DOI:10.13878/j.cnki.jnuist.2019.04.018

陈凯[1]　南东亮[2]　孙永辉[1]　夏响[1]

基于 XGBoost 算法的电网二次设备缺陷分类研究

摘要

电网二次设备缺陷严重程度的精确判断可为设备的运行和维护提供重要依据.针对电网二次设备缺陷数据特征量多、人为判断难度大、易出错等问题,提出基于 XGBoost(eXtreme Gradient Boosting)的二次设备缺陷分类方法,提高二次设备缺陷分类的准确率.首先,对二次设备历史缺陷数据进行去异常值、编码等一系列预处理工作,并筛选出与设备缺陷相关性高的特征建立特征指标集;然后,利用历史缺陷数据对 XGBoost 模型进行训练和参数寻优;最后,用训练好的分类模型实现二次设备缺陷的准确分类.本文采用某电厂二次设备缺陷数据对所提算法进行算例分析,并与传统分类器(决策树、逻辑回归等)进行比较,结果表明 XGBoost 可以实现对二次设备缺陷程度的精确判断,进而可以很好地辅助检修人员进行设备的维护与管理.

关键词

XGBoost 算法;二次设备;缺陷分类;机器学习

中图分类号 O429

文献标志码 A

0　引言

电网二次设备是智能变电站安全稳定运行的关键设备之一,其运行状态的好坏关乎电网能否可靠供电.近年来,随着科学技术的迅猛发展,电网规模不断扩大,电网中二次设备的数量也发生了跨越式的增长,“设备多,检修人员少”的矛盾给二次设备的运维人员带来了相当大的工作负担,同时也给电网运行带来了风险,二次设备的运维和管控水平亟待提高[1-3].根据国家电网 220 V 及以上电压等级系统继电保护装置缺陷分析报告,2018 年电网二次设备各类缺陷的缺陷率及所占比例如表 1 所示.

表 1　保护装置缺陷按缺陷程度分类统计情况

Table 1　Classification and statistics of protection device defects by defect degree

缺陷分类	缺陷率/(次/(万台·年))	所占比例/%
危急缺陷	80.1	40.73
严重缺陷	71.5	36.35
一般缺陷	45.1	22.92
总计	196.6	100.00

二次设备缺陷率逐年增长,严重影响了电网的稳定运行.随着保护设备数量的增多,缺陷发生时所要记录的相关数据量也随之增大,且各类数据之间有着或多或少的关联性,单凭运维人员的经验无法对缺陷严重程度进行准确的判断.

对于二次设备的状态评估与缺陷分类,一般多采用层次分析法、决策树[4]、C5.0 分类算法[5]、灰色定权聚类[6]、神经网络[7]等方法.文献[8]提出一种基于 Apriori 算法的二次设备缺陷数据挖掘与分析方法,通过建立基于关联规则的二次设备缺陷模型来寻找二次设备的薄弱环节及其诱因.文献[9]提出基于 ANN 的专家系统理论,并在此基础上设计了多模块协同互动的变电站故障信息分析决策系统,实现了变电站故障的自动和辅助决策处理,对电力系统的安全稳定运行有着重要的参考价值.随着大数据时代的到来,基于数据挖掘实现对二次设备的状态评价成为可能.文献[10]将粗糙集与神经网络有效结合,设计了一套改进算法用于将二次设备的基础数据加工处理成状态评价所需的状态量信息.文献[11]采用数据挖掘技术研制了一套保护设备故障信息管理与分析系统,为实现继电保护装置的状态检

收稿日期 2019-06-19

资助项目 国家自然科学基金(61673161);国网新疆电力有限公司电力科学研究院科技项目(SGXJDK00DJJS1900094)

作者简介

陈凯,男,硕士生,研究方向为二次设备状态评估.18305178588@163.com

孙永辉(通信作者),男,博士,教授,主要研究方向为电力系统分析与控制、负荷预测.sunyonghui168@gmail.com

1　河海大学　能源与电气学院,南京,210098

2　新疆大学　电气工程学院,乌鲁木齐,830047

修提供依据,为分析处理电网故障提供决策支持.以上文献通过采用各自的算法与理论,从大量历史数据中挖掘出与二次设备缺陷相关的特征量,进而分析得出设备缺陷成因,但并未对设备缺陷程度进行准确的判断与分类,从而导致设备检修不足或过度检修的问题,减少了设备的使用寿命.

