

DOI: 10.19783/j.cnki.pspc.201098

基于混合重抽样和 LightGBM 算法的配变低压跳闸预测

吴琼, 李荣琳, 洪海生, 罗锋, 黄锦增, 陆颖文

(广州供电局有限公司, 广东 广州 510620)

摘要: 针对配变台区在夏季用电高峰期易频繁跳闸的问题, 提出一种基于混合重抽样和 LightGBM 算法的配变低压跳闸预测模型。为了解决数据分布的边缘化问题, 首先采用隔离森林剔除样本中的离群值。其次采用 NCL 欠抽样与 SMOTE 过抽样相结合的混合重抽样方法处理训练样本的数据不平衡问题。然后采用混合重抽样算法产生的新样本对 LightGBM 分类器进行训练。最后利用训练好的模型对目标台区低压跳闸进行预测。通过算例仿真表明, 对比其他预测模型, 所提 iF-SMOTE-NCL-LightGBM 模型在低压跳闸预测中的各项评价指标均达到最高, 能有效预测低压跳闸事件。

关键词: 配变台区; LightGBM 算法; 混合重抽样; 隔离森林; 低压跳闸预测

Low-voltage tripping prediction of a distribution transformer based on hybrid resampling and a LightGBM algorithm

WU Qiong, LI Ronglin, HONG Haisheng, LUO Feng, HUANG Jinzeng, LU Haowen

(Guangzhou Power Supply Bureau Co., Ltd., Guangzhou 510620, China)

Abstract: There are frequent tripping faults in the distribution transformation area during the summer peak period. A low-voltage trip prediction model based on a hybrid resampling method and the LightGBM algorithm is proposed. First, an isolation forest is used to eliminate outliers in the samples to solve the problem of data distribution marginalization. Secondly, a mixed resampling method combining NCL under-sampling and SMOTE over-sampling is used to handle the data imbalance of training samples. Thirdly, the LightGBM classifier is trained by the new samples generated by the hybrid resampling algorithm. Finally, the probability of low-voltage tripping faults in the target station area is predicted by the well-trained classifier. The experimental results show that the proposed iF-SMOTE-NCL-LightGBM model achieves the highest performance evaluation indicators, among other prediction models, in low-voltage trip prediction, and can effectively predict low-voltage tripping events.

This work is supported by the Science and Technology Project of China Southern Power Grid Co., Ltd. (No. GZJKJXM20170049).

Key words: distribution transformation area; LightGBM algorithm; hybrid resampling; isolation forest; low-voltage tripping prediction

0 引言

随着经济发展和人们生活水平的日益提高, 用电水平也不断增长。尤其在夏季高温天气时段用电负荷大幅攀升, 部分配变会出现过负载的现象, 台区故障跳闸问题相应发生^[1], 由此也引发了众多供电类客服抱怨。目前, 相关部门对于配变低压跳闸的处理办法以更换开关、调整负荷、增容改造、公改

专等事后处理方式为主, 缺乏预判性的工作^[2]。因此, 开展低压跳闸的事先预测研究, 对于及时发现配变安全隐患, 提高供电可靠性与服务质量具有重要的意义, 有助于提高低压配电网的智能化水平^[3]。

目前国内外学者已针对配变台区运行状态展开了研究: 文献[4]针对重过载管理中存在的难题和业务空白, 提出了基于数据挖掘技术的重过载关联分析并实现了事前预测; 文献[5]基于数据统计技术分析行业类别、用电类别、外界天气等因素与配变重过载的关系, 并探究了各个因素的相对重要性; 文

基金项目: 南方电网公司科技项目资助(GZJKJXM20170049)

献[6]提出了采用重抽样技术处理重过载预测问题中正负样本数据不平衡问题,结合随机森林算法对重过载状态、程度及重过载时间进行预测,取得了较高的预测精度;文献[7]提出了一种春节期间配变负荷预测方法,通过最终预测值和配变的额定参数对比,进而对配变重过载情况进行预测;文献[8]针对重过载频繁发生的高速发展地区,从历史负荷、天气、设备台账和用户性质等数据出发,提出了一种基于逻辑回归的中长期重过载预警模型。

由此可见,当前研究多数以配变重过载预测为切入点。配变重过载是引起低压跳闸的主要原因之一,但仍有大量低压跳闸是由管理因素和设备因素引发^[9]。因此重过载预测结果本身不能确保对低压跳闸进行判断的准确率。

由于低压跳闸在日常配变运行中属于少数类,故原始数据为不平衡样本,需提前进行处理。当前有多种抽样方法处理不平衡样本,如随机过抽样(ROS)、随机欠抽样(RUS)、SMOTE^[10]过抽样以及领域清理(NCL)等^[11]。然而,采用单一的过抽样或欠抽样算法都可能会造成少数类样本过度拟合或信息丢失问题^[12-14]。

此外,配变台区数据由于人工记录时存在许多问题,导致记录存在异常值。而异常值会影响预测精度,为跳闸报警带来困难^[15]。因此,必须剔除样本中的异常数据,真实地反映跳闸信息。目前异常值检测的方法有:基于距离的检测方法、统计学方法^[16]、集成学习模型^[17]、神经网络^[18-19]等。

