DOI: 10.19783/j.cnki.pspc.191409

基于 AdaBoost 集成学习的窃电检测研究

游文霞[1], 申 坤[1], 杨 楠[1], 李清清[1], 吴永华[2], 李黄强[3]

(1.三峡大学电气与新能源学院，湖北 宜昌 443002；2.国网湖北省电力公司孝感供电公司，湖北 孝感 432000；
3.国网湖北省电力公司宜昌供电公司，湖北 宜昌 443002)

摘要：针对传统窃电检测中单一分类方法的不足，提出一种基于 AdaBoost 集成学习的窃电检测算法。首先利用训练集对决策树、误差逆传播神经网络、支持向量机和 k 最近邻四种方法进行训练对比，提出决策树作为 AdaBoost 集成学习算法的弱学习器。其次通过绘制不同学习率下的分类错误率曲线，确定 AdaBoost 集成学习算法的学习率和弱学习器个数。最后利用爱尔兰智能电表数据集中的居民用电数据对所提算法进行测试评估，将 AdaBoost 集成学习算法与决策树、k 最近邻、误差逆传播神经网络、支持向量机等各类单一强学习算法对比。结果表明基于 AdaBoost 集成学习的窃电检测算法在准确率、命中率、误检率等检测指标中最优，灵敏性分析验证了基于 AdaBoost 集成学习的窃电检测方法的有效性。

关键词：AdaBoost；窃电检测；集成学习；决策树；爱尔兰数据集

Research on electricity theft detection based on AdaBoost ensemble learning

YOU Wenxia[1], SHEN Kun[1], YANG Nan[1], LI Qingqing[1], WU Yonghua[2], LI Huangqiang[3]

(1. School of Electrical and New Energy, China Three Gorges University, Yichang 443002, China; 2. Xiaogan Power Supply Company, State Grid Hubei Electric Power Company, Xiaogan 432000, China; 3. Yichang Power Supply Company, State Grid Hubei Electric Power Company, Yichang 443002, China)

Abstract: There is a deficiency in the single classification method in traditional electricity thief detection. Thus a method based on AdaBoost ensemble learning is proposed. First, the training set is used to compare the decision tree, error backpropagation network, support vector machine and k-nearest neighbors, and the decision tree is adopted as the weak learner of the AdaBoost algorithm. Secondly, the learning rate and the number of weak learners of AdaBoost ensemble learning are determined by plotting the error rate curves under different learning rates. Finally, the proposed method is tested and evaluated on the Irish smart meter dataset. It is compared with the single strong learning algorithms, such as decision tree, error backpropagation network, support vector machine, k-nearest neighbors. The results show that electricity theft detection based on AdaBoost ensemble learning is the best among the indicators of accuracy, true positive rate and false positive rate. The sensitivity analysis shows the validity of the electricity theft detection method based on AdaBoost ensemble learning.

This work is supported by National Natural Science Foundation of China (No. 51607104) and 2019 Science and Technology Project of State Grid Hubei Electric Power Company (No. 5215K018006B).

Key words: AdaBoost; electricity theft detection; ensemble learning; decision tree; Irish data set

0 引言

电力系统在实现电能传输中存在能量损失，一是因为电网元件电阻或能源转化效率上限产生的技术性损失，二是因为电力用户的窃电等欺骗性用电行为产生的非技术性损失[1-2]。窃电由于会造成大量经济损失，因此一直受到供电企业和研究者的关注。随着高级量测体系(Advanced Metering Infrastructure, AMI)的逐渐建立和智能电表的不断普及，过去依靠破坏传统电表或私拉电线等窃电手段已转变为通过计算机技术和通信

基金项目：国家自然科学基金项目资助(51607104)；国网湖北省电力公司 2019 年科技项目资助(5215K018006B)

技术攻击智能电表，通过数据篡改等手段将用电量变小或直接归零。传统人工筛查进行窃电检测效率低下，无法满足窃电检测需求。充分利用海量数据对窃电用户进行筛查并开展窃电检测已成为国内外热点研究领域。

窃电检测主要有三类方法：基于系统状态、基于博弈论和基于分类[3]。基于系统状态的方法通过比较智能电表数据与其他仪器测量数据是否一致[4-6]，从而识别是否发生窃电，但需要额外投资。基于博弈论的方法将窃电检测问题描述为窃电者与电力公司之间的博弈[7-9]，但参与者的效用函数以及策略不易确定。基于分类的方法则只需要利用收集到的用户用电数据[10-15]，通过数据挖掘识别窃电[16-21]，目前已开展了广泛研究。

文献[10]提出基于误差逆传播(Back Propagation, BP)神经网络的反窃电方法，通过历史数据、当前数据、时间数据构建评价模型，采用遗传算法加快收敛速度，并在国网某省公司提供的数据集上得到了验证。文献[11]将决策树(Decision Tree, DT)与支持向量机(Support Vector Machine, SVM)结合，决策树计算结果作为支持向量机输入，从而判断用户属于窃电用户还是正常用户；该算法在爱尔兰智能电表数据集上进行测试，准确率达到了92.50%。文献[12]用k-近邻(K-nearest Neighbors, KNN)算法进行异常用电数据分类，利用某电力公司提供的数据对所提算法进行了验证。

上述方法都是建立强分类器进行电力用户行为模式辨识，计算成本较高且检测精度有待进一步提升。文献[22]和文献[23]的研究表明，与单一强分类器相比，多个弱分类器形成的弱分类器集群能够获得更好的计算效率。

