基于数据驱动的配电网停电预测模型

南东亮[1], 冯长有[1,2], 曹 晖[3], 王 昕[4], 李玉敦[4], 谭金龙[1]

(1.国网新疆电力有限公司电力科学研究院, 新疆 乌鲁木齐 830011;
2.国家电力调度控制中心, 北京 100031; 3.西安交通大学电气工程学院, 陕西 西安 710049;
4.国网山东省电力公司电力科学研究院, 山东 济南 250002)

摘要: 停电事故作为影响配电网供电可靠性重要的因素之一,其预测的准确性将给整个电力系统的可靠性带来积极影响。本文提出了一种基于数据驱动的配电网停电预测模型,能够有效地预测停电事故的发生。该模型首先采用一种基于 K-means 聚类的停电数据集欠采样方法降低原始数据集的不平衡比;然后在此基础上,提出了一种改进的 Adaboost 集成学习算法,在每次权值更新时,通过使用已经训练的弱分类下的分类误差进行权重更新,用于对后面的弱分类器进行训练,进而改善分类性能。某地区的实际数据测试结果表明,本文提出的基于数据驱动的配电网停电预测模型能够有效地预测配电网停电事故的发生,相比于传统预测方法具有更好的精确度、召回率、F1 值,停电预测性能得到明显提高。

关键词: 配电网; 数据驱动; 停电预测; 供电可靠性

DOI: 10.12067/ATEEE2104053 **文章编号**: 1003-3076(2021)12-0056-08 **中图分类号**: TM72

1 引言

随着我国智能电网建设的不断深入,各类状态监测设备和传感器得到广泛应用,从而积累了大量的数据[1],使得传统的数据挖掘算法无法满足实际业务需求,落后的数据分析处理能力与数据快速增长之间的矛盾突显。与此同时,大数据技术的发展为我们提供了新的思路和方法[2-5]。

配电网作为整个电力系统的终端,是整个系统与用户关联的重要组成部分。一旦配电网发生停电事故,将导致用户的电力供应出现问题,造成巨大的经济损失[6]。为了提高电力系统的可靠性,减少用户的停电投诉量,提前做好供电服务,国内外很多学者针对停电预测问题都开展了大量的研究工作。特别是随着计算机科学技术与物联网技术的迅速发展,针对停电预测问题的评估和预测获得了飞速的进展,研究方向包括对停电的影响因素、评估指标和评估模型等方面[7-9]。一些学者通过分析电力设备故障停电机理,提出相应的电力设备停电故障分析方法及其相应的概率预测模型[10]。文献[11]提出一种台风灾害下输电线路损毁的预警方法,采用正态分布函数及极值Ⅰ型分布函数对输电线路风荷载概率分布函数进行拟合,并基于应力强度干涉模型计算输电线路损毁概率,然而模型考虑的影响因素过于单一,实际中难以推广。因此,针对极端天气下电力设备停电问题,文献[12]除了考虑气象因素外,还考虑用户数、杆塔树、线路等电网因素,以及土壤含水量、经纬度、地表等地理因素,建立基于随机森林的用户停电区域预测评估方法,但在实际使用中均存在模型训练时间过长,缺乏实际数据验证等问题。文献[13]采用 XGBoost 算法,通过建立天气特征与配电故障线路数的预测模型,对各类天气对电力线路停电故障影响规律进行研究,但模型存在预测准确度低的问题。此外,文献[14]提出建立贝叶斯网络预测飓风情况下的电力停电概率,文献[15]考虑了电力设备下面的植被情况和雷达检测

收稿日期: 2021-04-22

基金项目: 国家重点研发计划(2016YFB0901100)

作者简介: 南东亮(1985-), 男, 新疆籍, 高级工程师, 硕士, 研究方向为电力系统运行控制与继电保护、数据分析与挖掘;

曹 晖(1978-), 男, 河南籍, 教授, 博士, 研究方向为电力大数据与人工智能(通信作者)。

数据,通过建立随机森林模型,来提高停电预测结果的准确性,但这些预测主要集中在极端环境,与实际配电网停电影响因素相距甚远。

针对目前配电网停电预测不准、难以实际推广的难题,本文提出基于数据驱动的配电网停电预测模型。配电网停电数据集是典型的不平衡数据集,即停电类数据在总数据集中所占比例非常小,针对不平衡数据集问题,本文从数据采样算法层面,提出一种基于 K-means 聚类的停电数据集欠采样方法。经实验证明,和传统的随机欠采样方法相比,本文提出的欠采样方法具有更好的采样效果,能够有效地解决数据集的不平衡问题。为提高配电网停电预测的准确性,本文提出了一种改进的 Adaboost 集成学习算法。实验结果表明,和传统的 Adaboost 算法相比,基于改进后的 Adaboost 算法的配电网停电预测模型具有更好的精确度、召回率、F1 值(精确率和召回率的调和均值),停电预测性能得到明显提高。

