DOI: 10.19666/j.rlfd.202306123

基于 LightGBM-VIF-MIC-SFS 的风电机组故障诊断输入特征选择方法

马良玉[1]，程东炎[1]，梁书源[1]，耿妍竹[1]，段新会[1,2]
（1.华北电力大学自动化系，河北　保定　071003；
2.保定华仿科技股份有限公司，河北　保定　071000）

［摘　　要］针对风电机组数据采集与监视控制（SCADA）系统数据维数较高、特征冗余、特征相关性高导致风电机组的故障诊断过程存在误差大、分类正确率低的问题，提出一种基于 LightGBM-VIF-MIC-SFS 的三段式特征选择方法。首先，根据 LightGBM 实现对所有特征的重要性计算，确定初步特征空间；其次，根据方差膨胀因子（VIF）和最大信息系数（MIC）构建相关性判别阵，据此评估一次筛选中重要性相近的特征，舍弃相似性高的输入特征；最后，使用序列前向搜索法对特征进行第 3 次处理，逐个输入前 2 次特征选择获得的特征，保留能提升系统性能的特征，从而实现最终特征的选取。在完成了模型的建立后，使用风电场真实 SCADA 系统数据进行性能评估，将所提方法与 2 种对比算法在 6 个数据集上进行对比，结果显示所提出的 LightGBM-VIF-MIC-SFS 相较 2 种对比特征选择算法有显著优势。对所提方法内部的 3 个模块进行了消融实验，有效验证了所提特征选取方法内部各个模块的有效性以及基于所提方法得到的最优特征空间的合理性及准确性。

［关　键　词］风电机组；特征选择；LightGBM；方差膨胀因子；最大信息系数；序列前向搜索

［引用本文格式］马良玉，程东炎，梁书源，等. 基于 LightGBM-VIF-MIC-SFS 的风电机组故障诊断输入特征选择方法[J]. 热力发电, 2024, 53(1): 154-164.　MA Liangyu, CHENG Dongyan, LIANG Shuyuan, et al. Input feature selection method for wind turbine fault diagnosis based on LightGBM-VIF-MIC-SFS[J]. Thermal Power Generation, 2024, 53(1): 154-164.

Input feature selection method for wind turbine fault diagnosis based on LightGBM-VIF-MIC-SFS

MA Liangyu[1], CHENG Dongyan[1], LIANG Shuyuan[1], GENG Yanzhu[1], DUAN Xinhui[1,2]
(1.Department of Automation, North China Electric Power University, Baoding 071003, China;
2.Baoding Huafang Technology Co., Ltd., Baoding 071000, China)

Abstract: In order to solve the problems of high error and low classification accuracy in the fault diagnosis process of wind turbines caused by the high dimension, feature redundancy and feature correlation of wind turbine supervisory control and data acquisition (SCADA) data, a three-stage feature selection method based on LightGBM-VIF-MIC-SFS is proposed. Firstly, based on the importance calculation of all features implemented by LightGBM, a preliminary feature space is determined. Secondly, a correlation discriminant matrix is constructed based on the variance inflation factor (VIF) and maximum information coefficient (MIC) to evaluate features with similar importance in a single screening, and discard input features with high similarity. Finally, the sequential forward search method is used to process the features for the third time, input the features obtained from the previous two feature selection one by one, and retain the features that can improve the system performance, so as to achieve the final feature selection. After the establishment of the model, the real SCADA data of the wind farm is used for performance evaluation, and the proposed algorithm is compared with the two comparison algorithms on six data sets. The results show that LightGBM-VIF-MIC-SFS has significant advantages over the two comparison feature selection algorithms. A ablation experiment was conducted on the three modules within the proposed algorithm,

收 稿 日 期：2023-06-23
基 金 项 目：河北省中央引导地方科技发展资金项目（226Z2103G）
Supported by：Hebei Province Central Leading Local Science and Technology Development Fund Project (226Z2103G)
第一作者简介：马良玉（1972），男，教授，硕士生导师，主要研究方向为人工智能在电站建模、控制和故障诊断中的应用，maliangyu@ncepu.edu.cn。
通信作者简介：程东炎（1997），男，硕士研究生，主要研究方向为风电机组状态监测及故障诊断，172313538@qq.com。

effectively verifying the effectiveness of each module within the proposed feature selection method and the rationality and accuracy of the optimal feature space obtained based on the proposed method.
Key words: wind turbine; feature selection; LightGBM; variance inflation factor; maximum information coefficient; sequence forward search

伴随着“双碳”目标的提出，风能作为一种清洁能源得到了广泛的应用。风电机组常年在条件恶劣的环境下运行，导致其故障率高于其他机电设备。风电机组相关故障及时有效的识别，对减少严重故障发生和降低风电场运行及维护费用有重要意义。

目前国内外研究人员在风电机组异常工况故障预警和故障诊断方面已进行了较多研究。文献[1]提出一种基于门控循环单元（gate recurrent unit，GRU）神经网络注意力机制的故障诊断方法，通过对数据参量中关联关系的挖掘，提高模型的精度。文献[2]提出了用最小角回归方法来对特征向量进行选择，并采用修正交互式多模型（hidden Markov model，HMM）建立故障诊断模型，提升了诊断精度。文献[3]针对基于单一模型、单一视角特征的齿轮箱故障诊断准确率相对低的问题，提出基于比例冲突分配规则的模型融合故障诊断方法。文献[4]提出一种基于快速密度峰值聚类（clustering by fast search and find of density peaks，CFSFDP）和 LightGBM 模型结合的风电机组异常状态监测方法。文献[5]提出一种双重改进的完全噪声辅助聚合经验模态分解（improved complete ensemble empirical mode decomposition with adaptive noise，ICEEMDAN）、主成分分析（principal component analysis，PCA）、GRU 神经网络的风电机组齿轮箱故障预警方法。文献[6]通过相关性分析实现数据降维，以简化径向基函数构建神经网络结构实现故障预警。文献[7]针对目前风机故障种类多、知识关联关系复杂、推理效率低等问题提出了风电机组故障知识的获取表达与推理框架。文献[8]提出一种主轴总成窜动在线监测技术，并开发了相应的监测装置和预警系统，实现了对风电机组主轴总成窜动位移的监测、预警。文献[9]针对风电机组激光修复主轴断裂的现象，进行了多种实验及分析，实现对故障原因的查明。

针对特征提取问题，国内外研究人员常使用数据采集与监视控制（supervisory control and data acquisition，SCADA）系统的海量数据，利用统计学习、机器学习等方法开展研究。文献[10]针对风力发电机组轴承故障振动信号，提出一种自适应经验小波变换（adaptive empirical wavelet transform，AEWT）与奇异值分解（singular value decomposition，SVD）的特征提取方法，并结合核极限学习机（kernel extreme learning machine，KELM）实现风电机组轴承的故障诊断。文献[11]提出一种基于深度特征融合网络的行星齿轮箱故障诊断方法。文献[12]提出基于关联度与自检验长短时记忆（self-checking long short-term memory，Sc-LSTM）神经网络的轴承寿命预测模型。以上研究均表明，数据和特征决定了机器学习效果的上限，选择不同模型的过程也是对这个上限的靠近过程，特征选择对于风电机组的故障预警和诊断模型的效果至关重要。

