函数表达式为：

$$h_2(w_t)=\begin{cases} w_t & w_t\leqslant\gamma \\ \gamma & w_t>\gamma \end{cases} \tag{2}$$

3）将随机时间段$(t_1,t_2)$内的所有电量$w_t$置0，可通过欠压法、欠流法和无表法实现。函数表达式为：

$$h_3(w_t)=\begin{cases}0 & t_1<t<t_2, t_2\geqslant t_1+4\\ w_t & \text{其他}\end{cases} \tag{3}$$

4)按照随机阈值$\gamma$可将任意的$w_t$置0,可通过欠压法、欠流法和无表法来实现,函数表达式为:

$$h_4(w_t)=\max\{w_t-\gamma,0\} \tag{4}$$

5)通过颠倒用电时序将电量$w_t$置于低电价时段,可通过移相法来实现,函数表达式为:

$$h_5(w_t)=w_{49-t} \tag{5}$$

6)取一天各时段用电量$w$的平均值,可通过欠流法、欠压法和移相法来实现,函数表达式为:

$$h_6(w_t)=\text{mean}(w) \tag{6}$$

式中:mean(·)表示求平均值。

### 2.2　窃电检测模型

基于多异学习器融合Stacking集成学习的窃电检测本质上是将用户历史用电数据作为输入,正常用户或窃电嫌疑用户作为输出的二分类模型,如图2所示。

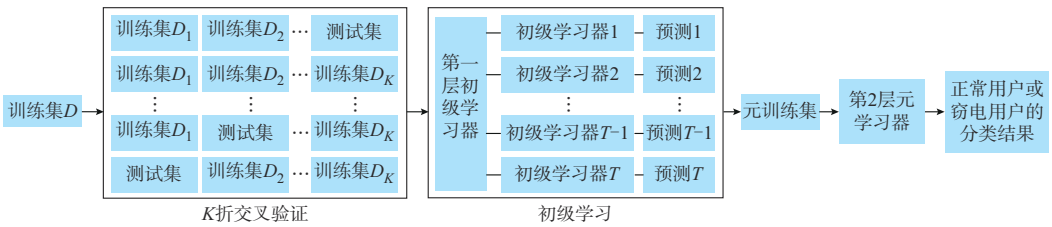


**图2　基于多异学习器融合Stacking集成学习的窃电检测**

**Fig. 2　Electricity theft detection based on multiple different learners fusion based on Stacking ensemble learning**

#### 2.2.1　$K$折交叉验证

模型在训练集上表现良好,但在测试集上表现却不理想,表明模型可能出现了过拟合。对于训练集$D=\{(x_m,y_m),m=1,2,\cdots,N\}$,其中$x_m$为第$m$个示例,$y_m$为对应示例的标记,$N$为示例总数。如果将该数据集同时用来训练初级学习器和元学习器,就会因用电数据被两层的学习器重复学习而造成很高的过拟合风险,导致对居民用户的用电行为判别不准确。因此,需要对平衡训练集$D$进行$K$折交叉验证。

如图2左边部分可知,平衡训练集$D$会被随机划分为$K$个同样大小的子集,有$D=D_1\cup D_2\cup\cdots\cup D_K$,$D_i\cap D_j=\varnothing(i\neq j)$,用某一个子集$D_l$当作第$l$次执行交叉验证时的测试集,余下子集的并集$D_{(-l)}=D\backslash D_l$当作训练集。这样就能对每个学习器进行$K$次的训练和测试,最后返回这$K$个测试结果的均值,以减小过拟合的风险[29]。给定$T$个初级学习器,使第$c$个初级学习器在交叉验证中的第$l$折训练集$D_{(-l)}$上进行训练,训练得出初级学习器$H_c^{(-l)}$。

#### 2.2.2　初级学习器和元学习器的选择

初级学习器的选择不仅可以从不同的空间和结构角度对居民用户历史用电数据进行数据挖掘,也会在实际的窃电检测中针对用电数据类别不平衡以及模型易陷入过拟合等问题实现不同学习器之间的优势互补,提高模型在窃电检测中的适应性。而相比于初级学习器的选取,单个元学习器的选取更加偏向于其分类过程中全方位的优化。附录A阐述了不同学习器在窃电检测中的分类机理以及优缺点。

利用Stacking结合策略,可以从不同的学习器中选择出优势互补的多异初级学习器和用于结合初级学习器的单个元学习器。由图2右边部分可知,在全部交叉验证过程结束时,将$T$个初级学习器的$K$次测试结果合并,得到元训练样例的示例部分$z_m=(z_{m1};z_{m2};\cdots;z_{mT})$,其中,$z_{mc}$为初级学习器$H_c^{(-l)}$在第$l$次执行的测试集中的样本$x_m$上的分类结果。将$z_m$与标记部分$y_m$一起生成元训练集$D'=\{(z_m,y_m)\}(m=1,2,\cdots,N)$,对应元学习器的输入向量。单个元学习器通过学习新生成的数据特征,输出正常用户或窃电用户的分类结果,以减小模型的过拟合风险。

