基于贝叶斯优化 XGBoost 算法的变压器故障诊断

贾皓阳[1],钱 宇[2]
(1.华北水利水电大学 能源与动力工程学院,河南 郑州 450045;2.河南中孚实业股份有限公司,河南 巩义 451261)

摘 要: 为提升对高能放电等小样本故障诊断的敏感度,提出基于贝叶斯优化极端梯度提升算法(BO-XGBoost)的变压器故障诊断模型。分析了贝叶斯优化 XGBoost 算法的基本原理和基于该算法进行变压器故障诊断的流程,选取 259 组故障样本,探讨了该模型的具体应用,并将其与 XGBoost、支持向量机(SVM)、随机森林(RF)、K 邻近法(KNN)等模型进行对比。结果表明,BO-XGBoost 模型在变压器故障诊断中的精度为 98.08%,比前述模型的诊断精度分别提高了 5.77%、27.42%、22.58%、19.5%。

关键词: 变压器故障诊断;贝叶斯优化算法;XGBoost 算法;油中溶解气体;故障类型;诊断流程;诊断精度;对比分析

中图分类号: TM41 **文献标识码:** A **doi:** 10.13681/j.cnki.cn41-1282/tv.2023.02.008

0 引言

变压器是电力系统的关键设备,其正常运行与生产生活息息相关,一旦发生故障,会带来许多不利影响[1]。因此,准确诊断变压器的故障,并对其进行处理,对电力系统的安全稳定具有重要意义[2]。目前,诊断变压器故障的方法主要是油中溶解气体分析法(dissolved gas analysis,简称 DGA)。该方法以 5 种烃类气体为特征输入量进行故障诊断[3]。随着人工智能技术的发展,一些学者将神经网络、支持向量机(SVM)、随机森林(RF)、深度残差收缩网络等与 DGA 法结合,进行变压器故障诊断,在一定程度上提高了诊断精度[4-7]。

极端梯度提升(Extreme Gradient Boosting,简称 XGBoost)算法是一种可进行并行计算的集成算法,也是一种灵活、高效和便捷的最优分布式算法[8]。国内外部分学者将 XGBoost 算法和智能算法结合起来,应用在设备故障诊断领域。如 WU Zhanhong 将改进的遗传算法与 XGBoost 算法结合,形成混合诊断网络,提高了遗传算法的全局搜索能力,从而也提高了故障诊断精度[9];RAICHURA MAULIK 等提出了卷积神经网络与 XGBoost 的组合算法,解决了故障数据缺少和数据特征单一等问题[10]。这些研究表明,应用智能算法与 XGBoost 算法的组合算法对原始模型进行优化,可以提高模型的诊断精度。

集成学习算法的性能取决于自身参数的选取,分类效果更加依赖自身的参数设定。常见的参数寻优方法有网格搜索法、随机搜索法、贝叶斯优化法等[11]。其中,贝叶斯优化法仅使用较少的计算机资源成本,就可对集成学习的众多超参数进行寻优。目前,贝叶斯优化 XGBoost 算法已在多个领域得到应用,且效果显著,但在变压器故障诊断领域应用较少[12-13]。笔者试对贝叶斯优化 XGBoost 模型在变压器故障诊断中的应用进行探讨,以期供相关技术人员参考。

1 基本理论

1.1 XGBoost 原理

XGBoost 算法是在 Gvadient Boosting 框架下实现的机器学习算法,是以分类回归树(Classification and Regression Tree,简称 CART)和随机森林为基础的一种扩展延伸。Boosting 是将多个弱分类器集成一个准确可靠的集成分类器的方法。XGBoost 训练始于一个常数预测,每次增加一个新的函数学习当前的树,找到当前最佳的树模型,并将其整合到模型中。这样,通过多轮整合,提升模型的准确率[14]。XGBoost 算法的原理如图 1 所示。

对于一个有 n 个样本、m 个特征的数据集,K 棵 CART 的最终预测输出的表达式为式(1)。

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

$$F=\{f(x)=\omega_{q(x)}\},q\cdot R^m\rightarrow T \tag{2}$$

式中:函数 f_k 指独立的树结构;q 为叶子标签;T 为叶子节点个数;F 为决策树结构的集合;$\omega_{q(x)}$ 为模型预测值,也是对样本 x 的打分。

