文章编号:1671-4598(2025)09-0027-09 DOI:10.16526/j.cnki.11-4762/tp.2025.09.004 中图分类号:TP391 文献标识码:A

基于 SMOTE 和 GWO-XGBoost 的变压器故障诊断研究

吴秋伶[1]，刘孙俊[1]，王　杰[2]，王琪凯[1]，李　刚[1]，何俊江[3]
(1. 成都信息工程大学 软件工程学院，成都　610200；
2. 国网四川省电力公司 电力科学研究院，成都　610041；
3. 四川大学 网络空间安全学院，成都　610065)

摘要：为提高变压器故障诊断的准确性及降低样本不平衡对模型识别准确率的影响，提出了基于 SMOTE 和 GWO-XGBoost 的变压器故障诊断方法；该方法利用 SMOTE 技术扩大少数派样本，采用非编码比例法构建多维特征参数，并引入邻域粗糙集优化策略以及灰狼优化算法对 XGBoost 参数进行优化；实验验证显示，该方法显著减少了少数派样本的误判，并展示出高精度、低误判率及稳定性，适用于实际变压器故障诊断应用。

关键词：故障诊断；变压器；过采样；邻域粗糙集；XGBoost；灰狼优化算法

Research on Transformer Fault Diagnosis Based on SMOTE and GWO-XGBoost

WU Qiuling[1], LIU Sunjun[1], WANG Jie[2], WANG Qikai[1], LI Gang[1], HE Junjiang[3]
(1. School of Software Engineering, Chengdu University of information technology, Chengdu　610200, China;
2. State Grid Sichuan Electric Power Company Electric Power Research Institute, Chengdu　610041, China;
3. College of Cyberspace Security, Sichuan University, Chengdu　610065, China)

Abstract: To enhance the accuracy of transformer fault diagnosis and reduce the effect of sample imbalance on the recognition accuracy of models, a transformer fault diagnosis method based on synthetic minority over-sampling technique (SMOTE) and gray wolf optimization and extreme gradient boosting (GWO-XGBoost) is proposed. This method utilizes the SMOTE technology to expand minority samples, employs a non-coding ratio method to construct multidimensional feature parameters, and introduces neighborhood rough set optimization strategies and a gray wolf algorithm to optimize the XGBoost parameters. Experimental validation demonstrates that the method significantly reduces the misjudgment of minority samples, with a high precision, low misjudgment rate, and stability, making it suitable for practical applications in transformer fault diagnosis.

Keywords: fault diagnosis; transformer; oversampling; domain rough set; XGBoost; GWO algorithm

0　引言

电力变压器是输变电系统中的关键设备，其运行状态关系到电力系统的稳定。当变压器发生故障时，如果不能及时做出准确的诊断，将会造成重大的经济损失[1]。因此，如何提高变压器故障诊断的准确性一直是学者们研究的热点问题。

随着变压器绝缘老化过程的进行，油浸式变压器将产生 H_2、CH_4、C_2H_6、C_2H_4、C_2H_2、CO_2 等气体，并溶解在绝缘油中。变压器的现状可以从油中这些溶解气体的浓度和组成推断出来。用于评估变压器状况的主要分析技术包括 IEC 三比值法、罗杰斯四比值法、杜瓦

收稿日期:2024-07-24；　**修回日期**:2024-09-14。
基金项目:国家自然科学青年基金项目(62101358)；四川省科技计划重点研发计划项目(2023YFG0294)。
作者简介:吴秋伶(1999-)，男，硕士研究生。
通讯作者:刘孙俊(1975-)，男，博士，副教授。
引用格式:吴秋伶，刘孙俊，王　杰，等. 基于 SMOTE 和 GWO-XGBoost 的变压器故障诊断研究[J]. 计算机测量与控制，2025，33(9)：27-35.

尔·五角大楼法、多恩伯格比法等[2]。这些传统的比值法在油浸式变压器故障诊断中各有优缺点。IEC 三比值法[3]简单直观，适用于初步快速诊断，尤其对局部故障敏感，然而，不能有效区分复杂的故障类型。罗杰斯四比值法[4]相比之下增强了故障类型诊断能力，但仍受比值法的局限。杜瓦尔·五角大楼法引入了 5 种气体的比值，提高了诊断精度，但计算复杂度较高。多恩伯格比法进一步综合了 6 种气体的比值，对复杂故障有更好的敏感性，但计算量大且需要高效的计算支持。因此，选择合适的方法应根据具体的变压器情况和诊断需求来决定。这些传统故障诊断方法虽然有一定的诊断效果，但不能充分利用溶解气体分析（DGA，dissolved gas analysis）数据与故障类型的内在联系。挖掘 DGA 数据的本质特征信息能发现早期故障和及时制定解决方案，对保护变压器安全运行有重大意义[5]。

支持向量机[6]、卷积神经网络[7]、自适应增强[8]、梯度增强决策树[9]等模型在分类识别方面都取得了显著的成功。然而，上述的故障诊断模型都是建立在有一个相对较大的数据集。然而，在实际操作中，变压器很少发生故障，不同类型故障发生的频率差别很大[10]。这使它变得难以满足数据样本的精度要求。因此，在实际的变压器故障诊断中，采样不平衡问题需要引起人们的高度重视[11]。

解决非均衡数据集的问题方法主要包括数据预处理与模型算法改进两个方面[12]。数据预处理主要包括数据欠采样删减多数类样本和数据过采样增加少数类样本。这种方法可能导致丢失关于大量样本类的关键信息，最终损害分类器的性能。过采样人为地增加有限的样本量以实现数据平衡。这可以通过合成少数类的过采样技术（SMOTE，synthetic minority over-sampling technique）[13]，SVM-SMOTE[14]，Borderline-SMOTE[15]，自适应合成采样（ADASYN，adaptive synthetic sampling)[16]。文献［17］提出了一种基于支持向量机和合成少数类过采样算法的电力变压器故障样本均衡化方法，并结合机器学习进行故障诊断，解决了不平衡数据集下变压器故障诊断整体精度低的问题。文献［18］提出了一种结合改进的 Crow 搜索算法和 XGBoost（XGBoost，extreme gradient boosting）的增强型变压器故障诊断模型，提高变压器故障诊断性能方面的有效性。

