DOI: 10.13382/j.jemi.B2205439

基于代价敏感型 LightGBM 的分子泵故障检测*

贾 凯[1,2] 江 明[1,2] 袁啸林[3] 左桂忠[3] 陈 跃[3]
(1. 安徽工程大学高端装备先进感知与智能控制教育部重点实验室 芜湖 241000;
2. 安徽工程大学电气工程学院 芜湖 241000;3. 中国科学院等离子体物理研究所 合肥 230031)

摘 要:针对 EAST 全超导托卡马克装置的分子泵在数据集不平衡条件下导致的故障识别率低,模型容易过拟合等问题,提出了一种基于时频域分析与改进的 LightGBM 算法相结合的方法。首先,利用在 EAST 搭建的分子泵实验平台采集正常与故障的振动数据,再对数据进行时频域特征提取。其次,通过优化误分类代价,建立了代价敏感型 LightGBM 故障检测架构。最后,将得到的特征量作为代价敏感型 LightGBM 算法的输入,实现分子泵故障检测。经实验验证,该方法的正确率达 99.4%,同时,所提出的方法在误报率和漏检率方面均优于传统分类算法与 LightGBM 算法。此方法能够有效解决模型过拟合问题,实现对分子泵故障的高准确率检测。

关键词: 故障检测;磁悬浮分子泵;时频域分析;LightGBM;真空泄漏
中图分类号: TB752+.27 **文献标识码**: A **国家标准学科分类代码**: 510.40

Fault detection of molecular pump based on cost-sensitive LightGBM

Jia Kai[1,2] Jiang Ming[1,2] Yuan Xiaolin[3] Zuo Guizhong[3] Chen Yue[3]
(1. Key Laboratory of Advanced Perception and Intelligence Control of High-end Equipment, Anhui Polytechnic University, Wuhu 241000, China; 2. School of Electrical Engineering, Anhui Polytechnic University, Wuhu 241000, China; 3. Institute of Plasma Physics, Chinese Academy of Science, Hefei 230031, China)

Abstract: Aiming at the problem of low accuracy and overfitting in the unbalanced data of molecular pump of EAST all-superconducting tokamak device, a method of time-frequency analysis and improved LightGBM algorithm is proposed. Firstly, the normal and fault vibration data are collected by the molecular pump experimental platform. Then, extract the time and frequency domain features. Moreover, the cost-sensitive LightGBM fault detection framework was established by optimizing the misclassification cost function. Finally, the obtained features are used as the input of the cost-sensitive LightGBM algorithm for molecular pump fault detection. The experimental results show that the fault detection accuracy is 99.4%. Meanwhile, the proposed method can consistently outperform traditional classifiers and LightGBM algorithms. This method can effectively solve the problem of overfitting and realize the detection of molecular pump fault with high accuracy.

Keywords: fault detection; magnetic molecular pump; time-frequency domain analysis; LightGBM; vacuum leak

0 引 言

EAST 全超导托卡马克是在长脉冲条件下进行高参数等离子物理实验的装置,并为未来建造稳定,先进的核聚变反应堆奠定了良好的工程与物理基础。分子泵作为 EAST 装置的高真空环境主要获得设备,在长期的核聚变实验中可能会发生真空泄漏等故障,若未能及时处理故障,将对 EAST 装置造成次生灾害[1-3]。因此,为了减少分子泵故障给 EAST 带来的伤害,对分子泵进行故障检测是必要的。

在旋转类机械的故障检测中,振动信号包含了其在

收稿日期: 2022-04-29 Received Date: 2022-04-29
*基金项目:国家自然科学基金(11905254,12105322)项目资助

运行时的丰富信息,因此振动信号也被大量作为旋转类机械故障检测时的模型输入[4-6]。故障检测分为如下几个步骤:1)采集数据;2)对原始数据特征提取;3)利用分类算法对特征进行分类识别。其中特征提取与分类算法对特征数据进行分类识别是极其重要的步骤,直接关系到了故障检测成功与否。特征提取方法通常有基于信号处理的方法与基于深度学习的方法。基于信号处理的特征提取方法包括时频域分析法,小波变换等,主要利用数学方法对原始信号进行处理,从而提取信号的特征[7-8]。基于深度学习的方法一般基于卷积神经网络、自动编码器等[9-10]。基于深度学习的方法虽然可以自动提取特征无需专业的数学知识,但也存在无法解释模型提取到的特征的含义以及不能确定提取的特征是否反应了原始数据的原貌、是否受到噪声严重干扰等问题,更重要的是,基于深度学习的方法使用多层神经网络训练,极大增加了模型训练时间,对于实时的分子泵故障检测来说是不利的。基于时频域分析的方法无需复杂的神经网路训练,可以解决模型训练时间过长的问题。Issam 等[11]利用经验模式分解(empirical mode decomposition,EMD)技术将振动轴承信号分解为有限数量的静止本征模式函数(intrinsic mode function,IMF),再用 FFT 算法计算不同的参数,分析表明,当轴承故障发生时,IMF 会在不同频段上发生变化,从而实现了轴承的特征提取与故障诊断。Cai 等[12]利用广义 S 变换将齿轮振动信号的时频能量分布准确呈现,实现了齿轮的故障诊断。Liu 等[13]通过时频域特征提取方法构建特征向量,利用回归树算法(classification and regression trees,CART)实现了电机轴承的故障诊断。郭远晶等[14]将振动信号 S 变换后圆整得到一个整数矩阵,再构建一个由 S 变换谱的时间和频率组成的二维随机变量,利用整数矩阵的元素值作为二位随机变量各个采样样本的个数,对二维随机变量进行核密度估计,此方法有效抑制了噪声,使特征更明显,从而实现了齿轮箱的故障诊断。王毅等[15]通过时频域统计特征的提取与随机森林相结合,实现了电器的故障电弧检测。机器学习一直是故障检测领域常用的方法,Gao 等[16]阐述了支持向量机(support vector machine,SVM)的缺陷,并提出了基于最小二乘法支持向量机的小波故障诊断法。Zheng 等[17]将一种改进的 AdaBoost-SVM 方法用于小波变换转换器的故障诊断,利用小波变换去除信号噪声,并将故障特征向量作为改进的 AdaBoost-SVM 分类器的输入,实现故障诊断。Zhang 等[18]提出了一种随机森林(randiom forest,RF)与 XGBoost 相结合的风力发电机组故障诊断方法,RF 根据重要性对数据的特征进行排序,XGBoost 为每种故障训练集成分类器,该方法具有良好的抗过拟合效果,在处理多维数据集时比 SVM 有更好的检测效果。Liu 等[19]利用 LightGBM 模型对经过卷积神经网络提取的多源信息特征进行分类,成功地实现了电机的故障诊断。Tang 等[20]提出了相关分析与 LightGBM 结合的风力发电机组变速箱故障检测方法,针对其大量数据,采用最大信息系数分析法进行特征选择,LightGBM 通过贝叶斯优化来选择最佳超参数,从而实现故障诊断,然而在数据集不平衡的条件下,故障诊断的性能还有待提高。