图 1 为基于数据诊断的跳闸预测模型的示意图。

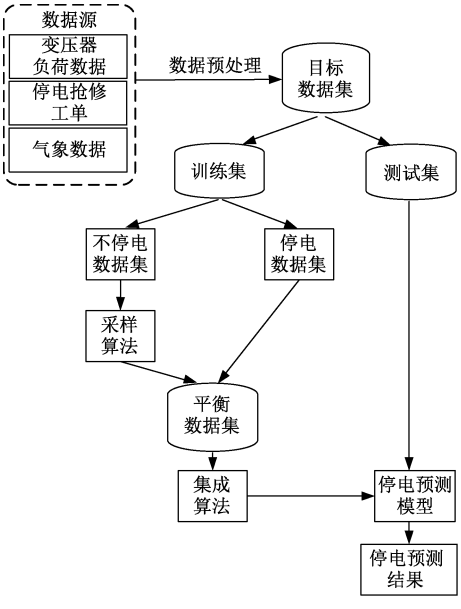


图 1　基于数据诊断的跳闸预测模型

Fig.1　Trip prediction model based on data diagnosis

2　数据集欠采样模型

配电网停电预测是典型的不平衡数据分类问题,和不停电类相比,停电类所占的比例是非常小的,即原始停电数据集样本分布极不平衡。在对不平衡数据集进行分类时,分类器的分类准确率比平衡数据集的分类准确率要低很多,造成这种结果的很大一部分原因是由于正类样本和少类样本的数据分布不均匀,导致分类器将正类(少数类)误判为噪声。这导致模型的预测性能较差,特别是针对样本较少类别的预测。针对不平衡数据分类问题,常用的方法之一就是对不平衡数据集进行采样使数据集达到平衡,进而提高分类器的分类准确率。欠采样是解决数据不平衡问题的主要方式之一。

然而,随机欠采样方法在采样过程中具有很大的随机性,没有充分考虑正负样本的数据分布,导致多类样本的数据会被过多删除。为了减缓随机欠采样导致过多删除重要信息的问题,D.L.Wilson 等提出了最近邻规则(Edited Nearest Neighbor,ENN)来解决该问题,然而此方法对多数类样本的删除十分有限。因此,J. Laurlkkal 等则提出了基于最近邻领域清理规则(Neighborhood CLeaning rule,NCL)[16]。领域清理规则方法的不足之处是对样本中出现的噪声没有进行考虑。为了克服随机欠采样的缺点,很多学者也提出了其他的欠采样方法,如 B.Sun 等提出的进化欠采样方法[17]、基于 K-近邻算法的欠采样技术[18]、基于遗传算法的欠采样方法(Genetic Algorithm based Under-Sampling,GAUS)[19] 等。其中,有研究学者考虑到数据的时空分布特性,为了减少多数类的相似样本点,提出利用聚类的方法对不平衡数据集进行欠采样[20]。聚类算法可以根据数据样本的空间结构信息,将数据集在其特征空间中进行聚类,得到一个最优划分,所以本文也采用基于聚类算法的欠采样技术对数据集中的多数类样本(不停电类)进行欠采样。

针对预测过程中由于类间不平衡导致少数类预测结果准确度低的问题,为了克服传统随机欠采样方法导致信息丢失的不足,本文提出了基于K-means聚类的停电数据集欠采样方法。该方法主要有两个过程:第一个过程是利用 K-means 聚类方法对多数类数据集(不停电的数据集)进行聚类,将数据集划分为 K 个簇;第二个过程是在每个簇中按照密度分布进行随机欠采样,具体根据每个簇中数据方差的大小进行排序,最先对方差小的簇以一定的采样率进行随机欠采样,将采样后的多类数据集和少类数据集相结合得到新的平衡数据集。

基于 K-means 聚类的采样可以避免在某个数据分布上删除过多的信息,防止出现欠采样不均匀导致数据失真的情况。其算法流程如下:

步骤 1:对原始配变数据集进行数据预处理,包

括基于邻近算法(K-Nearest Neighbor,KNN)的缺失值处理和基于分解检验异常检测算法(Standard Template Library Extreme Studentized Deviate,STL-ESD)的异常值处理,然后进行特征选择后得到带有停电和不停电标签的特征数据集 D:

$$D = \{(x_1,y_1),\cdots,(x_N,y_N),x_i \in \mathbf{R},y_i \in \{0,1\}\} \tag{1}$$

