

DOI: 10.13382/j.jemi.B2306901

# 变转速工况下基于角度重采样与 PCA-XGBoost 轴承故障诊断方法研究\*

刘馨雅<sup>1,2</sup> 马超<sup>2</sup> 黄民<sup>1,2</sup> 张占一<sup>3</sup>

(1. 北京信息科技大学机电工程学院 北京 100192; 2. 北京信息科技大学现代测控技术教育部重点实验室 北京 100192;  
3. 北京东方振动和噪声技术研究所 北京 100085)

**摘要:**针对变转速工况下,轴承振动信号容易发生信号特征混叠、频率偏移、信号截断和噪声污染问题,提出了一种结合角度重采样、主成分分析(PCA)和极端梯度提升树(XGBoost)的故障分类模型。首先,采用脉冲信号估计轴承转速的方法对轴承振动信号进行角度重采样并求取时频特征指标;其次,利用主成分分析(PCA)对时频特征参数进行降维核心提取,选取总贡献大于95%的主元作为XGBoost模型的输入样本;最后,利用网格搜索法对XGBoost进行主要参数调优,并划分训练集和测试集对该模型进行训练,验证其故障分类的准确性。结果表明该方法的故障诊断准确率为96.44%,相较于未降维后的数据运行时间缩短了27.24 s,且角度重采样后的诊断效果明显优于未角度重采样的诊断效果,故障识别率提高了7%以上,证明所提方法能够更加快速、准确的做出诊断。

**关键词:** 变转速;轴承;PCA;XGBoost

**中图分类号:** TH133.33 **文献标识码:** A **国家标准学科分类代码:** 460.2020

## Research on angle resampling and PCA-XGBoost bearing fault diagnosis method under variable speed working condition

Liu Xinya<sup>1,2</sup> Ma Chao<sup>2</sup> Huang Min<sup>1,2</sup> Zhang Zhanyi<sup>3</sup>

(1. College of Electromechanical Engineering, Beijing Information Science and Technology University, Beijing 100192, China;  
2. Key Laboratory of Modern Measurement and Control Technology Ministry of Education, Beijing Information Science and Technology University, Beijing 100192, China; 3. China Orient Institute of Noise and Vibration, Beijing 100085, China)

**Abstract:** Aiming at the variable speed condition, the bearing vibration signal is prone to signal feature aliasing, frequency shift, signal truncation and noise pollution, a fault classification model combining angular resampling, principal component analysis (PCA) and extreme gradient boosting tree (XGBoost) is proposed. Secondly, the time-frequency feature parameters are extracted by principal component analysis (PCA), and the main elements with total contribution greater than 95% are selected as input samples of XGBoost model; finally, the main parameters of XGBoost are tuned by grid search method, and the model is trained by dividing the training set and the test set to verify the accuracy of its fault classification. The results show that the accuracy of fault diagnosis is 96.44%, the running time is shortened by 27.24 s compared with that of the data without dimensionality reduction, and the diagnosis effect after angle resampling is obviously better than that of the diagnosis effect without angle resampling, and the fault recognition rate is improved by more than 7%, which proves that the proposed method can make diagnosis more quickly and accurately.

**Keywords:** bearings with variable speed conditions; bearings; PCA; XGBoost

## 0 引言

机械设备的正常运行对于工业生产和日常生活

至关重要,但长时间使用、磨损和环境因素会导致转动部件出现故障,严重影响设备运行甚至带来安全隐患<sup>[1]</sup>。转速轴承作为常见部件之一,承担着支撑轴的重要任务,由于其在工作中承受着连续旋转和

复杂载荷的作用,使得转速轴承容易受到损坏,而故障一旦发生,将会导致设备的性能下降、能源消耗增加、运行噪音加大,并且增加了设备的维修和更换成本。因此,及早发现和解决转速轴承故障对于提高设备的可靠性、减少停机时间、降低运维成本具有重要意义。

近年来,非平稳、非线性的轴承振动信号逐渐得到大家关注,针对轴承变转速工况问题,角度重采样是可用于降低变转速工况影响的行之有效的办法<sup>[2]</sup>。杨武成<sup>[3]</sup>利用计算阶次跟踪法根据转速脉冲计算旋转相位,再通过插值法对振动信号重采样得到角域信号,进一步利用 Hilbert 变换解调出低频信号,进而提取故障特征。万书亭等<sup>[4]</sup>采用了无转速阶次跟踪算法,利用时频脊线代入振动信号提取角域信号,降低了检测信号频谱模糊的影响。Urbanek 等<sup>[5]</sup>提出将相位解调和时频分析融合的方法来精确计算转速曲线。

借助人工智能技术,将机器学习、深度学习和模式识别等领域的知识相融合,研发了灵敏高效的轴承故障诊断模型,以应对传统方法的限制<sup>[6]</sup>。张志政等<sup>[7]</sup>通过构建 KPCA-SVM 模型对船舶燃油系统进行故障监测和诊断,利用特征空间中的  $T^2$  和  $SPE$  统计量来监测故障发生,并将监测到的故障样本作为 KPCA-SVM 的输入来实现故障分类。

何坤敏等<sup>[8]</sup>利用 WOA 优化调整后 ELM 模型实现风电齿轮箱故障分类识别,经实验测试验证该模型可取得较高得准确率。王桂兰等<sup>[9]</sup>构建 XGBoost 风机主轴承故障预测模型,对海量风机数据进行训练和预测,该模型分类精度较高且模型运算速度综合性较好。

极端梯度提升树 (extreme gradient boosting, XGBoost)<sup>[10]</sup>的主要目标是提供一种高效的算法来解决大规模稀疏数据上的分类和回归问题,通过将多个分类准确率较低的基分类器集成为一个准确可靠的集成分类器,来提升其预测速度与精度<sup>[11]</sup>。

本论文旨在研究基于角度重采样和机器学习结合的变转速工况轴承故障诊断方法。首先,根据轴承特性对振动信号进行重采样;其次,对角度重采样后的信号进行时频特征提取,并利用主成分分析 (principal component analysis, PCA)<sup>[12]</sup>方法对时频信号进行降维,提取核心信息;然后,选择经过参数优化后的 XGBoost 模型进行故障分类,将已经降维选择后的特征量作为模型输入,提高计算效率,至此建立起基于角度重采样-PCA-XGBoost 的变转速轴承故障诊断方法;最后,开展故障实验验证,并与其他故障分类方法进行了对比,实验结果表明,该方法在轴承故障诊断中具有明显的优势。

