

DOI:10.3969/j.issn.2097-0706.2024.01.010

基于 NNTR-SMOTE 与 GA-XGBoost 的变压器故障诊断方法研究

Transformer fault diagnosis method based on NNTR-SMOTE and GA-XGBoost

汪李忠¹, 池建飞¹, 丁叶强¹, 姚海燕², 唐志鹏², 吴同宇^{3*}
WANG Lizhong¹, CHI Jianfei¹, DING Yeqiang¹, YAO Haiyan²,
TANG Zhipeng², WU Tongyu^{3*}

(1. 国网浙江省电力有限公司杭州市临平区供电公司, 杭州 311199; 2. 杭州电力设备制造有限公司余杭群力成套电气制造分公司, 杭州 311100; 3. 东北电力大学 机械工程学院, 吉林 吉林 132012)

(1. Hangzhou Linping District Power Supply Company, State Grid Zhejiang Electric Power Company Limited, Hangzhou 311199, China; 2. Yuhang Qunli Complete Electric Manufacturing Branch, Hangzhou Electric Power Equipment Manufacturing Company Limited, Hangzhou 311100, China; 3. School of Mechanical Engineering, Northeastern Electric Power University, Jilin 132012, China)

摘要: 针对变压器故障诊断中故障样本数量少且分布不均导致诊断率低的问题, 提出了一种基于最近邻三角区域合成少数类过采样(NNTR-SMOTE)与利用遗传算法(GA)优化极端梯度提升(XGBoost)模型的变压器故障诊断方法。首先, 将采集到的变压器故障样本数据进行标准化处理, 使用 NNTR-SMOTE 方法得到平衡数据; 其次, 采用无编码比值法构造油中溶解气体的特征, 得到特征数据集并对特征数据集采用多维尺度分析(MDS)方法进行特征融合; 最后, 利用 GA 对 XGBoost 模型的参数进行优化, 构建变压器故障诊断模型。试验结果表明: 基于 NNTR-SMOTE 与 GA-XGBoost 的变压器故障诊断方法诊断准确率达 95.97%, 不仅解决了诊断模型对多数类的偏向问题, 还将模型的诊断精度进一步提高, 适用于变压器非均衡数据集的多分类故障诊断。

关键词: 变压器; 故障诊断; 不平衡小样本; 极端梯度提升; 遗传算法

中图分类号: TM 411 **文献标志码:** A **文章编号:** 2097-0706(2024)01-0084-10

Abstract: To address the low accuracy of transformer fault diagnosis caused by the insufficient number and uneven distribution of fault samples, a transformer fault diagnosis method based on nearest neighbour triangle regions synthetic minority oversampling technique (NNTR-SMOTE) and genetic algorithm optimized extreme gradient boosting (GA-XGBoost) is proposed. Firstly, the transformer fault sample data are collected and standardized, and then balanced data are obtained by NNTR-SMOTE. Secondly, the feature data of dissolved gas are categorized by non-coding ratio method, and then fused by multi-dimensional scaling (MDS) method. Finally, a new transformer fault diagnosis model is constructed based on the XGBoost model optimized by GA. The experimental results show that the diagnostic accuracy of the proposed diagnosis method based on NNTR-SMOTE and GA-XGBoost reaches as high as 95.97%. This method not only solves the bias towards the majority class during diagnosis modelling, but also improves the diagnostic accuracy of the model, making it suitable for multi-classification fault diagnosis for transformers with unbalanced data.

Keywords: transformer; fault diagnosis; unbalanced small sample; extreme gradient boosting; genetic algorithm

0 引言

油浸式变压器在输变电系统中具有至关重要的作用, 其健康状态直接决定着电力系统的安全与稳定^[1]。然而, 由于不能及时了解变压器的健康状

态, 变压器运行年限较长后, 存在绝缘老化等故障隐患^[2]。

油浸式变压器绝缘老化后, 会产生 H_2 , CH_4 , C_2H_4 , C_2H_2 以及 C_2H_6 等气体溶解在绝缘油中, 油中溶解气体的含量变化可以反映出变压器不同的故障类型^[3-4]。目前广泛采用的三比值法^[5]、大卫三角形法^[6]以及罗杰斯比值法^[7]等方法都存在编码不足、阈值不够精确以及界限区分绝对化等局限性。无

基金项目: 吉林省科技发展计划项目(20220508014RC)
Jilin Province Science and Technology Development Plan Project (20220508014RC)

编码比值法相比于传统的比值法具有更丰富的特征,可以有效弥补三比值法的编码缺陷,从而提高诊断精确度^[8]。

随着机器学习技术的发展,研究人员逐渐将机器学习引入到变压器故障诊断中。机器学习需要大量训练样本支撑,但实际工作中变压器发生故障的概率很低且不同故障发生的概率差异较大,导致采集的样本数量较少且分布不均衡,这种不平衡小样本难以满足智能诊断方法参数寻优及训练的需求,在识别过程中易造成对少数类样本误判问题。因此,数量充足且均衡的样本是变压器故障准确诊断的关键^[9-10]。

与传统模型相比,机器学习中一类以反向传播(Back Propagation, BP)神经网络^[11-13]、有向无环图模型^[14-15]和支持向量机(Support Vector Machine, SVM)^[16-18]为代表的非集成学习方法有效提升了模型的精确度以及运算速度,但存在一定的缺陷:BP神经网络学习效率高,但容易出现局部最佳解,而且还需要对模型的各项参数做出复杂的调节;SVM作为二分类器,在处理多分类问题时显得相对乏力^[19];有向无环图模型对条件属性有较高的要求。

另一类是以梯度提升树(Gradient Boosting Decision Tree, GBDT)模型和随机森林(Random Forest, RF)模型为代表的集成学习模型。RF作为一种组合模型,以多个决策树为基分类器,从而减小方差并有效提升预测准确度,但当其面临噪音干扰时,就会出现过拟合情况;GBDT可通过减小偏差来降低诊断系统总误差,并且不需要进行复杂的参数调节,但其个体学习器之间存在很强的依赖关系,使得它的训练效率相比RF略逊一筹。

