基于主成分分析与随机森林的电动矿卡电机故障诊断研究

彭 倩，杨晨瀚

（厦门理工学院 机械与汽车工程学院，福建 厦门 361024）

摘要： 针对纯电动矿山卡车作业过程中复杂因素交互干扰而引发的电机故障问题，提出了一种基于主成分分析（PCA）和随机森林（RF）的方法进行预测诊断。根据实际采集的电动矿卡电机故障构建数据集，利用PCA对故障数据进行特征值提取、降维，减少数据维度冗余；采用RF预测模型对降维后的数据进行训练测试，并对电机故障类别进行预测。结果表明：采用PCA-RF方法对电机故障类型进行诊断的准确率达到97%以上，对比未经降维处理方法的准确率明显提升。本文证实了以上方法对电动矿卡电机故障诊断的准确性。

关键词： 电动矿卡；主成分分析；随机森林模型；故障诊断

中图分类号： TM 307　　**文献标志码：** A　　**文章编号：** 1672-5581（2025）02-0329-05

Research on motor fault diagnosis of electric mining truck based on PCA-random forest

PENG Qian, YANG Chenhan

（School of Mechanical and Automotive Engineering, Xiamen University of Technology, Xiamen 361024, Fujian, China）

Abstract: Aiming at the problem of motor failure caused by complex factors interacting and interfering during the operation of pure electric mining trucks, a method based on principal component analysis (PCA) and random forest (RF) is proposed for predictive diagnosis. A dataset is constructed based on the actual collected motor failures of electric mining trucks, and the eigenvalue extraction and dimensionality reduction of the failure data are carried out using principal component analysis to reduce the dimensional redundancy of the data; the random forest prediction model is used to train and test the dimensionality-reduced data, and to predict the motor failure categories. The results show that the accuracy of motor fault type diagnosis using PCA-RF method reaches more than 97%, which is significantly improved compared with the accuracy of the method without dimensionality reduction processing. The accuracy of the above method for motor fault diagnosis of electric mining trucks is confirmed.

Key words: electric mining truck; principal component analysis; random forest model; fault diagnosis

现代化矿山生产过程中，绿色开采、智能开采是矿业领域发展的重要趋势[1]。其中，纯电动卡车不仅可以节约石油消耗，还可以实现减少行驶中的碳排放量，逐渐成为矿山的主要运输车辆。电机作为纯电动矿卡的唯一动力源，在长期连续作业中，由于受到外部环境以及电气系统的影响容易出现各类故障。这些故障严重影响电动矿卡的运行可靠性和安全性，如果不能及时诊断和处理故障，就可能造成严重的生产事故和财产损失[2]。因此，如何实现电机故障的快速识别与分类，对电动矿卡的安全可靠运行具有十分重要的工程价值。

目前，针对电机故障的诊断有定性和定量分

基金项目：福建省科技厅高校产学合作资助项目（2020H6021）；福建省科技厅STS计划资助项目（2023T3001）
作者简介：彭 倩（1982—），男，教授，博士。E-mail：pengqian603@163.com

析。定性分析有专家系统诊断和定性仿真,史强等[3]建立了完整的电机电源系统并构建了故障树模型,适用于电机电源系统中的逻辑循环截断问题;张栋良等[4]提出了一种模糊、本体和贝叶斯网络三者结合的故障诊断方法,较大程度上简化了贝叶斯网络模型,对电机组进行故障诊断获得了较高的准确性。此类定性分析虽然能够诊断大部分的故障类型并分析潜在原因,但是故障隐性影响因素无法明确指出。蒙康等[5]提出了一种基于单分类支持向量机的故障预警模型,此模型可以准确识别风电齿轮箱系统故障,并且建立了实际数据与物理模型的关系,具有较强的可解释性。王进花等[6]提出一种基于随机变分推理贝叶斯神经网络的故障诊断方法,根据少量数据获得较可靠的模型,具有较高的诊断性能。李兵等[7]利用改进随机森林算法对电机轴承故障进行诊断,并通过西储大学(Case Western Reserve University, CWRU)轴承数据集和现场诊断试验进行验证,在诊断准确率和漏报率上比传统算法更具优势。

