ELM 算法在用户用电行为分析中的应用①

胡殿刚[1], 李韶瑜[1], 楼 俏[2], 王 琼[2], 程淼海[2], 王国军[2], 李国辉[3,4]

[1](国网甘肃省电力公司, 兰州 730030)
[2](国网甘肃省电力公司兰州供电公司, 兰州 730050)
[3](福州大学 数学与计算机科学学院, 福州 350116)
[4](福建省网络计算与智能信息处理重点实验室, 福州 350116)

摘 要: 对于非法用电行为的检测, 电力企业通常采用传统的人工检查方式, 而这种方式的准确率和效率往往都比较低. 提出一种将极限学习机(ELM)应用于预测存在非法用电行为用户的方法. 首先, 在收集到的用户历史用电数据, 对原始数据进行预处理. 然后, 应用 ELM 算法建立异常用电行为的神经网络模型. 最后, 在真实用电数据上进行实证分析, 通过与随机森林算法建立的预测模型及预测结果的对比, 证明提出的方法具有较高的准确率和较好的性能.

关键词: 极限学习机; 特征选择; 用户用电行为

Application of ELM Algorithm in the Analysis of Customer Electrical Behavior

HU Dian-Gang[1], LI Shao-Yu[1], LOU Qiao[2], WANG Qiong[2], CHENG Miao-Hai[2], WANG Guo-Jun[2], LI Guo-Hui[3,4]

[1](State Grid Gansu Electic Power Company, Lanzhou 730030, China)
[2](State Grid Lanzhou Branch Electic Power Company of Gansu, Lanzhou 730050, China)
[3](College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)
[4](Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou 350116, China)

Abstract: In order to detect the illegal use of electricity, electrical enterprises generally adopt traditional manual examination ways. However, both of the accuracy and efficiency of the approaches are far from satisfaction. In this paper, an analysis method based on the Extreme Learning Machine (ELM) algorithm is proposed, which is used to predict the behavior of customers' illegal electric use. Firstly, it collects the historical electric usage data and preprocesses the data to make it suitable for analysis by the algorithms. Then, it applys an algorithm based on neural network model, which is called ELM, to build the model to describe the abnormal power utilization behavior of the customers. Finally, experiments on the real electrical consumption data are conducted to evaluate the proposed method. The experimental results demonstrate that the proposed method is accurate and efficient.

Key words: extreme learning machine; feature selection; customer electricity behavior

随着我国的经济的蓬勃发展, 电力企业成为了经济发展的重要组成部分. 用电检查环节是整个供电企业的主要组成部分, 也是保证居民正常用电的基础[1]. 我国的用电企业, 在对电力市场以及用电客户的安全工作上, 发挥了重要作用, 起到了验收反馈的主要功能. 供电企业为了保证居民能够正常、安全用电所进行了一系列检查、指导工作. 用电检查也是对用电营销后的售后服务, 这种服务, 对稳定顾客, 提高服务质量, 促进营销都起到了重要的作用[2]. 所以, 要想使得供电企业能够顺利工作, 不断加强用电检查, 提高用电检查的管理效率, 成为了重中之重.

国内外许多专家学者对用户用电行为特征进行了

① 基金项目:国家自然科学基金(61103175,61300104);教育部科学技术研究重点项目(212086);福建省科技创新平台建设(2009J1007);福建省自然科学基金(2013J01230);福建省高校杰出青年科学基金(JA12016);福建省高等学校新世纪优秀人才支持计划(JA13021)
收稿时间:2015-12-08;收到修改稿时间:2016-03-31 [doi:10.15888/j.cnki.csa.005305]

大量研究. 王继业等人针对智能配用电业务, 首先分析智能配用电大数据的特征, 然后分析数据融合后的智能配用电大数据整体业务需求和应用场景, 提出大数据环境下的研究思路和方法, 最后给出了智能配用电大数据应用技术架构[3]. 谢涛等人针对非法用电行为构建线性方程组数学模型, 基于智能电网中电表的可编程可计算特性, 提出了分布式的检测方法, 将各用户的非法用电行为检测交由附属智能电表就地计算解决[4]. 简富俊等人根据高级测量体系系统架构的特点,使用 One-class SVM 无监督机器学习架构对电力用户负荷异常进行检测,可以在小样本、样本分类不均衡环境下提高检测的准确性[5]. 冯晓蒲等人使用典型算法模糊 c 均值(FcM)对其进行聚类分析, 得到负荷簇和负荷代表曲线, 然后分析了属于各行业和电价类的用户负荷聚类结果, 显示了按负荷特性进行用户分类与现行按行业和电价的用户分类差异显著[6]. 林嘉晖构建了一个适用于电网企业的用户行为分析系统并实现了部分经典的数据挖掘算法, 该系统能够对现有信息管理系统留下来的大量用户数据进行分析, 挖掘出其中深层的关联规则, 并转变为决策型信息, 以辅助电网企业的市场营销决策并提高其客户服务水平[7].

