文章编号：0253-374X(2024)01-0035-06　　**DOI**：10.11908/j.issn.0253-374x.23387

基于主成分分析和深度森林算法的S700K转辙机故障诊断

胡小晨[1]，郭　宁[1]，沈　拓[2]，董德存[2]
（1.同济大学 道路与交通工程教育部重点实验室，上海 201804；2.同济大学 上海市轨道交通结构耐久与系统安全重点实验室，上海 201804）

摘要：针对目前转辙机故障诊断准确性不高、效率低等问题，提出了一种基于主成分分析(PCA)和深度森林(gcForest)算法的故障诊断方法。对于S700K转辙机11种故障模式下的电流、功率曲线，采用主成分分析进行电流特征值特征简约，然后使用嵌入简约特征值的改进深度森林模型提高数据处理能力，增强模型内在特征代表性。结果表明，改进深度森林模型故障诊断准确率为97.62%，验证了该方法的有效性和优越性。

关键词：故障诊断；S700K转辙机；主成分分析(PCA)；深度森林(gcForest)算法
中图分类号：U284.92　　**文献标志码**：A

PCA-gcForest-based Fault Diagnosis of S700K Switch Machine

HU Xiaochen[1], *GUO Ning*[1], *SHEN Tuo*[2], *DONG Decun*[2]
(1. Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 201804, China; 2. Shanghai Key Laboratory of Rail Infrastructure Durability and System Safety, Tongji University, Shanghai 201804, China)

Abstract: To overcome the shortage of the existing fault diagnosis methods such as low accuracy and efficiency, a fault diagnosis method based on principal component analysis (PCA) and multi-grain cascade forest (gcForest) algorithm was proposed. PCA was used to simplify the current eigenvalue for 11 fault modes of S700K switch machine. And an improved gcForest model with the simpler eigenvalue embedded was used to improve the data processing capability and enhance the inherent feature representativeness of the model. The experimental results show that the fault diagnosis accuracy of the improved gcForest model is 97.62%, which verifies the effectiveness and superiority of the method.

Keywords: fault diagnosis; S700K switch machine; principal component analysis (PCA); multi-grain cascade forest (gcForest) algrithm

在铁路第六次提速和道岔系统日益复杂的情况下，S700K转辙机作为铁路信号的重要设备面临严峻的挑战。目前，转辙机的故障诊断主要通过人工观察微机监测系统所采集的电流或功率曲线实现，现场人员必须具备全面分析信息的能力和快速维护的经验。因此，对转辙机的智能化诊断显得尤为重要。

国内外学者对转辙机故障诊断方法进行了大量的研究。黄世泽等[1]采用费雷歇距离定义的相似函数进行故障诊断。许庆阳等[2]采用Fisher准则函数和主成分分析(PCA)进行特征提取，通过建立不同故障分类下的隐马尔科夫模型(HMM)进行故障诊断。孔令刚等[3]通过概率神经网络(PNN)提取功率曲线多域特征数据实现故障诊断。Ou等[4]采用主成分分析和线性判别分析进行特征约简，并使用支持向量机(SVM)对故障进行分类。池毅等[5]将一维卷积神经网络应用于转辙机故障诊断。王瑞峰等[6]将灰色关联理论与神经网络相结合，实现转辙机故障识别。Ou等[7]基于不平衡监测数据，提出了一种基于贝叶斯估计的道岔在线故障诊断方法，在高准确率的前提下减少了训练时间。赵盼等[8]采用贝叶斯元学习方法，无须额外扩充小样本数据集，实现对多种型号的转辙机故障诊断。

收稿日期：2023-04-11
基金项目：国家重点研发计划(2022YFB4300501)
第一作者：胡小晨，博士生，主要研究方向为交通信息控制及系统安全。
E-mail：hxc20031514@163.com
通信作者：董德存，教授，博士生导师，主要研究方向为交通信息控制及系统安全。
E-mail：ddc@tongji.edu.cn