针对上述问题,本文以配变低压跳闸事件作为预测目标,提出一种基于混合重抽样和 LightGBM 算法的配变低压跳闸预测模型。为了解决数据边缘化问题,首先,采用隔离森林(Isolation Forest, IF)剔除数据的异常值。其次,使用 NCL 欠抽样与 SMOTE 过抽样相结合的混合重抽样方法对数据进行处理以解决数据不平衡问题。然后,使用新样本对 LightGBM 分类器进行训练,最后,根据训练好的分类器进行预测。算例结果验证了所提方法的有效性和准确性。

1 不平衡样本处理方法

针对样本集中离群值影响数据分布的问题,首先采用隔离森林剔除异常值。在过抽样算法中,当由边界的样本与其他样本进行过采样时,很容易生成一些噪音数据;在欠抽样算法中,可能造成多数类样本信息丢失。因此,为了解决上述问题,本文提出一种 NCL 欠抽样与 SMOTE 过抽样相结合的混合重抽样方法。

1.1 Isolation Forest 异常值检测

隔离森林算法是决策树算法基于隔离划分思想的一种扩展应用^[20]。与其他树集成方法一样,孤立森林建立在决策树的基础之上。在树中,通过首先随机选择特征然后在所选特征的最小值和最大值之间选择随机分割值来创建分区。最后通过计算建立分区的路径长度 $l(x)$ 判断是否为异常值, $l(x)$ 越小的数据越容易被隔离。其定义如式(1)所示。

$$s(x, n) = 2^{-\frac{E(l(x))}{c(n)}} \quad (1)$$

式中: $l(x)$ 为观测 x 的路径长度; $c(n)$ 为二叉搜索树中未成功搜索的平均路径长度; n 为外部节点数。每个样本都有一个异常得分:

- 1) $s(x, n)$ 接近1表示异常;
- 2) $s(x, n)$ 远小于 0.5 表示正常观察。

1.2 NCL 欠抽样

NCL 欠抽样的基本思想是保留所有少数类样本,并对其邻域上存在的多数类样本进行清理^[21]。该算法有以下两个步骤:(1)遍历训练样本集中的每个样本,找出其三个最邻近样本,若任意样本 x 属于多数类且其三个最邻近样本中至少两个是少数类,则 x 被认定为噪音数据,将其清理;(2)若样本 x 属于少数类且其三个最邻近样本中至少两个是多数类,则清理邻近样本中的多数类样本。

1.3 SMOTE 过抽样

SMOTE过抽样方法是Chawla等人提出的一种解决不平衡数据问题的过抽样方法^[22],其主要思想是通过随机线性插值方法在两个邻近的样本之间合成新样本,从而获得指定数目的样本。合成样本 X_{new} 计算方式如式(2)所示。

$$X_{\text{new}} = x + \text{rand}(0,1) \times (x' - x) \quad (2)$$

式中: $\text{rand}(0,1)$ 表示区间(0,1)的一个随机数; x 为任意少数类样本; x' 为 x 的 k 个最邻近样本中的随机一个。本文中 k 默认设置为5,过抽样数目为 N 。

1.4 混合重抽样

在进行混合重抽样之前,需要确认经重抽样形成的新数据集中少数类和多数类样本的目标比例,并由此计算过抽样和欠抽样添加和移除的数目^[23-24]。假设一个不平衡数据集中少数类样本的数量为 n_1 ,多数类样本的数量为 n_0 ,经重抽样形成的新数据集中少数类与多数类样本的目标占比为 $k_1:k_0$,过抽样和欠抽样需要添加和移除的样本数目 N 可由式(3)计算得到。

$$N = \text{round}[(k_1 \cdot n_0) - (k_0 \cdot n_1)] \quad (3)$$

式中, round 表示通过四舍五入对计算结果取整;

其中 $k_1 + k_0 = 1$, 因此仅需要设定参数 k_1 即可控制混合重抽样目标比例, 为确保新数据集中多数类样本数量不低于少数类和 N 值不为负数, k_1 的上限值设定为 0.5, 下限值为原始数据集中的少数类占比: $n_1 / (n_1 + n_0)$ 。

2 轻梯度提升机模型

GBDT(Gradient Boosting Decision Tree)是机器学习中一个长盛不衰的模型, 其主要思想是利用决策树^[25]迭代训练以得到最优模型, 该模型具有训练效果好、不易过拟合等优点。GBDT 不仅在工业界应用广泛, 通常被用于多分类、点击率预测、搜索排序等任务。LightGBM(Light Gradient Boosting Machine)是一个实现 GBDT 算法的框架^[26], 支持高效率的并行训练, 并且具有更快的训练速度、更低的内存消耗、更好的准确率, 支持分布式可以快速处理海量数据等优点。

