DOI: 10.19783/j.cnki.pspc.230205

基于样本扩充和特征优选的 IGWO 优化 SVM 的变压器故障诊断技术

欧阳鑫，李志斌

（上海电力大学自动化工程学院，上海 200090）

摘要：为了增强变压器故障诊断模型对不平衡样本的学习能力从而提高少数类故障样本的识别精度，提出了一种基于样本扩充和特征优选的融合多策略改进灰狼算法(improved grey wolf optimizer with multi-strategy, IGWO)优化支持向量机(support vector machine, SVM)的变压器故障诊断技术。首先，使用基于 K 最近邻过采样方法及核密度估计自适应样本合成算法的混合过采样技术对少数类样本进行扩充得到均衡数据集，并在此基础上采用方差分析对变压器候选比值征兆进行特征优选。然后，通过改进灰狼优化算法(grey wolf optimizer, GWO)初始化策略、参数及位置更新公式，并引入差分进化策略调整种群，提出了融合多策略的改进灰狼算法。最后，构建了一种基于混合过采样技术的 IGWO 优化 SVM 的变压器故障诊断模型，并通过多组对比实验验证了所提方法能够有效增强模型对少数类故障样本的识别能力，并提升模型的整体分类性能。

关键词：变压器故障诊断；不平衡数据集；混合过采样；特征优选；改进灰狼算法；支持向量机

Transformer fault diagnosis technology based on sample expansion and feature selection and SVM optimized by IGWO

OUYANG Xin, LI Zhibin

(College of Automation Engineering, Shanghai University of Electric Power, Shanghai 200090, China)

Abstract: To enhance the learning ability of a transformer fault diagnosis model for unbalanced samples and improve the recognition accuracy of minority fault samples, a transformer fault diagnosis technology based on sample expansion and feature optimization and support vector machine (SVM) optimized by improved grey wolf optimizer (GWO) with multi-strategy (IGWO) is proposed. First, the mixed oversampling technique based on K-nearest neighbor oversampling approach and kernel based adaptive synthetic algorithm is used to expand the minority samples to obtain the balanced datasets,, and analysis of variance (ANOVA) is used to select the transformer candidate ratio features. Then, by improving the initialization strategy and update formulas of parameters and positions of the GWO and introducing a differential evolution strategy to adjust populations, an improved GWO with multi-strategy is proposed. Finally, a transformer fault diagnosis model based on mixed oversampling technology and SVM optimized by IGWO is constructed, and experimental results show the method can enhance the recognition accuracy of the model for minority fault samples and improve the overall classification performance of the model effectively.

This work is supported by the National Natural Science Foundation of China (No. 51405286).

Key words: transformer fault diagnosis; unbalanced datasets; mixed oversampling; feature selection; improved grey wolf optimizer; support vector machine

0 引言

电力变压器一旦发生严重故障，不仅会影响电力系统安全稳定，而且会造成巨大的经济损失[1-3]，因此，建立高效、准确的变压器故障诊断模型有助于及时发现其潜在故障并制定合理、有效的应对策略，从而最大程度地保证电网的安全稳定运行[4]。

人工智能技术的快速发展不断推动着变压器故障诊断技术向智能化方向迈进。文献[5]采用集成学习思想构建了基于极端梯度提升树的变压器故障诊

基金项目：国家自然科学基金项目资助(51405286)；上海市青年科技英才杨帆计划资助(20YF1414800)；上海市电站自动化技术重点实验室项目资助(13DZ2273800)

断模型；文献[6-7]使用支持向量机(support vector machine, SVM)作为分类器建立了变压器特征量与状态量之间的映射关系，从而实现变压器故障诊断；文献[8-9]分别构建了基于改进深度耦合密集卷积神经网络及卷积门控循环单元的深度学习模型，并将其应用于变压器状态评估；文献[10]提出了一种基于随机森林特征优选，结合鲸鱼算法优化 SVM 的变压器故障诊断方法。上述方法的本质是通过机器学习模型分析变压器样本特征与运行状态之间的内在规律，从而提高变压器的故障诊断精度。然而，在实际工作中，电力变压器发生故障多为小概率事件，这将导致故障样本十分匮乏，出现非均衡数据集[11]。当训练数据高度不平衡时，分类模型通常会偏向于多数类样本而忽略少数类样本，导致少数类样本的识别精度降低，制约着机器学习模型的训练效果。

目前，针对非均衡数据集的处理问题，相关研究大多采用过采样算法。其中较具代表性的有合成少数类过采样技术(synthetic minority oversampling technique, SMOTE)[12]及其改进算法，如 ADASYN[13]、SVM SMOTE[14]、K-means-SMOTE[15]等。上述算法的应用在一定程度上实现了样本数据的均衡化，但都存在一定的缺陷。ADASYN 和 SVM SMOTE 算法专注于处理样本的边界区域，但并未考虑样本的类内分布情况和噪声混杂问题；K-means-SMOTE 算法基于 K-means 聚类簇进行样本扩充，可以有效增强样本类内聚合度，但并未考虑样本的边界特征，同时也并未对离群的噪声样本进行处理。

在分类器的选择方面，神经网络存在模型结构复杂、容易过拟合等缺点，相比之下，SVM 具有更强的泛化能力[16]。但 SVM 的分类性能与其相应参数的选取密切相关，因此如何利用智能优化算法对 SVM 模型进行参数调优成为目前研究的重点。灰狼优化算法(grey wolf optimizer, GWO)具有控制参数少、收敛速度快、寻优能力强等优点，近年来被广泛应用在故障诊断领域[17-18]。然而，GWO 存在全局勘探能力与局部搜索能力不协调、容易陷入局部最优解等固有缺陷。为此，国内外研究者提出了诸多改进策略[19-20]，并在一定程度上提升了 GWO 性能，但由于缺乏对这些固有缺陷的统筹改进，GWO 算法易陷入早熟停滞的问题仍然存在。

针对上述问题，本文提出了基于样本扩充和特征优选的 IGWO 优化 SVM 的变压器故障诊断技术。首先，利用基于 K 最近邻过采样方法及核密度估计自适应样本合成算法的混合过采样技术(k-nearest neighbor oversampling approach and kernel based adaptive synthetic, KNNOR-KernelADASYN)混合过采样技术对少数类样本进行扩充得到均衡数据集，同时采用方差分析(analysis of variance, ANOVA)对变压器候选比值征兆进行特征优选；然后选取 SVM 作为变压器故障诊断模型，并提出融合多策略的改进灰狼算法对 SVM 进行参数调优，以获得分类性能更强的诊断模型；最后，通过多组对比实验，验证本文所提方法的有效性和优越性。

