不均衡数据下基于CS-Boosting的故障诊断算法*

姚　培，　王仲生，　姜洪开，　刘贞报
（西北工业大学航空学院　西安，710072）

摘要　针对传统Boosting算法在训练样本不均衡数据情况下不能较好地实现转子系统故障诊断的问题，提出了一种基于代价敏感度框架的Boosting故障诊断算法CS-Boosting。该算法建立了一个代价敏感损失函数，通过先验概率公式计算正样本与负样本的惩罚因子，并通过决策规则的训练使代价损失函数最小化。将该算法应用到滚动轴承故障诊断中，并与传统的Adaboost算法进行对比。试验结果表明，在转子系统不能获取更多故障数据的情况下，该算法的故障诊断性能较其他算法有明显的提高。

关键词　代价敏感度；滚动轴承；Boosting算法；CS-Boosting；代价损失函数
中图分类号　TH17；TP306

引　言

滚动轴承是旋转机械中最常用、最易受损伤的零部件之一，其状态检测与诊断一直为大家所重视[1]。Adaboost算法因其良好的分类精度，已成功应用于机械故障诊断。在故障诊断应用领域，工程人员准备收集异常样本数据的时候，机械可能已经损坏而无法正常运行，故障样本数量相对于正常样本是相当少的，因此样本不平衡数据下的故障诊断方法由此产生[2]。传统的Adaboost故障诊断方法是在各个类别样本数量相同的前提条件下进行，对于利用代价敏感度Adaboost算法（CS-Boosting）的不均衡样本数据故障诊断，国内外尚无文献可查。传统的Adaboost算法对于均衡数据集合具有很好的分类精度，这在模式分类相关领域已经得到充分验证[3]。传统的AdaBoost算法开始时给每一个样本都赋予相同的权重，然后再用学习算法对训练集的样本集合进行迭代学习。如果将不平衡样本中的初始权值都设定成相同的，也就是说每一个样本都有相同的重要性，则分类曲线就会像样本点数少的一边移动。如何在不平衡样本数据下实现正确的故障诊断一直是学者关注的重点。

笔者通过将每一类样本赋予不同的权重，重新构造一种新的代价损失函数，通过决策规则的训练使代价损失函数最小化[4]，实现了滚动轴承系统样本不平衡下的故障诊断。试验采用美国凯斯西楚大学电气工程实验室轴承数据[5]，将本算法与其他算法进行比较。仿真结果表明，该算法在数据不均衡情况下的分类性能较传统方法有一定程度的提高。

1　传统Adaboost算法及描述

Adaboost是Boosting算法家族中的一种[6]，其基本思想是给定训练集合$(\boldsymbol{X}, y_i)$，其中，$\boldsymbol{X}=[x_1, x_2, \cdots, x_n]$，$y_i \in \{-1, +1\}$。首先，将每一组样本赋予相同的权重；然后，用该学习算法对训练样本训练T次，每次训练后，对训练失败的样本赋予较大的权重，相反则赋予较小的权重，从而得到1个预测函数序列$(h_1, h_2, \cdots, h_n)$，其中，每1个分类器对应1个权系数，预测结果好的其权系数就大；因此它又可以作为一种分类器的选择方法。具体算法如下：

1）输入训练样本集合$\boldsymbol{X}=[x_1, x_2, \cdots, x_n]$，$y_i \in \{-1, +1\}$，初始化样本权重$D_1(i)=\dfrac{1}{n}$

2）For $T=1, 2, \cdots, t$

在D_t下训练，得到预测函数（弱分类器）h_t

计算预测错误率

$$\mathrm{err}_t = \sum_{h_t(x_i) \neq y_i} D_t(i) \tag{1}$$

计算权系数

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\mathrm{err}_t}{\mathrm{err}_t}\right) \tag{2}$$