基于上述研究现状, 本文提出了一种非法用电行为用户预测的方法, 该方法基于 ELM 算法[8]来建立预测模型. ELM 算法与传统的前馈神经网络相比, 具有速度更快, 泛化能力更强等优点. 因此, 研究基于 ELM 算法的非法用电行为用户预测的方法对供电企业提高效率具有一定的参考价值.

全文组织结构如下: 第 1 节介绍了数据预处理方法. 第 2 节介绍了 ELM 算法和实验处理流程. 第 3 节介绍了实验数据以及实验结果. 第 4 节对本文所做的内容进行总结并对下一步研究提出展望.

1 数据预处理

数据挖掘是在大量的数据中挖掘出有用模式的过程, 数据源的质量直接影响到了挖掘的效果. 由于原始的负荷数据质量并不完美, 往往会存在一些数值的缺失, 还有离群值等. 因此, 在进行数据挖掘之前必须要对数据进行预处理. 数据预处理是数据挖掘的重要步骤, 它主要包括数据整合、数据填充、特征规范化、特征选择等步骤.

如果将用户的用电数据用曲线表示出来, 纵坐标表示用电负荷值, 横坐标表示时间, 可以发现, 每个正常用户对应的用电曲线都较相似. 而非法用电行为所检测出来的数据相比于正常用户, 波动很大.

1) 数据整合

整合包括硬件和软件的整合, 企业内部和跨企业的整合, 操作环境和业务流程的整合等等. 整合的首要问题是数据源的整合. 数据整合是把在不同数据源的数据收集、整理、清洗, 转换后加载到一个新的数据源, 整合后统一的数据平台可以用于业务分析和领导决策等[9].

2) 数据填充

原始数据通常存在缺失或错漏. 例如: 某些用户若干特征的值缺失(为空值)或明显异常(如用电量为负数). 需要对其处理, 才能保证后续算法的正常分析. 常见的处理方式包括: 填充为统一的默认值、填充为特征的统计量(如均值、最小值、中位数等)、删除包含异常值的记录等[10].

3) 特征规范化

原始数据不同特征的值域可能存在较大差异. 因此, 需要对原始数据做规范化. 常见的规范化方法包括: 区间规范化, 最大值规范化, 标准规范化[11].

4) 特征选择

原始数据的特征数量较多时, 一方面可能由于特征之间的相关性给分析带来困难, 另一方面也增加了算法运行的时间. 因此, 需要对原始数据的特征进行选择. 常见的特征选择方法包括: 基于信息熵的方法(InfoGain), χ^2 检验, 主成分分析(PCA), 基于特征相关性的方法(CFS)[12].

基于特征相关性的方法(CFS)[13]是一种经典的过滤器模式的特征选择算法, 它启发式地对单一特征对应于每个分类的作用来进行评价, 从而得出最终的特征子集, 其形式化的评估方法如下:

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (1)$$

其中, $Merit_s$ 表示一个包含 k 个特征的特征子集 S 的一个评价; $\overline{r_{cf}}$ 表示对应于该子集的特征-子集平均相关度, 其中 $f \in S$; $\overline{r_{ff}}$ 表示特征-特征的平均相关度. 公式(1)给出的所有的变量都是经过标准化的. 该评价指标能够有效地给出特征对于分类的贡献度, 并清除不相关的或者是贡献度很小的特征. 这些特征往往与其他特征相关度极高.

在公式(1)中, 所有的特征都必须是离散的随机变量,

如果是数值型变量, 必须首先对其进行离散化, 而且需要通过熵计算方式来对特征间的相关性进行评价.