1 基于 KNNOR-KernelADASYN 的混合过采样技术

1.1 KNNOR 算法

针对样本类内分布不均衡和类间噪声混淆等问题，文献[21]提出了 K 最近邻过采样技术(k-nearest neighbor oversampling appro-ach, KNNOR)，具体实现步骤如下：

1) 将数据集划分为多数类样本集 S_{maj} 和少数类样本集 $S_{\min}=\{x_1,x_2,\cdots,x_n\}$；

2) 通过 KNN 算法确定 $x_r(r=1,2,\cdots,n)$ 的 k 个近邻样本，将 $S_{\min}$ 中的所有样本按照其到其第 k 个近邻样本的距离进行升序排序，并保留排序后样本集的前 d %构成样本子集 $S'_{\min}=\{v_1,v_2,\cdots,v_m\}$，其中 d 为距离阈值；

3) 将 v_1 与其第 1 个近邻样本 $p_{1,1}$ 按照式(1)进行随机线性插值，形成合成样本 $v_{1,1}^{\text{new}}$，再将 $v_{1,1}^{\text{new}}$ 与 v_1 的第 2 个近邻样本 $p_{2,1}$ 进行相同操作，形成 $v_{2,1}^{\text{new}}$，如此迭代 k 次形成最终的合成样本 $v_{k,1}^{\text{new}}$，如图 1 所示。然后重复此操作，直到历遍 $S'_{\min}$ 中的所有样本。

$$\begin{cases} v_{0,I}^{\text{new}}=v_I \\ v_{J,I}^{\text{new}}=v_{J-1,I}^{\text{new}}+(p_{J,I}-v_{J-1,I}^{\text{new}})\cdot\theta_{J,I} \end{cases} \tag{1}$$

式中：$I=1,2,\cdots,m$；$J=1,2,\cdots,k$；$\theta_{J,I}\in[0,M]$，$M\leqslant 1$，M 的值取决于多数类中任意点和少数类中任意点间的最小距离。

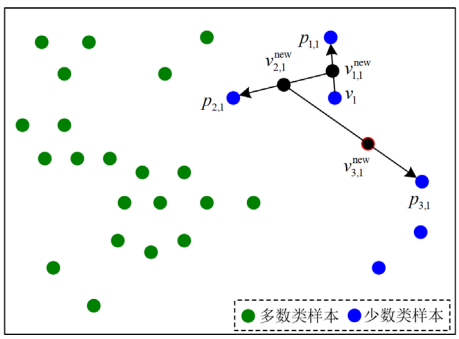

图 1 随机线性插值($k=3$)

Fig. 1 Random linear interpolation ($k=3$)

4) 利用 KNN 算法测试合成样本是否属于少数类样本，若属于，则保留该样本点，反之，则丢弃该样本点。

KNNOR 算法能够最大程度地保留原始数据的分布特征，通过步骤 2)可以避免选取离群噪声样本和分类边界处样本作为采样点，防止样本类间重叠和分类边界模糊，而步骤 3)的迭代过程可以有效解决样本的边缘分布问题，增强了样本的类内聚合性，同时结合步骤 4)可以避免生成新的噪声样本，最终增加了样本的多样性。

1.2 KernelADASYN 算法

文献[22]提出了基于核密度估计的自适应样本合成算法(kernel based adaptive synthetic, KernelADASYN)，其核心原理是应用核密度估计方法估计少数类样本的概率密度分布，并根据少数类样本的学习难易程度来度量它们的权值分布，处于决策边界处的少数类样本由于分布密度低、学习难度大而获得更高的权重，从而使合成的少数类数据更利于分类器在决策边界进行学习。

1.3 KNNOR-KernelADASYN 混合过采样

本文提出 KNNOR-KernelADASYN 混合过采样技术，首先通过 KNNOR 算法对同类样本内部进行扩充，增强样本的类内特征，然后使用 KernelADASYN 算法对不同类样本的边界进行扩充，增强样本的边界区分度，最终得到均衡化数据集。

2　融合多策略的改进 GWO

2.1 GWO

文献[23]提出了灰狼优化算法，将灰狼种群分为四个等级，包括α狼、β狼、δ狼和ω狼。种群的捕猎行为在α、β、δ狼的指导下进行，它们分别对应算法当前的最优解、次优解、第三优解，而ω狼则在它们的指挥下开展活动。灰狼群体的捕猎行为包括三个阶段：搜寻猎物、包围猎物、攻击猎物。其中包围猎物这一过程的数学模型表示为

$$\boldsymbol{X}(t+1)=\boldsymbol{X}_{\mathrm{P}}(t)-\boldsymbol{A}\cdot\left|\boldsymbol{C}\cdot\boldsymbol{X}_{\mathrm{P}}(t)-\boldsymbol{X}(t)\right| \tag{2}$$

式中：t、$\boldsymbol{X}_{\mathrm{P}}(t)$、$\boldsymbol{X}(t)$分别表示当前迭代次数、猎物的位置向量、灰狼个体的位置向量；$\boldsymbol{A}$、$\boldsymbol{C}$为系数向量，如式(3)所示。

$$\begin{cases}\boldsymbol{A}=2\boldsymbol{a}\cdot\boldsymbol{r}_1-\boldsymbol{a}\\ \boldsymbol{C}=2\cdot\boldsymbol{r}_2\end{cases} \tag{3}$$

式中：$\boldsymbol{r}_1$、$\boldsymbol{r}_2$为[0,1]间的随机向量；$\boldsymbol{a}$为收敛因子向量，其元素收敛因子a在迭代过程中由 2 线性递减到 0。

领头狼α、β、δ能够估计猎物的可能位置，并指导每一头ω狼进行位置更新，该过程可表示为

$$\begin{cases}\boldsymbol{X}'_\alpha=\boldsymbol{X}_\alpha-\boldsymbol{A}_\alpha\cdot\left|\boldsymbol{C}_\alpha\cdot\boldsymbol{X}_\alpha-\boldsymbol{X}\right|\\ \boldsymbol{X}'_\beta=\boldsymbol{X}_\beta-\boldsymbol{A}_\beta\cdot\left|\boldsymbol{C}_\beta\cdot\boldsymbol{X}_\beta-\boldsymbol{X}\right|\\ \boldsymbol{X}'_\delta=\boldsymbol{X}_\delta-\boldsymbol{A}_\delta\cdot\left|\boldsymbol{C}_\delta\cdot\boldsymbol{X}_\delta-\boldsymbol{X}\right|\end{cases} \tag{4}$$

$$\boldsymbol{X}(t+1)=(\boldsymbol{X}'_\alpha+\boldsymbol{X}'_\beta+\boldsymbol{X}'_\delta)/3 \tag{5}$$

式中，$\boldsymbol{X}_\alpha$、$\boldsymbol{X}_\beta$、$\boldsymbol{X}_\delta$及$\boldsymbol{X}'_\alpha$、$\boldsymbol{X}'_\beta$、$\boldsymbol{X}'_\delta$分别为α、β、δ狼的位置向量及对应的中间位置向量。