特征选择主要包含 2 个目标：1）减少特征数量、降低数据维数、增强模型的泛化能力，减少过拟合；2）去掉无助于判别的数据，消除其对模型的不良影响，提高模型的精度。

特征选择方法共有过滤法、封装法和嵌入法3类。过滤法通过计算待选变量和目标变量间的发散性和相关性来进行特征变量选择，皮尔森相关系数（Pearson correlation coefficient，PCC）[13]、互信息（mutual information，MI）[14]能评价非线性关系。卡方检验可以检验定性自变量和定性因变量间的相关性[15]。过滤法的计算过程相对简单，缺少对选择结果的验证部分，特征选择后的准确性有待研究。

封装法根据目标函数（预测或诊断评分函数）的评估结果进行特征选择，该方法需要对特征组合进行穷举，以分类器的准确率作为评价指标进行评估。递归特征消除法（recursive feature elimination，REF）是典型的封装法，通过反复构建模型，选出最优（差）特征，重复这个过程，直到历遍所有特征[16]。在高维的数据集下，该方法需要对大量特征组合进行穷举，运算量较大。SCADA 数据的特征高达上百种，不适合采用封装法。

嵌入法将特征选择嵌入模型的训练过程中，通过对基于机器学习算法的模型进行训练，得到各个特征的权重系数，根据系数大小进行特征选择。随机森林（random forest，RF）[17]、支持向量机（support vector machine，SVM）[18]是典型的嵌入法，凭借其

优良性能已广泛应用于风电机组的特征选择中。

本文在以往研究的基础上，提出一种应用于风电机组 SCADA 数据的三段式特征选择方法，用于实现风电机组故障诊断模型建立过程中的特征选择。此方法共进行 3 次特征筛选，适用于多输入特征、分类及回归模型。

1）用轻量级梯度提升机（light gradient boosting machine，LightGBM）[19]中的特征重要性计算方法，实现对特征重要性评估，计算出所有输入特征的重要性，删除重要性较低的特征。

2）构建特征关系评估矩阵，评估一次筛选中评估矩阵内部相对高的值所对应的 2 个特征，舍弃相似性高的 2 个特征中重要性相对低的输入特征。

3）基于一次和二次筛选选取的特征，使用前向搜索法对特征进行第 3 次处理，保留能提升模型精度的特征，实现最终特征的选取。

1 研究方法

1.1 LightGBM 算法

轻量级梯度提升机是由微软在 2017 年发布的一种基于梯度提升决策树（gradient boosting decision tree，GBDT）[20]的改进迭代提升树系统，引入了独立特征合并算法和单侧采样算法对 GBDT 进行优化。LightGBM 算法采用单边梯度下降（gradient-based one-side sampling，GOSS）和互斥特征捆绑（exclusive feature bundling，EFB）2 种独创技术使得算法更加高效。

不同特征对目标参数的贡献度有优劣之分，随着特征数目的增加，特征子集的冗余量会随之增加，这不仅会导致算法的计算量增加，还会对目标的识别效果产生消极影响，降低模型的性能。在 LightGBM 中，度量特征属性重要性的原理基于训练过程中每个特征被使用的频次（即特征的分裂次数）以及该特征对预测结果的影响程度 2 个方面。LightGBM 通过计算某个特征的平均信息增益来度量特征的重要性，特征对模型的影响越大，信息增益越大。

在计算特征重要性时，LightGBM 与重要性相关 2 个可选参数分别是 gain 和 split。其中，gain 指使用某个特征进行分裂后，所有子节点上的增益之和，即损失函数降低的程度。它反映了特征对模型目标函数变化的贡献度。gain 越大，表示该特征相对于其他特征来说更加重要。而 split 指使用某个特征进行分裂时，该特征被用了多少次作为分割点。它反映了该特征在模型中参与建模的次数。split 越大，表示该特征在构建决策树时参与的贡献度越高。

1.2 方差膨胀因子法和最大信息系数

1.2.1 方差膨胀因子法

方差膨胀因子（variance inflation factor，VIF）法[21]是一种用于评价模型输入变量之间线性关系的方法。它的基本思想是计算每个特征与其他特征的相关性，并使用一个方差膨胀因子值来表示每个特征的相关性强度。方差膨胀因子的计算方法是将每个特征作为因变量，其他特征作为自变量来拟合一个线性回归模型，然后计算自变量和因变量的均方误差比值，即：

$$\boldsymbol{V}_{\mathrm{IF}}=\frac{1}{1-R^2} \tag{1}$$

式中：R^2 为通过简单线性回归模型拟合自变量与其他自变量之间关系的拟合优度。

1.2.2 最大信息系数

最大信息系数（maximum information coefficient，MIC）[22]是一种用来测量 2 个变量之间关系的方法，它能够捕捉到多种形式的关系，包括非线性关系、周期性关系等。同时，MIC 还适用于包含噪声的数据，因此被广泛应用于统计学、数据挖掘、机器学习等领域。MIC 值越大，说明 2 个变量之间的关系越密切。

1.3 序列前向选择

序列前向选择（sequential forward selection，SFS）[23]是一种简单的贪心算法。SFS 确定一个空的特征子集 X，每次迭代选择一个特征 f 加入特征子集，评估性能指标，重复此步骤直到历遍所有特征。

SFS 方法存在局限性，如果特征存在相似性较高或相互依赖的特征，则会造成特征子集的冗余。即如果特征 A 依赖于特征 B、C，在加入特征 A 后再加入特征 B、C 会出现特征冗余的情况（特征子集内出现了冗余子集 A）。

