基于 FP-growth 算法的用电异常数据挖掘方法*

段晓萌[1],王 爽[1],赵 婷[1],丁徐楠[2]
(1.中国电力科学研究院有限公司,北京 100192;2.国网浙江省电力有限公司,浙江 杭州 310007)

摘 要:随着科学技术的不断进步,不法分子窃电手段日趋专业化多样化,而传统的防窃电技术实时性及可行性较低。研究对运行中智能电能表用电信息的数据采集及特征提取,分析异常用电数据,应用机器学习的方法对特征值进行学习,并推导出用电异常的判断阈值,采用关联规则数据挖掘方法对独立检测的结果进行融合,从而实现窃电数据的挖掘。最后验证了模型建立的准确性,并推导出用电异常案例的甄别方法。

关键词:电能表;用电异常;FP-growth 算法;数据挖掘
中图分类号:TN915;TM933 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.200073

中文引用格式:段晓萌,王爽,赵婷,等. 基于 FP-growth 算法的用电异常数据挖掘方法[J].电子技术应用,2020,46(10):47-50.
英文引用格式:Duan Xiaomeng,Wang Shuang,Zhao Ting,et al. Data mining method on abnormal electricity usage based on FP-growth algorithm[J]. Application of Electronic Technique,2020,46(10):47-50.

Data mining method on abnormal electricity usage based on FP-growth algorithm

Duan Xiaomeng[1],Wang Shuang[1],Zhao Ting[1],Ding Xunan[2]
(1.China Electric Power Research Institute,Beijing 100192,China;
2.State Grid Zhejiang Electric Power Co.,Ltd.,Hangzhou 310007,China)

Abstract: Because of the technology development, the means for stealing electricity becomes more specialized and diversified. The traditional anti-theft technology is less real-time and less feasible. This paper studied the intelligent diagnosis and characteristics extract method of electricity energy meter during online operation, analyzed the abnormal electricity consumption data, used machine learning abnormality judgment thresholds based on features, and used association rule data mining methods to fuse independent detection results, realizing the mining of power theft data. At last, this paper verified the accuracy of the model establishment, and deduced the screening method of power consumption abnormal cases.

Key words: energy meter; abnormal electricity usage; FP-growth algorithm; data mining

0 引言

电能表电能计量的准确性是电网公司与电力用户之间贸易结算及电网公司利润实现的最终环节,不法行为会严重伤害贸易关系的公平、公正、公开性,因此查处用电异常行为是电网公司一直以来的工作重点。随着电网公司对反窃电工作重视程度的增加,不法分子的手段也逐步变得隐蔽化与智能化[1]。近年来,随着用电信息采集系统的不断完善,已经能够按照业务需求广泛采集到电能表的大量数据,从大量无序数据中应用单一准则判断用电异常,容易产生误判情况,如由于环境或振动而引发的开表盖事件[2]。如何从大量的用电异常数据中提高辨别窃电数据的概率,从多组数据关联来推断是否窃电,是本文研究的重点。因此提出一种通过数据关联规则判断在运电能表用电异常行为的数据挖掘方法。

*基金项目:国家电网公司科技项目(5442PD180022)

1 研究现状

用电信息采集系统的推广应用使得研究在运电能表的工作状态成为了可能,近年来有很多研究是基于电能表在线监测[3]的,但是较多的研究是基于电能表是否故障,是否存在电网异常的,且受到通信技术的影响,应用效果没有达到预期。文献[4]运用了孤立森林算法及决策树算法来对异常用电数据进行检测。文献[5]研制了一套计量装置在线监测和智能诊断系统,可运用数据挖掘技术对用户违约用电窃电和计量装置故障进行智能诊断。文献[6]基于密度聚类算法的离群算法,通过提取用户用电数据的特征,应用聚类算法,计算出离散的数据点,实现了数据化检测用电异常行为。可见,相比传统采用单一维度的统计方法,采用数据挖掘的方法能够有效分辨出大数据之间的内在联系。

