DOI:10.20030/j.cnki.1000-3932.202203009

基于皮尔逊相关系数的滚动轴承混合域特征选择方法

肖 杨 李 亚 王海瑞 常梦容

(昆明理工大学信息工程与自动化学院)

摘 要 为解决单域特征难以全面准确反映轴承运行状态和故障诊断准确率低的问题，提出混合域特征集构建方法，分别提取时域、频域、时频域特征向量构建多元信息特征集。针对构建出来的混合域特征集线性相关性过高、维数高、冗余信息多及空间成本大等问题，运用皮尔逊相关系数(PCC)特征选择方法，对特征进行相关性计算，提取相关性较弱的低维特征输入随机森林分类器进行故障诊断。实验结果表明：基于PCC的滚动轴承混合域特征选择方法，不仅提高了分类性能，而且考虑了特征之间的相互作用，减少了信息的丢失，模型分类准确率可达97.32%，相比其他方法具有较为明显的优势。

关键词 特征选择 滚动轴承 混合域 冗余性 皮尔逊相关系数

中图分类号 TP14 **文献标识码** A **文章编号** 1000-3932(2022)03-0308-08

时域、频域、时频域各类特征作为具有显著类别差异信息的非平稳统计特征，能有效提高滚动轴承状态监测和故障诊断的性能和效率，因此研究人员在此方面开展了广泛的研究[1]。熊鹏博和王晓东提出了一种基于多时域特征与支持向量机(Support Vector Machines，SVM)的单向阀故障诊断方法[2]。马欣欣和郭敏将采集到的信号进行集合经验模态分解 (Ensemble Empirical Mode Decomposition，EEMD)，得到若干个固有模态函数分量(IMF)[3]，然后在前三阶IMF的基础上提取时域、频域和希尔伯特域特征，融合提取的特征组成多域特征向量组，最后送入支持向量机分类器中进行分类。彭涛等对原始信号分别生成时域、频域状态特征[4]，并利用多分辨率小波分解生成时频域状态特征，构建出144个表征原始振动信号特征的混合域特征集。李大江提出一种基于局部均值分解(LMD)和共空间模式(CSP)的时-频-空多域特征提取方法[5]。

上述研究中用于分析的数据集包含数百个特征(或属性)，其中大多数可能与故障诊断系统任务无关或冗余，因此，特征提取、特征降维和特征筛选显得尤其重要。戴豪民等采用加权最大相关最小冗余的特征选择方法[6]，选取7个有效特征向量，输入至SVM得到不错的效果。Tang X H等利用特征对特征的最大信息系数(MIC)得到的弱相关特征子集和特征对类别的最大信息系数(MIC)得到的强相关特征子集[7]，通过交集运算合并为最终的诊断特征集，在一定程度上减少了特征数量。白丽丽等利用拉普拉斯(LP)对能表征状态的特征进行选择[8]，将选择得到的数据输入到鲸鱼算法优化的SVM进行模式识别，证明了特征提取的有效性。但这些降维方法对轴承故障的特征集线性相关性过高且包含大量的冗余信息，所得到的低维空间对原始信号的解释具有一定的片面性。

为了解决单域特征难以表达原始信号的振

基金项目：国家自然科学基金项目(61863016)。

作者简介：肖杨(1994-)，硕士研究生，从事故障诊断的研究。

通讯作者：李亚(1978-)，副教授，从事计算机应用基础和工程技术的研究，59515091@qq.com。

引用本文：肖杨，李亚，王海瑞，等.基于皮尔逊相关系数的滚动轴承混合域特征选择方法[J].化工自动化及仪表，2022，49(3)：308~315.

动规律、高维特征容易发生过拟合并且容易引发维数灾难的问题，同时处理好冗余性、相关性问题，笔者提出了基于皮尔逊相关系数(Pearson Correlation Coefficient，PCC)的滚动轴承混合域特征选择方法，通过多元信息特征向量组确定一组高质量特征来进行稳定的预测。首先从原始信号中提取6个时域无量纲向量、10个时域有量纲向量、4个频域特征向量、6个小波变换特征向量和10个自适应噪声的完整集成经验模态分解(CEEMDAN)特征向量，结合提取出的特征参数，构造轴承故障混合域特征集。其次，运用PCC进行特征选择，对提取的混合域特征进行相关性分析，根据相关性，从特征集中剔除不相关和冗余的特征，提取出易于识别的低维主特征向量。最后将低维特征集导入到随机森林中作为模式识别的输入。

