

一种半监督 SVDD-KFCM 算法及其在轴承故障检测中的应用

李军利¹ 李巍华²

(珠海城市职业技术学院机电工程学院 珠海 519090)¹

(华南理工大学机械与汽车工程学院 广州 510641)²

摘 要 对机械设备故障诊断过程中故障样本较难提取和运行转速、载荷多变导致诊断方法的适用性不强、准确性不高等问题进行分析,结合支持向量数据描述(Support Vector Data Description, SVDD)算法与模糊核聚类(Kernel-based Fuzzy c-Means, KFCM)算法,提出一种基于半监督学习的 SVDD-KFCM(Semi-supervised SVDD-KFCM, SS-KFCM)方法用于轴承故障检测。实验表明,在复杂多载荷工况下该算法可有效检测轴承故障,诊断准确率较高。

关键词 SVDD, KFCM, 故障检测, 半监督学习

中图法分类号 TP206+.3 文献标识码 A

Semi-supervised SVDD-KFCM Algorithm and its Application in Bearing Fault Detection

LI Jun-li¹ LI Wei-hua²

(School of Mechanical and Electrical Engineering, Zhuhai City Polytechnic, Zhuhai 519090, China)¹

(School of Mechanical and Automotive Engineering, South China University of Technology, Guangzhou 510641, China)²

Abstract Machinery is always running under multiple operating regimes, and it is difficult to collect the specific fault samples to train the learning machine, which also leads to the low accuracy in fault detection and limits the generalization of the intelligent fault detection methods. Combining the support vector data description and the kernel-based fuzzy C-means clustering, a semi-supervised SVDD-KFCM algorithm for machine defect detection was proposed. Experiment results demonstrate that the proposed scheme is capable of detecting the incipient bearing fault effectively and correctly.

Keywords SVDD, KFCM, Fault detection, Semi-supervised learning

1 引言

模式识别诊断方法的研究大多在特定载荷下训练故障诊断模型,并识别设备状态,具有较大局限性^[1]。而在工程实践中,许多机械设备经常在不同载荷状况下切换工作,不仅故障样本较难获得,而且机械设备的工作条件复杂,很难保证转速、载荷等在运行过程中保持恒定,这就使得单一载荷下的诊断模型准确性不高,诊断方法适用性不强。

针对这种状况,张道强等提出一种半监督的模糊核聚类算法,通过运用已知和未知标签样本协同训练模式来提高性能^[2]。毕锦烟等将模糊核聚类算法运用于齿轮箱早期故障检测^[3]。Zhang 等添加已知和未知样本的分类错误率作为目标函数,并结合优化的核参数获得聚类的全局最优解^[4]。Xue 等利用部分标签样本,提出一种结合粗糙集的半监督模糊聚类异常检测方法^[5]。Luo 等建立了一种基于模糊集的支持向量数据描述(Support Vector Data Description, SVDD)算法^[6]。

支持向量数据描述^[7]是通过寻找支持向量来绘制数据的最小超球边界,落在边界内部的点即为目标样本点,位于边界以外的点就判为非目标样本点。聚类分析利用聚类算法使相

似度高的样本尽量聚集在一起,相似度低的样本尽量被区分开来。模糊 C 均值聚类算法是通过判断样本至聚类中心的隶属度实现分类,但该算法对初始聚类中心比较敏感^[8]。

为克服传统聚类方法初始聚类中心选择的盲目性和利用未知样本的学习性能提高检测效率,本文提出了一种将 SVDD 与模糊核聚类相结合的方法,并通过实验验证了本方法的有效性。

2 支持向量描述算法

为了描述一个给定的目标数据集 $\{x_i, i=1, 2, \dots, N\}$, 可以找出一个超球体,使得这个超球以最小的半径包含所有的目标数据。落在超球面上的数据点就称为支撑向量。为了减少偏离部分数据可能使描述的球体过大及增加异类点落入球内的机会,引入松弛变量 ξ_i , 则 SVDD 的目标函数为:

$$\begin{aligned} \text{Min } \ell(R^2, a, \xi) &= R^2 + P \sum_{i=1}^N \xi_i \\ \text{s. t. } (x_i - a)^T (x_i - a) &\leq R^2 + \xi_i, \forall i, \xi_i \geq 0 \end{aligned} \quad (1)$$

式中, R 和 a 分别为最小超球的半径和球心, P 为惩罚参数,表示对异常样本的惩罚程度。

式(1)是一个二次优化问题,结合约束条件构造拉格朗日

李军利(1970—),男,博士,副教授,主要研究方向为故障诊断、机电控制, E-mail: ljltiyut@163.com; 李巍华(1973—),男,博士,教授,主要研究方向为故障诊断等。

函数,得该优化问题的对偶形式:

$$\begin{aligned} \text{Max } L &= \sum_{i=1}^l a_i (x_i, x_i) - \sum_{i=1}^l \sum_{j=1}^l a_i a_j (x_i, x_j) \\ \text{s. t. } \sum_{i=1}^l a_i &= 1, 0 \leq a_i \leq c, i=1, \dots, l \end{aligned} \quad (2)$$

