

基于双近邻标准化的局部离群因子故障检测方法

冯立伟¹, 郭少锋^{2,3}, 吴弋飞^{2,3}, 邢宇^{2,3}, 李元⁴

(1. 沈阳化工大学理学院, 辽宁 沈阳 110142; 2. 沈阳化工大学计算机科学与技术学院, 辽宁 沈阳 110142;

3. 沈阳化工大学辽宁省化工过程工业智能化技术重点实验室, 辽宁 沈阳 110142;

4. 沈阳化工大学信息工程学院, 辽宁 沈阳 110142)

摘要:为实现对非线性多模态过程的更有效监测,本研究提出一种基于双近邻标准化的局部离群因子故障检测方法(DLNS-LOF)。该方法首先通过双近邻标准化(DLNS)技术,利用嵌套近邻集的均值和标准差对每个样本进行标准化处理,然后再利用局部离群因子(LOF)方法对过程数据进行故障检测。DLNS能够将非线性多模态过程数据转化为近似服从标准正态分布的单模态数据,并有效分离在线故障样本,从而显著提升LOF方法的故障检测性能。通过数值模拟与青霉素发酵过程的实验验证表明,DLNS-LOF方法具有良好的检测性能。

关键词:双近邻标准化;局部离群因子;多模态;非线性;故障检测

中图分类号:TP277

文献标志码:A

Fault detection based on double local neighborhood standardization and local outlier factor

FENG Liwei¹, GUO Shaofeng^{2,3}, WU Yifei^{2,3}, XING Yu^{2,3}, LI Yuan⁴

(1. College of Science, Shenyang University of Chemical Technology, Shenyang 110142, China;

2. College of Computer Science and Technology, Shenyang University of Chemical Technology, Shenyang 110142, China;

3. Key Laboratory of Intelligent Technology of Chemical Process Industry in Liaoning Province,

Shenyang University of Chemical Technology, Shenyang 110142, China;

4. College of Information Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China)

Abstract: To achieve effective monitoring of nonlinear multimodal processes, this paper proposed a fault detection method based on the double local neighborhood standardization and local outlier factor (DLNS-LOF). Firstly, the DLNS technology was adopted to standardize each sample by using the mean value and standard deviation of the embedded local neighborhood set. Secondly, the LOF method was used to detect faults in the process data. DLNS can transform multimodal process data with nonlinear characteristics into single-mode data that approximately obeys a standard normal distribution and separates online fault samples, significantly improving the fault detection capability of the LOF method. The experiments on a numerical simulation process and the penicillin fermentation process show that the DLNS-LOF method has a good performance.

Key words: double local neighborhood standardization; local outlier factor; multimodal; nonlinear; fault detection

尽早发现工业生产过程中的异常状况对保障系统的安全稳定运行具有关键作用。近年来,多元统计过程监测(multivariate statistical process monitoring, MSPM)技术作为一种数据驱动类的故障检测方法,因其高效性和强适应性得到广泛应用^[1-3]。其中,主成分分析(principal component analysis, PCA)和偏最小

收稿日期:2024-03-18

基金项目:辽宁省教育厅基础科研项目(LJKMZ20220792; LNYJG2022177)

作者简介:冯立伟(1980—),男,辽宁沈阳人,讲师,硕士,主要从事数据驱动的故障诊断方面的研究。

李元(1964—),女,辽宁沈阳人,教授,博士,主要从事过程控制、数据驱动的故障诊断方面的研究,本文通信作者。

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

二乘(partial least squares, PLS)是两种经典故障检测方法,分别利用霍特林 T^2 和平方预测误差(squared prediction error, SPE)统计量对主元空间和残差空间进行异常监测。

