

生成式人工智能算法的传感信息离群点自适应故障诊断方法

郑志娴¹

ZHENG Zhixian

摘要

工业传感数据流中早期故障信号微弱且易被复杂工况噪声淹没, 现有方法因缺乏对动态工况的自适应能力, 易受离群点干扰, 对微弱异常难以有效表征, 在数据分布时变、多模态噪声共存的场景下难以实现故障的实时检测与精准辨识, 导致故障漏检率高、类型误判严重。为解决这一问题, 文章提出一种基于生成式人工智能的传感信息离群点自适应故障诊断方法。采用基于局部密度与分离度融合的聚类中心优化方法, 通过综合评估数据点的局部密度指标和最小分离度, 实现鲁棒的聚类中心选取, 克服传统随机选择方法易受离群点干扰的问题; 基于实时数据分布特征, 设计自适应检测机制, 通过滑动窗口与分布漂移指数动态更新基准模型, 使检测阈值能够随工况波动自主调整; 引入去噪自编码器进行深层特征提取, 通过训练阶段的最小化重构误差建立正常模式基准, 在检测阶段结合隐空间编码与故障簇距离匹配实现异常检测与分层诊断。故障诊断测试结果表明, 所提方法在轴承内圈磨损、叶轮不平衡和机械密封泄漏三类故障场景下的真阳性率分别达到 94.3%、98.5% 和 89.6%, 故障类型平均识别准确率为 88.5%, 验证了其在多类故障检测与辨识任务中的有效性与工程适用性。

关键词

生成式; 传感信息; 自适应; 离群点; 人工智能算法

doi: 10.3969/j.issn.1672-9528.2026.06.041

0 引言

随着大数据时代的到来, 工业物联网等感知设备不断产生大量的监控数据。在此基础上, 将不同的数据按不同的时间顺序进行分类, 从而构成一种具有代表性的时序数据流。从理论上讲, 这些数据其总体变动趋势与次序紧密关联, 能够动态反映监控装置的工作状态和健康规律^[1]。所设计的传感器序列数据符合以下几个特点: 流式的、实时的、顺序的确定、与获取系统无关; 理论上, 数据的大小是无限的, 但由于内存和运算能力的限制, 一般情况下, 只能进行一次或多次扫描。

面对不断积累的海量时序数据, 传统的批量、离线处理方式难以满足需求^[2]。其核心难点是, 如何在“理论上只能读一次”的条件下, 准确、实时地识别出表征设备早期故障的异常模式, 即时序异常。这些异常值不只是噪音, 还包含着重要的“管理价值”和“知识模式”, 可以提示零件老化、系统失效或性能下降。如何从数据流特征中实时地提取失效信息, 并对其进行有效的故障诊断, 是目前工业生产智能化运营的迫切需求。

基于此, 本文开展了生成式人工智能算法的传感信息离

群点自适应故障诊断方法研究。

1 基于局部密度与分离度融合的聚类中心优化

在完成正常工况数据采集与训练样本集构建后, 如何从含噪声和离群点的传感数据流中准确选取代表性聚类中心, 成为影响后续正常模式建模精度的关键挑战。工业传感数据流中早期故障信号微弱且易被复杂工况噪声淹没, 加之数据分布时变、多模态噪声共存, 导致传统随机选择聚类中心的方法稳定性不足^[3]。初始聚类中心的随机选择不仅引发聚类结果波动, 更因离群点干扰造成中心偏移, 致使基准模型偏离真实正常工况。为此, 本节采用基于局部密度与分离度融合的聚类中心优化方法, 通过综合评估数据点的局部密度指标和最小分离度, 实现鲁棒的聚类中心选取, 为后续正常行为模式建模提供稳定可靠基准。局部密度指标的计算公式为:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \quad (1)$$

式中: ρ_i 表示数据点 i 的局部密度指标, d_{ij} 表示数据点 i 与数据点 j 之间的欧氏距离, d_c 表示一个预设的截断距离, $\chi(x)$ 表示一个符号函数, 当 $x < 0$ 时, 函数值为 1, 否则为 0。