引入核函数 $K(x, y)$, 将样本集映射至高维特征空间, 使非线性问题转化为线性问题。当核函数满足 Mercer 条件时, 它就对应于某一变换空间的内积^[9]: $K(x, y) = \varphi(x)^T \varphi(y)$, φ 为核映射, 且核函数满足: $K(x, y) = \exp\{-\|x - y\|^2 / \sigma_2^2\}$ 。该优化问题基于核的形式如下:

$$L = \sum_{i=1}^l a_i K(x_i, x_i) - \sum_{i=1}^l \sum_{j=1}^l a_i a_j K(x_i, x_j) \quad (3)$$

根据 Kuhn-Tucker 条件, 可求解以上优化问题得到 a_i 值; 当 $0 < a_i < P$ 时, 对应的样本点落在超球边界上, 即为该数据集的支持向量点; 当 $a_i = 0$ 时, 对应的样本点落在超球内部; 当 $a_i = P$ 时, 对应的样本点位于球体外部。球心的表达式为:

$$a = \sum_{i=1}^l a_i \varphi(x_i) \quad (4)$$

计算球心至任意一个支持向量点的距离就能求得球体的半径 R 。对应新样本 Z , 如果满足:

$$(\varphi(z) - a)^T (\varphi(z) - a) \leq R^2 \quad (5)$$

则 Z 属于目标样本, 否则为非目标样本。

3 模糊核聚类算法

模糊核聚类算法 (Kernel-based Fuzzy c-Means, KFCM) 算法^[11]描述为: 设一个数据集 $X = \{x_i, i=1, 2, \dots, n\}$, 每个样本含有 M 维特征向量。对其进行聚类, 假设 c 为预定的聚类数目, v_i 为输入空间的聚类中心, u_{ik} 表示第 k 个样本对第 i 类的隶属度, 且 $0 \leq u_{ik} \leq 1$ 。定义 KFCM 目标函数, 求解最优化问题:

$$\text{Min } J_m(u, v) = \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m \|\phi(x_k) - \phi(v_i)\|^2 \quad (6)$$

$$\text{s. t. } \sum_{i=1}^c u_{ik} = 1, k=1, 2, \dots, n$$

式中, m 为隶属度指数, 一般取 $1 \sim 3$; ϕ 为核映射。选取高斯核函数 $K(x, y) = \exp\{-\|x - y\|^2 / \sigma_2^2\}$, 代入目标函数式 (6), 展开化简可得:

$$\begin{aligned} J_m(u, v) &= \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m \{K(x_k, x_k) + K(v_i, v_i) - 2K(x_k, v_i)\} \\ &= 2 \sum_{i=1}^c \sum_{k=1}^n u_{ik}^m \{1 - K(x_k, v_i)\} \end{aligned} \quad (7)$$

在约束条件下, 对隶属度和聚类中心进行迭代优化:

$$u_{ik} = \frac{(1/(1 - K(x_k, v_i)))^{2/(m-1)}}{\sum_{j=1}^c (1/(1 - K(x_k, v_j)))} \quad (8)$$

$$v_i = \frac{\sum_{k=1}^n u_{ik}^m K(x_k, v_i) x_k}{\sum_{k=1}^n u_{ik}^m K(x_k, v_i)} \quad (9)$$

KFCM 算法步骤: (1) 确定初始聚类数目 c , 隶属度参数 m ; (2) 初始化各个聚类中心 v_i ; (3) 重复以下运算, 使得各个样本的隶属度达到稳定: 1) 用当前的聚类中心根据式 (8) 更新隶属度; 2) 用当前的聚类中心和隶属度根据式 (9) 更新各个聚类中心。

4 基于半监督学习的 SVDD-KFCM 故障检测方法

在机械设备故障诊断过程中, 通过将正常工作状况下的样本作为训练集来建立基于正常样本的 SVDD 数据描述模型, 实现对未知故障样本的检测。但在正常样本数较少或者代表的工况单一的情况下, 训练得出的模型适用性不强, 会将其他工况下的正常样本误判为故障, 引起误报, 造成不必要的停机和检修。为了避免这种损失, 本文提出一种基于半监督学习的 SVDD-KFCM 故障检测方法, 通过 SVDD 利用已知的正常样本建立超球边界, 再对未知样本进行判断。若未知样本含有潜在故障样本, 则可得到正常和潜在故障两类样本; 若未知样本均落在支持向量所形成的边界内部, 则说明未知样本均为正常样本, 此时可继续选择未知样本, 直到获得落在边界外的潜在故障样本, 进行下一步故障检测。选择正常和潜在故障样本在输入空间的中心作为 KFCM 的初始聚类中心, 再在测试过程中直接利用预训练得到的中心对样本进行聚类。算法流程如图 1 所示。

