

DOI:10.19651/j.cnki.emt.2107999

基于MDP-SVM的过程多类型故障诊断*

郭小萍 尹瑞琛 李元

(沈阳化工大学信息工程学院 沈阳 110142)

摘要: 针对工业过程多类型故障诊断率低的问题,提出一种边界判别投影(MDP)与支持向量机(SVM)相融合(MDP-SVM)的方法。边界判别投影常用于人脸识别领域,其可以将多类数据降维,获得不同类别清晰的边界。与主成分分析(PCA)和局部线性嵌入(LLE)算法相比,考虑了样本的局部结构和全局结构,避免了小样本问题。降维的数据通过SVM分类器进行类别判断,利用粒子群(PSO)算法得到最佳SVM分类器,实现故障诊断。仿真结果表明,相对于传统方法,所提方法故障识别准确率达到95.379%,而且可同时识别出多类故障。

关键词: 边界判别投影;支持向量机;多类型故障诊断

中图分类号: TQ015 **文献标识码:** A **国家标准学科分类代码:** 520.2040

Process multi-type fault diagnosis based on MDP-SVM

Guo Xiaoping Yin Ruichen Li Yuan

(School of Information Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China)

Abstract: To solve the problem of low diagnosis rate of multi-type faults in industrial processes, a method of boundary discriminant projection (MDP) and support vector machine (SVM) fusion (MDP-SVM) was proposed. Boundary discriminant projection is often used in the field of face recognition, which can reduce the dimensionality of multiple types of data to obtain clear boundaries of different categories. Compared with principal component analysis (PCA) and local linear embedding (LLE), the local and global structures of samples are considered and the problem of small samples is avoided. The classification of dimensionality reduction data is judged by SVM classifier, and the optimal SVM classifier is obtained by particle swarm optimization (PSO) algorithm to achieve fault diagnosis. The simulation results show that compared with the traditional method, the fault identification accuracy of the proposed method can reach 95.379%, and multiple faults can be identified simultaneously.

Keywords: margin discriminant projection; support vector machine; multi-type fault diagnosis

0 引言

工业生产过程日趋先进复杂,对于设备的安全性和可靠性要求也越来越高^[1]。为了尽可能地反映过程信息,通常安装了各种各样的传感器,因此可以收集到大量的反映过程信息的数据,同时存储了大量发生故障时的数据。数据维数高是工业过程数据的一大特征,目前通过降维方法进行特征提取是常用的方式,对故障类型数据进行降维,在保持数据的原有结构特征的基础上提高了对故障类型识别速度,同时也可以提高分类器的分类精度^[2]。

基于分类思想的故障诊断方法通常将生产过程中发生不同的故障或者进行不同的操作都可以视作不同的工况,检测这种故障可以视作多分类的问题。常用的方法有贝叶

斯网络(Bayesian network, BN)算法、K近邻(K-nearest neighbours, KNN)算法、支持向量机(support vector machine, SVM)等。其中,贝叶斯算法需要知道先验概率并且假定各变量间相互独立,KNN在样本不平衡时分类效率低而且计算量较大^[3]。而SVM可以解决非线性问题和小样本问题,同时因其泛化性能良好和精度高的优点所以被广泛使用^[4]。SVM模型的两个参数—惩罚参数 c 和径向基核函数参数 g 对模型分类效果有决定性的影响,目前,SVM在故障诊断的应用中参数寻优方法采用网格搜索法(grid search, GS)、遗传算法(genetic algorithm, GA)、粒子群(particle swarm optimization, PSO)算法等^[5-9]。GS本质上是一种枚举法,比较耗时^[10]。GA和PSO都属于仿生算法,但GA是无记忆的^[6]。PSO的粒子具有记忆特性,