对于具有最高局部密度的点, 其最小分离度则定义为所有点中最大的距离^[4]。最小分离度的计算公式为:

1. 福建船政交通职业学院 福建福州 350007

$$\delta_i = \min_{\{j: \rho_j > \rho_i\}} d_{ij} \quad (2)$$

对于密度最高的点, 满足 $\delta_i = \max_j d_{ij}$ 。为了综合评估一个点作为聚类中心的潜力, 将局部密度指标和最小分离度相乘, 得到一个综合显著性分数 $\Gamma_i = \rho_i \times \delta_i$ 。综合显著性得分高的点不仅处于数据密集区, 而且与其他高密集点之间有一定的距离, 是进行聚类的最佳候选。

通过上述综合显著性评估方法选取的聚类中心, 兼具数据分布代表性与抗离群点干扰的鲁棒性, 为系统建立了初始基准模型。该模型准确刻画了正常工况下的数据分布特征, 为后续的异常检测与模型自适应更新提供了可靠的参照基准。

2 基于分布漂移监测的自适应检测机制设计

在完成基于局部密度与分离度融合的聚类中心优化并建立初始基准模型后, 如何使该模型适应动态变化的工业现场环境成为关键^[5]。工业传感数据流中早期故障信号微弱, 易被复杂工况噪声淹没, 且数据分布随时间不断演变。若基准模型保持固定, 将导致模型性能随时间退化, 影响故障诊断准确性。为此, 本节旨在设计一种自适应检测机制, 通过对基准模型的动态更新, 使其能够持续反映系统近期的正常运行状态。

通过利用基准模型输出的重构误差分布特征, 设计基于数据分布密度的自适应检测机制。采用滑动窗口模型来管理实时到达的传感数据流^[6]。通过滑动窗口与分布漂移指数动态更新基准模型, 使检测阈值能够随工况波动自主调整。设滑动窗口的大小固定为 W , 在任意时刻 t , 窗口内包含的数据集合为 $\{X_{t-W+1}, X_{t-W+2}, \dots, X_t\}$ 。当新的数据点 X_{t+1} 到达时, 最旧的数据点 X_{t-W+1} 将被移出窗口, 新的数据点被纳入, 以此确保窗口内始终是最新的 W 个数据点, 该机制保证了用于模型更新和决策的数据能够反映系统近期的运行状态^[7]。

将数据空间划分为 M 个互不重叠的子区域 $\{S_1, S_2, \dots, S_M\}$ 。对于每个子区域 S_m , 计算其在当前滑动窗口内的数据分布密度。分布密度的计算公式为:

$$D_m^c = \frac{N_m}{V_m} \quad (3)$$

式中: D_m^c 表示子区域 S_m 的当前分布密度, N_m 表示落在 S_m 内的数据点数量, V_m 表示子区域 S_m 的体积。

同时, 记录生成式模型最后一次更新时所基于的训练数据在各个子区域的分布密度 D_m^{model} 。分布漂移指数 Δ 则通过计算当前分布密度与模型分布密度之间的加权差异来获得, 其计算公式为:

$$\Delta = \sum_{m=1}^M w_m |D_m^c - D_m^{\text{model}}| \quad (4)$$

式中: w_m 表示每个子区域的权重, 可以根据该区域在历史故障中的重要性来设定。

设定一个动态阈值 θ , 当计算出的分布漂移指数 Δ 超过 θ 时, 自适应机制被触发, 确保系统始终以反映近期运行状态的数据进行故障判断。

3 基于生成式人工智能算法的异常检测与分层诊断

早期故障信号常隐匿于复杂噪声背景中, 其特征一般不直接体现为数据分布的宏观变化, 而是存在于细微的特定信号模式中。对于工业传感数据流中这类微弱且易被噪声掩盖的故障特征, 仅依靠自适应检测机制难以有效识别。去噪自编码器作为生成式人工智能算法, 具备强大的特征提取能力, 能从含噪数据中还原纯净信号, 在训练过程中学会过滤随机噪声并捕捉数据内在规律。该算法可有效提取微弱故障信号的本质特征, 通过重构误差识别异常, 显著提升故障检测准确性, 降低漏检率, 弥补了自适应检测机制在特征提取方面的不足。