2.2 IGWO

1) Sobel 序列初始化种群

Sobel 序列产生的确定性拟随机数能够将尽可能均匀的点填充至多维超立方体，将其用于种群初始化，可以扩大算法的全局搜索范围，并提升算法的计算效率和规避局部极值的能力[24]。设算法全局解的取值范围为$[x_{\min},x_{\max}]$，由 Sobel 序列产生的第i个随机数为$S_i\subseteq[0,1]$，则种群初始位置可表示为

$$x_i=x_{\min}+S_i\cdot(x_{\max}-x_{\min}) \tag{6}$$

2) 非线性收敛因子调整控制参数

控制参数$\boldsymbol{A}$对于协调 GWO 全局搜索与局部开发能力起着至关重要的作用，当$|\boldsymbol{A}|>0$时，灰狼种群扩大搜索空间进行全局搜索；当$|\boldsymbol{A}|\leqslant 0$时，灰狼种群针对目标范围进行局部勘探。其中，收敛因子a采用线性时变更新策略，不能客观、真实地体现灰狼种群的实际搜索过程，容易导致算法陷入局部最优解。因此，本文引入非线性收敛因子更新策略[25]，如式(7)所示。

$$a=a_{\mathrm{state}}-(a_{\mathrm{state}}-a_{\mathrm{end}})\cdot\left(\frac{\mathrm{e}^{t/G}-1}{\mathrm{e}-1}\right)^u \tag{7}$$

式中：$a_{\mathrm{state}}=2$、$a_{\mathrm{end}}=0$分别为a的初始值和终止值；G为最大迭代次数；u为控制因子，用于控制a的衰减幅度。

3) 基于适应度比例权重的位置更新规则

GWO 使用当前种群中的三个最优个体来引导灰狼的位置更新，为了促使它们具有更明确的分工定位，同时进一步提升算法的寻优效率，本文引入一种基于适应度比例权重的位置更新规则[26]，如式(8)所示。

$$\begin{cases}f=f_\alpha+f_\beta+f_\delta\\ w_\alpha=\dfrac{f_\alpha}{f}\\ w_\beta=\dfrac{f_\beta}{f}\\ w_\delta=\dfrac{f_\delta}{f}\end{cases} \tag{8}$$

式中，f_α、f_β、f_δ及w_α、w_β、w_δ分别为α、β、

δ狼对应的适应度值及适应度比例权重。

4) 引入环境预警机制的位置更新公式

为了进一步扩大灰狼种群的搜索空间，加强其对潜在优良位置的探索，本文引入麻雀搜索算法[27]中的环境预警机制，当灰狼处于安全环境时，种群进行正常搜索，当灰狼处于危险环境时，种群重新寻找新的安全区域。灰狼个体位置更新公式最终表示为

$$X_{i,j}(t+1)=\begin{cases} w_\alpha \cdot X'_{\alpha,j}(t)+w_\beta \cdot X'_{\beta,j}(t)+w_\delta \cdot X'_{\delta,j}(t) & R<ST \\ X_{i,j}(t)+Q & R\geqslant ST \end{cases} \tag{9}$$

式中：$X_{i,j}$为个体i的第j维对应的位置信息；Q为服从标准正态分布的随机数；安全阈值$ST\in[0.5,1]$，预警值$R\in[0,1]$，当$R<ST$时，环境安全，反之，环境危险。

5) 差分进化调节种群

为了保证灰狼种群多样性并增强算法在迭代后期跳出局部最优解的能力，本文利用差分进化策略[28]在算法每次迭代环节末进行种群调节，其中变异、交叉和选择操作分别如式(10)—式(12)所示。

$$\boldsymbol{Y}_i(t)=\boldsymbol{X}_{\mathrm{r1}}(t)+Fr(t)\cdot(\boldsymbol{X}_{\mathrm{r2}}(t)-\boldsymbol{X}_{\mathrm{r3}}(t)) \tag{10}$$

$$Z_{i,j}(t)=\begin{cases} Y_{i,j}(t) & \text{if rand}<Cr(t)\ \text{or}\ j=j_{\mathrm{rand}} \\ X_{i,j}(t) & \text{otherwise} \end{cases} \tag{11}$$

$$\boldsymbol{X}_i(t+1)=\begin{cases} \boldsymbol{Z}_i(t) & \text{if } f_{\boldsymbol{Z}_i(t)}<f_{\boldsymbol{X}_i(t)} \\ \boldsymbol{X}_i(t) & \text{otherwise} \end{cases} \tag{12}$$

式中：$\boldsymbol{X}_{\mathrm{r1}}(t)$、$\boldsymbol{X}_{\mathrm{r2}}(t)$、$\boldsymbol{X}_{\mathrm{r3}}(t)$是从当前种群中随机选择的三个互不相同的个体(排除目标个体$\boldsymbol{X}_i(t)$)；j_{rand}表示一个随机的维度；$Fr(t)$为变异缩放因子，$Cr(t)$为交叉概率因子，本文参考文献[29]中的方式对Fr、Cr进行自适应调整，如式(13)所示。

$$\begin{cases} Fr(t)=0.5\cdot(\sin(2\pi l\cdot t)\cdot(t/G)+0.85) \\ Cr(t)=0.5\cdot(\sin(2\pi l\cdot t+\pi)\cdot(t/G)+1) \end{cases} \tag{13}$$

式中，l为进化调节参数。

综上所述，IGWO 的基本流程如图 2 所示。

3 方法实现过程

3.1 数据集详情

1) 变压器故障征兆

变压器发生故障时，绝缘油与固体绝缘材料会发生裂解并产生油中溶解气体，主要包括氢气(H_2)、甲烷(CH_4)、乙烷(C_2H_6)、乙烯(C_2H_2)、乙炔(C_2H_2)等。为了保证机器学习模型更准确地反映气体数据与故障类型的特征关联性，在参考传统比值法及相应文献采用的故障征兆的基础上[29-30]，本文共选取了 22 种变压器故障候选比值征兆，如表 1 所示，其中：$D=CH_4+C_2H_2+C_2H_4$；$T_G=H_2+CH_4+C_2H_2+C_2H_4+C_2H_6$；$T_H=CH_4+C_2H_2+C_2H_4+C_2H_6$。