收稿日期:2021-10-01

* 基金项目:国家自然科学基金项目(61490701,61673279)、辽宁省教育厅项目(LJ2020021)资助

信息共享等特性,收敛速度比 GA 快,所以 PSO 更适合 SVM 参数寻优^[7]。由于建模数据维数高导致 SVM 输入变量较多,目前常采用的方式是通过降维方法提取特征作为新的输入。文献[11]提出 PCA-SVM 故障诊断方法,首先使用主成分分析(principle component analysis, PCA)对参数进行优化和选择,利用降维思想进行变量的选择得到综合参数特征值,将综合参特征值输入到 SVM 分类器模型中对故障进行定位,提高了故障分类效率。文献[2]等将线性判别分析(linear discriminant analysis, LDA)对原始数据进行预处理,降低样本维数,使用低维故障样本作为 SVM 的输入,从而建立了故障诊断模型。文献 [12]通过局部线性嵌入(locally linear embedding, LLE)对数据进行降维,再通过 SVM 分类器实现了手部动作的识别,LLE 是一种流形学习算法,通过对样本局部特征的映射来还原样本的整体结构。但是由于只利用了样本数据的局部结构,在样本数据较少时不能很好的反映样本全局结构。文献[13]提出有监督的边界判别投影(margin discriminant projection,MDP)对数据进行降维,实现了对人脸数据的识别。文献[14]通过利用类样本均值与类内距离和类间距离定义了类内边界和类间边界,结合 K 近邻分类算法获得了具有更好信息的边界判别表示,从而实现了人脸识别。边界判别投影降低数据维度的同时保持了数据流形的几何形状,与传统基于距离的降维算法相比同时保持了样本的局部特性和全局特性,但目前未查到该方法在故障诊断领域的应用。

本文提出一种基于边界判别投影与支持向量机融合的故障诊断方法。通过边界判别投影方法进行降维,采用累计贡献率确定投影维数,将样本数据映射到低维空间,然后将降维后的数据采用 SVM 算法进行故障特征提取。采用 PSO 算法结合 K-CV 进行 SVM 模型惩罚系数 c 和核函数参数 g 的优化,增强模型分类有效性以及提高诊断准确度。最后通过田纳西-伊斯曼(Tennessce-Eastman, TE)过程验证所提方法的有效性。

1 基础理论

1.1 边界判别投影

边界判别投影的核心在于通过降低同类样本边界距离和增大异类样本边界距离来实现类内距离最小化和类间距离最大化,进而提高分类准确率^[15]。经过边界判别投影的变换后,类内样本更加紧凑,类间距离变大。如此使得不同类别样本判别边缘清晰^[14]。

设 x_i 与 x_j 为数据中任意的两个样本,它们之间的距离定义为:

$$d(x_i, x_j) = \|x_i - x_j\|_2 \quad (1)$$

设 x_p^i 为第 i 类样本的第 p 个数据, x_q^j 为第 j 类样本的第 q 个数据,同类样本之间的距离定义为:

$$d(y_i) = \|x_p^i - x_q^i\|_2 \quad (2)$$

设 x_m^i 为第 i 类样本的第 m 个数据, x_n^j 为第 j 类样本的第 n 个数据,类间距离定义为:

$$d(y_i, y_j) = \|x_m^i - x_n^j\|_2 \quad (3)$$

设投影矩阵为 V ,为保持样本的几何结构,引入正交约束 $V^T V = I$ 。设 δ 为投影后的距离,定义投影后的类间距离和类内距离如式(4)、(5)所示。

$$\delta(y_i, y_j) = \|V^T x_m^i - V^T x_n^j\|_2 \quad (4)$$

$$\delta(y_i) = \|V^T x_p^i - V^T x_q^i\|_2 \quad (5)$$

实现类内距离最小和类间距离最大即实现了在投影空间中不同类别样本之间边界最大,通过式(4)和(5)可得到边界判别投影的目标函数为式(6),求解即可得到投影矩阵 V ,进而使得样本在投影后判别更为清晰,有利于分类^[13]。目标函数如式(6)所示。

$$\max_{V^T V = I} \sum_{i \neq j} \delta(y_i, y_j) - \sum_{i=1} \delta(y_i) \quad (6)$$