论文拓展介绍

作为机器学习方法,集成学习的主要思想是通过特定的规则整合各种学习结果,从而获得比单一学习者更好的学习性能。在综合学习和深度学习的基础上,Zhou等[9]提出了深度森林(gcForest)模型,并将其引入级联框架,生成具有更丰富学习能力的深度森林模型。目前深度森林算法已经广泛用于图像处理、时间序列预测等领域。Liu等[10]将深度森林算法应用于水轮机故障诊断。结果表明,此方法的诊断精度优于现有方法,对噪声具有较好的鲁棒性,并且不受训练数据量的限制。Qin等[11]将XGBoost(extreme gradient boosting)和LightGBM(light gradient boosting machine)替代级联森林的原有分类器,对滚动轴承进行故障诊断。结果表明,该方法能准确识别出不同类型故障,同时具有非常少的超参数和非常低的计算机硬件要求。Zhang等[12]采用改进深度森林算法和案件推理对ZYJ7道岔进行故障诊断。由于一些故障集具有相似的特征,因此基于案件推理能更好地区分故障。实验结果表明,在数据有限时精度优于现有方法。

采用主成分分析对S700K转辙机的三段电流曲线进行特征提取,然后将简约特征嵌入功率曲线深度森林模型中进行故障诊断。经过现场实际数据验证,该方法可有效提高故障诊断的精度与效率。

1 S700K转辙机动作过程分析

S700K转辙机是高速铁路和提速段常用的交流式转辙机。不同类型转辙机的输出功率规律相似,S700K转辙机具有一定的代表性。

1.1 S700K转辙机正常工作状态

S700K转辙机正常运行时的电流、功率曲线如图1所示。S700K转辙机运行包括3个阶段,依次为启动、转换、指示。第一阶段是开关机启动,道岔开锁需要克服较大阻力;第二阶段是开关轨道移动到另一个基本轨道的过程以及道岔锁紧;第三阶段是使用低功耗指示电路指示转换。

1.2 S700K转辙机故障工作状态

S700K转辙机具有复杂的机电结构,长时间暴露在户外环境中并需要频繁拉动,根据前期研究[13]再结合现场数据,S700K转辙机的常见故障如表1所示。

2 基于主成分分析的特征参数提取

主成分分析[14]是一种常见的数据分析方法,常

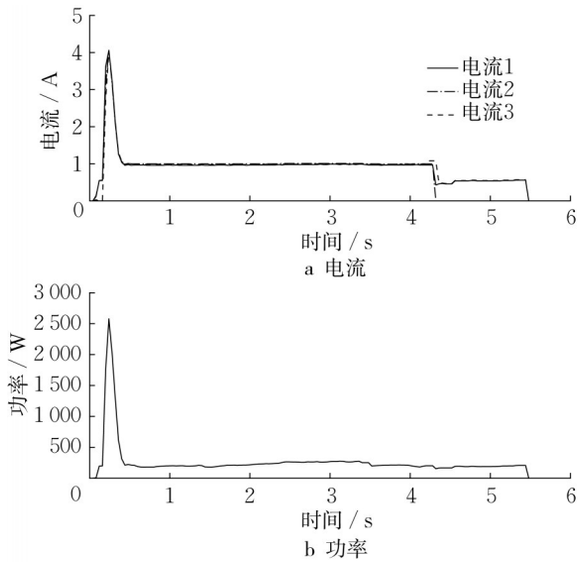

图1 S700K转辙机正常运行时的电流、功率曲线

Fig.1 Normal current and power curves of S700K switch machine

表1 S700K转辙机故障类型

Tab.1 S700K switch machine fault types

故障编号	故障类型
1	机械堵塞
2	滑椅位置不正确
3	开关电路阻抗异常
4	开关电路触点不良
5	指示电路中缺相保护装置异常
6	指示电路阻抗异常
7	启动电路中的电气继电器开关失败
8	供电中断
9	缺相保护装置故障
10	无法锁闭
11	指示杆块在间隙中

用于高维数据的降维,以提取数据的主要特征分量。该方法在信号处理、模式识别、数字图像处理、故障诊断等领域已经得到了广泛应用。

主成分分析通过线性转换将现有特征转化为新特征。新特征根据数据集的方差进行排序,这意味着具有最高表示能力的特征被选择用于分类。

假设矩阵$\boldsymbol{X}=\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix}$中包含$m$个样本和$n$维特征,则主成分分析基本步骤如下:

第1步 将$\boldsymbol{X}$中心化,即分别求出每个特征的平均值,计算式如下所示:

$$\mu_i=\frac{1}{m}\sum_{j=1}^{m}x_{ji} \tag{1}$$

式中:μ_i为第i维特征在m个样本上的平均值;x_{ji}为第j个样本的第i维特征值。对于所有的样例都减去

对应的均值，得到中心化后的矩阵$\boldsymbol{X}_0$，计算式如下所示：

$$\boldsymbol{X}_0=\begin{bmatrix} x_{11}-\mu_1 & \cdots & x_{1n}-\mu_n \\ \vdots & & \vdots \\ x_{m1}-\mu_1 & \cdots & x_{mn}-\mu_n \end{bmatrix} \tag{2}$$

第2步 计算中心化后的矩阵$\boldsymbol{X}_0$的协方差矩阵

$$\boldsymbol{X}_{\mathrm{cov}}=\frac{1}{n}\boldsymbol{X}_0\boldsymbol{X}_0^{\mathrm{T}} \tag{3}$$

第3步 用奇异值分解(SVD)求出协方差矩阵的特征值及其对应的特征向量，如下所示：

$$\boldsymbol{X}_{\mathrm{cov}}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{4}$$

式中：$\boldsymbol{U}$是一个$m\times n$的方阵，由于内部向量是正交的，因此又称为左奇异矩阵；$\boldsymbol{\Sigma}$是一个$m\times n$的实对角矩阵，由于对角线上的元素为奇异值，因此又称为奇异值矩阵，；$\boldsymbol{V}$是一个$n\times n$的矩阵，由于内部向量也是正交的，因此又称为右奇异矩阵。

第4步 将特征值从大到小排序，选择其中最大的k_{p}个，然后将其对应的特征向量分别作为行向量组成特征向量矩阵$\boldsymbol{X}_{\mathrm{e}}$。

第5步 将原始数据转换到k_{p}个特征向量构建的新空间中，经主成分分析处理后得到具有k_{p}维特征的矩阵$\boldsymbol{X}_{\mathrm{new}}$，表达式如下所示：

$$\boldsymbol{X}_{\mathrm{new}}=\boldsymbol{X}_{\mathrm{e}}\boldsymbol{X} \tag{5}$$

在上述原理下，绘制降维后各成分的方差和随主成分个数的变化而变化的曲线，选取最优主成分个数。原始数据包含360维特征，计算不同降维维度k_{p}下的投影误差，得到如图2所示的曲线。可以看出，当方差和约为95%时，即降维后的特征值保留了原始特征值95%的信息，k_{p}的取值约为10，这样就完成了特征数据的最优主成分个数选取。

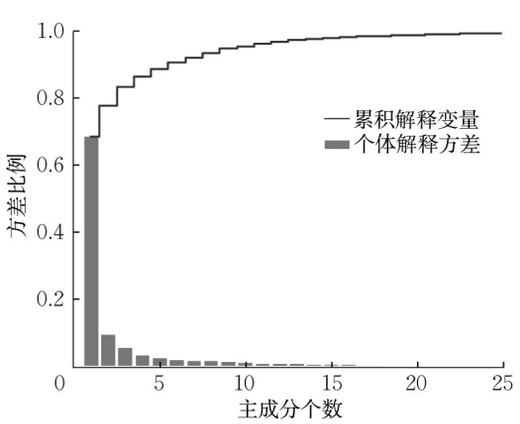

图2　主成分分析的最优主成分个数

Fig.2　Number of PCA optimal principal components

3　基于主成分分析的改进深度森林算法

3.1　多粒度扫描

多粒度扫描过程中采用不同大小的滑动窗口对原始输入特征进行提取，可以产生多个不同维度的特征，从而增强样本多样性。与原始输入特征相对应的特征实例经由一个完全随机森林和一个随机森林训练产生类概率向量，最后通过拼接得到转换特征向量。

如图3所示，多粒度扫描阶段分为特征扫描和特征转换2个过程。假设输入一个n维原始特征，滑动窗口大小为q维，滑动步长为k，滑动窗口扫描原始输入特征以提取特征信息，从而生成N个q维特征实例，计算式如下所示：