文献[20]提出了极端梯度提升(Extreme Gradient Boosting, XGBoost)算法,它与GBDT的区别在于,XGBoost为了有效提升模型的泛化能力并减少过拟合情况,在损失函数中添加了正则化项用以施加惩罚,以此来限制模型的复杂程度;同时,通过分布式计算与并行计算对模型的训练速度进行有效提升。

目前,XGBoost已在多个领域取得成效并得到广泛认可。文献[21]将XGBoost与多特征评估结合应用于高压断路器故障诊断,有效实现了对高压断路器机械振动信号的准确分类。文献[22]将XGBoost应用于无瓣海桑红树林地上生物量反演并表现出较好的应用能力。文献[23]将XGBoost应用于无线传感器网络冗余数据检测算法,仿真分析结果表明,该方法具有较好的检测效果,能够有效提升无线传感器网络运行的安全性。

基于上述研究,本文提出基于最近邻三角区域合成少数类过采样(The Nearest Neighbor Triangle Regions Synthetic Minority Oversampling Technique, NNTR-SMOTE)技术与遗传算法(Genetic Algorithm, GA)优化XGBoost模型的变压器故障诊断方法。首先,将采集到的变压器故障样本数据进行标准化处理,使用NNTR-SMOTE方法得到平衡数据;其次,采用无编码比值法构造油中溶解气体特征,得到特征数据集并对特征数据集采用多维尺度(Multidimensional Scaling, MDS)分析方法进行特征融合,消除冗余信息;最后,利用GA算法对XGBoost模型的参数进行优化,构建变压器故障诊断模型。

1 NNTR-SMOTE

当前,解决不平衡数据集的训练问题主要采用欠采样或过采样方法。欠采样算法易造成有效信息的丢失,较少应用于变压器故障诊断中^[24]。过采样算法是通过生成少数类样本来平衡各故障类别的样本数目,目前应用最为广泛的是SMOTE方法^[25],但该方法生成样例存在一定的缺陷,即生成的人工样例有一定概率位于类别界线上,甚至跨过类别界线入侵多数类样例的领域,如图1a所示。为了避免这种状况,研究人员提出了一种NNTR-SMOTE方法,该方法通过在最近的少数类样本形成的三角区域内构建少量样本,以达到避免出现跨界样本的目的,基于NNTR-SMOTE生成样例如图1b所示,具体公式为

$$y_{\text{new}} = y_2 + \text{rand}(0, 1) [y_0 + \text{rand}(0, 1) (y_1 - y_0) - y_2], (1)$$

式中: y_{new} 为合成的新的少数类样本; y_0 为随机选取的少数类样例; y_1, y_2 均为 y_0 的2个近邻少数类样例。

全部新的样例都生成后,再检测其最近邻样例的类别,如果为同类则保留,如果为异类则删除。

2 MDS分析方法

2.1 MDS原理

MDS的原理是利用成对样本之间的相似度来提取低维空间特征,使得该低维空间中,各样本之间的距离与其在高维空间中的距离保持高度的相似性。

一般来说,MDS降维技术的实现过程包括3个基本步骤:创建距离矩阵、计算内积矩阵和构建低维矩阵。

步骤1:计算 m 个 d 维样本之间的欧氏距离 σ ,生成距离矩阵 Δ 。假设有 m 个 d 维样本,样本间的

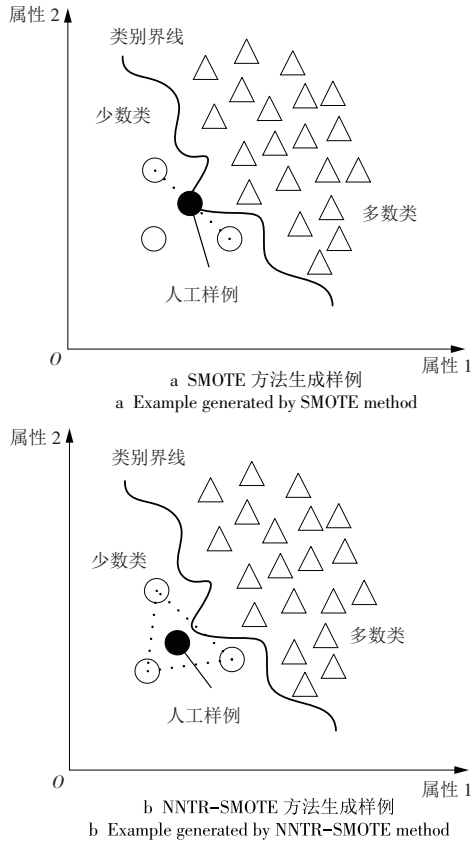


图 1 生成样例示意

Fig. 1 Sample extraction process

欧式距离定义为

$$\Delta = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{md} \end{bmatrix}, \quad (2)$$

$$\sigma_{ij} = x_i - x_j = \sum_{p=1}^d \sqrt{(r_{ip} - r_{jp})^2}, \quad (3)$$

式中: σ_{ij} 为第 i 个样本和第 j 个样本间的距离; x_i, x_j 分别为随机选取的第 i 个、第 j 个样本; r_{ip}, r_{jp} 分别为 x_i, x_j 中的第 p 维特征向量。

步骤 2: 计算 Δ 的双重中心化矩阵 $\hat{\Delta}$ 。

$$\hat{\Delta} = -\frac{1}{2} J \Delta^{(2)} J = X X^T, \quad (4)$$

式中: $\Delta^{(2)} = \sigma_{ij}^2$; $J = E - \frac{1}{m}$, E 为单位矩阵; X 为 m 个 d 维样本组成的样本矩阵。

则 $\hat{\Delta}$ 表达式为

$$\hat{\Delta} = -\frac{1}{2} (\sigma_{ij}^2 - \frac{1}{m} \sum_{j=1}^m \sigma_{ij}^2 - \frac{1}{m} \sum_{i=1}^m \sigma_{ij}^2 + \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m \sigma_{ij}^2) = x_i x_{j0} \quad (5)$$

步骤 3: 对内积矩阵 $\hat{\Delta}$ 进行特征值分解, 计算出 k 个特征值及其特征向量, 将特征值由高至低依次排列并求出降维后的样本 X' 。

由于矩阵 $\hat{\Delta}$ 具有对称性和半正定性, 可以通过

奇异值分解对其进行处理, 从而得到

$$\hat{\Delta} = U \Lambda U^T = U \Lambda^{\frac{1}{2}} \Lambda^{\frac{1}{2}} U^T, \quad (6)$$

