DOI：10.16183/j.cnki.jsjtu.2022.234

基于自步学习的刀具加工过程监测数据异常检测方法

张　建[1]，　胡小锋[1]，　张亚辉[2]
（1. 上海交通大学 机械与动力工程学院，上海 200240；
2. 上海交通大学 海洋装备研究院，上海 200240）

摘　要：针对零件加工过程的监控数据异常，导致刀具剩余寿命预测准确性下降的问题，提出一种基于自步学习的数据异常检测方法. 首先建立多层感知机模型关联刀具加工过程监测数据和所对应的刀具剩余寿命，其次在模型权重更新过程中，先固定模型权重参数，预测损失拟合高斯分布得到异常样本的损失阈值，然后构建基于自步学习方法的损失函数，迭代更新模型参数. 在模型训练结束后根据损失阈值划分出异常样本. 最后利用汽轮机转子轮槽的实际加工监测数据进行验证，并与局部异常因子算法、基于密度的聚类算法、K 均值算法、孤立森林算法、一分类支持向量机等方法进行对比分析，验证本方法的有效性.

关键词：刀具加工监测；数据质量；异常检测；自步学习
中图分类号：TH 166　　**文献标志码**：A

Abnormal Detection Method of Tool Machining Monitoring Data Based on Self-Paced Learning

ZHANG Jian[1]，　*HU Xiaofeng*[1]，　*ZHANG Yahui*[2]
（1. School of Mechanical Engineering，Shanghai Jiao Tong University，Shanghai 200240，China；
2. Institute of Marine Equipment，Shanghai Jiao Tong University，Shanghai 200240，China）

Abstract：Aiming at the problem that the accuracy of tool remaining life prediction was reduced due to the abnormal monitoring data in machining process，a data anomaly detection method based on self-paced learning was proposed. Firstly，a multi-layer perceptron model was established to correlate the tool processing monitoring data with the tool remaining life. Secondly，in the process of updating model weight，the model weight parameters were fixed first，and the loss threshold of abnormal samples was obtained by predicting loss fitting gaussian distribution. Then，the loss function based on self-paced learning method was constructed to update model parameters iteratively. At the end of the model training，abnormal samples were divided according to the loss threshold. Finally，the actual processing monitoring data of turbine rotor groove are used to verify the validity of the proposed method，and compared with local anomaly factor algorithm，density-based clustering algorithm，K-means algorithm，isolated forest algorithm and one-class support vector machines.

Key words：tool machining monitoring；data quality；anomaly detection；self-paced learning

收稿日期：2022-06-21
基金项目：国防基础科研计划项目(JCKY2021110B048)，国家重点研发计划资助项目(2018YFB1700502)
作者简介：张　建(1997-)，男，浙江省宁波市人，硕士生，现主要从事刀具磨损预测研究.
通信作者：胡小锋，男，研究员，博士生导师，电话(Tel.)：021-34205694；E-mail：wshxf@sjtu.edu.cn.

高附加值零件需要进行过程监控来保证其加工质量，而加工过程监测数据在采集过程中由于传感器、采集和传输设备受环境影响大，导致数据中存在异常值[1]. 这些异常数据与真实数据有显著差异，而刀具加工监测异常数据直接影响刀具剩余寿命预测的准确性.

数据异常检测算法主要分为有监督和无监督算法. 其中监督或者半监督的方法通过带标签的正常数据和异常数据来训练分类模型. 尚文利等[2]利用堆叠自编码神经网络(SAE)对工艺数据进行特征降维，然后设计长短期记忆神经网络(LSTM)来进行异常检测. 夏英等[3]提出一种融合了新型统计方法和双向卷积 LSTM 异常检测方法，能够处理多维时序数据. 孙滢涛等[4]对电力数据时间序列进行多域特征提取，并采用相关向量机和支持向量数据描述进行特征选择降维和异常检测. 傅世元等[5]提出一种基于元学习动态选择集成的电力调度数据异常检测方法. 有监督的异常检测依赖于已知的异常样本，但在加工监控过程中，首先实际生产加工的零件型号、使用的刀具多变，有异常标签的数据样本难以获取，其次异常信号的来源复杂，无法获取完备的异常数据来训练异常检测模型.

与有监督的算法不同，无监督异常检测方法从数据样本的统计规律[6]和样本间的距离[7]出发. 吴蕊等[8]结合数据对象的密集度与最大近邻半径，优化 K-means 初始聚类中心，在电力数据异常检测上取得了优异的效果. 吴金娥等[9]提出采用反向 k 近邻算法实现异常数据检测. 陈砚桥等[10]基于密度的聚类算法(DBSCAN)实现了多源数据异常检测. 宋丽娜等[11]将局部异常因子(LOF)算法与互补集成经验模态分解(CEEMD)法进行结合，识别监测数据的异常值. 王峰等[12]针对电力调度数据异常，提出基于对数区间隔离的检测方法. 王燕晋等[13]基于孤立森林算法提出了一种电力用户数据异常快速识别方法. 然而基于聚类的异常检测结果依赖聚类的效果[14]. 在实际加工监测过程中，采集的加工监测数据随刀具的剩余寿命减少而变化，导致正常和异常数据难以区分.

本文针对刀具加工监测异常数据无标签，加工监测数据随刀具性能衰退而变化，考虑刀具加工剩余寿命因素，提出基于多层感知机模型的预测偏差来实现异常数据样本的检测. 在多层感知机的训练过程中采用高斯分布来拟合训练样本损失，并融合自步学习框架来提升模型对正常样本的筛选能力. 最终将异常筛选前后的数据用于铣刀剩余寿命预测中，来验证异常数据检测的重要性和有效性.

