基于LOF+SVM的异常用电用户分阶段识别方法

顾臻[1]，庄葛巍[1]，贺青[1]，周磊[1]，安佰龙[1]，段艳[2]
（1.国网上海市电力公司 电力科学研究院，上海 200051；
2.同济大学 汽车学院 车辆工程系，上海 201804）

摘要：准确的电力异常用户识别方法能为供电企业锁定存在窃电行为或其他违规行为的电力用户提供参考。大多数基于机器学习的异常识别模型采用了无监督算法，但模型的准确度还较低。针对上述问题，提出一种结合无监督的局部离群因子（LOF）算法与有监督的支持向量机（SVM）算法的两阶段异常用电用户识别方法。基于分析异常电能表区别于正常电能表的电流电压表现，构建异常识别模型的输入特征；采用无监督的LOF算法进行采样，筛选出可疑样本交给人工进行标记，然后利用标记样本训练有监督的SVM模型；在之后的检测工作中，直接将LOF算法筛选出可疑样本交给SVM模型进行识别。实例结果表明，该方法对电力异常用户的识别准确度高，对供电企业的窃电稽查工作具有指导意义。

关键词：电力异常用户识别；机器学习；局部离群因子（LOF）；支持向量机（SVM）
中图分类号：TM28 **文献标识码**：A **DOI**：10.19457/j.1001-2095.dqcd23988

Phased Identification Method of Abnormal Electricity Users Based on LOF+SVM

GU Zhen[1]，ZHUANG Gewei[1]，HE Qing[1]，ZHOU Lei[1]，AN Bailong[1]，DUAN Yan[2]
（1. *Power Science Research Institute*，*State Grid Shanghai Electric Power Company*，*Shanghai* 200051，*China*；
2. *Department of Vehicle Engineering*，*School of Automobile*，*Tongji University*，*Shanghai* 201804，*China*）

Abstract: Accurate identification method of abnormal electricity users can provide reference for power supply enterprises to lock in electricity theft or other violations of power users. Most abnormal user identification models based on machine learning adopt unsupervised algorithms，but the accuracy of the models is low. To solve the above problems，a two-stage abnormal power user identification method combining unsupervised local outlier factor（LOF）algorithm and supervised support vector machine（SVM）algorithm was proposed. Based on the analysis of the current and voltage performance of the abnormal energy meter different from the normal energy meter，the input characteristics of the abnormal identification model were constructed. The LOF algorithm was used to sample，and the suspicious samples were selected and handed over to manual labeling. Then the supervised SVM model was trained by the labeled samples. In the subsequent detection work，the suspicious samples screened by LOF algorithm were directly sent to the SVM model for identification. The example results show that this method has high identification accuracy for power abnormal users，and has guiding significance for the power stealing inspection of power supply enterprises.

Key words: abnormal electricity user identification；machine learning；local outlier factor（LOF）；support vector machine（SVM）

随着智能电网的快速发展以及用户用电需求的不断增加，社会中出现了各种窃电行为，不仅给电网带来了直接经济损失，而且影响电网安全稳定运转[1]。

智能电网的推广应用过程中积累了海量的用户用电数据，伴随着机器学习在异常识别领域的兴起，如何利用机器学习方法识别非法窃电行为逐渐成为研究热点：曾虎[2]基于聚类分析与支持向量机回归的相关技术，提出了基于电流、电压、电量等电能计量数据的窃电检测模型，吴迪等[3]分析了电气特征参量、窃电方法及本质之间的关系，提出了基于大数据的防窃电模型与方法，李宁等[4]提出了基于电量离群点挖掘的窃电辨识方法。程俊文等[5]提出了一种基于K-means

作者简介：顾臻（1989—），男，硕士，工程师，Email：642768767@qq.com

聚类算法以及用电信息采集系统所采参数，建立多维特征因子关联模型，确定窃电嫌疑用户的方法。邓明斌等[6]提出了一种检测方法，该方法对异常用电的表现进行特征提取，基于每个特征采用机器学习方法学习用电异常的判决阈值，针对每一个具体特征均进行了独立的异常检测。

上述异常用户识别模型实现了对窃漏电行为的自动识别，但模型的准确度与人工检测的准确度还有一定差距。异常检测中的正常样本与异常样本的比例严重倾斜，机器学习中的无监督算法在处理类不平衡数据方面比有监督算法更具优势，所以相关研究多从无监督角度出发[7]。而有监督异常检测方法可以实现更高的准确率，前提是在训练模型时，保证训练样本中的异常样本占比足以让模型学到异常样本的特征。