XGBoost 模型不仅在算法准确度上较传统算法表现出色,同时,算法框架的可修正性好,可根据实际问题场景做出有针对性的优化.到目前为止,国内外将 XGBoost 算法应用到电网设备状态评估领域的并不多见,尤其是用来研究二次设备的缺陷分类则更少.基于此,本文提出基于 XGBoost 的二次设备缺陷分类算法,采用某电厂二次设备历史缺陷数据对模型进行训练与测试,并将 XGBoost 与决策树、Adaboost、随机森林、支持向量机等模型的分类效果进行对比.仿真实验结果表明 XGBoost 对设备缺陷的分类性能优于本文所采用的其他对比算法,能够根据二次设备缺陷相关数据实现对其当前状态的准确判断,辅助电网设备检修人员决策.

1 XGBoost 算法原理

XGBoost 全称叫极端梯度提升,是梯度提升机器算法(Gradient Boosting Machine)的扩展[12].其中 Gradient Boosting 属于集成算法的 Boosting 方法中的一种类别.Boosting 方法将多个弱学习器累加起来组成强学习器,进而使目标损失函数达到最小.XGBoost 是 Gradient Boosting Machine 的实现,能自动利用 CPU 多线程进行并行,并对算法加以改进以提高精度.XGBoost 的基学习器既有树(gbtree)又有线性分类器(gblinear),从而得到带 L_1+L_2 惩罚的线性回归或逻辑回归,其损失函数采用二阶泰勒展开,能分布式处理高维稀疏特征,具有高准确度、不易过拟合、可扩展性等特点[13].

XGBoost 的目标函数除损失函数项之外还包含了正则惩罚项将二者结合起来整体求最优解,用于权衡损失函数的下降和模型的复杂程度.正则化项的加入可以降低模型的方差,使得通过训练集学习得到的模型更加简单,从而防止产生过拟合.XGBoost 算法具体推导过程如下:设 $D=\{(x_i,y_i)\}$($|D|=n,x_i\in\mathbf{R}^m,y_i\in\mathbf{R}$)为一个拥有 n 个样本、每个样本有 m 个特征的数据集;x_i 表示第 i 个样本数据树的集成模型通过 K(树的数目)个相加函数来预测最终结果:

$$\hat{y}_i=\phi(x_i)=\sum_{k=1}^{K}f_k(x_i),f_k\in F, \tag{1}$$

其中,$F=\{f(x)=W_{q(x)}\}$($q:\mathbf{R}^m\to T,w\in\mathbf{R}^T$)($q$ 表示将样本实例 $\mathbf{R}^m$ 映射到相应叶索引的结构,T 表示叶子节点的数目,$\mathbf{R}^T$ 为叶子节点权重 w 的空间)代表了一个决策树的函数空间,样本 x_i 和预测值 $\hat{y}_i$ 的函数关系记为 θ;$W_{q(x)}$ 把每一个节点映射成一个值,即 $f(x)$ 的值;f_k 表示第 k 棵树的模型.每一个 f_k 对应着一个独立的树结构 q 和叶子节点的权值,w 为了学习模型中使用的函数集,故定义正则化目标函数如下:

$$\begin{cases}L(\phi)=\sum_i l(\hat{y}_i,y_i)+\sum_k\Omega(f_k),\\ \Omega(f)=\gamma T+\frac{1}{2}\lambda\|w\|^2,\end{cases} \tag{2}$$

其中,l 是一个用来衡量预测值 $\hat{y}_i$ 和真实值 y_i 之间差异的损失函数,Ω 表示模型复杂度的惩罚项,γ 表示叶子数目的正则化参数,用来抑制节点继续向下分裂,λ 表示叶子权重的正则化参数.