* 国家自然科学基金资助项目（51075330，50975231，61003137）
收稿日期：2012-04-10；修改稿收到日期：2012-05-31

根据错误率更新样本的权重

$$D_{t+1}(i)=\frac{1}{Z_t}D_t(i)\exp(-\alpha_t h_t y_i) \tag{3}$$

其中：$h_t\in\{-1,+1\}$为预测的结果；Z_t为正则化因子。

End

3) 最终分类器输出为

$$H(x)=\mathrm{sign}(\Sigma\alpha_t h_t) \tag{4}$$

2 CS-Boosting 算法

2.1 损失函数的建立

设计一个分类器$f(x)$，并希望该分类器对于正样本满足$I(y=1)yf(x)>0$，对于负样本满足$I(y=-1)yf(x)>0$，y为样本标签。该分类器对于整体样本集合应该满足

$$\min\{-I(y=1)yf(x)-I(y=-1)yf(x)\} \tag{5}$$

借助于指数函数的单调性，式(5)等价于

$$\min\{I(y=1)\mathrm{e}^{-yf(x)}+I(y=-1)\mathrm{e}^{-yf(x)}\}$$

由于训练集合存在不平衡性，因此建立一个系数约束条件，即

$$L=\begin{cases}0 & \text{if } y=f(x)\\ C_2 & \text{if } y=-1 \text{ and } f(x)=1\\ C_1 & \text{if } y=-1 \text{ and } f(x)=1\end{cases} \tag{6}$$

代价敏感Adaboost算法的代价函数可以写成

$$J(f)=E_{X,Y}\{I(y=1)\mathrm{e}^{-yC_1f(x)}+I(y=-1)\mathrm{e}^{-yC_2f(x)}\} \tag{7}$$

这里存在两个待定系数C_1，C_2。目前没有一个比较理想的数学表达式来估计这两个参数。定义表达式

$$C_1=C_2+(0.5-\mathrm{prior}(\mathrm{positive})) \tag{8}$$

通常情况下设定$C_2=1$，prior(positive)为正样本的先验概率，表示正样本空间占全部样本空间的比例。当正常样本数量与故障样本数量相同时，prior(positive)=0.5，此时$C_1=C_2$，表示正样本与负样本有相同的惩罚因子。

2.2 CS-Boosting 分类算法

给定训练样本$\{(x_i,y_i)\}_{i=1}^n$，并假设第m次迭代的分类结果是由最优步长α_m沿着损失函数的最速下降方向g_m构成。文献[7]令分类器$G(x)=\alpha g(x)$，则

$$\begin{aligned}J(F+\alpha g)&=E_{X,Y}[I(y=1)\exp(-C_1(f(x)+\alpha g))+I(y=-1)\exp(C_2(f(x)+\alpha g))]=\\&E_{X,Y}[I(y=1)\omega(x,1)\exp(-C_1\alpha g)+I(y=-1)\omega(x,-1)\exp(C_2\alpha g)]\end{aligned} \tag{9}$$

其中：$\omega(x,1)=\exp(-C_1(f(x)))$；$\omega(x,-1)=\exp(C_2(f(x)))$。

希望对于所有样本点x_i，该损失函数最小化，因此第m次迭代的最速下降方向g_m和最佳步长α_m为

$$\begin{aligned}(\alpha_m,g_m)=&\mathrm{argmin}E_{Y|X}[I(y=1)\omega(x,1)\times\exp(-C_1\alpha g)+I(y=-1)\omega(x,-1)\times\exp(C_2\alpha g)]=\\&\mathrm{argmin}_{\alpha,g(x)}[(\mathrm{e}^{C_1\alpha}-\mathrm{e}^{-C_1\alpha})b+\mathrm{e}^{-C_1\alpha}\Gamma_+ +(\mathrm{e}^{C_2\alpha}-\mathrm{e}^{-C_2\alpha})d+\mathrm{e}^{-C_2\alpha}\Gamma_-]\end{aligned} \tag{10}$$