1　混合域特征集的构成

1.1　时域特征集

时域信息是以时间为变量描绘出信号的波形，作为衡量信号特征的重要指标[9]。时域信号包括量纲特征参数和无量纲特征参数[10]。笔者主要引入6个时域无量纲参数、10个时域有量纲参数，组成16维时域特征向量构成时域特征集，包括最大值、最小值、峰值、峰峰值、平均值、绝对平均值、方根幅值、方差、标准差、有效值、峭度、偏度、波形因子、峰值因子、脉冲因子和峪度因子。

1.2　频域特征集

频域信息是以频率为变量描绘出频率信号的幅度，作为衡量信号特征的重要指标。笔者通过提取4个常用的频域特征向量来构成频域特征集，包括平均频率、重心频率、均方根频率和频率标准差。在构造频域特征集之前，采用傅里叶分析对原始信号进行处理。

1.3　时频域特征集

1.3.1　小波时频域特征集

小波分解主要是以短时傅里叶变换的理论为基础通过小波函数对时间序列进行细致描述[11]，因此可在不同维度进行信号分析[12]。若滚动轴承在某一时刻发生突变，单独依靠原始信号并不能对故障点进行准确描述，需要从时间序列的不同维度对信号进行分析，包括整体性分析和局部分析，因此通过小波分解的方法能够对时间序列进行全面的刻画，从而准确定位故障振动冲击时刻[13]。

1.3.2　CEEMDAN分解时频域特征集

CEEMDAN方法通过自适应加入白噪声，克服了EEMD方法的模态混叠问题，获得了较好的模态分离谱，同时提高了运算效率[14]，因此笔者提取CEEMDAN分解后的各分量的时频域特征，包括排列熵和瞬时能量，确保达到一个良好的特征分析效果。

2　特征选择

特征选择的目的不仅仅是为数据降维，还要消除冗余和无关的特性[15]。通过度量特征间的相关性，可以消除冗余特征。两个特征之间的相关性越强，它们之间的冗余性和可替代性就越强[16]。此外，通过测量特征与类别之间的相关性，可以消除不相关的特征。特征选择一般包括3个步骤：

a. 搜索。在特征空间中搜索特征子集，每个子集被称为一个状态，由选定的特性组成。

b. 评价。输入一个状态(子集)，通过评价函数或预置的阈值输出评价值，使评价值达到最优值。

c. 分类。使用最终的特征集完成分类算法。

皮尔逊相关系数是由卡尔·皮尔逊提出的，定义为秩变量之间的相关系数[17]。对于容量为n的样本，将n个原始数据转换为等级数据，相关系数为：

$$r_{xy}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})}}$$

其中，r_{xy}表示两个变量x、y之间的线性相关程度，r_{xy}的值在−1和+1之间；$x=[x_1,x_2,\cdots,x_n]$，$y=[y_1,x_2,\cdots,y_n]$；$\bar{x}$、$\bar{y}$分别为x_i、y_i的平均值。

若$r_{xy}>0$，表示两个变量正相关，即一个变量的值越大，另一个变量的值也越大；若$r_{xy}<0$，则表示两个变量负相关；当$r_{xy}=0$时，表示x和y不相关。相关系数绝对值越大，相关性越强；相关系数绝对值越接近0，相关性越弱[18]。一般情况下，变量的相关强度由以下取值范围来判断：相关系数绝对

值在0.8~1.0,非常强相关;在0.6~0.8,强相关;在0.4~0.6,中等相关;在0.2~0.4,弱相关;在0.0~0.2,非常弱相关或不相关。

3 实验

3.1 轴承故障数据说明

实验采用的是美国凯斯西储大学轴承数据中心采集的轴承故障数据集。数据集包含正常、内圈故障、滚动体故障和外圈故障(6点钟方向)4种不同状态的数据,除正常数据外每种状态有3种故障深度类型,直径分别为0.177 8、0.355 6、0.533 4 mm,即共10类故障类别。轴承电机载荷为0,轴承转速为1 797 r/min。每类数据划分为115个分类样本,10类总共1 150个样本。训练集大小为700,即每类70个,测试集为450个,每类45个。分类情况见表1,其中IF、RF、OF分别为内圈、滚动体、外圈故障(6点钟方向)。

