

旋转机械故障的拉普拉斯支持向量机诊断方法

郝腾飞 陈 果

南京航空航天大学,南京,211106

摘要:在旋转机械故障智能诊断中,收集大量的样本比较容易,而要对所有的样本进行类别标记却较为困难。针对这一问题,提出了一种基于拉普拉斯支持向量机的旋转机械故障智能诊断方法。滚动轴承故障诊断实例表明,有标记样本的数量较少时,与仅使用有标记样本进行学习的支持向量机相比,基于拉普拉斯支持向量机的诊断方法利用大量的无标记样本进行辅助学习,可以显著提高故障诊断的正确率。

关键词:故障诊断;滚动轴承;支持向量机;半监督学习;流形学习

中图分类号:TH17

DOI:10.3969/j.issn.1004-132X.2016.01.012

Fault Diagnosis of Rotating Machinery Based on Laplacian Support Vector Machines

Hao Tengfei Chen Guo

Nanjing University of Aeronautics and Astronautics, Nanjing, 211106

Abstract: In the intelligent fault diagnosis of rotating machinery, collecting a large number of data was relatively easy, but giving all collected data a label was often difficult. Aiming at this situation, an intelligent fault diagnosis approach for rotating machinery was proposed based on Laplacian support vector machines(LapSVM). The diagnosis example of rolling bearings shows that when the number of labeled data is limited, compared with the SVM that uses only labeled data for learning, the fault diagnosis approach based on LapSVM can improve the accuracy of fault diagnosis significantly by using a large amount of unlabeled data together with labeled data for learning.

Key words: fault diagnosis; rolling bearing; support vector machine(SVM); semi-supervised learning; manifold learning

0 引言

近年来,随着人工智能技术的发展,智能诊断成为旋转机械故障诊断研究的一个重要方向。旋转机械故障智能诊断在本质上是一个模式识别问题,支持向量机(support vector machines, SVM)作为一种推广性能优良的模式识别方法,目前已被广泛地应用于旋转机械故障智能诊断。文献[1]提出了一种基于SVM的航空发动机故障诊断方法。文献[2]研究了最小二乘SVM在滚动轴承故障诊断中的应用。文献[3]将集成学习方法AdaBoost和SVM相结合,利用多个单一SVM构造了SVM集成,并将其应用于滚动轴承故障诊断,获得了比单一SVM更高的故障诊断正确率。文献[4]与文献[5]对SVM在机械故障诊断中的应用进行了很好的综述。

以上研究充分展示了SVM在旋转机械故障智能诊断中的明显优势,然而SVM作为一种监督学习方法,要求所有的训练样本必须已知其类别标记。随着数据采集技术的发展,在旋转机械故障智能诊断中,收集大量的样本比较容易,而要

对所有的样本进行类别标记却较为困难,因为对样本进行类别标记通常需要花费一定的人力和物力才能完成。以滚动轴承故障诊断为例,采用相应的传感器可以获得大量的滚动轴承样本,但对一个样本进行类别标记则需要该领域的专家采用一定的设备对样本进行认真分析以后才能完成。在旋转机械故障智能诊断中,如果仅使用少量的有标记样本,那么利用它们所训练出的诊断模型很难获得较高的诊断正确率;另一方面,如果仅使用少量“昂贵的”有标记样本而不利用大量“廉价的”无标记样本,则是对数据资源的极大浪费。因此,在旋转机械故障智能诊断中,当有标记样本的数量较少时,研究如何利用大量无标记的样本来辅助提高故障诊断的正确率是一个非常有意义的问题。

半监督学习^[6-7]是一种可以同时利用有标记样本和无标记样本进行学习的方法,与仅利用有标记样本进行学习的监督学习方法相比,半监督学习方法可以利用无标记样本来辅助提高分类性能。流形学习^[8]是近年来发展起来的一种新型降维方法,其可从高维数据中恢复出数据内在的低维流形结构,从而实现维数约简,目前已有学者将

收稿日期:2014-06-05 修回日期:2015-07-30

基金项目:国家自然科学基金资助项目(61179057)