极端梯度提升是 GBDT（GBDT，extreme gradient lift）基础上对 boosting 思想的扩展[19]。XGBoost 是一种集成学习算法，通过结合多个弱分类器（通常是决策树）来构建一个强分类器[20]。它通过串行训练多个模型，并根据前一个模型的表现调整后续模型的预测结果，从而提升整体预测的准确性。优化 XGBoost 模型的超参数是提高模型性能的关键步骤之一。文献［21］提出了随机森林递归特征消除算法与改进麻雀算法优化极端梯度提升树的变压器故障诊断方法，有效地提高变压器故障诊断性能。文献［22］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的油浸式变压器故障诊断方法，验证了遗传算法对故障诊断模型的优化提升效果。

本工作的主要内容如下：

1）利用过采样方法提高了对不平衡和小样本数据的分类性能，避免了分类器过于关注多数样本而导致分类器的超平面向少数类样本偏移。

2）建立了油中溶解气体与故障类型的深层关系，利用邻域粗糙集[23]特征选择减少了特征之间的冗余，提高了诊断模型的计算效率。

3）灰狼优化算法对诊断模型进行参数优化，建立最优诊断模型。最后，通过不同的采样方法和不同的诊断模型验证了所提方法的有效性。

1 基于 SMOTE 采样的 GWO-XGBoost 故障诊断

1.1 SMOTE-GWO-XGBoost 算法整体框架

为了提升变压器故障诊断结果的准确性，本文采用了 SMOTE 对采集到的变压器不平衡数据进行预处理。这一方法有效地平衡了少数样本与多数样本之间的比例，从而增强了模型对不同故障类型的识别能力。接着，本文提出了一种基于灰狼优化算法的 XGBoost 模型，以更精准地识别变压器的各种故障类型。该模型的学习框架如图 1 所示，展示了其结构和工作流程。通过这一系列步骤，本文旨在实现更高的故障诊断准确性，为变压器的安全运行提供有力保障。

图 1 SMOTE-GWO-XGBoost 算法整体框架

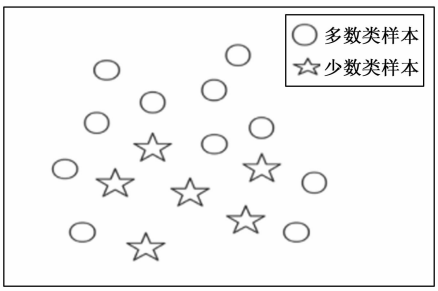
1.2 合成少数派过采样技术

SMOTE（合成少数类过采样技术）是一种广泛应用于机器学习和数据挖掘中的方法，旨在解决类别不平衡问题。在油浸式变压器的故障诊断任务中，故障数据通常存在显著的不平衡现象，某些类别的样本数量远远少于其他类别。这种不平衡会对模型的训练和预测性能产生负面影响，导致模型对少数类的识别能力不足。为了解决这一问题，SMOTE 通过合成新的少数类样本来增加其在数据集中的比例，从而实现类别平衡。其核心思想是随机选择一个来自多数类的样本，然后在其邻域内找到 k 个最近的邻居。接下来，根据一定的抽样概率，从这 k 个邻居中随机选择一个样本，并利用公式

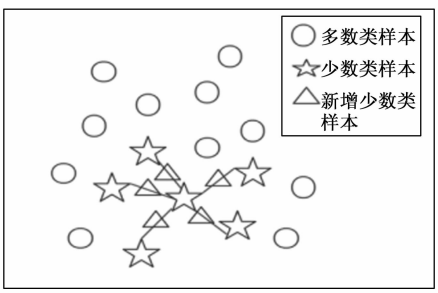
(1) 生成一个新的合成样本。通过这种方式，SMOTE 可以反复生成新的少数类样本，逐步平衡数据集中的各个类别。这种方法不仅能够有效增加少数类样本的数量，还能保持样本的多样性，从而提高模型在处理不平衡数据时的鲁棒性和准确性。因此，SMOTE 在油浸式变压器故障诊断等领域中，成为了一种重要的预处理技术，帮助研究人员和工程师更好地应对类别不平衡带来的挑战：

$$Y = Z_i + \text{rand} \times (Z_1 - Z_2) \tag{1}$$

其中：Z_1 为多数类样本；Z_2 是最接近 Z_i 的 k 个样本之一；Y 属于 [0，1] 的随机数；Y 代表新生成的少数族裔样本。SMOTE 过采样方法说明如图 2 所示。



(a) 未处理样本分布



(b) 处理后样本分布

图 2　SMOTE 过采样方法说明

1.3　XGBoost 算法

XGBoost 是一种高效的梯度提升算法，它在决策树的基础上进行模型训练。XGBoost 结合了多种优化技术，使其在处理大规模数据和提升模型性能方面表现优异。这里是 XGBoost 的基本原理及其核心公式。

XGBoost 通过将模型表示为一系列树的加权和来进行预测：

$$F(x) = \sum_{m=1}^{M} f_m(x) \tag{2}$$

其中：$F(x)$ 是最终的预测结果 $f_m(x)$ 是第 m 棵树。

XGBoost 的目标是最小化损失函数和正则化项（防止过拟合）：

$$L(\theta) = \sum^{i=1} Nl[y_i, F(x_i)] + \sum_{m=1}^{M} \Omega(f_m) \tag{3}$$

其中：$l[y_i, F(x_i)]$ 是损失函数，衡量预测值 $F(x_i)$ 和真实值 y_i 之间的差距；$\Omega(f_m)$ 是树模型的复杂度，通常定义为：

$$\Omega(f) = \gamma T + 1/2\ \lambda \|\omega\|^2 \tag{4}$$

这里，T 是树的叶子节点数，ω 是每个叶子节点的权重，γ 和 λ 是正则化参数。

每次迭代中，XGBoost 增加一颗新树，树的构建依赖于当前模型的负梯度，以此来拟合残差：

$$g_i = \frac{\partial l[y_i, F(x_i)]}{\partial F(x_i)} \tag{5}$$

$$h_i = \frac{\partial^2 l[y_i, F(x_i)]}{\partial F(x_i)^2} \tag{6}$$

其中：g_i 和 h_i 分别表示损失函数的一阶和二阶导数。

XGBoost 的主要参数包括：学习率（learning rate），控制每次迭代中模型参数更新的幅度；树的深度（max _ depth），控制每棵树的最大深度；树的个数（n _ estimators），决定了最终模型的复杂度；正则化参数（reg _ alpha，reg _ lambda），控制了模型的复杂度；样本采样参数（subsample，colsample _ bytree），控制了训练每棵树时的样本采样策略。

