DOI:10.20040/j.cnki.1000-7709.2025.20250723

基于SSA-KMIF的船闸人字门监测数据异常检测方法

肖于思[1],马翔宇[2],张燎军[1]

(1. 河海大学水利水电学院,江苏 南京 210098;2. 宿迁市港航事业发展中心,江苏 宿迁 223800)

摘要: 针对孤立森林算法固定阈值导致复杂工况下检测准确度降低的问题,提出一种基于奇异谱分析(SSA)与改进孤立森林(KMIF)的船闸人字门监测数据异常检测方法。利用SSA对监测数据进行分解与重构,分离趋势项和噪声项;引入K-Means++改进孤立森林算法(IF),动态设定不同监测数据集的异常阈值;将噪声项输入改进的孤立森林算法进行训练并检测异常值。以江苏船闸工程下闸首人字门的多测点应力、振动数据为对象进行实例验证。结果表明,提出的奇异谱分析—改进孤立森林方法(SSA-KMIF)在误检率、查准率、查全率和准确率指标上表现优异,具有较高准确性和灵活性,可为船闸人字门健康监测提供可靠技术支撑。

关键词: 奇异谱分析;孤立森林;K-Means++;异常检测;人字门健康监测

中图分类号: TV663^{+}.6 **文献标志码:** A **文章编号:** 1000-7709(2025)09-0119-04

1 引言

船闸人字门作为航运的关键启闭结构,其健康状态对航道通航效率及社会经济安全至关重要。然而,由于长期承受周期性启闭冲击与多源振动耦合作用,人字门易发生形变和磨损,引发安全隐患[1],其健康监测数据常受传感器漂移与机械振动噪声影响,呈现高频非平稳特性和局部稀疏异常,严重影响结构状态评估的可靠性。因此,检测和剔除人字门监测数据中的异常值,提升数据质量十分必要。孤立森林算法[2]因其无监督特性及较低的时间复杂度,在异常检测领域展现出潜力。徐浩等[3]利用孤立森林算法对供水企业取水量数据进行异常检测,结果相比传统箱线图法和k近邻算法具有更高的稳定性;赵新华等[4]结合SSA和孤立森林算法,避免了原始数据的趋势性干扰对检测结果的影响,但其在处理变化剧烈的数据时表现欠佳。另外,人字门在低频水流冲击和高频机械振动影响下频繁启闭运行,固定异常阈值无法适用不同构件的复杂工况,导致异常检测准确度下降。考虑到孤立森林在异常决策阶段仅评判正常与异常两类样本,将K-Means++聚类算法[5]与孤立森林算法结合,利用聚类算法动态划分异常阈值,可以更加灵活地适用于不同数据集。本文提出一种基于SSA-KMIF的异常检测方法,首先利用SSA对监测数据进行预处理,降低耦合干扰;然后,利用K-Means++聚类优化孤立森林算法的异常阈值,设计双簇聚类自适应阈值划分机制,提高异常检测的准确性;最后以江苏船闸工程下闸首人字门的多测点应力、振动数据为对象进行实例验证,证明了该方法的可行性和优越性。

2 理论与方法

2.1 奇异谱分析

SSA是一种适用于非线性、非平稳信号的分析方法,通过构造轨迹矩阵并进行奇异值分解,将时间序列分解为趋势和噪声分量,具有去噪、趋势提取功能[6]。将长度为N的一维时间序列嵌入到m维空间中,形成轨迹矩阵$\boldsymbol{X}$并通过奇异值分解为几个基本矩阵。对于每一个矩阵,通过对角平均方法可生成一个重构序列S,它反映了原始时间序列中特定的特征分量,可以表示为:

$$S_i(t)=\frac{1}{N-t+1}\sum_{j=1}^{N-t+1}X_{ij} \tag{1}$$

收稿日期: 2025-04-25,**修回日期:** 2025-05-25

基金项目: 江苏省交通运输科技项目(2020QD28)

