Computer Science and Application 计算机科学与应用, 2025, 15(5), 264-269
Published Online May 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.155148

Hans汉斯

基于机器学习与传感器数据的机器故障预测研究

黄子敬

宁波工程学院统计与数据科学学院，浙江 宁波

收稿日期：2025年4月25日；录用日期：2025年5月23日；发布日期：2025年5月30日

摘　要

本文研究了利用机器学习算法预测机器故障。作者使用了和鲸社区的传感器数据，包含footfall，tempMode，AQ，USS，CS，VOC，RP，IP，Temperature和fail等特征。利用这些数据，分别构建了XGBoost、随机森林和KNN三种机器学习模型进行故障预测，并比较了它们的性能。实验结果表明，XGBoost模型在AUC和准确率等指标上表现最佳，AUC值为0.9721，准确率为0.9120。机器故障预测在工业生产中具有重要意义，可以减少设备停机时间，提高生产效率。本文的研究成果可以为企业提供有效的机器故障预测方法，具有一定的实际应用价值。

关键词

机器学习，传感器，XGBoost

Research on Machine Fault Prediction Based on Machine Learning and Sensor Data

Zijing Huang

College of Statistics and Data Science, Ningbo University of Technology, Ningbo Zhejiang

Received: Apr. 25th, 2025; accepted: May 23rd, 2025; published: May 30th, 2025

Abstract

This paper studies the prediction of machine failures using machine learning algorithms. The author utilized sensor data from the whale community, including features such as footfall, tempMode, AQ, USS, CS, VOC, RP, IP, Temperature, and fail. Using these data, three machine learning models, namely XGBoost, Random Forest and KNN, were respectively constructed for fault prediction, and their performances were compared. The experimental resultss show that the XGBoost model

文章引用: 黄子敬. 基于机器学习与传感器数据的机器故障预测研究[J]. 计算机科学与应用, 2025, 15(5): 264-269.
DOI: 10.12677/csa.2025.155148

performs best in terms of indicators such as AUC and accuracy rate. The AUC value is 0.9721 and the accuracy rate is 0.9120. Machine failure prediction is of great significance in industrial production, which can reduce equipment downtime and improve production efficiency. The research results of this paper can provide enterprises with effective machine fault prediction methods and have certain practical application value.

Keywords

Machine Learning, Sensor, XGBoost

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

在现代工业体系下，机器故障预测已然成为推动生产效率提升与设备停机时长缩减的核心要素。精准预测机器故障，可助力企业制定科学合理的预防性维护规划，有效削减维护开支，延长设备使用寿命，保障生产线的稳定、持续运行。对影响机器故障的关键因素展开深度剖析，能让企业更透彻地洞察机器的运行状况。基于这一分析，企业能够制定更具针对性、更贴合实际需求的维护策略，进而全面提升生产效率，增强设备的可靠性。

随着人工智能技术的不断发展，机器故障诊断也使用了很多的机器学习技术。安会勇等[1]提出了一种基于机器学习算法的架空输电线路故障定位体系。针对光伏发电系统内故障频发的情况，赵海宝[2]提出了一种基于机器学习技术的故障诊断系统方案。郭广辉[3]针对运行状态监测与故障诊断问题，运用机器学习技术开展了算法优化与集成工作。傅闽豪[4]针对变压器故障诊断过程中输入特征集合内无关特征与冗余特征的问题，提出了一种基于机器学习技术的变压器故障诊断技术。左娟娟等[5]为实现对复杂电网连锁故障的及时监测，提高电力系统运行效率，提出了一种基于半监督机器学习的复杂电网连锁故障诊断技术。

2. 数据介绍

本文的实验数据是和鲸社区的传感器数据，数据由 footfall (经过机器的人数或物体数量)、tempMode (经过机器的人数或物体数量)、AQ (机器附近的空气质量指数)、USS (超声波传感器数据，表示接近度测量)、CS (当前传感器读数，表示机器的电流使用情况)、VOC (检测到的挥发性有机化合物水平)、RP (机器部件的旋转位置或每分钟转数)、IP (机器的输入压力)、Temperature (机器的运行温度)、fail (标签)10 列构成。在这组数据中，正常样本(标记为 0 类)的数量为 551 个，而故障样本(标记为 1 类)的数量是 393 个。

3. 算法简介

3.1. XGBoost 算法

在机器学习领域中，XGBoost 堪称梯度提升树(GBDT)算法的卓越优化版本。

XGBoost 算法其主流程如下：

假设输入 $I=\{(x_1,y_1),(x_2,y_2),\cdots(x_m,y_m)\}$，输出是 $f(x)$。

对迭代轮数 $t=1,2,\cdots T$ 有：

(1) 迭代通过 L 来计算第 i 个样本 $(i-1,2,\cdots m)$ 基于 $f_{t-1}(x_i)$ 的一阶导数 g_{ti} 和二阶导数 h_{ti}，进而求出一阶导数和 $G_t=\sum_{i=1}^{m}g_{ti}$，二阶导数和 $H_t=\sum_{i=1}^{m}h_{ti}$。