式中: Λ 为 $\hat{\Delta}$ 的特征值组成的矩阵; U 为 $\hat{\Delta}$ 的特征向量组成的矩阵。

为了更好地表达这 2 个矩阵, 需要按照 $\hat{\Delta}$ 的特征值由高至低依次排列。再通过式 (7) 结合前 a 个较大的特征值及其特征向量推算出降维后的样本 X' 。

$$X' = U \Lambda^{\frac{1}{2}}. \quad (7)$$

2.2 评价指标

评价降维能力时, 通常采用降维时间、降维损失以及降维质量作为评价指标。其中: 降维损失 S_{stress} 可用压力函数来表示, 具体数学表达式见式 (8); 降维质量 L_{LCMC} 可通过局部连续性元标准 (Local Continuity Meta-Criterion, LCMC) 来评价, 具体数学表达式见式 (9)。

$$S_{\text{stress}} = \sqrt{\frac{\sum_{i=1, j=1, i \neq j}^n (d_{ij} - \bar{d}_{ij})^2}{\sum_{i=1, j=1, i \neq j}^n d_{ij}^2}}, \quad (8)$$

$$L_{\text{LCMC}}(M) = \frac{1}{MN} \sum_{i,j \in 1, \dots, M} q_{i,j} - \frac{M}{N-1}, \quad (9)$$

式中: d_{ij} 为降维后的第 i, j 个对象间的空间距离; $\bar{d}_{ij}$ 为降维后的第 i, j 个对象间的平均空间距离; $q_{i,j}$ 为重叠数; M 为邻域的大小; N 为样本数。

LCMC 通过检查相邻的数据样本集合与其相应重叠镶嵌程度来评估降维算法的性能^[26]。输出参数包括具有最高 LCMC 值的数据坐标左侧的平均值 Q_L 、具有最高 LCMC 值的数据坐标右侧的平均值 Q_G 和 Q_L 及 Q_G 的平均值 Q_{LG} 。 Q_L 和 Q_G 越接近 1 表示降维质量越好, S_{stress} 越接近 0 表示拟合高维效果越好, 即降维损失越小。

3 GA 优化的 XGBoost 模型

3.1 XGBoost 原理

XGBoost 使用多棵分类与回归树 (Classification and Regression Tree, CART) 作为弱评估器, 并将它们通过梯度提升的方式进行集成。具体 XGBoost 模型构建如下。

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

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F, \quad (10)$$

式中: $\hat{y}_i$ 为第 i 个样本 x_i 的诊断结果; f_k 为第 k 个评估

器; F 为所有弱评估器的集合。

为对模型中的函数集合进行学习,最小化以下正态分布标准化目标。

$$L(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad (11)$$

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda w^2, \quad (12)$$

式中: ϕ 为模型中的函数集合; $\sum_i l(\hat{y}_i, y_i)$ 为损失函数,表示实际情况与 CART 预测情况的差异性; $\sum_k \Omega(f_k)$ 为正则项,通过其限制模型的复杂程度; γ, λ 均为调节模型复杂性的正态分布标准化参数,它们的数值越高,模型就更难过拟合; T 为叶子节点总数; w 为叶子权重。

通过引入正则项与损失函数,可以有效防止出现过拟合的情形。

由于传统方法具有局限性,无法优化上述以函数作为参数的目标函数,因此选择了一种更有效的方式——加法学习。一般情况下,模型学习由一个常数开始训练预测能力,在每轮迭代中添加新的函数对当前模型进行学习,则目标函数可以表示为

$$L^{(t)} = \sum_{i=1}^n (l(y_i, \hat{y}_i^{(t-1)}) + f_i(x_i)) + \Omega(f_i), \quad (13)$$

式中: $\hat{y}_i^{(t-1)}$ 为第 i 个样本在第 $t-1$ 次训练中的预测值。

为快速优化目标函数,对损失函数 $L^{(t)}$ 二阶泰勒展开,可得

$$L^{(t)} = \sum_{i=1}^n \left(l(y_i, \hat{y}_i^{(t-1)}) + g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right) + \Omega(f_i), \quad (14)$$

式中: g_i 为损失函数的一阶梯度; h_i 为二阶梯度。

将常数项移除,可以得到第 t 次训练的简化目标函数为

$$\tilde{L}^{(t)} = \sum_{i=1}^n \left(g_i f_i(x_i) + \frac{1}{2} h_i f_i^2(x_i) \right) + \Omega(f_i). \quad (15)$$

定义叶子节点 j 中的样本编号集合,以便更好地理解和分析数据。

$$I_j = \{i | q(x_i) = j\}, \quad (16)$$

式中: $q(x_i)$ 为 x_i 对应的叶子标签值。

最终得到叶子节点 j 的最优权重为

$$w_j^* = - \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}. \quad (17)$$

根据式(17),可以得出最优目标函数为

$$\tilde{L}^{(t)}(q) = - \frac{1}{2} \sum_{j=1}^T \frac{\left(\sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \quad (18)$$

3.2 基于 GA 优化的模型训练

本文将 GA^[27] 引入变压器故障诊断模型中,有效解决了传统优化方法,如交叉验证和网格搜索参数寻优的缺陷,并能够同时优化多个超参数,从而大大提升了变压器故障诊断的准确性。

在模型训练过程中,迭代次数、学习率 η 、决策树的最大深度 $d_{\max}$ 、随机样本的抽取比例 r_{ss} 、特征的抽取比例 r_{cs} 、决策树节点分裂标准 γ_s 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。

综上所述,GA 优化 XGBoost 模型的具体实现步骤如下。

- (1) 将降维后样本数据按一定比例分为训练集和测试集。
- (2) 利用训练集中数据构建 XGBoost 模型。
- (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(3)。
- (5) 测试模型的诊断效果,输出故障分类结果。

