DOI: 10.19650/j.cnki.cjsi.J1905320

多分类器集成加权均衡分布适配的滚动轴承寿命阶段识别*

陈仁祥[1,2],吴昊年[1],杨黎霞[1],唐林林[1],徐向阳[1]
(1. 重庆交通大学 交通工程应用机器人重庆市工程实验室 重庆 400074;
2. 重庆大学 机械传动国家重点实验室 重庆 400030)

摘 要:针对不同工况下样本有限不平衡造成滚动轴承寿命阶段识别中少数类样本无法被有效识别的问题,提出了多分类器集成加权均衡分布适配的滚动轴承寿命阶段识别方法。首先,采用随机抽样的方式获得源域多样本训练集,为目标域预测伪标签的同时赋予样本不同的初始权重,充分训练少数类样本;然后,在再生核希尔伯特空间训练各源域样本集的分类器,并通过迭代的方式优化伪标签、更新权重矩阵;最后,通过多分类器集成策略将合适的基分类器集成为强分类器,以获得最终识别结果。结合 F-score 评价标准,使用宏平均与微平均评价指标对多分类任务进行评价避免了准确率对识别结果的误导。在两组滚动轴承寿命阶段数据集上进行实验验证,证明了所提方法的可行性和有效性。

关键词: 样本有限不平衡;滚动轴承;寿命阶段识别;多分类器集成
中图分类号: TH165+.3 TP18 **文献标识码**: A **国家标准学科分类代码**: 510.40

Rolling bearing life stage recognition based on multi-classifier integration of the weighted and balanced distribution adaptation

Chen Renxiang[1,2], Wu Haonian[1], Yang Lixia[1], Tang Linlin[1], Xu Xiangyang[1]
(*1.Chongqing Engineering Laboratory for Transportation Engineering Application Robot, Chongqing Jiaotong University, Chongqing 400074, China; 2.The State Key Laboratory of Mechanical Transmission, Chongqing University, Chongqing 400030, China*)

Abstract: For rolling bearing life stage identification, a small number of samples cannot be effectively identified due to the limited sample imbalance under different working conditions. To solve this problem, a multi-classifier integration of the weighted and balanced distribution adaptation method is proposed. Firstly, the training set of multiple samples in source domain is obtained by random sampling, and different initial weights are given to the samples while predicting false labels in target domain. In this way, a few samples can be trained adequately. Then, the classifiers of sample sets in the source domain are trained in the reproducing kernel Hilbert space, and the pseudo labels are optimized. Meanwhile, the weight matrix is updated iteratively. Finally, the strategy of multi-classifier ensemble is achieved. The appropriate base classifier is integrated into a strong classifier to obtain final recognition results. Combining with F-score evaluation criteria, macro-average and micro-average evaluation indexes are used to evaluate multi-classification tasks, which can avoid misleading recognition results by accuracy. Experiments on two data sets of rolling bearing life stages verify that the proposed method is feasible and effective.

Keywords: sample finite unbalance; rolling bearing; life stage identification; multi-classifier integration

收稿日期:2019-06-29 Received Date:2019-06-29
*基金项目:国家自然科学基金(51975079)、国家自然科学基金(51975079,51975078)、重庆市教委科学技术研究项目(KJQN201900721)、交通工程应用机器人重庆市工程实验室开放基金(CELTEAR-KFKT-201803)、机械传动国家重点实验室开放基金(SKLMT-KFKT-201710)、重庆交通大学硕士研究生科研创新项目(2019S0109)资助

0 引　　言

对滚动轴承寿命阶段进行识别可以监测其性能衰退过程、准确评估其寿命阶段,有效防止重大事故的发生[1]。传统的寿命阶段识别方法把不同类型的数据同等看待,致力于提高整体的识别准确度[2]。但现实中滚动轴承受轴承制造精度、使用环境、工况变化等众多因素影响,不同寿命阶段的轴承样本数量不平衡(即不同类别样本数量差别较大),且受实验成本及周期的限制,只能获得少数型号轴承的有限寿命样本。所以在滚动轴承寿命阶段数据有限且数量不平衡时,正确识别少数类样本的意义高于整体分类准确率[3]。

针对机械状态识别中不平衡问题,相关学者已展开研究。如:何大伟等[4]通过建立基于代价敏感支持向量机的故障诊断模型以解决误分类代价不同的轴承故障识别问题;冯慧玲等[5]对超球大间隔支持向量机进行改进以进行不均衡下轴承故障诊断;赵帅等[6]结合随机森林与主成分分析模型进行样本不平衡下刀具磨损状态识别。这些方法均取得了较好结果,但所使用的传统机器学习方法受独立同分布条件限制,还需要足够多的已标注训练样本进行学习。

不同工况下训练样本与测试样本分布差异较大,且部分样本较难获得或收集代价很高[7],加大了样本有限不平衡下识别的难度。要解决样本有限不平衡下滚动轴承寿命阶段识别的问题,主要克服以下两个困难:1)训练数据不充分,少数类样本包含的信息淹没在多数类信息中,无法有效训练模型;2)弱分类器对少数类有较高的识别错误率,识别结果倾向于多数类。