2 基于 LightGBM-VIF-MIC-SFS 的特征选择模型

2.1 模型建立流程

基于 LightGBM-VIF-MIC-SFS 的特征选取流程如图 1 所示。

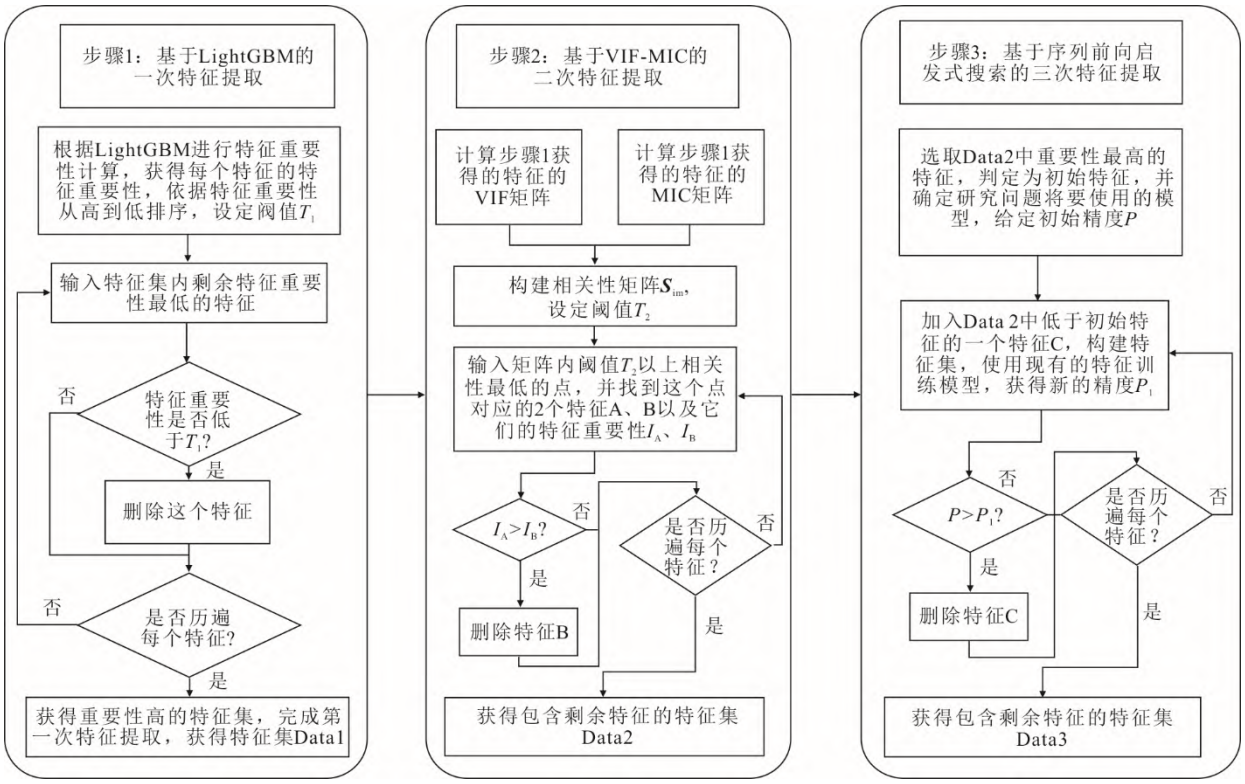

图 1 LightGBM−VIF−MIC−SFS 特征选取模型流程

Fig.1 Process of LightGBM−VIF−MIC−SFS feature selection model

由图 1 可见，LightGBM-VIF-MIC-SFS 特征选取模型流程详细步骤如下：

步骤 1　基于 LightGBM 的初次特征筛选。用 LightGBM 的 importance 函数计算特征重要性，可选用 LightGBM 特征重要性计算中的‘split’模块。LightGBM 中使用了一些随机技术，如随机数种子、子采样、随机特征选择等，这些随机技术可以提高模型的泛化能力，但也会导致模型随机性增加，从而每次训练的结果有所不同。所以基于 LightGBM 的特征重要性计算具有随机性，为保证结果的可靠性，求取输入特征重要性 N 次并计算均值作为最终特征重要性。设定阈值 T_1，删除特征重要性较低（低于 T_1）的特征，保留特征重要性相对高的特征进行下一步筛选。

步骤 2　基于相关性矩阵的二次特征筛选。进行相似性度量时，采用结合 VIF 和 MIC 的相关性矩阵对一次筛选获得的特征进行相关性分析，可以有效衡量特征间线性和非线性关系。设定阈值 T_2，删除特征相关性较高（高于 T_2）的 2 个特征中重要性相对低的特征，保留重要性相对高的特征进行下一步筛选。

二次特征筛选根据特征相似性矩阵，保证筛选后保留的特征具有更高的独特性和区分性。这也是特征选择的核心原则：保留对结果有最大贡献但重复度最低的特征，从而提高模型的泛化能力和性能。同时可以克服步骤 3 中 SFS 算法的局限性（当特征间存在依赖性时会产生特征冗余）。

步骤 3　SFS 对特征进行第 3 次筛选。

1）首先初始化特征子集，选取重要性最高的特征，判定为初始特征。

2）加入第 2 次特征筛选获得的比第 1 个特征重要性低的特征，若加入后系统评价值提升，将此特征加入模型，若评价值未提升，不加入此特征。

3）重复 2），直至历遍二次筛选获得的所有特征。返回获得的特征子集作为整个特征选择模型的最终特征子集。

2.2 输入特征重要性求取次数确定

特征重要性计算是三段式特征选择模型中第 1 段的关键，由于 LightGBM 特征重要性计算模块采用了随机划分节点、随机抽样的方法增加模型的泛化能力，为保证训练结果的有效性，采用训练 N 次求平均值的方法。训练次数 N 的增加会提升模型求取特征重要性的稳定性，但 N 过高会降低模型的运行速度。

对于训练次数 N 的确定，使用包含齿轮箱油池温度高于上限值、齿轮箱入口压力低于下限值、齿轮箱油池油位故障共 3 类故障、94 个特征的风电机组齿轮箱故障诊断数据集进行研究，与这 3 类故障相关性最高的特征分别是齿轮箱油池温度和齿轮箱油路入口油压，绘制特征重要性值前 12 个特征及其重要性的柱状图，结果如图 2—图 5 所示。

