

基于遗传算法优化XGBoost的油浸式变压器故障诊断方法

张又文,冯斌,陈页,廖伟涵,郭创新

(浙江大学电气工程学院,浙江杭州 310027)

摘要:为了提高油浸式变压器故障诊断的精度及可靠性,研究了一种基于遗传算法优化极端梯度提升(XGBoost)的油浸式变压器故障诊断方法。首先,以油中溶解气体分析(DGA)为依据,采用无编码比值方法提取油浸式变压器的9维故障特征,并对数据样本进行归一化处理;以归一化样本为输入建立基于XGBoost的故障诊断模型,并采用遗传算法对模型中的多个超参数同时进行优化。在算例部分,收集547例故障类型确定的DGA数据进行对比实验,结果表明与现有传统方法相比,所提方法的诊断精度和稳定性有显著提升;同时验证了遗传算法对故障诊断模型的优化提升效果。

关键词:油中溶解气体分析;变压器;故障诊断;极端梯度提升;遗传算法

中图分类号:TM 41

文献标志码:A

DOI:10.16081/j.epae.202012021

0 引言

油浸式变压器是输变电系统的关键设备之一,其运行状态决定着电力系统能否安全可靠运行。然而,我国目前尚在投运的油浸式变压器中,有很大一部分运行年限较长,存在绝缘劣化等故障隐患^[1]。为了保障电力系统运行的可靠性和经济性,必须建立高效准确的变压器故障实时诊断模型。

油浸式变压器发生绝缘老化时会产生少量气体溶解在绝缘油中,油中溶解气体的组成成分及其比例关系能够反映变压器的运行工况。正常运行的油浸式变压器,由于绝缘老化裂解等原因会产生极少量的气体,其主要成分为氢气(H_2)、甲烷(CH_4)、乙烷(C_2H_6)、乙烯(C_2H_4)、乙炔(C_2H_2)、一氧化碳(CO)和二氧化碳(CO_2)^[2]。当变压器发生不同类型的故障时,特定的气体组分会迅速增加,例如绝缘油过热时, CH_4 和 C_2H_6 的比例迅速增大;高能放电时, H_2 和 C_2H_2 含量升高。变压器故障类型和气体成分的变化呈现较强的相关性,因此对气体组分、含量进行分析的油中溶解气体分析DGA(Dissolved Gas Analysis)方法是目前最为重要的变压器状态检测和故障诊断方法之一^[3]。

以DGA为基础形成的三比值法、Rogers比值法、Duval三角形法等规则简单,曾发挥了重要作用,但均存在着状态编码不完备、编码界限过于绝对的问题,在实际应用中存在局限性^[4]。近年来,随着人工智能的不断发展,不少学者将机器学习的方法应用于变压器故障诊断建模中,并取得了较好的诊断效果。当前主流的变压器故障诊断模型分为非集成

学习模型和集成学习模型。其中非集成学习的方法包括BP神经网络、支持向量机(SVM)、贝叶斯网络等,上述方法改进了传统模型的计算速度和精度,但均存在局限性:BP神经网络学习能力强,但是易陷入局部最优,需要人为调整大量参数,且收敛速度较慢;SVM本质为二分类器,在处理变压器故障诊断等多分类问题时效率较低;贝叶斯网络要求条件属性较多等。基于集成学习的方法主要有随机森林RF(Random Forest)和梯度提升树GBDT(Gradient Boosting Decision Tree) 2个分支。RF以多个决策树为基分类器建立组合模型,通过减小方差提高预测精度,但在处理噪声较大的分类和回归问题时易产生过拟合的问题;GBDT通过降低偏差的方式减小总误差,且对调参的要求较低,具有更好的鲁棒性,但由于其个体学习器之间存在强依赖关系,无法并行训练,因此模型的训练速度较慢。

文献[5]提出了一种基于梯度提升(gradient Boost)的高拓展性方法——极端梯度提升XGBoost(eXtreme Gradient Boosting),其基本思想与GBDT相同,即基于Boosting思想,以损失函数的负梯度为依据训练多个个体学习器,并按照一定策略组合成一个准确可靠的集成学习器。与GBDT不同的是,XGBoost在损失函数中加入正则化项对模型的复杂程度施加惩罚,以提高模型的泛化能力,防止过拟合;并行和分布式的计算模式也大幅提高了其训练速度。文献[6]从梯度提升形式和正则化方法两方面进行阐述,论证了XGBoost相比于GBDT具有更高的适应性和更强的鲁棒性。目前XGBoost已在国内外多个领域得到了广泛认可。文献[7]将XGBoost和RF应用于海上和陆地风力涡轮机的故障分类,仿真结果统计表明,与SVM相比,该方法具有较强的抗过拟合能力,且在处理多维信号时具有显著优势。文献[8]将XGBoost应用于基于软件定义网络SDN

收稿日期:2020-03-05;修回日期:2020-10-29

基金项目:国家电网公司科技项目(52110418000T)

Project supported by the Science and Technology Project of SGCC(52110418000T)