迁移学习打破传统机器学习独立同分布的假设,利用少量有标记数据,挖掘不同工况数据的有价值信息共同训练模型,实现跨领域、跨任务学习。目前已有迁移学习方法用于进行不同工况或样本不平衡下识别问题。文献[2,7]以迁移学习 TrAdaboost 算法为基础进行改进以解决二分类的样本不平衡问题。加权均衡分布适配方法[8](weighted balanced distribution adaptation, W-BDA)利用类先验精确地逼近目标类别条件分布,并以此训练分类器以提升性能,在一定程度上解决了样本数量不平衡的问题,但面对不平衡程度高、少数类样本数量少时,表现乏力。

针对以上问题,提出了多分类器集成加权均衡分布适配(multi-classifier integration weighted balanced distributed adaptation, MC-W-BDA)的滚动轴承寿命阶段识别方法。首先在保证少数类样本被充分选择的情况下随机抽样多数类样本获得源域多样本训练集,充分训练少数类样本,提高少数类样本权重,并为没有标签的目标域测试样本直接预测得到多个目标域伪标签;然后依据源域多样本训练集与已获得的目标域标记样本集,采用类先验概率逼近条件分布概率的策略建立基分类器分别获得各自样本的识别结果;最后,结合多分类器集成策略整合基分类器信息形成强分类器,完成多分类识别任务,很好地解决了样本有限不平衡下滚动轴承寿命阶段识别的问题。采用两组滚动轴承数据进行验证,证明了本文方法的可行性和有效性。

1 加权均衡分布适配算法原理

假设源域 $D_S=\{x_{si},y_{si}\}_{i=1}^{n}$,$x_i$ 是源域特征样本集,$\boldsymbol{y}_i$ 是样本标签向量。没有标记的目标域 $D_t=\{x_{tj}\}_{j=1}^{m}$。假定源域与目标域的特征空间相同 $X_s=X_t$,并且他们的类别空间也相同 $Y_s=Y_t$,当源域与目标域的边缘分布不同时,$P_s(x_s)\neq P_t(x_t)$,条件概率分布也不相同 $P(y_s\mid x_s)\neq P(y_t\mid x_t)$。

W-BDA 的原理就是寻找一个变换矩阵 $\boldsymbol{A}$,使得经过变换后,源域与目标域的条件分布 $P(y_s\mid \boldsymbol{A}^{\mathrm{T}}x_s)$ 和 $P(y_t\mid \boldsymbol{A}^{\mathrm{T}}x_t)$ 的最大均值差异(maximum mean discrepancy, MMD)[9]最小。由于目标域里没有类别向量 $\boldsymbol{y}_t$ 无法直接求目标域的条件分布。一般都是根据贝叶斯公式忽略 $P(y_t)$,用 $P(x_t\mid y_t)$ 来近似 $P(y_t\mid x_t)$ 两域条件分布差异如下:

$$\left\|P(y_s\mid x_s)-P(y_t\mid x_t)\right\|_H^2=\left\|\frac{P(y_s)}{P(x_s)}P(x_s\mid y_s)-\frac{P(y_t)}{P(x_t)}P(x_t\mid y_t)\right\|_H^2 \tag{1}$$

式(1)隐含地假设了这个类别在每个域中的概率都是相似的。在不平衡样本下这显然是不合理的。所以 W-BDA 先假定 $P(x_s)$ 和 $P(x_t)$ 是不变的,采用类先验概率直接逼近条件分布概率从而避免了对上式条件分布散度的计算。将最终所求通过矩阵技巧和正则化形式化为:

$$\begin{aligned}&\min tr\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\left((1-\mu)\boldsymbol{M}_0+\mu\sum_{c=1}^{c}\boldsymbol{W}_c\right)\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}\right)+\lambda\left\|\boldsymbol{A}\right\|_{\mathrm{F}}^2\\&\text{s.t. }\boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{H}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{A}=\boldsymbol{I},\quad 0\leqslant\mu\leqslant 1\end{aligned} \tag{2}$$

式中:λ 是正则化项 $\|\cdot\|_{\mathrm{F}}^2$ 的 Frobenius 系数;平衡因子 $\mu\in[0,1]$ 用于适配两域分布;X 表示由 x_s 和 x_t 组成的矩阵;$\boldsymbol{A}$ 表示变换矩阵;$\boldsymbol{I}\in R^{(n+m)\times(n+m)}$ 是单位矩阵;$\boldsymbol{H}$ 是中心矩阵,即 $\boldsymbol{H}=\boldsymbol{I}-(1/n)\boldsymbol{1}$,$\boldsymbol{1}$ 是全为 1 的列向量;$\boldsymbol{M}_0$ 是 MMD 矩阵;$\boldsymbol{W}_c$ 为权重矩阵。

$$(\boldsymbol{M}_0)_{ij}=\begin{cases}\dfrac{1}{n^2}, & x_i,x_j\in s\\[2mm]\dfrac{1}{m^2}, & x_i,x_j\in t\\[2mm]-\dfrac{1}{mn}, & \text{其他}\end{cases} \tag{3}$$

$$(\boldsymbol{W}_c)_{ij}=\begin{cases}\dfrac{P(y_s^{(c)})}{n_c^2}, & x_i,x_j\in D_s^{(C)}\\ \dfrac{P(y_t^{(c)})}{m_c^2}, & x_i,x_j\in D_t^{(C)}\\ -\dfrac{\sqrt{P(y_s^{(c)})P(y_t^{(c)})}}{m_c n_c}, & \begin{cases}x_i\in{}_s^{(c)},x_j\in{}_t^{(c)}\\ x_i\in{}_t^{(c)},x_j\in{}_s^{(c)}\end{cases}\\ 0, & \text{其他}\end{cases} \tag{4}$$