表1 轴承数据集描述

类别	故障直径/mm	训练集/测试集	标签
正常	0	70/45	1
IF1	0.177 8	70/45	2
IF2	0.355 6	70/45	3
IF3	0.533 4	70/45	4
RF1	0.177 8	70/45	5
RF2	0.355 6	70/45	6
RF3	0.533 4	70/45	7
OF1	0.177 8	70/45	8
OF2	0.355 6	70/45	9
OF3	0.533 4	70/45	10

3.2 实验特征提取

本实验中,每类原始时间序列的总长度为117 760,将其切分为115份,每份长度为1 024,分别提取每段时间序列的时域、频域以及时频域共36个特征。原始特征向量记为$A_1,A_2,\cdots,A_n$,n=1024,提取出新的特征向量表示为$B_1,B_2,\cdots,B_m$,m=36(m<n),f_i为其对应映射函数,可表示为:

$$B_i=f_i(A_1,A_2,\cdots,A_n),i\in[1,m]$$

1~10为有量纲向量时域特征,分别为最大值、最小值、峰值、峰峰值、平均值、绝对平均值、方根幅值、方差、标准差和有效值;11~16为无量纲向量时域特征,分别为峭度、偏度、波形因子、峰值因子、脉冲因子和裕度因子;17~20为频域特征,分别为平均频率、重心频率、频率均方根和频率标准差。

时频特征主要提取小波变换和CEEMDAN相关特征。其中,小波变换将原始振动信号进行3层分解,划分成8个子频带。由于前4个子频带包含了原始信号的大部分能量,因此提取前4个子频带小波尺度熵,再提取信号的小波能量谱熵和小波奇异熵,构成其中一个时频特征子集。

分别对10类信号原始信号进行CEEMDAN分解,得到8个固有模态分量,如图1所示。第1个分量信号的振动频率比其他分量的大,而第2~6分量相比其他分量振动频率更符合高次谐波的特征,所有分量能量大多集中在第2~6个分量中,并且在时间周期范围内具有正弦波的特性,属于有效分量。而剩余分量含有的能量信息较少,振动冲击特征不突出,与原始信号无太大关系,可视为无效分量。因此,选取第2~6分量,并提取出各个模态的排列熵和瞬时能量,作为整个故障诊断数据的一个子集。

3.3 实验分析

本次实验将提取到的混合域特征进行相关性分析,分别计算每个特征向量与其他特征的皮尔逊相关系数值,并求其绝对值。由于离群值对皮尔逊相关性分析较为敏感,若特征中存在离群点,则计算结果将小于实际计算结果,从而对特征分析产生错误的判断,图2为随机抽取部分特征散点分布图。

本实验用中位数对离群点进行替换处理,以确保皮尔逊相关系数的有效性。图2a显示未处理前皮尔逊相关系数值为0.434,由图明显可得出两个特征向量存在相关关系,但由于离群值的干扰,相关系数减小,使得计算结果与实际分布产生较大误差,由中位数替代之后,重新计算皮尔逊相关系数为0.632,计算结果符合特征规律。表2为图2a~d未处理和已处理的相关系数值变化情况。