其中

$$\begin{cases}\Gamma_+=\{i\,|\,y_i=1\}\\ \Gamma_-=\{i\,|\,y_i=-1\}\end{cases} \tag{11}$$

$$\begin{cases}b=\sum\limits_{i\in\Gamma_+}\omega_i^{(m)}[1-I(y_i=g(x_i))]\\ d=\sum\limits_{i\in\Gamma_-}\omega_i^{(m)}[1-I(y_i=g(x_i))]\end{cases} \tag{12}$$

$$\omega_i^{(m+1)}=\begin{cases}\omega_i^m\exp(-C_1\alpha_m g_m(x_i)),i\in\Gamma_+\\ \omega_i^m\exp(C_2\alpha_m g_m(x_i)),i\in\Gamma_-\end{cases} \tag{13}$$

令$\dfrac{\partial(\alpha_m,g_m(x))}{\partial\alpha}=0$，计算可得$\alpha$的方程为

$$2C_1b\cos(C_1\alpha)+2C_2d\cosh(C_2\alpha)=C_1\Gamma_+\mathrm{e}^{-C_1\alpha}+C_2\Gamma_-\mathrm{e}^{-C_2\alpha} \tag{14}$$

$$g_m=\mathrm{argmin}_g(\mathrm{e}^{C_1\alpha}-\mathrm{e}^{-C_1\alpha})\cdot b+\mathrm{e}^{-C_1\alpha}\Gamma_+ +(\mathrm{e}^{C_2\alpha}-\mathrm{e}^{-C_2\alpha})\cdot d+\mathrm{e}^{-C_2\alpha}\Gamma_- \tag{15}$$

具体的算法流程如下。

1) 输入：训练集合为$\{(x_i,y_i)\}^n$，$y=\{-1,+1\}$；代价因子为C_1，C_2；弱分类器类型；最大迭代次数为m。

2) 初始化：每类样本集合选择均匀分布

$$\omega_i=\frac{1}{|\Gamma_+|},\forall\, i\in\Gamma_+,\omega_i=\frac{1}{|\Gamma_-|},\forall\, i\in\Gamma_-$$

For $m=\{1,\cdots,M\}$

For $k=\{1,\cdots,K\}$

计算式(11)、式(12)，利用牛顿迭代法解式(14)，求α。

利用式(15)计算g_m

End for

选择最小损失的弱分类器$\{g_m,\alpha_m\}$，并根据式(13)更新权系数

End for

3) 输出：强分类器$H(x)=\mathrm{sgn}\left[\sum\limits_{m=1}^{M}\alpha_m g_m(x)\right]$

3　滚动轴承样本数据不均衡下的故障诊断

3.1　不均衡数据下的基于CS-Boosting 的故障检测模型

不均衡数据下的基于CS-Boosting 的故障识别流程如图 1 所示。本试验的目的是为了展示 CS-Boosting 算法对于处理不均衡训练样本有较强能力。试验数据来自美国 Case Western Reserve University 电气工程实验室。振动信号由安装在风扇端振动加速度传感器获取。轴承型号为SKF6203，故障是通过电火花加工的单点损伤，切割深度为 0.177 8 mm，采样频率为 12 kHz。本次试验模拟了滚动轴承在1.73 kr/min 时的4 种工作状态：正常工作状态、内圈故障状态、滚动体故障状态和外圈故障状态[8]。时域波形如图 2 所示。