式中：$P(y_s^{(c)})$ 和 $P(y_t^{(c)})$ 表示源域和目标域中类 C 各自的类先验。再引入拉格朗日乘子 $\boldsymbol{\Phi}=(\Phi_1,\Phi_2,\cdots,\Phi_d)$，并由式(2)、(3)得到拉格朗日函数为：

$$(\boldsymbol{X}((1-\mu)\boldsymbol{M}_0+\mu\sum_{c=1}^{c}\boldsymbol{W}_c)\boldsymbol{X}^{\mathrm{T}}+\lambda\boldsymbol{I})\boldsymbol{A}=\boldsymbol{XHX}^{\mathrm{T}}\boldsymbol{A}\Phi \tag{5}$$

求取式(5)的目标函数，在式(2)的约束条件下即可得到最佳的映射变换矩阵 A，用以构建基分类器。

2 多分类器集成加权均衡分布适配算法实现

2.1 多分类器集成方法

欲采用加权均衡分布适配方法解决样本有限不平衡问题，必须充分训练少数类样本，以建立有效的训练模型。同时弱分类器无法很好地识别少数类样本，识别结果倾向于多数类的问题同样不容忽视。为此采用如图 1 所示的方式构建源域多样本训练集，并通过弱分类器加权集成的方式提升分类器性能，以解决样本有限不平衡下滚动轴承寿命阶段识别。由于 K 最邻近(k-nearest neighbor，KNN)分类器具有计算简单、无需参数估计与预训练等优点，本文所有基分类器均采用 KNN 构建。

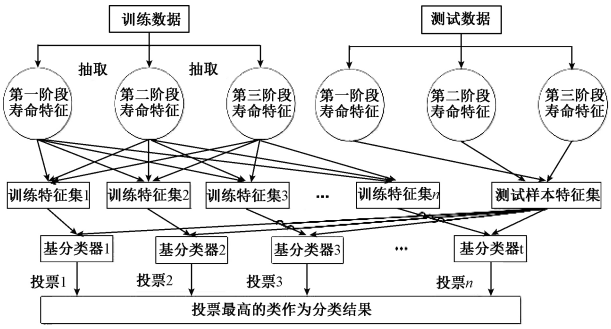


图 1 多分类器集成流程

Fig.1 The flowchart of multi-classifier integration

设 X_S 为 M 个不同寿命阶段滚动轴承特征样本，寿命阶段为 φ_i，$\forall i\in\Lambda=\{1,\cdots,M\}$，其中 φ_1 代表第一寿命阶段样本。对 X_S 各阶段寿命样本中的多数类样本进行无重复随机抽样，抽样数为 n，$n\leqslant\min(\varphi_1,\varphi_2,\cdots,\varphi_M)$，组成源域单样本训练集 x_s；抽取 k 次直到多数类样本全部抽完，得到 k 个源域单样本训练集 $\{x_{s1},x_{s2},\cdots,x_{sk}\}$；在保证少数类样本被充分选择的情况下随机抽样多数类样本获得源域多样本训练集，充分训练少数类样本，提高少数类样本权重，并获得多个目标域伪标签。然后利用式(6)计算伪标签分类器的错误率，将错误率最小的作为基分类器 $h_c(t=1,\cdots,k)$，$\omega_j(j=1,\cdots,M)$ 是第 j 类的标记，$h_c(j)$ 表示第 c 个分类器对第 j 类样本的预测。

$$\varepsilon=\sum_{k=1}^{n}\frac{w_{ck}\,|h_c(x_{sk}-y_{si})|}{\sum_{k=1}^{n}W_{ck}} \tag{6}$$