LightGBM 的实现如下。

1) 每一次迭代的目的是获得一个弱学习器 $h_t(x)$, 使迭代的损失函数 $L(y, F_t(x))$ 最小。

$$L(y, F_t(x)) = L(y, F_{t-1}(x) + h_t(x)) \quad (4)$$

式中, $F_{t-1}(x)$ 和 $L(y, F_{t-1}(x))$ 分别为上一次迭代获得的强学习器和损失函数。

2) 利用式(4)的负梯度拟合本次迭代损失的近似值。

$$r_{ii} = \frac{\partial L(y_i, F_{t-1}(x_i))}{\partial F_{t-1}(x_i)} \quad (5)$$

3) 使用平方差近似拟合 $h_t(x)$ 。

$$h_t(x) = \arg \min_{h \in H} \sum_{i \in H} (r_{ii} - h(x))^2 \quad (6)$$

4) 本次迭代获得的强学习器如式(7)。

$$F_t(x) = h_t(x) + F_{t-1}(x) \quad (7)$$

3 配变低压跳闸预测模型

3.1 预测模型介绍

综合前两节对预测模型以及数据处理方法的介绍, 本文所提出的配变低压跳闸预测模型如图 1 所示。

模型具体实现步骤如下:

1) 将原始的数据样本分为训练集和测试集;

2) 采用本文所提 iF-SMOTE-NCL 不平衡样本处理方法训练样本进行异常值剔除和混合重抽样, 从而获取新的平衡训练样本;

3) 使用平衡训练样本对 LightGBM 进行训练, 获得训练好的分类器;

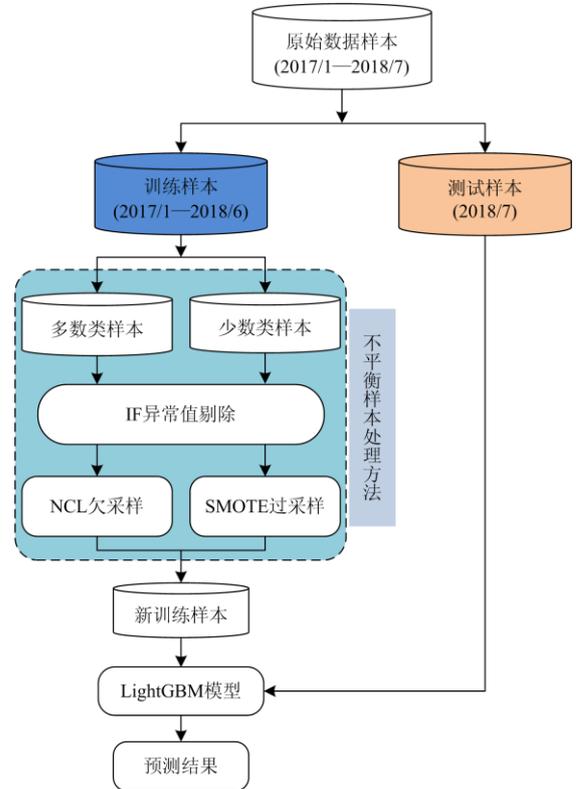


图 1 配变低压跳闸预测模型

Fig. 1 Prediction model of low voltage tripping in distribution transformer

4) 将测试样本代入分类器模型中进行跳闸预测, 最终获得各配变的跳闸预测结果。

3.2 模型性能评价

针对不平衡数据, 常用的评价方法大多建立在混淆矩阵的基础上^[27-28]。如表1所示, TP 和 TN 表示正确分类的少数类和多数类的数量, FP 和 FN 表示错误分类的少数类和多数类的数量。在本文中, 多数类表示配变正常运行, 少数类表示配变发生跳闸。

表 1 二元分类混淆矩阵

状态	少数类	多数类
	(预测发生跳闸)	(预测正常运行)
少数类(实际发生跳闸)	TP	FN
多数类(实际正常运行)	FP	TN

本文选用以下5项模型性能评价指标对两个标签的预测结果分别进行评价。

1) 查全率(Recall)表示正确预测的少数类个数占实际少数类总数的比例, 如式(8)所示。

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

在本文中, 模型的查全率指标越高, 说明模型

覆盖的实际跳闸样本数最多。

2) 查准率(Precision)表示正确预测的少数类占测试集中所有少数类的比例, 如式(9)所示。

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

本文中模型的查准率指标越高, 说明模型能正确预测出越多的跳闸事件。

3) F1-score是查准率和查全率的调和平均数, 如式(10)所示。

$$F1-score = \frac{2recall \times precision}{recall + precision} \quad (10)$$

使用该指标衡量模型对不平衡数据集整体分类性能, F1-score越高, 说明模型预测性能越好。

4) ROC-AUC, ROC为接收者操作特征曲线(Receiver Operating Characteristic Curve), 是由TPR和FPR描绘的, 但计算ROC曲线下的面积AUC(Area Under the ROC Curve)作为衡量分类器好坏的度量更为直观。同时AUC在数据不平衡时对评估模型的泛化能力很有参考价值。

5) 马修斯相关系数(Matthews Correlation Coefficient, MCC)是用来描述分类器的预测性能, 当取值的绝对值越大表示分类器预测能力越强, 如式(11)所示。