本文提出基于自适应提升(Adaptive Boosting, AdaBoost)的集成学习(ensemble learning)算法对窃电行为进行检测。首先介绍AdaBoost集成学习；然后在Boosting集成框架下考虑多种模型的性能指标，建立基于AdaBoost集成学习的窃电检测模型；最后在爱尔兰智能电表数据集的居民用电数据上，将AdaBoost集成学习算法与BP、DT、SVM、KNN等算法进行对比实验。

1 基于AdaBoost集成学习算法

1.1 集成学习

集成学习通过集合多个学习器来完成学习任务，也称基于委员会的学习[24]。图1是集成学习的一般结构：先通过训练确定多个个体学习器，再用某种策略将这些个体学习器集成起来。集成学习将多个学习器结合，常可获得比单一学习器更优越的性能。

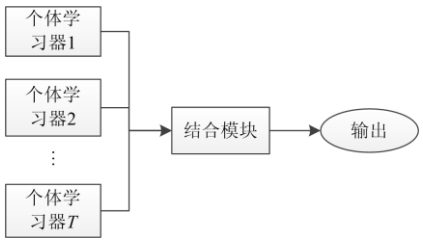


图1 集成学习的一般结构

Fig. 1 General structure of the ensemble learning

根据个体学习器的生成方式，集成学习方法大致分为两类：1) 个体学习器间存在强依赖关系的集成学习，各个体学习器串行生成，代表是基于Boosting的集成学习算法；2) 个体学习器间不存在强依赖关系的集成学习，各个体学习器并行生成，代表是基于Bagging的集成学习算法。

1.2 基于AdaBoost集成学习算法

基于Boosting的集成学习算法中最常用的是AdaBoost集成学习算法。其核心思想是训练一系列弱分类器，然后将弱分类器加权联合，构成一个强分类器，图2表示其一般结构。

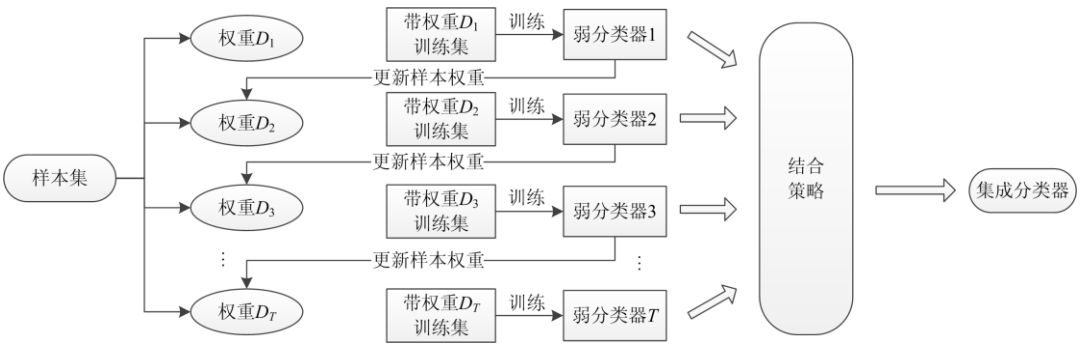


图2 AdaBoost集成学习的一般结构

Fig. 2 General structure of the AdaBoost ensemble learning

给定数据集：$(\boldsymbol{x}_1, y_1),(\boldsymbol{x}_2, y_2),\text{L},(\boldsymbol{x}_N, y_N)$，其中 $y_i \in \{-1,1\}$，用于表示样本的类别标签，$\boldsymbol{x}_i$ 表示样本的特征向量，$i = 1,\text{L},N$，N 为样本总数。集成学习的具体步骤如下。

Step 1：初始化数据的权值分布向量 $\boldsymbol{D}_1$。

$$\boldsymbol{D}_1 = (w_{11}, w_{12},\text{L},w_{1N}) = (1/N, 1/N,\text{L},1/N) \tag{1}$$

$$\sum_{i=1}^{N} w_{1i} = 1 \tag{2}$$

式中，w_{1i} 表示第 1 次迭代时第 i 个样本的权值。

Step 2：进行迭代运算，达到设定值时停止。对于第 t 次迭代，进行如下步骤($t = 1,2,\text{L},T$，T 为总迭代次数)。

(1) 选取一个当前误差率最低的弱分类器 h_t，并计算该弱分类器在分布 $\boldsymbol{D}_t$ 上的预测误差率 e_t 为

$$e_t = P(h_t(x_i) \neq y_i) = \sum_{i=1}^{N} w_{ti} I(h_t(x_i) \neq y_i) \tag{3}$$

式中，$h_t(x_i)$ 表示弱分类器对样本 x_i 的分类，若分类错误则 $I(\cdot)$ 值取 1，反之取 0。

(2) 计算该弱分类器在集成分类器中所占权重 α_t 为

$$\alpha_t = \frac{1}{2}\left(\frac{1-e_t}{e_t}\right) \tag{4}$$

(3) 更新训练样本的权值分布 $\boldsymbol{D}_{t+1}$。

$$\boldsymbol{D}_{t+1} = \frac{\boldsymbol{D}_t \exp(-\alpha_t y_i h_t(x_i))}{Z_t} \tag{5}$$

式中，Z_t 为归一化常数 $Z_t = 2\sqrt{e_t(1-e_t)}$。

Step 3：按弱分类器权值 α_t 组合各个弱分类器，得到集成学习分类器。

$$H(x) = \text{sign}(f(x)) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{6}$$

