第46卷 第1期 2022年1月10日
DOI：10.7500/AEPS20200204001
电力系统自动化
Automation of Electric Power Systems
Vol. 46 No. 1 Jan. 10，2022

数据驱动窃电检测方法综述与低误报率研究展望

金 晟[1]，苏 盛[1]，薛 阳[2]，杨艺宁[2]，刘 厦[2]，曹一家[1]
（1. 智能电网运行与控制湖南省重点实验室（长沙理工大学），湖南省长沙市 410114；
2. 中国电力科学研究院有限公司，北京市 100192）

摘要：配电系统窃电是造成电网非技术损失的主要原因，是供电企业运营管理中长期面对的痼疾。用电信息采集系统采集的海量用户数据使得开展数据驱动的用电异常检测、准确识别窃电用户成为可能。受用户用电行为多样性影响，数据驱动的窃电检测方法的误报率在某些场景下尚难以满足实践需求，严重制约了该类方法的工程应用。首先，介绍了窃电实现手法；然后，梳理了在实践中得到工程应用的窃电检测方法以及数据驱动窃电检测方法的基本思路和局限性；在此基础上，结合工程应用对窃电检测评价指标的差异性需求，分析指出提取的可用信息不足、特征指标项灵敏性和可靠性不高是阻碍数据驱动窃电检测方法走向工程实用的主要原因。最后，从算法设计、状态空间细分以及特征指标项设计选择等不同层面对低误报率窃电检测进行了展望。

关键词：窃电检测；低误报率；数据驱动；特征工程；状态空间

0 引言

为促进经济发展、降低实体企业用能成本，近年来，中国在持续推进新一轮电力市场化改革的基础上多轮次大幅调降一般工商业电价，供电企业承担降价总金额近3 000亿元。同时，外部环境日益复杂和全社会用电量增速放缓，也对供电企业的精益化运营提出了重大挑战[1-2]。为维系供电企业长远发展，亟待开展内部挖潜，提高经营收益。用户窃电直接造成供电企业净收益流失，近年来出现的比特币挖矿窃电等新现象，使得用电管理形式更为严峻。2019年，中国破获多起比特币挖矿窃电案件，其中仅江苏省镇江市破获的一起案件涉案金额就高达2 000万元。因此，研究满足工程应用要求的窃电检测方法为开展指向性的窃电稽查提供决策支持，是当前亟待解决的问题[3]。

智能电表和用电信息采集系统的普及应用，使得供电企业掌握的用户用电数据从月结抄表电量显著增长为15 min/30 min间隔的高密度计量数据，并能记录电表开盖次数和开盖时间等辅助信息，为分析用户行为、识别异常用户提供便利。电力工程技术人员从基本物理规律出发，提出了一些基于简单规则的窃电检测方法，能准确识别窃电行为，并在工程实践中得到推广应用。由于窃电的现象表征与电表接线形式和窃电方式紧密关联，该类方法仅适合于采用特定手法窃电的用户。高校院所科研人员主要从用电异常识别的角度探索了数据驱动的窃电检测方法。目前，该类研究主要基于特定数据集进行测试分析，在工程应用中能否满足生产需要尚待实践验证。由于用电信息采集系统和营销系统中记录有丰富的用户数据，在深入分析不同类型用户行为特性的基础上，有可能通过模型和算法等层面的优化设计，提出能够满足工程应用需求的窃电检测方法。

本文首先结合电表接线方式介绍了不同窃电手法的实现方式和现象表征；然后，梳理了工程技术人员总结的窃电检测方法的基本思路，进而分析数据驱动窃电检测方法异常识别思路及存在的问题。在此基础上，结合供电企业在人力资源约束下对窃电检测评价指标的差异性要求，分析指出不平衡样本和跨类杂糅用户的用电行为特性对应状态空间太过庞大，导致数据驱动窃电检测方法特征指标项的灵敏性和可靠性难以满足要求等问题。最后，从算法设计、特征指标选取以及用电异常的状态空间细分等层面对低误报率（false positive rate，FPR）窃电检测进行展望。

收稿日期：2020-02-04；修回日期：2020-06-23。
上网日期：2021-03-09。
国家自然科学基金资助项目（51777015）；国家电网有限公司总部科技项目“反窃电及稽查监控关键技术研究”；湖南省自然科学基金资助项目（2020JJ4611）。

1　窃电手法分析

常见的窃电手法按是否涉及计量装置分为2类：与计量装置无关的窃电手法主要包括绕表用电和用户私自增容偷逃基本电费2类；与计量装置有关的窃电手法主要包括欠压型、欠流型、移相型和扩差法窃电[4]。欠压型、欠流型及移相型窃电手法通过改变电压、电流及电压与电流之间的相角达到窃电的目的，扩差法窃电通过改变电表内部结构或干扰电表运行扩大电表计量误差，从而达到窃电的目的。4类窃电方式的具体实现手法如图1所示[5]。

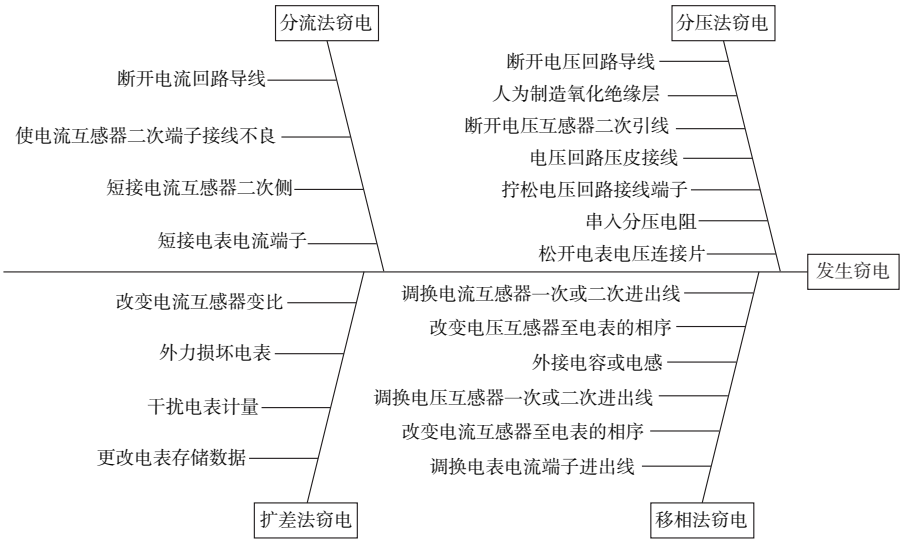

图1　窃电手法鱼骨图

Fig. 1　Fishbone diagram of electricity theft means

根据用户类型和电表接线方式的差异，各种窃电方式可以表现出不同的指征。如在低压单相用户中采用分流法使得电流不经零线回流，将使得低压用户出现零序电流。工商业用户一般采用三相三线制或三相四线制接线。其中，三相三线接线方式下不存在零序电流，常采用两元件电表计量（计量两相电流），而三相四线用户一般采用三元件电表计量（计量三相电流）。对应于涉及电流互感器的窃电手法中表现出的指征也有所差异。

根据供电企业稽查确认窃电用户的案例分析，按照窃电持续性和窃电程度对用电量时序的改变方式，可将窃电时电量变化分为表1所示的8类[6-9]。

第1类持续按照固定比例缩小计量电量，在实际中多对应单相/两相电流分流、单相电流反向或更换互感器等。为逃避稽查，第2类窃电用户在窃电回路中采用可控开关间断性地将电量降为0，可在每天的负荷高峰时段或无稽查风险时段窃电。第3类窃电用户表现为全时段零用电，多为数量庞大、难以上门稽查的低压居民用户。第4类和第5类为持续或断续按时变比例$f(t)$随机减少用电量的窃电用户，在实践中表现为篡改电表软件的智能化窃电或工业企业中部分车间整体绕表用电的窃电用户。第6类窃电用户电量表现为在用户电量均值基础上持续或按时变比例$f(t)$减少用电量。第7类窃电用户电量则表现为用电量均值。第8类窃电用户的总用电量不变，通过颠倒用电时序降低电费成本。