收稿日期: 2023-01-08

作者简介: 贾皓阳 (1997—),男,河南焦作人,硕士研究生,主要从事电力变压器状态诊断及故障检修的学习与研究。

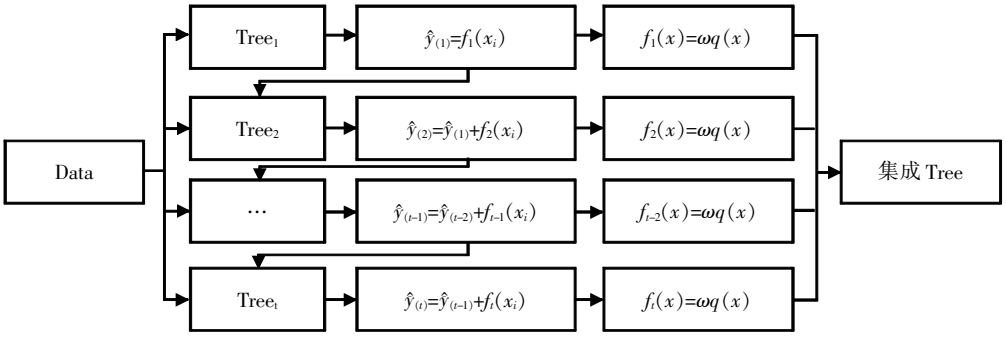

图 1 XGBoost 算法的原理图

Fig.1 Principle of XGBoost algorithm

模型损失函数如式(3)所示。定义 $L^{(t)}$ 的一阶导数 g_i、二阶导数 h_i 分别为式(4)和式(5),并将其带入损失函数 $L^{(t)}$,可得到最终的目标函数,如式(6)所示。

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

式中:$\Omega(f_t)$为正则化惩罚函数。

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

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

$$L^{(t)}=\sum_{i=1}^{n}\left[l(y_i,\hat{y}^{(t-1)})+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\right]+\Omega(f_t)+C \tag{6}$$

式中:C 为常数项。

由式(6)可知,最终的目标函数只依赖于节点在误差函数上的一阶导数和二阶导数。将式(6)中的常数项移除,得到第 t 轮的简化损失函数,如式(7)所示。

$$L^{(t)}=\sum_{i=1}^{n}(g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i))+\Omega(f_t) \tag{7}$$

1.2 正则化惩罚函数

XGBoost 算法需要不断地向模型中添加决策树,随着决策树的添加,预测效果会越来越接近真实值。但是,叶子节点和树过多,会增加过拟合的风险。因此,需要在该模型的目标函数中引入惩罚项函数,如式(8)所示。

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{8}$$

式中:T 为树的个数;γ 为 T 的惩罚系数;ω_j 为第 j 个叶子节点的权重值;λ 为正则化惩罚项系数。

1.3 XGBoost 算法的目标函数

XGBoost 算法的目标函数由损失函数和正则化项组成。为了使模型不陷入过拟合,同时使目标函数尽可能达到最小值,缩小预测值与真实值的差距,选择对目标函数进行最大化的降低,如式(9)所示。

$$O_{bj}(t)=\sum_{i=1}^{n}[2(\hat{y}_i^{(t-1)}-y_i)f_i(x_i)+f_i(x_i)^2]+\Omega(f_t)+C_0 \tag{9}$$

式中:$2(\hat{y}_i^{(t-1)}-y_i)$为残差,即预测值与真实值的差距;y_i 和 C_0 为计算过程中的常数。

在求导计算过程中,常数项无影响,所以可以直接省略。令节点 j 的一阶偏导数累加和二阶偏导数累加分别为 $G_j=\sum_{i\in I_j}g_i$,$H_j=\sum_{i\in I_j}h_i$。为确定目标函数的最小值,对目标函数进行求导,得式(10)。

$$\frac{\partial_j(f_i)}{\partial\omega_j}=G_j+(H_j+\lambda)\omega_j=0 \tag{10}$$

将 ω_j 的最优解带入目标函数中,得到化简后的目标函数,如式(11)所示。

$$O_{bj}(t)=-\frac{1}{2}\sum_{i=1}^{T}\frac{G_j^2}{H_j+\lambda}+\gamma T \tag{11}$$