故障诊断流程如图 2 所示。

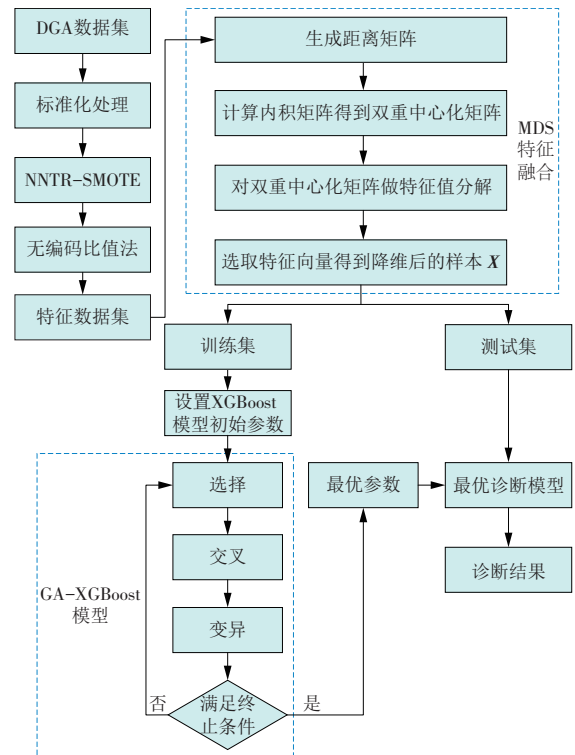


图 2 故障诊断流程

Fig. 2 Fault diagnosis process

3.3 评价指标

传统变压器故障诊断中,通常采用准确率作为

单一的诊断结果,无法区分误判与漏判,因此引入 2 类问题(正类、负类)的混淆矩阵,见表 1,表中: n_{TP} 表示样本本身为正类,分类器预测仍为正类; n_{FP} 表示样本本身为负类,分类器预测为正类; n_{TN} 表示样本本身为负类,分类器预测也为负类; n_{FN} 表示样本本身为正类,分类器预测为负类。

表 1 混淆矩阵

Table 1 Fusion matrix

项目	预测正类	预测负类
实际正类	n_{TP}	n_{FN}
实际负类	n_{FP}	n_{TN}

基于混淆矩阵分类,本文引入查全率 R_{re} 、查准率 R_{pr} 、误报率 R_{in} 、漏报率 R_{fp} 及 Kappa 系数 k_{Kappa} 作为变压器故障诊断综合评价指标。

$$R_{re} = \frac{n_{TP}}{n_{TP} + n_{FN}}, \quad (19)$$

$$R_{pr} = \frac{n_{TP}}{n_{TP} + n_{FP}}, \quad (20)$$

$$R_{in} = \frac{n_{FN}}{n_{TN} + n_{FN}} = 1 - R_{re}, \quad (21)$$

$$R_{fp} = \frac{n_{FP}}{n_{TP} + n_{FP}} = 1 - R_{pr}, \quad (22)$$

$$k_{Kappa} = \frac{P_0 - P_e}{1 - P_e}, \quad (23)$$

式中: P_0 为混淆矩阵对角线元素之和除以整个矩阵元素之和; P_e 为混淆矩阵第 i 行元素之和乘第 i 列元素之和再除以整个矩阵元素之和。

通常,Kappa 系数在作为模型评价指标时,计算值越靠近 1 表明模型的诊断效果越好。

4 算例分析

4.1 样本数据分布

本文收集某供电公司的变压器故障样本 425 例,样本的类别标签、数量以及占比见表 2。其中,局部放电类、中低温过热类和放电兼过热类样本数量占总样本数的 6.82%,4.00%,7.77%,体现了样本的不均衡性。

4.2 数据预处理及特征构造

能够反映变压器故障类型的气体主要包括 H_2 , CH_4 , C_2H_4 , C_2H_2 和 C_2H_6 ,这些气体的含量差异很大,如果直接将它们作为样本进行检测,可能会对诊断精度造成不利影响^[28]。因此,本文采用式(24)对采集到的样本数据进行标准化处理。

$$x' = \frac{x - \mu}{\sigma}, \quad (24)$$

式中: μ 为 x 数据集的均值; σ 为数据集元素的标准差; x' 为 x 标准化之后的结果。

表 2 样本类别标签及分布

Table 2 Categories of samples and their propositions

状态类型	样本数量	占比/%	类别标签
正常状态	91	21.41	1
中温过热	46	10.82	2
中低温过热	33	7.77	3
高温过热	53	12.47	4
放电兼过热	17	4.00	5
局部放电	29	6.82	6
低能放电	72	16.95	7
高能放电	84	19.76	8

对标准化处理后的故障样本数据采用无编码比值法构造油中溶解气体的 18 维特征,具体的油中溶解气体特征与各特征量对应的特征编码见表 3,表中:THC = $CH_4 + C_2H_2 + C_2H_4 + C_2H_6$, ALL = $H_2 + CH_4 + C_2H_2 + C_2H_4 + C_2H_6$ 。

表 3 油中溶解气体特征编码及特征量

Table 3 Characteristic quantities of dissolved gases and their codes

特征编码	特征量	特征编码	特征量
1	CH_4/H_2	10	C_2H_4/THC
2	C_2H_2/H_2	11	C_2H_6/THC
3	C_2H_2/C_2H_4	12	C_2H_2/THC
4	C_2H_4/C_2H_6	13	$(CH_4+C_2H_4)/THC$
5	C_2H_6/CH_4	14	H_2/ALL
6	C_2H_2/CH_4	15	CH_4/ALL
7	C_2H_4/CH_4	16	C_2H_2/ALL
8	H_2/THC	17	C_2H_4/ALL
9	CH_4/THC	18	C_2H_6/ALL

4.3 数据平衡处理

非均衡数据集会导致诊断结果更偏向于多数类,致使少数类的识别精度低,从而导致故障诊断准确率降低。因此,本文采用 NNTR-SMOTE 算法得到相对平衡的样本数据。处理前后的样本数量对比见表 4。

表 4 样本数量处理前后对比

Table 4 Sample size before and after processing

运行状态	处理前数量	处理后数量
正常状态	91	92
中温过热	46	92
中低温过热	33	93
高温过热	53	93
放电兼过热	17	93
局部放电	29	93
低能放电	72	93
高能放电	84	90