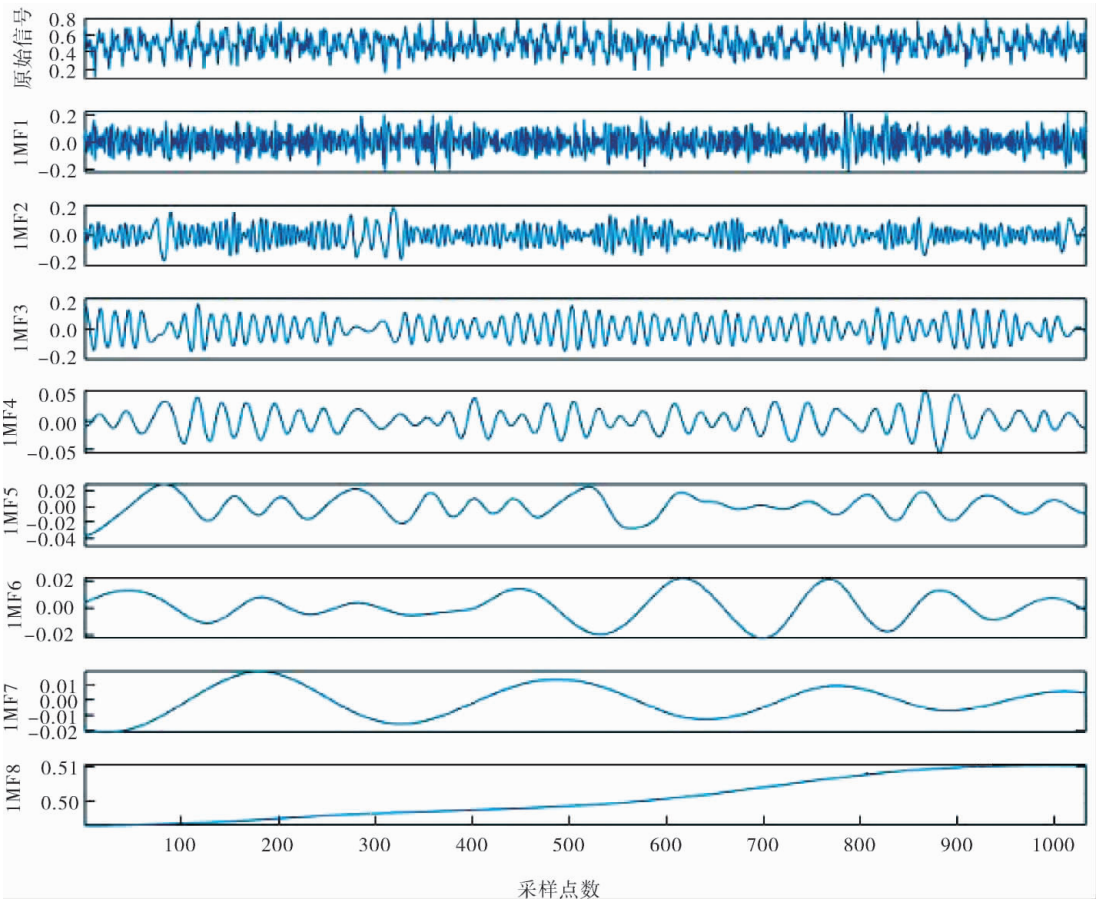


图1　CEEMDAN分解图

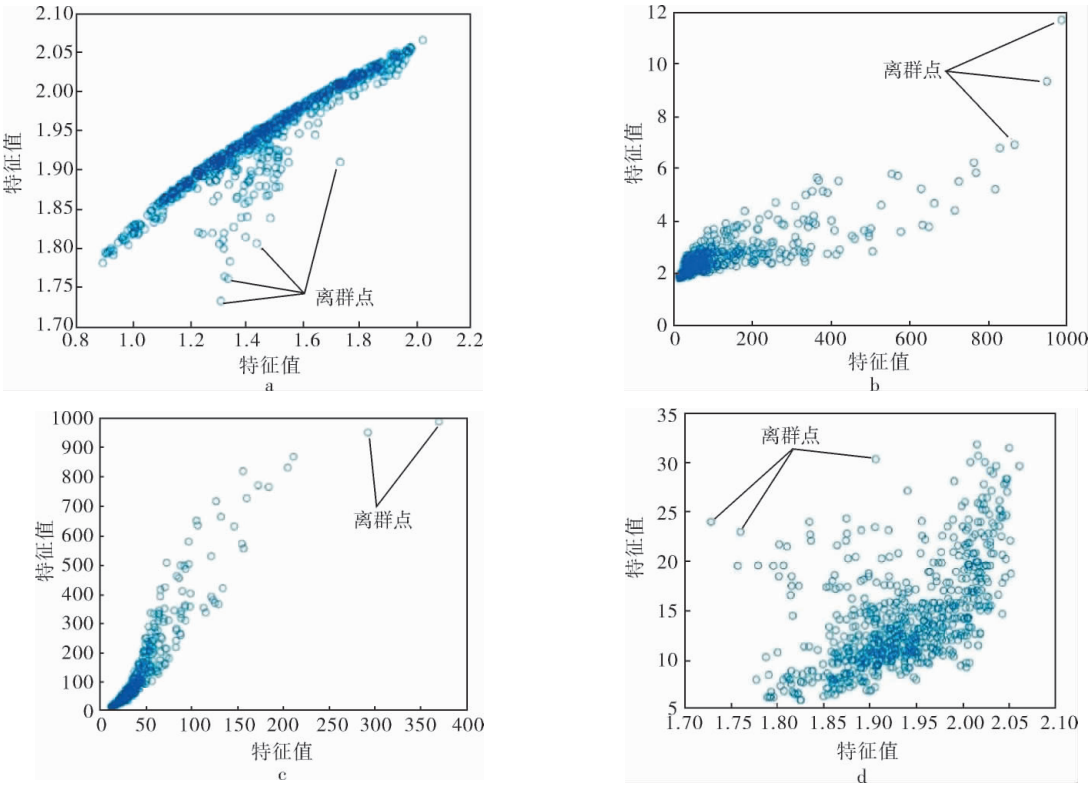


图2　随机抽取部分特征散点分布

表2 图2a~d相关系数在离群点处理前后的变化

图序号	相关系数	
	处理前	处理后
2a	0.434	0.632
2b	0.229	0.366
2c	0.245	0.378
2d	0.032	0.074

设置皮尔逊相关系数阈值为0.450,再分别统计每类特征与其他特征的相关性的强弱，大于0.450说明具有强相关性，小于0.450说明特征之间弱相关,并根据弱相关性大小进行排序。表3为计算的平均皮尔逊相关系数值,按由弱到强进行排序。

表3 计算的平均皮尔逊相关系数值

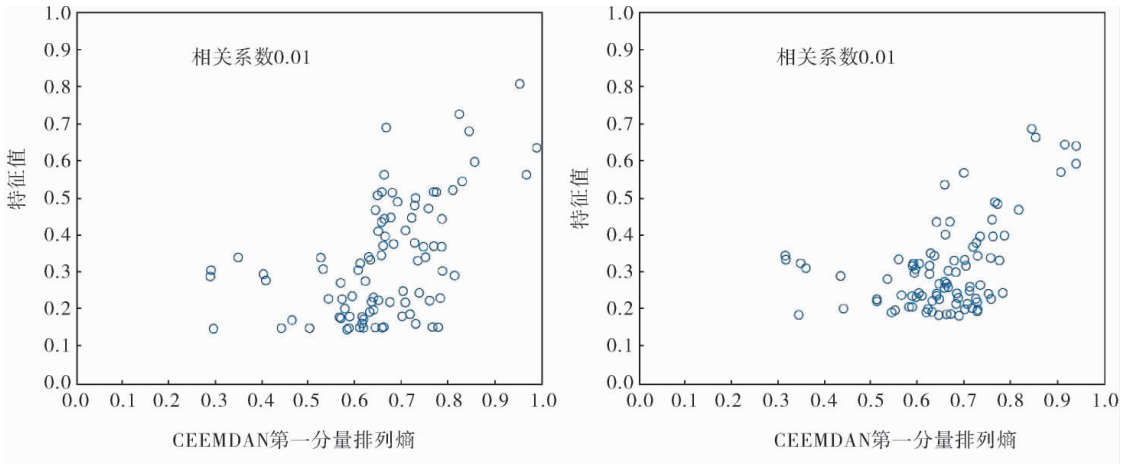
序号	数值	序号	数值
1	0.036	19	0.295
2	0.055	20	0.297
3	0.076	21	0.302
4	0.096	22	0.309
5	0.110	23	0.316
6	0.126	24	0.328
7	0.149	25	0.339
8	0.161	26	0.351
9	0.185	27	0.374

（续表3）

序号	数值	序号	数值
10	0.198	28	0.386
11	0.200	29	0.395
12	0.222	30	0.419
13	0.235	31	0.431
14	0.237	32	0.435
15	0.247	33	0.447
16	0.253	34	0.450
17	0.258	35	0.455
18	0.269	36	0.462

皮尔逊相关系数越大相关性越强,对于故障诊断越不利,因此需要筛选出相关性较弱的特征向量。第1个特征向量与其他特征向量的相关性最弱,相关系数平均值仅为0.036,接近于零,对于分类具有较好的表现。而相关性最高的几个特征向量的值已经超过阈值,表明其本身包含的特征信息与其他特征向量重复概率较大，可以剔除。

为了更直观地比较特征间的相关关系，同时选取具有最小相关关系的特征向量(CEEMDAN第一分量的排列熵值）和具有最大相关关系的特征向量(峰峰值)进行相关性实验,如图3、4所示,每个特征随机抽取4个特征与之进行分析,绘制散点图。为了排除随机实验的影响,本次实验共进行了5次,统计每次实验的平均相关系数值,结果见表4。