实际工业生产过程常呈现非线性与多模态特性,这也使得故障检测更为困难。PCA 与 PLS 方法在故障检测中构建 T^2 和 SPE 统计量时,均假设过程数据服从单模态多元高斯分布,但若实际数据偏离该假设,其检测性能将受到限制。为此,研究人员提出多种改进策略。针对非线性问题,一些基于核函数的扩展方法被提出,如核主成分分析法^[4]、关键性能指标式核偏最小二乘法^[5]和高效核偏最小二乘法^[6]等,均借助核映射将原始非线性数据投影到高维特征空间,以实现线性可分,但该方法未考虑多模态情形。多模态特征是指生产过程在不同运行区域下产生具有不同分布特性的数据,常与非线性并存。为此,He 等^[7]提出基于 k 近邻(k nearest neighbor, k NN)的故障检测方法,但该方法对疏密程度具有显著差异的多模态过程进行故障检测时存在局限性。冯立伟等^[8]进一步提出基于标准距离 k 近邻(standard distance k nearest neighbor, SD- k NN)的故障检测方法,通过构建标准距离统一不同模态样本间近邻距离的量纲,克服疏密程度的影响。Breunig 等^[9]提出基于局部离群因子(local outlier factor, LOF)的故障检测方法,Ma 等^[10]在其基础上提出基于马氏距离局部离群因子的故障检测方法。LOF 通过利用待测样本的近邻样本与其自身的相对密度构建统计量,能够对非线性的多模态过程进行故障检测。但上述 k NN 和 LOF 的扩展方法在处理微弱型故障时敏感度不高。李元等^[11]提出基于局部近邻标准化的 k 近邻(local neighborhood standardization and k nearest neighbor, LNS- k NN)故障检测方法,通过 LNS 技术预处理数据以削弱非线性和多模态特征,并分离故障样本。但 LNS 技术在处理部分位于两个模态之间的故障样本时,由于近邻样本可能来自不同模态,因此故障样本可能位于其近邻样本的中心位置,经 LNS 处理后会该部分故障样本混入正常样本中,影响检测准确性。

为实现对非线性多模态过程更有效地监测,本研究提出一种基于双近邻标准化的局部离群因子(double local neighborhood standardization and local outlier factor, DLNS-LOF)故障检测方法。DLNS 技术通过嵌套近邻集的均值与标准差对每个样本进行标准化,避免了 LNS 所用的近邻样本可能来自不同模态的问题,从而有效防止模态交界处故障样本在标准化后被误判为正常样本的现象。此外,经 DLNS 处理后正常样本聚集于原点附近,而故障样本则远离原点。LOF 方法通过计算待测样本近邻的平均局部密度与其自身密度之比构建统计量。正常样本因分布集中,其相对密度值在 1 附近小幅波动;故障样本偏离正常集群,其相对密度值显著大于 1。因此,结合 DLNS 预处理后,LOF 能更有效识别故障样本。

1 局部离群因子

设有训练集 $X \in \mathbf{R}^{l \times z}$, l 和 z 分别表示样本和变量个数。下面以样本 u 为例对 LOF 方法的计算步骤进行说明。

首先,在 X 中计算 u 的前 k 个近邻样本 u_f , $f=1, 2, \dots, k$, 并求出 u 到其每个近邻样本的可达距离 $D_k(u, u_f)$:

$$D_k(u, u_f) = \max(d_k(u_f), d(u, u_f)). \quad (1)$$

式中: $d_k(u_f)$ 是 u_f 到其第 k 近邻样本的距离,也称 u_f 的第 k 距离; $d(u, u_f)$ 表示 u 到 u_f 的欧式距离。

然后,计算 u 的局部可达密度:

$$l_{rd}(u) = \frac{k}{\sum_{f=1}^k D_k(u, u_f)}. \quad (2)$$

最后,计算 u 的局部离群因子:

$$L_{OF}(u) = \frac{1}{k} \sum_{f=1}^k \frac{l_{rd}(u_f)}{l_{rd}(u)}. \quad (3)$$

局部离群因子反映了近邻样本密度的均值与待测样本自身密度的比值。若待测样本是正常样本,其 LOF 值在 1 附近小幅波动;若为故障样本,其离群程度越大,LOF 的值越大。通过核密度估计技术可确定

LOF 的控制限,进而依据在线样本是否超出控制限判断过程是否发生故障。

2 双近邻标准化局部离群因子的故障检测

2.1 双近邻标准化

DLNS 是一种在 LNS 基础上的改进方法^[12]。LNS 方法利用待测样本在训练集中近邻集的均值和标准差对自身进行标准化。然而,当故障样本位于两个模态之间时,其在训练集中的近邻样本可能来自两个模态,且近邻集的均值可能近似等于该故障样本的坐标,如图 1 所示。此时,标准化过程中的去均值操作会使故障样本被移至原点附近,从而混入正常样本中,影响故障检测结果。DLNS 采用双层近邻结构,首先确定待测样本的近邻样本,接着利用每个近邻样本的近邻集的均值和标准差标准化自身,从而避免了标准化自身的近邻样本来自不同模态的现象。最后,DLNS 对每个嵌套近邻集标准化后的结果进行平均,以消除统计波动。