### 2.3　窃电检测流程

采用Stacking集成学习融合多异学习器的窃电检测流程如下。

步骤1:根据“好而不同”的原则确定模型的初级学习器。初级学习器中的单一学习器初步考虑为分别来自机器学习里符号主义、连接主义和统计学习中常见的5种单一学习器$k$-最邻近(KNN)[9]、逻辑回归(LR)[10]、决策树(DT)[13]、反向传播(BP)[12]和支持向量机(SVM)[11]。集成学习器初步考虑为分别来自集成学习中用于降低方差的Bagging并行集成方式和用于减小偏差的Boosting串行集成方式为代表的4种集成学习器RF[14]、AdaBoost[15]、梯度提升树(GBDT)[16]和XGBoost[17]。在包含正常样本和窃电样本的多个测试集上使用评价指标和多样性度量对比分析这9个学习器,充分考虑预测能力较强和差异度较大的学习器,确定最终的初级学

习器。

步骤2：在步骤1的基础上，将以上9个学习器分别作为元学习器进行对比分析，确定最终的元学习器。

步骤3：基于步骤1和步骤2确定最终用于融合的初级学习器和元学习器，训练出基于多异学习器融合Stacking集成学习的窃电检测模型。

步骤4：在训练好的模型中输入用户用电数据，输出正常用户和窃电用户的分类结果。

### 2.4　评价指标和多样性度量

本文采用混淆矩阵衍生出来的准确率$e_{ACC}$、F1分数$e_{F1}$和受试者工作特征（ROC）曲线下面积$e_{AUC}$对模型的性能进行对比分析[1]，见附录B。同时，利用成对度量指标——双误（double failure，DF）度量和Q统计量从不同的角度衡量集成中学习器的多样性[30]，见附录C。

## 3　算例分析

### 3.1　数据集

本试验数据集采用爱尔兰智能电表数据集，其中包括爱尔兰地区6 000多户居民和企业用户长达535 d的连续用电数据，每条数据以30 min为单位记录了用户一天中48个时段的用电量[31]，单位为kW·h。从剔除了异常数据和缺失数据后的数据集中选取具有良好数据质量的1 000名居民用户的用电数据进行实验。由于每个用户家中都装有智能电表，并且愿意提供他们的用电数据以用作研究，本文认为所有的用电数据均为正常数据。

为提供足够的窃电数据，按照2.1节中的6种窃电模式将随机选择10%的正常数据修改为窃电数据。将生成的这6种窃电数据分别与正常数据进行混合，得到ET1、ET2、ET3、ET4、ET5和ET6共6种混合数据集，数据样本已共享。同时，从中任意选取正常数据和窃电数据混合，得到MIX混合数据集（即包含6种窃电数据）。本文提出的窃电检测方法适用于包含这6种窃电模式的数据集。

对于以上7个混合数据集中的每一个数据集，将其中的全部数据按6∶2∶2的比例划分为训练集、验证集和测试集。采用SMOTE算法对用电数据进行过采样，使正常用户和窃电用户两个类别的用电数据平衡，再用平衡的训练集训练模型，用验证集调整参数，而用测试集进行模型的评估。附录D为验证采用SMOTE算法前后基于多异学习器融合Stacking集成学习的窃电检测模型在MIX混合数据集上的对比分析，由结果可知，采用SMOTE处理不平衡数据集可使窃电检测性能得到提升。同时，从居民用户一天48个用电量特征中提取最大值、最小值、平均值和标准差4个综合特征。附录E所示为最终基于多异学习器融合Stacking集成学习的窃电检测模型在增加这4个综合特征前后的性能变化，模型性能提升得到了验证。

本文采用的样本数据以及研究分析结果数据已共享，见支撑数据。

### 3.2　初级学习器的选取

为构建基于多异学习器融合Stacking集成学习的窃电检测模型，需要利用评价指标和多样性度量从初步考虑的初级学习器中选出最终用于模型融合的初级学习器。因此，首先考虑上述9个学习器在7个混合数据集上的$e_{ACC}$、$e_{F1}$和$e_{AUC}$值，如图3所示。图4为这9个学习器在MIX混合数据集上的ROC曲线。其中，$e_{TPR}$和$e_{FPR}$分别为命中率和误检率，具体含义见附录B。