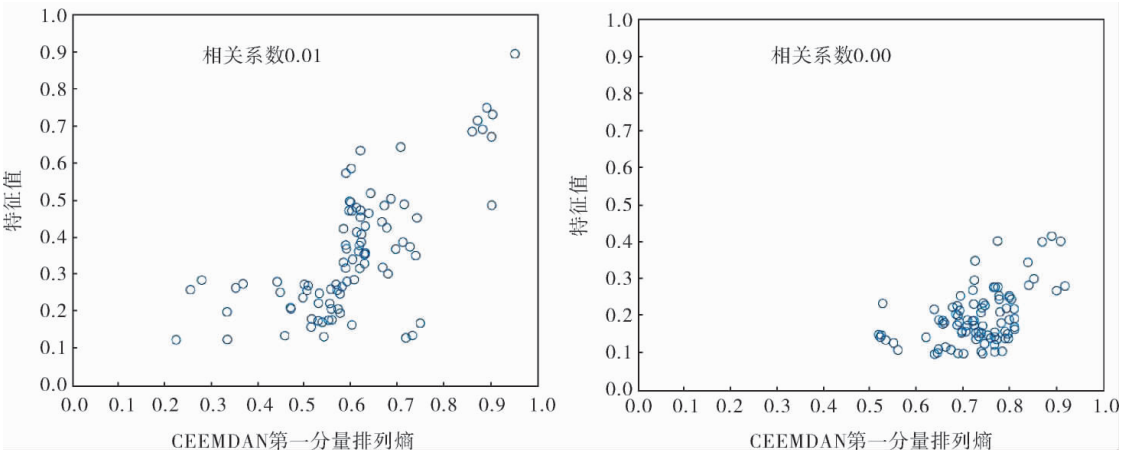


图3　CEEMDAN第一分量排列熵与随机特征相关关系散点图

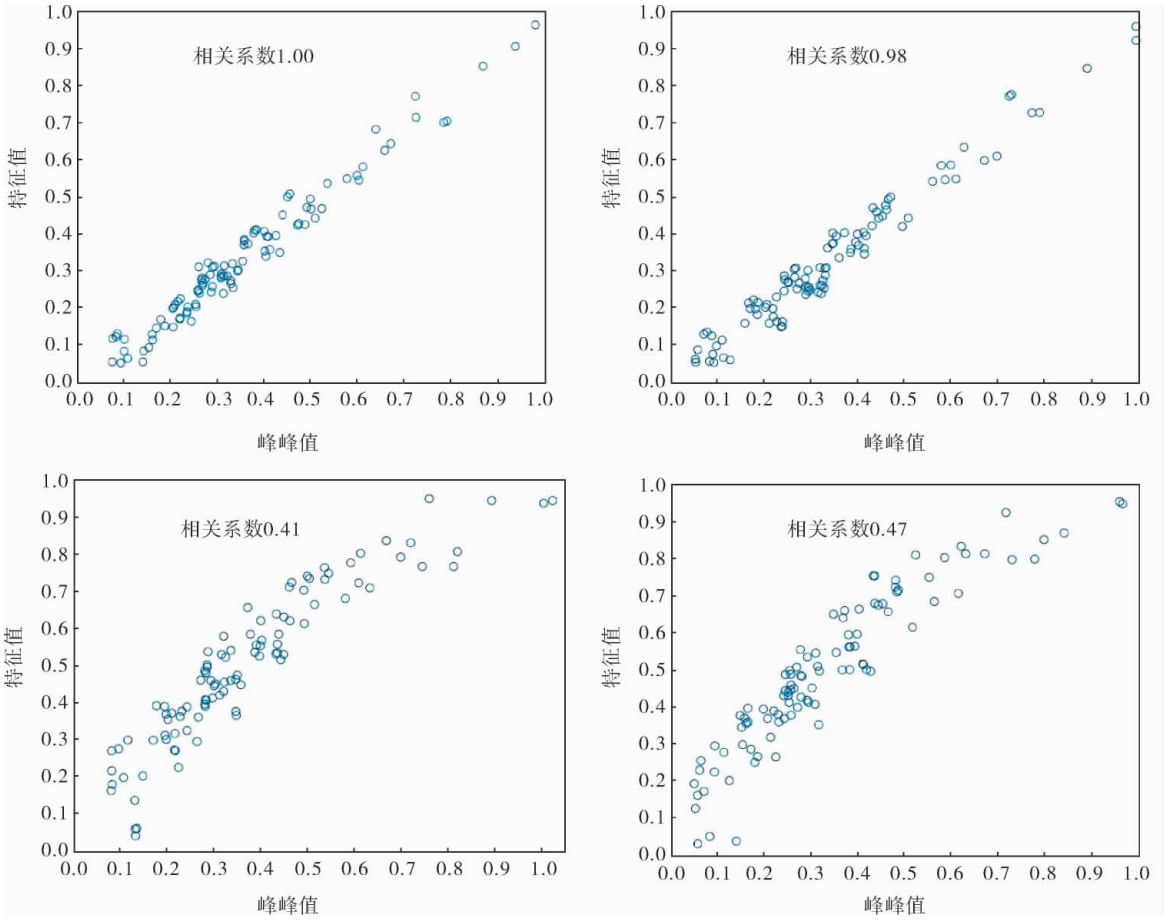


图4　峰峰值与随机特征相关关系散点图

由图3、4可知，基于CEEMDAN第一分量的排列熵与其他特征均无明显的相关性特点，点的分布规律较为均匀，大多数表现为弱相关关系，是较为理想的分类特征。而峰峰值与其他特征的分布特点是由左下角分布到右上角，呈现较强的正相关关系，点的分布拟合接近一条直线，不具有分类利用价值。

由表4数据可计算出峰峰值与其余全部特征

表4 5次实验平均相关系数

实验序号	平均相关系数	
	CEEMDAN第一分量排列熵	峰峰值
1	0.023 1	0.652
2	0.041 4	0.439
3	0.051 6	0.544
4	0.012 2	0.485
5	0.032 3	0.398

的平均相关系数S_{AVG}为0.462，已经大于0.450的阈值，且在进行随机实验时，相关系数浮动较大。而CEEMDAN第一分量排列熵的S_{AVG}仅为0.036 1，远低于峰峰值的，并且在进行随机抽取时，其实验结果均在平均值上下较小范围内浮动，具有较强鲁棒性。同时，将筛选后的特征向量集进行故障诊断，建立7个诊断模型进行准确率对比，模型分别为基于原始时域的故障诊断方法、基于原始频域的故障诊断方法、基于原始时频域的故障诊断方法、基于最大信息系数(MIC)的故障诊断方法、基于PCA降维的故障诊断方法以及基于原始混合域特征的故障诊断方法，分类方法采用随机森林(rf)进行分类，MIC-rf模型和PCA-rf模型均在本实验特征向量集下进行特征降维或筛选，具体数据见表5。

表5 7种分类模型性能比较

分类模型	准确率/%	运行时间/s
PCC-rf	97.32±0.35	2.735
时域-rf	93.72±0.13	3.781
频域-rf	81.32±0.75	2.312
时频域-rf	95.19±0.33	3.812