传统的梯度提升树(GBDT)在求解的过程中只利用了一阶导数的信息,而 XGBoost 对于损失函数做了二阶的泰勒展开使得求解的精度更高.其第 t 次的损失函数为

$$L^{(t)}=\sum_{i=1}^{n}l(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_t), \tag{3}$$

式中:$L^{(t)}$ 表示第 t 棵树的目标函数;$\hat{y}_i^{(t-1)}$ 表示前 $t-1$ 棵树的输出之和,构成前 $t-1$ 棵树的预测值;f_t 表示第 i 棵树的模型,$f_t(x_i)$ 表示第 t 棵树的输出结果,相加构成最新的预测值.其一阶导数 g_i,二阶导数 h_i 分别为

$$g_i=\frac{\partial l(y_i,\hat{y}_i^{(t-1)})}{\partial\hat{y}_i^{(t-1)}}, \tag{4}$$

$$h_i=\frac{\partial^2 l(y_i,\hat{y}_i^{(t-1)})}{\partial^2\hat{y}_i^{(t-1)}}. \tag{5}$$

将损失函数在 $\hat{y}_i^{(t-1)}$ 处利用泰勒公式展开,在经过 t 次迭代后损失函数变为

$$L^{(t)}=\sum_{i=1}^{n}\left[g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\right]+\Omega(f_t). \tag{6}$$

定义 $I_j=\{i\mid q(x_i)=j\}$ 作为叶子节点 j 的实例集,根据式(6)得:

$$L^{(t)}=\sum_{j=1}^{T}\left[(\sum_{i\in I_j}g_i)w_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)w_j^2\right]+\gamma T, \tag{7}$$

式中:w_j 表示叶子节点 j 的权重.对于固定的决策树的结构 $q(x)$,可以计算得出叶子节点 j 的最优权重 w_j^*,并代回目标函数得:

$$L^{(t)}(q)=-\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j}g_i)^2}{\sum_{i\in I_j}h_i+\lambda}+\gamma T. \tag{8}$$

为了获取最优分割,通过计算分裂后的值减去分裂前的值,从而计算其得到的增益.假设 I_L 和 I_R 分别是划分后左右子树叶子节点的集合,即 $I=I_L\cup I_R$,则划分后损失函数如下:

$$L_{spit}=\frac{1}{2}\left[\frac{(\sum_{i\in I_L}g_i)^2}{\sum_{i\in I_L}h_i+\lambda}+\frac{(\sum_{i\in I_R}g_i)^2}{\sum_{i\in I_R}h_i+\lambda}-\frac{(\sum_{i\in I}g_i)^2}{\sum_{i\in I}h_i+\lambda}\right]-\gamma. \tag{9}$$

2 基于 XGBoost 的二次设备缺陷分类模型

2.1 特征指标集的构建

电网二次设备类型多数量大,具备点多、面广、要素复杂等特点.其指标集的建立,需要从二次设备运行状态实际情况出发.电网二次设备缺陷分类指标集如图 1 所示.本文将与二次设备缺陷有关的特征量分为两大类:类别型和数值型.类别型特征由设备制造厂家、保护类别、设备电压等级等类别型变量组成,为离散值;数值型特征由缺陷累计时间、累计缺陷次数、电源输出电压等数值型变量组成,为连续值.

