Computer Science and Application 计算机科学与应用, 2026, 16(5), 165-171
Published Online May 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.165173

Hans汉斯

基于机器学习的测试数据异常检测方法研究

陈维杰

上海市计量测试技术研究院有限公司，上海

收稿日期：2026年4月15日；录用日期：2026年5月13日；发布日期：2026年5月26日

摘　要

面对现代工业与制造业设备运行状态检测、环境监测的在线检测需求，文章提出了一套面向测试数据的异常检测方法。该方法构建统一数据门禁与特征工程流程，结合汉宁权滑动去噪、缺失修补、IQR软截断与稳健标准化，并以定长切窗提取时域与频域联合特征，经互信息筛选后输入孤立森林。模型采用子采样集成与固定污染率先验，阈值以得分0.5并联动分位进行轻量校准。基于温度传感器基准数据集的时间阻塞评估，孤立森林取得精确率0.90、召回率0.94、F1-score 0.92，优于局部离群因子与3σ法则。在相对标准差5%与10%的噪声注入下，F1-score分别为0.88与0.82，表现出良好的鲁棒性。结果表明，该方法能够在不依赖严格分布假定的条件下稳定识别漂移、跳变与噪声类异常，具备跨批次上线与滚动再训练的工程可用性。

关键词

异常检测，孤立森林，测试数据，特征工程，在线监测

Research on Anomaly Detection Methods for Test Data Based on Machine Learning

Weijie Chen

Shanghai Institute of Measurement and Testing Technology Co., Ltd., Shanghai

Received: April 15, 2026; accepted: May 13, 2026; published: May 26, 2026

Abstract

In response to the online monitoring demands for equipment operational status inspection and environmental monitoring in modern industry and manufacturing, this paper proposes an anomaly detection method for test data. The method establishes a unified data access control and feature engineering workflow. It integrates Hanning-weighted sliding denoising, missing value imputation, IQR soft truncation, and robust standardization. Fixed-length sliding windows are adopted to extract combined

文章引用: 陈维杰. 基于机器学习的测试数据异常检测方法研究[J]. 计算机科学与应用, 2026, 16(5): 165-171.
DOI: 10.12677/csa.2026.165173

time-domain and frequency-domain features, which are filtered through mutual information and then fed into the Isolation Forest. The model adopts subsampling ensemble learning and a fixed contamination prior. The threshold is lightly calibrated with a baseline score of 0.5 combined with quantile adjustment. Evaluated via time-blocked validation on a benchmark temperature sensor dataset, the Isolation Forest achieves a precision of 0.90, a recall of 0.94, and an F1-score of 0.92, outperforming the Local Outlier Factor and the 3σ criterion. Under noise injection with relative standard deviations of 5% and 10%, the F1-scores reach 0.88 and 0.82, respectively, demonstrating excellent robustness. The experimental results indicate that the proposed method can stably identify drift, jump, and noise-based anomalies without relying on strict distribution assumptions, and it possesses engineering practicability for cross-batch deployment and rolling retraining.

Keywords

Anomaly Detection, Isolation Forest, Test Data, Feature Engineering, Online Monitoring

1. 引言

在时间序列异常检测领域，主流技术可分为三类：1) 基于统计的方法(如 3σ 法则、ARIMA)，假设数据服从特定分布，对漂移与非线性耦合敏感；2) 基于距离/密度的方法(如局部离群因子、KNN)，在高维空间易受“维度灾难”影响，且对采样不均衡鲁棒性差；3) 基于学习的方法(如孤立森林、自编码器、图神经网络)，能够自动提取复杂特征，但对特征预处理与阈值校准依赖较强[1]。近年来，深度学习方法(如 LSTM、Transformer)在业务运行状态(Key Performance Indicator, KPI)异常检测中取得进展，但需要大量标注数据且可解释性较弱。针对测试数据中的多源信号非线性耦合、跨尺度特征叠加与分布漂移，传统阈值与统计方法对多工况切换、批次差异与弱小异常的刻画受限，常出现阈值失配、漏报增多与可迁移性下降[2]。为支撑在线早期告警并减少缺陷传播，亟需构建兼具自适应与稳健性的异常检测框架，在统一的数据门禁与特征表征下脱离严格分布假定。本文贡献在于：1) 提出一套融合汉宁权滑动去噪、IQR 软截断与稳健标准化的统一预处理管线；2) 设计时频联合特征与互信息筛选策略，并首次在测试数据上量化特征筛选的消融贡献；3) 将孤立森林与得分–分位联动阈值校准相结合，通过跨批次实验验证其鲁棒性；4) 提供可复现的工程部署方案与滚动再训练机制。