表1　窃电时电量变化趋势

Table 1　Trends of electricity changes with electricity theft

种类	形式
1	$X_t=ax_t$，其中$0.2<a<1$，a为随机数
2	$X_t=f(t)x_t$，其中$f(t)=\begin{cases}0 & t_1<t<t_2\\1 & \text{其他}\end{cases}$，$t_1$和$t_2$分别为窃电时段起止时间
3	$X_t=0$，全时段持续为零电量
4	$X_t=f(t)x_t$，其中$0<f(t)<1$
5	$X_t=f(t)x_t$，其中$\begin{cases}0<f(t)<1 & t_1<t<t_2\\f(t)=1 & \text{其他}\end{cases}$
6	$X_t=f(t)\bar{x}$，其中$0<f(t)<1$
7	$X_t=\bar{x}$
8	$X_t=x_{24-t}$

注：X_t为窃电后显示用电量；x_t为正常用电量；$\bar{x}$为用电量均值；x_{24-t}为颠倒用电时序后的电量。

2　工程应用中的窃电检测方法

供电企业用电管理人员数量有限，现场稽查困难，窃电检测是经营管理中长期存在的问题。智能电表得到普及应用前，工程技术人员在可用数据极

度匮乏的条件下，摸索总结出了不依赖用户高密度计量数据的正常用户排除方法[10]，具体思路如下。

1）信用过滤：根据行业和客户信用评级过滤极不可能窃电的高信用等级客户。

2）线损过滤：计算客户接入线路/台区的线损率，过滤线损低于阈值的线路/台区下的接入用户。

3）功率因数过滤：一般窃电难以在窃电的同时保持功率因数不变，可以根据功率因数统计分布是否明显偏离过去的统计分布，过滤功率因数无明显变化的正常客户。

4）电量纵向比较过滤：以用户窃电导致用电量下降为前提，将用户电量与去年同期及近期电量相比，过滤电量无明显突变的客户。

采用上述方法，可在有限数据支持下将用电稽查人力资源靶向性地聚焦于高危窃电用户，提高窃电检测命中率（true positive rate，TPR），相关思路对于当下也极具参考价值。

智能电表的普及应用极大地丰富了可供挖掘分析的用户用电数据。电力工程技术人员根据生产实践经验，总结了一些具有物理意义的窃电检测实用性指标[11-13]，主要包括：

1）单相低压用户剩余电流：单相用户电流绕过入户中性线而经外接中性线入地时可减少电度计量，根据单相电表的剩余电流可准确识别窃电用户。

2）功率反向或缺相：绝大多数用户未配置分布式电源，一般用电表计量下网电量。用户窃电可能造成单相或三相功率反向或缺相，可根据是否存在明显的反向功率和数据缺相识别异常用户。

3）功率因数突变：部分窃电手法通过改变电表接线调整电流、电压的夹角实现窃电，窃电时伴有功率因数突变，可将功率因数突变作为辅助判据。

4）增量物理指标：电力装备企业根据多数窃电方式需要改动电表或外置强磁场干扰的特点，研制了具有开盖检测和强磁检测能力的防窃电电表，可根据记录的电表开盖次数、时间以及强磁场干扰，识别窃电用户及窃电起止时间。

以上几种在工程实际中得到推广应用的窃电检测方法所用特征指标项均具有明确的物理意义，能直观准确标识窃电行为，满足生产实践的准确性要求。由于窃电的现象表征与电表接线形式和窃电手法紧密关联，虽然该类方法可准确识别采用特定手法窃电的某些类型用户，但也存在适用范围有限的缺陷。

3　数据驱动的窃电检测方法

用电信息采集系统、电力营销系统及配电自动化系统可提供完整的用户台账、电量计量及接入配电线路和台区信息，有力地支撑了数据驱动窃电检测技术的发展[14-17]。数据驱动窃电检测方法根据实现机理可分为基于无监督学习的聚类分析、基于有监督学习的分类和基于配电网状态估计，如图2所示。下面分别综述这3类方法研究现状与存在的问题。

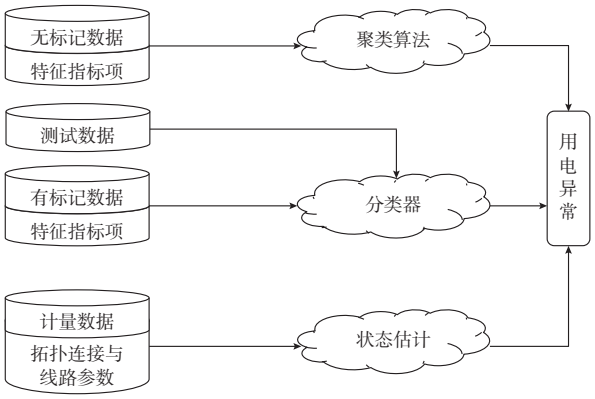

图2　数据驱动窃电检测流程

Fig. 2　Detection process of data-driven based electricity theft

3.1　基于聚类的窃电检测

由于同类型用户应具有相近的用电模式，有可能利用计量和营销数据，采用无监督的方式对用户用电特征指标项进行聚类，然后将不符合多数用户用电行为模式特征的少数用户识别为异常用户[1,8,18-22]。在异常用电检测过程中，随着用户数据采集频率的不断提高及数据分析维度的扩展，客户用电行为表现出的模式特征愈加复杂[23-27]，需要先根据用户行为模式来提炼和筛选可有效表征用户用电行为特性的特征指标项，再采用聚类算法找出显著偏离正常用户聚类簇的异常用户。

基于聚类的窃电检测方法的核心在于特征指标项选择和算法设计两方面。在特征指标选择上，由于窃电常表现为用电量的趋势性下降、日负荷曲线的异常以及报装容量利用率偏低等形式，多以上述表征为依归，设计特征指标项。在算法层面上，常见的聚类算法可分为基于划分、基于层次、基于密度和基于网络等，应用于异常检测的聚类方法主要有基于划分和基于密度2类。基于划分的聚类方法将用户的特征集合经过划分后，使得子集合中离中心较远的离群点作为异常点；基于密度的聚类则认为远离高密度点且自身处于低密度区域的点为异常点[19]。基于聚类的窃电检测方法对比如表2所示，详述如下。

表2　基于聚类的窃电检测方法对比
Table 2　Comparison of electricity theft detection methods based on clustering

文献	方法	特征指标	用户类别	数量	评价指标	计算方式
[1]	FCM	月均用电量,月最大用电量,用电量标准差,现场稽查记录次数,区域平均用电量	居民,商业,工业,政府	20 126	精确率,TPR	离线
[8]	CFSFDP	负荷时间序列	居民,中小企业	5 000	AUC,MAP	离线
[18]	离群点检测	电压、电流不平衡率	低压台区用户	1 278		离线
[19]	OPF	需求负荷,基本负荷,最大负荷,无功功率,变压器容量,功率因数,报装容量,负荷系数	工业,商业	8 126	准确率	离线
[20]	网格化LOF	变动性、波动性、趋势性及其他指标		6 200	ROC,AUC,CR	离线
[21]	GKLOF	变动性、波动性、趋势性及其他指标	居民,商业	5 000	TPR,FPR	离线
[22]	流式DBSCAN	实时电流,用电量		百万级		在线
[23]	DBSCAN	波动区间	商业,居民		ROC	在线