为将 MDP 算法的目标函数更为简洁的表示出来,将其归纳到图嵌入框架下^[13],引入类间相似性权重 W_{ij}^B 和类内相似性权重参数 W_{ij}^W ,如式(7)和(8)所示。

$$W_{ij}^B = \begin{cases} 1, & \text{样本点为 } x_m^i, x_n^j \\ 0, & \text{其他} \end{cases} \quad (7)$$

$$W_{ij}^W = \begin{cases} 1, & \text{样本点为 } x_p^i, x_q^i \\ 0, & \text{其他} \end{cases} \quad (8)$$

式(6)中目标函数就可以表达为:

$$\begin{aligned} & \max_{V^T V = I} \sum_{i \neq j} \delta(y_i, y_j) - \sum_{i=1} \delta(y_i) \\ & \max_{V^T V = I} \sum_{i \neq j} \|V^T x_m^i - V^T x_n^j\|_2 W_{ij}^B - \\ & \sum_{i=1}^m \|V^T x_p^i - V^T x_q^i\|_2 W_{ij}^W = \\ & \max_{V^T V = I} \text{tr}(V^T (X L^B X^T - X L^W X^T) V) \end{aligned} \quad (9)$$

其中 L 为 Laplacian 矩阵,计算公式如式(10)和(11)所示。

$$L^B = D^B - W_{ij}^B \quad (10)$$

$$L^W = D^W - W_{ij}^W \quad (11)$$

这里 D 为对角矩阵,且 $D_{ii} = \sum_{j=1}^n W_{ij}$ 。

MDP 方法的实现过程如算法 1 所示。

算法 1 MDP 算法

输入:数据 X 及对应的类别标签。

输出:低维投影 Y 。

- 1)计算 Laplacian 矩阵 L ;
 - 2)对矩阵 X 进行不完全 Cholesky 分解,得到 $X = QR$;
 - 3)对 $Z = RLR^T$ 作特征分解,获得特征值及对应的特征向量。计算累积贡献率,将对应的特征向量组成矩阵 U ;
 - 4)计算投影矩阵 $V = QU$;
 - 5)计算得到低维投影 $Y = V^T X$ 。
-

1.2 SVM 原理

支持向量机(SVM)是 1995 年以统计学习理论为基础

提出的一种机器学习算法^[16]。神经网络算法学习能力强,但是泛化能力差,结果容易出现过拟合现象。SVM基于结构风险最小化理论和VC维理论,在保持较强学习能力的同时也保持了泛化能力^[2]。因此,SVM算法在故障诊断领域得到了广泛的应用。

1) 基本原理

SVM分类的基本思想是:借助核函数将样本数据从低维空间映射到高维空间,非线性转化为线性,在高维线性空间中构造出最优超平面来得到分类器的判别函数。

给定数据 $D\{(x_i, y_i)\}, y_i \in \{-1, +1\}$, 设 w 为法向量, b 为位移项, $\varphi(x)$ 为数据在高维空间的映射, 则 SVM 的超平面用式(12)来描述。

$$w \cdot \varphi(x) + b = 0 \quad (12)$$

其判别函数如式(13)所示。

$$y(x) = \text{sgn}(w \cdot \varphi(x) + b) \quad (13)$$

求解最优超平面如式(14)所示。

$$\begin{cases} \min(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i) \\ \text{s. t. } y_i(w \cdot \varphi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0 \\ i = 1, 2, \dots, m \end{cases} \quad (14)$$

其中, C 为惩罚因子, ξ_i 为松弛变量, 利用 Lagrange 乘数法和 KKT(Karush-Kuhn-Tucker) 条件得到这个二次规划问题的对偶问题如式(15)所示。

$$\begin{cases} \max L = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j (\varphi(x_i) \cdot \varphi(x_j)) = \\ \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{s. t. } \sum_{i=1}^m y_i \alpha_i = 0 \\ 0 \leq \alpha_i \leq C, i = 1, 2, \dots, m \end{cases} \quad (15)$$

最优超平面表示如式(16)所示。

$$g(x) = \sum_{i=1}^m \hat{\alpha}_i y_i (\varphi(x_i) \cdot \varphi(x)) + \hat{b} = \sum_{i=1}^m \hat{\alpha}_i y_i K(x_i, x) + \hat{b} \quad (16)$$