综合上述分析,目前关于电机故障的诊断研究可通过不确定性问题分析来推导可能出现的故障原因,也可以通过机器学习从分析数据与故障的映射关系来建立模型,对故障进行准确诊断。然而,矿山工程车辆的使用环境相对复杂、严苛,故障影响因素与城市道路车辆存在较大的差异,加上电机故障数据获取复杂,且故障类型几乎不存在单一诱因,收集的数据维度较大,特征提取困难。因此,需要对采集的故障数据进行特征提取,然后基于处理后的特征数据进行故障分类模型训练。

本文提出了基于主成分分析(principal components analysis, PCA)和随机森林(random forest, RF)——PCA-RF方法对电动矿卡的电机故障类型进行诊断预测,通过主成分分析对收集的真实工况数据进行降维处理;在信息损失极少的情况下选择降维后的特征集,结合RF方法对分类问题求解精度高的特性,建立了一种电机故障分类诊断模型。实验结果表明,该模型能够精确地预测测试集样本中的电机故障类型。

1 数据集概况与数据处理

1.1 数据来源

实验数据来源于真实工况下的纯电动矿山卡车,收集的数据按照国家颁布的标准文GB/T 32960.3—2016[8]进行采集。试验矿山卡车电机的额定功率430 kWh,峰值功率为55 kWh,采集周期为210 d,分别在两种不同地区工况下进行行驶数据采集。数据采集频率为1 Hz,符合国家标准文件的采集要求。数据通过控制器局域网总线(controller area network, CAN)传输,实际产生行驶数据多达几百万条,其中电机故障数据有2 546组。经过诊断分析,故障大致可以分为驱动绝缘栅双极晶体管(insulated gate bipolar transistor, IGBT)电流过流、驱动IGBT电压欠压、常规过流故障、低压欠压故障、逐波限流故障、直流母线欠压。

1.2 PCA数据降维

由于数据量较大,且收集的类型维度较高,存在冗余信息,因此在尽量不损失数据信息的情况下,提取重要特征减少数据维度就成为首要解决的目标。PCA是一种常用的数据降维和特征提取方法[9]。每个主成分都是原始特征的线性组合,它们的重要性按照其对总体方差的贡献程度排序。使用PCA减少需要分析的指标,同时尽量减少原指标包含信息的损失,使数据降维,去除噪声。PCA的相关数学模型如下:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{N}x_i \tag{1}$$

$$S^2=\frac{1}{n-1}\sum_{i=1}^{N}(x_i-\bar{x})^2 \tag{2}$$

$$\mathrm{Cov}(X,Y)=E[(X-E(X))(Y-E(Y))]=\frac{1}{n-1}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) \tag{3}$$

式中:$\bar{x}$为样本的平均值;S^2为样本方差;$\mathrm{Cov}(X,Y)$为协方差。

对实验数据进行预处理,先对原始数据进行维度选择,在保持数据关系的基础上,让数据向量之和为0:

$$x_i \leftarrow x_i-\bar{x} \tag{4}$$

计算相关系数矩阵。所有元素x_i在该标准化方向v上的投影为

$$v^{\mathrm{T}}x_1, v^{\mathrm{T}}x_2, \cdots, v^{\mathrm{T}}x_n \tag{5}$$

将投影到标准化方向v上的元素带入方差中得到

$$\sigma^2=v^{\mathrm{T}}\left(\frac{1}{N}\sum_{i=1}^{N}x_ix_i^{\mathrm{T}}\right)v \tag{6}$$