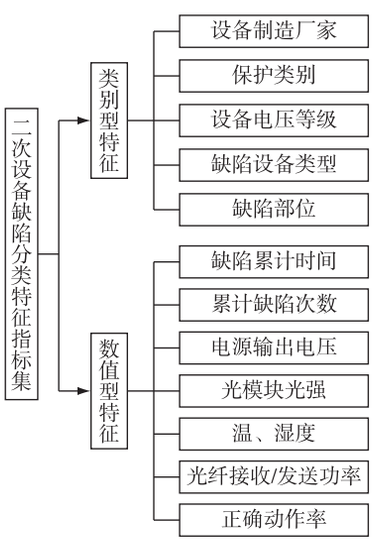

图 1 二次设备缺陷分类特征指标集

Fig. 1 Characteristic indicators set of defect classification for secondary equipment

2.2 特征及标签编码

由于 XGBoost 分类器的输入只能是数值型数据,需要对类别型特征进行相应的编码,将类别型特征转化为数值型特征.目前,常用的编码方式有序列编码(ordinal encoding)、独热编码(one-hot encoding)和二进制编码(binary encoder).

对于输入的类别型特征,由于每个特征中所包含的属性数量不多,本文对此采用独热编码的方式,即使用 0 和 1 表示这些类别型特征,用 N 位状态寄存器来对 N 个状态进行编码,每个状态都有独立的寄存器位,并且在任意时候只有一位有效.

对于缺陷标签编码,依据《国家电网公司继电保护和安全自动装置缺陷管理办法》,二次设备缺陷按严重程度共分为三级:危急缺陷、严重缺陷、一般缺陷.对此三类缺陷等级划分依据如下:

1) 危急缺陷是指继电保护和安全自动装置自身或相关设备及回路存在问题导致失去主要保护功能,直接威胁安全运行并须立即处理的缺陷;

2) 严重缺陷是指继电保护和安全自动装置自身或相关设备及回路存在问题导致部分保护功能缺失或性能下降,但在短时内尚能坚持运行,需尽快处理的缺陷;

3) 一般缺陷是指除上述危急、严重缺陷以外的不直接影响设备安全运行和供电能力,继电保护和安全自动装置功能未受到实质性影响,性质一般、程度较轻,对安全运行影响不大,可暂缓处理的缺陷.

本文对类别型特征中的保护类别及输出标签编码分别如表 2、表 3 所示.

表 2 类别型特征独热编码示例

Table 2 Examples of unithermal coding for categorical features

保护类别	独热编码
变压器保护	[1,0,0,0,0,0,0]
电抗器保护	[0,1,0,0,0,0,0]
电容器保护	[0,0,1,0,0,0,0]
断路器保护	[0,0,0,1,0,0,0]
过电压及远方跳闸保护	[0,0,0,0,1,0,0]
母线保护	[0,0,0,0,0,1,0]
线路保护	[0,0,0,0,0,0,1]

表 3 标签序列编码对照

Table 3 Label sequence coding contrast table

缺陷分类	序列编码
危急缺陷	2
严重缺陷	1
一般缺陷	0

2.3 样本数据分布

本文以某电厂 2016—2018 年间二次设备缺陷记录数据为例,共计 556 条典型缺陷信息,其中:一般缺陷 147 例、严重缺陷 256 例、危机缺陷 153 例.将数据按 4∶1比例随机划分为训练集和测试集,训练集和测试集中各类型缺陷样本个数如表 4 所示.

表 4 样本数据具体分布

Table 4 Specific distribution of sample data

缺陷分类	训练集	测试集
危急缺陷	122	31
严重缺陷	205	51
一般缺陷	118	29

2.4 缺陷分类模型应用步骤及架构

基于 XGBoost 算法的电网二次设备缺陷分类模型如图 2 所示.具体应用步骤如下:

1) 从二次设备缺陷记录中筛选出与设备缺陷相关性大的特征作为分类模型的输入;

2) 对输入数据中的连续型特征进行去除异常值、处理缺失值等预处理操作;

3) 对输入数据中的类别型特征及标签分别进行编码;

4) 按比例将处理好的缺陷记录数据集划分为训练集和测试集;

5) 将训练数据集输入 XGBoost 模型进行训练,通过参数调优,实现模型参数最优化;

6) 利用调优后的模型对测试集数据进行缺陷分类测试.