(Software Defined Network)的云中分布式拒绝攻击DDoS(Distributed Denial of Service attack)的检测,结果验证了该方法快速且可扩展,具有较高的准确性和较低的误报率。国内已经有部分学者将XGBoost应用于电力系统负荷预测和暂态稳定性评估的研究中^[9-11],但在设备故障诊断和状态评估领域的应用研究仍有欠缺。

基于上述研究,本文提出了一种基于XGBoost的油浸式变压器故障诊断方法:采用无编码比值法提取变压器实时运行的特征;采用归一化方法处理提取后的特征数据;将处理后的数据作为样本训练XGBoost模型,并采用遗传算法(GA)对模型的多个超参数同时进行优化,最终建立基于XGBoost的变压器故障诊断模型,实现对油浸式变压器运行状态的实时精准诊断。算例分析结果表明,所提方法及模型能够稳定可靠地提升变压器故障诊断的准确率和稳定性。

1 XGBoost

XGBoost以分类回归树CART(Classification And Regression Tree)为基分类器,采用集成学习中梯度提升的方法进行加法训练,将多个个体分类器组合成一个集成分类器,以提升其预测速度与精度。

1.1 基分类器与集成学习

集成学习是一种数据挖掘算法,本质上是将多个学习器进行有效融合,形成一个精度和泛化性显著优于单一学习器的集成学习器^[12]。集成学习主要分为个体学习器的构建和个体学习器集成方法的确定2个阶段。

XGBoost模型的基学习器为CART,对于每棵CART,其复杂度由结构和叶子节点输出值决定,对于一个确定的输入,存在一个叶子节点输出值与之对应,表征CART对当前输入的预测结果。只包含1棵CART的分类模型过于简单,往往无法对多种故障进行有效分类,为了提升模型的可信度,在模型中再加入1棵CART,通过不同的分类规则进行预测,最终预测结果为2棵CART的预测分数之和。2棵CART根据不同的特征进行分类,不仅提高了预测的可信度,也提升了计算速度。在实际研究中,为了处理更加复杂的分类问题,通常向模型中加入多棵CART以不断提升学习效果。

集成学习的性能同样依赖集成策略的选择。在集成学习中,Bagging算法与Boosting算法是2种典型的集成方式^[13]。Bagging算法通过放回抽样方式对训练样本进行采样,生成 n 个不同子集,利用每个样本子集训练生成一个基学习器,再利用组合策略将基学习器构建为集成学习器;Boosting算法也是一种基于数据集的重采样算法,与Bagging算法不同的是,Boosting算法要求基学习器能够对指定的数据分

布进行学习。指定的数据分布通过重赋权法实现,即在每轮训练中,根据基学习器的表现对训练样本的权重分布进行调整,每一次迭代增加不能正确学习的样本权重,降低能正确学习的样本权重,使得先前训练中结果错误的样本在后续训练获得更多关注。对无法接受带权样本的算法,则可通过重采样法实现,即在每一轮学习中,根据样本分布对训练集重新进行采样,再利用重采样所得样本集对基学习器进行训练^[14]。Boosting算法在每轮都要检查当前生成的基学习器是否满足基本条件,一旦不满足则抛弃当前基学习器,学习过程也随之停止^[15]。基于重赋权法的Boosting算法流程如图1所示。

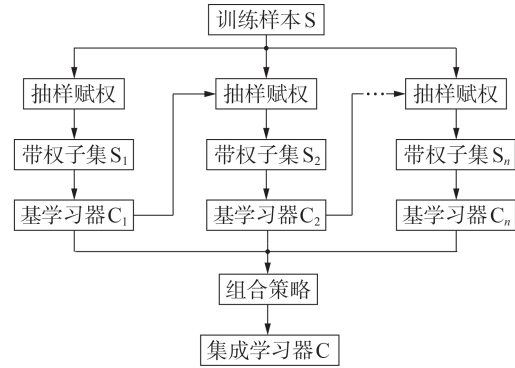


图1 Boosting算法流程

Fig.1 Flowchart of Boosting algorithm

1.2 XGBoost原理

XGBoost以CART为基分类器,通过梯度提升的方法进行集成,其基本思想是将多个分类准确率较低的基分类器集成为一个准确可靠的集成分类器。XGBoost的训练始于一个常数预测,每次迭代生成1棵CART拟合现有模型的残差,通过多轮迭代提升模型的准确率。XGBoost模型构建如下。

对于一个有 n 个样本、 m 个特征的数据集 $D = \{(\mathbf{x}_i, y_i)\} (\mathbf{x}_i \in \mathbf{R}^m, y_i \in \mathbf{R})$, K 棵CART最终预测输出 $\hat{y}_i$ 为:

$$\hat{y}_i = \sum_{k=1}^K f_k(\mathbf{x}_i) \quad f_k \in F \quad (1)$$

$$F = \{f(\mathbf{x}) = \omega_{q(\mathbf{x})}\} \quad q: \mathbf{R}^m \rightarrow T, \omega \in \mathbf{R}^T \quad (2)$$