1 相关技术

1.1 多层感知机

多层感知机包含输入层、输出层以及多个隐藏层. 相邻层之间的神经元节点进行全连接，即上一层的每个神经元都与下一层的所有神经元连接，同时同一层的神经元节点没有连接. 前一层的输出通过激活函数与下一层的输入进行关联.

多层感知机模型的训练包含两个部分，分别是前向传递和反向传播. 前向传递过程中，训练样本数据从输入层输入，通过一个或者多个全连接层，每两个神经元之间的参数包含一个权重，从而对输入的数据进行拟合，最后通过输出层将数据进行输出. 反向传播过程则由输出值与样本的真实值构建损失函数，通过反向传播的梯度下降算法对模型的参数进行更新，当模型的损失函数降到最小值时，多层感知机模型就能够拟合样本特征. 第 I 层到第 J 层神经元的前向传递和反向传播过程如下：

$$y_j = \varphi\Big(\sum_{i=0}^{l-1} w_{ij} x_i\Big) \tag{1}$$

$$w_{ji}(t) = w_{ji}(t-1) + \Delta w_{ji}(t-1) = w_{ji}(t-1) - \alpha \frac{\partial \varepsilon(t)}{\partial w_{ji}(t-1)} \tag{2}$$

式中：l 表示第 I 层神经元的个数；w_{ij} 表示第 I 层第 i 个神经元和第 J 层第 j 个神经元之间的权重；x_i 表示第 I 层第 i 个神经元的输入；y_j 表示第 J 层第 j 个神经元的输入；φ 表示激活函数；t 表示迭代次数；α 为梯度下降的学习率；$\varepsilon(t)$ 为多层感知机的输出和真实值之间的损失函数.

1.2 自步学习

Bengio 于 2009 年在 ICML 上首次提出课程学习[15]，即让模型先学习简单的知识，然后逐渐增加难度，过渡到更复杂的知识上去. 而自步学习[16]在课程学习的基础上进行了改进，模型在每一步的迭代过程中来决定下一步学习的样本.

传统机器学习方法的目标函数如下所示，要求出使得目标函数最小的权重值：

$$w_t = \arg\min\Big(\sum_{i=0}^{m-1} f(x_i, y_i; w_{t-1})\Big) \tag{3}$$

式中：x_i 为第 i 个样本；y_i 为第 i 个样本的标签值；w_t、w_{t-1} 分别为第 t 和第 $t-1$ 次迭代过程中模型的权重；m 为样本个数.

不同于传统的机器学习，自步学习在每一次的迭代过程中会倾向于从所有样本中选择具有较小训

练误差的样本，然后更新模型参数. 因此在每一次的迭代过程中，并不是所有的样本都参与了模型参数的更新. 自步学习在传统机器学习的目标函数中引入了二分变量 v_i，该变量用于表征每个样本是否被选择参与训练，其目标函数改写为

$$w_t, v_t = \arg\min\left(\sum_{i=0}^{m-1} v_i f(x_i, y_i; w_{t-1}) - \lambda \sum_{i=0}^{m-1} v_i\right) \quad (4)$$

式中：λ 为样本难易程度的筛选阈值. 当损失值 $f(x_i, y_i; w_{t-1}) < \lambda$ 时，v_i 取 1，而当损失值 $f(x_i, y_i; w_{t-1}) \geqslant \lambda$ 时，v_i 取 0. 自步学习中 λ 的选取往往需要人为给定，本文通过高斯分布来拟合训练样本的误差，从而自适应地选取 λ，将高于阈值的样本作为异常样本.

2　融合高斯分布和自步学习的多维数据异常检测

2.1　异常检测模型建立

刀具的性能状态需通过加工监测信号间接反映，并且随着刀具的磨损程度增加刀具的剩余寿命会降低[17].

首先针对刀具监测信号无异常数据标签问题，将监测信号和刀具剩余寿命进行关联，建立多层感知机模型对两者进行拟合，以监测信号为输入，以刀具剩余寿命为输出. 加工过程的异常数据受外界干扰与真实数据存在差异，无法反映刀具加工过程的真实状态，剩余寿命的预测误差会大于正常数据. 其次，采用全量样本训练多层感知机模型会引入异常样本，因此，在多层感知机模型每一步的训练迭代过程中，引入自步学习框架，选择预测误差小的样本来更新模型权重，防止异常样本的干扰. 针对自步学习的步长大小难以确定的问题，提出利用高斯分布来设置误差阈值作为自步学习的步长. 最后，模型收敛后，利用最后一次计算的误差阈值以及更新完成的多层感知机模型来对所有样本进行筛选. 图 1 展示了异常检测方法的流程.

多层感知机的实现包含输入层、两个隐藏层以及输出层，每一层之间通过激活函数相连. 其中第 1 个隐藏层、第 2 个隐藏层的输出通过 ReLU 激活函数连接，输出层后连接 Sigmoid 激活函数，最终输出刀具的剩余寿命值. 为了防止模型过拟合，在第 1 个隐藏层后连接一个 LayerNorm 正则化层. 剩余寿命预测模型各层结构如图 2 所示. 模型的损失函数为均方差函数：

$$\mathrm{MSE} = \frac{1}{m}\sum_{i=0}^{m-1}(\hat{y}_i - y_i)^2 \quad (5)$$

式中：y_i、$\hat{y}_i$、$\bar{y}$ 分别为真实的磨损、预测的磨损、真实磨损的平均值.