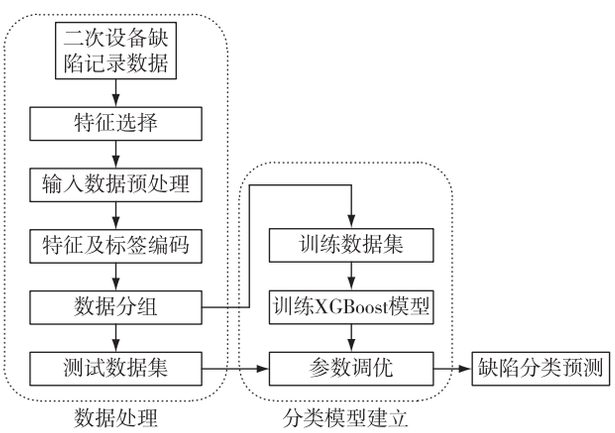

图 2 基于 XGBoost 的电网二次设备缺陷分类模型

Fig. 2 Defect classification model of secondary equipment in power grid based on XGBoost

3 仿真结果与性能比较

3.1 分类模型评估指标

多分类问题类似于二分类问题,能够通过混淆矩阵对模型进行性能评估.根据真实类别与预测类别可以分为真正类(TP,其量值记为 N_{TP})、真负类(TN,其量值记为 N_{TN})、假正类(FP,其量值记为 N_{FP})和假负类(FN,其量值记为 N_{FN}).采用准确率(Precision,其量值记为 P)、召回率(Recall,其量值记为 R)[14] 及 F_1 值 3 个指标测试分类精确度:

$$P=\frac{N_{TP}}{N_{TP}+N_{FP}}\times 100\%, \tag{10}$$

$$R=\frac{N_{TP}}{N_{TP}+N_{FN}}\times 100\%, \tag{11}$$

$$F_1=\frac{2\times P\times R}{P+R}\times 100\%. \tag{12}$$

准确率 P 是预测为正的样本数与所有的实际为正的样本数之比;召回率 R 是预测为正的样本数与该类实际样本数之比;F_1 是综合准确率和召回率考虑的模型分类精确度.通过准确率、召回率、F_1 评估每一类别的分类性能.

因为本文缺陷被划分为 3 类,采用宏平均(Macro-averaging),即对所有类别的准确率、召回率、F_1 取平均值,以评估缺陷分类的总体性能.F_1 值会随着准确率、召回率的提高而提高,值越大,说明模型分类效果越好.其具体计算参照下式:

$$P_{Macro}=\frac{1}{n}\sum_{i=1}^{n}P_i, \tag{13}$$

$$R_{Macro}=\frac{1}{n}\sum_{i=1}^{n}R_i, \tag{14}$$

$$F_{1,Macro}=\frac{2\times P_{Macro}\times R_{Macro}}{P_{Macro}+R_{Macro}}\times 100\%. \tag{15}$$

在实际数据集中经常出现样本不均衡的问题,ROC(Receiver Operating Characteristic Curve)曲线作为模型评估的指标能够在数据集中的正负样本分布变换的时候保持不变,因此本文另采用 ROC 和 AUC 对模型分类性能进行评估.AUC 是 ROC 曲线下面积(Area Under roc Curve)的简称.通常,AUC 的值介于 0.5 到 1.0 之间,AUC 值越大,诊断准确性越高.

3.2 参数设置

XGBoost、决策树、支持向量机等分类算法都有不同的参数需要调节,来使训练出的模型分类效果达到最优,本文使用 Gridsearch 网格搜索法对算法

部分参数进行寻优.Gridsearch 网格搜索法,就是对算法中需要调节的参数取一个范围与步长,通过遍历在其中选出一个或一组最佳值.GridsearchCV 是 Scikit-Learn 算法库中的一个子模块,用于系统地遍历多种参数组合,通过交叉验证确定最佳效果参数,本文所用算法部分参数如表 5 所示.

