

DOI:10.19853/j.zgjsps.1000-4602.2023.11.009

# 基于K-shape聚类的连续水位监测数据异常检测方法

何黎<sup>1</sup>, 陈磊<sup>2</sup>, 纪莎莎<sup>1</sup>, 陈泽伟<sup>1</sup>, 宋晨曦<sup>1</sup>

(1. 上海市城市建设设计研究总院<集团>有限公司, 上海 200125; 2. 重庆大学 环境与生态学院, 重庆 400045)

**摘要:** 受制于排水管网监测起步较晚、监测环境恶劣等因素,目前城市排水管网运行数据质量不容乐观,直接影响其有效应用及价值挖掘。而异常检测作为数据有效应用的第一步,在排水系统中尚未有效开展。以K-shape聚类算法为基础,提出了一种排水监测数据异常检测流程。首先,对特征序列进行提取并进行聚类分析,以确定描述时间序列集合的整体特征或平均特征的序列,从而降低异常检测的误报和漏报率。然后对识别的异常序列进行整体性判断,以提高异常检测算法的查全率。结果表明,基于K-shape的排水监测数据异常检测算法的查全率和查准率分别可以达到0.891 7和0.812 7。此外,与暴力算法(BF)的对比显示,采用固定长度的时间序列切分方式会导致误报和漏报率增加,其效果劣于K-shape聚类算法。

**关键词:** K-shape聚类; 排水管网数据; 异常检测; 时间序列

**中图分类号:** TU992 **文献标识码:** A **文章编号:** 1000-4602(2023)11-0056-06

## Abnormal Detection of Continuous Water Level Monitoring Data Based on K-shape Clustering

HE Li<sup>1</sup>, CHEN Lei<sup>2</sup>, JI Sha-sha<sup>1</sup>, CHEN Ze-wei<sup>1</sup>, SONG Chen-xi<sup>1</sup>

(1. Shanghai Urban Construction Design and Research Institute, Shanghai 200125, China;

2. College of Environment and Ecology, Chongqing University, Chongqing 400045, China)

**Abstract:** Due to factors such as the late start of drainage network monitoring and the harsh monitoring environment, the current quality of urban drainage network operation data is not optimistic, which directly affects its effective application. However, abnormal detection, as the first step in the effective application of data, has not been effectively carried out in the drainage system. Based on the K-shape clustering algorithm, an abnormal detection process of drainage monitoring data was proposed. First, the feature sequence was extracted and clustered to determine the sequence describing the overall feature or average feature of the time series, thereby reducing the false positive and false negative rates of abnormal detection. Then, a holistic judgment was made on the identified abnormal sequences to improve the recall rate of abnormal detection algorithms. The experimental results showed that the recall rate and precision rate of the drainage monitoring data abnormal detection algorithm based on K-shape could reach 0.891 7 and 0.812 7 respectively. In addition, through a comparative study with the brute force algorithm (BF), it was found that the use of a fixed-length time series segmentation method would lead to an increase in false positive rates, and its effect was inferior to the K-shape clustering algorithm.

