

DOI: 10.7652/xjtuxb202002001

采用单类随机森林的异常检测方法及应用

张西宁, 张雯雯, 周融通, 向宙

(西安交通大学机械制造系统工程国家重点实验室, 710049, 西安)

摘要: 针对随机森林算法不能处理异常检测问题的局限,提出了一种基于改进格雷厄姆扫描法的单类随机森林,实现了随机森林在只有单类样本时的分类应用。在格雷厄姆扫描法的算法原理以及操作流程的基础上,引入了边界软化率的概念,增大了数据点外边界的柔性。利用射线法生成与输入样本反分布的数据集,使得传统的随机森林模型经过训练后成为拥有精细决策边界的单类随机森林。训练好的模型输出待测数据的异常概率。在 XJTU-SY 轴承数据集上验证了所提方法对于滚动轴承状态监测的有效性。分析结果表明,单类随机森林能准确分离正常运行数据与退化数据,并且可以通过调整边界软化率来实现异常值检测准确率与召回率的平衡。当预测概率阈值取 0.5 时,硬边界的单类随机森林可以实现 98.37% 的检测准确率以及 0.972 的召回率,而 0.05 的边界软化率可以得到最小的退化后期预测均方根误差,较硬边界降低 1.01%。随着阈值的增大,边界软化率提供了有力的召回率保障。

关键词: 格雷厄姆扫描法;射线法;单类随机森林;特征融合;轴承状态监测

中图分类号: TH17 **文献标志码:** A **文章编号:** 0253-987X(2020)02-0001-08

Anomaly Detection Method Based on One-Class Random Forest with Applications

ZHANG Xining, ZHANG Wenwen, ZHOU Rongtong, XIANG Zhou

(State Key Laboratory for Manufacturing Systems Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Aiming at the limitation that the random forest algorithm cannot deal with the problem of anomaly detection, a one-class random forest based on improved Graham scanning method is proposed, which realizes the classification application of the random forest with only a single class of samples. Following the principle and process of Graham scanning method, the concept of boundary softening ratio is introduced to increase the flexibility of the outer boundary of data points. Ray casting is used to generate data set with inverse distribution of input samples to make the traditional random forest model become a one-class random forest with fine decision boundary after training, which outputs the abnormal probability of the data to be tested. The effectiveness of this method for the condition monitoring of rolling bearings is verified on the XJTU-SY bearing data set. The results show that the one-class random forest can accurately separate normal data from degradation data. Adjusting the boundary softening ratio, the balance between outlier monitoring accuracy and true positive rate can be realized. The one-class random forest with hard boundary can achieve detection accuracy of 98.37% and recall rate of 0.972 with threshold for

0.5. At the boundary softening rate of 0.05, the minimum RMSE for post-degradation prediction can be obtained, which gets 1.01% lower than that of the hard boundary. With the increasing threshold, the boundary softening rate provides a strong recall rate guarantee.

Keywords: Graham scanning; ray casting; one-class random forest; feature fusion; bearing condition monitoring

异常检测问题是机器学习算法的一个常见应用,在网络入侵、恶意欺诈等方面得到了广泛应用^[1]。与垃圾邮件识别等监督学习方法存在区别的是,异常检测问题中只含有非常少量的正类(异常)样本和大量的负类(正常)样本,并且可能存在许多不同种类的已知或未知的异常^[2]。

对于结构化数据的异常检测方法主要是通过对正常样本的描述,给出一个正负样本在特征空间中的决策边界,将边界外的数据视为异常^[3]。常用方法有基于图形位置分布的箱线图法、基于统计与数据分布的方法、基于密度检测的局部异常因子算法^[4]、基于划分思想的孤立森林法^[5]以及单类支持向量机^[6]等方法。随机森林在很多领域的多分类问题中取得显著的效果,大多对于随机森林的改进措施从提高决策树精度、增大决策树间差异和优化投票方式的角度入手来提升集成效果。但是,由于随机森林构造过程基于不纯度的减小而无法处理单分类问题^[7],因此在异常检测领域应用受限。

为保证现代机械装备安全可靠运行,降低维护成本,预测与健康的管理技术越来越受到企业的重视^[8]。通过获取反映轴承运行状态的信息并提取性能特征,可以实现状态监测与寿命预测^[9]。在工程实际中,往往难以获取滚动轴承的全寿命周期监测数据。在可能发生的故障种类未知的前提下,只根据正常状态下的振动数据来训练模型更加符合实际的监测条件。将滚动轴承的在线数据作为测试数据输入到训练好的正常模型中,输出的异常概率值越小,说明该振动数据是由正常状态下的滚动轴承产生的概率越大。当滚动轴承的性能开始退化时,模型输出的异常概率值会增大,从而反映滚动轴承所处的状态,实现轴承的状态监测。