针对上述问题，提出结合无监督局部离群因子（local outlier factor，LOF）和有监督支持向量机（support vector machine，SVM）的异常用户识别方法：在模型构建阶段，首先利用无监督的LOF算法进行采样，将采集到的异常样本交由人工进行检测并标记，以获得有监督模型的训练样本；然后利用有标记的样本训练有监督式SVM异常识别模型；在应用阶段，采用分阶段识别方法，首先利用LOF算法筛选出异常样本，然后将异常样本交给有监督式SVM异常识别模型进一步识别。通过实例比对了不同算法的准确度、召回率等指标，最终验证了结合无监督LOF与有监督SVM模型的分阶段识别方法的有效性。

1　算法简介

1.1　LOF算法

LOF算法是一种无监督异常检测算法。该算法通过计算给定数据点相对于其邻域的局部密度偏差而实现异常检测[8-9]。

LOF算法中核心定义如下：

1）点p与点o之间的距离：$d(p,o)$。

2）第k距离：$k\_distance$。对于数据集D中的点o，点p的第k距离$d_k(p)$满足如下条件：

$$d_k(p)=d(p,o) \tag{1}$$

在数据集D中至少有不包括点p在内的k个点$o'\in D\{x\neq p\}$，满足：

$$d(p,o')\leqslant d(p,o) \tag{2}$$

在数据集D中最多有不包括点p在内的$k-1$个点$o'\in D\{x\neq p\}$，满足：

$$d(p,o')<d(p,o) \tag{3}$$

图1为点p的$d_k(p)$，k=4时候的示意图。

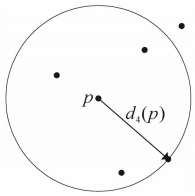

图1　点p的$d_4(p)$示意图

Fig.1　The schematic for $d_4(p)$of point p

3）点p的第k距离邻域：$N_k(p)$，即点p的第k邻域内（包括第k距离）的所有点的集合，$N_k(p)$中点的个数$|N_k(p)|\geqslant k$。

4）可达距离：$reach\_distance$。点p到点o的可达距离定义为

$$reach\_distance_k(p,o)=\max\{k\_distance(p),d(p,o)\} \tag{4}$$

图2为点p_1到点o的第5可达距离为$d(p_1,o)$，点p_2到点o的第5可达距离为$d_5(p_2)$，即

$$\begin{cases}reach\_distance_k(p_1,o)=d(p_1,o)\\reach\_distance_k(p_2,o)=d_5(p_2)\end{cases}$$

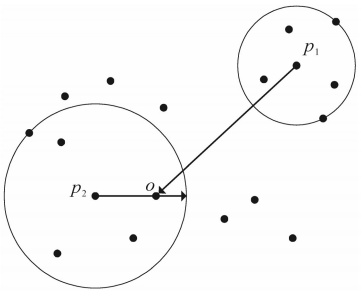

图2　点p_1与点p_2的可达距离

Fig.2　The reach distance of point p_1 and p_2

5）局部可达密度。点p的局部可达密度定义为点p的第k邻域内的点到点p的平均可达距离的倒数，公式为

$$lrd_k(p)=1/\frac{\sum_{o\in N_k(p)}reach\_distance_k(p,o)}{|N_k(p)|} \tag{5}$$

点p的局部离群因子定义为点p的邻域点的局部可达密度与点p的局部可达密度之比的平均数，公式为

$$LOF_k(p)=\frac{\sum_{o\in N_k(p)}\frac{lrd_k(o)}{lrd_k(p)}}{|N_k(p)|} \tag{6}$$

当$LOF_k(p)$的比值接近1，说明点p可能与邻域点同属一簇；如果此比值小于1，表明p可能为密集点；若此比值大于1，表明点p为异常点的可能性增加。

1.2 SVM算法

SVM算法是一种有监督分类算法[10]，其基本思想是在样本空间内寻找一个划分超平面，将不同类别的样本分开。

在样本空间中将给定的训练样本集：

$$O=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}\quad y_i\in\{-1,+1\} \tag{7}$$