## 1 变转速轴承振动信号角度重采样

对于变速旋转轴承而言,高转速下采集到的数据包含的信息量少,同时可能因不满足采样定理而发生混叠现象。除此外,由于转速的变化,采集到的信号不是周期信号,故在进行傅立叶变换时会导致信号泄露,使得频谱拖尾严重,频谱分析结果出现失真,无法准确提取出频域内的信号特征。

等角度重采样<sup>[13]</sup>是一种预处理方法,通过使经过预处理的振动信号在相同旋转角度上采样点数相等,从而使信号具有周期性,并能够获取更准确的频谱计算结果。这种方法调整了采样点的分布,消除了非均匀采样引起的频谱分析误差,有助于提高振动信号分析的准确性。其计算过程如下:

由转速脉冲信号计算得出转速曲线,并由处理过后的转速曲线中的转速值和时间间隔,计算出相邻时间点之间的角度变化量,得出角度变化曲线,同时确定每圈的采样点数。

通过角度变化曲线计算角速度(转速对应的角度变化率),由等角度间隔除以角速度获得等角度间隔对应的时刻;

根据角度对应的时刻对原时域振动信号进行线性插值计算,最终得到角域信号。

算法流程如图 1 所示。

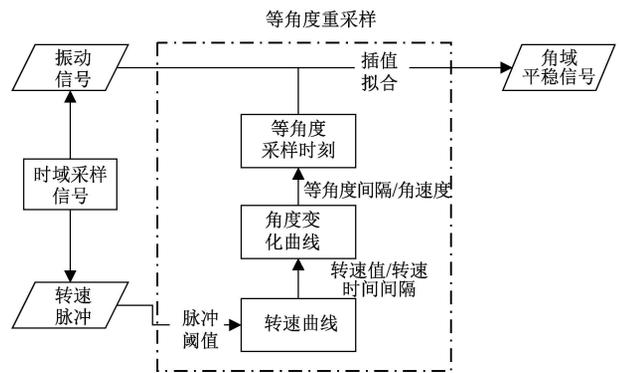


图 1 角度重采样

Fig. 1 Angular resampling

## 2 主成分分析特征降维

大型复杂数据集的特征通常很多且相关性高。对时频域故障数据集进行特征提取后,得到的多维特征会使分析复杂且耗时。为简化分析过程,可以进行特征降维,减少原始特征维度并进一步提取特征用于运算。PCA 可通过分解特征矩阵的方式,将原有数据的  $n$  维特征经过正交变换,构造出新的  $k$  维特征,去除数据中不重要的特

征量,降低数据的维度<sup>[14]</sup>。

通过特征降维,可以保留前  $k$  个含有大部分方差的主元,用它们代表整体数据特征,减少数据数量。这样可以提高模型运算速度,降低计算和存储成本,同时保持数据重要信息。具体步骤为:

根据提取的时域、频域和时频域的特征指标构建原始矩阵  $\mathbf{X}$ 。假设由  $m$  个样本  $p$  个指标,那么  $\mathbf{X}$  的大小为  $m$  行  $p$  列。

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \vdots & \cdots \\ x_{m1} & x_{m2} & \cdots & x_{mp} \end{bmatrix} = (\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p) \quad (1)$$

对故障特征矩阵  $\mathbf{X}$  做标准化处理,使得每个指标在不同样本之间具有相同的尺度。

对每个指标进行标准化处理,计算每列的均值和标准差。令  $\mathbf{X}_s$  表示标准化后的矩阵,其元素满足:

$$x_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \quad (2)$$

其中,  $x_{ij}$  是原始矩阵  $\mathbf{X}$  中第  $i$  行第  $j$  列的元素,  $\bar{x}_j$  和  $\sigma_j$  分别是  $\mathbf{X}$  在第  $j$  列上的均值和标准差。

根据标准化后的矩阵  $\mathbf{X}_s$  建立协方差矩阵  $\mathbf{C}_x$ 。

$$\mathbf{C}_x = \frac{1}{m} \sum_{i=1}^m (x_i - \bar{\mathbf{X}})(x_i - \bar{\mathbf{X}})^T \quad (3)$$

其中,  $m$  是样本数量,  $x_i$  是矩阵  $\mathbf{X}_s$  中第  $i$  个样本,  $\bar{\mathbf{X}}$  是所有样本的均值。

对协方差矩阵  $\mathbf{C}_x$  进行特征值分解,可以得到特征值和特征向量。记特征值为  $\lambda_i$ , 对应的特征向量为  $\mathbf{v}_i$ , 那么有  $\mathbf{C}_x \mathbf{v}_i = \lambda_i \mathbf{v}_i$

计算各个特征值贡献率  $\eta_k$ , 选择总贡献率大 95% 的前  $k$  个主元。

$$\eta_k = \frac{\lambda_i}{\sum_{j=1}^p \lambda_j}$$

再将该  $k$  个特征值对应特征向量按特征值从大到小排列作为主成分矩阵  $\mathbf{U}$ , 根据降维转换公式  $\mathbf{Z} = \mathbf{X}\mathbf{U}$  得到包含  $k$  个主元的故障特征提取矩阵  $\mathbf{Z}$ , 算法结束。

### 3 基于 XGBoost 模型的故障诊断

XGBoost 基于梯度提升决策树 (gradient boosting decision tree, GBDT) 框架, 是一种集成学习方法, 通过反复迭代, 每一轮用新建立的回归树模型拟合前一轮树模型的残差, 来减少训练误差, 这种方式使 XGBoost 模型适用于数据量较大的情况, 便于对大型复杂的故障数据集进行处理。其基本思想是对目标函数做二阶泰勒展开, 利用函数的二阶导数信息来训练树模型, 并把树模型复