(2) G 和 H 是分裂节点的一阶导数加上二阶导数。

对特征序号 $k=1,2...K$：

a) $G_L=0,H_L=0$

b) 1)取出第 i 个样本：

$$G_L=G_L+g_{tt},G_R=G-G_L \tag{3-1}$$

$$H_L=H_L+h_{tt},H_R=H-H_L \tag{3-2}$$

b) 2)尝试更新最大的分数：

$$S=\max\left(\text{score},\frac{1}{2}\frac{G_L^2}{H_L+\lambda}+\frac{1}{2}\frac{G_R^2}{H_R+\lambda}-\frac{1}{2}\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}-\gamma\right) \tag{3-3}$$

(3) 找到该进程中的分裂特征和特征值。

(4) 继续迭代下一轮弱学习器。如果 S 的最大值不是 0，就重复第二步。

3.2. 随机森林算法

对于机器学习分类问题，数据集的输入假设为 D = {(x, y_1), (x_2, y_2), ...(x_m, y_m)}，那么弱分类器的迭代次数为 T，强分类器就是 $f(x)$。

1) 对于 t = 1, 2, ..., T：

a) 针对训练数据做随机采样得到集合 D_t。

b) 用第 t 次采样得到的数据集 D_t 去训练第 t 个模型 $G_t(x)$，在数据中选择样本的一部分特征来训练决策树的节点。

2) T 个弱学习器的结果中票数最多的为最终类别。

3.3. KNN 算法

在分类场景里，KNN 算法的运行流程是这样的：当给定一个待判定类别的样本点时，算法会在训练数据集合里搜寻与该样本点距离最为接近的 K 个样本。最后再根据这 K 个相邻样本中数量最多的类别归属来判断待分类样本的类型。

4. 基于机器学习与传感器数据的机器故障预测

4.1. 模型构建

在本文里，首先对数据进行了预处理，将预测目标标签(fail)与建模所使用的特征数据进行了分离。随后，采用随机抽样的方法将完整数据集划分为训练集和测试集，划分比例为 7:3，即训练集占总样本的 70%，测试集占总样本的 30%。本研究采用三种经典机器学习算法——XGBoost 算法、随机森林算法以及 K 近邻(KNN)算法，分别构建了用于机器故障诊断的预测模型。具体参数设置如下：XGBoost 模型中，将随机种子设为 2025 以保障实验可重复性，树的最大深度(max_depth)设为 6，决策树数量(n_estimators)设为 100，学习率(learning_rate)设为 0.1，其余参数采用默认值；随机森林模型中，随机种子同样设为 2025，树的最大深度设为 2，决策树数量设为 20，其余参数保持默认；KNN 模型中，主要调整了邻居数量(n_neighbors)参数，设为 10，其余参数为默认设置。

4.2. 结果比较

本文采用计算准确率、召回率、精确率、F1 综合得分以及 AUC 这五项核心性能评估指标，对基于三种不同机器学习算法构建的机器故障预测模型进行比较分析。

3 个模型的测试集混淆矩阵如图 1，图 2 和图 3 所示。

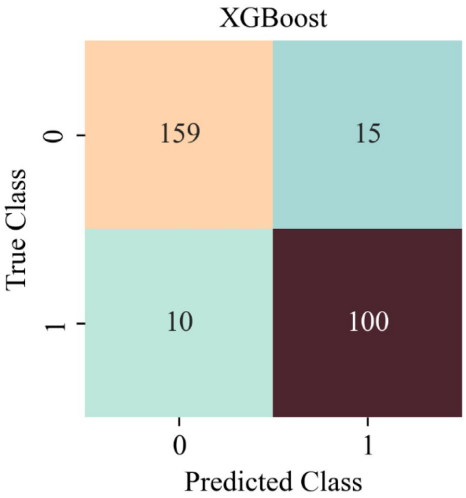

Figure 1. XGBoost confusion matrix
图 1. XGBoost 混淆矩阵

图 1 是 XGBoost 模型在传感器数据测试集上预测得到的混淆矩阵，图中 0 代表正常机器样本，1 代表故障机器样本。XGBoost 模型在测试集上预测对了 259 个样本，预测错了 25 个样本，其准确率为 91.20%。