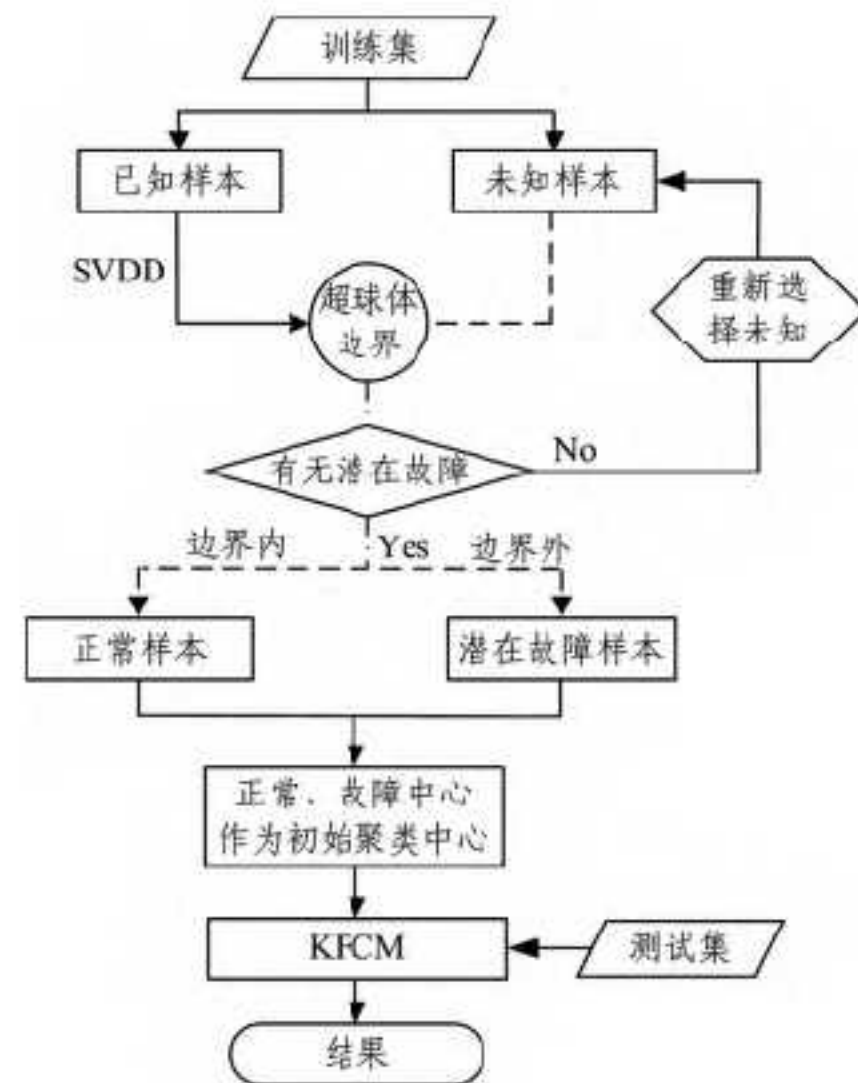


图 1 SS-KFCM 算法流程图

5 实验研究

本文实验数据来自美国凯斯西储大学电子电气工程实验室^[12]。实验台如图 2 所示, 包括一个 0.746kw 电动机、一个扭矩传感器、译码器和一个测功计。驱动端轴承型号为 SKF6205, 风扇端轴承型号为 SKF6203。采用的电火花加工轴承单点故障分别为: 0.1778mm、0.3556mm、0.5334mm、1.016mm。加速度传感器采集振动信号, 扭矩传感器和译码器测量功率和转速。



图 2 滚动轴承故障模拟实验台

实验模拟轴承分别在正常、内环故障、外环故障、滚动体

故障 4 种状态下的运行状况。由于篇幅所限,以驱动端轴承为例,采样频率为 12kHz,对正常状态轴承和故障程度最小(0.1778mm)的内环故障轴承、外环故障轴承及滚动体故障轴承的振动信号进行分析。

载荷为 0kw 时,轴转速为 1797rpm,经计算,轴承内环通过频率为 162.19Hz,外环通过频率为 107.36Hz,滚动体通过频率为 70.58Hz。载荷为 0kw 时,正常与各类故障信号的时域波形、包络解调谱,分别如图 3、图 4 所示。