式中,x_i 为原始数据集;y_i 为分类标签。

步骤2:将特征数据集 D 进行切分,选取其中的80%为训练数据集 D_1,20%为测试数据集 D_2。

步骤3:对训练数据集 D_1 中的不停电数据集 T_0 进行聚类,随机选取 k 个聚类质心点为:

$$\mu_1,\mu_2,\cdots,\mu_k \in \mathbf{R} \tag{2}$$

式中,μ 为随机选择的数据点。

步骤4:对于 $(x_i,y_i) \in T_0$,计算所属簇。

$$c^{(x_i)} = \arg\max_j \| x_i - \mu_j \|^2 \quad j = 1,2,\cdots,k \tag{3}$$

式中,$c^{(x_i)}$ 为数据 (x_i,y_i) 所属的簇;j 为聚类中心点。

步骤5:对于每一个簇 j,重新计算该簇的质心。

$$\mu'_j = \frac{\sum_{(x_i,y_i)\in T_0} \mathrm{I}(c^{(x_i)} = j)x_i}{\sum_{(x_i,y_i)\in T_0} \mathrm{I}(c^{(x_i)} = j)} \tag{4}$$

步骤6:计算簇心最大移动距离。

$$d = \max \| \mu'_i - \mu_j \|_2 \tag{5}$$

式中,d 为数据之间的距离。若 ε 为预设距离,且 $d > \varepsilon$,则更新 $\mu_j = \mu'_j$,跳到步骤7执行。

步骤7:对上述聚类结果的每一簇按照比例 α 进行随机欠采样,最终得到平衡训练数据集 D'_0。

3 基于集成学习算法的配电网停电预测方法

作为常用的机器学习算法之一,集成学习算法(Adaboost)是一种迭代算法,其核心思想是针对同一个训练集训练不同的分类器(弱分类器),然后把这些弱分类器集合起来,构成一个更强的最终分类器(强分类器),从而提高算法的整体性能。

作为一种强分类器,Adaboost 主要是将多个弱分类器进行训练来划分类别,同时通过投票的方式决定样本的类别。因此,Adaboost 算法的分类精度主要也是通过弱分类器之间的互补关系进行提高。和单个弱分类器对比,基于 Adaboost 算法的分类性能在一定程度得到了提高,然而也存在明显的缺点:在样本权重更新时,Adaboost 算法是将全部分类正确(或错误)的样本同等看待,并且迭代更新的权值只由上一次的训练结果决定。同时,一味降低分类样本的权值、提高错分样本的权值,容易导致噪声样本权值无限增大,从而使非噪声样本被选中的概率降低,最终的分类准确率也可能随之降低。

针对 Adaboost 权重更新只关注上一个弱分类器所带来的问题,本文提出了一种改进权值更新的 Adaboost 集成算法,在每次权值更新的时候,都考虑已经训练的弱分类下的分类误差,根据已训练的弱分类器对数据集的综合误差评估结果对训练数据集进行权值分布的更新。具体实施过程如下:

假设初始训练集为:

$$\begin{aligned} D &= \{(x_1,y_1),\cdots,(x_N,y_N),\ x_i \in X,y_i \in Y\} \\ &= (-1,+1) \end{aligned} \tag{6}$$

式中,$i = 1,2,\cdots,N$,共有 N 个样本。

(1)初始化训练样本的权重分布:一开始将数据集中所有样本权重设置为 $1/N$,得到初始权值向量 $\boldsymbol{\omega}$。

(2)对于第 $t(t=1,2,\cdots,T)$ 次训练:

1)T 为迭代次数,也是需要训练的弱分类器的数量。按照训练数据集的权重分布进行数据采样,有放回地随机抽取 N 个训练样本,得到的训练数据将服从权重分布,将其作为第 t 个弱分类器 $G_t(x)$ 的训练集。

2)利用采样得到的训练集 D_t 训练得到弱分类器 $G_t(x)$。

3)分别计算 $G_t(x)$ 的分类错误率 ε_t 和权重 α_t 为:

$$\varepsilon_t = P[G_t(x_i \neq y_i)] = \sum_{i=1}^{N} \omega_{ti}\mathrm{I}[G_t(x_i \leqslant y_i)] \tag{7}$$

$$\alpha_t = \frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t} \tag{8}$$

式中,P 为正确分类的概率;$\boldsymbol{\omega}$ 为初始权值向量;ω_{ti} 为第 t 次训练初值权重;$G_t(x)$ 为第 t 个弱分类器的训练集。

4)如果 $\varepsilon_t < 0.5$,则返回步骤1,重新训练 $G_t(x)$。

5)更新样本权重,根据以下公式统计第 n 个样本在前 t 个弱分类器的组合下能正确分类的概率。

$$p_t(n) = \frac{\sum_{K=1}^{t} \mathrm{I}[G_K(x_n = y_n)]\alpha_K}{\sum_{K=1}^{t}\alpha_K} \tag{9}$$