本文分析了异常用电数据,通过关联学习算法分析研究造成窃电的关联事件。关联规则学习是一种在大型

数据库中发现变量之间的相互联系的数据挖掘方法，通过寻找频繁项寻找频发信息，并挖掘数据项集之间隐藏的关系，找出项集与项集之间的关联关系，并判断是否发生事件的概率。本文尝试使用 FP-growth 算法对异常用电数据进行分析，在建立异常用电数据窃电特征的基础上，利用 FP-growth 算法进行了窃电数据的挖掘和分析，挖掘用电异常信息以及窃电和窃电原因的关联规则，并对挖掘结果展开分析。

2 异常用电数据特征的选取

通过异常用电数据挖掘进行窃电行为识别，其关键在于提取窃电情况下的电气特征信息。目前在用电信息采集系统中，已经对负荷数据、用电量、电表事件等信息进行采集和存储，因此可从中提取特征信息[7-9]。

本文从电能表记录的电气特征和电力用户的用电行为来对不法分子用电异常特征进行建模，具体描述如下：

(1)电能表倒走、停走。电能表用电量是随着电气使用而增加的，当采集的电量出现倒走或停走，需判断是安装时人为接线错误还是窃电行为。

①电能表倒走故障模型：电能表记录的正、反向有功总电能(三相用户可能判断组合有功总电能)小于前一天电能表冻结的数据。

②电能表停走故障模型：在 2 日之内监测到电能表的电量累计为 0，同时，在同一时间段内有 3 次测量出任意一相电流值大于 0.1 A。

(2)电流相关异常

电流相关异常主要考核三相上负载平衡程度，电流相关异常包括电流失流、潮流反向。不法分子通过修改 CT 变比，更改或破坏电流回路接线，导致电能表实际测量的电流值变小，使得三相负载失流或不平衡，达到少计电量的目的。

①电流失流故障模型：首先电能表电压应大于触发下限，默认值为标称值的 70%。对于三相三线的失流，判断是否 A、C 两相中任一相的电流小于 $0.005I_b(I_n)$，另外一相的电流大于等于 $0.05I_b(I_n)$。对于三相四线失流，判断是否有任意一相的电流值小于 $0.005I_b(I_n)$，其他相中至少有一相的电流大于等于 $0.05I_b(I_n)$。

②潮流反向故障模型：三相中任意一相有功功率方向发生改变，同时功率值大于设定值(默认设定值为0.5%单一相基本功率)，且持续事件大于 60 s 时，电能表可以记录潮流反向事件记录。

(3)电压相关异常

通过改变电能表的外部接线，导致电能表能够计量的电压值减小至无法正常工作，简称电压失压。使电压减小导致三相电压不平衡，中性点发生偏移，产生用电异常行为，简称电压不平衡。使电压增大导致存在损害电能表的可能，或者电压小幅减少导致计量电能变少，简称电压过压或欠压。

①电压失压故障模型

三相三线失压：A、C 两相中任意一相的电压值小于临界电压值，默认临界电压值为 78%标称电压。另外一相的电压值大于等于标称电压。

三相四线失压：三相中任意一相或两相的电压值小于临界电压值。另外有一相的电压值大于等于标称电压。

三相三线：$U_a<K\times U_n$，$U_c\geqslant K\times U_n$ 或 $U_c<K\times U_n$，$U_a\geqslant K\times U_n$

三相四线：$U_a<K\times U_n$，$U_c|U_b\geqslant K\times U_n$ 或 $U_b<K\times U_n$，$U_c|U_a\geqslant K\times U_n$ 或 $U_c<K\times U_n$，$U_a|U_b\geqslant K\times U_n$($K$ 建议值为 78%)。

考虑到电网的电压波动，需要监测出多次电压失压异常才可生成异常事件，一般选择每天至少 3 次，持续至少 3 天才会生成异常事件。

②电压过压或欠压故障模型

任意一相电压大于临界电压值，默认临界电压值为120%标称电压；或者任意一相电压大于 78%标称电压，且小于 90%标称电压。考虑到电网的电压波动，需要监测出多次电压过压或欠压异常才可生成异常事件，一般选择每天至少 3 次，持续至少 3 天才会生成异常事件。

