

轴承的剩余使用寿命预测

贾小林, 刘 佳

青岛大学数学与统计学院, 山东 青岛

收稿日期: 2025年2月11日; 录用日期: 2025年3月4日; 发布日期: 2025年3月11日

摘 要

轴承在现代机械设备中具有广泛的应用, 其主要功能是支持旋转或线性运动部件, 减少摩擦, 承载负荷, 确保设备平稳运行。对轴承进行剩余使用寿命预测, 不仅可以减少故障风险和不必要的停机时间, 还能降低维护成本, 提高设备的可靠性与生产效率, 是企业管理设备资产、提升生产水平的重要手段。本文对一轴承的振动频率数据集进行特征提取, 进而分析提取到的特征使用不同的分类方法对轴承的剩余使用寿命进行预测。

关键词

轴承, 预测, 分类器

Prediction of the Remaining Useful Life of the Bearing

Xiaolin Jia, Jia Liu

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Feb. 11th, 2025; accepted: Mar. 4th, 2025; published: Mar. 11th, 2025

Abstract

Bearings have a wide range of applications in modern machinery and equipment, and their main function is to support rotating or linear moving parts, reduce friction, carry loads, and ensure the smooth operation of equipment. Predicting the remaining service life of bearings can not only reduce the risk of failure and unnecessary downtime, but also reduce maintenance costs, improve equipment reliability and production efficiency, and is an important means for enterprises to manage equipment assets and improve production levels. In this paper, the vibration frequency dataset of a bearing is extracted, and then the extracted features are analyzed, and the remaining service life of the bearing is predicted by different classification methods.

Keywords

Bearing, Prediction, Classifiers

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

轴承作为易损伤部件, 并且长期在恶劣的工作环境中高负荷运行, 不可避免地会出现故障, 这会导致经济损失, 引发安全事故[1]。因此, 为了保障现代机械设备的安全平稳运行, 对轴承零件进行剩余使用寿命预测是非常必要的[2]。在本文中, 我们从初始数据中提取特征, 再分别使用 LightGBM 和 Catboost 分类器进行分类, 从而对轴承的剩余使用寿命进行预测。

LightGBM (Light Gradient Boosting Machine)是一种高效的分类器, 是由微软研究院开发, 旨在提高传统 GBDT 模型的训练速度和性能, 特别是在大数据和高维数据场景下[3]。LightGBM 被广泛应用于各种机器学习任务中, 可以用于解决金融风险预测, 欺诈预测, 文本分类, 图像分类, 个性化推荐, 广告点击率预测等各种问题。Catboost (Categorical Boosting)是一个能够高效处理类别数据的分类器, 由俄罗斯 Yandex 公司开发, 基于梯度提升树(Gradient Boosting Tree)算法的实现。可以应用于信用评分预测, 用户推荐系统, 疾病预测, 自然语言处理等众多方面。

2. 预备知识

2.1. 梯度提升决策树算法(Gradient Boosting Decision Trees, GBDT)

梯度提升决策树(GBDT)是以决策树为基础学习器的一种提升算法, 它在每一轮迭代中都建立一个决策树, 使得当前模型的残差在梯度方向上逐步减少, 然后将这个决策树与当前模型进行线性组合得到一个新的模型, 根据这个迭代不断进行循环, 直到决策树数目达到预期的值, 得到最终的学习器[4]。GBDT 经过这一系列的迭代得到了性能更强, 准确性更高, 适用性更强, 鲁棒性更高的学习器。并且在金融领域可以解决信用评分, 反欺诈检测, 风险预测等问题。在自然语言处理领域, 可以解决文本分类, 情感分析等问题。

GBDT 回归算法:

输入: 训练集 $D = \{(x_i, y_i)\}_{i=1}^m$, 其中 $x_i \in X \subseteq R^d, y_i \in Y \subseteq R$, 损失函数 L 。

过程:

1. 初始化模型 $H_0(x)$, 估计使损失函数最小化的常数 γ , 初始模型是只有一个根节点的树。

$$H_0(x) = \arg \min_{\gamma} \sum_{i=1}^m L(y_i, \gamma)$$

2. 对迭代轮次 $t = 1, 2, \dots, T$ 。

对样本 $i = 1, 2, \dots, m$, 计算当前模型的广义残差:

$$r_{it} = - \left[\frac{\partial L(y_i, H(x_i))}{\partial H(x_i)} \right]_{H(x)=H_{t-1}(x)}$$

利用 $(x_i, r_{ii}), i=1, 2, \dots, m$ 拟合回归树, 得到第 t 棵树的叶结点区域 $R_{tj}, j=1, 2, \dots, J$ 。

对每个叶结点区域 $R_{tj}, j=1, 2, \dots, J$, 计算能使区域 R_{tj} 损失函数最小化的最佳预测值 γ_{tj} :

$$\gamma_{tj} = \arg \min_{\gamma} \sum_{x_i \in R_{tj}} L(y_i, H(x_i) + \gamma)$$

得到本轮最佳拟合回归树:

$$h_t(x) = \sum_{j=1}^J \gamma_{tj} I(x \in R_{tj})$$

更新本轮迭代的加法模型:

$$H_t(x) = H_{t-1}(x) + h_t(x) = H_{t-1}(x) + \sum_{j=1}^J \gamma_{tj} I(x \in R_{tj})$$

得到最终的增强学习器:

$$H(x) = H_T(x) = \sum_{t=1}^T \sum_{j=1}^J \gamma_{tj} I(x \in R_{tj})$$

输出: 回归树 $H(x)$ 。

2.2. 网格搜索(Grid Search)

网格搜索是一种用于超参数优化的技术, 广泛应用于机器学习模型的调优过程中。其主要目标是通过遍历给定的超参数空间, 选择最优的超参数组合, 从而提升模型的性能[5]。网格搜索不仅简单而且易于理解, 还可以保证找到超参数空间中所有可能的组合, 从而确保不会遗漏任何最佳参数。

假设有一个模型有两个参数需要进行调参, 分别为 C 和 γ , C 的值可以是[0.1, 1, 10], γ 的值可以是[0.01, 0.1, 1]。我们可以把参数的所有可能取值列出来, 用表格进行表示, 循环过程就是在每个网格里遍历, 搜索。表 1 是参数所有可能出现的组合情况。

Table 1. Possible values for parameters

表 1. 参数可能取值

	$C = 0.1$	$C = 1$	$C = 10$
$\gamma = 0.01$	$C = 0.1, \gamma = 0.01$	$C = 1, \gamma = 0.01$	$C = 10, \gamma = 0.01$
$\gamma = 0.1$	$C = 0.1, \gamma = 0.1$	$C = 1, \gamma = 0.1$	$C = 10, \gamma = 0.1$
$\gamma = 1$	$C = 0.1, \gamma = 1$	$C = 1, \gamma = 1$	$C = 10, \gamma = 1$

3. 方法

本文共介绍了两种方法进行轴承的剩余使用时间预测, 第一可以通过计算均方根(root mean square, RMS)来寻找轴承发生故障的时间节点, 第二可以考虑更多的特征, 计算更多的特征进行整合, 寻找出最重要的几个特征, 使用分类器进行分类从而对轴承的剩余使用时间进行预测。

3.1. RMS 方法

这是一种异常值检测的基线方法, 我们知道加速度的 RMS 是检测轴承故障的一个很好的指标[6], 所以我们可以计算每 15 步长的 RMS 值, RMS 的计算公式为:

$$x_{rms} = \sqrt{\frac{1}{N} \sum_{i=0}^N x_i^2}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=0}^N (x_i - \mu)^2}$$

计算完 RMS 值之后, 我们可以根据 RMS 值的变化趋势确定一个轴承是否健康的阈值, 然后得到阈值内 RMS 值的范围, 超出范围的值我们视为轴承出现了故障。