文献[1]采用模糊C均值(fuzzy C-means,FCM)算法对包括居民、商业、工业及政府单位在内的用户数据集进行聚类分析,除使用连续6个月的月均用电量、最大月用电量、月电量方差及当地平均用电量等指标外,还将标识用户历史用电异常度的现场稽查次数纳入特征指标集进行聚类,分析得到能代表各聚类簇的典型向量。最后计算用户当前用电信息与所属聚类簇典型向量间的欧氏距离,评估用户异常风险。文献[8]将线损相关性分析与针对日负荷曲线的密度峰值快速搜索聚类(clustering by fast search and find of density peaks,CFSFDP)结合,采用爱尔兰居民用户及中小型企业500日的用电量数据进行用电异常检测。仿真结果表明,模型的曲线下面积(area under curve,AUC)指标和平均精度均值(mean average precision,MAP)指标在不同窃电形式下均可达70%以上。文献[18]利用分压或分流窃电时计量的电压、电流数据形成离群点的特征,对高损台区的低压用户的电流、电压数据进行聚类分析,识别采用欠压法和欠流法窃电的用户。文献[19]采用最优路径森林(optimum-path forest,OPF)聚类方法,按有功电量、报装容量、最大需量、无功电量、专变容量、功率因数和负荷系数(单位时间平均电量与最大需量之比)等用户用电行为特征指标项,对8 126户商业和工业用户数据进行聚类分析,仿真分析结果表明OPF聚类、K均值聚类、高斯混合模型(Gaussian mixture model,GMM)及近邻传播(affinity propagation,AP)算法识别工业和商业用户用电异常的准确率均在60%左右。

特征指标项之间可能存在强关联性,为了降低特征指标项的信息重叠,提高异常检测效率,文献[20]在用户多日负荷均值差值与变化斜率的变动性指标、前后多日负荷标准差的波动性指标及负荷升降趋势性指标的基础上,采用主成分分析提取主成分因子后,利用网格化局部离群因子(grid-based local outlier factor)算法检测离群的用电异常用户,能够筛选出低密度区域的数据点,提高算法效率。本文采用省级电网6 200个用户18个月的负荷数据,根据受试者工作特性(receiver operating characteristic,ROC)曲线、AUC及累计查全率(cumulative recall,CR)曲线对模型进行综合评价。文献[21]提出一种基于高斯核函数改进的电力用户用电数据离群点检测方法,针对文献[8]相同的用户数据,首先采用FCM聚类方法将用户聚成多个类簇,然后采用文献[20]相同的指标项进行特征集降维处理后,用高斯核密度局部离群因子(Gaussian kernel density based local outlier factor,GKLOF)算法进行聚类分析,识别离群的异常用户。文中采用TPR和FPR进行算法评价,仿真结果表明所提算法在TPR和FPR上均优于原始局部离群因子(local outlier factor,LOF)算法。

前述研究多以静态离线数据分析为主,难以适用于大数据流量和储存海量数据的实际生产系统。文献[22-23]提出结合实时数据进行同步异常检测,以快速发现用户用电异常。文献[22]针对在大规模数据处理平台上进行海量用户异常识别的需要,结合分布式流式计算平台,基于用户个体在纵向时间和横向空间上的聚类特性,设计并实现了流式基于密度的含噪声空间聚类(density based spatial clustering of applications with noise,DBSCAN),对比分析表明所提算法在异常检测上较原始DBSCAN算法更有效。文献[23]加入关联分析思想构造关联规则,以用户单位时间内的用电量波动作为特征指标项进行密度聚类,并在用户波动区间分簇后对标准化的分簇对象计算离群对象得分,所提算法能及时分析异常用电,其单位时段$[t_{i-1},t_i]$内波动量$b_{i-1,i}$的计算公式为:

$$b_{i-1,i}=a_i-a_{i-1} \tag{1}$$

式中：a_i为单位时段$[t_{i-1},t_i]$内的用电量。若包含m个样本点的单用户用电量序列$P=\{p_{t,1},p_{t,2},\cdots,p_{t,m}\}$，$p_{t,m}$为用户在对应时刻$t$的用电量，则$a_i=p_{t,i}-p_{t,i-1}$，以此构建单位时段的波动量序列$B=\{b_{0,1},b_{1,2},\cdots,b_{m-1,m}\}$。

综上，基于聚类的窃电检测方法一般使用表征用电量的趋势性下降、日负荷曲线的异常以及报装容量利用率偏低等特征指标项进行特征优选和聚类分析[19-20]。需要指出的是，不同行业用户用电行为特性存在显著差异，部分行业用电需求直接取决于订单需量，用户用电量的大幅或趋势性波动是常态。此外，窃电多发的工程建筑类用户的用电行为极不规律，基于日负荷曲线聚类识别用电异常的实用性也需要在工程实践中进行考验。

3.2　基于分类的窃电检测

分类属于有监督学习[28-35]，其采用已知类别标签的样本训练分类器，由经过训练的分类器根据输入特征量将无标签样本分类。与基于聚类的窃电检测方法类似，基于分类的异常用电检测需要采用适当的算法，根据选择的特征指标项将用户划分为正常和异常2类，差异之处在于后者可以利用大量已知类别的样本训练分类器。实际窃电样本往往明显少于正常用电样本，这将带来突出的不平衡样本集问题。基于分类的窃电检测方法对比如表3所示，详述如下。

表3　基于分类的窃电检测方法对比

Table 3　Comparison of electricity theft detection methods based on classification

文献	方法	特征指标	用户类别	数量	异常占比/%	评价指标	计算方式
[28]	相关系数，贝叶斯网络，决策树	月最大/最小用电量，信息获取次数，有功无功电量比，报装容量利用率		38 575	2.44	TPR	离线
[29]	SVM	月平均日负荷曲线，信用等级		186 968	14.99	TPR	离线
[30]	稀疏随机森林	日用电量	工业，商业，非工业，居民	5 690		TPR，FPR	离线
[31]	KNN	异常数量，毛刺宽度，重复出现的读数最大值	工业，居民	193		FPR	离线
[32]	ELM	负荷时间序列	商业	1 500		准确率	离线
[33]	MLP	需求负荷，基本负荷，最大负荷，无功功率，变压器容量，功率因数，装机负荷，负荷系数	工业，商业	9 131		MSE，准确率	离线
[34]	CNN	负荷时间序列		42 372	8.53	AUC，MAP	离线
[35]	堆叠去自相关编码器，SVM	负荷时间序列	居民，商业	5 000	20.00	TPR，FPR，BDR	离线

文献[28]以窃电将造成用电量明显下降为指引，将过去2年的用电数据划分为3个时间窗，第1个时间窗为最早16个月的用电数据，标识用户在正常状态下的用电行为；随后2个月为第2个时间窗，标识窃电导致电量下降过程；最后6个月标识窃电之后达到的稳定状态。在此基础上，定义特征指标项，采用皮尔森相关系数识别用电量突降的异常用户。除此以外，还根据月最大/最小用电量、月无功有功电量比、报装容量利用率等特征指标项，采用贝叶斯网络和决策树识别用电异常的用户。仿真结果表明，所提组合检测方法能覆盖不同类型窃电手法，提高检测准确率。文献[29]使用支持向量机(support vector machine，SVM)作为分类器，除使用用户每个月平均日负荷曲线外，还将用户信用评级作为特征输入，并对SVM进行了参数优化。本文测试数据集包括186 968户电力用户，通过TPR评价模型性能。文献[30]提出基于稀疏随机森林模型的用电异常检测方法，该方法利用窃电导致用电量下降的特性，采用日用电量为特征指标项。首先利用时间窗函数与有放回重采样，建立用电行为模式信息簇，然后基于随机权网络得到随机森林模型并稀疏化来识别用电异常。此外，通过用户异常度累积的方式，可在恰当阈值设置下避免干扰因素造成的短期负荷骤降引起的误报。该方法使用数据集为5 690个城镇用电客户的负荷数据，涵盖大工业、商业、非居民照明、非工业及居民5个类别，最后以TPR、FPR衡量模型性能好坏。文献[31]借鉴工程实践的判断规则，将根据功率反向等判据检测的指标用作特征项，将用户用电量时序数据异常分为normal、change和complex这3种类型，其中normal类型的异常体现为毛刺，change类型的异常体现为曲线有大幅度的下移，而complex类型则体现为杂乱无章且无规律，利用K近邻(K-nearest neighbors，KNN)算法对193个居民和工业用户异常数据进行