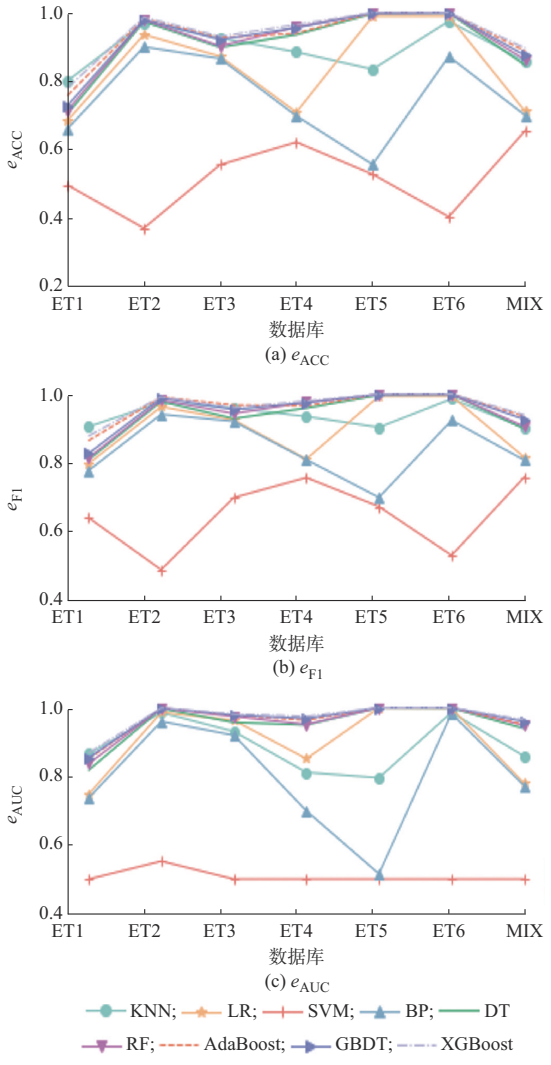


**图3　9个学习器在7个混合数据集上的评价指标**
**Fig. 3　Evaluation indices of nine learners for seven mixed data sets**

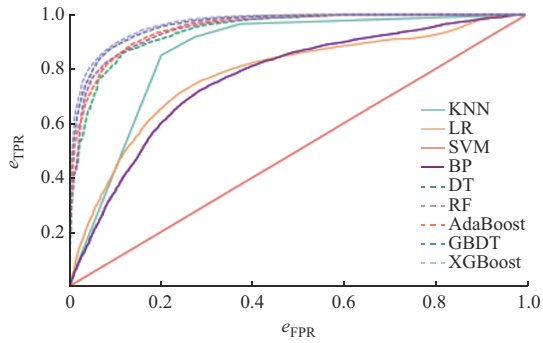


**图4　9个学习器在MIX混合数据集上的ROC曲线**
**Fig. 4　ROC curves of nine learners on MIX data set**

从图3(a)和(b)可明显看出，SVM在7个混合数据集上表现最不佳，其$e_{ACC}$和$e_{F1}$在部分数据集上低于0.5，而其余8个学习器的$e_{ACC}$和$e_{F1}$均超过了0.5。同时，除SVM外，BP的$e_{ACC}$和$e_{F1}$均最小。从图3(c)可知，SVM在7个混合数据集上的$e_{AUC}$均不高(在0.4～0.6之间)。而BP的$e_{AUC}$在7个混合数据集上的浮动很大，在ET5上甚至低于0.6，说明其在这7个数据集上表现不稳定。此外，由于MIX数据集包含了6种类型的窃电样本，故相对于只含一种类型的窃电样本数据集而言，其观测结果更具说服力。因此，通过在MIX混合数据集上的ROC曲线可知，SVM的ROC曲线几乎与对角线重合，而其余8个学习器的曲线都在对角线上方，这说明SVM的性能与随机猜测的学习器的性能基本无异。

综上所述，SVM和BP识别窃电用户的效果差，既会将窃电用户判别为正常用户，又会将正常用户判别为窃电用户，即误判的概率很大，这样不仅会遗漏窃电用户而且会干扰正常用户。因此，初步考虑的初级学习器首先排除SVM和BP。

其次，对于Stacking集成学习来说，不同学习器的差异程度越大，元学习器可以改进的地方就越多，因此模型的分类性能就越好。所以在选出分类性能优异的学习器后，还需考察各个学习器的多样性，尽可能选择差异性大的学习器。图5是7个学习器在MIX数据集上的DF值和Q值。

由于KNN和LR与其他学习器的训练机理差距较大，从而相关性较小，因此它们的DF值和Q值较其他学习器要小得多。并且对于单一学习器而言，虽然DT的$e_{ACC}$、$e_{F1}$和$e_{AUC}$值大部分都为最高，但是它的DF值和Q值也最高。同时，与RF、AdaBoost、GBDT和XGBoost集成学习器相比，DT的分类性能相对较差。因此，单一学习器中选择