作者简介: 肖于思(2000-),女,硕士研究生,研究方向为水工金属结构在线监测及健康诊断,E-mail:231302050011@hhu.edu.cn

通讯作者: 张燎军(1962-),男,教授、博导,研究方向为水工结构数值模拟及安全监测,E-mail:ljzhang@hhu.edu.cn

式中，$S_i(t)$为第 i 个重构序列；t 为时间索引；X_{ij} 为轨迹矩阵的元素。

选择累积贡献率为 85%的前 k 个成分重构时间序列，以保留主要特征并去除噪声。

2.2 孤立森林算法

孤立森林算法是一种基于集成学习的异常检测算法，属于无监督机器学习算法[7]，其原理见图 1。孤立森林算法的主要步骤为：①随机选择样本子集作为树的根节点；②利用递归的方式构建孤立树，直到每个叶子节点只有一个样本或达到预设的树高度，重复此过程，构造孤立森林；③计算每个样本的平均路径长度及异常分数；④根据异常分数判断每个样本是否为异常点。

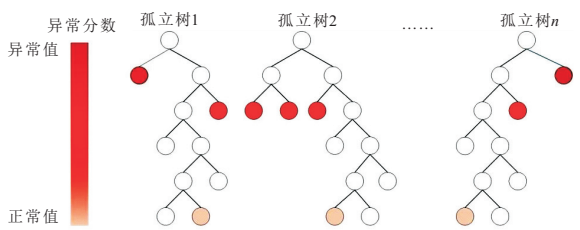

图 1 孤立森林算法原理图

Fig. 1 Principle diagram of the isolation forest algorithm

异常点因分布稀疏且离密度高的群体远，可较早被孤立。因此根据离群点到根节点的路径长度来寻找异常值，其通常具有较短的路径长度，而正常值则在树的深处。路径长度 $h(x)$和异常分数 $s(x,m)$为：

$$s(x,m)=2^{-\frac{E[h(x)]}{c(m)}} \tag{2}$$

$$c(m)=2H(m-1)-2(m-1)/m \tag{3}$$

$$H(m-1)=\ln(m-1)+\gamma \tag{4}$$

式中，$s(x,m)$为 m 个样本中获得的 x 的异常分数，取值范围为[0,1]；$E(h(x))$为数据 x 在孤立树集合中路径长度的均值；$c(m)$为用于标准化查询数据 x 的路径长度 $h(x)$；$H(m-1)$为调函数；γ 为欧拉常数。

2.3 SSA-KMIF 异常检测方法

人字门监测数据中仅包含正常值和异常值，其中异常值数量通常较少。为有效识别这些异常值，引入 K-Means++聚类算法，将异常分数分为正常簇、异常簇两个簇，其中异常簇数据量较少。

传统的 K-Means 聚类算法在初始聚类中心的选择上存在随机性，这可能导致算法收敛到局部最优解，影响聚类结果的准确性和稳定性。K-Means++算法采用一种距离最大化的策略选取初始聚类中心，使得距离已选择聚类中心较远的数据点有更高的概率被选为新的聚类中心。这种选取策略减少了初始聚类中心的随机性，收敛速度更快，聚类结果更稳定，能够有效避免局部最优问题。

孤立森林算法通过随机生成的分割树结构，利用路径长度差异识别异常数据，增强算法的稳定性和抗噪能力。然而，其完全随机性可能导致异常分数不稳定，且在不同工况下，固定阈值易产生误判。为适应人字门多工况运行的情况，结合 K-Means++聚类改进孤立森林算法，设计双簇聚类自适应阈值划分机制。具体步骤为：①输入时序数据，采用 SSA 对其分解并提取趋势项；②将提取趋势项后的剩余项输入孤立森林算法，进行训练和异常检测，计算样本的异常分数；③采用 K-Means++算法对异常分数进行双簇聚类（$K=2$），利用距离最大化的策略选取初始质心位置[5]，使质心分别靠近高密度区和低密度区的数据分布中心，更新簇划分至质心稳定；④定义样本数量较多的簇为正常簇 C_1，另一簇为异常簇 C_2，取异常簇中的最低分数作为异常阈值 T，即 $T=\min(C_2)$，当新样本的异常分数超过 T 时，标记为异常样本。

3 实例验证

3.1 实例数据集

选取江苏船闸工程[8]下闸首人字门右门叶作为研究对象，进行实例验证。该人字门设置了 23 个监测点，包括运行姿态监测点、主梁应力监测点及背拉杆应力监测点等。数据集涵盖人字门 5 次启闭过程中的应力和振动加速度数据，主要监测内容包括主梁腹板、横隔板、上下翼缘、背拉杆、AB 杆和推拉杆的应力，以及门体的振动加速度等力学性能指标。监测点的布置情况见图 2。