模型迭代过程中主要包括两个部分：

（1）误差阈值计算. 将所有样本通过异常检测模型预测的剩余寿命与真实值比较，基于高斯分布拟合误差分布，进而获取误差阈值，作为自步学习的步长参数.

（2）自适应训练. 使用误差阈值计算得到的结果代入自步学习的 λ 参数中，建立自步学习损失函数，通过梯度下降算法对模型进行权重更新，学习正常样本的数据特征.

最后当训练迭代结束时，以最后一次迭代得到的误差阈值为标准，将样本划分为正常样本和异常样本.

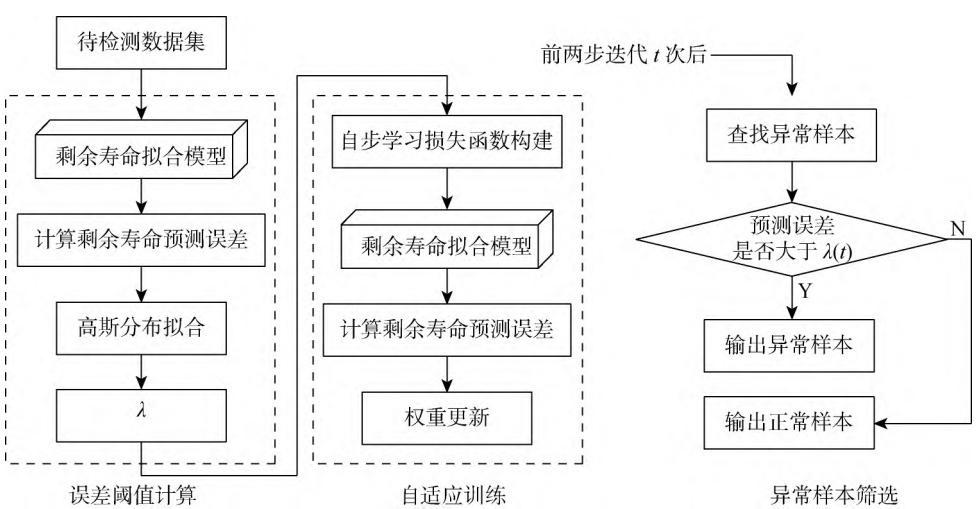

图 1　异常检测模型迭代流程图

Fig. 1　Iterative flow chart of anomaly detection model

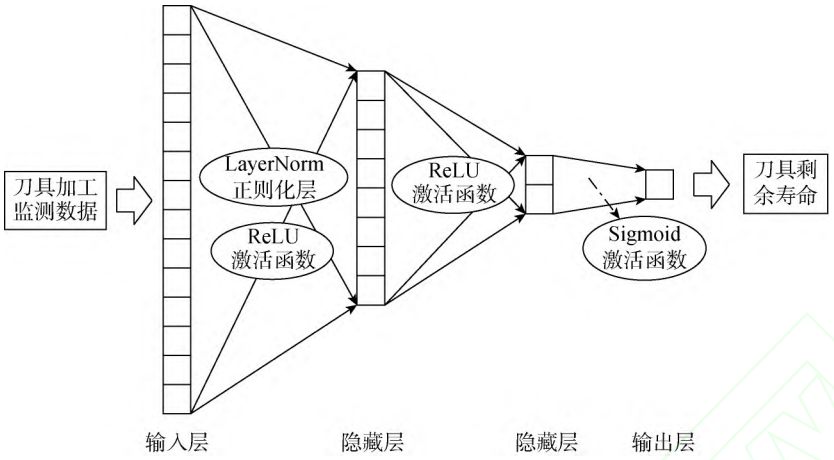

图 2　剩余寿命预测网络模型结构

Fig. 2　Residual life prediction network model structure

2.2　模型训练过程

2.2.1　基于高斯分布的误差阈值计算　剩余寿命预测模型回归加工监测数据的特征，通过式(5)计算得到的预测值和真实值的误差的平方数来衡量回归的精度，理想情况下加工监测数据经过剩余寿命预测模型计算后得到的误差为 0，但数据样本的质量问题使得模型的始终存在一定的误差. 误差越大发生的概率越小，误差越小发生的概率越大，且对于大部分正常样本其误差维持在一个较小的水平. 选择高斯分布来映射预测误差平方和数据的准确率，高斯分布定义随机变量 X 服从一个数学期望为 u、方差为 σ 的正态分布，记为 $N(u,\sigma^2)$. 高斯分布概率密度函数 $p(x)$ 为

$$p(x)=\frac{1}{\sqrt{2\Pi}\sigma}\mathrm{e}^{-\frac{(x-u)^2}{2\sigma^2}} \tag{6}$$

误差平方越接近 0 时数据的分散程度越小，映射后的数据正常的概率越大. 为了满足映射关系取高斯分布概率密度函数对称轴的右半轴为实际的映射对象. 其中期望大小为 0，标准差通过样本标准差 s 来估计：

$$\left.\begin{aligned} s&=\sqrt{\frac{1}{m-1}\sum_{i=0}^{m-1}(e_i-\bar{e})^2}\\ k&=3s \end{aligned}\right\} \tag{7}$$

式中：e_i 为第 i 个样本预测值和真实值的平方误差；$\bar{e}$ 为所有样本预测值和真实值的平方误差的均值；k 为误差阈值. 针对以上误差分布规律，常用 3σ 统计量来进行粗差探测[18-19]，即取 $3s$ 作为误差阈值 k，用作后续自步学习的步长参数 λ.

