文章编号:1671-4598(2018)02-0053-04 DOI:10.16526/j.cnki.11—4762/tp.2018.02.014 中图分类号:TP242 文献标识码:A

基于 ELM—AdaBoost.M2 的污水处理过程在线故障诊断

谭承诚[1]，于广平[2]，邱志成[1]
(1. 华南理工大学 机械与汽车工程学院，广州 510640；
2. 中国科学院 沈阳自动化研究所广州分所，广州 511458)

摘要：污水处理存在着强非线性和非稳态运行等特征，对其运行过程进行在线故障诊断在减少污染和保障生产过程安全方面具有重大意义；针对污水处理过程运行状态的不平衡分布造成故障诊断准确率下降的问题，提出一种基于极限学习机（ELM）和 AdaBoost.M2 算法的在线故障诊断方法；该模型以 ELM 为弱分类器，利用 AdaBoost.M2 将多个弱分类器集成，实现了强分类器；仿真结果表明，该模型在线故障诊断精度高，学习速度快，泛化性能好，相较于传统故障诊断方法，综合性能较为突出，较好地实现了污水处理的在线故障诊断。

关键词：污水处理；故障诊断；极限学习机；AdaBoost.M2；在线建模

Online Fault Diagnosis of Wastewater Treatment Process Based on ELM—AdaBoost.M2

Tan Chengcheng[1], Yu Guangping[2], Qiu Zhicheng[1]
(1. School of Mechanical & Automotive Engineering, South China University of Technology, Guangzhou 510640, China;
2. Shenyang Institute of Automation in Guangzhou. Chinese Academy of Science, Guangzhou 511458, China)

Abstract: Wastewater treatment exists strong nonlinearity, unsteady operation and other characteristics, the online fault diagnosis of wastewater treatment process in reducing pollution and ensure the safety is of great significance. Concerning the low accuracy of fault diagnosis induced by the unbalanced distribution of wastewater treatment process' running state, an online fault diagnosis model based on extreme learning machine (ELM) and AdaBoost.M2 is proposed. Firstly, set ELM as weak classifier and then use AdaBoost.M2 to embody several weak classifiers into strong classifier. The simulation experiments demonstrated that this online diagnosis model has higher precision, faster speed, better generalization ability, and outstanding performance, while comparing to the traditional ones. Therefore, the proposed model can meet the requirements of online fault diagnosis of wastewater treatment process.

Keywords: wastewater treatment; fault diagnosis; extreme learning machine; AdaBoost.M2; online modeling

0 引言

污水处理是一个典型的流程行业过程，具有数据量大、强耦合性、工业噪声和过程干扰多、动态性强等特点[1]，对其过程的实时监控，在避免环境污染、降低经济损失、保障生产安全等方面具有重大意义。污水处理既有复杂机械电气系统的参与，又包含众多的生化反应过程，对其系统内在的运行原理进行研究，成本过高，然而污水处理过程会产生海量隐藏着工艺变动和设备运转的数据，于是如何利用好这些数据来提高污水处理过程的效率与安全，降低二次污染成为了一个急需解决的问题。因此基于数据驱动的污水处理过程故障诊断方法得到了广泛的研究。文献［2］利用 BP 神经网络对污水厂进行了实际污水处理过程运行状态的监控，显示了良好的效果和稳定性，文献［3］提出了一种基于支持向量机的污水处理过程故障诊断模型，文献［4］根据污水处理厂的实际情况，提出了一种改进的支持向量机故障诊断方法，结果表明该方法可以较好的满足污水厂对安全生产的要求。上述故障诊断模型都取得了一定了成果，但神经网络具有在学习过程易陷入局部极小值、过拟合、训练时间长等缺点[5]，支持向量机随着样本量的增多，训练时间会变长，支持向量增多，模型的稀疏性渐失[6]。与此同时，污水处理过程的运行状态具有较强的不平衡分布，即正常运行状态所占比例远高于故障状态的所占比例，传统的神经网络与支持向量机故障诊断模型对少数类（故障类）的识别率极不理想。

根据现有成果与存在的问题，本文提出一种基于极限学习机（extreme learning machine，ELM）和 AdaBoost.M2 相结合的污水处理过程在线故障诊断模型。ELM 在学习过程中随机产生输入层与隐含层之间的连接权值和隐含层神经元的阈值，无需调整参数，仅需设置隐含层神经元的个数，便可将传统单隐层前置神经网络参数训练问题转化为求解线性方程组[7]，最终得到全局最优解。考虑到故障诊断是多分类问题，利用 AdaBoost.M2 算法的集成提升作用可将基于 ELM 的弱分类器构造为强分类器。通过分层组合，使迭代权重的重点聚焦于少数类或极少数类的困难数据上，提高了分类器的准确性和泛化性能，亦满足在线污水处理过程故障诊断对于实时性的要求，并通过仿真实验得到了验证。