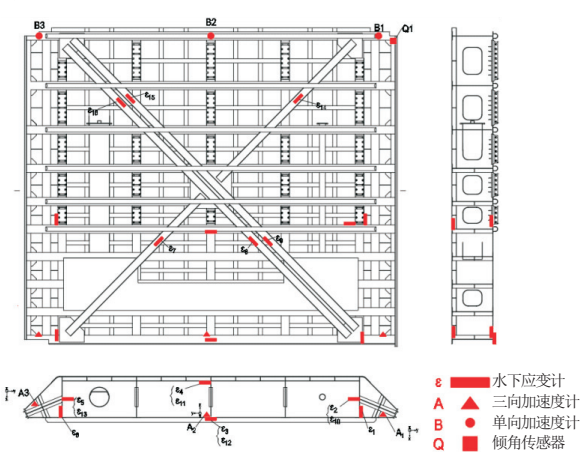

图 2 监测点布置图

Fig. 2 Layout of the monitoring points

3.2 基于 SSA 的人字门监测时序数据预处理

人字门的监测时序数据呈现非平稳特性，利

用 SSA 对其分解和重构，能够提取趋势成分，降低非平稳性的影响。以人字门 A 杆侧表面应变传感器采集的某组应力时序数据为例，设定窗口长度为 200 个采样点，并选取累积贡献率达到 85%的重构分量作为趋势项进行重构。经 SSA 处理后的应力时序见图 3，SSA 能够有效分离趋势项和噪声项，为后续异常检测提供更准确的样本。

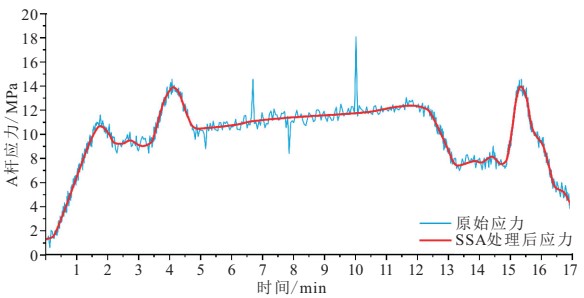

图 3　A 杆原始应力和 SSA 处理后的应力

Fig. 3　Original stress of A pole and stress after SSA

3.3　基于 SSA-KMIF 的监测数据聚类

通过结合孤立森林算法与 K-Means＋＋聚类算法，自适应划分异常阈值，改进孤立森林算法在复杂工况下固定阈值的检测局限。以右门叶背拉杆应力时序数据为例，其聚类结果见图 4。K-Means＋＋算法将计算出异常分数的数据聚为两簇，其中异常点聚在右侧。通过分析聚类结果，取异常簇中的分数下界 0.445 作为背拉杆应力时序数据的异常阈值。结果表明，SSA-KMIF 方法不仅可以准确地识别出异常值，还能够根据不同的监测数据集进行动态调整，自适应地为不同工况下的不同构件划分异常阈值，具有灵活性和适应性。

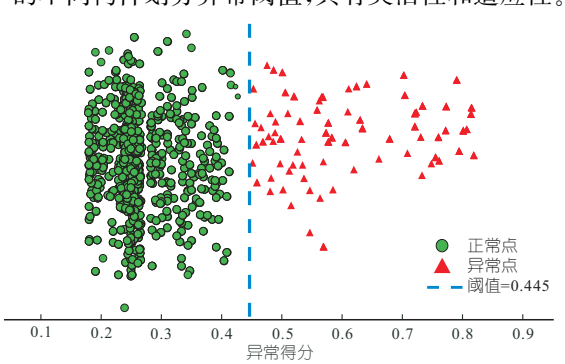

图 4　应力时间序列聚类结果图

Fig. 4　Clustering results of stress time series

3.4　试验对比分析

3.4.1　异常数据设置

为全面评估所提方法在人字门健康监测数据集中异常值检测的性能，结合该项目技术专家的工程经验，针对人字门各部位原始数据的复杂性及其承受荷载的不确定性，采用随机添加异常值的方式进行模拟试验，并分别设置 1%、5%、10%的异常率，以评估在不同异常率条件下的检测性能。