下面以样本 u 为例对 DLNS 方法的计算步骤进行说明。

首先,计算 u 在训练集中的前 k 个近邻样本集 $N(u)$,并进一步计算每个近邻样本 u_f 的前 K 个近邻样本集 $N(u_f)$:

$$N(u) = \{u_1, u_2, \dots, u_k\}, \quad (4)$$

$$N(u_f) = \{u_f^1, u_f^2, \dots, u_f^K\}. \quad (5)$$

式中: u_f 表示 u 的第 f 近邻样本, $f=1, 2, \dots, k$; u_f^j 表示 u_f 的第 j 近邻样本, $j=1, 2, \dots, K$ 。需要注意的是,参数 k 和 K 的选取均不宜过大。当参数最大为样本个数时,DLNS 就退化为了 Z-score 标准化,因此应选取较小的参数值。

然后,求出 K 个近邻样本集 $N(u_f)$ 的均值 $m(N(u_f))$ 和标准差 $s(N(u_f))$:

$$m(N(u_f)) = \frac{1}{K} \sum_{j=1}^K u_f^j, \quad (6)$$

$$s(N(u_f)) = \sqrt{\sum_{j=1}^K \frac{(u_f^j - m(N(u_f)))^2}{K-1}}. \quad (7)$$

最后,使用 k 个均值和标准差对样本 u 进行标准化:

$$\bar{u} = \frac{1}{k} \sum_{f=1}^k \frac{u - m(N(u_f))}{s(N(u_f))}. \quad (8)$$

2.2 双近邻标准化局部离群因子的故障检测步骤

假设已有正常状态下的训练集和引入故障的测试集,DLNS-LOF 的故障检测步骤如下。

1) 离线建模阶段:首先,计算每个训练样本的嵌套空间近邻集,获取均值和标准差,进而对每个训练样本进行双近邻标准化(DLNS);然后,通过式(1)~(3)计算每个训练样本的局部离群因子;最后,使用核密度估计技术确定局部离群因子的控制限。

2) 在线检测阶段:首先,对每个测试样本,基于其近邻信息进行双近邻标准化;然后,计算每个测试样本的局部离群因子;最后,用测试样本的局部离群因子(LOF)与离线建模阶段获取的控制限进行比较,若超出控制限则判定为故障样本,否则为正常样本。

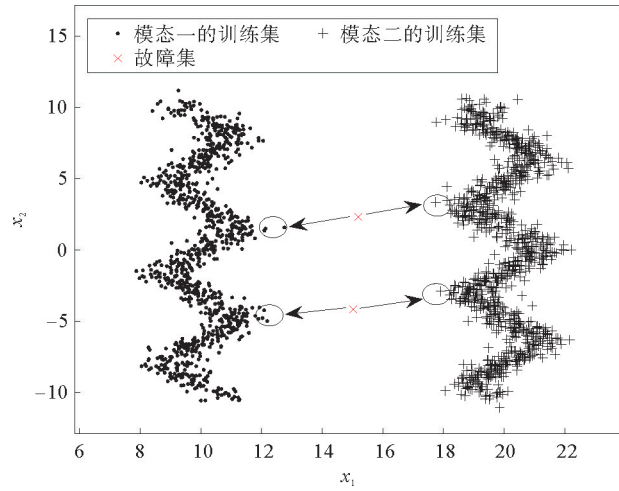


图 1 故障样本的近邻信息

Fig. 1 Neighbor information for fault samples

3 实验过程与分析

为验证所提方法的故障检测性能,分别使用数值模拟过程和青霉素发酵过程对 DLNS-LOF 方法进行性能检验,并与一些经典方法进行比较。

3.1 数值模拟过程

为验证 DLNS-LOF 方法的有效性,设计了一个具有非线性特征的多模态数值模拟过程,其过程模态一、模态二的表达式分别为:

$$\begin{cases} x_1 = e^{0.04s_1} \sin s_2 + e, \\ x_2 = -0.5s_1 + e; \end{cases} \quad (9)$$

$$\begin{cases} x_1 = e^{0.02s_1} \cos s_2 + 25 + e, \\ x_2 = -0.5s_1 + e. \end{cases} \quad (10)$$

