第 50 卷 第 12 期：5289-5301 高电压技术 Vol.50, No.12: 5289-5301
2024 年 12 月 31 日 High Voltage Engineering December 31, 2024
DOI: 10.13336/j.1003-6520.hve.20231800

基于改进 LightGBM 混合集成模型的变压器故障识别方法

荆澜涛[1]，张 野[1]，张 彬[2]，姚 晔[3]，许 东[4]，王 亮[1]
（1. 沈阳工程学院电力学院，沈阳 110136；2. 国网辽宁省电力有限公司电力科学研究院，沈阳 110000；3. 国网辽宁省电力有限公司检修分公司，沈阳 110003；4. 国网辽宁省电力有限公司实业分公司，沈阳 110004）

摘 要：针对变压器故障识别方法在处理不均衡故障数据时存在较大偏差的问题，构建了一种基于改进轻量级梯度提升机的混合集成模型，用以变压器故障识别。首先，提出一种结合梯度调和损失函数和交叉熵损失函数的改进轻量级梯度提升机(gradient harmonizing mechanism loss and cross entropy loss improved light gradient boosting machine，GCLightGBM)，提升模型对数据集中少数样本的关注度。然后，针对 GCLightGBM 中参数特异性取值影响模型识别能力的问题，提出一种基于 GCLightGBM 的混合集成模型，进一步提高其准确率的同时，确保模型对现实多变不均衡数据集依然保持良好的准确率。实验结果表明，GCLightGBM 可有效解决少数类样本准确率低的问题，整体准确率高达 0.911。且针对其他多变不均衡数据集，基于 GCLightGBM 混合集成模型故障识别方法平均准确率高达 0.988。

关键词：变压器；不均衡样本；故障识别；梯度调和损失函数；Stacking 集成框架

Transformer Fault Identification Method Based on Improved LightGBM Hybrid Integration Model

JING Lantao[1], ZHANG Ye[1], ZHANG Bin[2], YAO Ye[3], XU Dong[4], WANG Liang[1]
(1. School of Electric Power, Shenyang Institute of Engineering, Shenyang 110136, China; 2. Electric Power Research Institute of State Grid Liaoning Electric Power Co., Ltd., Shenyang 110000, China; 3. State Grid Liaoning Electric Power Co., Ltd. Maintenance Branch, Shenyang 110003, China; 4. State Grid Liaoning Electric Power Co., Ltd. Industrial Branch, Shenyang 110004, China)

Abstract：There is a significant bias when imbalanced fault data are dealt with by transformer fault identification methods . To solve this problem, we have built a hybrid ensemble model based on an improved lightweight gradient boosting machine. First, we propose a gradient harmonizing mechanism loss and cross entropy loss improved light gradient boosting machine (GCLightGBM). This approach enhances the model's attention to minority samples in the dataset. Then, to address the issue of parameter specificity affecting the model's recognition capability in GCLightGBM, we propose a hybrid ensemble model based on GCLightGBM. This model further improves accuracy while ensuring good performance on real-world varied and imbalanced datasets. The experimental results show that GCLightGBM can effectively solve the problem of low accuracy of minority samples, and the overall accuracy is as high as 0.911. In addition, for other variable and unbalanced datasets, the fault identification method based on GCLightGBM hybrid ensemble model has an average accuracy of 0.988.

Key words：transformer; unbalanced samples; fault identification; gradient harmonic loss function; Stacking integration framework

0 引言

电力变压器作为电能传输与转换的关键设备，其健康运行与否直接关系到电力系统的稳定[1]。目前在运变压器大多是油浸式变压器，变压器发生故障时会产生溶解于变压器油的气体，如氢气(H_2)、甲烷(CH_4)、乙烷(C_2H_4)、乙烯(C_2H_2)、乙炔(C_2H_6)等，且油中气体浓度与变压器故障类型具有强相关性[2-3]。因此，可以通过油中溶解气体分析法

基金资助项目：创新能力提升联合基金 (2022-NLTS-16-05)；辽宁省教育厅基本科研项目 (LJKMZ20221707)。
Project supported by Innovation Capability Enhancement Joint Fund (2022-NLTS-16-05), Basic Research Project of Liaoning Provincial Department of Education (LJKMZ20221707).

(dissolved gas analysis，DGA)检测油中气体浓度，用以识别故障特征[4]。随着人工智能的发展，将DGA与深度学习、机器学习、神经网络等智能算法相结合被广泛应用于油浸式变压器故障识别中，能大幅提升故障识别的准确性[5-6]。

为实现准确的预测结果，人工智能模型拟合过程依赖于充足且均衡的训练数据。然而，在变压器实际运行工况中，发生各类故障的概率随机性强，导致采集的各故障样本数量差异较大，总体数据不均衡，极大程度弱化了训练网络的泛化能力，造成训练模型存在较大偏差[7]。目前，业界针对不平衡数据集分析方法主要采用以下两类：数据预处理和优化分类模型。数据预处理主要包括数据增强以及欠采样两大类。数据增强主要采用过采样、变分自编码器(variational auto-encoder，VAE)等样本生成方法，生成少数类相似样本以达到数据均衡，但增加的数据与原始数据相似性过高，容易造成模型过拟合[8]。而欠采样主要剔除多数类中的部分样本数据，容易造成数据丢失。优化分类模型主要包括对神经网络、机器学习等现有模型的改进及优化。文献[9]采用改进变分自编码器生成少数类故障样本解决数据不均衡，并提出一种改进1维卷积神经网络(one-dimensional improved convolutional neural network，1D-ICNN)作为故障识别分类器，准确度提升了0.135，但其未考虑生成数据导致模型过拟合问题，且大多数神经网络分类模型常使用全连接层搭配Softmax等激活函数实现多分类的概率输出，其线性分类能力一定程度上不如轻量级梯度提升机(light gradient boosting machine，LightGBM)、支持向量机(support vector machine，SVM)、极致梯度提升机(xtreme gradient boosting，XGBoost)等非线性映射方法。文献[10]提出了一种基于遗传算法优化XGBoost的油浸式变压器故障诊断方法。以9维故障特征为输入，建立基于XGBoost的故障诊断模型，并采用遗传算法对模型中的多个超参数同时进行优化，诊断精度和稳定性高于现有SVM、卷积神经网络(convolutional neural network，CNN)等人工智能模型。文献[11]中Liu等采用多层卷积神经网络对变压器故障数据进行多元信息特征提取，利用LightGBM模型对上述特征进行分类识别，并在原文数据集上验证了LightGBM性能优于XGBoost、SVM等传统机器学习模型。但上述机器学习模型在数据均衡情况下完成，规避了数据样本不均衡所导致准确率下降的问题，因此无法适用于实际运行工况。文献[12]针对上述问题，采用聚焦损失函数(focal loss function，Focal Loss)对LightGBM模型进行优化。但Focal Loss性能在参数选取方面存在局限性，使得上述模型实际应用较差。针对Focal Loss的不足，文献[13]中Li B等提出了梯度调和损失函数(gradient harmonizing mechanism loss function，GHM Loss)，通过在梯度分布上添加一个渐进的分桶函数，来确定每个样本的权重，进而弱化多数样本对损失函数的贡献，提高模型对少数样本关注度。与Focal Loss损失函数相比，GHM Loss不需要手动调节超参数或使用特定的类别平衡策略，更简单高效。但GHM Loss针对样本总数较少的小样本数据集，会出现误判的现象。

算法层面的另一种改进方式是集成学习。集成学习基于弱学习器的输出结果，通过元学习器对上述结果进行融合，是提高模型整体泛化性和鲁棒性的一种模型融合方法。文献[14]提出一种基于Stacking元学习策略的组合方式，构建了一种多分类器组合模型。通过实验证明，该组合分类器精度优于单个分类器精度。