1.4 贝叶斯优化算法

贝叶斯优化(Bayesian Optimization,简称 BO)算法是基于概率分布的全局优化算法,多用于超参数确定,即通过对目标函数形状的学习,找到对全局提升最大的超参数。该算法的计算步骤为:先在整个区域随机均匀选点,每增加一个点,开始循环,找到 N 个候选解后,建立高斯回归模型。再采用该模型计算每个点的后验概率,并找出极大点作为下一个点。找出所有点中的极大值就是最优解[5]。相比于常规的网格搜索和随机搜索,利用已经搜索过的点的信息可以提高寻优过程的速度以及结果的质量。同时,贝叶斯优化算法不依赖高性能设备,迭代次数少,对于分类问题有更好的鲁棒性。

2 故障类型和诊断流程

2.1 故障类型

变压器的油中溶解气体为 H_2、CH_4、C_2H_6、C_2H_4、C_2H_2。溶解气体浓度数据不准确会导致无法表达设备的个性化、差异化特征,进而影响评价准确率[15]。因此,需对气体浓度数据进行归一化处理。本研究采用无编码比值法对故障数据进行处理,并将其作为样本输入,再利用油中溶解气体与故障特征的关联性,对故障进行识别。表 1 为故障类型编码。

表 1　故障类型编码

Tab.1　Coding of fault type

工作状态	状态编码	工作状态	状态编码
正常	0	局部放电	3
中低温过热	1	低能放电	4
高温过热	2	高能放电	5

2.2　诊断流程

基于贝叶斯优化 XGBoost 算法的变压器故障诊断步骤为:(1) 对采集到的样本数据进行归一化、标准化处理, 并将数据集按照 4:1 比例分为训练集与测试集。(2)对模型初始化后,采用高斯过程回归计算 AC 函数最大值,若满足目标值,则输出;不满足,返回高斯过程继续计算。(3)利用贝叶斯优化算法对模型的学习率、树深和分类器数目进行优化,得到最优超参数,并对 XGBoost 算法的参数进行设置。(4)判断超参数是否达到目标最大精度,若达到,则将模型的参数设定为最优超参数, 否则, 返回步骤(2)、(3)重复进行。(5)通过测试集对模型的诊断效果进行测试,确定该模型的诊断精度,对模型做出评价。具体流程如图 2 所示。