2.2.2　基于自步学习的自适应训练方法　建立其优化函数如下：

$$w_t,\ v_t=\arg\min\Big(\sum_{i=0}^{m-1}v_i f(x_i,\ y_i;\ w_{t-1})-\lambda\sum_{i=0}^{m-1}v_i\Big) \tag{8}$$

自步学习的步长大小记为 λ，等于式(7)计算得到的 k，表示样本难易程度的筛选阈值. 训练过程主要分为两步，第 1 步是 v 值的计算，第 2 步是 w 值的更新. v 值的计算需要先固定 w 权重值，通过对比预测误差和 λ 的关系进行确定，当样本通过剩余寿命预测模型后的得到误差小于 λ 时，上述的上述优化函数取 $v=1$ 时达到最小. 最终可以得到每个数据样本对应的 v 值. $v=0$ 则该样本不参与训练，$v=1$ 则该样本参与训练：

$$\left.\begin{aligned} v&=0,\quad f(x_i,\ y_i;\ w_{t-1})\geqslant\lambda\\ v&=1,\quad f(x_i,\ y_i;\ w_{t-1})<\lambda \end{aligned}\right\} \tag{9}$$

同理，w 权重的更新过程需要固定 v 值，采用梯度下降法进行更新：

$$w_t=w_{t-1}-\alpha\frac{\partial\sum_{i=0}^{m-1}v_i f(x_i,\ y_i;\ w_{t-1})}{\partial w_{t-1}} \tag{10}$$

当完成 t 次迭代训练后，取 λ_t 作为误差筛选阈值 k，将所有样本的预测误差与 k 进行对比，大于等于 k 的样本归为异常样本. 结合上述融合高斯分布的误差阈值的自步学习迭代过程得到算法如下.

算法　基于自步学习的刀具加工过程监测数据异常检测算法.

输入　待检测的样本集 $D(x_i,y_i)$，$1\leqslant i\leqslant m$ 训练迭代次数 epochs，学习率 α，样本批次大小 batch-size，多层感知机模型 $M(w,b)$，样本权重向量 V.

输出　多层感知机模型 $M(w,b)$ 和误差阈值 k.

步骤 1　随机初始化网络权重 w,b，样本权重 $V=I_m$；

步骤 2　foreach t in epochs do；

步骤 3　建立误差集合 $E(e_i)$，$1 \leqslant i \leqslant m$；

步骤 4　foreach (x_i, y_i) in D do；

步骤 5　多层感知机模型前馈计算，得到预测值 $\hat{y}_i$；

步骤 6　计算误差 e_i，加入 E：$e_i(\hat{y}_i - y_i)^2$；

步骤 7　end；

步骤 8　根据高斯分布计算误差阈值：

$$k = \lambda = 3\sqrt{\frac{1}{m-1}\sum_{i=0}^{m-1}(e_i - \bar{e})^2}\ ;$$

步骤 9　foreach batch(x_i, y_i) in D do；

步骤 10　多层感知机模型前馈计算，得到预测值 btach$(\hat{y}_i)$；

步骤 11　计算误差 batch(e_i)；

步骤 12　① 固定模型权重，更新样本权重 $V_{\text{batchsize}}$；

步骤 13　if $e_i < \lambda$ then；

步骤 14　$v_i = 1$；

步骤 15　else；

步骤 16　$v_i = 0$；

步骤 17　end；

步骤 18　② 固定样本权重，更新模型权重

$$w_t = w_{t-1} - \alpha \frac{\partial \sum_{i=0}^{\text{batchsize}-1} v_i e_i}{\partial w_{t-1}}\ ;$$

步骤 19　end；

步骤 20　end；

步骤 21　获得异常样本的误差筛选阈值：$k = \lambda_{t=\text{epochs}}$。

3　实验及其结果分析

3.1　数据集介绍

本文选用数据采集自某汽轮机厂的汽轮机转子轮槽铣削加工过程，一共使用了 15 把 J1 型精铣刀. 在轮槽加工过程中，采用 PCI-2AE 采集声发射信号，采样的频率为 1 MHz.

上述加工过程产生的样本数据集一共包含了 170 条轮槽的加工监控数据，每条轮槽的加工持续时间内能得到 10 000 条数据记录，每条数据记录涵盖 14 种 AE 属性信息：上升时间、计数、能量、幅值、平均频率、均方根值、平均信号电平、峰值频率、反算频率、初始频率、信号强度、绝对能量、中心频率、峰频. 将每条轮槽的 10 000 条 AE 数据进行均值化后得到 14 个维度的特征向量，每个特征向量作为一个数据样本，一共包含 170 个数据样本，对应于 170 条轮槽. AE 信号能表征精刀加工过程的健康情况，反映刀具的剩余使用寿命，因此，每个数据样本均会对应一个刀具剩余使用寿命，与加工刀具的实际加工情况有关，如表 1 所示.

表 1　汽轮机轮槽铣削加工刀具使用情况

Tab. 1　The use of cutting tools in turbine groove milling

铣刀编号	轮槽编号	铣刀编号	轮槽编号	铣刀编号	轮槽编号
1	1～13	6	65～75	11	119～128
2	14～26	7	76～86	12	129～139
3	27～38	8	87～96	13	140～149
4	39～50	9	97～107	14	150～159
5	51～64	10	108～118	15	160～170

汽轮机转子轮槽产品成本高、质量要求严格，加工过程复杂. 刀具剩余寿命预测影响换刀决策时间，对于保证加工质量和生产效率具有重要意义[20]. 本文使用采用高斯径向基核的支持向量回归机算法作为验证算法，该算法能够通过核函数实现高维空间的非线性映射，同时具有较好的鲁棒性，适用于加工过程的刀具剩余寿命预测.