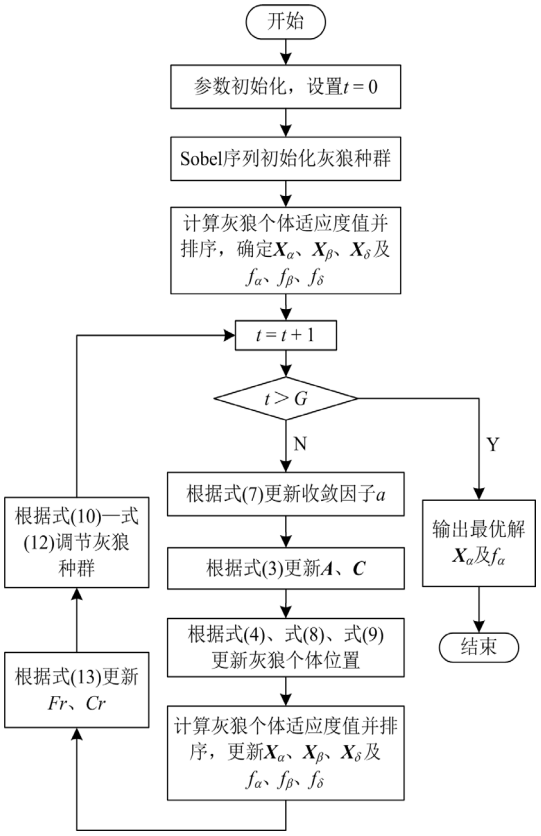

图 2 IGWO 流程图

Fig. 2 Flowchart of IGWO

表 1 变压器故障候选比值征兆

Table 1 Alternative ratios for transformer faults

特征编码	比值征兆	特征编码	比值征兆
S_1	CH_4/H_2	S_{12}	C_2H_2/D
S_2	C_2H_2/C_2H_6	S_{13}	H_2/T_G
S_3	C_2H_4/C_2H_6	S_{14}	CH_4/T_G
S_4	C_2H_6/CH_4	S_{15}	C_2H_6/T_G
S_5	C_2H_2/C_2H_4	S_{16}	C_2H_4/T_G
S_6	C_2H_2/CH_4	S_{17}	C_2H_2/T_G
S_7	C_2H_2/H_2	S_{18}	CH_4/T_H
S_8	H_2/D	S_{19}	C_2H_6/T_H
S_9	C_2H_6/D	S_{20}	C_2H_4/T_H
S_{10}	CH_4/D	S_{21}	C_2H_2/T_H
S_{11}	C_2H_4/D	S_{22}	H_2/T_H

2) 数据集来源

本文在 IEC TC 10 数据库[31]、文献[32-33]及华北某市电厂中共收集到的 875 组样本数据作为数据集，并按照 7:3 的比例将其分层划分为训练集、测

试集，如表 2 所示。

表 2 样本数据分布及标签

Table 2 Distribution and labels of samples

样本类型	类型编码	训练集/组	测试集/组	总计/组
中低温过热	1	81	35	116
高温过热	2	93	41	134
局部放电	3	49	21	70
低能放电	4	39	18	57
高能放电	5	95	42	137
正常	6	195	84	279
低能放电兼过热	7	28	12	40
电弧放电兼过热	8	29	13	42

3.2 数据平衡化处理

本文采用 KNNOR-KernelADASYN 混合采样算法按照逐次采样的方式对数据集进行平衡化处理，其中距离阈值 d 决定了样本的扩充范围，近邻样本数 k 决定了样本的局部增强程度，通过参考文献[21]并进行预实验，最终选取 $k=5$ 、$d\%=60\%$ 。首先以正常样本为负类样本、中低温过热样本为正类样本构建初始样本空间，利用混合采样算法对正类样本进行平衡化处理，然后将正常样本和处理后的中温过热样本作为新的负类样本，之后每次增加一类故障样本作为正类样本并对其进行平衡化处理，直到实现不平衡数据集的均衡化。训练数据采样前后的样本分布情况如表 3 所示。

表 3 训练数据采样前后的分布情况

Table 3 Distribution of the training data before and after sampling

样本类型编码	训练样本数量	
	采样前	采样后
1	81	191
2	93	188
3	49	192
4	39	190
5	95	191
6	195	195
7	28	192
8	29	190

3.3 样本特征选择

不同比值征兆丰富了故障特征的选择范围，但也增加了特征间的冗余性，降低诊断模型的计算效率，因此本文采用 ANOVA 对候选比值征兆进行优选，其主要原理可参考文献[34]。首先，利用 ANOVA 计算候选比值征兆的统计量 $F_{\text{-value}}$ ，并进行降序排序，如表 4 所示；然后，基于相同的训练集和测试集，按照逐维诊断的方式每次增加一个比值征兆组成候选子集，并以 SVM 的 5-折交叉验证识别率为优化目标，将其输入 IGWO-SVM 故障诊断模型中进行诊断；最后，选取最大目标值对应的候选子集作为最终的征兆子集。

表 4 变压器故障候选比值征兆的 $F_{\text{-value}}$

Table 4 $F_{\text{-value}}$ of the alternative ratios for transformer faults

排序	特征编码	$F_{\text{-value}}$	排序	特征编码	$F_{\text{-value}}$
1	S_{21}	289.24	12	S_{14}	67.87
2	S_{12}	272.81	13	S_2	49.98
3	S_{16}	235.70	14	S_{19}	44.02
4	S_{13}	229.14	15	S_1	27.15
5	S_{17}	202.18	16	S_9	17.49
6	S_5	172.25	17	S_3	12.18
7	S_{20}	154.02	18	S_7	9.58
8	S_{11}	137.42	19	S_{22}	9.56
9	S_{10}	126.88	20	S_6	8.45
10	S_{18}	100.46	21	S_8	7.69
11	S_{15}	77.81	22	S_4	2.64

如图 3 所示，当比值征兆维数 n 为 14~16 时，模型识别率达到最大值(93.99%)，考虑到模型诊断效率，本文选取 n 为 14 时所对应的候选子集(S_{21}、S_{12}、S_{16}、S_{13}、S_{17}、S_5、S_{20}、S_{11}、S_{10}、S_{18}、S_{15}、S_{14}、S_2、S_{19})作为最终的征兆子集。

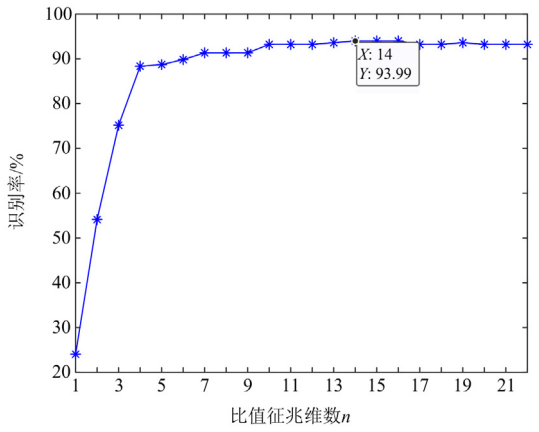

图 3 基于不同候选子集的故障诊断模型识别率

Fig. 3 Recognition rates of fault diagnosis model based on different alternative subsets