3.2. 分类器方法

使用原始振动信号并不是我们可以采取的最佳方法。传统观点认为, 在波形中应用信号处理技术, 压缩、分析和解释数据, 并提取有用的信息以用于进一步的诊断和预测。我们可以通过提取时域和频域中的特征来分析波形[7]。

时域特征:

1. 统计矩(均值, 方差, 偏度, 峰度)。
2. 香农熵:

$$H(x) = -\sum_{i=1}^n P(x_i) \log_2 P(x_i)$$

$P(x_i)$ 是事件 x_i 发生的概率。将输入的数据分成 500 份, 计算每个区间中数据出现的频次。

3. 平滑性和均匀性

$$S(f) = \int_a^b \left(\frac{d^n}{dx^n} f(x) \right)^2 dx$$

$$U(x) = \frac{1}{n} \sum_{i=1}^n |x_i - \mu|$$

4. 绝对值, 峰值到峰值

$$P2P = \max(x_1, x_2, \dots, x_n) - \min(x_1, x_2, \dots, x_n)$$

5. 形状因子, 峰值因子, 脉冲因子, 余量因子:

$$\text{形状因子} = \frac{\text{RMS}}{\frac{1}{T} \sum_{i=1}^n |x_i|}$$

$$\text{峰值因子} = \frac{\max(|x_1|, |x_2|, \dots, |x_n|)}{\text{RMS}}$$

$$\text{脉冲因子} = \frac{\max(|x_1|, |x_2|, \dots, |x_n|)}{\frac{1}{T} \sum_{i=1}^n |x_i|}$$

$$\text{余量因子} = \frac{\max(|x_1|, |x_2|, \dots, |x_n|)}{\text{RMS}}$$

6. 自回归系数

频域特征:

1. 机器特征频率下的振动水平;

2. 频率中心;
3. 均方频率和均方根频率。

f_i 是第 i 个频率点的频率, $S(f_i)$ 是频率 f_i 处的功率谱密度值(或振幅的平方)。

$$f_{MS}^2 = \frac{\sum_i f_i^2 S(f_i)}{\sum_i S(f_i)}$$

$$f_{RMS} = \sqrt{f_{MS}^2}$$

4. 频率的方差, 频率的根方差

$$\sigma_f^2 = \frac{\sum_i (f_i - f_{mean})^2 S(f_i)}{\sum_i S(f_i)}$$

5. 谱偏度, 谱峰度, 谱熵

$$\gamma = \frac{\sum_i (f_i - f_{mean})^3 S(f_i)}{\left(\sum_i (f_i - f_{mean})^2 S(f_i) \right)^{\frac{3}{2}}}$$

$$\beta = \frac{\sum_i (f_i - f_{mean})^4 S(f_i)}{\left(\sum_i (f_i - f_{mean})^2 S(f_i) \right)^2}$$

$$H(f) = -\sum_i \frac{S(f_i)}{S_{total}} \log_2 \left(\frac{S(f_i)}{S_{total}} \right)$$

对于时域特征根据计算公式进行计算, 对于频域特征我们使用包装函数来进行特征提取, 以谱密度为例, 首先计算谱密度, 然后从谱中返回统计矩和一些其他的特征, 比如说谱的熵, 信号的功率等, 然后根据转速频率将谱分成三部分, 低于转速频率的是低频区域, 在转速频率和 10 倍转速频率之间的是中频区域, 高于 10 倍转速频率的是高频区域, 最后将这一方法推广到所有频域特征中。我们将提取到的上述特征放在一个包装器中, 然后添加一些标签到数据集中, 有正常、内圈损坏和外圈损坏三个特征, 并且将他们储存在一个文件中。然后训练一个随机森林模型用来提取特征的重要性, 我们得到前 10 的重要特征, 根据这前 10 重要的特征, 分别使用 LightGBM 和 Catboost 分类器对数据进行分类, 从而确定滚动轴承的健康状态。