本文提出了一种基于软边界的单类随机森林方法,仅需正常数据作为模型输入,通过调整边界软化率得到合适的软凸包作为外边界,生成反分布的数据点构成正样本集,与凸包内的正常数据构成的负样本集共同训练传统的二分类随机森林。训练得到的模型即为单类随机森林模型,可以输出测试数据的异常概率。在滚动轴承状态监测中的应用验证了

该方法的有效性。

1 理论基础

1.1 随机森林

随机森林是以分类回归树为基学习器的装袋集成学习方法^[10],通过自举采样和随机子空间分割赋予的双重随机性解决了决策树泛化能力弱的缺点。在决策树的生长过程中,节点以基尼系数表示的最小化不纯度函数为分裂标准,选择基尼系数最小的特征及其对应的特征值^[11]。对于给定的样本集合 D ,基尼系数表达为

$$G_{\text{ini}}(D) = 1 - \sum_{k=1}^K \left(\frac{|C_k|}{|D|} \right)^2$$

式中: C_k 是 D 中属于第 k 类的样本子集; K 是类的个数。

当训练集中所有的样本都为同一类时,数据的不纯度指标始终为0,无法通过降低不纯度指标来构建决策树,故无法解决单类样本下的分类问题。

1.2 空间数据处理

凸包是计算几何(图形学)中的一个概念。给定二维平面上的点集,凸包指连接最外层的点而构成的凸边形,它包含了点集中的所有点。常用的有两种算法来计算点集的凸包,即格雷厄姆扫描法和贾维斯步进法^[12]。两者均按逆时针方向输出凸包上的点,从时间复杂度来看,后者不适于点数过多的情况。

处理空间数据是经常需要判断一个点是否是多边形内,常用的方法有:①射线法,从待判断点向某一方向作射线,依据和边界的交点个数奇偶性来判断^[13];②转角法,根据多边形顶点和待判断点连线方向的正负求和来判断;③夹角和法,依据多边形内部的点与所有边的夹角和等于 360° ;④面积和法,依据多边形内部的点与所有边组成的三角形的面积和等于多边形面积。与多个面积的计算和涉及反三角函数的角度计算相比,运用求交运算与坐标对比的射线法更为简单。

2 单类随机森林方法

2.1 数据预处理

单类随机森林仅通过对正常数据的训练即可判

断待测数据发生异常的概率。输入模型的特征分布在 $[0,1] \times [0,1]$ 的平面范围内,训练数据与测试数据均需经过同样的预处理方法,处理过程如图 1 所示。

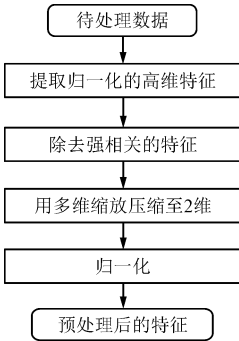


图 1 数据预处理步骤

出于充分描述样本的目的,往往需要提取高维特征,本文在去除其中强相关的特征后,选用多维缩放(MDS)^[14]的降维方法进行特征约简,从而得到 2 维特征。最后进行归一化,即可得到符合模型输入条件的特征。

2.2 基于格雷厄姆扫描法的软边界提取方法

格雷厄姆扫描法主要用一个栈来解决凸包问题。遍历点集 P 将符合条件的压栈,算法终止时,栈中的点就是凸包的顶点。该算法具体步骤如下。

- (1) 选出点集 P 中 y 坐标最小的点记为 p_1 , 如果 y 坐标相同,则相同点中 x 坐标最小的点。
- (2) 以 p_1 为极点,把其余点按照极角从小到大排序,极角相同的点按照到 p_1 的距离从小到大排序。
- (3) 把 p_1, p_2, p_3 压入栈,栈顶元素记作 s_t ,次顶元素记作 s_{nt} 。

(4) 逆时针遍历剩下的点,对 $i=4,5,\dots,n$,弹出没有左转(由 s_{nt}, s_t, p_i 构成的两个向量乘积大于零)的栈顶元素。

(5) 得到逆时针排序的凸包顶点构成的栈。

考虑到正常数据中也会混入一些噪声或者不足以表征正常状态的数据,对于用来提取外边界的点进行了筛选。通过引入边界软化率 r ,去掉点集中距离均值向量最远的 $n \times r$ 个点(n 为总点数),对保留下的点提取外边界。边界软化率为 0 时即为普通的格雷厄姆扫描法,用所有输入的点生成凸包。

