文章编号：1674-2974（2022）08-0138-11 DOI：10.16339/j.cnki.hdxbzkb.2022233

基于 CNN-LG 模型的窃电行为检测方法研究

卿柏元，陈珏羽†，李金瑾，蒋雯倩
（广西电网有限责任公司 计量中心，广西 南宁 530023）

摘 要：针对当前电网单一学习器窃电检测方法准确率低、实时性差且无特征提取的问题，提出一种基于卷积神经网络轻梯度提升机（CNN-LG）模型的窃电行为检测方法．通过卷积神经网络（CNN）提取用户用电数据电力特征，将提取特征输入以决策树为基学习器的轻梯度提升机（LG）分类器对数据进行训练，据此建立基于卷积神经网络轻梯度提升机模型的窃电行为检测方法．采用基于卷积神经网络轻梯度提升机模型对国家电网和爱尔兰智能能源径（ISET）数据集分别进行窃电行为检测．实验结果表明，本文提出方法可快速准确实现电网中各类窃电行为检测，相比于现有检测方法具有更高准确度、更优泛化性能和实时性．

关键词：窃电；决策树；用电数据；卷积神经网络；轻梯度提升机
中图分类号：TM715 **文献标志码**：A

Research on Detection Method of Electricity Theft Behavior Based on CNN-LG Model

QING Boyuan，CHEN Jueyu†，LI Jinjin，JIANG Wenqian
（Measurement Center of Guangxi Power Grid Co.，Ltd.，Nanning 530023，China）

Abstract：Focusing on the problems of low accuracy，poor real-time performance，and no feature extraction in the current grid single learner power-theft detection method，a power-theft behavior detection method based on the Convolutional Neural Network-Light Gradient Boosting Machine（CNN-LG）model is proposed．First，the power features of user electricity data are extracted through the Convolutional Neural Network（CNN），and the extracted features are input into the Light Gradient Boosting Machine（LightGBM，LG）classifier based on the decision tree in order to train the data．On this basis，a detection method of electricity theft based on the CNN-LG model is established．Finally，the State Grid Corporation of China and Irish Smart Energy Trail（ISET）datasets are used to conduct experiments to verify the accuracy and effectiveness of the method proposed in this paper．The experimental results show that the method proposed in this paper can quickly and accurately realize the detection of various power theft behaviors in the power grid．Compared with the existing detection methods，it has higher accuracy，better generalization performance，and real-time performance．

Key words：electricity theft；decision trees；electricity data；convolutional neural networks；Light Gradient Boosting Machine（LightGBM，LG）

＊ **收稿日期**：2021-09-03
基金项目：广西电网科技项目（GXKJXM20200020），Science and Technology Project of Guangxi Power Grid Co.，Ltd.（GXKJXM20200020）；国家自然科学基金资助项目（51777061），National Natural Science Foundation of China（51777061）
作者简介：卿柏元（1982—），男，湖南洞口人，广西电网有限责任公司高级工程师
† 通信联系人，E-mail：jueyuchen@qq.com

电力系统中配电网的电能损失分为技术损失(Technical Loss,TL)和非技术损失(Non-technical Loss,NTL)[1],造成NTL的原因多数与用户侧窃电相关.窃电用户的窃电非法行为不仅对电力公司造成巨大的经济损失、扰乱电力市场供电秩序,而且对电网安全稳定运行造成巨大风险[2].因此,研究如何提高窃电检测准确率,降低非技术损失,对于电力公司运营和社会发展具有重大价值与意义.

传统的窃电检测是利用人工现场稽查的方式对可疑用户进行排查,需耗费巨大人力成本且效率低.随着智能电网的发展,能源互联网作为智能电网逐渐演变的产物,拥有能源和信息流双向性的特征,而由智能电表、集中器、通信网络及数据管理系统组成的高级量测体系(Advanced Metering Infrastructure,AMI)作为能源互联网信息流的主要组成部分正逐步在电网建立与完善.随着AMI的快速发展,使得利用智能电表的海量数据进行窃电检测成为更加高效的检测方式[3].在AMI下利用智能电表数据对窃电行为进行检测的方法可分为以下3类[1].

第1类为基于电网状态的检测方法.该类方法通过分析配电网的拓扑结构,结合网络潮流计算、系统状态等理论,计算用户数据的理论值,再与实际量测值比较,实时检测窃电用户[4].文献[5]提出使用电力用户的有功和无功功率归一化残差检测和定位配电网中的异常用电.文献[6]提出基于状态估计和电源管理单元的窃电用户检测和定位模型,通过分析功率和电压的量测值偏差对窃电嫌疑用户定位.当前实际电网结构和设备种类较多、数据复杂、计算难度大,电网完整的网络拓扑和参数往往难以获取,且在配电网中安装额外设备辅助检测,不仅安装困难且需额外设备支出[7].

第2类为基于博弈论的检测方法.该类方法认为窃电用户与电力公司之间存在博弈,且可从博弈均衡中获得窃电用户和正常用户消费的不同分布[8].文献[9]构建纳什均衡模型,建立窃电用户集与供电企业间的Stackelberg博弈.文献[10]提出使用博弈论解决新型智能家居环境下的电力市场模型构建问题,由于基于博弈论的方法侧重于具有强大假设的理论分析,尚未得到实证的检验[8].

第3类为基于人工智能检测方法.该类方法是AMI下最为广泛的窃电行为检测方法,可基于用户负荷曲线和用电量的特征对窃电用户识别[11].针对无标签的用户数据,可采用基于聚类等无监督学习的方法对窃电行为检测,该类方法通过分析用户间的用电关系发现离群点,以此作为依据对窃电行为进行检测[12-13].而现阶段基于无监督学习的窃电检测方法因其参数难以设置,从而无法达到较高的检测精度,且难以处理大规模的高维数据[14].为克服无监督学习方法不足,通过采用有监督学习对带有标签的用户历史用电量数据进行学习,寻找异常用电模式,再对其他用户进行窃电检测.神经网络因其网络训练次数过多易出现过拟合[15],而当采用SVM或决策树方法时,若用电数据集中含有数据缺失等噪声时检测结果较差[16-18],对于用户用电高维数据,基于浅层结构的模型无法有效检测[19].上述基于有监督学习方法均采用单一学习器进行窃电检测,由于不同学习器预测结果可能存在差异,因此,基于单一学习器无法通过训练获取准确的检测模型[20-21].文献[22]采用XGboost集成学习方法检测窃电行为,但是该方法在处理海量用电数据时,无法实现准确预测分类,且消耗内存资源大,尤其在遍历分割点时,需进行分裂增益计算,导致模型训练时间较长.