划分开的划分超平面的线性方程为

$$\boldsymbol{\omega}^{\mathrm{T}}x+b=0\quad \boldsymbol{\omega}=\{\omega_1,\omega_2,\cdots,\omega_d\} \tag{8}$$

式中：$\boldsymbol{\omega}$为划分超平面的法向量；b为位移项。

记超平面为$(\boldsymbol{\omega},b)$，则样本空间中任意点x到超平面$(\boldsymbol{\omega},b)$的距离为

$$r=|\boldsymbol{\omega}^{\mathrm{T}}x+b|/\|\boldsymbol{\omega}\| \tag{9}$$

对于$(x_i,y_i)\in O$，若$y_i=+1$，则有$\boldsymbol{\omega}^{\mathrm{T}}x_i+b>0$；若有$y_i=-1$，则有$\boldsymbol{\omega}^{\mathrm{T}}x_i+b<0$，令：

$$\begin{cases}\boldsymbol{\omega}^{\mathrm{T}}x_i+b\geqslant+1 & y_i=+1\\ \boldsymbol{\omega}^{\mathrm{T}}x_i+b\leqslant-1 & y_i=-1\end{cases} \tag{10}$$

离超平面最近的训练样本点使得上述公式等号成立，它们即为支持向量(support vector)，两个异类支持向量到超平面的距离之和定义为间隔γ，公式为

$$\gamma=2/\|\boldsymbol{\omega}\| \tag{11}$$

当间隔γ取最大值，即

$$\begin{aligned}&\max_{\boldsymbol{\omega},b}\ 2/\|\boldsymbol{\omega}\|\\ &\text{s.t.}\ \ y_i(\boldsymbol{\omega}^{\mathrm{T}}x_i+b)\geqslant1\quad i=1,2,\cdots,m\end{aligned} \tag{12}$$

则可找到“最大间隔”的划分超平面。

当样本空间内的样本参数x的维度扩展到高维，SVM算法同样适用。

2 异常用电用户分阶段识别方法构建

2.1 构建用于识别用户类别的特征

2.1.1 异常用户用电特征分析

针对每个用户的电能表，电力部门每隔15 min对其A，B，C三相的电压、电流值进行采集，所以在一天的采集数据中，该电能表的三个相位都将预期获得96个电流采集数据和96个电压采集数据。用户的电流、电压值是由电力部门直接采集的数据，但在实际采集过程中不可避免存在缺失值，基于不完整的电流和电压数据计算而来的用电量和功率等特征可能与实际值存在误差，不能准确反映用户的用电特征，而直接针对电流、电压数据进行用电特征分析的过程能够避免上述误差的产生。在对异常用户的用电特征进行分析时，选取任意一天采集数据较完整的数据作为分析源(每一相的电流或电压采集数据的个数不低于90个)，通过分析用户用电时的电流电压数据是否表现异常，判定用户的窃电嫌疑。

1)电压分析。利用三相电压的采集数据作为用户用电特征参量，通过判断当前电能表是否处于欠压或失压状态，来判定用户的窃电嫌疑系数：电能表在正常用电状态下的电压采集值近似等于额定电压，若电能表的电压采集值低于设定的阈值电压(一般为额定电压的90%)，表明该电能表处于欠压或失压状态，该用户可能存在窃电嫌疑。

2)电流分析。利用三相电流的采集数据作为用户用电特征参量，通过判断当前电能表是否处于欠流或失流状态，来判定用户窃电嫌疑系数：电能表在正常用电状态下的三相电流值一般不为零，而异常用户电能表的三相电流值出现连续多个零值或负值，即该电能表处于失流状态；正常电能表在任意时刻的任意一相的电流值与该时刻的三相电流的平均值的比值接近1，而异常用户电能表的该项比值一般大于1.5及以上，即该电能表处于欠流状态，该对应用户可能存在窃电嫌疑。

2.1.2 基于统计的数据特征提取

基于异常用户与正常用户的用电特征有所不同，针对某用户电能表在某一天的电流及电压采集数据信息，构建新特征以反映该用户电能表在该天的电流电压表现，所构建的新特征如表1所示。