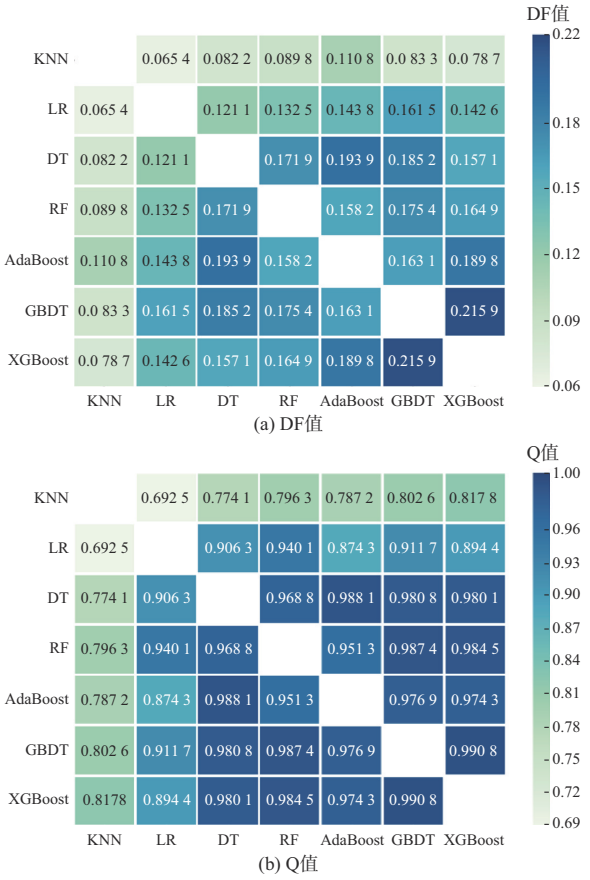


**图5　7个学习器的多样性度量**
**Fig. 5　Diversity measurement of seven learners**

KNN和LR作为模型的初级学习器。

RF、AdaBoost、GBDT和XGBoost这4种集成学习器都是以DT作为基学习器，数据观测方式存在较强相似性，所以DF值和Q值都较高。其中，RF使用了用于减少方差的并行集成方式，AdaBoost、GBDT和XGBoost使用了用于降低偏差的串行集成方式。从图3至图5来看，XGBoost的预测性能和多样性表现的都比AdaBoost和GBDT要好，故集成学习器中选择RF和XGBoost作为模型的初级学习器。

综上所述，DT在识别窃电用户的效果不如其他集成学习器的同时相关性也很高，而AdaBoost和GBDT在相关性高的同时分类性能也不如XGBoost，这样容易造成融合后的模型多样性较低从而不能更加准确地识别窃电用户。因此，基于多异学习器融合Stacking集成学习的窃电检测模型最终采用了KNN、LR、RF和XGBoost作为初级学习器。表1为选择性集成前后各个学习器和系统的多样性度量指标值的表现情况。

**表1　选择性集成前后各个学习器和系统的多样性度量指标值**
**Table 1　Diversity measurement indices of each learner and system before and after selective ensemble**

| 学习器 | DF值 | | | | Q值 | | | |
|---|---|---|---|---|---|---|---|---|
| | 选择性集成前 | | 选择性集成后 | | 选择性集成前 | | 选择性集成后 | |
| | 平均值 | 系统 | 平均值 | 系统 | 平均值 | 系统 | 平均值 | 系统 |
| KNN | 0.085 0 | | 0.078 0 | | 0.778 4 | | 0.768 9 | |
| LR | 0.127 8 | | 0.113 5 | | 0.869 9 | | 0.842 3 | |
| DT | 0.151 9 | | | | 0.933 0 | | | |
| RF | 0.148 8 | 0.142 2 | 0.129 1 | 0.112 3 | 0.938 2 | 0.903 9 | 0.907 0 | 0.854 3 |
| AdaBoost | 0.159 9 | | | | 0.925 4 | | | |
| GBDT | 0.164 1 | | | | 0.941 7 | | | |
| XGBoost | 0.158 2 | | 0.128 7 | | 0.940 5 | | 0.898 9 | |

由表1可知，KNN、LR、RF和XGBoost这4个学习器在选择性集成后的DF值和Q值较之前均有所减小，即选择性集成后的多样性程度更大。因此，通过评价指标和多样性度量优选好而不同的初级学习器能够使基于多异学习器融合Stacking集成学习的窃电检测模型更加有效地从多个视角开展窃电识别。

### 3.3　元学习器的选取

Wolpert早在提出Stacking时就认为元学习器的类型非常重要，因为元学习器既可以改善各个学习器的偏差，又可以保证一定的泛化能力以缓解过拟合。所以针对窃电检测二分类问题，需要选择用于结合多异初级学习器的元学习器。