收稿日期:2017-07-03； **修回日期**:2017-07-31。
基金项目:广东省科技项目(2016B090918113)。
作者简介:谭承诚(1994-),男,四川自贡人,硕士研究生,主要从事水质预测与故障诊断方向的研究。

1 基于 ELM－AdaBoost. M2 的污水处理在线故障诊断模型

1.1 极限学习机 ELM

Huang[8]在单隐层前馈神经网络（single－hidden layer feedforward neural network，SLFN）框架的基础上提出了极限学习机 ELM 算法。给定 N 个不同样本 $(x_i,t_i)_{i=1}^{N}$，$x_i \in R^m$，$t_i=[t_{i1},\cdots,t_{ik}]^T$，$m$ 为样本维数。设一个单隐层前馈神经网络有 K 个隐藏节点，则此输出模型可以表示为：

$$\sum_{i=1}^{K}\beta_i g(w_i,x_i,b_i)=o_j,\qquad j=1,2,\cdots,N \tag{1}$$

其中：x_i 表示数据集中的第 i 组样本数据，β_i 表示第 i 个隐含层神经元到输出层的连接权值，$g(x)$ 表示隐含层神经元的激活函数，w_i 表示输入层到第 i 个隐含层神经元到输入层的连接权值，b_i 表示第 i 个隐含层神经元的偏置。

若隐含层神经元个数与训练样本个数相等[11]时，对于任意的 w 和 b，SLFN 均可以零误差逼近训练样本集，即 $\sum_{j=1}^{K}\|o_j-t_j\|=0$，继而有：

$$\sum_{i=1}^{K}\beta_i g(w_i,x_i,b_i)=t_j,\qquad j=1,2,\cdots,N \tag{2}$$

式（2）可以表示为：

$$\boldsymbol{H\beta}=\boldsymbol{T} \tag{3}$$

其中：

$$\boldsymbol{H}=\boldsymbol{H}(w_1,\cdots,w_K,b_1,\cdots,b_K,x_1,\cdots,x_N)=\begin{bmatrix} g(w_1x_1+b_1) & \cdots & g(w_kx_1+b_K)\\ \vdots & \ddots & \vdots\\ g(w_1x_N+b_1) & \cdots & g(w_kx_N+b_K)\end{bmatrix}=\begin{bmatrix}h(x_1)\\ \vdots\\ h(x_N)\end{bmatrix},$$

$$\boldsymbol{\beta}=\begin{bmatrix}\boldsymbol{\beta}_1^T\\ \vdots\\ \boldsymbol{\beta}_K^T\end{bmatrix}_{K\times m},\boldsymbol{T}=\begin{bmatrix}\boldsymbol{t}_1^T\\ \vdots\\ \boldsymbol{t}_K^T\end{bmatrix}_{N\times m}$$

其中：$\boldsymbol{H}$ 为隐含层输出矩阵，$\boldsymbol{\beta}$ 为输出权值矩阵，$\boldsymbol{T}$ 为输出层输出矩阵。

在实际应用中，当激活函数 $g(x)$ 无限可微时，ELM 隐含层节点数小于训练样本数 N，SLFN 的隐含层参数 w 和 b 无需全部调整，且在训练中保持不变。因此此时，隐含层输出矩阵不是方阵，根据广义逆定理[12]，上述线性系统的可以通过求解 $\min_{\beta}\|\boldsymbol{H\beta}-\boldsymbol{T}\|$ 获得隐含层与输出层间的连接权值 β，其唯一且最小解为：

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{+}\ \boldsymbol{T} \tag{4}$$

其中：$\boldsymbol{H}^{+}$ 为隐含层输出矩阵 $\boldsymbol{H}$ 的 Moore－Penrose 广义逆。

具体地，ELM 算法有以下几个步骤[9]：

1）确定隐含层神经元个数，随机设定输入层与隐含层的连接权值 w 和偏置 b；

2）选择一个无限可微的激活函数，计算隐含层输出矩阵 $\boldsymbol{H}$；

3）据式（4）计算输出层权值 $\hat{\beta}$。

1.2 AdaBoost. M2 算法

AdaBoost 算法在解决二类分类问题时，只需弱分类器对任意样本分类准确率比 0.5 略高，然而，若直接将 AdaBoost 应用于多分类问题，这一条件过强，同时，要求弱分类器比随机猜测准确率略高的条件又过弱，因此会导致集成的强分类器准确率较低[10]。针对多分类问题，可以选择 AdaBoost 的扩展算法 AdaBoost. M2。