相关系数C为

$$C=\frac{1}{N}\sum_{i=1}^{N}x_ix_i^{\mathrm{T}}=\frac{1}{N}[x_1,x_2,\cdots,x_i]\begin{bmatrix}x_1\\x_2\\\vdots\\x_i\end{bmatrix} \tag{7}$$

$$\sigma^2=v^{\mathrm{T}}Cv \tag{8}$$

在数据集中寻找贡献较大的特征向量,并将特征值按降序排序,同时计算贡献率和累积贡献率。

可得到特征向量 $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_k$ 对应的特征值 $\lambda_1, \lambda_2, \cdots, \lambda_k$：

$$C\boldsymbol{v}=\lambda\boldsymbol{v} \tag{9}$$

最后，将所有类别的数据投影到相应的特征向量，得到处理后的降维数据为

$$x_i=\sum_{j=1}^{k}(x_i^{\mathrm{T}}\boldsymbol{v}_j)\,\boldsymbol{v}_j \tag{10}$$

1.3　随机森林算法

RF是一种集成学习算法，主要用于分类和回归问题[10]。它由多个决策树组成，通过对每棵决策树的预测结果进行平均或投票生成最终的预测结果。该方法在分类问题中取得不错效果，对于分类问题采用投票的方法，得票最多的子模型的分类即为最终的类别，如图1所示。

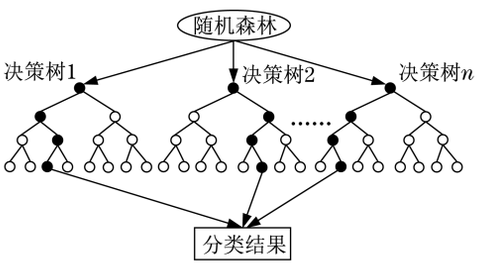

图1　随机森林分类预测

Fig.1　Random forest classification prediction

首先从随机向量中抽取训练集，确定输入向量 $\boldsymbol{X}$ 与输出向量 $\boldsymbol{Y}$[11-12]，将输入作为模型的基础，建立多棵决策树，通过对 n 棵决策树 $\{g(\theta_n, \boldsymbol{X}_N)\}$ 取平均数，则预测输出为 $g(\boldsymbol{X})$，令 $g(\boldsymbol{X})$ 的均方泛化误差为

$$E_{X,Y}[\boldsymbol{Y}-g(\boldsymbol{X})]^2 \tag{11}$$

当建立的决策树数量接近于无穷大时，泛化误差为

$$E_{X,Y}[\boldsymbol{Y}-\bar{g}(\boldsymbol{X},\theta_n)]^2 \to E_{X,Y}[\boldsymbol{Y}-E_\theta(\boldsymbol{X},\theta_n)]^2 \tag{12}$$

式(12)右边表示为随机森林的泛化误差，每棵决策树的平均泛化误差 PE^* 表示如下：

$$PE^*=E_\theta E_{X,Y}[\boldsymbol{Y}-g(\boldsymbol{X},\theta_n)]^2 \tag{13}$$

$$PE^*\leqslant\frac{\rho(1-s^2)}{s^2} \tag{14}$$

式中：s 为每棵决策树的分类能力；ρ 为不同决策树之间的关联性。

对于回归问题，采用上述公式进行计算并输出结果。而本文的故障诊断属于分类问题，随机森林算法可以通过训练多个基分类器，结合少数服从多数的投票原则，最终确定票数最多的类别作为分类结果。

RF的主要思想是引入了随机性，从而减小了模型的方差，并且通过模型的平均化，降低了模型的偏差，能够有效处理高维数据和特征较多的情况，同时具有很好的抗噪能力和准确性。