其中，$D_j(j=1,\cdots,M)$ 为不同类别得票数目。

$$D_j=\{t\mid h_c(j)=\varphi_i\} \tag{7}$$

将 c 个分类器识别结果代入一致性判别公式(见式(8))：

$$f_E(x_{sk})=\operatorname{argmax}(D_j) \tag{8}$$

利用多分类器集成的方式将多个弱分类器集成为强分类器判别获得最终识别结果。

2.2 MC-W-BDA 算法实现

根据上文论述，所提 MC-W-BDA 算法的滚动轴承寿命阶段识别算法实现流程如图 2 所示。该算法实现主要包括以下步骤。

1)训练样本构建：在保证少数类样本被充分选择的情况下随机抽样多数类样本，获得源域多样本训练集，充分训练少数类样本；

2)权重系数设置：在目标域上直接为测试样本预测伪标签，并经过多次迭代得到可靠的目标域标签以类先验概率逼近条件分布概率的策略建立基分类器；

3)多分类器集成：利用多分类器集成的方式将多个基分类器识别结果集成获得最终识别结果。

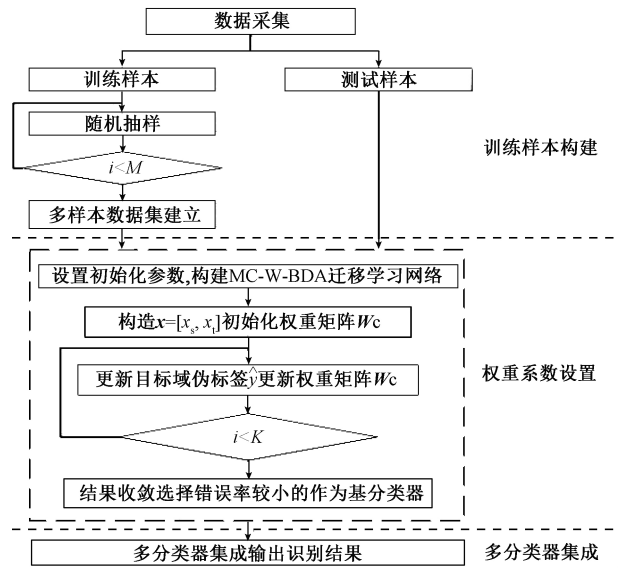


图 2 MC-W-BDA 算法流程

Fig.2 The flowchart of MC-W-BDA algorithm

3　不平衡样本下分类器性能的评价方法

准确率是一般分类器非常直观的度量指标,然而在样本有限不平衡下对滚动轴承寿命阶段进行识别时,准确率难以有效反映分类器的性能。例如在 99:1的不平衡样本比例下,只要将 99 个多数类全部识别达到 99%的识别准确率,就能掩盖少数类被忽略的事实。因此针对不平衡样本分类器性能的评测,本文引入混淆矩阵,利用查准率和召回率建立的不平衡样本的综合评价指标 F-score 进行评判,当 F-score 得分大于 0.6 时,说明分类器识别性能良好。并采用宏平均[10]和微平均作为多类别识别任务衡量分类器性能的评价指标。

表 1　二分类问题的混淆矩阵

Table 1　Confusion matrix for the binary classification problem

样本数	预测为正样本数	预测为反样本数
实际正类样本数	*TP*	*FN*
实际反类样本数	*FP*	*TN*

查准率(Precision,P):

$$P = TP/TP + FP \tag{9}$$

召回率(Recall,R):

$$R = TP/TP + FN \tag{10}$$

F-score:

$$F_\beta = \frac{(1+\beta^2) \times P \times R}{(\beta^2 \times P) + R} \tag{11}$$

其中,宏平均是先对每一个类统计指标值,然后在对所有类求算术平均值。可以清晰地展示不平衡状态下不同类别样本的识别效果。

$$Macro - P = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{12}$$

$$Macro - R = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{13}$$

$$Macro - F = \frac{(1+\beta^2) \times Macro\_P \times Macro\_R}{(\beta^2 \times Macro\_P) + Macro\_R} \tag{14}$$

式中:$P_i = \frac{TP_i}{TP_i + FP_i}$,$R_i = \frac{TP_i}{TP_i + FN_i}$,$\beta$ 值一般取 1。

微平均,是对数据集中的每一个实例不分其类别进行统计建立全局混淆矩阵,然后计算得到的指标。它更多关注的是样本整体的识别效果。

$$Micro - P = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \tag{15}$$

$$Micro - R = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \tag{16}$$

$$Micro - F = \frac{(1+\beta^2) \times Micro\_P \times Micro\_R}{(\beta^2 \times Micro\_P) + Micro\_R} \tag{17}$$

值得注意的是,当 β 取 1 时,微平均与准确率结果一致。本文微平均与宏平均计算中 β 均等于 1,即所计算的微平均与准确率结果一致。

4　性能对比实验分析

4.1　寿命阶段数据采集与特征的选取

实验采用 C36018 型角接触球轴承,节径 15 mm,包含 7 个滚动体,接触角度 15°。在相同转速、不同负载下运行相同圈数,分别以 T1、T2、T3 表示。在运行圈数和转速相同的条件下,负载越大,寿命损耗越多,故 3 种不同负载对应 3 个不同寿命阶段。具体寿命状态信息如表 2 所示。

表 2　寿命状态数据信息

Table 2　Life state data information

寿命阶段	工作状态	运行圈数
T1	4 kg 负载	1.44×10^7
T2	5 kg 负载	1.44×10^7
T3	6 kg 负载	1.44×10^7

采用不同工况条件对处于不同寿命阶段的滚动轴承进行振动信号采集,采集时运行转速分别为:1)A 工况 500 r/min;2)B 工况 1 000 r/min;3)C 工况 1 500 r/min;4)D 工况 2 000 r/min,加载负荷均为 1 kg。采样频率均为 25.6 kHz,采样长度均为 102 400,每种寿命阶段采样 2 次,对每种工况下各寿命阶段数据以 2 048 为分析点数。共 4 种不同工况寿命阶段数据,其中每个阶段 100 个样本,即每种工况共 300 样本。