综上所述，应用于不平衡数据的变压器故障识别方法仍存在改进空间：1）仅对算法整体性能改进优化，而忽略故障数据的不平衡特性会导致模型不适用少数类样本的识别分析，进而影响识别结果。2）GHM Loss处理不平衡数据能效显著，但在变压器故障识别领域尚未应用。为规避样本总数较少导致GHM Loss误判，应进一步挖掘GHM Loss的应用特点，结合故障数据不平衡特性，设计重视少数类样本的改进方法。3）例如采用Focal Loss的改进策略，针对给定的数据集，往往需要实验得出重要的参数取值，其特定取值无法广泛应用于现实复杂多变的变压器样本，实用性较差，亟需采取某些策略提高识别能力的稳定性。基于上述改进方向，提出一种融合损失函数对LightGBM进行改进，其中融合损失函数可结合GHM Loss和采用交叉熵损失函数(cross entropy loss function，CE Loss)的优点，加大模型对少数类样本的关注度(如局部放电样本)，使模型在每次迭代更新时更关注少数类样本，进而提高模型在不平衡变压器数据集下对少数样本的分类性能。针对变压器实际中多种运行工况，为使GCLightGBM无需实验确定融合损失函数比例系数最优取值依然具备良好性能，对Stacking框架提出

全新的用法，构建基于 GCLightGBM 的混合集成模型，提升模型整体准确率和应用可行性。最后实验证明，在不平衡数据集下，所提识别方法具有优越性和鲁棒性。

1　算法原理

1.1　油中溶解气体数据

油浸式变压器内部发生诸如短路等故障后，会出现升温或绕组放电等现象，引发变压器内部固体绝缘材料和变压器油之间化学反应，生成 C_2H_2、C_2H_4 等不溶解于变压器油的烃类气体[15]。烃类气体浓度、种类随变压器的劳损程度、运行时间动态变化。对于不同的变压器运行工况，分解产生的气体速率、种类大不相同。《变压器油中溶解气体分析和判断导则》[16]中给出气体比值编码组合与故障类型的映射关系，将变压器内部故障分为热故障(中低温过热、高温过热)和电故障(局部放电、低能放电、高能放电)。

然而，上述气体映射关系难以蕴含变压器所有故障工况，无法精确映射变压器溶解气体数据与故障种类之间的关联性。文献[17]通过大量实验得出：基于 5 种气体浓度不同组合形式的九比值法，如表 1 所示。其中，n 为气体浓度，C1 代表 CH_4 气体；C2 代表 C_2H_2、C_2H_4 和 C_2H_2 气体总和。

随着人工智能的发展，结合气体比值技术的深度学习、机器学习、集成模型等智能模型被广泛应用于变压器故障识别中，故障识别能力得到极大提高。因此，本文选取集成模型中性能较好的 LightGBM 与气体比值法相结合，使用不同气体比值特征作为输入，预测对应的故障类型，构建变压器识别模型。

例如：LightGBM、XGBoost、SVM 等人工智能算法依赖均衡的数据样本，但变压器实际运行中各类故障发生概率随机，导致故障数据不均衡。本文数据分布情况如图 1 所示，其中局部放电故障仅占总样本的 6.2%，人工智能模型倾向于学习数量多的样本，忽略对数量少样本的学习，进而导致少数类样本分类准确率降低。因此，如何在不平衡变压器故障数据下提高人工智能算法故障识别性能是亟须解决的问题。

1.2　传统 LightGBM

LightGBM 是由微软开发的一种高效且快速的梯度增强框架[18]。该算法的核心思想是通过特征分

表 1　变压器油中气体比值特征

Table 1　Characteristics of gas ratio in transformer oil

特征参量编号	气体比值特征
1	$n(CH_4)/n(H_2)$
2	$n(C_2H_4)/n(C_2H_2)$
3	$n(CH_4)/n$(C1+C2)
4	$n(C_2H_2)/n$(C1+C2)
5	$n(C_2H_4)/n(C_2H_6)$
6	$n(C_2H_4)/n$(C1+C2)
7	$n(H_2)/n(H_2+$C1+C2)
8	$n(CH_4+C_2H_4)/n$(C1+C2)
9	$n(C_2H_6)/n$(C1+C2)

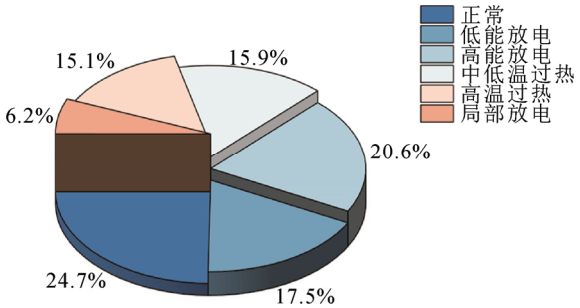


图 1　数据分布情况

Fig.1　Data distribution

裂构建新树，基于残差拟合方法迭代计算预测值与实际值之间的残差。最终，将所有树的预测结果累加得到最终的预测结果。

假设由 N 个变压器数据样本组成的训练集 Q，$Q=\{(x_i,y_j)\}_{i=1\,j=0}^{N\ \ 5}$，其中 $\boldsymbol{x}_i=\left[x_i^1,x_i^2,x_i^3\cdots x_i^9\right]\in \boldsymbol{X}=[x_1,x_2,\cdots,x_N]^{\mathrm{T}}$ 代表气体比值数据，$x_i^1,x_i^2,x_i^3\cdots x_i^9$ 依次表示表 1 中 9 种气体比值数据。$y_j\in\boldsymbol{Y}=[y_0,y_1,y_2,y_3,y_4,y_5]$ 表示类别标签，采用 one-hot 编码形式，具体如下：

$$\boldsymbol{Y}=\left[y_0,y_1,y_2,y_3,y_4,y_5\right]=\begin{bmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{bmatrix}\tag{1}$$

式中：$[y_0,y_1,y_2,y_3,y_4,y_5]$ 的 6 列数据依次表示正常、低能放电、高能放电、中低温过热、高温过热、局部放电共 6 个数据类型。

则集成模型 LightGBM 可表示为：

$$\hat{y}_i=\sum_{i=1}^{k}f_i\left(x_i\right),\ f_i\in F\tag{2}$$

式中：$f_i\left(x_i\right)$ 为第 i 棵决策树；x_i 为变压器第 i 个气体比值数据；$\hat{y}_i$ 为预测样本类别；F 为树模型；k

为决策树总数。

由此可构建目标函数为：

$$G(x_i)=\sum_{t=1}^{k}\phi\left(y_i,F_{t-1}(x_i)+f_t(x_i)\right)+\sum_{k}\Omega(f_k) \quad (3)$$

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\|w\|^2 \quad (4)$$

式中：$\phi\left(y_i,F_{t-1}(x_i)+f_t(x_i)\right)$为损失函数；$\Omega(f_k)$为正则化；$F_{t-1}(x_i)$为前 t–1 棵树模型；$f_t(x_i)$为第 t 棵树模型；w 为每棵树叶子权重；λ、γ 为超参数；T 为树模型中的叶子节点数；y_i 为第 i 个气体比值数据的真实类别。

为使目标函数最小，LightGBM 模型使用牛顿法寻求一个映射关系 $\bar{G}(x)$，来近似目标函数 $G(x)$，式(3)可推导为：

$$\bar{G}(x_i)\approx\sum_{i=1}^{n}\left(g_i f_t(x_i)+\frac{h_i f_t^2(x_i)}{2}\right)+\sum_{k}\Omega(f_k) \quad (5)$$

式中：g_i 为损失函数的一阶导数；h_i 为损失函数的二阶导数。

对于 LightGBM 分类模型，常采用 CE Loss。以二分类为例，交叉熵损失函数可表示为：

$$L_{\mathrm{CE}}(x_i,y_i)=\frac{-\sum_{i=1}^{N}\left(y_i\ln P(\hat{y}_i)+(1-y_i)\ln\left(1-P(\hat{y}_i)\right)\right)}{N} \quad (6)$$