其中, 位移项 \hat{b} 计算公式如式(17)所示。

$$\hat{b} = y_j - (w \cdot \varphi(x_j)) = y_j - \sum_{i=1}^m \hat{\alpha}_i y_i (\varphi(x_i) \cdot \varphi(x_j)) = y_j - \sum_{i=1}^m \hat{\alpha}_i y_i K(x_i, x_j) \quad (17)$$

2) 参数更新

支持向量机分类器惩罚参数 c 决定分类器的泛化能力, 径向基核函数参数 g 决定了样本在特征空间中的分布, 这些参数决定了支持向量机性能的好坏^[17-18]。使用粒子群参数寻优方法找到最优参数, 同时为了避免样本数据集在划分数据时的影响, 采用粒子群优化算法与 K 折交叉验证结合的方法。

粒子群(PSO)算法是1995年提出的一种群体智能算法, 其原理是根据个体和群体的最优位置来更正粒子速度和位置来寻找全局最优解。假设在 n 维搜索空间中种群大小为 m 的粒子群进行粒子群寻优, 设粒子移动速度为 v_{id} , 位置为 x_{id} , 其中个体最优位置为 $pbest$, 群体最优位置为 $gbest$ 。粒子移动速度 v_{id} 和位置 x_{id} 如式(18)、(19)所示。

$$v_{id}(t+1) = \omega v_{id}(t) + c_1 r_1 (pbest_{id} - x_{id}) + c_2 r_2 (gbest_{id} - x_{id}) \quad (18)$$

$$x_{id}(t+1) = x_{id}(t) + H_0 (1 - \frac{t}{t_{max}}) v_{id}(t+1) \quad (19)$$

其中, x_{id} 是粒子当前的位置; t 是迭代次数; ω 是惯性权重系数; 常量 c_1 和 c_2 代表学习因子; r_1 和 r_2 是位于 $[0, 1]$ 均匀分布的随机数; H_0 表示控制量。在设定好粒子的初始值以后便可以通过式(18)和(19)迭代找到粒子的最佳位置。

K 交叉验证(K-fold cross validation, K-CV)是一种可以充分利用训练样本对分类算法进行测试的方法^[19]。将数据随机的分为 K 组子集, 每次利用其中一组子集最为测试集, 剩下的 $K-1$ 组子集作为训练集进行训练^[2]。这样可以平均性能得到最佳参数。最后将 K 组测试结果中平均误差最小的参数返回。

2 MDP-SVM 多类型故障诊断方法

本文所提故障诊断方法分为建模和检测两大部分, 流程如图1所示。建模主要步骤如下所示。

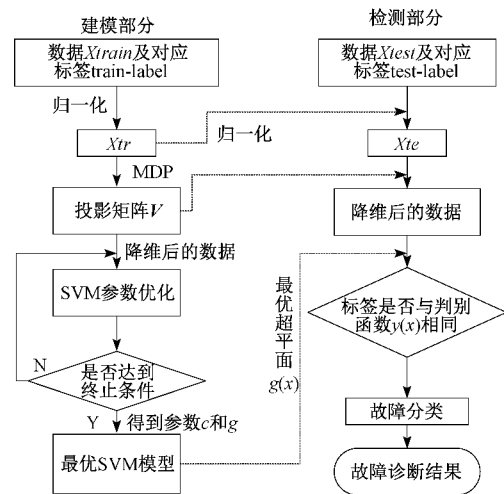


图1 MDP-SVM 多类型故障诊断方法流程

1) 将带有标签的样本数据 X_{train} 及对应标签作为输入数据。

2) 将数据集进行归一化处理为 X_{tr} 。

3) 通过 MDP 计算投影矩阵 V , 其中投影维数采用累计贡献率的方式确定。

4) SVM 模型的训练, 其中 SVM 参数 c 和 g 通过 PSO 和 K 折交叉验证来寻优选取。

检测主要步骤如下所示。

- 1) 测试数据归一化得 X_{tle} 。
- 2) 采用投影矩阵 V 对进行降维。
- 3) 利用 SVM 模型计算最优分类超平面 $g(x)$ 与判别函数 $y(x)$ 。
- 4) 判断样本数据是否在超平面同侧, 实现多故障分类。

