计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20180216
55(8):1599-1608,2018

基于用电特征分析的窃电行为识别方法

史玉良[1,2] 荣以平[3] 朱伟义[3]
[1](山东大学软件学院 济南 250100)
[2](山大地纬软件股份有限公司 济南 250100)
[3](国网山东省电力公司 济南 250001)
(shiyuliang@sdu.edu.cn)

Stealing Behavior Recognition Method Based on Electricity Characteristics Analysis

Shi Yuliang[1,2], Rong Yiping[3], and Zhu Weiyi[3]
[1](*School of Software, Shandong University, Jinan* 250100)
[2](*Dareway Software Co., Ltd., Jinan* 250100)
[3](*State Grid Shandong Electric Power Company, Jinan* 250001)

Abstract Anti-stealing electricity is an indispensable component of electricity enterprise management. In view of the current problems such as the large number of users, the wide distribution area, the increasing year-on-year power stealing, and the lack of supervision personnel, this paper analyzes and handles the data of power stealing behavior, and proposes a stealing behavior recognition method which can identify the stealing users. First, based on the collected samples, this method adopts a filtering algorithm and a regular threshold to implement feature extraction, so as to improve the effectiveness of the collected data. Then, the user's stealing behavior diagnosis model is constructed based on the logistic regression algorithm to realize the determination of suspected users. In addition, this paper uses the closed-loop working mechanism which continues to update data by the way of pushing, shooting, processing and feedback to continuously optimize the model. According to the collected data provided by the power consumption information collection system and marketing business application system of State Grid Shandong Electric Power Company, the experimental results prove the feasibility and applicability of the method.

Key words filter algorithm; rule threshold setting; feature extraction; stealing behavior diagnosis model; closed-loop feedback optimization

摘 要 反窃电工作是实现电力企业用电管理不可或缺的环节.针对山东省用电用户数量多、分布面积广、窃电现象逐年上升、检测人员不足等特点,对获取的用户窃电行为数据进行合理的分析、处理,提出一种基于用电特征分析的窃电行为识别方法,实现对窃电嫌疑用户的筛查.该方法首先基于采集样本,以过滤式算法和规则阈值设定的方式,实现采集样本数据的特征提取,从而提高采集数据的有效性;

收稿日期:2018-04-01;修回日期:2018-06-01
基金项目:山东省泰山产业领军人才工程专项经费(tscy20150305);山东省重点研发计划(2016GGX101008,2016ZDJS01A09);山东省自然科学基金重大基础研究项目(ZR2017ZB0419)
This work was supported by the TaiShan Industrial Experts Programme of Shandong Province (tscy20150305), the Primary Research and Development Plan of Shandong Province (2016GGX101008, 2016ZDJS01A09), and the Major Basic Research Project of Natural Science Foundation of Shandong Province (ZR2017ZB0419).

随后以逻辑回归算法构建用户窃电行为诊断模型,实现对窃电嫌疑用户的判定;此外,采用推送、排查、处理和反馈的闭环工作机制不断优化模型,并以国网山东省电力公司用电信息采集系统、营销业务应用系统提供数据进行算例分析,验证了所述方法的可行性与适用性.

关键词 过滤式算法;规则阈值设定;特征提取;用户窃电行为诊断模型;闭环反馈优化

中图法分类号 TP391

近年来,我国窃电相关的纠纷案件数量逐年上升,窃电手段不断变化发展,使得窃电范围不断扩大,对电力系统的正常运行造成了阻碍.为规范管理生产生活用电、提高电能利用率、推进电力企业健康稳定发展,开展用电检查与反窃电工作是电力企业的一项迫切工作.

早期,传统的电费收取方式为人工手抄表,体力劳动强且涉外人员数量不足,窃电查处问题也由涉外人员承担,用电检查人员数量难以满足需求;其次,由于传统用电管理思想观念根深蒂固,供电企业对窃电行为不够重视,造成用电检查及打击窃电行为力度低,助长了窃电行为.随着社会进步、经济发展,用户用电数量不断增多,社会生产生活对电能的需求量越来越大,电力企业也随之引进智能电表,以用电信息采集系统和SG186营销业务应用系统有效实现远程费控,然而窃电查处问题并未随之解决,且在经济利益的驱使下,窃电者不再局限于过去的居民、个体等,逐渐发展成为了集体企业、中外合作企业等,发展速度十分快,严重干扰了电力企业供电安全与秩序.此外,随着科学技术日新月异,高科技手段被窃电人员广泛应用,随着窃电技术的智能化、科技化发展,使得高科技含量的窃电方式越来越多,如无线遥控、有线远方控制等,此类窃电手段往往十分隐蔽,传统用电检查方法根本无法检测,且用电检查人员综合素质较低,难以满足用电检查和反窃电工作的现代化需求[1].

因此,亟需采取有效方法,利用电力企业现有系统所提供数据实现反窃电分析,对窃电嫌疑用户行为进行概率推测和诊断,精准识别重大窃电嫌疑用户,提高反窃电工作成效,加强我国电力企业对电能输出的高效监管力度.通过采用强有力的窃电监控识别手段,加大窃电的查处惩治力度,维护正常的供用电秩序,保障公司经营效益.

本文的主要贡献有3个方面:

1) 在用电特征提取阶段.一方面基于过滤式算法筛选窃电特征数据项,另一方面深化特征数据项的有效性,针对动态数据以规则与阈值结合的方式识别特征异常类型,并以日、周、月、季、年5类数据,实现当前用电数据与历史用电数据的对比,从而实现对人工判别异常数据的模拟.