$$N=\frac{n-q}{k}+1 \tag{6}$$

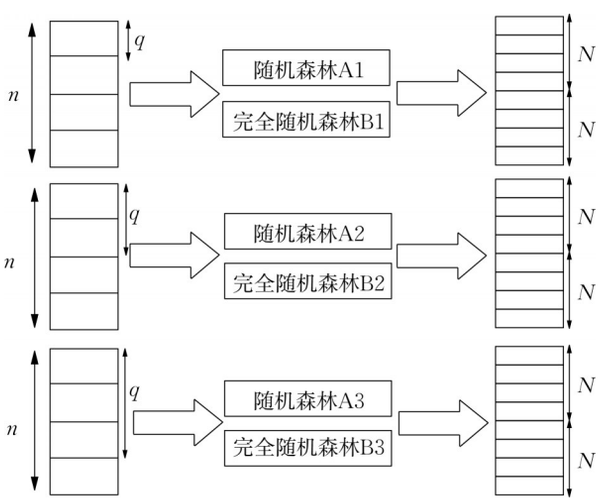

图3　多粒度扫描

Fig.3　Multi-grain scanning

经过随机森林和完全随机森林训练后，每个森林输出s维类概率向量，然后将所有类概率向量连接为L维转换特征向量，L的计算式如下所示：

$$L=2\left(\frac{n-q}{k}+1\right)s \tag{7}$$

通过多粒度扫描得到的转换特征向量规模高于原始输入特征，可以提取更多的特征信息。

3.2　级联森林

级联森林逐层训练可以增强特征信息的表征能力，每一层采用不同的分类器对深度森林算法集成学习十分重要。因此，级联森林的每层结构包含完全随机森林和随机森林2种不同的基础森林分类器。不同类型的森林分类器结合可以充分学习输入

特征向量的特征信息,从而提高模型的整体性能。 图4为级联森林过程。

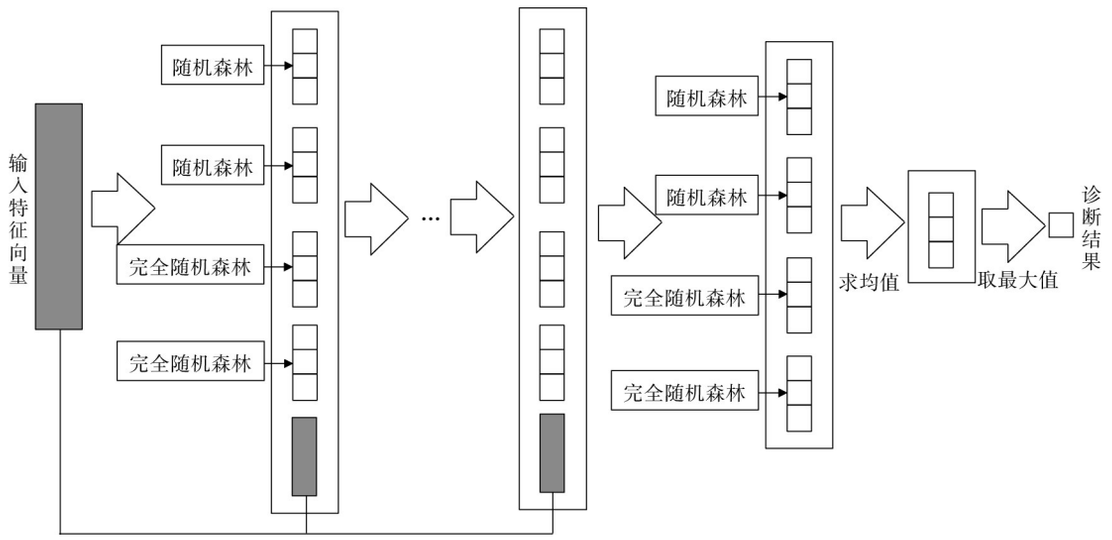

图4 级联森林

Fig.4 Cascade forest

级联森林的输入特征向量是多粒度扫描阶段最终产生的特征向量,然后在级联层学习并训练,输出的类概率向量在逻辑回归前都没有合并,最终产生的类概率向量和原始特征向量拼接作为下一层的输入。逐层训练后,最后一层级联森林产生的所有类概率向量通过逻辑回归产生最终类概率向量,从中取最大值,得到原始输入特征向量的最终分类。

为了避免级联森林训练产生过拟合现象,每个完全随机森林和随机森林的训练都通过K折交叉验证后产生类概率向量。由于模型的级联层数可以自适应确定,因此每个级联层生成的类概率向量是动态更新的。如果模型在连续3层训练中没有明显的性能改进,级联过程就自动终止。此过程可以提高故障诊断准确率和减少训练时间。

3.3 改进深度森林模型原理

级联森林每一层级的特征变化太小,一些重要的特征可能会被削弱,而且训练也需要大量的级联层[15]。为了解决该问题,将主成分分析后的电流特征值嵌入模型并融合到传统深度森林算法生成的变换特征中以提高模型性能。