4. 实例验证

4.1. 数据集

数据集来源 kaggle, 链接为 <https://www.kaggle.com/datasets/vinayak123tyagi/bearing-dataset>。该数据集描述了一个轴承从开始测试到测试失败的实验, 四个轴承安装在一根轴上, 旋转速度恒定在 2000 RPM, 且所有轴承都是强制润滑的。数据集由单独的文件组成, 这些文件是以特定间隔记录的 1 秒振动信号快

照, 每个文件由 20,480 个点组成, 文件名指示数据收集的时间, 我们使用的数据集 1 共有 2156 个文件, 记录了 8 个频道的数据。图 1 是 2003 年 10 月 22 日的一个时刻的八个频道的数据:

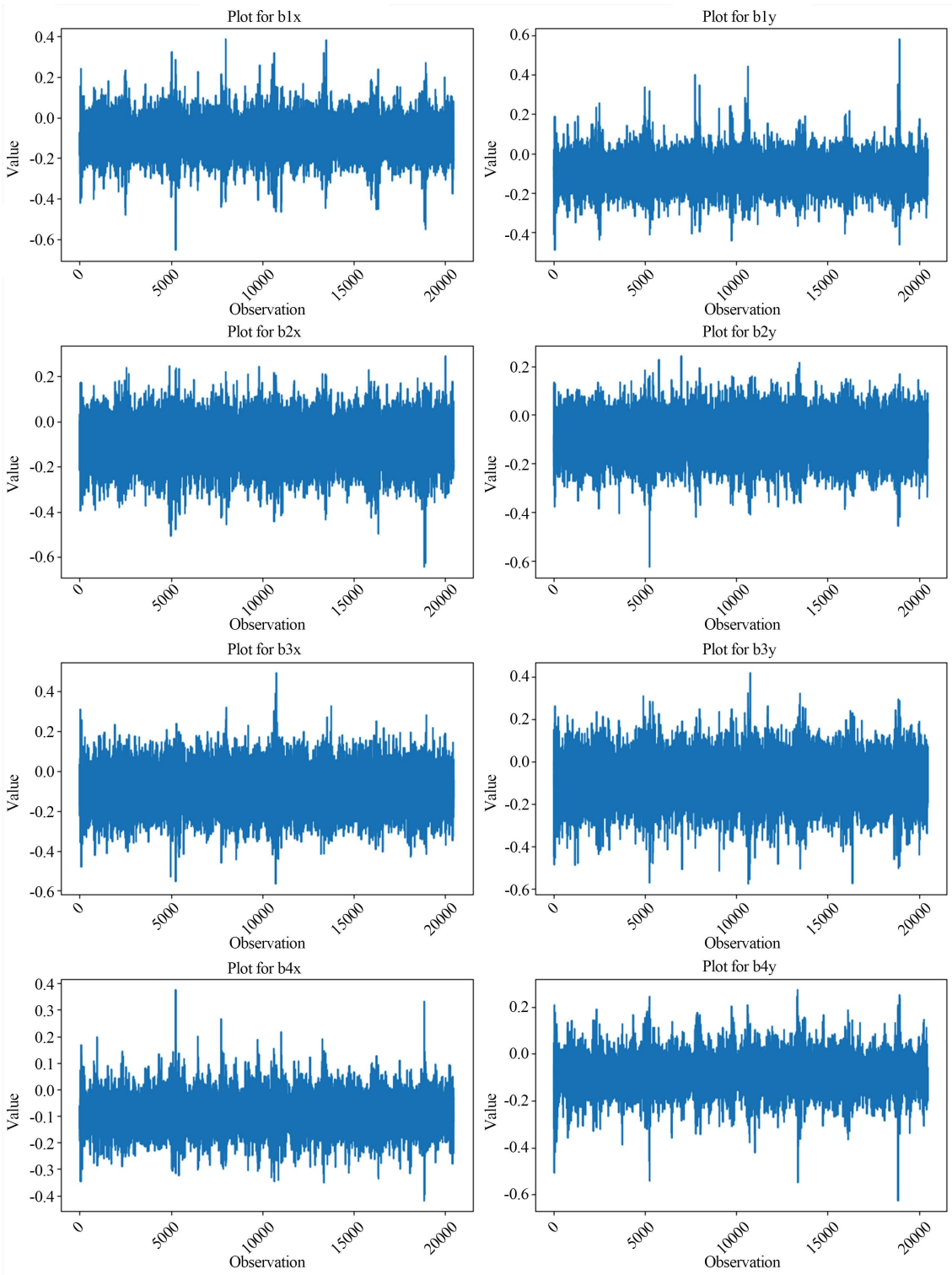


Figure 1. Data for eight channels on 22 October 2003
图 1. 2003 年 10 月 22 日八个频道数据

4.2. 结果

4.2.1. RMS 方法

首先以 15 个数据点为一个区间计算加速度的 RMS 值, 得到的结果如图 2 所示, 由 RMS 值的趋势, 轴承 3 和轴承 4 是非正常的, 由此我们将前 25 天看做正常, 计算前 25 天最大标准差出现的次数, 结果如图 3 所示, 得到最大的 RMS 值为 1.2562, 将这个值设为正常的阈值, 超过这个值的部分我们定义为故障。从图 4 我们看到轴承 4 在 30 天之前(即实验结束前 5 天)可靠地处于检测到的异常值状态。对于轴承 3, 在 23.1 天和 26.1 天后可能会有两次误报, 但在 30.9 天后, 我们得到了我们认为是正确的结果。