杂度作为正则项加到优化目标中, 控制模型的复杂度, 避免过拟合<sup>[15]</sup>。

#### 3.1 XGBoost 回归树拟合:

XGBoost 树集成模型可表示为:

$$\hat{y}_i = \sum_{t=1}^n f_t(x_i), f_t \in F \quad (4)$$

式中:  $n$  为回归树模型的数目;  $i$  为样本的个数;  $x_i$  为第  $i$  个样本的数据;  $f_t(x_i)$  为第  $t$  次迭代计算中输入数据  $x_i$  所在的树模型;  $\hat{y}_i$  为第  $i$  个样本的预测值;  $F$  为回归树模型集合。

所有树的结果累加起来得到模型对样本的预测值, 并通过定义目标函数在每一步生成一个较优的树, 目标函数由总样本损失、正则项两部分构成, 可表示为:

$$\begin{cases} \Omega(f_k) = \gamma T_k + \lambda \frac{1}{2} \sum_{j=1}^{T_k} \omega_j^2 \\ Obj = \sum_{i=1}^n l(y_i \hat{y}_i) + \sum_{k=1}^l \Omega(f_k) \end{cases} \quad (5)$$

式中:  $T_k$  为第  $k$  棵树叶子节点的个数;  $\omega$  为第  $k$  棵树叶子节点的分数;  $\gamma$  和  $\lambda$  皆为比例因子参数, 其中  $\gamma$  控制叶子节点的个数;  $\lambda$  控制叶子节点的分数不宜太大。在目标函数  $Obj$  中,  $\sum_{k=1}^l \Omega(f_k)$  为正则项,  $\sum_{i=1}^n l(y_i \hat{y}_i)$  则为误差项。

为了训练误差函数, 采用前向分布算法, 即在每一轮迭代中添加新的函数到模型中, 新生成的树拟合上一轮树模型预测的残差。向前分布算法可表示为:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \quad (6)$$

将其代入目标函数中, 迭代更新后的目标函数可表示为:

$$Obj^{(t)} = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + C \quad (7)$$

通过二阶泰勒展开计算逼近目标函数, 可得:

$$Obj^{(t)} \approx \sum_{i=1}^n \left[ l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (8)$$

式中:  $g_i$  为损失函数的一阶导数,  $g_i = \partial \hat{y}_i^{t-1} l(y_i, \hat{y}_i^{t-1})$ ;  $h_i$  为损失函数的二阶导数,  $h_i = \partial^2_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^{t-1})$ 。

目标函数可进一步化简为:

$$Obj(t) \approx \sum_{i=1}^n \left[ G_i w_j + \frac{1}{2} (H_i + \lambda) \omega_i^2 \right] + \gamma T \quad (9)$$

$$G = \sum_{i \in I} g_i \quad H = \sum_{i \in I} h_i$$

此时目标函数可看作按叶节点累加的形式,  $I$  为每个叶子上面样本的集合。对  $\omega$  求偏导求得最优解, 并将其代入原式中可以的到最优的  $\omega$  和目标函数分别如下所示:

$$\omega_i^* = -\frac{G_j}{H_j + \lambda} \quad (10)$$

$$Obj^{(i)} = -\frac{1}{2} \sum_{i=1}^T \left( \frac{G_i^2}{H_i + \lambda} \right) + \gamma T \quad (11)$$

通过迭代生成优化后的新树,以减少模型的偏差,从而改善模型的性能。

对于 XGBoost 算法中每次尝试将一个新的分割加入已有叶子节点时,都会计算出相应的参数增益,以用于选择最佳的切分点和参数。其采用贪心法对目标函数最小、增益最大的子树进行划分,增益表达式为:

$$\Theta = \frac{1}{2} \left[ \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma \quad (12)$$

式中:下标 L 代表左子树、R 代表右子树。故等式后第 1 项代表左子树的得分,第 2 项代表右子树的得分,第 3 项代表当前不分割节点的得分。

根据上述过程, XGBoost 对故障特征提取数据进行学习训练,在训练过程中利用梯度提升策略,逐步优化模型,从而实现了对轴承故障的高效检测和诊断,同时通过正则项控制模型复杂度,有效避免过拟合问题。

### 3.2 参数优化

XGBoost 模型共包含 3 类参数:通用参数、模型参数和学习任务参数。超参数直接影响 XGBoost 模型的性能和预测的效果,通常选择超参数是手动调节的,依靠有限次数的实验,得到一组相对合适的超参数<sup>[16]</sup>。

网格搜索法是一种穷举遍历算法,它将所有可能的参数组合在空间中划分成若干个网格,遍历网格中所有交点,对每个参数集合应用交叉验证来计算误差,得到误差最小的为全局最优解<sup>[17]</sup>。但由于遍历全部参数最优解需要耗费大量时间,本文只选取了 6 个主要超参数进行寻优。

## 4 诊断方法建立

本文提出的诊断方法主要研究流程如图 2 所示。其具体步骤为:

对原始信号数据进行消噪和清洗,消除噪声污染。采用加速度传感器收集数据时常夹杂着测量噪声,在进行特征提取前需要对原始数据进行预处理操作。使用邻近数据的平均值填补缺失值,剔除异常值,采用窗函数对信号进行滤波处理,并对滤波后的信号进行希尔伯特解调,对变速振动信号进一步地降噪和清洗,使得后续的分析更加准确。