传统深度森林模型使用功率特征值数据进行故障诊断,改进的深度森林模型通过主成分分析将三段电流特征值作为一个独立输入并与其他向量拼接,以此作为级联森林输入,避免了传统深度森林的削弱。

主成分分析可以有效地减少电流故障特征值的特征个数,可以很好地解决深度森林算法在处理具有长特性单样本数据时的特征冗余、算法运行效率低等问题。改进的深度森林模型如图5所示。

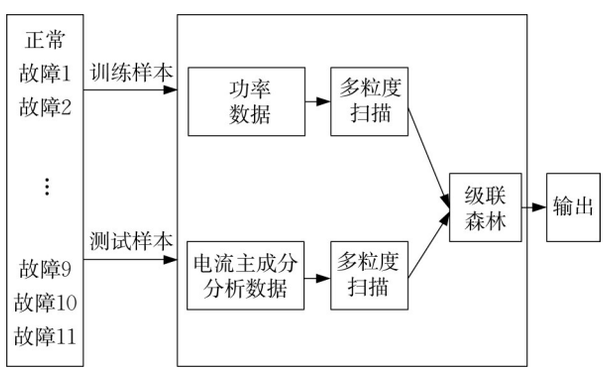

图5 改进深度森林模型

Fig.5 Structure of improved gcForest model

3.4 性能指标

通过分类精度(E_{accuracy})、查准率($E_{\text{precision}}$)、查全率(E_{recall})和F_1对故障诊断进行评价。其中,F_1为查准率和查全率的均值。E_{accuracy}、$E_{\text{precision}}$、E_{recall}和F_1的计算式如下所示:

$$E_{\text{accuracy}}=\frac{E_{\text{TP}}+E_{\text{TN}}}{E_{\text{TP}}+E_{\text{TN}}+E_{\text{FP}}+E_{\text{FN}}} \tag{8}$$

$$E_{\text{precision}}=\frac{E_{\text{TP}}}{E_{\text{TP}}+E_{\text{FP}}} \tag{9}$$

$$E_{\text{recall}}=\frac{E_{\text{TP}}}{E_{\text{TP}}+E_{\text{FN}}} \tag{10}$$

$$F_1=\frac{2E_{\text{TP}}}{2E_{\text{TP}}+E_{\text{FP}}+E_{\text{FN}}} \tag{11}$$

式中:E_{TP}、E_{TN}、E_{FP}、E_{FN}分别为真正例、真负例、假正例、假负例。

4　实验

为了验证该算法的有效性，以2016年—2018年广州铁路多个站点储存的S700K转辙机电流、功率曲线为实验数据，原始数据集由正常样本和11种故障类型样本组成。数据集共1 200个样本，即每种故障100个样本，每个样本长为360。

在改进深度森林算法中，首先需要确定如表2所示的5个参数，分别是训练样本大小、多粒度扫描时级联森林的决策树数量、多粒度扫描时扫描窗口大小、数据切片步长、每个级联森林中包含的决策树数量。调整步长是有意义的，因为它直接决定了森林层数和训练时间，从而影响模型的准确性。在S700K转辙机动作功率曲线的局部特征中，5个样本点是最小识别粒度，所以实验中步长设为5。此外，对于3个窗口的多粒度扫描，即40、90、120，在窗口的大小为40和120时模型可以识别锁阶段和转换阶段的特点，与窗口的大小为90时呈现完全不同的特性[12]。

表2　参数选择

Tab.2　Parameter selection

参数	数值
训练样本大小	360
多粒度扫描时级联森林的决策树数量	50
多粒度扫描时扫描窗口大小	[40,90,120]
数据切片步长	5
每个级联森林中包含的决策树数量	180

为了保证实验的准确性，交叉验证时采用取平均值的方法，每个案例运行10次。当训练数据占总数据的10%、20%、30%、40%时，验证12类数据的诊断效果，得到的结果如图6所示。

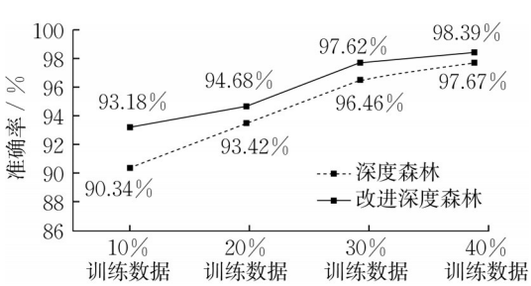

图6　10%~40%训练数据下S700K转辙机故障诊断精度对比

Fig.6　Comparison of fault diagnosis accuracy for S700K switch machine between 10%~40% training data