去噪自编码器是一种能从局部损伤或含噪声的数据恢复出纯净、完整的数据的神经网络。训练阶段, 明确训练目标是最小化重构序列 X' 与原始正常序列 X 之间的差异, 通常使用均方误差作为损失函数:

$$L_{\text{train}} = \frac{1}{N} \sum_{i=1}^N (X_i - X'_i)^2 \quad (5)$$

通过训练过程, 去噪自编码器能够掌握过滤随机噪声并提取正常传感数据内在规律的能力。在检测阶段, 将实时采集到的传感序列 Y 直接输入到已经训练好的去噪自编码器, 编码器-解码器网络输出重构序列 Y' 。

基于重构误差识别异常序列后, 进一步结合隐空间编码与故障簇距离匹配实现分层诊断。将异常序列输入去噪自编码器进行隐空间编码, 获得高度抽象的特征表示。计算异常点潜在编码 Z_m 到每个已知故障簇中心 C_f 的欧氏距离 D_f :

$$D_f = \|Z_m - C_f\|_2 \quad (6)$$

将距离最小的故障类型作为首要的诊断结果, 完成从异常检测到故障辨识的全流程诊断。

4 故障诊断测试

4.1 测试环境

为验证本文上述提出的故障诊断方法在实际工业场景中的可行性与有效性, 设计并执行一系列故障诊断测试。

测试环境旨在模拟一个真实的工业设备监控场景, 以获取具有实际物理意义的传感数据流。测试采用了一个工业水泵机组状态监测实验台架作为数据来源。

该平台包括一台额定功率 5.5 kW 的离心水泵, 一套控制

水泵转速的可编程逻辑控制器 (PLC)，一套含有振动、温度和压力传感器的数据采集系统和上位机工作站。所述的振动传感器（型号：ICP 加速度仪，灵敏度 100 mV/g），其上装有一种用来获取径向振动信号的装置；在所述轴承外环上设置有用于监控所述轴承温度的温度传感器（PT100）；在泵的出口处装有一个压力传感器，用来监控出水压力。为了适应振动信号的分析要求，采用 24 位高精度数字采集卡，以 25.6 kHz 的采样频率对各传感器信号进行同步采集。测试数据集构成如表 1 所示。

表 1 测试数据集构成表

数据类别	数据量 / 万条	描述	用途
正常状态训练集	96	前 6 h 正常数据	模型初始训练
正常状态验证集	24	后 2 h 正常数据	模型验证与阈值设定
故障状态 F1	30	轴承内圈磨损数据	故障诊断测试
故障状态 F2	30	叶轮不平衡数据	故障诊断测试
故障状态 F3	30	机械密封泄漏数据	故障诊断测试

表 1 中故障状态 F1、F2、F3 分别对应轴承内圈轻微磨损、叶轮轻微气蚀不平衡、机械密封轻微泄漏。图 1 所示为测试数据集的具体分布情况示意图。

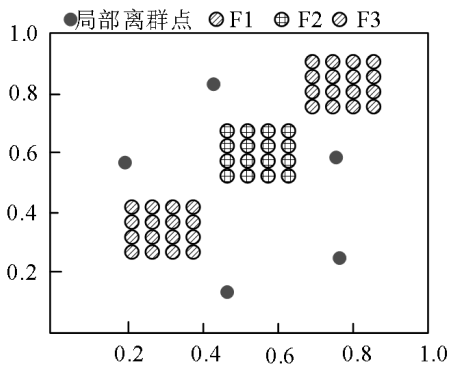


图 1 测试数据集的具体分布示意图

整个测试过程共采集包含正常与三种故障状态在内的约 210 万条多维传感序列数据，构成本次测试的数据基础。

4.2 参数配置

对于生成式人工智能模型，其结构配置为：输入层维度为 3，对应振动、温度、压力三个传感维度。编码器部分由两个全连接层构成，神经元数量分别为 64 和 16，激活函数采用 ReLU。在训练时，人工增加高斯噪音均值为 0.05，批量数据规模设定为 128，采用 Adam 寻优算法，其初值为 0.001。模型的训练，一共进行 200 轮。