通信作者: 纪莎莎 E-mail: jishasha@sucdri.com

**Key words:** K-shape clustering; urban drainage network data; abnormal detection; time series

城市排水管网的全面建设和安全稳定运行是保障和改善人居环境的重要途径<sup>[1-2]</sup>。2020年住建部发布的《关于加强城市地下市政基础设施建设的指导意见》指出了“提高地下市政基础设施安全韧性,补齐排水短板,推进排水管网改造和建设”的重要性。《城镇水务2035年行业发展规划纲要》也强调了“城镇排水设施信息化建设以及智能化管理”的必要性。排水管网设施监测数据的挖掘分析能够提高其在城市内涝预警和城市排水泵站优化调度中的作用<sup>[3]</sup>,是智慧水务建设过程中的必要方式。根据监测任务的具体要求,需做如下考虑:①综合考虑监测指标、测量范围、测量精度、成本效益以及运维成本等因素,从而进行排水监测设备的合理选型,是提高数据质量的前提;②确保监测设备位置适宜、管道稳固、连接牢靠、通信顺畅,是确保数据合理准确的基础;③定期对监测设备进行检查和维护,及时处理设备故障等,是维护数据有效连续的保障。因此,监测设备的专业化选型、准确安装、可靠运维对保障监测数据质量具有重要意义。但是,由于城市排水管网深埋在地下、传输介质复杂,以及数据传输网络波动等原因,导致排水监测数据中含有大量异常,使用人工识别的方式进行异常检测效率较低<sup>[4]</sup>。这些异常会对监测数据在水力建模和数据建模中的应用造成很大的影响<sup>[5]</sup>。因此,开发高效的排水监测数据的异常自动识别技术对于我国排水行业的智能化发展具有重要意义。

近年来诸如自回归模型分析法(AR)、密度聚类(DBC)、支持向量机(SVM)等机器学习方法成功应用到交通、电力、网络安全等领域的异常识别与分析中<sup>[6-8]</sup>。同时在供水领域,此类机器学习方法也同样应用于高频压力异常模式识别<sup>[9]</sup>、用水的异常识别<sup>[10]</sup>以及数据的异常清洗<sup>[11]</sup>等场景,而对于城市排水管网数据的异常识别方面的系统研究尚未有效开展。一般来说,数据异常识别多基于预测策略,其需要大量历史数据支撑,而排水领域的监测系统建设起步较晚,且由于其监测环境的复杂性导致现阶段采集数据质量不高,难以直接用于搭建高精度预测模型,从而进一步限制了排水管网运行数据异

常检测的发展。前池水位数据作为排水系统监测数据的重要组成部分,决定了排水泵的启闭状态,进而影响整个排水系统的运行。因此,选择泵站前池的液位变化数据作为研究对象开展先期研究。

针对上述问题,拟构建基于K-shape聚类算法的排水数据异常检测流程,通过无监督学习方式研究上海某排水泵站前池液位监测数据,分析K-shape聚类异常识别模型在数据异常检测中的应用成效,并与传统的暴力算法(BF)模型结果进行对比。

## 1 研究方法

### 1.1 K-shape聚类算法

K-shape聚类算法是由Paparrizos和Gravano提出的一种基于形状的时间序列聚类方法,用于将相似的时间序列分为同一类<sup>[12]</sup>。其基本思想是通过将时间序列进行自动化预处理和转换,然后使用基于形态的距离(SBD)来计算时间序列之间的相似度,从而实现聚类<sup>[13]</sup>。K-shape聚类算法的优点在于它能够有效地处理长时间序列,且能够在处理具有不同长度的时间序列时提供较高的准确度。SBD距离的数学表达式见文献<sup>[12]</sup>,K-shape通过互相关方法来计算两个时间序列 $\vec{x}$ 和 $\vec{y}$ 的距离,当 $SBD(\vec{x}, \vec{y})=0$ 时,则说明 $\vec{x}$ 与 $\vec{y}$ 序列的形状完全相同。基于K-shape聚类的异常检测流程见图1。

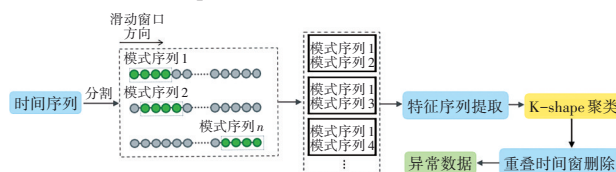


图1 基于K-shape聚类的异常检测流程

Fig.1 Flow chart of abnormal detection based on K-shape clustering

研究中先基于样本数据通过时间窗口滑动<sup>[14]</sup>进行重采样,扩大时间序列特征,经过特征序列提取实现模型序列的提取,最终经过K-shape聚类算法对样本数据进行聚类分类,将聚类样本中数据少的定义为异常数据类,具体过程如下:

① 对于时间序列 $T = t_1, t_2, \dots, t_n$ ,通过设置合适的滑动窗口长度 $m$ ,按照滑动步长为1进行重采样,得到形状为 $(n - m + 1, m)$ 的序列 $D$ 。

② 将序列 $D$ 经过中心序列算法计算生成长度为 $m-1$ 的若干个特征序列,每个特征序列样本的形状为 $(m-1, m, 1)$ ,以此构成序列集合 $R$ 。

③ 对序列集合中的每个子序列样本 $R_i$ 使用K-shape聚类方法进行聚类,在一定的范围内选择合适的聚类数量,并通过对每个类别数量计算轮廓系数以确定最优聚类数量,从而获得最优类别数量下的聚类中心,并以此作为子序列样本 $R_i$ 的模式序列;重复此过程以获得序列集合 $R$ 中每个子序列样本的模式序列,以此构成模式序列集合 $M$ 。

④ 对于模式序列集合 $M$ 中的每个样本 $M_i$ 同样使用K-shape聚类方法进行聚类,聚类类别范围为 $[2, 6]$ ,通过对每个类别数量计算轮廓系数以确定最优聚类数量。

⑤ 序列集合 $M$ 聚类的少数类即为异常类别,并获得其类别中每个样本在原始序列中的位置,在设置合适决策长度judge\_length的基础上,获取每个子异常序列的位置。

## 1.2 暴力算法

暴力算法(BF)是一种普通的模式匹配算法<sup>[15]</sup>,其思想是,通过长度为 $m$ 的滑动窗口在长度为 $n$ 的时间序列上获取 $n-m+1$ 个子序列后,得到大小为 $n-m+1$ 的子序列集合,在这个集合中,将所有的子序列两两相比较,将欧氏距离计算值最大的子序列输出为异常子序列。其计算流程如表1所示。

表1 BF算法流程

Tab.1 Flow chart of BF algorithm

输入	时间序列 $T = t_1, \dots, t_n$ , 滑动窗口长度 $m$
输出	异常子序列位置loc, 异常子序列最邻近距离dist
Step1	设置当前最邻近距离best_so_far_dist = 0, 当前异常子序列位置best_so_far_loc = NaN
Step2	For $p = 1$ 到 $n - m + 1$ Do
	设置当前最邻近距离nearest_dist = infinty
	For $q = 1$ 到 $n - m + 1$ Do
	If $ p - q  \geq m$ Do
	If $\text{Dist}(t_p, \dots, t_{p+m-1}, t_q, \dots, t_{q+m-1}) < \text{nearest\_dist}$ Then
	nearest_dist = $\text{Dist}(t_p, \dots, t_{p+m-1}, t_q, \dots, t_{q+m-1})$
	End For
Step3	If nearest_dist > best_so_far_dist Then
	best_so_far_dist = nearest_dist; best_so_far_loc = $p$
End For	
输出best_so_far_dist, best_so_far_loc	

## 1.3 异常检测效果评价指标

当使用异常检测算法处理液位监测数据时,可

以将算法对每个时间点数据的判断结果与真实情况下的数据状态的组合划分为4种情况:真正例TP(true positive)、假正例FP(false positive)、真反例TN(true negative)以及假反例FN(false negative)。TP指算法正确预测出数据表示的是异常状态的次数;FP指算法错误地预测出数据表示的是异常状态的次数,而实际上数据表示的是正常状态;TN指算法正确预测出数据表示的是正常状态的次数;FN指算法错误地预测出数据表示的是正常状态的次数,而实际上数据表示的是异常状态。同时,使用混淆矩阵来计算各种性能指标,如表2所示。

表2 异常检测结果混淆矩阵

Tab.2 Confusion matrix of abnormal detection

真实数据状态	算法判断该时刻数据情况	
	异常	正常
异常	TP(真正例)	FN(假反例)
正常	FP(假正例)	TN(真反例)