式中: s_1 为服从区间 $[0, 72]$ 的均匀分布, s_2 为服从区间 $[0, 10\pi]$ 的均匀分布, e 为独立的随机噪声变量。在数值模拟过程的正常状态下,先分别从模态一和模态二中采集 1 000 个样本,共同构成训练集;再从两模态中采集 100 个样本,共同构成测试集。最后,在过程中引入图 2 所示的 9 个故障样本。

分别使用 LOF^[9]、LNS- k NN^[11] 和 DLNS-LOF 方法对该数值模拟过程进行故障检测,设定置信度为 97% 以确定控制限。LOF 方法的近邻参数 k 设为 3; LNS- k NN 中 LNS 的近邻参数 K 设为 5, k NN 的近邻参数 k 设为 4; DLNS-LOF 中 DLNS 的两个近邻参数 k 和 K 分别设为 5 和 12, LOF 的近邻参数 k 为 7。3 种方法的故障检测结果如图 3 所示。

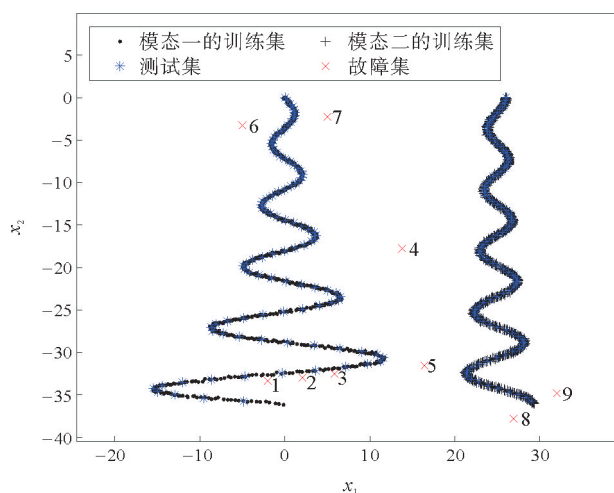


图 2 过程数据的空间散点图

Fig. 2 Spatial scatterplot of process data

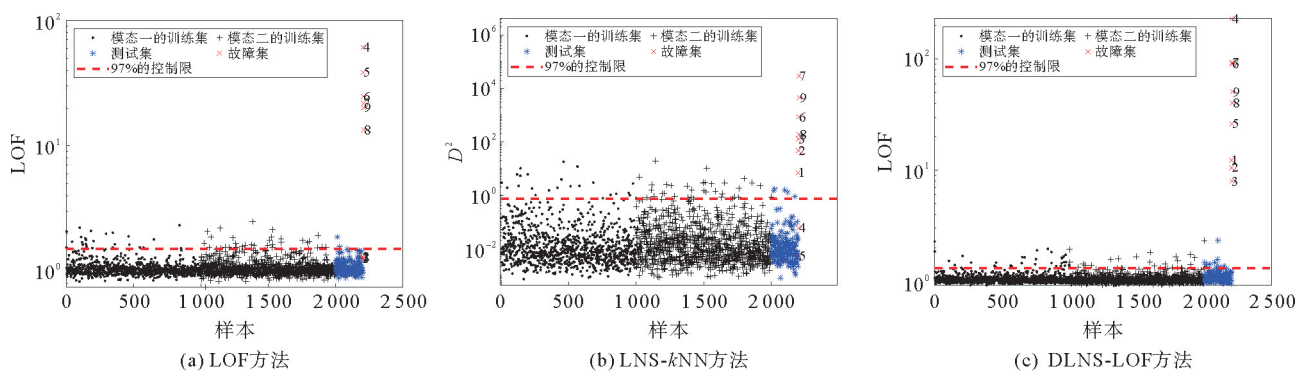


图 3 故障检测图

Fig. 3 Fault detection diagrams

从图 3 可以看出, DLNS-LOF 方法取得了最好的故障检测结果。故障 1、2 和 3 是位于模态一的微弱故障,其故障样本与训练样本的距离和部分训练样本间的距离大小相似。LOF 方法在处理此类故障时,故障样本到近邻样本的可达距离将部分取值为近邻样本的第 k 距离,而非其到近邻样本的距离。上述现象导致故障样本和近邻样本的局部可达密度非常接近,使得故障局部离群因子在 1 附近波动,难以有效