表 5　算法部分参数设置

Table 5　Arithmetic partial parameter settings

分类模型	部分参数设置
XGBoost	max_depth = 2, learning_rate = 0.1, n_estimators = 100, booster = gbtree
决策树	criterion = gini, min_samples_split = 3, min_samples_leaf = 2
逻辑回归	penalty = l2
多项式贝叶斯	alpha = 0.9
K 近邻	n_neighbors = 5, leaf_size = 30
支持向量机	kernel = rbf, degree = 3, cache_size = 200
Adaboost	n_estimators = 50, learning_rate = 1.2

3.3　模型仿真对比

本文所使用的 XGBoost 分类算法,在经过训练集训练及网格搜索参数寻优后,得出一个分类效果较好的模型.图 3 为测试集混淆矩阵.本文将二次设备缺陷按严重程度划分为 3 类,混淆矩阵下方为模型预测标签,左侧为实际标签.从该混淆矩阵可以清楚地统计出预测标签与实际标签相符的缺陷记录共 103 条,预测标签与实际标签不符的缺陷记录共 9 条,其在测试集上的分类准确率可达 92%.

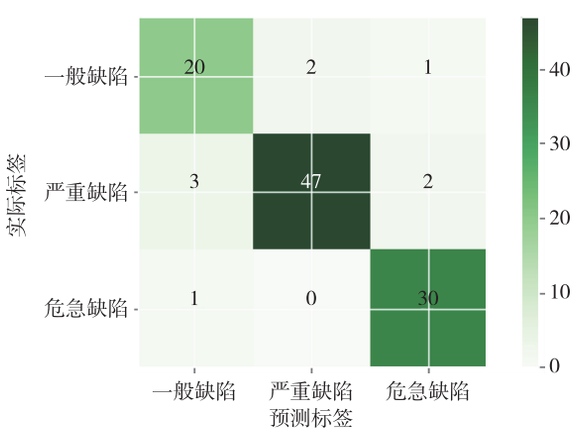

图 3　二次设备缺陷数据测试集混淆矩阵

Fig. 3　Obfuscation matrix of test set for secondary equipment defect data

为了评估 XGBoost 模型在缺陷分类上的性能,本文采用十折交叉验证(10-fold cross validation)的方法,即将训练数据集分成 10 份,轮流将其中 9 份作为训练数据,1 份作为验证数据进行试验.每次试验都会得出相应的评价指标,最终以 10 次评价指标的平均值作为对模型精度的估计,并与决策树、逻辑回归、K 近邻等算法作对比,详细的对比结果如表 6 所示.

表 6　分类模型性能比较

Table 6　Performance comparison of classification models

%

分类模型	P_{Macro}	R_{Macro}	$F_{1,\text{Macro}}$
XGBoost	92.23	93.64	92.85
决策树	84.48	84.92	84.67
逻辑回归	78.96	76.59	77.53
多项式贝叶斯	90.78	88.68	89.55
K 近邻	86.94	89.01	87.4
支持向量机	86.16	85.85	85.93
Adaboost	86.12	87.56	86.73

从表 6 可以看出,基于 XGBoost 的二次设备缺陷分类的各类评价指标值均要优于本文所采用的其他算法,准确率比逻辑回归提高了近 14 个百分点,而与其他较为先进的多项式贝叶斯、K 近邻、Adaboost 相比,也提高了 2~5 个百分点的缺陷分类准确率.由此可见,XGBoost 算法模型具有更强的泛化能力,在二次设备缺陷分类问题上的表现优于其他算法.

最后通过绘制出各种分类算法的 ROC 曲线来对分类器的分类性能做出直观的评判.从图 4 中可以清楚地看到 XGBoost 的 ROC 曲线位于其他算法的上方,说明其分类性能的优越性.而对于本文所使用的其他算法,由于各自的 ROC 曲线有所交叉,不能很好地区分开,需要通过计算 ROC 曲线下方的面积,即模型的 AUC 值来进行比较,AUC 值越大则模型的分类性能越好.