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (11)$$

4 算例分析

本文以广州市某地区在2017年—2018年7月, 共19个月的实际采集数据验证所提预测方法。数据采集点为每月1个点, 观测对象为583个发生过低压跳闸事件的配变台区, 实验样本数量共11 077个, 去除存在数据质量问题的样本, 剩余7 152个数据齐全且质量较好的样本, 其中有872个低压跳闸事件和143个供电类客服抱怨事件(事件的计算方法以每月为一个点统计, 如: 当月发生了低压跳闸事件3次, 则记为3, 表示该月发生了低压跳闸事件)。

本文所选取的特征变量包含配变的设备台账数据、供电环境属性数据以及台区用户特征数据, 可分为静态特征、时序特征两大类。

1) 静态特征包含: 配变容量、低压用户数、投运时长、用电性质(包含居民、工业、商业、综合四类)、地区特征(包含城区、乡镇)、当日是否越限。

2) 时序特征: 负载率、月平均最高温度、月平均最低温度、节假日信息。

具体的特征描述与相应的预处理方式如表2所

示。对表2中的各项数据采用相应的处理方式进行处理后, 均作为特征变量输入模型。

表2 特征选取与数据预处理

特征名	特征类型	处理方式
配变容量	数值型	分箱
配变低压用户数	数值型	分箱
配变投运时长	数值型	分箱
月平均最高温度	数值型	标准化
月平均最低温度	数值型	标准化
月最高负载率	数值型	标准化
季节	字符型	编码
用电性质	字符型	编码
地区特征	字符型	编码
当月是否越限	字符型	编码

实验以2017年—2018年6月的数据作为模型训练样本, 以2018年7月的数据作为测试样本对预测模型性能进行测试。预测模型中隔离森林剔除的离群值参数及混合重抽样参数根据相应实验的最优结果进行设置。

4.1 数据处理环节参数设置

数据处理环节有两个参数需要确定, 经过混合重采样处理后少数类样本目标占比 k_1 , 隔离森林剔除的离群值数量占全部新样本集的比值 K 。本节实验中, 设置参数 k_1 和 K 的取值分别为0.01~0.06和0.3~0.6, 以ROC-AUC值为评价指标, 以Logistic回归LR为基础分类器进行实验, 进而确定参数 k_1 和 K 的最佳取值大小。实验结果如表3—表6及图2所示。

表3 不同K值下ROC-AUC值的变化($k_1=0.3$)

K	0.01	0.02	0.03	0.04	0.05	0.06
ROC-AUC	0.903	0.911	0.903	0.910	0.929	0.900

表4 不同K值下ROC-AUC值的变化($k_1=0.4$)

K	0.01	0.02	0.03	0.04	0.05	0.06
ROC-AUC	0.944	0.941	0.924	0.922	0.943	0.925

表5 不同K值下ROC-AUC值的变化($k_1=0.5$)

K	0.01	0.02	0.03	0.04	0.05	0.06
ROC-AUC	0.934	0.921	0.912	0.928	0.950	0.908

表6 不同K值下ROC-AUC值的变化($k_1=0.6$)

K	0.01	0.02	0.03	0.04	0.05	0.06
ROC-AUC	0.902	0.900	0.898	0.910	0.912	0.876

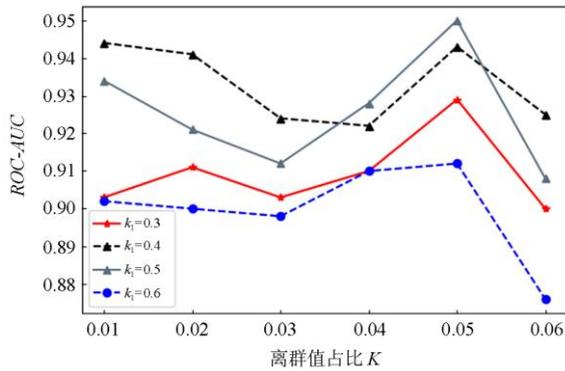


图 2 不同 k_1 和 K 参数预测效果 $ROC-AUC$ 评价对比
Fig. 2 $ROC-AUC$ evaluation comparison with different k_1 and K values in prediction

从图 2 可以直观地看出, 随着 k_1 的增大, AUC 的值变大, 预测效果逐步变好。而当 k_1 继续增大到 0.6 时, AUC 下降, 说明过采样添加了过多的合成样本, 造成信息冗余和导致模型过拟合, 影响预测结果。当 k_1 达到 0.5 时 AUC 值最大, 为 0.950, 证明当数据集趋于平衡时, 预测效果最好。同时也证明本文所提出的混合重抽样算法的有效性。而离群值占比 K 设置为 0.05 时, 预测结果最好。