2　基于 AdaBoost 集成学习的窃电检测

2.1 检测流程

基于 AdaBoost 集成学习的窃电检测流程如图 3 所示，具体步骤如下。

Step 1：用生成的训练集对常见的 BP、DT、SVM、KNN 四种算法进行训练对比，确定 AdaBoost 集成学习的弱学习器。

Step 2：控制 AdaBoost 集成学习的学习率(Learning Rate, LR)，绘制弱学习器个数–分类错误率曲线，确定最佳弱学习器个数(迭代次数)和 LR。

Step 3：根据 Step 1、Step 2 得到 AdaBoost 集成学习的参数，对 AdaBoost 分类器进行训练。

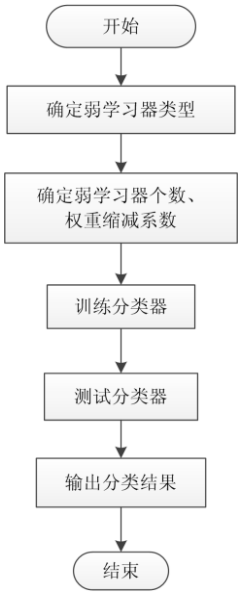


图 3 基于 AdaBoost 集成学习的窃电检测流程

Fig. 3 Process of electricity theft detection based on AdaBoost ensemble learning

Step 4：将测试集输入到 Step 3 训练好的 AdaBoost 分类模型中，输出分类结果。

2.2 评价指标

窃电检测问题本质是将用电数据进行正常和异常用电的二元分类。二元分类可将样本根据其真实类别与学习器预测类别的组合划分为真正例(True Positive)、假正例(False Positive)、真反例(True Negative)、假反例(False Negative)四种情形，令 TP、FP、TN、FN 分别表示其对应的样本数，则显然有 $TP + FP + TN + FN =$ 样本总数。分类结果的“混淆矩阵”如表 1 所示。

表 1 混淆矩阵

Table 1 Confusion matrix

用户	检测为正常用户	检测为异常用户
实际正常用户	TP	FN
实际异常用户	FP	TN

本文采用准确率(Accuracy, ACC)、命中率(True Positive Rate, TPR)、误检率(False Positive Rate, FPR)、ROC 曲线下面积(Area Under ROC Curve, AUC)4 个分类检测评价指标，定义分别如下。

准确率(ACC)为

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{7}$$

准确率表示总样本中有多少被正确预测。但当正负样本数量严重失衡时，单纯使用准确率作

为分类模型评价指标缺乏可信度[25]，因此还需综合其他指标来评价。

命中率(TPR)和误检率(FPR)定义为

$$TPR = TP/(TP+FN) \tag{8}$$

$$FPR = FP/(FP+TN) \tag{9}$$

TPR 和 *FPR* 的取值为[0, 1]，*TPR* 越接近 1，*FPR* 越接近 0 说明检测效果越好。

受试者工作特征(Receiver Operating Characteristic, ROC)曲线描述 *FPR* 和 *TPR* 两个指标变化的相对关系，如图 4 所示。对于二元分类模型输出的连续数值，将大于阈值的样本划为正类。进行学习器比较时，若一个学习器的 ROC 曲线被另一个学习器的曲线完全“包住”，则可判定后者的性能优于前者；若两个学习器的 ROC 曲线发生交叉时，则较为合理的是比较 ROC 曲线覆盖下的面积 AUC。较大 *AUC* 代表了较好性能表现，*AUC*=1 对应理想分类器。

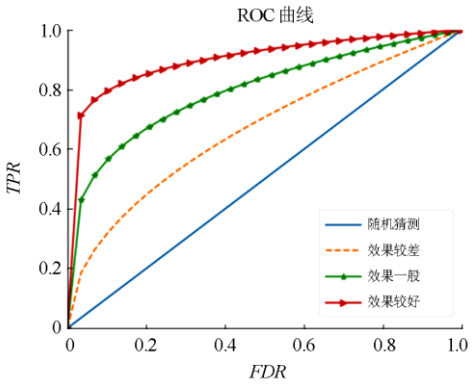


图 4 四种检测效果的 ROC 曲线

Fig. 4 ROC curves of four detection effects

3 算例分析

本节展示了算例分析和对比实验，实验使用 Core-TM i5-3470@3.20 GHz 处理器在 Anaconda(基于 python 3.6)环境下进行。

3.1 数据集

本实验选用爱尔兰智能电表数据集，该数据集含有爱尔兰 6 000 多户家庭和商业用户连续 535 天的用电记录(每 30 min 采集一次数据)[26]。选用其中 1 000 户居民用户进行实验。因此一共得到 535×1000=535 000 条用电记录。由于数据集中用户均同意将其用电记录作为研究使用，因此假设所有用户均属于正常用电用户。在这 535 000 条用电记录中，随机选择 10%作为窃电样本，并结合文献[3]中的窃电样本生成方法修改了这部分样本。给定正常用电记录 $\boldsymbol{x}=\{x^1,x^2,\mathrm{L}\ ,x^{48}\}$，6 种窃电样本分别按照以下方式生成。

(1) $h_1(x^t)=\alpha x^t,\alpha\in(0.2,0.8)$；

(2) $h_2(x^t)=\begin{cases}x^t, & x^t\le\gamma\\ \gamma, & x^t>\gamma\end{cases}$；