分类训练，分类准确率接近80%。

文献[32]首先根据用户多日平均的日负荷曲线确定工作日和节假日的高峰负荷上下限，然后使用极限学习机(extreme learning machine，ELM)作为分类器，将标幺化的用户日负荷曲线作为特征输入，判断用户是否窃电。所使用数据集包括1 500户商业用户，分类准确率可达54.61%。文献[33]将多层感知机(multi-layer perceptron，MLP)神经网络作为分类器，所使用数据集分为2个部分，前者为3 486户工业用户，后者为5 645户商业用户，所使用的特征指标集与文献[19]相同，最后以均方误差(mean-square error，MSE)和准确率衡量方法有效性。

上述基于聚类和分类的窃电检测方法所采用的特征指标项，多根据研究人员对于窃电用户在用电行为特性上的先验知识来设计和选择。近年来，机器学习领域的研究进展表明，深度学习可以从输入的低阶特征中提取高阶特征，从而显著提高系统的泛化能力。基于此，文献[34-35]分别采用深度卷积神经网络(convolutional neural network，CNN)和基于堆叠去相关自编码器的SVM，利用负荷时间序列数据识别窃电用户，在测试所用数据集上取得了较好的识别效果。

从可用信息多寡角度来看，基于分类的异常检测方法增加了带分类标签样本信息，检测效果有可能优于基于聚类的异常检测方法。基于分类的窃电检测中，除正面与负面样本不平衡特性影响外，正面与负面样本之间行为特性的差异性以及负面样本代表性的影响更为突出。已有分类检测文献多围绕用电量突降设计特征指标项，这实际上隐含正常用户用电基本平稳的假设。但相当一部分行业电力用户的用电量可能本身就并不平稳，而设备大修、停产改造等原因也会造成用户持续低电量。用户正常用电行为下的电量波动和窃电造成的电量波动之间的区别并不明显，容易混淆。仅从算法层面进行优化改进，不一定能取得满意的检测效果。

3.3　基于状态估计的窃电检测

电网本身的物理特性决定了各节点的节点电压和注入功率等状态量具有强耦合性，服从潮流方程约束，可以采用状态估计的方式根据其他节点的量测量估计目标用户的状态量[36]。用户窃电时篡改的只是本地的计量数据，实际上很难通过多个用户协同篡改各自的计量数据实现满足潮流约束的窃电。因此，有可能在用户计量数据的基础上，结合配电网的节点电压等数据，开展基于配电网状态估计的窃电检验[37-39]。

基于配电网状态估计的窃电检测除要求掌握详细的网络拓扑结构和参数外，还要求线变户表关系正确、与实际系统严格一致，工程应用中局限性较强。尽管采用文献[40-41]提出的方法可借助无线传感器和射频识别标签提高状态数据采集精度，但也需额外增加硬件和运维成本。

4　窃电检测评价指标与工程应用需求分析

进行异常检测时，往往采用表4所示混淆矩阵及其衍生指标评价检测效果。混淆矩阵将所有用户按照其实际归属和检测归属分为真阳性(true positive，TP)、真阴性(false negative，FN)、假阳性(false positive，FP)和真阴性(true negative，TN)，其中：TP和TN为正确分类的部分，比例越高说明检测效果越好；FP为误报而FN为漏报。

表4　异常用电检测中应用的混淆矩阵
Table 4　Confusion matrix applied in anomaly electricity detection

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

以混淆矩阵为基础，可以推导出多个分类器的评价指标。常用指标主要有准确率[19,32-33]、精确率[1]、TPR[21,28-30,35]和FPR[21,30-31,35]等，其中：准确率为所有预测样本中预测正确的比例；精确率为所有预测为异常用户的样本中实际异常用户的占比；TPR为实际异常用户中预测正确样本的占比；FPR为误检为异常用户的正常用户在所有正常用户中的占比。各指标准确定义如式(1)至式(4)所示。

$$k_{\mathrm{Accuracy}}=\frac{P_{\mathrm{TP}}+P_{\mathrm{TN}}}{P_{\mathrm{TP}}+P_{\mathrm{TN}}+P_{\mathrm{FN}}+P_{\mathrm{FP}}} \tag{1}$$

$$k_{\mathrm{Precision}}=\frac{P_{\mathrm{TP}}}{P_{\mathrm{TP}}+P_{\mathrm{FP}}} \tag{2}$$

$$k_{\mathrm{TPR}}=\frac{P_{\mathrm{TP}}}{P_{\mathrm{TP}}+P_{\mathrm{FN}}} \tag{3}$$

$$k_{\mathrm{FPR}}=\frac{P_{\mathrm{FP}}}{P_{\mathrm{FP}}+P_{\mathrm{TN}}} \tag{4}$$

式中：k_{Accuracy}、$k_{\mathrm{Precision}}$、k_{TPR}、k_{FPR}分别为准确率、精确率、TPR、FPR指标；P_{TP}为实际异常用户被检测为异常用户的数量；P_{FN}为实际异常用户被检测为正常用户的数量；P_{FP}为实际正常用户被检测为异常用户的数量；P_{TN}为实际正常用户被检测为正常用户的数量。

除了静态和单一指标外，还可以采用多种或动态指标衡量模型整体可信度。附录A表A1所列为

从混淆矩阵衍生出的多种模型评价曲线。

1）精确率-召回率（precision-recall，P-R）曲线[42]的横坐标和纵坐标分别为TPR和精确率，二者的值越趋近于1则模型的效果越好，故P-R曲线凸向（1，1）。

2）ROC曲线[43]的横坐标为FPR，纵坐标为TPR。FPR越趋近于0且TPR越趋近于1时模型的检测效果越好，故ROC曲线的图像凸向（0，1）。

3）提升曲线[44]与增益曲线[45]的横坐标均为d_{depth}，其计算公式为：

$$d_{\text{depth}}=\frac{P_{\text{TP}}+P_{\text{FP}}}{P_{\text{TP}}+P_{\text{FP}}+P_{\text{TN}}+P_{\text{FN}}} \tag{5}$$