1.4　灰狼优化算法

灰狼优化算法（GWO，grey wolf optimization）是一种受到灰狼社会行为启发的优化算法，用于解决优化问题。它模拟了灰狼群体中个体之间的社会行为，包括追逐、狩猎和群体协作，通过模拟这些行为来调整参数以寻找最优解。

XGBoost 是一种梯度提升算法，被广泛应用于机器学习中的分类和回归问题。XGBoost 的性能高度依赖于其超参数的设置，包括树的深度、学习率、子样本比例等。

首先需要建立灰狼社会等级层次模型，计算每个个体的适应度，然后从种群中选择适应度最好的三只灰狼，分别标记为 α、β、δ。GWO 的优化过程主要由这 3 个最佳解（即 α、β、δ）来指导完成每一代种群的演化。

灰狼搜索猎物时会逐渐地接近猎物并包围它，该行为数学模型如下：

$$D = |\ C\boldsymbol{X}_p(t) - X(t)\ | \tag{7}$$

$$X(t+1) = \boldsymbol{X}(t) - AD \tag{8}$$

$$\boldsymbol{A} = 2\alpha \cdot \boldsymbol{r}_1 - \alpha \tag{9}$$

$$\boldsymbol{C} = 2 \cdot \boldsymbol{r}_2 \tag{10}$$

式中，t 表示当前的迭代数；$\boldsymbol{A}$ 和 $\boldsymbol{C}$ 表示协同系数向量；$\boldsymbol{X}_p$ 表示猎物的位置向量；$\boldsymbol{X}(t)$ 表示灰狼的位置向量；D 表示灰狼与猎物之间的坐标距离；在整个迭代过程中 α 由 2 线性降到 0；$\boldsymbol{r}_1$、$\boldsymbol{r}_2$ 是 [0，1] 的随机数向量。

灰狼具备识别潜在猎物位置的能力，其搜索过程主要依赖于 α、β、δ 的引导。由于猎物的具体坐标未知，

其最优位置也未知。为了模拟灰狼的搜索行为，假设 α、β、δ 具有较强的潜在猎物识别能力。因此，在每次迭代过程中，我们保留当前种群中的最佳 3 只灰狼（α、β、δ），然后利用它们的位置信息来更新其他搜索代理的位置。在位置的迭代计算公式中引入了一个随机项 rand，该项取值范围为 [0. 01，0. 1]，以模拟灰狼的随机搜索行为。本研究中所有参数已归一化至 [0，1] 区间内，这种随机性的数学模型可以表述如下：

$$D_\alpha = | C_1 X_\alpha(t) - X(t) | \tag{11}$$

$$D_\beta = | C_2 X_\beta(t) - X(t) | \tag{12}$$

$$D_\delta = | C_3 X_\delta(t) - X(t) | \tag{13}$$

$$X_1 = X_\alpha(t) - A_1 D_\alpha + \text{rand} \tag{14}$$

$$X_2 = X_\beta(t) - A_2 D_\beta + \text{rand} \tag{15}$$

$$X_3 = X_\delta(t) - A_3 D_\delta + \text{rand} \tag{16}$$

$$X(t+1) = (X_1 + X_2 + X_3)/3 \tag{17}$$

式中，X_α（t）、X_β（t）、X_δ（t）分别为 t 代时 α、β、δ 的位置。

将灰狼优化算法用于优化 XGBoost 多分类问题的原理是通过调整 XGBoost 模型的超参数，如学习率、树的数量、树的深度等，来优化模型的性能。灰狼优化算法可以帮助搜索最优的超参数组合，从而提高 XGBoost 模型的分类准确度和泛化能力。具体步骤包括：

1）初始化一群灰狼，每只灰狼表示一个可能的超参数组合。

2）计算每只灰狼的适应度，即 XGBoost 模型在训练集上的分类准确度。

3）根据适应度，更新每只灰狼的位置，模拟狼群的行为。

4）不断迭代，直到找到最优的超参数组合，使 XGBoost 模型在验证集上的分类准确度最高。

通过灰狼优化算法优化 XGBoost 多分类问题，可以更快速地找到最优的超参数组合，提高模型性能，减少模型训练时间，提高模型的泛化能力。

1.5 GWO-XGBoost 算法构建

XGBoost 模型的超参数众多，导致同时优化这些参数变得相对困难。为了解决这一问题，可以采用灰狼优化算法（GWO）来优化模型的超参数，从而进一步提升 XGBoost 模型的预测能力。在 XGBoost 模型的参数优化阶段，狼群中每个个体的适应度函数定义为该个体对应参数下，XGBoost 模型预测结果与真实数据之间的均方根误差（RMSE，root-mean-square error）。通过不断迭代和更新狼群的位置，GWO 能够有效地搜索参数空间，并在 RMSE 达到全局最小值时，确认已找到 XGBoost 的最佳参数组合。这一过程不仅提高了模型的准确性，还增强了其在实际应用中的可靠性。因此，结合 GWO 的优化方法，可以显著提升 XGBoost 模型在复杂数据集上的预测性能。