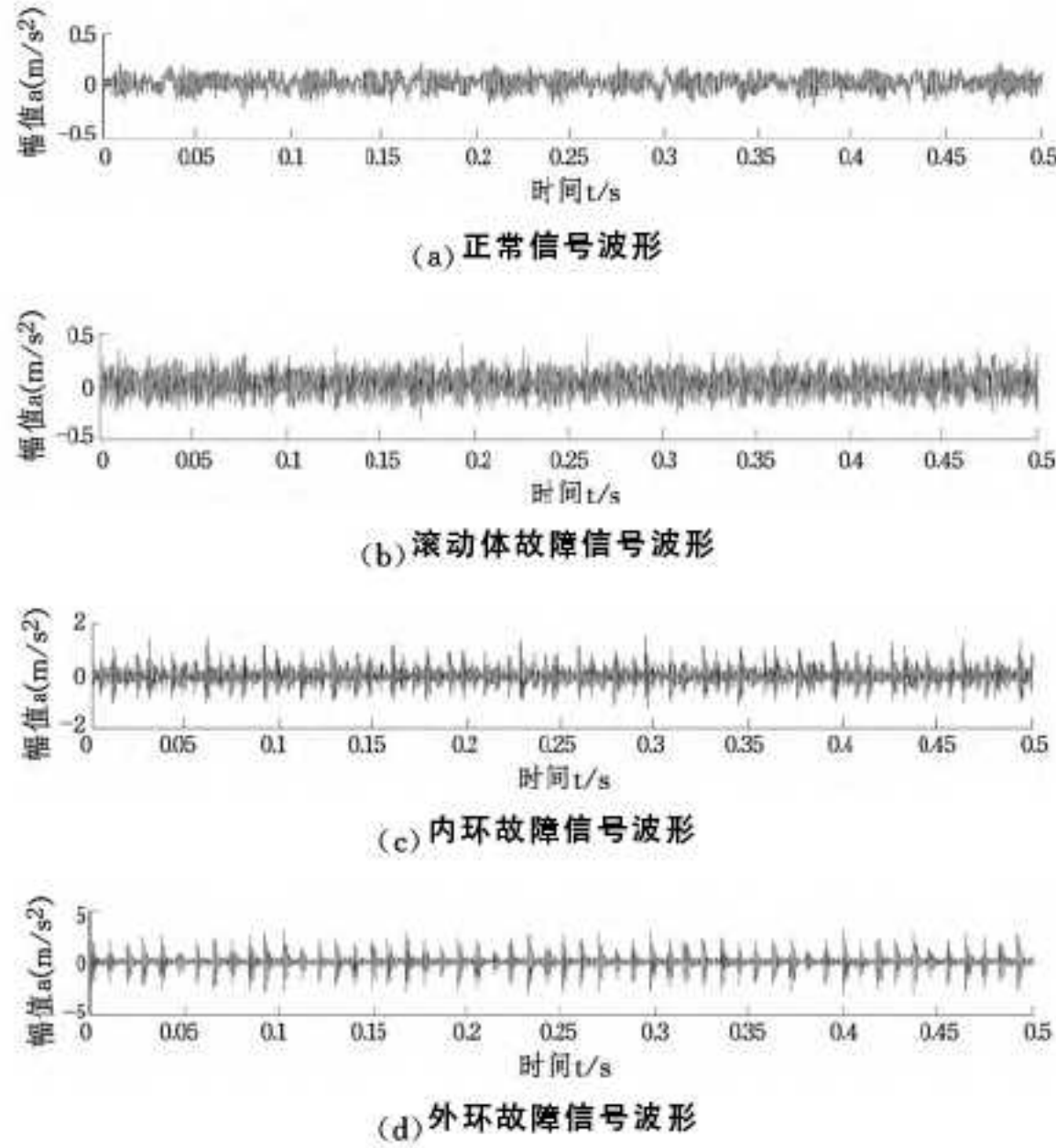


图 3 载荷为 0kw 时轴承正常、滚动体故障、内环故障、外环故障信号的时域波形

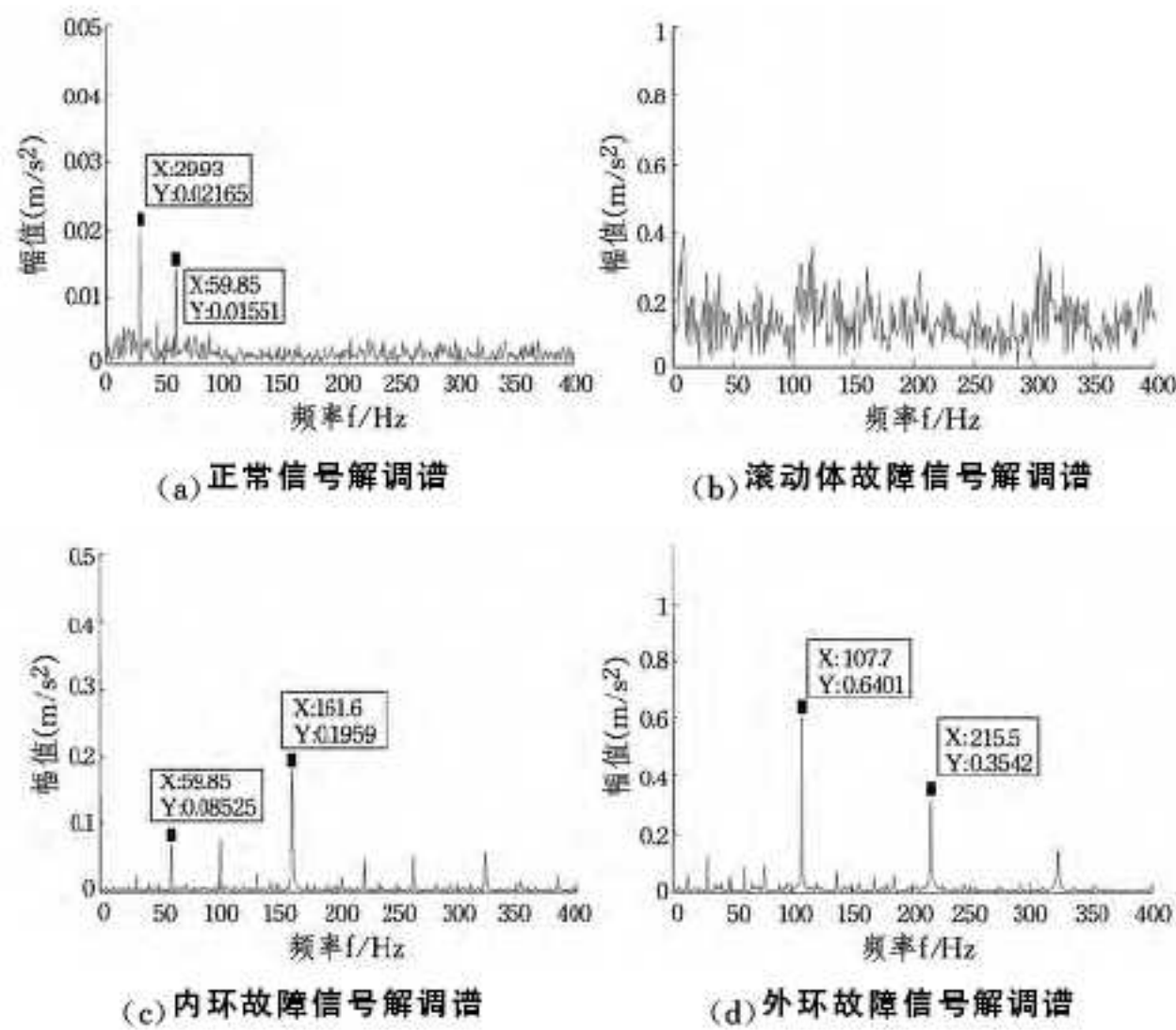


图 4 载荷为 0kw 时轴承正常、滚动体故障、内环故障、外环故障信号的包络谱

SS-KFCM 分类器中待确定的参数有:(1)惩罚参数 P ;(2)SVDD 核参数 σ_1 ;(3)隶属度指数 m ;(4)KFCM 核参数 σ_2 。

设定 $P=1, m=2$ 。对于 σ_1 的选择,考虑到超球边界随着 σ_1 的增大由紧致变得松弛,趋于球状,当训练样本点数比较少、分布散度大时, σ_1 可选择较大的值;当训练样本点数多、分布集中时, σ_1 可选择较小的值。

为验证算法的有效性,以滚动体轻微故障检测为例,在已知样本数较少且多载荷工况下,分析 σ_1 对支持向量所形成边界的影响。选取已知 4 种载荷下共 40 个正常样本作为训练集对滚动体故障进行检测,测试样本含正常和故障样本各 20 个。将含有 14 维特征的数据利用主成分分析降至二维, σ_1 取值变化时的检测结果如图 5 所示,错分点数如图 6 所示。

由图 5 可得,当 σ_1 取值过小时,正常状态域被分为多个互不连通的封闭区域,部分正常样本被错判为故障; σ_1 取值过大时,虽然仍可以得到紧致的正常域边界,但却造成某些故障样本被错判为正常的情况。当 $\sigma_1=4$ 时,此时错分样本数为 0,边界松弛适中,能较好表达数据分布。