识别故障样本。

故障 4 和 5 在训练集中的近邻样本来源于两个模态,且其均值近似等于故障样本的坐标值,因此故障 4 和 5 经 LNS 标准化处理后,被移动到原点附近,混入正常样本中,如图 4(a)所示。而 DLNS 通过利用嵌套近邻集的均值和标准差进行标准化,避免了因近邻样本跨模态导致的偏差,从而有效分离出故障 4 和 5,如图 4(b)所示。LNS- k NN 在处理 LNS 标准化后的数据时,由于故障 4 和 5 与近邻样本的距离和正常样本相似,故无法检出故障 4 和 5。而 LOF 在处理 DLNS 后的数据时,正常样本在训练集中近邻样本的均值密度与自身密度大小接近,统计值接近于 1。故障样本在训练集中近邻样本的均值密度比自身密度大的多,因此统计值显著大于 1。故 DLNS-LOF 能够检出所有故障样本。

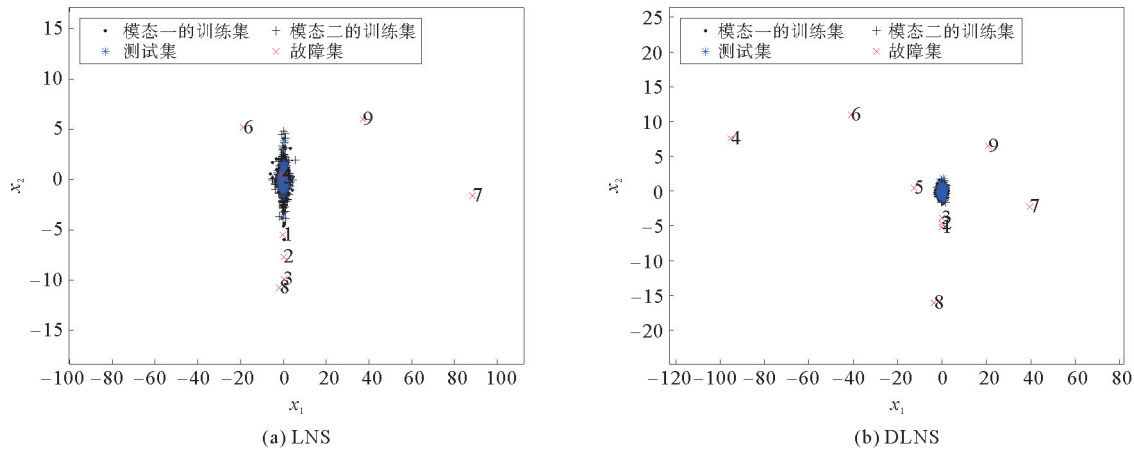


图 4 LNS 和 DLNS 后过程数据的空间散点图

Fig. 4 Spatial scatterplots of process data after LNS and DLNS

3.2 青霉素发酵过程

青霉素发酵过程是典型的具有非线性特征的多模态工业生产过程。本节实验数据来源于青霉素发酵过程的仿真平台^[13],该平台已被多项研究^[14-16]用于验证过程监控算法性能。实验所选取的变量见表 1。

仿真时长设置为 400 h,并间隔 1 h 采样一次。首先在正常过程运行下采集一批数据作为训练集,随后在通风率、搅拌功率和底物流速率中分别添加阶跃或斜坡故障,设置如表 2 所示的 6 类故障,并相应生成 6 批测试集。每批测试集的故障皆从第 51 h 开始持续到过程结束。

表 1 青霉素发酵过程所选变量

Table 1 Selected variables for the penicillin fermentation process

序号	变量名	序号	变量名
1	通风率	7	反应器温度
2	搅拌功率	8	底物浓度
3	底物流速率	9	溶解氧浓度
4	底物流温度	10	菌体浓度
5	反应器体积	11	青霉素浓度
6	CO ₂ 浓度	12	反应热

