引用格式：李友才，彭威龙，周玉明，等.基于Borderline SMOTE与NGO优化的双层XGBoost算法的变压器故障模式识别研究[J].电力科学与技术学报,2025,40(6):32-42.

Citation: LI Youcai, PENG Weilong, ZHOU Yuming, et al. Research on transformer fault pattern recognition based on double-layer XGBoost algorithm optimized by Borderline SMOTE and NGO[J]. Journal of Electric Power Science and Technology,2025,40(6):32-42.

基于Borderline SMOTE与NGO优化的双层XGBoost算法的变压器故障模式识别研究

李友才[1]，彭威龙[1]，周玉明[1]，李奇艳[1]，曾　成[2]，杨　鑫[2]

（1.湖南五凌电力工程有限公司，湖南 长沙 410000；2.长沙理工大学电网防灾减灾全国重点实验室，湖南 长沙 410114）

摘　要：为提升变压器故障诊断精度，提高模型识别的效率，改善不平衡样本对变压器故障识别模型的不良影响，提出一种基于边界线合成少数类过采样技术（borderline synthetic minority over-sampling technique，BSMOTE）与北方苍鹰优化（northern goshawk optimization，NGO）算法的双层极限梯度提升（extreme gradient boosting，XGBoost）的变压器故障模式识别模型。首先，利用BSMOTE对少数类样本进行扩充，得到平衡的数据集。其次，通过无编码比值法建立多维特征量，并采用XGBoost确定最优特征子集。再次，通过NGO对XGBoost进行参数寻优，得到变压器故障诊断模型，实现对变压器故障的准确识别。最后，采用实例对所提方法进行了仿真分析。所提方法的诊断准确率比递归特征消除法（recursive feature elimination，RFE）、随机森林（random forest，RF）特性筛选、分类提升（categorical boosting，CatBoost）特征提取与19维特征分别提高了2.88%、4.03%、4.44%与7.47%。研究结果表明，所提方法故障辨识精度高、误判率低、性能稳定。

关　键　词：变压器；故障诊断；过采样；极限梯度提升算法；北方苍鹰优化算法

DOI：10.19781/j.issn.1673-9140.2025.06.004　　**中图分类号**：TM411　　**文章编号**：1673-9140(2025)06-0032-11

Research on transformer fault pattern recognition based on double-layer XGBoost algorithm optimized by Borderline SMOTE and NGO

LI Youcai[1], PENG Weilong[1], ZHOU Yuming[1], LI Qiyan[1], ZENG Cheng[2], YANG Xin[2]

(1. Hunan Wuling Electric Power Engineering Co., Ltd., Changsha 410000, China; 2. State Key Laboratory of Disaster Prevention and Reduction for Power, Changsha University of Science & Technology, Changsha 410114, China)

Abstract: To enhance the diagnostic accuracy of transformer fault, raise the efficiency of model recognition, and mitigate the adverse effects of imbalanced samples on transformer fault recognition models, a transformer fault pattern recognition model with double-layer extreme gradient boosting (XGBoost) based on borderline synthetic minority over-sampling technique (BSMOTE) and northern goshawk optimization (NGO) algorithms is proposed. Firstly, BSMOTE is used to expand minority class samples, and a balanced dataset is obtained. Secondly, the noncoding ratio method is used to establish multidimensional feature quantities, and XGBoost is used to determine the optimal feature subset. Then, an NGO algorithm is used to optimize the XGBoost parameters, and a transformer fault diagnosis model is obtained, achieving accurate recognition of transformer faults. Finally, practical cases are adopted to make a simulation analysis of the proposed method. The diagnostic accuracy of the proposed method is 2.88%, 4.03%, 4.44%, and 7.47% higher than that of the recursive feature elimination (RFE), random forest (RF) feature screening, categorical boosting (CatBoost) feature extraction, and the 19-dimensional feature, respectively. The results show that the method proposed in this article has higher fault recognition accuracy, lower misjudgment rate, and stable performance.

Key words: transformer; fault diagnosis; over-sampling; extreme gradient boosting algorithm; northern goshawk optimization algorithm

收稿日期：2025-04-20；**修回日期**：2025-02-27

基金项目：国家自然科学基金（52177015）；国家电投五凌电力有限公司2023年C类科技项目（WLKY-C-2023-FD06）

通信作者：李友才（1980—），男，高级工程师，主要从事电力技术及其自动化、电气设备试验技术等方面的研究；E-mail：87179595@qq.com

作为电气系统中起枢纽作用的主设备，电力变压器的运行状态在很大程度上影响着电力系统的安全与稳定[1-4]。如果不能及时发现并诊断变压器的故障，就会导致大规模停电等民生问题，造成大范围的经济损失。因此，建立一套高效、全面、准确的变压器故障诊断模型对保证电力系统安全与稳定至关重要[5-6]。

变压器的绝缘老化会使绝缘油发生分解，H_2、CH_4、C_2H_6、C_2H_4与C_2H_2等气体会溶解于绝缘油中。作为最常用的变压器状态评估与故障诊断方法，油中溶解气体分析（dissolved gas analysis，DGA）通过分析绝缘油中溶解的各气体的成分与含量比值来判别变压器内部故障类型与严重程度。常见的DGA方法有：三比值法[7-8]、四比值法[9]、三角形法[10]等。该类方法原理简单，易于实现，但是其存在边界比值僵硬、误判率高等缺点[11]。随着人工智能的发展，机器学习的方法逐步应用在变压器故障模式识别上，如支持向量机[12-13]（support vector machine，SVM）、神经网络[14-15]、云模型[16-17]等。但这些方法也有一些缺点。SVM采用二次规划方法来确定支持向量，当数据量较大时，SVM的计算效率会降低。神经网络虽然具有较强的学习能力，但其可能出现过拟合现象，增加模型复杂性。云模型要求有大量的数据供其训练，实际情况不一定能满足该条件。

与传统的机器学习模型不同，集成学习通过结合多个模型的力量，提高模型的预测精确度与稳定性，且其具备更好的可解释性与适应性[18-20]。文献[21]将处理完备的样本数据输入到梯度提升树（gradient boosting decision tree，GBDT）中，有效提升了变压器故障诊断效果。但是GBDT模型的训练速度与效率还有待提高，且其在诊断的时效性上也存在不足。近年来，不少学者对GBDT进行了优化改进。其中，极端梯度提升（extreme gradient boosting，XGBoost）由于极致的训练效率而活跃在机器学习领域，展现了较好的训练精准度与出色的普适性。文献[22]探索了XGBoost算法在变压器故障检测中的应用，并借助遗传算法对相关参数进行了优化调整，但该研究并未借助XGBoost模型对多维特征的重要性进行排序与特征关联性分析。文献[23]通过应用粒子群优化技术来提升极限学习机（extreme learning machine，ELM）的诊断准确度，并取得了一定的成效。该研究也指出，粒子群优化算法在初始种群设置和避免陷入局部最优解的方面仍有待改进。