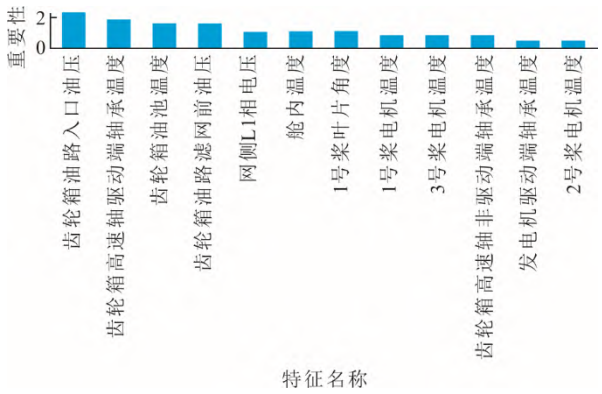

图 2 特征重要性计算 5 次求均值

Fig.2 Feature importance by 5 times calculation to find the mean

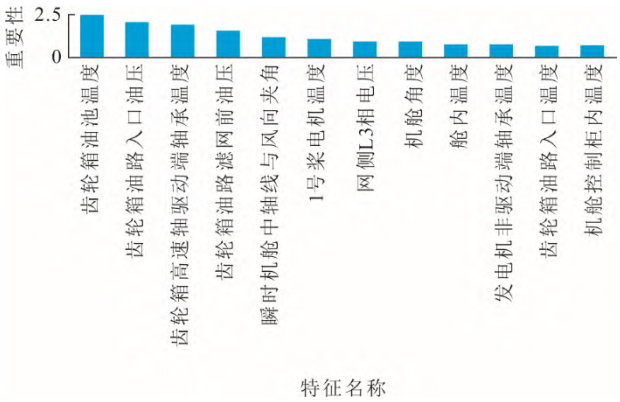

图 3 特征重要性计算 10 次求均值

Fig.3 Feature importance by 10 times calculation to find the mean

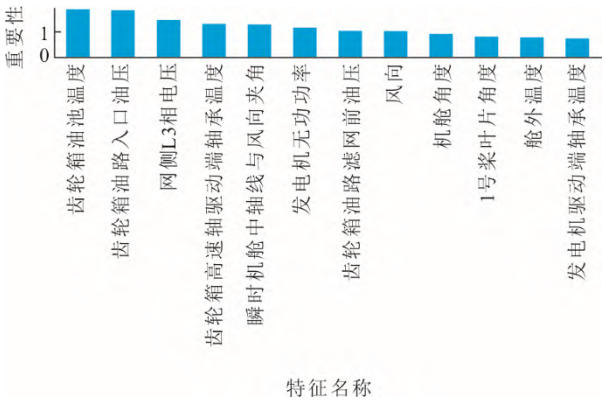

图 4 特征重要性计算 30 次求均值

Fig.4 Feature importance by 30 times calculation to find the mean

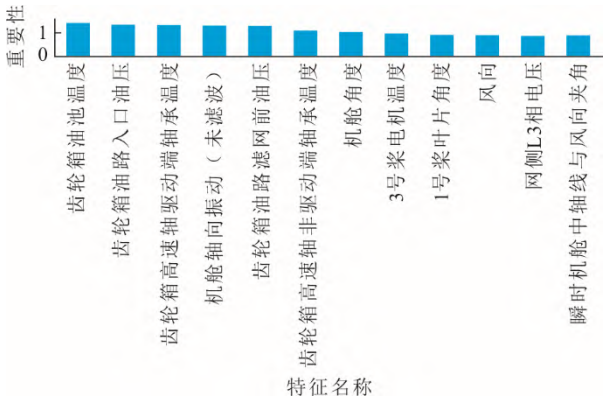

图 5 特征重要性计算 50 次求均值

Fig.5 Feature importance by 50 times calculation to find the mean

观察图 2—图 5 可发现：在特征重要性计算次数达到 10 次以上时，可以稳定获得与研究故障相关性最高的特征。伴随着计算次数增加，相同特征的重要性数值可能会有变化，但其大小不影响模型的使用。不同特征间大小关系会影响模型最后选取的特征类型。图 3—图 5 可以稳定获得齿轮箱相关故障最重要的 2 个特征：齿轮箱油池温度和齿轮箱油路入口油压，且大小顺序一致，故选用 N=10 作为特征重要性训练次数。

2.3 相关系数权重确定

相关系数 $\boldsymbol{S}_{\mathrm{im}}$ 由 $\boldsymbol{V}_{\mathrm{IF}}$ 和 $\boldsymbol{M}_{\mathrm{IC}}$ 构成，相关系数矩阵的计算公式如下：

$$\boldsymbol{S}_{\mathrm{im}}(X_1,X_2)=a*\boldsymbol{V}_{\mathrm{IF}}(X_1,X_2)+b*\boldsymbol{M}_{\mathrm{IC}}(X_1,X_2) \quad (2)$$

式中：X_1、X_2 分别为需要衡量相关性的 2 个特征；a 为线性相关性判别矩阵 $\boldsymbol{V}_{\mathrm{IF}}$ 的权重；b 为非线性相关性判别矩阵 $\boldsymbol{M}_{\mathrm{IC}}$ 的权重。

衡量线性相关性的 $\boldsymbol{V}_{\mathrm{IF}}$ 的取值范围是[1,+∞]，其中 1 表示没有多重共线性，多重共线性的程度越高，$\boldsymbol{V}_{\mathrm{IF}}$ 值越大。一般而言，$\boldsymbol{V}_{\mathrm{IF}}$ 值在[1,5]时认为特征间存在轻微共线性，$\boldsymbol{V}_{\mathrm{IF}}$ 值在(5,10]时认为特征间存在中度的共线性，$\boldsymbol{V}_{\mathrm{IF}}$ 值大于 10 则认为特征间存在严重的共线性。

衡量非线性相关性的 $\boldsymbol{M}_{\mathrm{IC}}$ 的取值范围是[0,1]。如果 2 个特征之间存在强烈的非线性关系，则 $\boldsymbol{M}_{\mathrm{IC}}$ 值可能更接近 1，反之，则 $\boldsymbol{M}_{\mathrm{IC}}$ 值较低。

为了均衡相关系数 $\boldsymbol{S}_{\mathrm{im}}$ 的线性判别能力和非线性判别能力，$\boldsymbol{S}_{\mathrm{im}}$ 矩阵内部的 $\boldsymbol{V}_{\mathrm{IF}}$ 参数的权重 a 设定为 0.05，$\boldsymbol{M}_{\mathrm{IC}}$ 参数的权重 b 设定为 0.5，使得权重加入后 $\boldsymbol{V}_{\mathrm{IF}}$、$\boldsymbol{M}_{\mathrm{IC}}$ 对于 $\boldsymbol{S}_{\mathrm{im}}$ 的贡献尽量各自保持在[0,0.5]，即模型对于线性关系和非线性关系有着均衡的衡量效果，$\boldsymbol{S}_{\mathrm{im}}$ 值尽量维持在[0,1]，在 2 个特征参数相关性过大

时会出现少数 S_{im} 值大于 1 的情况，由于特征筛选阈值 T_2 会设定在[0,1]，当特征相关性参数大于 1 时，会被特征筛选阈值 T_2 成功识别并筛选，所以这部分的数据不会影响模型特征选取性能。

2.4 阈值的设定

步骤 1 中根据特征重要性筛选和步骤 2 中根据相关系数矩阵 S_{im} 筛选的本质是通过阈值 T_1、T_2 的设定实现特征筛选，阈值 T_1 和 T_2 的选择将影响模型最终的选择结果。T_1 为特征重要性筛选阈值，步骤 1 删除的特征数量随 T_1 增大而增多；T_2 为相关性判别阈值，步骤 2 中如果 2 个特征相关系数值大于 T_2，则对这 2 个特征重要性进行比较，删除二者间特征重要性相对低的特征，所以步骤 2 删除的特征数量随 T_2 的增大而减小。

2.4.1 阈值 T_1 的选取

对于 T_1 的选取，需要根据 LightGBM 计算得出的特征重要性值域进行区间划分，并对每个区间内特征的个数进行分析。选用风电机组齿轮箱故障数据集和风电机组变桨系统故障数据集进行特征重要性计算。

由于步骤 1 中筛选是根据 T_1 删除特征重要性相对低的特征，需要关注重要性相对低的特征的数量。选定适量的低重要性特征进行去除，保留相对多的特征，以便进行后续特征选择。

选用 94、62、20 个特征的齿轮箱故障数据集和变桨故障数据集，通过特征重要性计算模块得出特征重要性值，划分 0~0.1、0.1~0.3、0.3~0.4、0.4~0.6、0.6~0.8、0.8~1.0、＞1.0 共 7 个重要性区间，绘制不同区间重要性特征个数的柱状图，结果如图 6、图 7 所示。

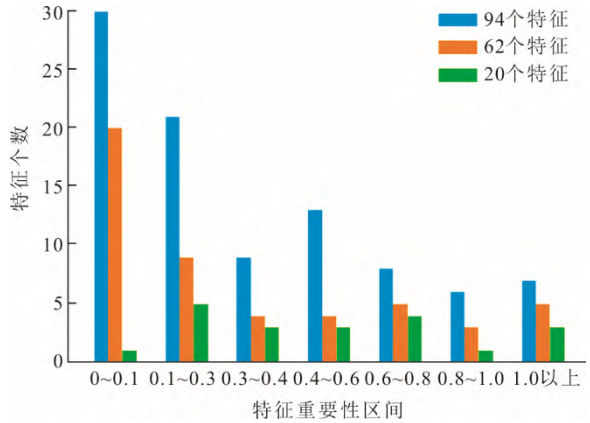

图 6 齿轮箱故障不同特征重要性区间特征个数

Fig.6 Features number within the different importance intervals of gear box faults

