计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(9):232−236 [doi: 10.15888/j.cnki.csa.008157]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于采样技术和 LightGBM 的用户用电异常检测模型①

刘中强[1,2], 邹维维[2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
通讯作者: 刘中强, E-mail: 823795956@qq.com

摘　要: 在大数据的时代背景下, 我国电力事业信息化的发展日趋重要, 尤其是需要使用计算机技术对用电数据进行分析. 对于用户用电异常的分析问题, 传统方法既耗时又耗力, 这就需要引入机器学习的相关方法自动的识别异常信息. 现阶段, 用电异常分析主要基于传统的异常检测算法或深度神经网络, 传统异常检测算法运行精度不足而深度神经网络计算速度又过慢. 针对目前存在的不足, 本分采用了基于采样技术和 LightGBM 的用户用电异常检测模型, 把用电异常检测问题看作分类问题, 并使用当前流行的分类模型 LightGBM 进行训练, 在保证速度快的前提下提高了检测的准确率.

关键词: 机器学习; 用电异常; 采样技术; 分类模型

引用格式: 刘中强,邹维维.基于采样技术和 LightGBM 的用户用电异常检测模型.计算机系统应用,2021,30(9):232–236. http://www.c-s-a.org.cn/1003-3254/8157.html

Anomaly Detection Model of Consumer Power Consumption Based on Sampling Technology and LightGBM

LIU Zhong-Qiang[1,2], ZOU Wei-Wei[2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)

Abstract: In the context of big data, the informatization of China's power industry is becoming more important, especially the analysis of power consumption data with computer technology. For the analysis of abnormal user power consumption, traditional methods are time-consuming and labor-intensive. This requires the introduction of machine learning related methods to automatically identify anomaly information. At this stage, the analysis of abnormal power consumption is mainly based on traditional anomaly detection algorithms or deep neural networks. Anomaly detection algorithms have insufficient accuracy and calculations with deep neural networks are quite slow. In response to the current shortcomings, this study adopts an anomaly detection model of user power consumption based on sampling technology and LightGBM. The detection of abnormal power consumption is regarded as a classification problem, and the popular classification model LightGBM is applied to training. The detection accuracy is improved while fast speed is maintained.

Key words: machine learning; abnormal electricity consumption; sampling technique; classification model

国家电网公司通过十多年的信息化建设, 积累了海量生产运行和经营管理数据. 随着物联网技术的发展, 数据量还会继续增大. 按照十四五规划, 智能化和数字化必将是下一阶段的重点发展方向. 用电信息采

① 收稿时间: 2020-12-14; 修改时间: 2021-01-11, 2021-02-23; 采用时间: 2021-03-03; csa 在线出版时间: 2021-09-02

集系统的发展和推广, 为电力大数据分析提供了数据基础, 可以用来做用户用电异常行为分析. 所谓异常用电数据, 就是由于环境、系统或人为的因素导致的计量设备故障、异常及违规用电等行为, 可以通过电能表所反应的电压、电流、功率及线损率等指标进行反应. 但是目前大多数电力部门仅使用传统的统计方法进行异常分析, 费时费力且成效低. 本文将基于用户用电数据采集记录, 利用过采样技术对数量较少的异常类样本数据进行增多, 这样就可以把用户用电异常检测的问题看作是一个是否异常的二分类问题, 然后使用当前流行的机器学习模型, 对用户用电是否异常进行分类. 训练该模型的目的是帮助检测系统更加快速、准确地识别异常用电行为.