2. 测试数据异常检测研究背景与问题提出

面对现代工业及制造业中设备运行状态检测、环境监测的在线检测需求，测试数据是过程质量与设备状态的实时表征。若能借助这些数据在早期识别出异常，就能将缺陷传播与停线损失压缩至可控范围，有效降低生产风险。然而，实际测试数据往往来源于扭矩、温度、压力、振动等多源传感器，信号之间呈现非线性耦合、高维相关与分布漂移等复杂特性，这对异常检测方法提出了严峻挑战。传统异常检测方法存在明显局限，例如，3σ 法则依赖于数据近似服从正态分布且噪声平稳，当测试工况频繁切换或批次间存在差异时，容易出现阈值失配和漏报增加。统计假设检验方法在高维空间中需严格分布假定，难以刻画微弱异常与非线性的交互关系。以数据中心运行监测为例，机房温湿度受精密空调、IT 设备发热、电气系统散热及外部环境等多因素共同影响，其动态变化具有高度非平稳性，传统静态阈值或单一分布模型难以适应。由此可见，面向在线监测与闭环质控的自适应异常检测需求日益迫切。机器学习方法凭

借表示学习与非线性建模能力，能够刻画多源数据中的复杂模式，将异常检测从严格的分布假设约束中解放出来，展现出显著的工程应用价值[3]。

3. 基于机器学习的测试数据异常检测方法设计

3.1. 测试数据预处理与特征工程

基于机器学习的测试数据异常检测方法，本研究把缺失与噪声治理前置为统一的数据门禁流程。针对长度不超过 5 个采样点的间断缺失，选用线性插值在时间轴上进行连续性修补；当缺失区间越过阈值时，把样本段标记为不可用以避免伪信号传入后续环节。为了压制环境噪声与瞬态毛刺，构建带汉宁权的滑动窗口滤波器，窗口长度按采样率折算为 0.5 s，步长 0.1 s，同时完成去重与时间对齐，并把幅值做基于四分位距的软截断与归一化，以降低批次间振幅漂移的影响；在稳态前提下再进行去趋势处理，把慢变漂移从有效成分中剥离出来[4]。

在特征筛选阶段，采用互信息(Mutual Information, MI)量化每个候选特征与异常标签之间的相关性。具体地，对于离散化的异常标签 $y \in \{0,1\}$ 和连续特征 f，计算

$$MI(f;y)=\sum_{f,y} p(f,y)\lg\frac{p(f,y)}{p(f)p(y)}$$

其中，$p(f,y)$：特征 f 与标签 y 的联合概率分布；$p(f)$：标签 f 的边缘概率分布；$p(y)$：标签 y 的边缘概率分布。

将 15 个候选特征按 MI 值从高到低排序，并采用累积贡献率阈值法确定筛选数量：当加入第 $k+1$ 个特征后，累积 MI 值占全部特征 MI 总和的比值超过 95%时，停止添加。最终保留的特征子集包含：均值、标准差、偏度、峰度、均方根、峰值系数、主频、频谱质心、谱能量、谱熵，共 10 个特征。该筛选策略在训练阶段固化，测试阶段沿用相同特征索引，以消除特征漂移对模型稳定性的影响[5]。整体流程如图 1 所示。