其应用于机械故障特征提取^[9-10]。拉普拉斯支持向量机^[11-12] (Laplacian support vector machines, LapSVM) 是一种将流形学习思想和支持向量机相结合的半监督分类方法, 它将无标记样本的内在流形结构信息融入到分类器的设计中, 在保持分类器光滑性的同时, 亦保持数据的流形结构, 从而可以有效地利用无标记样本来提高分类性能。

基于以上分析, 本文提出了一种利用 LapSVM 进行旋转机械故障智能诊断的方法。首先, 利用双螺旋仿真数据和双月形仿真数据直观地表明 LapSVM 的优势; 然后, 以滚动轴承故障诊断为例, 研究了基于 LapSVM 的旋转机械故障诊断方法的优越性。

1 拉普拉斯支持向量机

1.1 基本原理

LapSVM 是流形正则化^[11] 的一个具体实例, 而流形正则化是在传统正则化算法框架的基础上提出的一个半监督学习框架。下面首先介绍正则化算法和流形正则化的框架, 然后在此基础上介绍 LapSVM 的原理。

许多监督学习方法, 例如支持向量机、正则化最小二乘等, 均可解释为采用不同损失函数和正则化项的正则化算法。给定一组有标记的样本 (x_i, y_i) ($i=1, 2, \dots, l$) 以及核函数 K , 则传统正则化算法的框架可以表示为

$$f^* = \arg \min_{f \in \mathbf{H}_K} \frac{1}{l} \sum_{i=1}^l V(x_i, y_i, f) + \gamma \|f\|_K^2 \quad (1)$$

其中, $V(x_i, y_i, f)$ 为损失函数, 用以度量训练样本 x_i 的期望输出 y_i 和实际输出 $f(x_i)$ 之间的误差损失; $\|f\|_K^2$ 为再生核希尔伯特空间 \mathbf{H}_K 中函数 f 的某个特定范数, 用以度量函数的复杂性; γ 为正则化参数, 用以平衡经验损失和函数的复杂性。对于上述正则化算法的框架, 取损失函数为铰链损失函数就成为标准的支持向量机。

为利用生成数据的概率分布的几何结构来辅助有标记的样本进行学习, Belkin 等^[11] 在传统正则化算法框架的基础上, 通过引入一个额外的流形正则化项, 提出了流形正则化。流形正则化是一个基于流形假设的半监督学习框架, 其基本思想是, 假设数据分布在外围空间的一个子流形上, 然后利用大量无标记的数据估计出数据内在的流形结构, 进而通过一个流形正则化项将数据的流形结构信息融入到分类器的设计之中, 在保持分类器光滑性的同时, 亦保持数据的流形结构。给定 l 个有标记的样本 (x_i, y_i) 和 u 个无标记的样

本 x_j ($j=l+1, l+2, \dots, l+u$), 则流形正则化的框架可以表示为

$$f^* = \arg \min_{f \in \mathbf{H}_K} \frac{1}{l} \sum_{i=1}^l V(x_i, y_i, f) + \gamma_\Lambda \|f\|_K^2 + \gamma_l \|f\|_1^2 = \arg \min_{f \in \mathbf{H}_K} \frac{1}{l} \sum_{i=1}^l V(x_i, y_i, f) + \gamma_\Lambda \|f\|_K^2 + \frac{\gamma_l}{(u+l)^2} \mathbf{F}^T \mathbf{L} \mathbf{F} \quad (2)$$

$$\mathbf{F} = [f(x_1) \quad f(x_2) \quad \dots \quad f(x_{l+u})]^T$$

式中, $\|f\|_1^2$ 为流形正则化项; γ_l 为流形正则化参数; \mathbf{L} 为数据邻接图的拉普拉斯矩阵。

拉普拉斯支持向量机就是上述流形正则化的框架取特定损失函数时的一个具体例子。具体来说, 在上述流形正则化的框架中, 取损失函数为铰链损失函数, 即令 $V(x_i, y_i, f) = (1 - y_i f(x_i))_+ = \max(0, 1 - y_i f(x_i))$, $y_i \in \{-1, 1\}$, 则拉普拉斯支持向量机的学习问题可以表示为以下优化问题:

$$f^* = \arg \min_{f \in \mathbf{H}_K} \frac{1}{l} \sum_{i=1}^l (1 - y_i f(x_i))_+ + \gamma_\Lambda \|f\|_K^2 + \frac{\gamma_l}{(u+l)^2} \mathbf{F}^T \mathbf{L} \mathbf{F} \quad (3)$$