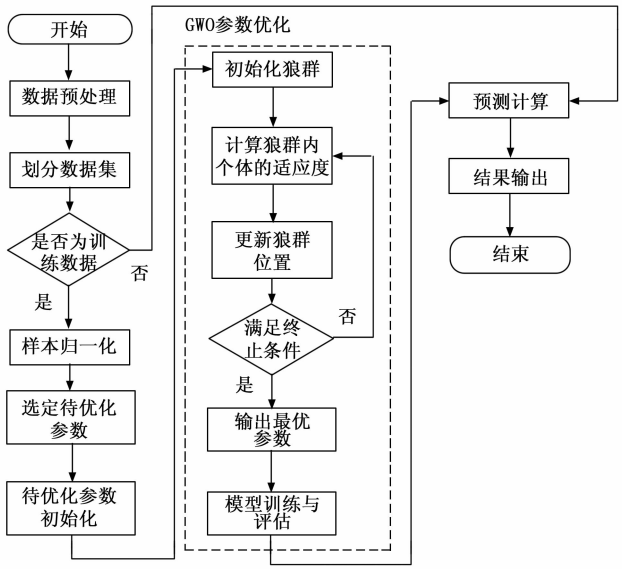


图 3 GWO 算法优化 XGBoost 模型流程示意图

2 基于 SMOTE 采样的 GWO-XGBoost 故障诊断

2.1 数据分析与处理

据 DL/T722-2014《变压器油中溶解气体分析判断准则》[24]，根据变压器是否发生故障和故障类型，将变压器分为 6 种类型，用标签 0～5 表示，分别为低能放电（D1）、高能放电（D2）、中低温放热（T1&T2）、高温放热（T2）、局部放热（PD）、正常（N）。本文选取某供电公司提供的 1 287 组监测数据作为故障样本集。样品集中的每一种工作状态包括 H_2、CH_4、C_2H_6、C_2H_4、C_2H_2 5 种特征气体。原始数据集分布表如表 1 所示。

表 1 数据集分布表

状态类型	数量	占比/%	标签
N	27	2	0
D1	373	28	1
D2	234	18	2
T2	286	22	3
T1&T2	147	11	4
PD	220	17	5

分析可知，中低温过热故障仅占 11%，在所有情况下，低能放电故障仅 28%。如果用于模型构建的数据集是不平衡的，模型可能无法对某些样本类型获得足够的熟练度，导致在识别这些样本类型时出错的可能性增加，从而影响模型的分类精度。在本研究中，我们使用 SMOTE 来平衡数据集。SMOTE 过采样后的样分布如图 4 所示。

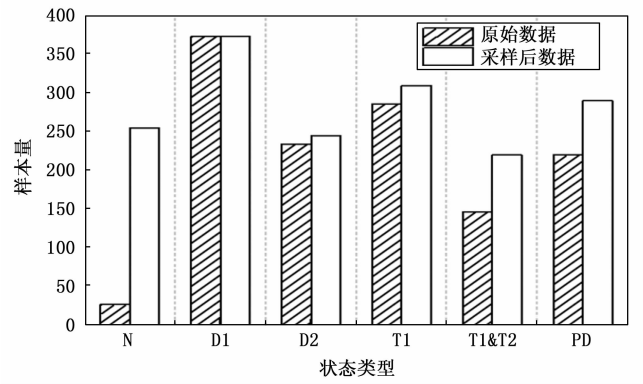


图 4　原始数据与采样后数据对比表

2.2　特征优化

传统油浸式变压器故障诊断方法表明，油浸式变压器发生故障时，产生的故障气体之间的比值与所对应的故障类型之间有很强的关联性。经分析，故障气体中的 H_2、CH_4、C_2H_6、C_2H_4、C_2H_2 这 5 种气体之间的比值对判断油浸式变压器内部运行状态较为重要。本文采用无编码比值法构建了 25 种气体比值，具体结果见表 2。

表 2　特征气体比值表

编号	比值	编号	比值	编号	比值	编号	比值
1	CH_4/H_2	8	C_2H_4/C_2H_2	15	C_2H_6/C_2	22	$CH_4/(C_1+C_2+H_2)$
2	C_2H_2/H_2	9	C_2H_6/C_2H_2	16	$H_2/(C_1+C_2)$	23	$C_2H_2/(C_1+C_2+H_2)$
3	C_2H_4/H_2	10	C_2H_6/C_2H_4	17	$CH_4/(C_1+C_2)$	24	$C_2H_4/(C_1+C_2+H_2)$
4	C_2H_6/H_2	11	H_2/C_2	18	$C_2H_2/(C_1+C_2)$	25	$C_2H_6/(C_1+C_2+H_2)$
5	C_2H_2/CH_4	12	CH_4/C_2	19	$C_2H_4/(C_1+C_2)$		
6	C_2H_4/CH_4	13	C_2H_2/C_2	20	$C_2H_6/(C_1+C_2)$		
7	C_2H_6/CH_4	14	C_2H_4/C_2	21	$H_2/(C_1+C_2+H_2)$		

尽管本文构建了一个 25 维的特征空间以充分利用数据信息，但也可能存在信息冗余问题，这些冗余特征会增加模型的计算负担。因此，有必要减少数据维度，简化数据模型的复杂度。为此，引入邻域粗糙集方法对这 25 种特征进行了优化。通过删除重要性较低的特征气体比率，最终得到了与故障标签高度相关的 10 种重要特征气体比率，具体结果见表 3。其中 C_1 代表一阶碳氢化合物，C_2 代表二阶碳氢化合物之和（即 $C_2H_6+C_2H_4+C_2H_2$）。

表 3　删减后的特征气体比值表

编号	比值	编号	比值
2	C_2H_2/H_2	17	$CH_4/(C_1+C_2)$
4	C_2H_6/H_2	19	$C_2H_4/(C_1+C_2)$
11	H_2/C_2	20	$C_2H_6/(C_1+C_2)$
13	C_2H_2/C_2	22	$CH_4/(C_1+C_2+H_2)$
14	C_2H_4/C_2	23	$C_2H_2/(C_1+C_2+H_2)$