2) 在模型算法选择阶段.以逻辑回归算法实现窃电行为的数据挖掘,始终着眼于整体数据,对全局数据的综合性把握较高,特别随着新数据的参与,模型可基于反馈数据快速调整其输入特征及参数变化.

3) 在模型训练阶段.基于用电信息采集系统和SG186营销业务应用系统积累的大量客户用电信息,结合大量典型窃电案例,综合考虑各种窃电因素,依据事物发展变化的因果关系来识别数据的异常走势,是一种从定量至定性的诊断方法,具有模拟人工识别异常数据和多维数据的综合信息挖掘的优点.

1 反窃电相关研究与分析

目前,反窃电工作越来越受到各级电力企业的重视,并引起社会各界的广泛研究与关注,反窃电手段亦得到不断发展.如针对当前常见的窃电技术和存在的缺陷,采用4G通信模块、智能视频取证等关键技术,集成防窃电和实时视频监控的智能视频监控终端的防窃电监测方式[2];针对电力传输过程中产生的线损数据,对高线损异常用户进行识别,从而实现反窃电监测的方式[3];针对异常用电用户,提出基于无监督学习的异常用电模式检测模型,主要以特征提取、主成分分析、网格处理、计算局部离群因子等建模,输出所有用户用电行为的异常度及疑似概率排序,以检测异常度排序靠前的少数用户查出异常用户[4].

由参考文献[2-4]可以看出,反窃电工作越来越具体化与目标明确化,且主要以远程侦查为发展目标,由此减轻涉外人员的工作强度;同时带来的问题是,如何在不增加电力企业经济压力的前提下,实现对窃电行为的全面侦查,特别随着用电信息采集系统和SG186营销业务应用系统的全覆盖应用,“数据海量,信息匮乏”的现象正反映了反窃电工作的尴尬处境.针对窃电问题带来的电力企业用电监管

问题,本文基于当前电力系统积累的大量客户用电信息,综合考虑各种因素,建立窃电行为识别方法[5],对窃电嫌疑行为进行概率推测和诊断[6],精准识别重大窃电嫌疑户.

2 窃电行为识别方法的整体设计思路

本方法旨在构建有效可行的窃电用户行为识别模型.在该模型特征数据输入阶段,为有效提取窃电行为的相关数据和异常数据,一方面基于过滤式算法筛选窃电特征数据项,一方面基于反窃电领域专家经验设定动态数据曲线异常识别规则及阈值,从而提高了模型输入特征数据的有效性;随后以逻辑回归算法[7]构建用户窃电行为诊断模型,以样本特征数据为输入,输出样本分类的方式实现对用户窃电行为的识别,整体采用一种闭环工作的反窃电诊断机制[8],满足精准识别重大窃电嫌疑户的业务需求,整体方法流程步骤如图 1 所示:

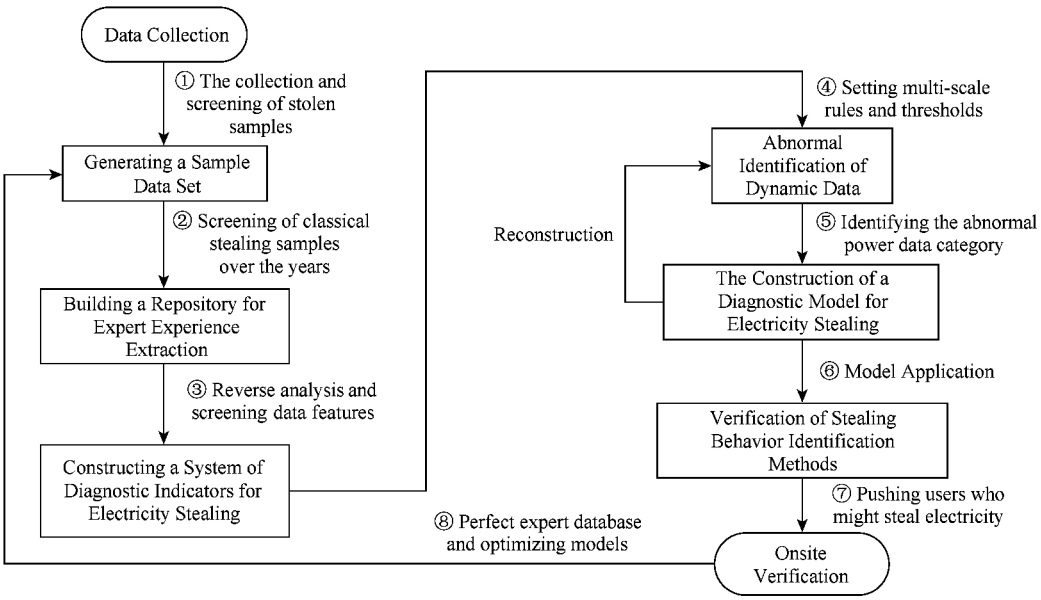

Fig. 1 The flow chart of stealing behavior diagnosis method

图 1 用户窃电行为诊断方法的总体流程图

① 自 SG186 营销业务应用系统搜集窃电样本、无窃电正常样本、用户档案类数据,从用电信息采集系统获取用户计量指标数据,窃电样本数据与正常样本数据构成数据样本集;

② 基于窃电样本,对样本进行典型窃电案例精简,初始化典型窃电案例的专家样本库;

③ 基于典型窃电案例反向分析,基于过滤式算法筛选窃电特征数据,建立基于窃电行为的特征数据库,即窃电行为诊断指标体系,初始化特征数据库;

④ 基于特征数据项提取模型输入数据信息,选取多时间域的用户用电数据动态曲线,基于规则和阈值识别特征异常类别;

⑤ 基于用户基本特征数据项和异常数据类别,利用逻辑回归算法构建用户窃电行为诊断模型,实现对窃电用户行为的识别,若识别正确率后期降低,则将信息反馈至④,从而对新特征异常类别加以重构;

⑥ 以某一时间段内的用户数据构建验证数据样本集,对本文构建的用户窃电行为识别方法进行验证;

⑦ 根据验证输出的窃电嫌疑用户生成窃电排查工单,现场进行检查取证、查处工作,对现场排查确认的窃电用户,确认窃电行为及采用的窃电方式;

⑧ 现场核实结果反馈至 SG186 营销业务应用系统,提取有效的窃电数据作为案例加入专家样本库中,完善窃电行为诊断指标体系,根据反馈的窃电案例对窃电行为识别方法不断修正优化.