对降噪解调后的变转速轴承信号进行等角度重采样。通过转速脉冲信号计算转速曲线并平滑滤波,进一

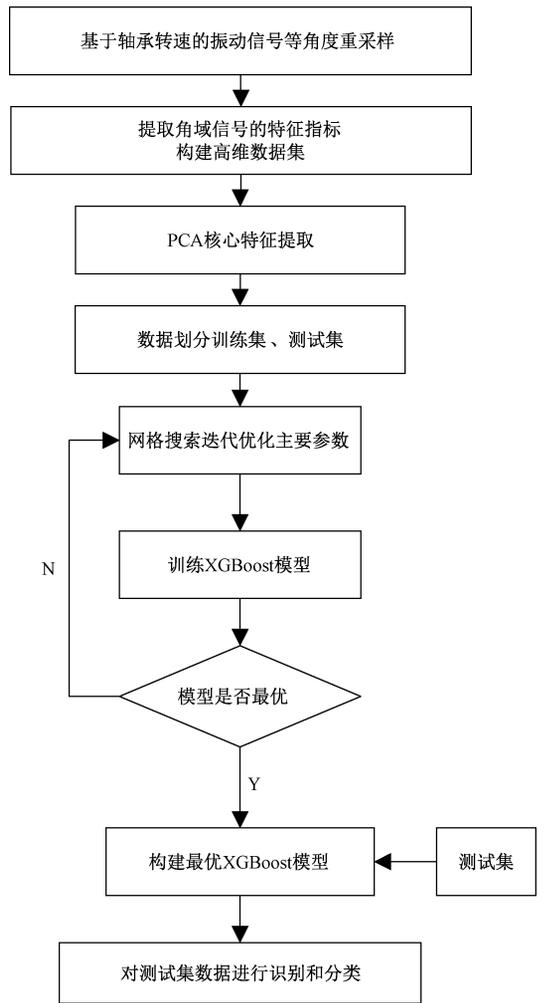


图 2 变转速轴承故障诊断流程

Fig. 2 Fault diagnosis flow of variable speed bearing

步计算角度变化曲线,根据角度变化曲线计算角速度,最后通过线性插值得到等角度间隔的角域信号。

对重采样后的角域信号数据求取时域、频域、时频域特征信息,从 3 种特征域获取轴承信息,构建高维故障数据集。

对高维故障数据集数据降维,采用主成分分析法对归一化处理后的数据降维提取,对构建的核矩阵计算特征值和特征向量,最后计算各个特征值贡献率,选择大于 95% 的各主成分建立模型样本。

将降维后构建的新样本数据作为 XGBoost 模型的输入,划分训练集、测试集,并采用网格搜索优化最佳参数。利用测试集对训练后的最优模型进行性能评价。

## 5 实验验证及结果分析

### 5.1 数据集介绍

实验数据来源于渥太华大学的 SpectQuest 机械故障

模拟器(MFS-PK5M)<sup>[18]</sup>。试验台由电机、AC驱动器、编码器、联轴器、转子和轴承组成。实验采用ER16K球轴承,振动信号采样频率为200 kHz,每次实验采样持续时间为10 s。

该数据集轴承状态共包括5种:健康、内圈故障、外圈故障、滚动体故障、内外圈及滚动体的复合故障(内圈、外圈、滚动体)。并模拟了4种不同工况:升速、减速、升速再减速、减速再升速。

### 5.2 故障信号角度重采样

由图3可知轴承故障信号受变转速工况影响,在时域上表现为非平稳性信号,其振动幅值随转速波动而变化。

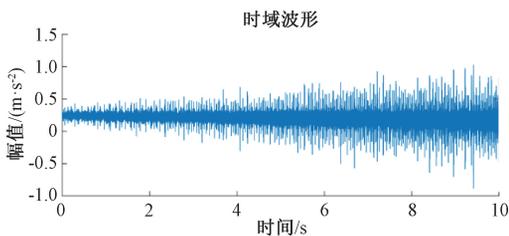


图3 轴承故障时域信号

Fig.3 Bearing failure time domain signal

为了保证信号的周期性,满足数据在相同旋转角度内的采样点数相等,根据轴承状态的不同,本文将该数据集划分为5个子集,每个子集内包含该状态下的升速、减速、升速再减速共3种工况数据,且每组工况数据包含3组,每个子集共9组数据,整个数据集45组数据。对每组数据进行等角度重采样,设置等角度采样每圈的采样点数为2 048,则每个等间隔角度为 $\frac{2\pi}{2\ 048}$ ,角域与时域概念对比如表1所示。

表1 角域与时域概念对比

Table 1 Angular domain and time domain concept comparison

	角度域	频域
采样长度	$\theta / (^\circ)$	$t/s$
采样间隔	$\Delta\theta_s / (^\circ)$	$\Delta t_s / s$
采样率	$O_s / (\text{点数}/\text{转})$	$f_s / (\text{点数}/\text{s})$

基于轴承转速信号对故障时域信号进行角域重采样,可获得加速状态下单次实验的内圈轴承故障的转速曲线如图4所示,前1 s内的时域信号和角域信号如图5所示。在变转速情况下,由于采样频率未能根据转速变化动态调整,传统的等时间间隔采样可能会导致频谱中的特征混叠和频率分量的偏移。通过按照角度对信号进行重采样,即使在转速发生变化的情况下,也可以保持特征在频谱中的位置稳定,从而减少频率偏移和混叠,原始

信号的部分频谱图和重采样后的阶次谱如图6所示。

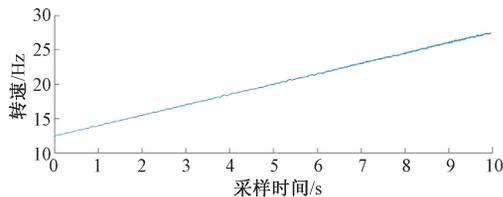


图4 轴承转速曲线

Fig.4 Bearing speed curve

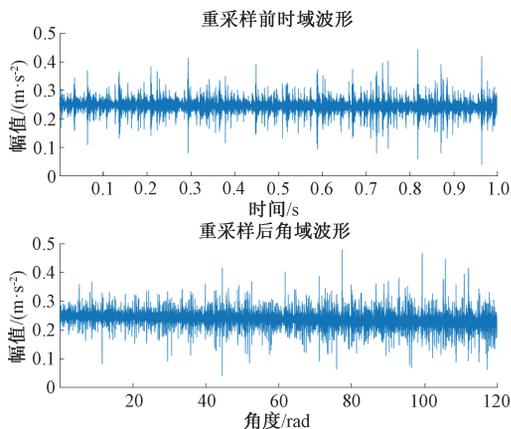


图5 时域信号与角域信号对比

Fig.5 Comparison of time-domain and angle-domain signals

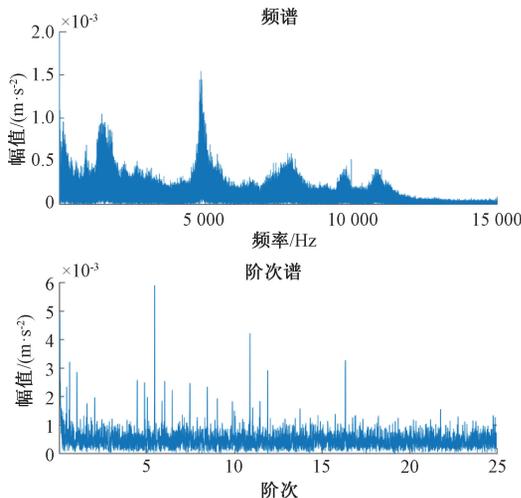


图6 频谱与阶次谱

Fig.6 Spectrum and order spectrum

通过进行等角度重采样,将信号从时域转换到角域,从而实现平稳性的改善,有效消除了转速波动对轴承时频特征提取的影响,能够更加清晰地提取出故障信号,提高故障诊断的准确性和可靠性。