3.4.2　评价指标

试验结果采用误检率(F)、查准率(P)、查全率(R)及准确率(A)等常用评价指标进行量化分析：

$$F = F_{FP}/(T_{TN} + F_{FP}) \tag{5}$$

$$P = T_{TP}/(T_{TP} + F_{FP}) \tag{6}$$

$$R = T_{TP}/(T_{TP} + F_{FN}) \tag{7}$$

$$A = (T_{TP} + T_{TN})/(T_{TP} + F_{FP} + F_{FN} + T_{TN}) \tag{8}$$

式中，F_{FP} 为实际正常但被识别为异常的样本数；T_{TN} 为实际正常且被正确分类的样本数；T_{TP} 为实际异常且被正确识别的样本数；F_{FN} 为实际异常但未被识别出的样本数。

3.4.3　鲁棒性试验

为验证算法的鲁棒性能，采用人字门 23 个监测点的真实数据，并依据异常率添加异常值。取异常率为 1%、5%、10%时的评价指标平均值作为衡量算法鲁棒性的指标，结果见表 1。由表 1 可知，本文算法在不同异常率下均展现出良好的鲁棒性。在低异常率(1%)时，误检率极低，查准率和查全率接近 0.9，准确率高达 0.974，表明算法能够高效且精准地检测异常值。即使在异常率升至 10%时，算法仍能保持较高的准确率(0.858)，说明其对不同异常情况均有较好的鲁棒性。

表 1　不同异常率的指标值

Tab. 1　Indicator values at different anomaly rates

异常率	F	P	R	A
1%	0.10	0.82	0.85	0.974
5%	0.08	0.78	0.80	0.925
10%	0.15	0.72	0.74	0.858

3.4.4　消融性试验

为验证算法各个模块对整体性能的贡献，设计消融性试验，逐步移除关键模块，评估各模块对异常检测性能的影响，从而证明各模块的有效性及其协同作用。以人字门的 23 个测点的应力与振动数据集为基础，进行 IForest、K-Means＋＋算法以及它们二者组合(IF＋＋)与 SSA-KMIF 算法的对比分析。各算法的评价指标值见表 2。

表 2　各部分算法的评价指标值

Tab. 2　Evaluation metric values for different parts of the algorithms

检测算法	F	P	R	A
IForest	0.13	0.70	0.70	0.875
K-Means＋＋	0.18	0.75	0.78	0.862
IF＋＋	0.09	0.76	0.75	0.891
SSA-KMIF	0.05	0.83	0.82	0.925

由表 2 可知，SSA-KMIF 算法表现出更好的性能和更稳定的检测效果，其中误检率与查准率显著优于对比组。在检测过程中，利用 SSA 提取趋势项后使真正的异常更突出，减少误判和漏报

的风险。仅使用孤立森林和 K-Means＋＋二者的简单叠加(IF＋＋)难以有效区分工况切换与真实异常。相比之下,SSA-KMIF 算法通过先分离数据的趋势与噪声,再根据不同构件的不同工况自适应地调整异常检测的阈值,具有更高的准确性和灵活性,表明其对人字门监测数据的应用具有实践价值。

3.4.5 优越性试验

为进一步验证算法在人字门健康监测异常检测中的优越性,将其与四分位数法[9]、局部异常因子(LOF)[10]和 K 最近邻(KNN)这三种异常检测方法进行对比分析。各方法的评价指标值见表 3。由表 3 可知,在异常率与异常分布一致的情况下,SSA-KMIF 在误检率、查准率、查全率及准确率 4 项指标值均优于其他算法,表明该方法在人字门健康监测数据检测中的优越性,能够更全面、更准确地检测出异常值,可为船闸人字门健康监测提供更可靠的技术支持。

表 3 各算法评价指标值

Tab. 3 Evaluation metric values for different algorithms

检测算法	*F*	*P*	*R*	*A*
四分位数法	0.21	0.60	0.58	0.862
LOF	0.18	0.72	0.70	0.885
KNN	0.16	0.68	0.66	0.874
SSA-KMIF	0.08	0.78	0.80	0.925

以人字门 A 杆侧表面应变传感器所采集的某组应力时间序列为例,通过人工添加 10 个异常数据的方式(图 5),详细直观地分析 SSA-KMIF 算法的优越性。各算法的检测结果见图 6(a)～(d)(图中不同标识分别代表正确识别的异常值、错误识别的异常值以及漏识别的异常值)。由图 6 可知,这 4 种方法都能够检测出离群程度较高的异常值,但四分位数法和 KNN 法分别漏检 3 个、2 个异常值,LOF 法则出现连续的 4 个误检。相较于其他三种算法,SSA-KMIF 算法的漏判、误判数量明显减少,表现出更优越的检测性能。