2 基于ELM算法的用电行为分析

2.1 ELM 算法介绍

ELM 算法是由黄广斌教授提出来的求解单隐层前馈神经网络(SLFNs)的算法. 该算法的特点是在确定网络参数的过程中可以随机确定输入权重和偏置, 在训练过程中无需调节, 只需要设置隐层神经元的个数, 便可以获得唯一的最优解. 而网络的输出权值是通过最小化平方损失函数得到的最小二乘解(最终转化成求解一个矩阵的 Moore-Penrose 广义逆问题), 这样在确定网络参数的过程中就无需进行任何迭代步骤, 从而大大降低了网络参数的调节时间. 对比大量的传统神经网络, 尤其是 SLFNs, 该方法在保证精度的前提下, 具有比其他传统方法学习速度更快、泛化性能更好等优点[14].

以下介绍 ELM 算法的一些概念和算法描述:

对于一个单隐层神经网络, 设有 N 个任意的样本 (x_i,t_i), $x_i=[x_{i1},x_{i2}\cdots,x_{in}]^T\in R^n$, 表示一个包含. 对于有 $L(L<N)$个隐层结点的 SLFNs 可以用模型形式表示为:

$$\sum_{i=1}^{L}\beta_i f(w_i\cdot x_j+b_i)=o_j,j=1,\cdots,N \tag{2}$$

其中, $w=(w_{ij})_{L*n}$ 表示隐含层与输入神经元之间的权重, $w_i=[w_{i1},w_{i2}\cdots,w_{in}]^T$ 为连接隐含层第 i 个结点和输入神经元之间的输入权重, $b=[b_1,b_2,\cdots,b_L]^T$ 为偏置值, $f(x)$ 为激活函数, $w_i\cdot x_j$ 表示两者之间的内积.

单隐含层神经网络的目标是使得输出的误差最小, 可以表示为

$$\sum_{i=1}^{L}\|o_j-t_j\| \tag{3}$$

也就是说, 存在 β_i,w_i,b_i 使得

$$\sum_{i=1}^{L}\beta_i f(w_i\cdot x_j+b_i)=o_j,j=1,\cdots,N \tag{4}$$

用矩阵表示为

$$H\beta=T \tag{5}$$

其中, H 表示隐含层的输出.

$$H=\begin{pmatrix} f(w_1\cdot x_1+b_1) & \cdots & f(w_L\cdot x_1+b_L)\\ \vdots & \ddots & \vdots\\ f(w_1\cdot x_N+b_1) & \cdots & f(w_1\cdot x_N+b_L)\end{pmatrix} \tag{6}$$

表示输出权重, $\beta=(\beta_1,\cdots,\beta_L)^T$. **T** 为期望输出, $T=(t_1,\cdots,t_N)^T$.

传统的神经网络大多基于梯度下降法, 但是基于梯度的学习需要在迭代过程中不断调整所有的参数. 而在 ELM 算法中, 一旦输入权重和隐含层的偏置值被随机确定, 它就可以转化为求解一个线性系统公式(4), 这样, 输出权重就可以确定了, $\beta=H^+T$, H^+ 为矩阵 H 的 Moore-Penrose 广义逆[15]. ELM 的算法如下.

算法 1

输入	样本集N个任意的样本(x_i,t_i), 激活函数$f(x)$, 隐含层结点数L
输出	输出权值β
步骤 1	随机生成输入权值w和b
步骤 2	计算隐含层输出矩阵H
步骤 3	计算输出权值β, $\beta=H^+T$, $T=(t_1,\cdots,t_N)^T$

2.2 总体流程处理

总体上, 数据处理分为用电数据模型训练和模型预测两个主要流程. 用电数据模型训练指的是根据输入的用于训练的原始用电数据得到数学模型(根据不同的算法有神经网络、决策树、回归方程等不同形式), 输入原始用户用电量数据后, 对其进行预处理, 将其产生的数据运用 ELM 算法, 得到模型. 模型预测指的是将训练得到的数学模型应用于输入的需要预测的原始用电数据, 并对预测结果进行检验. 对要预测的用电数据进行预处理后, 将其输入到之前所建立的模型, 对数据进行验证. 如图 1 所示.