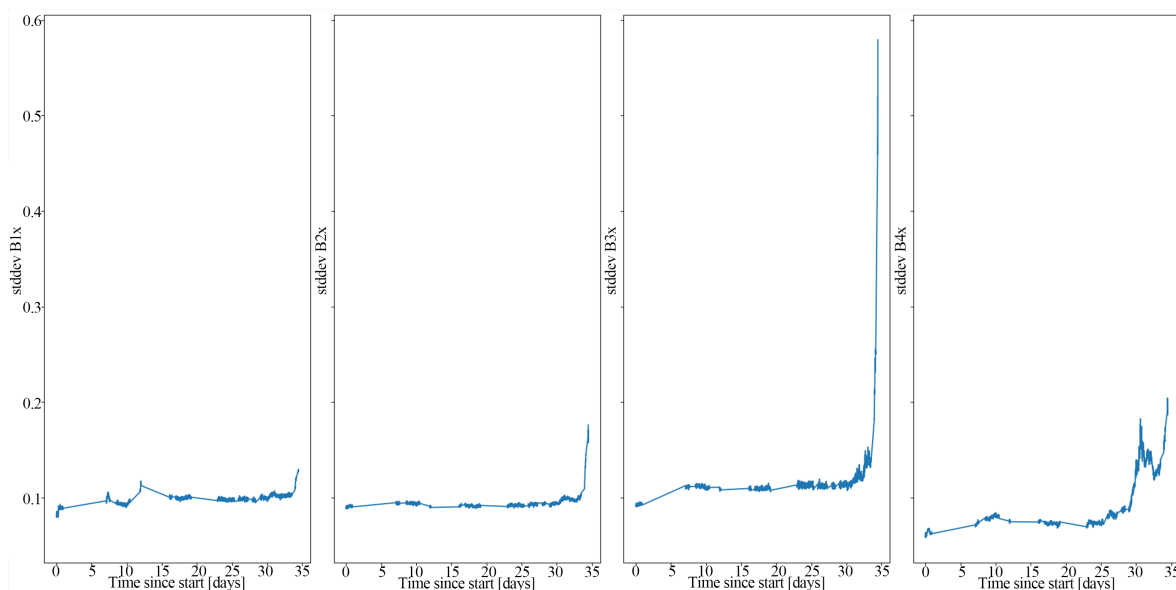


Figure 2. RMS value

图 2. RMS 值

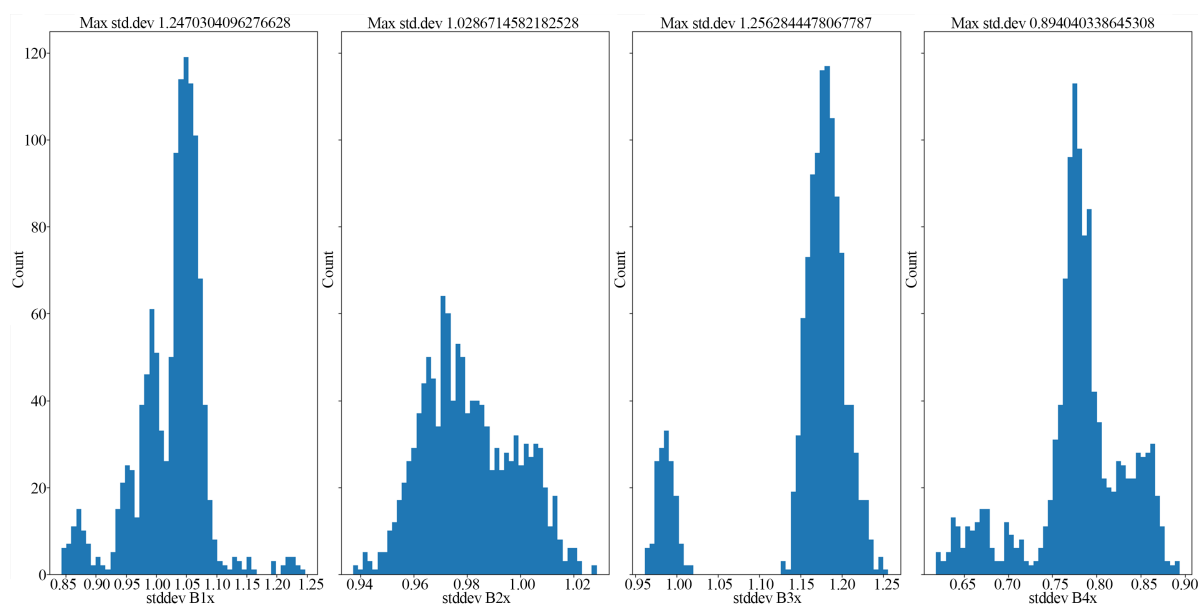


Figure 3. The number of times different RMS values occur

图 3. 不同 RMS 值出现的次数

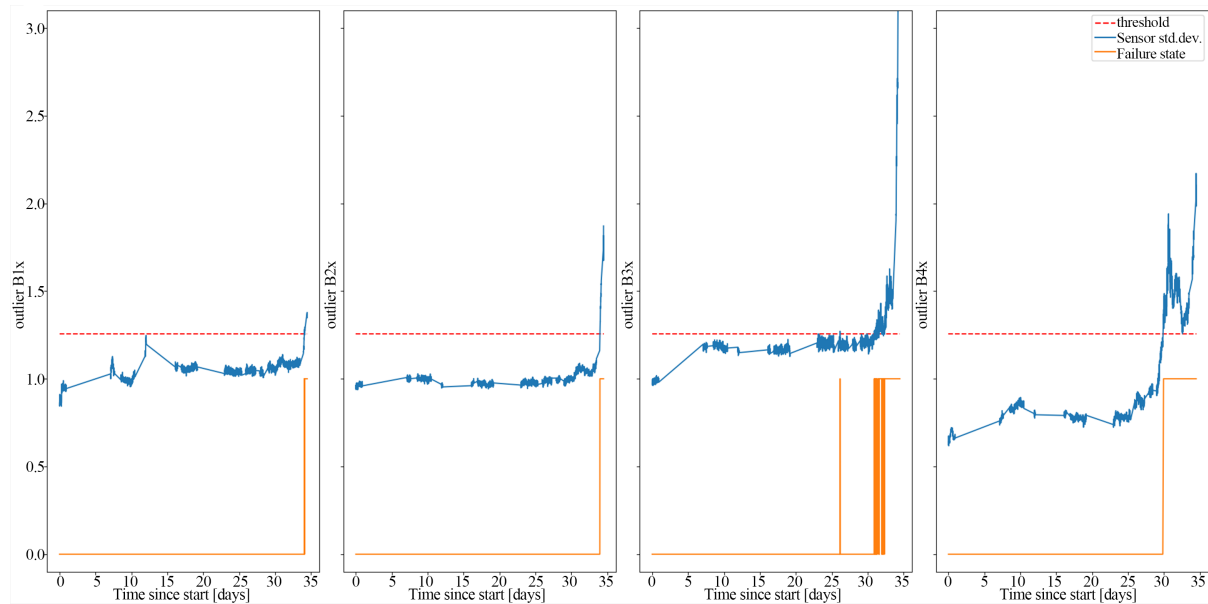


Figure 4. Bearing failure results
图 4. 轴承故障结果

4.2.2. 分类器方法

使用随机森林的方法，我们得到的重要性排名前 20 的特征如图 5 所示。

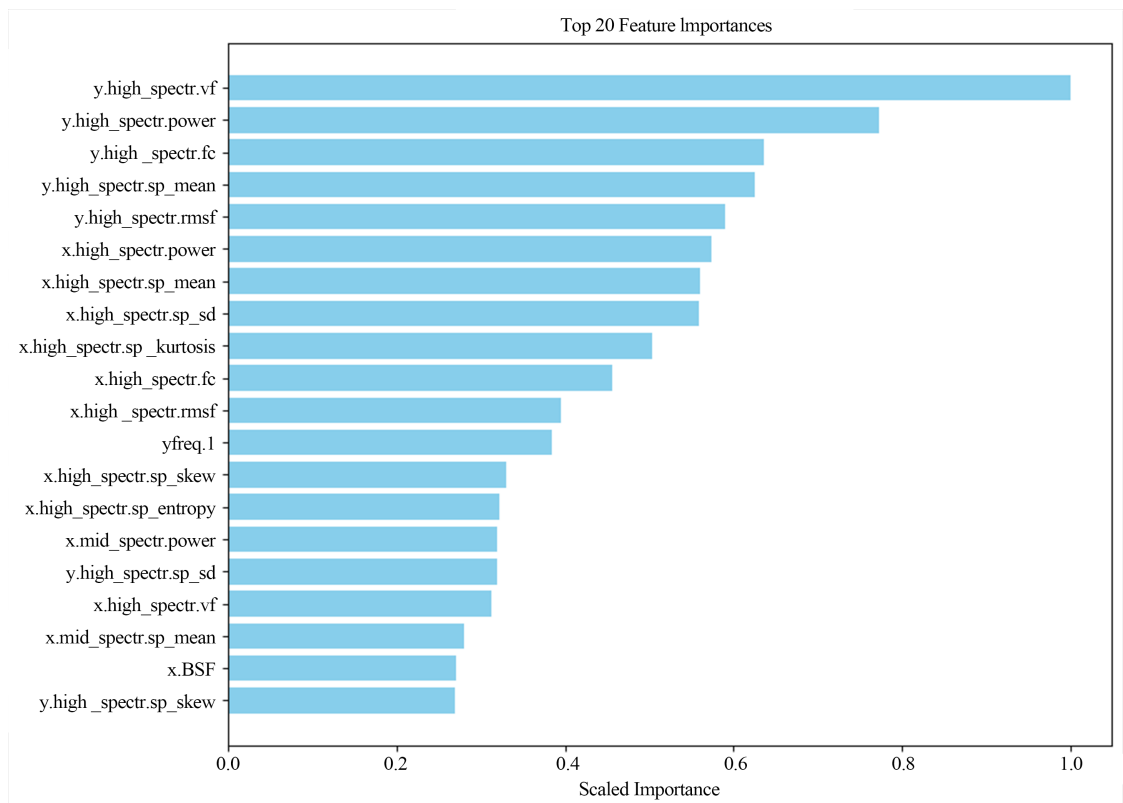


Figure 5. Top 20 characteristics in importance
图 5. 重要性排名前 20 特征