本文主要对窃电行为识别方法的构建加以阐述,并以 matlab 仿真对模型的训练与测试效果进行验证,实践应用环节还有待开展,并可基于后期反窃电开展工作不断优化提高窃电行为识别方法的性能.

3 窃电行为识别方法的构建过程

本文构建的窃电行为识别方法主要包含 2 部分:

1)基于多尺度识别用电特征异常,实现对人工识别窃电异常数据的模拟;2)基于用户基本数据和特征异常识别数据,采用逻辑回归算法[9]构建用户窃电行为诊断模型,从而完成嫌疑用户窃电行为识别.

3.1 窃电特征数据提取

本文对历年窃电用户在采集、营销系统的电量、电压、电流、报警、信用等数据进行反向分析,构建本次建模的指标体系[10],主要分为 2 部分:1)基于过滤式算法筛选窃电特征数据项;2)基于规则和阈值设定识别动态用电异常特征数据.

1) 基于过滤式算法筛选窃电特征数据项

本文首先对样本特征数据进行规范化处理,特征数据取值为$[-1,1]$;随后,基于过滤式算法对样本数据进行特征选择,主要方法为针对每一个初始特征,以特征相关性度量特征对分类结果的重要性.

设 C_1 为窃电类别,C_2 为非窃电类别即正常用电,给定窃电样本集$\{(x_1,C_1),(x_2,C_1),\cdots,(x_I,C_1)\}$,随后根据窃电样本集用户,基于采集其非窃电的用电数据,构建其正常用电特征数据样本$\{(x'_1,C_2),(x'_2,C_2),\cdots,(x'_I,C_2)\}$,对每一个样本 $x_i,i\in\{1,2,\cdots,I\}$,x_i 包含 $j\in\{1,2,\cdots,J\}$个特征属性. x_i 先在其同类样本中寻找其最近邻样本 $x_{i,\mathrm{nh}}$ 作为其猜中近邻,再从异类样本中锁定 $x'_{i,\mathrm{nh}}$ 作为 $x_{i,\mathrm{nm}}$即猜错近邻,同理,分别找到 I 个样本的猜中近邻与猜错近邻. 随后,针对特征 j 计算其相关统计量 δ^j,计算为

$$\delta^j=\sum_{i=1}^{I}\left[-diff(x_i^j,x_{i,\mathrm{nh}}^j)^2+diff(x_i^j,x_{i,\mathrm{nm}}^j)^2\right],\quad(1)$$

其中,x_i^j 表示样本 x_i 在属性 j 上的取值,同理 $x_{i,\mathrm{nh}}^j$ 和 $x_{i,\mathrm{nm}}^j$,$diff(x_i^j,x_{i,\mathrm{nh}}^j)$取决于属性 j 的类型. 若属性 j 为离散型,则 $x_i^j=x_{i,\mathrm{nh}}^j$时 $diff(x_i^j,x_{i,\mathrm{nh}}^j)=0$,否则为 1;若属性 j 为连续型,则 $diff(x_i^j,x_{i,\mathrm{nh}}^j)=|x_i^j-x_{i,\mathrm{nh}}^j|$.

由式(1)可得,对于属性 j,若 x_i 与其猜中近邻 $x_{i,\mathrm{nh}}$的距离越小,与其猜错近邻 $x_{i,\mathrm{nm}}$的距离越大,则其相关统计量 δ^j 越大,说明特征属性 j 的区分窃电与非窃电类别的能力越强,将 δ^j 进行降序排列,设定阈值 τ,将相关统计量大于阈值 τ 的特征作为筛选特征.

2) 基于规则识别动态用电异常特征数据

在传统的反窃电侦查过程中,工作人员往往通过用电信息采集系统、营销业务应用系统提供的采集数据和历史数据对比进行人工识别,由于用电数据动态变化往往包含清晰的窃电识别信息,反窃电专家往往基于此结合丰富的反窃电经验对此作出甄别,故本文针对此类用电动态数据制定数据异常判别规则[11],结合阈值的方式识别特征数据异常并给出异常类别,如电流三相不平衡、电量突减、相位角反极性等.

基于反窃电专家历年窃电诊断经验,本文对电流数据、电压数据、电量数据和相位角数据进行特征异常类别甄别. 由于反窃电诊断是基于当前采集数据进行,故识别数据曲线为当日采集数据、前推一周采集数据、前推一月采集数据、前推一季度采集数据和前推一年采集数据. 判别规则分类制定,图 2 为由采集数据获取的电压失压断相实际曲线,其规则及阈值设定如下:

三相四线断相:任一相电压小于 $K\times$额定电压,另两相电压中任一相电压不小于 $K\times$额定电压.

若上述规则成立,则判定其为电压失压断相.