目前大多数数据驱动的机器学习方法都假设正常样本与故障样本数量接近,但在实际工业现场中,正常样本高于故障样本数量,这导致很多机器学习算法在处理不平衡数据集时对多数类的识别率较高,对少数类的识别率较低。分子泵的故障发生时间短,大部分时间处于正常工作状态,故障样本属于小样本。传统的机器学习算法也没有考虑到分类器的误报率与漏检率过高而造成的损失,以基尼系数和信息增益率为优化目标,在基分类器中并没有引入误分类代价指标,因此故障检测的性能不是很好。基于上述问题,本文考虑了误分类代价指标,以平均总代价最小化为优化目标,有效的提高了故障检测的正确率,在 LightGBM 的权值公式中引入代价函数来替代信息增益,使算法在每次迭代更新时都更关注少数类,从而提高在不平衡数据集上的分类性能。

1 时频域分析

基于时频域分析的信号处理方法一般基于数学方法提取时域与频域的统计特征,公式如表 1 所示。表 1 中 M 代表一个周期内采样点数,x_i 表示当前周期内第 i 个数据样本。

本文选取了 12 种时域统计特征,其中包括最大值、最小值、均值、方差、标准差、均方根、峰度、峭度、波形因子、峰值因子、脉冲因子、裕度因子。这些时域数据的特征反应了分子泵在不同工作条件时的状态。

通常仅有时域统计特征往往不能够达到完整体现分子泵在不同条件下的工作状态。而频谱图显示了在一个频率范围内每个给定频带的信号量,其横轴表示频率,纵轴表示该频率信号的幅度。在有故障发生时,故障特征频率处有明显的幅度变化。快速傅里叶变换(fast Fourier transform,FFT)在故障检测中有广泛的应用,FFT 可以将时域信号转换成易于分析的频域信号,FFT 的数学表达式如(1)所示:

$$X(k)=\sum_{n=0}^{N-1}x(n)W_N^{k_n} \tag{1}$$

其中,$k=0,1,\cdots,N-1$,$W_N=e^{-j\frac{2\pi}{N}}$ 是旋转因子,$X(k)$ 表示频域值,$x(n)$ 表示时域采样点,n 代表时域采样点的序列索引,k 代表频域值的索引,N 代表进行转换的采样点数量。将采集到的原始数据进行 FFT 变换后,对其频

域数据进行特征提取,选取了频率均值、频率重心、频率标准差这 3 种频域特征。其数学公式如表 2 所示。表 2 中 L 代表一个周期 FFT 得到的频谱长度,f_j 代表基波对应的频谱幅度。

表 1　时域统计特征

Table 1　Time domain statistical features

时域特征	公式
最大值	$X_{\max} = \max(x_i)$
最小值	$X_{\min} = \min(x_i)$
平均值	$X_{mean} = \frac{1}{M}\sum_{i=1}^{M} x_i$
方差	$X_{var} = \frac{1}{M-1}\sum_{i=1}^{M}(x_i - X_{mean})^2$
标准差	$X_{std} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(x_i - X_{mean})^2}$
均方根	$X_{RMS} = \sqrt{\frac{1}{M}\sum_{i=1}^{M} x_i^2}$
峰度	$X_{skew} = \sum_{i=1}^{M}\frac{(x_i - X_{mean})^3}{(M-1)X_{var}^3}$
峭度	$X_{kurt} = \sum_{i=1}^{M}\left(\frac{(x_i - X_{mean})^4}{(M-1)X_{var}^4}\right)$
波形因子	$X_{sf} = \frac{X_{RMS}}{X_{mean}}$
峰值因子	$X_{cf} = \frac{X_{\max}}{X_{RMS}}$
脉冲因子	$X_{pf} = \frac{X_{\max}}{X_{mean}}$
裕度因子	$X_{mf} = \frac{X_{\max}}{\left(\frac{1}{M}\sum_{i=1}^{M}\sqrt{\lvert x_i\rvert}\right)}$

表 2　频域统计特征

Table 2　Frequency domain statistical features

频域特征	公式
频谱均值	$F_{mean} = \frac{1}{L}\sum_{j=1}^{L} f_j$
频谱重心	$F_{jc} = \frac{\sum_{j=1}^{L}\bar{f}_j f_j}{2\pi\sum_{j=1}^{L} f_j^2}$
频谱标准差	$F_{std} = \sqrt{\frac{\sum_{j=1}^{L}(f_j - F_{mean})^2}{L}}$

2　改进的 LightGBM 算法

2.1　LightGBM 算法原理

轻型梯度提升机是一种基于梯度单边采样和独特特征捆绑而提出的决策树算法,优化方向为损失函数负梯度方向[21]。LightGBM 算法在处理高维大数据方面更有效率,这是因为 LightGBM 中的独特特征捆绑(EFB)算法和基于梯度的单侧采样(GOSS)算法。GOSS 方法引入了一个具有恒定乘数和小梯度的数据实例,可以从数据集中抽取与原始数据具有相同分布和特征的数据,在提高分类速度的同时保证了分类精度。EFB 方法是使两个特征形成一个新的特征,这样可以减少数据样本。GDBT 算法使用梯度下降法来近似拟合每个决策树,在训练过程中,每一轮训练都是在上一轮训练结果与残差叠加后继续拟合,每次迭代时新的决策树会沿着损失函数减少最快的方向减少,这个方向就是损失函数负梯度减少的方向,这会使预测的精度更高。虽然 GBDT 算法已经具备了较好的准确性,但仍然存在效率低下、可扩展性差等问题,因为 GBDT 算法需要遍历所有数据的信息增益才能找到合适的分割点。为了解决上述问题,LightGBM 算法主要采取了两种策略来优化。