在运用机器学习与集成学习工具对变压器故障进行模式识别时，充分平衡的数据集是不可或缺的。而实际运行状态中的变压器故障率较低，且故障类别并不平衡，这就必须对原始数据集进行提前处理。处理不平衡数据集的方法主要有欠采样和过采样2种。欠采样方法[24-25]一般从多数类样本中直接删去部分数据，这可能导致多数类中具有重要信息的样本被删去。在过采样方法中，合成少数类过采样技术（synthetic minority over-sampling technique，SMOTE）、边界线合成少数类过采样技术（borderline synthetic minority over-sampling technique，BSMOTE）[26]、自适应合成过采样方法（adaptive synthetic sampling，ADASYN）是过程简单且常见的算法。文献[27]采用SMOTE处理原始数据集，并结合随机森林对变压器故障进行诊断，得到了不错的效果。然而，SMOTE主要基于相邻少数样本的假设，并未深入考虑邻近样本的分布特征，这可能导致类间出现较高的重复性。文献[28]提出自适应合成过采样技术（adaptive synthenic sampling，ADASYN），该算法考虑了在少数类样本周围多数类的分布密度，特别是在密度较高的区域，该算法会采集更多的样本。但其对异常值较为敏感。

鉴于这些问题，本文将过采样技术与集成学习相结合，提出一种基于BSMOTE与NGO优化的双层XGBoost的变压器故障模式识别模型。首先，利用BSMOTE对原始不平衡数据进行处理。其次，通过无编码比值法建立多维特征量，并采用优选特征策略确定较优特征子集。再次，搭建NGO-XGBoost变压器故障诊断模型，实现对变压器故障的准确识别。最后，在实例中进行仿真分析，验证所提算法的可行性与有效性。

1　算法原理

1.1　边界线合成少数类过采样技术

基于SMOTE的BSMOTE，通过生成合成的少数类样本来过采样少数类。对每一个少数类样本，该算法先寻找其同类别的k个最近邻；再根据过采样率从中随机选择一些样本；最后，在少数类样本和其选定的最近邻样本间的连线上生成新的合成样本。SMOTE算法主要基于相邻少数样本的假设，未深入考虑邻近样本的分布特征，这可能导致类间出现较高的重复性。BSMOTE算法通过识别危险类别的边界样本，有效减少了此类重复现象的发

生。BSMOTE只对边界少数类样本(危险少数类样本)进行过采样。该算法先找出危险少数类样本,再将这些样本创造出的合成样本整合进基础的训练资料库中。令整个训练集为T,少数类集为P,多数类集为N,则有

$$P=\{p_1,p_2,\cdots,p_r\} \tag{1}$$

$$N=\{n_1,n_2,\cdots,n_q\} \tag{2}$$

式中,p_i为第i个少数类;r为少数类样本个数;n_i为第i个多数类;q为多数类样本个数。

BSMOTE算法可分为4步。

1) 对少数类集P中的每一个p_i, $i=1,2,\cdots,r$,在整个训练集T中选取其m个最近邻,这m个最近邻中有$m'(0\leqslant m'\leqslant m)$个多数类样本。

2) 如果$m=m'$,也就是p_i周围的全部m个最近相邻点均为多数类别样本,则将p_i判定为噪声点,并在后续处理过程中予以忽略。如果$\frac{m}{2}\leqslant m'<m$,也就是$p_i$的最近邻中的多数类点的数量多于其最近邻少数类点的数量,则p_i被认为是容易误分类的,并被放入一个名为D的集合中。如果$0\leqslant m'<\frac{m}{2}$,也就是$p_i$的最近邻中的多数类点的数量少于其最近邻少数类点的数量,则p_i被认为是安全的,在后续处理过程中予以忽略。

3) 集合D所包含的样本为边界数据,故有$D\subseteq P$,令

$$D=\{p_i'|\,p_i'\in P, i=1,2,\cdots,d\}, 0\leqslant d\leqslant r \tag{3}$$

对D中的所有样本,从P中寻找其k个最近邻。

4) 从D中生成$s\times d$个合成正例样本,s为$[1,k]$的整数。对$\forall p_i'\in D, i=1,2,\cdots,d$,先随机抽取其$k$个近邻中的$s$个$p_j, j=1,2,\cdots,s$;再计算$p_i'$与这些每一个最近邻$p_j, j=1,2,\cdots,s$之间的差异$\alpha_j$;然后,将$\alpha_j$乘以随机数$\gamma_j, j=1,2,\cdots,s, \gamma_j\in[0,1]$;最后,生成新的合成少数类样本$\tilde{p}_{ij}$,即

$$\tilde{p}_{ij}=p_i+\alpha_j\gamma_j, j=1,2,\cdots,s \tag{4}$$

1.2 XGBoost原理

梯度提升法依次结合弱分类器的决策,是一种决策树(decision tree, DT)的有效机器学习算法。XGBoost是梯度提升机器学习集合方法的扩展,是一种有效的分类器方法,其降低了梯度提升方法的计算复杂度,节约了计算时间。

设包含n个实例和m个特征的给定数据集为$Q=\left\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\middle|\,\boldsymbol{x}_i\in\boldsymbol{R}^m,\boldsymbol{y}_i\in\boldsymbol{R}^n, i=1,2,\cdots,n\right\}$,则具有$k$个加法函数的树提升模型的输出为

$$\begin{cases}\hat{y}_i=\sum_{k=1}^{K}f_k(x_i), \quad f_k\in F\\ F=\{f(x)=\omega_q(x)\,|\,q:R^m\rightarrow T,\omega\in R^t\}\end{cases} \tag{5}$$

式中,$\hat{y}_i$为样本x_i的预测值;F为分类和回归树的空间;T为特定树的叶子;q为树的结构部分;ω为叶子重量;f_k为包括q与ω的函数。

为更好地分类,使用树中的决策规则。该规则通过将各个叶子的分数相加来估计最终预测,函数f_k通过最小化目标函数O来学习,即

$$\min O=\sum_{i=1}^{n}L(y_i,\hat{y}_i)+\sum_{k=1}^{K}W(f_k) \tag{6}$$

$$W(f)=\gamma T^2+\frac{1}{2}\lambda\|\omega\|^2 \tag{7}$$

式中,L为训练损失函数或误差估计函数,即测量预测项$\hat{y}_i$和实际对象y_i之间的距离;W为树模型复杂度惩罚项,即额外正则化项,其有助于平滑最终的学习权重,以防止模型出现过度拟合等情况;n为样本总数;K为函数总数;T为树叶的数量;γ、λ为控制参数,分别用于控制叶子节点的数量及权重分数。