2.3　评价指标

在数据分类不平衡的情况下，来自优势类的过多样本可能导致模型倾向于过度关注大多数类别，从而忽略了少数选择。这可能导致分类器的平面向这些类别内的样本子集移动。为了有效评价转换后的变压器故障诊断模型的有效性，本文选择了基于混淆矩阵的多分类评价指标体系，以准确率、召回率、F_1 值、g 均值和 *Kappa* 系数作为模型评价指标。

准确度是预测阳性样本与实际阳性样本的比例。召回率表示预测阳性样本占实际阳性样本结果的比例：

$$P=\frac{TP}{TP+FP} \tag{15}$$

$$R=\frac{TP}{TP+FN} \tag{16}$$

其中，P 表示准确率；R 表示召回率；TP 是分类时的情况的阳性样本是正确的；FP 是反例样本被错误分类的情况；FN 是阳性样本被错误分类的情况。

F_1 值代表准确率和召回率的调和平均值：

$$F_1=\frac{2PR}{P+R} \tag{17}$$

Kappa 系数反映了实际分类与预测分类的一致性，是评价故障诊断准确性的常用指标之一：

$$Kappa=\frac{p_0-p_e}{1-p_e} \tag{18}$$

其中：P_0 为正确预测样本数除以样本总数。

假设每个类别的真实样本分别为 a_1，a_2，…，a_n，预测的未分类样本分别为 b_1，b_2，…，b_n：

$$p_e=\frac{a_1\times b_1+a_2\times b_2+\cdots+a_n\times b_n}{n\times n} \tag{19}$$

Kappa 系数取值范围为 [0，1]，一般分为 5 组，分别代表不同的一致性水平：0～0.20（极低一致性）、0.21～0.40（一般一致性）、0.41～0.60（中等一致性）、0.61～0.80（高一致性）、0.81～10（几乎相同）。即 *Kappa* 系数越接近 1，诊断效果越好。

3　故障诊断结果对比分析

本文的仿真实验均在 AMD Ryzen 5 5600 6－Core Processor、3.5 GHz 内存 32 GB 环境下，利用 Pycharm（Professional Edition）作为实验仿真平台和 Python3.9 进行编程。

3.1　样本均衡效果的对比分析

为了验证诊断模型处理不平衡数据的有效性，利用

XGBoost 故障诊断模型，从数据处理层面采用随机过采样和 ADASYN 过采样方法对采用 SMOTE 采样的诊断结果进行比较。

将通过不同采样方法采样后的数据划分为无编码比值作为特征参量输入 XGBoost 模型，在默认参数下对不同采样方法采样后的 DGA 数据进行训练。4 组不同采样方法所得到的测试集故障诊断详细结果表见表 4。根据诊断结果，可以看出过采样后，模型的诊断准确率和 *Kappa* 系数均得到提高。在相同实验条件下，采用 SMOTE 采样诊断准确率为 96.45%，采用 ADASYN 采样诊断准确率为 94.83%；采用随机采样诊断准确率为 94.68%；采用不平衡数据集直接诊断准确率为 86.21%。在本文中使用 SMOTE 进行数据增强之后，诊断每种抽样方法的准确性和综合指标均优于其他抽样方法，进一步验证了该方法的有效性。该方法在处理不平衡数据方面具有优越性。

表 4 不同采样方法下诊断结果

数据采样方法	准确率/%	*Kappa* 系数
SMOTE 采样	96.45	0.973 4
ADASYN 采样	94.83	0.938 6
随机采样	94.68	0.925 1
不平衡数据集	86.21	0.836 0

3.2 不同优化算法诊断效果对比分析

为了验证本文提出的改进方法在变压器故障识别中能够有效地提高准确率，通过多个评价指标将 GWO-XGBoost 与 DBO-XGBoost、NGO-XGBoost、BO-XGBoost 和 XGBoost 进行了比较，以验证模型的诊断效果。

超参数进行优化调整是改善 XGBoost 诊断性能的有效方法。利用灰狼优化算法对 n _ estimators、learning _ rate，max _ depth 共 3 个关键参数进行寻优，以正确率作为优化的目标函数。为了测试 GWO 算法优化 XGBoost 超参数的性能，对比 DBO 和 BO 对 XGBoost 的优化效果。

将经过 SMOTE 采样方法采样后的数据按比例分为训练集、测试集和验证集 6∶2∶2。具体分布如表 5 所示。

表 5 数据集分布表

状态类型	训练集	测试集	验证集
N	159	53	53
D1	213	71	71
D2	153	51	51
T2	177	59	59
T1&T2	129	43	43
PD	177	59	59

利用经过 SOMTE 采样处理过的数据集构建不同优化算法优化的 XGBoost 模型，测试结果见图 5。

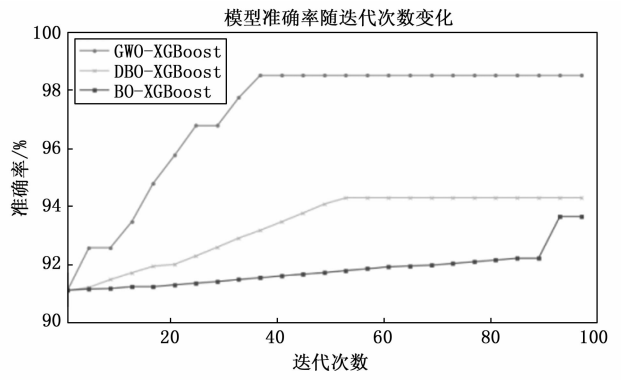


图 5 原始数据与采样后数据对比表

从图 5 可以看出，在 100 次迭代中，GWO-XGBoost 算法在第 38 次迭代正确率达到最大值 98.51%；DBO-XGBoost 算法在 56 次迭代正确率达到最大值 94.30%；BO-XGBoost 算法在 93 次迭代正确率达到最大值 93.66%。虽然 3 种算法都能在 100 次迭代里趋近全局最优，但是 GWO-XGBoost 算法在寻优能力方面强于另外两种算法。

