本文引用格式：范国栋,李博涵.基于机器学习的工业机械设备故障预测方法[J].自动化与信息工程,2023,44(4):13-18;50.
FAN Guodong, LI Bohan. Fault prediction method of industrial machinery equipment based on machine learning[J]. Automation & Information Engineering, 2023,44(4):13-18;50.

基于机器学习的工业机械设备故障预测方法

范国栋　李博涵
（重庆交通大学机电与车辆工程学院，重庆 400074）

摘要：为提高工业生产效率和安全性，研究基于机器学习的工业机械设备故障预测方法。首先，利用斯皮尔曼等级相关系数分析工业机械设备故障特征之间的相关性，并过滤冗余特征；然后，采用随机森林算法筛选影响工业机械设备故障的 3 个核心特征；最后，基于逻辑回归、朴素贝叶斯、XGBoost、决策树等机器学习算法分别建立工业机械设备的故障预测模型和故障类型预测模型。经实验验证，基于 XGBoost 算法构建的工业机械设备故障预测模型和决策树训练出来的工业机械设备故障类型预测模型具有较高的准确性。该方法具有实际的应用价值，可有效地预测不同工业机械设备的故障类型，为工业安全生产提供技术支持。

关键词：机器学习；工业机械设备；故障预测；斯皮尔曼相关性分析；随机森林算法；预测模型
中图分类号：TP399　　**文献标志码：**A　　**文章编号：**1674-2605(2023)04-0003-07
DOI：10.3969/j.issn.1674-2605.2023.04.003

Fault Prediction Method of Industrial Machinery Equipment Based on Machine Learning

FAN Guodong　LI Bohan
(School of Electromechanical and Vehicle Engineering, Chongqing Traffic University, Chongqing 400074, China)

Abstract: To improve industrial production efficiency and safety, a machine learning based fault prediction method for industrial machinery and equipment is studied. Firstly, the Spearman rank correlation coefficient is used to analyze the correlation between fault features of industrial machinery equipment, and redundant features are filtered; Then, the random forest algorithm is used to screen the three core features that affect the faults of industrial machinery and equipment; Finally, based on machine learning algorithms such as logistic regression, naive Bayes, XGBoost, and decision tree, a fault prediction model and a fault type prediction model for industrial machinery equipment are established. Through experimental verification, the industrial machinery equipment fault prediction model constructed based on XGBoost algorithm and the industrial machinery equipment fault type prediction model trained from decision trees have high accuracy. This method has practical application value and can effectively predict the fault types of different industrial machinery and equipment, providing technical support for industrial safety production.

Keywords: machine learning; industrial machinery and equipment; fault prediction; Spearman correlation analysis; random forest algorithm; prediction model

0　引言

工业机械设备故障的突发性和不可预见性，会影响生产效率和生产成本。通过对工业机械设备进行预测性维护，可减少故障损失、提高生产效率、降低生产成本。传统的工业机械设备故障预测大多通过专业的传感器进行监测和分析，如高海军[1]利用电气类机械设备运行过程中产生的异常声音和表面温度升高进行故障诊断；张益沛[2]利用振动监测仪和温度传感器等，提高旋转类机械设备的故障检测效率；马梁[3]采用状态检测和故障诊断平台对煤矿机电设备进行

故障预测。以上方法主要基于声发射、热成像、振动分析、超声波检测等技术，存在成本高、动态响应差等问题。

随着人工智能技术的不断进步，越来越多的学者将其应用于工业机械设备故障预测领域。李玉吉等[4]利用机器学习算法诊断煤矿汽车机械设备的故障，实验结果表明，故障诊断的准确性和效率都优于传统方法。