式中：L_{CE} 为 CE Loss；$\hat{y}_i$ 为第 i 个样本的正确类别；P 为计算预测样本为正类的概率。则式(5)中对应的一阶、二阶导数可表示为：

$$g_i=\left(y_i-P(\hat{y}_i)\right) \quad (7)$$

$$h_i=\left(1-P(\hat{y}_i)\right)P(\hat{y}_i) \quad (8)$$

由于变压器样本不均衡，如本文中正常样本数是局部放电样本数的 4 倍。正常样本的损失值在总体样本损失中占比大，为使式(5)最小，模型会倾向于预测占比大的类别，过多学习正常样本的特征，忽略局部放电样本类别。通过多数类的正确性，隐藏少数样本的错误率，导致模型总体分类结果呈现较高的“正确率”，但实际中不能很好地识别局部放电样本类别。

1.3 改进 LightGBM 模型

在本文变压器故障数据样本中，正常、低能放电、高能放电等样本占比较大，且不同数据样本之间不均衡。使用 CE Loss 进行训练会导致模型过多关注多数类样本。文献[13]提出一种 GHM Loss，通过使用渐进式加权直方图将样本的梯度进行分档，然后为每个档位分配一个动态权重，使得难分样本(如极小的正样本或相似度高的样本)获得更高的权重，从而促进模型对这些样本的学习。可有效解决上述问题，其计算式为：

$$L_{\mathrm{GHM}}=L_{\mathrm{CE}}\beta_i \quad (9)$$

式中：L_{GHM} 为 GHM Loss；β_i 为第 i 个样本的权重。

β_i 表示每个样本的重要性，通过第 i 个样本的梯度密度 G_g 进行计算，计算式如下：

$$\beta_i=\frac{N}{G_g} \quad (10)$$

G_g 的计算式如下：

$$G_g=\frac{R_{\mathrm{ind}(g_i)}}{\epsilon} \quad (11)$$

式中：$R_{\mathrm{ind}(g_i)}$ 表示以 g_i 为中心、长度为 ϵ 的区域中的样本数。

然而，通过实践发现在故障识别分类任务中，直接使用 GHM Loss 替换 CE Loss，并不能使网络分类性能达到最佳。经过分析可知，一些故障数据被误判为离群点，权重降低，导致最终的分类性能不佳。而 CE Loss 逐数据进行交叉熵损失计算。因此，本文将 GHM Loss 和 CE Loss 进行融合，作为模型的总损失函数。融合损失函数以 GHM Loss 为主，解决数据不均衡以及样本难分问题，CE Loss 为辅，降低 GHM Loss 对离群点误判。

融合损失函数表示为：

$$L=(1-\alpha)L_{\mathrm{CE}}+\alpha L_{\mathrm{GHM}} \quad (12)$$

式中：α 为融合损失函数比例系数。

用融合损失函数替换原本 LightGBM 中的 CE Loss。根据式(5)，GCLightGBM 的目标函数为：

$$\bar{G}(x_i)\cong\sum_{i=1}^{n}\left(g_i' f_t(x_i)+\frac{h_i' f_t^2(x_i)}{2}\right)\sum_{k}\Omega(f_k) \quad (13)$$

式中：g_i' 表示损失函数的一阶导数；h_i' 表示损失函数的二阶导数。计算式如下：

$$g_i'=((1-\alpha)+\alpha\beta_i)\left(y_i-P(\hat{y}_j)\right) \quad (14)$$

$$h_i'=((1-\alpha)+\alpha\beta_i)\left(1-P(\hat{y}_j)\right)P(\hat{y}_j) \quad (15)$$

综上所述，针对数据集中的少数样本，如局部放电样本，其导数在附近区域 ϵ 内样本数较少，由式(11)可知，将获得较小的梯度密度 G_g，由式(9)、(10)可知其损失函数将获得更大的权重 β_i。相反地，针对多数样本，将获得较小的权重。因此，为使式(13)最小，模型将加大对少数类样本关注度，进而减少少数类别的损失值。

面对变压器不平衡故障数据，融合损失函数比例系数 α 的取值是影响 GCLightGBM 模型识别效果的重要因素，α 取值应兼顾少数类、整体准确率均较高。为探讨 α 对少数类样本、以及整体准确率的影响，本文将变压器气体比值数据作为输入，对应故障标签作为输出，准确率作为评价指标。选取 α 分别为 0、0.2、0.4、0.6、0.8、1 的 6 个 GCLightGBM 模型作为分类模型，测试 α 对模型少数类以及整体准确率的影响，实验结果如图 2 所示。其中，本文局部放电样本为少数类样本。

由图 2 可知，当 α 取 0、0.2 时，GHM Loss 占比较低，对网络性能影响较小，整体准确率较低，随着 α 取值逐步增大，网络性能先升高后降低，当 α 取 0.8 时，GCLightGBM 模型的整体准确率为 0.911，高于不使用 GHM Loss(α=0)和仅使用 GHM Loss(α=1)的模型准确率，表明混合损失函数的故障识别性能优于 GHM Loss 和 CE Loss。当 α=0.6、α=0.8 时，局部放电样本准确率为 1.00，但 α=0.8 时模型整体准确率最高，因此，针对本文数据集，最佳比例系数 α 取 0.8。

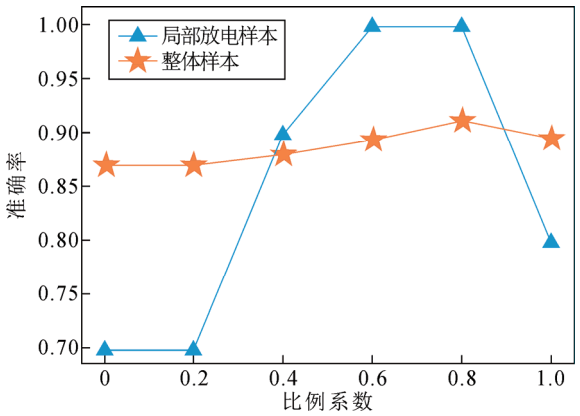


图 2　α 对 GCLightGBM 模型准确率的影响

Fig.2　Influence of α on the accuracy of GCLightGBM model

1.4　基于 GCLightGBM 混合集成学习模型

融合损失函数的比例系数 α 随数据样本改变而改变，本文最佳比例系数未必适用于其他数据集。针对实际多变的运行工况，为使 GCLightGBM 无需实验确定 α 最优取值依然具备良好性能，本文基于 Stacking 框架创新性地对不同特异性取值的 GCLightGBM 进行集成，提出一种基于 GCLightGBM 混合集成模型，更加适用于实际多变运行工况。

Stacking 也称堆叠泛化[19]，由 Wolpert 提出，是一种异构串行学习方法，它由多个基学习器组成第一层预测模型，并使用一个元学习器作为第二层的融合模型。基学习器的预测结果用作元学习器的输入，并通过再次训练来生成最终的预测结果。Stacking 的核心思想是利用多个基学习器的预测结果，通过元学习器的融合来提高整体预测性能[20]。Stacking 框架性能与基、元学习器选取有关，本文模型选取如下：

1）基学习器选取

在 Stacking 模型的基学习层中，选择差异较大的模型能够增加模型的多样性，并降低基学习层输出标签之间的相关性。通过使用差异性模型，Stacking 模型可以综合各个模型的优势，以多重视角对数据进行建模，更好地适应不同的数据分布和特征。这种多样性和降低相关性的组合可为 Stacking 模型带来更强大的学习能力和识别性能。因此，在选择基学习层模型时，优先选择差异性较大的模型，以凸显模型的多样性，并降低输出标签之间的相关性。这样的策略可以更好地利用集成模型的优势，提升整体性能[21]。