3.4 变压器故障诊断

本文采用径向基核函数作为 SVM 的核函数，并使用 IGWO 对 SVM 的惩罚因子 c 和核参数 σ 进行寻优。其中，IGWO-SVM 模型的参数设置情况如表 5 所示。为提高模型对少数类样本的识别率，本文选取模型 5-折交叉验证的平均 Kappa 系数作为 IGWO 的优化对象，选用(1-Kappa 系数)作为 IGWO 极小化的适应度函数。

表 5 IGWO-SVM 模型参数

Table 5 Parameters of IGWO-SVM model

参数名	数值
种群规模 N	100
搜索维度 D	2
最大迭代次数 G	100
控制因子 u	6
进化调节参数 l	0.45
惩罚因子 c	[1,100]
核参数 σ	$[10^{-3},1]$

3.5 变压器故障诊断模型架构

根据以上流程，基于混合过采样技术的 IGWO-SVM 变压器故障诊断模型结构如图 4 所示。

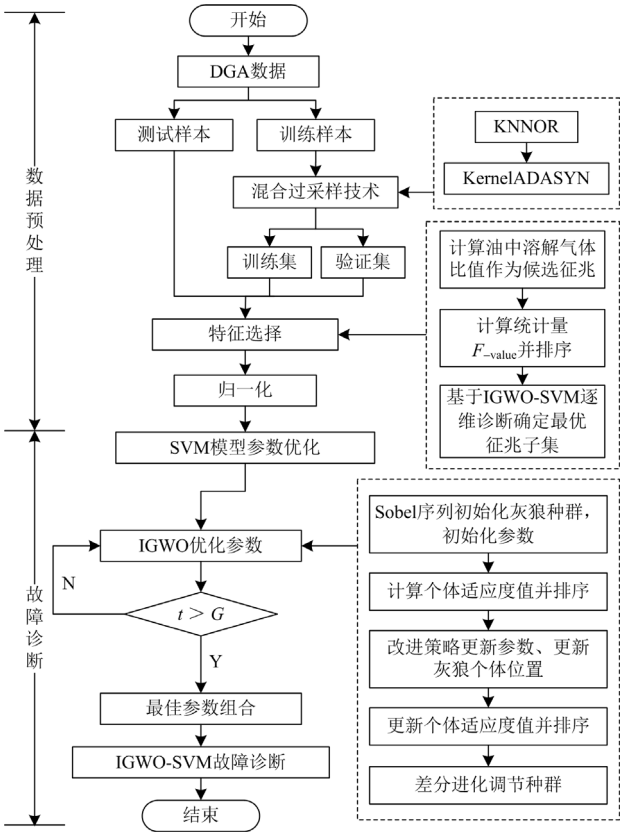

图 4 变压器故障诊断模型结构

Fig. 4 Structure of the transformer fault diagnosis model

3.6 模型性能评价指标

针对不平衡数据集，由于样本分布不均衡，少数类样本对总体识别率的影响较小，即使分类模型将全部样本视为多数类样本，诊断模型仍然可以获得较高的识别率。因此，本文引入基于混淆矩阵的分类模型性能评价指标，选取 R_{ecall}、P_{recision}、$F_{\text{1-score}}$ 来评估模型的分类性能。以二分类问题为例，令少数类样本为正类样本，多数类样本为负类样本，混淆矩阵如表 6 所示。

表 6 混淆矩阵

Table 6 Confusion matrix

样本类别	正类	负类
正类	*TP* (真正例)	*FN* (假负例)
负类	*FP* (假正例)	*TN* (真负例)

相应的评价指标如下：

$$\begin{cases} R_{\text{ecall}} = \dfrac{TP}{TP+FN} \\ P_{\text{recision}} = \dfrac{TP}{TP+FP} \\ F_{\text{1-score}} = \dfrac{2 \cdot R_{\text{ecall}} \cdot P_{\text{recision}}}{R_{\text{ecall}} + P_{\text{recision}}} \end{cases} \tag{14}$$

R_{ecall} (查全率)用来表征分类模型对正类样本识别的敏感度，P_{recision} (查准率)用来表征分类模型对正类样本判断的可信度，$F_{\text{1-score}}$ 为 P_{recision} 和 R_{ecall} 的调和平均值，用来表征模型的综合性能。

4 实验结果对比与分析

4.1 本文方法性能评价

为了验证本文方法的可行性和有效性，按照第 3 节进行仿真实验。

改进前后的灰狼适应度迭代曲线如图 5 所示。其中，GWO 在第 17 代时陷入早熟收敛，寻优效果并不理想，相比之下，IGWO 不仅具有更优的初始适应度值，而且在迭代前期能够多次跳出局部极值，最终在第 45 代时收敛，对应的最佳 Kappa 系数为 0.977，可以验证本文针对 GWO 提出的改进策略能够生成质量更高的初始种群，并有效平衡算法的全局探索与局部开发能力、避免算法陷入局部最优并获得更好的全局解。

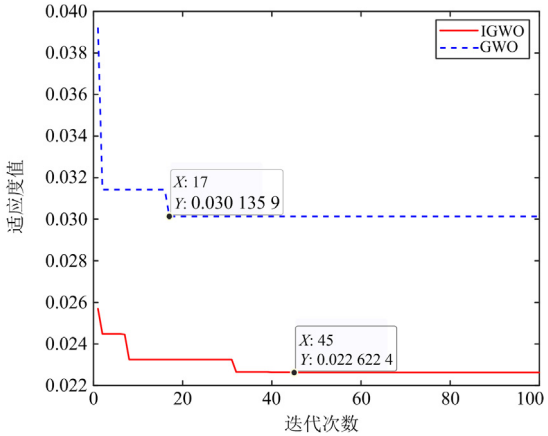

图 5 灰狼适应度迭代曲线

Fig. 5 Fitness iteration curves of grey wolf

图 6 及表 7 分别展示了 IGWO-SVM 模型仿真结果的可视化混淆矩阵及评价指标，从少数类样本的泛化特性上看，本文模型的查全率及查准率在 7 种少数类样本上的分布区间较为稳定，均高于 91%，同时各类样本的 $F_{1\text{-score}}$ 均高于 0.94，模型整体的 F_1 分数(各类样本 $F_{1\text{-score}}$ 均值)高达 0.963，诊断精度为 96.2%，Kappa 系数为 0.954，可以验证该模型的少数类泛化性能强、故障识别敏感度及判断可信度高、样本整体分类能力强。