图 3 为电量趋势突减异常采集数据曲线,电量趋势下降指标作为模型的异常特征指标,部分行业的用户在春节及长假数据可能对结果造成误判,需要剔除,故其量化公式为电量趋势判断规则如下:

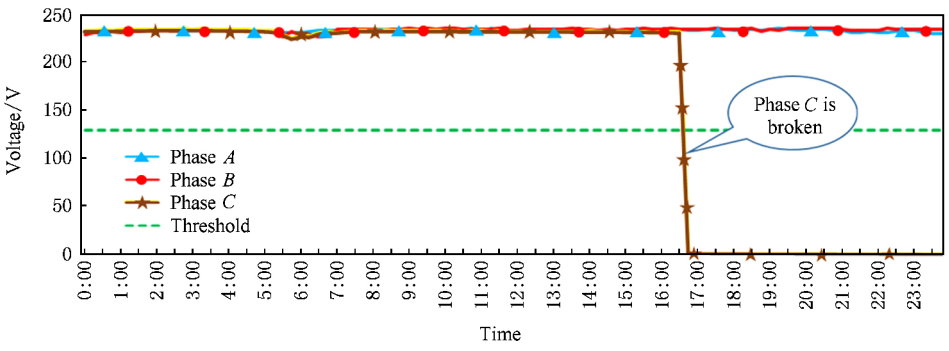

Fig. 2 Discriminating curve of voltage phase failure

图 2 电压失压断相判别数据曲线

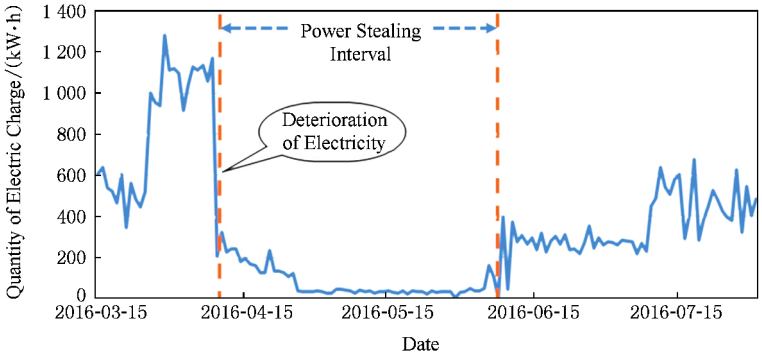

Fig. 3 Discriminating curve of electricity trend

图 3 电量趋势判断数据曲线

$$k_l = \sum_{l=r-d}^{r-1} \alpha_r \times \frac{g_r - g_l}{g_l} + \sum_{l=r+1}^{r+d} \alpha_r \times \frac{g_l - g_r}{g_r}, \quad (2)$$

其中,k_l 为当天下降趋势指标,g_r 为当天电量,g_l 为前后几天电量,α_r 为权重,d 为前后天数.

图 2 为单日电压数据采集曲线,反窃电专家从该数据变化判别该用户电压数据异常,即电压失压断相;然而,仅依赖于单日采集数据进行特征异常判别,往往导致局部视角狭窄,如图 3 所示,在 2016-04-15—2016-06-15 区间的单日电量数据难以察觉其用电量异常,需结合前推一季度乃至前推一年的用电量加以判别. 本文基于反窃电专家的判别经验,尽可能地以动态规则的视角模拟人工识别视角,以多尺度的规则及阈值设定实现单类特征异常识别,从而最大程度地拟合人工识别,提高窃电类型识别的精准度. 具体异常特征识别类型如表 1 所示:

Table 1 Abnormality Feature Identification Type

表 1 异常特征识别类型

Feature Category	Abnormal Type
Electric Current	Abnormal current circuit (such as sudden increase or decrease); Unbalanced three-phase current; Negative current; Current loss
Voltage	Voltage loss; Abnormal voltage circuit (such as sudden increase or decrease)
Quantity of Electric Charge	Abnormal trend; The meter goes backwards; Abnormal line loss rate (such as increase); Transformer light load
Phase Sequence	Abnormal polarity; Trend reverse; Abnormal power factor; Disturbed phase sequence data

3.2 用户窃电行为诊断模型

用户窃电行为诊断模型是本方法的核心内容,以逻辑回归算法为数据处理原理,主要包含模型训练与模型测试 2 部分,具体模型构建流程如图 4 所示,包括 8 个步骤:

步骤 1. 基于用户用电数据选取等比例窃电样本数据与正常用电样本数据作为模型样本集,从模型样本集内分别随机抽取相应比例的数据构成训练集数据及测试集数据;

步骤 2. 将训练集数据进行用户样本集定义 $X=\{x_1,x_2,\cdots\}$,特征权重向量 $\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots)$,则目标函数 $f(\boldsymbol{\theta})=\boldsymbol{\theta}^{\mathrm{T}}\times X$,类别集合 $C\in\{C_1,C_2\}$,初始化迭代次数 $k=0$,允许误差 $\varepsilon>0$,基于一定范围对 $\boldsymbol{\theta}$ 随机赋值;

步骤 3. 进入迭代求解过程,$k=k+1$;

步骤 4. 采用拟牛顿法对目标函数进行最优求解,目标函数的梯度

$$\nabla f = \frac{\sum\limits_{x_l\in X,C=C_1} p_{l1}\,p_{l2}\,x_l}{\sum\limits_{x_l\in X,C=C_1} p_{l1}+N_1} - \frac{\sum\limits_{x_l\in X,C=C_1} p_{l1}\sum\limits_{x_l\in X,C=C_1} p_{l1}\,p_{l2}\,x_l}{\left(\sum\limits_{x_l\in X,C=C_1} p_{l1}+N_1\right)^2}. \quad (3)$$

