

DOI:10.19651/j.cnki.emt.2417055

基于实域粗糙集和 NRBO-XGBoost 的变压器故障诊断^{*}

杨勇 胡东 代浩 唐超 谢菊芳

(西南大学工程技术学院 重庆 400715)

摘要: 针对基于油中气体分析(DGA)数据的传统变压器智能诊断模型准确率容易受到输入特征维度以及超参数的选择影响,本研究提出了基于实域粗糙集和 NRBO-XGBoost 相结合的变压器智能故障诊断模型。首先,基于实域粗糙集的概念提出了一种具有自适应性能的特征提取算法用于对变压器初始故障特征进行特征提取;其次针对变压器故障诊断中 XGBoost 受超参数选择影响的缺陷,本研究利用 NRBO 算法高收敛速度和有效避免局部最优的特点对 XGBoost 的超参数进行全局寻优,从而提出 NRBO-XGBoost 模型进行变压器故障诊断;最后通过多组实验对比,相较于其他传统特征,使用本研究所提取特征在多种分类器中的性能都得到了提升,证明了本文所提特征提取算法能有效提取特征中的信息增强模型表现性能。并且 NRBO-XGBoost 在变压器故障诊断领域相较于其他对比模型仅收敛 20 次的同时就达到了 92.09% 的准确率,拥有更优越的表现性能。

关键词: 变压器智能诊断模型;实域粗糙集;超参数寻优

中图分类号: TM41;TN06 **文献标识码:** A **国家标准学科分类代码:** 416.50

Research on transformer fault diagnosis method based on real domain rough set and NRBO-XGBoost

Yang Yong Hu Dong Dai Hao Tang Chao Xie Jufang

(School of Engineering and Technology, Southwest University, Chongqing 400715, China)

Abstract: In response to the fact that the accuracy of traditional transformer intelligent diagnosis models based on dissolved gas analysis (DGA) data is easily affected by the selection of input feature dimensions and hyperparameters, this study proposes a transformer intelligent fault diagnosis model based on the combination of real domain rough sets and NRBO-XGBoost. Firstly, a feature extraction algorithm with adaptive performance is proposed based on the concept of real domain rough set for extracting initial fault features of transformers; secondly, in response to the limitation of XGBoost being affected by hyperparameter selection in transformer fault diagnosis, this study utilizes the high convergence speed and effective avoidance of local optima of the NRBO algorithm to globally optimize the hyperparameters of XGBoost, and proposes the NRBO-XGBoost model for transformer fault diagnosis; finally, through multiple experimental comparisons, compared with other traditional features, the performance of the feature extraction algorithm proposed in this study has been improved in various classifiers, proving that the feature extraction algorithm proposed in this paper can effectively extract information from the features to enhance the performance of the model. Moreover, NRBO-XGBoost achieves an accuracy of 92.09% in transformer fault diagnosis with only 20 convergences compared to other comparative models, demonstrating superior performance.

Keywords: transformer intelligent diagnosis model; real domain rough set; hyperparameter optimization

0 引言

电力变压器作为电力系统中的核心设备,其正常运行

对于整个电力系统的稳定至关重要,所以在线监测和分析变压器运行状态并保障其稳定运行对电力系统的可靠性具有重大意义^[1]。

收稿日期:2024-10-09

^{*} 基金项目:国家自然科学基金(51977179)项目资助

目前在电网中,由于大多数运行中的变压器是油浸式的,变压器油中溶解气体分析(dissolved gas analysis, DGA)已成为有效的实时监测变压器运行状态的方法^[2]。DGA 通过检测变压器油中溶解的特定气体,包括甲烷(CH_4)、乙炔(C_2H_2)、乙烯(C_2H_4)、乙烷(C_2H_6)、氢气(H_2)、一氧化碳(CO)和二氧化碳(CO_2)等 7 种碳氢化合物,来评估变压器的健康状况,分析可能存在的潜在故障^[3]。一些学者在深入研究油中气体分析原理的基础上,提出了一系列科学有效的方法,诸如 IEC 三比值法^[4]、Duval 三角形法^[5]以及 Roger's 方法^[6]等。但这些传统的变压器故障诊断方法都有其缺陷和条件,例如:IEC 三比值法,因为存在编码缺损和临界值判据缺损问题,容易导致误判漏判^[7];Duval 三角形法,必须要求关键气体浓度超过允许限值^[8]。近年来,随着电力系统和人工智能的发展,众多学者提出了不同的变压器故障诊断模型:比如有学者基于混合核极限学习提出了变压器故障诊断模型^[9],引入核函数到极限学习机中,加强了极限学习机的学习速度和泛化能力。文献[10]使用长短期记忆网络通过对变压器时序数据 DGA 来预测变压器未来运行状态。文献[11]使用概率神经网络进行变压器故障诊断,利用卷积层强大的特征提取能力在变压器故障诊断中表现优异。XGBoost(extreme gradient boosting)是一种优化的分布式梯度增强库,旨在实现高效、灵活且便携的机器学习,在故障检测问题上表现良好^[12],并且对于标签不平衡数据表现更好,非常适配于变压器 DGA 数据。文献[13]通过 XGBoost 创建了一种双层的机器学习模型分别用来检测变压器是否故障和故障类型。文献[14]指出 XGBoost 的分类效果非常容易受输入特征的维度和超参数的影响。因此本文分别从特征提取和超参数寻优方面对 XGBoost 做出改进。

特征提取方面:特征选取是机器学习中非常重要的步骤,它可以提取特征中有效信息,降低模型复杂程度,增强模型表现性能。特征选取根据不同的方法大致可以分为 3 类:嵌入式特征选择,包裹式特征选择和过滤式特征选取^[15]。其中,嵌入式特征选择和包裹式选择的的思想都是将特征选择过程与机器学习过程相结合的技术,例如文献[16]采用无监督 SVM 的方式给数据打上标签,再使用 SVM 的分类效果进行特征选择,这种使用分类器的分类效果作为特征选择依据的方法其泛化性能不强。过滤式是指通过不同数据处理方法在数据层面上进行特征筛选,具有较强的通用性和较高的计算效率。在过滤式特征选择上:文献[17]采用核主成分分析法(kernel principal component analysis, KPCA)对变压器故障特征进行提取,该方法利用核函数来对原本的主成分分析法(principal component analysis, PCA)进行升维,从而避免故障特征集在向量空间中线性不可分的特点。但是核函数的选择需要经验,且不同的核函数对提取的效果也有不同;文献[18]采用邻域粗糙集对特征集合进行属性约简,其核心思想就是