表 6 展示了不同模型的诊断结果。通过优化算法对 XGBoost 模型的超参数进行优化，实验结果表明，该优化算法具备强大的寻优能力，能够显著提升模型的诊断性能。具体而言，GWO-XGBoost 相较于 DBO-XGBoost、BO-XGBoost 和传统的 XGBoost 模型，其诊断准确率分别提高了 4.21%、4.85%和 7.39%。通过对精度、召回率、F_1 值和 *Kappa* 系数等关键指标的深入分析，我们进一步验证了该方法的有效性。结果表明，本文提出的 GWO-XGBoost 模型在诊断效果上优于其他模型，充分证明了其在实际应用中的优越性。这一发现不仅增强了对模型性能的理解，也为未来的研究提供了新的方向。

表 6 模型对比结果分析

模型	精度	*Kappa*	准确率/%
GWO-XGBoost	0.985 8	0.981 9	98.51
DBO-XGBoost	0.9427	0.941 7	94.30
BO-XGBoost	0.937 2	0.935 2	93.66
XGBoost	0.912 5	0.911 8	91.12

在相同实验条件下，采用 GWO-XGBoost 与 DBO-XGBoost、BO-XGBoost 和 XGBoost 诊断拟合图如图 6。

如图 6 所示，使用 SMOTE 采样故障数据后，不同模型在不同故障类型上的分类情况。其中倒三角图标表示实际故障，五角星图标表示预测故障。若两种图标完全重合，则表示分类结果准确，反之分类结果错误。由

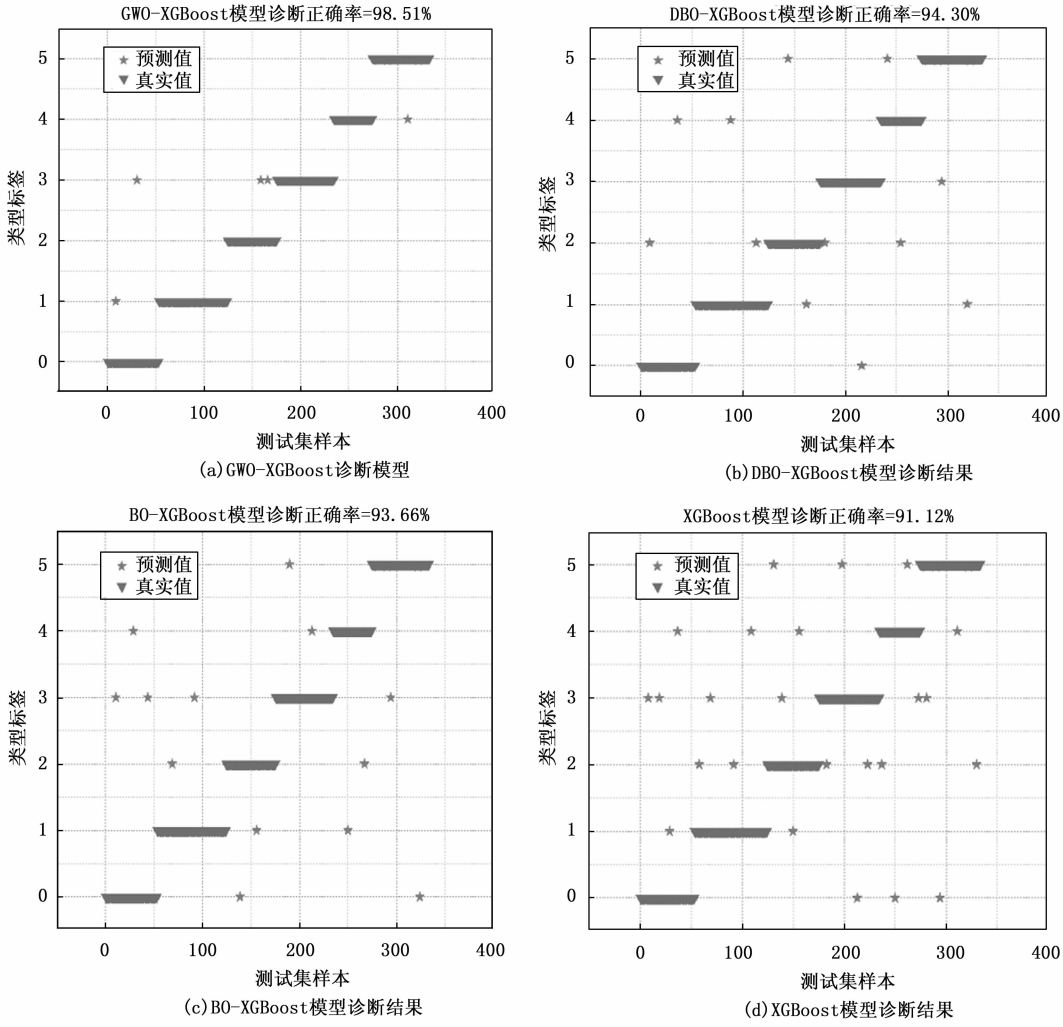


图 6　基于 SMOTE 采样的不同模型诊断结果比较

图 7 可知，GWO-XGBoost 模型诊断正确率最高，出现误判率最低。

DBO-XGBoost 模型较 XGBoost 模型诊断正确率提升 3.28 个百分点。BO-XGBoost 模型较 XGBoost 模型提升 2.54 个百分点。

3.3　故障诊断结果分析

图 7 展示了基于 SMOTE 和 GWO-XGBoost 的变压器故障诊断结果的混淆矩阵。图中的蓝色对角线表示真实样本中正确预测的数量，而每行数据的和则代表该类样本的总数。在图 7 的 336 个测试样本中，共有 5 个故障样本被误判，总体诊断准确率达到了 98.51%。在故障样本的识别中，模型成功地对低能放电样品、高温放热样品和中低温放热样品进行了正确的鉴定。值得注意的是，高能放电和局部放电样品的误判率分别仅为 3.92%和 1.69%，这表明本文提出的模型具有较高的稳定性和可靠性。