计算海森矩阵 $\boldsymbol{H}_{k+1}$:

$$\boldsymbol{H}_{k+1}=\boldsymbol{H}_k+\left(1+\frac{\boldsymbol{q}^{(k)^{\mathrm{T}}}\boldsymbol{H}_k\boldsymbol{q}^{(k)}}{\boldsymbol{p}^{(k)^{\mathrm{T}}}\boldsymbol{q}^{(k)}}\right)\frac{\boldsymbol{p}^{(k)}\boldsymbol{p}^{(k)^{\mathrm{T}}}}{\boldsymbol{p}^{(k)^{\mathrm{T}}}\boldsymbol{q}^{(k)}}-\frac{\boldsymbol{p}^{(k)}\boldsymbol{q}^{(k)^{\mathrm{T}}}\boldsymbol{H}_k+\boldsymbol{H}_k\boldsymbol{q}^{(k)}\boldsymbol{p}^{(k)^{\mathrm{T}}}}{\boldsymbol{p}^{(k)^{\mathrm{T}}}\boldsymbol{q}^{(k)}}, \quad (4)$$

其中,$\boldsymbol{p}^{(k)}=\boldsymbol{\theta}^{(k+1)}-\boldsymbol{\theta}^{(k)}$,$\boldsymbol{q}^{(k)}=\nabla f(\boldsymbol{\theta}^{(k+1)})-\nabla f(\boldsymbol{\theta}^{(k)})$,$p_{i2}$ 表示被正确分类为 C_2 类的实例个数,置海森矩阵 $\boldsymbol{H}_1$ 为单位矩阵.

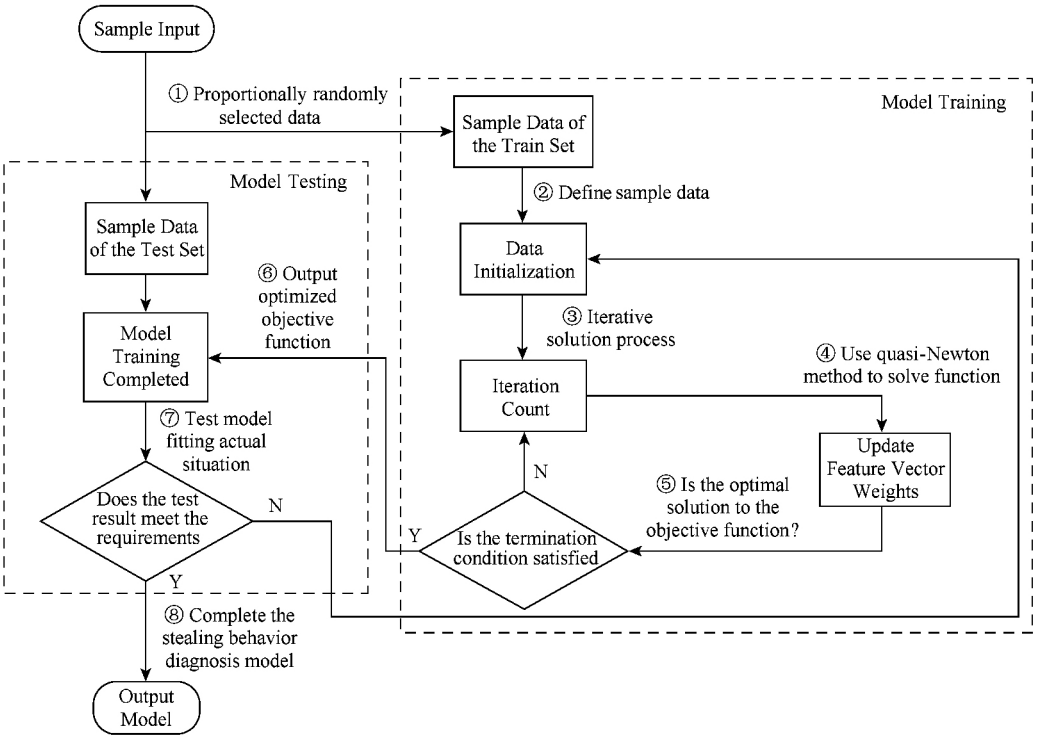

Fig. 4 The flow chart of the stealing behavior diagnosis model

图 4 构建用户窃电行为诊断模型流程图

梯度下降方向

$$\boldsymbol{d}^{(k)}=-\boldsymbol{H}_k\ \nabla f. \tag{5}$$

进而从 $x_{(k)}$ 出发，延方向 $\boldsymbol{d}^{(k)}$ 搜索，求步长 λ_k，求解方程满足如下：

$$f(\boldsymbol{\theta}^{(k)}+\lambda_k\boldsymbol{d}^{(k)})=\min_{\lambda\geqslant 0}f(\boldsymbol{\theta}^{(k)}+\lambda_k\boldsymbol{d}^{(k)}). \tag{6}$$

更新特征权重向量 $\boldsymbol{\theta}=(\theta_1,\theta_2,\cdots)$，

$$\boldsymbol{\theta}^{(k+1)}=\boldsymbol{\theta}^{(k)}+\lambda_k\boldsymbol{d}^{(k)}. \tag{7}$$

步骤 5. 将特征权重向量 $\boldsymbol{\theta}$ 代入目标函数 $f(\boldsymbol{\theta})$，判断式(8)是否成立

$$\|\nabla f(\boldsymbol{\theta}^{(k+1)})\|>\varepsilon. \tag{8}$$

若成立，则返回步骤 3 继续进入迭代求解过程；若不成立，则获得本次计算所得的最优化目标函数，进入步骤 6.