在异常检测中，d_{depth}代表预测为异常用户的样本占整体样本数的比例。提升曲线与增益曲线的纵坐标分别为L_{Lift}和G_{Gain}。

$$L_{\text{Lift}}=\frac{P_{\text{TP}}}{P_{\text{TP}}+P_{\text{FP}}}\frac{1}{d_{\text{depth}}}=\frac{k_{\text{Precision}}}{k_{\text{Accuracy}}} \tag{6}$$

$$G_{\text{Gain}}=\frac{P_{\text{TP}}}{P_{\text{TP}}+P_{\text{FP}}} \tag{7}$$

假设随着模型中设定阈值减小，更多用户被划分为异常样本，即d_{depth}（检测为异常用户占所有检测用户的比例）值变大，L_{Lift}衡量了0与不利用窃电检测模型相比、使用窃电检测模型带来的检测效果提升程度。当阈值为0且不使用检测模型时，d_{depth}为1，此时$L_{\text{Lift}}=[P_{\text{TP}}/(P_{\text{TP}}+P_{\text{FP}})](P_{\text{TP}}+P_{\text{FP}}+0+0)/(P_{\text{TP}}+0)=1$。提升曲线偏离1越远，表示与不使用检测模型相比，使用检测模型提升检测效果的程度越大，效果越好；增益曲线越接近（0，1）时，表明只需检测较少的样本即可取得较高的准确率，模型检测效果越好。文献[20]对模型评价所使用的CR曲线横坐标也是d_{depth}，纵坐标为TPR。曲线越接近（0，1），模型检测效果越好，检测较小比例样本即可取得较高的检出率。

4）柯尔莫可洛夫-斯米洛夫（Kolmogorov-Smirnov，KS）曲线[46]的横坐标为检测模型设定的阈值，纵坐标为TPR与FPR的差值，TPR与FPR的差值越大表明检出率越高而误检率越低，检测模型的性能越好。KS值的计算公式为$|\max(k_{\text{TPR}}-k_{\text{FPR}})|$，即TPR与FPR差值绝对值的最大值。当模型取得KS值处对应的横坐标阈值时，检测效果最佳。

除前述模型评价曲线外，诸如AUC[8,21,34]、MSE[33]、MAP[34]和贝叶斯检出率（Bayesian detection rate，BDR）[35]等指标在窃电检测模型评价中也有应用。基于混淆矩阵及衍生评价指标在应用中主要存在以下2点问题。

1）从工程技术人员应用反馈来看，阻碍数据驱动窃电检测方法走向推广应用的主要瓶颈在于FPR偏高。由于供电企业用电管理人力资源有限，加之地方政府为改善营商环境限制供电企业频繁进行现场稽查，FPR高的窃电检测方法难以得到推广应用。

2）从表2和表3可以看出，既有研究多侧重提高窃电检测准确率，与工程应用要求的低FPR并不一致，在不均衡样本中以准确率高为目标进行异常检测，本身就是一种误导。一般认为电力用户中窃电用户占比不高，是典型的不平衡样本。当窃电用户占比为1%时，窃电检测只需要将所有样本判定为正，检测准确率就可达到99%，但此指标实际上不具有参考价值。

从供电企业角度来看，杜绝窃电现象并不需要查处全部窃电用户，准确检测部分窃电用户进而震慑其他用户，同样可以达到目的。供电企业用户数量庞大，其中相当比例为窃电异常用户。从实际工作出发，漏报部分窃电用户对于开展用电稽查影响不大，但出现误报将使用电稽查失去靶向性，进而导致稽查人员放弃使用数据驱动窃电检测方法。综上，工程应用对于窃电检测的要求是可容忍一定程度的漏报率，并尽可能降低FPR。

5　低FPR检测方法研究展望

从工程应用角度来看，既有数据驱动的窃电检测方法在正负样本的不平衡特性和基于特征指标项的检测判据可靠性等方面存在较明显缺陷。除在样本不平衡条件下以最大化检出率和检测准确率为目标进行算法优化设计、导致窃电检测FPR偏高以外，更突出的问题集中在特征指标的选择上。数据驱动的窃电检测往往以窃电导致低电量或用电量突降异常为依归，根据泛化的窃电表征（如用电量陡降或趋势性下降、报装容量利用率低和日负荷曲线离群异常等）设计和选取特征指标项，此时实际上隐含正常用户用电基本平稳的假设。但从图3中实际工业用户的用电数据来看，相当部分行业的用户按订单安排生产，用电量并不具有平稳特性。正常情况下，用户日电量波动可达30%～50%，而单相分流窃电时电量波动也在30%左右，很容易和正常波动混淆导致误判。用户设备大修、停产改造、消防整改、环保检查和安全检查等原因造成持续的用电量突降，也可导致误判。

跨行业杂糅用户用电行为特性对应的状态空间太过庞大，是导致难以设置能准确刻画用户用电行为特性的特征指标项，进而影响异常检测灵敏性和

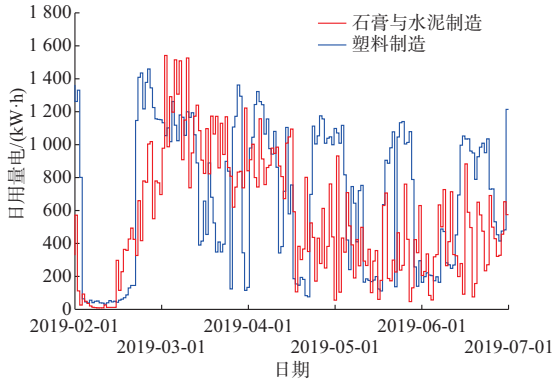

图3　典型行业日用电量曲线
Fig. 3　Daily power consumption curves of typical industry

可靠性的关键。因为正常和异常用户的电量波动本身就容易混淆，所以围绕电量设计的特征指标项存在可用信息不足的缺陷。仅根据电量信息，很难判断电量异常是否为窃电所致。实际上，负荷数据可用作用户的负荷指纹，从中分析提炼用户所处生产经营状态的增量信息，从而识别低电量异常是否对应窃电。此外，利用用户窃电与接入线路/台区线损之间的关联性，引入台区/线路线损增量信息，也可提高窃电检测的靶向性。

根据以上分析，可按图4所示，从不平衡样本的算法设计、基于生产经营状态识别的窃电二次筛查、考虑行业用电特性差异的特征指标项提取以及高损线路/台区线损电量归因分析4个方面研究低FPR窃电检测方法。

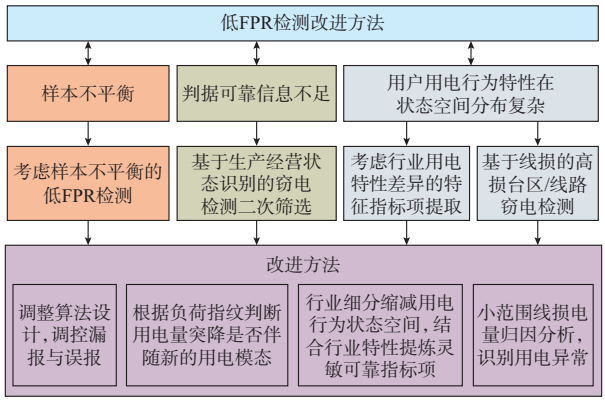

图4　低FPR检测改进方法
Fig. 4　Improved method for detection with low FPR

5.1　考虑样本不平衡的低FPR窃电检测算法设计

样本不平衡是基于分类的异常检测的常见问题。采用不平衡样本训练分类器时，以提高分类准确率为目标设计算法会过多关注多数类样本而忽略少数类样本，从而降低少数类异常样本的分类性能。

利用供电企业进行窃电检测时要求FPR尽可能低而可以接受一定比例漏报的特点，基于分类的窃电检测算法可从算法层面进行优化设计，调控漏报率与FPR之间的平衡，降低由样本不平衡导致的高FPR，具体措施如下。

1)文献[47]按最大化分类准确率为目标设计算法，根据不同窃电手法的数学描述产生大量训练样本来消除不平衡水平，有助于提高异常检测的准确率。

2)采用随机欠采样(random under sampling，RU)技术，在训练样本中多次有放回地随机抽取少量样本作为弱分类器中的训练样本[48]，也可提高不平衡样本条件下的异常检测效果。