树提升模型以加法的方式训练。若直接将式(6)的树提升模型作为目标函数,则其不能在欧几里得空间中借助传统优化方法来优化。梯度树提升是树提升方法的增强版本。如果$\hat{y}_i^{(t)}$为第i个实例的第t次迭代的预测项,则在第t次迭代时,添加f_t,即$\hat{y}_i^{(t-1)}$。故式(6)可改写为

$$O^{(t)}=\sum_{i=1}^{n}L\left(y_i,\hat{y}_i^{(t-1)}+f_t(x_i)\right)+W(f_t) \tag{8}$$

XGBoost技术通过二阶泰勒展开来近似,式(8)的第t次迭代的泰勒展开式$\tilde{O}^{(t)}$为

$$O^{(t)}\simeq\tilde{O}^{(t)}=\sum_{i=1}^{n}\left[L\left(y_i,\hat{y}_i^{(t-1)}+g_if_t(x_i)+\frac{1}{2}b_if_t^2(x_i)\right)\right]+W(f_t) \tag{9}$$

式中,g_i和b_i分别为损失函数的一阶和二阶梯度统计量。

继续简化,则有

$$\tilde{O}^{(t)}=\sum_{j=1}^{T}\left[\left(\sum_{i\in I_j}g_i\right)\omega_j+\frac{1}{2}\left(\sum_{i\in I_j}b_i+\lambda\right)\omega_j^2\right]+\gamma T \tag{10}$$

式中,ω_j为第j片树叶权重;I_j为第j片树叶的实例,即$I_j=\{i|q(x_i)=j\}$。

对于树的固定结构$q(x)$,叶j的解权重w_j^*为

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j}(b_i+\lambda)} \tag{11}$$

式(11)又被认为是用于$q(x)$树结构估计的评分函数，其可获得用于分类的最佳树结构。通常，不可能评估所有树结构$q(x)$。本文采用一种有效的算法，该算法从单个叶子开始，迭代地将分支附加到树中。假设I_L和I_R分别是分裂后左节点和右节点集，并且有$I=I_L\Delta I_R$。则分裂后的损失最小化目标函数O_{split}为

$$O_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L}(b_i+\lambda)} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R}(b_i+\lambda)} + \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I}(b_i+\lambda)} - \lambda\right] \tag{12}$$

1.3　NGO原理

NGO优化算法[29-30]具有较高的收敛精度和稳定性。NGO模仿苍鹰的狩猎过程，其包括2个主要阶段：识别并攻击猎物，追逐与逃逸。

1.3.1　识别并攻击猎物

这一阶段的苍鹰随机选择目标并迅速攻击，其数学模型为

$$Z_i = Y_k, i=1,2,\cdots,N \tag{13}$$

$$Y_{i,j}^{new,z1} = \begin{cases} y_{i,j} + r\left(z_{i,j} - Ix_{i,j}\right), W_i^p < W_i \\ y_{i,j} + r\left(x_{i,j} - p_{i,j}\right), W_i^p \geqslant W_i \end{cases} \tag{14}$$

$$Y_i = \begin{cases} Y_i^{new}, W_i^{new,p1} < W_i \\ Y_i, W_i^{new,p1} \geqslant W_i \end{cases} \tag{15}$$

式中，Z_i为第i只苍鹰选择的猎物的位置；Y_k苍鹰种群中另一只苍鹰的位置；k为[1,n]中的随机自然数；$Y_{i,j}^{new,z1}$为第一阶段第i只苍鹰在j维上的新位置；$y_{i,j}$为更新前第i只苍鹰在j维上的位置；r为[0,1]的随机数；$z_{i,j}$为模拟捕食行为时，第i只猎物在第j维上的位置；I为随机数，$I\in\{1,2\}$；W_i^p为猎物位置的适应度；W_i为苍鹰当前位置适应度；Y_i为第i只苍鹰位置；Y_i^{new}为第i只苍鹰第一阶段产生的新位置；$W_i^{new,p1}$为第i只猎物新位置目标函数值；r为搜索和更新中产生自发NGO行为的随机数。

1.3.2　追逐与逃逸

苍鹰接近猎物后，猎物会试图逃跑，鹰会继续追逐，形成你追我赶的局面。鹰可全局搜索，捕获猎物。其数学模型为

$$Y_{i,j}^{new,z2} = y_{i,j} + R(2r-1)y_{i,j} \tag{16}$$

$$R = 0.02\left(1-\frac{t}{T}\right) \tag{17}$$

$$Y_i = \begin{cases} Y_i^{new,z1}, W_i^{new,z2} < W_i \\ X_i, W_i^{new,z2} \geqslant W_i \end{cases} \tag{18}$$

式中，$Y_{i,j}^{new,z2}$为第2阶段第i只苍鹰在j维上的新位置；R为苍鹰攻击系数，其值由式(21)确定；r为[0,1]的随机数；t，T分别为当前迭代次数和最大迭代次数；$Y_i^{new,z2}$为第i只苍鹰第二狩猎阶段产生的新位置；$W_i^{new,z2}$为第i只苍鹰在新状态下的适配值；W_i为第i只苍鹰在当前位置的适应度。

2　双层XGBoost变压器故障诊断模型

为对变压器的特性优选与故障诊断这两个方面进行优化，本研究建立双层XGBoost模型。首先，利用BSMOTE对样本进行扩充；其次，采用XGBoost对多维变压器特征量进行优选；最后，使用NGO-XGBoost模型对故障模式进行识别，并利用测试集对模型进行了综合评价。本模型旨在提升变压器故障样本不均匀情况下故障模式的识别精度。双层XGBoost变压器故障诊断模型结构如图1所示。