针对粗糙集只能应用于离散系统这一特点,通过邻域这一数学概念来对数据进行离散化。但离散化处理会改变数据的原始特征,且算法的截至是人为主观决定的,无法保证能提取所有的重要特征;胡寿松教授提出的实域粗糙集^[19]不同于邻域粗糙集仅仅将数据离散化的思想,实域粗糙集通过属性特征定义属性的广义重要度,以重新衡量样本空间中的相似距离。基于 DGA 数据的变压器故障诊断也是通过标签对连续数据进行分类学习的过程,所以通过实域粗糙集可以有效衡量 DGA 数据中每一个特征属性对于决策属性的分类能力大小,从而达到提取特征有效信息,进行特征约简的目的。

超参数寻优方面:对于分类器效果容易受到输入超参数的影响,一些智能算法被运用到了参数的寻优中。例如麻雀搜索算法(sparrow search algorithm, SSA)^[20],粒子群优化算法(particle swarm optimization, PSO)^[21],蜣螂优化算法(dung beetle optimizer, DBO)^[22]等,这些元启发式算法具有稳定性强且易于实现的特点。但是,元启发式算法也存在一定缺陷,例如粒子群算法容易陷入局部最优等。NRBO(newton-raphson-based optimizer, NRBO)是一种新颖牛顿拉夫逊算法的优化算法^[23],通过模拟自然界中资源的分配和利用来优化问题,他对寻优算法的收敛速度和容易陷入局部最优的两个问题针对性的提出了牛顿拉夫逊搜索规则(Newton-Raphson search rule, NRSR)和陷阱避免算子(trap avoidance operator, TAO),具有收敛速度快和不容易陷入局部最优的优点,适用于 XGBoost 中多个超参数寻优这种多维度解问题。

针对上诉两个方面的不足:在特征提取方面,本文提出了一种基于实域粗糙集的变压器故障特征过滤式特征提取算法,该算法不被主观经验参数限制,能较好的提取变压器故障特征气体集合中相对重要的特征,并且通过真实故障数据实验表明使用该算法提取后的特征能够有效提高模型诊断的准确率;在超参数寻优方面,本文使用牛顿拉夫逊优化算法优化 XGBoost 中多个超参数的值,从而提出牛顿拉夫逊优化 XGBoost(NRBO-XGBoost)的变压器故障诊断方法。通过与粒子群算法优化 XGBoost(PSO-XGBoost)、麻雀搜索算法优化 XGBoost(SSA-XGBoost)、蜣螂优化算法优化 XGBoost(DBO-XGBoost)等模型进行对比实验,证明所提出模型具有收敛速度快特点的同时,可以跳出局部最优解;此外,将 NRBO-XGBoost 与 SVM、ELM 和 lightGBM 等常用分类模型进行变压器故障诊断对比实验,结果验证本文所提模型具有更高的评估指标和更好的性能。

1 实域粗糙集理论

变压器 DGA 故障诊断的本质是对特征集进行学习和分类的过程,所以可以将整个过程看作一个信息或决策系统,由有序四元组 $S = \{U, C, V, f\}$ 表示。即对于变压器 DGA 故障数据集, $U = \{X_1, X_2, \dots, X_n\}$ 为论域也就是数

据中的所有样本。 $C = \{a_1, a_2, \dots, a_n\}$ 为数据中气体特征集合也叫条件属性, 其中 a_n 为气体特征集合中的第 n 个特征。 $V = \{d_1, d_2, \dots, d_{r(d)}\}$ 为标签属性即故障类型。 f 为一个信息函数, 用来确定 U 中每一个对象 X_n 对应条件属性下 a_n 的取值, 记作 $f(X_n, a_n)$ 。对于变压器故障诊断实域粗糙集模型, 有以下定义:

1) 对于变压器故障诊断信息/决策系统, $S = \{U, C, V, f\}$, 定义特征气体属性 a 的广义重要度:

$$\sigma_g(a) = \begin{cases} 1 - \frac{1}{C_{r(d)}^2} \sum_{i \neq j, i, j=1}^{r(d)} \frac{a(d_i) \cap a(d_j)}{\maxcross(a(d))} \\ 1, \forall i, j, a(d_i) \cap a(d_j) = \emptyset \end{cases} \quad (1)$$

其中, $C_{r(d)}^2$ 表示总共 $r(d)$ 数量故障特征中取 2 的组合, $a(d_i) \cap a(d_j)$ 表示特征气体属性 a 对应的故障特征 d_i 的属性子集与对应故障特征 d_j 的属性子集的交集部分。 $\maxcross(a(d))$ 表示特征气体属性 a 对应每两个故障特征的属性子集的所有交集所包围的最大区间。

2) 对实域决策系统 $S = \{U, C, V, f\}$, 实域知识空间 $B \subseteq C$ 。对任意的 $x, y \in U$ 和 $a_i \in B$, x 和 y 在实域空间上的广式重要度欧氏 $d_B(x, y)$ 距离为:

$$d_B(x, y) = \sqrt{\sum_{i=1}^{|B|} \sigma(a_i) (f(x, a_i) - f(y, a_i))^2} \quad (2)$$

3) 对实域决策系统 $S = \{U, C, V, f\}$, 实域知识空间 $B \subseteq C$, 阈值 $\delta \geq 0$, 对任意 $x \in U$, x 在实域空间中 B 的广义近邻 $[x]_B^\delta$ 定义如下:

$$[x]_B^\delta = \{y \in U \mid d_B(x, y) \leq \delta\} \quad (3)$$

由式(3)可以看出, 在广义近邻中, 邻域半径的大小对邻域信息粒的规模起着决定性作用, 进而影响实域决策信息系统的邻域粒化结果。当邻域半径较大时, 可能会增加样本误分量, 导致类的欠划分; 而当邻域半径趋近于 0 时, 样本邻域信息粒中的样本元素过少, 又可能导致类的过度划分。而且对于上述变压器决策系统多属性, 非线性特点, 采用单一邻域会造成较大误差, 因此参考文献[24], 本文采用多半径邻域的方法, 具体来说, 对于实域知识空间 $B \subseteq C$, $\delta(B)$ 定义如下:

$$\delta(B) = \frac{\sum_{i=1}^B (std(a_i) \div \lambda)}{N} \quad (4)$$

其中, N 表示知识空间 B 中条件属性的数量, $std(a_i)$ 表示属性 a_i 的标准差, λ 表示特征参数, 本文取 2。

4) 对实域决策系统 $S = \{U, C, V, f\}$, 实域知识空间 $B \subseteq C$, 对任意 $X \subseteq U$, X 在实域空间 B 上的 δ 下近似 $\underline{B}^\delta(X)$ 和 δ 上近似 $\bar{B}^\delta(X)$ 定义如下:

$$\underline{B}^\delta(X) = \{x \in U \mid [x]_B^\delta \subseteq X\} \quad (5)$$

$$\bar{B}^\delta(X) = \{x \in U \mid [x]_B^\delta \cap X \neq \emptyset\} \quad (6)$$