本文针对电网中跨类杂糅窃电行为的快速、准确检测难题,首先,对AMI下采集的用户原始用电数据进行预处理,通过CNN对预处理用电数据进行特征提取;再以决策树为基学习器的LightGBM集成学习方法对数据训练获得窃电检测模型,据此建立基于卷积神经网络轻梯度提升机(Convolutional Neural Network-Light Gradient Boosting Machine, CNN-LG)模型的窃电行为检测方法;最后通过国家电网数据集和爱尔兰智能能源径(Irish Smart Energy Trail, ISET)数据集分别对本文提出方法的准确性和有效性进行验证与分析.

1　基于CNN-LG模型窃电行为检测算法

1.1　卷积神经网络

用户用电数据时间序列特征的准确提取是实现窃电用户识别的关键环节.卷积神经网络由输入、卷积层、池化层、全连接层及输出层组成[23].CNN模型框架如图1所示.CNN拥有表征学习能力,对输入数据能按其网络结构层层学习,基于CNN提取特征效果明显,对数据没有额外的特征工程要求.因此,本文采用CNN对用户用电数据特征自适应提取.

由图1可知,卷积层为CNN的核心组成模块,由一组平行特征图组成,通过卷积核对输入特征图进行卷积运算,得到输出特征图,该特征图中所有元素均通过同一个卷积核计算,即权值和偏置项共享.卷

积运算如下：

$$\boldsymbol{x}_j^r = \boldsymbol{f}\left(\sum \boldsymbol{x}_i^{r-1} \times \boldsymbol{k}_{i,j}^r + \boldsymbol{b}_j^r\right) \tag{1}$$

式中：$\boldsymbol{x}_j^r$为通过第r层卷积运算所得第j个输出特征图；$\boldsymbol{b}_j^r$表示第r层网络第j个卷积核的偏置；$\boldsymbol{k}_{i,j}^r$表示第r层与第i个输入特征图运算的第j个卷积核；$\boldsymbol{f}$为非线性激活函数. 为了提高网络的拟合能力与稀疏性，在激活函数的选择上，ReLU 函数相比 Sigmoid 函数有防止梯度弥散和计算速度快等优点，故模型中选择 ReLU 函数作为激活函数，其表达式为：

$$R_{\mathrm{eLU}}(\boldsymbol{x}) = \begin{cases} \boldsymbol{x}, & \boldsymbol{x} > 0 \\ 0, & \boldsymbol{x} \leqslant 0 \end{cases} \tag{2}$$

式中：$\boldsymbol{x}$为卷积运算后得到的数据.

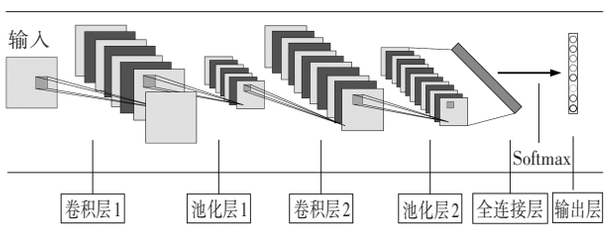


图 1 CNN模型框架

Fig.1 Architecture of the CNN model

池化层在CNN中用于缩小模型体积，提高计算速度，同时提高所提取特征的鲁棒性，在减少冗余特征量同时，保留用电行为主要特征，通过减少计算参量以达到降维效果，防止过拟合现象，提高模型泛化能力. 实际上池化操作为一种下采样操作，其操作包括最大池化、均值池化、随机池化等. 池化操作计算式为：

$$p(i,j) = \frac{1}{w^2} \sum_{u=(i-1)w+1}^{iw} \sum_{v=(j-1)w+1}^{jw} a(u,v) \tag{3}$$

式中：$a(u,v)$表示池化层输入矩阵中行列的值；$p(i,j)$表示池化层输出矩阵第i行j列的值；w表示参与集合区域的边值.

全连接层将CNN中最后一个池化层的所有神经元进行全连接操作，其模型可表示为：

$$\boldsymbol{y} = \boldsymbol{w}\boldsymbol{x} + \boldsymbol{b} \tag{4}$$

式中：$\boldsymbol{x}$为全连接层的输入；$\boldsymbol{w}$为权值矩阵；$\boldsymbol{b}$为偏置向量. 全连接层起到所学到的分布式特征映射到样本标记空间的作用.

1.2 LightGBM算法

LightGBM由Ke等于2017年提出[24]，该方法为Boosting算法重要成员，属于轻量级的提升决策树(Gradient Boosting Decision Tree，GBDT)算法，以使用决策树为学习算法的基分类器. LightGBM主要提升GBDT在处理高维度大数据时算法训练效率和准确度，采用分布式的算法框架，支持高效率并行训练，具有训练速度快、内存消耗低、准确度高及支持分布式计算以达到快速处理海量用户用电数据的优点. 算法主要通过基于直方图(Histogram)的决策树算法、带深度限制的按叶生长(Leaf-wise)策略、基于梯度的单边采样(Gradient-based one-side Sampling，GOSS)算法以及互斥特征捆绑(Exclusive Feature Bundling，EFB)算法进行优化.

直方图算法也称为Histogram算法，先把连续的浮点特征值离散化成k个整数，同时构造一个宽度为k的直方图. 直方图算法示意图如图2所示.

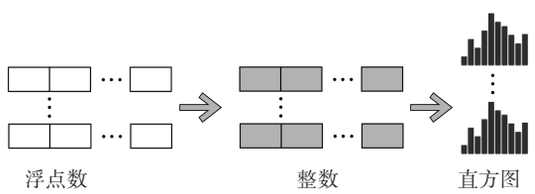


图 2 直方图算法示意图

Fig.2 Schematic diagram of Histogram algorithm

由图2可知，在遍历数据时，根据离散化后的值作为索引在直方图中累积统计量，当遍历一次数据后，在直方图中累积需要的统计量，再根据直方图的离散值，遍历寻找最优的分割点.

LightGBM算法使用按叶生长(Leaf-wise)策略，如图3所示. 每次在当前叶子节点中，寻找出分裂增益最大的叶子节点进行分裂，而其他结点不再分裂，这样可以提高精度，但缺点是可能会长出较深的决策树，产生过拟合. 为此，在Leaf-wise上增加max-depth参数进行限制，以控制模型的复杂度，同时防止过拟合现象发生.