本文将 α 在[0、0.2、0.4、0.6、0.8、1]中不同取值的 6 个 GCLightGBM 作为第一层基学习器。上述模型全部基于树模型构建，为增加第一层基学习器差异性，选择非线性模型 SVM 也作为第一层基学习器。SVM 可使用非线性核函数 $\varphi(x)$，将样本映射到高维特征空间，使得原本在低维空间中不可分的变压器数据样本，在高维空间中变得可分，其原理可表示为：

$$\begin{cases}\min \dfrac{1}{2}\|\omega\|^2 + c\sum\limits_{i=1}^{N}\xi_i \\ \text{s.t.}\, y_{si}(\boldsymbol{\omega}^{\mathrm{T}}\varphi(\boldsymbol{x}_{si})+b) \geq 1-\xi_i, \xi_i \geq 0, i=1,2,\cdots,R\end{cases} \tag{16}$$

式中：$\boldsymbol{\omega}$ 为超平面法向量；c 为惩罚系数；ξ_i 为松弛系数；b 为偏置参数；$\varphi(\boldsymbol{x}_{si})$ 为非线性核函数；$\boldsymbol{x}_{si}$ 为输入到 SVM 中第 i 个样本的特征向量；y_{si} 为输入到 SVM 中第 i 个样本的标签；R 为输入到 SVM 的样本总数量。

2）元学习器选取

Stacking 框架需要选择泛化性强、学习能力强的模型作为元学习器[21]，因此，本文选取极限学习机(extreme learning machine，ELM)作为元学习器。

ELM 是一种基于前馈神经网络的机器学习算法，具有泛化性强、学习能力强的优点[22]。单隐层 ELM 的结构如图 3 所示，其具体原理为：

$$\sum_{j=1}^{o}\theta_j[\mu(\boldsymbol{a}_j\bullet\boldsymbol{x}_{ei}+\boldsymbol{\gamma}_j)]^{\mathrm{T}}=\hat{y}_{ei},i=1,2,\cdots,M \quad (17)$$

式中：$\boldsymbol{x}_{ei}$为输入到 ELM 中的第 i 个样本的特征向量；y_{ei}为 ELM 输出的预测样本类别；M为输入到 ELM 的样本总数量；μ为激活函数；o为隐含层神经元个数；$\boldsymbol{a}_j$为权重向量；$\boldsymbol{\gamma}_j$为偏置向量；“$\bullet$”为向量内积；θ_j为隐含层到输出层权重参数。

综上所述，本文基学习器选择 GCLi-ghtGBM(α=0)、GCLightGBM(α=0.2)、GCLi-ghtGBM(α=0.4)、GCLightGBM(α=0.6)、GCLi-ghtGBM(α=0.8)、GCLightGBM(α=1)、SVM 共 7 个模型，元学习器采用 ELM。所提出的基于 GCLightGBM 混合集成模型框架图，如附录图 A 所示。

2 基于 GCLightGBM 混合集成模型的变压器故障识别

2.1 基于 GCLightGBM 混合集成模型识别流程

本文所提基于 GCLightGBM 混合集成模型的变压器故障识别包括浓度比值矩阵构建、基学习器特征学习、元学习器故障识别 3 个阶段，流程图如图 4 所示。结合附录图 A 和图 4 可得，基于 GCLightGBM 混合集成模型的变压器故障识别具体流程如下：

1）构建变压器气体浓度比值数据。将油中气体浓度数据按表 1 进行折算，将折算后的 9 组气体浓度比值数据作为模型输入，对应故障类型作为模型输出。

2）将气体比值数据划分成数据集和测试集。

3）选择第一个模型，如 GCLightGBM(α=0)模型等。

4）将变压器训练集样本拆分为互不交叉的 5 份，记为训练 l—训练 5。

5）交叉验证。将训练 l—训练 4 作为训练集，训练 5 作为测试集，预测结果为预测值 5。以此类推，将训练 l—训练 5 全部预测，结果为预测值 1—预测值 5。

6）将步骤 4）的预测结果(预测值 1—预测值 5)补充在对应的位置，将预测结果生成的数据作为元学习层模型的训练集。

7）用交叉验证中产生的 5 个模型分别对测试集进行预测，对得到的 5 个预测值取平均值，由该平均值生成新的特征作为元学习层模型测试集的特征数据。

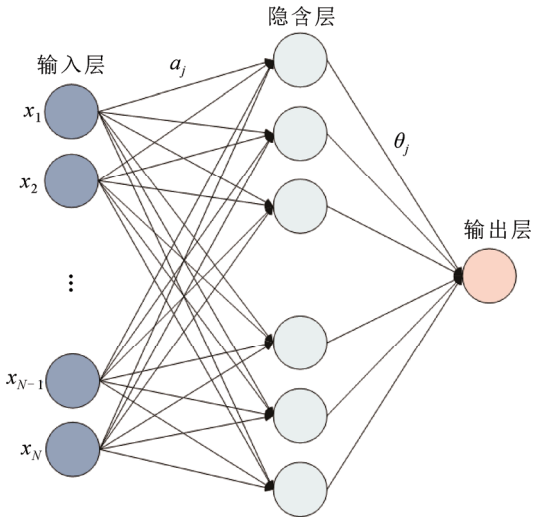


图 3 单隐层 ELM 结构图

Fig.3 Single hidden layer ELM structure diagram

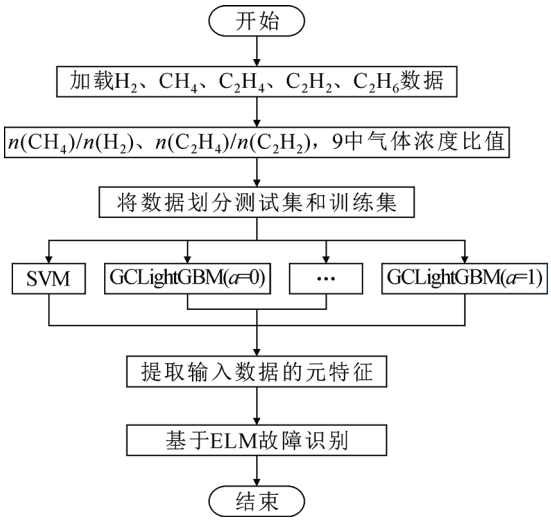


图 4 改进 LightGBM 的混合集成模型流程图

Fig.4 Flow chart of hybrid integration model for improving LightGBM

8）选取下一个基学习器模型，如 GCLightGBM(α=0.2)，重复步骤 2)—步骤 5)。遍历 7 个基学习器模型后，会在元学习器模型的训练集上生成 7 组特征数据。同理，在元学习模型的测试集上也会生成 7 组特征数据。

9）基于获得的新训练集及测试集特征数据，构建 ELM 元学习器模型对故障进行识别。

2.2 评价指标

在不平衡变压器故障识别中，由于数据类别的不均衡，会使训练出的模型向多数类样本偏移，因此仅使用准确率对模型进行评估缺乏可信度[23]。

本文基于变压器混淆矩阵，引入检测准确率 λ_A、漏报率 λ_L、误报率 λ_W、查全率 λ_R、查准率 λ_P 以及

K_{appa}作为识别的评价指标。具体计算式如下：

$$\lambda_A=\frac{n}{N}\times100\% \tag{18}$$

$$\lambda_R=\frac{n_T}{n_R}\times100\% \tag{19}$$

$$\lambda_P=\frac{n_T}{n_P}\times100\% \tag{20}$$

$$\lambda_W=\frac{n_F}{n_P}\times100\% \tag{21}$$

$$\lambda_L=\frac{n_F}{n_R}\times100\% \tag{22}$$

$$K_{appa}=\frac{\lambda_A-\sum_{i=1}^{6}n_{Ri}n_{Pi}\Big/N^2}{1-\sum_{i=1}^{6}n_{Ri}n_{Pi}\Big/N^2}\times100\% \tag{23}$$

式中：n 为识别中预测正确的样本数；n_T 为某种故障样本正确预测个数；n_R 为某种故障样本总数；n_P 表示预测为某种故障样本的总数；n_F 表示某种故障样本错误预测个数。