### 5.3 故障特征提取及降维

构建高维故障特征指标数据集,将4种故障信号和

正常信号共 45 组数据划分单个角域信号数据样本长度为 4 000,分别求取单个样本的时域、频域、时频域指标 17 个,其中包含时域特征最大值、峰峰值、均方根、绝对平均值、方差、峰值因子、脉冲因子、裕度因子、波形因子共 9 个,频域特征重心频率、幅值最大值、幅值平均值、幅值中位数、幅值方差、功率最大值、功率平均值、功率中位数共 8 个。

上文中提到高维的数据直接用来当作模型的输入数据往往会导致模型过拟合,且数据集数量大太会使运算效率降低,所以在划分故障特征指标数据集前对其采用 PCA 进行主元分析,选定累积贡献率大于 95%的主元。经过 PCA 处理后,图 7 所示前 4 个主成分的累积贡献率高达 96.43%,显示出轴承数据集中的 17 个特征指标的绝大部分信息被有效地集中在这 4 个主成分中。此外,这 4 个主成分的贡献率相对集中,保留了原始数据的大部分信息,有效实现了降维的目的。

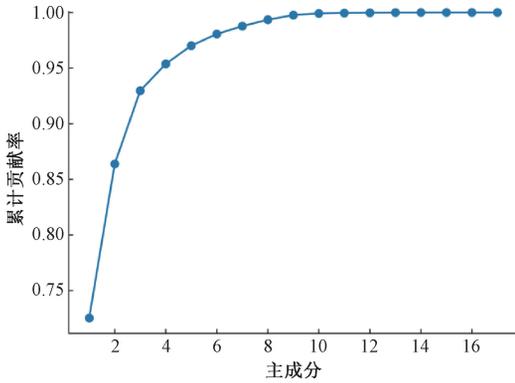


图 7 PCA 累计贡献率

Fig. 7 Cumulative PCA contribution

将降维后的数据集每种状态按照 8 : 2 的样本比重划分了训练集和测试集用于故障分类模型的训练和测试。

故障分类标签对应关系如表 2 所示。

表 2 轴承故障分类标签对应关系

Table 2 Bearing fault classification label correspondence

轴承状态	健康	内圈故障	外圈故障	滚动体故障	复合故障
标签	0	1	2	3	4

### 5.4 XGBoost 故障诊断模型结果分析

将划分好的训练集与测试集输入采用经过网格搜索优化后的 XGBoost 进行故障诊断,优化后的参数设置如表 3 所示。

经过角域转换和 PCA 的特征提取后,降低故障特征数据的维度,成功减少了故障诊断模型所需要的输入数据量。这一改进带来了两方面的优势:时间和精度的提

表 3 XGBoost 主要参数优化

Table 3 XGBoost main parameters optimisation

参数名称	取值范围	步长	调优值	
n_estimators	迭代次数	[100,600]	50	450
learning_rate	学习率	[0.1,0.5]	0.01	0.1
max_depth	树最大深度	[3,10]	1	4
colsample_bytree	树特征抽样比例	[0.5,1.0]	0.1	0.8
subsample	每棵树样本采样比例	[0.5,1.0]	0.1	0.8
min_child_weight	叶节点最小的样本权重和	[0,5]	1	1

升。首先,由于输入数据的减少,故障诊断的处理时间得以缩短。其次,由于保留了与故障相关的主要特征,模型在识别和分类故障时的准确度也有所提高。这一改动使得故障诊断更加高效和准确,提升了整体的性能和效果。实验结果如图 8 所示。

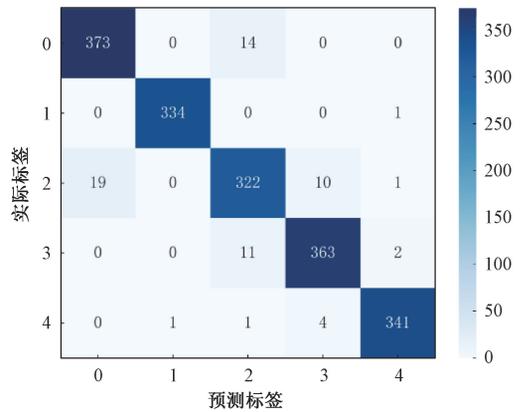


图 8 故障分类混淆矩阵

Fig. 8 Fault classification confusion matrix

结果表明,利用该模型进行故障诊断时,内圈故障诊断准确率最高,有 1 个诊断为复合故障,准确率为 99.70%;内圈、外圈和滚动体的复合故障诊断准确率其次,有 1 个诊断为内圈故障、1 个诊断为外圈故障、4 个诊断为滚动体故障,准确率为 98.84%;诊断准确率取得最低的为外圈故障,其中有 19 个诊断为正常、10 个诊断为滚动体故障,1 个诊断为复合故障,准确率为 91.48%;另外,健康状态有 14 个诊断为外圈故障,准确率为 96.38%;滚动体故障有 11 个诊断为外圈故障、2 个诊断为复合故障,准确率为 96.54%。综合故障诊断率为 96.44%。

### 5.5 不同诊断模型实验对比分析

为验证等角度重采样方法的有效性,分别对未进行角度重采样的数据集和已进行角度重采样的数据集进行特征提取。按照相同步骤进行 PCA 降维和数据集划分,

分别输入 XGBoost、KNN、LR、SVM、贝叶斯(NB)5种诊断模型进行训练,训练集与测试集划分比重均为8:2。同时,通过与KNN、LR、SVM、贝叶斯(NB)4种经典分类模型进行对比来验证 XGBoost 模型适应复杂大规模数据的适应能力。