表1 用户在某一天的用电特征

Tab.1 Power consumption characteristics of users on a certain day

特征类	特征参数	说明
I/I_{average}	I_A/I_{average} I_B/I_{average} I_C/I_{average}	针对每一相：计算当前数据采集时刻的该相电流值与三相电流平均值的比值，然后取一天中不同时刻计算值的最大值作为该特征值
$Ratio_{I=0}$	$Ratio_{I_A=0}$ $Ratio_{I_B=0}$ $Ratio_{I_C=0}$	针对每一相：计算一天内电流采集数值为0的数据个数占该天电流数据总采集个数的比例
$Ratio_{I<0}$	$Ratio_{I_A<0}$ $Ratio_{I_B<0}$ $Ratio_{I_C<0}$	针对每一相：计算一天内电流采集数值小于0的数据个数占该天电流总采集个数的比例
$Ratio_{U=0}$	$Ratio_{U_A=0}$ $Ratio_{U_B=0}$ $Ratio_{U_C=0}$	针对每一相：计算一天内电压采集数值为0的数据个数占该天电压总采集个数的比例
$Ratio_{U<90\%U_0}$	$Ratio_{U_A<90\%U_0}$ $Ratio_{U_B<90\%U_0}$ $Ratio_{U_C<90\%U_0}$	针对每一相：计算一天内电压采集数值小于额定电压U_0的90%的数据个数占该天电压总采集个数的比例

根据表1中的特征构建方式，某用户电能表在某一天的用电情况可以由5×3=15个特征表示。假定对应同一用户电能表的每一天的用

电特征值具有相同的分布情况，在所有用户的电能表均处于正常状态且状态一致时，同一电能表在同一特征类里的A，B，C三相位的特征参数值无较大差异，应该具有相同的分布特征，例如对于同一用户的I/I_{average}特征类而言，其I_A/I_{average}，I_B/I_{average}，I_C/I_{average}三个特征参数在理想情况下应该满足同分布，所以在分析的过程中忽略相位差异，将I_A/I_{average}，I_B/I_{average}，I_C/I_{average}三个特征参数视为I/I_{average}的三个计算数值，并不做区别对待。

电力部门对用户窃电行为进行排查的工作周期一般为一个月，由于电能表在一个月内的采集数据是海量的，为了在海量的特征数据中提取分布特征，对同一用户电能表的特征数据进行相关统计量的计算：针对同一电能表，首先基于以天为单位的电流电压采集数据，计算该电能表与每天的电流电压采集数据相对应的15个用电特征参数值，得到一个月内的特征参数集，然后基于该数据集，忽视相位差别，分别统计I/I_{average}，$Ratio_{I=0}$，$Ratio_{I<0}$，$Ratio_{U=0}$，$Ratio_{U<90\%U_0}$这5个特征类的最大值Max( )、最小值Min( )、期望E( )和中位数Median( )这4个统计量，用于描述该用户在该月的用电情况，得到以月为单位对用户进行窃电嫌疑判断的5×4=20个特征值，即对于用户X，其特征向量$\boldsymbol{T}$表示为

$$\boldsymbol{T}=\begin{bmatrix}\text{Max}(I/I_{\text{average}}) & \text{Max}(Ratio_{I=0}) & \text{Max}(Ratio_{I<0}) & \text{Max}(Ratio_{U=0}) & \text{Max}(Ratio_{U<90\%U_0})\\ \text{Min}(I/I_{\text{average}}) & \text{Min}(Ratio_{I=0}) & \text{Min}(Ratio_{I<0}) & \text{Min}(Ratio_{U=0}) & \text{Min}(Ratio_{U<90\%U_0})\\ \text{E}(I/I_{\text{average}}) & \text{E}(Ratio_{I=0}) & \text{E}(Ratio_{I<0}) & \text{E}(Ratio_{U=0}) & \text{E}(Ratio_{U<90\%U_0})\\ \text{Median}(I/I_{\text{average}}) & \text{Median}(Ratio_{I=0}) & \text{Median}(Ratio_{I<0}) & \text{Median}(Ratio_{U=0}) & \text{Median}(Ratio_{U<90\%U_0})\end{bmatrix} \tag{13}$$

2.2　识别方法构建

异常识别模型的训练过程如图3所示，大量的无标记样本进入无监督采样阶段，通过LOF模型筛选出可疑样本集，这些可疑样本中既有被LOF模型正确识别的异常样本，也有被LOF模型错误识别的正常样本，可疑样本集中异常样本占比相比最初的无标记样本集高，减轻了类的不平衡问题。为了获得正确标签，将这些可疑样本交给专家进行人工标记，专家标记的过程可以被视为人工检查的过程，即对可疑用户进行窃电行为的排查，该过程可以降低人工检查的工作量，同时能够获取用于训练SVM模型的标记样本。