AdaBoost 算法的错分概率为：

$$\frac{1}{2}(1-h(x_i,y_i)+h(x_i,y)) \tag{5}$$

其中：x_i 为样本数据，y_i 为样本类标签，y 表示除 y_i 之外的类标签。对于 $k(k>2)$ 类问题，$k-1$ 个不同的 y，定义不同的 y 的权重为 $q(i,y)$，且 $\sum_{y\neq y_i}q_t(i,y)=1$，代入式（5）得到：

$$\frac{1}{2}(1-h(x_i,y_i)+\sum_{y\neq y_i}q_t(i,y)h(x_i,y)) \tag{6}$$

从而可得 AdaBoost. M2 算法的伪误差：

$$\varepsilon_t=\frac{1}{2}\sum_{i=1}^{N}D_t(i)(1-h_t(x_i,y_i)+\sum_{y\neq y_i}q_t(i,y)h_t(x_i,y)) \tag{7}$$

式中，$h_t(x_i,y)$ 表示弱分类器 h_t 将 x_i 分为 y 的置信度；$q_t(i,y)$ 为标签加权函数，表示将样本 x_i 错误分为类别 y 的概率，值越大表示样本被错分的概率越大，则该样本在下一次迭代的时候会得到重点学习。$q_t(i,y)$ 在多次的循环迭代过程中不断的改变样本的权重，对易错分样本进行重点学习，继而提高了弱分类器的泛化能力和学习能力，提高了分类的准确率和稳定性[11]。

1.3 ELM－AdaBoost. M2 算法流程

针对污水处理过程运行状态分布不平衡，其本质是一个多分类的问题，ELM－AdaBoost. M2 算法将多个 ELM 弱分类器集成为强分类器，在集成为强分类器的过程中对被错分样本赋予更大权值，使错误分类的样本在下一次迭代中被重点学习，最终可以提高模型的分类准确率和泛化能力。ELM－AdaBoost. M2 算法的实现流程如下所示。

对于污水处理过程原始数据进行归一化处理，得到个样本的训练集：(x_i,y_i)，$x_i\in R^m$，$y_i\in(1,2,\cdots,k)$，其中为类别数。定义 D 为样本上的分布，T 为弱分类器的个数，即循环迭代的次数。

1）初始化：设置权值向量 $w_{i,y}^t=D(i)/(k-1)$；$i=1,2\cdots,N$，$y\in Y-y_i$，$D(i)=1/N$；

2）for i=1：T

（1）令 $W_i^t=\sum_{y\neq y_i}w_{i,y}^t$，$q_t(i,y)=w_{i,y}^t/W_{i,y}^t$，$(y\neq y_i)$，样本分布权值 $D_t(i)=W_i^t/\sum_{i=1}^{N}W_i^t$，其中，$D_t(i)$ 为当前样本的权值占所有样本集的比重，$q_t(i,y)$ 为类别加权函数，表示样本被错误分为类别 y 的概率，W_i^t 为当前样本的各类权值总和；

（2）根据样本分布 $D_t(i)$ 选择新样本训练极限学习机 ELM，得到弱分类器 $h_t:X*Y\rightarrow[0,1]$；

（3）根据式（7）计算伪误差 ε_t，若 $\varepsilon_t\geqslant 0.5$：转到步骤 3）；

（4）令 $\beta_t=\varepsilon_t/(1-\varepsilon_t)$，更新权值向量：

$$w_{i,y}^{t+1}=w_{i,y}^t\beta_t \tag{8}$$

其中：$b_t=(1/2)(1-h(x_i,y_i)+h(x_i,y))$；

end

3）结束循环，输出强分类器 H：

$$H(x)=\underset{y\in Y}{\operatorname{argmax}}\sum_{1}^{T}(\log\frac{1}{\beta_t})h_t(x,y) \tag{9}$$

由式（8）和式（9）可知，ε_t 越小，$\log\frac{1}{\beta_t}$ 越大，亦即弱分类器的置信度越高，在组成强分类器时的影响也越大。污水处理在线故障诊断模型建模流程如图 1 所示。