3 TE 过程仿真实验

TE 过程是一个实际化工过程的仿真数据集, 主要由反应器、产品冷凝器、气液分离塔、离心压缩机和汽提塔 5 个部分组成^[2]。TE 过程可以仿真 21 种不同的故障, 其中前 160 个观测值为正常数据, 从 161 采样时刻引入故障。21 种故障类型中, 故障 1~7 为阶跃故障; 故障 8~12 为随机变化型干扰; 故障 13 为缓慢漂移干扰; 故障 14 和 15 为堵塞型干扰; 故障 16~21 为未知型干扰^[20]。

3.1 样本的选取

为符合实际生产情况避免单一理想情况, 本文选用 TE 过程正常工况、故障 1、5、7、12 和 17 作为样本数据来进行算法有效性验证实验。样本运行模式如表 1 所示。训练集中, 正常工况无故障共 960 组数据, 故障运行的仿真周期为 48 h, 故障在 8 h 引入, 间隔为 3 min, 共采集 960 组数据。其训练数据大小是 $4\ 800 \times 52$, 其中 1~800 为正常样本, 801~1 600 是故障 1, 1 601~2 400 是故障 5, 2 401~3 200 是故障 7, 3 201~4 000 是故障 12, 4 001~4 800 是故障 17; 测试集中, 正常工况仿真周期为 25 h, 共 500 组数据, 故障运行的仿真周期为 24 h, 共 480 组数据。测试数据大小是 $2\ 900 \times 52$, 其中 1~500 是正常数据, 501~980 是故障 1, 981~1 460 是故障 5, 1 461~1 940 是故障 7, 1 941~2 420 是故障 12, 2 421~2 900 是故障 17。

表 1 样本运行模式

运行模式	故障描述	类型
正常工况	—	—
故障 1	A/C 进料流量比变化, 组分 B 含量保持不变	阶跃
故障 5	冷凝器冷却水入口温度发生变化	阶跃
故障 7	物料 C 压力损失	阶跃
故障 12	冷凝器冷却水入口温度发生变化	随机变化
故障 17	未知	未知

3.2 实验过程

1) 数据的采集及预处理。TE 过程采集训练数据及测试数据, 将相应故障数据的标签与数据集进行组合。得到训练数据 $4\ 800 \times 52$, 测试数据 $2\ 900 \times 52$ 。

2) 数据的特征提取及降维。将数据经过 MDP 算法映射到低维空间, 进行故障特征的提取。为表明 MDP 降维的优势, 选取 3 类测试数据进行测试。图 2 是降维前的 2D

散点图, 图 3 是经过 MDP 降维后的 2D 散点图。由图 2 和 3 可见, 降维后的 3 类清晰可分, 类内之间更加紧凑。这意味着 MDP 算法在故障识别中有较好的性能。