本文基于机器学习技术，利用工业机械设备作业的信息数据进行故障预测和故障类型诊断，不仅能提高设备的安全性和可靠性，还能实现更精准的故障预测和诊断。

1　数据描述及预处理

1.1　数据预处理

本文使用的数据集是由某行业协会提供的工业机械设备故障预测数据集和工业机械设备故障类型预测数据集。工业机械设备故障预测数据集主要包括机器编码（工业机器人型号、电动机序列号等）、统一规范代码、机器质量等级（机械、电气、液压等机器的性能指标和品质等级）、厂房室温（整个厂房内的平均温度，在数据集中用室温（K）表示）、设备室温（设备存放和工作的环境温度，在数据集中用室温（K）.1 表示）、转速、扭矩、使用时长、是否发生故障、具体故障类型等 10 个数据标签。其中，机器编码、厂房室温、设备室温、转速、扭矩、使用时长 6 个数据标签是连续变量；统一规范代码、机器质量等级、是否发生故障、具体故障类型 4 个数据标签是离散变量。因为机器编码和统一规范代码这 2 个数据标签与设备故障无关，所以排除在设备故障预测的相关变量之外[5]。

利用统计分析软件（SPSS、Excel 等）可了解连续变量的数据分布和集中程度。箱型图可清晰地展示数据的分布情况，包括中位数、四分位数、极值和异常值等信息。利用箱型图对厂房室温、设备室温、转速、扭矩的异常值进行可视化处理，分别如图 1~ 4 所示。

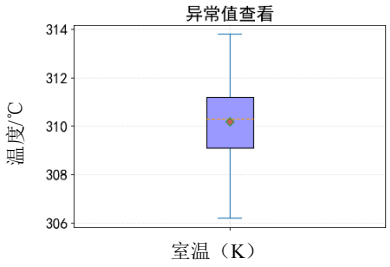

图 1　厂房室温异常值箱型图

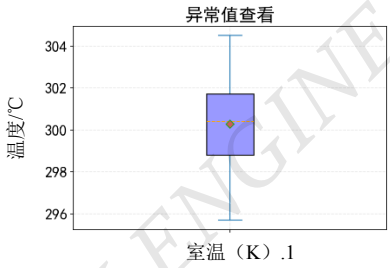

图 2　设备室温异常值箱型图

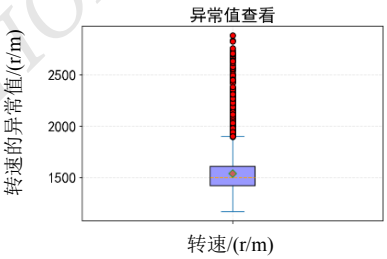

图 3　转速异常值箱型图

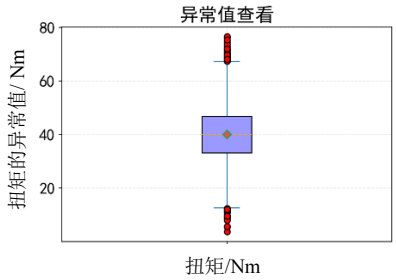

图 4　扭矩异常值箱型图

由图 1~4 可知，转速和扭矩的个别异常值超出了箱型图下界。通过分析设备的运行状态和性能，进一步判断这些异常值是因为不同机械设备的转速和扭矩存在差异而导致的，因此不删除这些异常值。

1.2　数据类型转换

机器质量等级（L 级、M 级、H 级）是离散数据，采用文字描述表示。然而，机器学习模型在训练和预

测过程中，只能处理连续数据。因此，需将离散数据转换为连续数据。本文采用独热编码技术，将离散数据转换为二元数据，即用 0、1、2 分别替换 L 级、M 级、H 级，转换后的数据如表 1 所示。

表 1　机器质量等级离散数据转换为连续数据

机器质量等级	室温（K）/℃	室温（K）.1/℃	转速/（r/m）	扭矩/Nm	使用时长/ min	是否发生故障	具体故障类型
1	296.4	307.4	2 833	5.6	213	1	PWF
2	295.8	307.3	1 235	76.2	89	1	PWF
1	295.7	307.2	2 270	14.6	149	1	PWF
2	296.3	307.1	1 534	33.8	151	0	Normal
0	296.3	307.1	1 774	25.9	154	0	Normal
2	296.2	307.0	2 119	18.3	159	0	Normal
1	296.2	307.0	1 414	48.3	162	0	Normal
2	296.1	307.0	1 523	42.0	164	0	Normal
1	296.1	307.1	1 651	35.7	167	0	Normal
2	296.1	307.1	1 485	36.0	169	0	Normal
2	296.2	307.2	1 168	63.4	172	0	Normal
1	296.3	307.3	1 566	35.8	175	0	Normal