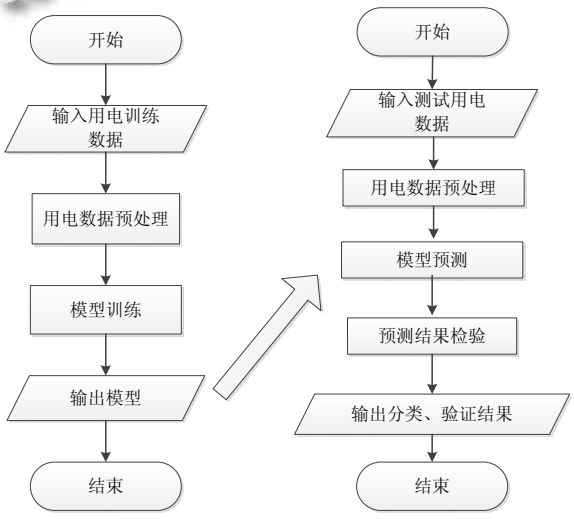

图 1 数据处理总体流程

3 实验与分析

3.1 实验数据及其预处理

本实验数据来自某电力企业, 我们对其进行了预处理, 产生的数据作为算法的输入数据. 每条记录共有 26 个字段, 分别为用户编号, 一年中的平均抄见电量和抄见电量最大值, 如表 2 所示. 将 3 个商业电价编码为 200, 300, 370 的数据与真实违规用户数据合并成 1 个训练集, 3 个商业电价编码数据对应 3 个训练集. 训练得到模型后, 将电价编码为 001 的数据集作为测试集. 原始数据表字段说明如表 1 所示.

表 1 原始数据表字段说明

字段名	字段说明
CONS_NO	用户编号
CP_NO	采集点编号
MPED_ID	测量点编号
DATA_DATE	数据日期
TRADE_CODE	行业编号
KWH	抄见电量

3.1.1 数据整合

由于低压用户用电量统计表较大, 且许多字段在用电量分析中不需要用到, 对这些表进行过滤, 只保留 CONS_TG_NO(用户台区编号)、MPED_ID(测量点标识)、PR_ORG(分区单位)、DATA_DATE(数据日期)、KWH(抄见电量)、行业编码(TRADE_CODE)等关键字段. 此外, 由于原始数据是按天统计用电量, 全部作为特征会导致特征膨胀, 影响数据挖掘的质量. 我们根据原始用电量信息组合出 2 个关键特征: 每日用电峰值、每月用电总额. 整合后的数据表中所有特征值均为 0 的记录相当于空记录, 对分析作用不大, 也一并删除. 如表 2 所示.

表 2 整合后数据表字段说明

字段名	字段说明
CONS_NO	用户编号
KWH_01_SUM	1 月份抄见电量总和
⋮	⋮
KWH_12_SUM	12 月份抄见电量总和
KWH_01_MAX	1 月份抄见电量最大值
⋮	⋮
KWH_12_MAX	12 月份抄见电量最大值

3.1.2 数据填充

对于缺失值, 填充为 0; 对于异常值, 由于包含异常值的记录很少, 直接删除包含异常值的记录.

3.1.3 特征规范化

例如: 大型企业客户的用电量可能达到 10000 度以上, 而居民客户的用电量一般在 500 度以下. 如果直接在原始数据上分析, 数值大的特征将湮没数值小的特征, 使催收次数这样的特征无法得到有效利用. 根据“(原始值-最小值)/(最大值-最小值)”, 将 KWH_01_SUM 至 KWH_12_SUM 以及 KWH_01_MAX 至 KWH_12_MAX 等 24 个负荷值归一化到[0,1]区间. 如果某个特征的取值全为 0, 将导致规范化公式的分母为 0. 此时, 不对该特征规范化, 即保持原始值 0.

3.1.4 特征选择

由于基于特征相关性的方法使用较为普遍, 并且效果较好, 因此, 采用第一节介绍的基于特征相关性的方法. 得到如表 3 所述字段.

表 3 特征选择后数据表字段说明

字段名	字段说明
KWH_05_SUM	5 月份抄见电量总和
KWH_06_SUM	6 月份抄见电量总和
KWH_07_SUM	7 月份抄见电量总和
KWH_09_SUM	9 月份抄见电量总和
KWH_10_SUM	10 月份抄见电量总和
KWH_12_SUM	12 月份抄见电量总和
KWH_01_MAX	1 月份抄见电量最大值

3.2 实验结果

使用 ELM 算法和随机森林算法[16]对实验数据进行实验.