其中,每个函数 f_k 对应一棵独立的树结构向量 q 和叶子权重 ω , q 由样本指向相应的叶子标签,每棵CART的每个叶子节点对应一个连续分数值,即权重,第 i 个节点的分数为 ω_i ; T 为叶子节点个数; F 为CART构成的集合; $\omega_{q(\mathbf{x})}$ 为对样本 $\mathbf{x}$ 的打分,即模型预测值。对于每个样本,各CART依据不同分类规则将其分类到叶子节点中,通过累加对应叶子的分数 ω 获得最终的预测结果。

为了学习模型中的函数集合,最小化下列正则化目标:

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \quad (3)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \omega^2 \quad (4)$$

其中, ϕ 为模型中的函数集合; γ, λ 为控制模型复杂程度的正则化参数, 参数值越大, 模型越不容易过拟合。式(3)为可微的凸损失函数, 用于度量预测值与标签值的差距; 式(3)等号右侧第1项为损失函数, 用于度量 CART 预测值与真实值之间的差距; 式(3)等号右侧第2项为正则项, 用于对模型的复杂程度进行惩罚, 该附加的正则项能够平滑各叶节点的权重, 避免过拟合。显然, 正则化的目标函数倾向于选择结构简单、预测性强的模型。当正则化参数为0时, 目标函数就退化为传统的 GBDT。

上述目标函数以函数作为参数, 无法采用传统方法进行优化, 故采用加法学习方式(additive manner)进行训练。训练由一个常数预测开始, 每次向模型中加入一个新函数学习当前的模型。记 $\hat{y}_i^{(t)}$ 为第 i 个样本在第 t 次训练中的预测值, 寻找一个最优的新函数 f_t , 以最大限度地降低目标函数。第 t 次训练中的预测值等于第 $t-1$ 次训练中的预测值与 f_t 之和。

$$\begin{cases} \hat{y}_i^{(0)} = 0 \\ \hat{y}_i^{(1)} = f_1(\mathbf{x}_i) = \hat{y}_i^{(0)} + f_1(\mathbf{x}_i) \\ \vdots \\ \hat{y}_i^{(t)} = \sum_{k=1}^t f_k(\mathbf{x}_i) = \hat{y}_i^{(t-1)} + f_t(\mathbf{x}_i) \end{cases} \quad (5)$$

$$L^{(t)} = \sum_{i=1}^n \left(l(y_i, \hat{y}_i^{(t-1)}) + f_t(\mathbf{x}_i) \right) + \Omega(f_t) \quad (6)$$

为了快速优化目标函数, 对损失函数 $L^{(t)}$ 进行二阶泰勒展开, 可得:

$$L^{(t)} = \sum_{i=1}^n \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t) \quad (7)$$

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad (8)$$

$$h_i = \partial_{\hat{y}_i^{(t-1)}}^2 l(y_i, \hat{y}_i^{(t-1)}) \quad (9)$$

其中, g_i 和 h_i 分别为损失函数的一阶、二阶梯度。将常数项移除, 则可以得到第 t 次训练的简化目标函数为:

$$\tilde{L}^{(t)} = \sum_{i=1}^n \left(g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right) + \Omega(f_t) \quad (10)$$

将叶子节点 j 中的样本编号集合定义为:

$$I_j = \{i | q(\mathbf{x}_i) = j\} \quad (11)$$

其中, $q(\mathbf{x}_i)$ 为 $\mathbf{x}_i$ 对应的叶子标签值。

则根据式(11), 可以将式(10)写为:

$$\begin{aligned} \tilde{L}^{(t)} &= \sum_{i=1}^n \left(g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right) + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 = \\ &= \sum_{j=1}^T \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T \end{aligned} \quad (12)$$

对于一个固定的结构 q , 其叶子节点 j 的最优权重如式(13)所示。

$$\omega_j^* = - \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \quad (13)$$

则对应的最优目标函数如式(14)所示, 其可用于评价 CART 结构 q 的质量。

$$\tilde{L}^{(t)}(q) = - \frac{1}{2} \sum_{j=1}^T \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \quad (14)$$

通常情况下, 列举所有可能的 CART 结构 q 是不可能的, 因此采用精确贪心算法, 从单叶节点开始迭代地向树中加入分支。设 I_L, I_R 分别为分裂后左、右子分支节点的样本编号集, $I = I_L \cup I_R$, 则分裂后损失的减小值如式(15)所示, 其可用于评估候选分裂点。

$$L_{\text{split}} = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i \right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i \right)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i \right)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \quad (15)$$

1.3 分裂点的查找

在 XGBoost 的训练中, 最关键的问题是根据式(15)找到最佳分裂点。为此, 利用精确贪心算法枚举出所有特征的所有可能分裂点。遍历连续特征的所有可能分裂点对计算的要求很高, 为了提高计算效率, 精确贪心算法必须首先对每个特征的所有数据进行排序, 并且按照顺序访问数据, 以获得式(15)所示的结构评分的梯度数据。精确贪心算法的伪代码见附录 A。

精确贪心算法遍历了所有可能的分裂点, 但当数据量过大时, 无法将全部的数据加载入内存, 为此提出了一种近似算法^[5]。近似算法首先根据特征的百分比分布提出候选分裂点, 然后将连续的特征值映射到由候选分裂点分割出的桶(buckets)中, 遍历所有的候选分裂点找出最佳的分裂点。近似算法的伪代码见附录 B。