1.3　过采样处理

本文使用的数据集包含了 9 000 条工业机械设备的信息，其中无故障和有故障的工业机械设备信息分别有 8 697 条和 303 条。无故障的工业机械设备信息数量远多于有故障的工业机械设备信息数量，导致训练后的模型偏差较大。

利用合成少数类过采样技术（synthetic minority oversampling technique, SMOTE）算法，通过样本合成的方法，生成与原始样本相似的新样本，达到扩充数据集的目的。利用 SMOTE 算法的过采样方法，将有故障和无故障的工业机械设备信息数量均衡化，提升预测模型的准确性。均衡和扩充后的数据集中，无故障和有故障的工业机械设备信息数量均为 6 300 条。

2　机械设备故障预测相关变量分析与筛选

2.1　是否故障的相关变量分析与筛选

斯皮尔曼相关性分析是一种非参数检验方法，用于评估两个连续变量之间的相关性。通过热力图，可直观地看出变量之间的相关性指数，颜色越深表示相关性越强，颜色越浅表示相关性越弱。对均衡和扩充后的无故障和有故障的工业机械设备信息进行相关变量影响因素的可视化处理，如图 5 所示。

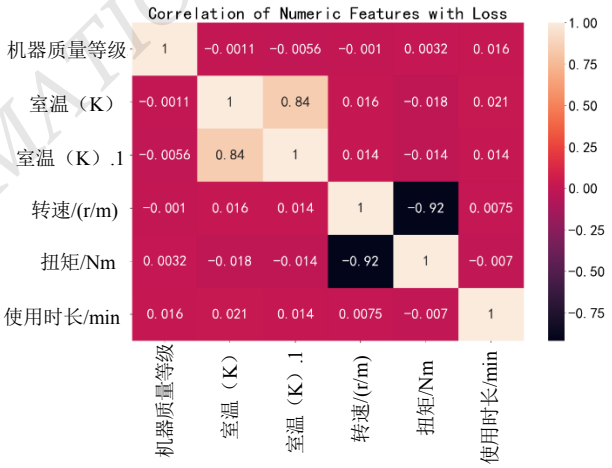

图 5　是否故障热力图

由图 5 可知：室温（K）（厂房室温）和室温（K）.1（设备室温）的相关性指数为 0.84；转速和扭矩的相关性指数为−0.92，说明它们之间的相关性较强，删除室温（K）（厂房室温）和扭矩这 2 个变量，以避免信息冗余和多重共线性的问题。

根据每个变量在随机森林中对模型预测结果的影响程度，得出变量的重要性评分[6]如图 6 所示。

由图 6 可知，机器质量等级变量对工业机械设备是否发生故障的影响较低，对该变量进行删除处理。

通过对厂房室温、设备室温、扭矩、转速、使用

时长和机器质量等级等变量进行斯皮尔曼相关性分析和随机森林重要性评分后，本文选择室温（K）.1（设备室温）、转速、使用时长 3 个变量为预测工业机械设备故障的指标。