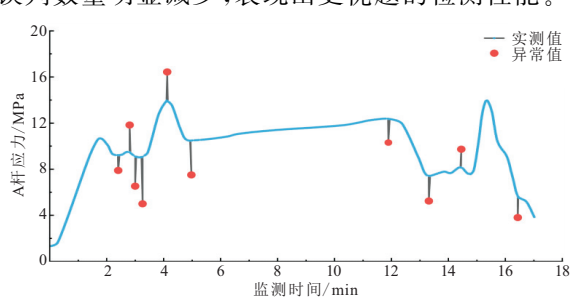

图 5 人字门 A 杆应力时序曲线

Fig. 5 Stress time series curve of A pole of the miter gate

4 结论

a. 本文提出基于 SSA-KMIF 的异常检测方法,利用 SSA 分解、重构监测时序数据,降低复杂工况干扰,结合 K-Means＋＋优化孤立森林算法的异常阈值,实现各构件在不同工况下阈值的动态调整,能够精准识别异常值,可靠性高,灵活性强。

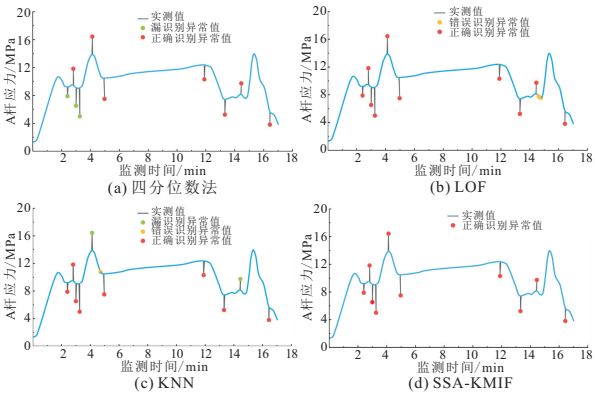

图 6 各方法异常检测效果图

Fig. 6 Detection effect of anomalies using different methods

b. 对比试验表明,SSA-KMIF 在误检率、查准率、查全率和准确率等指标上表现优异,能有效区分工况切换与真实异常,为人字门状态评估提供高质量数据支持,为船闸健康监测提供了重要参考。

参考文献:

[1] 万可,喻瑾,于俊生. 航运枢纽工程船闸闸门启闭机设计分析[J]. 珠江水运,2024(3):82-84.

[2] LIU F T, TING K M, ZHOU Z H. Isolation forest [C]//2008 Eighth IEEE International Conference on Data Mining, 15-19 December 2008, Pisa, Italy. 2008: 413-422.

[3] 徐浩,刘怀利,瞿暄. 基于孤立森林的取水数据异常值检测[J]. 水电能源科学,2024,42(9):29-32,59.

[4] 赵新华,范振东,何宇,等. 基于数据重构与孤立森林法的大坝自动化监测数据异常检测方法[J]. 中国农村水利水电,2021(9):174-178.

[5] ARTHUR D, VASSILVITSKII S. K-Means＋＋: The advantages of careful seeding [C] // Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007.

[6] 闫孟婷,陶湘明,王胜军,等. 基于 SSA-LSTM 模型的水电站能效综合评价方法[J]. 水电能源科学,2024,42(2):177-182.

[7] LIU F T, TING K M, ZHOU Z-H. Isolation-based anomaly detection[J]. ACM transactions on knowledge discovery from data, 2012, 6(1): 1-39.

[8] 张燎军. 船闸人字门全工况仿真模拟与计算理论深化研究[R]. 南京:河海大学,2025.

[9] 杭震,彭浩,曹文卓,等. 船闸浮式系船柱运行状态检测方法研究[J]. 机电工程技术,2022,51(12):156-159,193.

[10] 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.