根据表示定理, 上述优化问题的解可以表示为

$$f^*(x) = \sum_{i=1}^{l+u} \alpha_i^* K(x, x_i) \quad (4)$$

令 $\boldsymbol{\alpha}^* = [\alpha_1 \quad \alpha_2 \quad \dots \quad \alpha_{l+u}]^T$, 则有

$$\boldsymbol{\alpha}^* = [2\gamma_\Lambda \mathbf{I} + 2 \frac{\gamma_l}{(u+l)^2} \mathbf{L} \mathbf{K}]^{-1} \mathbf{J}^T \mathbf{Y} \boldsymbol{\beta}^* \quad (5)$$

其中, \mathbf{I} 为 $(l+u) \times (l+u)$ 的单位矩阵; \mathbf{L} 为 $(l+u) \times (l+u)$ 的拉普拉斯矩阵; \mathbf{K} 为 $(l+u) \times (l+u)$ 的核矩阵; $\mathbf{J} = [\mathbf{I}_{l \times l} \quad \mathbf{0}]$ 为 $l \times (l+u)$ 的矩阵; $\mathbf{I}_{l \times l}$ 为 $l \times l$ 的单位矩阵; $\mathbf{Y} = \text{diag}(y_1, y_2, \dots, y_l)$; $\boldsymbol{\beta}^*$ 可以通过求解以下二次规划问题得到:

$$\boldsymbol{\beta}^* = \max_{\boldsymbol{\beta} \in \mathbf{K}^l} \sum_{i=1}^l \beta_i - \frac{1}{2} \boldsymbol{\beta}^T \mathbf{Q} \boldsymbol{\beta} \quad (6)$$

$$\text{s. t. } \sum_{i=1}^l \beta_i y_i = 0$$

$$0 \leq \beta_i \leq 1/l \quad i = 1, 2, \dots, l$$

$$\mathbf{Q} = \mathbf{Y} \mathbf{J} \mathbf{K} [2\gamma_\Lambda \mathbf{I} + 2 \frac{\gamma_l}{(l+u)^2} \mathbf{L} \mathbf{K}]^{-1} \mathbf{J}^T \mathbf{Y}$$

1.2 仿真实验

为直观地展示 LapSVM 的优越性, 本小节使用仿真数据研究 LapSVM 的特点。首先使用双螺旋仿真数据研究无标记的样本在 LapSVM 训练中的作用。双螺旋仿真数据如图 1 所示, 图中的正方形和圆形表示两类样本, 在每一类中随机选择一个样本赋予其类别标记 (在图中用实心表示), 剩下的样本作为无标记的样本 (在图中用空心表示)。实验中, SVM 仅使用两个有标记的样本进行训练, LapSVM 同时使用两个有标记的样

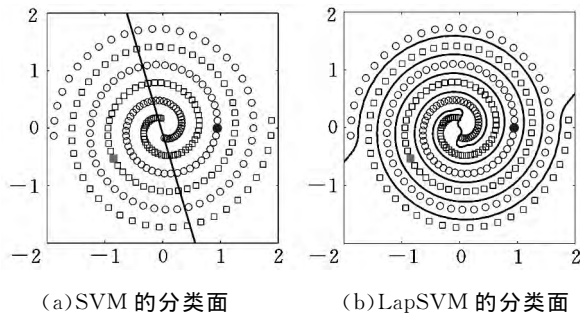


图 1 SVM 和 LapSVM 在双螺旋仿真数据上的分类面

本和剩下的无标记样本进行训练。从图 1a 可以看出,有标记的样本太少,无法反映出数据的真实分布,在这种情况下,SVM 训练得到的分类面对于两个有标记的样本来说虽然是最优的,但使用该分类面对原来的数据集进行分类将会造成很多的误分。从图 1b 可以看出,LapSVM 的分类面将两类样本完美地划分开来,这是由于 LapSVM 在训练中不仅利用了有标记样本的类别信息,同时考虑了无标记样本的流形结构信息,因此最终得到了理想的分类面。该仿真实验直观地表明了 LapSVM 可以有效地利用无标记的样本来辅助提高分类性能。