采用查准率(Precision)、查全率(Recall)、正确率(Accuracy)、F1值(F1-score)评估算法对数据异常与否判断的准确率<sup>[16]</sup>,计算方法见式(1)~(4)。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (1)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (3)$$

$$F1 = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

查准率是指算法对于所有判断为异常的数据中,真实为异常的数据占比,它反映了算法在判断异常数据时的准确度;查全率是指所有真实异常数据中,被算法正确判断为异常的数据占比,它反映了算法在检测异常数据时的召回率;正确率是指算法对所有数据的判断结果均正确的数据占比,它反映了算法对所有数据的判断能力;F1值是在同样重视查准率和查全率的基础上,考虑算法在异常检测方面的整体表现的指标。这些指标的取值均在0~1之间,越接近1则表示算法对异常检测越准确。

## 2 案例分析

### 2.1 案例数据介绍

使用带有标签的2020年10月18日—11月18日的某城市排水泵站前池液位数据进行测试,该时段数据采样间隔为5 min,共获得9 216个时刻的连



续液位数据。其实际数据及数据标签如图2所示。其中异常数据总量为3 454个,按照时段划分,异常时段共计7段,a~g段的异常液位数据量分别为52、121、219、730、612、123和1 597个。

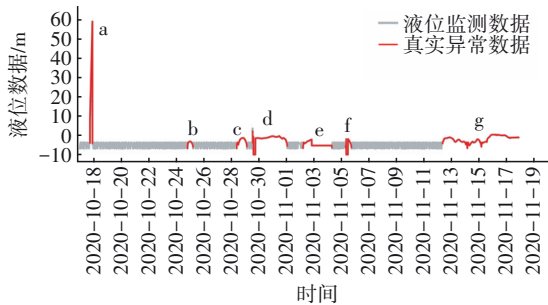


图2 前池液位监测数据及真实异常标注

Fig.2 Forebay water level monitoring data and abnormal marking

## 2.2 数据归一化

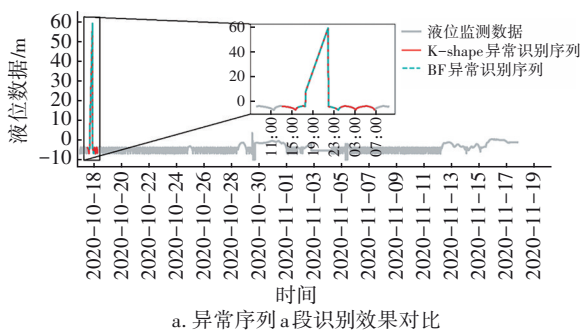
数据归一化是机器学习中的一项挖掘数据的基础工作,其可以缩小数量之间的相对关系以及消除指标之间的量纲影响,从而解决数据指标之间的可比性。使用标准分数归一化(ZSN)进行数据归一化,其计算如式(5)所示。

$$Y = \frac{X - \mu}{\sigma} \quad (5)$$

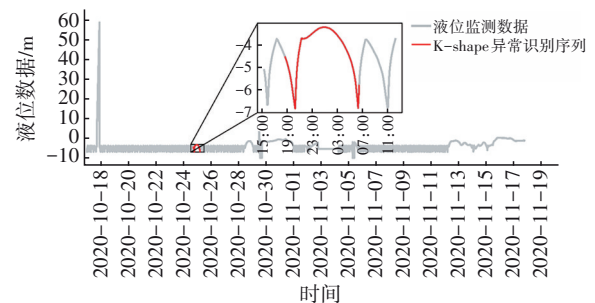
式中: $Y$ 为归一化时间序列; $X$ 为原始时间序列; $\mu$ 为原始时间序列均值; $\sigma$ 为原始时间序列标准差。经过标准分数归一化处理后的数据符合标准正态分布。