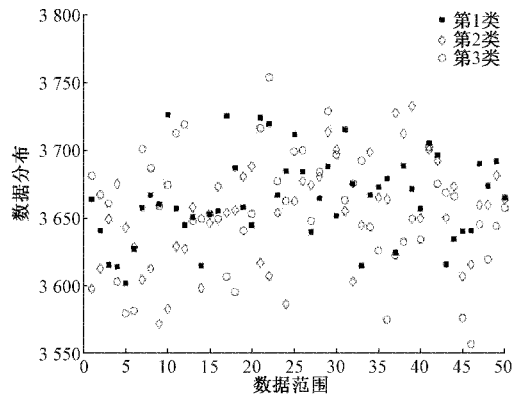


图 2 降维前部分测试数据可视化图

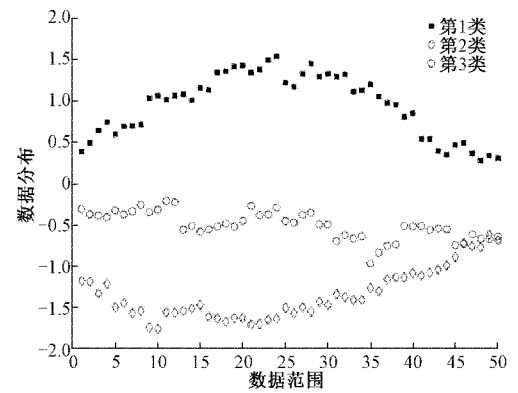


图 3 降维后部分测试数据可视化图

3) 建立 SVM 分类模型, 通过训练数据拟合出 SVM 分类模型。根据累计贡献率选取训练数据维数, 确定 MDP 算法降维后维度为 36。

4) 确定最佳性能 SVM 分类模型。将降维后的 36 个变量作为 SVM 模型的输入, SVM 模型的输出为 6 类数据的标签, 记为正常工况为 0, 故障类型为 1、5、7、12 和 17。通过 PSO 寻优算法及 K 折交叉验证, 确定最佳 SVM 模型的参数, 其中惩罚系数 $c=50$, 核函数参数 $g=0.124\ 4$, 寻优适应度曲线图如图 4 所示。其中种群数量为 20, 共 200 代。

5) 将经过 MDP 算法处理的测试集数据输入到优化后的 SVM 分类器中, 未产生明显的过拟合和欠拟合现象。

3.3 实验结果及分析

通过测试数据对模型进行测试, 故障诊断结果使用混淆矩阵表示, 如图 5 所示, 该图横纵坐标标为 0、1、5、7、12、17 代表 TE 过程 6 类数据, 对正常工况以及五类故障可以进行联合诊断。图中每一行按顺序代表真实的类别情况, 每一列数据代表预测情况。以第 3 行为例, 纵坐标标为 5, 代表 TE 故障 5, 对其 480 个测试样本, 判断为故障 5 的有

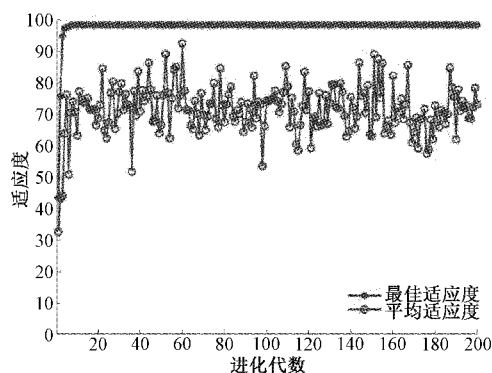


图4 粒子群适应度曲线

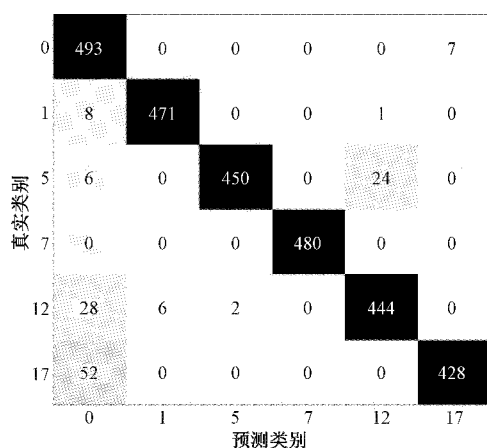


图5 诊断结果的混淆矩阵

459个(即第3列显示的数值459),准确率为95.625%。

为了表明本文所提MDP-SVM方法具有良好的检测效果,采用SVM、PCA-SVM、LLE-SVM算法对TE过程6种故障类型的分类准确率的对比,结果如表2所示。由表2可以看出MDP-SVM方法故障诊断效果最好,准确率在95%以上,表明了该方法有效性。

表2 分类准确率统计

方法	分类准确率/%
SVM	37.379 3
PCA-SVM	94.275 9
LLE-SVM	89.500 0
MDP-SVM	95.379 3

4 结 论

本文提出基于边界判别投影和SVM相融合进行故障诊断的方法,主要通过MDP方法使类内距离减小,类间距离增大使得不同类型故障在低维空间中更易于区分。通过采用PSO与K折交叉验证得到SVM模型最佳参数。通过仿真实验并与SVM、LLE-SVM、PCA-SVM进行对比,结果表明,在不同类型故障诊断方面,本文方法具有较高的检测能