表 2 故障信息

Table 2 Description of faults

批次	故障变量	故障类型	故障幅度	批次	故障变量	故障类型	故障幅度
F1	1	阶跃	-0.5%	F4	1	斜坡	-2.5 L/h
F2	2	阶跃	-1.5%	F5	2	斜坡	-4.5 W
F3	3	阶跃	2.5%	F6	3	斜坡	0.02 L/h

分别使用 k NN^[7]、SD- k NN^[8]、LOF^[9] 和 DLNS-LOF 方法对青霉素发酵过程进行故障检测,置信度均设为 98%。 k NN 和 LOF 均设置近邻参数 k 为 4;SD- k NN 设置疏密信息参数 K 为 10,近邻参数 k 为 4;DLNS-LOF 设置 DLNS 的近邻参数 k 和 K 分别为 2 和 7, k NN 的近邻参数 k 设为 4。各方法的故障检测率如表 3 所示。

由表 3 可知,DLNS-LOF 方法故障检测率最高,尤其在 F3 批次的测试集上表现突出。

从图 5 过程数据中变量 1、2、3 的空间散点图可以发现,故障样本距正常样本的距离与正常样本间的距离相近。 k NN、LOF 和 SD- k NN 方法本质上都是基于近邻距离的故障检测方法,难以识别此类故障样本。LOF 和 SD- k NN 因不受样本疏密程度影响,其故障检测率高于 k NN。DLNS 技术利用嵌套近邻集的均值和标准差对过程数据进行标准化,有效分离了此类微弱故障。图 6 和图 7 分别展示了 DLNS 处理后变量 1、2、3 的空间散点图与频率分布图,可见原具有非线性、多模态特征的过程数据被转化为近似服从标准正态分布的单模态数据,且故障样本得到明显分离。故障样本的有效分离使得故障样本的近邻样本与自身的相对密度值增大,即局部离群因子增大,从而提高了 DLNS-LOF 的检测率,结果如图 8 所示。

表 3 所用方法的故障检测率

Table 3	Fault detection rate of the method used %			
批次	k NN	SD- k NN	LOF	DLNS-LOF
F1	88.57	98.29	96.86	98.57
F2	56.57	91.43	88.57	99.43
F3	1.14	7.14	6.00	99.14
F4	100.00	100.00	100.00	100.00
F5	98.00	98.57	98.57	100.00
F6	84.29	88.86	86.29	98.86

