

DOI:10.15923/j.cnki.cn22-1382/t.2018.5.01

基于 LMD 与随机森林的滚动轴承故障诊断

秦喜文^{1,2}, 郭宇¹, 郭佳静¹, 董小刚¹,
冯阳洋¹, 王强进¹

(1.长春工业大学 数学与统计学院, 吉林 长春 130012;

2.长春工业大学 汽车工程研究院, 吉林 长春 130012)

摘要:提出了一种基于 LMD 和随机森林相结合的滚动轴承故障诊断方法,该方法融合了 LMD 和随机森林算法,利用 LMD 分解方法将滚动轴承振动信号分解成若干个 PF 分量,轴承发生不同故障时,不同 PF 分量内的统计特征和频带能量会发生变化,从不同的 PF 分量中计算出其对应的均值、标准差与信息熵等特征值,将这些特征值构造成特征向量,最后运用随机森林算法实现分类滚动轴承类型。结果表明,利用 LMD 和随机森林相结合算法具有较高的识别精度,可以有效识别滚动轴承的故障类型。

关键词:滚动轴承;故障诊断;局部均值分解;特征向量;随机森林

中图分类号: TP 206.3 **文献标志码:** A **文章编号:** 1674-1374(2018)05-0417-07

The fault diagnosis of rolling bearing based on local mean decomposition and random forest

QIN Xiwen^{1,2}, GUO Yu¹, GUO Jiajing¹,
DONG Xiaogang¹, FENG Yangyang¹, WANG Qiangjin¹

(1.School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, China;

2.School of Automotive Engineering Research, Changchun University of Technology, Changchun 130012, China)

Abstract: This paper proposes a fault diagnosis method for rolling bearing based on LMD combined with random forest. This method integrates LMD and random forest algorithm. The LMD method is used to decompose the vibration signal of rolling bearing into several Product Functions. When the bearing experiences different failures, different Product Functions are used. The statistical characteristics and frequency band energy within the components will change, and their corresponding mean values, standard deviations, and information entropy are calculated from different Product Functions. These eigenvalues are constructed as feature vectors. Finally, a random forest algorithm is used to implement classification. Rolling bearing type. The results show that the combination of LMD and random forest has higher recognition accuracy and can effectively identify the type of failure of

收稿日期: 2018-07-30

基金项目: 国家自然科学基金资助项目(11301036,11226335,11571051); 吉林省教育厅科研基金资助项目(JJKH20170540KJ)

作者简介: 秦喜文(1979-),男,汉族,吉林梅河口人,长春工业大学教授,博士生导师,主要从事数据分析与统计建模方向研究,
E-mail: qinxuwen@ccut.edu.cn.

rolling bearings.

Key words: rolling bearing; fault diagnosis; local mean decomposition; eigenvector; random forest.

0 引言

滚动轴承故障诊断过程主要包括故障特征提取和分类,其本质是模式识别的过程^[1]。机械设备一旦发生故障,不但会消耗大量的维修时间和维修费用,而且会给企业带来巨大的经济损失^[2]。因此滚动轴承的状态监测和故障诊断都有非常重要的研究意义。

滚动轴承故障振动信号多数为非平稳信号,所以在故障诊断中必须运用可以处理非平稳信号的特征提取方法。时频分析是一种处理非平稳信号的有力方法,在对滚动轴承信号处理方面具有广泛应用^[3-4]。常见的时频分析有 Wigner-Ville 分布、短时傅里叶变换(Gabor 变换)、希尔伯特黄变换(Hilbert-Huang Transform, HHT)。这些方法都有各自的缺点和不足。如 Wigner-Ville 分布中具有交叉干扰项存在,短时傅里叶变换窗函数的时频分辨率不能达到最优,HHT 在处理复杂信号时体现出了长时间的计算缺点^[5]。为了解决传统时频分析法的不足,黄锬等^[6]提出了一种较新的时频分析方法,称为经验模态分解(Empirical Mode Decomposition, EMD)方法。但 EMD 方法存在局限性,算法本身存在模态混叠与端点效应,对噪声敏感以及依赖于插值方法的选择等问题。局部均值分解(Local Mean Decomposition, LMD)方法是由学者 Smith^[7]提出的一种处理非平稳信号的自适应时频分析方法,LMD 在某些方面的性能要优于 EMD 方法,例如在抑制模态混叠、端点效应、避免过包络、欠包络、迭代次数等方面。LMD 方法自适应的将一个复杂的高频信号分解成若干个瞬时频率的具有物理意义的乘积函数 PF(Product Function)分量之和。每一个 PF 分量是由一个纯调频信号和一个包络信号直接求出,其中,包络信号是 PF 分量的瞬时幅值,而纯调频信号可以求出 PF 分量的瞬时频率,将全部 PF 的瞬时频率和瞬时幅值组合便可以得到原始信号完整的时频分布,对此进行更精准有效的处理原始信号特征信息。