步骤 6. 基于最优化目标函数构成用户窃电行为诊断模型，

$$P(C=C_1\,|\,x_j)=\frac{1}{1+\mathrm{e}^{-\boldsymbol{\theta}^{\mathrm{T}}x_j}}. \tag{9}$$

对用户窃电行为诊断模型进行样本测试，比较概率，概率与类别比例相比较获取对应类标号，进行测试样本分类，其中，x_j 为测试样本数据，属于步骤 1 所得的测试集数据.

步骤 7. 计算测试集数据的测试参数，判断是否满足用户窃电判别要求，若不满足，则返回步骤 2 对 $\boldsymbol{\theta}$ 更新随机赋值，若满足，则进入步骤 8.

步骤 8. 构建完成用户窃电行为诊断模型，并输出本次更新模型.

4 实验与结果

为验证本方法的可行性和有效性，本文基于国网山东省电力公司用电信息采集系统为背景实施平台，并作为基础数据来源. 其中，窃电行为诊断包含与外围系统连接的输入输出信息模块，存储单元包含 3 类数据库：1)数据库存储输入信息与信息预定义；2)数据库存储解决方案与测试结果；3)数据库存储专家样本与样本特征.

4.1 模型数据特征提取

自 SG186 营销业务应用系统搜集从 2009—2016 年的窃电样本 1 万例，对应相关窃电用户从用电信息采集系统获取用户计量指标数据、SG186 营销业务应用系统获取用户档案类数据、采集终端获取异常事项数据，此外，从山东省各地市全面地抽取无窃电记录用户 1 万例，其对应相关数据作为正常用户数据，窃电样本数据与正常用户数据构成数据样本集，基于 1 万余例可用窃电样本，对样本进行典型窃电案例精简，初始获取 77 例典型窃电案例的专家

样本库,进而建立基于窃电行为的特征数据库,即反窃电预警诊断指标体系.

基于获取的1万余例可用窃电样本数据,采用过滤式算法筛选窃电特征数据项,如表2所示:

Table 2 Results on Feature Data Screening

表2 特征数据筛选结果

First-Level Attributes	Variable
Basic Information of Users	User tag, User ID, User classification, Address, Electricity category, User status, Power supply company ID
Electricity Price	Power point ID, Electricity price ID, Electricity category, Electricity price code
Measuring Point	Measurement point ID, Voltage level, Measurement method, Line ID, Regional ID
Meter	Meter ID, Comprehensive Rate
Charges Receivable	Monthly electricity in the past 1 year, Monthly electricity bill in the past 1 year, Accumulated arrears in the past 1 year, Accumulated amount of arrears in the past 1 year
Credit Information	Stealing records
Current Stealing Factor	Current data
Voltage Stealing Factor	Voltage data
Electricity Stealing Factor	Electricity data
Line Loss Rate Data	Line loss rate data
Phase Stealing Factor	Phase data
Metering Device Events	The meter had been opened, Voltage/current imbalance, Phase sequence abnormal, Abnormal voltage circuit

随后,针对筛选特征项内的动态变化数据,采用基于规则和阈值设定的方式识别用电异常特征数据,并以部分识别的异常数据为例进行展示,结果如表3所示:

Table 3 Results on Partial Abnormal Data Feature Recognition

表3 部分异常数据特征识别结果

User ID	Date	Electricity Trend	Trend Reverse	Voltage Loss	Current Loss	…	Whether Theft Occurred
1	2012-07-03	−0.732 1	0.865 3	0	0	…	yes
2	2013-05-06	−0.365 3	0.893 6	0	0	…	yes
3	2015-03-17	−0.689 1	0	0.792 0	0	…	yes
4	2015-04-16	−0.301 2	0	0.980 0	0	…	yes
5	2015-03-02	−0.467 3	0.803 7	0	0.756 9	…	yes
6	2016-08-03	−0.286 6	0	0	0.902 1	…	yes
7	2017-05-02	−0.479 3	0	0.872 2	0	…	yes
8	2013-04-01	−0.179 8		0	0	…	no
9	2014-06-05	−0.200 6	0.154 6	0	0.127 8	…	no
10	2015-05-02	−0.199 8	0	0.098 8	0	…	no
11	2016-04-29	−0.674 5	0	0	0.091 8	…	no
12	2017-06-12	−0.561 4	0	0	0	…	no
13	2017-06-24	−0.230 0	0	0	0	…	no
⋮	⋮	⋮	⋮	⋮	⋮		⋮

综上所述,针对于用户窃电行为诊断模型的训练和测试,本文采用77例典型窃电案例实现特征异常数据规则与阈值的设定,采用2万例样本集数据用于实现对逻辑回归概率预测的构建,并以此作为模型特征数据的筛选与提取.

4.2 过程及结果分析

基于2万例样本集数据[12]构建用户窃电行为诊断模型,为获取更优化的逻辑回归概率预测,本文采用重复3次训练过程优化逻辑回归概率预测,训练样本为1.6万例,累计获取8次特征权重向量重新

赋值,迭代次数阈值设定为 200 次,模型优化准确率目标为 98%,模型分别在第 3 次、第 5 次、第 8 次迭代过程中取得满足准确率识别要求的参数解,其迭代次数分别为 165 次、158 次和 200 次,其历次准确率与迭代次数变化如图 5 所示,其中 OI(output iteration)表示输出迭代值:

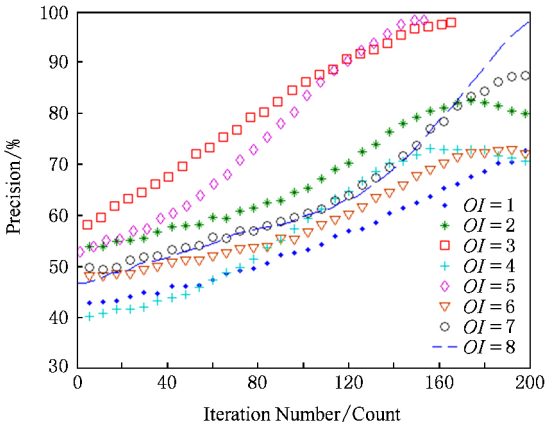

Fig. 5 Iterative solution process for model building

图 5 构建模型的迭代求解过程

根据图 5 的模型训练结果,本文选取准确率满足要求的 3 例模型,以 4 000 例测试样本对其进行测试,最终获取测试结果的正确率、召回率及精准率如表 4 所示:

Table 4 Test Results on the Model of Stealing Behavior Recognition

表 4 用户窃电行为诊断模型测试结果

Test Item	Model 1	Model 2	Model 3
Accuracy P/%	90.48	94.28	91.70
Recall Rate R/%	97.80	93.60	95.20
True Positive	1 956	1 872	1 904
False Positive	337	101	236
True Negative	44	128	96
False Negative	1 663	1 899	1 764

表 4 中,True/False 表示样本实际是否是窃电数据;Positive/Negative 表示数据通过模型后输出结果是否是窃电(模型判断是窃电的为 Positive,反之为 Negative).

由于本文为窃电监测类识别,故需在保证准确率的情况下提高召回率,即尽可能地识别实际窃电的用户,故采用综合评价指标(F-Measure),在准确率和召回率出现矛盾的情况,通过加权调和评价,计算为

$$F\text{-Measure}=\frac{2\times P\times R}{P+R}. \tag{10}$$

由综合评价指标计算可得,Model 1,Model 2 和 Model 3 分别为 94%,93.94% 和 93.4%,故选择 Model 1 为最终用户窃电行为诊断模型.

4.3 实验结果对比和分析

为说明本文所使用方法的合理性,本文基于 4.2 节实验数据,选择当前反窃电研究热点领域的 2 种方法作为窃电行为识别的对比方案,并将本文所述方法作为第 3 种方案,开展对比实验并对实验结果加以说明,具体方案如下:

1) Options 1. 基于采集数据,建立基于正态分布离群点算法的窃电行为识别方法[13].

2) Options 2. 基于本文的特征提取数据,建立基于无限深度神经网络的窃电行为识别方法[14].

3) Options 3. 本文基于用电特征分析的窃电行为识别方法.

最终实验结果如表 5 所示:

Table 5 Comparison of Experimental Results for Three Options

表 5 3 种方案实验对比结果

Options	Training Time for Model Reconstruction	Test Results	
		F-Measure/%	Average Time Consumption/s
Options 1	<2 min	72.33	0.34
Options 2	>24 h	87.2	0.62
Options 3	<2 min	94.0	0.36

由表 5 可知,由于 Options 1 基于采集数据的离群点检测构建正态分布概率统计模型,与本文所述方法的用电异常特征数据筛选环节原理相似,但本文一方面结合了反窃电专家的数据甄别经验实现异常特征规律的人工判别模拟,另一方面以多特征而非单一异常特征实现窃电行为识别,从而提高了模型的综合性和整体性识别水平,故本文所用方法虽然平均时耗高于 Options 1,但在综合评价指标方面体现出较大优势. 随着机器学习的发展,无限深度神经网络开始成为各领域的研究热点,Options 2 亦基于此方法开展窃电行为识别的训练与测试,该模型构建须基于海量数据支撑,且在数据质量不稳定时存在较高风险,尤其对于当前处于发展阶段的预测分类,其输入/输出数据的变动,往往导致模型产生高重构代价的风险,虽然在再训练过程中可借鉴历史经验,但是其调参复杂迭代次数往往为 5 000~10 000 次,甚至再训练时间多达几天,且基于当前的

有限采集窃电样本,构建的模型综合评价指标测试结果不理想,仅为87.2%,故在时效性、精确性和合理性方面,Options 2在本文所述应用环境中均受限.

银行行业的预测模型中,80%是采用逻辑回归算法构建[15],可见逻辑回归方法在模型构建、数据处理稳定性方面具有显著优势,通过表5的测试结果亦可得,基于现有的历史数据,本文所述方法的窃电行为识别的综合评价指标均高于Options 1和Options 2.且反窃电工作属于发展阶段,随着科技技术的日新月异,窃电技术亦不断发展,故在输入/输出数据均存在变动性的情况下,本文所述方法可快速实现模型的再训练,适应性更强.由以上数据显示,本方法对窃电概率预警具有高效预测能力,可有效辅助国网山东省电力公司相关工作人员开展反窃电工作.

5 结束语

本文以国网山东省电力公司集成化数据平台为背景,对用电信息采集系统、SG186营销业务应用系统及采集终端可提供的用户数据进行分析处理,构建了一种基于用电特征分析的窃电行为识别方法.首先基于窃电样本可用数据对窃电行为用电特征数据进行筛选,随后基于窃电样本筛选典型窃电案例构建专家样本库,并基于此设定窃电行为导致用电数据异常的规则与阈值,从而提取出用电特征中的异常数据类别,上述数据作为模型输入.基于逻辑回归算法构建用户窃电行为诊断模型,并以拟牛顿算法求解最优目标函数有效减少迭代求解次数,上述模型实现对窃电嫌疑用户的筛查.