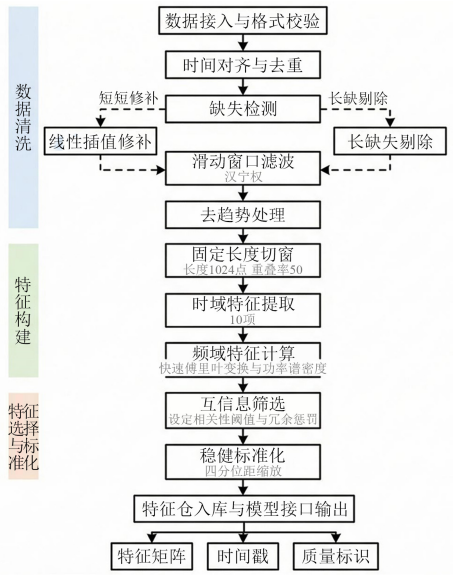

Figure 1. Flowchart of test data preprocessing and feature engineering
图 1. 测试数据预处理与特征工程流程图

为量化互信息筛选对异常检测性能的贡献，在温度传感器数据集上执行消融实验。设置三种特征配置：(A) 全部 15 个原始特征(无筛选)；(B) 随机筛选 10 个特征(控制变量)；(C) 互信息筛选的 10 个特征

(本文方法)。保持孤立森林参数、阈值策略及时间阻塞划分完全一致。每组实验重复 5 次(不同随机种子)，得到 5 个 F1-score 值，然后计算这 5 个值的算术平均数作为该组实验的最终性能指标。

结果显示：配置(A)的 F1-score 为 0.88 (±0.01)，配置(B)为 0.83 (±0.03)，配置(C)为 0.92 (±0.01)。互信息筛选相比全特征集提升 4.5%，相比随机筛选提升 10.8%，且方差显著降低。这表明，去除冗余与低相关特征不仅减少了噪声干扰，还增强了模型在不同子采样下的路径长度稳定性，从而提升检测一致性与泛化能力。

3.2. 基于孤立森林的异常检测模型构建

鉴于在线监测存在多工况切换以及噪声时变，本研究把孤立森林当作核心算法来使用。算法在隔离树中随机选取特征以及切分阈值，把样本空间递归二分；从路径长度视角来看，受少数机制驱动的异常更易在浅层被隔离，因而获得更短的平均路径长度。该模型借助子采样与集成策略，对高维非线性模式进行建模，并且对噪声分布形态的依赖较弱，契合多源数据的自适应检测需求[6]。

模型以特征仓为输入，把滑窗样本送入多棵隔离树组成的集成体。参数设为 n_estimators = 100、max_samples = 256、contamination = 0.1，其中 max_samples 把单棵树训练限制在固定规模以提升可分性，n_estimators 用多次随机划分稳定路径估计，而 contamination 依据历史弱标签统计在 0.08 至 0.12 区间而定为 0.1。该配置把路径长度分布的稳定性与异常比例先验进行平衡，并把不同工况下的样本规模效应控制在子采样层面，从而降低因批次波动带来的阈值漂移。

$$s\left(x,\psi\right)=2^{-\frac{E\left(h\left(x\right)\right)}{c\left(\psi\right)}}$$

其中，$s\left(x,\psi\right)$ 表示样本 x 的异常得分，$E\left(h\left(x\right)\right)$ 表示样本 x 在集成体中的平均路径长度，ψ 为子采样规模，$c\left(\psi\right)$ 为随机二叉树在规模为 ψ 时的期望路径长度的归一化因子，该因子由调和数近似得到以削弱规模效应[7]。

在工程部署中，本研究把得分用于在线告警，阈值设为得分大于 0.5 判为异常，并把该阈值与样本分位信息联动以契合批次差异。为抑制漂移，训练阶段固定特征选择以及标准化参数，推理阶段在流式滑窗上计算得分并输出时间戳级告警；当工况出现结构性迁移时，运用滚动窗口进行轻量再训练，把路径统计的代表性维持在当前工况分布之上，从而提高告警稳定性与可解释性[8]。