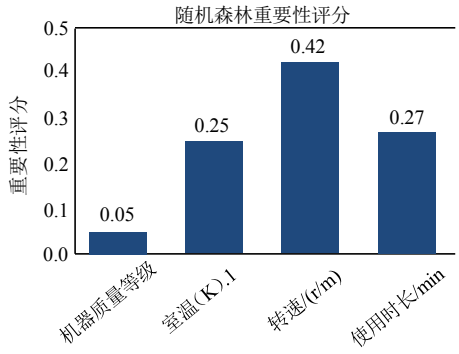

图 6　是否故障重要性评分

2.2　故障类型的相关变量分析与筛选

工业机械设备故障类型预测数据集主要包括扭矩扳手故障（torque wrench fault, TWF）、高频设备故障（high-frequency device fault, HDF）、电源故障（power supply fault, PWF）、规格超标故障（oversized specification fault, OSF）、随机非重复故障（random non-repetitive fault, RNF）等 5 种故障类型。为预测工业机械设备故障类型，需删除没有故障的机械设备信息，保留有故障的机械设备信息，并将离散的具体故障类型转换为连续数据，即用 0、1、2、3、4 分别替换 TWF、HDF、PWF、OSF、RNF。通过斯皮尔曼相关性分析和随机森林重要性评分，选出与故障类型预测相关的变量，如图 7、图 8 所示。

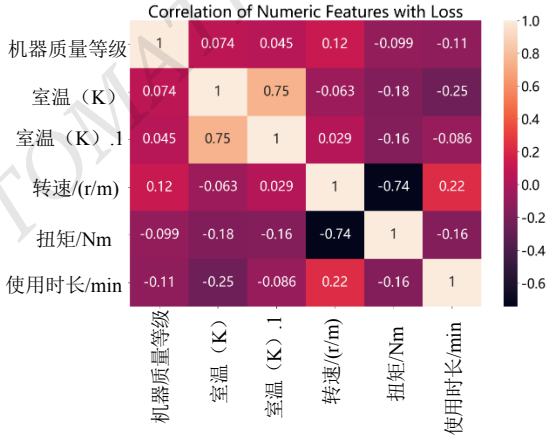

图 7　具体故障类型热力图

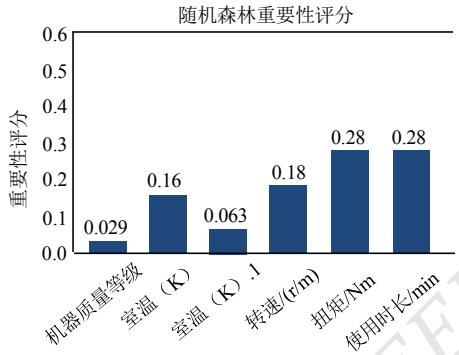

图 8　故障类型重要性评分

由图 7 可知，室温（K）（厂房室温）和室温（K）.1（设备室温）的相关性较高，删除室温（K）（厂房室温）。

由图 8 可知，机器质量等级变量对工业机械设备故障类型预测的影响程度最小[7]，对其进行删除处理。

3　模型建立及评估

3.1　建立故障预测模型

将工业机械设备故障预测数据集按 1:1 的比例随机划分为训练集和测试集。其中，训练集用于故障预测模型的训练和优化；测试集用于评估故障预测模型的性能和泛化能力。采用交叉验证的方法进行多次实验，以减少随机误差，提高模型的稳定性。利用随机森林、XGBoost、逻辑回归和朴素贝叶斯模型对故障预测模型进行性能评估。

利用训练好的模型进行预测，采用准确率、精确率、召回率和 F1 值等 4 个指标来评价模型的预测性能[7]。混淆矩阵可直观地表现预测模型的误差。随机森林、XGBoost、逻辑回归和朴素贝叶斯模型的混淆矩阵分别如图 9~12 所示，评价指标如表 2 所示。