1)LightGBM 算法选择了具有深度限制的按叶生长策略。GBDT 算法选择了按层生长策略,按层生长策略的优点在于数据可以同时对同一层的叶子进行分割,减少过拟合。但是,由于按层生长策略不加区分的对同一层级的叶子进行处理,导致了处理难度增大,算法效率随之降低。采用按叶生长策略的 LightGBM 算法可以从当前所有的叶子中找到分裂信息增益最大的叶子,然后再依次进行分裂和循环,这种方法减少了训练带来的误差,提高了算法的分类精度。但是,这种方法容易造成树的深度过大,容易造成过拟合,因此 LightGBM 在训练过程中通过限制树的最大深度来预防过拟合问题。图 1 表示了按层生长策略模式,图 2 表示了按叶生长策略模式。

2)LightGBM 算法选择了直方图算法搜索最优分割点。GBDT 算法选择了预排序方法,预排序法在排序后的特征上枚举所有特征点,再根据信息增益搜索最优分割点,此方法效率低下。而 LightGBM 采用了直方图算法,以浮点数据为例,直方图算法的核心是将连续的特征值离散成 p 个整数,并构建宽度为 p 的直方图,再将离散值作为索引进行累计统计,在连续遍历数据集后,根据直方图累积的统计量确定最优分割点。由于直方图的数量比数据的数量少的多,所以直方图算法寻找最优分割点所需的时间和占用的内存比预排序方法更小。图 3 显示了直方图算法的核心思路。

假设一个由 N 个样本组成的训练集 Q,$Q=\{(x_i,y_i)\}_{i=1}^{N}$,其中 $x_i \in X=\{x_1,x_2,\cdots,x_k\}$ 代表数据,X 表示 k 维向量空间,$y_i \in Y=\{0,1\}$ 代表类别标签,$y_i=1$ 表示故障样本。LightGBM 算法的目的是找到一个映射关系 $\bar{G}(x)$ 来近似函数 $G(x)$,从而使损失函数 $\phi(y,G(x))$ 最小。目标函数可以表示为:

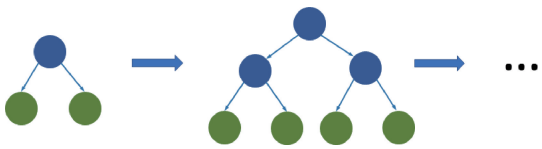


图 1 按层生长策略

Fig. 1 Level-wise tree growth

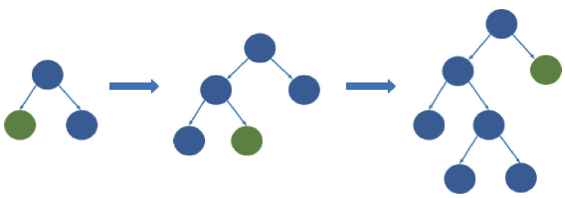


图 2 按叶生长策略

Fig. 2 Leaf-wise tree growth

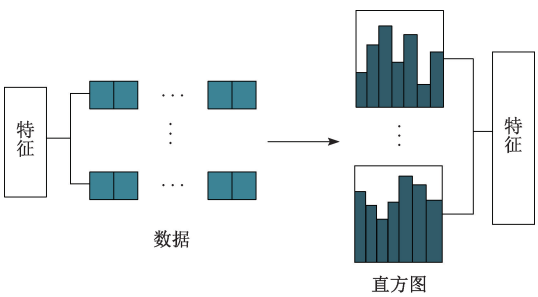


图 3 直方图算法

Fig. 3 Histogram algorithm

$$O^{(t)} = \sum_{i=1}^{n} \phi(y_i, F_{t-1}(x_i) + f_t(x_i)) + \sum_k \Omega(f_k) \quad (2)$$

其中，$\phi(y_i, F_{t-1}(x_i)+f_t(x_i))$是损失函数，$\Omega(f_k)$表示正则项，不同于 GBDT 的快速下降法，LightGBM 使用牛顿法快速逼近目标函数，式(2)可推导为：

$$O^{(t)} \cong \sum_{i=1}^{n} \left(g_i f_t(x_i) + \frac{h_i f_t^2(x_i)}{2} \right) + \sum_k \Omega(f_k) \quad (3)$$

其中，g_i 代表一阶损失函数，h_i 代表二阶损失函数。公式如下：

$$g_i = \sigma_{F_{t-1}(x_i)} \phi(y_i, F_{t-1}(x_i)) \quad (4)$$

$$h_i = \sigma^2_{F_{t-1}(x_i)} \phi(y_i, F_{t-1}(x_i)) \quad (5)$$

LightGBM 中的信息增益如下：

$$H = \frac{1}{2} \left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} - \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda} \right] \quad (6)$$

2.2 改进的 LightGBM 算法

传统的 LightGBM 算法在不平衡数据集上的分类效果较差，因此提出了一种成本敏感方法以解决此类问题[22]。即引入误分类代价指标代替信息增益等指标，目的是使平均总成本最小化，以此提高分类算法对小样本的预测能力。引入成本矩阵表示误分类成本，如表 3 所示。

表 3 成本矩阵

Table 3 Cost matrix

类型	C_N	C_F
C_N	$F(C_N, C_N)$	$F(C_N, C_F)$
C_F	$F(C_F, C_N)$	$F(C_F, C_F)$

表 3 中 C_F 表示故障类别，C_N 表示正常类别，$F(C_N, C_N)$表示正常类别被正确分类的成本，$F(C_N, C_F)$表示正常类别错误分类的成本，$F(C_F, C_N)$表示故障类别被错误分类的成本，$F(C_F, C_F)$表示故障类别被正确分类的成本。