将编号为 13、14、15 号精铣刀作为测试集，编号为 1～12 号的刀具作为训练集，对训练集进行异常样本的检测和剔除，然后参与刀具剩余寿命预测模型的训练，最后通过测试集的预测结果对比数据异常检测前后效果.

3.2　实验结果评价指标介绍

汽轮机轮槽铣刀监测数据在实际生产加工中采集得到，没有区分样本正常和异常的标签值，但每个样本都具有相应的刀具剩余寿命. 因此通过对比异常检测前后，支持向量回归机在测试集上刀具剩余寿命预测的表现来反映所提出的数据异常检测算法的有效性. 平均绝对误差（MAE）、均方根误差（RMSE）被广泛用于回归问题中，本文采用上述两种指标表示模型预测效果的优劣.

（1）平均绝对误差反映实际预测误差的大小. 其计算公式如下：

$$\text{MAE} = \frac{1}{m}\sum_{i=0}^{m-1} |\ \hat{y}_i - y_i\ | \tag{11}$$

（2）均方根误差的作用是衡量预测值与真实值之间的偏差大小. 其计算公式如下：

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=0}^{m-1}(\hat{y}_i - y_i)^2} \tag{12}$$

3.3　实验分析

首先建立第 2 章的多层感知机模型，输入层大小为 14，对应样本的 14 个维度. 第 1 层隐藏层大小为 8，第 2 层隐藏层大小为 2，输出层大小为 1，对应

刀具的剩余寿命值. 然后对 15 把刀中编号为 1～12 的铣刀加工监测数据样本进行训练，训练轮次为 500 次. 由于模型权重参数是随机初始化的，因此选取了 5 次不同的随机种子下异常检测模型训练的结果.

在异常检测模型剔除异常数据样本后，将剩余正常样本用于训练支持向量回归机，其中 80%作为训练集，20%作为验证集，进行刀具的剩余寿命预测回归. 采用网格搜索法搜索支持向量回归机的最佳参数，其中惩罚系数 C 的取值范围为 [0.001：0.001：0.01，0.01：0.01：0.1，0.1：0.1：1：1，1：1：10]，γ 取值范围为 [0.001：0.001：0.01，0.01：0.01：0.1，0.1：0.1：1：1]. 最后将测试样本的预测结果进行对比，得到不同异常检测算法的性能.

异常检测算法将多层感知机作为刀具剩余寿命预测模型，预测误差结果用高斯分布进行拟合，并应用 3σ 方法进行粗差探测，融合自步学习框架对多层感知机模型进行逐步训练调整. 为了验证式(7)计算的阈值 k 的有效性，改写式(7)为 $k=\beta s$，分别取 $\beta=1.5, 2.0, 2.5, 3.0, 3.5$ 作为系数，对刀具训练集数据进行异常检测，得到的测试集预测结果如表 2 所示. 图 3 展示了在不同的系数下，测试集的

表 2　不同系数下的测试结果

Tab. 2　Test results under different coefficients

实验编号	$\beta=1.5$		$\beta=2.0$		$\beta=2.5$		$\beta=3.0$		$\beta=3.5$	
	MAE	RSME	MAE	RSME	MAE	RSME	MAE	RSME	MAE	RSME
1	1.386	1.608	1.138	1.334	1.169	1.417	1.013	1.223	1.315	1.572
2	1.125	1.303	1.286	1.533	1.214	1.391	1.122	1.365	1.424	1.630
3	1.170	1.410	1.333	1.513	1.431	1.738	1.122	1.365	1.424	1.630
4	1.170	1.410	1.214	1.391	1.431	1.738	1.122	1.365	1.459	1.674
5	1.232	1.388	1.305	1.679	1.214	1.391	0.965	1.202	1.424	1.630
平均值	1.216 6	1.423 8	1.255 2	1.490 0	1.291 8	1.535 0	1.068 9	1.304 0	1.409 2	1.627 2

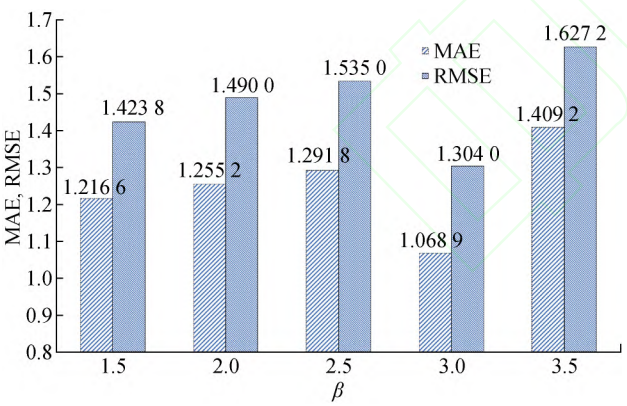

图 3　不同系数 β 下的测试平均结果

Fig. 3　Average test results at different coefficients

MAE 和 RMSE 平均结果.

从实验结果可以看出，不同系数的取值会对测试集最终预测结果产生影响. 其中在 $\beta=3.0$ 时，经过异常样本检测和剔除后测试集预测结果最佳，验证了 3σ 方法的有效性. 表 3 显示在最佳系数取值下，5 组实验检测出的异常样本，在剔除异常样本之后，剩余样本在测试集上获得的平均测试结果 MAE 值为 1.069，RMSE 值为 1.304.

取实验编号 5 得到训练过程中 λ、模型预测平均损失值(MSE)、m 随 t 变化的曲线如图 4 所示.