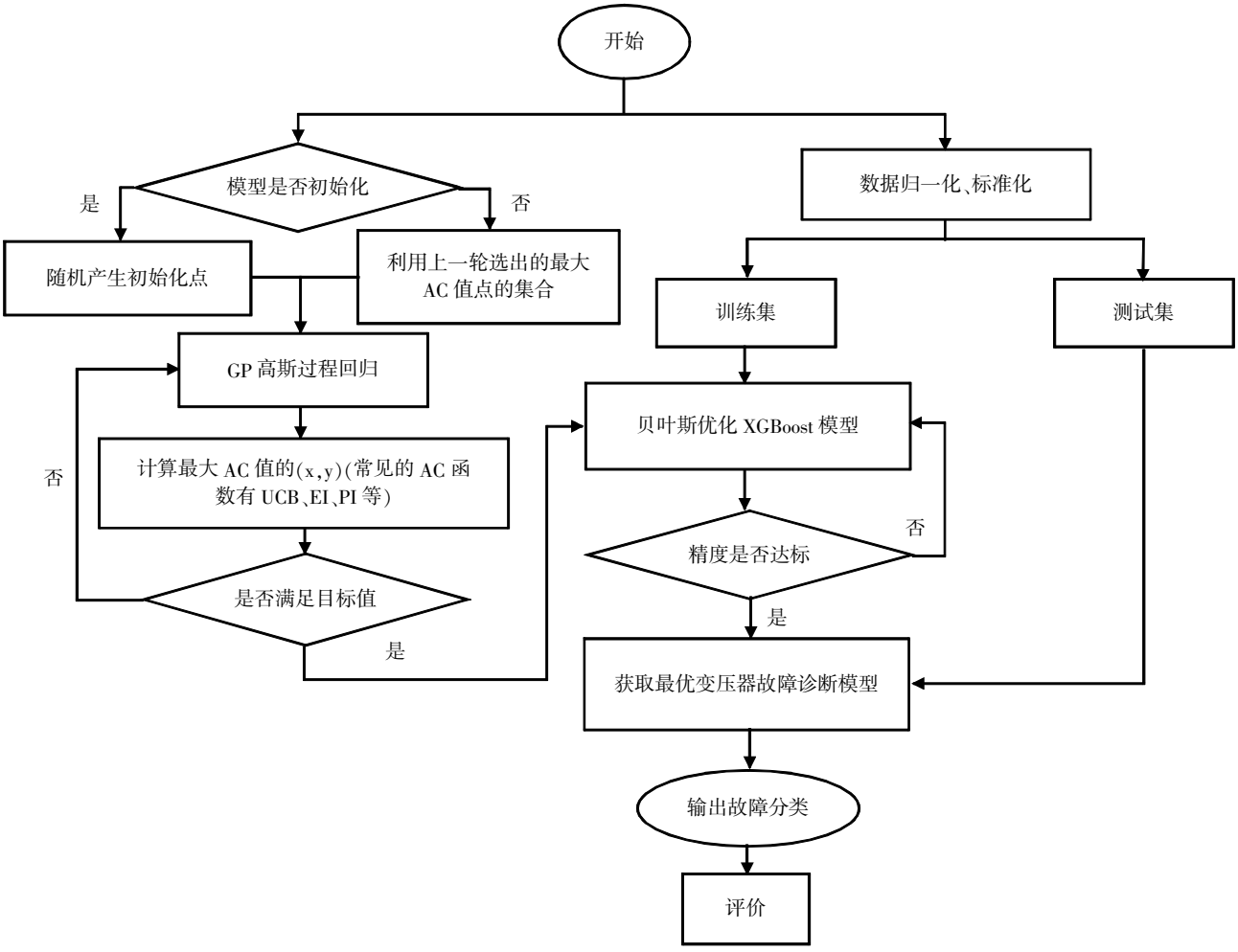

图 2　贝叶斯优化 XGBoost 模型的流程图

Fig.2　Bayesian optimization process of XGBoost model

3　算例分析

3.1　故障样本选择

本文选取 259 组故障类型确定的变压器 DGA 数据进行分析, 训练样本与测试样本比值为 4:1,52 组数据作为测试样本,207 组数据作为训练样本[16]。训练集与测试集数据分布如表 2 所示。

3.2　参数设置

利用 Anaconda 编程, 在 PyCharm 环境下运行,

表 2 样本分布

Tab.2 Sample distribution

工作状态	样本总数	训练集样本数	测试集样本数
正常	35	28	7
中低温	45	36	9
高温	45	36	9
局部放电	45	36	9
低能放电	45	36	9
高能放电	44	35	9
总计	259	207	52

表 3 XGBoost 诊断结果

Tab.3 XGBoost diagnostic result

故障类型	精度/%	召回/%	F1 分数/%
0	70.00	100.00	82.35
1	100.00	100.00	100.00
2	100.00	100.00	100.00
3	100.00	100.00	100.00
4	87.50	77.70	82.35
5	100.00	77.78	87.50
准确度	–	–	92.31
微平均	92.92	92.59	92.03
宏平均	93.80	92.31	92.41

XGBoost 算法的初始参数预设为：最大深度 max_depth=8、分类器数量 n_estimators=770、学习率 learning_rate=0.55，其余参数均为默认值[15]。采用训练样本对 XGBoost 模型进行训练，结果如表 3 所示。

由表 3 可知，中低温过热、高温过热、局部放电的故障样本全部被正确识别，2 个低能放电和 1 个高能放电被错判为正常状态，1 个高能放电被错判为低能放电，3 个正常状态、2 个低能放电、2 个高能放电的故障样本存在错误识别。从整体来看，在不同故障状态下，模型表现稳定，且诊断准确率为 92.31%。图 3 为模型的混淆矩阵。