为了保证计算性能的稳定性,5种方法均运行了15次,取平均诊断准确率做最终评价指标。结果如表4所示。

表4 不同模型准确率与运算时间对比

Table 4 Comparison of accuracy and computing time of different models

	对比算法	分类准确率/%	运算时间/s
未角度 重采样	PCA-XGBoost	78.98	13.52
	PCA-SVM	78.23	16.23
	PCA-LR	76.13	15.63
	PCA-KNN	71.00	13.18
	PCA-NB	69.00	11.46
角度 重采样	PCA-XGBoost	96.44	5.22
	PCA-SVM	94.10	7.48
	PCA-LR	92.20	5.97
	PCA-KNN	81.00	4.35
	PCA-NB	77.91	4.86

经过对比可知,在进行特征提取前未经过角度重采样会导致特征混叠,使得准确率普遍低,效果差。XGBoost、KNN、LR、SVM、贝叶斯(NB)5种诊断模型在输入数据未进行角度重采样的情况下取得的分类准确率均低于80%。融合角度重采样后各分类模型的准确率有明显的提升,其中 XGBoost、SVM、LR 3种模型提升效果更为明显,达到90%以上。

为验证 PCA 主成分提取的降维效果,将特征数据分别进行 PCA 和 KPCA 降维,选择为经过降维的数据和经过两种降维后的数据分别输入上述实验取得效果较好的 XGBoost、SVM、LR 模型进行验证。实验模型、参数、步骤等保持不变。结果如表5所示。

表5 不同降维算法准确率与运算时间对比

Table 5 Comparison of accuracy and computing time of different dimensionality reduction algorithms

	对比算法	分类准确率/%	运算时间/s
	PCA-LR	92.20	5.97
	KPCA-LR	91.27	5.97
	PCA-SVM	94.10	7.48
	KPCA-SVM	92.34	7.39
	XGBoost	97.82	32.46
	KPCA-XGBoost	94.86	5.39
	PCA-XGBoost	96.44	5.22

由表5可知,未进行 PCA 主成分提取直接进行 XGBoost 分类准确率最高但运算需要 32.46 s,融入 PCA

主成分分析的 PCA-XGBoost 分类方法只需要 5.22 s,准确率为 96.44%,由此可知融合 PCA 降维可在保证准确率精度的同时,极大地提高运算效率。从表5中也可以看出,KPCA-LR 比 PCA-LR 分类准确率低 0.93%,KPCA-SVM 比 PCA-SVM 分类准确率低 1.76%,KPCA-XGBoost 比 PCA-XGBoost 分类准确率低 1.58%这说明 PCA 方法更适合该类降维问题。

在分类准确率上,XGBoost 模型表现出良好的优越性,处理复杂大规模数据的优势更加明显,与其他4种模型相比具有最高的准确率,最少的运算时间。

## 5.6 验证分析

为验证本文所提方法的适用性,采用某机务段轴承实验数据和西安交大机械故障综合模拟试验台采集的变转速轴承数据进一步分析。

验证数据集 A:某机务段轴承振动数据。实验采用 NJ2232WB 轴承,振动信号采样频率为 20 kHz,每次实验采样持续时间为 10 s。本文选取内圈故障、外圈故障、滚动体故障、健康状态共 4 种状态下的数据进行实验验证。将 4 种工况下的振动信号按照单个样本长度 1 024 进行划分,每种工况包含 191 个样本,总体样本 764 个,构建为验证数据集 A。

验证数据集 B:西安交大轴承变转速数据。实验采用 NSK6203 轴承,振动信号采样频率为 25.6 kHz,每次实验采样持续时间为 15 s,每次实验都包含一个完整的从静止状态逐渐加速至 3 000 rpm,后保持稳定,最后逐渐减速为 0 的加减速过程<sup>[19-20]</sup>。本文选取内圈轻度故障、内圈中度故障、内圈重度故障、外圈轻度故障、外圈中度故障、外圈中度故障、健康状态共 7 种状态下的数据进行试验验证。将 7 种工况下的变转速振动信号做等角度重采样处理转换为平稳角域数据,图 9 展示了内圈中度故障数据时域信号与角域信号对比结果。对重采样后的角域信号按照单个样本长度 1 024 进行划分,采用滑动窗口重叠采样对角域数据进行增强,滑动窗口长度设为 1 024,移动步长 300,共划分总体样本 1 410 个。其中包含健康状态样本 203 个;内圈轻度故障样本 195 个;内圈中度故障样本 194 个;内圈重度故障样本 193 个;外圈轻度故障样本 193 个;外圈中度故障样本 200 个;外圈重度故障样本 232 个,构建为验证数据集 B。

验证数据集 A 的故障分类标签对应关系如表 6 所示,和验证数据集 B 的故障分类标签对应关系如表 7 所示。

表6 某机务段轴承数据分类标签对应关系

Table 6 Machine bearing data classification label correspondence

轴承状态	健康	内圈故障	外圈故障	滚动体故障
标签	0	1	2	3

表 7 西安交大变转速轴承数据分类标签对应关系

Table 7 Xi'an Jiaotong University variable speed bearing data classification label correspondence

轴承状态	故障损伤程度		
	轻度	中度	重度
	标签		
健康	0		
内圈故障	1	2	3
外圈故障	4	5	6

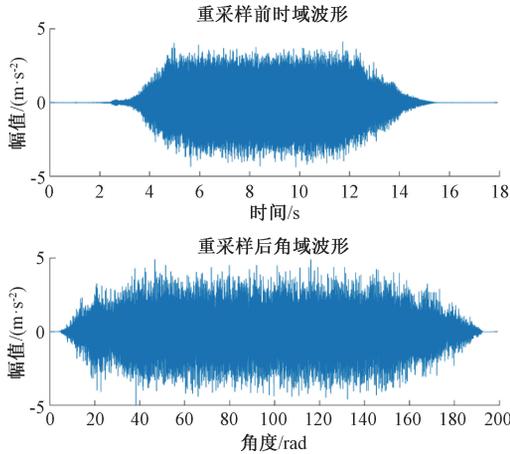


图 9 西安交大时域信号与角域信号对比

Fig. 9 Xi'an Jiaotong University comparison of time-domain and angle-domain signals

在 5.3 节中提到的 17 个特征指标被用于对验证数据集 A 和验证数据集 B 中的单个样本进行计算,从而构建两个高维故障特征数据集。进而采用 PCA 算法分别对两个高维故障特征数据集进行降维处理,所得两种情况下的主元累积贡献率如表 8 所示。