(3) $h_3(x^t)=x^{48-t}$；

(4) $h_4(x^t)=\max\{x^t-\gamma,0\}$；

(5) $h_5(x^t)=\begin{cases}0, t_1<t<t_2\\ x^t, \text{其他}\end{cases}$；

(6) $h_6(x^t)=mean(\boldsymbol{x})$。

其中，$h_1(\cdot)$ 为将所有样本乘以 0.2 到 0.8 之间相同的随机数；$h_2(\cdot)$ 为随机选择阈值 $\gamma(0<\gamma<\max(x_t))$，如果实际用电量大于 γ，则将用电量替换为 γ，否则保持不变；$h_3(\cdot)$ 为将用电记录倒序排列；$h_4(\cdot)$ 为将用电记录减去阈值 γ，并取其差值与 0 之间的最大值；$h_5(\cdot)$ 定义时间段 (t_1,t_2)，当用电记录位于该区间时记为 0，其中 $t_1,t_2\in(0,48)$，且 $t_2\ge t_1+4$；$h_6(\cdot)$ 为用电记录的平均值。

窃电样本生成后，为了验证所用模型在不同环境中的可行程度，分别将 6 种窃电样本与正常用电样本混合，形成 6 个窃电样本数据集，分别记为 ET1、ET2、ET3、ET4、ET5 和 ET6。并随机选择窃电样本与正常样本混合形成第 7 个窃电样本集(即包含所有 6 种窃电方式)，记为 MIX。每个含窃电样本数据集可表示为 $\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\mathrm{L}\ ,(\boldsymbol{x}_N,y_N)\}$，$N$ 是所有用户总用电记录的条数，$y_i\in\{-1,1\}$ 代表每条记录标签(值为−1 代表发生窃电)。最后，所有样本被随机分配到训练集(60%)、验证集(20%)和测试集(20%)。训练集和验证集分别用于训练模型和调整 AdaBoost 的超参数，测试集用于评估模型。图 5 表示用户正常用电及其对应生成的 6 种窃电行为。横轴表示时间点(每 30 min 为一个点，共 48 个点)，纵轴表示用电量。

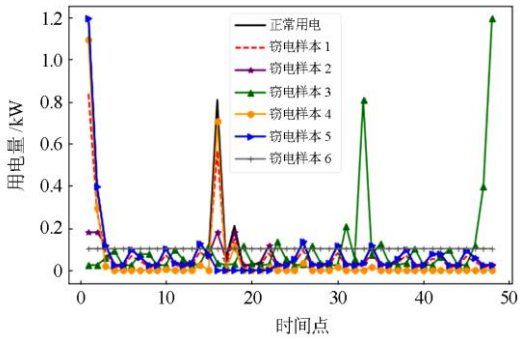


图 5 正常用电及其对应的 6 种窃电记录

Fig. 5 Normal power consumption and its corresponding six electricity theft records

3.2 模型超参数确定

3.2.1 弱学习器类型确定

为确定弱学习器类型，采用 BP、DT、SVM、KNN 四种分类器进行对比，并依据经验选取部分模型参数。图 6 分别表示四种不同分类器在 7 个含窃电样本数据集训练集上的 *ACC*、*AUC*、*TPR* 和 *FPR* 指标。