5) 对实域决策系统 $S = \{U, C, V, f\}$, 实域知识空间 $B \subseteq C$, B 对 d 的近似分类质量 $\gamma_B^\delta(\{d\})$ 为:

$$\gamma_B^\delta(\{d\}) = \frac{|\bigcup_{X_i \in U/d} B^\delta(X_i)|}{|U|} \quad (7)$$

6) 对实域决策系统 $S = \{U, C, V, f\}$, 实域知识空间 $B \subseteq C$, $a \in C$, 定义 a 相特征属性在实域空间 B 中对于决策属性 d 的分类能力大小 $ability(\{a\}, B, \{d\})$ 为:

$$ability(\{a\}, B, \{d\}) = 1 - \frac{|\bigcup_{X_i \in U/d} B^\delta(X_i)|}{|\bigcup_{X_i \in U/d} B \cup \{a\}^\delta(X_i)|} \quad (8)$$

上述 $ability(\{a\}, B, \{d\})$ 定义表明在变压器 DGA 故障数据中, 故障特征气体属性 a 相较于实域知识空间 B 对故障类型 d 的分类能力大小, 其值越大, 表明其对于故障结果影响越大, 即出现在约简后特征集的概率越高。

2 NRBO-XGBoost

XGBoost 是基于提升决策树 (gradient boosting-decision tree, GBDT) 的树状集成算法, 以 CART 树为基础, 通过改进每一个弱分类器的效果搜索最佳分类模型^[25]。在 XGBoost 中, 假定输入样本为 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n]^\top$, 其中 \mathbf{x}_i 表示每一个弱评估器的叶子节点的权重相加之和, $\mathbf{x}_i = [x_{i,1}, x_{i,2}, \dots, x_{i,m}]$, 样本编码为 $\mathbf{Y} = [y_1, y_2, \dots, y_n]$, 其中 n 为样本总数, m 的大小为特征维度, 假设一共有 M 个弱评估器, 则评估结果可以表示为:

$$\hat{y}_i = \sum_{k=1}^M f_k(x_i), \quad f_k \in P \quad (9)$$

其中, \hat{y}_i 表示第 i 个样本的诊断结果, f_k 为第 k 个评估器, P 为所有评估器的集合。模型的目标函数由损失函数和正则化系数构成, 可以表示为:

$$Obj = \underbrace{\sum_{i=1}^n L(y_i, \hat{y}_i)}_{\text{Loss Function}} + \underbrace{\sum_{k=1}^M \Omega(f_k)}_{\text{Regularization}} \quad (10)$$

$$\Omega(f) = \gamma T + \|\omega\|^2 / 2\lambda \quad (11)$$

其中, $\Omega(f)$ 表示正则项惩罚系数, 用来防止模型过拟合, T 为叶子节点的总数, γ 和 λ 是手动调整的系数。损失函数在本文中选择交叉熵函数^[26], 交叉熵函数使用概率解释并且可以有效避免过饱和问题, 适用于本文中的多分类模型。对于多分类模型的激活函数, 本文选择 Softmax 函数^[27], Softmax 将一组数值转换成概率分布, 适用于多分类任务的概率解释。

牛顿拉夫逊优化算法 (NRBO) 是一个新的元启发式算法, 它使用两个规则来探索整个搜索过程, 牛顿拉夫逊搜索规则 (NRSR) 和陷阱避免算子 (TAO), 并通过几组矩阵进一步探索最佳搜索结果, 具有高收敛速度和有效避免局部最优的优点。算法大致的更新规则如下:

$$x_{n+1} = x_n - NRSR \quad (12)$$

其中, x_{n+1} 表示下一次位置, NRSR 表示搜索规则, 其

中有两步核心思想,第一步倾向于保留更优的适应度解,即 $X_b = x_n - \Delta x$, X_b 表示 x_n 邻域周边更好的位置;第二步思想是保留部分劣质解,增加种群多样性,以避免陷入局部最优,即 $X_w = x_n + \Delta x$, X_w 此时就为 x_n 邻域周边更差的位置。整个规则的计算公式如下:

$$NRSR = randn \times \frac{(X_w - X_b) \times \Delta x}{2 \times (X_w + X_b - 2 \times x_n)} \quad (13)$$

其中, $randn$ 表示具有平均值 0 和方差 1 的正态分布随机数 dim 表示维度数,在本文中表示超参数的个数, $\Delta x = rand(1, dim) \times |X_b - X_n^{IT}|$, 表示扰动量, X_n^{IT} 表示迭代次数为 IT 次当前的位置。除了更新规则,算法还设计了陷阱避免算子来跳出局部最优。算法被验证在较高的探索能力和有效避免局部最优能力,这表明算法特点非常适配于本文的 XGBoost 超参数寻优。

选择表 1 中的 5 个 XGboost 模型的主要参数进行优化,并将模型诊断的错误率 E_{error} 作为寻优算法的迭代适应度函数,其中 $A_{accuracy}$ 为表示模型诊断的准确率。

$$E_{error} = 1 - A_{accuracy} \quad (14)$$

表 1 XGBoost 主要超参数

Table 1 Main hyperparameters of XGBoost		
超参数名称	参数意义	寻优范围
max leaves	树最大叶子节点数	[1,50]
learning_rate	学习率	[0.01,0.2]
subsample	树数据样本比例	[0.01,1]
colsamplebytree	构建树特征子集比例	[0.01,0.2]
max_depth	单棵树的最大深度	[1,15]

3 基于实域粗糙集和 NRBO-XGBoost 的变压器故障诊断

基于实域粗糙集和 NRBO-XGBoost 的变压器故障诊断整体流程如图 1 所示,步骤如下:

- 1)按照上述准则构造变压器故障诊断的初始样本集合。
- 2)对初始样本集合基于实域粗糙集进行特征提取,输出提取后的特征。
- 3)NRBO-XGBoost 对输入提取后的特征进行参数寻优和故障诊断,最终输出模型最优超参数和分类结果。