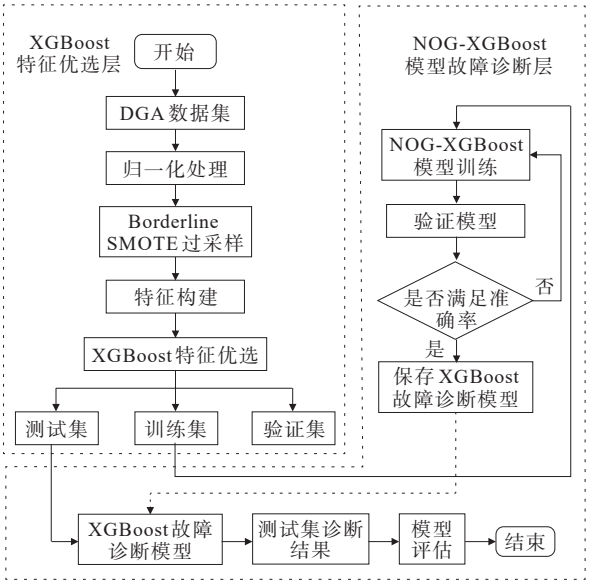

图1　双层XGBoost变压器故障诊断模型结构

Figure 1　Structure of double-layer XGBoost transformer fault diagnosis model

2.1　特征优选层

气体比例的多样性扩展了特征参数的选取范围，但这也导致了特征之间的重复性，降低了模型处理效率。基于特征在作为分裂点时所贡献的平

均增益值，XGBoost为每个特征赋予一个重要性评分，并按该评分对特征进行排序，识别与预测目标密切相关的特征。利用这一特征排序，可剔除不必要的特征，降低数据冗余，简化模型结构，提升模型的运算速度与泛化能力，有效防止过拟合现象的发生。

在模型训练完成后，先按重要性从高到低对特征分数进行排序，并设定不同阈值，进行特征筛选，形成多个特征组合；再依次评估使用这些特征组合的模型在测试集上的诊断表现；最后，将在诊断准确性方面表现最佳的特征组合作为最终采纳的特征集合。

2.2 故障诊断层

2.2.1 基于NGO优化XGBoost的变压器故障诊断

在对变压器故障数据特征进行优化选择后，将经过预处理的样本数据导入至下层诊断模型。XGBoost的参数众多且设置烦琐，若选择不当，极易造成过拟合或欠拟合问题，损害模型的精度及其泛化能力。故本研究运用NGO算法对XGBoost模型的超参数进行调整优化。以XGBoost的5折交叉验证中的系数平均I_{kappa}作为NGO算法的优化目标，同时将$(1\text{-}I_{\text{kappa}})$作为最小化的适应度值。经过连续的迭代训练，构建一个在变压器故障诊断中具有较高诊断精度的超参数配置，并使用这组最优的超参数配置对测试数据集进行诊断。通过基于混淆矩阵的评价标准对测试诊断结果进行分析，检验模型的泛化性能。

NGO与XGBoost模型参数：NGO参数最大迭代次数G为100，搜索维度D为8，种群规模N为20；XGBoost参数惩罚项γ为(0,10]，样本采样率r_{sub}为[0.5,1]，学习率η为[0.01,0.50]，最小样本权重$w_{\min}$为[1,10]，特征采样率r_{col}为[0.5,1.0]，L2正则项λ为(0,10]，树的最大深度$d_{\max}$为[1,10]，迭代轮数R_{num}为[50,100]

2.2.2 评估函数

1）I_{kappa}系数。

I_{kappa}系数是衡量分类算法预测结果与实际结果匹配程度的量化指标，其被广泛应用于评价分类算法的表现。$I_{\text{kappa}}\in[-1,1]$。若$I_{\text{kappa}}=-1$，则此时的预测与实际完全相反；若$I_{\text{kappa}}=0$，则此时的预测与实际结果的匹配纯属偶然；若$I_{\text{kappa}}=1$，则此时的预测与实际完全吻合。一个更高的$I_{\text{kappa}}$值通常意味着更好的分类器准确度。

2）基于混淆矩阵的评估指标。

混淆矩阵是机器学习领域评价模型性能的重要手段之一，其不仅记录了分类器在多个类别上的判别情况，还能通过图形化的方式展示分类效果，提供了一个全面且易于理解的模型效能分析途径。以正负二分类问题为例，其混淆矩阵为$\boldsymbol{A}_{2\times2}=\begin{pmatrix} x_{\text{TP}} & x_{\text{FN}} \\ x_{\text{FP}} & x_{\text{TN}} \end{pmatrix}$。其中，$x_{\text{TP}}$为本来是正类，也被判别为正类的真正例的个数；$x_{\text{FN}}$为本来是正类，但被判别为负类的假负例的个数；$x_{\text{FP}}$为本来是负类，但被判别为正类的假正例的个数；$x_{\text{TN}}$为本来是负类，也被判别为负类的真负例的个数。

具体的评价指标分别为

$$\begin{cases} I_{\text{recall}}=\dfrac{x_{\text{TP}}}{x_{\text{TP}}+x_{\text{FN}}} \\ I_{\text{precision}}=\dfrac{x_{\text{TP}}}{x_{\text{TP}}+x_{\text{FP}}} \\ F_{1-\text{score}}=\dfrac{2I_{\text{recall}}I_{\text{precision}}}{I_{\text{recall}}+I_{\text{precision}}} \end{cases} \tag{19}$$

式中，I_{recall}为分类模型在识别正类样本时的敏感程度的召回率；$I_{\text{precision}}$为模型对于正类样本判定的信赖度的精确率；$F_{1-\text{score}}$为F1分数，其作为精确率和召回率的加权平均，综合体现了模型的性能。

3 试验结果与分析

将变压器的状态分为6类：中低温放热（T1&T2）、高温放热（T3）、正常（N）、低能放电（D1）、高能放电（D2）和局部放电（PD）。本研究采用中国某家供电企业提供的480组监测数据，构建故障样本库。在该故障样本库中，不同运行状态由H_2、CH_4、C_2H_6、C_2H_4及C_2H_2这5种标志性气体组成。数据分布与样本标签见表1。

表1 数据分布与样本标签

Table 1 Data distribution and sample labeling

状态类型	标签	占比/%	数量
N	1	16.3	78
D2	2	21.9	105
T1&T2	3	30.6	147
T3	4	15.8	76
PD	5	9.4	45
D1	6	6.0	29

3.1 数据预处理

当变压器发生故障时，其绝缘油内溶解的气体

种类及其浓度会相应地改变。对这些标志性气体进行归一化处理，即

$$x'=\frac{x_i-x_i^{\min}}{x_i^{\max}-x_i^{\min}} \tag{20}$$

式中，x_i和x'分别为某特征在归一化前、后的状态值；$x_i^{\max}$与$x_i^{\min}$分别为在进行归一化之前的特征数据中的最小值和最大值。

3.2　数据平衡化处理

由表1可知，中、低温过热型故障的比例为30.6%，而局部放电型故障仅占6.0%。这说明该数据集是不平衡的。如果使用不平衡的数据集来训练模型，可能导致模型对这些少数类别的样本学习不足；在后续的分类识别过程中，模型很可能误判少数类别的样本，影响其识别准确度。本文运用BSMOTE对数据集中的少数类样本进行过采样处理。数据在BSMOTE平衡化前、后的分布情况见表2。由表2可知，各类别的样本数量在处理后实现了均衡，为后续进行特征筛选、模型训练及诊断打下了良好基础。

