文章编号:1671-4598(2018)06-0223-04 DOI:10.16526/j.cnki.11—4762/tp.2018.06.057 中图分类号:TP274 文献标识码:A

基于样本优化选取的支持向量机窃电辨识方法

卢 峰[1],丁学峰[1],尹小明[1],陈洪涛[2],王 颖[2]
(1.国网浙江长兴县供电有限公司,浙江 湖州 313100;2.中国计量大学 机电工程学院,杭州 310018)

摘要:目前窃电行为普遍存在,如何提高用户用电系统的窃电辨识能力是电力公司一直关注的热点问题;随着智能电表在各地区的普及,数据挖掘等大数据分析技术在用电数据处理上的应用越来越受到重视;针对电力公司亟待解决的反窃电问题,在研究支持向量机原理和分析用电数据特性的基础上,将One-class SVM算法引入到疑似窃电判断当中,提出了一种将电量波动特征和One-class SVM结合的窃电辨识模型;首先提出改进的电量数据波动系数来表征电量波动,然后设计了基于One-class SVM窃电辨识方案;提出一种以电量波动系数作为指标选取训练样本的方法,训练得到相应分类模型,通过该模型分析用电数据从而辨别出是否存在窃电行为;算法验证结果表明该方法能提高窃电辨识的准确率和效率,具备一定的实用性。

关键词:窃电行为;一类支持向量机;电量波动;反窃电

Electricity Theft Identification Using Support Vector Machine Based on Sample Optimization Selection

Lu Feng[1], Ding Xuefeng[1], Yi Xiaoming[1], Chen Hongtao[2], Wang Ying[2]
(1. State Grid Zhejiang Changxing Power Supply Company Limited, Huzhou 313100, China;
2. College of Mechanical and Electrical Engineering, China Jiliang University, Hangzhou 310018, China)

Abstract: Nowadays, Energy theft is widespread, how to improve the electricity theft identification of the user's power system has been concerned about the hot issues by the power company. With the popularity of smart meters in all regions, data mining and other large data analysis technology in the application of electricity data processing has been receiving increasing attention. Focused on the problem of anti-stealing electricity which power companies are concerned, and based on the study of the principle of support vector machine and the analysis of the characteristics of electricity data, the One-class SVM algorithm is introduced into the judgment of suspected energy theft, and an electricity theft identification model combining power fluctuation feature and One-class SVM is proposed. This paper first proposes an improved power data fluctuation coefficient to characterize the fluctuation of electricity, and then designs an electricity theft identification scheme based on One-class SVM. This method combines the fluctuation characteristics of electricity to select the load data samples, and constructs the detection model of energy theft based on the electricity data, then identifies whether there is electricity theft behavior. Experiments show that this method can improve the accuracy and efficiency of electricity theft identification and it has certain practicability.

Keywords: energy theft; one-class SVM; fluctuation of electricity; anti-electricity stealing

0 引言

如今,窃电现象时常发生,导致用电台区线损率一直偏高,已经严重损坏了电力公司的利益,扰乱了供用电秩序,影响了国家的经济建设和社会稳定[1]。近几年,电力公司开始高度关注窃电问题,并进行了不少反窃电相关的工作。国家电网公司在2018年度总部科技项目申报指南中就包含了反窃电及稽查监控相关技术研究的项目,说明现在反窃电问题仍然是电力公司亟待解决的关键问题,研究反窃电技术具有很好的理论意义和实际应用价值。

现在智能电表的迅速普及,带来了大量的用电数据。这些用电数据数量大,种类多且复杂,其中蕴藏着巨大的研究价值,对于用户窃电分析很有帮助[2]。随着大数据时代的到来,数据挖掘等数据处理技术也开始应用于用电数据处理。文献[3]应用了BP神经网络建立用户窃电嫌疑分析模型,具有一定的窃电嫌疑分析能力。但该模型需要大量的正常数据和窃电数据作为训练数据,而且实际分析过程中往往由于样本不平衡的问题,模型分类效果不理想。文献[4]提出了一种基于树形结构的电能表层次模型的电能表管理系统架构。其采用的用电数据偏少,当数据增多时需要对层次模型进一步优化。文献[5]应用One-class SVM算法进行了用电异常检测,采用了单一电表的数据,样本数量偏少,验证的有效性不足,且没有提出作为SVM的核心问题之一的训练样本选取方法。