关键参数配置如表 2 所示。

表 2 关键参数配置

参数模块	参数名称	取值	说明
生成式模型	编码器层 1	64	第一层神经元数
	编码器层 2	16	第二层神经元数
	潜在空间维度	8	特征编码维度
	噪声标准差	0.05	训练时添加噪声强度
	批次大小	128	训练批处理量
	初始学习率	0.001	Adam 优化器参数
自适应机制	滑动窗口大小	5 000	用于模型更新的数据量
	子区域数量	6	数据空间划分个数
	分布漂移阈值	0.18	触发模型更新的阈值
异常诊断	误差计算窗口	100	局部误差平均窗口
	敏感度参数	2.5	异常判定阈值系数
	历史最大误差	0.32	用于严重程度归一化

如表 2 中所示，对于自适应检测机制，滑动窗口的大小 W 设置为 5 000 个数据点，这相当于在 1 Hz 的数据流更新频率下约 83 min 的数据量，能够平衡对工况变化的响应速度与模型的稳定性。

4.3 故障诊断效果分析

在利用本文提出的方法对传感信息离群点自适应故障诊断时，得到图 2 所示的聚类结果。

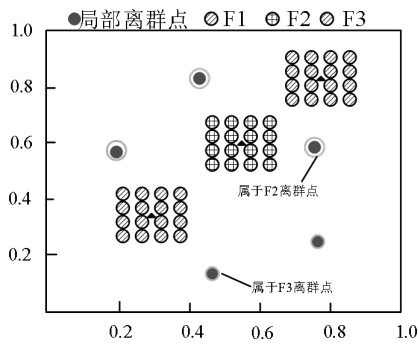


图 2 聚类结果

从图 2 可以看出，正常工况数据形成了稳定致密的主聚类簇，三种故障状态的数据分别形成了独立的子簇，且各簇之间保持了良好的分离度。基于局部密度与分离度融合的聚类中心优化方法有效识别出了具有代表性的聚类中心，这些中心点均位于各簇的密度核心区域，为正常行为模式建模提供了稳定的基准参考。在此基础上，评估方法在三种故障状态下检测离群点的能力。计算了每个故障数据集中，被正确标记为异常的数据点比例（真阳性率），以及正常验证集中被误判为异常的数据点比例（假阳性率），结果如表 3 所示。

表 3 离群点检测性能分析

故障状态	真阳性率	假阳性率	平均重构误差
正常状态（验证集）	—	0.021	0.045
F1：轴承内圈磨损	0.943	0.036	0.178
F2：叶轮不平衡	0.985	0.029	0.254
F3：机械密封泄漏	0.896	0.041	0.152

在轴承内圈磨损（F1）、叶轮不平衡（F2）和机械密封泄漏（F3）三种故障场景下，真阳性率分别达到 94.3%、98.5% 和 89.6%，表明所提方法对不同类型故障均具备较高的检测灵敏度。正常验证集的假阳性率为 2.1%，说明系统在保持高检测率的同时具有较好的特异性。平均重构误差值显示，故障状态下的重构误差显著高于正常状态，其中叶轮不平衡工况的平均重构误差最大，这与该类故障引起的振动特征变化最为明显具有直接关联。

基于此，本文方法针对故障类型识别能力进行系统评估与分析。对于每个被检测出的异常序列，提取其潜在空间编码和误差特征向量，并与已知的 F1、F2、F3 故障特征库进行匹配。统计故障类型识别的混淆矩阵如表 4 所示。

表 4 故障类型识别的混淆矩阵

实际类型	单位：%			
	诊断结果：F1	诊断结果：F2	诊断结果：F3	诊断结果：未知
F1：轴承内圈磨损	88.5	5.2	2.1	4.2
F2：叶轮不平衡	3.8	91.7	1.5	3.0
F3：机械密封泄漏	4.5	2.8	85.3	7.4