2　实验仿真与结果分析

2.1　实验数据

本文所用污水处理过程数据来源于加州大学欧文分校机器学习数据库，该数据集共包含 527 个样本，每个样本具有 38 个属性，其中无缺失值的样本 380 个，具有 13 种运行状态，为简化分类难度，污水处理过程的运行状态分为 4 类，其中类别 1 为正常运行状态，类别 2 为运行性能为高于均值的正常状态，类别 3 为进水量较少的正常运行状态，类别 4 表示运行故障状态。4 中运行状态的比率为 23.7∶8.29∶4.64∶1，属于典型的不平衡分布数据集。

2.2　性能评价指标

污水处理过程的故障诊断是一个多分类问题，具有不平衡分布的数据集是分类问题的一个特例。因此，本文中采用 F_1 度量（$F_1-score$）作为衡量故障诊断模型性能的主要指标，模型诊断准确率 acc 和仿真实验总时间 Time 作为辅助评价指标。多分类问题的混淆矩阵如表 1 所示，表中 f_{ij} 表示第 i 类被分到第 j 类的个数。模型的故障诊断准确率 acc、第 i 类的诊断精度 precision 以及召回率 recall 分别定义为 $acc=\sum_{i=1}^{k}f_{ii}/N$，$p_i=f_{ii}/\sum_{j=1}^{k}f_{ji}$，$r_i=f_{ii}/\sum_{j=1}^{k}f_{ij}$。$F_{1i}$是精度 p_i 和召回率 r_i 的调和平均，定义为 $F_{1i}=2r_ip_i/r_i+p_i$。故模型整体的性能指标 $F_1-score$ 计算公式为：

$$F_1-score=(\prod_{i=1}^{k}F_{1i})^{\frac{1}{k}} \tag{10}$$

其中：N 表示样本总数，k 表示类别数。

表 1　多分类问题的混淆矩阵

Item	Predictive class 1	Predictive class 2	…	Predictive class k
Actual class 1	f_{11}	f_{12}	…	f_{1k}
Actual class 2	f_{21}	f_{22}	…	f_{2k}
…	…	…	…	…
Actual class k	f_{k1}	f_{k2}	…	f_{kk}

2.3　在线仿真实验

仿真实验中，仅提取出数据集中具有完整属性的 380 个样本，按 2：1 的比例随机分层抽样，得到训练集 Xtr 和测试集 Xte。对训练集 Xtr 进行归一化处理，将处理后的数据分别输入 BP 神经网络模型，支持向量机（SVM）模型，相关向量机（RVM）模型，极限学习机（ELM）模型以及本文提出的基于 ELM－AdaBoost. M2 的故障诊断模型中进行离线建模和故障诊断测试。其中 BP 神经网络为三层结构，隐含层节点数由 5 折交叉验证法在一定的范围内搜索，最终确定 BP 神经网络采用 38－10－4 结构；SVM 模型采用径向基（RBF）核函数，利用遗传算法来搜索惩罚参数 c 和径向基函数参数 g，采用一对一（one－versus－one）的分类方法实现。ELM 模型和 ELM－AdaBoost. M2 模型选用 sigmoid 函数为激活函数，隐含层节点个数使用 5 折交叉验证法在［10，200］内择优选取。ELM－AdaBoost. M2 模型迭代次数 T 初始化为 10。

选择不同的隐含层节点个数对 ELM 模型和 ELM－AdaBoost. M2 模型的交叉验证准确率均有影响，如图 1 和图 2 所示。从图 1 可以看出，ELM 模型的节点在 45 个时，交叉验证的准确率最高，随着节点个数的增加，交叉验证准确率整体呈下降趋势，