准确率表示变压器故障识别模型分类预测的综合指标，其数据越高表明模型对样本整体识别能力越强。查准率是模型对某种单一故障识别的指标，查准率越高表明该类故障识别可信度越高。查全率是模型对某种单一故障漏判的指标，查全率越高表明该类故障识别灵敏度越高[23]。漏报率是模型将某种故障类型识别为其他类故障类型的指标。误报率是模型将其他类故障类型识别为某种故障类型的指标。K_{appa}表示模型的识别结果与实际类别之间的一致性程度。较高的 K_{appa} 意味着模型的分类结果更准确，且与实际类别之间一致性更高。文献[12]将分类错误最多的类别定义为难分样本。由图 5(a)可知，本文实验中高能放电样本误判个数最多，因此将高能放电故障定义为本文的难分样本。本文部分实验将列写局部放电样本准确率(少数样本)以及高能放电样本准确率，以此对本文所提模型性能做出有效评估。其中局部放电样本、高能放电样本准确率分别用 A_J、A_G 表示。

3　变压器故障识别实验分析

3.1　实验数据

本文数据来源于文献[24]、IEC TC 10、中国电力行业标准 DL/T722—2014 的故障案例数据库，数据集包含变压器低能放电、高能放电、高温过热、中低温过热、局部放电和正常 6 种运行工况，每种运行工况包含 H_2、C_2H_2、C_2H_4、C_2H_6、CH_4 这 5 种气体特征，采用文献[17]所提九比值方法，对 5 种气体浓度进行折算，获得 9 维变压器浓度比值数据。样本数据共 485 组，并按 7:3 比例进行训练与测试。训练及测试样本数据组成如表 2 所示。其中低能放电 85 例，高能放电 100 例，局部放电 30 例，高温过热 73 例，中低温过热 77 例，正常数据 120 例。正常数据最多，局部放电数据较少。

3.2　GCLightGBM 有效性和优越性验证

为验证 GCLightGBM 有效性，选取不改进 LightGBM 与 GCLightGBM(α=0.8)进行对比。人工智能模型超参数的合理取值可影响模型识别性能，本文超参数取值主要通过多次实验比对获得，超参数选择如表 3 所示，其中，learning rate 代表学习率，其设定过小会导致梯度下降较慢；num_boost_round 确定了要训练的提升树的数量，取值可以增加模型的复杂度和学习能力，提高模型的性能；num_class 表示故障类别总数；max_depth 参数控制了树结构的复杂程度。当 max_depth 设置较小时，树结构简单可能会欠拟合数据，导致性能较差；max_depth 设置较大时，树结构变得复杂，可以更好地拟合训练数据，但可能导致过拟合问题；num_leaves 用于控制 GCLightGBM 中每棵树的叶子节点数，num_leaves 值越大表示模型越复杂；n_estimators 指定集成模型中基础模型(比如决策树)的数量。

表 2　不同数据类型分布情况

Table 2　Distribution of different data types

数据类型	数据编号	训练集样本数	测试集样本数	总数
正常	0	84	36	120
低能放电	1	60	25	85
高能放电	2	70	30	100
中低温过热	4	54	23	77
高温过热	3	51	22	73
局部放电	5	20	10	30

表 3　GCLightGBM 超参数设置

Table 3　GCLightGBM hyperparameter settings

超参数	数值
learning_rate	0.5
num_boost_round	1 000
num_class	6
max_depth	13
num_leaves	45
n_estimators	1 000
bins	10

为验证参数取值的正确性以及验证超参数对计算结果的影响，选取对模型影响较大的参数 learning rate、max_depth 进行实验，对应取值如下：learning rate 为 0.3、0.5、0.7，max_depth 为 8、10、13、15。实验结果如表 4 所示。

由表 4 可知，超参数取值会对模型性能产生影响，模型识别性能可随典型参数 learning rate、max_depth 取值增大先升高后降低。

图 5 给出当 α=0 时和当 α=0.8 时，变压器故障识别结果的可视化混淆矩阵，矩阵中对角线元素表示某故障类样本正确预测的个数，每行数据的加和表示该故障类样本总数，每列数据的加和表示被预测为该故障类的样本总数。

由图 5(a)可知，当 α=0 时，对于少数样本故障 5 错判个数为 3，难分样本故障 2 错判个数为 8。由图 5(b)可知，当 α=0.8 时对于少数样本故障 5 错判个数为 0，难分样本故障 2 错判个数为 2。错判个数显著降低，表明融合损失函数可有效解决样本不均衡及样本难分问题。

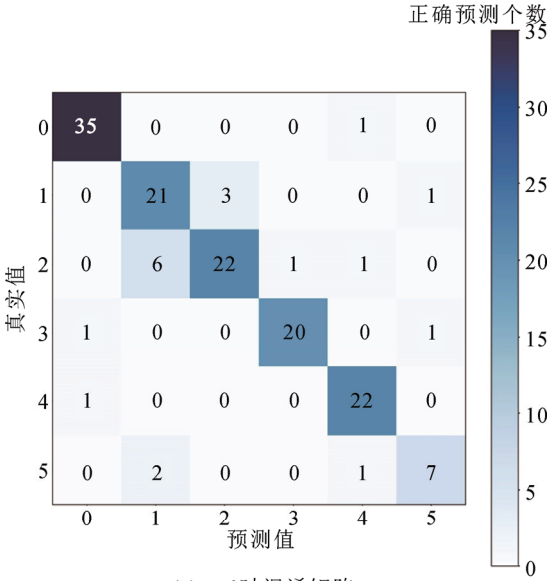
为验证 GCLightGBM 优越性，本文选取 LightGBM、ELM、RF、SSDAE、1D-CNN、SVM、采用 Focal Loss 改进的 FLightGBM、以及处理不平衡数据常采用的过采样技术 Borderline-SMOTE(borderline synthetic minority oversampling technique)、欠采样技术进行对比。FLightGBM 中 Focal Loss 的参数为 g=2.5、a=0.75。BSLightGBM 采用 Borderline-SMOTE 对原始数据集按最大样本数进行采样，即输出的所有样本数均为 120，样本总和为 720，并使用 LightGBM 对扩充后样本进行识别。QLightGBM 表示采用欠采样技术对原始数据按最小样本数进行欠采样，即输出的所有样本数均为 30，样本总和为 180，并使用 LightGBM 对扩充后样本进行识别。上述模型均基于九比值法，为进一步验证九比值法的优越性，与传统基于 DGA 的变压器故障诊断方法(IEC 三比值法)进行对比。将各个 DGA 方法的特征输入 GCLightGBM 进行故障诊断。本文选取各模型评价指标如表 5 所示。

由表 5 可知，在本文数据集中 GCLightGBM 性能最佳，优于 FLightGBM、1D-CNN、SVM 及过采样、欠采样等数据处理策略，整体准确率、K_{appa}、A_J、A_G 分别可达 0.911、0.889、1.00、0.966。相较于 FLightGBM 整体准确率、K_{appa}、A_J、A_G 分别提高了 0.012、0.010、0.100、0.100，表明针对本文数据集 GHM Loss 性能优于 Focal Loss。BSLightGBM 相较于 LightGBM 准确率、A_J 均有所增加，表明过