由于真实违规用电用户的数据是由电力执法人员现场调查得到的. 受到人力限制, 调查人员只能对很小一部分可疑人员进行调查. 因此, 我们能够进行比对的真实违规用户数据集只是实际真实违规用户数据集上一个极小子集. 为了验证我们的分析结论, 对算法预测的违规用户信息与正常用户信息进行比对, 得到如图 2~7 所示的用户月平均用电量与月最大用电量对比图.

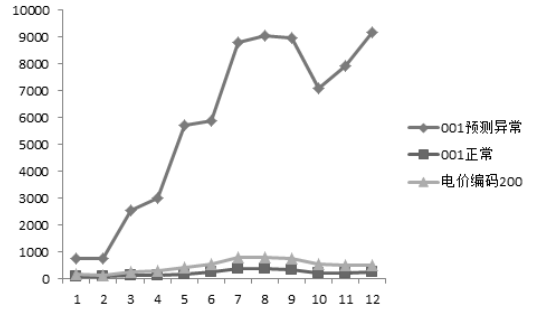

图 2 200 在 001 上用户月平均用电量(ELM)

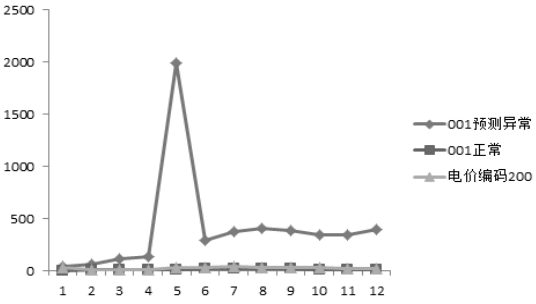

图 3 200 在 001 上用户月最大用电量(ELM)

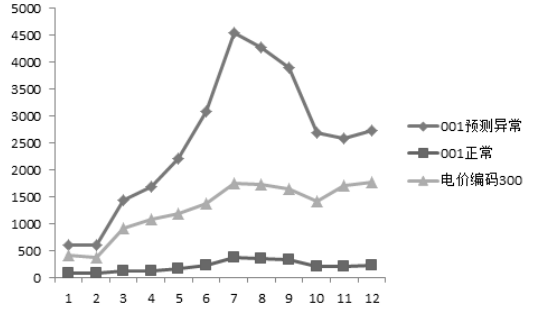

图 4 300 在 001 上用户月平均用电量(ELM)

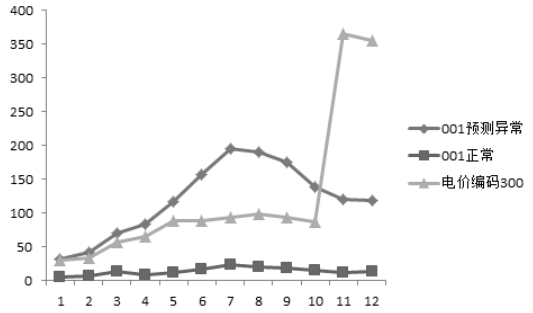

图 5 300 在 001 上用户月最大用电量(ELM)

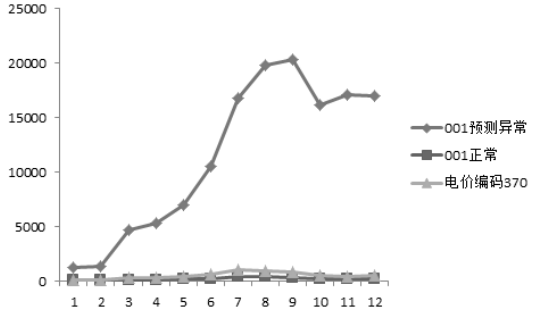

图 6 370 在 001 上用户月平均用电量(ELM)

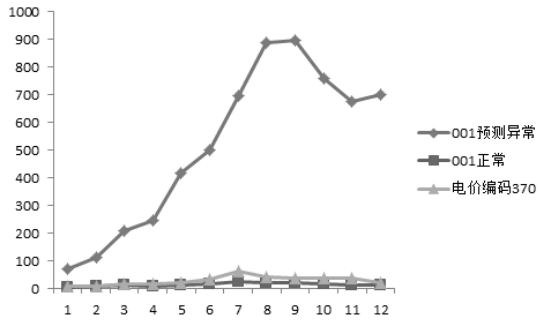

图 7 370 在 001 上用户月最大用电量(ELM)

从上面的实验结果分析图可以看出:

1) 算法预测得到的违规用户的用电量与正常用户的用电量有显著区别. 例如, 违规用户在 7、8、9 这三个月份的月平均用电量均在 5000 度以上, 而正常用户的月平均用电量不超过 500 度.