下面使用重要性排名前十的特征进行后续的验证。首先, 将数据集分为训练集和测试集, 其中训练集占总数据集的 70%, 测试集占总数据集的 30%, 然后使用网格搜索进行了 10 倍三次重复交叉验证找到最优的梯度增强分类器。下面分别使用上面得到的最优的梯度增强分类器, LightGBM 分类器和迭代次数为 1000 次, 树的深度为 10, 学习率为 0.05 的 Catboost 分类器分别进行预测。表 2 是三种不同分类器预测的精确度以及运行的时间比较。

Table 2. The accuracy of the forecast
表 2. 预测的精确度

	梯度增强分类器	LightGBM	Catboost
Accuracy	0.94	0.9417	0.9471
Time	3.5 h	9 s	1 m 29 s

梯度增强分类器是一个基础的算法, 通过迭代构建决策树, 然后每次迭代都试图更改之前的错误, 最小化损失函数, 但是梯度增强算法的时间成本更多, 实现起来更加缓慢。LightGBM 分类器是一个基于直方图的决策树算法, 这个算法通过构建直方图, 加速算法的实现, 这个算法支持特征和数据并行计算, 所以运行效率更高, 更加适合大规模的数据。Catboost 分类器也是一种梯度提升算法, 但其针对类别特征进行了优化, 能够处理类别特征有效防止过拟合, 这个算法使用有序提升算法, 可以有效防止信息泄露, 提高训练速度。