接下来使用双月形仿真数据研究流形正则化参数 γ_1 对 LapSVM 的影响。双月形仿真数据如图 2 所示,图中符号的含义与双螺旋仿真数据相同。使用图中仿真数据对 LapSVM 进行训练,图 2 所示为 γ_1 不同、其他参数保持不变的分

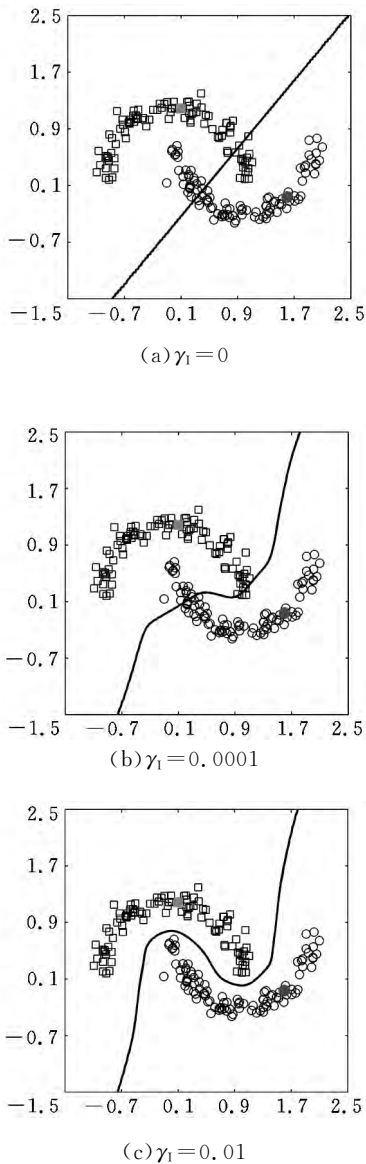


图 2 流形正则化参数 γ_1 对 LapSVM 的影响

式(4)可知, $\gamma_1=0$ 时,LapSVM 在训练中将忽略无标记样本的流形结构信息,此时 LapSVM 自动退化为普通的 SVM,因此其分类面正好就是两个有标记样本连线的垂直平分线,如图 2a 所示。 $\gamma_1=0.0001$ 时,LapSVM 训练得到的分类面开始根据双月形数据的分布发生变化,但仍有部分样本发生误分,如图 2b 所示。 $\gamma_1=0.01$ 时,LapSVM 训练得到的分类面进一步根据双月形数据的分布发生变化,将两类样本正确地划分开来,如图 2c 所示。从该仿真实验可以直观地看出,流形正则化参数 γ_1 可用于衡量训练时考虑数据的流形结构信息的程度,该参数越大,构造分类面时将越多地考虑数据的流形结构信息。

2 基于 LapSVM 的旋转机械故障诊断方法

上一节的仿真实验直观地表明,LapSVM 可以有效地利用无标记的样本来辅助提高分类性能。基于此,本文提出了一种基于 LapSVM 的旋转机械故障智能诊断方法,其流程如图 3 所示。该诊断方法包括模型训练和故障诊断两个阶段。在模型训练阶段,首先使用相应的传感器采集能够反映诊断对象状态的振动信号;然后采用某种信号处理方法提取所采集信号的特征,由故障诊断专家选择一些样本进行认真分析后确定其类别标记,并将这些样本作为有标记样本,同时将其余的样本作为无标记样本;最后将有标记样本和无标记样本一起输入 LapSVM 进行训练得到诊断模型。在故障诊断阶段,首先采用和模型训练阶段相同的特征提取方法形成诊断向量,然后利用诊断模型进行故障诊断,最后根据诊断结果决定是继续监控还是进行相应维修。