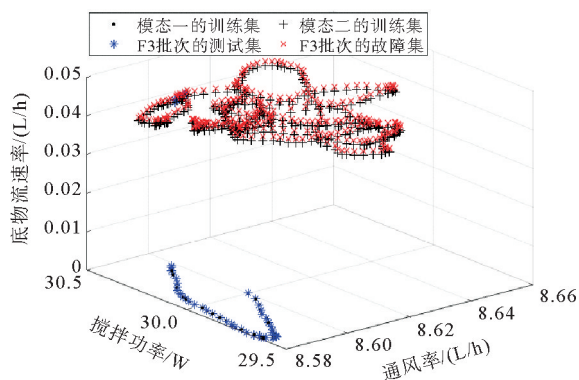


图 5 过程数据变量 1、2 和 3 的空间散点图

Fig. 5 Spatial scatterplot of variables 1, 2 and 3 in process data

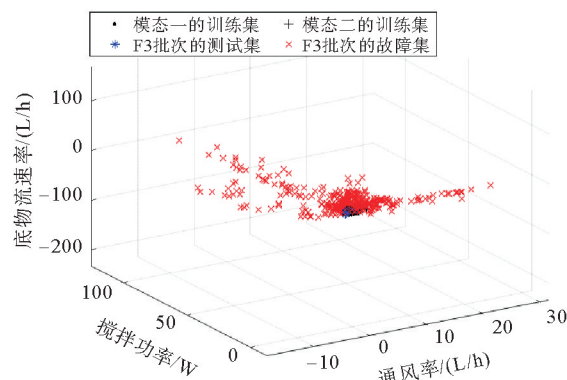


图 6 DLNS 处理后变量 1、2 和 3 的空间散点图

Fig. 6 Spatial scatterplot of variables 1, 2 and 3 after DLNS

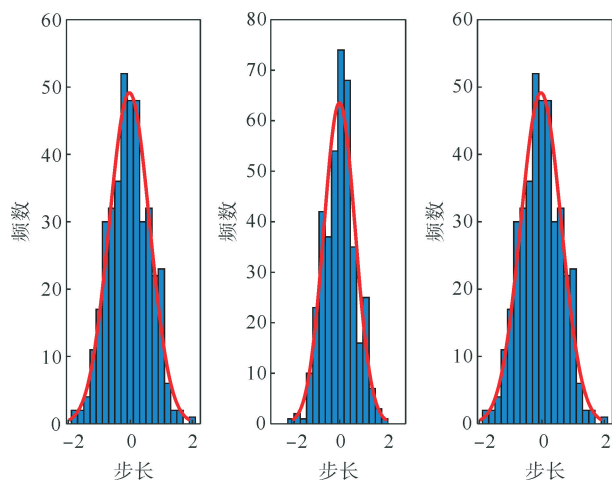


图 7 DLNS 处理后变量 1、2 和 3 的频率分布图

Fig. 7 Frequency distribution of variables

1, 2, and 3 after DLNS

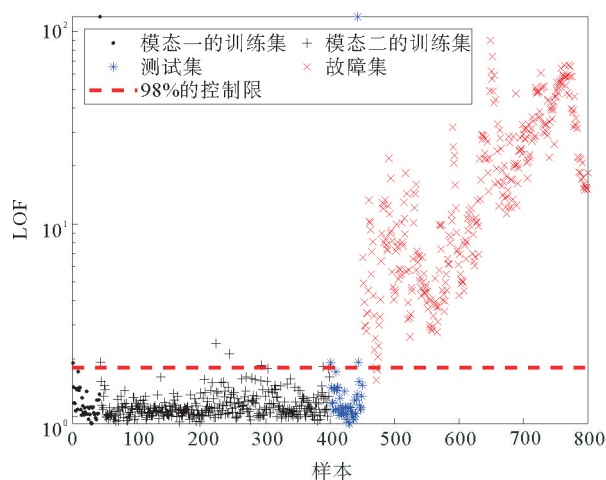


图 8 DLNS-LOF 方法对 F3 批次测试集的故障检测图

Fig. 8 Fault detection plot of DLNS-LOF method for F3 batch test set

DLNS 技术能够成功转化非线性多模态过程数据为近似服从标准正态分布的单模态数据,并成功分离出空间上发生偏离的故障样本。鉴于许多经典算法对于过程数据基于单模态和高斯分布的假设,DLNS 可以帮助其实现数据转换以剔除不利特征的影响,显著提升其故障检测能力。

4 结论

现实中的多数大型工业生产过程具有非线性和多模态特征。DLNS-LOF 方法采用 DLNS 技术,能够将具有非线性特征的多模态过程数据转化为近似服从标准正态分布的单模态数据,并在处理过程中有效分离故障样本,从而显著提升 LOF 方法的故障检测性能。与传统方法相比,DLNS-LOF 展现出更优的故障识别能力,有助于更好地保障系统的安全性与可靠性。然而,DLNS 技术仅从空间角度考虑故障样本的分离,对于因过程动态性导致的故障样本,即在空间分布上与正常样本重叠但发生时间不同的情况难以实现有效区分。这意味着 DLNS 目前仅适用于稳态过程数据的故障检测,未来将针对动态过程下的故障检测问题进一步研究。

参考文献:

- [1] 彭家浩,邱爱兵,缪杰,等. 基于递推 PCA 的变工况中央空调系统故障诊断[J]. 电子测量与仪器学报,2024,38(1):134-144.
PENG Jiahao, QIU Aibing, MIAO Jie, et al. Fault diagnosis of HVAC in variable operating mode based on recursive PCA [J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(1): 134-144.
- [2] 张谦谦,王文标,郝友维. 基于动态加权差分主成分分析的故障检测[J]. 科学技术与工程,2023,23(36):15522-15529.
ZHANG Qianqian, WANG Wenbiao, HAO Youwei. Fault detection based on dynamic weighted differential principal component analysis[J]. Science Technology and Engineering, 2023, 23(36): 15522-15529.
- [3] 张珂,杨鹏宇,石怀涛,等. 基于改进动态偏最小二乘法故障检测方法[J]. 沈阳建筑大学学报(自然科学版),2024,40(1):167-178.
ZHANG Ke, YANG Pengyu, SHI Huaitao, et al. Fault detection method based on improved dynamic partial least square method[J]. Journal of Shenyang Jianzhu University (Natural Science), 2024, 40(1): 167-178.
- [4] SCOLKOPF B, MIKA S, BURGESS C J C, et al. Input space versus feature space in kernel-based methods[J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1000-1017.
- [5] SI Y B, WANG Y Q, ZHOU D H. Key-performance-indicator-related process monitoring based on improved kernel partial least squares[J]. IEEE Transactions on Industrial Electronics, 2020, 68(3): 2626-2636.
- [6] 孔祥玉,周红平,罗家宇,等. 基于高效核偏最小二乘的质量相关故障检测[J]. 控制理论与应用,2023,40(4):683-692.
KONG Xiangyu, ZHOU Hongping, LUO Jiayu, et al. Quality-related fault detection based on efficient kernel PLS[J]. Control Theory and Applications, 2023, 40(4): 683-692.
- [7] HE Q P, WANG J. Fault detection using the k -nearest neighbor rule for semiconductor manufacturing processes[J]. IEEE Transactions on Semiconductor Manufacturing, 2007, 20(4): 345-354.
- [8] 冯立伟,张成,李元,等. 基于标准距离 k 近邻的多模态过程故障检测策略[J]. 控制理论与应用,2019,36(4):553-560.
FENG Liwei, ZHANG Cheng, LI Yuan, et al. Fault detection strategy of standard-distance-based k nearest neighbor rule in multimode processes[J]. Control Theory and Applications, 2019, 36(4): 553-560.
- [9] BREUNIG M M, KRIEGEL H P, NG R T, et al. LOF: Identifying density-based local outliers[J]. ACM SIGMOD Record, 2000, 29(2): 93-104.
- [10] MA H H, HU Y, SHI H B. Fault detection of complex chemical processes using Mahalanobis distance-based local outlier factor[J]. CIESC Journal, 2013, 64(5): 1674-1682.

- [11] 李元,李美萱,张成,等. 基于局部临近标准化的 FD-KNN 故障检测[J]. 山东科技大学学报(自然科学版),2017,36(5):1-6.
LI Yuan,LI Meixuan,ZHANG Cheng,et al. FD-KNN Fault detection based on local nearest neighborhood standardization [J]. Journal of Shandong University of Science and Technology(Natural Science),2017,36(5):1-6.
- [12] 冯立伟,张成,李元,等. 基于双近邻标准化和 PCA 的多阶段过程故障检测[J]. 化工学报,2018,69(7):3159-3166.
FENG Liwei,ZHANG Cheng,LI Yuan,et al. DLNS-PCA-based fault detection for multimode batch process[J]. CIESC Journal,2018,69(7):3159-3166.
- [13] BIROL G, ÜNDEY C, CINAR A. A modular simulation package for fed-batch fermentation: Penicillin production[J]. Computer and Chemical Engineering,2002,26(11):1553-1565.
- [14] VANLAER J, GINS G, IMPE J V. Comparison of MPCAV, ARPCA, and BDPCA fault detection performance in a fed-batch penicillin fermentation process[J]. IFAC Proceedings Volumes,2009,42(8):840-845.
- [15] 赵小强,张妍. 基于 NPE-ELM 的间歇过程质量相关故障检测[J]. 控制工程,2022,29(12):2284-2292.
ZHAO Xiaoqiang,ZHANG Yan. Quality-related fault detection of batch process based on NPE-ELM[J]. Control Engineering of China,2022,29(12):2284-2292.
- [16] 苑忠帅,孙四通. 基于局部信息的 LNS-PCA 的多模态过程故障监测[J]. 过程工程学报,2023,23(5):790-798.
YUAN Zhongshuai,SUN Sitong. Multimodal process fault monitoring of LNS-PCA based on local information[J]. The Chinese Journal of Process Engineering,2023,23(5):790-798.

(责任编辑:齐敏华)