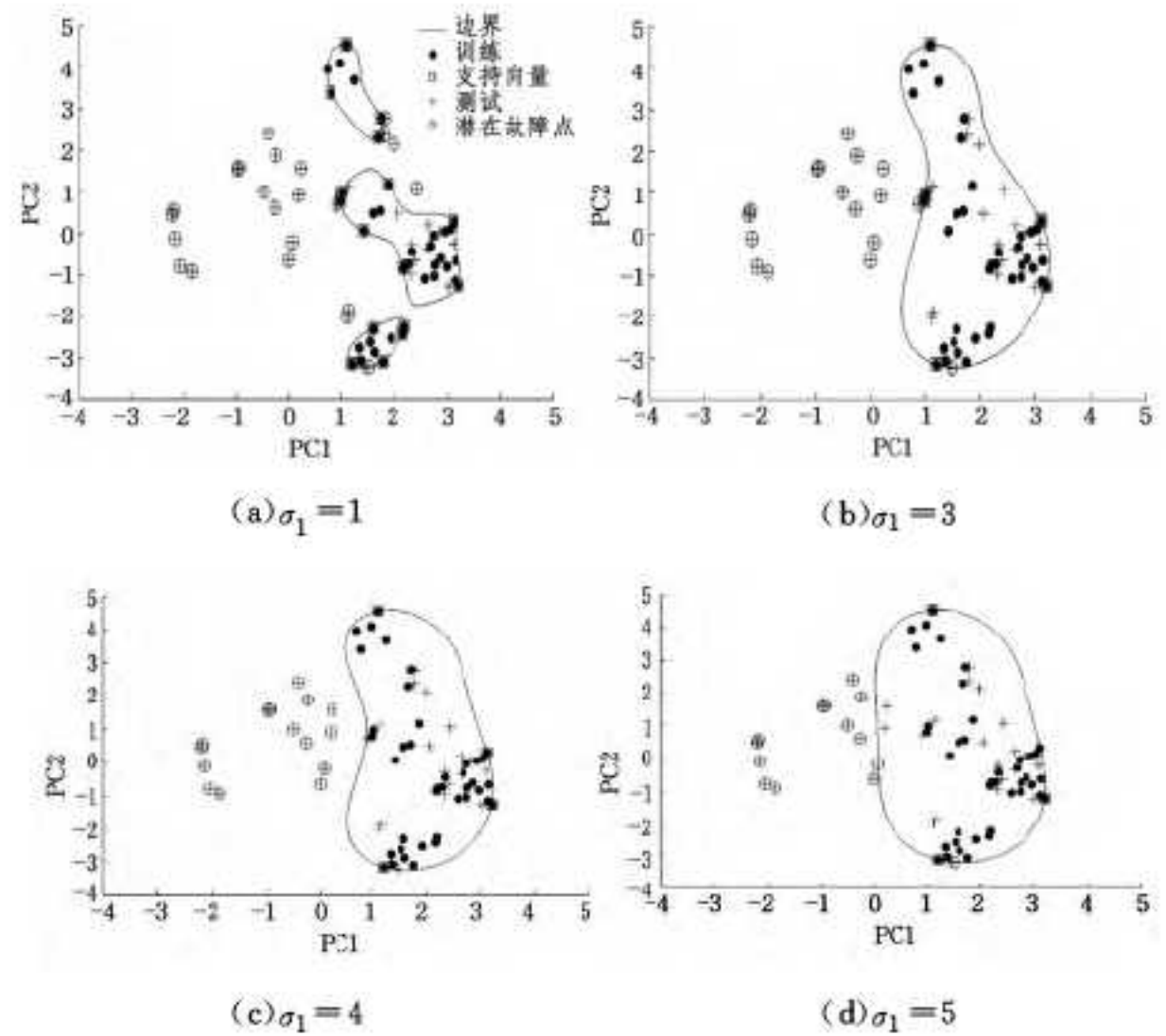


图 5 不同 σ_1 对描述边界的影响

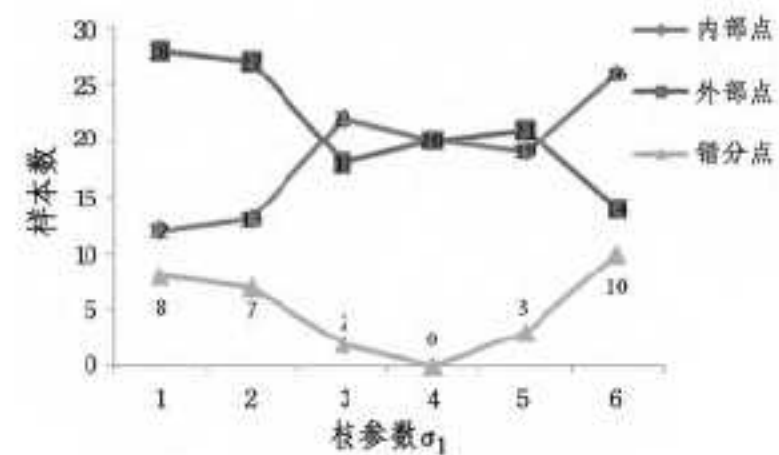


图 6 不同 σ_1 下边界内外点及错分点数

由图 6 可得,随着 σ_1 的增大,内部点增多,错分点数先减少后增大。在 σ_1 取值为 4 时,错分样本个数为 0;当其取值增大时,错分样本个数增多。随着边界的扩大,部分故障样本也被错判为正常,从而验证了图 5 所得的结果。

为进一步验证算法的有效性,在拟定参数 $\sigma_1=4, \sigma_2=4.5$ 的基础上,选择不同的训练样本数,对比监督式 SVDD、无监督 KFCM、半监督 SS-KFCM 3 种分类器的检测性能。利用 3 种分类器分 20 次对内环故障、外环故障、滚动体故障进行检测,统计错分点数,以平均准确率作为检测性能的评价指标。考虑在训练样本不充分情况下的检测效果,训练集所含样本由 16 依次增加到 48,训练集 16(4+12)表示训练集共 16 个样本,由 4 个正常样本与 12 个未知样本组成,其余同理。测试集所含样本个数不变,由 50 个正常和 50 个故障样本组成,检测结果如表 1 所列。