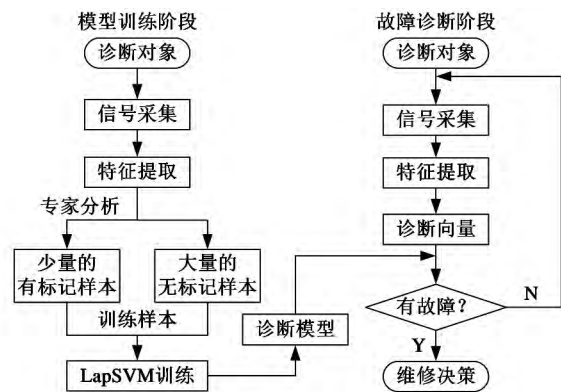


图 3 基于拉普拉斯支持向量机的旋转机械故障诊断流程

3 故障诊断实例

滚动轴承是旋转机械中的一种重要支承部

件,本节以滚动轴承故障诊断为例,研究基于 LapSVM 的旋转机械故障诊断方法的优越性。实验数据采用美国 Case Western Reserve University 电气工程实验室的滚动轴承试验台数据^[13]。在该试验台中,试验轴承支承电机转轴,在其内圈、外圈和滚动体上分别用电火花技术加工了单点损伤,以模拟内圈故障、外圈故障和滚动体故障。滚动轴承的振动信号通过安装在轴承座上方的加速度传感器进行采集,采样频率为 12 kHz。滚动轴承在不同状态下的典型时域波形如图 4 所示。

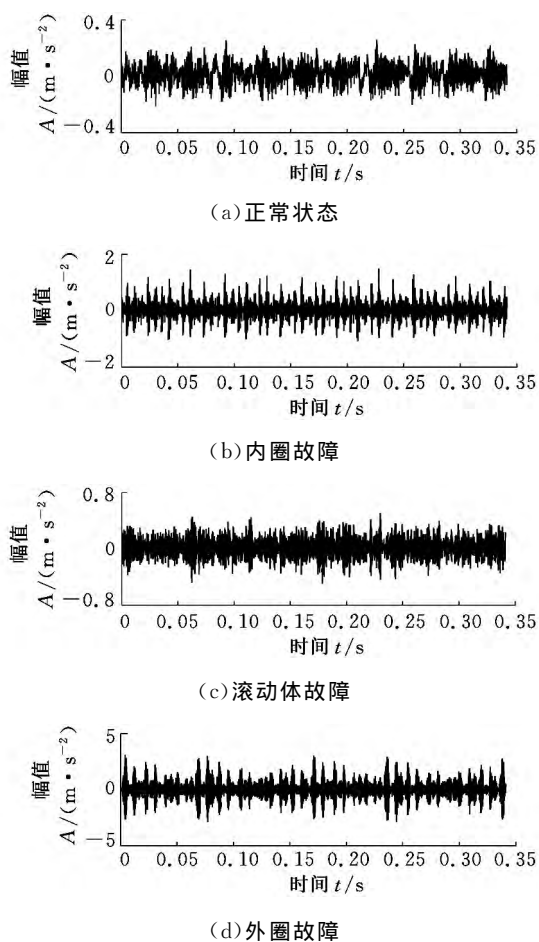


图 4 滚动轴承原始振动信号的时域波形

3.1 特征提取

特征提取是实现智能诊断的关键步骤之一。当滚动轴承发生局部损伤时,受载运行过程中产生的冲击会激起系统的高频固有振动,从而会导致不同频带的能量分布发生变化,因此可根据滚动轴承振动信号不同频带能量的分布来诊断滚动轴承的不同故障。为提取滚动轴承振动信号的能量特征,首先需要将滚动轴承的振动信号分解为一系列不同频率成分的信号。由 Huang 等^[14]提出的经验模式分解(empirical mode decomposition, EMD)方法可以将一个信号自适应地分解为