2　基于优化数据维度随机森林的电动矿卡电机故障分类流程

为准确识别电动矿卡的电机故障，将经过PCA降维后的数据通过随机森林算法建立故障分类模型，对电机故障的不同信息特征进行融合，并对故障类型分类，流程如图2所示。

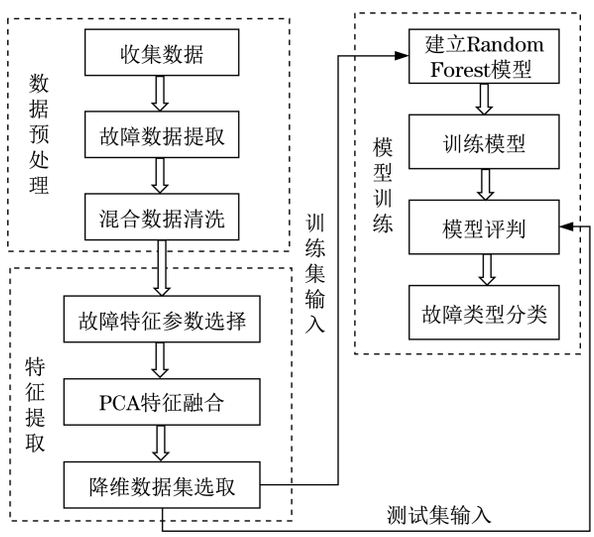

图2　基于优化数据维度随机森林的电动矿卡电机故障分类流程

Fig.2　Flowchart of electric mining truck motor fault classification based on random forest with optimized data dimensions

主要的诊断步骤如下：

步骤1　选择电动矿卡中的电机故障样本作为数据集，选取能够反映电机相应故障的特征数据作为故障分类特征变量。

步骤2　对清洗后的特征数据集进行主成分分析，计算出贡献度高的特征，并选择特征累计贡献率超过95%以上的向量。

步骤3　将处理好的特征向量进行随机排序，并随机提取70%数据作为模型的训练集，剩余的数据作为考核模型准确度的测试集以验证模型的准确率与鲁棒性。

步骤4　将没有经过降维处理的数据进行模型训练，对比两者之间的差异。

3　实际应用与结果分析

3.1　基于PCA的数据降维及特征选择

本研究采集到的故障类型为6种，故障类型数

据分别为IGBT电流过流、驱动IGBT电压欠压、常规过流故障、低压欠压故障、逐波限流故障、直流母线欠压。将这些故障定义为故障1～6,数据集维度为46,样本数量为2 546组。根据故障特性分析故障产生的原因与海拔、电流、电压、温度等因素有关。删除重复值筛选相关故障特征得到12维数据集,样本量为1 297。利用PCA算法对故障数据集进行降维,将故障特征维度从原有的12个维度降至6个维度。部分数据经PCA降维后的特征分布如图3所示。