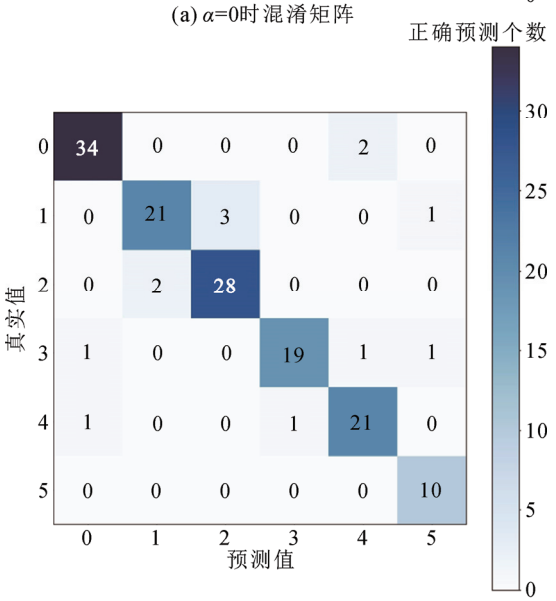
表 4 超参数取值对模型的影响

Table 4 Influence of hyperparameter values on the model

超参数	取值	准确率
learning rate	0.3	0.900
learning rate	0.5	0.911
learning rate	0.7	0.900
max_depth	8	0.900
max_depth	10	0.911
max_depth	13	0.911
max_depth	15	0.900



(a) α=0时混淆矩阵



(b) α=0.8时混淆矩阵

图 5 不同 α 取值的混淆矩阵

Fig.5 Confusion matrix with different α values

采样可以解决样本不均衡，但其 A_G 小于 GCLightGBM，总体性能低于 GCLightGBM，表明 GHM Loss 可增加模型对难分样本关注度，进而提高整体准确性。QLightGBM 相较于 LightGBM 的 A_J 增加了 0.100，表明欠采样可以解决样本不均衡导致准确率下降的问题，但欠采样破坏了数据完整性，使模型不能完整识别其他类别数据，整体准确率下降了 0.027。LightGBM 整体准确率相较于 1D-CNN 提高了 0.018，表明对于多分类问题 LightGBM 线性分类能力优于 1D-CNN，更适合变压器故障诊断。基于九比值识别的 GCLightGBM 相较于基于三比值模型准确率提升了 0.069，结果表明基于九比值的特征筛选法较传统方法能显著提升模型故障诊断精度，验证了九比值方法的优越性。LightGBM 识别准确率、K_{appa}、A_J、A_G 均高于 ELM、RF、SSDAE、SVM 等其他人工智能模型，进一步验证了本文选择 LightGBM 进行改进的合理性。

上述 BSLightGBM 性能取决于 Borderline-SMOTE 生成数据好坏，为验证本文对比实验中 Borderline-SMOTE 的可靠性，以局部放电样本为例，原始数据和生成数据的典型数据如表 6 所示。采用主成分分析法(principal component analysis，PCA)将溶解气体数据映射到二维空间进行可视化处理，如图 6 所示。

由表 6、图 6 可知，实际数据分布较为分散，实际数据中心与生成数据中心接近，且生成数据均在实际数据边界以内，因此生成数据可以作为样本扩充案例。

3.3　基于 GCLightGBM 混合集成模型结果分析

为验证集成模型的可行性，本文采用基于 GCLightGBM 的混合集成模型对 3.1 节所提实验数据进行分析，实验结果如图 7 和表 7 所示。

由图 7 和表 7 可知，146 个测试样本中，共有 5 个故障类型错判(低能放电样本 3 个、高能放电样本 1 个、中低温过热样本 1 个)。本文模型整体准确率、K_{appa} 可达 0.966、0.958。针对少量局部放电故障样本，本文方法可达到 0.909 的查准率以及 1.00 的查全率。表明本文方法可良好解决样本不均衡导致少数类准确率低的问题。针对难分的高能放电故障样本，文献[23]所提处理不均衡数据集方法的查准率、查全率达到 0.900 以上，本文方法可达到 0.935 的查准率以及 0.967 的查全率，表明本文方法对难分样本也具有良好的识别性能。综上所述，本文所提故障识别模型可信度高，对难分以及少数样本查全率、查准率高，可适用于不平衡数据集下变压器故障识别。

表 5　不同不平衡方法的变压器故障识别结果

Table 5　Transformer fault identification results using different unbalanced methods

不平衡方法	准确率	K_{appa}	A_J	A_G
GCLightGBM	0.911	0.889	1.000	0.966
FLightGBM	0.899	0.879	0.900	0.866
BSLightGBM	0.891	0.862	1.000	0.766
LightGBM	0.869	0.869	0.700	0.766
ELM	0.863	0.853	0.800	0.800
1D-CNN	0.851	0.832	0.600	0.833
GCLightGBM-三比值	0.842	0.852	0.700	0.812
QLightGBM	0.842	0.856	0.800	0.800
SVM	0.839	0.826	0.700	0.812
SSDAE	0.801	0.793	0.600	0.733
RF	0.774	0.765	0.600	0.636

表 6　局部放电样本的实际数据与生成数据

Table 6　Actual and generated data of PD samples

故障类型	气体体积分数/%				
	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2
原始	87.3	6.5	1.1	5.2	0
	480.5	37	30	6	0
生成	91.07	3.36	8.67	5.32	0.003
	616.25	48.27	10.27	6.93	0.56

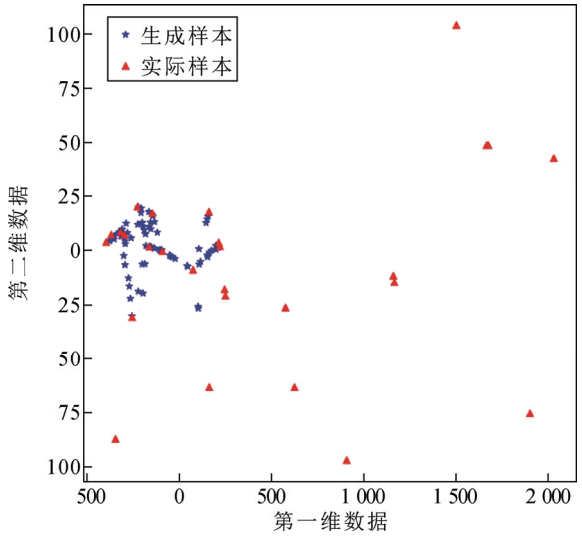


图 6　数据生成前后分布图

Fig.6　Distribution of data before and after data generation

3.4　基于 GCLightGBM 混合集成模型优越性和鲁棒性验证

为验证基于 GCLightGBM 混合集成模型的优越性，本文以准确率作为评价指标，选取 Stacking1、

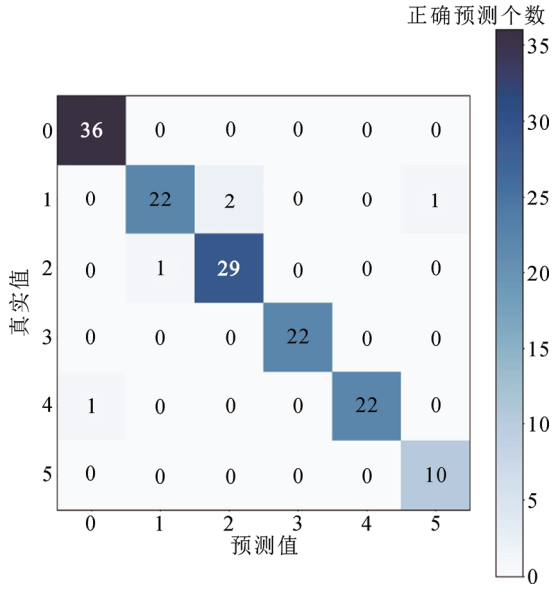


图 7 故障识别分类混淆矩阵

Fig.7 Fault diagnosis classification confusion matrix

Stacking2、Stacking3、Xgboost、ELM、随机森林(random forest，RF)、以及多层感知器(multi-layer perceptron，MLP)与本文方法进行对比，实验结果如表 8 所示。

Stacking1、Stacking2、Stacking3 是目前准确率较高的集成方法。由表 8 可知，本文方法、Stacking1、Stacking2、Stacking3 等采用集成方式的识别模型可以整合各个基学习器优点，相较于单一模型，识别精度更高。且本文所提方法识别性能最高，优于传统集成方式。