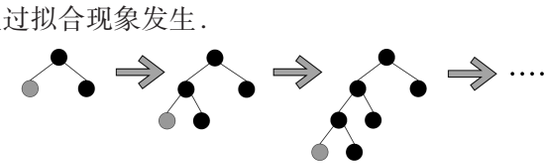


图 3 按叶生长(Leaf-wise)策略示意图

Fig.3 Schematic diagram of Leaf-wise tree growth strategy

LightGBM通过基于梯度的单边采样算法减少数据量和互斥特征捆绑算法减少特征量以优化模型训练效率. 基于梯度的单边采样算法，通过对样本采样的方法减少计算目标函数增益时的复杂度，在计算信息增益时，梯度更大的样本点占有更重要的作

用;在对样本进行下采样时,保留梯度较大的样本点,并随机去除梯度较小的样本点.具体做法:首先,将样本按照梯度排序,选出梯度最大的$a\times100\%$个样本;在剩下小梯度数据中随机选取$b\times100\%$个样本,在计算信息增益时,将选出来的$b\times100\%$小梯度样本的信息增益扩大$1-a/b$的倍数.互斥特征捆绑算法是将互斥特征绑在一起以减少特征维度,该算法可有效减少用于构建直方图的特征数量,降低计算复杂度,尤其当特征中包含大量稀疏特征时,LightGBM算法训练速度提升更为明显.

针对单一卷积神经网络模型,在窃电用户分类预测应用中存在功能单一导致准确率不足的问题,本文提出2种模型的融合算法,通过LightGBM代替卷积神经网络中的Softmax层,使网络中最后一层归一化处理,变成对用电特征集成学习分类的优化处理,从而实现窃电行为的准确识别.

1.3　CNN-LG窃电行为检测

为实现窃电行为准确检测,采用CNN提取时间序列的关联特征,将CNN结构中Softmax层用LightGBM代替,构建基于CNN-LG的窃电行为检测方法.CNN-LG窃电行为检测模型如图4所示,实现步骤如下.

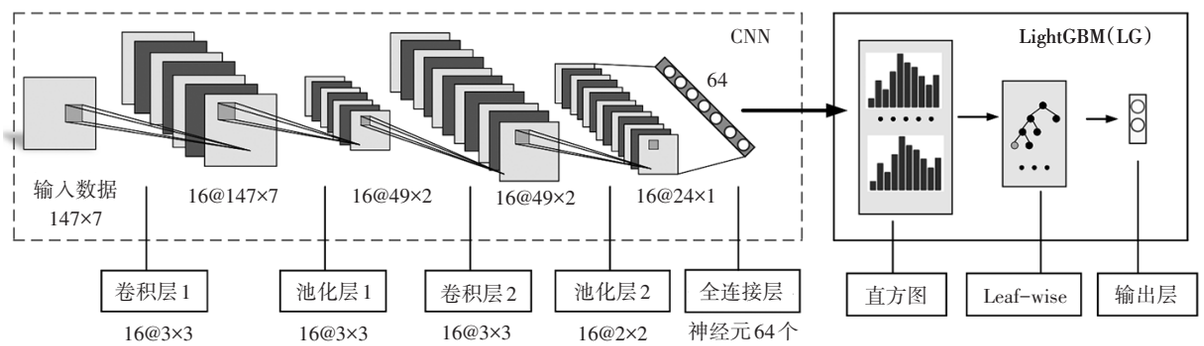


图4　CNN-LG窃电行为检测模型

Fig.4　CNN-LG electricity theft behavior detection model

1)将初始化卷积神经网络用预处理后的数据集通过两个卷积层和两个池化层进行预训练,并将训练好的权重固定,保存CNN模型参数.

2)将预处理后数据分为训练集、验证集和测试集.

3)设置网络训练迭代次数,利用训练集对网络进行训练,输出每次迭代的准确率,并与全局准确率比较,若准确率更高则更新权重,否则不更新.

4)利用已经训练完成的卷积神经网络对用电数据集进行特征提取.

5)将步骤4)中提取的特征输入至LightGBM模型,首先初始化n棵分类决策树,其中训练样例的权重为$1/n$;训练弱分类器$f(x)$,根据训练误差确定当前弱分类器$f(x)$的权重λ;当达到最大迭代次数,训练得到最终分类器,如式(5)所示.

$$f_n(x)=\lambda_0 f_0(x)+\lambda_1 f_1(x)+\lambda_2 f_2(x)+\cdots+\lambda_i f_i(x)+\cdots+\lambda_n f_n(x) \tag{5}$$

式中:n为算法迭代次数;i为第i次迭代,$0\leqslant i\leqslant n$.将测试集输入至训练完成的CNN-LG模型中,以此获得检测结果.

本文提出的CNN-LG算法流程图如图5所示.该算法有效利用卷积神经网络可自适应提取特征,且对数据无额外特征工程要求的优点,结合LightGBM具有并行训练效率高、训练速度快、内存消耗低、准确度高的优势,对国家电网中海量用户用电数据进行窃电行为检测.CNN-LG模型参数设置如表1所示.

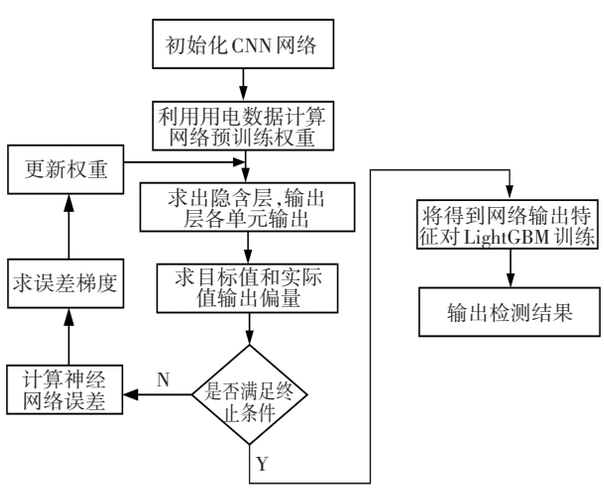


图5　CNN-LG算法流程图

Fig.5　CNN-LG algorithm flow chart

表1 CNN-LG模型参数设置

Tab.1 CNN-LG model parameter setting

参数名称	参数值
卷积核大小	3×3
卷积步长 stride	1
池化层1大小	3×3
池化层1填充 padding(即 max pooling 层1)	1
池化层2大小	2×2
池化层2填充 padding(即 max pooling 层2)	1
全连接层神经元个数	64
叶子数	25
树模型最大深度 max-depth	5
学习率	0.1