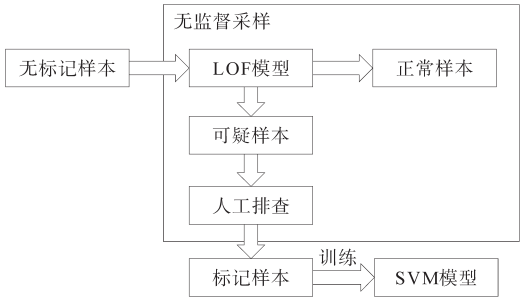

图3　异常用户识别模型的训练过程

Fig.3　Training process of abnormal user identification model

异常识别模型的应用过程如图4所示，首先利用LOF模型对待检测样本进行初步识别，将检测样本划分为正常样本和可疑样本，然后利用有监督式SVM异常识别模型对可疑样本进行下一阶段的识别，将可疑样本标记为正常或异常样本。

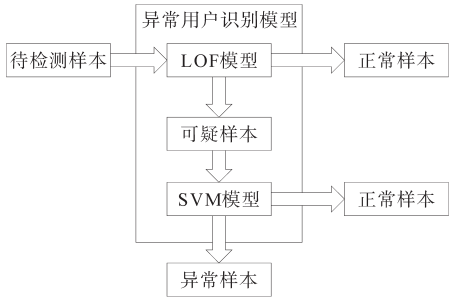

图4　异常用户识别模型的应用过程

Fig.4　Application process of abnormal user identification model

3　算例分析

3.1　数据特征分析及数据预处理

数据源自国家电网对上海地区用户电能表的电流电压采集数据，单个用户电能表在一个月内的电流电压采集数据为一个样本信息，实验过程随机选取了3 000名用户于2020年6月份的电流电压采集数据，即选取了3 000个实验样本，其中异常用户样本占比为0.055。

基于统计方法，对每个样本的采集数据进行特征构建，异常样本和正常样本的特征矩阵示例如表2和表3所示：表2为异常样本的特征示例，用户表号为“6 000 279 226”，该样本输入模型的特征值为4×5的特征矩阵，经过人工核实，该用户电能表出现异常的原因为B相熔丝坏掉导致失压；表3为正常样本的特征示例，用户表号为“1 188 125 655”，该样本输入模型的特征值如表3所示，经过人工核实，该用户为正常用户。

表2 异常用户样本特征示例

Tab.2 Example of abnormal user sample characteristics

	最大值	最小值	期望	中位数
I/I_{average}	1.022 236	0.000 000	0.594 711	0.690 099
$Ratio_{I=0}$	0.729 167	0.000 000	0.105 769	0.000 000
$Ratio_{I<0}$	0.000 000	0.000 000	0.000 000	0.000 000
$Ratio_{U=0}$	0.000 000	0.000 000	0.000 000	0.000 000
$Ratio_{U<90\%U_0}$	1.000 000	0.000 000	0.384 615	0.000 000

表3 正常用户样本特征示例

Tab.3 Example of normal user sample characteristics

	最大值	最小值	期望	中位数
I/I_{average}	1.393 414	0.591 050	0.920 289	0.897 789
$Ratio_{I=0}$	0.031 250	0.000 000	0.005 859	0.000 000
$Ratio_{I<0}$	0.000 000	0.000 000	0.000 000	0.000 000
$Ratio_{U=0}$	0.000 000	0.000 000	0.000 000	0.000 000
$Ratio_{U<90\%U_0}$	0.000 000	0.000 000	0.000 000	0.000 000

在模型训练前，为了减小离群点对输入的干扰，对样本特征数据进行0～1标准化处理[11]。

3.2 不平衡样本情况下的无监督识别结果

在模型训练与测试阶段，抽取3 000个样本构建2∶1的训练集和测试集，为了得到无偏的模型测试结果，采用分层抽样的方法保证训练集和测试集的异常样本占比都为5.5%。采用上述样本数据，分别尝试了目前在异常检测领域流行且识别效果较好的三种无监督算法——局部异常因子（LOF）、孤立森林[12]（isolation forest，Iforest）和一类支持向量机[13]（one-class support vector machine，OC-SVM），分别构建了三个无监督异常用户识别模型，采用网格参数寻优策略调整模型参数。