为验证基于 GCLightGBM 混合集成模型的鲁棒性，本文从抗干扰性、适普性 2 个维度实验。针对抗干扰性，基于 3.1 节所提数据随机更改训练集 20 组数据类别作为噪声数据(如将某一低能放电样本故障类别改为高能放电)添加到原始数据，构造 5 种含噪声数据实验样本。针对适普性，选取文献[25]中 256 组故障样本数据作为备用样本集。对上述数据集进行实验，实验结果如表 9 所示。

由表 9 可知，针对多变的含噪声样本，本文方法仍有高达 0.958 的准确率。本文方法在备用数据集中也具有良好的适配性，相较于原文方法，总体准确率提高 0.048。表明其具有良好的鲁棒性。

3.5 其他数据集准确性验证

为验证本文所提模型对于不同故障样本数据集都有较好的应用能力，本文按 7:3 随机 5 次划分文献[26]所提实际 413 组故障样本数据。此数据集中低能放电、高能放电、中低温过热、高温过热、局部放电样本数依次为 59、96、86、114、58，数据分布不均衡，实验结果如表 10 所示。

由表 10 可知，本文方法在文献[26]所提实际数据中具有良好的准确性，5 次随机划分原始数据总准确率均高于原文方法的 0.968，平均准确率为 0.989，表明本文模型适用于实际多变的变压器环境。少数类样本平均准确率为 0.977，高于原文方法的 0.943，表明本文方法对少数类样本具有良好的识别能力，可规避样本不均衡导致模型发生偏差。综

表 7 基于 GCLightGBM 的混合集成模型评价指标

Table 7 Evaluation index of hybrid integrated model based on gclightgbm

数据类型	查全率	查准率	漏报率	误报率	K_{appa}	准确率
正常	1.000	0.973	0.000	0.027	0.958	0.966
低能放电	0.880	0.957	0.120	0.043		
高能放电	0.967	0.935	0.033	0.065		
高温过热	1.000	1.000	0.000	0.000		
中低温过热	0.957	1.000	0.043	0.000		
局部放电	1.000	0.909	0.000	0.091		

表 8 不同模型评价指标

Table 8 Evaluation indicators of different models

实验方法	准确率
本文方法	0.966
Stacking1	0.922
Stacking2	0.916
Stacking3	0.892
ELM	0.873
Xgboost	0.862
MLP	0.832
RF	0.823

注：Stacking1 以 LightGBM、SVM、RF、MLP 为基学习器，ELM 为元学习器；Stacking2 以 LightGBM、CNN、SVM 为基学习器，ELM 为元学习器；Stacking3 以 LightGBM、SVM 为基学习器，ELM 为元学习器

表 9 原始数据与噪声数据识别结果

Table 9 Identification results of raw data and noise data

实验方法	总准确率
本文方法-原始数据	0.966
本文方法-噪声数据 1	0.966
本文方法-噪声数据 2	0.958
本文方法-噪声数据 3	0.966
本文方法-噪声数据 4	0.952
本文方法-噪声数据 5	0.958
原文方法-备用数据集	0.949
原文方法-备用数据集	0.979

表 10　本文方法与原论文识别结果

Table 10　Diagnostic results of the method used in this paper and the original paper

实验方法	总准确率	少数类样本准确率
原文方法	0.968	0.943
本文方法-数据集 1	0.984	0.999
本文方法-数据集 2	0.992	0.943
本文方法-数据集 3	0.992	0.999
本文方法-数据集 4	0.999	0.999
本文方法-数据集 5	0.976	0.943

上所述，本文所提方法在不平衡多变数据集下，总体样本、少数类样本均可达到 0.943 以上的准确率，更契合实际应用。

4　结论

1）针对变压器溶解气体数据不均衡导致少数类样本准确率低的问题，提出了结合 GHM Loss 和 CE Loss 的融合损失函数，该函数可改变 LightGBM 的损失权重，加大 LightGBM 对不均衡数据中少数样本的关注度，整体准确率可达 0.911。

2）由于融合损失函数受参数 α 影响较大，为进一步提升所提方法的准确率以及实际应用能力，提出一种基于 GCLightGBM 混合集成模型。模型基于 Stacking 框架，结合多个 α 取值的 GCLightGBM 优点，识别性能高于其他智能算法，总体准确率高达 0.966。针对多变的数据集，无需设置 α 取值，依然保持良好的准确率、鲁棒性，适用于实际不平衡数据集下变压器故障识别。

附录见本刊网络版(http://hve.epri.sgcc.com.cn)。

参考文献　References

[1] 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.

[2] LI J Z, ZHANG Q G, WANG K, et al. Optimal dissolved gas ratios selected by genetic algorithm for power transformer fault diagnosis based on support vector machine[J]. IEEE Transactions on Dielectrics and Electrical Insulation, 2016, 23(2): 1198-1206.

[3] 朱保军，咸日常，范慧芳，等. WRSR 与改进朴素贝叶斯融合的变压器故障诊断技术研究[J]. 电力系统保护与控制，2021，49(20)：120-128.
ZHU Baojun, XIAN Richang, FAN Huifang, et al. Transformer fault diagnosis technology based on the fusion of WRSR and improved naive Bayes[J]. Power System Protection and Control, 2021, 49(20): 120-128.

[4] 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.

[5] 荣智海，齐　波，李成榕，等. 面向变压器油中溶解气体分析的组合 DBN 诊断方法[J]. 电网技术，2019，43(10)：3800-3807.
RONG Zhihai, QI Bo, LI Chengrong, et al. Combined DBN diagnosis method for dissolved gas analysis of power transformer oil[J]. Power System Technology, 2019, 43(10): 3800-3807.

[6] 蒋　晨，王　渊，陈　珉，等. 基于 Kbert 文本聚类模型的变压器故障识别[J]. 高电压技术，2022，48(8)：2991-3000.
JIANG Chen, WANG Yuan, CHEN Min, et al. Transformer fault recognition based on kbert text clustering model[J]. High Voltage Engineering, 2022, 48(8): 2991-3000.

[7] 刘云鹏，许自强，和家慧，等. 基于条件式 Wasserstein 生成对抗网络的电力变压器故障样本增强技术[J]. 电网技术，2020，44(4)：1505-1513.
LIU Yunpeng, XU Ziqiang, HE Jiahui, et al. Data augmentation method for power transformer fault diagnosis based on conditional Wasserstein generative adversarial network[J]. Power System Technology, 2020, 44(4): 1505-1513.

[8] 李　昂，韩　萌，穆栋梁，等. 多类不平衡数据分类方法综述[J]. 计算机应用研究，2022，39(12)：3534-3545.
LI Ang, HAN Meng, MU Dongliang, et al. Survey of multi-class imbalanced data classification methods[J]. Application Research of Computers, 2022, 39(12): 3534-3545.

[9] 李　平，胡根铭. 基于数据增强型一维改进卷积神经网络的变压器故障诊断方法[J]. 电网技术，2023，47(7)：2957-2966.
LI Ping, HU Genming. Transformer fault diagnosis based on data enhanced one-dimensional improved convolutional neural network[J]. Power System Technology, 2023, 47(7): 2957-2966.

[10] 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.

[11] LIU S Q, JI Z S, WANG Y, et al. Multi-feature fusion for fault diagnosis of rotating machinery based on convolutional neural network[J]. Computer Communications, 2021, 173: 160-169.

[12] 武天府，刘　征，王志强，等. 基于 Focal 损失 SSDAE 的变压器故障诊断方法[J]. 电力工程技术，2021，40(6)：18-24.
WU Tianfu, LIU Zheng, WANG Zhiqiang, et al. Transformer fault diagnosis method based on Focal loss SSDAE[J]. Electric Power Engineering Technology, 2021, 40(6): 18-24.

[13] LI B Y, LIU Y, WANG X G. Gradient harmonized single-stage detector[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI, 2019: 8577-8584.