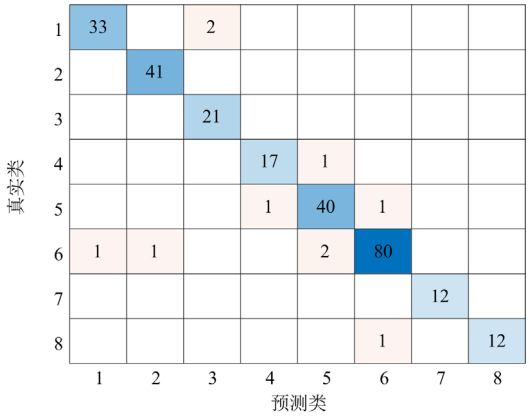

图 6 故障诊断混淆矩阵

Fig. 6 Confusion matrix of fault diagnosis

表 7 IGWO-SVM 模型评价指标

Table 7 Evaluation metrics of IGWO-SVM model

故障类型编码	评价指标				
	R_{ecall}/%	P_{recision}/%	$F_{1\text{-score}}$	识别率/%	Kappa 系数
1	94.3	97.1	0.957		
2	100	97.6	0.988		
3	100	91.3	0.955		
4	94.4	94.4	0.944		
5	95.2	93.0	0.941	96.2	0.954
6	95.2	97.6	0.964		
7	100	100	1.000		
8	92.3	100	0.960		
平均值	96.4	96.4	0.963		

4.2 不同方法性能对比

表 8 展示了采用不同采样方法进行样本扩充后的数据集及未进行样本扩充的原始数据集在 IGWO-SVM 模型下的诊断结果，其中所有算例的输入特征均采用表 1 中未进行特征优选的 22 种候选比值征兆。

如表 8 所示，采用本文方法进行样本扩充后的数据集在 IGWO-SVM 模型的诊断下，其识别率高达 95.1%，Kappa 系数高达 0.940，相比于 K-means-SMOTE、SVM SMOTE、SMOTE 方法及原始数据集，其识别率分别提升了 2.2%、3.8%、5.6%、8.3%，Kappa 系数分别提高了 0.027、0.045、0.068、0.101。

表 8 不同采样方法的诊断结果

Table 8 Diagnosis results of different sample ways

不同采样方法	IGWO-SVM	
	识别率/%	Kappa 系数
原始数据集	86.8	0.839
SMOTE	89.5	0.872
SVM SMOTE	91.3	0.895
K-means-SMOTE	92.9	0.913
本文方法	95.1	0.940

图 7 和图 8 分别展示了基于不同采样方法的 IGWO-SVM 模型在各类样本中的查全率和查准率变化曲线，可以看出基于本文采样方法的模型查全率及查准率在各类样本上的分布区间更稳定，模型对各类样本的识别能力更为均衡。综上所述，使用本文采样方法进行样本扩充能够有效提升少数类样本的识别精度及模型的整体诊断性能。

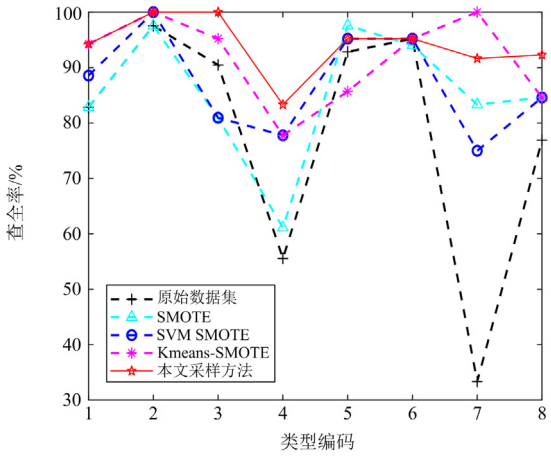

图 7 不同采样方法的查全率曲线

Fig. 7 Recall curves of different sample ways

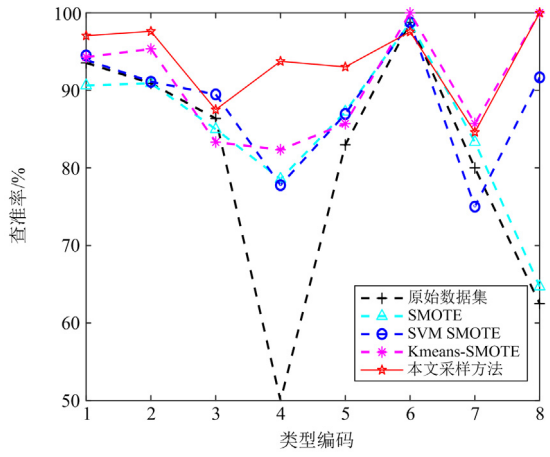

图 8 不同采样方法的查准率曲线

Fig. 8 Precision curves of different sample ways

表 9 展示了采用本文采样方法进行样本扩充的数据集在不同输入特征及不同优化算法(基本参数设置情况相同)下的 SVM 模型诊断结果。其中，关键气体法、Rogers 比值法、IEC 三比值法的输入特征分别为(H_2、CH_4、C_2H_4、C_2H_2、C_2H_6)、(CH_4/H_2、C_2H_6/CH_4、C_2H_4/C_2H_6、C_2H_2/C_2H_4)、(CH_4/H_2、C_2H_4/C_2H_6、C_2H_2/C_2H_4)。

表 9 不同输入特征及不同优化算法下的的诊断结果

Table 9 Diagnosis results of different feature inputs and optimization algorithms

输入特征	GA-SVM		PSO-SVM		GWO-SVM		IGWO-SVM	
	识别率/%	Kappa 系数	识别率/%	Kappa 系数	识别率/%	Kappa 系数	识别率/%	Kappa 系数
关键气体法	39.8	0.146	16.7	0.109	15.4	0.102	43.6	0.218
Rogers 比值法	61.7	0.526	65.4	0.573	65.8	0.577	77.1	0.715
IEC 三比值法	60.5	0.511	62.0	0.533	62.4	0.536	76.7	0.710
本文优选征兆	86.5	0.835	92.5	0.908	94.0	0.927	96.2	0.954

通过纵向对比可以看出，以本文优选征兆为输入特征能够显著提升模型的诊断识别率，且其对应的各模型 Kappa 系数均大于 0.81，说明模型预测结果和实际分类结果几乎完全一致。

通过横向对比可以看出，无论使用哪种组合形式作为输入特征，IGWO-SVM 模型诊断识别率及 Kappa 系数都要高于其他 3 种模型，其中以本文优选征兆为输入特征的 IGWO-SVM 模型的诊断识别率和 Kappa 系数在所有算例中均为最高，相比于同样输入特征的 GWO-SVM、PSO-SVM、GA-SVM 模型，其识别率分别提升了 2.2%、3.7%、9.7%，Kappa 系数分别提高了 0.027、0.046、0.119。