MIC-rf	85.57±0.64	161.083
PCA-rf	86.41±0.31	3.313
混合域-rf	97.13±0.62	5.812

首先从准确率方面进行分析，基于原始频域的故障诊断方法准确率约为81.00%，说明频域特征表现力不强，对故障不能进行很好的识别。而基于MIC的故障诊断方法和基于PCA的故障诊断方法由于特征选择错误，导致重要性较高的特征被剔除，准确率仅约为85.00%。识别准确率最高的为笔者所提出的方法和基于原始混合域的故障诊断方法，准确率可达约97.00%，因此可看出混合域特征集经过PCC筛选后，重要性较高的特征向量得以保留，相关性较强的特征被剔除。

再从运行时间角度进行分析，通过表5可看出基于MIC的故障诊断方法由于其近似算法计算时间较长，导致其诊断时间远远超过其他模型。剩余模型的诊断时间相比MIC模型较好，但与笔者提出的方法也有一定差距。同时，从表中可看出虽然基于原始混合域的故障诊断模型准确率较高，但笔者提出的方法仅需约2 s便达到较高准确率，与原始混合域的故障诊断模型相比诊断时间缩短近2倍。因此，从诊断准确率和运行时间进行综合考虑，笔者提出的基于皮尔逊相关系数的研究方法具有更大的优势。

4 结论

4.1 针对单个特征故障诊断精度不高、特征提取和特征集构建困难的问题，分别从原始信号的时域、频域和时频域提取各个维度的综合特征参数，充分利用了不同维度的有效信息。

4.2 为解决高维数据冗余性过高、相关性较强的缺点，提出利用皮尔逊相关系数对混合域特征集进行特征筛选，降低特征集相关性，为后续故障识别提供较为合理干净的数据。

4.3 所提的基于皮尔逊相关系数的混合域轴承故障诊断方法，分类准确率可达97.32%，相比其他方法有较为明显的优势，在进行特征选择后，准确率未出现明显下降，具有较高的工程应用价值。

参 考 文 献

[1] 孔子迁，邓蕾，汤宝平，等.基于时频融合和注意力机制的深度学习行星齿轮箱故障诊断方法[J].仪器仪表学报，2019，40(6)：221~227.