由于初级学习器的预测各不相同且各有优缺点，这时需要选择合适的元学习器才能使最终Stacking集成学习的分类效果达到最优。因此，本文在选定的初级学习器的基础上，将最初进行对比的9个学习器分别作为元学习器进行训练，验证训练得到的模型在7个混合数据集上指标$e_{ACC}$、$e_{F1}$、$e_{AUC}$的平均值和平均运行时间，结果见表2。

**表2　9个学习器分别作为元学习器在7个混合数据集上指标的平均值**
**Table 2　Averages of indices on seven mixed data sets by using nine learners as meta-learner respectively**

| 元学习器 | $e_{ACC}$/% | $e_{F1}$/% | $e_{AUC}$/% | 运行时间/s |
|---|---|---|---|---|
| Stacking-KNN | 95.05 | 97.27 | 87.53 | 530.14 |
| Stacking-LR | 94.21 | 96.80 | 95.67 | 464.15 |
| Stacking-SVM | 93.56 | 96.44 | 93.84 | 494.41 |
| Stacking-BP | 95.11 | 97.29 | 95.54 | 512.04 |
| Stacking-DT | 95.16 | 97.31 | 91.85 | 483.22 |
| Stacking-RF | 95.24 | 97.37 | 96.23 | 543.00 |
| Stacking-AdaBoost | 95.28 | 97.40 | 96.09 | 763.03 |
| Stacking-GBDT | 95.36 | 97.44 | 95.80 | 716.23 |
| Stacking-XGBoost | 95.60 | 97.58 | 96.36 | 440.71 |

在选定初级学习器后，当元学习器分别为KNN、LR、SVM、BP、DT、RF、AdaBoost、GBDT和XGBoost时，Stacking融合后的$e_{ACC}$和$e_{F1}$值都超过了0.93，由此可见元学习器可以充分发挥不同学习器的优势。但是针对窃电检测二分类问题，为了使初级学习器的优势发挥至极致，元学习器应选择使最终的融合结果达到最佳的学习器。由表2可知，当元学习器为XGBoost时，Stacking集成模型表现最好，其$e_{ACC}$、$e_{F1}$和$e_{AUC}$值最高。

由表2可以看到，除了XGBoost外，集成学习器作为元学习器的运行时间都比单一学习器的要长，特别是以AdaBoost和GBDT作为元学习器的运行时间都超过了700 s，这是因为集成学习器的内部结构比单一学习器要复杂。但当XGBoost作为元学习器时，运行时间仅有440.71 s，这比将单一学习器作为元学习器时的运行时间还要短，证明了当XGBoost作为元学习器时，模型的复杂程度被降低。所以，基于多异学习器融合Stacking集成学习的窃电检测模型最终采用XGBoost作为元学习器。这样不仅能最大限度地避免误判和漏判，还能快速精准检测出窃电用户，减少供电公司的经济损失并提高检测效率。因此，通过选择最优的元学习器可以使基于多异学习器融合Stacking集成学习的窃电检测模型的分类性能达到最优，从而辅助供电企业进行用电稽查工作。附录F为在保证“好而不同”的条件以及元学习器不变的基础上，当初级学习器的数量分别为2、3和4时不同初级学习器组合方式在MIX混合数据集上的对比分析。由结果可知，本文所提出的方法能够使模型的检测性能得到较大的提升。

### 3.4　对比分析

为了验证Stacking结合策略下融合多异学习器的有效性，对于每种窃电方式将Stacking集成学习方法分别与采用多数投票法（majority voting，MV）、

加权投票法(weighted voting,WV)和改进加权投票法(improved weighted voting,IWV)[28]作为结合策略的集成学习方法进行比较,结果如表3所示。

**表3　不同结合策略的对比**
**Table 3　Comparison of different combination strategies**

| 数据集 | $e_{ACC}$/% | | | | $e_{F1}$/% | | | | $e_{AUC}$/% | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MV | WV | IWV | Stacking | MV | WV | IWV | Stacking | MV | WV | IWV | Stacking |
| ET1 | 90.14 | 90.95 | 91.34 | 93.71 | 92.90 | 93.47 | 94.01 | 95.13 | 93.62 | 94.13 | 94.66 | **95.07** |
| ET2 | 96.58 | 96.74 | 96.85 | 97.14 | 98.78 | 98.94 | 99.15 | 99.36 | 96.28 | 96.35 | 96.44 | **96.60** |
| ET3 | 94.19 | 94.38 | 95.02 | 95.65 | 95.93 | 96.26 | 96.51 | 96.92 | 97.03 | 97.24 | 97.35 | **97.56** |
| ET4 | 93.16 | 93.41 | 93.75 | 94.22 | 95.51 | 95.69 | 95.80 | 96.04 | 92.83 | 93.02 | 93.37 | **93.72** |
| ET5 | 94.31 | 94.62 | 95.76 | 96.47 | 97.26 | 97.44 | 97.74 | 98.15 | 96.33 | 96.46 | 96.69 | **97.21** |
| ET6 | 97.05 | 97.13 | 97.30 | 97.30 | 98.65 | 98.72 | 98.79 | 98.80 | 97.59 | 97.72 | 97.86 | **97.96** |
| MIX | 92.68 | 92.92 | 93.55 | 94.71 | 97.13 | 97.87 | 98.23 | 98.66 | 95.48 | 95.67 | 95.92 | **96.40** |