给定一个误分类成本矩阵 $\boldsymbol{C}$，若实际类别 j 与预测类别 i 相同，则预测正确。样本 x 的最佳预测结果应该是使期望总样本最小化的类别：

$$Z(C_I \mid x) = \sum P(C_j \mid x) F(C_j, C_i) \quad (7)$$

$P(C_j|x)$是将样本 x 分类为 C_j 的后验概率。

对于二分类问题，LightGBM 的对数损失函数表达式如下：

$$\log L(x_i, y_i) = \frac{-\sum_{i=1}^{N} (y_i \log P(x_i) + (1 - y_i) \log(1 - P(x_i)))}{N} \quad (8)$$

其中，P 表示后验概率。在对数损失函数中将 $P(x_i)$替换为：

$$P(x_i) = \frac{1}{1 + e^{-2\mu(x_i) - 2\tau}} \quad (9)$$

其中，$\mu = \frac{F(C_N, C_F) + F(C_F, C_N)}{2}$，$\tau = \frac{1}{2} \log \frac{F(C_N, C_F)}{F(C_F, C_N)}$，那么代价敏感的对数损失函数可以简化为：

$$CS\log L(x_i, y_i) = \frac{\log \frac{P(c = F \mid x_i) F(C_N, C_F)}{P(c = N \mid x_i) F(C_F, C_N)}}{F(C_N, C_F) + F(C_F, C_N)} \quad (10)$$

其中，$P(c=F|x_i)$代表将样本划分为故障类别的后验概率，$P(c=N|x_i)$代表将样本划分为正常类别的后验概率，显然 $P(c=F|x_i)=1-P(c=N|x_i)$。

根据式(2)，改进的 LightGBM 的目标函数可以写为：

$$O^{(t)} = \sum_{i=1}^{n} \phi(y_i, F_{t-1}(x_i) + f_t(x_i)) + \sum_k \Omega(f_k) \quad (11)$$

其中，ϕ 是损失函数，Ω 是正则项。根据二阶泰勒式展开，目标函数可以写为：

$$O^{(t)} \cong \sum_{i=1}^{n}\left(\phi(y_i,F_{t-1}(x_i)) + g_i f_t(x_i) + \frac{h_i f_t^2(x_i)}{2}\right) + \sum_k \Omega(f_k) \tag{12}$$

一阶损失函数和二阶损失函数如下所示：

$$g_i(x_i) = 2\tau(y - P(x_i)) \tag{13}$$

$$h_i(x_i) = -4\tau^2 P(x_i)(1 - P(x_i)) \tag{14}$$

给定树的结构，得到每个叶子节点的最优权重如下：

$$w_j^* = \frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda} \tag{15}$$

3　分子泵故障检测架构

3.1　故障检测流程

为了最大限度的减少分子泵故障检测在数据集不平衡情况下造成的损失，建立了基于时频域分析与改进的 LightGBM 算法的分子泵故障检测架构。改进的 LightGBM 算法步骤如算法 1 所示。

算法 1　改进的 LightGBM 算法

1. 输入分子泵特征数据集，代价敏感损失函数 $\phi(y,F(x))$，正则项 Ω，迭代次数 T

2. 计算一阶损失函数 $g_i(x_i) = 2\tau(y - P(x_i))$ 的梯度

3. 计算二阶损失函数 $h_i(x_i) = -4\tau^2 P(x_i)(1 - P(x_i))$ 的海森矩阵

4. 通过最大化公式(6)确定树的结构 $q(x)$

5. 通过

$$w_j^* = \mathrm{argmin}_{w_i}\{\sum_{i=1}^{n}[\phi(y_i,F_{t-1}(x_i)) + g_i f_i(x_i) + \frac{h_i f_i^2(x_i)}{2}] + \sum_k \Omega(f_k)\}$$

确定最优的叶子权重

6. $w_j^* = \dfrac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}$

7. 计算式(12)的最优目标函数

8. 输出改进的 LightGBM 模型

如图 4 所示，实线框表示了分子泵故障检测的主要架构，可分为两个部分：离线建模与在线检测。其中离线建模部分包括：数据采集，特征提取以及模型性能评估，其主要流程为：首先将从分子泵实验平台获取的原始振动信号划分为故障数据集和正常数据集，并提取表 1 所示的相应的时域特征，再对原始数据进行 FFT 变换后，提取表 2 所示相应的频域特征，将得到的时域与频域特征组合形成新的数据集后划分为训练集与测试集，再经过标准化处理后，训练集作为代价敏感型 LightGBM 算法的输入，经过误分类代价函数对小样本进行最优权重调整后，得到代价敏感型 LightGBM 模型，最后将测试集导入模型中，进行模型性能评估与验证。在线检测部分流程为：将在线采集的分子泵振动原始数据经特征提取后输入已有的模型中，经过模型判断，输出分类标签与概率，从而得到 EAST 分子泵的运行状态，实现分子泵故障检测。

为了提高模型的分类性能，首先需要对输入数据进行标准化，公式如下所示：

$$y = \frac{x - mean(x)}{var(x)} \tag{16}$$

其中，$mean(x)$ 表示输入数据的均值，$var(x)$ 表示输入数据的方差，缺失值也会对模型的性能产生影响，而处理缺失值的方法通常包括删除和填补法。

3.2　故障实验与分析

EAST 核聚变装置的分子泵使用爱德华公司生产的 XA4503C 型磁悬浮分子泵，磁悬浮分子泵由电机驱动叶轮高速旋转，将气体分子从 EAST 装置内部抽走，从而形成实验所需的高真空环境。在开启分子泵前，需要先开启真空泵将装置内的真空度抽到低于 10 Pa，此时才可以打开分子泵继续抽气。如果在装置真空度远高于 10 Pa 时打开分子泵，这将会使分子泵的电机过载运行，此时分子泵在超负荷情况下运行一段时间，容易造成电机熔毁，从而给 EAST 装置造成次生灾害。设置采样率为 1 KHz，在一次 EAST 实验中，当装置真空度高于 50 Pa 时，提前打开分子泵进行抽气，采集分子泵处于真空泄露状态的故障数据。此外还采集了分子泵在装置真空度低于 10 Pa 时正常运行的数据。