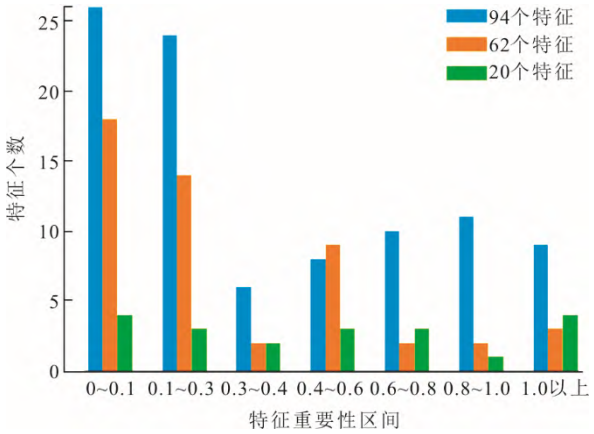

图 7 变桨系统故障不同特征重要性区间特征个数

Fig.7 Features number within the different importance intervals of pitch system faults

由图 6、图 7 可知，经过特征重要性计算后，较多的低重要性特征集中在 0~0.3，此区间具有最多的特征量，故选取 0.3 作为 T_1，在模型特征选择过程中删除重要性低于 0.3 的特征，保留重要性在 0.3 以上的特征作为步骤 2 特征筛选模块的输入特征集。

2.4.2 阈值 T_2 的选取

阈值 T_2 出现在步骤 2 特征选择模型的第 2 阶段中，在第 1 阶段基于特征重要性实现初次筛选后，应用相关性阈值 T_2 对筛选后的特征进行二次筛选。由 2.3 节可知，相关性阈值 T_2 在[0,1]。对于 T_2 的具体取值，使用初次特征筛选后的 94 特征的齿轮箱故障数据集进行研究。为了便于观察，绘制特征重要性靠前的 14 个特征，结果如图 8 所示。

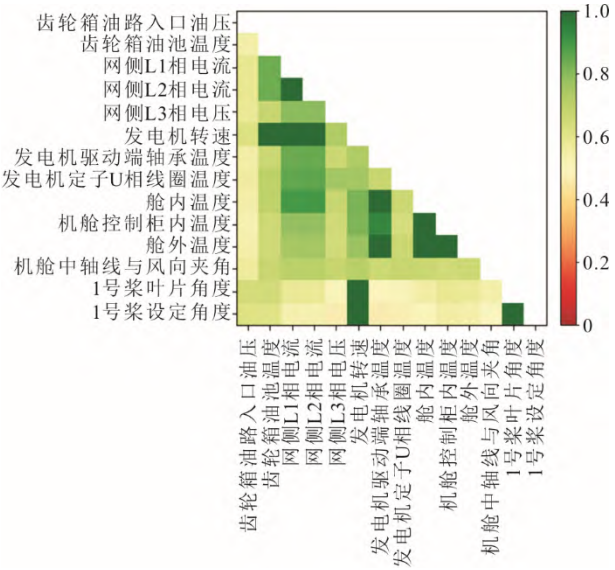

图 8 数据集经初次筛选后部分特征相关性热力图

Fig.8 Partial feature correlation thermal map of data set after initial screening

由图 8 可知，经过初次筛选后，获得的特征数据集内部特征间的相关性多在 0.4~0.7，此部分的特征间相关性相对低，可以进行保留；特征间相关性在 0.7 以上的特征组合相对少，且这部分的相关性高，有必要对这部分的特征组合进行分析与筛选，故选定阈值 T_2 为 0.7。在 2 个特征间相关性高于 T_2 时判定 2 个特征为高相关性特征，此时根据特征重要性矩阵对 2 个特征重要性进行比较，删除低重要性特征，保留高重要性特征，对所有相关性高于阈值 T_2 的特征完成删减后，将剩余的特征子集作为步骤 3 的输入特征子集。

2.5 超重要特征的设定

在风电机组实际研究过程中，会存在部分学者想要研究某些特定特征对于某些问题的影响。若仅使用特征重要性对特征进行评估，可能会出现步骤 1 中评估重要性过低或步骤 2 中进行特征重要性对比时重要性相对低，导致研究需要保留的特征被误删除的情况。为避免这种情况的出现，在 LightGBM-VIF-MIC-SFS 特征集成选取模型中引入超重要特征的设定。当想要研究某一特定特征时，可在模型输入此特征名称，将其设定为超重要特征，在模型的三段式筛选的过程中将提高此特征的重要性。在保证精度的情况下，实现此特征在模型筛选过程的保留以及对于此特征有依赖的特征处理。当学者不需要研究特定特征时，可以不输入任何特征，模型将按照原有三段式筛选进行特征选择。

3 特征选择实例

3.1 数据集组成与对比算法

本文选取 3 个风场时间跨度为 1 年的预处理后风电机组数据进行研究，数据采样间隔为 1 min，数据集中包含风速、发电机转速、有功功率、给定桨距角、实际桨距角、偏航角度、舱内温度、齿轮箱油路入口油压、齿轮箱油温等诸多特征参数。选用风电机组齿轮箱故障数据集和变桨系统故障数据集对算法进行测试。风电机组故障分别包含齿轮箱油池温度高于上限值、齿轮箱入口压力低于下限值、齿轮箱油池油位故障共 3 种齿轮箱系统相关的故障数据，以及某一桨叶 90º 位置传感器故障重复出现、某一变桨电机堵转、某一桨叶 3º 位置传感器故障重复出现共 3 种风电机组变桨系统相关的故障。表 1 为这些数据集的名称、样本数以及特征数。

表 1 数据集详细信息

Tab.1 Specific information of the selected data sets

序号	数据集	特征数/个
1	风场 1 风电机组齿轮箱故障	94
2	风场 1 风电机组变桨系统故障	94
3	风场 2 风电机组齿轮箱故障	62
4	风场 2 风电机组变桨系统故障	62
5	风场 3 风电机组齿轮箱故障	20
6	风场 3 风电机组变桨系统故障	20

本文风电场 1 的 SCADA 数据共有 94 个特征。风电机组齿轮箱故障数据集包含共计 9 365 组数据，变桨系统故障数据集包含共计 2 912 组数据。

风电场 2 的 SCADA 数据共有 62 个特征。风电机组齿轮箱故障数据集包含共计 3 509 组数据，该风场变桨系统故障数据集包含共计 4 048 组数据。

风电场 3 的 SCADA 数据共有 20 个特征。风电机组齿轮箱故障数据集包含共计 6 686 组数据，该风场变桨系统故障数据集包含共计 2 543 组数据。

为验证本文所提出的 LightGBM-VIF- MIC-SFS 方法的优越性，引入 2 种近年来具有代表性的特征选择方法（引力搜索特征选择算法（gravitational search algorithm，GSA）[24]和蚁狮优化（ant lion optimization, ALO）算法[25]）作为对比。