表2　数据在Borderline SMOTE平衡化前、后的分布情况

Table 2　Data distribution before and after Borderline SMOTE balance

数据类型	不同样本标签的数量					
	1	2	3	4	5	6
原始数据	78	105	147	76	45	29
Borderline SMOTE处理后	147	147	147	147	147	147

3.3　变压器故障特征优选

变压器发生故障时，其内部的绝缘介质受到电场、热和机械应力等影响，发生分解或击穿，从而释放出一定量的H_2、CH_4、C_2H_6、C_2H_4与C_2H_2等气体，这些气体溶解于绝缘油介质中。通过研究溶解于油中的气体种类及其浓度，能推断变压器的内部工作状况与可能出现的故障种类。

在DGA故障检测研究中，常规的IEC三比值法和Rogers四比值法被作为基准，广泛采用。但这些方法在特征选择上存在局限性，不能充分利用数据，无法充分揭示故障与特征之间的内在联系。鉴于此，本文选取H_2、CH_4、C_2H_6、C_2H_4与C_2H_2这5种主要特征气体，构建一个19个维度的比值特征体系。变压器候选比值特征见表3。在表3中，THC是CH_4、C_2H_6、C_2H_4与C_2H_2的混合气体。

表3　变压器候选比值特征

Table 3　Alternative ratio features of transformer

特征编码	特征量	特征编码	特征量
T1	H_2	T11	CH_4/H_2
T2	CH_4	T12	C_2H_2/H_2
T3	C_2H_6	T13	C_2H_4/THC
T4	C_2H_4	T14	C_2H_6/THC
T5	C_2H_2	T15	CH_4/THC
T6	CH_4/C_2H_6	T16	C_2H_2/THC
T7	CH_4/C_2H_4	T17	H_2/THC
T8	C_2H_6/C_2H_4	T18	$(CH_4+C_2H_4)$/THC
T9	C_2H_2/C_2H_4	T19	H_2/全部5种气体
T10	C_2H_4/C_2H_6		

本文构建的19维特征向量拓展了特征维度，能更全面、充分地利用了数据信息。但这一过程容易造成信息重叠。重叠的信息特性会增加模型的计算复杂度。因此，采用了XGBoost对这19个特征量进行重要性评估与排序。特征重要度排序结果见表4。

表4　特征重要度排序结果

Table 4　Feature importance ranking result

特征编码	特征量	特征重要度值
T3	C_2H_6	0.168
T11	CH_4/H_2	0.110
T9	C_2H_2/C_2H_4	0.095
T5	C_2H_2	0.055
T1	H_2	0.054
T7	CH_4/C_2H_4	0.053
T17	H_2/THC	0.051
T4	C_2H_4	0.047
T19	H_2/ALL	0.039
T12	C_2H_2/H_2	0.037
T15	CH_4/THC	0.034
T18	$(CH_4+C_2H_4)$/THC	0.031
T16	C_2H_2/THC	0.029
T13	C_2H_4/THC	0.027
T6	CH_4/C_2H_6	0.025
T10	C_2H_4/C_2H_6	0.022
T14	C_2H_6/THC	0.015
T8	C_2H_6/C_2H_4	0.014
T2	CH_4	0.012

依次将经过XGBoost算法重要性排序的特征输入至第2层，进行模式识别。为降低结果偶然性，本文采用10折交叉验证方法对数据样本进行验证，并以其平均准确率作为评判标准，模型诊断准确率平均值与特征子集的数量关系如图2所示。从图2中可以看出，随着诊断模型的输入特征个数从1逐步增加到8，XGBoost的识别精度也随之提高，当输入诊断模型的特征个数为8时，该模型的平均诊断准确率达到最高的93.68%，随后，特征数量的增加并不会进一步提升模型诊断准确率。这是因为过多的特征增加了模型的复杂度。综上所述，本文采用XGBoost排序后的前8个关键特征来进行模型的训练与诊断。

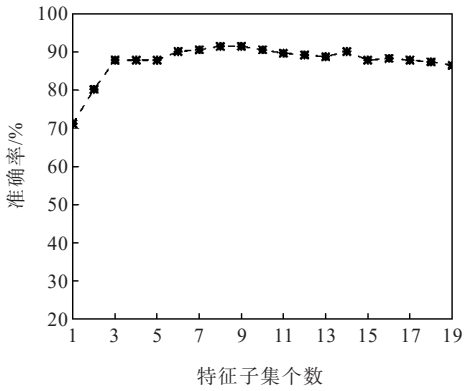

图2 模型诊断准确率平均值与特征子集的数量关系

Figure 2 Relationship between mean value of model's diagnostic accuracy and number of feature subsets

3.4 NGO-XGBoost模型故障诊断结果

按照6∶2∶2的比例，将优选的特征子集分为训练集、验证集与测试集。样本数据的分布见表5。

表5 样本数据的分布

Table 5 Distribution of sample data

状态类型	训练集样本个数	测试集样本个数	验证集样本个数
1	89	29	29
2	89	29	29
3	89	29	29
4	89	29	29
5	89	29	29
6	89	29	29

测试集样本混淆矩阵如图3所示。图3展示了结合BSMOTE与NGO-XGBoost的变压器故障识别的混淆矩阵。在该矩阵中，对角线部分的区块数字为实际样本中被准确识别的数量，各行的数据总和为相应类别的样本总量。从图3中可以看出，在174个测试样本中，7个故障样本被误判，该变压器故障检测的准确率为95.98%。这进一步验证了该模型在识别准确率方面的优势。在正常类别与中低温过热类别中，所有样本均判别正确；在高能放电类别中，3个故障样本被错误归类。这可能是在平衡化处理过程中，故障特征的权重有所提升，使得部分高效能放电类边缘样本被误判。该模型的$I_{precision}$=0.959 8，I_{recall}=0.960 1，$F_{1-score}$=0.959 9，I_{kappa}=0.952 1。这些均反映该模型的实际分类结果与预测结果高度一致，充分说明了本文所提模型在故障辨识与分类方面的优势。