表 8 主元累积贡献率

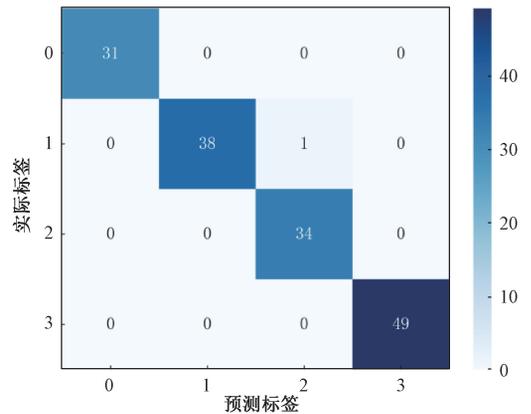
Table 8 Cumulative contribution of the principal element

主元个数	累积贡献率/%	
	验证数据集 A	验证数据集 B
1	69.83	69.83
2	85.33	86.01
3	91.29	91.84
4	94.24	94.60
5	96.40	96.58
6	98.01	98.08

根据表 8 的结果显示,当主成分数量达到 5 个及以上时,两个高维故障特征数据集的主元累积贡献率均已超过 95%。因此,选择保留每个高维故障特征数据集的前 5 个主成分作为代表性特征指标,并将它们作为 XGBoost 模型的输入。

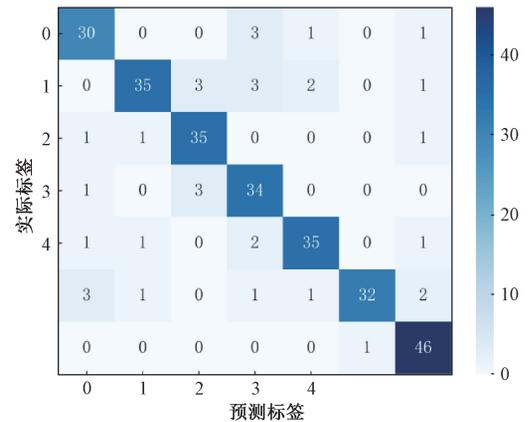
训练集与测试集划分比重依旧保持 8 : 2,测试结果

如图 10 所示。从图中可以看出,在处理某机务段实际工况故障数据时,本文所提方法对健康、外圈中度故障、滚动体轻度故障的故障分类完全正确,内圈故障有 1 个被错分为外圈故障,整体准确率高达 99.35%。在处理西安交大变转速轴承故障数据时,本文所提方法对外圈重度故障的分类效果最好,仅有一个被错分为外圈中度故障,准确率为 97.87%;内圈中度故障只有少数被错分,准确率为 92.11%;健康、内圈重度、外圈轻度和外圈中度故障准确率均在 80% 以上,分别为 85.71%、89.47%、87.5%、80%;内圈轻度故障分类效果最差,有 8 个被错分为其他轴承状态,准确率为 79.55%;整体分类的综合故障诊断准确率为 87.59%,可以达到良好的诊断目的。



(a) 某机务段轴承实验故障分类

(a) Experimental failure classification of bearing in a machine section



(b) 西安交大变转速轴承实验故障分类

(b) Experimental failure classification of variable speed bearing at Xi'an Jiaotong University

图 10 两种验证实验故障分类混淆矩阵

Fig. 10 Two validation experiments fault classification confusion matrix

## 6 结 论

针对变转速工况轴承故障诊断精度较低问题,本文

建立了基于角度重采样-PCA-XGBoost 的诊断模型,并通过故障注入实验来验证方法的有效性。

该方法首先对轴承变转速振动信号进行角度重采样,使用脉冲信号估计轴承转速,准确地获取轴承转速信息;再利用 PCA 算法对特征指标进行了加权融合,在保留主要特征信息的同时优化数据结构使其更加精简;然后,引入网格搜索技术对 XGBoost 模型的超参数进行了寻优,能够有效提高诊断模型的泛化能力和精度。实验结果表明,本文所提方法诊断内圈故障表现最好,精度可达 99.70%,而处理外圈故障的表现最差,精度为 92.53%;引入 PCA 对特征指标降维后使运算时间从 32.46 s 提升到 5.22 s,大大节约了时间成本。同时对比其他 4 种故障分类模型,本文所提方法综合故障诊断准确率最高,在处理紧急变工况的实际背景下具有较强的应用价值。