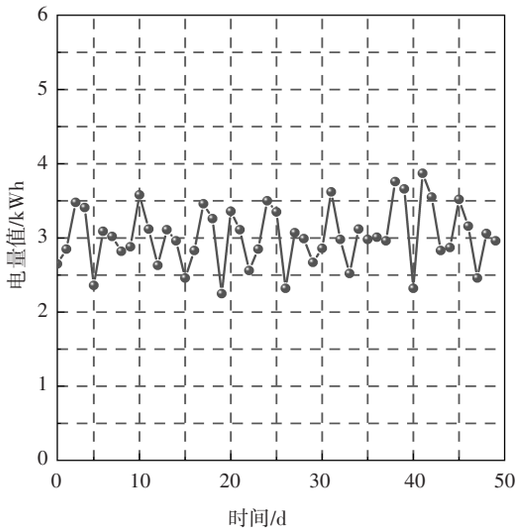
2 数据分析与预处理

2.1 数据分析

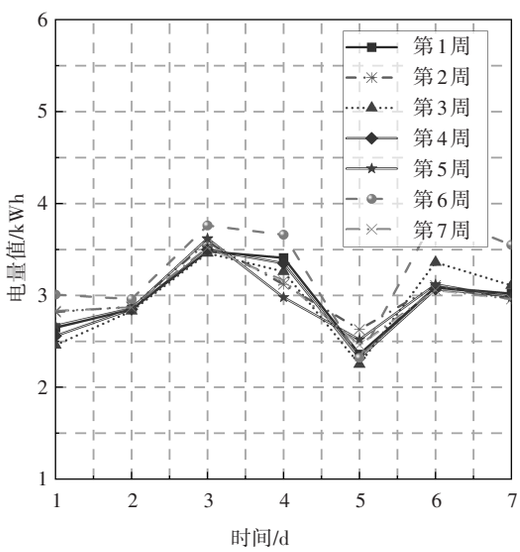
采用2种不同数据集验证本文提出方法的有效性和准确性. 第1种是国家电网(State Grid Corporation of China, SGCC)公开数据集,该数据集包含正常用户和窃电用户,提供是否窃电的标签[25-26]. 第2种为ISET数据集,该数据集被认为只包含正常用户. 本文通过篡改用电数据以模拟用户窃电,其中选取6种模式对正常数据进行模拟窃电攻击模式.

SGCC数据集由中国国家电网提供某地区的用户用电数据,该数据集包含从2014年1月—2016年10月,近147周42 372个用户每天的用电量. 该数据集分为正常用户和窃电用户,其中窃电用户为3 615个,占总用户数的8.53%;正常用户为38 757个,占总用户数的91.47%.

对SGCC数据集进一步分析得到正常用户和窃电用户电量值分别如图6和图7所示. 由图6(a)和7(a)可知,很难发现以天为单位的正常用户日用电量和窃电用户日用电量的用电规律;由图6(b)和7(b)可知,正常用户日用电量趋势大致相同,即第3 d用电量为峰值,第5 d用电量为谷值;窃电用户在前些周日用电量呈一定幅度的周期性波动,而从某周开始窃电用户的日用电量随时间变化呈下降趋势,并维持在较低用电量水平,且该147周数据呈现类似规律. 进一步提炼窃电用户行为特征可知,窃电用户初始用电量相比正常用户更多(正常用户日用电量为2~4 kWh,窃电用户日用电量为10~40 kWh),由此可知,窃电用户窃电行为收益更大、窃电动机更足.



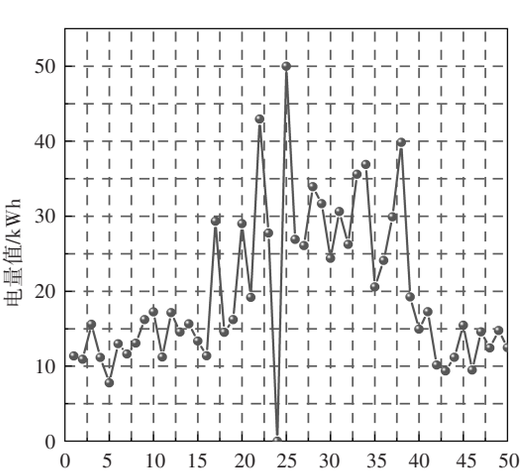
(a)以天为单位



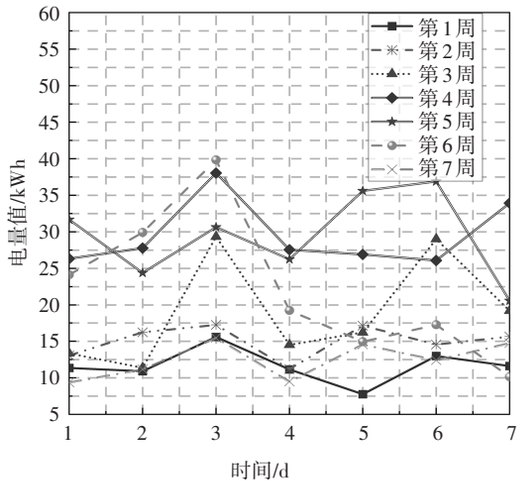
(b)以周为单位

图6 正常用户的电量图

Fig.6 Power consumption graph of normal users



(a)以天为单位



(b)以周为单位

图7　窃电用户的电量图

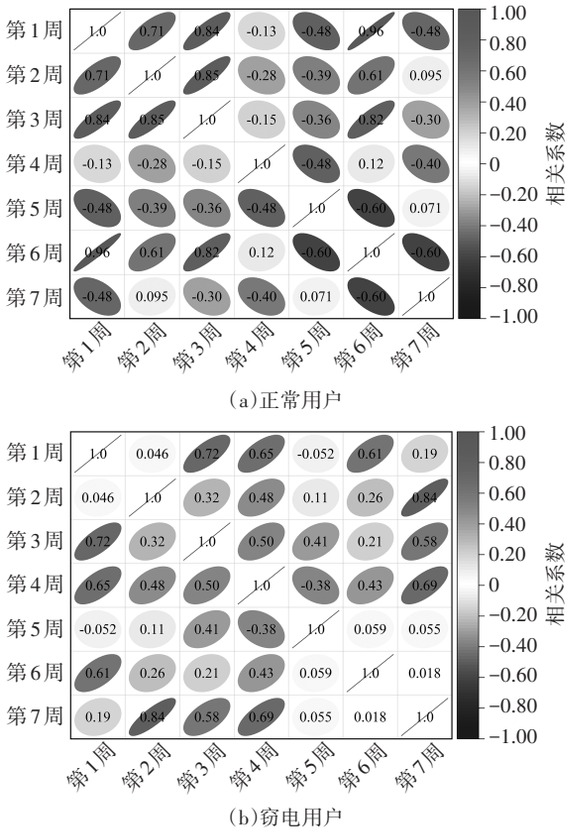
Fig.7　Power consumption graph of theft users

为进一步挖掘窃电用户和正常用户间用电量的区别,本文采用Pearson相关系数进行分析,计算式如下:

$$r=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\Big/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\bar{y})^2} \tag{6}$$

式中:$\bar{x}$、$\bar{y}$为样本均值.|r|值越大,其相关程度越高.图8为2种用户的相关系数矩阵.由图8可知,窃电用户的相关系数大多为正,且相关程度高,而正常用户则相反.