## 2.3 结果分析

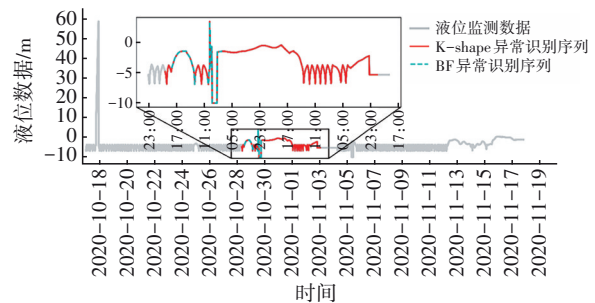
同时使用 K-shape 聚类算法和 BF 算法对同期的泵站前池液位监测数据进行异常检测,以评估这两种方法在检测城市排水管网运行数据中的异常情况时的效果,结果如图3所示。可以看出,在对案例数据进行异常检测时,K-shape 聚类算法比BF算法更加有效。



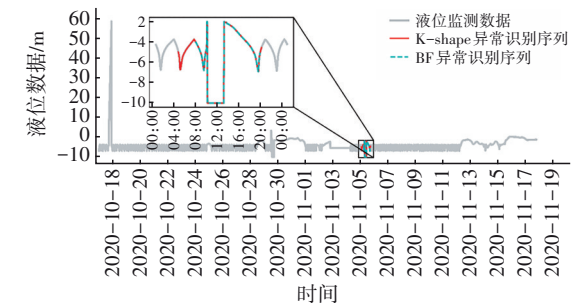
a. 异常序列a段识别效果对比



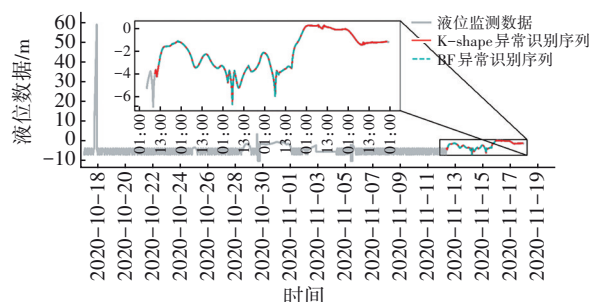
b. 异常序列b段识别效果对比



c. 异常序列c~e段识别效果对比



d. 异常序列f段识别效果对比



e. 异常序列g段识别效果对比

图3 K-shape算法与BF算法在液位数据异常识别中的对比

Fig.3 Comparison of K-shape and BF methods in abnormal identification of water level data

从整体上看,K-shape 聚类算法能够有效检测出所有的异常时间段,而BF算法仅能检测出部分。在每个异常时间段的细节方面,K-shape 算法也比BF算法更加优秀。例如对于异常时间段g,BF算法只能检测出其中的1/4,而K-shape 算法则能够完全检测出。另外,BF算法也只能检测出异常时间段f的3/5,而K-shape 算法则能够完全检测出。分析认

为,BF算法的漏报率较高的原因在于,它采用了固定长度的滑动窗口来计算不同数据段之间的距离,人为设定的长度大小缺乏理论基础。当滑动窗口长度较小,而异常时间段序列的长度较大时,过长的异常序列被滑动窗口切分成较多的片段,由于异常片段间的距离较小,导致异常序列后段不容易被检测出。当滑动窗口长度过大,而异常时间段序列的长度较小时,由滑动窗口切分出的片段所包含的正常数据远多于异常数据,从而导致异常片段被掩盖,如异常数据段b无法被BF算法检测出。

因此,BF算法在时间序列异常检测中的应用受到滑动窗口长度选择的限制,而K-shape聚类算法则通过使用自适应的滑动窗口长度来提取模式序列,并将这些模式序列用于异常检测,有效地避免了固定滑动窗口长度所造成的影响,使得K-shape聚类算法在时间序列异常检测中的应用更加灵活。K-shape聚类算法使用自适应的滑动窗口长度来提取模式序列,这有助于避免固定滑动窗口所造成的影响,但同时也略微提高了异常检测的误报率。从图3(a)和图3(c)中可以看出,K-shape聚类算法会将较多的正常数据识别为异常数据。然而,微增的误报率仅会增加数据异常修复过程中的工作量,而较高的漏报率则会影响数据异常修复的正确率和有效性,从而降低已修复数据的应用价值。