力。但目前算法只能判断哪类故障发生,不能准确定位哪里发生了故障,因此,今后将在故障定位方面进行一些研究。

参考文献

- [1] 华丽,于海晨,邵诚,等.基于SVM-BOXPLOT的乙烯生产过程异常工况监测与诊断[J].化工学报,2018,69(3):1053-1063.
- [2] 冀丰偲,余云松,张早校.LDA_SVM方法在化工过程故障诊断中的应用[J].高校化学工程学报,2020,34(2):487-494.
- [3] 杨剑锋,乔佩蕊,李永梅,等.机器学习分类问题及算法研究综述[J].统计与决策,2019,35(6):36-40.
- [4] 孙扩,李文海,王怡苹.基于SVM专家系统的模拟电路PCB故障诊断研究[J].国外电子测量技术,2018,37(9):22-26.
- [5] 朱哈娜,刘慧明.基于改进VMD与GS_SVM的轴承故障诊断[J].电子测量技术,2020,43(21):71-76.
- [6] 陈昭明,徐泽宇,赵迎.DPCA与GA-SVM融合的智能台车液压系统故障诊断[J].控制工程,2020,27(11):1980-1986.
- [7] 刘冬梅,霍龙龙,王浩然,等.基于PSO-SVM的电流放大器故障诊断研究[J].传感器与微系统,2021,40(8):50-52,56.
- [8] 陶志勇,于子佳,林森.PSO_SVM算法在太阳能电池板裂缝缺陷检测研究[J].电子测量与仪器学报,2021,35(1):18-25.
- [9] 刘伟,韩彦华,王荆,等.基于粒子群算法优化支持向量机的变压器绕组变形分类方法[J].高压电器,2020,56(3):72-78.
- [10] 焦卫东,林树森.整体改进的基于支持向量机的故障诊断方法[J].仪器仪表学报,2015,36(8):1861-1870.
- [11] 吉敏.基于PCA-SVM的轴承故障诊断研究[J].电子设计工程,2019,27(17):14-18.
- [12] 伍吉瑶,王璐,程正南,等.基于LLE和SVM的手部动作识别方法[J].传感器与微系统,2016,35(8):4-7.
- [13] 何进荣,丁立新,李照奎,等.基于边界判别投影的数据降维[J].软件学报,2014,25(4):826-838.
- [14] 魏义康,金聪.边界判别投影在人脸识别上的应用[J].电子测量技术,2019,42(4):104-109.
- [15] 张素智,陈小妮,杨芮,等.基于类内和类间距离的主成分分析算法[J].计算机工程与设计,2020,41(8):2177-2183.
- [16] 汪海燕,黎建辉,杨风雷.支持向量机理论及算法研究综述[J].计算机应用研究,2014,31(5):1281-1286.
- [17] 王二化,刘颀.基于PCA和改进型SVM的齿轮裂纹故障诊断方法[J].机械设计与研究,2021,37(2):83-87,91.
- [18] 徐可,陈宗海,张陈斌,等.基于经验模态分解和支持向量机的滚动轴承故障诊断[J].控制理论与应用,2019,36(6):915-922.

- [19] 王钰,赵晓艳,杨杏丽,等. 基于 K 折交叉验证 Beta 分布的 AUC 度量的置信区间[J]. 系统科学与数学, 2020,40(9):1564-1577.
- [20] 张鑫,胡瑾秋,张来斌,等. 基于 RS 和 SVM 的化工过程高精度故障诊断方法[J]. 石油学报(石油加工), 2017,33(4):777-784.

作者简介

郭小萍,博士,教授,主要研究方向为基于数据驱动的复

杂过程故障诊断。

E-mail:gxp2001@sina.com

尹瑞琛,硕士研究生,主要研究方向为基于数据驱动的复杂过程故障诊断。

E-mail:2271655107@qq.com

李元,博士,教授,主要研究方向为基于数据驱动的复杂过程故障诊断。

E-mail:li-yuan@mail.tsinghua.edu.cn