表 3　多随机种子下异常检测方法结果

Tab. 3　Results of anomaly detection method under multiple random seeds

实验编号	MAE	RMSE	异常样本数量/个	异常样本编号
1	1.013	1.223	3	10、43、107
2	1.122	1.365	3	9、10、11
3	1.122	1.365	3	9、10、11
4	1.122	1.365	3	9、10、11
5	0.965	1.202	3	10、11、12
平均值	1.069	1.304		
标准差	0.067	0.075		
优化比/%	26.284	28.194		

在训练过程中，λ 始终大于 MSE，其中 e 超过 λ 的样本即会被归为异常样本，不参与此轮模型权重的训练. 在训练初始阶段，模型对于训练样本的拟合程度较差，因此，训练样本的 MSE 较大，计算得到的 λ 也较大. 随着训练次数的增加，模型渐渐收敛，大部分样本都能够被模型进行拟合，此时，仍然 $e>\lambda$ 的样本数量渐渐稳定. 最终选取训练结束后，将 e 与此时 λ 进行比较来确定异常样本.

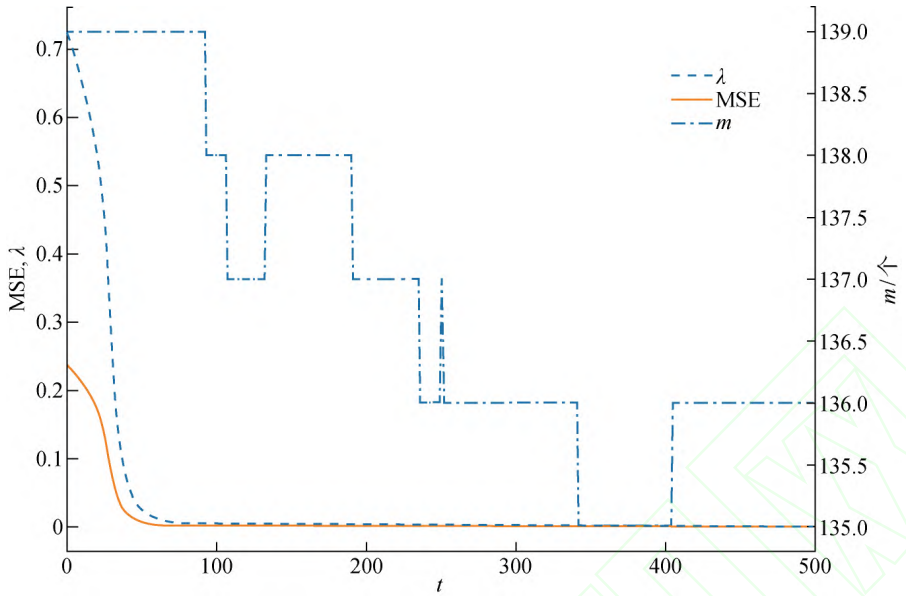

图 4　异常检测方法的训练过程

Fig. 4　Training process of anomaly detection method

为了对照所提出的异常检测模型的有效性，选取了其他 5 种异常检测算法. 局部异常因子算法(LOF)是一种基于密度的异常检测算法，通过计算数据样本周围样本的局部密度以及自身的所在位置的密度来衡量样本的异常程度. DBSCAN 是一种基于密度的聚类算法，可以在数据空间中发现噪声点. 孤立森林根据数据点被孤立的难易程度将噪声点和正常点进行区分. K 均值法是应用广泛的基于距离的聚类算法. 一分类支持向量机(One-Class SVM)与本文方法均为基于分类的异常检测方法[14]，其利用核函数将数据样本拟合到一个超平面上，远离超平面上的样本为异常样本. 此外增加了随机抽样方法，随机选取 3 个样本作为异常样本，其余作为正常样本，与我们的方法进行对比. 不同的方法对比结果如表 4 所示. 分别从 4 个指标维度进行比较，分别是测试集的 MAE、RMSE、MAE 相较于全量训练集得到的提升比例(MAE')、RMSE 相较于全量训练得到的提升比例(RMSE').

本文方法获得的测试集预测结果平均值在 MAE 上的提升比例为 26.28%，在 RMSE 上提升比例为 28.19%，优于其他方法的结果. 在其他的异常检测算法中，LOF 取得了次优的结果，在 MAE 和 RMSE 上的提升比例分别为 15.93% 和 18.17%，LOF 能通过衡量样本间的聚集程度来判断离群点，鲁棒性较好. 孤立森林的提升比例与 LOF 方法近似，在 MAE 和 RMSE 上的提升比例分别为 12.90%和 15.91%. One-Class SVM 在 MAE 和 RMSE 上的提升比例分别为 11.72% 和 11.62%，差于前 3 种方法. 两种基于聚类的方法得到的 MAE 提升百分比和 RMSE 提升比例均为负数. 聚类算法受限于聚类结果，而刀具加工过程监测数据受刀具性能衰退而变化，因此基于聚类的异常检测方法难以准确区分正常和异常样本. 随机采样法是样本选择中一个比较常见的方法，该方法便于操作，但不同的随机采样过程得到的结果差异很大. 从结果表中看到，随机采样的 MAE 和 RMSE 的标准差分别达到了 0.327 和 0.456. 图 5 直观地显示了各个算法在测试集上指标的提升比例.