为更好地表征不同工况下,滚动轴承寿命阶段特征。本文提取 56 维多域特征构建高维特征集,包括最大值、均值、平均幅值、方根幅值、均方根值、均方值、偏斜度、峭度、标准差、方差、波形因数、峰值因数、脉冲因数、裕度指标、峭度指标、偏度指标共 16 个时域特征;频域幅值平均值、频域幅值偏度指标、频域幅值峭度指标、重心频率、均方频率、频率标准差、频率方差、频域频率歪度、均方根频率、频域频率峭度、频域幅值方差、平方根比率共 12 个频域特征(上述 28 个特征计算公式见参考文献[11])。另外还提取了 8 维 db3 小波 3 层小波包能量特征,8 维 db3 小波 3 层小波包相对能量特征,8 维 db3 小波 3 层小波包能量谱熵以及振动信号幅值谱熵、倒谱熵、自相关谱熵、奇异值谱熵共 28 维特征。

4.2　不平衡样本对 W-BDA 算法分类器性能影响实验

本节实验的目的是为了展示不同工况下样本数据集

不平衡时，对 **W-BDA** 算法识别性能的影响。设置 **T1**、T2、T3 寿命阶段样本不平衡比例为 1∶10∶1的 D 工况为训练集。A 工况 T1、T2、T3 各 100 样本，共 300 样本数据集作为测试集。实验进行 500 次，随机抽样产生训练样本数据集，求平均值作为最后识别结果。其中 Z 值代表随机抽样次数，评价指标采用 F-score。其中 W-BDA 采用，$\mu=0.3$，迭代次数 $N=10$，正则化参数 $\lambda=0.001$。实验结果如图 3 所示。从实验结果中可以发现不平衡训练数据对 W-BDA 算法的性能有较大影响。在上述不平衡比例下，W-BDA 识别效果一般且识别结果波动较大。500 次 F-score 性能评价值的平均值约为 0.513，小于 0.6，说明其分类性能一般。

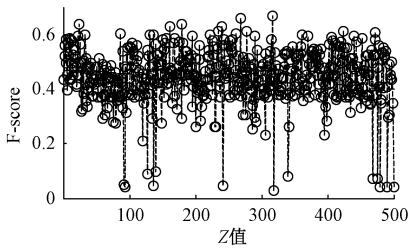


图 3 500 次随机抽样 F-score 值分布

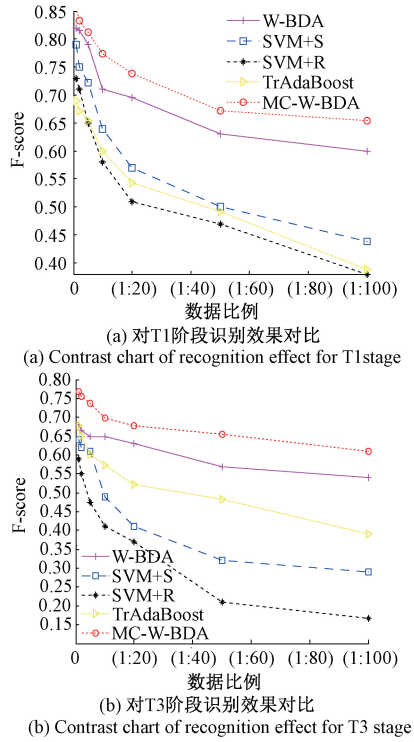
Fig.3 F-score value distribution of 500 random samples

这是因为 A 工况（1 kg，500 r/min）到 D 工况（1 kg，2 000 r/min）转速跨度较大，且不平衡比例的增加也使得为目标域直接预测伪标签的效果较差，无法更精确识别少数类样本。然而在滚动轴承寿命阶段识别中，不同工况寿命阶段样本已很难收集，造成的不均衡数据的比例势必会很大，如何解决样本有限不平衡下基于 W-BDA 算法的寿命阶段识别问题就显得尤为突出。

4.3 不同组合策略下样本有限不平衡的对比试验

为验证本文提出的多分类器集成策略在采样方式上解决数据不均衡情况下识别的优势，并证明本文方法的可行性。从二分类问题入手，依次采用寿命阶段 T1 和 T3 样本为少数类，T2 为多数类，设置多数类与少数类不平衡比例为 100∶1、50∶1、20∶1、10∶1、5∶1、2∶1、1∶1的 D 工况为训练集。测试样本仍选择 C 工况 T1 和 T2 寿命阶段（T2 和 T3）各 100 个，共 200 样本的数据集。对比基于支持向量机（support vector machine，SVM）和少数类过采样方法（synthetic minority oversampling technique，SMOTE）相结合的方法[12]（SVM+S），SVM 和随机采样（random sampling，RS）平衡样本相结合方法（SVM+R）以及迁移学习 TrAdaBoost[13] 方法。

实验中 SVM 采用高斯核函数，宽度为 1；SMOTE 算法 K 值取 5；TrAdaBoost 辅助样本从源域选取，采用迭代次数 $N=20$；W-BDA 采用迭代次数 $N=10$，正则化参数 $\lambda=0.01$。实验结果如图 4 所示。