引入了边界软化率的格雷厄姆扫描法在提取软边界时的具体步骤如图 2 所示。

2.3 单类随机森林模型

单类随机森林模型分为软边界生成,反分布数据生成,传统随机森林训练 3 个部分,模型见图 3。

模型仅需要正常数据作为输入。软边界生成时

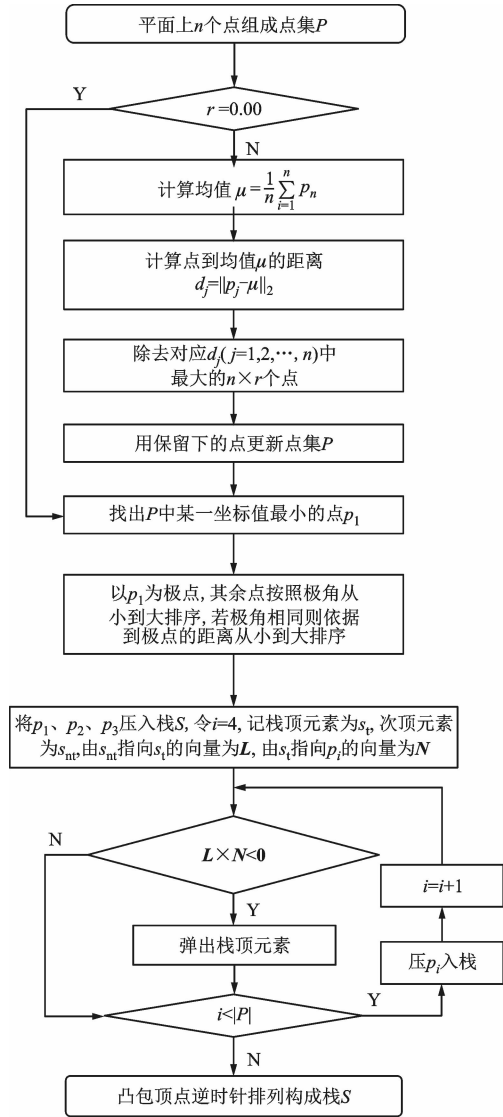


图 2 软边界格雷厄姆扫描法步骤

可设置参数边界软化率 r ,默认为 0.05。反分布生成部分无需任何参数,可以生成数量与用于生成软边界的点数相同的点集。反分布点的生成过程采用射线法的思想,即边界外的点与多边形相交次数为偶数,边界内的点与多边形相交次数为奇数。软边界内的点构成负样本集 N_{eg} ,反分布生成的点构成正样本集 P_{os} ,将两者用于训练传统的随机森林模型,即可输出单类随机森林模型。

将待测数据输入训练好的单类随机森林模型,可以得到 $[0,1]$ 内的异常概率,概率越大说明异常的可能性越大。

3 轴承故障状态监测

3.1 XJTU-SY 数据集

本文使用的 XJTU-SY 轴承数据集^[15]是由西安交通大学与晟阳科技联合发布的滚动轴承加速寿命

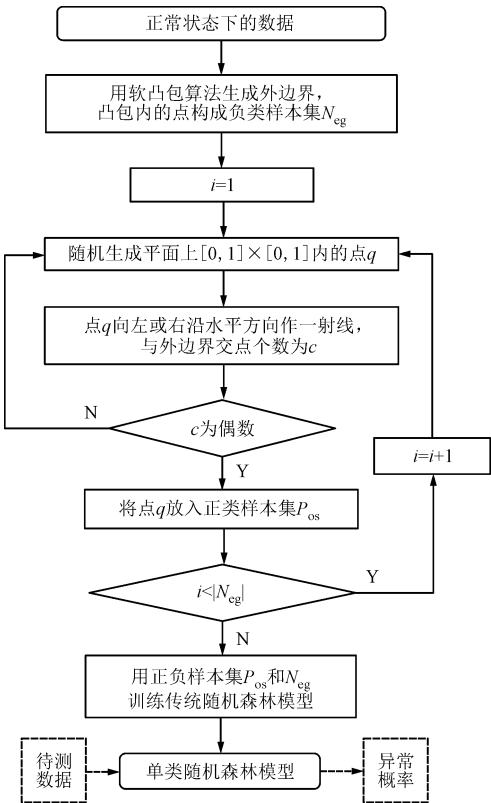


图 3 单类随机森林模型

实验数据。实验轴承为带同心套外球面轴承 LDK UER204,使用 PCB 352C33 加速度传感器,采样频率为 25.6 kHz,每段采样时长为 1.28 s,采样点数为 32 768,采样间隔为 1 min,见图 4。本次测试使用的轴承发生了外圈剥落,共采样 123 段,转速为 35 Hz,负载为 12 kN。

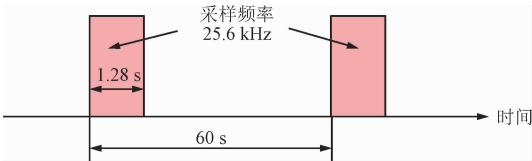


图 4 数据采样设置示意图

将 123 段数据拼接起来得到图 5 所示近似完整的加速退化全寿命数据,可以看到故障逐渐扩大的趋势,直至最后完全失效。对每一段数据进行统计分析后得到图 6 所示箱线图,可以观察到,从第 79 段数据开始,数据点分布范围变广,四分位数变大,