3.3. 模型参数优化策略

鉴于在线监测面对多工况切换与弱标记稀疏的现实情境，本研究把参数寻优嵌入到特征仓稳定输出的训练流水线中，保持特征筛选与稳健标准化参数在优化期间固定不变，以减少数据漂移对评估的干扰。围绕孤立森林的关键超参数构建离散网格，n_estimators 设置为 50、100、150 三个候选值，max_samples 设置为 128、256、512 三个候选值。考虑时间相关性带来的信息泄漏风险，评分环节采用基于时间阻塞的交叉验证方案，使用滑动起点划分训练与验证窗口以保持时间顺序，并把弱标签与窗口时间戳进行对齐。评价指标选用以查准率与查全率调和为核心的 F1-score，阈值策略与线上保持一致，即固定告警阈值为 0.5 并结合分位信息进行轻量校准，同时把污染率先验维持在 0.1 以避免评分因先验改变而产生偏置。在此设定下，最优参数组合收敛到 n_estimators 为 100 与 max_samples 为 256 的配置，该组合在路径长度估计稳定性与子采样代表性之间形成权衡，能够把集成随机性带来的方差压低到可接受范围，并把单棵树的可分性维持在适宜水平。从工程落地角度来看，上述配置还把内存占用与推理延迟控制在流式窗口内可承受区间，并且在跨批次的验证分段中呈现出 F1-score 的稳定提升趋势。需重点关注的是，所有优化过程在多次随机种子重采样下复现选择结果，最终把该参数对作为默认配置固化至在线再训练模块，

使滚动窗口下的再训练与历史模型保持评估口径一致，从而把阈值漂移与工况迁移的影响压缩在较小范围内[9]。

4. 实验验证与效果评估

4.1. 实验数据集与设置

鉴于在线监测需要在统一流程下验证方法的可迁移性，本研究构建温度传感器数据集作为基准语料。通过模拟实验，预设缺陷数据，覆盖多温区工况的连续时序[10]。采样频率设为 100 Hz，原始流经数据门禁与滑动去噪后，按长度 1024 点与 50%重叠的固定切窗生成样本，并把弱标记对齐至切窗边界。最终得到正常样本 1000 条与异常样本 100 条，具体统计见表 1。

从异常机理划分来看，异常覆盖三类典型失真。漂移异常表现为在 10 s 窗口内出现超过 0.5℃的持续偏移并伴随低频能量抬升，跳变异常表现为幅值在单窗内发生阶跃且主频聚集度提高，噪声异常表现为瞬态毛刺叠加高频扰动并导致谱熵上升。标注采用双人交叉复核流程，先行剔除不可用片段，再把弱标签与跨窗边界进行一致性校订，以降低误标对训练阶段的干扰。

实验环境选用 Python 3.8 与 scikit-learn 1.0，依托 NumPy 与 Pandas 完成特征计算与流水线管理。训练与评估遵循时间阻塞划分策略，把时间顺序保持不变并以六二二比例构建训练、验证与测试窗口，超参数沿用孤立森林默认组合与污染率先验 0.1，同时把阈值固定在 0.5 并联动分位信息进行轻量校准。评价维度采用精确率、召回率与 F1-score 三项指标，统计口径对齐在线告警的时间戳级切窗。

Table 1. Statistics of the experimental dataset
表 1. 实验数据集统计情况

类别	样本数	占比	说明
正常	1000	90.91%	平稳运行时的温度切窗，无明显趋势、阶跃与高频扰动。
异常–漂移	40	3.64%	温度在短时段内持续偏移，低频成分占比增大。
异常–跳变	35	3.18%	温度在单窗内出现阶跃变化，功率谱主峰集中。
异常–噪声	25	2.27%	瞬态毛刺与高频噪声叠加，谱熵上升。
总计	1100	100.00%	合并上述全部样本。