4　结束语

本文采用目前较为流行的 XGBoost 算法对电网二次设备缺陷实现精确分类,通过对分类模型的训练及参数调优,其测试准确率最高可达 93.75%,十折交叉验证平均准确率达 92%,分类性能好.最后与传统的机器学习算法的比较结果,展现了 XGBoost 性能的优越性及良好的泛化能力.基于 XGBoost 的二次设备缺陷分类模型,可以准确有效地判断出当前二次设备的缺陷程度,避免了人为判断的主观性,为二次设备的

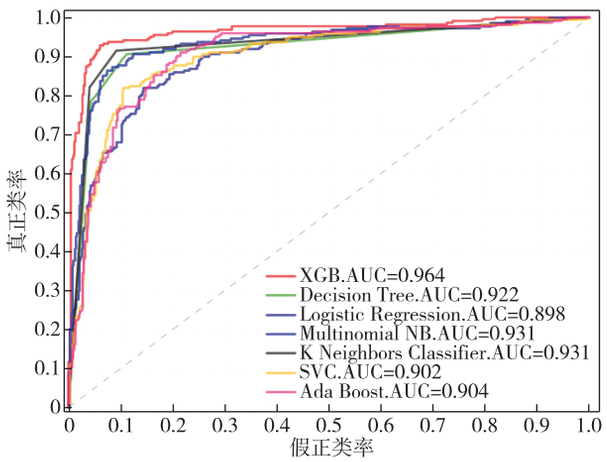

图 4　分类模型 ROC 曲线及对应的 AUC 值

Fig. 4　ROC curve of classification model and corresponding AUC value

维修与管理提供辅助性决策.但由于缺陷记录数据集较小,模型的学习能力还有很大的提升空间.

参考文献

References

[1] 郭创新,陆海波,俞斌,等.电力二次系统安全风险评估研究综述[J].电网技术,2013,37(1):112-118
GUO Chuangxin,LU Haibo,YU Bin,et al.A survey of research on security risk assessment of secondary system [J].Power System Technology,2013,37(1):112-118

[2] 曹楠,王芝茗,李刚,等.智能变电站二次系统动态重构初探[J].电力系统自动化,2014,38(5):113-121
CAO Nan,WANG Zhiming,LI Gang,et al.Study on dynamic reconfiguration in secondary system of intelligent substation[J].Automation of Electric Power Systems,2014,38(5):113-121

[3] 袁浩,屈刚,庄卫金,等.电网二次设备状态监测内容探讨[J].电力系统自动化,2014,38(12):100-106
YUAN Hao,QU Gang,ZHUANG Weijin,et al.Discussion on condition monitoring contents of secondary equipment in power grid[J].Automation of Electric Power Systems,2014,38(12):100-106

[4] 史逸民,史达伟,郝玲,等.基于数据挖掘 CART 算法的区域夏季降水日数分类与预测模型研究[J].南京信息工程大学学报(自然科学版),2018,10(6):760-765
SHI Yimin,SHI Dawei,HAO Ling,et al.Model prediction of regional summer precipitation days based on CART algorithm[J].Journal of Nanjing University of Information Science & Technology(Natural Science Edition),2018,10(6):760-765

[5] 黄志刚,刘虹,刘娟,等.基于 C5.0 算法的胃癌生存预测模型研究[J].南京信息工程大学学报(自然科学版),2017,9(4):406-410
HUANG Zhigang,LIU Hong,LIU Juan,et al.Gastric cancer prediction model based on C5.0 classification algorithm[J].Journal of Nanjing University of Information Science & Technology(Natural Science Edition),2017,9(4):406-410

[6] 徐沛勤,姚天祥.基于灰色定权聚类的江苏省工业节能减排评价研究[J].南京信息工程大学学报(自然科学版),2014,6(4):374-379
XU Peiji,YAO Tianxiang.Research of industrial energy conservation and emission reduction in Jiangsu province based on grey fixed weight cluster[J].Journal of Nanjing University of Information Science & Technology(Natural Science Edition),2014,6(4):374-379