通过混淆矩阵中的信息，我们可以进一步计算出诊断模型的精度、召回率和 F_1 值，分别为 0.985 8、

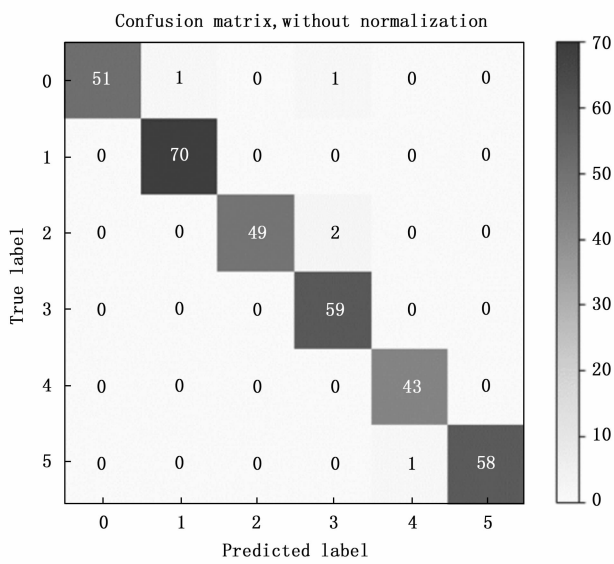


图 7　基于 SMOTE 采样的 GWO-XGBoost 算法诊断结果混淆矩阵

0.984 3 和 0.991 2。此外，模型的 *Kappa* 系数为 0.981 9，这意味着模型的真实分类与预测分类几乎完全一致，进一步证明了该模型在故障识别和分类方面的强大能力。总体而言，这些结果表明，基于 SMOTE 和 GWO-XGBoost 的诊断模型在变压器故障诊断中表现出色，具有较高的准确性和稳定性。

3.4 消融实验结果分析

在本节中，我们对消融实验的结果进行详细分析，以探讨各个组件对模型性能的影响。

为进行消融实验，设计如下几组实验。

1）基线模型：直接使用原始数据训练 XGBoost 模型，不进行任何采样处理或参数优化。

2）SMOTE 采样：使用 SMOTE 对数据进行过采样，然后训练 XGBoost 模型，但不进行参数优化。

3）GWO 优化：在原始数据上直接使用 GWO 优化 XGBoost 模型的参数，但不进行数据采样处理。

4）SMOTE 采样与 GWO 优化：使用 SMOTE 进行数据采样处理，然后使用 GWO 优化 XGBoost 模型的参数。

消融实验结果如表 7 所示。

表 7 消融实现结果对比分析

模型	准确率/%	*Kappa*	F_1 值
基线模型	86.21	0.836 0	0.882 4
SMOTE 采样	94.68	0.925 1	0.956 3
GWO 优化	91.12	0.911 8	0.923 7
SMOTE 采样与 GWO 优化	98.51	0.981 9	0.991 2

基线模型的准确率为 86.21%，而应用 SMOTE 采样后，准确率显著提升至 94.68%。这一结果表明，SMOTE 技术有效地缓解了数据不平衡问题，增强了模型对少数类样本的学习能力。通过生成合成样本，模型能够更好地捕捉到少数类的特征，从而提高了整体性能。

在与基线模型进行比较时，GWO 优化后的模型取得了 91.12%的准确率。尽管这一结果相比基线模型有所提升，但未能达到 SMOTE 采样模型的水平。这表明，虽然 GWO 优化在参数调整上起到了一定的作用，但单独依赖参数优化无法充分解决数据不平衡问题。

结合 SMOTE 采样与 GWO 优化后，模型的准确率进一步提升至 98.51%。这一结果清楚地表明，数据采样和参数优化的结合能够产生协同效应。SMOTE 采样为模型提供了更多的训练样本，而 GWO 优化则确保了模型在这些样本上的学习更加高效。因此，二者的结合显著提升了模型的性能。

最后，将基线模型与 SMOTE + GWO 模型进行比较，后者的准确率达到 98.51%，相较于基线模型的 86.21%有了显著的提升。这一结果进一步验证了数据采样和优化策略在提升模型性能方面的重要性。通过消融实验，我们可以得出结论：在处理不平衡数据集时，单独的优化或采样策略可能不足以达到理想效果，而两者的结合能够显著提高模型的准确性和稳定性。

通过消融实验的结果分析，我们可以清晰地看到每个组件对模型性能的贡献。SMOTE 采样在解决数据不平衡方面表现突出，而 GWO 优化则在提升模型学习效率上发挥了重要作用。二者的结合不仅提升了模型的准确率，也增强了其对少数类样本的识别能力，为今后的研究提供了重要的参考。

4 结束语

针对变压器故障数据不平衡导致的少数类样本误分类问题，提出一种基于 SMOTE 和 GWO-XGBoost 的变压器故障诊断方法。根据实证数据，得出以下结论：

1）利用邻域粗糙集特征选择方法选择最优特征子集，避免了冗余信息，有效地增强了变压器故障识别的准确性。

2）该研究在数据处理层面解决了故障样本的不平衡问题。SMOTE 过采样方法解决了数据不足和不平衡的问题，从而降低了诊断模型的误诊率。

3）与其他集成学习模型相比，构建了一个基于 XGBoost 的变压器故障诊断模型，具有较高的诊断精度。准确率、召回率、F_1 分数等评价指标进一步验证了所提方法的优越性。

综上所述，本文提出的策略能够实现电力变压器的在线诊断，提高变压器管理的运行效率；一定程度上解决了实际运行中故障样本发生的稀缺性和不平衡性。然而，SMOTE 仅通过插值生成新样本，没有考虑不同类别之间的真实分布和重叠情况。在某些情况下，合成样本可能会增加不同类别之间的重叠，导致分类器难以区分不同类别，影响了模型的诊断能力。本文对油中溶解气体的研究不够深入，忽略了两种不同的气体 CO 和 CO_2 对变压器故障的影响。深入分析和完善这些问题，需要进一步的研究。

参考文献：

[1] JIN Y，WU H，ZHENG J，et al. Power transformer fault diagnosis based on improved BP neural network [J]. Electronics，2023，12 (16)：3526.