③电压不平衡故障模型

在非电压断相情况下，即任意一相电压均大于 0 V，以三相电压的最大值与最小值之间的差值与最大值之间比值来代表电压不平衡率，当电压不平衡率大于 30%时判断为电压不平衡。即，$\mathrm{Max}(U_a,U_b,U_c)-\mathrm{Min}(U_a,U_b,U_c)/\mathrm{Max}(U_a,U_b,U_c)>30\%$；考虑到电网的电压波动，需要监测出多次电压不平衡异常才可生成异常事件，一般选择每天至少 10 次，持续至少 3 天才会生成异常事件。

(4)异常用电相关异常

不法分子通过打开或破坏计量装置铅封即外壳，更改电源回路接线，实现窃电的目的，简称开盖记录异常。通过外接强磁场，影响电能表计量性能或电能表重要芯片工作，从而影响计量准确性。

①开盖记录故障模型：电能表开盖的主要原因为：人为原因开启表盖，因表盖未紧固等其他问题导致误报。

读取电能表状态字，判断电能表开盖时间与电能表安装于时间阈值。排除正常电能表开盖，例如初次工单的情况排除开盖时间逻辑错误的情况。

②恒定磁场故障模型：三相电能表会记录大于特定值(100 mT)的恒定磁场干扰事件，通过超读电能表恒定磁场干扰事件记录进行判断。

3 关联规则算法的案例研究

3.1 数据的选择

本文共选取数据1 726 条，其中 126 条数据为存在窃电的数据，8 768 条数据为不存在窃电及不确定是否存在窃电数据。数据中存在众多问题及无关项，过滤某些空值、无效值及检定人员等无关项，凭借异常用电数据窃电特征选取依据，选出可能存在窃电的电能表检定项目：开盖记录、恒定磁场事件、电表倒走、电表停走、电

压过压或欠压、电压不平衡、电流失流、电压失压、潮流反向，且事件发生判定为1，未发生判定为0。经审核，原始数据中对这9个检测项目的记录完整，且只存在少量的无效值及空值，基于完整度的前提下，将这9项试验的结果作为评价电能表是否窃电的指标。因需挖掘的对象为是否窃电，所以理论上抽取的数据记录为窃电的记录。

如表1所示，抽取9条数据进行数据挖掘研究，M_1~M_9分别表示：开盖记录异常、恒定磁场事件、电表倒走、电表停走、电压过压或欠压、电压不平衡、电流失流、电压失压、潮流反向，R表示窃电。

表1 用电异常相关事件数据

M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	R
1	0	1	0	0	0	0	1	1	1
1	0	0	1	0	0	1	1	0	0
1	0	0	1	0	1	0	1	0	1
1	0	1	0	1	0	1	0	0	1
1	0	1	0	0	1	0	1	0	1
1	1	0	1	0	0	1	1	0	1
1	1	0	1	0	0	1	1	0	1
0	1	0	0	0	1	0	1	0	0
0	0	1	0	1	0	1	0	0	1
0	0	0	0	0	1	1	0	0	0

3.2 FP-growth算法应用

FP-growth算法的基本思想是将待分析数据集中的事务映射至FP-tree，并根据FP-tree寻找频繁项集。无论数据集大小，FP-tree的构建过程中仅需要对数据集进行两次扫描，FP-tree的算法步骤主要包含：FP-tree的构建和FP-tree上频繁项集的挖掘。

(1)FP-growth算法首先扫描数据集，根据事件相关数据按其支持度计数降序排列出1-项集，如表2所示。

表2 项集及其支持度计数

项集	M_1	M_8	M_7	M_3	M_4	M_6	M_2	M_5	M_9
支持度计数	7	7	6	4	4	4	3	2	1

(2)根据表2中的项集及其支持度计数，定义最小支持度，重新将事件相关数据按项集频繁程度进行降序排序，并删去小于最小支持度的项目，数据排序后如表3所示。

(3)再次扫描，将事务数据库DB中排序和删除后的事务进行创建频繁项头表(从上往下降序)并构建FP树。FP树是一种树结构，由标记为空(NULL)的根节点和项目前缀子树组成。项目前缀子树中的每一个节点由两个域组成：项目名和支持计数。其中，项目名表示节点代表哪个项目，支持度计数表示到目前节点为止路径中的事务数。按此构建如图1所示FP树。