1 用户用电异常检测

基于机器学习的用户用电异常检测已经有很多研究成果: 张荣昌[1] 选择孤立森林算法应用到用户用电异常检测问题, 该方法使用随机切分的办法划分数据集来得到一组决策树从而构成孤立森林, 根据叶子结点距离根节点的距离判断该叶子结点上的样本是否为异常样本, 该方法规则简单易于实现并且速度非常快, 但是检测精度并不够高; 张小秋等人[2] 提出了一种基于逻辑回归的增量式异常用电行为检测方法, 该方法把每天的用户用电数据的序列都看作一个单独的数据集来训练逻辑回归模型, 然后利用所有的逻辑回归模型完成增量式的学习; 张颖等人[3] 采用最小二乘法和聚类算法来对用户用电数据进行分析, 由于正常用户的用电数据曲线基本符合正态分布, 所以该算法使用最小二乘法近似逼近正态分布来求出每个用户用电数据 曲线, 再用 K-means 聚类算法找出异常用电数据; 郝方舟等人[4] 提出基于高维随机矩阵的用电行为分析方法, 该方法将用户用电数据做归一化后看作高维随机矩阵, 利用随机矩阵的一些特性和原理对用电是否异常做判定, 但是该方法在处理完数据之后需要人工对数据的统计特征进行分析, 效率不够高; 林女贵等人[5] 通过改进深度自编码网络的方式进行用电异常检测, 在自编码网络中引入稀疏约束和噪声编码, 提高了计算效率和模型的鲁棒性, 该方法准确率较高但由于神经网络层数较深, 所以运行速度较慢. 针对上述现阶段算法存在的问题, 本文采用了魏志强等人[6] 在解决 Web 异常检测问题时使用的 LightGBM 模型, 该模型结合 SMOTE 和 Tomeklinks 采样算法将异常检测问题转化为了一个分类问题. 与 Web 异常检测问题相比, 用户用电异常检测问题同样存在数据量大、正负样本不平衡等问题, 所以使用该模型可以保证训练速度较快的同时提高了异常用电用户检测的准确率.

2 基本算法介绍

2.1 采样技术

类别不平衡问题往往会导致模型训练结果出现较大的偏差, 所以对于正、负类别样本数量差距比较大的情况一般会使用采样技术对原始数据进行增加或删减来构建新的数据集, 这样做可以使模型的训练结果更加稳定.

2.2 SMOTE 算法

SMOTE 是过采样算法, 通过对少数类样本数量增多来达到样本平衡的目的. 图 1 是 SMOTE 过采样算法的做法: 随机取出一个少数类样本的 k 近邻样本, 然后与原始样本做线性组合来生成新的样本.

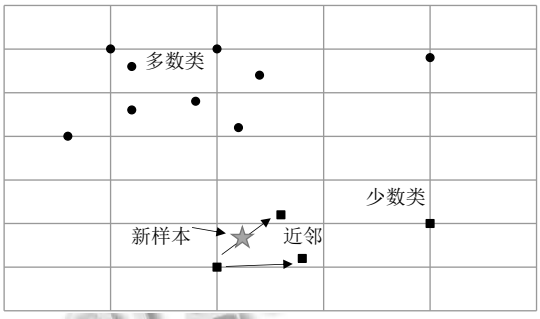

图 1 SMOTE 过采样算法

SMOTE 过采样算法的流程如算法 1.

算法 1. SMOTE 过采样算法

1) 输入数据集, 其中正样本为多数类, 负样本为少数类, 设置 k 近邻的 k 值;
2) 利用欧氏距离, 计算出每个负样本的 k 近邻, 这里可以使用 KD 树等方法降低求两个样本之间距离的计算量;
3) 对于每一个负样本 X, 从它的 k 近邻中随机取出一个样本 O;
4) 利用原来的样本 X 和随机取出的样本 O, 可以通过如下公式构建出新的样本 Y, 其中 $rand$ 是生成随机数的函数: $Y=X+rand(0,1)|O-X|$
5) 输出一个新的负样本 Y;
6) 重复 3)~5) 的操作, 直至正负样本比例平衡;

2.3 TomekLinks 算法

TomekLinks 是欠采样算法, 对于任意的两个样本 X 和 Y, 如果 X 和 Y 互为最近邻样本, 那么这两个样本则称为 TomekLink. 对于任意的 TomekLink, 它们都应

该属于同一个类别, 如果不是这样, 那么这两个样本就至少有一个划分到了错误类别, 删除错误分类样本就可以达到欠采样的目的.