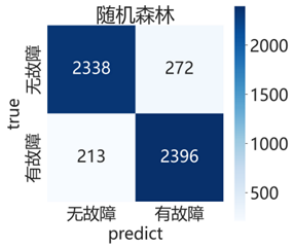

图 9　随机森林混淆矩阵

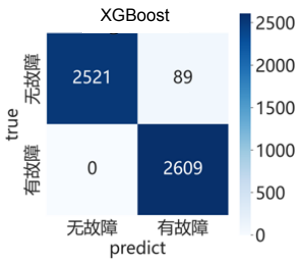

图 10　XGBoost 混淆矩阵

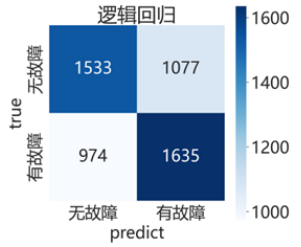

图 11　逻辑回归混淆矩阵

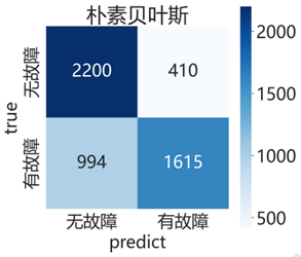

图 12　朴素贝叶斯混淆矩阵

表 2　4 个模型的评价指标

模型	准确率	精确率	召回率	F1 值
随机森林	0.907	0.898	0.918 3	0.908
XGBoost	0.982	0.967	1.000 0	0.983
逻辑回归	0.607	0.602	0.626 0	0.614
朴素贝叶斯	0.730	0.797	0.619 0	0.697

ROC 曲线下面积（area under curve, AUC）是评估分类器性能的一个指标，取值范围为 0.5~1，指标数值越接近 1，说明分类器的性能越好。根据 4 个模型的混淆矩阵绘制 ROC 曲线，可直观地看出模型效果，如图 13 所示。

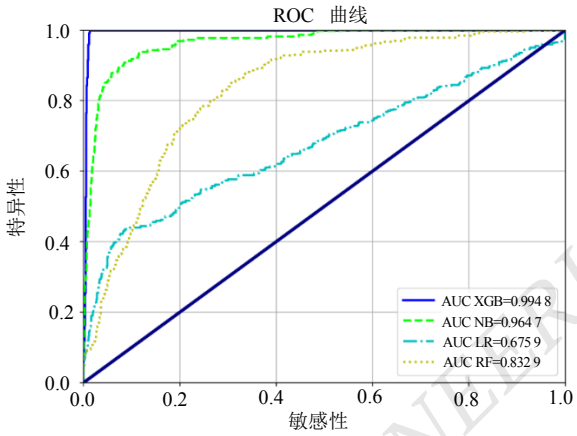

图 13　4 个模型的 ROC 曲线

由图 13、表 2 可知，XGBoost 训练出来的故障预测模型准确率、精确率、召回率、F1 值较高，AUC 值也最高，说明该模型的预测效果最好[8]。

3.2　建立故障类型预测模型

工业机械设备故障类型预测数据集按 1:1 的比例随机划分为训练集和测试集。工业机械设备故障类型预测模型的性能评估利用决策树、梯度提升树、支持向量机等模型，采用准确率、精确率、召回率和 F1 值作为评价指标。

决策树、梯度提升树、支持向量机 3 个模型的混淆矩阵分别如图 14~16 所示，评价指标如表 3 所示。

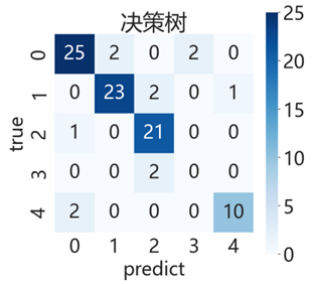

图 14　决策树混淆矩阵

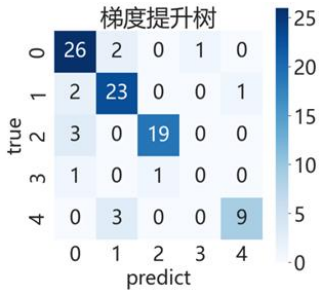

图 15　梯度提升树混淆矩阵

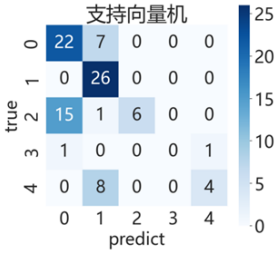

图 16　支持向量机混淆矩阵

表 3　3 个模型的评价指标表

评价指标	决策树	梯度提升树	支持向量机
准确率	0.868 13	0.846 15	0.637 36
精确率	0.870 35	0.841 97	0.708 62
召回率	0.868 13	0.846 15	0.637 36
F1 值	0.867 95	0.841 67	0.593 43