[7] 陈龙龙,王波,袁玲.一种电力变压器神经网络故障诊断方法[J].南京信息工程大学学报(自然科学版),2018,10(2):199-202
CHEN Longlong,WANG Bo,YUAN Ling.A neural network-based method for fault diagnosis of power transformer[J].Journal of Nanjing University of Information Science & Technology(Natural Science Edition),2018,10(2):199-202

[8] 张延旭,胡春潮,黄曙,等.基于 Apriori 算法的二次设备缺陷数据挖掘与分析方法[J].电力系统自动化,2017,41(19):147-151,163
ZHANG Yanxu,HU Chunchao,HUANG Shu,et al.Apriori algorithm based data mining and analysis method for secondary device defects[J].Automation of Electric Power Systems,2017,41(19):147-151,163

[9] 孙金莉,李煜磊,冯凝,等.智能变电站二次设备缺陷分析专家系统的研究与应用[J].电网与清洁能源,2016,32(10):94-98
SUN Jinli,LI Yulei,FENG Ning,et al.Research and application of the expert system for defect analysis of the secondary equipment in smart substations[J].Power System and Clean Energy,2016,32(10):94-98

[10] 王师霜.二次设备状态评价数据挖掘技术的研究与应用[D].保定:华北电力大学,2013
WANG Shishuang.Research and application on data mining technology in secondary device status evaluation [D].Baoding:North China Electric Power University,2013

[11] 李勋,龚庆武,杨群瑛,等.基于数据挖掘技术的保护设备故障信息管理与分析系统[J].电力自动化设备,2011,31(9):88-91
LI Xun,GONG Qingwu,YANG Qunying,et al.Fault information management and analysis system based on data mining technology for relay protection devices[J].Electric Power Automation Equipment,2011,31(9):88-91

[12] Chen T,He T.Higgs boson discovery with boosted trees [J].JMLR:Workshop and Conference Proceedings,2015,42:69-80

[13] 叶倩怡,饶泓,姬名书,等.基于 Xgboost 的商业销售预测[J].南昌大学学报(理科版),2017,41(3):275-281
YE Qianyi,RAO Hong,JI Mingshu,et al.Sales prediction of stores based on xgboost algorithm[J].Journal of Nanchang University(Natural Science),2017,41(3):275-281

[14] 李伟,王丽霞,李广野,等.基于极限梯度提升树的输电线路缺陷风险预报[J].控制工程,2018,25(7):

1172-1178
LI Wei, WANG Lixia, LI Guangye, et al. Prediction of transmission line defects risk based on extreme gradient boosting tree[J]. Control Engineering of China, 2018, 25(7):1172-1178

Defect classification of secondary equipment in power grid based on XGBoost

CHEN Kai[1] NAN Dongliang[2] SUN Yonghui[1] XIA Xiang[1]
1 College of Energy and Electrical Engineering, Hohai University, Nanjing 210098
2 School of Electric Engineering, Xinjiang University, Urumqi 830047

Abstract Accurate determination of the severity of secondary equipment defects in power grid can provide an important basis for the operation and maintenance of equipment. Therefore, in this paper, to address problems such as large quantity of defective data features, and the great difficulty of using error-prone human judgment as an evaluation parameter, a defect classification method based on XGBoost (eXtreme Gradient Boosting) is proposed to improve the accuracy of defect classification of secondary equipment. First, a series of pre-processing work, such as removing outliers and coding, is performed on the secondary equipment historical defect data, and the characteristics highly correlated with equipment defects are extracted to establish the feature index set. Subsequently, the XGBoost model is trained and optimized using historical defect data. Finally, the trained classification model is used to realize the accurate classification of secondary equipment defects. Based on the secondary equipment defective data of a power plant, simulation results are presented to illustrate the effectiveness of the proposed algorithm and are compared with those of traditional classifiers (decision tree, logistic regression, etc.). Simulation results show that XGBoost can accurately determine the defect degree of secondary equipment, to assist the maintenance and management of equipment.

Key words XGBoost; secondary equipment; defect classification; machine learning