5. 总结

我们使用了三种不同的分类器, 以此来确定滚动轴承的健康状况, 并且识别故障类型。我们在特征提取和减少方面进行了很多实验, 这使得我们在测试集上取得了高于 90% 的准确率, 这说明我们的工作是有意义的。除此之外在测试集上的预测存在一部分无意义的结果, 但这部分的存在是合理的, 这是从正常到故障的过渡部分, 如何减少没有意义的预测结果使得正常与故障的界限更加清晰是我们需要解决的一个问题。

参考文献

- [1] 郭玉荣, 茅健, 赵嫚. 基于 CNN 与注意力 BiLSTM 的轴承剩余使用寿命预测方法[J]. 上海工程技术大学学报, 2023, 37(1): 96-104.
- [2] 范强飞, 廖爱华, 丁亚琦. 基于 RVM 和 WPHM 的滚动轴承剩余寿命预测[J]. 上海工程技术大学学报, 2019, 33(4): 334-338.
- [3] 胡澜也, 蒋文博, 李艳婷. 基于 LightGBM 的风力发电机故障诊断[J]. 太阳能学报, 2021, 42(11): 255-259.
- [4] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- [5] 赵玉程, 李英建, 沈世民, 等. 基于网格搜索和投票分类模型的喷油器故障诊断研究[J]. 机床与液压, 2024, 52(5): 213-220.
- [6] Qiu, H., Lee, J., Lin, J. and Yu, G. (2006) Wavelet Filter-Based Weak Signature Detection Method and Its Application on Rolling Element Bearing Prognostics. *Journal of Sound and Vibration*, **289**, 1066-1090. <https://doi.org/10.1016/j.jsv.2005.03.007>
- [7] 李立森. 基于时域和频域信息融合的股票趋势预测模型[D]: [硕士学位论文]. 济南: 山东大学, 2023.