分子泵故障检测实验平台如图 5 所示，图 5(a)中左上方图中实线框标出的是分子泵主体，虚线框标出的是加速度计传感器。右上方图片中实线框标出的是 NI CDAQ 9185 以太网机箱，可控制 C 系列 I/O 模块与外部主机之间的时序同步，完成数据传输等功能，虚线框标出的是 NI 9230 采集板卡，最高采样率 12 800 Hz。右下方图中实线框标出的是上位机界面，数据采集软件为基于 LabVIEW 的测试软件。左下方图片中标出了加速度计传感器放置的方位。图 5(b)中，分子泵振动数据经过加速度传感器与采集板卡采集后，由以太网机箱经过交换机通过 TCP/IP 协议方式传输到上位机进行存储与处理。

分子泵开启时稳定以 24 000 r/min 的转速运行，振动传感器安装在分子泵电机处(3 轴向)全方位的采集分子泵运行时的振动数据。当出现真空泄露故障时，电机过载运行，通过分析振动数据，可以了解分子泵运行状态，故障与正常态的 X 轴向的时域图如图 6 所示，频域图如图 7 所示。由图 6(a)可以看出，正常情况下的分子泵振动幅度较小且平稳，当发生真空泄漏故障时，分子泵的

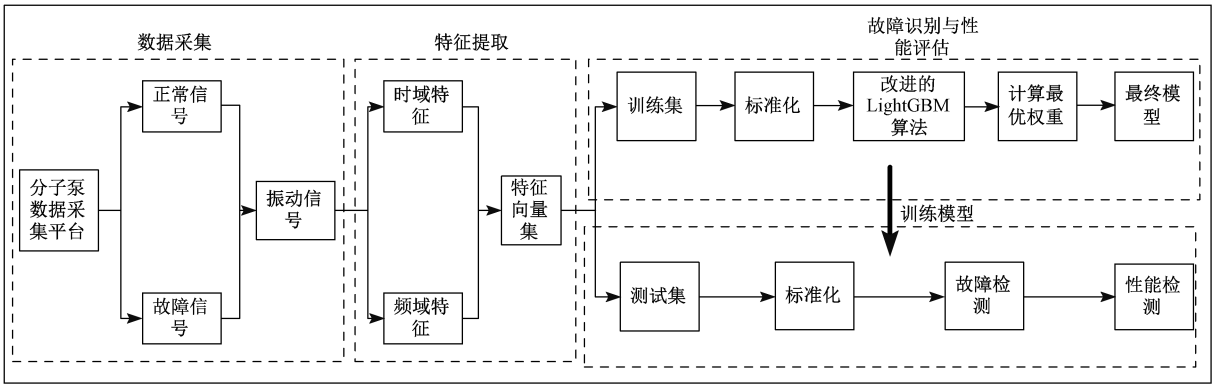
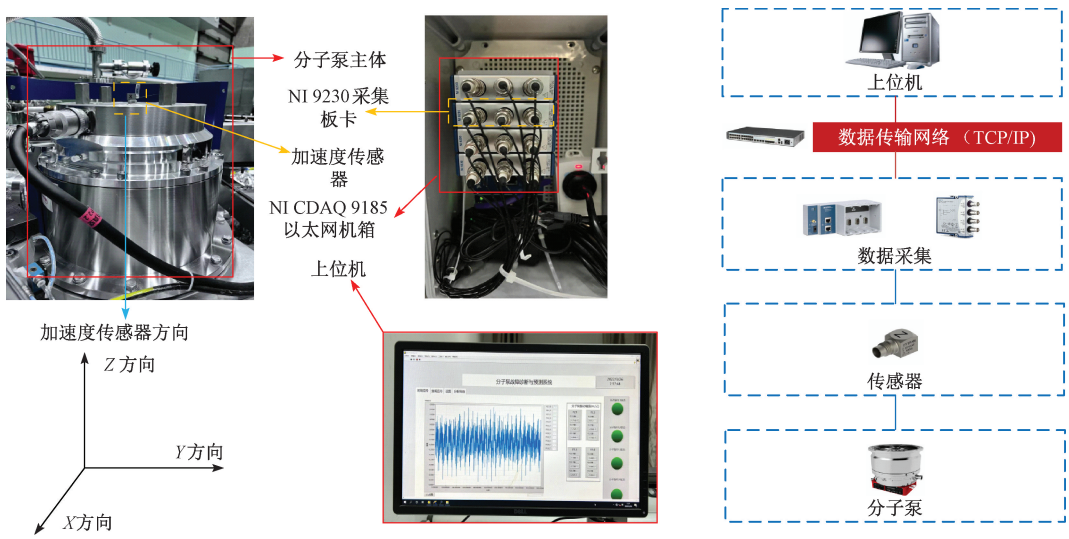