此外,本方法采用预警、排查和处理反馈的闭环工作机制不断丰富专家样本库,模型根据反馈案例持续进行学习训练、优化重构,不断提高模型的精度和泛化能力,提高识别窃电嫌疑用户的精准度.由算例分析数据可得,本方法是高效可行的,可精准识别窃电嫌疑用户,它提供一种强有力的反窃电监控预警手段,有助于加大反窃电的查处惩治力度,维护正常的供用电秩序.

参 考 文 献

[1] Wang Yingtao. A study on the classification and management of severity and disability of beitai district [D]. Shanghai: Shanghai Jiaotong University, 2012 (in Chinese)
(王颖韬. 市北台区窃电严重程度分级界定及管理办法研究[D]. 上海: 上海交通大学, 2012)

[2] Xiong Dezhi, Chen Yilei, Yang Jie, et al. Design of preventing electricity-stolen intelligent video surveillance terminal based on 4G network [C] //Proc of 2017 IEEE Conf on Energy Internet and Energy System Integration (EI2). Piscataway, NJ: IEEE, 2017: 26-28

[3] Bais V, Dongre K, Bhagat A P, et al. Network analysis model based on canny communication system for theft detection [C] //Proc of 2016 Online Int Conf on Green Engineering and Technologies (IC-GET). Piscataway, NJ: IEEE, 2017: 1-6

[4] Zhuang Chijie, Zhang Bin, Hu Jun, et al. Detection of abnormal electricity model of power users based on unsupervised learning [J]. Journal of China Electro-mechanical Engineering, 2016, 36(2): 379-387 (in Chinese)
(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)

[5] Li Zhijie, Li Yuanxiang, Wang Feng, et al. Online learning algorithms for big data analytics: A survey [J]. Journal of Computer Research and Development, 2015, 52(8): 1707-1721 (in Chinese)
(李志杰, 李元香, 王峰, 等. 面向大数据分析的在线学习算法综述[J]. 计算机研究与发展, 2015, 52(8): 1707-1721)

[6] Fu Yiqi, Dong Wei, Yin Liangze, et al. Software defect prediction model based on the combination of machine learning algorithms [J]. Journal of Computer Research and Development, 2017, 54(3): 633-641 (in Chinese)
(傅艺绮, 董威, 尹良泽, 等. 基于组合机器学习算法的软件缺陷预测模型[J]. 计算机研究与发展, 2017, 54(3): 633-641)

[7] Guo Huaping, Dong Yadong, Wu Changan, et al. Logistic regression method for class imbalance problem [J]. Pattern Recognition and Artificial Intelligence, 2015, 28(8): 686-693 (in Chinese)
(郭华平, 董亚东, 邬长安, 等. 面向类不平衡的逻辑回归方法[J]. 模式识别与人工智能, 2015, 28(8): 686-693)

[8] Mirzaee A, Salahshoor K. Fault diagnosis and accommodation of nonlinear systems based on multiple-model adaptive unscented Kalman filter and switched MPC and H-infinity loop-shaping controller [J]. Journal of Process Control, 2012, 22(3): 626-634

[9] Zhang Yazhou, Zhang Yanxia, Meng Gaopeng, et al. A wide area information based clustering recognition method of coherent generators [J]. Power System Technology, 2015, 39(10): 2889-2893 (in Chinese)
(张亚洲, 张艳霞, 蒙高鹏, 等. 基于广域信息的同调机群聚类识别方法[J]. 电网技术, 2015, 39(10): 2889-2893)

[10] Gang Luo, Frey L J. Efficient execution methods of pivoting for bulk extraction of entity-attribute-value-modeled data [J]. IEEE Journal of Biomedical and Health Informatics, 2016, 20(2): 644-654

[11] Niu Xinzheng, Wang Chongqi, Ye Zhijia, et al. An efficient association rule hiding algorithm based on cluster and threshold interval [J]. Journal of Computer Research and Development, 2017, 54(12): 2785-2796 (in Chinese)

(牛新征, 王崇屹, 叶志佳, 等. 基于簇和阈值区间的高效关联规则隐藏算法[J]. 计算机研究与发展, 2017, 54(12): 2785-2796)

[12] Zhang Zhongheng, Trevino V, Hoseini S S, et al. Variable selection in logistic regression model with genetic algorithm [J]. Chinese Journal of Electronics, 2015, 24(4): 813-817

[13] Wang Xinxia, Wang Ke, Jiao Dongxiang, et al. Research on anti-stealing based on normal distribution outlier algorithm [J]. Electrotechnical Application, 2017, 36(7): 60-65 (in Chinese)
(王新霞, 王珂, 焦东翔, 等. 基于正态分布离群点算法的反窃电研究[J]. 电气应用, 2017, 36(7): 60-65)

[14] Zhang lei, Zhang Yi. Big data analysis by infinite deep neural networks [J]. Journal of Computer Research and Development, 2016, 53(1): 68-79 (in Chinese)
(张蕾, 章毅. 大数据分析的无限深度神经网络方法[J]. 计算机研究与发展, 2016, 53(1): 68-79)

[15] Li Zhanjiang. Small enterprise credit evaluation model based on hierarchical logistic regression [J]. Statistics & Decision, 2016(7): 178-182 (in Chinese)
(李战江. 基于分层逻辑回归的小企业信用评价模型[J]. 统计与决策, 2016(7): 178-182)

Shi Yuliang, born in 1978. PhD, professor. His main research interests include cloud computing, database and privacy preserving.

Rong Yiping, born in 1977. Senior engineer. His main research interests include electricity marketing, power information.

Zhu Weiyi, born in 1971. Senior engineer. His main research interests include electricity marketing, power information.