2 基于 XGBoost 的变压器故障诊断模型

2.1 特征提取与数据预处理

油浸式变压器油中溶解特征气体为 H_2 、 CH_4 、 C_2H_6 、 C_2H_4 、 C_2H_2 , 上述 5 种气体含量分散性较大, 直接作为样本输入会对诊断精度产生不良影响^[16]。因此, 在 DGA 故障诊断领域, 一般采用特征气体含量比值而非单一气体含量作为样本输入。本文按照式(16)对特征气体含量进行归一化处理。

$$x'_{ij} = \frac{x_{ij}}{\sum_{k=1}^5 x_{ik}} \quad (16)$$

其中, x_{ij} 为第 i 个样本中第 j 个特征气体含量; x'_{ij} 为归一化的特征气体含量。

油浸式变压器故障诊断中, DGA 气体的特征编码方式较多, 如 ICE 比值法 (包含 3 个维度的特征)、Rogers 比值法 (包含 4 个维度的特征) 等, 不同的编码方式可提取变压器内部不同的故障信息, 从而对后续故障诊断的结果产生影响。文献[17]通过大量的变压器模拟实验数据总结出不同工况下变压器气体产生规律, 结合特征气体的含量及其比值关系, 提出了利用无编码比值法分析判断变压器故障性质的诊断方法。无编码比值法包含 9 个维度的特征: $C(\text{CH}_4)/C(\text{H}_2)$, $C(\text{C}_2\text{H}_4)/C(\text{C}_2\text{H}_2)$, $C(\text{C}_2\text{H}_4)/C(\text{C}_2\text{H}_6)$, $(C(\text{CH}_4)+C(\text{C}_2\text{H}_4))/(C_1+C_2)$, $C(\text{H}_2)/(C(\text{H}_2)+C_1+C_2)$, $C(\text{C}_2\text{H}_4)/(C_1+C_2)$, $C(\text{CH}_4)/(C_1+C_2)$, $C(\text{C}_2\text{H}_6)/(C_1+C_2)$, $C(\text{C}_2\text{H}_2)/(C_1+C_2)$ 。其中, $C(\cdot)$ 表示气体含量; C_1 为 CH_4 所代表的一阶碳氢化合物气体含量之和; C_2 为 C_2H_6 、 C_2H_4 、 C_2H_2 所代表的二阶碳氢化合物气体含量之和。

本文采用上述无编码比值作为油浸式变压器故障诊断模型的特征输入, 利用油中溶解特征气体含量比值与故障类型的相关性来诊断变压器故障类型。与传统的比值法相比, 无编码比值法包含了更多维度的特征信息, 解决了三比值法编码不完备的问题, 能够提升诊断结果的准确性^[18]。

2.2 故障状态编码

在对故障模型进行训练前, 应对变压器故障类型进行编码。根据 DL/T 722—2014《变压器油中溶解气体分析和判断导则》, 油浸式变压器故障类型分为低温过热、中温过热、高温过热、局部放电、低能放电、高能放电、低能放电兼过热、高能放电兼过热 8 种。结合正常工况, 油浸式变压器工作状态及相应状态编码如表 1 所示。

表 1 设备状态编码

Table 1 Device status code

工作状态	状态编码	工作状态	状态编码
正常	0	低能放电	5
低温过热	1	高能放电	6
中温过热	2	低能放电兼过热	7
高温过热	3	高能放电兼过热	8
局部放电	4		

2.3 基于 GA 优化的模型训练

为了克服交叉验证以及网格搜索参数寻优等传统优化方法的缺点, 本文将 GA^[19] 引入变压器故障诊

断模型中, 实现对多个超参数的同时优化, 以提高模型诊断准确率。

2.3.1 优化参数选取

XGBoost 采用集成学习中梯度提升的方法进行加法训练, 其由一个常数预测开始, 在每次迭代过程中, 以最小化正则化目标函数为依据生成新的弱学习器并将其加入现有模型中, 以提升模型的预测精度。

在模型训练中, 迭代次数过少易造成欠拟合, 导致模型对问题的求解能力不足, 而迭代次数过多又可能造成过拟合, 使得模型对于训练数据表达能力过强而泛化性能低下; 学习率 η 过小时, 梯度下降很慢, 过大时又可能跨过最优值, 产生振荡; 同时模型中其他参数, 如决策树的最大深度 $d_{\max}$ 、随机样本的抽取比例 $r_{\text{subsample}}$ 、特征的抽取比例 $r_{\text{colsample}}$ 、决策树节点分裂标准 γ_{split} 都会对模型学习能力和分类性能产生影响。因此本文选取以上 6 个超参数, 采用 GA 进行优化, 以提高诊断模型的性能。

2.3.2 GA 优化的 XGBoost 模型

基于 XGBoost 的油浸式变压器故障诊断模型流程图见附录 C, 其具体实现步骤如下:

- (1) 对采集到的样本数据进行归一化处理;
- (2) 选取无编码比值作为模型的特征输入, 将样本数据按比例随机分为训练集和测试集;
- (3) 设置 XGBoost 模型的初始参数并进行预训练, 利用 GA 不断对模型参数进行调整;
- (4) 判断是否达到最大迭代次数或者终止条件, 若达到则将此时的训练参数值作为模型的最优参数, 否则返回步骤 (3);
- (5) 利用测试集对模型的诊断效果进行测试, 输出故障分类结果。

3 算例分析

3.1 特征提取与数据预处理

本文从文献[20-22]中选取 547 个故障类型确定的典型油浸式变压器 DGA 数据进行分析, 其中测试集样本数占样本总数的 20%。训练集和测试集样本分布具体情况见表 2。XGBoost 的初始参数预设为: $n=500$; $\eta=0.3$; $\gamma_{\text{split}}=0$; $d_{\max}=6$; $r_{\text{subsample}}=1$; $r_{\text{colsample}}=1$;

表 2 样本分布

Table 2 Sample distribution

工作状态	样本总数	训练集样本数	测试集样本数
正常状态	52	41	11
低温过热	52	43	9
中温过热	58	41	17
高温过热	119	80	39
局部放电	55	43	12
低能放电	118	79	39
高能放电	93	65	28
总计	547	392	155

定义损失函数为线性回归(reg:linear);其余参数均设置为默认值。

3.2 XGBoost的诊断结果

在Inter(R) Core(TM) i5、2.70 GHz、内存8 GB的环境下,利用python3编程,输入预设参数和训练样本对所提模型进行训练。基于XGBoost的油浸式变压器故障诊断结果如表3所示。数据显示,测试样本中表征中温过热、高温过热、局部放电和高能放电的故障样本全部被正确识别;低温过热的故障样本误判1个;正常状态样本误判4个;低能放电的故障样本误判3个。除正常状态外,本文方法的诊断准确率较高,且在不同故障类型下表现稳定。

表3 XGBoost的诊断结果

Table 3 Diagnosis results of XGBoost

工作状态	测试样本数	诊断正确样本数	正确率
正常状态	11	7	0.636
低温过热	10	9	0.900
中温过热	13	13	1
高温过热	22	22	1
局部放电	11	11	1
低能放电	26	23	0.885
高能放电	17	17	1
总计	110	102	0.927

3.3 多模型诊断结果对比

选取线性判别分析(LDA)、K近邻(KNN)、SVM、RF、GBDT这5种主流监督学习模型在经验参数下进行训练,并将测试结果与XGBoost模型进行对比,结果如表4所示。由表可见,XGBoost模型的诊断精度在6种模型中最高,并且在分类正确率较低的状态中也有很好的表现。

表4 不同模型的诊断结果

Table 4 Diagnosis results of different models

状态编码	诊断正确率					
	LDA	KNN	SVM	RF	GBDT	XGBoost
0	0.455	0.636	0.273	0.545	0.455	0.636
1	0.300	0.500	0.700	0.900	0.900	0.900
2	1	0.692	0.692	0.923	0.923	1
3	0.818	0.909	0.909	0.955	0.955	1
4	0.455	0.909	0.909	0.818	0.909	1
5	0.538	0.769	0.692	0.923	0.885	0.885
6	0.471	0.765	0.706	1	1	1
平均	0.600	0.764	0.718	0.891	0.882	0.927

为了对XGBoost的计算稳定性进行说明,连续对547个样本组成的样本集进行100次随机抽样,每次抽取20%的样本作为测试集,其余80%的样本作为训练集,利用得到的训练样本对不同模型进行训练,并对诊断结果进行统计,结果如表5所示。由表可见,基于XGBoost的变压器故障诊断模型的100次诊断平均正确率为0.917,正确率均方差为0.018;与5种主流监督学习模型相比,基于XGBoost的变压器

故障诊断模型具有最高的平均正确率,并且与准确率较高的RF和GBDT模型相比正确率均方差更小,说明本文所提方法能够稳定保持高精度的计算输出。

表5 不同模型重复训练诊断结果

Table 5 Diagnosis results of different models with repeated training

模型	最高正确率	最低正确率	平均正确率	正确率均方差
LDA	0.709	0.518	0.611	0.038
KNN	0.845	0.700	0.741	0.044
SVM	0.800	0.718	0.724	0.035
RF	0.936	0.855	0.873	0.032
GBDT	0.945	0.818	0.887	0.026
XGBoost	0.982	0.891	0.917	0.018

GBDT、XGBoost、LightGBM都是目前机器学习中的流行的梯度提升方法,为了对它们的计算量进行对比,读取上述100次随机抽样训练所用的平均训练时间和平均测试时间如表6所示。由表可见,本文所采用的XGBoost与GBDT、LightGBM相比,平均测试时间没有显著差异,但其平均训练时间大幅减少,仅为GBDT平均训练时间的35.6%、LightGBM平均训练时间的16.5%。由于测试阶段的计算量与训练阶段相比可以忽略不计,综合来看XGBoost模型的收敛速度快,训练效率较高。