为了进一步评估K-shape聚类算法和BF算法在城市排水管网运行数据异常检测中的效果,表3展示了这两种方法的异常检测评估指标的情况。

表3 K-shape聚类算法及BF算法的异常检测评价指标  
Tab.3 Abnormal detection evaluation indicators of K-shape clustering algorithm and BF algorithm

指标	K-shape	BF	指标	K-shape	BF
TP	3 080	3 020	Precision	0.812 7	0.754 8
FP	710	981	Recall	0.891 7	0.862 1
TN	5 052	4 732	Accuracy	0.882 4	0.841 1
FN	374	483	F1-score	0.850 4	0.804 9

从表3可以看出,K-shape聚类算法的4种评价指标值均高于0.8,并且明显优于BF算法。这表明K-shape聚类算法在城市排水管网运行数据异常检测中具有较高的异常检出率,同时具有较低的误报率和漏报率。其中,K-shape聚类算法的查准率为0.812 7,虽然相对较低,但仍远高于BF算法的0.754 8;查全率为0.891 7,说明K-shape聚类算法

对于异常数据的检出率较高,漏报率较低。综合这些定性和定量分析,可以看出在城市排水管网运行数据异常检测中,K-shape聚类算法优于BF算法,K-shape聚类算法具有较高的异常检出率和较低的误报率和漏报率,可以很好地应用于实际场景中。

### 3 结论与展望

① 无监督学习方式不仅可有效地避免现有数据质量参差不齐的问题,还可实现长时间异常数据自动识别的过程。

② K-shape聚类方法可有效考虑长时间异常发生情况,既可降低异常检测的误报率及漏报率,也可提高查全率;相比BF等算法,其计算效率高且无需手动设置参数即可进行扩展应用。

③ 在后续工作中应进一步将识别的异常数据结合外部条件进行甄别分析,针对由于真实排水异常事件导致的异常数据进行标记存入数据库,以供预报预警相关研究使用;针对单纯的数据异常则根据不同的异常情况,进行相应数据补全,从而保障后续分析处理的有效性。

### 参考文献:

- [1] 孙雪梅,刘全海,冉慧敏,等.城市排水管网设施智能管理解决方案研究[J].城市勘测,2022(3):48-52.  
SUN Xuemei, LIU Quanhai, RAN Huimin, et al. Research on intelligent management of urban drainage facilities [J]. Urban Geotechnical Investigation & Surveying, 2022(3): 48-52(in Chinese).
- [2] 张莹.城市排水管网运行风险评估研究进展[J].城市道桥与防洪,2022(6):104-109.  
ZHANG Ying. Study on risk assessment in operation of urban drainage network [J]. Urban Roads Bridges & Flood Control, 2022(6): 104-109 (in Chinese).
- [3] 叱华娟.城市排水泵站调度系统管理模式探讨[J].价值工程,2015,34(29):106-108.  
CHI Huajuan. Management mode of urban drainage pumping station dispatching system [J]. Value Engineering, 2015,34(29): 106-108 (in Chinese).
- [4] CHEN L, YAN H, YAN J, et al. Short-term water demand forecast based on automatic feature extraction by one-dimensional convolution [J]. Journal of Hydrology, 2022, 606: 127440.
- [5] 陈盛达,冯一军,吴荣波,等.排水防涝规划水力模型应用探讨及案例分析[J].中国给水排水,2022,38