表 4　不同异常检测算法对比结果

Tab. 4　Comparison results of different anomaly detection algorithms

方法	MAE	RMSE	MAE'/%	RMSE'/%
未筛选	1.45	1.816	—	—
本文方法	1.069±0.067	1.304±0.075	26.28	28.19
LOF	1.219	1.486	15.93	18.17
DBSCAN	1.693	2.137	−16.76	−17.68
孤立森林	1.263	1.527	12.90	15.91
K 均值	1.844	2.215	−27.17	−21.97
One-Class SVM	1.28	1.605	11.72	11.62
随机采样	1.485±0.327	1.877±0.456	−2.41	−3.37

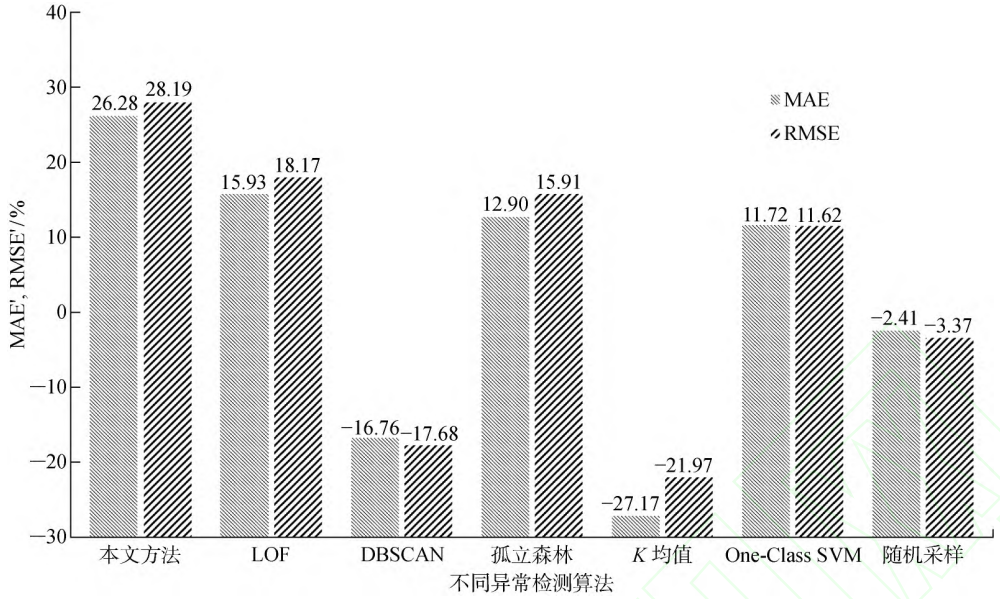

图 5　测试集预测结果在不同异常检测算法应用前后提升比例

Fig. 5　The optimized percentage of test set prediction results before and after application of different anomaly detection algorithms

4　结语

本文面向刀具加工过程监测，基于高斯分布和自步学习框架提出一种数据异常检测方法. 利用多层感知机模型学习监测数据样本的整体特征，分离监测数据集中的异常数据. 在模型的更新过程中，用模型预测误差来拟合高斯分布，并设定误差阈值，同时结合自步学习框架，优选高质量数据样本参与权重的更新. 最终训练结束后通过多层感知机模型的预测误差有效检测异常数据样本.

通过与多种异常检测算法的对比实验可见，融合高斯分布和自步学习框架的数据异常检测方法能够有效地区分加工监测数据中的异常样本. 多层感知机模型通过高斯分布计算样本误差阈值，与自步学习框架结合，针对性地选取样本对模型权重进行更新，保证模型具备对异常样本的判别能力. 综合上述分析和实验结论，本文所提出的面向刀具加工监测数据的异常检测方法相比其他方法能更适用于刀具加工监测数据样本的异常检测.

参考文献：

[1] 孙宇. 电力设备监测数据处理和数据库设计[D]. 浙江：浙江大学，2022.
SUN Yu. Data processing and database design of power equipment monitoring [D]. Zhejiang: Zhejiang University, 2022.

[2] 尚文利，石贺，赵剑明，等. 基于 SAE-LSTM 的工艺数据异常检测方法[J]. 电子学报，2021，49(08)：1561-1568.
SHANG Wenli, SHI He, ZHAO Jianming, *et al*. An anomaly detection method of process data based on SAE-LSTM[J]. **Acta Electronica Sinica**, 2021, 49(08): 1561-1568.

[3] 夏英，韩星雨. 融合统计方法和双向卷积 LSTM 的多维时序数据异常检测[J]. 计算机应用研究，2022，39(05)：1362-1367.
XIA Ying, HAN Xingyu. Multi-dimensional time series data anomaly detection fusing statistical methods and bidirectional convolutional LSTM[J]. **Application Research of Computers**, 2022, 39(05): 1362-1367.

[4] 孙滢涛，张锋明，陈水标，等. 基于多域特征提取的电力数据异常检测方法[J]. 电力系统及其自动化学报，2022，34(06)：105-113.
SUN Yingtao, ZHANG Fengming, CHEN Shuibiao, *et al*. Power data anomaly detection algorithm based on multi-domain feature extraction[J]. **Proceedings of the CSU-EPSA**, 2022, 34(06): 105-113.

[5] 傅世元，高欣，张浩，等. 基于元学习动态选择集成的电力调度数据异常检测方法[J]. 电网技术，2022，46(08)：3248-3261.
FU Shiyuan, GAO Xin, ZHANG Hao, *et al*. Anomaly detection for power dispatching data based on meta-learning dynamic ensemble selection [J]. **Power System Technology**, 2022, 46(08): 3248-3261.

[6] 刘鑫. 无监督异常检测方法研究及其应用[D]. 成都：电子科技大学，2018.
LIU Xin. Research on unsupervised anomaly detec-

tion algorithm and application[D]. Chengdu: University of Electronic Science and Technology of China, 2018.