ISET数据集由爱尔兰CER(The Commission for Energy Regulation)组织的Electricity Customer Behaviour Trial提供,该组织通过智能电表记录居民和商业共5 000个用户,从2009年—2010年共533 d的用电数据[27].该数据集提供各用户每天每半小时用电量,可用向量$\boldsymbol{X}=[x_1, x_2, \cdots, x_{48}]$代表某个用户一天的用电量情况,该数据集被认为全部是正常用户的用电量数据.为了对窃电检测模型进行训练,本文采用6种攻击模式对该数据集进行篡改,模拟产生窃电行为[28].该6种篡改模式数学式如表2所示.其中,类型1表示所有读数乘以相同的随机生成的参数(0.2~0.8);类型2表示电表读数乘以不同的随机数α_t;类型3表示电表在t_1-t_2时间段内发送其抄表数,并在其他时间段发送零,t_1-t_2是一个随机定义的超过6 h的时间段;类型5表示电表将当天计量数据的平均值发送到数据管理系统;类型4在类型5的基础上乘以随机因子α_t;类型6表示窃电用户颠倒一天中的抄表顺序.



(a)正常用户

(b)窃电用户

图8　2种用户的相关系数矩阵

Fig.8　Pearson correlation coefficient of two kinds of users

表2　6种篡改模式

Tab.2　Six types of malicious samples

攻击类型	攻击方式
1	$Y_t=\alpha x_t,\ 0.2<\alpha<0.8$
2	$Y_t=\alpha_t x_t,\ 0.2<\alpha_t<0.8$
3	$Y_t=\beta x_t,\ \beta=\begin{cases}1, & t_1<t<t_2\\0, & \text{其他}\end{cases}$
4	$Y_t=\alpha_t\overline{X},\ 0.2<\alpha_t<0.8$
5	$Y_t=\overline{X}$
6	$Y_t=x_{48-t}$

注:Y_t为窃电后实际计量电量;x_t为正常用电量;$\overline{X}$为用电量均值;α为0.2~0.8的随机数;β为0或1;x_{48-t}中48-t中48-t表示1 d中以0.5 h为一个数据点(即48个点),减去当下时间点,x_{48-t}即可实现1 d中抄表的顺序的颠倒.

对任何企图窃电的用户来说,其目的是减少或消除自己所需支付的电费.设在t时段,经过篡改后电表所记录的用户用电量为x_t,对应时刻的单位电价为p_t,而用户实际的用电量为x_t^*,则

$$\sum_{t\in T}p_t x_t\leqslant\sum_{t\in T}p_t x_t^* \tag{7}$$

即篡改后的电费较原来更低.

用电用户可能会将某些时刻的电表读数直接篡

改为0,或按一定比例削减自己的用电量,也有可能在不改变总体用电量的同时对用电曲线进行移峰,以取得如式(7)所示的效果.6种攻击模式下产生的用电量曲线如图9所示.

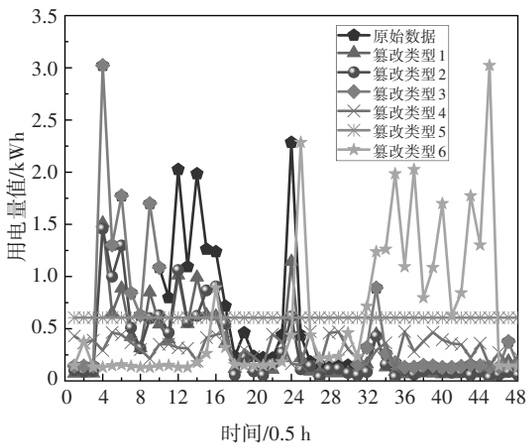


图9 6种攻击模式下产生的用电量曲线

Fig.9 Electricity curve generated by six attack modes

2.2 数据预处理

智能电能表采集的用户用电量数据中可能包含错误数据或有数据缺失,因此需对缺失数据进行补值处理,本文采用牛顿差值法对采集数据的缺失值进行处理.已知n个点对$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$的所有阶差商式分别为:

$$f[x_1,x]=\frac{f[x]-f[x_1]}{x-x_1}=\frac{f(x)-f(x_1)}{x-x_1} \tag{8}$$

$$f[x_2,x_1,x]=\frac{f[x_1,x]-f[x_2,x_1]}{x-x_2} \tag{9}$$

$$f[x_3,x_2,x_1,x]=\frac{f[x_2,x_1,x]-f[x_3,x_2,x_1]}{x-x_3} \tag{10}$$

$$\vdots \qquad\qquad \vdots$$

$$f[x_n,x_{n-1},...,x_1,x]=\frac{f[x_{n-1},x_{n-2},\cdots,x_1,x]-f[x_n,x_{n-1},...,x_2,x_1]}{x-x_n} \tag{11}$$

联立式(8)~式(11),建立差值多项式$f(x)$,有

$$\begin{aligned}f(x)=&f(x_1)+(x-x_1)f[x_2,x_1]+(x-x_1)(x-x_2)f[x_3,x_2,x_1]+\\&(x-x_1)(x-x_2)(x-x_3)f[x_4,x_3,x_2,x_1]+\cdots+\\&(x-x_1)(x-x_2)\cdots(x-x_{n-1})f[x_n,x_{n-1},\cdots,x_2,x_1]+\\&\underbrace{(x-x_1)(x-x_2)\cdots(x-x_n)f[x_n,x_{n-1},\cdots,x_1,x]}_{R(x)}=\\&P(x)+R(x)\end{aligned} \tag{12}$$

式中:$P(x)$表示牛顿差值逼近函数;$R(x)$表示误差函数.将缺失点x代入$f(x)$求得缺失值.

针对智能电能表采集的错误值(即离群点值),本文采用3σ定律对离群值进行修复,计算式如下:

$$f(x_i)=\begin{cases}\dfrac{x_{i-1}+x_{i+1}}{2}, & 若x_i>3(\sigma \boldsymbol{X}_i)且x_{i-1},x_{i+1}\neq \mathrm{NaN}\\ x_i, & 其他\end{cases} \tag{13}$$

式中:$\sigma(\boldsymbol{X}_i)$为向量$\boldsymbol{X}_i$的标准差;x_i为某用户在一个周期内的用电量值;NaN表示x_i为非数值符号或0时的情况.