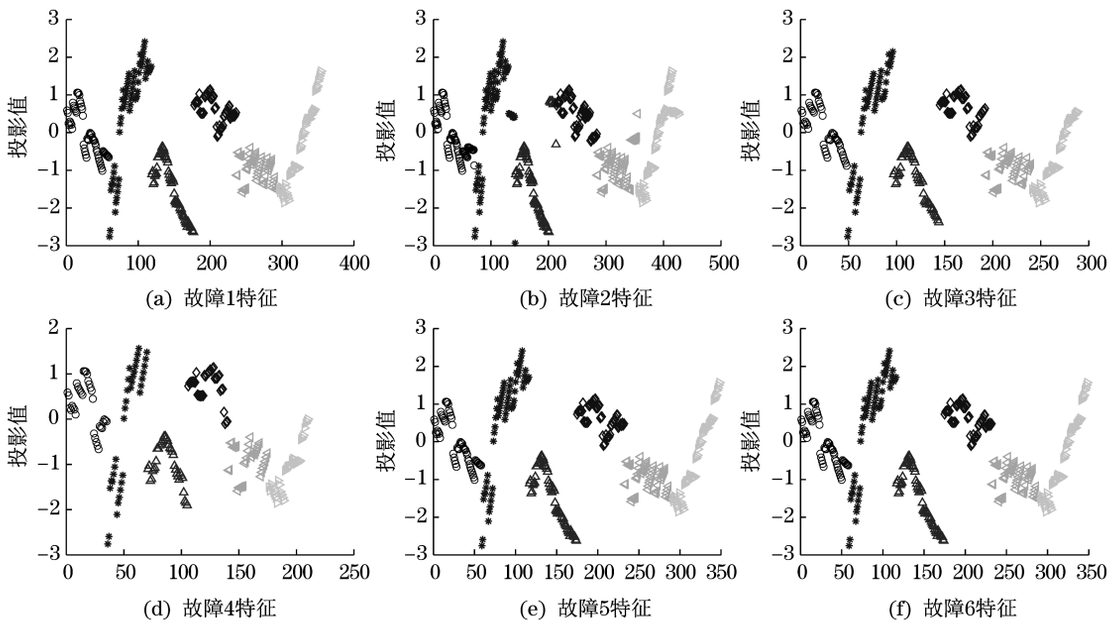

图3 故障降维特征分布

Fig.3 Distribution of fault downgrading features

根据图3的结果可知,进行PCA数据处理后,6种故障类型的各个特征数据比较分散,且特征数据分布大致相同。由于经过处理后的数据没有了实际的物理意义,所以将从左至右将特征属性定义为1～6,特征属性1、2、3、6数据出现明显的区分,特征属性4、5出现比较严重的重叠。根据PCA贡献率的计算结果显示,当主成分为6时累计贡献率已经达到95%以上,具体贡献率见表1。

表1 主成分特征值和贡献率

Tab.1 Principal component eigenvalues and contributions

主成分	初始特征值		
	特征值	贡献率/%	累计贡献率/%
1	3.491 1	29.09	29.09
2	3.204 2	26.70	55.79
3	2.699 2	22.49	78.29
4	1.325 6	11.05	89.33
5	0.452 5	3.77	93.11
6	0.306 4	2.55	95.66
7	0.240 3	2.00	97.66
8	0.136 3	1.14	98.80

3.2 故障类型预测

通过PCA降维后的数据集按照7:3分为训练集和测试集,将训练数据集输入到RF中进行模型构建,测试集数据则作为模型诊断故障的评判指标。RF中不同决策树对误差的影响如图4所示。

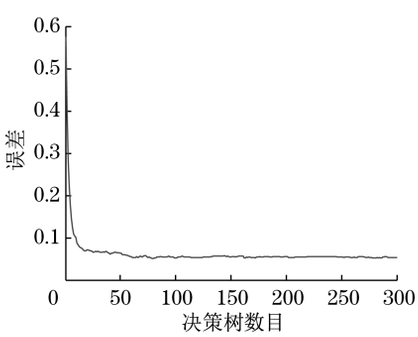

图4 决策树与误差的关系

Fig.4 Decision tree versus error

对不同决策树数量进行测试比对,以模型的识别结果误差和分类过程的计算时间作为指标,经过对比得到最优的决策树数量为150。将测试集输入构建完成的模型中结果表明,分类准确率为97.63%。其中,IGBT电流过流、常规过流故障、低压欠压故障、直流母线欠压4种故障的分类准确率均为97%以上,驱动IGBT电压欠压和逐波限流故

障则比较容易诊断为其他故障，但总体准确率也达到95%。有多个驱动IGBT电压欠压和逐波限流故障样本被误诊断为其他故障。具体分类情况如图5所示。

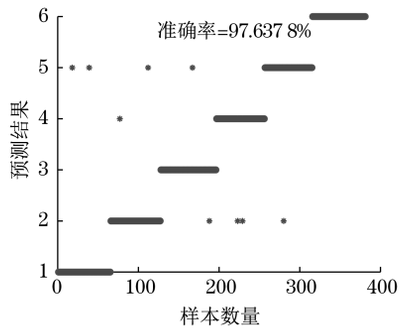

图5　故障分类的预测结果

Fig.5　Predicted results of fault classification

3.3　对比分析

为了验证该方法的准确性和有效性，将未经过PCA降维的数据输入随机森林模型进行预测，数据集与上文一样按照7∶3划分训练集和测试集。分类结果如图6所示。由图可知，没经过降维处理的数据诊断准确率只有85.5%，出现了较多次数的故障类型的误判。