表6 不同模型训练时间对比

Table 6 Comparison of training time among different models

模型	平均训练时间/ms	平均测试时间/ms
GBDT	845.9	3.36
XGBoost	302.5	3.08
LightGBM	1 837.0	3.72

3.4 GA寻优效果

由于XGBoost模型的超参数对其训练学习效果有着显著的影响,因此在预训练结束后,采用GA对设定的6个超参数同时进行优化。GA参数设置如下:采用实数编码方式;种群个数为50;适应度函数为XGBoost模型诊断正确率。当遗传代数达到第6代时,适应度达到0.981并在之后保持不变。相比于初始模型,诊断正确率提升了5.4%。此时输出XGBoost模型的最优参数,如表7所示。

表7 最优参数

Table 7 Optimal parameters

参数	最优值	参数	最优值
η	0.55	$r_{\text{cossample}}$	0.80
d_{max}	8	n	770
$r_{\text{subsample}}$	0.64	γ_{split}	0.23

选取3.3节多模型诊断结果中精度较高的2种监督学习模型RF和GBDT,在上述GA参数设置下,对两者的超参数进行优化,结果如图2所示。可以

看出,RF和GBDT优化模型精度分别在第6代和第10代达到了0.910和0.905,并在之后保持不变,相比于初始模型,诊断精度分别提升了3.7%和1.8%。与GA优化的XGBoost模型相比,RF和GBDT对参数设置更加敏感,GA的优化能够提升两者的诊断效果,防止不良参数设置对模型学习能力的负面影响,但最终精度仍低于优化后的XGBoost模型。

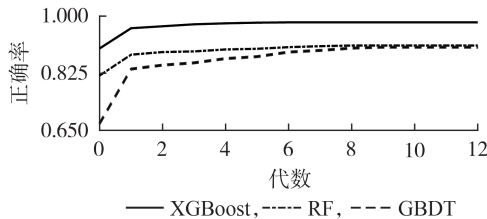


图2 GA的优化效果

Fig.2 Optimization result of GA

3.5 实例计算

利用上述模型对来自文献[23]的10组变压器检测异常数据进行故障识别,样本数据详见附录D。基于XGBoost的诊断模型和改良ICE比值法对10组样本的诊断结果比较如表8所示。由表可见,由于ICE比值法的编码边界太过于绝对,编码不完备,当样本处于不同编码区间边界时往往无法正确识别故障类型。而采用所提XGBoost模型可以对处于模糊区间的样本进行学习和正确分类,基本弥补了ICE比值法编码缺失、诊断正确率低等不足。

表8 诊断结果

Table 8 Diagnosis results

序号	实际故障 类型编码	改良ICE比值法		XGBoost 诊断结果
		编码	诊断结果	
1	0	000	无编码	0
2	1	001	1	1
3	1	001	1	1
4	2	021	2	2
5	2	021	2	2
6	3	002	3	3
7	3	002	3	3
8	4	010	4	4
9	5	121	8*	5
10	6	101	7*	5*

注:“*”表示诊断错误。

4 结论

本文针对现阶段油浸式变压器故障多样性和复杂性导致的诊断精度不高的问题,提出了一种基于XGBoost的油浸式变压器故障诊断模型,结合算例数据分析,得到以下结论:

(1)对样本数据进行归一化处理,并采用无编码比值作为输入特征,深层地挖掘了油中溶解气体与故障类型间的内在联系;

(2)在采用无编码比值作为特征输入时,相比于

LDA、KNN、SVM、RF和GBDT,XGBoost具有更高的精度和稳定性;

(3)将GA引入XGBoost诊断模型,选取对模型影响较大的6个超参数进行同时优化,有效地提升了模型诊断精度。

附录见本刊网络版(<http://www.epae.cn>)。

参考文献:

- [1] 廖伟涵,郭创新,金字,等. 基于四阶段预处理与GBDT的油浸式变压器故障诊断方法[J]. 电网技术,2019,43(6):2195-2203.
LIAO Weihang, GUO Chuangxin, JIN Yu, et al. Oil-immersed transformer fault diagnosis method based on four-stage preprocessing and GBDT[J]. Power System Technology, 2019, 43(6): 2195-2203.
- [2] 刘一,倪远平. 三比值灰关联度分析的变压器故障诊断法[J]. 高电压技术,2002,28(10):16-17,27.
LIU Yi, NI Yuanping. A transformer faults diagnosis method based on the grey correlation theory[J]. High Voltage Engineering, 2002, 28(10): 16-17, 27.
- [3] 刘云鹏,付浩川,许自强,等. 基于AdaBoost-RBF算法与DSmT的变压器故障诊断技术[J]. 电力自动化设备,2019,39(6):166-172.
LIU Yunpeng, FU Haochuan, XU Ziqiang, et al. Transformer fault diagnosis technology based on AdaBoost-RBF algorithm and DSmT[J]. Electric Power Automation Equipment, 2019, 39(6): 166-172.
- [4] 石鑫,朱永利,宁晓光,等. 基于深度自编码网络的电力变压器故障诊断[J]. 电力自动化设备,2016,36(5):122-126.
SHI Xin, ZHU Yongli, NING Xiaoguang, et al. Transformer fault diagnosis based on deep auto-encoder network[J]. Electric Power Automation Equipment, 2016, 36(5): 122-126.
- [5] CHEN Tianqi, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2016: 785-794.
- [6] NIELSEN D. Tree Boosting with XGBoost: why does XGBoost win “every” machine learning competition[D]. Trondheim, Norway: Norwegian University of Science and Technology, 2016.
- [7] ZHANG Dahai, QIAN Liyang, MAO Bajin, et al. A data-driven design for fault detection of wind turbines using random forests and XGboost[J]. IEEE Access, 2018, 6: 21020-21031.
- [8] CHEN Zhuo, JIANG Fu, CHENG Yijun, et al. XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud[C]//2018 IEEE International Conference on Big Data and Smart Computing. Shanghai, China: IEEE, 2018: 251-256.
- [9] 刘波,秦川,鞠平,等. 基于XGBoost与Stacking模型融合的短期母线负荷预测[J]. 电力自动化设备,2020,40(3):147-153.
LIU Bo, QIN Chuan, JU Ping, et al. Short-term bus load forecasting based on XGBoost and Stacking model fusion[J]. Electric Power Automation Equipment, 2020, 40(3): 147-153.
- [10] 张晨宇,王慧芳,叶晓君. 基于XGBoost算法的电力系统暂态稳定评估[J]. 电力自动化设备,2019,39(3):77-83,89.
ZHANG Chenyu, WANG Huifang, YE Xiaojun. Transient stability assessment of power system based on XGBoost algorithm[J]. Electric Power Automation Equipment, 2019, 39(3): 77-83, 89.
- [11] 赵洪山,闫西慧,王桂兰,等. 应用深度自编码网络和XGBoost的风电机组发电机故障诊断[J]. 电力系统自动化,2019,43(1):81-86.