表3 事务数据库

事务编号	事务	排序和删除后的事务
1	M_1, M_3, M_8, M_9	M_1, M_8, M_3
2	M_1, M_4, M_7, M_8	M_1, M_8, M_7, M_4
3	M_1, M_4, M_6, M_8	M_1, M_8, M_4, M_6
4	M_1, M_3, M_5, M_7	M_1, M_7, M_3
5	M_1, M_3, M_6, M_8	M_1, M_8, M_3, M_6
6	M_1, M_2, M_4, M_7, M_8	M_1, M_8, M_7, M_4, M_2
7	M_1, M_2, M_4, M_7, M_8	M_1, M_8, M_7, M_4, M_2
8	M_2, M_6, M_8	M_8, M_6, M_2
9	M_3, M_5, M_7	M_7, M_3
10	M_6, M_7	M_7, M_6

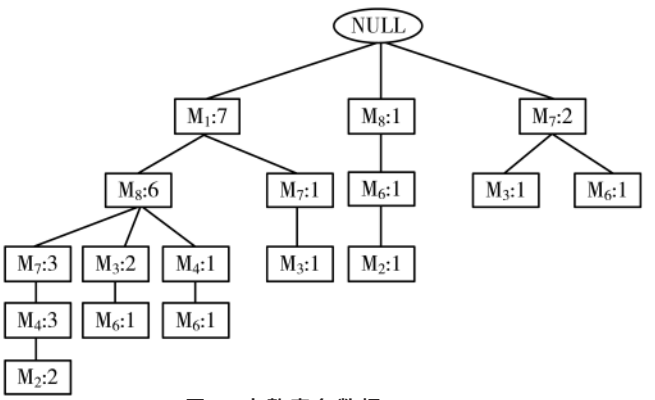

图1 完整事务数据FP-tree

据此FP树，可充分展现关联规则，依靠查找项目前缀子树与支持度计数，获取强关联结果。

关联规则生成有效规则需同时满足规则最小支持度和最小置信度阈值要求，规则$A \Rightarrow B$的支持度为事务T中包含$A \cup B$的比例，而$A \Rightarrow B$的置信度为事务T中包含A也同时包含B的比例，即：

$$\text{Support}(A \Rightarrow B)=P(A\&B)/T \quad (1)$$

$$\text{Confidence}(A \Rightarrow B)=P(A\&B)/P(A) \quad (2)$$

式中，A、B均为包含于T的项集，且$A \cap B=\Phi$。

其中支持度指的是9项故障中若干项不通过导致电能表窃电的概率，而置信度表现为若干项故障导致电表窃电结论的可能性大小。当最小支持度为30%，最小置信度为70%时，生成的规则如表4所示。

表4 关联规则与挖掘结果

关联规则	支持度/%	置信度/%
$M_1 \Rightarrow R_1$	60	85.7
$\{M_1, M_3\} \Rightarrow R_1$	30	100
$\{M_1, M_7\} \Rightarrow R_1$	30	75
$\{M_1, M_8\} \Rightarrow R_1$	50	83.3
$\{M_4, M_8\} \Rightarrow R_1$	30	75
$\{M_1, M_4, M_8\} \Rightarrow R_1$	30	75

表4中数据满足最小支持度及最小置信度阈值，可以输出为有效关联规则。

表 5 电能表窃电检测数据挖掘结果

关联规则		支持度/%	置信度/%	实例
关联项	关联结果			
开盖记录=1	窃电=1	49.21	100	62
电表停走=1	窃电=1	45.24	100	57
电压失压=1	窃电=1	22.22	100	28
电压不平衡=1、电表停走=1	窃电=1	21.43	100	27
开盖记录=1、电压失压=1、电流失流=1	窃电=1	16.67	100	21
电表倒走=1	窃电=1	7.14	100	9
恒定磁场事件=1	窃电=1	5.56	100	7
反向潮流=1	窃电=1	3.17	100	4