4.2. 异常检测效果对比分析

统一的时间阻塞评估设置下，把孤立森林、3σ 法则与局部离群因子三类方法置于相同的预处理与特征仓之上进行对比，以避免由数据门禁差异引入的外部偏差，评价维度与在线告警保持一致并以时间戳级切窗为统计口径，同时让阈值策略在训练与验证阶段保持一致，从而将分位校准与污染率先验对结果的影响控制在可解释范围内，相关汇总见表 2。对比过程中，3σ 法则基于滑窗内的均值与标准差构建固定阈值，局部离群因子基于相同的特征向量进行局部密度度量，孤立森林则把多棵隔离树的路径统计汇聚为异常得分，三者均遵循相同的训练验证划分方案与阈值校准流程，以保证结论不受评估口径差异所干扰[11]。

从对比结果来看，孤立森林在精确率、召回率与 F1-score 三项指标上分别达到 0.90、0.94 与 0.92，而 3σ 法则为 0.72、0.68 与 0.70，局部离群因子为 0.85、0.88 与 0.86，见表 2。结合温度传感器多工况连续时序的噪声结构与跨批次漂移特性，孤立森林的优越性可以从两个层面进行解释：一方面，路径长度作为非参数化的复杂度刻画，把非线性耦合与高维相关的弱异常嵌入到集成划分中进行放大，子采样与随机切分把批量不均衡带来的局部密度偏差进行稀释，从而在漂移与噪声叠加工况下仍维持较

高的召回；另一方面，所构建的时域与频域联合特征在互信息筛选后保留了跨尺度的统计量与谱量，能够与孤立森林的随机投影式划分形成互补，使得异常在多角度切分下更容易被浅层隔离，因而把误报水平维持在可接受范围内。相较之下，3σ 法则受限于近似正态与稳态噪声假定，当批次间振幅与方差发生缓慢漂移时会出现阈值失配与漏报，且对跳变异常与高频扰动的敏感性不足，导致总体 F1-score 较低。局部离群因子在同一特征仓中表现相对稳健，但其局部邻域密度对采样不均衡与簇边界样本较为敏感，高维空间中的邻域度量稳定性偏弱，叠加多工况切换后密度尺度难以全局对齐，进而拉低了整体的 F1-score。

需重点关注的是，孤立森林在告警阈值与分位信息联动的设定下，把批次差异对得分分布的影响压缩在有限区间，既避免了单一阈值在不同温区与流量状态下产生的系统性偏差，也让滚动再训练能够在不改变污染率先验的前提下保持告警口径稳定。综合对比表明，面对包含漂移、跳变与噪声三类失真的复合场景，本研究的模型在不依赖严格分布假定的条件下，对多源测试数据的非线性与跨尺度结构建立了更强的适配能力，从准确识别早期异常与降低跨批次阈值漂移的角度来看，更契合在线监测与闭环质控的部署需求[12]。

Table 2. Performance comparison of different anomaly detection methods
表 2. 不同异常检测方法性能对比表

方法	精确率	召回率	F1-Score	输入特征	关键参数	阈值策略
孤立森林	0.90	0.94	0.92	时域与频域联合特征经互信息筛选	n_estimators = 100, max_samples = 256, contamination = 0.1	得分大于 0.5 联动样本分位进行轻量校准
局部离群因子	0.85	0.88	0.86	与孤立森林一致的特征向量	n_neighbors = 20, leaf_size = 30	基于得分分布按污染率先验 0.1 设定分位阈值
3σ 法则	0.72	0.68	0.70	单通道幅值与滑窗	滑动窗口长度 1024 点，统计均值与标准差	在均值加减 3 倍标准差之外判定为异常