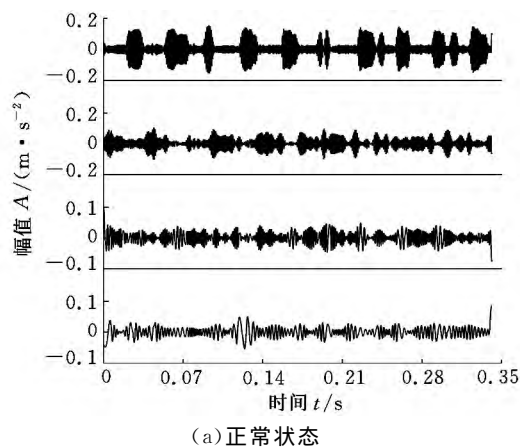
一系列本征模函数(intrinsic mode function, IMF)分量,这些分量分别包含了信号从高到低不同频率段的成分,因此计算每一个 IMF 分量的能量,然后将这些能量值组合起来就是滚动轴承振动信号在不同频率段的能量分布。

本文将 IMF 能量特征作为诊断输入的特征向量,其具体构造方法为^[15]:首先采用 EMD 将滚动轴承的原始振动信号分解为一系列 IMF 分量。由于滚动轴承的故障信息主要在高频带,因此选择前 4 个 IMF 分量进行计算。然后分别计算这 4 个 IMF 分量的能量,并对这些能量值进行归一化。最后将这 4 个归一化后的能量值组成特征向量,并作为接下来智能诊断实验的输入。

根据上述方法计算图 4 中 4 个信号的 IMF 能量特征,由 EMD 分解得到的前 4 个 IMF 分量如图 5 所示,计算得到的 IMF 能量值如图 6 所示。从图 6 可以看出,由于滚动轴承发生故障时会激起系统的高频固有振动,因此与正常状态相比,滚动轴承存在内圈故障、滚动体故障和外圈故障时高频成分的能量显著增加,低频成分的能量相应减少;而且不同故障类型引起的能量变化程度不同,其中,外圈故障的能量变化最为明显,接下来依次是滚动体故障和内圈故障。从该组滚动轴承振动信号的 IMF 能量特征可以看出,上述特征提取方法可有效地用于提取滚动轴承故障的特征。

3.2 智能诊断

为研究无标记的样本在智能诊断中的作用,同时采用 SVM 和 LapSVM 进行智能诊断实验。SVM 和 LapSVM 的核函数均选取高斯径向基核函数。SVM 和 LapSVM 本身都是针对两类分类问题设计的,而滚动轴承故障诊断是一个多类分



(a) 正常状态

类问题。目前,通过两类分类方法来解决多类分类问题的方法主要有一对一、一对多和有向无环图等方法,本文采用一对多方法^[16]来解决滚动轴

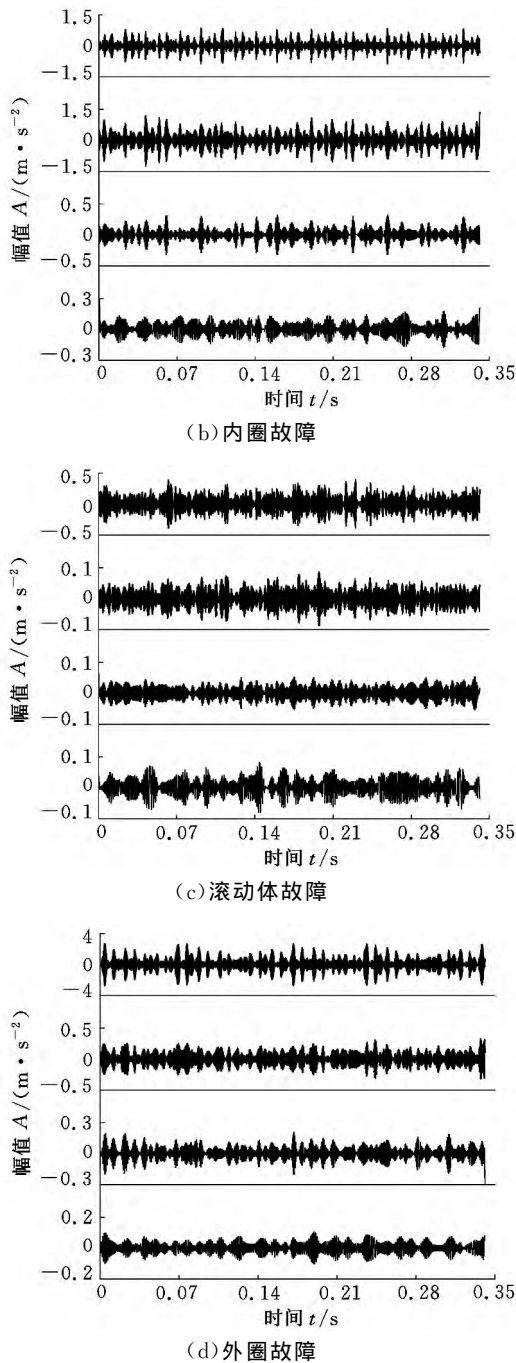


图 5 由 EMD 分解得到的前 4 个 IMF 分量 (由上到下分别为 IFM1~IFM2)