文中运用 LMD 方法对轴承信号进行分解,并从包含有主要故障信息的模态分量中提取特征

值作为故障特征向量,然后运用随机森林(RF)对轴承信号进行分类,进而得出此方法的有效性。实验证明,文中提出的 LMD 与 RF 相结合的故障诊断方法具有较高的准确率。

1 研究方法

1.1 局部均值分解(LMD)

LMD 方法的原理是从原始信号中分离出不同的包络信号和纯调频信号组合,将包络信号和纯调频信号相乘就可以得到一个具有物理意义瞬时频率的 PF 分量,迭代处理到原始信号的全部 PF 分量被分离出来,就能得出原始信号的时频分布情况。对于原始信号 $x(t)$,分解过程如下^[8]:

1)从原始信号中提取出所有的局部极值点 n_i ,对求出的相邻局部极值点求均值:

$$m_i = \frac{n_i + n_{i+1}}{2} \quad (1)$$

将所有相邻两个平均值点 m_i 用直线连接起来,得到局部均值线段,然后采用滑动平均法进行平滑处理,可得到局部均值函数 $m_{11}(t)$ 。

2)利用局部均值点,求出两相邻极值点之间的包络估计值

$$a_i = \frac{|n_i - n_{i+1}|}{2} \quad (2)$$

将全部相邻两个包络估计值 a_i 用直线连接,然后采用滑动平均方法进行平滑处理,得到包络估计函数 $a_{11}(t)$ 。

3)将局部均值函数 $m_{11}(t)$ 从原信号 $x(t)$ 中分离出来,得出

$$h_{11}(t) = x(t) - m_{11}(t) \quad (3)$$

4)用 $h_{11}(t)$ 除以包络估计函数 $a_{11}(t)$ 以对 $h_{11}(t)$ 进行解调,得出

$$s_{11}(t) = \frac{h_{11}(t)}{a_{11}(t)} \quad (4)$$

计算 $s_{11}(t)$ 的包络估计函数 $a_{12}(t)$,假如 $a_{12}(t)$ 不等于 1,说明 $s_{11}(t)$ 不是一个纯调频信号,需要重复上述迭代过程,使得 $s_{1n}(t)$ 成为一个纯调频信号为止,即 $-1 \leq S_{1n}(t) \leq 1$,且其包络函

数 $a_{1(n+1)}(t)=1$, 所以有

$$\begin{cases} h_{11}(t) = x(t) - m_{11}(t) \\ h_{12}(t) = s_{11}(t) - m_{12}(t) \\ \vdots \\ h_{1n}(t) = s_{1(n-1)}(t) - m_{1n}(t) \end{cases} \quad (5)$$

式中

$$\begin{cases} s_{11}(t) = \frac{h_{11}(t)}{a_{11}(t)} \\ s_{12}(t) = \frac{h_{12}(t)}{a_{12}(t)} \\ \vdots \\ s_{1n}(t) = \frac{h_{1n}(t)}{a_{1n}(t)} \end{cases} \quad (6)$$

迭代终止的条件为

$$\lim_{n \rightarrow \infty} a_{1n}(t) = 1 \quad (7)$$

迭代分解过程在实际应用中是无法实现的, 为了获得较为理想的纯调频信号, 有一个较为合理的迭代终止条件, 可以用

$$a_{1n}(t) \approx 1 \quad (8)$$

5) 在迭代过程中产生的全部包络估计函数相乘, 便可以得到包络信号

$$a_1(t) = a_{11}(t)a_{12}(t)\cdots a_{1n}(t) = \prod_{q=1}^n a_{1q}(t) \quad (9)$$