3.1 构建初始故障集合

本文采用某电力局变压器历史监测数据为研究对象,结合 IEC TC 10 标准和文献[28-31]所提供的变压器故障数据,共 1 181 组故障样本。根据 IEC TC 10^[32] 内容,变压器故障类型被划分为 6 种,分别为局部放电(PD),低能放电状态(LED),高能放电状态(HED),中低温过热状态(LMT),高温过热状态(HT)和正常状态(N)。按照 1 : 3 将样本划分为训练集和测试集,样本分布如表 2 所示。

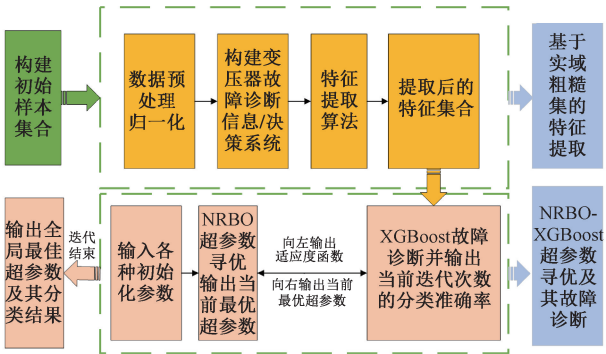


图 1 算法整体流程

Fig. 1 Overall algorithm process

表 2 故障样本分布情况

Table 2 Distribution of fault samples

故障类型	故障样本分布	
	训练集	测试集
LED	89	32
N	91	24
HED	221	72
HT	253	78
LMT	267	76
PD	259	72

变压器故障类型复杂,采取引言中所提到的单一烃类特征气体作为故障样本难以满足诊断的准确性要求。文献[33]表明 DGA 特征气体中的两种气体的相对比值与变压器运行状态更为相关。因此参考文献[34]特征气体比值选取并遵循相关比值法构造故障特征规则,选取了以下特征气体及其比值作为初始故障特征气体集合,如表 3 所示,其中 TH 为总烃。

表 3 初始故障特征气体集合

Table 3 Initial fault characteristics gas set

编号	特征气体	编号	特征气体
1	H ₂	11	H ₂ /C ₂ H ₆
2	CH ₄	12	H ₂ /C ₂ H ₄
3	C ₂ H ₆	13	C ₂ H ₂ /C ₂ H ₆
4	C ₂ H ₄	14	C ₂ H ₂ /C ₂ H ₄
5	C ₂ H ₂	15	C ₂ H ₂ /CH ₄
6	TH	16	H ₂ /(H ₂ +TH)
7	CH ₄ /H ₂	17	(CH ₄ +C ₂ H ₄)/TH
8	CH ₄ /C ₂ H ₆	18	C ₂ H ₄ /TH
9	CH ₄ /TH	19	C ₂ H ₄ /C ₂ H ₆
10	CH ₄ /C ₂ H ₄	20	C ₂ H ₆ /TH

3.2 基于实域粗糙集的特征提取

本文采用上述实域粗糙集的概率和定义,结合变压器故障类型与特征气体间的深层联系,引入具有自适应性能

的特征属性约简算法,算法步骤如下:

1) 以 1 181 组样本构建变压器故障诊断信息/决策系统 $S = \{U, C, V, f\}$, 其中 $C = \{a_1, a_2, \dots, a_n\}$ 条件属性由表 3 构成, $V = \{d_1, d_2, \dots, d_{r(d)}\}$ 决策属性由表 1 所示的故障类型构成。输入信息/决策系统 $S = \{U, C, V, f\}$, 输入特征参数 $\lambda = 2$, 全体样本归一化。

2) 初始化约简集合 $R = \emptyset$, $\bigcup_{X_i \in U/d} R^\delta(X_i) = 1, Y = C$, $\gamma_R^\delta(\{d\}) = 0$ 。

3) 对任意 $x \in U$, 利用式(4)计算 C 上的 $\delta(C)$, 根据式(3)计算 C 上的广义近邻 $[x]_C^\delta$ 。

4) 根据式(7)计算 S 决策系统中决策 d 相对条件属性集 C 的近似分类质量 $\gamma_C^\delta(\{d\})$ 。

5) 依据式(8)计算 Y 中每一个属性 a 的 $ability(\{a\}, R, \{d\})$, 选 $ability(\{a\}, R, \{d\})$ 最大的 a , 更新 $R = R \cup \{a\}, Y = Y \setminus \{a\}$ 。

6) 计算更新后的 $\gamma_R^\delta(\{d\})$, 判断 $\gamma_R^\delta(\{d\}) < \gamma_C^\delta(\{d\})$ 是否成立, 若成立就跳回步骤 5), 继续执行上述步骤; 若不成立立即跳出循环, 输出约简后的集合 R 。

3.3 NRBO-XGBoost 超参数优化及故障诊断算法流程

NRBO-XGBoost 以上述 E_{error} 为寻优算法每一次迭代时的适应度函数, 对特征提取之后的特征集合划分训练集和测试集进行训练验证。每一次迭代时故障诊断之后进行交叉验证, 整个模型的训练过程如图 2 所示, 其中 n 为模型迭代次数, 本文设置为 50。

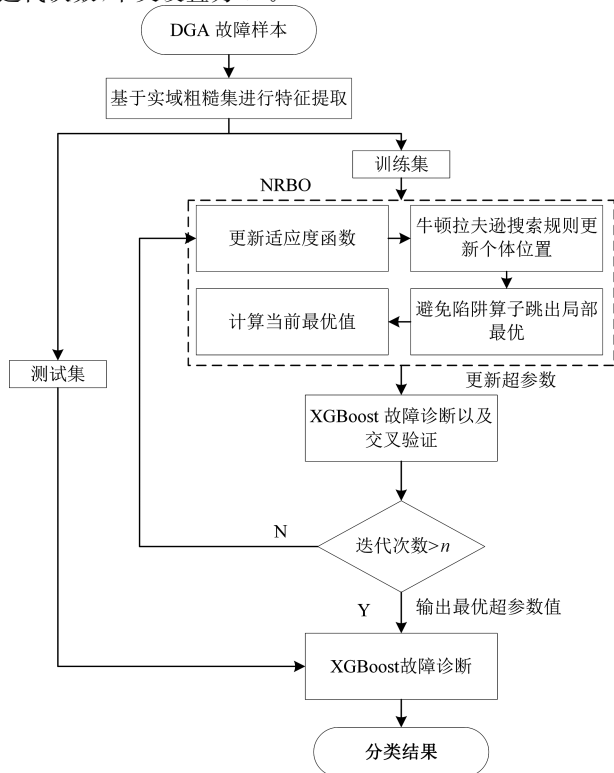


图 2 NRBO-XGBoost 训练流程

Fig. 2 NRBO-XGBoost training process

4 实例计算

4.1 评估指标

本文模型为复杂多分类模型, 所以通过准确率和 Kappa 系数从整体上评价分类器的分类效果, 且通过精确率, 召回率, F1 Score 来综合衡量评价器在各类故障中的诊断精度和诊断灵敏度。设分类器在每一类别中诊断结果为: 真阳性 (TP, 正确地将正类样本预测为正类), 假阳性 (FP, 错误地将负类样本预测为正类), 真阴性 (TN, 正确地将负类样本预测为负类), 假阴性 (FN, 错误地将正类样本预测为负类)。则式(13)~(17)为上述指标的表达式:

$$A_{accuracy} = \frac{\sum_{i=1}^K (TP_i + TN_i)}{n} \quad (15)$$

$$p_e = \frac{a_1 \times b_2 + a_2 \times b_2 + \dots + a_K \times b_K}{n \times n} \quad (16)$$

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

$$Precision = \frac{\sum_{i=1}^K TP_i}{\sum_{i=1}^K (TP_i + FP_i)} \quad (17)$$

$$Recall = \frac{\sum_{i=1}^K TP_i}{\sum_{i=1}^K (TP_i + FN_i)} \quad (18)$$

$$F1 \text{ Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (19)$$

其中, n 为样本总数, K 为样本类别数, a_1, \dots, a_K 为每个类别真实的样本数, b_1, \dots, b_K 为每个类别诊断结果样本数, p_0 等同于准确率 $A_{accuracy}$ 。

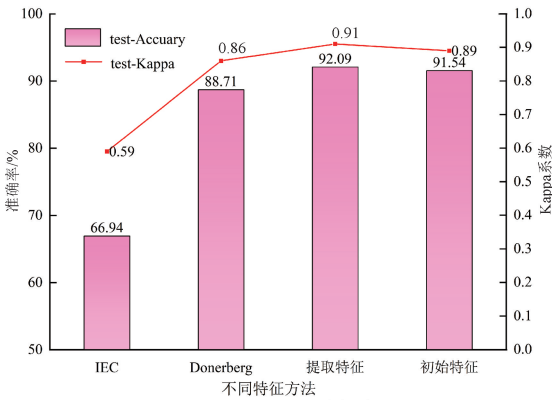
4.2 基于实域粗糙集的特征选取结果分析

以上述 1 181 组样本构建变压器故障诊断信息/决策系统, 基于实域粗糙集进行特征提取, 获得属性约简后的特征集合以及其每一个特征属性相对分类能力大小。如表 4 所示。由表 4 可知, 验证了相较于单一特征气体, 特征气体之间的比值与变压器运行状态更加紧密这一结论。从特征提取的结果来看, H_2/C_2H_4 , C_2H_4/C_2H_6 , C_2H_2/C_2H_4 这 3 个属性的相对分类能力最大, 即 IEC 导则推荐的三比值特征量, 符合传统的故障特征量选取导则。

为了验证本文算法所提取特征的有效性和合理性, 实现特征的综合性评价。以表 2 所展示的数据为基础, 将 IEC 推荐三比值特征, Donerberg 特征^[35], 初始特征与本文所提取特征分别作为 SVM, ELM, lightGBM 几个常用机器学习方法和本文所提模型 NRBO-XGBoost 的输入进行变压器故障诊断。其中图 3(a)~(d) 分别显示了本文所提取的特征和其他不同特征作为 SVM, ELM, lightGBM 和 NRBO-XGBoost 等模型的输入, 各个模型在测试集中达到的准确率和 Kappa 系数。

表 4 提取后的特征集合
Table 4 Extracted feature set

特征气体	分类能力	特征气体	分类能力
C_2H_2/C_2H_4	0.983	H_2/C_2H_6	0.935
C_2H_4/C_2H_6	0.983	CH_4/C_2H_6	0.932
H_2/C_2H_4	0.983	C_2H_4	0.887
$(CH_4+C_2H_4)/TH$	0.983	C_2H_6	0.887
C_2H_2/CH_4	0.971	CH_4	0.881
CH_4/C_2H_4	0.961	H_2	0.871
C_2H_2/C_2H_6	0.961	$H_2/(H_2+TH)$	0.861



(d) NRBO-XGBoost诊断结果
(d) NRBO-XGBoost diagnosis results

图 3 不同特征集合诊断准确率

Fig. 3 Diagnostic accuracy of different feature sets

由图 3 可知,实域粗糙集提取后的特征集合,从上图中可以看出,本文基于实域粗糙集提取的特征在 SVM、极限学习机(extreme learning machines, ELM)、轻量级梯度提升机器(light gradient boosting machine, lightGBM)、NRBO-XGBoost 中的表现都要好于三比值特征, Donerberg 特征,这说明本文所提取特征相较于传统特征在分类器中的表现性能更具有优势;在 4 个分类器中本文所提取的特征表现性能都优于初始特征,也验证了本文基于实域粗糙集的特征提取算法能够有效的提取特征中的信息减少冗余度,并且具有良好的泛化性能。

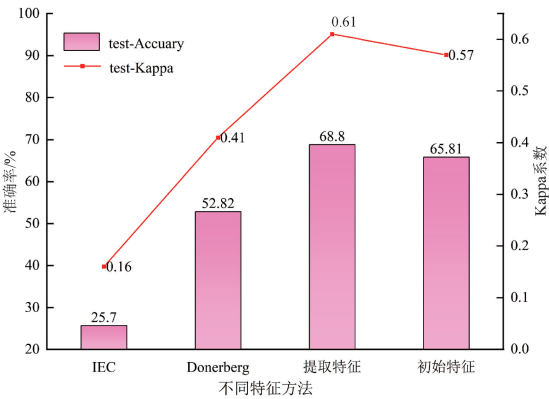
4.3 不同寻优算法比较

以本文表 4 提取后的特征集合为输入特征,分别采用不同的寻优算法例如 SSA, PSO, DBO, 对 XGBoost 进行超参数寻优从而进行变压器故障诊断。将 PSO-XGBoost, SSA-XGBoost, DBO-XGBoost 三种算法和本文中 NRBO-XGBoost 进行比较,每个算法的参数设置和 XGBoost 寻优到最佳超参数的值分别如表 5 和 6 所示,其中, V_{max} 为最大粒子速度, c_1 、 c_2 、 ω_1 和 ω_2 分别为相应的系数。 pop 为种群总数, lb 为最大种群上界, ub 为最小种群下界(每一个超参数的种群范围已在表 1 中给出), T_{max} 为最大迭代次数, dim 为种群维度,在本文也就是对应 XGBoost 的 5 个超参数。