[2] 熊鹏博，王晓东.多时域特征与SVM的隔膜泵单向阀故障诊断[J].机械科学与技术，2019，38(4)：538~543.

[3] 马欣欣，郭敏.基于EEMD和多域特征融合的手势肌电信号识别研究[J].云南大学学报(自然科学版)，2018，40(2)：252~258.

[4] 彭涛，杨慧斌，李健宝，等.基于核主元分析的滚动轴承故障混合域特征提取方法[J].中南大学学报(自

然科学版),2011,42(11):3384~3391.
[5] 李大江. 基于多域特征的滚动轴承故障检测和状态识别方法 [J]. 机械设计与制造工程,2021,50(3):55~58.
[6] 戴豪民,许爱强,李文峰,等.基于WMRMR的滚动轴承混合域特征选择方法 [J]. 振动与冲击,2015,34(19):57~61.
[7] Tang X H,Wang J C,Lu J G,et al.Improving Bearing Fault Diagnosis Using Maximum Information Coefficient Based Feature Selection [J].Applied Sciences,2018,8(11):2143.
[8] 白丽丽,韩振南,任家骏,等.基于拉普拉斯分值和鲸鱼寻优SVM的滚动轴承故障诊断[J].太原理工大学学报,2019,50(6):829~834.
[9] 王海瑞,张楠.基于KPCA-RVM的转子故障诊断[J].价值工程,2017,36(15):154~156.
[10] 张朝林, 范玉刚.CEEMD与卷积神经网络特征提取的故障诊断方法研究 [J]. 机械科学与技术,2019,38(2):178~183.
[11] Roffo G,Melzi S,Castellani U,et al.Infinite Feature Selection:a Graph-based Feature Filtering Approach [J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,43(12):4396~4410.
[12] 万晓静, 孙文磊, 陈坤. 小波包能量熵和改进的LSSVM在风力机轴承故障诊断中的应用[J].水电能源科学,2021,39(2):142~145.
[13] 周建民,徐清瑶,张龙,等.结合小波包奇异谱熵和SVDD的滚动轴承性能退化评估[J].机械科学与技术,2016,35(12):1882~1887.
[14] Zhao D,Li K C,Li H Y.A New Method for Separating EMI Signal Based on CEEMDAN and ICA[J].Neural Processing Letters,2021,53:2243~2259.
[15] Chen R,Chen S,Yang L,et al.Looseness Diagnosis Method for Connecting Bolt of Fan Foundation Based on Sensitive Mixed-Domain Features of Excitation-response and Manifold Learning [J].Neurocomputing,2017,219:376~388.
[16] Chen P,Li F,Wu C.Research on Intrusion Detection Method Based on Pearson Correlation Coefficient Feature Selection Algorithm [J].Journal of Physics:Conference Series,2021,1757(1):12054.
[17] Yu K,Liu L,Li J,et al.Multi-Source Causal Feature Selection [J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(9):2240~2256.
[18] Xue X,Zhou J.A Hybrid Fault Diagnosis Approach Based on Mixed-domain State Features for Rotating Machinery[J].ISA Transactions,2017,66:284~295.

(收稿日期:2021-08-30,修回日期:2022-03-12)

Feature Selection Method for Rolling Bearings in Mixed Domain Based on Pearson Correlation Coefficient

XIAO Yang, LI Ya, WANG Hai-rui, CHANG Meng-rong

(*Faculty of Information Engineering and Automation, Kunming University of Science and Technology*)

Abstract Considering single domain feature's difficulty in reflecting the running state of bearings comprehensively and accurately and the lower accuracy of fault diagnosis, a hybrid domain feature set construction method was proposed, which extracts feature vectors in time domain, frequency domain and time-frequency domain respectively to construct multi information feature set. Aiming at high linear correlation, high dimension, redundant information and large space cost of the feature set in the mixed domain, the feature selection method of Pearson correlation coefficient (PCC) was used to calculate the correlation of features, and the low dimension features with weak correlation were extracted and input into the random forest classifier for fault diagnosis. The experimental results show that, the hybrid domain feature selection method based on PCC can improve the classification performance, and it considers the interaction between features and reduces the loss of information. The classification accuracy of the model can reach 97.32%, which has obvious advantages compared with other methods.

Key words feature selection, rolling bearing, mixed domain, redundancy, PCC