6) 将纯调频信号 $s_{1n}(t)$ 和包络信号 $a_1(t)$ 相乘, 便可以得到原信号的 PF 分量, 即

$$PF_1(t) = a_1(t)s_{1n}(t) \quad (10)$$

7) 将第一个分量 $PF_1(t)$ 从原始信号 $x(t)$ 中分离出来, 求解出一个新的信号 $u_1(t)$, 将 $u_1(t)$ 作为原始数据重复以上步骤, 循环 k 次直到 u_k 为一个单调函数为止。

$$\begin{cases} u_1(t) = x(t) - PF_1(t) \\ u_2(t) = u_1(t) - PF_2(t) \\ \vdots \\ u_k(t) = u_{k-1}(t) - PF_k(t) \end{cases} \quad (11)$$

所有的 PF 分量和 u_k 重组, 可以求出原始信号, 即

$$x(t) = \sum_{p=1}^k PF_p(t) + u_k(t) \quad (12)$$

说明 LMD 分解没有造成原信号的丢失。

1.2 特征提取

在 LMD 分解之后, 对各 PF 分量进行特征值提取, 需要提取的特征值有均值、标准差、极差、变异系数、波动指数、能量熵和信息熵。其中, 均值、标准差和极差将不详述^[9]。

1.2.1 变异系数

对轴承故障信号进行分析时, 常用特征有平均值、方差、标准差和变异系数等, 其中变异系数可以衡量轴承信号的幅度大小变化, 是衡量滚动轴承信号中变异程度的一种统计量。

其定义为:

$$CV = \frac{\sigma^2}{\mu^2} \quad (13)$$

其中

$$\mu = \frac{1}{l} \sum_{j=1}^l |PF_{(j)}|$$

$$\sigma = \sqrt{\frac{1}{l} \sum_{j=1}^l \left(PF_{(j)} - \frac{1}{l} \sum_{j=1}^l PF_{(j)} \right)^2}$$

式中: l —— PF 分量的长度。

1.2.2 波动指数

波动指数可以衡量信号变化的强度。滚动轴承故障状态下信号的波动通常会比滚动轴承正常状态下信号的波动剧烈。

定义有:

$$F_i = \frac{1}{l} \sum_{j=1}^l |PF_{(j+1)} - PF_{(j)}| \quad (14)$$

式中: l —— PF 分量的长度。

1.2.3 能量熵

为了便于特征提取, 选择能量熵来表征不同类别 PF 分量特征的差别。根据信息论中能量熵的定义, 固有模态函数 $c_i(t)$ 的能量可采用下式计算:

$$E_1 = \int_{t_1}^{t_2} c_i^2(t) dt \quad (15)$$

式中: t_1, t_2 —— 分别为信号起始时间和信号结束时间。

其定义为:

$$H = - \sum_{i=1}^n P_i \log P_i \quad (16)$$

式中: P_i —— 第 i 个 PF 分量的能量占整个信号能量的比例, $P_i = E_i/E$, 其中 E 为整个轴承故障信号的能量。

1.2.4 信息熵

信息中排除了冗余后的平均信息量称为“信息熵”，轴承故障状态下信号的波动通常会比轴承正常状态下信号的信息熵值低。

其定义为：

$$H(x) = E[I(x_i)] = E\left[\log\left(2, \frac{1}{p(x_i)}\right)\right] = -\sum_{i=1,2,\dots,n} p(x_i) \log(2, p(x_i)) \quad (17)$$

式中： x —— 随机变量。

1.3 识别算法

1.3.1 随机森林

随机森林(Random Forest, RF)是近年来一种比较新的机器学习模型^[10]。机器学习模型中比较经典的是神经网络,但神经网络模型计算量非常大。20 世纪 80 年代, Breiman 等^[10]发明分类树的算法,在 2001 年, Breiman^[11]把分类树组合成随机森林,其实质是采用多棵分类树对样本进行训练并预测的一种新型分类方法。