为平衡样本数据,本文采用随机过采样方法,通过复制少数类示例来平衡数据,以消除数据不平衡带来的影响.

3 算例分析

3.1 模型评价指标构建

窃电行为检测本质上为二元分类问题,当算法完成对用户的分类后,需对检测方法的准确性进行评估.混淆矩阵是衡量方法优劣的重要工具,表3为窃电行为检测中的混淆矩阵.

表3 窃电行为检测中的混淆矩阵

Tab.3 Confusion matrix in the detection of electricity theft

用户	检测为异常用户	检测为正常用户
实际异常用户	TP(True Positive)	FN(False Negative)
实际正常用户	FP(False Positive)	TN(True Negative)

混淆矩阵将所有被检测用户按照实际归属和检测归属分为TP、FN、FP和TN这4类,TP和TN为模型检测下正确分类的部分,比例越高说明检测效果越好.命中率T_{PR}和误检率F_{PR}计算式分别如下:

$$T_{\mathrm{PR}}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{14}$$

$$F_{\mathrm{PR}}=\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{15}$$

由式(14)和式(15)可知,T_{PR}越接近1,F_{PR}越接近0,说明检测效果越好.通过表3的混淆矩阵定义召回率(R_{ecall})、精度(P_{recision})以及F_1值,对应式(16)~式(18)所示.

$$R_{\text{ecall}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{16}$$

式中:R_{ecall}表示在实际为正的样本中被预测为正样本的概率.

$$P_{\text{recision}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} \tag{18}$$

式中:P_{recision}表示被分为正例的样本中实际为正例的比例;F_1表示使用调和平均结合召回率和精度的指标.

ROC曲线下区域面积A_{UC}(Area Under ROC Curve)可通过接收者操作特征曲线(Receiver Operating Characteristic, ROC)下的各部分面积和求得,A_{UC}值越大越好,当A_{UC}=1时为理想分类器.A_{UC}计算式如下[25]:

$$A_{\text{UC}} = \frac{\sum_{i \in \text{正例}} R_{\text{anki}} - M(1 + M)/2}{MN} \tag{19}$$

式中:R_{anki}代表样本i的排序值;M为正样本的个数;N为负样本的个数.

平均精度均值M_{AP}(Mean Average Precision)用于评估模型检测性能.$M_{\text{AP@}N}$定义为在前N个嫌疑度最高的用户中,检测模型正确识别为窃电用户的平均精度均值[25],即

$$M_{\text{AP@}N} = \frac{\sum_{i=1}^{r} P@k_i}{r} \tag{20}$$

式中:r代表在前N个嫌疑度最高的用户中窃电用户的数量.$P@k_i$定义为:

$$P@k_i = Y_{k_i}/k_i \tag{21}$$

式中:Y_{k_i}表示在前k个嫌疑度最高的用户中正确识别窃电用户的数量;k_i(i=1, 2, 3, …, r)表示k的位置,本文采用$M_{\text{AP@100}}$和$M_{\text{AP@200}}$作为评价指标.

3.2　实验验证

为验证本文提出算法的有效性和准确性,实验平台采用64位6核心十二线程的Intel Core i7-8750H CPU@2.20 GHz,深度学习框架采用TensorFlow和Keras.实验数据为基于中国国家电网(SGCC)公开数据集和ISET公开数据集,具体介绍详见本文第2节.本文通过对CNN、LightGBM(该方法简称LG)、CNN+随机森林(CNN用于特征提取,随机森林用于分类,该方法简称CNN-RF)[29]、CNN+XGboost(CNN用于特征提取,XGboost用于分类,该方法简称CNN-XG)以及本文方法进行比较.

针对SGCC数据集的检测试验,各模型输入项为经预处理后的数据集.本文随机选取50%输入数据作为训练样本(其中40%作为训练集,10%作为验证集),余下50%数据作为测试样本.基于SGCC数据集下不同窃电检测方法的结果如表4所示.

表4　基于SGCC数据集下不同窃电检测方法的结果

Tab.4　Results of different electrical theft detection methods based on the SGCC dataset

方法	评价指标					
	A_{UC}	R_{ecall}	P_{recision}	F_1	$M_{\text{AP@100}}$	$M_{\text{AP@200}}$
CNN	0.792 08	0.871 43	0.613 29	0.850 61	0.861 21	0.830 36
LG	0.744 77	0.817 56	0.607 03	0.847 31	0.817 56	0.812 19
CNN-RF	0.796 24	0.879 87	0.796 95	0.849 19	0.927 44	0.910 29
CNN-XG	0.855 31	0.901 77	0.771 69	0.888 71	0.975 96	0.958 11
本文方法	0.863 33	0.904 18	0.796 95	0.894 17	0.990 29	0.972 09

由表4可知,本文采用的CNN-LG窃电行为检测模型在SGCC数据集下各项指标均优于其他几种方法,在CNN和LightGBM模型基础上检测精度均有一定提升,其中F_1值达到0.894 17,$M_{\text{AP@100}}$为0.990 29;A_{UC}值由原CNN的0.792 08和LightGBM的0.744 77提升至0.863 33.由此可知,本文提出方法有效利用二者优点,实现窃电行为准确检测,且CNN和LightGBM模型融合相比于CNN与其他两种集成学习方法融合检测效果更好.采用ROC曲线对所有方法的实验结果进行可视化描述,如图10所示,在ROC空间坐标中,越靠近左上的ROC曲线意味着在同样的检测命中率下造成的误检率(F_{PR})越低,检测效果越好;A_{UC}为ROC曲线下的各部分面积和,其值为窃电检测非常重要的评价指标,本文提出的基于CNN-LG模型的窃电行为检测方法A_{UC}值表现优于其他方法.

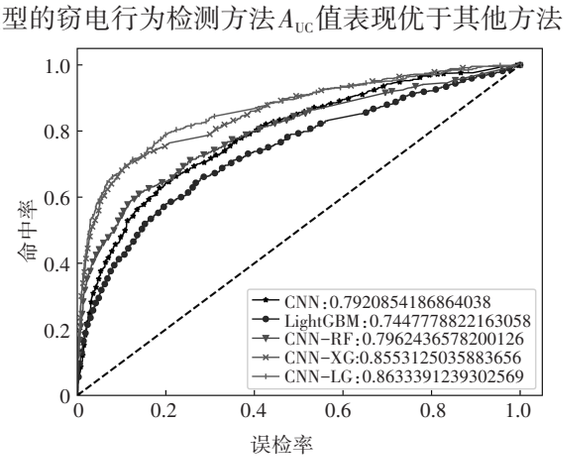


图10　SGCC数据集下不同方法的ROC曲线

Fig.10　The ROC curve of different methods under the SGCC dataset

窃电行为的快速检测为电力公司现场稽查提供依据.对本文提出的CNN-LG窃电行为检测模型的实时性进行验证,随机选取SGCC数据集50%作为实验数据,通过与CNN-XG、CNN-RF以及CNN-LG 3种融合模型训练时间进行比较,其中CNN训练模型的迭代次数为10.各方法的实验结果如图11所示,由图11可见,CNN-XG模型、CNN-RF模型、CNN-LG模型训练时间分别为93.86 s、81.74 s、42.47 s.由此可知,本文提出方法的模型训练时间远低于其他两种方法,在实际电网环境下的数据集中实时性表现更好.