- (24):17-22.
- CHEN Shengda, FENG Yijun, WU Rongbo, *et al.* Discussion on hydraulic model application for urban drainage and flood control planning and case analysis [J]. *China Water & Wastewater*, 2022, 38(24): 17-22 (in Chinese).
- [6] 刘涛,李英俊,邢峰,等. 基于多变量自动回归的电力大数据异常值检测平台设计[J]. *自动化技术与应用*, 2022, 41(10): 93-96.
- LIU Tao, LI Yingjun, XING Feng, *et al.* Design of power big data outlier detection platform based on multivariate automatic regression [J]. *Techniques of Automation and Applications*, 2022, 41(10): 93-96 (in Chinese).
- [7] 瞿迪庆,吕齐,杨怀仁,等. 基于机器学习的网络异常检测及安全威胁等级预测研究[J]. *电脑知识与技术*, 2021, 17(34): 10-12.
- QU Diqing, LÜ Qi, YANG Huairan, *et al.* Research on anomaly detection and security threat level prediction based on machine learning [J]. *Computer Knowledge and Technology*, 2021, 17(34): 10-12 (in Chinese).
- [8] 阮嘉琨,蔡延光,乐冰. 基于DBSCAN密度聚类算法的高速公路交通流异常数据检测[J]. *工业控制计算机*, 2019, 32(7): 92-94.
- RUAN Jiakun, CAI Yanguang, YUE Bing, *et al.* Highway traffic flow anomaly data detection based on dbscan density clustering algorithm [J]. *Industrial Control Computer*, 2019, 32(7): 92-94 (in Chinese).
- [9] 赵云璐,信昆仑,PIERRE Bonardet,等. 供水管网中高频压力数据异常模式识别的研究与案例分析[J]. *中国市政工程*, 2022(5): 108-111.
- ZHAO Yunlu, XIN Kunlun, PIERRE Bonardet, *et al.* Research & case analysis on abnormal pattern recognition of high-frequency pressure data in water supply network [J]. *China Municipal Engineering*, 2022(5): 108-111 (in Chinese).
- [10] 黄琛,李文婷,张旭,等. 城市供水管网片区用水异常模式识别[J]. *云南大学学报(自然科学版)*, 2018, 40(5): 879-885.
- HUANG Chen, LI Wenting, ZHANG Xu, *et al.* Abnormal patterns recognition of urban water distribution system [J]. *Journal of Yunnan University (Natural Sciences Edition)*, 2018, 40(5): 879-885 (in Chinese).
- [11] 黄国东. 基于预测的供水管网异常监测数据清洗研究[D]. 杭州:浙江大学, 2022.
- HUANG Guodong. Research on Abnormal Monitoring Data Cleaning of Water Supply Network Based on Prediction [D]. Hangzhou: Zhejiang University, 2022 (in Chinese).
- [12] PAPARRIZOS J, GRAVANO L. K-shape: efficient and accurate clustering of time series [C]//ACM. Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. New York: ACM, 2015: 1855-1870.
- [13] 李海林,贾瑞颖,谭观音. 基于K-shape的时间序列模糊分类方法[J]. *电子科技大学学报*, 2021, 50(6): 899-906.
- LI Hailin, JIA Ruiying, TAN Guanyin, *et al.* Fuzzy classification for time series data based on K-shape [J]. *Journal of University of Electronic Science and Technology of China*, 2021, 50(6): 899-906 (in Chinese).
- [14] 董文璐. 自适应与鲁棒的缺失值填充方法研究[D]. 镇江:江苏科技大学, 2021.
- DONG Wenlu. Research on Adaptive and Robust Missing Value Imputation Algorithm [D]. Zhenjiang: Jiangsu University of Science and Technology, 2021 (in Chinese).
- [15] 傅维翔. 时间序列的符号化与异常检测研究[D]. 重庆:重庆大学, 2017.
- FU Weixiang. Research on Symbolization and Discord Discovery of Time Series [D]. Chongqing: Chongqing University, 2017 (in Chinese).
- [16] 曹蕾. 基于数据挖掘的电信用户流失预测研究[D]. 济南:山东师范大学, 2022.
- CAO Lei. Research on Forecasting Telecom Customer Churn Based on Data Mining [D]. Jinan: Shandong Normal University, 2022 (in Chinese).

作者简介:何黎(1994-),女,四川广安人,硕士,工程师,主要从事城市排水数值模型及相关大数据算法研究。

E-mail:hl\_hitscut@163.com

收稿日期:2023-01-11

修回日期:2023-02-06

(编辑:李德强)