随机森林 $\{h(\mathbf{x}, \boldsymbol{\theta}_k), k=1, 2, \dots\}$ 是一个由树形分类器组合而成的集合,集合中的元分类器 $h(\mathbf{x}, \boldsymbol{\theta}_k)$ 是利用 CART 算法构造的没有剪枝的分类决策树。式中 \mathbf{x} 为输入向量, $\{\boldsymbol{\theta}_k\}$ 为独立同分布的随机向量,由此决定了单棵树的生长过程;森林的输出采用简单多数投票方法,或者是单棵树输出结果的简单平均值得到,具体方法^[12]如下:

1)用 Bagging 方法生成若干个别训练集,即每一个个别训练集都是出自原训练集的 n 个样品中,并且有放回地抽取其中 n 个样品;

2)对于每一个个别训练集,用如下方法生成一棵不剪枝的分类回归树;

假设一共有 M 个原始属性,给出一个正整数 m_{try} ,且满足 $m_{\text{try}} \leq M$ 。在每一个内部节点,从 M 个原始属性中随机抽取 m_{try} 个属性作为该分裂节点的候选属性。在整个森林的生成过程中, m_{try} 不变;选出一个最优的分类方式在 m_{try} 个候选属性之中对此节点进行分类;令每棵树充分成长,不再进行剪枝;

3)重复上述步骤,直到生成 n_{tree} 棵分类回归

树,并且 n_{tree} 足够大;

4)进行分类时,如果遇见未知的样本类别,输出的类别名称可以由森林中树的多数投票决定,即:

$$c = \operatorname{argmax}_c \left(\frac{1}{n_{\text{tree}}} \sum_{k=1}^{n_{\text{tree}}} I(h(\mathbf{x}, \boldsymbol{\theta}_k) = c) \right) \quad (18)$$

随机森林在 Bagging 的基础上采用了随机选择属性的方法,有效降低了树与树之间存在的相关性,同时建立的单棵不剪枝分类回归树能够得到较低的误差,从而确保了随机森林的分类准确性。

1.3.2 支持向量机

支持向量机(Support Vector Machine, SVM)于 1995 年正式发表,是数据挖掘中的一项新技术^[12],它是一种有坚实理论基础的新颖小样本学习方法,具有一定的鲁棒性,并且增删非支持向量机样本对模型没有任何影响。

支持向量机是根据给定的训练

$$T = \{(x_1, y_1), \dots, (x_l, y_l)\} \in (X \times Y)^l \quad (19)$$

其中, $x_i \in X = R^n$, X 为输入空间,输入空间中的每一个 x_i 有 n 个属性特征组成, $y_i \in Y = \{-1, 1\}$, $i=1, 2, \dots, l$ 。寻找 R^n 上的实值函数 $g(x)$,以使用分类函数

$$f(x) = \operatorname{sgn}(g(x)) \quad (20)$$

推导出任意一个模式 x 对应 y 的值为分类问题^[13]。支持向量机的关键在于核函数选择,文中选择高斯核函数。

2 实证分析

2.1 数据来源

文中数据来源于美国凯斯西储大学(Case Western Reserve University)轴承资料,数据是基于一个马达电机、一个转矩传感器、一个功率计及电子控制设备的实验平台获取。文中针对轴承在转速为 1 797 r/min 的情况下进行采集,采样频率为 12 kHz,对轴承出现的内圈故障状态、外圈故障状态、滚动体故障状态以及正常状态情况下进行分析,每种状态下采集 24 个长度为 5 000 的小样本,采集长度是从初始点到 120 000。每个状态下第一次采集的样本原始序列如图 1 所示。