图7和图8是2种故障诊断方法在30%训练数据下的混淆矩阵。图7中，0代表正常情况，1~11分别代表11种故障类型。改进深度森林算法的故障诊断精度为97.98%，深度森林算法的故障诊断精度为94.88%。可以看出，改进深度森林算法的诊断性能更优。

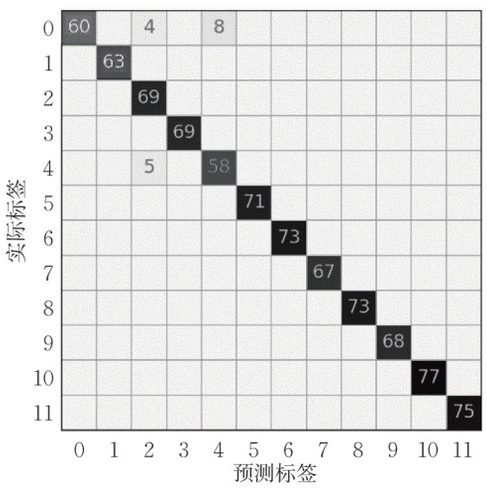

图7　30%训练数据下基于改进深度森林算法的S700K转辙机故障诊断方法的混淆矩阵

Fig.7　Confusion matrix of S700K switch machine fault diagnosis method based on improved gcForest algorithm at 30% training data

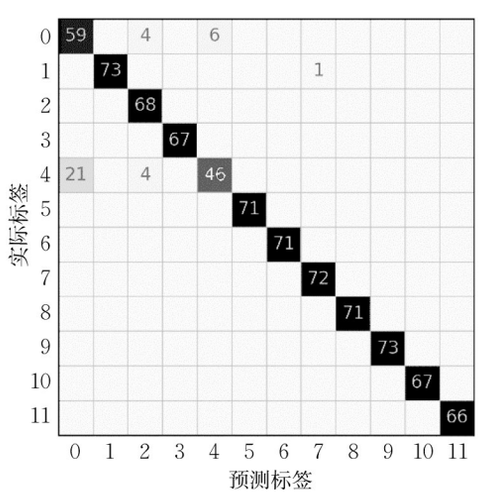

图8　30%训练数据下基于深度森林算法的S700K转辙机故障诊断方法的混淆矩阵

Fig.8　Confusion matrix of S700K switch machine fault diagnosis method based on gcForest algorithm at 30% training data

综上所述，基于改进深度森林算法的S700K转辙机故障诊断方法相比基于深度森林算法的S700K转辙机故障诊断方法在各比例训练数据、各健康状态下都有更高的诊断精度。

为了验证所提模型的有效性，将所提模型与现

有常用模型进行对比。表3为4种模型进行10次训练后得到的测试结果均值,分别为模型的精度和时间。与其他模型相比改进深度森林算法的故障诊断精度最高,达到了97.62%,并且在时间上也更短。

表3 改进深度森林模型故障诊断精度和其他模型对比

Tab.3 Comparison of fault diagnosis accuracy between improved gcForest model and other models

模型	精度/%	时间/s
改进深度森林	97.62	0.75
深度森林	94.46	0.89
SVM	93.33	1.20
随机森林	95.26	0.49

5 结语

针对高速铁路S700K转辙机的故障诊断问题,提出了一种基于主成分分析的改进深度森林故障诊断方法。首先使用主成分分析对电流数据进行特征简约,然后把简约后的特征值嵌入基于功率数据的深度森林模型。实验结果表明,与直接使用功率数据的深度森林故障诊断方法相比,提出的诊断方法具有更高的诊断精度,诊断精度达到97.62%。此外,还可以收集更多的数据以覆盖更多的故障类型,从而实现适用于现场环境的智能诊断方法,提高转辙机的维修效率。

作者贡献声明:

胡小晨:算法仿真与实验,论文写作。

郭 宁:思路梳理。

沈 拓:论文修改。

董德存:提出研究思路,参与研究方法。

参考文献:

[1] 黄世泽,陈威,张帆,等.基于弗雷歇距离的道岔故障诊断方法[J].同济大学学报(自然科学版),2018,46(12):1690.
HUAN Shize, CHEN Wei, ZHANG Fan, *et al*. Method of turnout fault diagnosis based on Frechet distance[J]. Journal of Tongji University (Natural Science), 2018,46(12):1690.