3.2 本文方法与 GSA 和 ALO 算法的对比实验

本文所提算法 LightGBM-VIF-MIC-SFS 与 2 种对比算法的参数设置见表 2。数据集划分方法均采用 80%作为训练集，20%作为测试集。

表 2 实验算法参数设置

Tab.2 Experimental algorithm parameter settings

算法	项目	数值
GSA	最大迭代次数(Max_iter)、种群大小（Pop_size）	100、50
ALO	最大迭代次数(Max_iter) 、蚂蚁数量（Num_ants）、蚁狮数量（Num_Antlions）、蚂蚁运动能力参数(Alpha)、蚁狮感知范围参数（Beta）、蚁狮蚂蚁适应度权重参数（Gamma）	100、10、4、0.5、0.8
LightGBM-VIF-MIC-SFS	特征重要性求取次数 N、$\boldsymbol{V}_{\mathrm{IF}}$ 参数的权重 a、$\boldsymbol{M}_{\mathrm{IC}}$ 参数的权重 b、特征重要性阈值 T_1、特征相关性阈值 T_2	10、0.05、0.5、0.3、0.7

特征选择模型的目标是筛选出对风电机组故障诊断最有效的特征子集，属于故障诊断中的数据预处理阶段，在完成数据预处理后，需要将处理好的数据输入故障诊断模型，最终实现风电机组故障

诊断，特性选择的最终目标是为故障诊断服务，通过提供更精简、更有效的数据集来提升故障诊断的效果。为了更直观地表现特征选择模型的效果，在采用特征选择模型选好特征后，使用选好的特征数据集进行故障诊断，用故障诊断评价指标进行评估。

选取 XGBoost 回归模型作为诊断模型，均方根误差（root mean square error，δ_{RMSE}）、特征维数缩减率（dimensionality reduction，δ_{DR}）作为评价指标。计算公式如下：

$$\delta_{\mathrm{RMSE}}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-f(x_i))^2} \tag{3}$$

$$\delta_{\mathrm{DR}}=1-N_{\mathrm{SF}}/N_{\mathrm{AF}} \tag{4}$$

式中：m 为样本数；y_i 为样本真实值（故障诊断特征选择前使用的样本包含故障类型的数字标签，这个值即为样本对应的真实值）；$f(x_i)$为使用选择后的特征数据集，应用故障诊断模型（XGBoost 回归模型）进行故障诊断后模型对于故障的判定值。N_{SF} 为选择的特征数；N_{AF} 为特征总数。

表 3、表 4、表 5 给出 6 个数据集上 LightGBM-VIF- MIC-SFS 与其余算法在分类准确率、特征维数缩减率和运行时长上的对比结果。红色数据为最优结果。为提高结果的可靠性，每个 XGBoost 模型均训练 5 次，取 5 次结果的平均值作为最终结果。

首先，观察表 3 的均方根误差结果可以发现，LightGBM-VIF-MIC-SFS 在 6 个数据集上分类准确率上均达到最高。因此，可以认为 LightGBM-VIF-MIC-SFS 特征选择方法在多维数据集特征选择分类效果上有着较好的性能。

观察表 4 中的特征维数缩减率可以发现，LightGBM-VIF-MIC-SFS 在 5 个数据集上维数缩减率达到最高，在未达到最高的数据集 5 上低于最优特征选择方法 GSA，但在该数据集 5 上的均方根误差远低于 GSA 特征选择方法，仍具有实用性。进一步观察可发现，在中等维数和高维的数据集上，LightGBM-VIF-MIC-SFS 特征选择方法可以在维持较高的特征维数缩减率的同时，获得较小的均方根误差。这再次证明了本文中所提的 LightGBM- VIF-MIC-SFS 在风电机组相关故障诊断特征选择问题方面的优越性。

观察表 5 中的运行时长，可以发现本文所提的 LightGBM-VIF-MIC-SFS 特征选择模型在高、中、低维数的数据集的训练速度均优于其余 2 种算法，且有着显著优势，具有良好的快速性。

表 3 3 种算法在均方根误差方面的对比
Tab.3 Comparison of the three algorithms in δ_{RMSE}

数据集	LightGBM-VIF-MIC-SFS	GSA	ALO
1	0.188	0.195	0.359
2	0.334	0.338	0.346
3	0.188	0.326	0.350
4	0.332	0.377	0.361
5	0.295	0.422	0.341
6	0.579	0.786	0.615

表 4 3 种算法在特征维数缩减率方面的对比　单位：%
Tab.4 Comparison of the three algorithms in DR

数据集	LightGBM-VIF-MIC-SFS	GSA	ALO
1	94.68	67.82	57.57
2	95.65	79.07	52.32
3	92.98	77.55	59.18
4	87.72	82.35	56.86
5	80.00	95.00	44.99
6	83.33	83.33	51.83

表 5 3 种算法在运行时长方面的对比　　单位：s
Tab.5 Comparison of the three algorithms in operation time

数据集	LightGBM-VIF-MIC-SFS	GSA	ALO
1	8.124	37.991	55.726
2	5.156	20.327	42.491
3	4.124	21.129	24.917
4	2.564	8.873	19.661
5	4.756	11.742	9.223
6	2.872	8.504	5.183

3.3 三段组合特征提取方法消融实验及结果分析

为验证三段组合式特征选择方法各个阶段对故障特征提取和降维的有效性和贡献，在风场 1 风电机组齿轮箱故障数据上进行消融实验。为此针对各模块不同组合的模型进行性能对比实验。各模型序号、名称以及在特征维数缩减率、均方根误差和运行时间上的对比见表 6。

表 6 各模型性能对比
Tab.6 Comparison of the models performances

各模型组合（序号:名称）	特征维数缩减率/%	均方根误差	运行时间/s
1:LightGBM	59.55	0.216	0.539
2:LightGBM-VIF-MIC	92.55	0.197	4.450
3:LightGBM-SFS	88.65	0.188	8.539
4:LightGBM-VIF-MIC-SFS	94.68	0.188	8.124

由表 6 可见，LightGBM-VIF-MIC-SFS 在 4 种模型上特征维数缩减率达到最高，具有实用性。步骤 1 中的根据 LightGBM 特征重要性对低重要性特征进行去除，该模块对特征维数缩减有着最高的贡

献率。在降低维数方面，LightGBM-VIF-MIC-SFS 模型内部的 3 个模块均具有一定作用。

观察运行时间可以发现，模型 4 在模型 3 的基础上增加了根据相关性对特征选择的 VIF-MIC 模块，但运行时间却缩短了。出现这种现象的原因为：三段筛选的前 2 个模块（LightGBM 模块、VIF-MIC 模块）不需要调用诊断模型，仅根据输入输出信息进行特征选择，训练时间会相对短；三段筛选的第 3 个模块（SFS 模块），在进行特征筛选的过程中需要引入诊断模型来实现模型训练及均方根误差的计算，耗时较多。这部分的训练若加入的特征较多，会使特征选择时间大大增加。实验中，在均方根误差方面，模型 3、模型 4 均达到了最优值，但在运行时间方面，模型 4 低于模型 3，这是因为在模型 4 中加入了 VIF-MIC 模块，实现了 SFS 模块执行前的输入特征数量的进一步缩减。

进一步观察可以发现，在均方根误差上，模型 3 和模型 4 具有相同的效果，但在特征维数缩减率方面，模型 4 高于模型 3，说明模型 3 得到结果中存在 1.3 节中阐述的冗余特征。VIF-MIC 模块的加入在不降低模型分类性能的前提下，消除了冗余特征，有利于降低诊断模型训练的时间。

上述实验结果表明，LightGBM-VIF-MIC-SFS 组合特征选择方法可以在维持较高的特征维数缩减率的同时更快获得较优的均方根误差，再次证明了该特征选择方法的优越性。