4 结果与分析

据关联规则算法，对 126 条窃电数据进行处理，并对很多实例数与支持度相同且冗余的规则进行整合、优化，综合对比不同阈值下多次处理结果的差异，最终设定最小支持度为 15%，最小置信度为 85%，结果如表 5 所示。

结果分析：(1)结果显示存在开盖事件而存在窃电结果的支持度最高，占 49.21%，其次是电表停走，占 45.24%，电能表存在窃电事件由这两项引起的案列最多，可以看出当电表停走与开盖记录异常同时出现时，可以很大程度上判断出存在用电异常行为。在所有参数里，失压与电压不平衡对于能否确定电能表存在用电异常行为的影响较小；(2)电表倒走(7.14%)、恒定磁场事件(5.56%)、反向潮流(3.17%)等项目结果显示支持度不高，表明由发生电表倒走、恒定磁场事件及反向潮流事件而产生窃电可能的情况很少；(3)在产生窃电结果的事件中，常常伴随着多项事件的发生，其结果更具有可信度；(4)如果检测到故障信息为开盖、电量倒走、电量停走、电流失流、电压失压，可以很大程度上判断出可能存在用电异常行为，应加紧实施现场勘察，确诊案例，并纳入数据库。

5 结论

本文通过调取筛选的用户异常用电数据，通过数据挖掘 FP-growth 算法构建 FP-tree，并运用关联规则分析用户异常用电中的窃电数据，分析产生窃电的可能性事件，提高对用电异常事件的判断。

通过对用电异常分析研究，大部分用电异常事件的产生伴随着电能表开盖记录异常事件，同时当存在电能表停走、电压失压、异常时，发生用电异常行为的可能性也比较大。电压过压或欠压以及电压不平衡独立出现时，发生用电异常行为的可能性较小。以上结论可以看出，基于 FP-growth 算法构建进行的数据挖掘结果符合实际应用时对于电能表用电异常的判断方法，并且验证了用电异常模型构建的准确性。

参考文献

[1] 潘明明，田世民，吴博，等.基于智能电能表数据的台区识别与窃电检测方法研究[J].智慧电力，2017，45(12)：80-84.

[2] 张晶，刘晓巍，张松涛.基于营销大数据的用电异常事件统计及窃电特征分析[J].供用电，2018(6)：77-82.

[3] 程瑛颖，杨华潇，肖冀，等.电能计量装置运行误差分析及状态评价方法研究[J].电工电能新技术，2014，33(5)：76-80.

[4] 张荣昌.基于数据挖掘的用电数据异常的分析与研究[D].北京：北京交通大学，2017.

[5] 肖坚红，严小文，周永真，等.基于数据挖掘的计量装置在线监测与智能诊断系统的设计与实现[J].电测与仪表，2014，51(14)：62-67.

[6] 蔡耀年，王明琪，刘建森，等.一种基于离群算法的窃电行为检测的研究[J].计算技术与自动化，2018，37(2)：73-77.

[7] 李端超，王松，黄太贵，等.基于大数据平台的电网线损与窃电预警分析关键技术[J].电力系统保护与控制，2018，46(5)：143-151.

[8] 于小青，齐林海.基于流数据聚类算法的电力大数据异常检测[J].电力信息与通信技术，2020，18(3)：8-14.

[9] 徐育涛.基于用电信息采集大数据的防窃电方法探讨[J].通讯世界，2019，26(12)：237-238.

(收稿日期：2020-02-07)

作者简介：

段晓萌(1989-)，通信作者，男，硕士，工程师，主要研究方向：电能计量新技术，E-mail：dxm_89@163.com。

(上接第 46 页)

Advances in Neural Information Processing Systems，Vancouver，2019：8024-8035.

(收稿日期：2020-05-06)

作者简介：

刘欣(1985-)，男，硕士，助理研究员，国家安全评价师，主要研究方向：计算机技术、煤矿自动化及安全。

李卫龙(1985-)，男，硕士，助理研究员，主要研究方向：煤矿机电与自动化。

张灿明(1984-)，通信作者，男，硕士，助理研究员，主要研究方向：深度学习、煤矿自动化及安全，E-mail：zhangcm0103@126.com。

版权声明