式中,α_K 为基分类器 G_K 线性组合中的权值系数,该系数反映了该基分类器在最终强分类器中的重要程度。根据 $P_t(n)$ 计算第 n 个样本第 $t+1$ 次的权值 $\omega_{t+1}(n)$,前 t 次的分类准确率越低,权值提升越大,计算方式如下:

$$\omega_{t+1}(n)=\frac{\omega_t(n)}{Z_t}\mathrm{e}^{-P_t(n)} \tag{10}$$

其中,归一化因子 Z_t 的计算公式为:

$$Z_t=\sum_{n=1}^{N}\omega_t(n)\mathrm{e}^{-P_t(n)} \tag{11}$$

(3)返回训练阶段得到的 T 个弱分类集合 $G=\{G_1(x),G_2(x),\cdots,G_T(x)\}$。

4　实验结果与分析

4.1　不平衡数据集采样实验及结果分析

结合基于 K-means 聚类的停电数据集欠采样方法算法过程,具体的实验流程如图 2 所示。首先,将配变数据集划分为训练集和测试集,利用K-means对训练集中的多数类进行聚类,然后进行随机欠采样,采样完成后得到新的平衡训练集;接着利用训练集对决策树(Decision Tree, DT)分类算法进行模型训练,最后利用测试集对上述训练得到的模型进行测试,利用相应的指标对其性能进行评测。

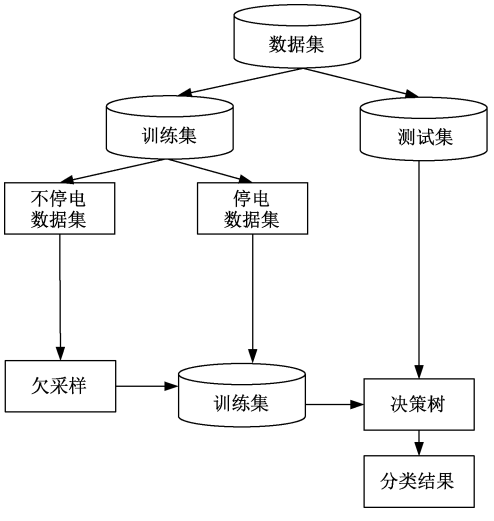


图 2　基于 K-means 聚类的停电数据集欠采样实验流程

Fig.2　Flow chart of under-sampling experiment of power outage data set based on K-means clustering

为了表述采样结果的好坏,需要一个分类器对其进行评估。本文选用 DT 作为分类器,并主要采用召回率对欠采样后分类器性能进行评估。召回率表征的是模型对停电的识别能力。召回率指的是预测正确的停电数据占实际中停电数据的比例,召回率越高,说明真实停电数据在该模型下得到了很好的识别,也表明了模型对停电的识别能力越强。

表 1 为基于同一个决策树分类器 DT 不进行欠采样、随机欠采样(绝对平衡 1∶1)和基于 K-means 聚类欠采样(绝对平衡 1∶1)后的召回率、精确率和 F1 值。可以看出,不采样+DT、随机欠采样+DT 和基于 K-means 欠采样+DT 采得出的结果精确率和 F1 值都比较低,所以本文主要以召回率作为性能评价指标来评价欠采样的结果。如果不进行欠采样就直接进行分类,召回率为 0,说明由于数据集不平衡性太大导致分类器不能识别出停电类;使用欠采样技术将数据集的不平衡性降到 1∶1,即绝对平衡后,采用随机欠采样进行分类后召回率为 0.60,说明通过欠采样使数据达到平衡后,可以提高分类器对停电类的识别能力;使用基于K-means聚类欠采样方法使数据集的不平衡性降到 1∶1 后,分类结果的召回率为 0.90,说明本文提出的基于 K-means 聚类欠采样方法比传统的随机欠采样方法更好,得出的平衡数据集可以保持原始数据分布,减少重要信息丢失,进而提高分类器的识别性能。