图 4 基于改进的 LightGBM 算法的分子泵故障检测流程

Fig. 4 Molecular pump fault detection procedure of improved LightGBM algorithm



(a) 分子泵故障检测平台
(a) Molecular pump fault detection platform

(b) 主要架构
(b) Main structure

图 5 分子泵故障检测架构

Fig. 5 Molecular pump fault detection framework

振动信号毛刺增加,振动幅度也显著增加,在图 6(b)中,分子泵发生真空泄漏故障时,频谱图在固定频率处有明显的幅度变化。

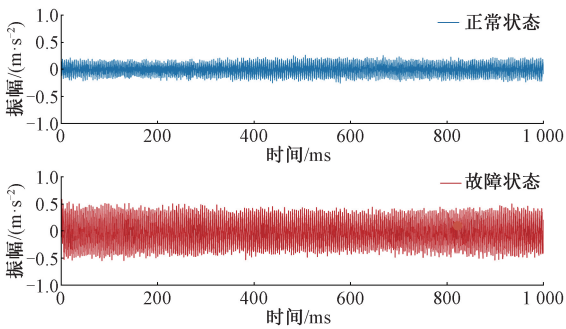


图 6 时域图

Fig. 6 Time domain diagram

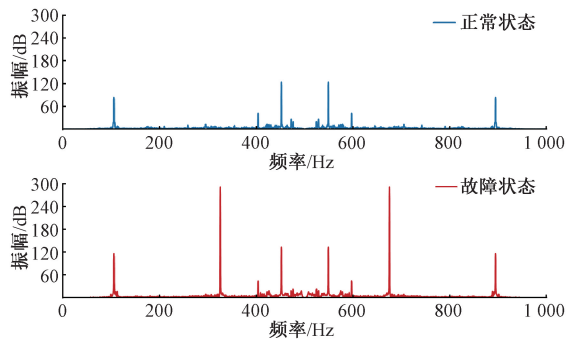


图 7 频域图

Fig. 7 Frequency domain diagram

4　结果与分析

4.1　误报率与漏检率

对采样得到的数据进行分组,选取分子泵正常运行时 16 000 个数据,故障运行时 4 000 个数据,采用交叉取点法,进行特征提取后,最终得到正常类型 1 600 组特征数据,故障类型 400 组特征数据,按照 80%:20%随机划分为训练集与测试集。二分类问题的混淆矩阵如表 4 所示。表 4 中 T_P 代表真阴性,表示被正确识别为正常类别的数量;TN 代表真阳性,表示被正确识别为故障类别的数量;F_N 代表假阴性,表示被错误识别为故障类别的数量;F_P 代表假阳性,表示被错误识别为正常类别的数量。

引入误报率(FAR)和漏检率(MDR)来分别评估分类器误报和漏检的概率。FAR 与 MDR 公式如下:

$$FAR = \frac{F_N}{F_N + T_P} \tag{17}$$

$$MDR = \frac{F_P}{F_P + T_N} \tag{18}$$

表 4　混淆矩阵

Table 4　Confusion matrix

实际分类	预测分类	
	正常	故障
正常	T_P	F_N
故障	F_P	T_N

在同一数据集上对 LightGBM、K-近邻(KNN)和逻辑回归(LR)等算法采用 10 折交叉验证法验证模型,如图 8 与 9 所示。FAR 与 MDR 越小代表算法性能越好,从图 8 和 9 的结果表明,改进的 LightGBM 算法的平均 FAR 与 MDR 分别为 0.002 17 与 0.009 56,LightGBM 的平均 FAR 与 MDR 分别为 0.008 54 与 0.056 09,KNN 的平均 FAR 与 MDR 分别为 0.027 82 与 0.076 9,LR 的平均 FAR 与 MDR 分别为 0.031 42 与 0.105 21,改进的 LightGBM 算法的 FAR 与 MDR 明显低于 LightGBM 算法,也低于 KNN 与 LR 算法,因此改进的 LightGBM 算法的性能优于 LightGBM,KNN 与 LR 算法,其泛化能力也强于 LightGBM,KNN 与 LR 算法。

为了衡量不平衡数据集的分类性能,引入精准率、召回率和 F_1 指标。精准率表示预测的结果中有多少是被正确分类的,召回率表示实际的样本中有多少样本被正确分类,F_1 指标则代表精准率和召回率的调和均值。4 种算法精准率、召回率以及 F_1 指标如表 5 所示。精准率,召回率以及 F_1 指标越大,代表算法对于不平衡数据集的分类效果越好,通过表 5 可以看出,改进的 LightGBM 算法的精准率,召回率以及 F_1 指标在测试集上均大于 LightGBM、KNN 与 LR 算法。LightGBM 算法在故障小样本上的精确率、召回率和 F_1 指标分别低于正常大样本 1.1%、4.7%、2.1%,KNN 算法在故障小样本上的精确率、召回率和 F_1 指标分别低于正常大样本 3.5%、5.2%、4.1%,LR 算法在故障小样本上的精准率,召回率以及 F_1 指标分别低于正常大样本 4.1%、9.5%、7.0%,而改进的 LightGBM 算法在故障小样本上的精准率,召回率以及 F_1 指标分别只低于正常大样本 0.7%、1.2%、1%,对比 LightGBM、KNN 与 LR 算法均有明显提升,因此改进的 LightGBM 算法在不平衡数据集的分类问题中有更好的表现,这说明了本文提出的方法很好地解决了分子泵故障检测中数据集不平衡的问题。从表 5 可以看出改进的 LightGBM 算法的分类准确率达 99.4%,比 LightGBM 算法高出 1.85%,比 KNN 算法高出 3.75%,比 LR 算法高出 5.55%。综合不同算法的 FAR、MDR 与精准率,召回率以及 F_1 指标对比可以看出,改进的 LightGBM 算法解决了分子泵在故障检测中的过拟合问题,提升了对不平衡数据的分类准确性,提高了模型分类的可靠性,因此改进的 LightGBM 方法在分子泵的故障检测中是有效的。

表 5　4 种算法的不同测试结果对比

Table 5　Test result of four different algorithms

分类模型名称	预测状态	准确率/%	精准率/%	召回率/%	F_1 指标/%
代价敏感型 LightGBM	正常	99.40	99.60	99.80	99.80
	故障		98.90	98.60	98.80
LightGBM	正常	97.55	98.7	99.1	98.7
	故障		96.6	94.4	95.6
K-近邻	正常	95.65	97.00	97.10	96.70
	故障		93.50	91.90	92.60
逻辑回归	正常	93.85	94.90	96.50	95.80
	故障		90.80	87.00	88.80