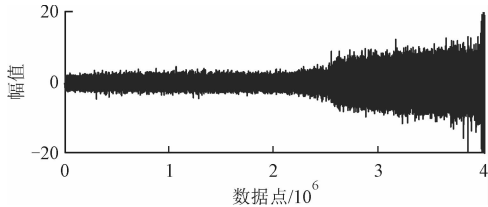


图 5 外圈剥落轴承加速退化全寿命数据

峰值上涨,因此可以认为从 79 段开始失效。

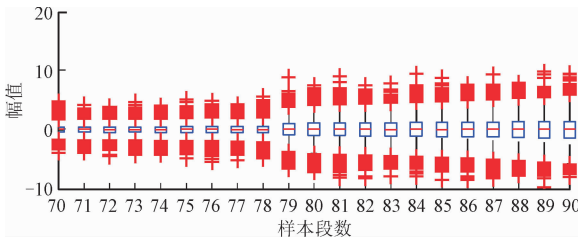


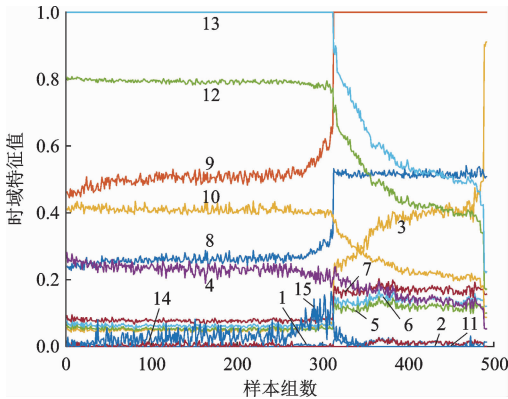
图 6 外圈故障轴承全寿命周期数据

3.2 特征工程

为了丰富样本数量,将每一段振动信号分为 4 组,从而构造了 492 组样本,从第 316 组开始表现出性能退化。本文构造了表 1 所示的基于时域、频域、时频域的 48 维混合特征集^[16-17]。归一化的各类特征见图 7~图 11。经过相关性分析后,取相关性最弱的 10 个特征通过多维缩放进行降维处理,得到全寿命周期的二维分布散点图 12。图 12 中以 316 组为界限区分正常与退化数据。

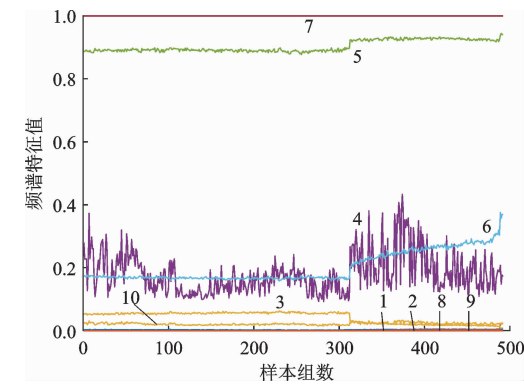
表 1 基于时域、频域、时频域的混合特征集

领域	特征	数量
时域	均值、均方值、方差、标准差、方根幅值、平均幅值、均方幅值、峰值、峰峰值、偏斜度、峭度、波形指标、峰值指标、脉冲指标、裕度指标	15
	频谱振动能量、分散程度、主频带位置等	10
频域	功率谱的平均频率、平均频度、稳定系数、变异系数、偏斜度、峭度、均方根比	7
时频	连续小波尺度能量谱	8
域	小波包分解多尺度空间能量特征	8



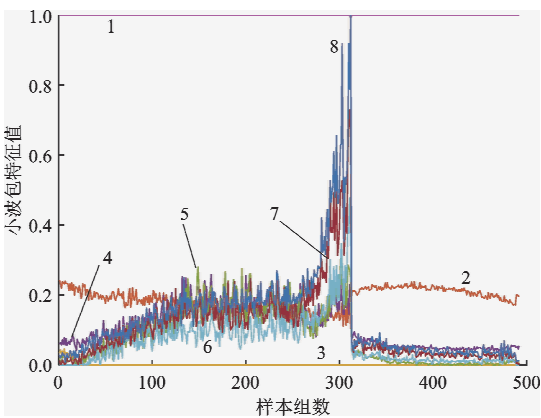
1—均值;2—均方值;3—方差;4—标准差;5—方根幅值;
6—平均幅值;7—均方幅值;8—峰值;9—峰峰值;
10—波形指标;11—峰值指标;12—脉冲指标;
13—裕度指标;14—偏斜度;15—峭度。