因此, 在接下来的实验中, 将参数 k_1 和 K 分别设置为 0.5 和 0.05。

4.2 iF-SMOTE-NCL 有效性验证

为了验证所提出混合重抽样算法的有效性, 本节采用: (1) 单一抽样算法 SMOTE 与 NCL; (2) 混合抽样算法 SMOTETomek、SMOTEENN、SMOTE-NCL; (3) 本文所提出混合抽样算法的 iF-SMOTE-NCL 对原始数据进行处理。用 Logistic 回归 LR 进行预测, 结果如表 7 及图 3、图 4 所示。其中 Raw 一列为原始数据不做任何处理, 直接用 LR 进行预测。

表 7 不同预处理算法的预测结果

Table 7 Forecast results of different preprocessing algorithms

预处理算法	Raw	SMOTE	NCL	SMOTETomek
$F1-score$	0.868	0.914	0.815	0.914
$recall$	0.767	0.842	0.933	0.842
$precision$	1.000	1.000	0.723	1.000
$ROC-AUC$	0.883	0.921	0.920	0.921
MCC	0.850	0.899	0.769	0.899

预处理算法	SMOTEENN	SMOTE-NCL	iF-SMOTE-NCL
$F1-score$	0.744	0.919	0.933
$recall$	0.883	0.850	0.875
$precision$	0.642	1.000	1.000
$ROC-AUC$	0.878	0.925	0.938
MCC	0.678	0.905	0.921

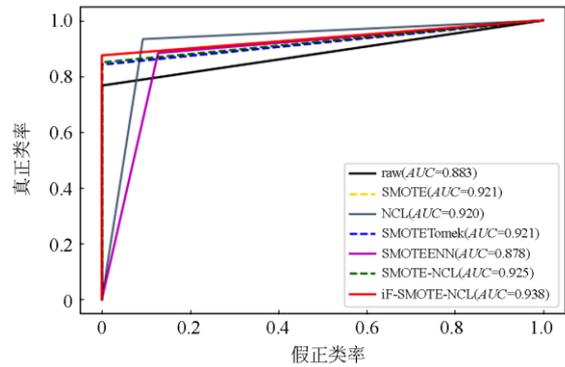


图 3 不同预处理算法的 ROC 曲线与 AUC 对比
Fig. 3 Comparison of the ROC curves and AUC in different preprocessing algorithms

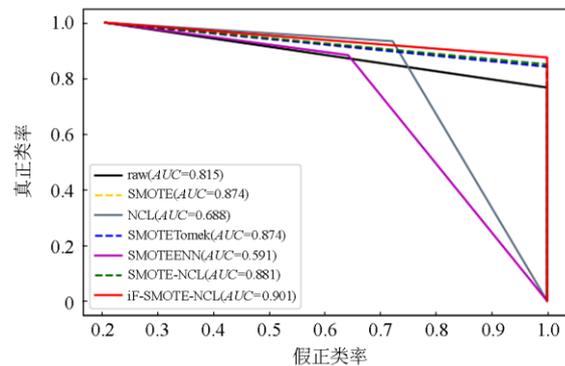


图 4 不同预处理算法的 PR 曲线与 AUC 对比
Fig. 4 Comparison of the PR curves and AUC in different preprocessing algorithms

根据表 7 及图 3、图 4, 可以得出以下结论:

1) 单一的过抽样算法可以提升整体预测精度。由于原来的数据极度不平衡, 分类器对低压跳闸事件的识别能力较差, 但经过 SMOTE 过抽样算法对少数类数据合成新数据, 使得数据基本达到均衡, 获得较好的预测效果。其中 $F1-score$ 提升了 5.30%, $recall$ 提升了 9.78%, $ROC-AUC$ 提升了 4.30%, MCC 提升了 5.76%。

2) 单一的欠抽样算法可以提升低压跳闸事件的识别率, 但整体预测精度下降。NCL 通过清理多数类样本使数据达到平衡, 但是丢失了多数类样本的大量信息, 导致对正常的事件识别率下降。其中 $F1-score$ 下降了 6.11%, $precision$ 下降了 27.70%, MCC 下降了 9.53%。

3) 混合抽样算法通过对样本清洗可以进一步提升预测精度, 其中本文所提出的 SMOTE-NCL 算法预测效果最好。对比 SMOTE 预测结果, 所提算法的 $F1-score$ 提升了 0.55%, $recall$ 提升了 0.95%, $ROC-AUC$ 提升了 0.43%, MCC 提升了 0.67%。

4) 通过隔离森林剔除离群值可以有效提升预测精度。与SMOTE-NCL比较, iF-SMOTE-NCL的 *F1-score* 提升了 1.52%, *recall* 提升了 2.94%, *ROC-AUC*提升了 1.41%, *MCC*提升了 1.77%。

4.3 LightGBM有效性验证

本文以 2017 年—2018 年 6 月的数据作为模型训练样本,以 2018 年 7 月的数据作为测试样本分别对以下 5 种预测模型进行对比分析: (1) Logistic 回归模型(LR); (2) 支持向量机(SVM); (3) 随机森林(RF); (4) 梯度提升决策树模型(XGBoost); (5) 轻梯度提升机(LightGBM)。预处理算法采用 iF-SMOTE-NCL, 各模型预测结果如表 8 和图 5、图 6。