3.4 本文方法在深度学习中的应用

深度学习模型能够通过多层非线性变换，从原始数据中学习到高级抽象的特征表示。在训练过程中，模型会自动发现并利用对目标有用的特征，并对其进行有效的编码和组合。这一过程与本文提出的特征选择类似，因此通常情况下深度学习相关模型是不需要在模型使用前进行显式的特征选择，可以应用原始数据直接进行训练。然而，如果输入数据具有大量冗余或噪声特征，可以进行特征选择以提高模型效率和准确性。

对于本文提出的特征选择模型在深度学习模型方面的实用性，基于风场 1 风电机组齿轮箱故障数据集，使用深度模型中的卷积神经网络分类模型进行分析。共分为 3 个实验组：

实验组 1：卷积神经网络分类模型。

实验组 2：应用三段式特征选择模型对特征进行选择，选取卷积神经网络作为特征选择模型中第三段（序列前向搜索）的诊断模型。

实验组 3：应用三段式特征选择模型前 2 段对特征进行选择，将选择后特征构成的数据集加入卷积神经网络。

为提高结果的可靠性，每个模型均训练 5 次，取 5 次结果的平均值作为最终结果，见表 7。其中运行时间 1 为本文提出的模型部分，运行时间 2 为卷积神经网络部分。模型 1 为使用原始 94 特征数据集进行训练的卷积神经网络模型，模型 2 为三段式特征选择模型将第 3 段序列前向搜索的内部诊断模型替换为卷积神经网络（此部分运行时间 2 为应用卷积神经网络进行序列前向搜索的时间），模型 3 先使用三段式特征选择模型前 3 段进行特征选择，之后将选择后的特征加入卷积神经网络。

表 7 各实验组在准确率、运行时间上的对比
Tab.7 Comparison of the models in accuracy and operation time

各模型组合（序号:名称）	准确率/%	运行时间 1/s	运行时间 2/s	总运行时间/s
1:卷积神经网络	91.05	0	8.726	8.726
2:LightGBM-VIF- MIC-SFS	93.04	1.907	22.650	24.557
3:LightGBM-VIF- MIC-卷积神经网络	92.57	1.999	5.429	7.428

通过对比表 7 中的模型 1 和模型 3 可知，本文的特征选择模型可以与深度学习中的卷积神经网络相结合，虽然卷积神经网络有隐式特征提取的能力，但在特征维数较高、存在较多冗余特征时，如果在使用卷积神经网络前使用本文的特征选择模型前 2 段还可以提升模型的精度和运行速度。这是因为特征选择前 2 段可以用较短的时间实现对特征的初步筛选，实现输入数据特征维数缩减、去除冗余特征的效果，再将缩减后的数据集输入卷积神经网络模型中，最终实现整体运行时间缩短的效果。通过对比表 7 中的模型 2 和模型 3 可知：在应用三段特征选择模型时，可以进一步提升诊断效果，但需要的运行时间会显著增加。快速性的提升经常需要牺牲一些准确性，而追求准确性则可能会增加计算或处理的时间，这 2 种模型有各自的优势，需要使用者根据自己的研究目标进行选择。

4 结 论

1）针对风电机组故障预警及诊断时存在难于从维数较高 SCADA 数据中选择有效特征的问题，提出了基于 LightGBM-VIF-MIC-SFS 的三段式组合特征选择方法。该方法无需指定特征数量，即可获得对

预警和诊断模型效果最佳的特征子集，避免在模型输入特征选取时，出现输入特征太多导致的维数灾难，或输入特征偏少而降低模型效果，有利于高维数据的分析。特征自动选取由于模型的建立不依赖于研究对象的特性，所以可在风电机组其他部件故障诊断或其他类似对象故障诊断研究中发挥作用，具有一定的泛用性。经过与 2 种近年来的特征选择算法在 6 个数据集上的比较实验、模型自身的消融实验以及模型在深度学习相关方法上的应用，验证了 LightGBM-VIF-MIC-SFS 算法的良好性能。

2）本文所提模型的运行速度还有进一步提升的空间：SFS 中每次特征的加入都伴随着模型的一次训练，如果前 2 段特征选择模块选择结束后获得的特征子集的特征量较多，第 3 段会需要较多次数的训练，导致特征选择模型耗费的时间大大增加，若想进一步提升三段式特征选择模型的运行速度，可以考虑提高阈值 T_1 和降低阈值 T_2，以实现效果更强的特征筛选，进而降低第 3 步的特征总数，实现特征选择模型速度的提升。阈值 T_1、T_2 的变化幅度需要基于具体数据集进行求取，这部分的研究需要选择多种类型、数量充足的故障诊断数据进行支持，今后对此会做进一步的研究。

［参 考 文 献］

［1］ 徐进, 汤海宁, 丁显. 基于改进 GRU 的海上风电机组齿轮箱故障诊断[J]. 船舶工程, 2022, 44(9): 167-173.
XU Jin, TANG Haining, DING Xian, et al. Fault diagnosis of offshore wind turbine gear box based on improved GRU[J]. Ship Engineering, 2022, 44(9): 167-173.

［2］ 孙群丽, 周瑛, 刘长良. 基于LARS特征选择的风电机组故障诊断的研究[J]. 可再生能源, 2020, 38(10): 1349-1354.
SUN Qunli, ZHOU Ying, LIU Changliang. Research on fault diagnosis of wind turbines based on LARS feature selection[J]. Renewable Energy, 2020, 38(10): 1349-1354.

［3］ 孙文卿, 邓艾东, 邓敏强, 等. 基于模型融合的风电机组齿轮箱故障诊断[J]. 太阳能学报, 2022, 43(1): 64-72.
SUN Wenqing, DENG Aidong, DENG Minqiang, et al. Fault diagnosis of wind turbine gearbox based on model fusion[J]. Acta Energiae Solaris Sinica, 2022, 43(1): 64-72.

［4］ 马良玉, 袁乃正. 基于 CFSFDP 与 LightGBM 的风电机组异常状态预警研究[J]. 太阳能学报, 2023, 44(5): 401-406.
MA Liangyu, YUAN Naizheng. Research on abnormal condition early warning for wind turbine based on CFSFDP and LightGBM[J]. Acta Energiae Solaris Sinica, 2023, 44(5): 401-406.

［5］ 马永光, 冯勇升. 基于IICEEMDAN-PCA-GRU的风电机组齿轮箱故障预警方法研究[J]. 太阳能学报, 2023, 44(4): 67-73.
MA Yongguang, FENG Yongsheng. Research on fault warning method of wind turbine gearbox based on IICEEMDAN-PCA-GRU[J]. Acta Energiae Solaris Sinica, 2023, 44(4): 67-73.

［6］ 符杨, 周全, 贾锋, 等. 基于 SCADA 数据图形化的海上风电机组故障预测[J]. 中国电机工程学报, 2022, 42(20): 7465-7475.
FU Yang, ZHOU Quan, JIA Feng, et al. Fault prediction of offshore wind turbines based on graphical processing of SCADA data[J]. Proceedings of the CSEE, 2022, 42(20): 7465-7475.

［7］ 朱俊杰, 任鑫, 郝延, 等. 风电机组故障知识的获取表达与推理框架[J]. 热力发电, 2023, 52(3): 73-80.
ZHU Junjie, REN Xin, HAO Yan, et al. Acquisition, expression and reasoning framework of wind turbine fault knowledge[J]. Thermal Power Generation, 2023, 52(3): 73-80.