图 7 归一化的时域特征



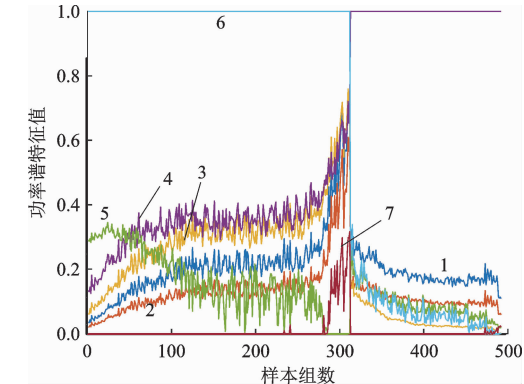
1—振动能量;2,3,4,6,8,9,10—分散程度;
5,7—主频带位置。

图 8 归一化的频谱特征



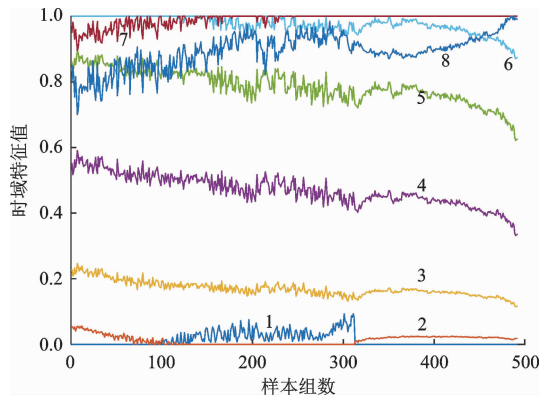
1—频段 1;2—频段 2;3—频段 3;4—频段 4;
5—频段 5;6—频段 6;7—频段 7;8—频段 8。

图 11 归一化的小波包特征



1—平均频率;2—平均频度;3—稳定系数;4—变异系数;
5—偏斜度,6—峭度;7—均方根比。

图 9 归一化的功率谱特征



1—尺度 1;2—尺度 2;3—尺度 3;4—尺度 4;
5—尺度 5;6—尺度 6;7—尺度 7;8—尺度 8。

图 10 归一化的小波能谱特征

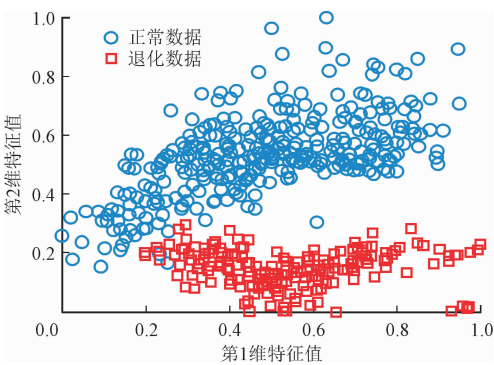
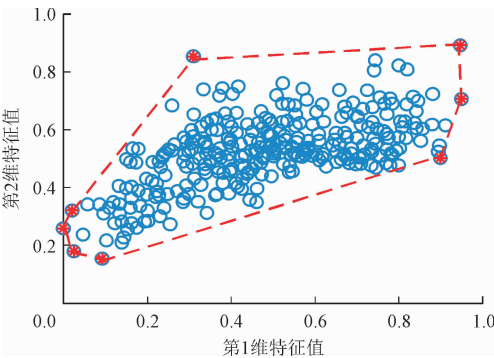
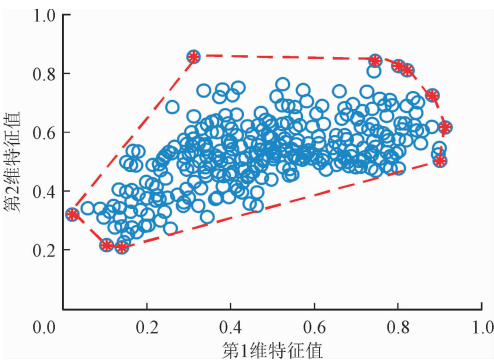


图 12 全寿命数据二维散点分布图



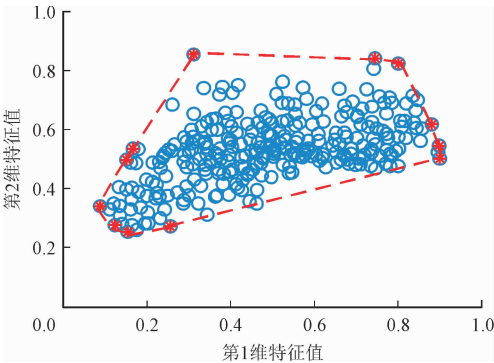
(a) $r=0.00$



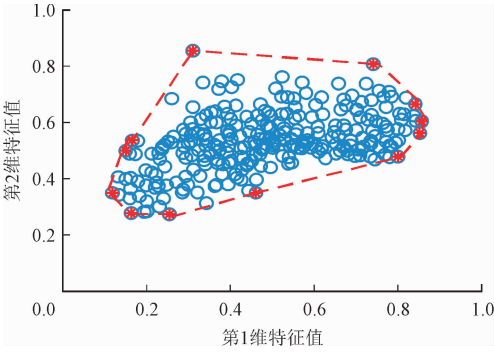
(b) $r=0.02$

3.3 模型训练与测试

训练时使用 1 至 300 组数据作为模型的输入,调整边界软化率 $r=0.00,0.02,0.05,0.10$,得到图 13 所示的边界。随着边界软化率的增大,构造软边