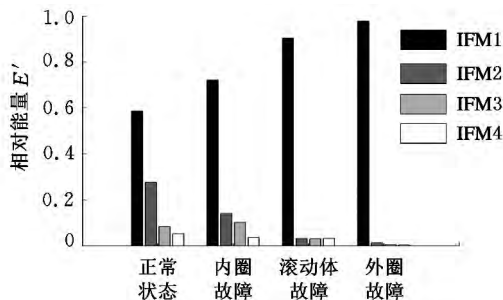


图 6 4 种状态下的 IMF 能量特征

承故障诊断问题。

实验中训练样本和测试样本的个数如表 1 所

示。在训练样本中,对于滚动轴承每一种状态的样本,随机选取 5% 的样本赋予其类别标记(作为有标记样本),剩下的作为无标记样本。SVM 仅使用少量的有标记样本进行训练,LapSVM 同时使用少量的有标记样本和大量的无标记样本进行训练。训练完成以后,将得到的模型对测试样本进行测试,结果如表 2 所示。从表 2 可以看出,LapSVM 的总诊断正确率比 SVM 提高了 6.5%,这表明在有标记样本较少时,LapSVM 可以有效地利用无标记样本来辅助提高诊断正确率。

表 1 滚动轴承故障诊断训练/测试样本个数

正常状态	内圈故障	滚动体故障	外圈故障
100	100	100	100

表 2 SVM 和 LapSVM 的故障诊断正确率 %

诊断方法	正常状态	内圈故障	滚动体故障	外圈故障	总诊断正确率
SVM	100	80	98	56	83.50
LapSVM	100	81	85	94	90.00

为进一步研究增加有标记样本的数量对 SVM 和 LapSVM 诊断正确率的影响,在每一种状态的训练样本中分别随机选取 10%、20%、30%、40%和 50% 的样本作为有标记样本进行故障诊断实验,结果如图 7 所示。从图 7 可以看出,随着有标记样本数量的增加,SVM 和 LapSVM 的诊断正确率均不断提高,但在每一种情况下,LapSVM 的诊断正确率仍然高于 SVM,这表明随着有标记样本数量的增加,LapSVM 仍然可以利用无标记的样本来进一步提高故障诊断的正确率。此外,从图 7 还可以看出,随着有标记样本数量的增加,LapSVM 的诊断正确率和 SVM 的诊断正确率的差距逐渐缩小。这是因为当有标记样本增加到一定数量时,已经基本上可以反映出数据的真实分布,这时无标记样本的作用逐渐减小,因此 SVM 的诊断结果逐渐接近 LapSVM 的诊断结果。

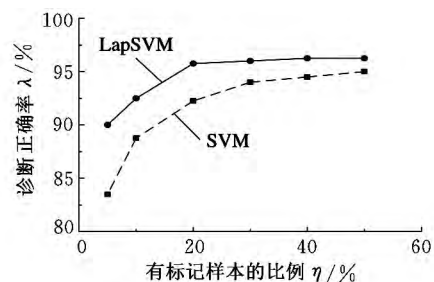


图 7 训练样本中有标记样本的比例对诊断正确率的影响

4 结语

本文提出了一种基于拉普拉斯支持向量机的

旋转机械故障智能诊断方法。该方法在智能诊断中可以有效地利用大量无标记的样本来辅助有标记的样本进行训练,即利用无标记的样本估计出数据的流形结构,在训练中不仅考虑有标记样本的类别信息,还考虑无标记样本的流形结构信息,从而在有标记样本数量较少的情况下,通过利用无标记样本进行辅助训练,获得比传统支持向量机更高的故障诊断正确率。滚动轴承故障诊断实例验证了所提方法的优越性。