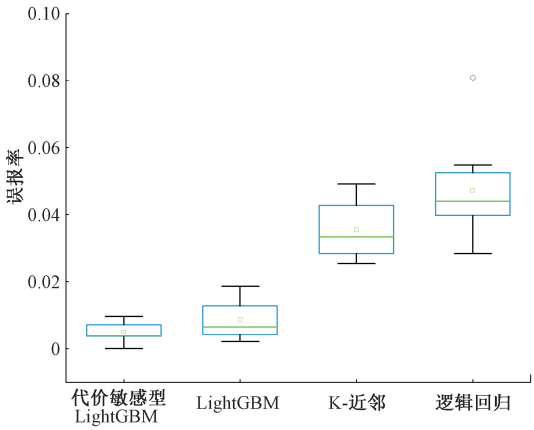


图 8　4 种算法的误报率

Fig. 8　False alarm rate of four algorithms

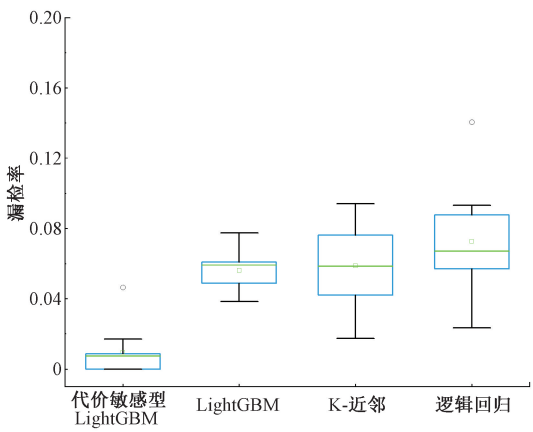


图 9　4 种算法的漏检率

Fig. 9　Missing detection rate of four algorithms

4.2　模型性能

ROC 曲线与 PR 曲线是判定分类器性能的常用指标。ROC 曲线的全称是接受者操作特征曲线，FPR 与 TPR 公式如下：

$$FPR = \frac{F_P}{F_P + T_N} \tag{19}$$

$$TPR = \frac{T_P}{T_P + F_N} \tag{20}$$

ROC 曲线给出的是当阈值变化时，假正率和真正率的变化情况，左下角的点对应的是将所有样例均判为反例的情况，右上角的点对应将所有样例判为正例的情况，对角虚线是参考线。ROC 曲线下的面积 AUC 代表分类器的性能，曲线下面积越大则分类器性能越好，图 10 中给出的是各分类器在测试集上的 ROC 曲线，由上至下分别代表改进的 LightGBM 算法、LightGBM 算法、K-近邻算法和逻辑回归算法，从图 10 可以看出改进的 LightGBM 算法的 AUC 最大，说明其分子泵故障检测正确率最高。

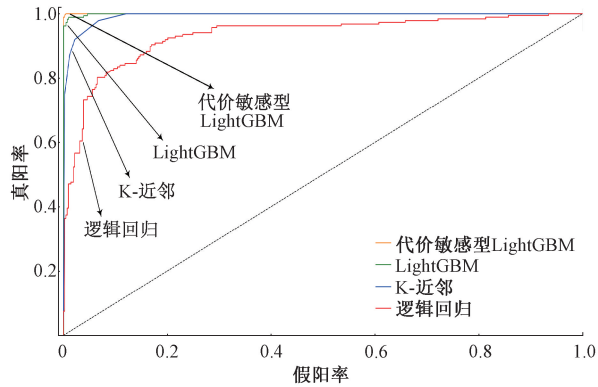


图 10　4 种算法的 ROC 曲线

Fig. 10　RCO curves of four algorithms

PR 曲线是精准率与召回率的比值，曲线上的各个取值是当前阈值下的样本判断结果，且随着阈值的降低，越来越多的样本会被判为正样本，召回率也会不断增加，由于 PR 曲线涉及到精准率的计算，所以更容易受到样本分布的影响，ROC 曲线本质上是正样本或者负样本的召回率计算，不受样本分布的影响。所以 PR 曲线可以更好的评判分类器对不平衡数据集的分类性能。各分类算法的 PR 曲线如图 11 所示，曲线下的面积越大代表分类器对于不平衡数据集的分类性能越好，图 11 中由上到下分别是改进的 LightGBM 算法、LightGBM 算法、K-近邻算法、逻辑回归算法，可以看出传统的分类器对于不平衡数据集的分类效果较差，而改进的 LightGBM 算法在不平衡数据集上的表现优于传统的分类算法，说明本文提出的方法在分子泵故障检测的不平衡数据集中有良好的性能。

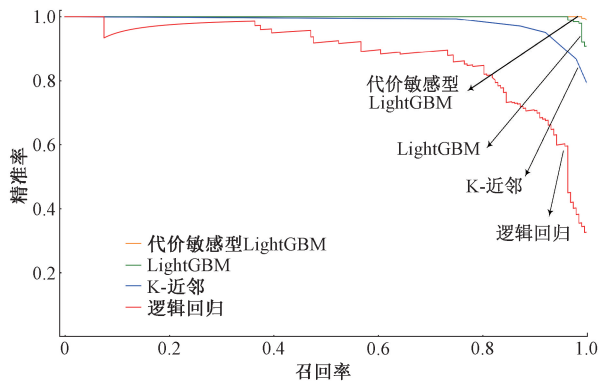


图 11　4 种算法的 PR 曲线

Fig. 11　PR curves of four algorithms

5　结　论

为了保证 EAST 全超导托卡马克装置的真空运行安

全,对真空系统分子泵进行了故障检测,提出了基于振动信号的时频域统计特征提取与代价敏感型 LightGBM 算法相结合的方法实现分子泵的故障检测。这种方法的优点是:1)基于振动信号的时频域特征提取可以快速的提取特征,无需复杂的过程。2)提取的特征可以全面的还原原始数据的全貌,代表了分子泵在正常与故障情况下的机械运行情况。3)引入误分类代价函数替代 LightGBM 算法原有的信息增益,使每种样本都获得了最优的权重。所提出的方法在误报率,漏检率以及精准率,召回率和 F_1 指标上都有良好的表现,解决了在不平衡数据集的情况下 LightGBM 模型容易过拟合的问题。实验表明,本文的方法可以在分子泵故障检测中达到 99.4% 的分类正确率,对不平衡数据具有良好的分类性能,从而全面的实现了 EAST 核聚变装置的分子泵故障检测。

参考文献

[1] 李建刚. 托卡马克研究的现状及发展[J]. 物理, 2016, 45(2):88-97.
LI J G. The status and progress of tokamak research[J]. Physics, 2016, 45(2):88-97.

[2] 袁啸林, 陈跃, 李建刚, 等. 基于实验物理与工业控制系统的 EAST 真空抽气及加料控制系统设计[J]. 核聚变与等离子体物理, 2019, 39(1):34-40.
YUAN X L, CHEN Y, LI J G, et al. Design of vacuum pumping and fueling control system based on experimental physics and industrial control system[J]. Nuclear Fusion and Plasma Physics, 2019, 39(1):34-40.

[3] 李加宏, 胡建生, 王小明, 等. EAST 超导托卡马克装置真空抽气系统[J]. 真空, 2010, 47(1):11-14.
LI J H, HU J SH, WANG X M, et al. Vacuum pumping system for experimental advanced superconducting tokamak (EAST)[J]. VACUUM, 2010, 47(1):11-14.