由表 3 可知，由决策树训练出来的故障类型预测模型的准确率、精确率、召回率、F1 值最高，说明该模型预测故障类型的效果最好[9]。

对基于决策树算法构建的故障类型预测模型进行调参时，易出现过拟合现象。因此，需先选择合适的正则化参数，本文通过交叉验证的方法确定了正则化参数为 0.07，再定义超参数搜索范围。本文定义 max_depth（表示决策树的最大深度）范围为 1~10，min_samples_split（表示节点在分裂之前所需的最小样本数）范围为 2~10，min_samples_leaf（表示叶节点上的最小样本数）范围为 1~5，max_features（表示在每个节点中考虑的最大特征数）范围为 1~10。将预测结果进行比较，具体结果如表 4 所示。

表 4　基于决策树算法构建的故障类型预测模型调参前后性能指标表

调参时间	准确率	精确率	召回率	F1 值
调参前	0.868 13	0.870 35	0.868 13	0.867 95
调参后	0.876 42	0.874 91	0.876 42	0.875 25

由表 4 可以看出，基于决策树算法构建的故障类型预测模型调参后，其性能指标均有所提升。调整和优化基于决策树算法构建的故障类型预测模型，能够提高模型的识别准确率、泛化能力和稳定性，降低误差率和资源占用率。

4　结论

本文基于机器学习算法建立了工业机械设备故障的预测模型和类型预测模型，具有较高的准确性，可为工业机械维护部门提供有效的参考。然而，本研究还存在不足之处：首先，只考虑了室温、转速、使用时长等少量特征，对其他可能影响工业机械设备故障的特征，如湿度、负载等没有进行探究；其次，仅针对单一类型的工业机械设备故障进行预测，对于不同类型的机械设备模型还需进一步探究[10]；最后，该研究可扩展到工业互联网领域，使各种工业设备实现数据的共享和交互，为工业设备的智能维护提供更多的可能性。

参考文献

[1] 高海军.化工电气常见故障分析及处理方法[C].//中国机电一体化技术应用协会.第七届全国石油和化工电气技术大会论文集.[出版者不详],2023:193-195.

[2] 张益沛.旋转机械故障诊断技术在炼钢设备中的运用分析[J].冶金与材料,2023,43(1):71-73.

[3] 马梁.煤矿机电设备实时监测故障诊断技术研究应用[J].煤炭科技,2023,44(1):64-68.

[4] 李玉吉,曹旭辉,王江宏,等.基于机器学习算法的煤矿汽车机械设备故障诊断模型[J].能源与环保,2021,43(10):241-245.

[5] 盛建龙,乔宇,王平,等.基于 LOF-SMOTE 算法的地下水影响下矿山岩溶塌陷风险预测研究[J].有色金属科学与工程,2023,14(3):372-380;399.

[6] 张文涛,龚振宇,令凡琳,等.基于随机森林算法的盾构改良渣土渗透系数预测及工程应用[J].隧道建设(中英文), 2022, 42(11):1863-1870.

[7] 刘偲,刘道星.XGBoost 算法在塔式起重机传感器故障诊断中的应用[J].建设机械技术与管理,2022,35(5):115-117.

[8] 陈天锴,王贵勇,申立中,等.基于 GBDT 算法的柴油机性能预测[J].车用发动机,2022(5):51-58.

[9] 蒋琳,徐猛.基于朴素贝叶斯分类的交通枢纽内移动时间估计——以北京南站为例[C]//中国科学技术协会,交通运输部,中国工程院,湖北省人民政府.2022 世界交通运输大会(WTC2022)论文集（运输规划与交叉学科篇）.人民交通出版社股份有限公司,2022:556-562.

[10] 任利娟.滚动轴承性能退化评估与剩余寿命预测[D].济南:山东大学,2019.

（下转第 50 页）