由于 SMOTE 过采样算法是随机的选取少数类样本周围的近邻样本进行线性组合得到新的样本, 所以有可能从负样本的区域扩散到正样本的区域, 所以在对原始数据进行过采样之后, 可以配合使用 TomekLinks 算法删除这种正负样本重叠的区域.

2.4 GBDT 算法

梯度提升树 (GBDT) 算法是基于分类回归树 (CART) 的加法模型. 每一轮训练拟合上一轮产生的残差来生成一棵 CART 树, 经过多轮迭代, 得到最终的模型. 这里拟合残差的思路类似梯度下降, 将每一轮的目标函数 f 看作梯度下降的参数来优化, 公式为:

$$f_{n+1} = f_n + \Delta f \tag{1}$$

设总体目标函数为 L, 对于一个样本 x, 它的标签为 y, 那么 Δf 的计算公式如下:

$$\Delta f(x) = -\left[\frac{\partial L(y, f(x))}{\partial f(x)}\right]_{f(x)=f_{n-1}(x)} \tag{2}$$

由于 f_n 的部分已经固定, 所以每次只迭代只需要学习 Δf 这部分, 这就是拟合梯度的残差.

2.5 LightGBM 算法

LightGBM 是基于 GBDT 实现的一个具体模型, 在原始的 GBDT 基础上做了一些优化. LightGBM 的目标函数改进为二阶泰勒展开并加入了正则化项. LightGBM 的目标函数公式如下所示:

$$L_n = \sum_{i=1}^{n} l\left(y^i, y_{n-1}^i + f_n\left(x^i\right)\right) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

其中, x^i 表示第 i 个样本, y^i 表示它的标签, l 表示原始的目标函数, L_n 表示添加正则化项后第 n 次迭代的目标函数, f_n 是第 n 次迭代的模型, γ 和 λ 是参数, T 是叶子结点数量, w_i 表示第 j 个叶子结点的输出值.

LightGBM 的分割点选择是基于直方图的算法, 大大提高了选择分割点的速度. 对于决策树的增长策略, LightGBM 是使用 leaf-wise 的策略, 每次分裂选择收益最大的点分裂, 可以提高模型的精度. 此外, LightGBM 在特征的处理上和并行计算上都做了很多的优化, 是当前流行的机器学习模型, 相对于深度神经网络模型拥有速度快的优势, 相对于传统机器学习模型又有着精度高的优势, 所以本文选择 LightGBM 模型进行分类.

3 基于采样技术和 LightGBM 的检测模型

实验首先对数据集进行预处理, 然后提取特征构建训练集, 在得到训练集之后对数据进行过采样使得正负样本数量平衡, 接下来使用欠采样算法过滤掉训练集中正负样本重合的部分, 最后将处理好的训练集输入到 LightGBM 模型中进行分类, 图 2 是实验的整体流程.