(c) $r=0.05$



(d) $r=0.10$

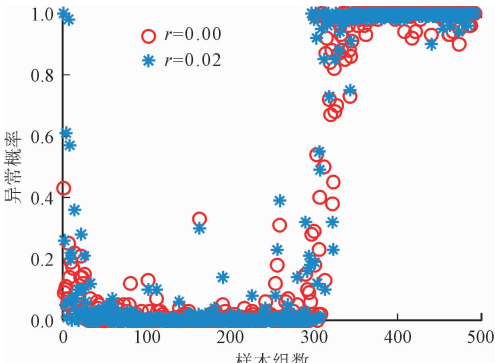
图 13 不同边界软化率下的凸包

界时逐渐排除了一些离群点,使边界更加圆滑。这样在生成反分布数据的时候,缩小了正负样本的间隔,训练模型时得到的决策边界划分得更加细致。

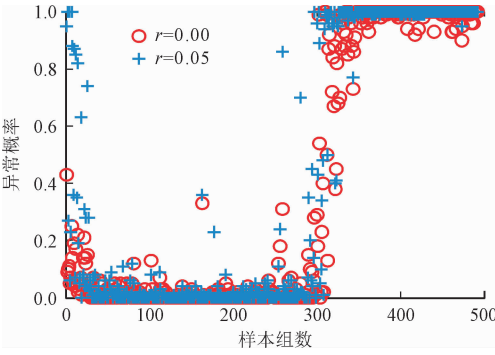
在二分类问题中,召回率^[18]指预测为正的正类样本占真实正样本的比例。召回率越大,漏报率越低。准确率指预测正确的样本占有所有样本的比例。不平衡问题中该指标过于片面,异常检测问题更重视将所有的正类样本都检测出来,即召回率要大。

将整个全寿命周期的数据看作测试数据,输入训练好的模型中,得出图 14 所示的检测结果。以 0.5 为阈值,边界软化率 $r=0.00$ 时,可以将所有的数据判断正确,过渡阶段呈现出异常概率上涨的趋势。对比软边界($r=0.02, 0.05, 0.10$)与硬边界($r=0.00$)得到的测试结果,直观来看,软边界较硬边界得到的结果有所上浮,对异常结果给出更明确的界定,对正常结果的置信度在小范围内有所下降。

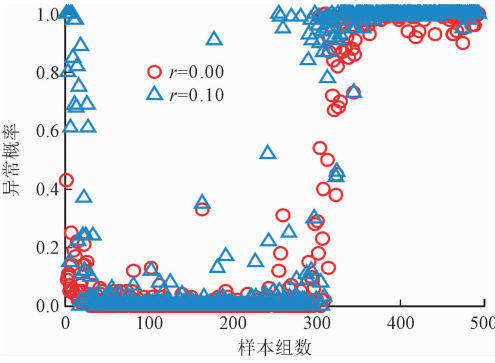
选取两个平稳阶段进行分析。第 100~300 数据段内, $r=0.02$ 的预测概率平均提高 0.935%, $r=0.05$ 的预测概率平均提高 2.02%, $r=0.10$ 的预测概率平均提高 5.995%。第 350~450 数据段内, $r=0.02$ 的预测概率平均提高 0.13%, $r=0.05$ 的预测概率平均提高 0.5%, $r=0.10$ 的预测概率平均