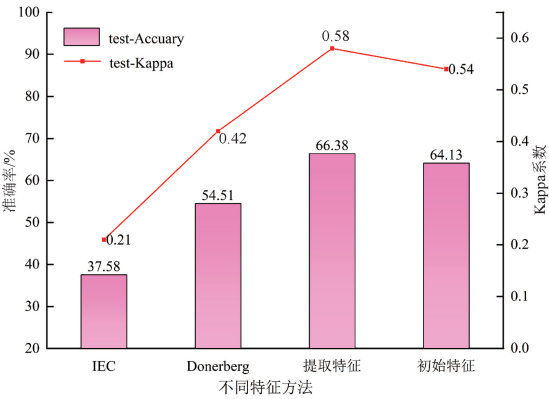
表 5 不同算法参数设置

Table 5 Different algorithm parameter settings

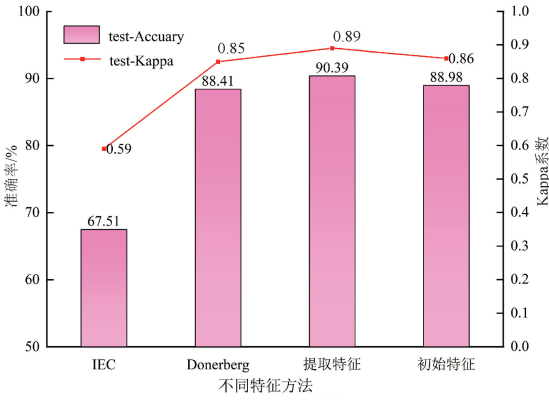
模型	PSO-XGBoost	SSA-XGBoost	DBO-XGBoost	NRBO-XGBoost
V_{max}	1 200	$pop=10$	$pop=10$	$pop=10$
c_1	2	$lb=50$	$lb=50$	$lb=50$
c_2	2	$ub=0.01$	$ub=0.01$	$ub=0.01$
ω_1	0.2	$T_{max}=50$	$T_{max}=50$	$T_{max}=50$
ω_2	0.4	$dim=5$	$dim=5$	$dim=5$



(a) SVM诊断结果
(a) SVM diagnosis results



(b) ELM诊断结果
(b) ELM diagnosis results



(c) LightGBM诊断结果
(c) LightGBM diagnosis results

表 6 不同算法最优超参数值

Table 6 Optimal hyperparameter values				
超参数	NRBO-XGBoost	DBO-XGBoost	SSA-XGBoost	PSO-XGBoost
max leaves	40	24	18	15
learning_rate	0.2	0.2	0.2	0.2
subsample	0.93	1	1	1
max_depth	0.2	0	0	0
Colsamplebytree	6	14	15	15

全体寻优算法都以上述模型在训练集中的 E_{error} 准确率差值作为适应度函数。通过比较结果做出寻优过程中平均适应度变化曲线如图 4。可以发现其中 NRBO 对于 XGBoost 的超参数寻优效果最好,在迭代到 20 次时准确率差值 E_{error} 就达到了 0.003 63。而 PSO-XGBoost 和 SSA-XGBoost 一直陷入了局部最优无法跳出,DBO-XGBoost 在第 5 次迭代时也陷入了局部最优,这表明 NRBO-XGBoost 相较于其他寻优算法具有更强的跳出局部最优能力;并且只需要 20 次就达到了最优,这也表明 NRBO-XGBoost 相较于其他优化算法具有更高的收敛精度和更快的收敛速度。

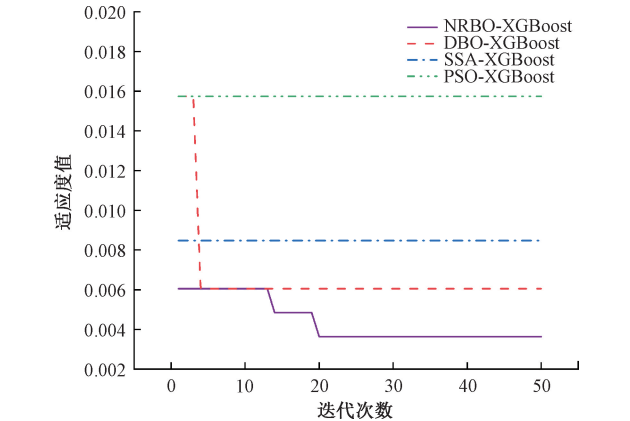


图 4 不同模型适应度迭代曲线

4.4 NRBO-XGBoost 诊断结果

根据表 6 得到的最优超参数值来设置 XGBoost 的超参数,NRBO-XGBoost 模型在测试集上的准确率、Kappa 系数、召回率、F1 Score、精确率,评估指标结果如表 7 所示。同时,模型对于不同故障类型的诊断结果也在图 5 的模型混淆矩阵中给出。可以看出,本文提出的模型在测试集上表现良好,在每一个指标上都取得了较高的得分,且对于每一个故障类型的诊断结果相对平衡。

4.5 不同诊断模型对比

为了验证本文所提模型的优越性,将 SVM,ELM,lightGBM 和 NRBO-XGBoost 基于相同的数据和特征输入进行故障诊断来对比测试集中诊断结果的召回率,F1

Score,精确率等参数指标。其中所有模型都以表 2 所划分的训练集和测试集进行训练和测试,都以本文所提算法提取后的特征作为模型输入。表 8 展示了各个模型在测试集中对于每一类故障类型中的诊断准确率对比。

表 7 NRBO XGBoost 诊断评估指标

Table 7 NRBO-XGBoost evaluation indicators	
评估指标	值
Accuracy	0.921
Kappa	0.910
Recall	0.925
F1 Score	0.925
Precision	0.927