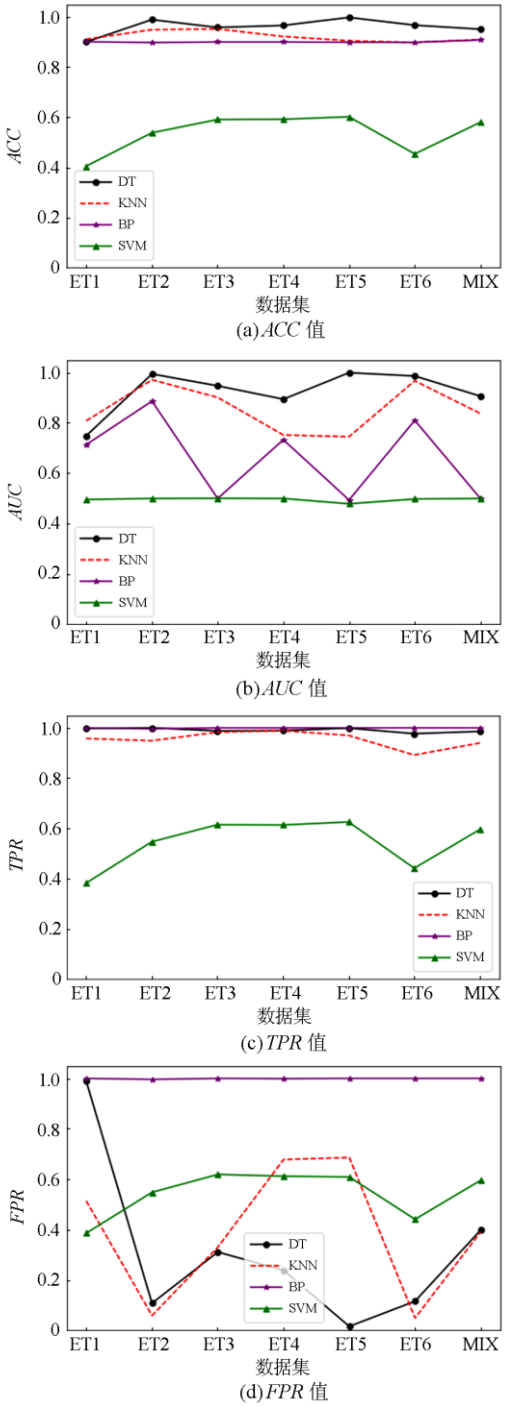


图 6 四种分类器在爱尔兰数据集上的平均测试结果

Fig. 6 Average test results of four classifiers on Irish data sets

由图 6 可知，DT 和 KNN 算法在四项评价指标中优于 BP 和 SVM 算法。四种分类器在数据集 MIX 上所花费时间如图 7 所示，KNN 算法所花费时间是 DT 算法的 10 倍以上。因此从评价指标和计算时间综合考虑，选择 DT 算法作为 AdaBoost 算法的弱分类器。

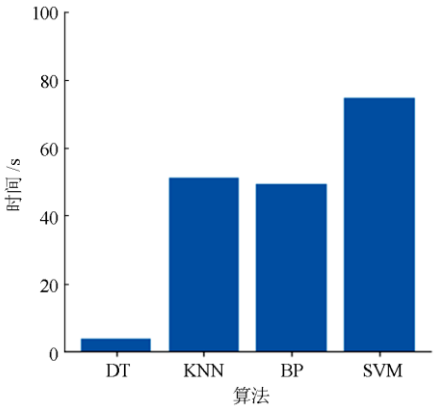


图 7 四种分类器花费时间

Fig. 7 Time spent by four classifiers

3.2.2 弱学习器个数和学习率确定

对于基于 AdaBoost 的集成学习算法，弱学习器个数和 LR 两个参数会对分类准确率有直接影响。为确定基于 AdaBoost 集成学习算法的最佳弱学习器个数，在 MIX 数据集的验证集上，分别将学习率取 0.2、0.3、0.5、0.8 和 1.0，得到弱学习器个数-分类错误率(与准确率之和为 1)曲线如图 8 所示。由图 8 可知，当弱学习器个数为 45 时，各曲线值开始趋于稳定，考虑图 9 所示训练时间和弱学习器个数几乎成正比，故将弱学习器个数设为 45。

由图 8 可知，当 *LR* 大于 0.5 时，其分类错误率更低，因此将 *LR* 取 0.6~1.0。改变 *LR* 的值，进一步得到不同 *LR* 值时的分类错误率，如图 10

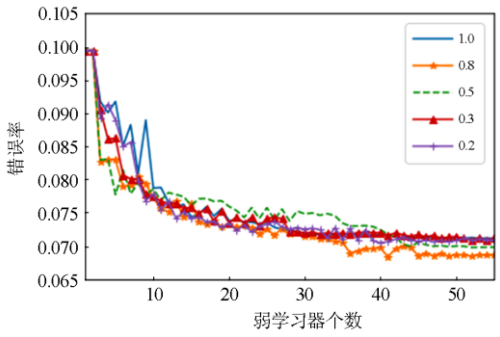


图 8 弱学习器个数-分类错误率曲线(0<*LR*≤1)

Fig. 8 Weak learner number-error rate curve (0<*LR*≤1)

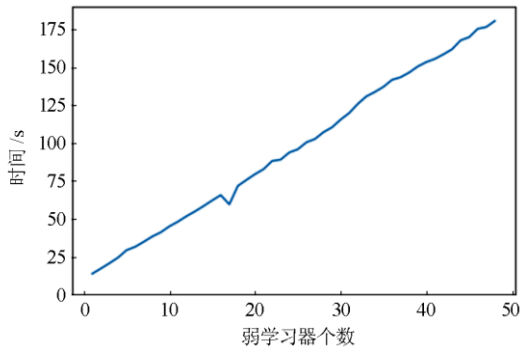


图 9 弱学习器个数-花费时间曲线

Fig. 9 Weak learner number/time spent curve

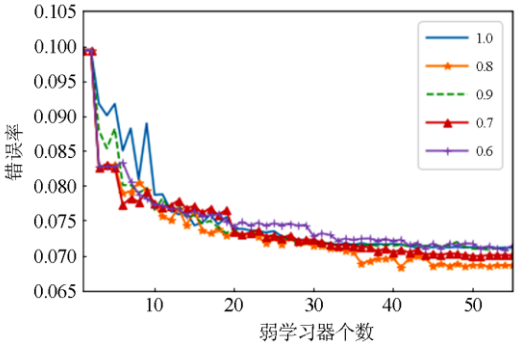


图 10 弱学习器个数-分类错误率曲线(0.5<*LR*≤1)

Fig. 10 Weak learner number-error rate curve (0.5<*LR*≤1)

所示。当 *LR* 取 0.8 时，算法具有最低分类错误率，故 *LR* 取为 0.8。

得到 AdaBoost 集成学习的参数后，为确定模型是否会过拟合，分别得到 AdaBoost 模型在 MIX 训练集和验证集上的误差，如图 11 所示。由图 11 可知，训练集和验证集上的误差比较接近，这说明基于 AdaBoost 集成学习的模型没有被过拟合。