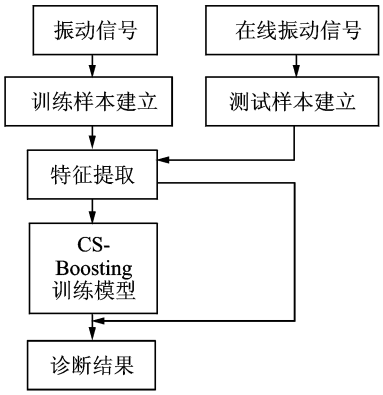


图 1　CS-Boosting 的故障诊断流程

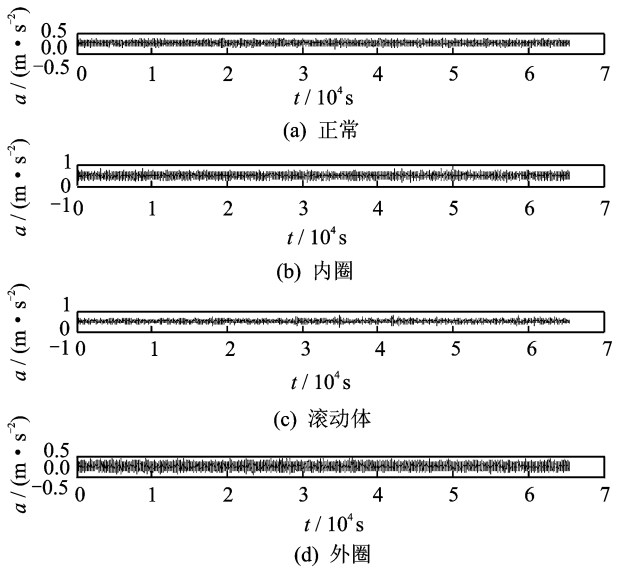


图 2　原始振动信号时域波形图

特征参数由时域特征参数（均值、中位数、方差、峭度、倾斜度、峰峰值、标准差、标准误差、最大梯度、最大值、最小值、总和）、频域峰值因子及时频域特征参数（小波包能量系数）组成。频域幅值因子表示在频域上的最大幅值与平均幅值之比

$$A = \frac{\max(\text{amplitude})}{\text{mean}(\text{amplitude})} \tag{16}$$

其中：max（amplitude）表示功率谱的最大幅值；mean(amplitude)表示功率谱的幅值的均值。

利用小波包能量系数作为频域特征参数，利用 db4 小波对轴承信号进行 3 层小波包分解，获取轴承振动信号在不同频带的能量，提取第 3 层从低频到高频的 8 个频率成分的信号特征。通过计算得到以下结论：轴承在振动时小波能量往往会集中在低频段，因此将第 3 层的前两个频带的小波能量成分作为故障特征参数[9]。

由于训练样本的特征量较多，且随着信号处理方法快速发展，训练样本的维数会越来越大，这会降低分类器的工作效率。为了增加分类器的工作效率，同时不对分类精度产生影响，选用局部保形映射（LPP）方法对这些特征量进行特征提取，将原始特征空间的维数降低到 2 维空间，结果如图 3 所示，详细算法见文献[10]。

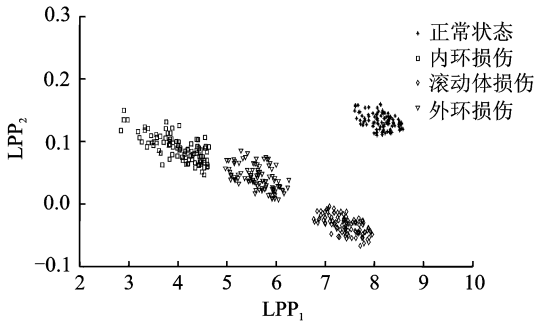


图 3　特征提取

3.2　试验结果及其评估准则

在故障诊断过程中，由于异常的故障样本难以收集，导致故障样本的数量远少于正常样本的数量。假设正常样本数量与故障样本数量之比为99∶1，当故障样本都被错分为正常样本时，得到的分类精度为99%，然而故障的漏检率为100%。因此，传统的性能评估准则已经不能够满足非均衡数据分布的情况。在此使用二分类方法，将训练集合分为少数类和多数类，并将少数类称为正类，多数类称为反类，构建一个二分类的混合矩阵，如表 1 所示。

表1 二分类的混合矩阵

实际样本数	预测为正类样本数量	预测为负类样本数量
实际正类样本数	TP	FN
实际负类样本数	FP	TN

分类器正确预测正类样本的比例TPR为

$$\mathrm{TPR}=\frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})} \tag{17}$$