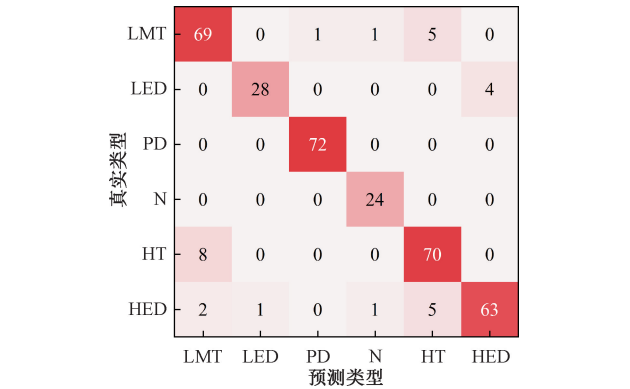


图 5 NRBO-XGBoost 混淆矩阵

表 8 不同模型对不同故障类型的准确率对比

Table 8 Comparison of accuracy of different models					%
故障类型	SVM	ELM	lightGBM	NRBO-XGBoost	
LMT	86.84	65.78	89.47	90.78	
LED	0.00	40.62	84.37	87.50	
PD	98.61	94.44	1.00	1.00	
N	0.00	0.00	1.00	1.00	
HT	62.85	60.25	88.46	89.74	
HED	65.27	69.44	83.33	87.5	

每个模型在测试集中的诊断评估指标的比较如图 6 所示,其中每个顶点表示一个参数指标,虚线边界表示指标大小。

结合表 8 和图 6,可以发现在准确率方面,NRBO-XGBoost 在每种类型中的诊断准确率都高于比较模型;在评价指标方面,可以看到 NRBO-XGBoost 绘制的图形“包裹”了其他模型对应的图形。验证了本文提出的模型的在变压器故障诊断中,表现优于 SVM、ELM 和 lightGBM 等常用模型。这表明本文提出的 NRBO-XGBoost 多分类模

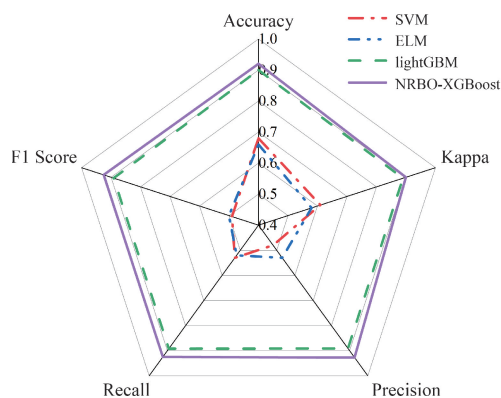


图 6 不同模型准确率分布图

Fig. 6 Accuracy distribution of different models

型对各种类型的不平衡数据仍具有良好的诊断性能。

5 结 论

针对变压器油中气体分析(DGA)的智能诊断模型准确率容易受输入特征维度和质量以及输入超参数影响,本文提出了基于实域粗糙集的自适应特征提取算法与NRBO-XGBoost的变压器故障诊断方法,具体总结如下:

针对 XGBoost 在变压器故障诊断领域中容易受到输入特征质量和维度影响而分类效果和泛化性能不佳问题:在 DGA 数据基础上构建变压器原始故障集合,以实域粗糙集概念构造了变压器故障诊断信息/决策系统,提出具有自适应的过滤式特征提取算法对重要特征进行提取。实验证明:当输入特征为本文所提取特征时,在 SVM、ELM、lightGBM 和本文所提模型 NRBO-XGBoost 测试集中的准确率分别为 68.8%、63.38%、90.39% 和 92.09%,全部高于 IEC 推荐三比值特征和 Donerberg 特征,证明了本文所提取特征的优越性;并且本文所提取特征在上述分类模型中准确率和 Kappa 系数也都高于本文所构建的初始特征,证明了本文基于实域粗糙集的特征提取算法能提取特征中的有效信息,增强模型的表现性能。

针对 XGBoost 在变压器故障诊断中受超参数参数影响的缺陷:引入牛顿拉夫逊寻优算法对 XGBoost 超参数进行寻优,通过与其他寻优算法对比寻优适应度变化曲线说明,牛顿拉夫逊寻优算法相较于其他模型在变压器故障诊断领域具有更快的收敛速度和更好的全局寻优能力。当输入特征为本文特征时,将 NRBO-XGBoost 与 SVM、ELM、lightGBM 分类模型进行变压器故障诊断对比实验,结果表明所提模型不仅在各个故障类型中的准确率都高于对比模型,而且在评估指标的雷达分布图中也优于其他模型,证明了所提模型在变压器故障诊断领域中的优越性。

参考文献

[1] 全兆景,乔征瑞,李金香,等. 基于 MPC 算法优化的贝叶斯网络变压器故障诊断[J]. 电子测量技术,2021,

44(17): 41-45.

TONG ZH J, QIAO ZH R, LI J X, et al. Transformer fault diagnosis based on MPC algorithm optimized by Bayesian network [J]. Electronic Measurement Technology, 2021, 44(17): 41-45.

- [2] 陈铁,陈卫东,李咸善,等. 基于油中溶解气体分析的变压器故障预测[J]. 电子测量技术,2021,44(22):25-31.
- CHEN T, CHEN W D, LI X SH, et al. Transformer fault prediction based on dissolved gas analysis in oil[J]. Electronic Measurement Technology, 2021, 44(22): 25-31.
- [3] XIANG CH M, HUANG ZH Y, LI J, et al. Graphic approaches for faults diagnosis for Camellia insulating liquid filled transformers based on dissolved gas analysis[J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2018, 25(5): 1897-1901.
- [4] Mineral oil-impregnated electrical equipment in service-guide to the interpretation of dissolved and free gases analysis; IEC 60599-2007[S]. Geneva, Switzerland: IEC, 2007.
- [5] DUVAL M. A review of faults detected by gas-in-oil analysis in transformers [J]. IEEE Electrical Insulation Magazine, 2002, 18(3): 8-17.
- [6] ROGERS R R. IEEE and IEC codes to interpret incipient faults in transformers, using gas in oil analysis [J]. IEEE Transactions on Electrical Insulation, 2007, 13(5):349-354.
- [7] SHI H Y, CHEN M X. A two-stage transformer fault diagnosis method based multi-filter interactive feature selection integrated adaptive sparrow algorithm optimised support vector machine[J]. IET Electric Power Applications, 2023, 17(3): 341-357.
- [8] DUVAL M. The duval triangle for load tap changers, non-mineral oils and low temperature faults in transformers[J]. IEEE Electrical Insulation Magazine, 2008, 24(6):22-29.
- [9] 谢国民,江海洋. 基于 Adaboost-INGO-HKELM 的变压器故障辨识[J]. 电力系统保护与控制,2024,52(5): 94-104.
- XIE G M, JIANG H Y. Transformer fault identification based on Adaboost-INGO-HKELM[J]. Power System Protection and Control, 2024, 52(5): 94-104.
- [10] WANG L, LITTLER T, LIU X Q. Dynamic incipient fault forecasting for power transformers using an LSTM model[J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2023, 30(3): 1353-1361.
- [11] 徐耀松,包力铭,管智峰,等. 基于 IPPA 优化 PNN 的