为直观展现出 NNTR-SMOTE 方法可以在保留少数类样本数据特征的前提下扩充样本数量,按类别采用 t 分布随机近邻嵌入^[29] (t-Distributed Stochastic Neighbor Embedding, t-SNE) 将采样前后数据映射到二维空间进行可视化处理,如图 3 所示。

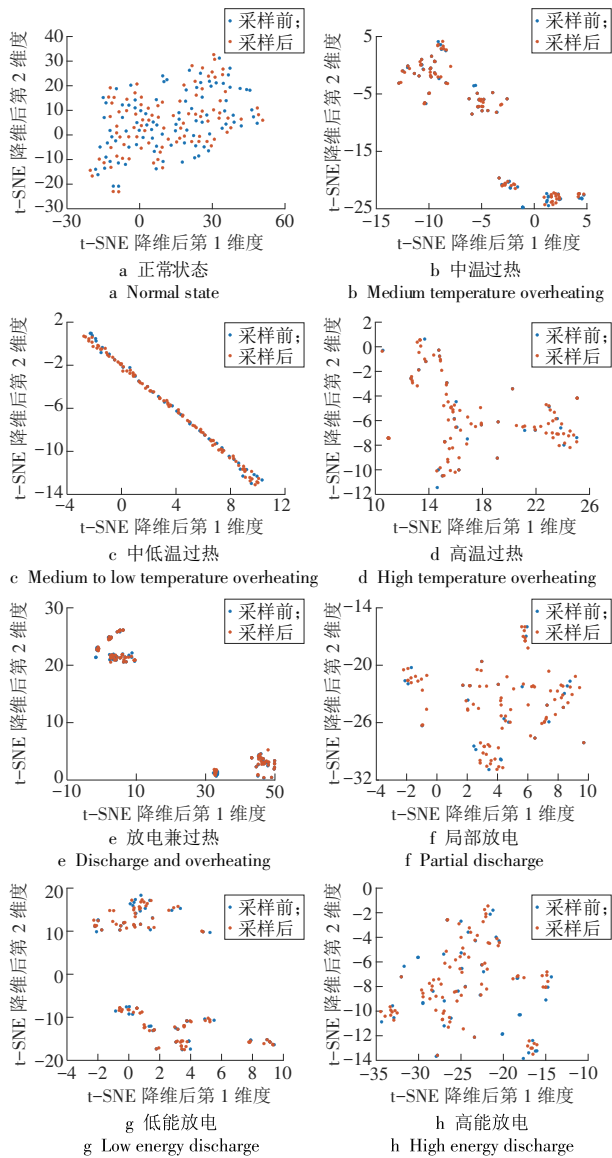


图3 不同故障类型样本数据平衡化前后分布趋势

Fig. 3 Distributions of sample data of different fault types before and after equalization

由图 3 可知,采用 NNTR-SMOTE 方法前、后,不同类型故障的数据分布趋势基本相同,表明本文采用的方法能较好地保留原始数据的特征。

4.4 数据特征融合

为避免故障特征中存在冗余信息,同时减少计算量,需采用 MDS 对构建的 18 维联合特征进行特征融合。MDS 对数据进行降维时要设定降维后的维数,因此对经设定降维后维数为 1—6 的数据进行对比,结果见表 5。

表 5 不同维数降维指标

Table 5 Dimension reduction indicators of different dimensionality

降维指标	6 维	5 维	4 维	3 维	2 维	1 维
降维时间/s	2.46	2.48	2.53	2.57	2.59	2.64
S_{stress}	0.014 7	0.015 3	0.016 4	0.017 8	0.019 1	0.021 6
Q_L	0.825 2	0.832 6	0.844 3	0.857 2	0.862 5	0.871 4
Q_G	0.403 2	0.416 4	0.422 0	0.430 2	0.438 5	0.449 2
Q_{LG}	0.614 2	0.624 5	0.633 2	0.643 7	0.650 5	0.660 3
准确率/%	91.84	92.36	93.52	95.45	95.97	90.63

由表 5 可见:由于 4—6 维降维质量不断提升,因此诊断准确率呈上升趋势;2,3 维的准确率达到最高且趋于平缓;1 维时由于降维损失过多导致准确率呈下降趋势。此外,对数据进行降维时要求降维质量高的同时要求降维损失小,因此以降维损失 $S_{\text{stress}} < 0.02$,降维质量 $Q_{LG} < 0.65$ 作为选择标准,最终设定 MDS 降维后维数为 2。

4.5 故障诊断结果

将经过 MDS 降维处理后的数据 70% 划分为训练集,30% 划分为测试集,具体分配情况见表 6。

表 6 降维后的数据分配

Table 6 Data allocation after dimension reduction

运行状态	样本总数	训练集样本数	测试集样本数
正常状态	92	64	28
中温过热	92	64	28
中低温过热	93	65	27
高温过热	93	65	27
放电兼过热	93	65	27
局部放电	93	65	27
低能放电	93	65	27
高能放电	90	64	28

XGBoost 的初始参数预设为:特征的抽取比例 $r_{\text{cs}} = 1$,决策树的最大深度 $d_{\text{max}} = 6$,学习率 $\eta = 0.3$,决策树节点分裂标准 $\gamma_s = 0$,随机样本的抽取比例 $r_{\text{ss}} = 1$ 。当得到最大迭代次数时,得到最优 GA-XGBoost 分类器,最后将测试集输入最优分类器,得到混淆矩阵,如图 4 所示。

基于混淆矩阵中数据,计算出各状态的查全率、查准率、误报率、漏报率与诊断模型的 Kappa 系数,见表 7。

由图 4 可知:测试集 149 个样本中被正确诊断的为 143 个,总正确率为 95.97%;中低温过热、局部放电、高能放电故障诊断准确率为 100%;诊断模型的 Kappa 系数为 0.954 2,进一步验证了本文所提模型故障识别准确率高。