表 1　基于 DT 不同采样方法后的召回率

Tab.1　Recall rate based on different sampling methods of DT

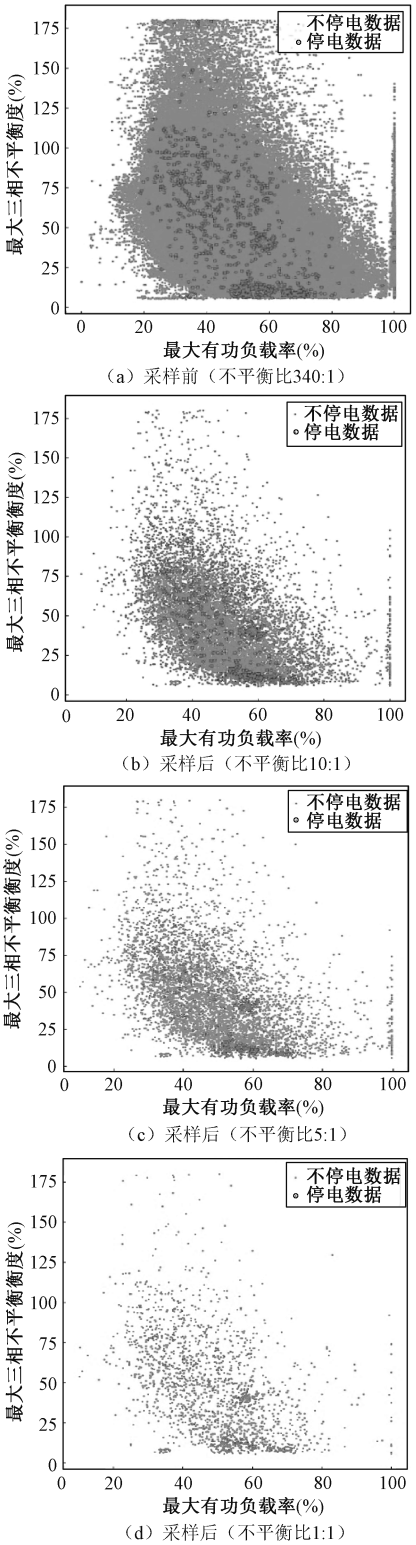
采样方法	不采样+DT	随机欠采样+DT	基于 K-means 欠采样+DT
召回率	0	0.60	0.90
精确率	0.01	0.01	0.01
F1 值	0.01	0.01	0.01

图 3 为使用基于 K-means 欠采样前后数据集的分布情况。图 3(a)是采样前数据集的分布情况(不平衡比为 340∶1);图 3(b)是按一定的采样率后得到不平衡比为 10∶1 的数据集;图 3(c)是按一定的采样率后得到不平衡比为 5∶1 的数据集;图 3(d)是按一定的采样率后得到不平衡比为 1∶1 的数据集。从图 3 也可以看出,基于 K-means 欠采样后数据的分布规则并没有发生改变。

基于 K-means 聚类的欠采样方法不仅可以对原始数据集进行欠采样以实现数据集的平衡性,而且可以选择性地删除数据,保持原始数据集的分布规律,避免重要信息的丢失,进而为后续的分类模型做支撑,提高分类能力。

4.2　停电预测实验及结果分析

停电预测实验主要包含两部分:①进行实验分析改进的 Adaboost 算法在不同采样率(数据不平衡



（a）采样前（不平衡比340:1）

（b）采样后（不平衡比10:1）

（c）采样后（不平衡比5:1）

（d）采样后（不平衡比1:1）

图3 使用基于K-means欠采样前后数据集的分布情况

Fig.3 Distribution of data set before and after K-means under-sampling

比)下的预测性能,并找出最佳的采样率(不平衡比);②将改进的Adaboost算法、决策树和传统的Adaboost作比较,分析各模型的预测性能。具体实验流程:首先基于Embedded方法得到特征子集;然后利用K-means欠采样方法得到有标签的平衡数据集,考虑非停电数据集和停电数据集的不平衡比较大,并通过设置K-means欠采样的取样比,分别得到不平衡比为100∶1,50∶1,10∶1和1∶1的数据集;接着利用采样后的平衡数据集训练决策树模型、Adaboost模型和改进后的Adaboost模型,分别得到不同的停电预测模型;最后利用测试集分别测试各个停电预测模型,并对输出结果进行性能评估。停电预测实验流程如图4所示。

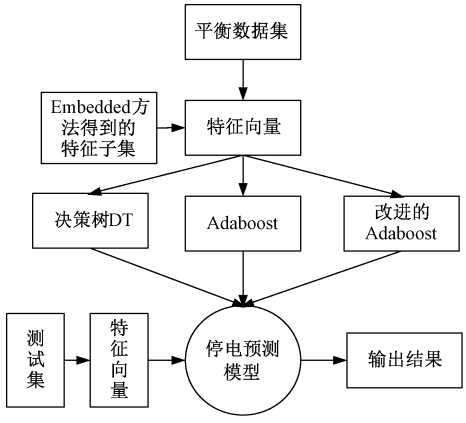


图4 停电预测实验流程

Fig.4 Outage prediction experiment process

各停电预测模型基于不同不平衡比数据集得到的非停电类和停电类的预测结果见表2和表3。从非停电类精确率、召回率和F1值的结果可以看出,决策树、Adaboost模型和改进的Adaboost模型对非停电类的预测结果都很好,其中精确率都在0.99以上,召回率和F1值都在0.8以上。随着数据集不平衡比的变化,模型对非停电类的预测结果变化不是很大,一个主要原因是非停电类在数据集中所占比例较大,另一个原因是各模型对非停电类的识别能力很强。