如表3所示,虽然在ET6数据集上采用IWV的集成学习方法的$e_{ACC}$与Stacking集成学习方法一样,但是在其余数据集上其$e_{ACC}$和$e_{F1}$都比Stacking集成学习方法要低。针对每种窃电方式,分别采用MV、WV以及IWV的集成学习方法,各方法对应$e_{ACC}$、$e_{F1}$和$e_{AUC}$的大小关系均为:IWV>WV>MV。由此看来,对于窃电检测分类问题,IWV比MV和WV效果更好。

同时,由于Stacking是利用元学习器XGBoost将不同学习器的优势发挥至极致,既能归纳并纠正不同学习器对于用电数据的偏置情况,又能保持较高的泛化能力来防止过拟合。所以,采用XGBoost作为元学习器的Stacking集成学习方法在除ET6外的6个数据集上的$e_{ACC}$、$e_{F1}$和$e_{AUC}$都比采用MV、WV以及IWV的集成学习方法要高,即有Stacking>IWV>WV>MV。特别地,对于采用IWV的集成学习方法而言[28],本文所提出的Stacking集成学习方法不仅可以充分发挥不同学习器的优势,还可以利用另一个学习器有效地综合这些优势。因此,通过选择最优的融合方式可以使不同学习器的优势发挥至极致,从而提升模型的检测性能。

综上所述,针对窃电检测二分类问题,本文所提出的利用Stacking集成学习融合多个不同学习器的窃电检测模型有以下几个方面的优势:一是对于数据而言,通过SMOTE算法平衡用电数据以避免分类结果出现偏倚;二是对于模型构建而言,采用$K$折交叉验证方法训练各个学习器以防止模型出现过拟合;三是对于融合对象而言,利用评价指标和多样性度量选择好而不同的多个学习器可以使模型能从多个视角识别窃电用户;四是对于融合方式而言,采用Stacking集成学习方式可以利用优选的元学习器有效融合多个不同的学习器以充分发挥它们的优势,提升模型检测性能。

## 4　结语

本文提出了一种基于多异学习器融合Stacking集成学习的窃电检测模型。针对用电数据类别不平衡以及采用投票法作为结合策略的集成学习方法无法充分发挥多个不同学习器优势等问题,本文利用SMOTE算法构造平衡的数据集,并采用Stacking结合策略融合多个不同学习器的优势和差异。在爱尔兰智能电表数据集上进行了验证,算例表明该模型可有效解决类别不平衡问题,且能够充分发挥不同学习器的优势。进一步将研究以下问题:一是窃电检测中的数据质量问题;二是利用电压等新特征开展窃电检测;三是当正常用户的用电行为模式和按窃电模式生成的用户用电模式相似时所造成的误判问题。

**支撑数据和附录见本刊网络版(http://www.aeps-info.com/aeps/ch/index.aspx),扫英文摘要后二维码可以阅读网络全文。**

## 参考文献

[1] 陈启鑫,郑可迪,康重庆,等.异常用电的检测方法:评述与展望[J].电力系统自动化,2018,42(17):189-199.
CHEN Qixin, ZHENG Kedi, KANG Chongqing, et al. Detection methods of abnormal electricity consumption behaviors: review and prospect [J]. Automation of Electric Power Systems, 2018, 42(17): 189-199.

[2] GAUR V, GUPTA E. The determinants of electricity theft: an empirical analysis of Indian states[J]. Energy Policy, 2016, 93: 127-136.

[3] 张宇帆,艾芊,李昭昱,等.基于特征提取的面向边缘数据中心的窃电监测[J].电力系统自动化,2020,44(9):128-134.
ZHANG Yufan, AI Qian, LI Zhaoyu, et al. Feature extraction based electricity theft detection for edge data center [J].

Automation of Electric Power Systems, 2020, 44(9): 128-134.