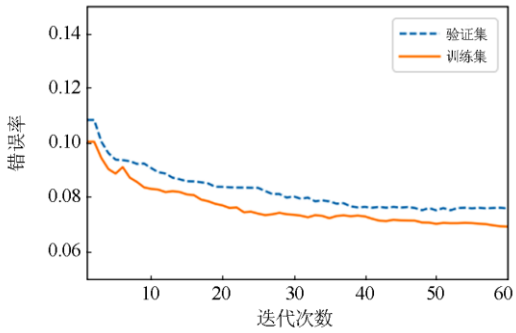


图 11 AdaBoost 在训练集和验证集上的误差

Fig. 11 Error of AdaBoost on training and validation sets

3.3 与其他算法的对比分析

对比基于 AdaBoost 集成学习与 BP[10]、DT[11]、SVM[11]和 KNN[12]算法的测试结果。

3.3.1 在含单一窃电样本数据集上的对比分析

将 BP、DT、SVM、KNN 以及基于 AdaBoost 集成学习 5 种算法，在 ET1、ET2、ET3、ET4、ET5 和 ET6 的 6 个只含单一窃电样本数据集的测试集上进行测试，结果如表 2 所示。

表 2 5 种算法在 6 个只含单一窃电样本数据集上的测试结果

Table 2 Test results of five algorithms on six data sets containing single tamper samples

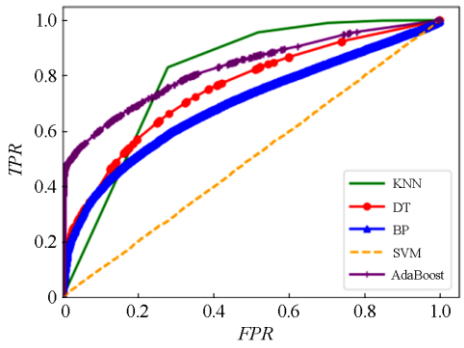
类型	*ACC*/%					*AUC*/%				
	DT[11]	KNN[12]	BP[10]	SVM[11]	AdaBoost	DT[11]	KNN[12]	BP[10]	SVM[11]	AdaBoost
ET1	89.8	**90.8**	89.8	57.8	89.8	75.1	80.5	70.0	49.9	**82.7**
ET2	98.9	95.1	89.9	51.4	**99.3**	99.5	97.2	50.1	45.9	**99.9**
ET3	95.8	95.4	60.8	60.4	**97.2**	94.7	90.4	50.0	49.7	**98.6**
ET4	96.6	92.4	89.9	46.6	**99.5**	89.7	74.6	64.8	50.5	**95.9**
ET5	99.8	90.3	89.7	58.9	**99.9**	99.8	74.4	54.1	50.8	**99.9**
ET6	96.6	89.9	89.0	34.6	**99.6**	98.5	96.7	83.7	50.1	**99.9**

由于部分样本之间不存在显著相关特性，因而 SVM 的分类正确率对于不同数据集波动较大。此外，虽然 BP 在 5 个数据集上的准确率都接近 0.9，但其 *AUC* 值较低。相比于其他算法，AdaBoost 在 6 个数据集上 *AUC* 值均为最高的同时，*ACC* 值也在 5 个数据集上达到了最大。对于数据集 ET1，由于其生成方式为实际用电数据乘以 0.2 到 0.8 之间的一个随机数，即二者在数值之间有一定的相似性，使算法不易区分，造成 DT、KNN 和 AdaBoost 三种算法在 ET1 数据集上的表现较其他五个数据集差。5 种算法在 6 个数据集的 ROC 曲线见图 12。

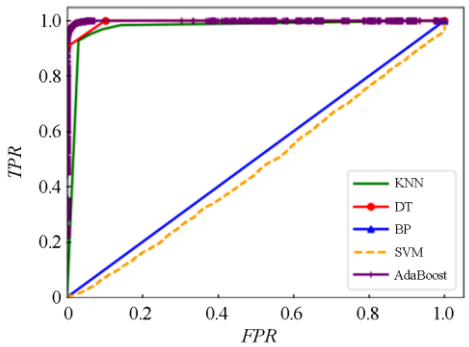
3.3.2 在 MIX 数据集上的对比分析

相比只含单一窃电样本数据集，包含混合窃

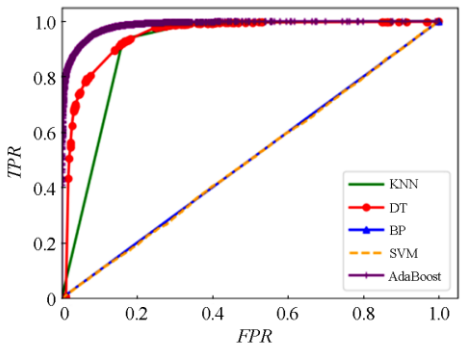
电样本数据集上的检测结果在实际应用中更具有意义。表 3 展示了 DT、KNN、BP、SVM 以及 AdaBoost 集成学习 5 种算法在数据集 MIX 测试集上的实验结果。