表 8 不同分类器的预测结果

Table 8 Forecast results of different classifiers

分类器	LR	SVM	RF	XGBoost	LightGBM
<i>F1-score</i>	0.933	0.862	0.781	0.948	0.992
<i>recall</i>	0.875	0.858	0.967	0.983	0.983
<i>precision</i>	1.000	0.866	0.655	0.915	1.000
<i>ROC-AUC</i>	0.938	0.912	0.917	0.980	0.992
<i>MCC</i>	0.921	0.826	0.734	0.935	0.989

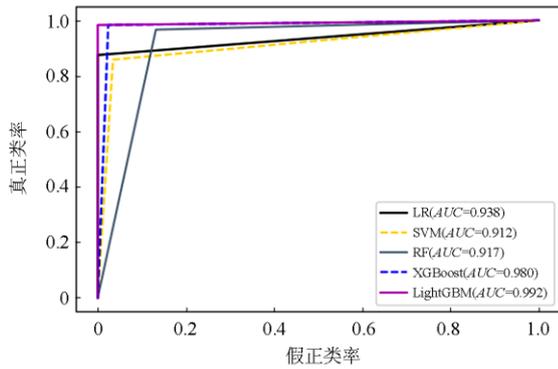


图 5 不同分类器的 ROC 曲线与 AUC 对比

Fig. 5 Comparison of the ROC curves and AUC in different classifiers

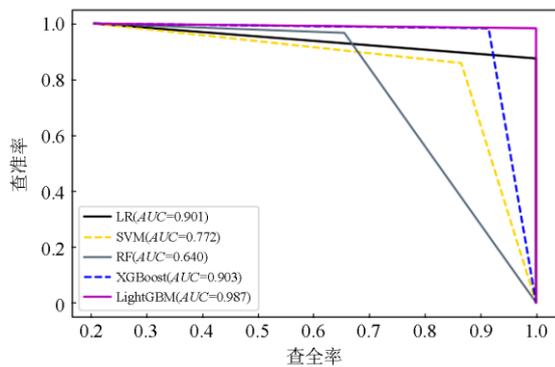


图 6 不同分类器的 PR 曲线与 AUC 对比

Fig. 6 Comparison of the PR curves and AUC in different classifiers

通过对比表 8 和图 5、图 6 各模型评价指标可以看出:

1) 集成学习算法预测结果优于传统的机器学习算法。集成学习算法通过组合许多分类器产生一个强的分类器组,然后通过投票决定获得分类结果,具有更强的鲁棒性。然而,RF 的整体预测结果较差,但其对低压跳闸识别的准确率有所提升。因此,直接采用传统的随机森林模型不能有效预测低压跳闸事件。

2) 对比其他 4 种预测模型,本文所提 iF-SMOTE-NCL-LightGBM 模型在低压跳闸预测中的各种评价均达到最高,说明该模型在少数类识别和整体分类性能上要优于其他方法,更好地预测出低压跳闸事件。同时,与 XGBoost 模型相比,引入带深度限制的 Leaf-wise 叶子生长策略及直方图作差加速之后,可以进一步提高低压跳闸预测模型的性能。该地区在 2018 年 6 月发生了低压跳闸事件的配变共有 120 个,采用本文所提 iF-SMOTE-NCL-LightGBM 模型对该月进行预测,预测结果如图 7 所示。由该图可见,在 120 个实际发生跳闸的配变中共有 118 个被正确预测,预测准确率(即模型对少数类样本的召回率)达 98.33%。

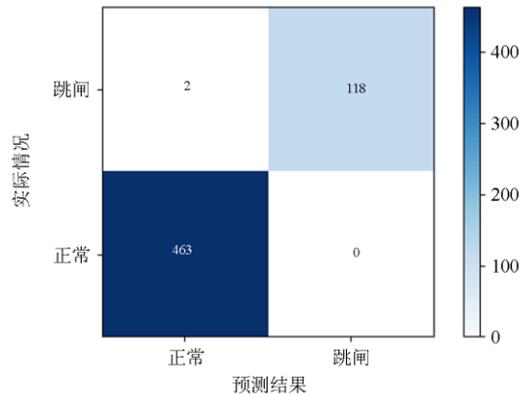


图 7 LightGBM 预测的混淆矩阵

Fig. 7 Predicted confusion matrix of LightGBM

5 结论

针对配变在夏季频繁出现低压跳闸故障问题,本文提出一种基于混合重抽样和 LightGBM 算法的配变低压跳闸预测模型。通过使用隔离森林可以剔除异常值避免数据分布边缘化。为了合理改善训练样本的不平衡性,提出 NCL-SMOTE 混合重抽样方法使分类器对少数类样本的识别能力进一步增强。同时使用 LightGBM 分类器进行分类,有效利用了各个变量与低压跳闸之间的相关性,使模型的性能进一步优化。