对于停电类预测结果的性能,从停电类预测的精确率可以看出,随着不平衡比的降低,决策树和Adaboost的精确度逐渐下降,改进的Adaboost算法的精确度从总体上也是下降。其中,改进的Adaboost算法的精确率在数据集不平衡比为50∶1、100∶1、200∶1和340∶1时都高于0.6以上,说明该情况下模型对停电类的区分能力很强;当数据集为绝对平衡时(不平衡比为1∶1),停电预测模型的精确率都比较低,说明欠采样过多,导致信息严重丢失,模型对停电类的区分能力下降。从停电类预测

表 2　各停电预测模型基于不同不平衡比数据集非停电类的各项性能指标

Tab.2　Performance indicators of non-blackout categories of power outage prediction models based on different imbalance ratio data sets

不平衡比	评价指标	决策树	Adaboost	改进的 Adaboost
1	召回率	0.80	0.79	0.80
	精确率	1.00	1.00	1.00
	F1 值	0.85	0.84	0.85
10	召回率	0.90	1.00	1.00
	精确率	1.00	1.00	1.00
	F1 值	0.90	1.00	1.00
50	召回率	1.00	1.00	1.00
	精确率	1.00	1.00	1.00
	F1 值	1.00	1.00	1.00
100	召回率	1.00	1.00	1.00
	精确率	1.00	1.00	1.00
	F1 值	1.00	1.00	1.00
200	召回率	1.00	1.00	1.00
	精确率	1.00	1.00	1.00
	F1 值	1.00	1.00	1.00
340	召回率	1.00	1.00	1.00
	精确率	1.00	1.00	1.00
	F1 值	1.00	1.00	1.00

表 3　各停电预测模型基于不同不平衡比数据集停电类的各项性能指标

Tab.3　Performance indicators of blackout categories of power outage prediction models based on different imbalance ratio data sets

不平衡比	预测方法	决策树	Adaboost	改进的 Adaboost
1	召回率	0.72	0.87	0.83
	精确率	0.01	0.01	0.01
	F1 值	0.02	0.03	0.05
10	召回率	0.70	0.87	0.70
	精确率	0.10	0.30	0.32
	F1 值	0.19	0.43	0.43
50	召回率	0.43	0.69	0.70
	精确率	0.37	0.62	0.70
	F1 值	0.39	0.63	0.70
100	召回率	0.40	0.60	0.70
	精确率	0.40	0.63	0.72
	F1 值	0.39	0.60	0.72
200	召回率	0.25	0.30	0.37
	精确率	0.58	0.80	0.83
	F1 值	0.38	0.40	0.51
340	召回率	0.25	0.27	0.29
	精确率	0.70	0.84	0.70
	F1 值	0.38	0.40	0.40

的召回率可以看出,决策树、Adaboost 和改进的 Adaboost 的召回率随着不平衡比的降低而提高的,改进的 Adaboost 算法的精确率在数据集不平衡比为 1∶1、50∶1 和 100∶1 时都高于 0.7 以上,说明该情况下模型对停电类的识别能力很强,同时也看出精确率和召回率互相矛盾,难以同时达到最大值。

因为 F1 值作为精确率和召回率的综合评价,所以对比不同模型对于相同分类任务的 F1 指标,从停电类预测的 F1 值可以看出,随着不平衡比的下降,各模型的 F1 值先提高后下降,说明绝对平衡时的预测结果并不是最好,当不平衡比为 100∶1 时,各评价模型的性能是最好的,这个时候既降低了不平衡比,又不过多进行欠采样导致丢失更多的重要信息。

不平衡比为 100∶1 时各停电预测模型预测结果性能指标见表 4。从表 4 中可以看出,集成学习方法 Adaboost 对停电预测的精确率、召回率、F1 值和 G-mean 比弱分类器决策树好,说明集成学习方法在处理不平衡数据集上,可以通过对各个弱分类器的互补关系来提高预测性能;而改进的 Adaboost 算法要比传统 Adaboost 的预测结果好,说明本文提出的改进的 Adaboost 算法能够通过改善算法权值更新的方式来有效提高停电预测性能。

表 4　采用不同预测方法的性能指标对比

Tab.4　Comparison of performance indicators using different forecasting methods

预测方法	决策树	Adaboost	改进的 Adaboost
召回率	0.40	0.60	0.71
精确率	0.39	0.60	0.71
F1 值	0.40	0.61	0.71
G-mean	0.63	0.78	0.83