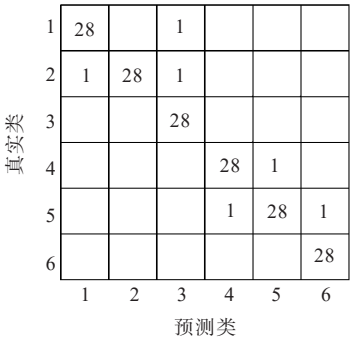

图3 测试集样本混淆矩阵

Figure 3 Test set sample confusion matrix

4 结果与讨论

4.1 多种特征优选方式对比研究

为检验所采纳的特征选择策略的优势，本文采用递归特征剔除（recursive feature elimination，RFE）、CatBoost、随机森林（random forest，RF）与19个维度的特性组合方法，并将其分别输入至NGO-XGBoost模型中。不同特征优选方法诊断结果对比如图4所示。RFE、RF、CatBoost与19维特征方法的诊断准确率分别为93.10%、91.95%、91.54%和88.51%，RFE、RF、CatBoost与19维特征方法的I_{kappa}值分别为0.918 3、0.911 7、0.900 0和0.865 2。XGBoost的诊断准确率比RFE、RF、CatBoost与19维特征方法的诊断准确度分别提高了2.88、4.03、4.44与7.47个百分点，XGBoost的I_{kappa}也比RFE、RF、CatBoost与19维特征方法的I_{kappa}分别增加0.033 9、0.047 2、0.052 1及0.086 9。这些都进一步说明了XGBoost特征选择策略的卓越性能。

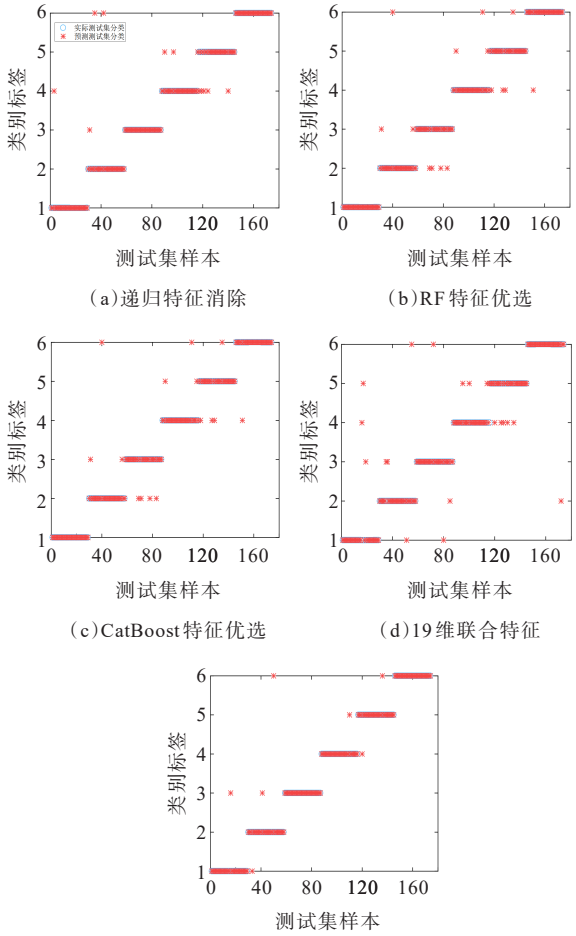

图 4　不同特征优选方法诊断结果对比

Figure 4　Comparison of diagnostic results of different feature optimization methods

4.2　样本均衡化效果对比分析

为检验诊断模型在应对数据失衡问题上的优势，本文在数据预处理阶段分别运用了SMOTE和ADASYN这2种过采样技术，并将其结果与直接使用未处理的原始数据集得到的诊断效果进行对照。样本均衡化混淆矩阵对比如图5所示，不同样本均衡方法下的诊断结果见表6。诊断结果表明，未经平衡处理的初始数据集的少数类样本训练不足，导致低能放电、局部放电和高温过热这3种类型的样本在分类时存在较高的误判率。通过采用过采样技术，模型的诊断准确率及I_{kappa}值均有所提高。本文运用BSMOTE进行少数类样本数据扩充，每种故障类型的诊断精确率与整体性能均超越了SMOTE和ADASYN的，这进一步证明了本文所提方法在应对数据不平衡问题上的高效性。

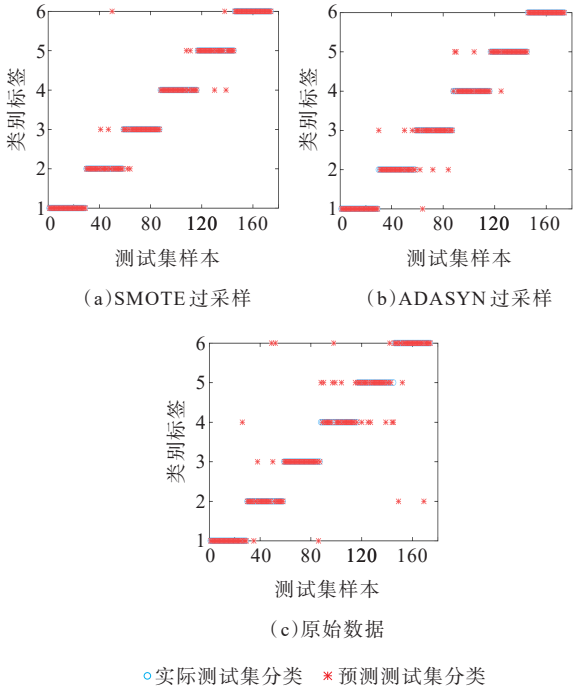

图 5　样本均衡化混淆矩阵对比

Figure 5　Comparison of confusion matrices with sample balancing

表 6　不同样本均衡方法下的诊断结果

Table 6　Diagnostic results under different sample balancing methods

数据采样方法	准确率/%	I_{kappa}
SMOTE	94.83	0.938 6
ADASYN	93.68	0.925 1
非均衡数据集	86.21	0.836 0

4.3　多模型诊断效果对比分析

为检验本文所提出的NGO-XGBoost在提高变压器故障识别准确率方面的有效性，本文对4种集成学习模型：DT、RF、XGBoost以及WOA-XGBoost进行分析与比较。

不同集成学习诊断模型下诊断准确率与I_{kappa}值如图6所示。不同集成学习诊断模型综合评价指标结果见表7。分析图6与表7可知，在单个诊断模型中，XGBoost模型的精确率上领先于RF与DT的。不论是在精确率、召回率还是F1分数等关键指标上，XGBoost均显著胜出，这证明其在变压器故障识别方面具有更高的适用性和效能。另外，本文通过采用NGO算法对XGBoost的关键参数进行优化搜索，所建立的最佳诊断模型在故障识别精确率上优于WOA-XGBoost模型的。该优化算法显示出较强的寻优能力，显著提升了变压器故障诊断的准确率。