[4] ISMAIL M, SHAHIN M, SHAABAN M F, et al. Efficient detection of electricity theft cyber attacks in ami networks[C]// 2018 IEEE Wireless Communications and Networking Conference, April 15-18, 2018, Barcelona, Spain: 1-6.

[5] ZHANG T, GAO R, SUN S. Theories, applications and trends of non-technical losses in power utilities using machine learning [C]// 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), May 25-27, 2018, Xi′an, China: 2324-2329.

[6] JIANG R, LU R, WANG Y, et al. Energy-theft detection issues for advanced metering infrastructure in smart grid [J]. Tsinghua Science and Technology, 2014, 19(2): 105-120.

[7] HE Y, MENDIS G J, WEI J. Real-time detection of false data injection attacks in smart grid: a deep learning-based intelligent mechanism [J]. IEEE Transactions on Smart Grid, 2017, 8(5): 2505-2516.

[8] AMIN S, SCHWARTZ G A, CÁRDENAS A A, et al. Game-theoretic models of electricity theft detection in smart utility networks: providing new capabilities with advanced metering infrastructure [J]. IEEE Control Systems Magazine, 2015, 35(1): 66-81.

[9] 沈海涛，秦靖雅，陈浩，等.电力用户用电数据的异常数据审查和分类[J].电力与能源，2016，37(1)：17-22.
SHEN Haitao, QIN Jingya, CHEN Hao, et al. Anomaly detection and category of electrical utilization data[J]. Power and Energy, 2016, 37(1): 17-22.

[10] 史玉良，荣以平，朱伟义.基于用电特征分析的窃电行为识别方法[J].计算机研究与发展，2018，55(8)：1599-1608.
SHI Yuliang, RONG Yiping, ZHU Weiyi. Stealing behavior recognition method based on electricity characteristics analysis [J]. Journal of Compute Research and Development, 2018, 55(8): 1599-1608.

[11] 胡天宇，郭庆来，孙宏斌.基于堆叠去相关自编码器和支持向量机的窃电检测[J].电力系统自动化，2019，43(1)：119-125.
HU Tianyu, GUO Qinglai, SUN Hongbin. Nontechnical loss detection based on stacked uncorrelating autoencoder and support vector machine [J]. Automation of Electric Power Systems, 2019, 43(1): 119-125.

[12] 王庆宁，张东辉，孙香德，等.基于GA-BP神经网络的反窃电系统研究与应用[J].电测与仪表，2018，55(11)：35-40.
WANG Qingning, ZHANG Donghui, SUN Xiangde, et al. Research and application of electricity anti-stealing system based on GA-BP neural network [J]. Electrical Measurement and Instrumentation, 2018, 55(11): 35-40.

[13] JINDAL A, DUA A, KAUR K, et al. Decision tree and SVM-based data analytics for theft detection in smart grid [J]. IEEE Transactions on Industrial Informatics, 2016, 12(3): 1005-1016.

[14] 许刚，谈元鹏，戴腾辉.稀疏随机森林下的用电侧异常行为模式检测[J].电网技术，2017，41(6)：1964-1973.
XU Gang, TAN Yuanpeng, DAI Tenghui. Sparse random forest based abnormal behavior pattern detection of electric power user side[J]. Power System Technology, 2017, 41(6): 1964-1973.

[15] 游文霞，申坤，杨楠，等.基于AdaBoost集成学习的窃电检测研究[J].电力系统保护与控制，2020，48(19)：151-159.
YOU Wenxia, SHEN Kun, YANG Nan, et al. Research on electricity theft detection based on AdaBoost ensemble learning [J]. Power System Protection and Control, 2020, 48(19): 151-159.

[16] RAZAVI R, GHARIPOUR A, FLEURY M, et al. A practical feature-engineering framework for electricity theft detection in smart grids [J]. Applied energy, 2019, 238: 481-494.

[17] BUZAU M M, TEJEDOR-AGUILERA J, CRUZ-ROMERO 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] CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: synthetic minority over-sampling technique [J]. Journal of artificial intelligence research, 2002, 16: 321-357.

[19] 王毅，谷亿，丁壮，等.基于模糊熵和集成学习的电动汽车充电需求预测[J].电力系统自动化，2020，44(3)：114-121.
WANG Yi, GU Yi, DING Zhuang, et al. Charging demand forecasting of electric vehicle based on empirical mode decomposition-fuzzy entropy and ensemble learning [J]. Automation of Electric Power Systems, 2020, 44(3): 114-121.

[20] ZHOU Zhihua. Ensemble methods foundations and algorithms [M]. Boca Raton, USA: CRC Press, 2012: 67-95.

[21] 史佳琪，张建华.基于多模型融合Stacking集成学习方式的负荷预测方法[J].中国电机工程学报，2019，39(14)：4032-4042.
SHI Jiaqi, ZHANG Jianhua. Load forecasting based on multi-model by Stacking ensemble learning [J]. Proceedings of the CSEE, 2019, 39(14): 4032-4042.