表1 检测结果对比

类型	训练样本数 (已知+未知)	平均准确率		
		SVDD	KFCM	SS-KFCM
内环故障检测	16(4+12)	96.3%	54.5%	100%
	32(12+20)	97.9%		100%
	48(16+32)	98.8%		100%
外环故障检测	16(4+12)	95.6%	50%	100%
	32(12+20)	96.8%		100%
	48(16+32)	97%		100%
滚动体故障检测	16(4+12)	90.7%	43.5%	91.1%
	32(12+20)	93.3%		93.5%
	48(16+32)	94.6%		95.5%

由表1可知,SVDD对于内环故障和外环故障的检测能力均在95%以上,在相同训练样本条件下对滚动体故障的检测能力比前两类故障的检测能力低4%左右;KFCM检测时不需样本进行训练,直接对数据集进行聚类,聚类结果具有盲目性,检测的准确率较低;SS-KFCM在检测内环故障和外环故障时的准确率高达100%,即使在小样本条件下(训练样本数为16),其检测能力也明显优于SVDD,对于较难检测的滚动体故障,虽然准确率有所下降,但仍略优于SVDD。

图7分别展示了表1中SS-KFCM分类器对正常样本和滚动体故障样本的训练与测试结果,图中以“X”表示错分样本。

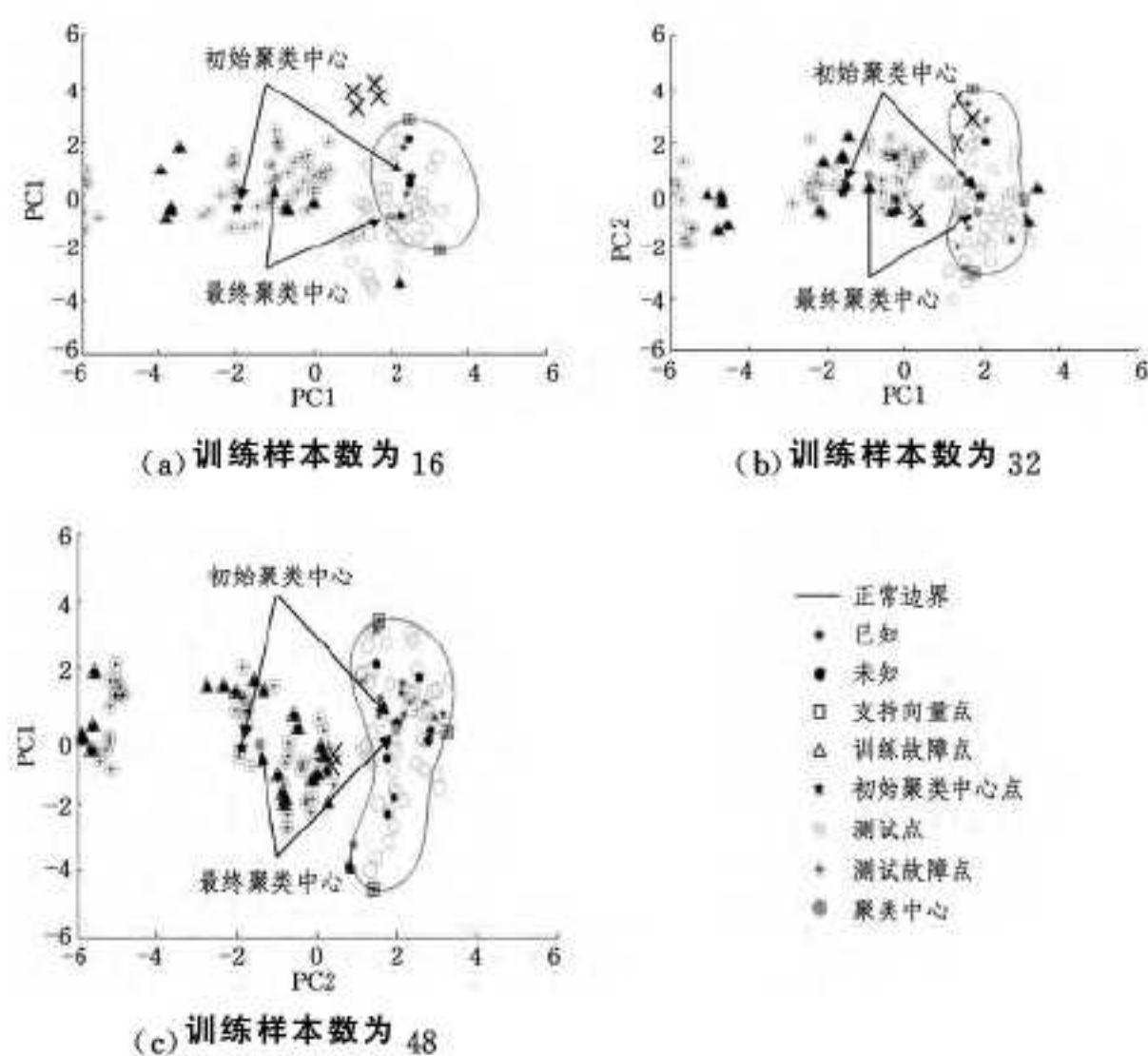


图7 训练样本数分别为16、32、48时的检测结果

由图7(a)知,在训练样本数为16时,训练获得的支持向量点较少,所形成的边界范围小,有4个正常样本错判为故障;测试获得正常样本数据较故障数据分布集中,聚类效果明显,这是因为正常情况下机械处于稳定运行状态。由图7可得,随着训练样本数的增加,经SVDD获得的初始聚类中心越接近最终聚类结果的两类样本中心,所描述的正常域边界越紧致,算法检测性能明显提高。由表1可得,在大的训练样本条件下,SS-KFCM的检测能力越接近SVDD。

综合可知,在核参数一定时,SS-KFCM的检测性能均有所提升,监督式SVDD方法的准确率与训练样本数相关,在一定范围内,训练样本数越多,错分点越少,准确率越高;无监督式KFCM方法由于初始类别中心是随机确定的,没有已知样本进行指导分类,因此导致分类效果较差;半监督式SS-KFCM算法在少量训练样本时即可达到很高的检测准确率,

且随着训练样本的增多,初始聚类中心越接近最终的聚类中心,降低了聚类过程的迭代次数,提高了计算效率。所以,半监督式SS-KFCM算法既可避免训练样本不足导致的监督式分类器无法良好训练问题,又克服了无监督时聚类的盲目性。

结束语 针对机械设备故障诊断中运行载荷多变难以获取故障样本进行训练的问题,提出了一种半监督的SVDD-KFCM故障检测算法。以多载荷工况下轴承故障实验为例,通过主成分分析对信号提取的特征实现降维,运用SS-KFCM算法分别对轴承内环故障、外环故障、滚动体故障进行检测,并将检测效果与监督式SVDD算法、无监督KFCM算法对比,验证了算法在训练样本较少的情况下的有效性和准确性。该算法只需少量训练样本,克服了无监督模糊核聚类算法初始聚类中心随机确定导致分类盲目性的不足,较大程度提高了故障检测的准确率。

参考文献

- [1] 王鑫,于洪亮,张琳,等.采用改进遗传神经网络的多载荷振动信号故障诊断[J].噪声与振动控制,2011,31(4):137-141
- [2] Zhang Dao-qiang, Tan Ke-ren, CHEN Songcan. Semi-supervised Kernel-Based Fuzzy C-Means[C]// Pal N R. et al. eds., ICONIP 3316. Berlin Heidelberg: Springer-Verlag, 2004, 1229-1234