分类器正确预测负类样本的比例TNR为

$$\mathrm{TNR}=\frac{\mathrm{TN}}{(\mathrm{TN}+\mathrm{FP})} \tag{18}$$

利用正、负样本比例的集合平均值GM作为评价指标，定义为

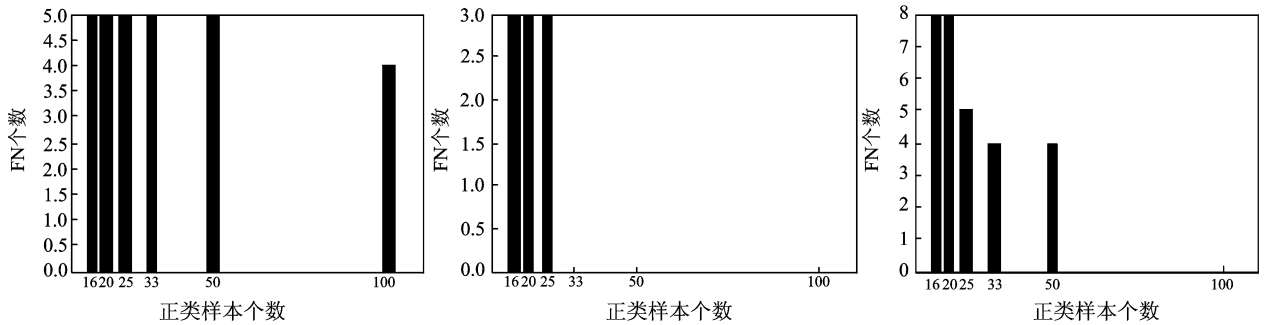
$$\mathrm{GM}=\sqrt{\mathrm{TPR}\cdot\mathrm{TNR}} \tag{19}$$

它综合考虑了两个类的分类性能。若分类器偏向于多数类一边，即若TPR=1，TNR=0，则GM=0，这说明不均衡样本数据对分类器精度的影响比较大；若GM值趋近1，说明该分类器对于不均衡样本集合有较强的处理能力。

3.3 不均衡数据的CS-Boosting分类器

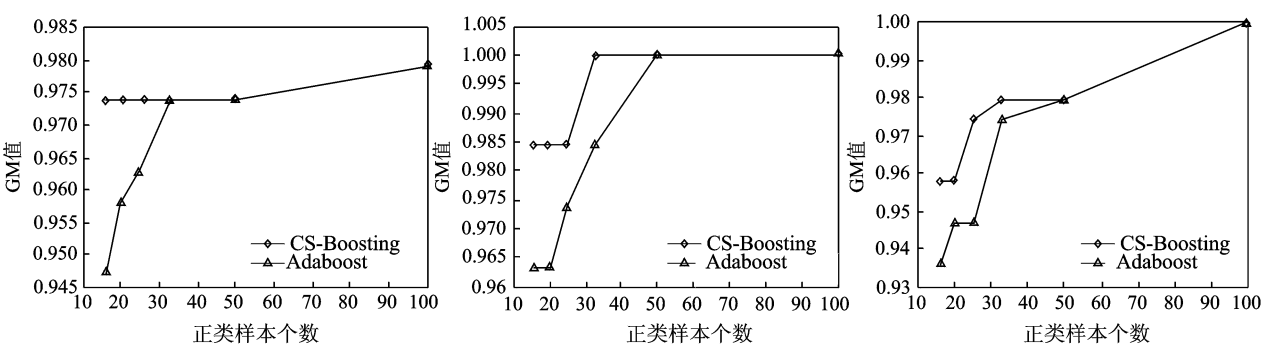
本试验采用197个正常状态样本，其中100个正常样本作为训练集合，内圈故障、滚动体故障和外圈故障的训练样本个数占正常样本个数的比例分别为6∶1，5∶1，4∶1，3∶1，2∶1，1∶1。测试样本分别为97个正常状态样本、97个外圈损伤样本、97个内圈损伤样本和97个滚动体损伤样本。试验采用CS-Boosting算法，以GM作为评价指标，Adaboost的最大迭代次数为20，选择CART作为弱分类器。图4为当正、负样本比例分别为1∶1，1∶2，1∶3，1∶4，1∶5，1∶6情况下FN的个数，即正类样本个数分别为100，50，33，25，20，16时FN的个数。从图4(a)可以看出，当两个样本集合均衡时，将内环损伤错分为正常状态的样本个数为4个。对于其他不均衡的样本状态，将内环损伤样本错分为正常状态样本的个数为5个。图4(b)为当训练样本的比例分别为1∶3，1∶2，1∶1时，滚动体损伤错分为正常状态的个数为0个，而当训练样本的比例为1∶6，1∶5，1∶4时，滚动体损伤错分为正常状态的个数为3个。图4(c)为训练样本均衡时外环损伤错分为正常状态的个数为0个，而当训练样本的比例为1∶3，1∶2时，外环损伤错分为正常状态的个数为4个，当训练样本比例为1∶4时，外环损伤错分为正常状态的个数为5个，训练样本的比例为1∶6，1∶5时，外环损伤错分为正常状态的个数为8个。