2) 算法预测得到的违规用户的用电量与真实违规用户的用电量比较接近. 这一点在图 2 中表现尤为明显, 两条曲线不但走势一致, 且几乎重合.

3) ELM 可以将不同月份用电特征的反映在神经网络结点连接边的权重上.

由此可以得出结论, 算法得到的结果是有比较高的可信度的.

使用随机森林算法, 得到如图 8~13 所示的用户月平均用电量与月最大用电量对比图.

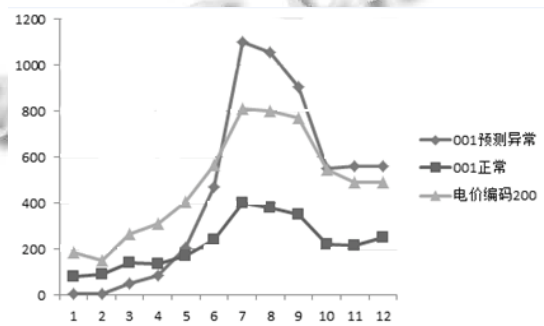

图 8 200 在 001 上用户月平均用电量(随机森林)

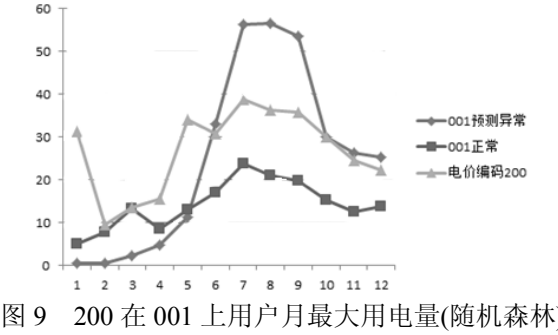

图 9 200 在 001 上用户月最大用电量(随机森林)

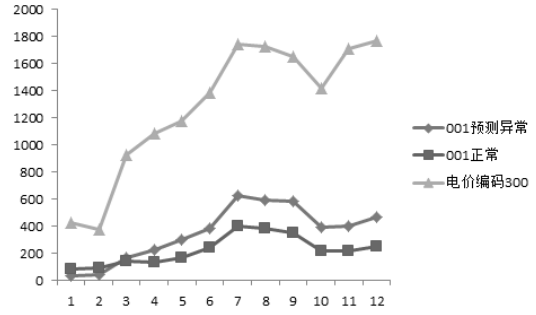

图 10 300 在 001 上用户月平均用电量(随机森林)

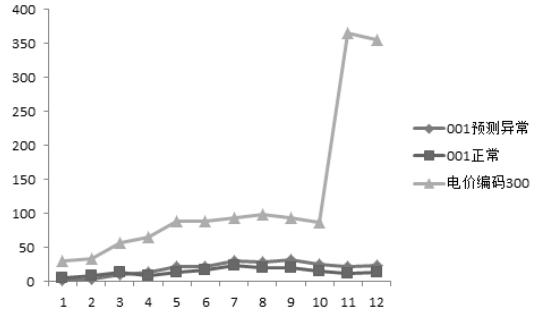

图 11 300 在 001 上用户月最大用电量(随机森林)

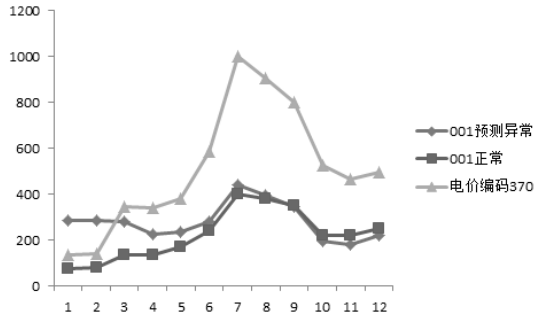

图 12 370 在 001 上用户月平均用电量(随机森林)