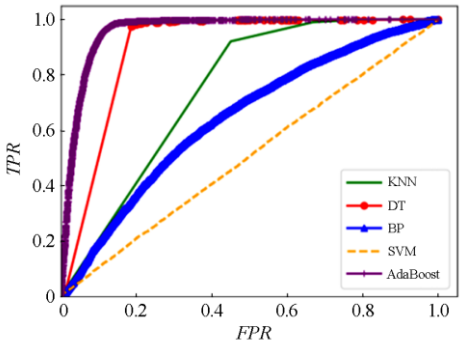
(a) ET1 上的 ROC 曲线



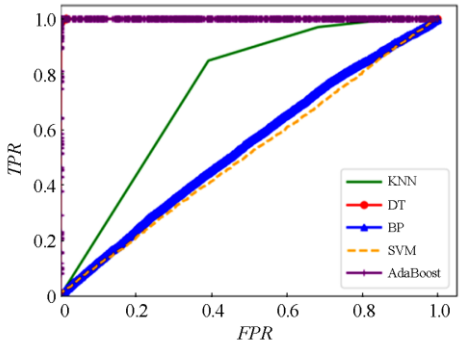
(b) ET2 上的 ROC 曲线



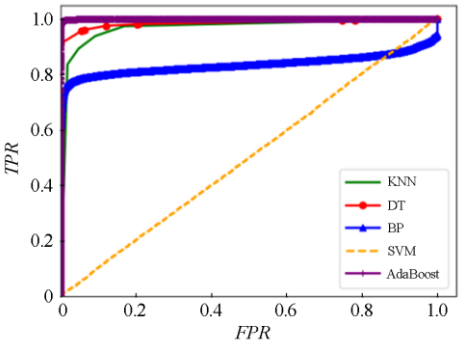
(c) ET3 上的 ROC 曲线



(d) ET4 上的 ROC 曲线



(e) ET5 上的 ROC 曲线



(f) ET6 上的 ROC 曲线

图 12　5 种算法在 6 个只含单一窃电样本数据集上的 ROC 曲线

Fig. 12 ROC curves of five algorithms on six containing single tamper samples

表 3　5 种算法在 MIX 上的测试结果

Table 3 Test results of five algorithms on MIX

类型	*ACC*/%					*AUC*/%				
	DT[11]	KNN[12]	BP[10]	SVM[11]	AdaBoost	DT[11]	KNN[12]	BP[10]	SVM[11]	AdaBoost
MIX	94.6	90.8	90.5	42.5	**96.5**	90.8	83.8	73.3	49.5	**96.5**

由表 3 可知，与只含单一窃电样本数据集相比，BP 和 SVM 算法在 MIX 上的 *ACC* 和 *AUC* 值几乎不变。DT、KNN 与基于 AdaBoost 集成学习 3 种算法的 *ACC* 和 *AUC* 值均出现了一定程度下降，但仍远高于 BP 和 SVM 算法。5 种算法在数据集 MIX 上的 ROC 曲线见图 13。

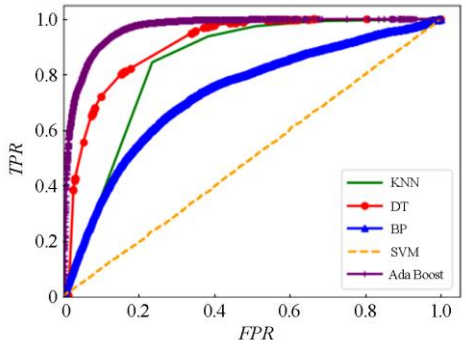


图 13 5 种算法在 MIX 上的 ROC 曲线

Fig. 13 ROC curves of five algorithms on MIX

3.4 灵敏性分析

为了说明窃电样本所占比例的不同对基于 AdaBoost 集成学习窃电检测模型的影响，在不同窃电样本占比下，分别得到 5 种算法的 *ACC* 和 *AUC* 值，结果分别如图 14 和图 15 所示。

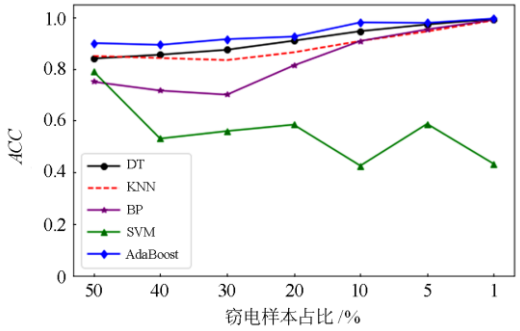


图 14 5 种算法在不同窃电样本占比上的 *ACC*

Fig. 14 *ACC* curves of five algorithms on MIX

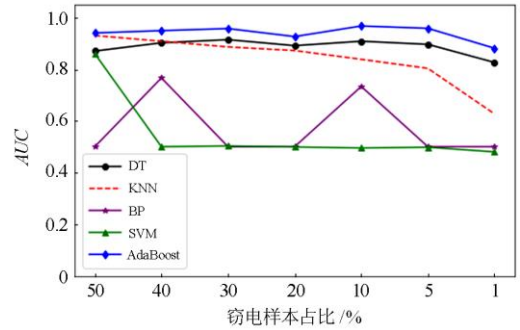


图 15 5 种算法在不同窃电样本占比上的 *AUC* 值

Fig. 15 *AUC* curves of five algorithms on MIX

由图 14 可知，随着窃电样本占比的减少，除 SVM 外其他 4 种算法的 *ACC* 值呈现明显地上升趋势。其中，AdaBoost、DT 和 KNN 3 种算法的值较为接近，且 AdaBoost 的 *ACC* 值始终最大。

图 15 显示了 5 种算法的 *AUC* 值。整体上，AdaBoost、DT 和 KNN 算法明显优于 BP 和 SVM 算法。随着窃电样本占比的减小，KNN 和 SVM 算法的 *AUC* 值呈下降趋势，而 AdaBoost 和 DT 算法的 *AUC* 值在 0.8~1.0 波动，且 AdaBoost 始终最大。

4 结论

本文提出了基于 AdaBoost 集成学习的窃电检测方法。采用决策树作为弱分类器，利用爱尔兰智能电表数据集进行对比，验证了本方法的精确性与有效性。后续可将多个类型弱学习器结合起来，进一步提高窃电检测模型的准确率与检测效率。

参考文献

[1] 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): 119-127.
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): 119-127.