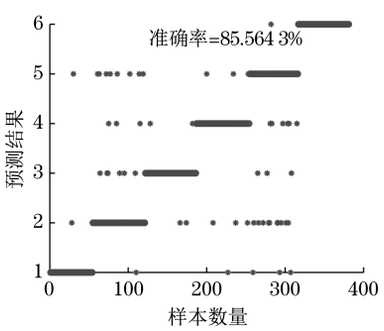

图6　随机森林的故障分类结果

Fig.6　Fault classification results for random forest

故障1～5（IGBT电流过流、驱动IGBT电压欠压、常规过流故障、低压欠压故障、逐波限流故障）的误判对比PCA-RF方法有较大差距。由此得出，对原始数据降维后构建的特征数据集，能够比较细致地反映电机故障类型，分类准确率明显提升。

4　结论

本文提出一种基于主成分分析优化维度，利用RF进行电机故障诊断的方法，通过实车故障数据验证了6种故障状态的识别，得到了以下的结论：

（1）基于PCA-RF的预测分类模型降低了原始数据的维度，并消除了特征之间的相关性，从而有效减少了噪声对模型的干扰，同时减少了单棵树过拟合的风险。这使得模型更加稳健，泛化能力较强，能够在复杂的分类问题中表现出色。

（2）实例分析结果表明，该方法可快速有效分类出驱动IGBT电流过流、驱动IGBT电压欠压、常规过流故障、低压欠压故障、逐波限流故障、直流母线欠压6种故障类型，分类准确率可达97.63%。

（3）经过验证，基于PCA-RF的预测分类模型适用于处理高维故障数据。与传统的RF算法相比，该方法在分类准确率和计算时间方面取得了更好的结果。因此，该模型可以进一步应用于电动矿卡电机在线故障诊断，为实现准确可靠的故障诊断提供支持。

参考文献：

[1]　顾清华，李学现，卢才武，等. “双碳”背景下露天矿智能化建设新模式的技术路径[J]. 金属矿山，2023(5)：1-13.

[2]　CUI S，YIN Y，WANG D，et al. A stacking-based ensemble learning method for earthquake casualty prediction[J]. Applied Soft Computing，2021，101(1)：107038.

[3]　史强，王晓峰，马帅，等. 丧失厂外电的柴油发电机故障树逻辑循环截断[J]. 核电子学与探测技术，2021，41(5)：774-782.

[4]　张栋良，汪刘峰，洪勤勤，等. 基于模糊贝叶斯网络的汽轮机组故障诊断研究[J]. 计算机仿真，2022，39(7)：476-481.

[5]　蒙康，滕伟，彭迪康，等. 运行机理与数据双驱动的风电齿轮箱系统故障预警[J]. 中国机械工程，2023，34(12)：1476-1485.

[6]　王进花，岳亮辉，曹洁，等. 基于随机变分推理贝叶斯神经网络的发电机轴承故障诊断[J]. 控制与决策，2023，38(4)：1015-1021.

[7]　李兵，韩睿，何怡刚，等. 改进随机森林算法在电机轴承故障诊断中的应用[J]. 中国电机工程学报，2020，40(4)：1310-1319.

[8]　电动汽车远程服务与管理系统技术规范 第3部分：通信协议及数据格式：GB/T 32960.3—2016[S]. 北京：中国标准出版社，2016.

[9]　丁敬国，郭锦华. 基于主成分分析协同随机森林算法的热连轧带钢宽度预测[J]. 东北大学学报(自然科学版)，2021，42(9)：1268-1274，1289.

[10]　吕何，孔政敏，张成刚. 基于混合优化随机森林回归的短期电力负荷预测[J]. 武汉大学学报(工学版)，2020，53(8)：704-711.

[11]　余嘉熹，李奇，陈维荣，等. 基于随机森林算法的大功率质子交换膜燃料电池系统故障分类方法[J]. 中国电机工程学报，2020，40(17)：5591-5598.

[12]　程养春，张振亮. 基于随机森林的变压器多源局部放电诊断[J]. 中国电机工程学报，2018，38(17)：5246-5256.