当不平衡比为 100∶1 时各停电预测模型预测结果的接收者操作特征(Receiver Operating Characteristic,ROC)曲线如图 5 所示。从图 5 中可以看出,集成学习 Adaboost 算法的 ROC 曲线在弱分类器决策树的左上方一点,说明 Adaboost 算法的性能比弱分类器决策树的性能好;相比之下可以看到,改进的 Adaboost 算法的 ROC 曲线位于传统 Adaboost 算法和决策树的 ROC 曲线的左上方,说明改进的 Adaboost 算法在不平衡数据的停电预测任务中又比传统的 Adaboost 算法预测结果更好。从 ROC 曲线下面积(Area Under Curve,AUC)可以发现,改进的 Adaboost 的 *AUC* 为 0.92,传统的 Adaboost 的 *AUC* 为

0.9,决策树的 *AUC* 为 0.89,进一步说明改进的 Adaboost 算法用于配电网停电预测时具有更好的性能。

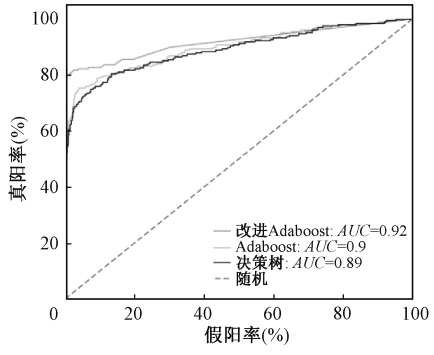


图5 各停电预测模型预测结果的ROC曲线

Fig.5 ROC curve of prediction results of each power outage prediction model

5 结论

本文提出了基于数据驱动的配电网停电预测模型。针对配电网停电数据集不平衡问题,从数据采样层面角度,提出了一种基于 K-means 聚类的停电数据集欠采样方法;针对配电网停电预测问题,构建了基于集成学习算法的停电决策模型,利用强分类器进行停电预测。基于 K-means 聚类的欠采样方法降低了数据集的不平衡比,虽然可能会有部分信息丢失,但仍能有效地实现重要信息的保留。通过实验证明,和传统的随机欠采样方法对比,本文提出的基于 K-means 聚类欠采样方法具有更好的性能,保证了后续的停电预测模型能够在极其不平衡的停电数据中学习到有用的信息,取得良好的预测效果。另外,与传统的预测算法相比,本文提出的算法预测准确度明显强于传统算法,能够为城市配电网的运维提供数据支撑,为检修工作的开展提供一定的理论指导。然而,影响配电网设备停电因素众多,原因错综复杂,且地域性明显,在本文的训练模型中,仅使用了某一地区的实际运行数据进行验证,为采用其他地区的数据进行进一步分析,所以模型的鲁棒性需要进一步验证。

参考文献(References):

[1] 王德文, 周青 (Wang Dewen, Zhou Qing). 一种电力设备状态监测大数据的分布式联机分析处理方法 (A method of distributed on-line analytical processing of status monitoring big data of electric power equipment) [J]. 中国电机工程学报 (Proceedings of the CSEE), 2016, 36 (19): 5111-5121.

[2] 刘峰, 李文敏 (Liu Feng, Li Wenmin). 电力大数据研究综述 (A survey of big data research in power industry) [J]. 电工电能新技术 (Advanced Technology of Electrical Engineering and Energy), 2020, 39 (12): 62-72.

[3] 张稳, 盛万兴, 杜松怀, 等 (Zhang Wen, Sheng Wanxing, Du Songhuai, et al.). 基于海量数据的配电网运行分析系统架构与技术实现 (Architecture and technology implementation of massive data based on distribution network operation analysis system [J]. 电力系统自动化 (Automation of Electricity Power System), 2020, 44 (3): 147-153.

[4] Wang G, Gunasekaran A, Ngai E W T. Distribution network design with big data: Model and analysis [J]. Annals of Operations Research, 2018, 270 (1-2): 539-551.

[5] Sun L, Hu M, Meng Q, et al. The solution for performance improvement of electic distribution network line loss based on Hadoop big data technology [A]. 2015 4th International Conference on Computer Science and Network Technology [C]. Shanghai, China, 2015. 178-182.

[6] 冷喜武, 陈国平, 蒋宇, 等 (Leng Xiwu, Chen Guoping, Jiang Yu, et al.). 智能电网监控运行大数据应用模型构建方法 (Model construction method of big data application for monitoring and control of smart grid) [J]. 电力系统自动化 (Automation of Electricity Power System), 2018, 42 (20): 115-122.

[7] 肖永立, 刘松, 见伟, 等 (Xiao Yongli, Liu Song, Jian Wei, et al.). 智能变电站二次设备多源数据建模与存储方法研究 (Multi-source data modeling and storage method for secondary equipment in intelligent substation) [J]. 计算机应用与软件 (Computer Applications and Software), 2019, 36 (9): 6-11, 126.

[8] 蔡煜, 蔡泽祥, 王奕, 等 (Cai Yu, Cai Zexiang, Wang Yi, et al.). 配电网广域保护控制通信网络建模与组网策略 (Modeling and networking strategy of communication network of wide-area protection and control for distribution network) [J]. 电力自动化设备 (Electric Power Automation Equipment), 2018, 38 (4): 183-190.