- ZHAO Hongshan, YAN Xihui, WANG Guilan, et al. Fault diagnosis of wind turbine generator based on deep autoencoder network and XGBoost[J]. Automation of Electric Power Systems, 2019, 43(1): 81-86.
- [12] 于玲, 吴铁军. 集成学习: Boosting 算法综述[J]. 模式识别与人工智能, 2004, 17(1): 52-59.
- YU Ling, WU Tiejun. Assemble learning: a survey of Boosting algorithms[J]. Pattern Recognition and Artificial Intelligence, 2004, 17(1): 52-59.
- [13] 沈学华, 周志华, 吴建鑫, 等. Boosting 和 Bagging 综述[J]. 计算机工程与应用, 2000, 36(12): 31-32, 40.
- SHEN Xuehua, ZHOU Zhihua, WU Jianxin, et al. Survey of Boosting and Bagging[J]. Computer Engineering and Applications, 2000, 36(12): 31-32, 40.
- [14] 马天男, 牛东晓, 黄雅莉, 等. 基于 Spark 平台和多变量 L2-Boosting 回归模型的分布式能源系统短期负荷预测[J]. 电网技术, 2016, 40(6): 1642-1649.
- MA Tiannan, NIU Dongxiao, HUANG Yali, et al. Short-term load forecasting for distributed energy system based on Spark platform and multi-variable L2-Boosting regression model[J]. Power System Technology, 2016, 40(6): 1642-1649.
- [15] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 173-177.
- [16] 汪可, 李金忠, 张书琦, 等. 变压器故障诊断用油中溶解气体新特征参量[J]. 中国电机工程学报, 2016, 36(23): 6570-6578.
- WANG Ke, LI Jinzhong, ZHANG Shuqi, et al. New features derived from dissolved gas analysis for fault diagnosis of power transformers[J]. Proceedings of the CSEE, 2016, 36(23): 6570-6578.
- [17] 杜洋. 用“无编码比值法”分析和判断变压器故障性质[J]. 变压器, 1999, 36(3): 32-36.
- DU Yang. Analysis and judgement for transformer fault by using “non-code ratio method”[J]. Transformer, 1999, 36(3): 32-36.
- [18] 薛华, 张宏鹏. 模糊无编码比值在变压器油中气体检测的应用[J]. 科学技术与工程, 2012, 12(33): 9033-9037.
- XUE Hua, ZHANG Hongpeng. Fuzzy no coding ratio in transformer oil gas detection in the application[J]. Science Technology and Engineering, 2012, 12(33): 9033-9037.
- [19] 马永杰, 云文霞. 遗传算法研究进展[J]. 计算机应用研究, 2012, 29(4): 1201-1206, 1210.
- MA Yongjie, YUN Wenxia. Research progress of genetic algorithm[J]. Application Research of Computers, 2012, 29(4): 1201-1206, 1210.
- [20] 李本铎. 智能算法在油浸式变压器故障诊断中的应用研究[D]. 南昌: 华东交通大学, 2015.
- LI Benxin. Application research of oil-immersed transformer fault diagnosis based on intelligent algorithms[D]. Nanchang: East China Jiaotong University, 2015.
- [21] 尹金良. 基于相关向量机的油浸式电力变压器故障诊断方法研究[D]. 北京: 华北电力大学, 2013.
- YIN Jinliang. Study on oil-immersed power transformer fault diagnosis based on relevance vector machine[D]. Beijing: North China Electric Power University, 2013.
- [22] 华丁剑. 基于支持向量机的油浸式变压器故障诊断研究[D]. 长沙: 长沙理工大学, 2012.
- HUA Dingjian. Study on oil-immersed power transformer fault diagnosis based on support vector machine[D]. Changsha: Changsha University of Science & Technology, 2012.
- [23] 曾成碧, 蒲维, 曾先鋒. BP 神经网络在油浸式变压器过热性故障诊断中的应用[J]. 四川电力技术, 2012, 35(4): 60-63.
- ZENG Chengbi, PU Wei, ZENG Xianfeng. Study on overheat fault of oil-immersed transformer based on BP neural network[J]. Sichuan Electric Power Technology, 2012, 35(4): 60-63.