[22] 邓威，郭钇秀，李勇，等.基于特征选择和Stacking集成学习的配电网网损预测[J].电力系统保护与控制，2020，48(15)：108-115.
DENG Wei, GUO Yixiu, LI Yong, et al. Power losses prediction based on feature selection and Stacking integrated learning [J]. Power System Protection and Control, 2020, 48(15): 108-115.

[23] HEATON J. An empirical analysis of feature engineering for predictive modeling [C]// IEEE SoutheastCon 2016, March 30-April 3, 2016, Norfolk, USA: 1-6.

[24] PUNMIYA R, CHOE S. Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing [J]. IEEE Transactions on Smart Grid, 2019, 10(2): 2326-2329.

[25] YANG Z, WEN H. Electricity theft detection base on extreme gradient boosting in AMI [J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-9.

[26] 程超，张汉敬，景志敏，等.基于离群点算法和用电信息采集系统的反窃电研究[J].电力系统保护与控制，2015，43(17)：69-74.
CHENG Chao, ZHANG Hanjing, JING Zhimin, et al. Study on the anti-electricity stealing based on outlier algorithm and the electricity information acquisition system [J]. Power System Protection and Control, 2015, 43(17): 69-74.

[27] 金晟,苏盛,薛阳,等.数据驱动窃电检测方法综述与低误报率研究展望[J].电力系统自动化,2022,46(1):3-14.
JIN Sheng, SU Sheng, XUE Yang, et al. Review on data-driven based electricity theft detection method and research prospect for low false positive rate[J]. Automation of Electric Power Systems, 2022, 46(1): 3-14.
[28] 游文霞,申坤,杨楠,等.基于Bagging异质集成学习的窃电检测[J].电力系统自动化,2021,45(2):105-113.
YOU Wenxia, SHEN Kun, YANG Nan, et al. Electricity theft detection based on Bagging heterogeneous ensemble learning [J]. Automation of Electric Power Systems, 2021, 45(2): 105-113.
[29] 周志华.机器学习[M].北京:清华大学出版社,2016.
ZHOU Zhihua. Machine learning [M]. Beijing: Tsinghua University Press, 2016.
[30] 孙博,王建东,陈海燕,等.集成学习中的多样性度量[J].控制与决策,2014,29(3):385-395.
SUN Bo, WANG Jiandong, CHEN Haiyan, et al. Diversity measures in ensemble learning[J]. Control and Decision, 2014, 29(3): 385-395.
[31] ISSDA. Data from the commission for energy regulation[EB/OL]. [2020-02-01]. http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

游文霞(1978—),女,博士,副教授,主要研究方向:机器学习在电力系统中的应用。E-mail:youwenxia@ctgu.edu.cn

李清清(1996—),女,通信作者,硕士研究生,主要研究方向:人工智能在电力系统中的应用。E-mail:516378642@qq.com

杨 楠(1987—),男,博士,副教授,主要研究方向:电力系统运行与控制、电力系统规划。E-mail:ynyyayy@ctgu.edu.cn

(编辑 代长振)

## Electricity Theft Detection Based on Multiple Different Learner Fusion by Stacking Ensemble Learning

*YOU Wenxia*, *LI Qingqing*, *YANG Nan*, *SHEN Kun*, *LI Wenwu*, *WU Zeli*

(School of Electrical and New Energy, China Three Gorges University, Yichang 443002, China)

**Abstract:** Aiming at the problems that the consumer power consumption data categories are unbalanced for electricity theft detection, and the ensemble learning method using voting as a combination strategy cannot give full play to the advantages of multiple different learners, a model using Stacking ensemble learning to fuse multiple different learners is proposed and applied to electricity theft detection. First, starting from the factors affecting electricity metering, six electricity theft behavior modes are simulated according to five common electricity theft methods. Secondly, synthetic minority oversampling technique (SMOTE) is used to process the unbalanced power consumption data, and *K*-fold cross-validation method is used to divide the balanced training sets to alleviate the overfitting caused by repeated learning. Then, the evaluation indicators and diversity metrics are employed to optimize different primary learners and meta-learners of the model, and a Stacking ensemble learning electricity theft detection model integrating the advantages and differences of different learners is constructed. Finally, the comparative analysis results of examples show that the proposed electricity theft detection model can effectively solve the imbalance of power consumption data categories, give full play to the advantages of different learners, and the evaluation index is good.

This work is supported by National Natural Science Foundation of China (No. 51607104).

**Key words:** Stacking combination strategy; ensemble learning; electricity theft detection; synthetic minority oversampling technique (SMOTE); *K*-fold cross-validation