(a) 检测 1



(b) 检测 2



(c) 检测 3

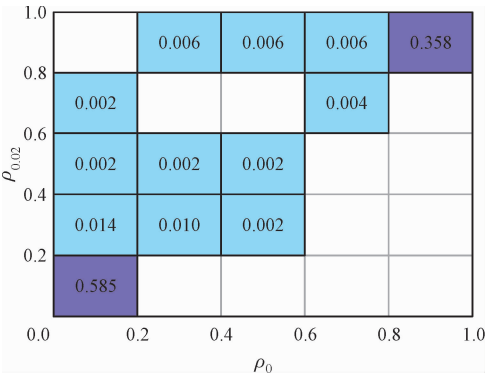
图 14 不同边界软化率下的检测结果

提高 0.32%。

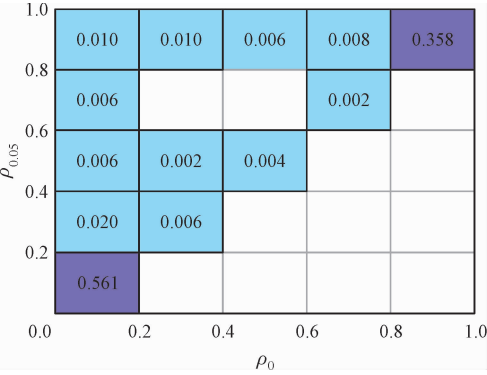
正常状态初期的某些数据出现了扩散现象,甚至误判,可能的原因是,运行初期轴承出于磨合阶段,与正常运行的状态略有不同,被软化后更加细致严格的边界识别出区别从而划分为正类。但是,运行初期的数据在实际应用中不会作为测试数据,因此不影响状态监测的实际准确率。

3.4 检测结果分析

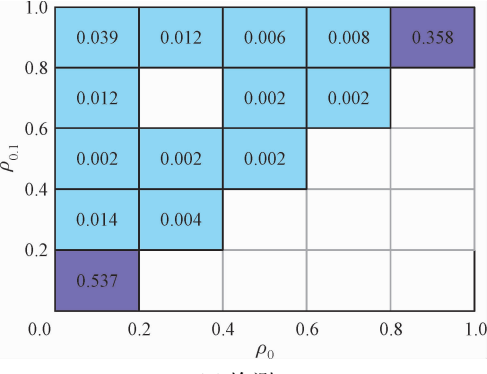
将 3 种软边界得到的检测结果与硬边界进行对比见图 15, ρ_r 表示边界软化率为 r 时检测得到的异常概率分布, $r=0.00, 0.02, 0.05, 0.10$, 同时得出以下结论。



(a)检测 1



(b)检测 2



(c)检测 3

图 15 不同边界软化率下的检测结果对比

- (1)原来以 0~0.2 的概率被判断为正常的数据随边界软化率的增加向更大的概率扩散,甚至在 $r=0.10$ 时有 3.9% 的数据以 0.8 以上的概率被判断为异常。
- (2)原来以 0.2~0.4 的概率被判断为正常的数据不会获得更低的异常概率。
- (3)原来得到 0.4~0.6 的异常概率的数据(绝大部分为过渡阶段数据)大部分出现了异常概率增大的趋势。
- (4)原来以 0.6~0.8 的概率被判断为异常的数据有一部分得到了更高的概率被判断为异常。
- (5)原来以 0.8~1.0 的概率被判断为异常的数

据量保持不变。

为观察全寿命周期上的误差分布情况,将第 1~300 组数据划分为正常阶段,第 301~340 组数据划分为退化初期(过渡阶段),第 341~492 组数据划分为退化后期。理想状态下,第 1~312 组的检测结果均为 0,第 313~492 组的检测结果均为 1。依据此理想状态的设定,硬边界时可得 0.972 的召回率以及 98.37% 的准确率,且第 330 组之后的准确率可达 100%。统计各阶段在不同边界软化率下的均方根误差,得到图 16 所示的误差分布,同时得到如下结论。

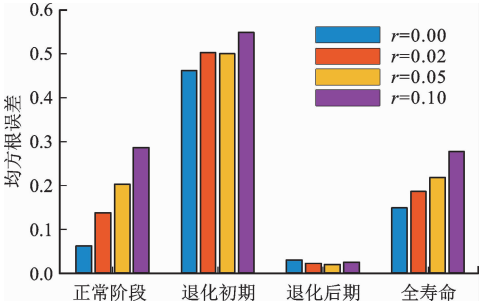


图 16 各阶段在不同边界软化率下的检测误差分布情况

- (1)检测误差主要存在于退化初期,正常阶段次之,退化后期误差最小。
- (2)随着边界软化率的增加,正常阶段、全寿命周期的误差均呈增加趋势。
- (3)退化初期的误差在边界软化率取 0.05 时有所降低,总体增大趋势比正常阶段和全寿命周期小。
- (4)随着边界软化率的增加,退化后期的误差总体呈下降趋势,在边界软化率取 0.10 时回升。

此前的报警阈值均设置为 0.5,即预测概率大于 0.5 认为是异常。实际应用中,在远离被监测设备寿命时可将阈值调大。由图 17 可知,当阈值大于 0.6 时,正类召回率随边界软化率的增大而增大,即边界越软,漏报率越低;阈值越大,软化率的贡献越大。