本文方法在进行 XGBoost 模型超参数寻优时的算法效率和选定寻优的超参数的数量还有待提升,未来将在这方面做进一步优化。

## 参考文献

- [ 1 ] 雷亚国,贾峰,孔德同,等. 大数据下机械智能故障诊断的机遇与挑战[J]. 机械工程学报,2018,54(5):94-104.  
LEI Y G, JIA F, KONG D T, et al. Opportunities and challenges of machinery intelligent fault diagnosis in bigdata era [ J ]. Journal of Mechanical Engineering, 2018,54(5):94-104.
- [ 2 ] 胡春生,李国利,赵勇,等. 变工况滚动轴承故障诊断方法综述[J]. 计算机工程与应用,2022,58(18):26-42.  
HU CH SH, LI G L, ZHAO Y, et al. Summary of fault diagnosis methods for rolling bearings under variable working conditions [ J ]. Computer Engineering and Applications, 2022,58(18):26-42.
- [ 3 ] 杨武成. 阶次跟踪和 Hilbert 包络解调在滚动轴承故障诊断中应用[J]. 机械科学与技术,2017,36(12):1873-1876.  
YANG W CH. Application of order tracking and Hilbert envelope demodulation in rolling bearing fault diagnosis[J]. Mechanical Science and Techn-ology for Aerospace Engineering, 2017,36(12):1873-1876.
- [ 4 ] 万书亭,王燕杰,张雄,等. 时变工况的风电机组齿轮箱无转速计阶次跟踪方法研究[J]. 振动工程学报,2023,36(1):266-279.  
WANG SH T, WANG Y J, ZHANG X, et al. Research on tachometer-less step tracking method of wind turbine gearbox under time-varying conditions [ J ]. Journal of Vibration Engineering, 2023,36(1):266-279.
- [ 5 ] URBANEK J, BARSZCZ T, ANTONI J. A two-step procedure for estimation of instantaneous rotational speed with large fluctuations [ J ]. Mechanical Systems and Signal Processing, 2013,38(1):96-102.
- [ 6 ] ZHOU H, WANG S, MIAO Z, et al. Review of the application of deep learning in fault diagnosis [ C ]. Technical Committee on Control Theory, Chinese Association of Automation, Chinese Association of Automation, Systems Engineering Society of China. Shanghai Systems Science Press, 2019:177-181.
- [ 7 ] 张志政,王冬捷,张勇亮. 基于 PSO 改进 KPCA-SVM 的故障监测和诊断方法研究[J]. 现代制造工程, 2020(9):101-107.  
ZHANG ZH ZH, WANG D J, ZHANG Y L. Research on fault monitoring and diagnosis method based on PSO improved KPCA-SVM [ J ]. Modern Manufacturing Engineering, 2020(9):101-107.
- [ 8 ] 何坤敏,王霄,杨靖,等. 基于 RF 特征优选和 WOA-ELM 的风电齿轮箱故障诊断[J]. 电子测量技术, 2023,46(5):57-64.  
HE K M, WANG X, YANG J, et al. Fault diagnosis of wind turbine gearbox based on RF feature optimization and WOA-ELM [ J ]. Electronic Measurement Technology, 2023,46(5):57-64.
- [ 9 ] 王桂兰,赵洪山,米增强. XGBoost 算法在风机主轴承故障预测中的应用[J]. 电力自动化设备, 2019,39(1):73-77,83.  
WANG G L, ZHAO H SH, MI Z Q. Application of XGBoost algorithm in prediction of wind motor main bearing fault[J]. Electric Power Automation Equipment, 2019,39(1):73-77,83.
- [ 10 ] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system [ C ]. 22nd ACM SIGKDD International Conference, 2016.
- [ 11 ] 王雨虹,王志中. 基于 RFRFE 与 ISSA-XGBoost 的变压器故障辨识方法[J]. 电子测量与仪器学报, 2021,35(12):142-150.  
WANG Y H, WANG ZH ZH. Transformer fault identification method based on RFRFE and ISSA-XGBoost [ J ]. Journal of Electronic Measurement and Instrumentation, 2021,35(12):142-150.
- [ 12 ] CHEN C, SEO H, JUN C H, et al. Pavement crack detection and classification based on fusion feature of LBP and PCA with SVM [ J ]. International Journal of Pavement Engineering, 2022, 23(9):3274-3283.
- [ 13 ] 魏义敏,刘辉,杨乐红. 基于角域重采样与 VMD 的电梯曳引轮轴承故障诊断方法[J]. 机电工程, 2023,40(8):1259-1266.

- WEI Y M, LIU H, YANG L H. Fault diagnosis method of elevator traction wheel bearing based on angular domain resampling and VMD[J]. Journal of Mechanical & Electrical Engineering, 2023, 40(8): 1259-1266.
- [14] 薛勇, 赵荣珍. 改进 LPCDA 算法及其在旋转机械故障诊断中的应用[J]. 振动. 测试与诊断, 2023, 43(1): 132-138, 202.
- XUE Y, ZHAO R ZH. Improved LPCDA algorithm and its application in fault diagnosis of rotating machinery [J]. Journal of Vibration, Measurement & Diagnosis, 2023, 43(1): 132-138, 202.
- [15] 姜少飞, 邬天骥, 彭翔, 等. 基于 XGBoost 特征提取的数据驱动故障诊断方法[J]. 中国机械工程, 2020, 31(10): 1232-1239.
- JIANG SH F, GUO T J, PENG X, et al. Data driven fault diagnosis method based on XGBoost feature extraction [J]. China Mechanical Engineering, 2020, 31(10): 1232-1239.
- [16] 徐博涵, 阮敬. 基于粒子群优化算法改进的 XGBoost 模型制备 C4 烯烃工艺条件优化[J]. 科学技术与工程, 2023, 23(5): 2016-2024.
- XU B Y, RUAN J. Optimization of processing conditions for C4 olefin production based on particle swarm optimization algorithm improved XGBoost model [J]. Science Technology and Engineering, 2023, 23(5): 2016-2024.
- [17] 潘少伟, 郑泽晨, 雷建永, 等. 基于混合优化 XGBoost 算法的砂岩储层孔隙度预测[J]. 计算机应用与软件, 2023, 40(5): 103-109, 206.
- PAN SH W, ZHENG Z CH, LEI J Y, et al. Porosity prediction of sandstone reservoirs based on hybrid optimization XGBoost algorithm [J]. Computer Applications and Software, 2023, 40(5): 103-109, 206.
- [18] HUANG H, BADDOUR N. Bearing vibration data collected under time-varying rotational speed conditions[J]. Data in Brief, 2018, 21: 1745-1749.
- [19] LIU S, CHEN J, HE S, et al. Subspace network with shared representation learning for intelligent fault diagnosis of machine under speed transient conditions with few samples [J]. ISA Transactions, 2022, 128: 531-544.
- [20] SHI Z, CHEN J, ZI Y, et al. A novel multitask adversarial network via redundant lifting for multicomponent intelligent fault detection under sharp speed variation [J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-10.

## 作者简介



**刘馨雅**, 2021 年于河北师范大学获得学士学位, 现为北京信息科技大学硕士研究生, 主要研究方向为智能运维与故障诊断。

E-mail: liuxinya1026@123.com

**Liu Xinya** received her B. Sc. degree from Hebei Normal University in 2017. Now she is a M. Sc. candidate at Beijing Information Science and Technology University. Her main research interests include intelligent operation and maintenance and fault diagnosis.



**马超** (通信作者), 2009 年于北京理工大学获得博士学位, 现为北京信息科技大学副研究员/副教授, 主要研究方向为机械系统动力学、机械系统信号监测分析、生物力学。

E-mail: mach2006@126.com

**Ma Chao** (Corresponding author) received his Ph. D. degree from Beijing Institute of Technology University in 2009. Now he is an associate researcher/associate professor at Beijing Information Science and Technology University. His main research interests include mechanical system dynamics, signal monitoring and analysis of mechanical systems, and biomechanics.