- [3] 毕锦烟,李巍华.基于半监督模糊核聚类的齿轮箱离群检测方法[J].机械工程学报,2009,45(10):48-52
- [4] Zhang Hua-xiang, Lu Jing. Semi-supervised fuzzy clustering: A kernel-based approach[J]. Knowledge-Based Systems, 2009, 22(6):477-481
- [5] Xue Zhen-xia, Shang You-lin, Feng Ai-fen. Semi-supervised outlier detection based on fuzzy rough C-means clustering[J]. Mathematics and Computers in Simulation, 2010, 80(9): 1911-1921
- [6] Luo Hui, Cui Jiang, Wang You-ren. A SVDD approach of fuzzy classification for analog circuit fault diagnosis with FWT as pre-processor[J]. Expert Systems with Applications, 2011, 38(8): 10554-10561
- [7] Tax D M J, Duni R P W. Support vector domain description[J]. Pattern Recognition Letters, 1999, 20(11/13): 1191-1199
- [8] James C B, Robert E, William F. FCM: The fuzzy c-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2/3): 911-203
- [9] Cao L J, Chua K S, Chong W K. A Comparison of PCA, KPCA and ICA for Dimensionality Reduction in Support Vector machine [J]. Neurocomputing, 2003, 55(1/2): 321-336
- [10] Macqueen J. Some methods for classification and analysis of multivariate observations[C]// LeCam L M, Neyman J. eds., Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume I, Statistics. Berkeley: University of California Press, 1967: 281-297
- [11] 张道强.基于核的联想记忆及聚类算法的研究与应用[D].南京,南京航空航天大学,2004
- [12] Loparo K A. Bearings vibration data set[DB/OL]. 2009. <http://csegroups.case.edu/bearingdatacenter/pages/welcome-case-western-reserve-university-bearing-data-center-website>