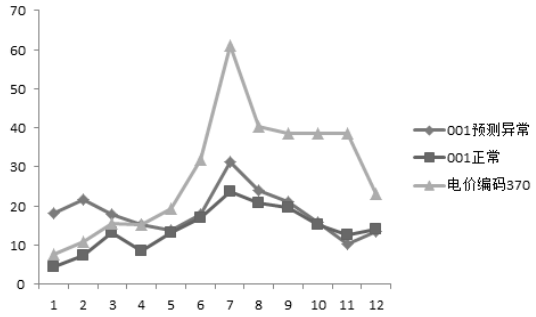

图 13 370 在 001 上用户月最大用电量(随机森林)

从图 8 至图 13 的结果可以得到和 ELM 实验相似的结论. 算法预测的违规用户月平均用电量显著高于正常用户, 而且通过算法预测得到的违规用户的用电量与真实违规用户的用电量较为相似. 随机森林生成的规则树能够有效提取最能反映不同用户用户量区别的特征月份.

比较两个算法, 我们发现ELM算法算法预测得到的违规用户的用电量与正常用户的用电量之间的差别更加显著, 而算法预测得到的违规用户的用电量与真实违规用户的用电量更加接近. 因此, 我们认为 ELM 算法本身的可信度是较高的.

4 总结

本文提出了一种基于 ELM 算法的非法用电行为用户方法. 首先, 采用数据整合, 数据填充, 数据特征规范化, 特征提取等方法对数据进行预处理, 然后, 采用了ELM算法对预处理后的数据进行了分类, 最后, 在真实的企业提供的用户用电数据集上进行了实验, 此外, 通过与随机森林算法的对比, 验证了提出的方法不仅具有较高的准确率, 而且具有很强的性能, 进一步提高了研究结果的效率和可信度. 通过本文提出的方法, 可以有效地识别出非法用电行为的用户, 指导用电检查人员有针对性地进行排查, 及时发现非法用电用户, 提高用电检查人员的工作效率. 下一步, 我们将就如何将算法并行化, 以进一步提高求解效率等问题进一步展开深入研究.

参考文献

1 魏瑶,朱伟义,龚桃荣,等.基于数据挖掘技术的用电异常分析系统设计.电力信息与通信技术,2014,12(5):70–73.
2 倪精华.供电企业用电检查工作中存在的问题及对策.企业家天地旬刊,2010:49–49.
3 王继业,季知祥,史梦洁,等.智能配用电大数据需求分析与应用研究.中国电机工程学报,35(8):1829–1836.
4 谢涛,靳丹,马志程,等.基于智能电网的分布式非法用电行为检测方法.微型电脑应用,2015,31(2):54–57.
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.电测与仪表,2014(06):64–69.
6 冯晓蒲.基于实际负荷曲线的电力用户分类技术研究[硕士学位论文].保定:华北电力大学,2011.
7 林嘉晖.基于数据挖掘的电网用户行为分析系统的设计与实现[硕士学位论文].广州:中山大学,2013.
8 Huang GB, Zhu QY, Siew CK. Extreme learning machine: a new learning scheme of feedforward neural networks. Proc. 2004 IEEE International Joint Conference on Neural Networks. IEEE. 2004. 985−990.
9 余晓平,瓮正科,张振宇,等.数据整合技术研究.兵团教育学

院学报,2006,1:32–33.
10 朱晓峰.缺失值填充的若干问题研究[硕士学位论文].桂林:广西师范大学,2007.
11 任家东,何海涛,郝忠孝.时态关系数据的特征及其规范化.小型微型计算机系统,2000,3:302–304.
12 王娟,慈林林,姚康泽.特征选择方法综述.计算机工程与科学,2005,12.
13 孙宁青.基于神经网络和 CFS 特征选择的网络入侵检测系统.计算机工程与科学,2010,32:37–39.
14 Huang GB, Zhu QY, Siew CK. Extreme learning machine: Theory and applications. Neurocomputing, 2006, 70: 489–501.
15 Prasad KM, Bapat RB. The generalized Moore-Penrose inverse. Linear Algebra & its Applications, 1992, 165(3): 59–69.
16 Liaw A, Wiener M. Classification and regression by randomForest. R news, 2002, 2(3): 18–2.