[2] 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.

[3] JOKAR P, ARIANPOO N, LEUNG V C M. Electricity theft detection in AMI using customers' consumption patterns[J]. IEEE Transactions on Smart Grid, 2016, 7(1): 216-226.

[4] 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]. IEEETransactions on Smart Grid, 2017, 8(5): 2505-2516.

[5] HAN Wenlin, XIAO Yan. NFD: non-technical loss fraud detection in smart grid[J]. Computers & Security, 2017, 6(5): 187-201.

[6] 王昕，田猛，赵艳峰，等. 一种基于状态估计的新型窃电方法及对策研究[J]. 电力系统保护与控制，2016, 44(23): 141-146.
WANG Xin, TIAN Meng, ZHAO Yanfeng, et al. A kind of electricity theft based on state estimation and countermeasure[J]. Power System Protection and Control, 2016, 44(23): 141-146.

[7] AMIN S, SCHWARTZ G A, CARDENAS A A. Game-theoretic models of electricity theft detection in smart utility networks: providing new capabilities with

advanced metering infrastructure[J]. IEEE Control Systems, 2015, 35(1): 66-81.

[8] MYERSON R B. Game theory: analysis of conflict[M]. Cambridge, USA: Harvard University Press, 1991.

[9] 吴浩可, 雷霞, 黄涛, 等. 价差返还机制下售电公司博弈模型[J]. 电力系统保护与控制, 2019, 47(12): 84-92.
WU Haoke, LEI Xia, HUANG Tao, et al. A game-theoretic model for retail companies under the spring_rebate mechanism[J]. Power System Protection and Control, 2019, 47(12): 84-92.

[10] 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 & Instrumentation, 2018, 55(11): 35-40.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 李春阳, 王先培, 田猛, 等. AMI 环境下异常用电检测研究[J]. 计算机仿真, 2018, 35(8): 66-70.
LI Chunyang, WANG Xianpei, TIAN Meng, et al. Abnormal power consumption detection in AMI enviroment[J]. Power Simulation, 2018, 35(8): 66-70.

[16] 杨茂, 张罗宾. 基于数据驱动的超短期风电功率预测综述[J]. 电力系统保护与控制, 2019, 47(13): 171-186.
YANG Mao, ZHANG Luobin. Review on ultra-short term wind power forecasting based on data-driven approach[J]. Power System Protection and Control, 2019, 47(13): 171-186.

[17] 张禄, 李国昌, 陈艳霞, 等. 基于数据挖掘的电动汽车用户细分及价值评价方法[J]. 电力系统保护与控制, 2018, 46(22): 124-130.
ZHANG Lu, LI Guochang, CHEN Yanxia, et al. Customer segmentation and value evaluation method based on data mining for electric vehicles[J]. Power System Protection and Control, 2018, 46(22): 124-130.

[18] SUN Q, SHI L, NI Y, et al. An enhanced cascading failure model integrating data mining technique[J]. Protection and Control of Modern Power Systems, 2017, 2(1): 19-28. DOI: 10.1186/s41601-017-0035-3.

[19] ASHA K S, JAYA L A. Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy[J]. Protection and Control of Modern Power Systems, 2018, 3(3): 303-314. DOI: 10.1186/s41601-018- 0103-3.

[20] ZHENG K, CHEN Q, WANG Y, et al. A novel combined data-driven approach for electricity theft detection[J]. IEEE Transactions on Industrial Informatics, 2019, 15(3): 1809-1819.

[21] 张良均. Python 数据分析与挖掘实战[M]. 北京: 机械工业出版社. 2013: 61-62.

[22] CAO F, TAN Y, CAI M. Sparse algorithms of random weight networks and applications[J]. Expert Systems with Applications, 2014, 41(5): 2457-2462.

[23] 王爱平, 万国伟, 程志全, 等. 支持在线学习的增量式极端随机森林分类器[J]. 软件学报, 2011, 22(9): 2059-2074.
WANG Aiping, WAN Guowei, CHENG Zhiquan, et al. Incremental learning extremely random forest classifier for online learning[J]. Journal of Software, 2011, 22(9): 2059-2074.

[24] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 171-173.

[25] 庄池杰, 张斌, 胡军, 等. 基于无监督学习的电力用户异常用电模式检测[J]. 中国电机工程学报, 2016, 36(02): 379-387.
ZHUANG Chijie, ZHANG Bin, HU Jun, et al. Anomaly detection for power consumption patternsbased on unsupervised learning[J]. Proceedings of the CSEE, 2016, 36(2): 379-387.

[26] ISSDA. Data from the commission for energy regulation[EB/OL]. [2019-07-01]. http: //www.ucd.ie/issda/data/commissionforenergyregulationcer/.

收稿日期：2019-11-11；　修回日期：2020-01-17

作者简介：

游文霞(1978—)，女，博士，副教授，主要研究方向为电力系统优化，电力系统人工智能；E-mail: youwenxia@ctgu.edu.cn

申　坤(1994—)，男，通信作者，硕士研究生，主要研究方向为电力系统人工智能。E-mail: 437472944@qq.com

(编辑　许　威)