[9] Liu Z C, Zou Y P. Research on distribution network operation and control technology based on big data analysis [A]. 2018 China International Conference on Electricity Distribution [C]. Tianjin, China, 2018. 138-145.

[10] 宋国兵, 高淑萍, 蔡新雷, 等 (Song Guobing, Gao Shuping, Cai Xinlei, et al.). 高压直流输电线路继电保护技术综述 (Survey of relay protection technology for HVDC transmission lines) [J]. 电力系统自动化 (Automation of Electricity Power System), 2012, 36 (22):

123-129.

[11] 黄勇, 魏瑞增, 周恩泽, 等 (Huang Yong, Wei Ruizeng, Zhou Enze, et al.). 台风灾害下输电线路损毁预警方法 (Early warning method of transmission line damage under typhoon disaster) [J]. 电力系统自动化 (Automation of Electricity Power System), 2018, 42 (23): 142-150.

[12] 彭寒梅, 郭颖聪, 昌玲, 等 (Peng Hanmei, Guo Yingcong, Chang Ling, et al.). 基于系统短期时序状态转移抽样法的孤岛运行微电网可靠性评估 (Short-term reliability evaluation of islanded microgrid based on system short-term sequential transition sampling) [J]. 电工电能新技术 (Advanced Technology of Electrical Engineering and Energy), 2018, 37 (1): 69-77.

[13] 严道波, 杨勇, 邱丹, 等 (Yan Daobo, Yang Yong, Qiu Dan, et al.). 基于天气因素和 XGBoost 算法的配电线路故障停电预测 (Failure prediction of distribution line based on weather factors and XGBoost algorithm) [J]. 电力与能源 (Power & Energy), 2019, 40 (2): 168-171.

[14] Mensah A F, Dueñas-Osorio L. Outage predictions of electric power systems under hurricane winds by Bayesian networks [A]. 2014 International Conference on Probabilistic Methods Applied to Power Systems [C]. Durham, 2014. 1-6.

[15] Wanik D W, Parent J R, Anagnostou E N, et al. Using vegetation management and LiDAR-derived tree height data to improve outage predictions for electric utilities [J]. Electric Power Systems Research, 2017, 146 (5): 236-245.

[16] Laurikkala J. Improving identification of difficult small classes by balancing class distribution [A]. Conference on AI in Medicine in Europe: Artificial Intelligence Medicine [C]. Cascals, Portugal, 2001. 112-119.

[17] Sun B, Chen H, Wang J, et al. Evolutionary under-sampling based bagging ensemble method for imbalanced data classification [J]. Frontiers of Computer Science, 2018, 12 (2): 331-350.

[18] Beckmann M, Ebecken N, Lima B. A KNN undersampling approach for data balancing [J]. Journal of Intelligent Learning Systems and Applications, 2015, 7 (4): 104-116.

[19] Ha J, Lee J S. A new under-sampling method using genetic algorithm for imbalanced data classification [A]. International Conference on Ubiquitous Information Management & Communication [C]. Amsterdam, Netherlands, 2016. 1-6.

[20] Rayhan F, Ahmed S, Mahbub A, et al. CUSBoost: Cluster-based under-sampling with boosting for imbalanced classification [A]. 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solutions [C]. Bengaluru, India, 2017. 1-5.

Data-driven predictive model of distribution system blackout

NAN Dong-liang[1], FENG Chang-you[1,2], CAO Hui[3], WANG Xin[4], LI Yu-dun[4], TAN Jin-long[1]

(1. State Grid Xinjiang Electric Power Co., Ltd., Electric Power Research Institute, Urumqi 830011, China;
2. National Electric Power Dispatching Control Centre, Beijing 100031, China;
3. School of Electrical Engineering, Xi'an Jiaotong University, Xi'an 710049, China;
4. State Grid Shandong Electric Power Company Electric Power Research Institute, Ji'nan 250002, China)

Abstract: Blackout is one of the most critical influence factors of distribution network reliability, whether to accurately predict in order to take measures, is an import way to improve the reliability of the power system. This paper presents a data-driven predictive model of power system blackout, which could predict the probability of blackout preciously. This paper firstly uses a sampling method based on K-means algorithm to solve the imbalance of the data set. In order to achieve outage prediction in distribution network, the improved integrated learning algorithm-Adaboost is proposed and the performance is improved by the weight update method with considering the error of the weak classifier. The test proves that the model based on the improved Adaboost algorithm has better accuracy, recall, and F1 value as compared with the Adaboost algorithm. The outage prediction performance is largely improved.

Key words: distribution network; data-driven; blackout prediction; power supply reliability