采取混淆矩阵评价指标来判定模型的优劣。混淆矩阵又称误差矩阵，是表示精度评价的一种标准格式，用n行n列的矩阵形式表示。混淆矩阵中的各列代表了预测的类别，各列样本数目之和为该类别的样本总数；各行代表样本的真实归属类别，各行样本数目之和为该类别样本的实例数目。

分类模型最终的分类结果可以在混淆矩阵中体现，混淆矩阵的行向量表示在进行分类时的倾向性[14]，当样本预测准确率为100%时，混淆矩阵表示为一个对角阵，即当混淆矩阵越“对角”，该分类模型的拟合优度越高。

图5～图7分别为LOF模型、OC-SVM模型及Iforest模型在测试集的混淆矩阵的训练结果评价图。

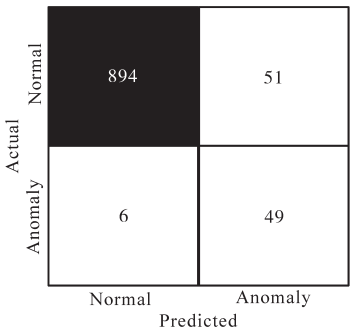

图5 LOF模型的训练结果评价

Fig.5 Evaluation of training results for LOF model

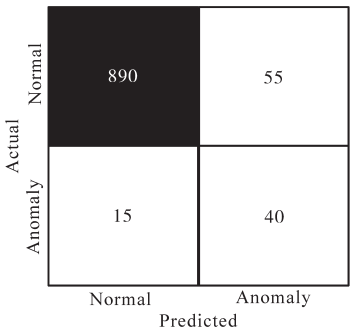

图6 OC-SVM模型的训练结果评价

Fig.6 Evaluation of training results for OC-SVM model

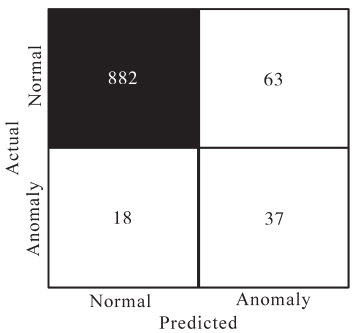

图7 Iforest模型的训练结果评价

Fig.7 Evaluation of training results for Iforest model

结合混淆矩阵，采用识别精确率、准确率、和召回率三项指标对无监督模型进行评价：分类精确率表示所有样本中被预测正确的样本比率，准确率表示预测为异常的样本中被预测正确的样本比率，召回率表示实际为异常的样本被模型预测正确的样本比率。三种无监督模型的评价结果如表4所示。

表4 无监督识别模型的效果对比

Tab.4 Comparison of unsupervised identification models

	精确率/%	准确率/%	召回率/%
LOF	94.3	49.0	89.1
OC-SVM	93.0	42.1	72.7
Iforest	91.9	37.0	67.3

LOF算法的识别精确率、准确率以及召回率均高于另外两种算法，LOF模型在样本数据集上

的表现效果最好，尤其是识别精确率达到94.3%，说明基于电流电压数据所构建的特征参数作为模型的输入是有效。但LOF算法的识别准确率不够理想，准确率低意味着模型出现误判的概率较高，即可疑样本中包含了较多实际为正常的样本，工作人员在排查可疑用户是否存在窃电行为时，其中的无效工作量较多，所以有必要提高模型的准确率，所以LOF模型仅作为分阶段识别模型的第一阶段筛选模型，保证高精度优先级，而将提高模型的识别准确率作为第二阶段识别模型的任务。

3.3　结合无监督采样的有监督异常识别结果

将LOF模型检测出的可疑样本交给人工进行标记，作为第二阶段模型的训练样本。在第二阶段识别任务中，将可疑样本集以2∶1的比例划分为训练集与测试集，采用机器学习中的有监督式二分类算法构建异常识别模型。为得到最佳识别效果，分别尝试多种有监督式二分类算法构建模型，包括逻辑回归[15-16]、支持向量机（SVM）及K近邻[17]、决策树[18]、随机森林[19-20]和极限梯度提升算法（eXtreme gradient boosting，XGB）[21]，仍选取识别精确率、准确率和召回率作为评价指标，实验结果如表5所示。