图 9 和图 10 分别展示了基于本文优选征兆为输入特征的 4 种模型在各类样本中的查全率和查准率变化曲线，可以看出 IGWO-SVM 模型除了在高能放电类样本上查准率略低于 GA-SVM 模型，其在各类样本中的查全率和查准率均为最高，且都稳定在 90%~100%范围内。可以证明，IGWO-SVM 模型的诊断性能更优且更稳定，IGWO 相对于 GA、PSO、GWO 而言具有更强的优化能力。

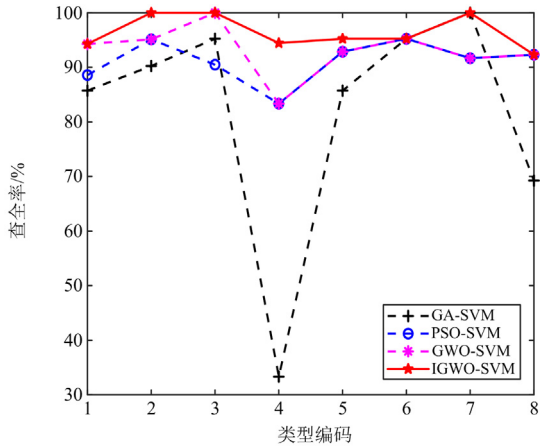

图 9 基于优选征兆的不同模型查全率曲线

Fig. 9 Recall curves of different models based on selective ratios

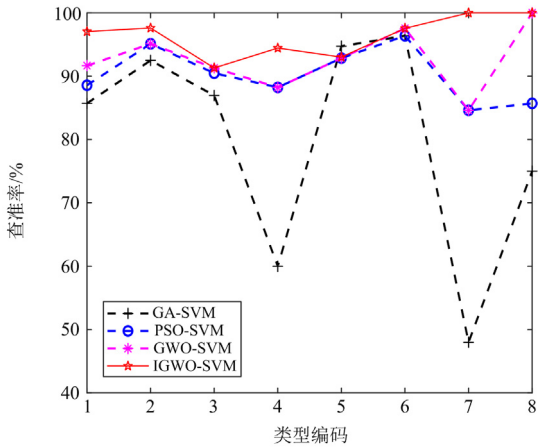

图 10 基于优选征兆的不同模型查准率曲线

Fig. 10 Precision curves of different models based on selective ratios

5 结论与展望

1) 相较于传统过采样算法，本文提出的混合过采样技术能够避免噪声样本生成并增强样本的类内聚合性及类间区分度，从而更有效地合成数据点，进而增强模型对少数类样本的识别能力。

2) 与关键气体法、Rogers 比值法、IEC 三比值法相比，采用 ANOVA 进行特征优选能够降低特征间的冗余性，并增强特征组合与故障类型间的关联性。

3) 相较于其他智能优化算法，本文针对 GWO 的改进策略能够有效平衡算法的全局探索与局部开发能力、避免其陷入早熟。经过 IGWO 优化后的 SVM 模型相比于 GA-SVM、PSO-SVM、GWO-SVM 模型具有更高的诊断识别率和更强的泛化性能。

4) 变压器样本数据不均衡时，本文所提方法能够在一定程度上缓解模型在各类样本间识别效果差距大的问题，而当样本数据极度不平衡时，如何在保证模型泛化性能的基础上提升少数类样本的识别率，需要进一步深入研究。

参考文献

[1] 唐文虎, 牛哲文, 赵柏宁, 等. 数据驱动的人工智能技术在电力设备状态分析中的研究与应用[J]. 高电压技术, 2020, 46(9): 2985-2999.
TANG Wenhu, NIU Zhewen, ZHAO Boning, et al. Research and application of data-driven artificial intelligence technology for condition analysis of power equipment[J]. High Voltage Engineering, 2020, 46(9): 2985-2999.

[2] 张宽, 吐松江·卡日, 高文胜, 等. 基于云模型和改进 D-S 证据理论的变压器故障诊断[J]. 高压电器, 2022, 58(4): 196-204.
ZHANG Kuan, TUSONGJIANG·Kari, GAO Wensheng, et al. Fault diagnosis of transformer based on cloud model and improved D-S evidence theory[J]. High Voltage Apparatus, 2022, 58(4): 196-204.

[3] 詹仲强, 陈文涛, 郝建, 等基于模糊逻辑和 D-S 证据理论的变压器故障诊断方法[J]. 高压电器，2022, 58(11): 160-166.
ZHAN Zhongqiang, CHEN Wentao, HAO Jian, et al. Fault diagnosis method of transformer based on fuzzy logic and D-S evidence theory[J]. High Voltage Apparatus, 2022, 58(11): 160-166.

[4] 石宜金, 谭贵生, 赵波, 等. 基于模糊综合评估模型与信息融合的电力变压器状态评估方法[J]. 电力系统保护与控制, 2022, 50(21): 167-176.
SHI Yijin, TAN Guisheng, ZHAO Bo, et al. Condition assessment method for power transformers based on fuzzy comprehensive evaluation and information fusion[J]. Power System Protection and Control, 2022, 50(21): 167-176.

[5] 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.

[6] 方涛, 钱晔, 郭灿杰, 等. 基于天牛须搜索优化支持向量机的变压器故障诊断研究[J]. 电力系统保护与控制, 2020, 48(20): 90-96.
FANG Tao, QIAN Ye, GUO Canjie, et al. Research on transformer fault diagnosis based on a beetle antennae search optimized support vector machine[J]. Power System Protection and Control, 2020, 48(20): 90-96.

[7] DING C, DING Q, WANG Z, et al. Fault diagnosis of oil-immersed transformers based on the improved sparrow search algorithm optimised support vector machine[J]. IET Electric Power Applications, 2022, 16(9): 985-995.

[8] LI Z, HE Y, XING Z, et al. Transformer fault diagnosis based on improved deep coupled dense convolutional neural network[J]. Electric Power Systems Research, 2022, 209.

[9] 杨威, 蒲彩霞, 杨坤, 等. 基于 CNN-GRU 组合神经网络的变压器短期故障预测方法[J]. 电力系统保护与控制, 2022, 50(6): 107-116.
YANG Wei, PU Caixia, YANG Kun, et al. Short-term fault prediction method for a transformer based on a CNN-GRU combined neural network[J]. Power System Protection and Control, 2022, 50(6): 107-116.

[10] 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.

[11] 刘云鹏, 许自强, 李刚, 等. 人工智能驱动的数据分析技术在电力变压器状态检修中的应用综述[J]. 高电压技术, 2019, 45(2): 337-348.
LIU Yunpeng, XU Ziqiang, LI Gang, et al. Review on applications of artificial intelligence driven data analysis technology in condition based maintenance of power transformers[J]. High Voltage Engineering, 2019, 45(2): 337-348.

[12] 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.