［8］ 汪臻, 邓巍, 赵勇, 等. 风电机组主轴总成窜动监测与故障预警[J]. 热力发电, 2022, 51(12): 141-148.
WANG Zhen, DENG Wei, ZHAO Yong, et al. Monitoring and fault warning of main shaft assembly runout of wind turbine[J]. Thermal Power Generation, 2022, 51(12): 141-148.

［9］ 史志刚, 冯铁玲, 刘雪峰, 等. 某风电机组主轴断裂原因分析[J]. 热力发电, 2022, 51(12): 186-192.
SHI Zhigang, FENG Tieling, LIU Xuefeng, et al. Cause analysis of main shaft fracture of a wind turbine[J]. Thermal Power Generation, 2022, 51(12): 186-192.

［10］ 齐咏生, 单成成, 高胜利, 等. 基于 AEWT-KELM 的风电机组轴承故障诊断策略[J]. 太阳能学报, 2022, 43(8): 281-291.
QI Yongsheng, SHAN Chengcheng, GAO Shengli, et al. Fault diagnosis strategy of wind turbines bearing based on AEWT-KELM[J]. Acta Energiae Solaris Sinica, 2022, 43(8): 281-291.

［11］ 李东东, 赵阳, 赵耀, 等. 基于深度特征融合网络的风电机组行星齿轮箱故障诊断方法[J]. 电力系统保护与控制, 2022, 50(10): 1-10.
LI Dongdong, ZHAO Yang, ZHAO Yao, et al. A fault diagnosis method for a wind turbine planetary gear box based on a deep feature fusion network[J]. Power System Protection and Control, 2022, 50(10): 1-10.

［12］ 兰孝升, 李云凤, 苏元浩, 等. 基于关联度与自检验长短期记忆网络的风电机组轴承寿命预测模型[J].高电压技术, 2023, 49(6): 2652-2661.
LAN Xiaosheng, LI Yunfeng, SU Yuanhao，et al. Wind turbine bearing life prediction model based on indexed relation and self-checking long short-term memory[J]. High Voltage Engineering, 2023, 49(6): 2652-2661.

［13］ 刘灏, 商峻, 毕天姝, 等. 基于实测数据的电网频率信号特征分析与提取方法[J]. 电力系统自动化, 2023, 47(10): 135-144.
LIU Hao, SHANG Jun, BI Tianshu, et al. Feature analysis and extraction method of power grid frequency signal based on measured data[J]. Automation of Electric Power Systems, 2023, 47(10): 135-144.

［14］ 曾祥军, 冯琛, 杨明, 等. 考虑运行状态相似性的风电机组数据异常检测方法[J]. 电力系统自动化, 2022, 46(11): 170-180.
ZENG Xiangjun, FENG Chen, YANG Ming, et al. Data anomaly detection method for wind turbines considering operation state similarity[J]. Automation of Electric Power Systems, 2022, 46(11): 170-180.

［15］ 甄志龙, 张居晓. 卡方统计中基于 KL 散度的高维文

本数据特征筛选[J]. 统计与决策,2022,38(17): 43-46.
ZHEN Zhilong, ZHANG Juxiao. Feature screening for high dimensional text data based on kl divergence in chi-squared statistics[J]. Statistics & Decision, 2022, 38(17): 43-46.
[16] 刘献礼, 秦怡源, 岳彩旭, 等. 递归特征消除与极端随机树在铣刀磨损监测中的研究[J]. 机械科学与技术, 2023, 42(6): 821-828.
LIU Xianli, QIN Yiyuan, YUE Caixu, et al. Research on recursive feature elimination and extra trees in milling cutter wear monitoring[J]. Mechanical Science and Technology for Aerospace Engineering, 2023, 42(6): 821-828.
[17] 李汪繁, 丁先, 方晶剑. 基于 GWO-RF 的凝汽器真空预测方法[J]. 动力工程学报, 2023, 43(4): 436-442.
LI Wangfan, DING Xian, FANG Jingjian. Prediction method of condenser vacuum based on GWO-RF[J]. Journal of Chinese Society of Power Engineering, 2023, 43(4): 436-442.
[18] 彭道刚, 姬传晟, 涂煊, 等. 基于 LSTM-SVM 的燃气轮机压气机故障预警研究[J]. 动力工程学报, 2021, 41(5): 394-399.
PENG Daogang, JI Chuansheng, TU Xuan, et al. Research on gas turbine compressor fault early warning based on LSTM-SVM[J]. Journal of Chinese Society of Power Engineering, 2021, 41(5): 394-399.
[19] 贾凯, 江明, 袁啸林, 等. 基于代价敏感型 LightGBM 的分子泵故障检测[J]. 电子测量与仪器学报, 2022, 36(10): 55-64.
JIA Kai, JIANG Ming, YUAN Xiaolin, et al. Fault detection of molecular pump based on cost sensitive LightGBM[J]. Journal of Electronic Measurement and Instrumentation, 2022, 36(10): 55-64.
[20] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.
[21] 朱佳慧, 于丽英. 我国科技创新与金融发展的耦合协同测度——基于 VIF-变异系数的筛选[J]. 上海大学学报(自然科学版), 2021, 27(4): 785-794.
ZHU Jiahui, YU Liying. Coupling synergy measure of sci-tech innovation and financial development in China: screening based on VIF-variation coefficient[J]. Journal of Shanghai University (Natural Science Edition), 2021, 27(4): 785-794.
[22] 崔树银, 汪昕杰. 基于最大信息系数和多目标 Stacking 集成学习的综合能源系统多元负荷预测[J]. 电力自动化设备, 2022, 42(5): 32-39.
CUI Shuyin, WANG Xinjie. Multivariate load forecasting in integrated energy system based on maximal information coefficient and multi-objective Stacking ensemble learning[J]. Electric Power Automation Equipment, 2022, 42(5): 32-39.
[23] 姚锐, 惠萌, 李俊, 等. 基于随机森林的局部放电特征提取和优选研究[J]. 华北电力大学学报(自然科学版), 2021, 48(4): 63-72.
YAO Rui, HUI Meng, LI Jun, et al. Feature extraction and optimal selection based on random forest for partial discharges[J]. Journal of North China Electric Power University (Natural Science Edition), 2021, 48(4): 63-72.
[24] 张雪峰, 杜孝平, 王晓健, 等. 基于引力搜索机制的数据聚类及特征选择算法[J]. 计算机工程与设计, 2021, 42(9): 2536-2544.
ZHANG Xuefeng, DU Xiaoping, WANG Xiaojian, et al. Data clustering and feature selection algorithm based on gravitational search mechanism[J]. Computer Engineering and Design, 2021, 42(9): 2536-2544.
[25] 李庚松, 刘艺, 郑奇斌, 等. 基于多目标混合蚁狮优化的算法选择方法[JL]. 计算机研究与发展, 2023, 60(7): 1533-1550.
LI Gengsong, LIU Yi, ZHENG Qibin, et al. Algorithm selection based on multi-objective hybrid ant lion optimizer[J]. Journal of Computer Research and Development, 2023, 60(7): 1533-1550.

（责任编辑 杜亚勤）