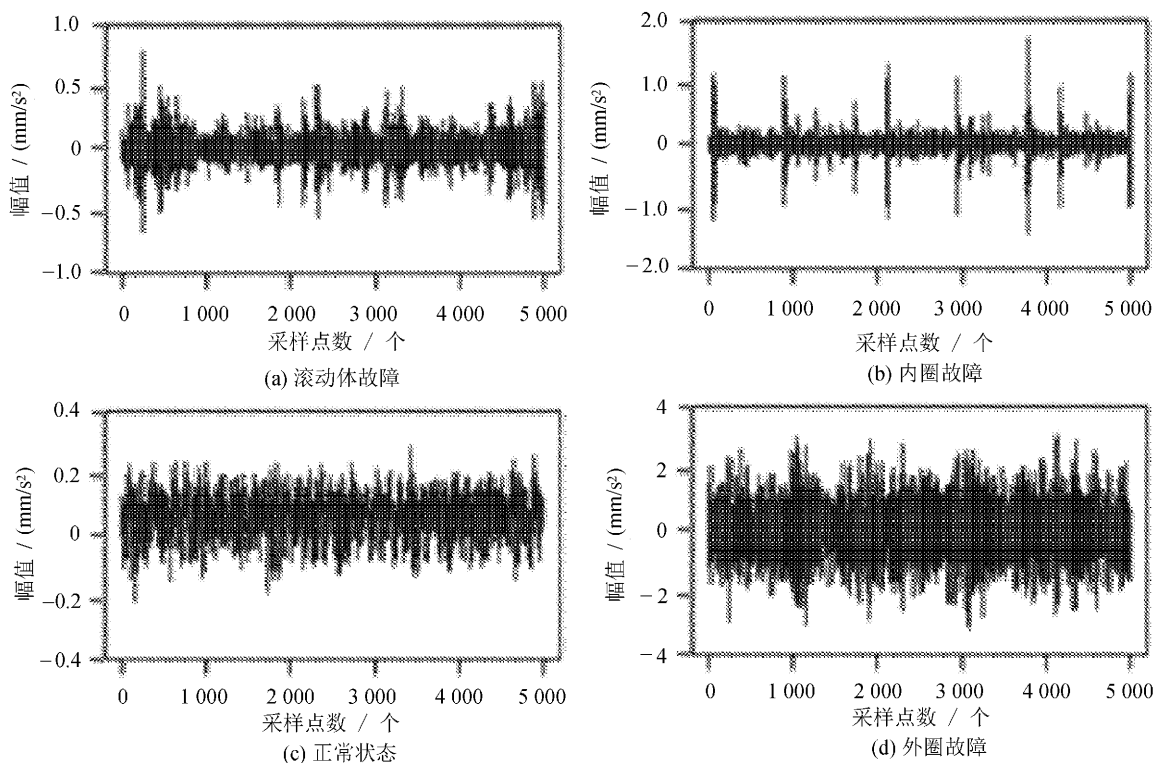
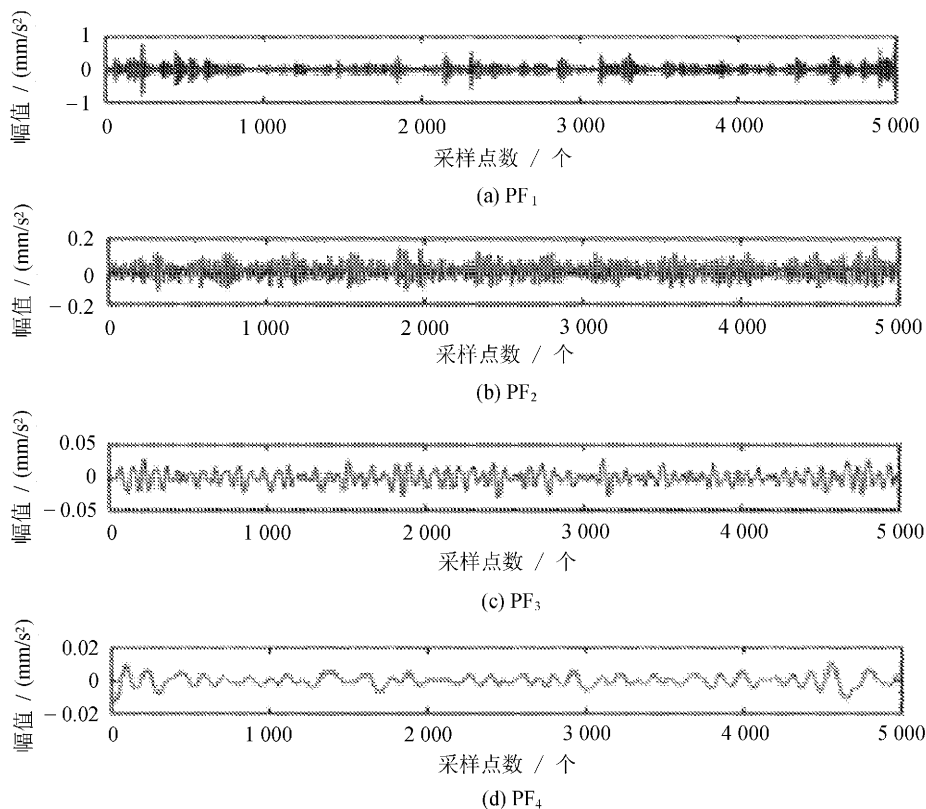