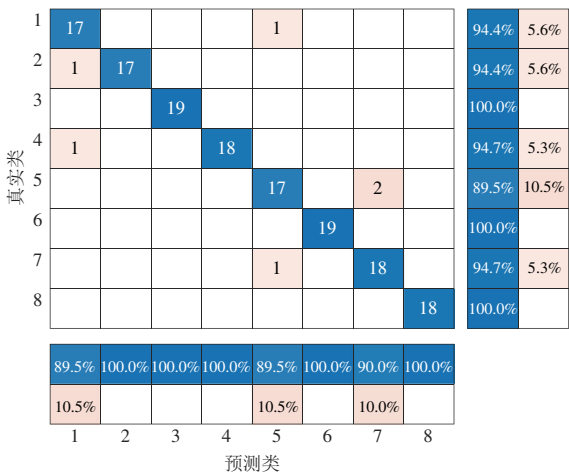


Fig. 4 Confusion matrix of fault diagnosis classification

表 7 GA-XGBoost 模型评价指标

Table 7 Evaluation indicators of the GA-XGBoost model

运行状态	查全率/%	查准率/%	误报率/%	漏报率/%	Kappa 系数	准确率/%
正常状态	94.4	89.5	5.6	10.5	0.954 2	95.97
中温过热	94.4	100.0	5.6	0		
中低温过热	100.0	100.0	0	0		
高温过热	94.7	100.0	5.3	0		
放电兼过热	89.5	89.5	10.5	10.5		
局部放电	100.0	100.0	0	0		
低能放电	94.7	90.0	5.3	10.0		
高能放电	100.0	100.0	0	0		

5 分析与讨论

5.1 不同数据均衡算法对比分析

将采集到的原始样本数据与分别采用 NNTR-SMOTE, SMOTE- 编辑最近邻 (Edited Nearest Neighbours, ENN), 自适应合成抽样 (Adaptive Synthetic Sampling, ADASYN), SMOTE 和随机过采样处理后的数据应用于故障诊断中, 最终的诊断结果见表 8。

表 8 不同采样方法的诊断结果

Table 8 Diagnostic results under different sampling methods

采样方法	准确率/%	Kappa 系数
NNTR-SMOTE	95.97	0.954 2
SMOTE-ENN	93.26	0.939 5
ADASYN	91.75	0.913 3
SMOTE	90.56	0.896 1
随机过采样	89.32	0.884 9
非均衡数据集	86.32	0.854 2

由表 8 可见: 由于不平衡小样本难以满足智能诊断方法参数寻优及训练的需求, 在识别过程中造成了对少数类样本的误判, 因此未经平衡处理的非

均衡数据集诊断准确率与 Kappa 系数最低; 使用不同过采样方法后, 诊断准确率与 Kappa 系数均有不同程度的提升, 但由于随机过采样随机复制少数类样本, 造成了过拟合现象; ADASYN 易受离群点的影响, 会生成多余样本; NNTR-SMOTE 相比于其他 SMOTE 及其改进方法, 避免了生成的样例位于甚至跨过类别界线侵入多数类样例领域的问题; NNTR-SMOTE 方法的诊断准确率和 Kappa 系数最高。

结果表明: NNTR-SMOTE 有效解决了数据分布不均衡的情况且效果优于其他过采样方法, 避免了非均衡数据集导致诊断结果更偏向于多数类, 从而降低故障诊断准确率的情况。

5.2 不同特征提取方法诊断结果对比分析

本文对 MDS 提取的融合特征、PCA 融合特征、线性判别分析 (Linear Discriminant Analysis, LDA) 融合特征、局部线性嵌入 (Locally Linear Embedding, LLE) 融合特征以及 18 维联合特征输入后的诊断结果进行对比。诊断结果如图 5 所示, 诊断结果及对应的 Kappa 系数见表 9。

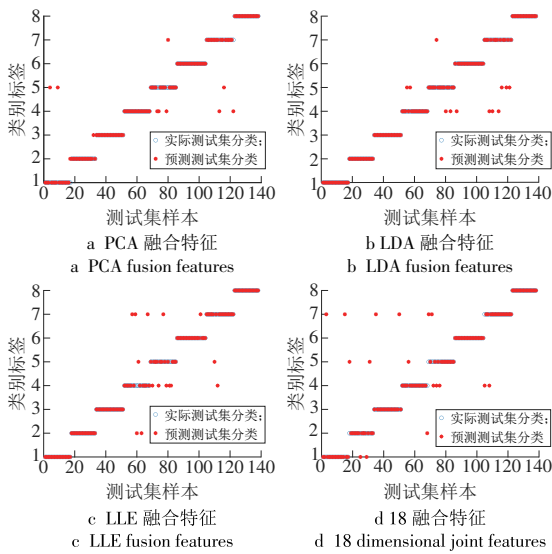


图 5 不同特征输入的诊断结果对比

Fig. 5 Diagnostic results with different feature inputs

表 9 不同特征的 Kappa 系数对比

Table 9 Kappa coefficients for different features

评价指标	MDS 融合特征	PCA 融合特征	LDA 融合特征	LLE 融合特征	18 维联合特征
诊断准确率/%	95.97	92.02	91.30	89.13	87.68
Kappa 系数	0.954 2	0.937 3	0.912 6	0.900 2	0.875 4

由表 9 可见: 由于未降维数据中含有冗余信息, 因此 18 维联合特征的诊断准确率与 Kappa 系数最低; 相比于 18 维联合特征, 其余 4 种特征诊断准确率与 Kappa 系数均有不同程度的提升, 但由于多维数据之间存在非线性关系, PCA 算法无法很好地排

除冗余数据;小样本问题导致边缘数据影响 LDA 算法对样本中心点的选择,从而影响降维效果;LLE 提取的是局部样本数据信息,对全局结构的识别不够全面;MDS 算法的诊断准确率和 Kappa 系数最高。

结果表明,MDS 算法可有效去除冗余信息且效果优于其他降维算法,有效提升故障诊断准确率。

5.3 不同故障诊断模型对比分析

为验证 GA-XGBoost 的分类识别性能,将 MDS 特征提取后数据输入至文献[30]提出的粒子群优化双向长短期记忆(Particle Swarm Optimization-Bidirectional Long Short-Term Memory, PSO-BiLSTM)、文献[31]提出的贝叶斯优化支持向量机(Bayesian Optimization Support Vector Machine, BO-SVM)和文献[32]提出的鲸鱼优化算法优化支持向量机(Whale Optimization Algorithm-Support Vector Machine, WOA-SVM)诊断模型中,诊断结果如图 6 所示。模型评价指标对比见表 10。