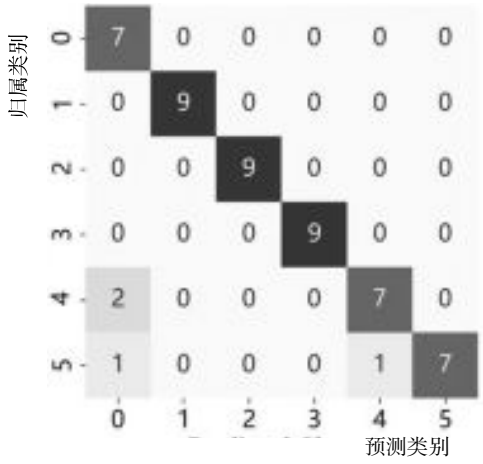

图 3 XGBoost 算法混淆矩阵

Fig.3 Confusion matrix of XGBoost algorithm

3.3 贝叶斯寻优

XGBoost 模型的超参数对其训练学习效果有显著的影响，采用贝叶斯对超参数进行优化，将超参数作为目标函数的输入。采用 INT 函数对最大深度和分类器数目 2 个超参数进行取值取整。为了能够复现此优化结果，定义随机树种子为 2022。由于目标函数为损失函数，贝叶斯优化只支持寻找最大值，此处需要设定目标函数的输出为负值。采用贝叶斯优化后的参数取值如表 4 所示。

由表 4 可知，优化后，学习率 learning_rate=0.06、最大深度 max_depth=5、分类器数量 n_estimator=134。采用优化后的参数，再对样本进行诊断，结果如表 5 所示。

由表 5 可以看出，与经验参数下的模型相比，BO-XGBoost 模型的诊断精确度有所提升，整体精度达到了 98.08%，图 4 为 BO-XGBoost 模型的混淆矩阵。由图 4 可知，1 个低能放电错判为正常状态，1 个正常状态和 1 个低能放电故障存在错误识别，经贝叶斯优化后的 XGBoost 算法，低能放电和高能放

表 4　贝叶斯优化后 XGBoost 算法各参数的取值

Tab.4　Parameters of XGBoost algorithm after Bayesian optimization

参数	优化前	优化后	优化区间
学习率	0.5	0.06	(0.05,0.5)
最大深度	4	5	(3,6)
分类器数	100	134	(100,150)

表 5　贝叶斯优化后 XGBoost 算法诊断结果

Tab.5　Diagnosis result of XGBoost algorithm after Bayesian optimization

故障类型	精度/%	召回/%	F1 分数/%
0	87.50	100.00	93.33
1	100.00	100.00	100.00
2	100.00	100.00	100.00
3	100.00	100.00	100.00
4	100.00	88.89	94.12
5	100.00	100.00	100.00
准确度	–	–	98.08
微平均	97.92	98.15	97.91
宏平均	98.32	98.08	98.08

电故障诊断精度得到明显提高。

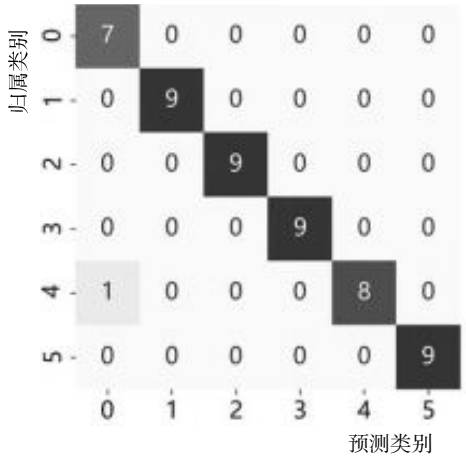

图 4　BO-XGBoost 模型混淆矩阵

Fig.4　Confusion matrix of BO-XGBoost model

3.4　多模型诊断结果对比

选取支持向量机(SVM)、随机森林(RF)、K 邻近(KNN)3 种主流监督学习模型在经验参数下进行训练,SVM 中超参数惩罚系数 C=1.74,核函数 kernel=rbf(高斯径向基核函数),gamma=30.554, 最大迭代次数 max_iter=-1[4];RF 中最大特征数 max_features=None,最小样本数 Tsplit=2,叶子节点最少样本数 Tleaf=1,最大深度 Dtre=10,分类器数 Nest=72[5];KNN 中取邻近点的个数 n_neighbors=5,叶子节点阈值 leaf_size=30,算法 algorithm=auto[15]。将测试结果与初始 XGBoost 模型以及贝叶斯优化后的 XGBoost 模型进行对比,模型运行时间如图 5 所示,诊断结果如表 5 所示。