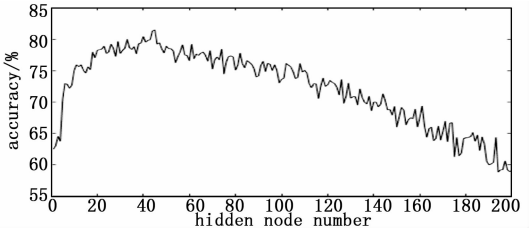

图 1　ELM 模型不同隐含层节点个数的验证准确率

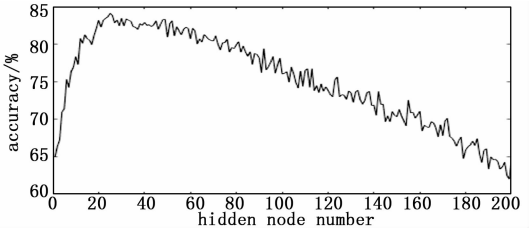

图 2　ELM－AdaBoost. M2 模型不同隐含层节点个数的验证准确率

可见对于 ELM 模型，隐含层节点的个数不是越多越好，过多的隐含层节点会造成过拟合现象。从图 2 可以看出，ELM－AdaBoost. M2 相对于 ELM 模型，模型整体的交叉验证准确率有明显的提升，同时该模型验证准确率最高是在 25 个节点左右，45 个节点左右的验证准确率在模型中并不突出，该模型对于交叉验证准确率较低的弱分类器提升效果明显，较强的弱分类器提升效果一般，由此可见不同的弱分类器对于模型最终准确率的提升效果不同。

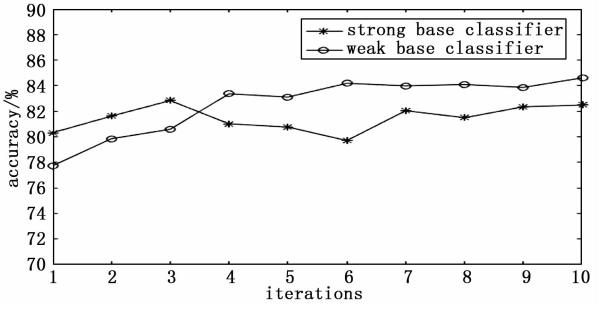

图 3　基分类器验证准确率

图 3 所示为弱分类器迭代 10 次的交叉验证准确率趋势图，可以看出，较强的弱分类器随着迭代次数的增加，不仅最终交叉验证准确率提升远低于较弱的弱分类器，同时鲁棒性也低于较弱的弱分类器，而较弱的弱分类器的交叉验证准确率稳中有升，最终迭代组成的强分类器的交叉验证准确率提升明显。综合对图 1～图 3 的分析可知，对于 ELM－AdaBoost. M2 模型而言，隐含层的节点数与弱分类器的个数对模型的性能均有较大的影响。所以利用 GA 算法对隐含层节点个数与迭代次数进

行寻优，隐含层节点的搜索范围为［10，200］和迭代次数的搜索范围为［5，20］。得到结果，当 ELM－AdaBoost. M2 模型在节点数为 23 和迭代次数为 15 次数时，交叉验证的准确率达到最佳。

模型进行在线故障诊断后，得到的每一组新样本都需加入模型进行更新。历史数据集通常采用限定记忆法来保持其容量，即当加入一组新数据时，便删除最早的一组样本数据，从而保证模型始终包含新数据的信息。本文中各模型的在线仿真实验均进行 10 次，得到的数据为 10 次试验的算术平均值，各模型的在线诊断结果如表 2 所示。

表 2 4 种模型的在线诊断结果 Model

Model	F1－score/%	acc/%	Time/s
BP	30.35	82.03	154.9
SVM	40.17	82.57	1310.64
ELM	36.52	83.10	12.53
ELM－A	52.02	86.56	97.44

从表 2 中可以看出，BP 神经网络模型的模型训练时间尚可，但准确率 *acc* 和 *F*1－*score* 皆为最低，原因在于 BP 神经网络在学习过程中陷入了局部最小点未达到全局最优，其中对少数类的识别率过低是造成其 F_1－*score* 最低的主要原因；相对于 BP 神经网络模型，SVM 模型的 *acc* 仅提高 0.54%，但 *F*1－*score* 却提高了 9.82%，原因在于 SVM 基于核函数将原始数据映射到高维特征空间后降低了原始数据的强非线性，对于少数类的识别率有所提升，但数据维数的增加也大幅的增加了模型的训练时间，所耗时间远大于文中其他模型；ELM 模型将传统的 SLFN 参数的训练转换成对线性方程组的求解问题，大大地减少了模型训练时间，相比于 BP 神经网络模型降低了 12 倍，另外 *acc* 提高了 1.07%，*F*1－*score* 提高了 6.17%，原因是 ELM 模型对于少数类的识别率高于 BP 模型；ELM－AdaBoost. M2 模型消耗的训练时间多于 ELM 模型是因为组合弱分类器的过程造成，集成的 ELM－AdaBoost. M2 强分类器模型，相比于基于 BP、SVM 和 ELM 的故障诊断模型，*acc* 分别提高了 4.53%、3.99%和 3.46%，*F*1－*score* 分别提高了 21.67%、11.85%和 15.5%，不仅提升了模型的准确率，也大幅提升了 F_1－*score*，可以看出该模型对于少数类的识别率较为理想，较好地克服了具有不平衡分布数据集对传统分类算法带来的影响，该模型一次更新和测试的总时间平均为 0.77 s，可以满足污水处理过程故障诊断对于实时性的要求。