表5　有监督识别模型的效果对比

Tab.5　Comparison of supervised identification models

	精确率/%	准确率/%	召回率/%
逻辑回归	73.0	27.3	68.1
K近邻	67.0	60.1	85.1
支持向量机	83.0	84.1	78.7
决策树	67.0	71.9	48.9
随机森林	74.0	66.7	89.4
XGB	75.0	71.2	78.7

经过对比，SVM算法在实验数据集上的测试效果优于其他算法，尤其是准确率较高，达到了84.1%。分阶段识别模型的最终判定模型，即第二阶段识别模型，应该具有较高的准确率，以此更加准确地锁定窃电用户，所以SVM算法符合作为异常用户第二阶段识别模型的要求。

基于“LOF+SVM”的分阶段识别模型的效果与采用单一的LOF异常检测模型的对比结果如表6所示，相比于单一的LOF模型，基于“LOF+SVM”的分阶段识别模型大幅度提高了异常用户识别的准确度，有效降低异常用户的误判率（实际为正常的用户被判定为异常用户），减少工作人员的无效工作量。

表6　单一的异常识别模型与分阶段识别模型的效果对比

Tab.6　Comparison between single anomaly identification model and phased identification model

	精确率/%	准确率/%
LOF	94.3	49.0
LOF+SVM	97.7	84.1

4　结论

基于现有异常用电用户识别方法的识别准确度较低等问题，提出了一种基于“LOF+SVM”的分阶段识别模型，并利用电网数据进行建模分析：对比了几种不同的无监督异常识别算法，证实了LOF算法的识别精度最优，但准确率较低。对比几种不同的有监督式二分类异常识别模型的效果，结果表明SVM算法在测试数据数据集的表现优于其他算法，证实了选择SVM模型作为第二阶段识别模型的优越性。通过对比基于LOF算法的异常识别模型和“LOF+SVM”识别模型的效果，结果表明“LOF+SVM”识别模型的准确率更高，显著降低无监督LOF异常检测模型的误判率，能够减少电网工作人员的无效工作量。

参考文献

[1] 耿俊成，张小斐，周庆捷，等. 基于局部离群点检测的低压台区用户窃电识别[J]. 电网与清洁能源，2019，35（11）：30-36.
GENG Juncheng，ZHANG Xiaofei，ZHOU Qingjie，et al. Identification of power stealing in low voltage station area based on local outlier detection[J]. Power Grid and Clean Energy，2019，35（11）：30-36.

[2] 曾虎. 电能计量数据聚类分析与窃电检测研究[D]. 昆明：昆明理工大学，2017.
ZENG Hu. Research on clustering analysis of energy metering data and power theft detection [D]. Kunming：Kunming University of Science and Technology，2017.

[3] 吴迪，王学伟，窦健，等. 基于大数据的防窃电模型与方法[J]. 北京化工大学学报（自然科学版），2018，45（6）：79-86.
WU Di，WANG Xuewei，DOU Jian，et al. Anti-theft model and method based on big data[J]. Journal of Beijing University of Chemical Technology（Natural Science Edition），2018，45（6）：79-86.

[4] 李宁，王伟峰，蔡慧，等. 基于电量离群点挖掘的窃电辨识方法研究[J]. 中国计量大学学报，2018，29（3）：331-337.
LI Ning，WANG Weifeng，CAI Hui，et al. Research on electricity stealing identification method based on power outlier mining [J]. Journal of China University of Metrology，2018，29（3）：331-337.

[5] 程俊文，李慧娟，曹志强．基于K-means算法和用电信息采集系统的防窃电研究[J]. 供用电，2019，36(1)：75-80.
CHENG Junwen，LI Huijuan，CAO Zhiqiang. Research on electricity theft prevention based on K-means algorithm and electricity information acquisition system[J]. Distribution & Utilization，2019，36(1)：75-80.

[6] 邓明斌，徐志淼，邓志飞，等．基于多特征融合的窃电识别算法研究[J]. 计算机与数字工程，2017，45(12)：2398-2401，2414.
DENG Mingbin，XU Zhimiao，DENG Zhifei，et al. Research on electricity stealing identification algorithm based on multi-feature fusion[J]. Computer & Digital Engineering，2017，45(12)：2398-2401，2414.