[4] ZHU J, ASHUTOSH S. Review on engine vibration fault analysis based on data mining [J]. Journal of Vibroengineering, 2021, 23(6):1433-1445.

[5] CRUZ F R G, BALLADO A H, VALMOCENA D P, et al. Vibration signal analysis for generator machine fault detection[C]. 2020 4rd International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM), 2020:205-209.

[6] AYAZ K, SUMAIR A, MUHAMMAD A, et al. An expert system for rotating machine fault detection using vibration signal analysis [J]. Sensors, 2021, 21(22):7587.

[7] 钱小毅, 张宇献. 基于动态特征矩阵的 k 近邻风电机组故障检测方法[J]. 仪器仪表学报, 2019, 40(6):202-212.
QIAN X Y, ZHANG Y X. Fault detection of wind turbines using k-nearest neighbor based on dynamic feature matrix [J]. Chinese Journal of Scientific Instrument, 2019, 40(6):202-212.

[8] 许洁, 胡寿松. 基于 KPCA 和 MKL-SVM 的非线性过程监控与故障诊断[J]. 仪器仪表学报, 2010, 31(11):2428-2433.
XU J, HU SH S. Nonlinear process monitoring and fault diagnosis based on KPCA and MKL-SVM[J]. Chinese Journal of Scientific Instrument, 2010, 31(11):2428-2433.

[9] 张婷, 王海淇, 张认成, 等. 基于自归一化神经网络的电弧故障检测方法[J]. 仪器仪表学报, 2021, 42(3):141-149.
ZHANG T, WANG H Q, ZHANG R CH, et al. An arc fault detection method based on the self-normalized convolutional neural network[J]. Chinese Journal of Scientific Instrument, 2021, 42(3):141-149.

[10] 林鹏飞, 陶继忠. 基于多样性特征和多源信息的分子泵故障诊断[J]. 真空科学与技术学报, 2020, 40(1):33-39.
LIN P F, TAO J ZH. Intelligent fault diagnosis method of turbo-molecular pump: An instrumentation study[J]. Chinese Journal of Vacuum Science and Technology, 2020, 40(1):33-39.

[11] ISSAM A, NADIR F, NADIR B, et al. Multiclass support vector machine based bearing fault detection using vibration signal analysis [C]. International Conference on Electrical Engineering and Control Applications, 2019:885-895.

[12] CAI J H, LI X Q. Gear fault diagnosis based on time-frequency domain de-noising using the generalized S transform[J]. Journal of Vibration and Control, 2018, 24(15):3338-3347.

[13] LIU Z, JU Y L, XIE Z H. Anti-noise motor fault diagnosis method based on decision tree and the feature extraction methods in the time domain and frequency domain [C]. 2021 International Conference on Communications, Information System and Computer Engineering (CISCE), 2021:71-75.

[14] 郭远晶, 魏燕定, 金晓航, 等. 基于 S 变换谱核密度估计的齿轮故障诊断[J]. 仪器仪表学报, 2017, 38(6):1432-1439.
GUO Y J, WEI Y D, JIN X H, et al. Gear fault diagnosis based on kernel density estimation of S transform spectrum[J]. Chinese Journal of Scientific Instrument, 2017, 38(6):1432-1439.

[15] 王毅, 陈进, 李松浓, 等. 基于时频域分析和随机森

林的故障电弧检测[J]. 电子测量与仪器学报, 2021, 35(5):62-68.
WANG Y, CHEN J, LI S N, et al. Arc fault detection based on time and frequency analysis and random forest[J]. Journal of Electronic Measurement and Instrumentation, 2021, 35(5):62-68.

[16] GAO Q W, LIU W Y, TANG B P, et al. A novel wind turbine fault diagnosis method based on integral extension load mean decomposition multiscale entropy and least squares support vector machine[J]. Renewable Energy, 2018, 116:169-175.

[17] ZHENG X X, PENG P. Fault diagnosis of wind power converters based on compressed sensing theory and weight constrained AdaBoost-SVM [J]. Journal of Power Electronics, 2019, 19(2):443-453.

[18] ZHANG D, QIAN L, MAO B, et al. A data-driven design for fault detection of wind turbines using random forests and XGboost [J]. IEEE Access, 2018, 6: 21020-21031.

[19] 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.

[20] TANG M Z, ZHAO Q, WU H W, et al. Cost-sensitive lightGBM-based online fault detection method for wind turbine gearboxes[J]. Frontiers in Energy Research, 2021, 9:1-12.

[21] KE G L, MENG Q, FINLEY T, et al. LightGBM: A highly efficient gradient boosting decision tree [C]. Advances in Neural Information Processing System, 2017:3147-3155.

[22] REN Z J, ZHU Y S, KANG W, et al. Adaptive cost-sensitive learning: Improving the convergence of intelligent diagnosis models under imbalanced data[J]. Knowledge-Based Systems, 2022, 241:108296.

作者简介

贾凯,现为安徽工程大学硕士研究生,主要研究方向为真空设备异常检测与故障诊断。
E-mail: kai. jia@ ipp. ac. cn

Jia Kai is a M. Sc. candidate at Anhui Polytechnic University. His main research interests include vacuum equipment anomaly detection and fault diagnosis.

江明(通信作者),1993 年于上海工业大学(现上海大学)获得硕士学位,现为安徽工程大学教授、硕士生导师,主要研究方向为机器人智能控制系统和先进检测技术。
E-mail: kjjm@ ahpu. edu. cn

Jiang Ming (Corresponding author) received his M. Sc. degree from Shanghai Technology University (now Shanghai University) in 1993. Now he is a professor and M. Sc. supervisor at Anhui Polytechnic University. His main research interests include robotic intelligent control system and advanced detection technology.

袁啸林(通信作者),2018 年于中国科学技术大学获得博士学位,主要研究方向为磁约束聚变装置真空系统研究、系统故障诊断及预测性维护研究。

Yuan Xiaolin (Corresponding author) received his Ph. D. degree from University of Science and Technology of China in 2018. His main research interests include vacuum system of magnetic confinement fusion device and system fault diagnosis and predictive maintenance.