结合以上分析，基于 ELM－AdaBoost. M2 的污水处理故障诊断模型综合性能优于其他模型，满足实际污水处理中对于实时性和准确性的要求。

3 结论

污水处理是一个复杂的生化反应过程，具有强非线性特征，对其运行过程进行在线故障诊断是减少污染和保障安全生产的重要方法之一。针对污水处理过程运行状态的不平衡分布造成故障诊断准确率下降的问题，本文提出一种基于 ELM－AdaBoost. M2 的污水处理过程在线故障诊断模型。该模型以 ELM 为弱分类器，利用 AdaBoost. M2 算法对弱分类器的集成提升作用，组成强分类器，建立了污水处理过程的在线故障诊断模型。仿真实验结果表明，相对传统模型，该模型兼具学习速度快、分类准确率高和泛化能力强等优点，克服了污水处理过程状态不平衡分布带来的不良影响，较好地实现对污水处理过程的在线故障诊断。

参考文献：

[1] 张瑞成，王 宇，李 冲. 基于 NW 型小世界人工神经网络的污水出水水质预测 [J]. 计算机测量与控制，2016，24 (1)：61－63.

[2] Fuente M J，Vega P. Neural networks applied to fault detection of a biotechnological process [J]. Engineering Application of Artificial Intelligence，1999，12 (5)：569－584.

[3] 王华忠，张雪申，俞金寿. 基于支持向量机的故障诊断方法 [J]. 华东理工大学学报：自然科学版，2004，30 (2)：179－182.

[4] 李晓东，曾光明，蒋 茹，等. 改进支持向量机对污水处理厂运行状况的故障诊断 [J]. 湖南大学学报：自然科学版，2007，34 (12)：68－71.

[5] Tian Y，Qiao J F. Neural network soft measurement of BOD based on genetic algorithm [J]. Computer Technology and Development，2009，19 (3)：127－133.

[6] Tipping M E. Sparse Bayesian learning and the relevance vector machine [J]. Journal of Machine Learning Research，2001，1 (3)：211－244.

[7] Huang G B，Zhu Q Y，SIEW C K. Extreme learning machine：a new learning scheme of feedforward neural networks [A]. Proceedings of 2004 IEEE International Joint Conference on Neural Networks [C]. 2004：985－990.

[8] Huang G B，Zhu Q Y，SIEW C K. Extreme learning machine：theory and application [J]. Neuro－computing，2006，70 (1)：489－501.

[9] 许有才，万 舟，汤 超. 基于 IFD 与 DE－ELM 的轴承故障诊断 [J]. 计算机测量与控制，2015，23 (12)：3990－3994.

[10] 曹 莹，苗启广，刘家辰，等 AdaBoost 算法研究进展与展望 [J]. 自动化学报，2013，39 (6)：745－758.

[11] Zhu J，Zou H，Rosset S，et al. Multi－class AdaBoost [J]. Statistic and Its Interface，2009，2 (1)：349－360.

(上接第 52 页)

[20] 倪安福. 基于包络谱分析的滚动轴承故障诊断方法研究 [J]. 煤矿机械，2017，38 (02)：155－159.

[21] 郭庆丰，王成栋，刘佩森. 时域指标和峭度分析法在滚动轴承故障诊断中的应用 [J]. 机械传动，2016，40 (11)：172－175.

[22] 康守强，王玉静，杨广学，等. 基于经验模态分解和超球多类支持向量机的滚动轴承故障诊断方法 [J]. 中国电机工程学报，2011，31 (14)：96－102.

[23] 崔克楠，张志鹏，朱 彤，等. 基于蛙跳算法优化 SVM 的小功率 LED 寿命预测模型 [J]. 半导体技术，2016，41 (9)：711－715，720.

[24] 王健峰，张 磊，陈国兴，等. 基于改进的网格搜索法的 SVM 参数优化 [J]. 应用科技，2012，39 (3)：28－31.

[25] 吕洪艳，刘 芳. 组合核函数 SVM 在特定领域文本分类中的应用 [J]. 计算机系统应用，2016，25 (5)：124－128.