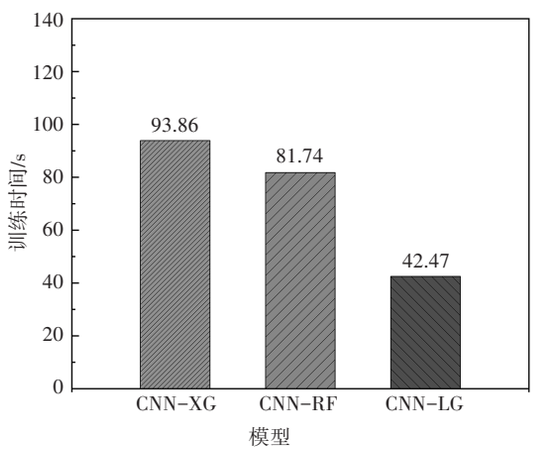


图11 SGCC数据集下不同方法的模型训练时间

Fig.11 The model training time of different methods under SGCC dataset

针对ISET数据集的检测试验,在用户533 d的样本中随机选择50%的样本,采用表2中的6种窃电手段,随机选择50%数据作为训练集(其中40%作为训练集,10%作为验证集),剩余50%数据作为测试集,以验证模型的有效性.基于ISET数据集下不同窃电检测方法的结果如表5所示.

表5 基于ISET数据集下不同窃电检测方法的结果

Tab.5 Results of different electrical theft detection methods based on the ISET dataset

方法	评价指标					
	A_{UC}	R_{ecall}	$P_{recision}$	F_1	$M_{AP@100}$	$M_{AP@200}$
CNN	0.837 31	0.760	0.913 58	0.796 91	0.922 56	0.897 05
LG	0.783 84	0.718	0.775 61	0.715 72	0.857 67	0.852 63
CNN-RF	0.935 25	0.876	0.860 91	0.875 88	0.996 91	0.984 31
CNN-XG	0.958 41	0.876	0.910 26	0.875 88	0.993 77	0.980 33
本文方法	0.963 44	0.878	0.925 11	0.877 75	1.000 00	0.992 16

由表5可知,本文采用的CNN-LG窃电行为检测模型在ISET数据集下各项指标均优于其他几种方法,在CNN和LightGBM模型基础上检测精度均有一定提升,其中F_1值达到0.877 75,$M_{AP@100}$为1,$M_{AP@200}$为0.992 16;A_{UC}值由原CNN的0.837 31和LightGBM的0.783 84提升至0.963 44,提升效果明显,而CNN-RF的A_{UC}值为0.935 25,CNN-XG的A_{UC}值为0.958 41,由此可知,CNN和LightGBM模型融合相比于CNN与其他两种集成学习方法融合检测效果更好.采用ROC曲线对所有方法的实验结果进行可视化描述,如图12所示,可明显看出,本文提出方法的A_{UC}值表现优于其他方法.

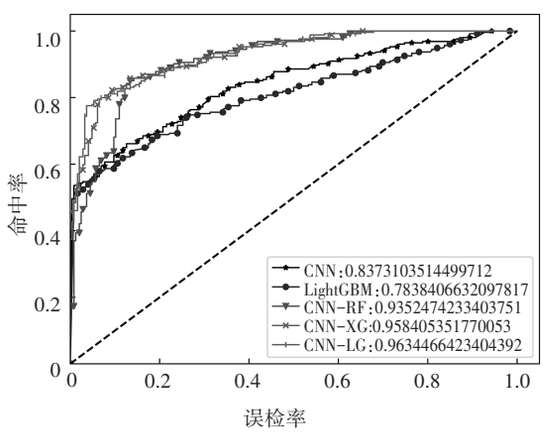


图12 ISET数据集下不同方法的ROC曲线

Fig.12 The ROC curve of different methods under the ISET dataset

对本文提出的基于CNN-LG模型窃电行为检测方法的实时性进行验证,随机选择经篡改后的ISET数据集50%数据作为实验数据,通过与CNN-XG、CNN-RF以及CNN-LG 3种融合模型训练时间进行比较,其中CNN训练模型的迭代次数为10.图13为ISET数据集下不同方法的模型训练时间,其中CNN-XG模型训练时间为16.37 s,CNN-RF模型训练时间为14.84 s,CNN-LG模型的训练时间为10.76 s.由此可知,本文提出方法的模型训练时间远低于其他两种方法,在ISET数据集中实时性表现更好.

通过上述2种不同数据集的实验可知,本文提出的基于CNN-LG模型窃电行为检测方法在实际电网数据集下检测准确度高,相比于其他几种方法,各项评价指标均更优,表现出良好的泛化性能,且该融合模型相比于其他融合模型实时性更好.

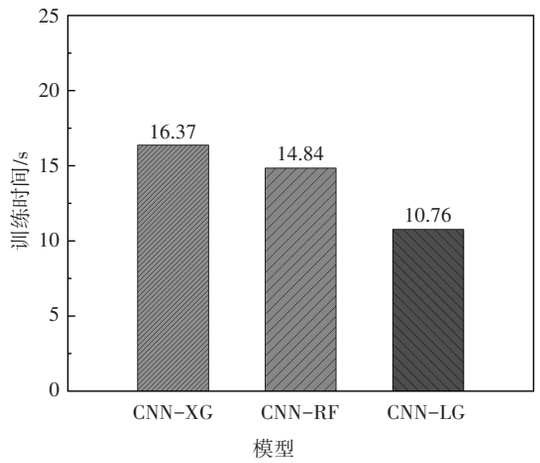


图 13　ISET数据集下不同方法的模型训练时间

Fig.13　The model training time of different methods under ISET dataset

4　结　论

本文提出基于CNN-LG模型的窃电行为检测方法,实现了窃电行为的快速准确检测,通过国家电网和ISET两种不同实际电网数据集对本文所提方法进行实验验证.结果表明,通过卷积神经网络和LightGBM的融合模型可有效利用二者优点对窃电行为进行快速准确检测,该模型可从用户用电数据准确提取电力特征,避免人为特征提取的不确定性和复杂性.LightGBM用于分类预测,进一步提高检测准确度,通过减少数据量和特征量提高检测效率,降低内存占用率以达到快速检测效果,且拥有更小的计算复杂度,在保证高效率的同时防止过拟合现象的出现,相比于现有单模型和融合模型方法,表现出更高准确度、良好的泛化性能以及更好的实时性.本文提出方法更适用于电网中各类窃电行为检测,有助于提高电力公司稽查效率,为电力公司在对非法用户窃电行为现场稽查取证时,提供有效的依据和可靠的目标.