为了验证该算法的优越性，将该算法与传统的Boosting算法比较，采用GM作为评价指标，仿真结果如图5所示。



(a) 正常状态与内环损伤的分类结果 (b) 正常状态与滚动体损伤的分类结果 (c) 正常状态与外环损伤的分类结果

图4 正类样本预测为负类样本的个数



(a) 正常状态与内环损伤的分类结果 (b) 正常状态与滚动体损伤的分类结果 (c) 正常状态与外环损伤的分类结果

图5 不均衡数据二分类的几何平均值

可以看出，在正、负样本比例分别为1∶6，1∶5和1∶4时，CS-Boosting算法的GM大于传统Adaboost算法的GM值，即CS-Boosting算法在处理不平衡数据，特别是不平衡度较大时该算法优于传统的Adaboost算法。

4　结束语

提出了一种基于CS-Boosting分类算法，推导了算法中惩罚因子的代数表达式。针对训练样本不均衡的问题，对比试验表明，该算法的分类精度要优于传统的Adaboost算法，使轴承的损伤状态能够更准确地检测出来。

参　考　文　献

[1] Li C, Wu S. Online detection of localized defects in bearing by pattern recognition analysis [J]. ASME Journal of Engineering for Industry, 1989, 111(4): 331-336.

[2] 陶新民，刘福荣. 不均衡数据下基于SVM的故障检测新算法[J]. 振动与冲击，2010，29(12)：8-12.
Tao Xinmin, Liu Furong. Novel fault detection method based on SVM with unbalanced datasets[J]. Journal of Vibration and Sltock, 2010,29(12):8-12. (in Chinese)

[3] Mease D, Wyner A. Boosted classification trees and class probability [J]. Machine Learning Research, 2007, 8:409-439.

[4] Friedman J, Hastie T. Adaptive logistic regression: a statistical view of boosting [J]. The Annals of Statistics, 2000, 38:337-374.

[5] Bearing data center of case western reserve university [DB/OL]. [2011-02-05]. http://csegroups.case.edu/bearingdatacenter/home.

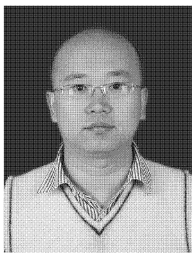
[6] Yoav F, Robert E. A decision-theoretic generalization of on-line learning and an application to boosting [J]. Journal of Computer and System Sciences, 1997, 55(1):119-139.

[7] Hamed M, Nuno V. Cost-sensitive boosting [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(2):294-309.

[8] Sugumaran V, Muralidharan V. Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing[J]. Mechanical Systems and Signal Processing, 2007,21:930-942.

[9] 苏文胜，王奉涛，朱泓，等. 基于小波包样本熵的滚动轴承故障特征提取[J]. 振动、测试与诊断，2011，31(2)：162-166.
Su Wensheng, Wang Fengtao, Zhu Hong, et al. Feature extraction of rolling element bearing fault using wavelet packet sample entropy[J]. Journal of Vibration, Measurement & Diagnosis, 2011, 31(2): 162-166. (in Chinese)

[10] He X, Niyogi P. Locality preserving projections[C]// 17th Annual Conference on Neural Information Processing Systems (NIPS). Canada, Vancouva: Neural Information Processing Systems Foundation, 2003, 16: 153-160.



第一作者简介：姚培，男，1982年11月生，博士。主要研究为故障诊断，模式识别。

E-mail：yaopei82@mail.nwpu.edu.cn