参考文献

- [1] 夏海峰, 刘一民, 董希建, 等. 稳定控制装置特高压交流线路无故障跳闸判据的改进[J]. 电力系统保护与控制, 2019, 47(22): 105-112.
XIA Haifeng, LIU Yimin, DONG Xijian, et al. Improvement of non-fault UHV AC line trip criterion in power system stability control equipment[J]. Power System Protection and Control, 2019, 47(22): 105-112.
- [2] 魏瑞增, 李敏, 王彤, 等. 雷暴天气识别及其在电网运行中的应用[J]. 电网与清洁能源, 2019, 35(9): 12-24.
WEI Ruizeng, LI Min, WANG Tong, et al. Thunderstorm weather identification and its application in power grid operation[J]. Power System and Clean Energy, 2019, 35(9): 12-24.
- [3] 李博, 高志远. 人工智能技术在智能电网中的应用分析和展望[J]. 中国电力, 2017, 50(12): 136-140.
LI Bo, GAO Zhiyuan. Analysis and prospect on application of artificial intelligence technologies in smart grid[J]. Electric Power, 2017, 50(12): 136-140.
- [4] 张国宾, 王晓蓉, 邓春宇. 基于关联分析与机器学习的配网台区重过载预测方法[J]. 大数据, 2018, 4(1): 105-116.
ZHANG Guobin, WANG Xiaorong, DENG Chunyu. Heavy overload prediction method for distribution network based on association analysis and machine learning[J]. Big Data, 2018, 4(1): 105-116.
- [5] 贺建章, 王海波, 季知祥, 等. 面向智能电网的配电变压器重过载影响因素分析[J]. 电网技术, 2017, 41(1): 279-284.
HE Jianzhang, WANG Haibo, JI Zhixiang, et al. Analysis of factors affecting distribution transformer overload in smart grid[J]. Power System Technology, 2017, 41(1): 279-284.
- [6] 贺建章, 王海波, 季知祥, 等. 基于随机森林理论的配电变压器重过载预测[J]. 电网技术, 2017, 41(8): 2593-2597.
HE Jianzhang, WANG Haibo, JI Zhixiang, et al. Heavy overload forecasting of distribution transformers based on random forest theory[J]. Power System Technology, 2017, 41(8): 2593-2597.
- [7] 史常凯, 闫文棋, 张筱慧, 等. 基于 BP 网络和灰色模型的春节配变重过载预测[J]. 电力科学与技术学报, 2016, 31(3): 140-145.
SHI Changkai, YAN Wenqi, ZHANG Xiaohui, et al. Heavy overload forecasting of distribution transformer during the spring festival based on BP network and grey model[J]. Journal of Electric Power Science and Technology, 2016, 31(3): 140-145.
- [8] LI M, ZHOU Q. Distribution transformer mid-term heavy load and overload pre-warning based on logistic regression[C] // 2015 IEEE Eindhoven PowerTech, June 29- July 2, 2015, Eindhoven, Netherlands.
- [9] 李成翔, 郭琦, 朱益华, 等. 新东直流送端被动孤岛运行的逻辑判别优化[J]. 电网与清洁能源, 2019, 35(9): 1-6, 11.
LI Chengxiang, GUO Qi, ZHU Yihua, et al. Optimization of distinguishing logic of xinsong-dongfang HVDC in passive isolated island operation[J]. Power System and Clean Energy, 2019, 35(9): 1-6, 11.
- [10] DOUZAS G, BACAO F. Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE[J]. Information Sciences, 2019, 501.
- [11] 孙腾飞, 程浩忠, 张立波, 等. 基于改进混合抽样与最小切负荷计算的电力系统可靠性评估方法[J]. 电力系统保护与控制, 2016, 44(13): 96-103.
SUN Tengfei, CHENG Haozhong, ZHANG Libo, et al. Power system reliability assessment method based on improved mixed sampling and minimum load shedding[J]. Power System Protection and Control, 2016, 44(13): 96-103.
- [12] YUAN X M, YANG M, YANG Y. An Ensemble Classifier based on structural support vector machine for imbalanced data[J]. Pattern Recognition & Artificial Intelligence, 2013, 26(3): 315-320.
- [13] 熊冰妍, 王国胤, 邓维斌. 基于样本权重的不平衡数据欠抽样方法[J]. 计算机研究与发展, 2016, 53(11): 2613-2622.
XIONG Bingyan, WANG Guoyin, DENG Weibin. Unbalanced data undersampling method based on sample weight[J]. Computer Research and Development, 2016, 53(11): 2613-2622.
- [14] 牛壮, 李凤莲, 张雪英, 等. 改进欠抽样方法及其在非平衡数据集分类中的应用[J]. 计算机工程, 2019, 45(6): 2018-2024.
NIU Zhuang, LI Fenglian, ZHANG Xueying, et al. Improved undersampling method and its application in classification of unbalanced data sets[J]. Computer Engineering, 2019, 45(6): 2018-2024.
- [15] 樊帅, 唐群先. 基于 AdaBoost-SAMME 的风力发电机组变桨异常识别系统[J]. 电力系统保护与控制, 2020, 48(21): 31-40.
FAN Shuai, TANG Qunxian. Wind turbine pitch anomaly recognition system based on AdaBoost-SAMME[J]. Power System Protection and Control, 2020, 48(21): 31-40.
- [16] 朱倩雯, 叶林, 赵永宁, 等. 风电场输出功率异常数据识别与重构方法研究[J]. 电力系统保护与控制, 2015,