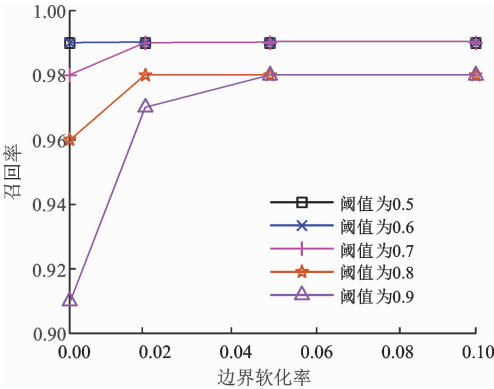


图 17 不同阈值下的召回率和软化率的关系

综合以上分析,对本文所使用的数据集和数据处理方法,取边界软化率为 0.05 可以得到最好的检测结果。

4 结 论

本文提出了基于改进格雷厄姆扫描法的单类随机森林方法,用于处理异常检测问题。为验证所提方法的有效性,在轴承全寿命周期数据上进行了测试,得到以下结论。

(1)预测概率阈值取 0.5 时,硬边界的单类随机森林可以实现 98.37% 的检测准确率以及 0.972 的召回率。

(2)阈值取 0.5 时,0.05 的边界软化率可以得到最小的退化后期预测均方根误差,较硬边界降低 1.01%。

(3)随着阈值的增大,边界软化提供了有力的召回率保障。0.05 的边界软化率可以在阈值为 0.9 时增强 7% 的召回率。

参考文献:

- [1] 邱卫,杨英杰,汪永伟,等. 基于改进遗传算法和隐 Markov 模型的协议异常检测方法 [J]. 计算机应用研究, 2016, 33(4): 1164-1168.
QIU Wei, YANG Yingjie, WANG Yongwei, et al. Protocol anomaly detection method based on improved genetic algorithms and hidden Markov model [J]. Application Research of Computers, 2016, 33(4): 1164-1168.
- [2] 曹玉苹,卢霄,田学民,等. 基于动态单类随机森林的非线性过程监控方法 [J]. 化工学报, 2017, 68(4): 1459-1465.
CAO Yuping, LU Xiao, TIAN Xuemin, et al. Non-linear process monitoring using dynamic one-class random forest [J]. CIESC Journal, 2017, 68(4): 1459-1465.
- [3] 史忠奎,李培军,罗伦,等. 基于形态学属性剖面 and 单类随机森林分类的道路路域新增建筑物提取方法 [J]. 北京大学学报(自然科学版), 2018, 54(1): 105-114.
SHI Zhongkui, LI Peijun, LUO Lun, et al. A method for extraction of newly-built buildings in road region using morphological attribute profiles and one-class random forest [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2018, 54(1): 105-114.
- [4] 周绪川,蔡利平. 基于局部异常因子的无监督学习集成异常检测 [J]. 西南民族大学学报(自然科学版),

2012, 38(3): 444-448.

ZHOU Xuchuan, CAI Liping. Ensemble anomaly detection with unsupervised learning method based on local outlier factor algorithm [J]. Journal of Southwest University for Nationalities (Natural Science Edition), 2012, 38(3): 444-448.

- [5] LIU F T, TING K M, ZHOU Z H. Isolation forest [C] // 2008 Eighth IEEE International Conference on Data Mining. Piscataway, NJ, USA: IEEE, 2008: 413-422.
- [6] 尚文利,李琳,万明,等. 基于优化单类支持向量机的工业控制系统入侵检测算法 [J]. 信息与控制, 2015, 44(6): 678-684.
SHANG Wenli, LI Lin, WAN Ming, et al. Intrusion detection algorithm based on optimized one-class support vector machine for industrial control system [J]. Information and Control, 2015, 44(6): 678-684.
- [7] DÉSIR C, BERNARD S, PETITJEAN C, et al. One class random forests [J]. Pattern Recognition, 2013, 46(12): 3490-3506.
- [8] 胡昌华,施权,司小胜,等. 数据驱动的寿命预测和健康管理技术研究进展 [J]. 信息与控制, 2017, 46(1): 72-82.
HU Changhua, SHI Quan, SI Xiaosheng, et al. Data-driven life prediction and health management: state of the art [J]. Information and Control, 2017, 46(1): 72-82.
- [9] 张小丽,王保健,马猛,等. 滚动轴承寿命预测综述 [J]. 机械设计与制造, 2015(10): 221-224.
ZHANG Xiaoli, WANG Baojian, MA Meng, et al. A review of life prediction for roller bearing [J]. Machinery Design & Manufacture, 2015(10): 221-224.
- [10] BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1): 5-32.
- [11] 李航. 统计学习方法 [M]. 北京: 清华大学出版社, 2012: 67-71.
- [12] 毛鹏. 快速凸包计算实现及其应用 [D]. 西安: 西安电子科技大学, 2013: 15-20.
- [13] 向俊,王静,夏幼明. 判断点与多边形拓扑关系的改进算法 [J]. 计算机工程与设计, 2014, 35(5): 1732-1737.
XIANG Jun, WANG Jing, XIA Youming. Improved algorithm to determine topological relationship between point and polygon [J]. Computer Engineering and Design, 2014, 35(5): 1732-1737.
- [14] 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2016: 225-240.

(下转第 157 页)