4.3. 模型鲁棒性验证

本研究把加性高斯噪声注入温度传感器基准语料来检验模型对随机扰动的适应能力，噪声强度设置为相对样本幅值标准差的 5%与 10%，并将预处理、特征仓、阈值联动以及污染率先验保持不变，使评估只反映扰动对判别边界与得分分布的影响。在与前述时间阻塞划分一致的设置下，孤立森林在 5%噪声下的 F1-score 为 0.88，在 10%噪声下为 0.82，而 3σ 法则在相同扰动下降至 0.55，体现出所提方法在随机噪声增大的情形下仍可维持相对稳定的识别能力。进一步观察发现，路径长度驱动的非参数刻画把稀疏毛刺与轻度频谱扩展压缩在浅层隔离中，子采样把局部密度失衡进行稀释，叠加互信息筛选后保留下来的跨尺度统计量与谱量，使异常在多角度随机切分下仍能维持可分性[13]。需重点关注的是，数据门禁中的汉宁权滑动滤波、四分位距软截断以及稳健标准化，已经把幅值漂移与离群脉冲的能量外溢控制在较低水平，从而减少噪声注入对阈值分布的拉动；在线阶段把得分与分位信息进行联动，也把批次差异引起的尾部扩张抑制在稳定区间。相较之下，3σ 法则依赖固定的均值与方差进行刻画，噪声强度抬升会把窗口内的标准差抬高并导致阈值上移，进一步诱发漏报，同时对高频扰动不敏感，难以对跳变与毛刺的组合失真给出一致响应。由此推导，在设备运维与流程质控的常见扰动水平下，所构建模型能够把随机噪声的影响约束在可接受的精度退化范围之内，从工程部署角度为跨批次上线提供更稳定的告警口径与可复现实验评估路径。

5. 结语

本文构建了面向测试数据的异常检测框架，形成统一数据门禁与时频联合特征仓，并以孤立森林为

核心实现非参数化建模与在线阈值联动。基于基准数据集的验证显示，该框架在跨批次与噪声扰动下保持较高的精度与稳定性，满足流式监测与滚动再训练的部署需求。当前工作的局限性在于弱标记稀疏与单一基准语料，未来将扩展至多场景多通道融合，探索自监督表示与概念漂移检测，加强告警解释性与可维护性。

参考文献

[1] 陈向效, 崔鑫, 杜秦, 唐浩耀. 基于机器学习的异常流量检测模型优化研究[J]. 计算机科学, 2024, 51(S1): 982-986.

[2] 严文洁, 张阳. 基于深度学习的试飞数据异常检测方法[J]. 中国科技信息, 2025(23): 35-37.

[3] 尚书一, 李宏佳, 宋晨, 卢至彤, 王利明, 徐震. 互联网服务场景下基于机器学习的 KPI 异常检测综述[J]. 计算机研究与发展, 2025, 62(1): 207-231.

[4] 赵海燕, 吴思雨, 曹健, 陈庆奎. 面向主动学习的异常检测方法: 现状与展望[J]. 小型微型计算机系统, 2025, 47(2): 361-369.

[5] 蔡晓华. 基于机器学习的异常流量检测在智慧审计中的应用研究[J]. 网络安全和信息化, 2025(5): 54-56.

[6] 曾君, 童英华, 王得芳. 基于累积概率波动和自动化聚类的异常检测方法[J]. 计算机应用, 2025, 45(12): 3864-3871.

[7] 杨宏宇, 张豪豪, 胡泽, 成翔. 基于深度学习的网络异常流量检测研究综述[J]. 武汉大学学报(理学版), 2025, 71(2): 159-172.

[8] 陈红松, 刘新蕊, 陶子美, 王志恒. 基于深度学习的时序数据异常检测研究综述[J]. 信息网络安全, 2025, 25(3): 364-391.

[9] 彭易简, 田梦忻, 句媛媛, 吴刘仓. 数据流的异常值在线检测方法[J]. 系统科学与数学, 2025, 46(4): 1311-1324.

[10] 沈夏闰, 李若楠, 张昊田. 基于 CVAE-LSTM 的服务器 KPI 异常检测[J]. 系统工程与电子技术, 2025, 47(3): 1019-1027.

[11] 杨海明, 刘莹. 基于大数据技术的网络流量异常检测算法研究[J]. 黑龙江科学, 2025, 16(10): 62-65.

[12] 王婕婷, 张泽珑, 李飞江, 钱宇华. 基于图神经网络的时序信号异常检测方法[J]. 西北大学学报(自然科学版), 2025, 55(2): 343-354.

[13] 徐登彬, 袁立宁, 吴沛宸, 刘钊. 图神经网络驱动的图异常检测研究综述[J]. 计算机科学与探索, 2025, 19(5): 1123-1138.