本文将One-class SVM算法引入到疑似窃电判断当中,提出了一种将电量波动特征和One-class SVM结合的窃电辨识模型。利用电量数据波动指标为One-class SVM选取相对优化的训练样本,训练得到相应分类模型。通过该模型对用户用电数据进行分类,将结果进行分析处理从而辨别出是否存在窃电行为。

1 用电用户电量波动分析

目前电量数据处理多只采用平均数或方差等分析指标,但这些单独的指标无法满足对不同时间段电量波动情况进行比较的要求。因此,在对用电数据特征进行深入分析的过程中,总

收稿日期:2017-10-17; **修回日期**:2017-11-06。
基金项目:浙江省自然科学基金青年科学基金项目(LQ17E070003)。
作者简介:卢 峰(1976-),男,浙江长兴人,硕士,高级工程师,主要从事营销管理工作方向的研究。

结得到了描述电量数据波动的指标 CV（电量波动系数），它用于分析统计期间用户电量数据异常波动的程度，模型定义为：

$$CV = \frac{\sigma}{\mu} = \frac{1}{\overline{d}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i - \overline{d})^2} \tag{1}$$

式中，d_i 为用户单日电量，$\overline{d}$ 为日电量平均值，N 为累计天数，σ 为标准差，μ 为均值。CV 是单位均值上的离散程度，CV 越大，反映样本偏离度越大，即电量波动程度越大。

2 One－class SVM 算法原理

支持向量机（support vector machine，SVM）是一种二类分类模型。它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知器[6]。其核心思想是建立一个超平面作为决策面，使得样本数据当中正例和反例的间隔达到最大化。这样求最大间隔的问题其实就等价于求最优分类超平面，其中超平面是对任意维度下线性函数的统称[7]。其原理如图 1 所示。

图 1 中，圆形和三角形代表整个训练样本，圆形代表正例，三角形代表反例，虚线将两种类别区分开，L_1 与 L_2 之间的间隔为数据类间隔，中间的虚线为分离超平面。L_1 和 L_2 上的圆形和三角形即为相应的支持向量。对于上述的训练样本类型可视为线性可分训练数据集，通过支持向量求其间隔最大化或求其等价的凸二次规划问题可以得到使数据类间隔最大的分类超平面。

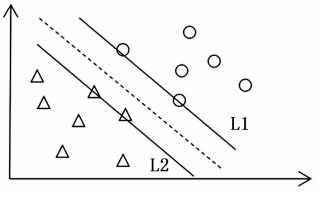

图 1 支持向量机原理

在应用中，支持向量机一般将数据映射到高维空间，使原本线性不可分的样本在高维空间线性可分，通过核函数构造最优超平面完成分类[8]。

SchöLkopf 等人经过对 SVM 算法的研究，开发出了 One－class SVM 算法[9]，它的核心思想是通过 SVM 训练得到具有最大分类间隔的超平面，从而把一分类问题转化成一个特殊的二值分类问题。实际在采用训练数据集进行训练的时候，One－class SVM 只选取一类具有相似特征的数据集合进行训练，得到的模型其基于的分类规则只有一类数据的特性 A，然后在分类的时候模型就将测试数据集分为属于 A 类和不属于 A 类两种类型，公式如下所示。

模型优化函数：

$$\min_{w,\zeta,\rho}\frac{1}{2}\|w\|^2 - \rho + \frac{1}{vl}\sum_{i=1}^{l}\zeta_i \tag{2}$$

决策函数：

$$\text{s.t}\ (w\cdot\varphi(x_i)+b) \geqslant \rho - \zeta_i, \zeta_i \geqslant 0 \tag{3}$$

式中，w 和 ρ 为超平面的法向量和截距，ζ_i 为松弛变量，v 是惩罚参数，φ 为非线性映射，即核函数。

使用 One－class SVM 的关键之一在于选择选择上述模型优化函数中的 v 值和合适的核函数。v 是一个比例值，其范围是 0 到 1 之间。其体现为所选取的训练集数据中规定的异类数据的比例。本文选取高斯核函数，其定义为空间中任一点 x_1 到某中心点 x_2 之间欧氏距离的单调函数[10]。其公式如下：

$$\kappa(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\beta^2}\right) \tag{4}$$