(下转第 105 页)

(自然科学与工程技术版),2024,57(2):174-185.
[7] 江星星,宋秋昱,杜贵府,等.变分模式分解方法研究与应用综述[J].仪器仪表学报,2023,44(1):55-73.
[8] 娄革伟,郑永煌,陈均,等.混合多策略改进的蜣螂优化算法[J].计算机工程与应用,2024,60(24):97-109.
[9] 彭继慎,郝茗,宋立业,等.基于 TSSA-SVR 算法的 TBM 掘进速度预测[J].辽宁工程技术大学学报(自然科学版),2023,42(5):634-640.
[10] 杜庆峰, 张双俐, 张晨曦, 等. 基于均值滤波去噪和 XGBoost 算法的泥水平衡盾构掘进速度预测方法[J]. 现代隧道技术, 2022, 59(6): 14-23.
[11] 杨辉斌, 郑德仁, 王贺龙, 等. 基于 GA-SVR 的管网异常漏损检测[J]. 水电能源科学, 2024, 42(3): 133-136, 53.
[12] 柏晓鹏,南瑞川,杜甫,等.基于随机森林的地下水溶质运移替代模型研究[J].水电能源科学,2024,42(11):60-63.
[13] 李晨阳, 郑东健. 基于多层次数据处理的 NGO-XGBoost 大坝变形预测模型及其应用[J]. 水电能源科学, 2023, 41(11): 77-81.
[14] 杨爽,薛晔.基于 GGRA-GPR 模型的洪涝灾害直接经济损失预评估[J].水电能源科学,2023,41(10):67-71.
[15] 张建明,侍克斌,贾运甫,等.基于 VMD 与加权 RF 的 TBM 掘进速度预测 SHAP 解释模型[J].隧道建设(中英文),2024,44(5):1012-1028.

Prediction Model of Shield Tunneling Speed Based on VMD-DBO-Stacking Ensemble Learning

DENG Zi-ang, ZHANG Yu-xian, ZHANG Ji-xun

(College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China)

Abstract: Addressing the issues of single model algorithm, low accuracy, and poor generalization in existing shield tunneling speed prediction methods, this study proposes a shield tunneling speed prediction approach to improve prediction accuracy based on Variational Mode Decomposition (VMD), Dung Beetle Optimizer (DBO), and Stacking ensemble learning. Firstly, to obtain more effective data, VMD is applied to decompose and reconstruct the original data to obtain denoised construction parameter data for subsequent model prediction. Secondly, based on the ensemble learning strategy, Support Vector Regression (SVR), Random Forest (RF), and Extreme Gradient Boosting (XGBoost) models are selected as base learners, while Gaussian Process Regression (GPR) is chosen as the meta-learner to construct a Stacking ensemble learning prediction model with higher prediction accuracy and stronger generalization ability. Thirdly, to further enhance prediction accuracy, DBO is employed to optimize the hyperparameters of the ensemble learning model. Finally, this prediction method is applied to the shield tunneling construction of a water diversion tunnel project in Henan Province and compared with other prediction methods. Compared to other single models (SVR, RF, XGBoost), the results indicate that the proposed method achieves higher prediction accuracy, with average accuracy improvements of 7.76%, 6.70%, and 4.97%, respectively, providing a new approach for shield tunneling speed prediction.

Key words: shield tunneling machine; tunneling speed; variational mode decomposition; Dung Beetle Optimizer; Stacking ensemble learning

(上接第 122 页)

Anomaly Detection Method for Ship Lock Miter Gate Monitoring Data Based on SSA-KMIF

XIAO Yu-si[1], MA Xiang-yu[2], ZHANG Liao-jun[1]

(1. College of Water Conservancy and Hydropower Engineering, Hohai University, Nanjing 210098, China;
2. Suqian Port and Shipping Development Center, Suqian 223800, China)

Abstract: To address the issue of reduced detection accuracy under complex working conditions due to the fixed threshold of the isolation forest algorithm, an anomaly detection method for ship lock miter gate monitoring data based on singular spectrum analysis (SSA) and an improved isolation forest (KMIF) is proposed. The SSA is employed to decompose and reconstruct the monitoring data, and separate the trend and noise components. The isolation forest algorithm is improved by incorporating K-Means++ clustering to dynamically set anomaly thresholds for different monitoring datasets. The noise component is then fed into the improved isolation forest algorithm for training and anomaly detection. Taking the stress and vibration data from multiple measuring points of the lower lock miter gate in Jiangsu ship gate project as an example for validation, the results show that the proposed SSA-KMIF method performs excellently in terms of false positive rate, precision, recall ratio, and accuracy. It demonstrates high accuracy and flexibility, which provides a reliable technical support for health monitoring of ship lock miter gates.

Key words: singular spectrum analysis; isolation forest; K-Means++; anomaly detection; health monitoring of miter gates