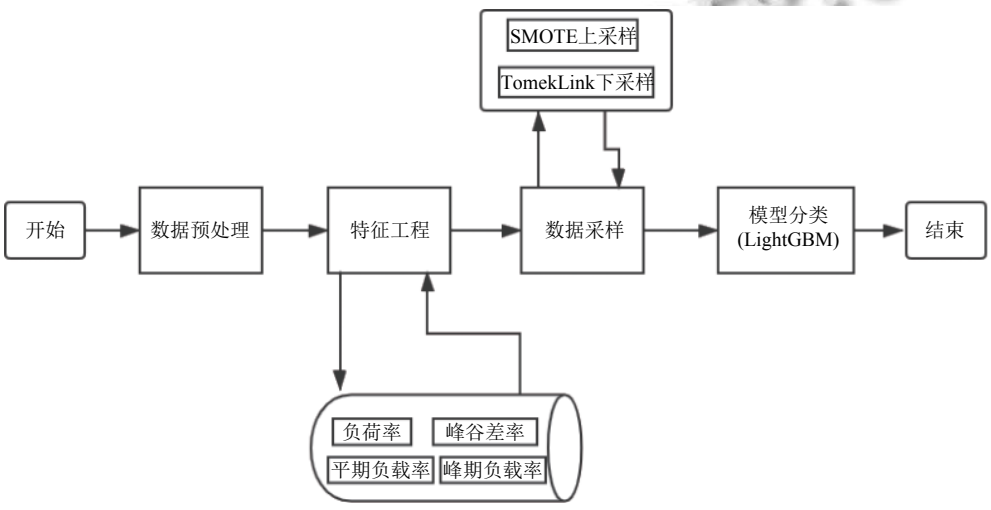

图 2 实验的整体流程

3.1 数据集

本文数据采用某地区采集到的用户用电数据, 数据每隔 15 分钟采集一次, 采集的数据有电压、电流、电量等信息. 每一条数据表示一个用户一天 24 小时的

用电情况, 这样每条数据就有 96 个数据点. 该数据集已经标注好了用电是否异常的标签, 其中异常数据占比约 1.5%.

3.2 数据预处理和提取特征

首先对数据缺失值和数据冗余情况进行处理, 经过预处理的数据再提取特征. 对于每条用户用电数据主要提取用电的统计特征. 本文从全天、峰期 (用电高峰时间段)、谷期 (用电低谷时间段)、平期这 4 个时间段的特征来反应用户全天的用电特征, 对于每一个时段单独计算一组如下特征.

平均负荷类特征, 计算电压、电流或功率在该时段的平均值除以最大值, 用来反应整个时段用电数据变化情况.

$$a1 = P_{\text{av}}/P_{\max} \tag{4}$$

差值类特征, 计算电压、电流或功率在该时段的最大值与最小值的差值除以最大值, 用来反应用电峰值与谷值的差异程度.

$$a2 = (P_{\max} - P_{\min})/P_{\max} \tag{5}$$

TOP 值类特征, 计算该时段用电数据前三的时间点的平均值, 用以反应用电高峰时期一段时间内的统计信息.

$$a3 = \sum_{i=top3} P_i/3 \tag{6}$$

计算好每一个用户的多组统计信息作为这个用户的全部特征, 然后将计算好的特征数据作为模型的训练集.

3.3 数据采样

由于数据的正负样本极度不平衡. 所以使用 SMOTE 算法对数量较少的负样本进行过采样, 采样之后的数据正负样本数量相等. 由于 SMOTE 过采样算法生成的负样本会与原来的正样本区域产生重叠, 所以使用 TomekLinks 算法将所处区域极度接近但是不属于同一类别的样本删除, 这样既可以减少训练数据的规模也可以去除掉难以判断的样本, 提高训练速度和准确度. 经过两个算法对数据的处理之后采样结果如表 1 所示.

表 1 采样结果

类型	正样本数量	负样本数量
原数据	415 432	6225
过采样后数据	415 432	415 432
欠采样后数据	394 660	394 660

3.4 评价指标

评价分类结果的 4 个基本指标是: *TP*(真正例)、*FP*(假正例)、*TN*(真负例)、*FN*(假负例). 这 4 个基本指标主要用于度量预测结果中正、负样本分类正确与错误的数量. 本文主要使用 3 个评价指标: 准确率、*F*1_*score*、AUC 值, 这些评价指标的公式如下:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{7}$$

$$F1\_score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{8}$$

$$TPR = \frac{TP}{TP+FN} \tag{9}$$

$$FPR = \frac{FP}{FP+TN} \tag{10}$$

其中, *Precision* 是准确率, *Recall* 是召回率, 而 AUC 值是 ROC 曲线下面积, 将预测样本根据概率大小依次作为阈值划分正、负样本, 然后计算每组阈值下的 *TPR* 和 *FPR*. 将 *TPR* 作为纵坐标, *FPR* 作为横坐标, 就可以画出一条 ROC 曲线, 计算这条曲线下面积就能得到 AUC 值.