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 Q X,ZHENG K D,KANG C Q,*et al*. Detection methods of abnormal electricity consumption behaviors: review and prospect [J]. Automation of Electric Power Systems,2018,42(17):189-199.(In Chinese)

[2]　王德文,杨凯华.基于生成式对抗网络的窃电检测数据生成方法[J].电网技术,2020,44(2):775-782.
WANG D W,YANG K H. A data generation method for electricity theft detection using generative adversarial network[J]. Power System Technology,2020,44(2):775-782.(In Chinese)

[3]　ANGELOS E W S,SAAVEDRA O R,CORTÉS O A C,*et al*. Detection and identification of abnormalities in customer consumptions in power distribution systems [J]. IEEE Transactions on Power Delivery,2011,26(4):2436-2442.

[4]　HUANG S C,LO Y L,LU C N. Non-technical loss detection using state estimation and analysis of variance[J]. IEEE Transactions on Power Systems,2013,28(3):2959-2966.

[5]　RAGGI L M R,TRINDADE F C L,CUNHA V C,*et al*. Non-technical loss identification by using data analytics and customer smart meters[J]. IEEE Transactions on Power Delivery,2020,35(6):2700-2710.

[6]　CARQUEX C, ROSENBERG C. Multi-timescale electricity theft detection and localization in distribution systems based on state estimation and PMU measurements[C]//Proceedings of the Ninth International Conference on Future Energy Systems. New York, USA: Association for Computing Machinery, 2015: 282-290.

[7]　KRISHNA V B,GUNTER C A,SANDERS W H. Evaluating detectors on optimal attack vectors that enable electricity theft and DER fraud[J]. IEEE Journal of Selected Topics in Signal Processing,2018,12(4):790-805.

[8]　AMIN S, SCHWARTZ G A, CARDENAS 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]　CÁRDENAS A A,AMIN S,SCHWARTZ G,*et al*. A game theory model for electricity theft detection and privacy-aware control in AMI systems[C]//2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton). Monticello, IL, USA: IEEE,2012:1830-1837.

[10]　LIU Y,HU S Y,HUANG H,*et al*. Game-theoretic market-driven smart home scheduling considering energy balancing[J]. IEEE Systems Journal,2017,11(2):910-921.

[11]　ZHENG K D, CHEN Q X, 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.

[12]　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(2):379-387.
ZHUANG C J, ZHANG B, HU J, *et al*. Anomaly detection for power consumption patterns based on unsupervised learning[J]. Proceedings of the CSEE,2016,36(2):379-387.(In Chinese)

[13]　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 C, ZHANG H J, JING Z M, *et al*. Study on the anti-electricity stealing based on outlier algorithm and the electricity in-

formation acquisition system[J]. Power System Protection and Control,2015,43(17):69-74.(In Chinese)

[14] 金晟, 苏盛, 曹一家, 等. 基于格兰杰归因分析的高损台区窃电检测[J]. 电力系统自动化, 2020, 44(23): 82-89.
JIN S, SU S, CAO Y J, *et al*. Electricity-theft detection for high-loss distribution area based on granger causality analysis[J]. Automation of Electric Power Systems, 2020, 44(23): 82-89. (In Chinese)

[15] 李晓峰,刘刚,卫晋,等. 基于卷积神经网络与特征选择的医疗图像误差预测算法[J]. 湖南大学学报(自然科学版), 2021, 48(4): 90-99.
LI X F, LIU G, WEI J, *et al*. Error prediction algorithm of medical image based on convolution neural network and feature selection [J]. Journal of Hunan University (Natural Sciences), 2021, 48(4):90-99. (In Chinese)

[16] NAGI J, YAP K S, TIONG S K, *et al*. Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system[J]. IEEE Transactions on Power Delivery, 2011, 26(2): 1284-1285.

[17] 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.

[18] 李明俊,张正豪,宋晓琳,等. 基于一种多分类半监督学习算法的驾驶风格分类模型[J]. 湖南大学学报(自然科学版), 2020, 47(4): 10-15.
LI M J, ZHANG Z H, SONG X L, *et al*. Driving style classification model based on a multi-label semi-supervised learning algorithm [J]. Journal of Hunan University (Natural Sciences), 2020, 47(4): 10-15. (In Chinese)

[19] HU T Y, GUO Q L, SHEN X W, *et al*. Utilizing unlabeled data to detect electricity fraud in AMI: a semisupervised deep learning approach[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(11): 3287-3299.

[20] 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 W X, SHEN K, YANG N, *et al*. Research on electricity theft detection based on AdaBoost ensemble learning[J]. Power System Protection and Control, 2020, 48(19): 151-159. (In Chinese)

[21] 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 W X, SHEN K, YANG N, *et al*. Electricity theft detection based on Bagging heterogeneous ensemble learning[J]. Automation of Electric Power Systems, 2021, 45(2): 105-113. (In Chinese)

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

[23] 周飞燕,金林鹏,董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251.
ZHOU F Y, JIN L P, DONG J. Review of convolutional neural network [J]. Chinese Journal of Computers, 2017, 40(6): 1229-1251. (In Chinese)

[24] KE G, MENG Q, FINLEY T, *et al*. LightGBM: a highly efficient gradient boosting decision tree[C]//Advances in Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017: 1-9.

[25] ZHENG Z B, YANG Y T, NIU X D, *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.

[26] TAKIDDIN A, ISMAIL M, NABIL M, *et al*. Detecting electricity theft cyber-attacks in AMI networks using deep vector embeddings [J]. IEEE Systems Journal, 2021, 15(3): 4189-4198.

[27] Irish Smart Energy Trial. Data from the commission for energy regulation (CER)-smart metering project[EB/OL]. [2012-01-13]. http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

[28] 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.

[29] LI S, HAN Y H, YAO X, *et al*. Electricity theft detection in power grids with deep learning and random forests[J]. Journal of Electrical and Computer Engineering, 2019: 4136874.