3)基于合成少数类过采样技术(synthetic minority oversampling technique，SMOTE)，从每个少数类样本的最近邻中随机选择1个样本，然后在2个样本间的连线上随机选择一点作为新合成的样本，也是一种有效的方法[49]。文献[50]采用改进SMOTE，通过合成小类别边界数据样本来加强对样本边界的分类能力，防止分类界面模糊和类别边界样本分类困难。

4)优化分类算法的分类阈值。窃电检测为二分类检测，一般采用0.5作为正常和异常的分类界限。使用训练集训练分类器时，搜索以取得分类器分类准确率最高/FPR最低时对应的分类阈值作为最优阈值，也可降低FPR。

5.2　基于生产经营状态识别的窃电二次筛查

基于电量异常识别窃电的突出问题是无法识别是用户低电量生产经营状态还是窃电造成用户低电量异常，容易将用户正常生产经营状态变化误判为窃电。

对于特定工商业用户而言，其用电设备构成基本固定。在特定生产经营状态下使用的设备组合和对应的用电模式是基本确定的。每种生产经营状态下投运电气设备的组合决定其对应的用电模式。正常节假日时，用户投运设备不同，负荷明显低于工作日，对应的用电模式也有显著差异。正常用户经过一段时间之后，将会遍历自身各种生产和经营状态对应的用电模式。

利用用户用电行为对应一定的生产经营状态的特点，可以将用户一日的三相计量数据作为负荷指纹，标识用户当天的用电行为模式及所属的生产经营状态。对于正常用户而言，电量突降只是从正常的高电量生产经营状态转换进入低电量生产经营状态。而窃电用户电量突降时对应的用电模式与生产经营状态都可能与正常的生产经营状态不同。因此，根据负荷指纹识别用户的生产经营状态，可以提

供电量之外的增量信息，从而规避仅基于电量的窃电检测存在的信息不足问题，降低窃电检测FPR。

5.3　考虑行业用电特性差异的特征指标项提取

既有基于分类或聚类的窃电检测方法多倾向于跨行业进行窃电检测，一般仅按照居民、商业、工业等大类来分类窃电检测。由于跨行业混杂的用户在用电行为特性上具有显著的差异性，一方面在选择特征指标项时无法针对个别行业用电行为特性设计和选取指标，而只能选择用电量突降、报装容量利用率偏低和日负荷曲线异常等脱离行业特性的泛化指标；另一方面，不同行业用户的行为模式差异巨大，又在灵敏性和可靠性上对选出的指标造成了负面影响。

实际上，细分行业的用户在用电行为特性上具有强相似性。如图5所示的路灯负荷就具有明确的日负荷曲线特征。由于路灯供电线路散布于城市四处，邻近居民店面盗接路灯线路的窃电案例屡见不鲜。路灯一般采用小容量专变供电，只有供电电量而没有用电电量计量，不能根据线损率识别异常。但结合路灯负荷的运行时间及点亮模式特点，有可能采用负荷波动性作为特征指标项来设计检测模型，识别用电异常的路灯专变用户。

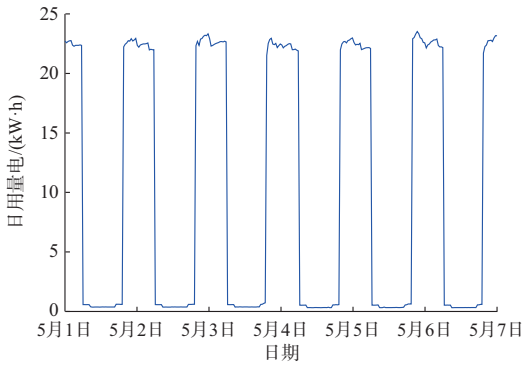

图5　典型路灯负荷曲线

Fig. 5　Typical street light load curve

按细分行业设置特征指标项的实质是充分挖掘和利用对特定行业用电行为的先验知识。与前述路灯专变用户类似的，其他行业也可以根据行业用电行为特性提取能有效刻画用电行为特性和识别用电异常的特征指标项。

5.4　高损台区/线路下的线损电量归因分析

绝大多数既有窃电检测方法共同的缺陷是需要检测到窃电导致的用电行为突变，但供电企业一般仅保留近几年的用电数据，不一定能覆盖窃电导致突变的时间点。此外，部分用户甚至可能在入网时就配置了错误的电流、电压互感器，无从检测到窃电导致的用电行为突变。

由于配电线路/台区下一般仅接入有限数量的用户，非技术线损电量由接入用户造成。为缩小目标、简化问题，传统上在极度匮乏用户用电数据的条件下，主要围绕高损台区/线路进行针对性的用电稽查，是缩小状态空间、降低窃电检测FPR的实用方法，有可能为高效准确检测包括成比例无突变窃电在内的用电异常提供解决思路。

由于台区/线路线损电量与用户窃电电量之间存在对应关系[51-55]，技术上可以采用归因分析方法，识别造成台区线损率升高的异常用户，从而通过确定台区下属用户用电量时间序列与台区损失电量时间序列之间的因果关系，识别台区下属窃电异常用户。

6　结语

用户窃电直接造成供电企业净收益流失，依托营销与电能计量数据进行数据驱动的窃电检测，是提高供电企业经营收益的重要途径。本文首先介绍了各种窃电手法及对应的外在表征；然后，分别分析了现场工程应用和理论研究中常用窃电检测方法的基本思路及存在的缺陷。在此基础上，结合供电企业工程应用需求分析指出低FPR是推动数据驱动窃电检测方法走向工程实用的关键。最后，从算法设计、状态空间细分和特征指标项的设计选择等方面展望了实现低FPR窃电检测的研究方向。

本文抛砖引玉，讨论了几种可能推进低FPR窃电检测的思路。实际上，窃电的表现形态随着用户侧电源、负荷以及用电模式而变化，电能替代战略的推进和分布式电源、储能系统的普及应用，都将对窃电检测技术不断提出新的要求。而某些用户如低压居民用户中还存在相当数量的零电量用户，很难通过提高计量数据采集频率提供窃电检测的有效信息，需要集思广益，多视角地研究提出满足工程应用需求的实用方法。