[14] 刘　惠，董锡耀，杨志涵. 融合 Stacking 框架的 BiGRU-LGB 云负载预测模型[J]. 西安电子科技大学学报，2023，50(3)：83-94，104.
LIU Hui, DONG Xiyao, YANG Zhihan. BiGRU-LGB cloud load prediction model incorporating Stacking framework[J]. Journal of Xidian University, 2023, 50(3): 83-94, 104.

[15] 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(04)：1752-1762.
Liao Caibo, Yang Jinxin, Qiu Zhibin, et al. An interpretable transform-

er fault diagnosis method based on Sharpley value and dissolved gas analysis in oil[J]. Power System Technology, 2024, 48(04): 1752-1762.

[16] 中华人民共和国国家经济贸易委员会. 变压器油中溶解气体分析和判断导则：DL/T 722—2000[S]. 北京：中国电力出版社，2001.
State Economic and Trade Commission of the People's Republic of China. Guide to the analysis and the diagnosis of gases dissolved in transformer oil: DL/T 722—2000[S]. Beijing, China: China Electric Power Press, 2001.

[17] 吕　忠，周　强，周　琨，等. 基于遗传算法改进极限学习机的变压器故障诊断[J]. 高压电器，2015，51(8)：49-53.
LYU Zhong, ZHOU Qiang, ZHOU Kun, et al. Fault diagnosis of transformer based on extreme learning machine optimized by genetic algorithm[J]. High Voltage Apparatus, 2015, 51(8): 49-53.

[18] 庞传军，余建明，张　波，等. 基于梯度提升树计及非线性的电力负荷影响因素分析[J]. 电力系统保护与控制，2020，48(24)：71-78.
PANG Chuanjun, YU Jianming, ZHANG Bo, et al. Nonlinear correlation analysis of influence factors of a power load based on a gradient boosting decision tree[J]. Power System Protection and Control, 2020, 48(24): 71-78.

[19] NAIMI A I, BALZER L B. Stacked generalization: an introduction to super learning[J]. European Journal of Epidemiology, 2018, 33(5): 459-464.

[20] GUO L, WANG S F, CAO Z C. An ensemble classifier based on stacked generalization for predicting membrane protein types[C]//Proceedings of the 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics. Shanghai, China: IEE, 2017: 1-6.

[21] 李驰运，缪建明，沈丙振. 基于改进 Stacking 集成学习方法的武器装备体系作战效能预测[J]. 兵工学报，2023，44(11)：3455-3464.
LI Chiyun, MIAO Jianming, SHEN Bingzhen. Prediction of combat effectiveness of weapon and equipment system based on improved stacking ensemble learning method[J]. Acta Armamentarii, 2023, 44(11): 3455-3464.

[22] 曾林俊，许加柱，王家禹，等. 考虑区间构造的改进极限学习机短期电力负荷区间预测[J]. 电网技术，2022，46(7)：2555-2563.
ZENG Linjun, XU Jiazhu, WANG Jiayu, et al. Short-term electrical load interval forecasting based on improved extreme learning machine considering interval construction[J]. Power System Technology, 2022, 46(7): 2555-2563.

[23] 王　艳，李　伟，赵洪山，等. 基于融合少数类过采样均衡多分类数据的改进极限学习机的变压器故障诊断方法[J]. 电网技术，2023，47(9)：3799-3807.
WANG Yan, LI Wei, ZHAO Hongshan, et al. Transformer fault diagnosis fused with synthetic minority over-sampling balanced multi-classification data based on improved extreme learning machine[J]. Power System Technology, 2023, 47(9): 3799-3807.

[24] SHRIVASTAVA K, CHOUBEY A. Data mining approach with IEC based dissolved gas analysis for fault diagnosis of power transformer[J]. International Journal of Engineering Research & Technology, 2013, 2(3): 1-11.

[25] 康佳宇，张沈习，张庆平，等. 基于 ANOVA 和 BO-SVM 的变压器故障诊断方法[J]. 高电压技术，2023，49(5)：1882-1891.
KANG Jiayu, ZHANG Shenxi, ZHANG Qingping, et al. Fault diagnosis method of transformer based on ANOVA and BO-SVM[J]. High Voltage Engineering, 2023, 49(5): 1882-1891.

[26] 朱　莉，汪小豪，李　豪，等. 不平衡样本下基于变异麻雀搜索算法和改进 SMOTE 的变压器故障诊断方法[J]. 高电压技术，2023，49(12)：4993-5001.
ZHU Li, WANG Xiaohao, LI Hao, et al. Transformer fault diagnosis method based on variation sparrow search algorithm and improved SMOTE under unbalanced samples[J]. High Voltage Engineering, 2023, 49(12): 4993-5001



荆澜涛
1985—，女，博士，副教授
研究方向为电力设备在线监测与故障诊断
E-mail: letalaura@163.com

JING Lantao
Ph.D.
Associate professor



王　亮(通信作者)
1983—，男，博士，副教授
研究方向为电力设备在线监测与故障诊断
E-mail: 2569065200@qq.com

WANG Liang
Ph.D.
Associate professor
Corresponding author

收稿日期　2023-10-09　修回日期　2024-03-14　　编辑　卫李静

附录 A

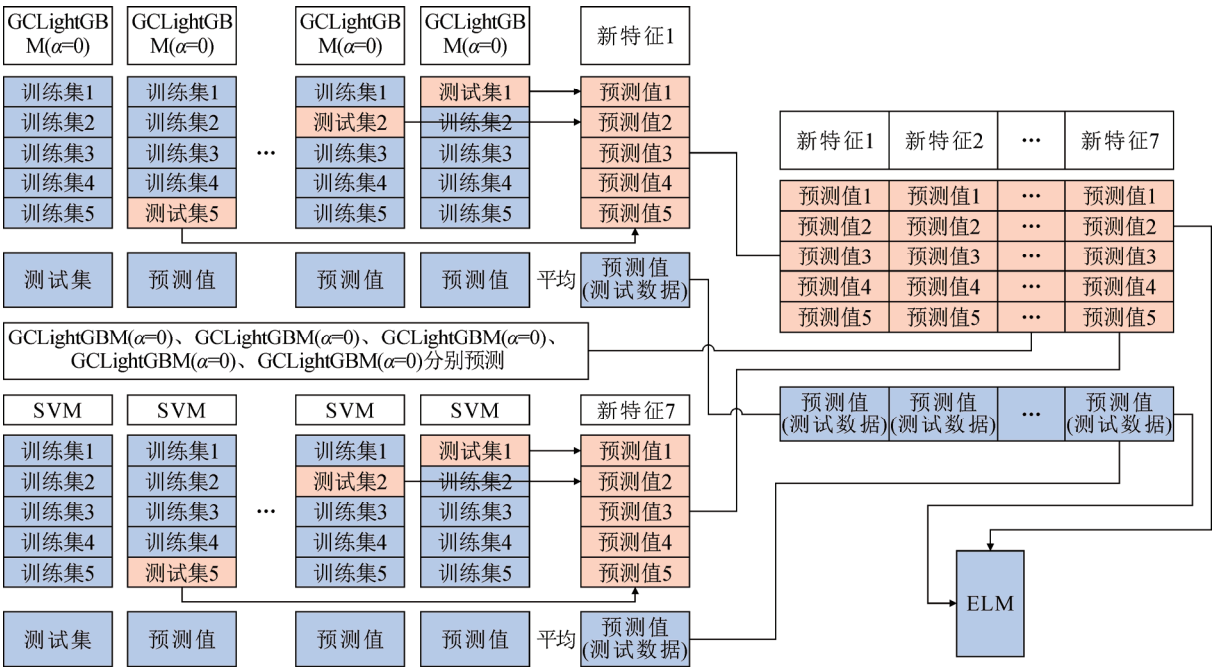


图 A　GCLightGBM 的混合集成模型框架图

Fig.A　Hybrid integrated model framework for GCLightGBM