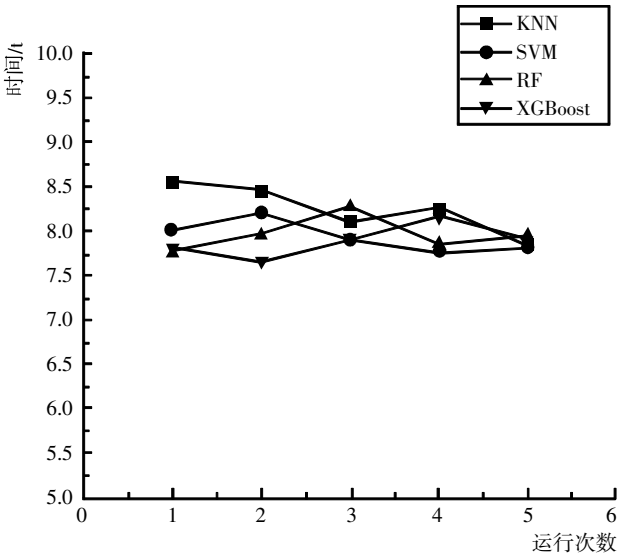

图 5　各模型运行时间

Fig.5　Running time of each model

由图 5 可以看出,5 模型采用同组数据多次运行时间进行对比分析,平均运行时间较长的是 KNN 模型,其次是 SVM 和 RF 模型,表现最佳的是 XGBoost 模型。

由表 5 可知,5 种模型对正常状态的诊断精度分别为 47.50%、55.40%、66.45%、84.11%、93.61%,对高温过热故障与局部放电故障诊断精度整体较高;XGBoost 模型对正常状态和低能放电诊断精度相对较低, 采用贝叶斯优化后的模型对正常状态的

表 5 多模型诊断结果对比

Tab.5 Comparison of multiple model diagnosis results

故障类型	SVM/%	RF/%	KNN/%	XGBoost/%	BO-XGBoost/%
0	47.50	55.40	66.45	84.11	93.61
1	66.90	89.07	74.23	100.00	100.00
2	70.20	92.23	91.29	100.00	100.00
3	91.20	79.80	89.92	100.00	100.00
4	79.82	91.70	74.29	82.52	94.34
5	68.35	93.20	75.30	88.43	100.00
平均	70.66	75.50	78.58	92.31	98.08

诊断精度提升了 9.5%,低能放电故障诊断精度提高了 11.58%，高能放电故障诊断精度提高了 11.52%，BO-XGBoost 模型的诊断精度达到了 93.61%。该模型在中低温过热、高温过热、局部放电和高能放电故障类型中均保持 100%的诊断精度,在正常状态和低温过热故障类型中仍可以保持较高的诊断正确率。贝叶斯算法改善了 XGBoost 算法超参数的寻优能力,构建的BO-XGBoost 诊断模型提高了故障诊断率,且高于 XGBoost 和 SVM 等模型的诊断率,从诊断的总体性能来比较,其诊断结果整体性能最好,能够较好地满足变压器故障诊断精度的要求。

4 结语

综上所述，贝叶斯优化算法改善了 XGBoost 算法超参数的寻优能力，与经验参数下的模型相比，BO-XGBoost 诊断模型可以有效提高变压器的故障诊断精度。尤其是对于高能放电故障,其诊断精度明显提高。与支持向量机(SVM)、随机森林(RF)、K 邻近(KNN)3 种主流监督学习模型相比,BO-XGBoost 模型诊断结果的整体性能最好。

参考文献:

[1] HU Hao, MA Xin, SHANG Yizi. A novel method for transformer fault diagnosis based on refined deep residual shrinkage network[J]. IET Electric Power Applications,2021(2):206-223.

[2] 林凡勤,李明明,郭红.变压器故障诊断技术综述[J].计算机与现代化,2022(3):116-126.

[3] 任双赞,徐尧宇,李元,等.应用于油中溶解气体分析的深度信念网络与典型神经网络对比研究[J].高压电器,2020,56(9):39-45.

[4] 张玉欣,程志峰,徐正平,等.参数寻优支持向量机在基于光声光谱法的变压器故障诊断中的应用[J].光谱学与光谱分析,2015,35(1):10-13.

[5] 王雪,韩韬.基于贝叶斯优化随机森林的变压器故障诊断[J].电测与仪表，2021，58(6)：167-173.

[6] 张镱议,彭鸿博,李昕,等.基于 DGA 特征量优选与改进磷虾群算法优化支持向量机的变压器故障诊断模型[J].电测与仪表,2019,56(21):110-116.

[7] 马鑫,尚毅梓,胡昊,等.基于数据特征增强和残差收缩网络的变压器故障识别方法[J].电力系统自动化,2022,46(3):175-183.