本文在撰写过程中得到国家自然科学基金委员会-国家电网公司智能电网联合基金项目(U19266207)资助，特此感谢！

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

参 考 文 献

[1] DOS-ANGELOS E W, SAAVEDRA O R, CORTES 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.
[2] XIA X, XIAO Y, LIANG W. ABSI: an adaptive binary splitting algorithm for malicious meter inspection in smart grid [J]. IEEE Transactions on Information Forensics and Security, 2019, 14(2): 445-458.
[3] 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 Single Processing, 2018, 12(4): 790-805.
[4] 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):78-86.
JIN Sheng, SU Sheng, CAO Yijia, et al. Electricity-theft detection for high-loss distribution area based on Granger causality analysis [J]. Automation of Electric Power Systems, 2020, 44(23): 78-86.
[5] 窦健,刘宣,卢继哲,等.基于用电信息采集大数据的防窃电方法研究[J].电测与仪表,2018,55(21):43-49.
DOU Jian, LIU Xuan, LU Jizhe, et al. Research on electricity anti-stealing method based on power consumption information acquisition and big data [J]. Electrical Measurement & Instrumentation, 2018, 55(21): 43-49.
[6] 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):189-198.
CHEN Qixin, ZHENG Kedi, KANG Chongqing, et al. Detection method of abnormal electricity consumption: comment and prospect[J]. Automation of Electric Power Systems, 2018, 42(17): 189-198.
[7] JIANG R, LU R, WANG Y, et al. Energy-theft detection issues for advanced metering infrastructure in smart grid [J]. Tsinghua Science and Technology, 2014, 19(2): 105-120.
[8] 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.
[9] 刘亮,苏盛,钱斌,等.计量自动化系统卫星时间同步攻击危害与防护[J].南方电网技术,2020,14(1):3-9.
LIU Liang, SU Sheng, QIAN Bin, et al. Impact and protection of satellite time synchronization attacks in advanced metering infrastructure[J]. Southern Power System Technology, 2020, 14(1): 3-9.
[10] 方兆本,杨培洁,彭甘霖,等.电力企业客户信用风险管理实证研究[J].电力系统自动化,2005,29(1):61-64.
FANG Zhaoben, YANG Peijie, PENG Ganlin, et al. Empirical research on credit risk management for electric power clients [J]. Automation of Electric Power Systems, 2005, 29(1): 61-64.
[11] 韩谷静,殷小贡,秦亮,等.电能计量设备防电流法窃电新技术[J].电测与仪表,2004,44(10):29-32.
HAN Gujing, YIN Xiaogong, QIN Liang, et al. A novel technique of preventing electricity-stealing in current method for electric power measuring equipment[J]. Electrical Measurement & Instrumentation, 2004, 44(10): 29-32.
[12] 李大勇,王瑜,黎灿兵,等.基于无线射频技术的防窃电开箱记录仪设计[J].电测与仪表,2008,45(10):51-55.
LI Dayong, WANG Yu, LI Canbing, et al. A design of an box-opening recorder for anti-power-stealing based on RFID [J]. Electrical Measurement & Instrumentation, 2008, 45(10): 51-55.
[13] 周末,朱瑞德,王金全.基于GSM网络的防窃电实时监控系统方案探讨[J].电力自动化设备,2004,24(2):64-69.
ZHOU Wei, ZHU Ruide, WANG Jinquan. GSM-based monitoring and control system against electricity stealing [J]. Electric Power Automation Equipment, 2004, 24(2): 64-69.
[14] 张铁峰,顾明迪.电力用户负荷模式提取技术及应用综述[J].电网技术,2016,40(3):804-811.
ZHANG Tiefeng, GU Mingdi. Overview of electricity customer load pattern extraction technology and its application [J]. Power System Technology, 2016, 40(3): 804-811.
[15] 栾文鹏,余贻鑫,王兵.AMI数据分析方法[J].中国电机工程学报,2015,35(1):29-36.
LUAN Wenpeng, YU Yixin, WANG Bing. AMI data analytics [J]. Proceedings of the CSEE, 2015, 35(1): 29-36.
[16] 刘道新,胡航海,张健,等.大数据全生命周期中关键问题研究及应用[J].中国电机工程学报,2015,35(1):23-28.
LIU Daoxin, HU Hanghai, ZHANG Jian, et al. Research on key issues of big data lifecycle and its applications [J]. Proceedings of the CSEE, 2015, 35(1): 23-28.
[17] 黄彦浩,于之虹,谢昶,等.电力大数据技术与电力系统仿真计算结合问题研究[J].中国电机工程学报,2015,35(1):13-22.
HUANG Yanhao, YU Zhihong, XIE Chang, et al. Study on the application of electric power big data technology in power system simulation[J]. Proceedings of the CSEE, 2015, 35(1): 13-22.
[18] 程超,张汉敬,景志敏,等.基于离群点算法和用电信息采集系统的反窃电研究[J].电力系统保护与控制,2015,43(17):69-74.
CHENG Chao, ZHANG Hanjing, JING Zhimin, et al. Study on the anti-electricity stealing based on outlier algorithm and the electricity information acquisition system [J]. Power System Protection and Control, 2015, 43(17): 69-74.
[19] JUNIOR L A P, RAMOS C O, RODRIGUES D, et al. Unsupervised non-technical losses identification through optimum-path forest [J]. Electric Power Systems Research, 2016, 140: 413-423.
[20] 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 Chijie, ZHANG Bin, HU Jun, et al. Anomaly detection for power consumption patterns based on unsupervised learning [J]. Proceedings of the CSEE, 2016, 36(2): 379-387.
[21] 孙毅,李世豪,崔灿,等.基于高斯核函数改进的电力用户用电数据离群点检测方法[J].电网技术,2018,42(5):1595-1604.
SUN Yi, LI Shihao, CUI Can, et al. Improved outlier detection method of power consumer data based on Gaussian kernel function [J]. Power System Technology, 2018, 42(5): 1595-1604.
[22] 王桂兰,周国亮,赵洪山,等.大规模用电数据流的快速聚类和异常检测技术[J].电力系统自动化,2016,40(24):27-33.
WANG Guilan, ZHOU Guoliang, ZHAO Hongshan, et al. Fast clustering and anomaly detection techniques for large-scale power data streams[J]. Automation of Electric Power Systems, 2016, 40(24): 27-33.
[23] 田力,向敏.基于密度聚类技术的电力系统用电量异常分析算法[J].电力系统自动化,2017,41(5):64-70.

TIAN Li, XIANG Min. Analysis algorithm of power consumption anomaly of power system based on density clustering technology [J]. Automation of Electric Power Systems, 2017, 41(5): 64-70.

[24] 郝然，艾芊，肖斐.基于多源大数据平台的用电行为分析构架研究[J].电力自动化设备，2017，37(8)：20-27.
HAO Ran, AI Qian, XIAO Fei. Research on the analysis framework of electricity behavior based on multi-source big data platform[J]. Electric Power Automation Equipment, 2017, 37(8): 20-27.

[25] 罗滇生，杜乾，别少勇，等.基于负荷分解的居民差异化用电行为特性分析[J].电力系统保护与控制，2016，44(21)：29-33.
LUO Diansheng, DU Qian, BIE Shaoyong, et al. Analysis of the characteristics of differentiated electricity consumption behavior of residents based on load decomposition [J]. Power System Protection and Control, 2016, 44(21): 29-33.

[26] 何永秀，王冰，熊威，等.基于模糊综合评价的居民智能用电行为分析与互动机制设计[J].电网技术，2012，36(10)：247-252.
HE Yongxiu, WANG Bing, XIONG Wei, et al. Analysis and interactive mechanism of intelligent electricity consumption behavior of residents based on fuzzy comprehensive evaluation [J]. Power System Technology, 2012, 36(10): 247-252.

[27] 王成亮，郑海燕.基于模糊聚类的电力客户用电行为模式画像[J].电测与仪表，2018，55(18)：77-81.
WANG Chengliang, ZHENG Haiyan. A portrait of electricity consumption behavior mode of power users based on fuzzy clustering [J]. Electrical Measurement & Instrumentation, 2018, 55(18): 77-81.

[28] MONEDERO I, BISCARRI F, LEON C, et al. Detection of frauds and other non-technical losses in a power utility using pearson coefficient, Bayesian networks and decision trees [J]. Journal of Electrical Power and Energy Systems, 2012, 34: 90-98.

[29] NAGI J, YAP K S, TIONG S K, et al. Nontechnical loss detection for metered customers in power utility using support vector machines [J]. IEEE Transactions on Power Delivery, 2010, 25(2): 1162-1171.

[30] 许刚，谈元鹏，戴腾辉.稀疏随机森林下的用电侧异常行为模式检测[J].电网技术，2017，41(6)：1964-1971.
XU Gang, TAN Yuanpeng, DAI Tenghui. Sparse random forest based abnormal behavior pattern detection of electric power user side[J]. Power System Technology, 2017, 41(6): 1964-1971.

[31] 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. Abnormal data review and classification of power user data [J]. Power and Energy, 2016, 37(1): 17-22.

[32] NIZAR A H, DONG Z Y, WANG Y. Power utility nontechnical loss analysis with extreme learning machine method [J]. IEEE Transactions on Power Systems, 2008, 23(3): 946-955.