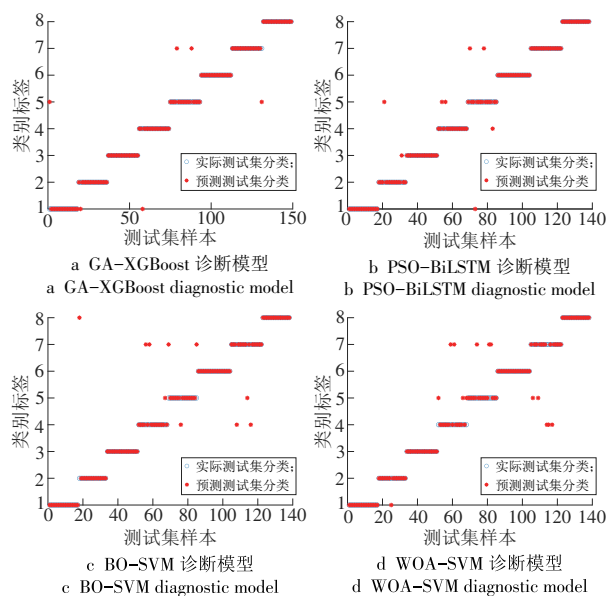


图 6 不同模型诊断结果对比
Fig. 6 Diagnostic results with different diagnosis models

表 10 模型评价指标对比

Table 10 Evaluation indicators of different models				
评价指标	GA-XGBoost	PSO-BiLSTM	BO-SVM	WOA-SVM
查全率	0.944 5	0.937 6	0.913 7	0.893 3
查准率	0.943 2	0.936 2	0.914 1	0.895 6
Kappa 系数	0.954 2	0.946 3	0.928 4	0.906 9
诊断准确率/%	95.97	94.20	92.75	89.85

结果表明:在输入特征相同的情况下,GA-XGBoost 诊断模型相比 PSO-BiLSTM,BO-SVM 和 WOA-SVM 诊断模型准确率均有所提升;从查全率和查准率等评价指标方面看,本文所提方法相较于

其他模型表现较为稳定;从 Kappa 系数角度看,本文所提方法达到了 0.954 2,进一步说明了本文所提特征提取方法和故障诊断模型的有效性。

6 结论

(1)本文所采用的 NNTR-SMOTE 少数类过采样方法能够有效解决变压器故障诊断中样本不平衡导致诊断准确率低的问题。

(2)利用 MDS 进行特征融合,有效避免了冗余信息出现,同时减少了计算量,进一步提高了故障诊断模型的精度。

(3)相比于其他诊断模型,本文所提 GA-XGBoost 模型准确率达到 95.97%,Kappa 系数达到 0.954 2,表明 GA-XGBoost 模型具有更好的准确性。

参考文献:

[1]YAN P C, ZHANG C Y, MEI K F, et al. Research on fault diagnosis of transformer based on laser induced fluorescence technology [J]. Journal of Molecular Structure, 2022, 1258: 132645.

[2]廖伟涵,郭创新,金宇,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 Weihao, 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.

[3]MERVE D, HALUK G, CENGİZ M T. Improvement of power transformer fault diagnosis by using sequential Kalman filter sensor fusion [J]. International Journal of Electrical Power and Energy Systems, 2023, 149:109038.

[4]肖宏磊,留毅,夏红军,等. 基于PCA与SSA-LightGBM的油浸式变压器故障诊断方法[J]. 综合智慧能源, 2023, 45(3): 9-16.

XIAO Honglei, LIU Yi, XIA Hongjun, et al Oil-immersed transformer fault diagnosis method based on PCA and SSA-LightGBM[J]. Integrated Intelligent Energy, 2023, 45(3): 9-16.

[5]Mineral oil-impregnated electrical equipment in service-guide to the interpretation of dissolved and free gases analysis: IEC 60599—2007[S].

[6]DAI J J, SONG H, SHENG G H, et al. Dissolved gas analysis of insulating oil for power transformer fault diagnosis with deep belief network [J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2017, 24 (5) : 2828-2835.

[7]ROGERS R R. IEEE and IEC codes to interpret incipient faults in transformers, using gas in oil analysis [J]. IEEE