[7] DU H, ZHAO S, ZHANG D, *et al*. Novel clustering-based approach for local outlier detection[C]// **International Conference on Computer Communications Workshops**. San Francisco, CA, USA: IEEE, 2016: 802-811.

[8] **吴蕊, 张安勤, 田秀霞, 等. 基于改进 *K*-means 的电力数据异常检测算法[J]. 华东师范大学学报(自然科学版)**, 2020(04): 79-87.
WU Rui, ZHANG Anqin, TIAN Xiuxia, *et al*. Anomaly detection algorithm based on improved *K*-means for electric power data[J]. **Journal of East China Normal University (Natural Science)**, 2020(04): 79-87.

[9] **吴金娥, 王若愚, 段倩倩, 等. 基于反向 k 近邻过滤异常的群数据异常检测[J]. 上海交通大学学报**, 2021, 55(05): 598-606.
WU Jine, WANG Ruoyu, DUAN Qianqian, *et al*. Collective data anomaly detection based on reverse *k*-nearest neighbor filtering[J]. **Journal of Shanghai Jiao Tong University**, 2021, 55(05): 598-606.

[10] **陈砚桥, 孙彤, 张侨禹. 基于 DBSCAN 的智能机舱多源数据异常检测方法[J]. 舰船科学技术**, 2021, 43(17): 156-160.
CHEN Yanqiao, SUN Tong, ZHANG Qiaoyu. Intelligent engine room multi-source data detecting method based on DBSCAN cluster algorithm[J]. **Ship Science and Technology**, 2021, 43(17): 156-160.

[11] **宋丽娜, 刘淼, 秦韬, 等. 基于 LOF 与 CEEMD 的城镇取用水监测数据异常值识别[J]. 水利信息化**, 2022(02): 33-40.
SONG Lina, LIU Miao, QIN Tao, *et al*. Outlier identification of urban water intake monitoring data based on LOF and CEEMD[J]. **Water Resources Informatization**, 2022(02): 33-40.

[12] **王锋, 高欣, 贾欣, 等. 一种基于对数区间隔离森林的电力调度数据异常检测集成算法[J]. 电网技术**, 2021, 45(12): 4818-4827.
WANG Feng, GAO Xin, JIA Xin, *et al*. An anomaly detection ensemble algorithm for power dispatching data based on log-interval isolation[J]. **Power System Technology**, 2021, 45(12): 4818-4827.

[13] **王燕晋, 易忠林, 郑思达, 等. 基于孤立森林算法的电力用户数据异常快速识别研究[J]. 电子设计工程**, 2022, 30(03): 11-14.
WANG Yanjin, YI Zhonglin, ZHENG Sida, *et al*. Research on fast identification of power user data abnormal based on isolation forest algorithm[J]. **Electronic Design Engineering**, 2022, 30(03): 11-14.

[14] **卓琳, 赵厚宇, 詹思延. 异常检测方法及其应用综述[J]. 计算机应用研究**, 2020, 37(S1): 9-15.
ZHUO Lin, ZHAO Houyu, ZHAN Siyan. Anomaly detection and its application[J]. **Application Research of Computers**, 2020, 37(S1): 9-15.

[15] BENGIO Y, LOURADOUR J, COLLOBERT R, *et al*. Curriculum learning[C]// **Proceedings of the 26th Annual International Conference on Machine Learning**. New York, NY, USA: Association for Computing Machinery, 2009: 41-48.

[16] KUMAR M P, PACKER B, KOLLER D. Self-paced learning for latent variable models[C]// **Proceedings of the 23rd International Conference on Neural Information Processing Systems-Volume 1**. Red Hook, NY, USA: Curran Associates Inc., 2010: 1189-1197.

[17] **王艺玮, 邓蕾, 郑联语, 等. 基于多通道融合及贝叶斯理论的刀具剩余寿命预测方法[J]. 机械工程学报**, 2021, 57(13): 214-224.
WANG Yiwei, DENG Lei, ZHENG Lianyu, *et al*. A multi-channel signal fusion and Bayesian theory based method for tool remaining useful life prediction [J]. **Journal of Mechanical Engineering**, 2021, 57(13): 214-224.

[18] **吴浩, 卢楠, 邹进贵, 等.** GNSS **变形监测时间序列的改进型** 3σ **粗差探测方法[J]. 武汉大学学报(信息科学版)**, 2019, 44(09): 1282-1288.
WU Hao, LU Nan, ZOU Jingui, *et al*. An improved 3σ gross error detection method for GNSS deformation monitoring time series[J]. **Geomatics and Information Science of Wuhan University**, 2019, 44(09): 1282-1288.

[19] **徐洪钟, 吴中如, 李雪红, 等. 基于小波分析的大坝变形观测数据的趋势分量提取[J]. 武汉大学学报(工学版)**, 2003(06): 5-8.
XU Hongzhong, WU Zhongru, LI Xuehong, *et al*. Abstracting trend component of dam observation data based on wavelet analysis[J]. **Engineering Journal of Wuhan University**, 2003(06): 5-8.

[20] **党英, 吉卫喜, 陆家辉, 等. 基于深度学习的铣刀剩余寿命预测方法研究[J]. 现代制造工程**, 2021(12): 79-87.
DANG Ying, JI Weixi, LU Jiahui, *et al*. Research on prediction method of remaining useful life of milling cutter based on deep learning[J]. **Modern Manufacturing Engineering**, 2021(12): 79-87.

(本文编辑:王一凡)