其中：β 为函数的宽度参数，为简化公式令 $\gamma = \frac{1}{2\beta^2}$，通过对 γ 的调整以获得更好的分类结果。

结合用户的实际用电特征来考虑，由于正常用户数据往往远大于窃电用户数据，所以会导致两类数据数量不平衡的情况。因此，解决实际分类过程中的数据类型不平衡的问题非常重要。相比其他类型的 SVM，One－class SVM 能更好地处理此类问题，这是由于上文提到的算法性质所决定的。因此选择 One－class SVM 来训练模型。

3 基于 One－class SVM 窃电辨识方案

3.1 总方案设计

本文设计了一种基于 One－class SVM 的用户异常数据检测模型，其包括训练样本采集、模型选取、数据预处理、参数优化、分类器分类、决策报警等部分。其计算分析是在 MATLAB 环境下进行的，流程如图 2 所示。

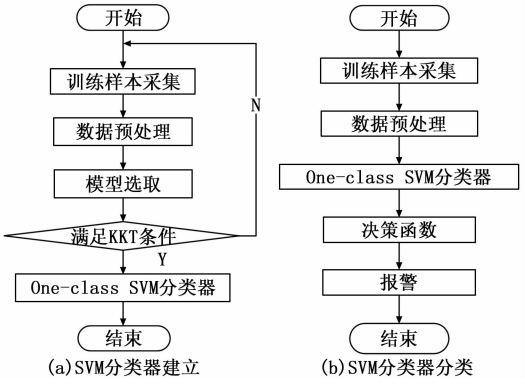

图 2 基于 One－class SVM 的数据异常检测框图

本方案主要分为两个步骤，首先建立 SVM 分类模型，通过训练样本采集和数据预处理得到特征向量，并根据负荷类型选取模型进行训练，寻找最优超平面，当求得的解满足 KKT 条件时，即可得到最优超平面，从而得到 One－class SVM 分类模型；其次采集测试数据并进行预处理，用得到的 One－class SVM 分类器对该数据进行处理，将结果导入决策函数分析，如发现窃电行为则进行报警。

3.2 训练样本选取

本文提出了一种新的支持向量机的样本选取方法，即利用上文提到的电量波动系数作为选取样本的指标。

因为电量根据功率和时间的乘积，而功率是电压和电流的乘积，因此当电压恒定和时间相同的情况下，可以根据电量情况来反映电流情况。而电量波动情况容易分析，所以可通过电量来分析电流，从而得到相应的三相电流数据样本。

根据已有的研究进行分析，规定当某月的电量波动系数 CV 满足 $0<CV<0.2$ 时，则视为该月的用电情况正常，该月的负荷数据也是正常的，该月的负荷数据可以作为训练样本进行训练。规定正常用电数据用标签＋1 表示，异常数据用标签－1 表示，训练样本数据全部用标签＋1 表示。

3.3 模型选取

可按工作日和节假日的负荷加以区分，对于两种负荷类

别，在进行分类前分别选取工作日和节假日的正常负荷数据进行训练，得到相应的分类模型。一般工作日选择 5－10 天的负荷数据作为训练样本，节假日选择 4 天以上的负荷数据作为训练样本。

3.4　数据预处理

为了防止某些偏差过大的值对模型分类的准确性产生不良影响，需要对数据进行预处理。采用线性函数法，即：

$$y(k)=\frac{|x(k)-\min(x(n))|}{|\max(x(n))-\min(x(n))|} \tag{5}$$

其中：$x(k)$ 代表任意一个样本值，$\min(x(n))$ 代表样本最小值，$\max(x(n))$ 代表样本最大值。这种归一化处理方法一般是将 $y(k)$ 转化为介于 0 和 1 之间的数。