图 1 四种状态部分样本集时序图

2.2 LMD 分解及特征提取

由于样本数据集较大,文中选择其中一个样本集展示 LMD 分解过程。选择滚动体故障中的

第一个小样本,用 LMD 对样本序列进行分解。该样本的 LMD 分解结果如图 2 所示。



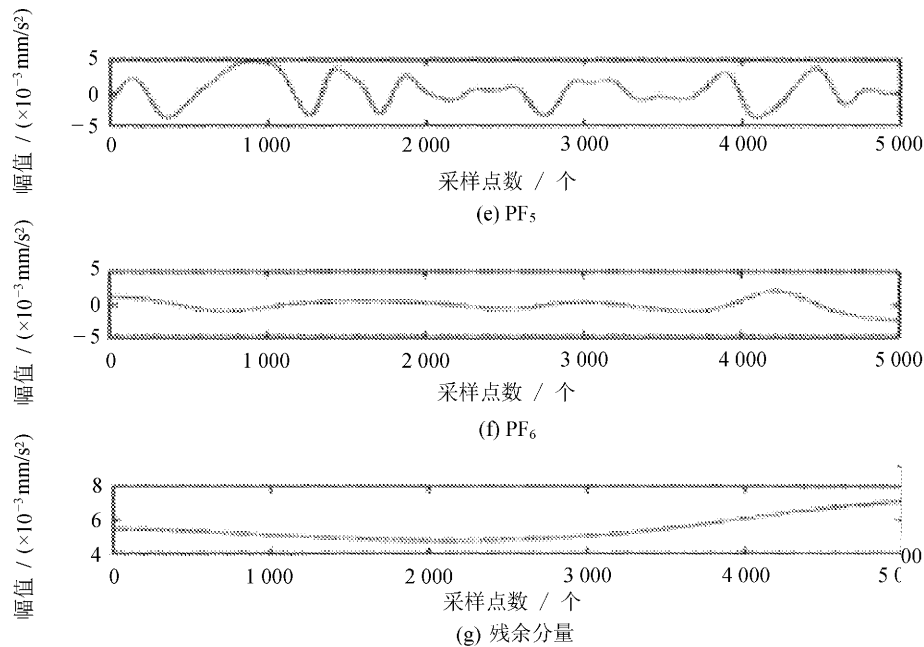


图 2 内圈故障样本经 LMD 分解结果

分解后每一类故障状态都有不同的 PF 分量。对得到的 216 个分量按照 1.2 中的方法计算其对应的均值、标准差、极值、变异系数、波动指数、信息熵和能量熵,得到一个维数为 216×7 的特征向量。

2.3 故障识别分析

从计算得到的 216×7 特征向量矩阵中随机抽取 156 个特征向量作为训练集,剩余的 60 个特征向量作为测试集。将训练集矩阵输入到 RF 中,得到一个模型。再将测试集矩阵输入到该模型中进行分类。

RF 分类结果见表 1。

表 1 测试集随机森林分类结果

轴承状态	正常状态	内圈故障	外圈故障	滚动体故障	误判率
正常状态	18	0	0	0	0.000
内圈故障	0	14	0	0	0.000
外圈故障	1	0	16	0	0.058
滚动体故障	0	0	0	11	0.000