[2] 许庆阳,刘中田,赵会兵.基于隐马尔可夫模型的道岔故障诊断方法[J].铁道学报,2018,40(8):98.
XU Qingyang, LIU Zhongtian, ZHAO Huibing. Method of turnout fault diagnosis based on hidden Markov model [J]. Journal of the China Railway Society , 2018,40(8):98.

[3] 孔令刚,焦相萌,陈光武,等.基于多域特征提取与改进PSO-PNN的道岔故障诊断[J].铁道科学与工程学报,2020,17(6):1327.
KONG Linggang, JIAO Xiangmeng, CHEN Guangwu, *et al*. Turnout fault diagnosis based on multi-domain feature extraction and improved PSO-PNN [J]. Journal of Railway Science and Engineering, 2020,17(6):1327.

[4] OU D, XUE R, CUI K. A data-driven fault diagnosis method for railway turnouts [J]. Transportation Research Record: Journal of Transportation Research Record, 2019, 2673(4):448.

[5] 池毅,陈光武.基于一维卷积神经网络的实时道岔故障诊断[J].计算机工程与应用,2022,58(20):293.
CHI Yi, CHEN Guangwu. Real-time turnout fault diagnosis based on one-dimensional convolutional neural network [J]. Computer Engineering and Applications,2022,58(20):293.

[6] 王瑞峰,陈旺斌.基于灰色神经网络的S700K转辙机故障诊断方法研究[J].铁道学报,2016,38(6):68.
WANG Ruifeng, CHEN Wangbin. Research on fault diagnosis method for S700K switch machine based on grey neural network [J]. Journal of the China Railway Society, 2016,38(6):68.

[7] OU D, JI Y, ZHG R, *et al*. An online classification method for fault diagnosis of railway turnouts[J]. Sensors, 2020, 20(16):4627.

[8] 赵盼,王小敏,傅美君.基于贝叶斯元学习的小样本转辙机故障诊断[J].铁道科学与工程学报,2023,20(10):4008.
ZHAO Pan, WANG Xiaomin, FU Meijun. Few-shot switch machine fault diagnosis based on Bayesian meta-learning [J]. Journal of Railway Science and Engineering,2023,20(10):4008.

[9] ZHOU Z H, FENG J. Deep forest: towards an alternative to deep neural networks[C]// Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI). Melbourne: [s. n.], 2017:3553-3559.

[10] LIU X, TIAN Y, LEI X, *et al*. Deep forest based intelligent fault diagnosis of hydraulic turbine [J]. Journal of Mechanical Science and Technology, 2019, 33(5):1976.

[11] QIN Xiwen, XU Dingxin, DONG Xiaogang, *et al*. The fault diagnosis of rolling bearing based on improved deep forest [J]. Shock and Vibration, 2021, 2021:9933137.

[12] ZHANG Yao, XU Tianhua, CHEN Cong, *et al*. A hierarchical method based on improved deep forest and case-based reasoning for railway turnout fault diagnosis [J]. Engineering Failure Analysis, 2021,127:105446.

[13] 张钉,李国宁.基于改进WNN分析功率曲线的S700K转辙机故障诊断[J].铁道科学与工程学报,2018,15(8):2123.
ZHANG Ding, LI Guoning. Fault diagnosis of S700K switch machine based on improved WNN analyses power curve [J]. Journal of Railway Science and Engineering, 2018,15(8):2123.

[14] 邵怡韦,陈嘉宇,林翠颖,等.小训练样本下齿轮箱故障诊断:一种基于改进深度森林的方法[J].航空学报,2022,43(8):118.
SHAO Yiwei, CHEN Jiayu, LIN Cuiying, *et al*. Gearbox fault diagnosis under small training samples: an improved deep forest based method [J]. *Acta Aeronautica et Astronautica Sinica*, 2022,43(8):118.

[15] HU G Z, LI H F, XIA Y Q, *et al*. A deep Boltzmann machine and multi-grained scanning forest ensemble collaborative method and its application to industrial fault diagnosis [J]. Computers in Industry, 2018, 100:287.