[7] CAMPOS G O，ZIMEK A，MEIRA W. An unsupervised boosting strategy for outlier detection ensembles[C]//Proc of Pacific-Asia Conference on Knowledge Discovery & Data Mining，2018：564-576.

[8] REUNIG M M，KRIEGEL H P，NG R T，et al. LOF：identifying density-based local outliers[C]//Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data，2000：93-104.

[9] WANG Weixing，LI Hongxia，WANG Kevin，et al. Pavement crack detection on geodesic shadow removal with local oriented filter on LOF and improved level set[J]. Construction and Building Materials，2020，237 (Mar.20)：117750.1-117750.9.

[10] 周志华．机器学习[M]. 北京：清华大学出版社，2016：121-145.
ZHOU Zhihua. Machine learning[M]. Beijing：Tsinghua University Press，2016：121-145.

[11] 伍世虔，卢宇，方志军，等．一种加权的红外图像线性归一化方法[J]. 武汉理工大学学报，2010，32(20)：1-5.
WU Shiqian，LU Yu，FANG Zhijun，et al. A weighted linear normalization method for infrared images[J]. Journal of Wuhan University of Technology，2010，32(20)：1-5.

[12] 李新鹏，高欣，阎博，等．基于孤立森林算法的电力调度流数据异常检测方法[J]. 电网技术，2019，43(4)：1447-1456.
LI Xinpeng，GAO Xin，YAN Bo，et al. Anomaly detection method of power dispatching flow data based on isolated forest algorithm[J]. Power System Technology，2019，43(4)：1447-1456.

[13] 张瑜，罗可．基于OC-SVM的大型数据集分类方法[J]. 计算机工程与应用，2011，47(4)：131-133.
ZHANG Yu，LUO Ke. Classification method of large data set based on OC-SVM[J]. Computer Engineering and Applications，2011，47(4)：131-133.

[14] 张静，宋锐，郁文贤，等．基于混淆矩阵和Fisher准则构造层次化分类器[J]. 软件学报，2005(9)：1560-1567.
ZHANG Jing，SONG Rui，YU Wenxian，et al. Constructing hierarchical classifier based on confusion matrix and fisher criterion[J]. Journal of Software，2005(9)：1560-1567.

[15] 李长山．基于Logistic回归法的企业财务风险预警模型构建[J]. 统计与决策，2018，34(6)：185-188.
LI Changshan. The construction of enterprise financial risk early warning model based on Logistic regression method[J]. Statistics & Decision，2018，34(6)：185-188.

[16] HORYAINOV O M，ALPEEVA O V. Work of trucks and electric vehicles in the logistics system[J]. Nauka ta Progres Transportu，2008(25)：122-124.

[17] 谢娟英，高红超，谢维信．K近邻优化的密度峰值快速搜索聚类算法[J]. 中国科学：信息科学，2016，46(2)：258-280.
XIE Juanying，GAO Hongchao，XIE Weixin. K-nearest neighbor optimization density peak fast search clustering algorithm [J]. Scientia Sinica(Informationis)，2016，46(2)：258-280.

[18] 王道明，鲁昌华，蒋薇薇，等．基于粒子群算法的决策树SVM多分类方法研究[J]. 电子测量与仪器学报，2015，29(4)：611-615.
WANG Daoming，LU Changhua，JIANG Weiwei，et al. Research on decision tree SVM multi-classification method based on particle swarm optimization algorithm[J]. Journal of Electronic Measurement and Instrumentation，2015，29(4)：611-615.

[19] 吴潇雨，和敬涵，张沛，等．基于灰色投影改进随机森林算法的电力系统短期负荷预测[J]. 电力系统自动化，2015，39(12)：50-55.
WU Xiaoyu，HE Jinghan，ZHANG Pei，et al. Short-term load forecasting of power system based on grey projection improved random forest algorithm[J]. Automation of Electric Power Systems，2015，39(12)：50-55.

[20] PAVLOV Yu L. Distributions of the number of vertices in strata of a random forest[J]. Theory of Probability & Its Applications，1988，33(1)：96-104 .

[21] 李占山，刘兆赓．基于XGBoost的特征选择算法[J]. 通信学报，2019，40(10)：101-108.
LI Zhanshan，LIU Zhaogeng. Feature selection algorithm based on XGBoost[J]. Journal on Communications，2019，40(10)：101-108.

收稿日期：2021-09-18
修改稿日期：2021-10-28