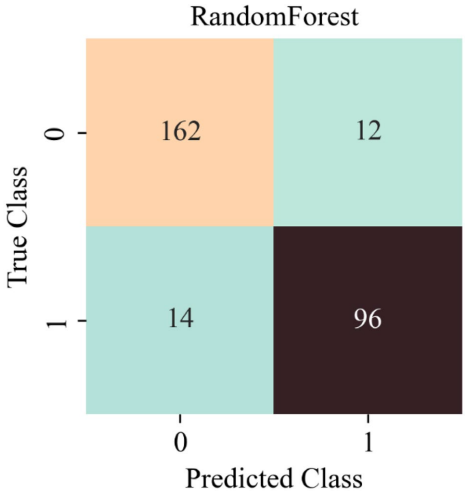

Figure 2. Random forest confusion matrix
图 2. 随机森林混淆矩阵

图 2 是随机森林模型在传感器数据测试集上预测得到的混淆矩阵。随机森林模型在测试集上预测对了 258 个样本，预测错了 26 个样本，其准确率为 90.85%。

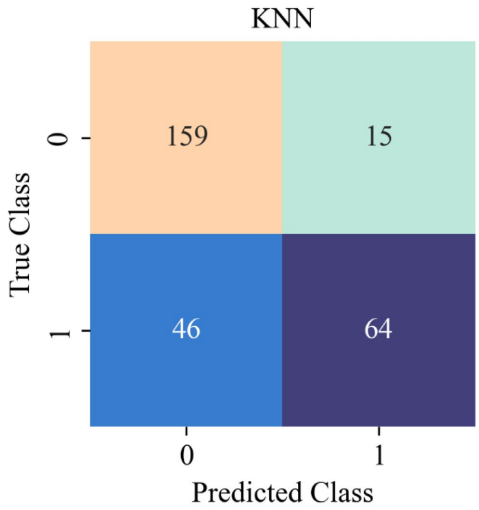

Figure 3. KNN confusion matrix
图 3. KNN 混淆矩阵

图 3 是 KNN 模型在传感器数据测试集上预测得到的混淆矩阵。随机森林模型在测试集上预测对了 223 个样本，预测错了 61 个样本，其准确率为 78.52%。

XGBoost 预测模型、随机森林预测模型和 KNN 预测模型的五项性能指标如表 1 所示。

Table 1. Comparison of model results
表 1. 模型结果比较

	AUC	准确率	精确率	召回率	F1 得分
XGBoost	0.9721	0.9120	0.8696	0.9091	0.8889
RF	0.9614	0.9085	0.8889	0.8727	0.8807
KNN	0.8124	0.7852	0.8101	0.5818	0.6772

通过对比三个模型在测试集上的各项性能指标数据，可以清晰看出，采用 XGBoost 算法构建的机器故障预测模型在评估关键指标上展现出显著优势。具体而言，该模型在 AUC 与准确率两项核心评估维度上均取得了领先表现，其得分均突破 90%的优异水准：AUC 值高达 0.9721，准确率值达到 0.9120。同时，该模型在精确率(0.8696)、召回率(0.9091)及 F1 综合得分(0.8889)等维度亦表现稳健。综合各项指标评估结果，XGBoost 模型展现出卓越的预测性能，被确立为本研究的最优预测模型。

5. 结语

本文选取和鲸社区丰富的传感器数据作为核心研究对象，分别运用 XGBoost 算法、随机森林算法以及 K 近邻(KNN)算法构建了三种机器故障预测模型。经过严谨的实验验证，结果表明所构建的机器故障预测模型整体展现出较高的预测精度与稳定性。尤为突出的是，最优的 XGBoost 模型在测试数据集上取得了令人瞩目的成绩，其 AUC 值高达 0.9721，准确率值也达到了 0.9120。本文的研究成果不仅丰富了机器故障预测的理论体系，更对机器故障诊断提供了切实有效的参考依据，具有较高的应用价值。

参考文献

[1] 安会勇, 孔庆绿, 刘硕. 基于机器学习算法的架空输电线路故障定位系统设计与实现[J]. 电气技术与经济, 2025(3): 1-3.

[2] 赵海宝. 基于机器学习的光伏发电系统故障诊断系统研究[J]. 自动化应用, 2025, 66(4): 27-29.

[3] 郭广辉. 基于机器学习的智能变电站运行状态监测与故障诊断方法研究[J]. 电力设备管理, 2025(2): 216-218.

[4] 傅闽豪. 基于机器学习的变压器故障诊断及预警研究[D]: [硕士学位论文]. 南昌: 南昌大学, 2024.

[5] 左娟娟, 朱红杰, 杨继党, 等. 基于半监督机器学习的复杂电网连锁故障诊断方法[J]. 自动化技术与应用, 2024, 43(12): 47-50+92.