- 43(3): 38-45.
ZHU Qianwen, YE Lin, ZHAO Yongning, et al. Methods for elimination and reconstruction of abnormal power data in wind farms[J]. Power System Protection and Control, 2015, 43(3): 38-45.
- [17] 连超, 李华, 刘亚, 等. 一种基于 DBN-LR 集成学习的异常检测模型[J]. 小型微型计算机系统, 2019, 40(12): 2637-2643.
LIAN Chao, LI Hua, LIU Ya, et al. Anomaly detection model based on DBN-LR integrated learning[J]. Journal of Chinese Computer Systems, 2019, 40(12): 2637-2643.
- [18] 向玲, 邓泽奇, 赵玥. 基于 SCADA 数据的风电机组异常识别方法[J]. 太阳能学报, 2020, 41(11): 278-284.
XIANG Ling, DENG Zeqi, ZHAO Yue. Anomaly recognition method for wind turbines based on SCADA data[J]. Acta Energetica Solaris Sinica, 2020, 41(11): 278-284.
- [19] MAJD AA, SAMET H, GHANBARI T. k-NN based fault detection and classification methods for power transmission systems[J]. Protection and Control of Modern Power Systems, 2017, 2(4): 359-369. DOI: 10.1186/s41601-017-0063-z.
- [20] LIN Z, LIU X, COLLU M. Wind power prediction based on high-frequency SCADA data along with isolation forest and deep learning neural networks[J]. International Journal of Electrical Power and Energy Systems, 2020, 118.
- [21] 牛壮, 李凤莲, 张雪英, 等. 改进欠抽样方法及其在非平衡数据集分类中的应用[J]. 计算机工程, 2019, 45(6): 218-224.
NIU Zhuang, LI Fenglian, ZHANG Xueying, et al. Improved under-sampling method and its application in the classification of imbalanced data sets[J]. Computer Engineering, 2019, 45(6): 218-224.
- [22] 石洪波, 陈雨文, 陈鑫. SMOTE 过采样及其改进算法研究综述[J]. 智能系统学报, 2019, 14(6): 1073-1083.
SHI Hongbo, CHEN Yuwen, CHEN Xin. Summary of research on SMOTE oversampling and its improved algorithms[J]. CAAI Transactions on Intelligent Systems, 2019, 14(6): 1073-1083.
- [23] 高锋, 黄海燕. 基于邻域混合抽样和动态集成的不平衡数据分类方法[J]. 计算机科学, 2017, 44(8): 225-229.
GAO Feng, HUANG Haiyan. Imbalanced data classification method based on neighborhood hybrid sampling and dynamic ensemble[J]. Computer Science, 2017, 44(8): 225-229.
- [24] 尚旭. 不平衡数据集的混合采样方法[J]. 数字技术与应用, 2016(12): 68-71.
SHANG Xu. Hybrid sampling method for unbalanced data sets[J]. Digital Technology and Application, 2016(12): 68-71.
- [25] KIRANMAI S A, LAXMI A J. Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy[J]. Protection and Control of Modern Power Systems, 2018, 3(3): 303-314. DOI: 10.1186/s41601-018-0103-3.
- [26] 许立武, 李开成, 罗奕, 等. 基于不完全 S 变换与梯度提升树的电能质量复合扰动识别[J]. 电力系统保护与控制, 2019, 47(6): 24-31.
XU Liwu, LI Kaicheng, LUO Yi, et al. Classification of complex power quality disturbances based on incomplete S-transform and gradient boosting decision tree[J]. Power System Protection and Control, 2019, 47(6): 24-31.
- [27] ANDREAS H, PETER R, HOLGER S, et al. Confusion flow: a model-agnostic visualization for temporal analysis of classifier confusion[J]. IEEE Transactions on Visualization and Computer Graphics, 2020.
- [28] 宋亚飞, 王晓丹, 雷蕾, 等. 基于证据理论和混淆矩阵的传感器可靠性评估[J]. 控制与决策, 2015, 30(6): 1111-1115.
SONG Yafei, WANG Xiaodan, LEI Lei, et al. Evaluating dynamic reliability of sensors based on evidence theory and confusion matrix[J]. Control and Decision, 2015, 30(6): 1111-1115.

收稿日期: 2020-09-07; 修回日期: 2020-10-30

作者简介:

吴琼(1972—), 女, 通信作者, 硕士, 高级工程师, 研究方向为企业运营监控、电力数据挖掘; E-mail: xmzy_2015@163.com

李荣琳(1988—), 女, 硕士, 助理工程师, 研究方向为企业运营监控、营配专业协同分析; E-mail: li_ronglin@163.com

洪海生(1984—), 男, 硕士, 工程师, 研究方向为企业运营监控、基建管理及配网自动化. E-mail: honeyhycere@gmail.com

(编辑 许威)