3.5 实验结果分析

由于使用 SMOTE 算法需要寻找 *k* 近邻, 所以 *k* 就是一个超参数, 我们首先对参数 *k* 进行实验, 结果如图 3 所示, 当 *k* 等于 5 的时候模型准确率最高.

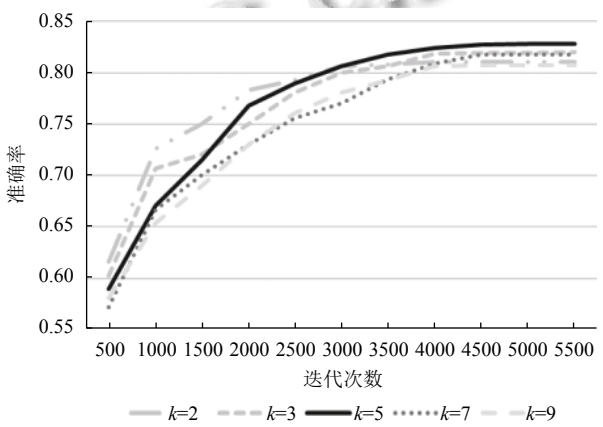

图 3 不同 *k* 值下的准确率

模型的最终实验结果对比如表 2 所示. 单一的 LightGBM 模型在准确率上优于其他模型, 但是由于数据不平衡问题, AUC 值和 *F*1 值表现不佳. 而使用了采样技术的 LightGBM 模型 3 个指标都高于其他模型, 在模型精度上最优.

表 2 实验结果

算法	AUC	准确率	*F*1值
采样+LightGBM	0.9297	0.8281	0.8271
LightGBM	0.9066	0.8171	0.8198
孤立森林	0.9078	0.8011	0.8027
聚类	0.8416	0.7523	0.7407

4 结论与展望

在本文中, 我们应用了一种基于采样技术和 LightGBM 的模型来解决用户用电异常检测问题 该模型首先使用 SMOTE 算法对少数类样本进行过采样, 从而使得正负样本数量平衡, 然后使用 TomekLinks 算法删除掉不同类别距离十分接近的样本, 把采样好的数据输入到 LightGBM 模型中进行二分类训练, 最终输出的预测结果采用不同的评价指标进行评估. 本文采用的模型在用户用电异常检测问题上, 具有高精确率和速度快的优势. 由于模型中使用的特征没有进行详细的筛选, 所以模型还有很大的提升空间, 下一步我们将根据不同特征对模型的重要性进行模型优化, 对特征空间进行更加细致的处理, 进一步提升模型的准确率. 并且目前模型主要是检测用电数据是否异常, 没有分析产生异常的原因, 所以下一步工作还将对异常用电数据进行分类, 确定产生异常的原因.

参考文献

1 张荣昌. 基于数据挖掘的用电数据异常的分析与研究 [硕士学位论文]. 北京: 北京交通大学, 2017.
2 张小秋, 周超, 徐晴. 基于逻辑回归的增量式异常用电行为检测方法. 科学技术与工程, 2019, 19(29): 144–149. [doi: 10.3969/j.issn.1671-1815.2019.29.023]
3 张颖, 王琳, 王丽华, 等. 基于最小二乘法和聚类的用电数据异常分析算法. 河北电力技术, 2019, 38(5): 4–6, 9. [doi: 10.3969/j.issn.1001-9898.2019.05.003]
4 郝方舟, 孙奇珍, 沈超, 等. 基于高维随机矩阵的配电网用户侧用电行为分析. 广东电力, 2019, 32(11): 111–119. [doi: 10.3969/j.issn.1007-290X.2019.011.014]
5 林女贵, 洪兰秀, 黄道姗, 等. 基于改进深度自编码网络的异常用电行为辨识. 中国电力, 2020, 53(6): 18–26.
6 魏志强, 张浩, 陈龙. 一种采用 SmoteTomek 和 LightGBM 算法的 Web 异常检测模型. 小型微型计算机系统, 2020, 41(3): 587–592. [doi: 10.3969/j.issn.1000-1220.2020.03.024]