[33] PEREIRA L A M, AFONSO L C S, PAPA J P, et al. Multilayer perceptron neural networks training through charged system search and its application for non-technical losses detection [C]// 2013 IEEE PES Conference on Innovative Smart Grid Technologise, February 24-27, 2013, Washington, USA.

[34] ZHENG Z, YANG Y, NIU X, 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.

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

[36] 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): 2959-2996.

[37] LEITE J B, SANCHES M J R. Detecting and locating non-technical losses in modern distribution networks [J]. IEEE Transactions on Smart Grid, 2018, 9(12): 1023-1032.

[38] EACA N, COELHO J. Probabilistic methodology for technical and non-technical losses estimation in distribution system [J]. Electric Power Systems Research, 2013, 97(1): 93-99.

[39] HE Y, MENDIS G J, WEI J. Real-time detection of false data injection attacks in smart grid: a deep learning-based intelligent mechanism [J]. IEEE Transactions on Smart Grid, 2017, 8(5): 5205-2516.

[40] LO C, ANSARI N. CONSUMER: a novel hybrid intrusion detection system for distribution networks in smart grid [J]. IEEE Transactions on Emerging Topics in Computing, 2013, 1(1): 33-44.

[41] KHOO B, CHENG Y. Using RFID for anti-theft in a Chinese electrical supply company: a cost-benefit analysis [C]// IEEE Wireless Telecommunications Symposium, April 13-15, 2011, New York, USA.

[42] DAVIS J, GOADRICH M. The relationship between precision-recall and ROC curves [C]// 23rd International Conference on Machine Learning, June, 2006, New York, USA.

[43] FAWCETT T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861-874.

[44] 邓小龙，王柏，吴斌，等.遗传演化SPA流失预测算法及并行化[J].计算机科学与探索，2011，5(5)：433-445.
DENG Xiaolong, WANG Bai, WU Bin, et al. Genetic evolution based parallelized SPA churn prediction algorithm[J]. Journal of Frontiers of Computer Science and Technology, 2011, 5(5): 433-445.

[45] 李航.统计学习方法[M].北京：清华大学出版社，2012.
LI Hang. Statistical learning method [M]. Beijing: Tsinghua Press, 2012.

[46] 勾婷婷.BP神经网络和Logistic回归在信用评级上的应用与模型对比[D].重庆：重庆大学，2012.
GOU Tingting. The application and model of BP neural network and logistic regression on credit rating are compared with the model[D]. Chongqing: Chongqing University, 2012.

[47] 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.

[48] 张承智，肖先勇，郑子萱.基于实值深度置信网络的用户侧窃电行为检测[J].电网技术，2019，43(3)：1083-1091.

ZHANG Chengzhi, XIAO Xianyong, ZHENG Zixuan. Electricity theft detection for customers in power utility based on real-valued deep belief network[J]. Power System Technology, 2019, 43(3): 1083-1091.

[49] 董明刚,姜振龙,敬超.基于海林格距离和SMOTE的多类不平衡学习算法[J].计算机科学,2020,47(1):102-109.
DONG Minggang, JIANG Zhenlong, JING Chao. Multi-class imbalanced learning algorithm based on hellinger distance and SMOTE algorithm [J]. Computer Science, 2020, 47 (1): 102-109.

[50] 刘洋,高丽霞,刘璐.考虑样本不平衡的并行化用户负荷类型辨识方法[J].电网技术,2020,44(11):4310-4317.
LIU Yang, GAO Lixia, LIU Lu. Parallel load type identification algorithm considering sample class imbalance[J]. Power System Technology, 2020, 44(11): 4310-4317.

[51] 赵磊,栾文鹏,王倩.应用AMI数据的低压配电网精确线损分析[J].电网技术,2015,39(11):3189-3194.
ZHAO Lei, LUAN Wenpeng, WANG Qian. Accurate line loss analysis of distribution network using AMI data[J]. Power System Technology, 2015, 39(11): 3189-3194.

[52] 李亚,刘丽平,李柏青,等.基于改进*K*-means聚类和BP神经网络的台区线损率计算方法[J].中国电机工程学报,2016,36(17):4543-4551.
LI Ya, LIU Liping, LI Baiqing, et al. Calculation of line loss rate in transformer district based on improved *K*-means clustering algorithm and BP neural network[J]. Proceedings of the CSEE, 2016, 36(17): 4543-4551.

[53] 李昀昊,王建学,王秀丽.基于混合聚类分析的电力系统网损评估方法[J].电力系统自动化,2016,40(10):60-65.
LI Yunhao, WANG Jianxue, WANG Xiuli. The method of network loss assessment of power system based on mixed cluster analysis [J]. Automation of Electric Power Systems, 2016, 40(10): 60-65.

[54] 王奇,庄远灿,阎帅,等.基于随机矩阵理论的交直流输电通道线损大数据关联特性分析[J].电力自动化设备,2018,38(5):70-76.
WANG Qi, ZHUANG Yuancan, YAN Shuai, et al. Relevance characteristic analysis of line loss big data in AC and DC transmission channels based on random matrix theory [J]. Electric Power Automation Equipment, 2018, 38(5): 70-76.

[55] 刘健,段璟靓.配电网极限线损分析及降损措施优化[J].电力系统保护与控制,2013,41(12):27-35.
LIU Jian, DUAN Jingjing. Line loss limitation analysis and optimal planning of loss reduction for distribution grids [J]. Power System Protection and Control, 2013, 41(12): 27-35.

金　晟(1996—),男,硕士研究生,主要研究方向:电力系统反窃电技术。E-mail:1467445457@qq.com

苏　盛(1975—),男,通信作者,博士,教授,博士生导师,主要研究方向:电力系统网络安全防护与电力系统大数据技术应用。E-mail:eessheng@163.com

薛　阳(1986—),男,高级工程师,主要研究方向:电力系统反窃电。E-mail:xueyang3@epri.sgcc.com.cn

(编辑　王梦岩)

Review on Data-driven Based Electricity Theft Detection Method and Research Prospect for Low False Positive Rate

JIN Sheng[1], *SU Sheng*[1], *XUE Yang*[2], *YANG Yining*[2], *LIU Sha*[2], *CAO Yijia*[1]

(1. Hunan Key Laboratory of Smart Grid Operation and Control (Changsha University of Science and Technology), Changsha 410014, China; 2. China Electric Power Research Institute Co., Ltd., Beijing 100192, China)

Abstract: Electricity theft in the power distribution system is the main cause of non-technical loss of power grids, and it is a chronic problem in operation and management of power utilities. The electricity information acquisition system collects massive user data, which makes it possible to carry out data-driven abnormal electricity detection and accurately pinpoint electricity theft consumers. Affected by the diversity of electricity consumption behaviors of users, the false positive rate of data-driven based electricity theft detection method is still difficult to meet the practical demands in some scenarios, which seriously restricts the engineering application of this method. Firstly, this paper describes the implementation measures of electricity theft, and then sorts out the basic ideas and limitations of the electricity theft detection methods applied in engineering practice and the data-driven based electricity theft detection methods. On this basis, combining with different requirements of engineering application for the evaluation index of electricity theft detection, it is pointed out that the lack of useful extracted information, the low sensitivity and reliability of the characteristic index items are the major reasons that hinder the data-driven based electricity theft detection methods from being practical in engineering. Finally, the electricity theft detection with low false positive rate is prospected from different levels such as algorithm design, state space subdivision, and design and selection of characteristic index items.

This work is supported by National Natural Science Foundation of China (No. 51777015), State Grid Corporation of China and Hunan Provincial Natural Science Foundation of China (No. 2020JJ4611).

Key words: electricity theft detection; low false positive rate; data-driven; feature engineering; state space