3.5　参数优化

对于最优参数的选择，有两种方法，一种是根据实际经验进行选择，模型优化函数中的参数 v 一般取 0.01，0.001，0.000 1，核函数中 γ 一般取 10。这主要是因为在选择训练样本集时尽可能采用正常的用电数据作为训练数据集样本，因此 v 的值也就是异常用电数据占训练样本集的比例会很小。

还有一种方法是利用程序自身寻找最优参数。一般采用的是网格参数寻优。网格参数寻优核心的思想是 k 折交叉验证。即随机选取一部分样数据作为训练数据，其他作为测试数据检验，经过 k 次循环找到最优参数。这种方法的好处在于其随机性和重复性，可以有效地减小误差。

第一种方法得到的参数训练出来的模型准确度高，但是每次都需要人工寻找，比较繁琐。第二种方法虽然免去了人工，但是求得的参数训练出来的模型准确率较低，且随着数据的增大，计算机的运算量也会增大，使得运行时间过长。本文主要采用第一种方法。

3.6　One－class SVM 分类器

One－class SVM 作为一种分类器，是以对训练数据训练得到的模型作为分类规则。其输入可以是多维的数据，但是输出是一维的。规定 One－class SVM 只输出＋1 和－1 两种数据来进行数据的分类。其输出＋1 代表其所对应的负荷数据是正常用电数据，输出－1 代表其所对应的负荷数据是异常数据。

3.7　决策报警

由于模型分类的结果一定存在误差和窃电问题的特殊性，不能把每个时间点检测出来的异常数据都当成是窃电数据，像某些时间点的数据异常可能是其他非窃电行为如跳闸，设备检修等原因导致的，不能将其纳入窃电行为的范畴。

经查证有关电力公司的资料和根据数据规律分析，采用以连续 3 天发现异常数据作为警报的触发条件。也就是说，当每天 96 个检测时间点，连续 3 天，共计 288 个检测时间点均为异常数据时，可以认定其存在窃电行为，并报告首次检测到异常数据的时间和窃电报警的时间。当然，具体的判别标准也可根据实际情况而定。通过这样设置就可以有效的防止个别异常数据对分类结果的影响，排除误报。

4　算法验证

为了检验样本选取方法的可行性和算法在实际窃电辨识中的准确率，从用电数据采集系统中提取某地区针织厂的用电数据，并在 Matlab 环境下对其进行分析验证。

根据公式（1）计算某地区针织厂的电量数据波动系数，如图 3 所示。其中已知该用户 7、8 两个月存在窃电行为，从

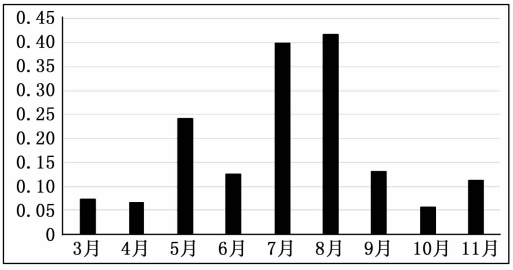

图 3　某针织厂每月电量波动系数

图 3 中可以看出这两个月份的波动系数较其它月份大。根据电量波动系数的不同将 5 月份和 3 月份的三相电流数据分别作为训练样本进行分析处理，其结果如图 4 和图 5 所示。