- Transactions on Electrical Insulation, 1978, 13 (5): 349-354.
- [8] 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.
- [9] QIN J, YANG D S, WANG N, et al. Convolutional sparse filter with data and mechanism fusion: A few-shot fault diagnosis method for power transformer[J]. Engineering Applications of Artificial Intelligence, 2023, 124: 106606.
- [10] 龙思成, 黄志鸿. 基于多尺度极限融合网络的电力变压器故障诊断方法研究[J]. 综合智慧能源, 2022, 44(9): 78-83.
- LONG Sicheng, HUANG Zhihong. Research on fault diagnosis method of power transformers based on multi-scale extreme fusion network[J]. Integrated Intelligent Energy, 2022, 44(9): 78-83.
- [11] 朱永利, 吴立增, 李雪玉. 贝叶斯分类器与粗糙集相结合的变压器综合故障诊断[J]. 中国电机工程学报, 2005, 25(10): 159-165.
- ZHU Yongli, WU Lizeng, LI Xueyu. Synthesized diagnosis on transformer faults based on Bayesian classifier and rough set[J]. Proceedings of the Csee, 2005, 25 (10): 159-165.
- [12] 姜寅, 任荭葳, 陈凯, 等. 基于 BP 神经网络的风电系统控制器 I/O 硬件故障自诊断方法[J]. 华电技术, 2020, 42 (5): 55-60.
- JIANG Yin, REN Hongwei, CHEN Kai, et al. Self-detecting method for faults in wind turbine controller I/O hardware based on BP neural network algorithm[J]. Huadian Technology, 2020, 42(5): 55-60.
- [13] 李金阔, 王秀平. 基于循环 BP 模型的变压器状态数据清洗方法[J]. 电力学报, 2023, 38(2): 101-109.
- LI Jinkuo, WANG Xiuping. Online detection method of transformer capacity based on cyclic BP model[J]. Journal of Electric Power, 2023, 38(2): 101-109.
- [14] 李勇琦, 雷旗开, 王浩, 等. 基于 BP 神经网络的梯次利用电池健康状态诊断[J]. 华电技术, 2021, 43(7): 42-46.
- LI Yongqi, LEI Qikai, WANG Hao, et al. State of health estimation for echelon-used batteries based on BP neural network[J]. Huadian Technology, 2021, 43(7): 42-46.
- [15] 王永强, 律方成, 李和明. 基于粗糙集理论和贝叶斯网络的电力变压器故障诊断方法[J]. 中国电机工程学报, 2006, 26(8): 137-141.
- WANG Yongqiang, LV Fangcheng, LI Heming. Synthetic fault diagnosis method of power transformer based on rough set theory and Bayesian network[J]. Proceedings of the CSEE, 2006, 26 (8): 137-141.
- [16] 陈欢, 彭辉, 舒乃秋, 等. 基于蝙蝠算法优化最小二乘双支持向量机的变压器故障诊断[J]. 高电压技术, 2018, 44(11): 3664-3671.
- CHEN Huan, PENG Hui, SHU Naiqiu, et al. Fault diagnosis of transformer based on LS-TSVM optimized by bat algorithm[J]. High Voltage Engineering, 2018, 44 (11): 3664-3671.
- [17] 魏伟, 高赐威, 宋梦, 等. 基于 HHO-SVM 的电力线路故障检测分类系统[J]. 综合智慧能源, 2023, 45(3): 1-8.
- WEI Wei, GAO Ciwei, SONG Meng, et al. Power line fault detection and classification system based on HHO-SVM[J]. Integrated Intelligent Energy, 2023, 45(3): 1-8.
- [18] 李赢, 舒乃秋. 基于模糊聚类 and 完全二叉树支持向量机的变压器故障诊断[J]. 电工技术学报, 2016, 31(4): 64-70.
- LI Ying, SHU Naiqiu. Transformer fault diagnosis based on fuzzy clustering and complete binary tree support vector machine[J]. Transactions of China Electrotechnical Society, 2016, 31(4): 64-70.
- [19] ZHOU Y, LIU Y Y, WANG N, et al. Partial discharge ultrasonic signals pattern recognition in transformer using BSO-SVM based on microfiber coupler sensor[J]. Measurement, 2022, 201: 111737.
- [20] CHEN T, TONG H. Higgs boson discovery with boosted trees [C]// International Conference on High-energy Physics & Machine Learning, 2015: 69-80.
- [21] 裴瑶, 武建文, 马速良, 等. 基于多特征评估与 XGBoost 的高压断路器故障诊断[J]. 高压电器, 2023, 59(4): 1-9.
- CHANG Yao, WU Jianwen, MA Suliang, et al. Fault diagnosis of high voltage circuit breaker based on multi-feature assessment and XGBoost[J]. High Voltage Apparatus, 2023, 59 (4): 1-9.
- [22] 谭雨欣, 田义超, 黄卓梅, 等. 北部湾茅尾海无瓣海桑红树林地上生物量反演——基于 XGBoost 机器学习算法[J]. 生态学报, 2023, 43(11): 4674-4688.
- TAN Yuxin, TIAN Yichao, HUANG Zhuomei, et al. Aboveground biomass of Sonneratia apetala mangroves in Maowei Sea of Beibu Gulf based on XGBoost machine learning algorithm[J]. Acta Ecologica Sinica, 2023, 43 (11): 4674-4688.
- [23] 商俊燕, 丁辉, 胡学龙. 基于 XGBoost 的无线传感器网络冗余数据检测算法[J]. 传感技术学报, 2022, 35(11): 1568-1572.
- SHANG Junyan, DING Hui, HU Xuelong. Redundant data detection algorithm for wireless sensor network based on XGBoost[J]. Chinese Journal of Sensors and Actuators, 2022, 35 (11): 1568-1572.
- [24] TAN X M, QI J, GAN J Q, et al. Multi-filter semi-

- supervised transformer model for fault diagnosis [J]. Engineering Applications of Artificial Intelligence, 2023, 124:106498.
- [25] CHAWLA N V, BOWYER K W, HALL L O, et al. Smote: Synthetic minority over-sampling technique [J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.
- [26] DEBABRATA D, TAMÁS S. Multidimensional stationary time series: Dimension reduction and prediction [M]. CRC Press: 2021.
- [27] 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.
- [28] 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.
- [29] MAATEN L, HINTON G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9(2): 2579-2605.
- [30] 樊清川, 于飞, 宣敏. 基于优化 Bi-LSTM 模型的电力变压器故障诊断[J]. 计算机仿真, 2022, 39(11): 136-140.
FAN Qingchuan, YU Fei, XUAN Min. Power transformer fault diagnosis based on optimized Bi-LSTM model [J]. Computer Simulation, 2022, 39(11): 136-140.
- [31] 康佳宇, 张沈习, 张庆平, 等. 基于 ANOVA 和 BO-SVM 的变压器故障诊断方法[J]. 高电压技术, 2023, 49(5): 1882-1891.
KANG Jiayu, ZHANG Shenxi, ZHANG Qingping, et al. Fault diagnosis method of transformer based on ANOVA and BO-SVM [J]. High Voltage Engineering, 2023, 49(5): 1882-1891.
- [32] 安国庆, 史哲文, 马世峰, 等. 基于 RF 特征优选的 WOA-SVM 变压器故障诊断[J]. 高压电器, 2022, 58(2): 171-178.
AN Guoqing, SHI Zhewen, MA Shifeng, et al. Fault diagnosis of WOA-SVM transformer based on RF feature optimization [J]. High Voltage Apparatus, 2022, 58(2): 171-178.

(本文责编: 刘芳)

收稿日期: 2023-08-22; 修回日期: 2023-10-10
上网日期: 2023-12-05; 附录网址: www.ienergy.cn

作者简介:

汪李忠(1977), 男, 高级经济师, 从事电气工程及其自动化技术的研究, 252751929@qq.com;

吴同宇*(1998), 男, 硕士生, 从事输配电设备智能化、机电设备故障诊断等方面的研究, 1647851897@qq.com。

*为通信作者。