(a) 对T1阶段识别效果对比

(a) Contrast chart of recognition effect for T1stage

(b) 对T3阶段识别效果对比

(b) Contrast chart of recognition effect for T3 stage

图 4 不平衡样本学习方法对比结果

Fig.4 Comparison results of the balanced sample learning method

观察图 4 发现，随不平衡比例增加 F-score 值均有所下降。其中 SVM+R 方法采用欠采样方式丢弃了很多反例，会造成重要信息的丢失，导致 SVM 分类器识别结果倾向于多数类，性能下降明显。SVM+S 方法通过在少数类样本之间进行插值来平衡样本，SVM 识别效果受新合成样本影响较大，对少数类样本识别性能较差。且 SVM 作为传统机器学习分类器，不同工况下受两个假设限制也是重要原因。迁移学习 TrAdaBoost 算法采用正负样本同样的权值更新策略，未赋予少数类样本不同权重，且其辅助数据中往往存在大量冗余，增大了识别难度。W-BDA 方法在不平衡比例达到 1∶10 时，F-score 得分已经下降至 0.6 左右，对于不平衡比例较大的数据集表现乏力（其中 W-BDA 为 500 次取平均值结果）。MC-W-BDA 算法使用两域类先验概率近似逼近条件概率，算法初始就为不平衡样本多数类与少数类赋予了不一样的权重，多分类器集成策略整合基分类器信息形成强分类器，使得其 F-score 均大于 0.6，对少数类样本识别效果良好。

4.4 多类别不平衡情况下分类器性能实验

MC-W-BDA 算法可以将多类别同时识别完成多分类任务。本节实验仍采用 3.1 节中采集的不同工况下滚动轴承寿命阶段样本集，以验证样本有限不平衡下多类别分类时 MC-W-BDA 算法分类器的性能。设置 D 工况不

平衡数据集(如表 3 所示)为训练集。A(B、C)工况 T1、T2、T3 各 100 样本,共 300 样本作为测试集。

表 3　不同工况下多分类器集成 W-BDA 算法性能(F-score)

Table 3　Performance of W-BDA algorithm for multi-classifier integration under different operating conditions(F-score)

测试	D 工况		
	1:100:5	1:100:10	5:100:5
A 工况	0.688 4	0.692 1	0.724 8
B 工况	0.879 2	0.836 2	0.883 3
C 工况	0.890 5	0.850 1	0.876 8
平均	0.819 4	0.792 8	0.828 3

由表 3 可知,样本有限不平衡时,MC-W-BDA 可以有效识别少数类样本。在只有 1 个训练样本(占整个样本集 0.34%),且不平衡比例达到 1:100 的严格条件下,依旧可以得到很高的 F-score。

上述实验的目标域样本采取 T1、T2、T3 寿命阶段各 100 个,共 300 个为测试集,而实际工程条件下,往往获得的目标域待识别样本也是不平衡状态,即训练样本不平衡,测试样本也不平衡。图 5 详细展示了此种条件下各工况的 F-score 得分情况。训练集选择 D 工况数据,设置 T1、T2、T3 不平衡比例依次 1:100:5的数据集,测试样本选择 A(B、C)工况数据集以 1:N:1的方式设置,结果如图 5、6 所示。

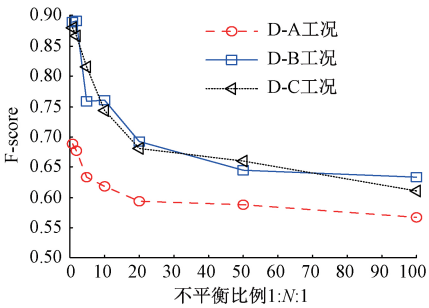


图 5　测试样本不平衡下评价得分

Fig.5　Evaluation scores under unbalanced test samples

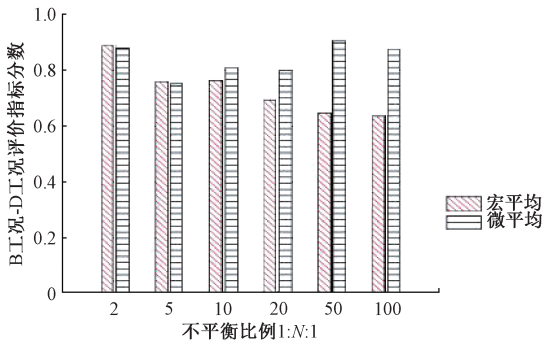


图 6　D-B 工况下多分类评价指标

Fig.6　Multi-classification evaluation index under D-B conditions

由图 5 和 6 可以看出随着测试样本不平衡比例的增加,MC-W-BDA 方法宏平均值逐渐在下降,不平衡比例达到 1:100:1时,依旧可以有效识别测试样本也不平衡状态下仅存的少数类样本。

对比图 5 和 6 可知,F-score 平均达到 0.8 左右。证明本文方法能够较好地解决样本有限不平衡的问题,在保证整体识别精度的前提下,有效识别了少数类样本。

5　全寿命数据实验分析

5.1　全寿命周期数据实验参数

本节实验是为了证明 MC-W-BDA 算法完成多类别不平衡样本识别任务的泛化能力,进一步验证本文方法的有效性。实验采用 PRONOSTIA 实验台采集的文献[14]中加速寿命实验振动信号数据进行实验验证和分析。此数据采样频率为 25.6 kHz,采样间隔为 10 s,每个样本采样时间为 0.1 s。该数据集包含多个工况条件下的全寿命周期实验数据,工况信息如表 4 所示。

表 4　工况信息

Table 4　Working condition information

工况	转速/(r·min^{-1})	载荷/N	样本个数	寿命阶段个数
Data1	1 650	4 200	911	3
Data2	1 500	5 000	1 637	3
Data3	1 800	4 000	2 259	3