参考文献:

- [1] 徐启华,师军. 基于支持向量机的航空发动机故障诊断[J]. 航空动力学报,2005,20(2):298-302.
Xu Qihua, Shi Jun. Aero-engine Fault Diagnosis Based on Support Vector Machine[J]. Journal of Aerospace Power,2005,20(2):298-302.
- [2] 王太勇,何慧龙,王国峰,等. 基于经验模式分解和最小二乘支持向量机的滚动轴承故障诊断[J]. 机械工程学报,2007,43(4):88-92.
Wang Taiyong, He Huilong, Wang Guofeng, et al. Rolling Bearings Fault Diagnosis Based on Empirical Mode Decomposition and Least Square Support Vector Machine[J]. Chinese Journal of Mechanical Engineering,2007,43(4):88-92.
- [3] Hu Qiao, He Zhengjia, Zhang Zhousuo, et al. Fault Diagnosis of Rotating Machinery Based on Improved Wavelet Package Transform and SVMs Ensemble [J]. Mechanical Systems and Signal Processing, 2007,21(2):688-705.
- [4] 袁胜发,褚福磊. 支持向量机及其在机械故障诊断中的应用[J]. 振动与冲击,2007,26(11):29-35.
Yuan Shengfa, Chu Fulei. Support Vector Machines and Its Application in Machine Fault Diagnosis[J]. Journal of Vibration and Shock,2007,26(11):29-35.
- [5] Widodo A, Yang B S. Support Vector Machine in Machine Condition Monitoring and Fault Diagnosis [J]. Mechanical Systems and Signal Processing, 2007,21(6):2560-2574.
- [6] Chapelle O, Schölkopf B, Zien A. Semi-supervised Learning[M]. Cambridge, MA, USA: MIT Press, 2006.
- [7] Zhu Xiaojin, Goldberg A B. Introduction to Semi-supervised Learning [M]. San Rafael: Morgan & Claypool Publishers,2009.
- [8] Zhang Zhenyue, Wang Jing, Zha Hongyuan. Adaptive Manifold Learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,34(2):253-265.
- [9] 蒋全胜,贾民平,胡建中,等. 基于拉普拉斯特征映射的故障模式识别方法[J]. 系统仿真学报,2008,20(20):5710-5713.
Jiang Quansheng, Jia Minping, Hu Jianzhong, et al. Method of Fault Pattern Recognition Based on Laplacian Eigenmaps[J]. Journal of System Simulation,2008,20(20):5710-5713.
- [10] Li Benwei, Zhang Yun. Supervised Locally Linear Embedding Projection (SLLEP) for Machinery Fault Diagnosis[J]. Mechanical Systems and Signal Processing,2011,25(8):3125-3134.
- [11] Belkin M, Niyogi P, Sindhvani V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples[J]. Journal of Machine Learning Research,2006,7(11):2399-2434.
- [12] Melacci S, Belkin M. Laplacian Support Vector Machines Trained in the Primal[J]. Journal of Machine Learning Research,2011,12(3):1149-1184.
- [13] The Case Western Reserve University Bearing Data Center. Bearing Data Center Fault Test Data[EB/OL]. [2012-03-01]. <http://www.eecs.cwru.edu/laboratory/bearing/>.
- [14] Huang N E, Shen Z, Long S R, et al. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis[J]. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences,1998,454(1971):903-995.
- [15] Yang Yu, Yu Dejie, Cheng Junsheng. A Roller Bearing Fault Diagnosis Method Based on EMD Energy Entropy and ANN[J]. Journal of Sound and Vibration, 2006,294(1/2):269-277.
- [16] Rifkin R, Klautau A. In Defense of One-vs-All Classification[J]. Journal of Machine Learning Research,2004,5(1):101-141.

(编辑 张 洋)

作者简介:郝腾飞,男,1983年生。南京航空航天大学民航学院博士研究生。主要研究方向为航空发动机状态监测与故障诊断。发表论文4篇。陈果,男,1972年生。南京航空航天大学民航学院教授、博士研究生导师。