- 变压器故障诊断研究[J]. 电子测量与仪器学报, 2022, 36(10):138-145.
- XU Y S, BAO L M, GUAN ZH F, et al. Research on transformer fault diagnosis based on IPPA optimized PNN[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(10):138-145.
- [12] YE ZH Y, LI K, CHEN G L, et al. Accelerated design and fabrication of thermal protection coating via high-throughput experiments and machine learning[J]. Corrosion Science, 2024, 238: 112388.
- [13] ZHANG D, LI C B, SHAHIDEHPOUR M, et al. A bi-level machine learning method for fault diagnosis of oil-immersed transformers with feature explainability[J]. International Journal of Electrical Power & Energy Systems, 2022, 134: 107356.
- [14] YANG J X, LIAO C B, HU X, et al. Transformer fault diagnosis based on DGA and TPE-LightGBM[J]. Journal of Electric Power Science and Technology, 2024, 39(4): 70-77.
- [15] GHOUALMI L. Feature selection based on machine learning algorithms: A weighted score feature importance approach for facial authentication[C]. 2022 3rd International Informatics and Software Engineering Conference(IISEC), IEEE, 2022: 1-5.
- [16] TAN X M, GUO CH, WANG K, et al. A novel two-stage dissolved gas analysis fault diagnosis system based semi-supervised learning [J]. High Voltage, 2022,7(4):676-691.
- [17] 商立群,侯亚东,黄辰浩,等. 基于 IDOA-DHKELM 的变压器故障诊断[J]. 高电压技术, 2023 (11): 4726-4735.
- SHANG L Q, HOU Y D, HUANG CH H, et al. Transformer fault diagnosis based on IDOA-DHKELM[J]. High Voltage Engineering, 2023(11): 4726-4735.
- [18] 吴振. 基于邻域粗糙集与 GWO-SVM 的变压器故障诊断方法研究[D]. 长春: 长春工业大学, 2022.
- WU ZH. Research on transformer fault diagnosis method based on neighborhood rough set and GWO-SVM [D]. Changchun: Changchun University of Technology, 2022.
- [19] 王芳元. 基于实域粗糙集的图像认知词袋模型与缺陷车牌识别研究[D]. 合肥: 合肥工业大学, 2017.
- WANG F Y. Research on image cognitive bag of words model and defect license plate recognition based on real domain rough sets [D]. Hefei: Hefei University of Technology, 2017.
- [20] 李黄曼,张勇,张瑶. 基于 ISSA 优化 SVM 的变压器故障诊断研究[J]. 电子测量与仪器学报, 2021, 35(3):7.
- LI H M, ZHANG Y, ZHANG Y. Research on transformer fault diagnosis based on ISSA optimization SVM[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(3): 7.
- [21] 肖云波,范菁,张宜,等. 基于改进粒子群算法与油中溶解气体的变压器故障诊断的研究[J]. 电子测量技术, 2021, 44(18): 122-128.
- XIAO Y B, FAN J, ZHANG Y, et al. Research on transformer fault diagnosis based on improved particle swarm optimization algorithm and dissolved gas in oil[J]. Electronic Measurement Technology, 2021, 44(18): 122-128.
- [22] XUE J K, SHEN B. Dung beetle optimizer: A new meta-heuristic algorithm for global optimization[J]. The Journal of Supercomputing, 2023, 79 (7): 7305-7336.
- [23] SOWMYA R, PREMKUR M, JANGIR P. Newton-raphson-based optimizer: A new population-based metaheuristic algorithm for continuous optimization problems[J]. Engineering Applications of Artificial Intelligence, 2024, 128: 107532.
- [24] WAN J H, CHEN H M, YUAN ZH, et al. A novel hybrid feature selection method considering feature interaction in neighborhood rough set[J]. Knowledge-Based Systems, 2021, 227: 107167.
- [25] CHEN T Q, GUESTIN C. XGBoost: A scalable tree boosting system[C]. 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016: 785-794.
- [26] SEMENOV A, BOGINSKI V, PASILIAO E L. Neural networks with multidimensional cross-entropy loss functions [C]. Computational Data and Social Networks: 8th International Conference, 2019: 57-62.
- [27] CHEN K J, HU J, HE J L. Detection and classification of transmission line faults based on unsupervised feature learning and convolutional sparse autoencoder[J]. IEEE Transactions on Smart Grid, 2016, 9(3): 1748-1758.
- [28] ILLIAS H A, ZHAO L W. Identification of transformer fault based on dissolved gas analysis using hybrid support vector machine-modified evolutionary particle swarm optimisation[J]. PLoS One, 2018, 13 (1): e0191366.
- [29] LI EN W, WANG L N, SONG B. Fault diagnosis of power transformers with membership degree[J]. IEEE Access, 2019, 7: 28791-28798.
- [30] GOUDA O E. Proposed heptagon graph for DGA

- interpretation of oil transformers[J]. IET Generation, Transmission & Distribution, 2018, 12(2): 490-498.
- [31] GOUDA O E. Proposed three ratios technique for the interpretation of mineral oil transformers based dissolved gas analysis [J]. IET Generation, Transmission & Distribution, 2018, 12(11): 2650-2661.
- [32] 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.
- [33] KIM S W, KIM S J, SEO H D, et al. New methods of DGA diagnosis using IEC TC 10 and related databases Part 1: Application of gas-ratio combinations[J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2013, 20(2): 685-690.
- [34] 周秀,田天,怡恺,等. 基于邻域粗糙集与 AMPOS-ELM 的变压器 DGA 故障诊断方法[J]. 电力科学与技术学报, 2022, 37(3):8.
- ZHOU X, TIAN T, YI K, et al. Transformer DGA fault diagnosis method based on neighborhood rough set and AMPOS-ELM[J]. Journal of Electric Power Science and Technology, 2022, 37(3): 8.
- [35] DORNENBURG E, STRITTMATTER W. Monitoring oil-cooled transformers by gas-analysis[J]. Brown Boveri Review, 1974, 61(5): 238-247.

作者简介

杨勇,硕士研究生,主要研究方向为变压器在线监测和故障诊断、边缘计算。

E-mail:1451754339@qq.com

谢菊芳(通信作者),博士,主要研究方向为农业信息化和电力系统自动化等。

E-mail:xjufang0311@sina.com