滚动轴承从全新装配到完全失效的整个寿命周期共经历 3 种寿命阶段:磨合期、有效工作期和衰退期。实验设置 3 种工况条件下数据样本集:1)Data1 为 1 650 r/min、4 200 N 数据样本集;2)Data2 为 1 500 r/min、5 000 N 数据样本集;3)Data3 为 1 800 r/min、4 000 N 数据样本集。

为了更好地完成寿命阶段识别实验,利用文献[15]的方法划分 3 种寿命阶段,截取阶段明显的寿命阶段样本,把不明确样本进行了少量剔除。其中 Data1 包含磨合期样本 60 个、有效工作期样本 800 个和衰退期样本 37 个;Data2 依次是 100 个、1 100 个、100 个;Data3 依次为 150 个、1 375 个、180 个。

5.2　全寿命周期数据验证实验

本节实验采用表 4 所示 3 组数据,设置训练集与测试集具体信息如表 5 所示。

从表 5 中可以看出不均衡比例最小约为 1:10,最大达到了 1:1 000。将表 4 中 3 组数据交叉验证,用不同评价方法(宏平均与微平均)得到 F-score 评分对比结果如图 7 所示,其中 Data2-1 代表使用表 5 中 Data2 数据集做训练,Data1 数据集做测试,其他同理。

表 5 实验数据信息

Table 5 Experimental data information

数据集	D_S			数据集	D_t		
	多数类样本	少数类样本 1	少数类样本 2		多数类样本	少数类样本 1	少数类样本 2
Data2	1 100	100	100	Data1	800	20	20
Data3	1 375	150	180	Data1	800	60	37
Data1	800	60	37	Data2	500	5	25
Data3	1 375	150	180	Data2	1 100	100	100
Data1	800	60	37	Data3	500	25	25
Data2	1 100	100	100	Data3	1 000	50	1

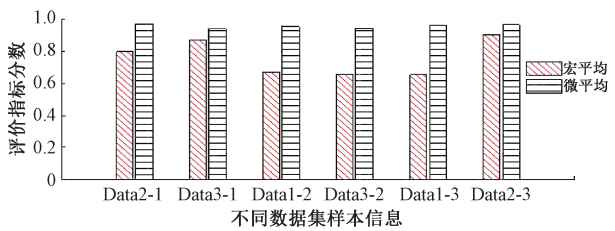


图 7 不同数据集下评价标准结果

Fig.7 Results of evaluation criteria under different data sets

观察图 7 可知,微平均得分全部在 0.9 以上,这意味着在多分类任务中本文方法整体识别精度较高,整体寿命阶段识别准确率可以达到 90%以上。与此同时由每一类统计指标宏平均看到 3 个不同工况下不平衡样本交叉验证 F-score 评分均达到 0.6 以上,平均 0.73,说明本文方法可以很好地识别各阶段的少数类寿命样本。

5.3 迁移学习对比试验

为对比本方法与其他迁移学习算法相比的优势。本文选择在解决不平衡样本上应用广泛的 TrAdaBoost 算法作对比,结合决策树将 TrAdaBoost 改为多分类模型。

对比结果如表 6 所示,其中 TrA(s&t)代表辅助样本使用源域样本, TrA(t&t)代表辅助样本使用目标域样本。仍以宏平均作为评价指标。TrAdaBoost 对于多类别不平衡样本识别能力有限,且受分类器性能影响识别结果倾向于多数类导致宏平均得分较低。本文方法在 F-score 评价指标下都在 0.6 以上很好地完成了少数类样本识别任务,其中最高得分达到 0.899,证明了所提方法的有效性。

表 6 迁移学习算法对比结果

Table 6 Comparison results of transfer learning algorithms

方法	数据					
	Data 2-1	Data 3-1	Data 1-2	Data 3-2	Data 1-3	Data 2-3
本文方法	0.799 2	0.866 1	0.665 7	0.654 8	0.658 2	0.899 1
TrA(s&t)	0.511 8	0.567 2	0.433 2	0.618 1	0.550 1	0.416 9
TrA(t&t)	0.405 5	0.477 2	0.218 6	0.371 5	0.423 2	0.377 1

6 结 论

本文提出了多分类器集成加权均衡分布适配(MC-W-BDA),用于样本有限不平衡下滚动轴承寿命阶段识别问题。使用随机抽样的方式保证少数类样本充分参与训练,更好地赋予目标域样本合适的伪标签。采用类先验更精确地逼近条件分布概率,在提高少数类样本关注度的同时在多分类器集成策略下兼顾了多数类样本的有效识别。实验结果表明,MC-W-BDA 较其他算法在样本有限不平衡下保证整体精度的基础上,有效识别了少数类样本。

参考文献

[1] CHEN R X, CHEN S Y, HE M, et al. Rolling bearing fault severity identification using deep sparse auto-encoder network with noise added sample expansion [J]. Proceedings of the Institution of Mechanical Engineers Part O:Journal of Risk and Reliability. 2017, 231(6): 666-679.

[2] 么素素,王宝亮,侯永宏.绝对不平衡样本分类的集成迁移学习算法[J]. 计算机科学与探索, 2018, 12(7): 130-138.
MO S S, WANG B L, HOU Y H. Integrated migration learning algorithm for absolutely unbalanced sample classification [J]. Computer Science and Exploration, 2018, 12 (7): 130-138.