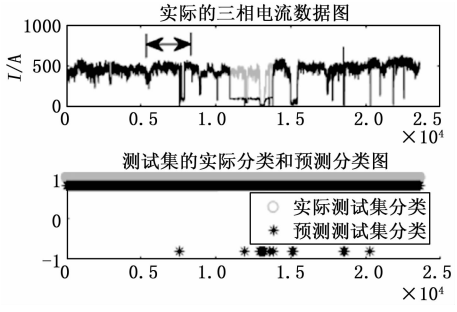

图 4　某针织厂 5 月份三相电流特征及分析结果

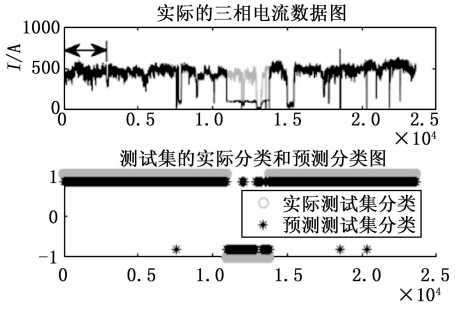

图 5　某针织厂 3 月份三相电流特征及分析结果

图 4 和图 5 中，上子图中黄、绿、红三线分别代表中 A、B、C 三相电流（单位为 A）；横坐标代表数据的序号，对应各个负荷数据采样点；双箭头标注的区域为样本数据选取范围。下子图纵坐标代表分类的类别，1 代表正常数据，－1 代表异常数据；蓝色部分代表实际的数据分类，而红色的部分代表模型分类的结果（注：为了在图上以示区分，将预测测试集分类的值乘以系数 0.8）。

将分析结果中正确分类的个数占总测试数据个数的比值称作分类准确率，用以衡量分类结果的好坏。从图 4 和图 5 中可以看出，选取 3 月份负荷数据作为训练样本得到的分类效果要明显好于选取 5 月份的负荷数据作为训练样本的分类效果，它们的分类准确率分别为 97.85％和 87.90％，且前者发出窃电警报的时间与实际窃电时间相符。同时，若分别以窃电发生之前的 4、6 月份的负荷数据作为训练样本，得到的分类结果与图 3 相似，分类准确度分别为 97.79％和 97.82％。

以上分析说明根据电量波动系数选取样本的方法是可行的。根据现有研究，选取电量波动系数小于 0.2 时的负荷数据作为训练样本选择依据，这个阈值可根据负荷类型情况作相应

调整。

从所分析的样本数据中提取了同一地区 6 个用户的窃电数据，将使用算法检测得到的窃电时间和电力公司实际查证的窃电时间相对比。为了减少误报，连续 3 天检测到异常数据时才发出窃电警报。对算法发出窃电警报时间和实际查证的窃电时间进行比较，如表 1 所示。

表 1 各用户实际窃电时间与检测窃电时间对比

用户名称	实际窃电时间	警报时间
纺织公司 a	06－17 16:00:00	06－20 16:00:00
纺织公司 b	06－16 17:30:00	06－19 17:30:00
某针织厂	07－03 12:00:00	07－06 18:15:00
某酒店	07－04 07:30:00	07－07 11:15:00
某化纤公司	07－04 07:30:00	07－07 11:15:00
丝绸厂 a	05－01 07:45:00	05－04 08:00:00
丝绸厂 b	05－29 12:00:00	06－01 12:00:00

表 1 中表明算法警报时间与实际窃电实际基本吻合，其中存在的误差主要原因可分为：1）一些用电数据的缺失；2）样本数据的选取导致训练模型有偏差；3）参数还可以进一步优化。

5 结论

本文基于窃电现状和一些反窃电的研究成果，在目前用电信息采集系统数据处理分析不够充分的情况下，提出了一种利用电量波动系数来优化选取样本，然后利用支持向量机算法对用户用电信息进行处理分析的窃电辨识方法。通过数据采集、数据预处理、参数优化、决策函数及报警等步骤，最终得到的算法结果能够满足要求。经过实际的检验分析可知，这种方法对于窃电问题的分析处理效果比较理想，能基本实现区分正常用电数据和窃电数据的功能。该方法为防窃电工作提供了一种新的思路。

但是由于负荷类型的多样性，以及有些窃电手法的隐蔽性，不能说某一种窃电辨识方法可以识别所有的窃电行为，也不可避免地存在误报的现象。需要在今后进一步研究，完善本文所提算法，或者将文中的算法与其它数据挖掘算法融合，以进一步提高窃电辨识的准确性。

参考文献：

[1] 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.

[2] 吴倩红，高 军，侯广松，等. 实现影响因素多源异构融合的短期负荷预测支持向量机算法 [J]. 电力系统自动化，2016，40 (15)：67－72.

[3] 周文婷，顾 楠，王 涛，等. 基于数据挖掘算法的用户窃电嫌疑分析 [J]. 河南科学，2015，33 (10)：1767－1772.

[4] 谢晶晶. 基于层次模型的电能表管理与数据分析方法研究 [D]. 南京：南京邮电大学，2016.