[8] CHEN Tianqi,GUESTRIN C.XGBoost:a scalable tree boostingsystem [C].Proceedings of the 22nd ACM SIGKDD Inter-national Conference on Knowledge Discovery and Data Mining.New York,USA:ACM,2016:785-794.

[9] WU Zhanhong. Improved Genetic Algorithm and XGBoost Classifier for Power Transformer Fault Diagnosis[J]. Frontiers in Energy Research, 2021(10):56-67.

[10] RAICHURA MAULIK, CHOTHANI NILESH, PATEL DHARMESH. Efficient CNN - XGBoost technique for classification of power transformer internal faults against various abnormal conditions [J]. IET Generation, Transmission & Distribution, 2021, 15(5): 972-985.

[11] 江敏,陈一民.贝叶斯优化算法的发展综述[J].计算机工程与设计,2010,31(14):3 254-3 259.

[12] 孙斌,储芳芳,陈小惠.基于贝叶斯优化 XGBoost 的无创血压预测方法[J].电子测量技术,2022,45(7):68-74.

[13] 周旭,王认卓,代亚勋,等.基于 BO-XGBoost 的煤自燃分级预警研究[J].煤炭工程,2022,54(8):108-114.

[14] 张又文,冯斌,陈页.基于遗传算法优化 XGBoost 的油浸式变压器故障诊断方法[J].电力自动化设备,2021,41(2):200-206.

[15] 张鹏,齐波,李成榕,等.电力变压器油中溶解气体特性影响因素的量化分析[J].中国电机工程学报,2021,41(10):3 620-3 631.

[16] 田晓飞.基于改进蝙蝠算法优化支持向量机的变压器故障诊断研究[J].黑龙江电力,2019,41(1):11-15.

[责任编辑 荆旭春]

Diagnosis on Transformer Fault Based on Bayesian Optimization XGBoost Algorithm

JIA Haoyang[1], *QIAN Yu*[2]
(1.North China University of Water Resources and Electric Power, Zhengzhou 450045, Henan, China;
2. Henan Zhongfu Industrial Co., Ltd., Gongyi 451261, Henan, China)

Abstract: In order to improve the sensitivity of small sample fault diagnosis, such as high energy discharge, a transformer fault diagnosis model is proposed based on Bayesian optimization extreme gradient lifting algorithm (BO-XGBoost). The basic principle of Bayesian optimization XGBoost algorithm and the flow of transformer fault diagnosis based on this algorithm are analyzed. Two hundred and fifty-nine groups of fault samples are selected. The specific application of this model is discussed. The model is compared with XGBoost, Support Vector Machine (SVM), Random Forest (RF) and K proximity method (KNN). The results show that the accuracy of BO-XGBoost model in transformer fault diagnosis is 98.08%, which is 5.77%, 27.42%, 22.58% and 19.5% higher than that of the aforementioned model, respectively.

Key Words: Transformer fault diagnosis; Bayesian optimization algorithm; XGBoost algorithm; dissolved gas in oil; fault type; diagnosis process; diagnosis accuracy; comparison of results

(上接第 29 页)

[10] 吴志刚,江滔,樊艳磊,等.基于 Landsat8 数据的地表温度反演及分析研究:以武汉市为例[J].工程地球物理学报,2016,13(1):135-142.

[责任编辑 杨明庆]

Research on Mitigation Effect of Urban Park on Heat Island Effect in Kaifeng City

CUI Yaohui[1,2], *HUA Naixin*[1], *ZHANG Dan*[1,2]
(1.Yellow River Conservancy Technical Institute, Kaifeng 475004, Henan, China; 2.Henan Engineering Research Center of Surveying and Mapping Live-action 3D Technology, Kaifeng 475004, Henan, China)

Abstract: Urban heat island effect is one of the important factors affecting urban ecosystem and urban ecological civilization construction. As an important part of the urban ecosystem, the park has a high ecological service function. Its layout and quality reflect the quality of life in a city, and play an irreplaceable role in promoting the sustainable development of the city and guaranteeing the security of the city. The land surface temperature was obtained by using remote sensing image inversion technology, and the mitigation effect of park on urban heat island effect in Kaifeng City is discussed by comparing the distribution of land surface temperature caused by different periods and different surface features.

Key Words: Kaifeng city; urban park; urban heat island effect; mitigation effect; remote sensing technology; geographic information system