[3] HE H, GARCIA E A. Learning from imbalanced data[J]. IEEE Transactions on Knowledge & Data Engineering, 2009, 21(9):1263-1284.

[4] 何大伟,彭靖波,胡金海.基于改进 FOA 优化的 CS-SVM 轴承故障诊断研究[J]. 振动与冲击, 2018, 37(18):113-119.
HE D W, PENG J B, HU J H, et al. CS-SVM bearing fault diagnosis based on improved FOA optimization [J]. Vibration and Shock, 2018, 37 (18): 113-119.

[5] 冯慧玲,常国权,孔娟.基于拉普拉斯分值和超球支持向量机的轴承故障诊断方法设计[J].计算机测量与控制, 2015,23(4):1102-1105.
FENG H L, CHANG G Q, KONG J. Design of bearing fault diagnosis method based on Laplace score and superball support vector machine [J]. Computer Measurement and Control, 2015, 23 (4):1102-1105.

[6] 赵帅,黄亦翔,王浩任,等. 基于随机森林与主成分分析的刀具磨损评估[J].机械工程学报,2017,53(21): 192-200.
ZHAO SH, HUANG Y X, WANG H R, et al. Tool wear assessment based on random forest and principal

component analysis [J]. Journal of Mechanical Engineering, 2017,53(21): 192-200.

[7] 陈琼,徐洋洋,陈林清.不平衡数据的迁移学习分类算法[J]. 华南理工大学学报(自然科学版), 2018, 46(1):122-130.
CHEN Q, XV Y Y, CHEN L Q. Migration learning classification algorithm for unbalanced data [J]. Journal of South China University of Technology (Natural Science Edition), 2018, 46 (1): 122-130.

[8] WANG J D, CHEN Y Q, HAO SH J, et al. Balanced distribution adaptation for transfer learning [C]. IEEE International Conference on Data Mining, 2017, doi: 10.1109/ICDM.2017.150.

[9] PAN S J, TSANG I W, KWOK J T, et al. Domain adaptation via transfer component analysis. [J]. IEEE Transactions on Neural Networks, 2011, 22 (2): 199-210.

[10] PILLAI I, FUMERA G, ROLI F. F-measure optimization in multi-label classifiers [C]. 21st International Conference on Pattern Recognition (ICPR), 2012, 2424-2427.

[11] 陈仁祥, 陈思杨, 杨黎霞,等. 基于振动敏感时频特征的航天轴承寿命状态识别方法[J]. 振动与冲击, 2016, 35(17):134-139.
CHEN R X, CHEN S Y, YANG L X, et al. Life state recognition method for space bearings based on sensitive time-frequency features of vibration. [J]. Vibration and Shock, 2016, 35(17):134-139.

[12] LIU Y, AN A J, HUANG X J. Boosting prediction accuracy on imbalanced datasets with SVM ensembles[C]. Advances in Knowledge Discovery & Data Mining, Pacific-Asia Conference, 2006, doi: 10.1007/11731139_15.

[13] DAI W Y, YANG Q, XUE G R, et al. Boosting for transfer learning[C].International Conference on Machine Learning, 2007, doi:10.1145/1273496.1273521.

[14] LI X, LU W F, ZHAI L Y, et al. Predictive Modeling for Life Cycle Reliability Analysis and Machine Health Condition Prediction in Remanufacturing [M]. London: Springer, 2014.

[15] 阙子俊,金晓航,孙毅.基于 MKF 的轴承剩余寿命预测方法研究[J].仪器仪表学报, 2016, 37(9): 2036-2043.
KAN Z J, JIN X H, SUN Y. Research on the prediction method of bearing residual life based on KF [J]. Journal of Instruments and Instruments, 2016, 37 (9): 2036-2043.

作者简介

陈仁祥,分别在 2007 年和 2012 年于重庆大学获得学士学位和博士学位,现为重庆交通大学教授、硕士生导师,主要研究方向为智能测试技术与信号处理。
E-mail: manlou.yue@126.com

Chen Renxiang received his B. Sc. and Ph. D. degrees both from Chongqing University in 2007 and 2012, respectively. He is currently a professor and master supervisor at Chongqing Jiaotong University. His main research interests include intelligent testing technology and signal processing.

吴昊年,2016 年于重庆交通大学获得学士学位,现为重庆交通大学硕士研究生,主要研究方向为机电装备故障诊断及其安全服役、迁移学习。
E-mail: 296018167@qq.com

Wu Haonian received his B. Sc. degree from Chongqing Jiaotong University in 2016. He is currently a M. Sc. student at Chongqing Jiaotong University. His main research interests include fault diagnosis of mechanical and electrical equipment, safety service and transfer learning.

徐向阳(通信作者),2004 年于郑州大学获得学士学位,分别在 2009 和 2012 年于重庆大学获得硕士学位和博士学位,现为重庆交通大学教授、硕士生导师,主要研究方向为机械设计及理论。
E-mail: 296018167@qq.com

Xu Xiangyang (Corresponding author) received his B. Sc. degree from Zhengzhou University in 2004, his M. Sc. and Ph. D. degrees both from Chongqing University in 2009 and 2012, respectively. He is currently a professor and master supervisor at Chongqing Jiaotong University. His main research interest is mechanical design and theory.