[13] HE H, BAI Y, GARCIA E A, et al. ADASYN: adaptive synthetic sampling approach for imbalanced learning[C] // 2008 IEEE International Joint Conference on Neural Networks, June 1-8, 2008, Hong Kong, China: 1322-1328.

[14] 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.
LIU Yunpeng, HE Jiahui, XU Ziqiang, et al. Equalization method of power transformer fault sample based on SVM SMOTE[J]. High Voltage Engineering, 2020, 46(7): 2522-2529.

[15] DOUZAS G, BACAO F, LAST F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE[J]. Information Sciences, 2018, 465: 1-26.

[16] 叶远波, 李端超, 谢民, 等. 基于 SSA-SVM 的继电保护装置状态评估方法研究[J]. 电力系统保护与控制, 2022, 50(8): 171-178.
YE Yuanbo, LI Duanchao, XIE Min, et al. A state evaluation method for a relay protection device based on SSA-SVM[J]. Power System Protection and Control, 2022, 50(8): 171-178.

[17] ZHANG L, GAO T, CAI G, et al. Research on electric vehicle charging safety warning model based on back

propagation neural network optimized by improved gray wolf algorithm[J]. Journal of Energy Storage, 2022, 49.

[18] 张振海, 王维庆, 王海云, 等. 基于 HCS-GWO-MSVM 的风电机组齿轮箱复合故障诊断研究[J]. 太阳能学报, 2021, 42(10): 176-182.
ZHANG Zhenhai, WANG Weiqing, WANG Haiyun, et al. Research on compound fault diagnosis of wind turbine gearbox based on HCS-GWO-MSVM[J]. Acta Energiae Solaris Sinica, 2021, 42(10): 176-182.

[19] 龙文, 伍铁斌, 唐明珠, 等. 基于透镜成像学习策略的灰狼优化算法[J]. 自动化学报, 2020, 46(10): 2148-2164.
LONG Wen, WU Tiebin, TANG Mingzhu, et al. Grey wolf optimizer algorithm based on lens imaging learning strategy[J]. Acta Automatica Sinica, 2020, 46(10): 2148-2164.

[20] BANAIE-DEZFOULI M, NADIMI-SHAHRAKI M H, BEHESHTI Z, et al. R-GWO: representative-based grey wolf optimizer for solving engineering problems[J]. Applied Soft Computing, 2021, 106: 107328.

[21] ISLAM A, BELHAOUARI S B, REHMAN A U, et al. KNNOR: an oversampling technique for imbalanced datasets[J]. Applied Soft Computing, 2022, 115: 108288.

[22] TANG B, HE H. KernelADASYN: kernel based adaptive synthetic data generation for imbalanced learning[C] // 2015 IEEE Congress on Evolutionary Computation, May 25-28, 2015, Sendai, Japan: 664-671.

[23] MIRJALILI S, MIRJALILI S M, LEWIS A. Grey wolf optimizer[J]. Advances in Engineering Software, 2014, 69: 46-61.

[24] 何庆, 罗仕杭. 混合改进策略的黑猩猩优化算法及其机械应用[J]. 控制与决策, 2023, 38(2): 354-364.
HE Qing, LUO Shihang. Chimp optimization algorithm based on hybrid improvement strategy and its mechanical application[J]. Control and Decision, 2023, 38(2): 354-364.

[25] 刘成汉, 何庆. 融合多策略的黄金正弦黑猩猩优化算法[J/OL]. 自动化学报: 1-14[2023-04-25]. https://doi.org 10.16383/j.aas.c210313.
LIU Chenghan, HE Qing. Golden sine chimp optimization algorithm integrating multiple strategies[J/OL]. Acta Automatica Sinica: 1-14[2023-04-25]. https://doi.org/10.16383/j.aas.c210313.

[26] 刘志强, 何丽, 袁亮, 等. 采用改进灰狼算法的移动机器人路径规划[J]. 西安交通大学学报, 2022, 56(10): 49-60.
LIU Zhiqiang, HE Li, YUAN Liang, et al. Path planning of mobile robot based on TGWO algorithm[J]. Journal of Xi'an Jiaotong University, 2022, 56(10): 49-60.

[27] XUE J, SHEN B. A novel swarm intelligence optimization approach: sparrow search algorithm[J]. Systems Science & Control Engineering, 2020, 8(1): 22-43.

[28] DRAA A, BOUZOUBIA S, BOUKHALFA I. A sinusoidal differential evolution algorithm for numerical optimization[J]. Applied Soft Computing, 2015, 27: 99-126.

[29] 张育杰, 冯健, 李典阳, 等. 基于油色谱数据的变压器故障征兆新优选策略[J]. 电网技术, 2021, 45(8): 3324-3332.
ZHANG Yujie, FENG Jian, LI Dianyang, et al. New feature selection method for transformer fault diagnosis based on DGA data[J]. Power System Technology, 2021, 45(8): 3324-3332.

[30] 吐松江·卡日, 高文胜, 张紫薇, 等. 基于支持向量机和遗传算法的变压器故障诊断[J]. 清华大学学报(自然科学版), 2018, 58(7): 623-629.
TUSONGJIANG Kari, GAO Wensheng, ZHANG Ziwei, et al. Power transformer fault diagnosis based on a support vector machine and a genetic algorithm[J]. Journal of Tsinghua University (Science and Technology), 2018, 58(7): 623-629.

[31] DUVAL M, DEPABLO A. Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases[J]. IEEE Electrical Insulation Magazine, 2001, 17(20): 31-41.

[32] 李恩文. 基于重构聚类分析方法的油浸式变压器故障诊断研究[D]. 武汉: 武汉大学, 2019.
LI Enwen. Research on fault diagnosis of oil-immersed transformer based on reconstructed clustering analysis[D]. Wuhan: Wuhan University, 2019.

[33] 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.
LI Lei, WANG Tingtao, HE Jianan, et al. Transformer fault diagnosis strategy considering parameter optimization of oversampler and classifier[J]. Electric Power Automation Equipment, 2023, 43(1): 209-217.

[34] 卢发兴, 姚鸿鹤, 史浩然. 基于方差分析变量约减的指令制导回路误差分配方法[J]. 系统工程与电子技术, 2020, 42(5): 1131-1138.
LU Faxing, YAO Honghe, SHI Haoran. Error allocation method of instruction guidance loop based on variance analysis variable reduction[J]. Systems Engineering and Electronics, 2020, 42(5): 1131-1138.

收稿日期：2023-03-01； 修回日期：2023-06-27

作者简介：

欧阳鑫(2000—)，男，硕士研究生，研究方向为电力变压器状态监测与故障诊断；E-mail: eathon_oyx@163.com

李志斌(1974—)，男，通信作者，博士，教授，研究方向为测控技术与自动化装置。E-mail: thermal_li@163.com

(编辑 姜新丽)