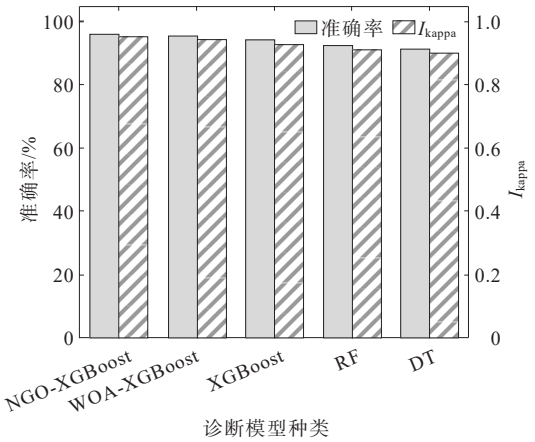

图6 不同集成学习诊断模型下诊断准确率与I_{kappa}值

Figure 6 Diagnostic accuracy and I_{kappa} value under different ensemble learning diagnostic models

表7 不同集成学习诊断模型综合评价指标结果

Table 7 Comprehensive evaluation index results of different ensemble learning diagnostic models

数据采样方法	精确率	召回率	F1值
NGO-XGBoost	0.959 8	0.960 1	0.959 9
WOA-XGBoost	0.954 0	0.957 3	0.955 6
XGBoost	0.942 5	0.942 5	0.942 5
RF	0.925 3	0.926 9	0.926 1
DT	0.913 8	0.919 3	0.916 5

5 结语

本文聚焦于电力变压器的故障检测问题，提出了一种结合BSMOTE与NGO优化的双层XGBoost算法的故障诊断方法。实证分析表明，该策略可行、高效。本文得出了以下结论：

1）XGBoost特征选择模型能够筛选出最优特征集，并在此基础上提高变压器故障诊断精度。

2）本文在数据处理环节针对故障样本的不均衡问题进行了优化，采用BSMOTE算法进行样本的过采样处理，有效克服了样本量不足及分布不平衡所带来的诊断准确率下降的难题，显著降低了诊断模型的误判可能性。

3）相较于传统的机器学习架构，NGO优化的双层XGBoost算法展现出了更优异的泛化能力和预见性。在应对多样本类别以及较小规模数据集的情境中，其依旧能够保持稳健的分类准确率。

参考文献：

[1] 何宁辉，朱洪波，李秀广，等. 基于贝叶斯网络和假设检验的变压器故障诊断[J]. 电力科学与技术学报，2021, 36(6): 20-27.
HE Ninghui, ZHU Hongbo, LI Xiuguang, et al. Transformer fault diagnosis based on Bayesian network and hypothesis testing[J]. Journal of Electric Power Science and Technology, 2021, 36(6): 20-27.

[2] 陆云才，廖才波，李群，等. 基于声纹特征和集成学习的变压器缺陷诊断方法[J]. 电力工程技术，2023, 42(5): 46-55.
LU Yuncai, LIAO Caibo, LI Qun, et al. Transformer fault diagnosis method based on voiceprint feature and ensemble learning[J]. Jiangsu Electrical Engineering, 2023, 42(5): 46-55.

[3] 王辉东，姚海燕，郭强，等. 基于多尺度卷积神经网络的变压器故障诊断方法[J]. 电力科学与技术学报，2023, 38(4): 104-112.
WANG Huidong, YAO Haiyan, GUO Qiang, et al. A transformer fault diagnosis method based on multiscale 1DCNN[J]. Journal of Electric Power Science and Technology, 2023, 38(4): 104-112.

[4] 李元，李星辉，孙渭薇，等. 基于多模型级联的油浸式电力变压器故障诊断方法[J]. 智慧电力，2023, 51(6): 86-92.
LI Yuan, LI Xinghui, SUN Weiwei, et al. Fault diagnosis method of oil-immersed power transformer based on multi-model cascade fusion[J]. Smart Power, 2023, 51(6): 86-92

[5] 丁学辉，许海林，罗颖婷，等. 基于随机森林特征优选与MAEPSO-ELM算法的变压器DGA故障诊断[J]. 电力科学与技术学报，2022, 37(2): 181-187.
DING Xuehui, XU Hailin, LUO Yingting, et al. Transformer DGA fault diagnosis based on the random forest feature optimization and MAEPSO-ELM algorithm[J]. Journal of Electric Power Science and Technology, 2022, 37(2): 181-187.

[6] 周秀，怡恺，李刚，等. 基于邻域粗糙集与AMPOS-ELM的变压器DGA故障诊断[J]. 电力科学与技术学报，2022, 37(3): 157-164.
ZHOU Xiu, YI Kai, LI Gang, et al. Atransformer DGA fault diagnosis approachbased on neighborhood rough set and AMPSO-ELM[J]. Journal of Electric Power Science and Technology, 2022, 37(3): 157-164.

[7] Mineral oil-impregnated electrical equipment in service - Guide to the interpretation of dissolved and free gases analysis: DS/EN 60599/A1: 2007[S]. Danish Standards, 2007.

[8] 杨金鑫，廖才波，胡雄，等. 基于DGA与TPE-LightGBM的变压器故障诊断[J]. 电力科学与技术学报，2024, 39(4): 70-77.
YANG Jinxin, LIAO Caibo, HU Xiong, et al. Transformer fault diagnosis based on DGA and TPE-LightGBM[J]. Journal of Electric Power Science and

Technology, 2024, 39(4): 70-77.

[9] 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, EI-13(5): 349-354.

[10] DUVAL M, DEPABLA A. Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases[J]. IEEE Electrical Insulation Magazine, 2001, 17(2): 31-41.

[11] 廖才波, 杨金鑫, 邱志斌, 等. 一种基于夏普利值及油中溶解气体分析的可解释变压器故障诊断方法[J]. 电网技术, 2024, 48(4): 1752-1762.
LIAO Caibo, YANG Jinxin, QIU Zhibin, et al. Interpretable transformer fault diagnosis based on SHAP value and dissolved gas analysis of transformer oil[J]. Power System Technology, 2024, 48(4): 1752-1762.

[12] 余松, 胡东, 唐超, 等. 基于TLR-ADASYN平衡化数据集的MSSA-SVM变压器故障诊断[J]. 高电压技术, 2021, 47(11): 3845-3853.
YU Song, HU Dong, TANG Chao, et al. MSSA-SVM transformer fault diagnosis method based on TLR-ADASYN balanced data set[J]. High Voltage Engineering, 2021, 47(11): 3845-3853.

[13] 马宏忠, 肖雨松, 孙永腾, 等. 基于ICEEMDAN和时变权重集成预测模型的变压器油中溶解气体含量预测[J]. 高电压技术, 2024, 50(1): 210-220.
MA Hongzhong, XIAO Yusong, SUN Yongteng, et al. Prediction of dissolved gas concentration in transformer oil based on ICEEMDAN and time-varying weight integrated prediction model[J]. High Voltage Engineering, 2024, 50(1): 210-220.