作者简介:



张又文

张又文(1996—),女,河南郑州人,硕士研究生,主要研究方向为大数据与人工智能在电力系统中的应用(E-mail: 21810164@zju.edu.cn);

冯斌(1997—),男,浙江杭州人,博士研究生,主要研究方向为大数据与人工智能在电力系统中的应用(E-mail: 11910028@zju.edu.cn);

陈页(1996—),女,江西南昌人,硕士研究生,主要研究方向为综合能源系统优化运行(E-mail: ye_chen@zju.edu.cn)。

(编辑 任思思)

Fault diagnosis method for oil-immersed transformer based on XGBoost optimized by genetic algorithm

ZHANG Youwen, FENG Bin, CHEN Ye, LIAO Weihang, GUO Chuangxin
(College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China)

Abstract: In order to improve the accuracy and reliability of fault diagnosis for oil-immersed transformer, the fault diagnosis method for oil-immersed transformer based on XGBoost (eXtreme Gradient Boosting) optimized by GA (Genetic Algorithm) is studied. Firstly, based on DGA (Dissolved Gas Analysis), the non-coding method is used to extract the 9-dimensional fault characteristics of oil-immersed transformer, and the data samples are normalized. The fault diagnosis model based on XGBoost is built with the normalized samples as inputs, and the hyperparameters in the model are simultaneously optimized by GA. In case study, 547 samples of DGA data determined by fault types are collected for comparison experiments. Results show that the diagnosis accuracy and stability of the proposed method are significantly improved compared with the existing traditional methods. The optimization effect of GA on the fault diagnosis model is verified at the same time.

Key words: dissolved gas analysis; electric transformers; fault diagnosis; XGBoost; genetic algorithms

附录 A

输入: I , 当前节点的样本编号集合

输入: d , 特征维度

gain $\leftarrow 0$

$$G \leftarrow \sum_{i \in I} g_i, H \leftarrow \sum_{i \in I} h_i$$

for $k=1$ to m do

$$G_L \leftarrow 0, H_L \leftarrow 0$$

for j in I do

$$G_L \leftarrow G_L + g_j, H_L \leftarrow H_L + h_j$$

$$G_R \leftarrow G - G_L, H_R \leftarrow H - H_L$$

$$\text{score} \leftarrow \max(\text{score}, \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{G^2}{H + \lambda})$$

end

end

输出: 得分最大的分裂点

附录 B

for $k=1$ to m do

对特征 k 按照百分比划分 $S_k = \{s_{k1}, s_{k2}, \dots, s_{kl}\}$

end

for $k=1$ to m do

$$G_{kv} \leftarrow \sum_{j \in \{j | s_{k,v} \geq x_{jk} \geq s_{k,v-1}\}} g_j$$

$$H_{kv} \leftarrow \sum_{j \in \{j | s_{k,v} \geq x_{jk} \geq s_{k,v-1}\}} h_j$$

end

其中, x_{jk} 为特征 k 中第 j 个样本数据。

附录 C

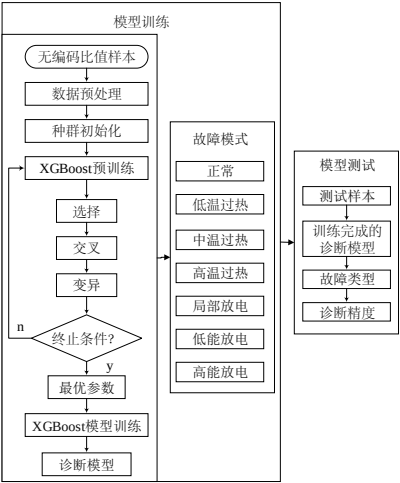


图 C1 诊断模型流程图
Fig.C1 Flowchart of diagnosis model

附录 D

表 D1 变压器异常数据

Fig.D1 Abnormal data of transformer

序号	油中溶解气体含量/ (uL L ⁻¹)				
	H ₂	CH ₄	C ₂ H ₆	C ₂ H ₄	C ₂ H ₂
1	31.0	23.0	10.6	7.1.0	0.4
2	160.0	130.0	33.0	96.0	0
3	78.0	56.0	18.0	21.0	0
4	181.0	262.0	210.0	528.0	0
5	27.0	90.0	42.0	63.0	0.2
6	20.6	19.9	7.4	61.3	1.5
7	70.4	68.5	28.9	241.0	10.4
8	650.0	53.0	34.0	20	0
9	176.0	205.9	47.7	75.7	68.7
10	115.9	75.0	14.7	25.5	6.8