[2] ALI M S，OMAR A，JAAFAR A S A，et al. Conventional methods of dissolved gas analysis using oil-immersed power transformer for fault diagnosis：A review [J]. Electric Power Systems Research，2023，216：109064.

[3] HONG L，CHEN Z，WANG Y，et al. A novel SVM-based decision framework considering feature distribution

for Power Transformer Fault Diagnosis [J]. Energy Reports, 2022, 8: 9392 - 9401.

[4] TAHA I B M, HOBALLAH A, GHPNEIM S S M. Optimal ratio limits of rogers'four-ratios and IEC 60599 code methods using particle swarm optimization fuzzy-logic approach [J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2020, 27 (1): 222 - 230.

[6] ZHANG H, SUN J, HOU K, et al. Improved information entropy weighted vague support vector machine method for transformer fault diagnosis [J]. High Voltage, 2022, 7 (3): 510 - 522.

[7] ZHAO H, YANG Y, LIU H, et al. Hierarchical spiking neural network auditory feature based dry-type transformer fault diagnosis using convolutional neural network [J]. Measurement Science and Technology, 2023, 35 (3): 036104.

[8] WANG L, CHI J, DING Y, et al. Transformer fault diagnosis method based on SMOTE and NGO-GBDT [J]. Scientific Reports, 2024, 14 (1): 7179.

[9] MENEZES A G C, ARAUJO M M, ALMEIDA O M, et al. Induction of decision trees to diagnose incipient faults in power transformers [J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2022, 29 (1): 279 - 286.

[10] 苗思雨，夏经德，邵文权，等. 基于铁磁特性虚拟相位在变压器保护中的研究 [J]. 自动化仪表，2023，44 (3)：26 - 33.

[11] 郭方洪，刘师硕，吴　祥，等. 基于联邦学习的含不平衡样本数据电力变压器故障诊断 [J]. 电力系统自动化，2023，47 (10)：145 - 152.

[12] 李艳霞，柴　毅，胡友强，等. 不平衡数据分类方法综述 [J]. 控制与决策，2019 (4)：16.

[13] CAMACHO L, DOUZAS G, BACAO F. Geometric SMOTE for regression [J]. Expert Systems with Applications, 2022, 193: 116387.

[14] BAGUI S S, MINK D, BAGUI S C, et al. Determining resampling ratios using BSMOTE and SVM-SMOTE for identifying rare attacks in imbalanced cybersecurity data [J]. Computers, 2023, 12 (10): 204.

[15] Al MAOUB H, ELGEDAWY I, AKAYDM Ö, et al. HCAB-SMOTE: A hybrid clustered affinitive borderline SMOTE approach for imbalanced data binary classification [J]. Arabian Journal for Science and Engineering, 2020, 45 (4): 3205 - 3222.

[16] RAMADHAN N G. Comparative analysis of ADASYN-SVM and SMOTE-SVM methods on the detection of type 2 diabetes mellitus [J]. Scientific Journal of Informatics, 2021, 8 (2): 276 - 282.

[17] 刘云鹏，和家慧，许自强，等. 基于 SVM SMOTE 的电力变压器故障样本均衡化方法 [J]. 高电压技术，2020，46 (7)：2522 - 2529，6520.

[18] JIANG J, LIU Z, WANG P, et al. Improved crow search algorithm and XGBoost for transformer fault diagnosis [C] //Journal of Physics: Conference Series, IOP Publishing, 2023, 2666 (1): 012040.

[19] WU Z, ZHOU M, LIN Z, et al. Improved genetic algorithm and xgboost classifier for power transformer fault diagnosis [J]. Frontiers in Energy Research, 2021, 9: 745744.

[20] 史佳琪，张建华. 基于多模型融合 Stacking 集成学习方式的负荷预测方法 [J]. 中国电机工程学报，2019，39 (14)：4032 - 4042.

[21] 王雨虹，王志中. 基于 RFRFE 与 ISSA-XGBoost 的变压器故障辨识方法 [J]. 电子测量与仪器学报，2021，35 (12)：142 - 150.

[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.

[23] BEI T, XIAO J, WANG X. Transient stability assessment of power systems based on the transformer and neighborhood rough set [J]. Electronics, 2024, 13 (2): 270.

[24] 栗　磊，王廷涛，赫嘉楠，等. 考虑过采样器与分类器参数优化的变压器故障诊断策略 [J]. 电力自动化设备，2023，43 (1)：209 - 217.

(上接第 26 页)

[11] 古莹奎，吴　宽，李　成. 基于格拉姆角场和迁移深度残差神经网络的滚动轴承故障诊断 [J]. 振动与冲击，2022，41 (21)：228 - 237.

[12] 王佳诺，王　奇，韩　健，等. 考虑弹性轮对的车辆－轨道耦合动力学建模及减振特性研究 [J]. 机械工程学报，2023，59 (8)：235 - 244.

[13] 丁春嵘，周雨轩，胡　浩，等. 基于深度特征提取神经网络的滚动轴承故障诊断 [J]. 北京化工大学学报：自然科学版，2022，49 (1)：106 - 112.

[14] 陈向民，韩梦茹，舒文伊，等. 基于 VMD 与多尺度一维卷积神经网络的故障诊断方法 [J]. 现代电子技术，2023，46 (9)：103 - 109.

[15] 薛富春，张建民，马建林，等. 运行条件下车辆－轨道－路基－地基强耦合大系统振动变形精细化分析方法 [J]. 应用基础与工程科学学报，2023，31 (1)：81 - 102.

[16] 王　博，左强新，刘伯奇，等. 高速铁路客站雨棚围护结构现状及列车风致振动检测 [J]. 铁道建筑，2023，63 (4)：99 - 103.

[17] 王道忠，李　广，宋亚东，等. 基于构架状态测量的轨道车辆车体横向振动估计 [J]. 振动与冲击，2023，42 (7)：162 - 169.