由表 4 可知，该方法对三种故障类型的平均识别准确率达到 88.5%，其中轴承内圈磨损和叶轮不平衡的识别准确率分别为 88.5% 和 91.7%，机械密封泄漏的识别准确率为 85.3%。识别错误主要发生在轴承内圈磨损与叶轮不平衡之间，这与两者引发的振动特征具有一定相似性相关。少量故障样本被标记为“未知”类型，主要集中在故障特征尚未充分发展的初期阶段。

5 结语

本文针对工业传感数据流中早期故障信号微弱、易受工况噪声干扰且数据分布动态演变的技术难题，提出一种基于生成式人工智能的传感信息离群点自适应故障诊断方法。通过构建多层次的基准模型与动态更新机制，实现了从正常模式建模、工况自适应到故障精准诊断的全流程技术闭环。主要研究成果包括：

（1）提出基于局部密度与分离度融合的聚类中心优化方法，通过综合评估数据点局部密度与最小分离度，实现鲁棒的聚类中心选取，建立准确刻画正常工况分布的初始基准

模型，克服传统随机选择方法因离群点干扰导致的基准偏移问题。

（2）设计基于分布漂移监测的自适应检测机制，采用滑动窗口管理实时传感数据流，通过分布密度计算与动态阈值比较，实现基准模型的在线更新与检测阈值自主调整，使系统适应慢变工况与多工况切换的复杂工业环境。

（3）引入去噪自编码器进行深层特征提取，通过最小化重构误差建立正常模式基准，结合隐空间编码与故障簇距离匹配，实现微弱故障特征增强提取与故障类型层次化诊断，提升早期故障检测灵敏度与分类精度。

实验结果表明：在轴承内圈磨损、叶轮不平衡和机械密封泄漏三类典型故障场景下，该方法真阳性率分别达 94.3%、98.5% 和 89.6%，故障类型平均识别准确率为 88.5%，正常工况误报率控制在 2.1% 以内，验证了其在多类故障检测与辨识任务中具有较高可靠性与稳定性。

本研究建立了完整的自适应故障诊断技术体系，解决了传统方法在动态工况下模型性能退化与微弱故障特征提取不足的问题，为工业设备智能运维提供了有效技术方案，具有明确的工程应用价值。

参考文献：

[1] 陈锦铭, 蒋玮, 袁宇波, 等. 基于电压一阶差分分离群点关联分析的配电网线变关系诊断 [J]. 电力系统保护与控制, 2024, 52(18):158-168.

[2] 贺忠海, 朱温涵, 陈旭旺, 等. 基于自适应密度聚类的多准则主动学习方法 [J]. 仪器仪表学报, 2024, 45(3):179-187.

[3] 袁卫声, 王运辉, 罗力维, 等. 基于半监督学习的无线传感网离群值检测算法 [J]. 南昌大学学报 (工科版), 2023, 45(2):189-194.

[4] 江军, 张文乾, 李波, 等. 电力变压器油中溶解气体离群值识别和数据重构 [J]. 电工技术学报, 2024, 39(17):5521-5533.

[5] 陈明锐, 朱楚伟, 陈辉, 等. 基于机器学习与预测分析的直流光测量系统远程模块的寿命预测方法 [J]. 传感技术学报, 2023, 36(3):397-403.

[6] 徐升, 佟佳睿, 胡祥恩. 下一代个性化学习: 生成式人工智能增强智能辅导系统 [J]. 开放教育研究, 2024, 30(2):13-22.

[7] 唐伦, 李质萱, 文雯, 等. 基于智能分层切片技术的数字孪生传感信息同步策略 [J]. 电子与信息学报, 2024, 46(7): 2793-2802.

【作者简介】

郑志嫻（1979—），女，福建福州人，硕士，副教授，研究方向：移动互联开发、Java 开发、人工智能。

（收稿日期：2025-10-23 修回日期：2026-06-04）