其中行数据为真实故障类型,前四列数据为预测类型,最后一列是误判率,总的误判率为 1.67%。重复上述方法,将训练集矩阵输入到

SVM 中,得到另一分类结果,SVM 分类结果见表 2(SVM 分类的总的误判率为 13.33%)。

表 2 测试集支持向量机分类结果

轴承状态	正常状态	内圈故障	外圈故障	滚动体故障	误判率
正常状态	17	0	1	0	0.056
内圈故障	0	12	0	2	0.142
外圈故障	0	2	15	0	0.117
滚动体故障	0	0	3	8	0.272

从表1和表2中可以发现,针对同一组数据集RF和SVM分类故障状态效果相差很大。其中RF分类将1个外圈故障误判为正常状态,而其余状态全部判断正确;SVM分类将1个正常状态误判为外圈故障,2个内圈故障误判为滚动体

故障,2个外圈故障误判为内圈故障,3个滚动体故障误判为外圈故障。

对测试样本的错误分类个数和正确率做了总体统计,将表1与表2进行整理得到表3。

表3 测试集分类结果汇总

轴承状态	样本量	LMD-RF		LMD-SVM	
		错误样本个数/个	准确率/%	错误样本个数/个	准确率/%
正常状态	18	0	100.0	1	94.4
内圈故障	14	0	100.0	2	85.7
外圈故障	17	1	94.1	2	88.2
滚动体故障	11	0	100.0	3	72.7
总体统计	60	1	98.3	8	86.7

从表3可以更清晰地看到,RF分类的准确率在各种状态下都高于SVM分类。

通过以上分析对比,RF分类的精度更高,文中所提出的故障诊断方法能够有效识别出滚动轴承的工作状态和故障类型。

3 结 语

基于滚动轴承的状态监测和故障诊断的现实需要,提出了局部均值分解与随机森林相结合的故障诊断方法。针对实验数据,对滚动轴承信号进行LMD分解,将分解后的PF分量提取均值、标准差、极差等特征值,构造成特征向量,用以随机森林和支持向量机分类。上述实验过程得出结论:文中提出的LMD与随机森林结合的方法具有相对较高的识别精度,可以有效识别出滚动轴承工作状态及故障类型。

参考文献:

- [1] 何正嘉,陈进,王太勇,等.机械故障诊断理论及应用[M].北京:高等教育出版社,2010.
- [2] 唐贵基,王晓龙.参数优化变分模态分解方法在滚动轴承早期故障诊断中的应用[J].西安交通大学学报,2015,49(5):73-81.
- [3] Baydar N, Ball A. Detection of gear failures via vibration and acoustics signals using wavelet transform [J]. Mechanical Systems and Signal Processing, 2003, 17(4): 787-804.
- [4] Zheng H, Li Z, Chen X. Gear fault diagnosis based

on continuous wavelet transform [J]. Mechanical Systems and Signal Processing, 2002, 16(2/3): 447-457.

- [5] 程军圣,史美丽,杨宇.基于LMD与神经网络的滚动轴承故障诊断方法[J].振动与冲击,2010,29(8): 141-144.
- [6] Huang N E, Shen Z, Long S R. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis [J]. Royal Society of London Proceedings, 1998, 454(A): 903-995.
- [7] Smith S J. The local mean decomposition and its application to EEG perception data [J]. Journal of the Royal Society Interface, 2005, 2(5): 443-454.
- [8] 程军圣,张亢,杨宇,等.局部均值分解与经验模式分解的对比研究[J].振动与冲击,2009,28(5): 13-16.
- [9] 张澎涛,刘晋浩.连续小波变换在机械故障特征提取中的应用[J].吉林大学学报:信息科学版,2014,32(2): 172-176.
- [10] Breiman L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.
- [11] Breiman L, Friedman J, Olshen R, et al. Classification and regression trees [M]. New York: Chapman & Hall, 1984.
- [12] 黄文,王正林.数据挖掘:R语言实战[M].北京:电子工业出版社,2014: 242-245.
- [13] 张鑫,王宏志.支持向量机在短期负荷预测中的应用[J].长春工业大学学报:自然科学版,2009,30(4): 412-416.