[14] 刘展程, 王爽, 唐波. 基于布谷鸟搜索算法和DBN模型的变压器故障识别[J]. 电力科学与技术学报, 2022, 37(2): 3-11.
LIU Zhancheng, WANG Shuang, TANG Bo. Transformer fault identification based on the cuckoo search algorithm and DBN model[J]. Journal of Electric Power Science and Technology, 2022, 37(2): 3-11.

[15] 杜江, 范志远, 范仲华, 等. 电力变压器油中溶解气体异常数据识别与含量预测研究[J]. 电网技术, 2025, 49(2): 844-853.
DU Jiang, FAN Zhiyuan, FAN Zhonghua, et al. Research on abnormal data identification and content prediction of dissolved gas in power transformer oil[J]. Power System Technology, 2025, 49(2): 844-853.

[16] 张重远, 林志锋, 刘栋, 等. 基于正态云模型与改进贝叶斯分类器的变压器故障诊断[J]. 电测与仪表, 2017, 54(4): 50-56.
ZHANG Zhongyuan, LIN Zhifeng, LIU Dong, et al. Transformer fault diagnosis based on normal cloud model and improved Bayesian classifier[J]. Electrical Measurement & Instrumentation, 2017, 54(4): 50-56.

[17] 邹阳, 林锦茄, 李安娜, 等. 基于灰色关联分析和聚类云模型的变压器油纸绝缘状态评估[J]. 电力系统保护与控制, 2023, 51(21): 35-43.
ZOU Yang, LIN Jinjia, LI Anna, et al. Evaluation of transformer oil-paper insulation status based on grey relational analysis and a cluster cloud model[J]. Power System Protection and Control, 2023, 51(21): 35-43.

[18] 陆云才, 廖才波, 李群, 等. 基于声纹特征和集成学习的变压器缺陷诊断方法[J]. 电力工程技术, 2023, 42(5): 46-55.
LU Yuncai, LIAO Caibo, LI Qun, et al. Transformer fault diagnosis method based on voiceprint feature and ensemble learning[J]. Electric Power Engineering Technology, 2023, 42(5): 46-55.

[19] 徐耀松, 李佳旺, 段彦强. 基于相似度机制AdaBoost-DBN的变压器故障层级诊断[J]. 高压电器, 2023, 59(6): 154-164.
XU Yaosong, LI Jiawang, DUAN Yanqiang. Fault hierarchical diagnosis of transformer based on AdaBoost-DBN similarity mechanism[J]. High Voltage Apparatus, 2023, 59(6): 154-164.

[20] 张国治, 陈康, 方荣行, 等. 基于DGA与鲸鱼算法优化Logit Boost-决策树的变压器故障诊断方法[J]. 电力系统保护与控制, 2023, 51(7): 63-72.
ZHANG Guozhi, CHEN Kang, FANG Rongxing, et al. Transformer fault diagnosis based on DGA and a whale algorithm optimizing a Logit Boost-decision tree[J]. Power System Protection and Control, 2023, 51(7): 63-72.

[21] 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 Weihan, 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.

[22] 张又文, 冯斌, 陈页, 等. 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.
ZHANG Youwen, FENG Bin, CHEN Ye, et al. Fault diagnosis method for oil-immersed transformer based on XGBoost optimized by genetic algorithm[J]. Electric Power Automation Equipment, 2021, 41(2): 200-206.

[23] 黄新波, 王享, 田毅, 等. 基于PSO-ELM融合动态加权AdaBoost的变压器故障诊断方法[J]. 高压电器, 2020, 56(5): 39-46.
HUANG Xinbo, WANG Xiang, TIAN Yi, et al. Transformer fault diagnosis algorithm based on PSO-ELM fusion dynamically weighted Ada boost[J]. High Voltage Apparatus, 2020, 56(5): 39-46.

[24] 任艳平, 郑重, 江一飞, 等. 融合后验概率和密度的不平衡数据欠采样方法[J]. 计算机工程与应用, 2022, 58(23): 268-277.
REN Yanping, ZHENG Zhong, JIANG Yifei, et al. Posterior probability and density-based imbalanced data undersampling[J]. Computer Engineering and Applications, 2022, 58(23): 268-277.

[25] 朱雅庆, 纪荣祎, 董登峰, 等. 欠采样全相位FFT鉴相方法仿真与实现[J]. 红外与激光工程, 2023, 52(11): 223-233.
ZHU Yaqing, JI Rongyi, DONG Dengfeng, et al. Simulation and implementation of undersampling all-phase FFT phase discrimination method[J]. Infrared and Laser Engineering, 2023, 52(11): 223-233.

[26] 马贺, 宋媚, 祝义. 改进边界分类的Borderline-SMOTE过采样方法[J]. 南京大学学报(自然科学), 2023, 59(6): 1003-1012.
MA He, SONG Mei, ZHU Yi. Improved Borderline-SMOTE oversampling method for boundary classification[J]. Journal of Nanjing University (Natural Science), 2023, 59(6): 1003-1012.

[27] 刘磊, 李龙飞, 韩雪峰, 等. 基于SMOTE和随机森林的变压器故障诊断研究[J]. 山东电力技术, 2023, 50(11): 11-19.
LIU Lei, LI Longfei, HAN Xuefeng, et al. Research on transformer fault diagnosis based on SMOTE and random forest[J]. Shandong Electric Power, 2023, 50(11): 11-19.

[28] HE H B, BAI Y, GARCIA E A, et al. ADASYN: Adaptive synthetic sampling approach for imbalanced learning[C]//2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Hong Kong, China. IEEE, 2008: 1322-1328.

[29] 李思尧, 董章, 陈雅旎, 等. 基于NGO-VMD-HHT的电缆局部放电信号特征量提取方法[J]. 电气技术, 2023, 24(11): 35-41+47.
LI Siyao, DONG Zhang, CHEN Yani, et al. Feature extraction method of cable partial discharge signal based on NGO-VMD-HHT[J]. Electrical Engineering, 2023, 24(11): 35-41+47.

[30] 任志玲, 张景智. 多策略融合LSSVM-NGO的滑动电接触失效诊断[J]. 电子测量与仪器学报, 2023, 37(12): 37-47.
REN Zhiling, ZHANG Jingzhi. Multi-strategy fusion of LSSVM-NGO for slidingelectrical contact failure diagnosis[J]. Journal of Electronic Measurement and Instrumentation, 2023, 37(12): 37-47.