[5] 简富俊，曹 敏，王 磊，等. 基于 SVM 的 AMI 环境下用电异常检测研究 [J]. 电测与仪表，2014，51 (6)：64－69.

[6] 张晓宇，付 林，沈 炯，等. 基于在线支持向量机的锅炉动态建模方法研究 [A]. 中国电机工程学会年会 [C]. 2016.

[7] 朱雪芳. 改进支持向量聚类算法的研究 [J]. 计算机测量与控制，2006，14 (12)：1732－1735.

[8] 杨锡运，孙宝君，张新房，等. 基于相似数据的支持向量机短期风速预测仿真研究 [J]. 中国电机工程学报，2012，32 (4)：35－41.

[9] SchöLkopf B，Smola A J，Williamson R C，et al. New Support Vector Algorithms [J]. Neural Computation，2000，12 (5)：1207.

[10] 舒胜文，阮江军，黄道春，等. 基于电场特征量和 SVM 的空气间隙击穿电压预测 [J]. 中国电机工程学报，2015，35 (3)：742－750.

(上接第 222 页)

5 结语

本文针对不同冗余度连接方式下的多自治域网络，分析了其拓扑结构，建立了级联失效模型并进行仿真分析，观察不同冗余度下发包率、转发能力和容量对级联失效的影响，仿真结果表明，随着发包率的增加，网络失效规模增大；冗余度大的网络其结构鲁棒性好，发生级联失效时网络性能较好；增大转发能力比增大节点容量可以更有效的控制级联失效的传播。

在现实网络环境中，网络规划人员可以结合多自治域网络的吞吐率和平均负载情况，选择合理的冗余度接入方式，防止网络级联失效的发生。一旦级联失效发生，网络管理人员可以根据网络性能恶化的不同方面采取针对性的补救措施。

参考文献：

[1] 黎 松，诸葛建伟，李 星. BGP 安全研究 [J]. 软件学报，2013，24 (1)：121－138.

[2] Li Q，Zhang X W，Zhang X，et al. Invalidating idealized BGP security proposals and countermeasures [J]. IEEE Trans on Dependable and Secure Computing，2015，12 (3)：298－311.

[3] Tanenbaum A S. 计算机网络 [M]. 潘爱民，译. 4 版. 北京：清华大学出版社，2004：387.

[4] 沈 迪，李建华，等. 一种基于介数的双层复杂网络级联失效模型 [J]. 复杂系统与复杂性科学，2014，11 (3)：12－18.

[5] 袁 铭. 带有层次结构的复杂网络级联失效模型 [J]. 物理学报，2014，63 (22)：220501.

[6] 陈世明，邹小群，等. 面向级联失效的相依网络鲁棒性研究 [J]. 物理学报，2014，63 (2)：028902.

[7] 王正武，王 杰，等. 控制城市道路交通网络级联失效的关闭策略 [J]. 系统工程，2016，34 (2)：103－108.

[8] 尹洪英，权小锋. 交通运输网络级联失效影响规律及影响范围 [J]. 系统管理学报，2013，22 (6)：869－875.

[9] 邱 菡，李玉峰，等. 域间路由系统的级联失效攻击及检测研究 [J]. 中国科学：信息技术，2017，47 (12)：1715－1729.

[10] 陆余良，杨 斌. 域间路由级联失效分析与建模 [J]. 系统工程与电子技术，2016，38 (1)：172－178.

[11] 苗 甫，王振兴，等. BGP－SIS：一种域间路由系统 BGP－LDoS 攻击威胁传播模型 [J]. 计算机应用研究，2017，34 (12)：3735－3739.

[12] Magoni D，Pansiot J J. Internet topology modeler based on map sampling [C]. Piscataway，NJ：IEEE，2002：1021－1027.

[13] Waxman B M. Routing of multipoint connections [J]. IEEE Journal on Selected Areas in Communications，1988，6 (9)：1617－1622.

[14] 赵 娟，郭 平，邓宏钟，等. 基于信息流动力学的通信网络性能可靠性建模与分析 [J]. 通信学报，2011，8 (32)：159－164.

[15] Brandes，U. A faster algorithm for betweenness centrality [J]. Journal of Mathematical Sociology，2001，25 (2)：163－177.