Characteristic space-based algorithm for detecting abnormal monitoring data in underground engineering

Chenyang Wang¹, Zixin Zhang¹, Xin Huang¹†, Lunfei Che², Jun Liu²

¹Department of Geotechnical Engineering, Tongji University, Shanghai 200092, China
²China Railway Siyuian Survey and Design Group Co., Ltd., Wuhan 430063, China

*Corresponding author: xhuang@tongji.edu.cn; ORCID ID: https://orcid.org/0000-0002-3859-6247

Abstract. Abnormal points in underground space monitoring data contain valuable information. An anomaly detection algorithm based on characteristic space is proposed in this document. First, based on a temporal edge operator, the edge amplitude of each point in time series data is calculated. The monitoring data curve is linearly segmented by selecting the points with larger edge amplitudes. Second, the data characteristics of each subsegment are extracted, and the original monitoring curve is mapped into the characteristic space. Finally, the abnormal features are identified by the local anomaly factor outlier algorithm, from which abnormal data points can be further obtained. The effectiveness of the proposed algorithm was first verified by identifying artificial abnormal data serials or discrete data point from a normal dataset. The feasibility and applicability of the algorithm was further verified by applying the developed algorithm to detect the anomaly of monitoring data for a real tunnel project.

1. Introduction

Monitoring data reflect the operating status of underground engineering, which provides important information for project management and decision-making. However, because of sensor monitoring errors and environmental interference, the monitoring data will inevitably have abnormal points throughout the collection and transmission process. Compared with most normal data points, the abnormal points contain more critical and valuable information[1]. Therefore, abnormal detection of underground space monitoring data is of great significance to ensure effective operation and maintenance of underground engineering.

Monitoring data are time-varying. Hence, mining of abnormal values of monitoring data in underground engineering is essentially the abnormal detection of time series data. Existing abnormal detection methods for time series data can be divided into three categories[2]. The first type is the method based on the characteristic space. Such method first segments the time series and extracts data characteristics of each segmental subsequence. Then, the anomaly detection of the original data is realized by detecting abnormal features in the characteristic space. Xiao[3] proposed a time series method for anomaly detection based on pattern density. The dynamic model distance is used to measure the distance between various features, and model density in the overall characteristic space is used as an indicator to measure data anomalies. Zhou et al.[4] combined the idea of time series segmentation with the $k$ nearest neighbor algorithm. The important points of the sequence are used as the segmentation points to segment the time series. The $K$ nearest neighbor distance between each
subsequence is calculated as the anomaly measure of the subsequence, and the anomaly degree of the original data points is inversely deduced in turn. Ren et al. [5] proposed piecewise aggregation model representation of the time series, which has higher robustness than the piecewise linear representation method. The second type is based on the prediction method, which judges a data anomaly by the deviation between the predicted and true values. This approach is limited by the accuracy of the prediction model and has poor timeliness. Keogh et al. [6] used a sliding window to predict data streams. By calculating the confidence interval of the prediction results, data outside the interval are taken as alternative anomaly data. Keiichi [7] combined the symbolic representation algorithm and prediction algorithm of the time series. In the symbolic representation process of the original time series, the data change rule is found and the data trend is predicted. Anomaly detection is conducted by analyzing the deviation between the predicted and monitoring values. Sun [8] introduced the idea of grid sliding window on the basis of the strong search algorithm. The original time series was segmented; and the abnormal eigenvalue of the subsequence was calculated based on distance and density, through which the abnormal mode of the time series was identified. The third method is frequency-based, which needs many sample data for model training and has a relatively high time complexity. A representative algorithm is the hidden Markov model. Li [9] proposed an incremental HMM-learning algorithm, which cuts the training set into several smaller subsets for training, and finally merges the trained submodels into the final model. Experiments show that the method can greatly reduce the computational time. Sun [10] proposed a Markov chain anomaly detection model based on the feature mode, which can avoid repeated detection. The experiment shows that the algorithm is simple and real time, and has a high detection rate and low false alarm rate. However, the algorithm is sensitive to parameter settings, which can directly affect the detection effect of the algorithm. Wang et al. [11] proposed a behavior modeling and anomaly detection method based on semi-supervised learning. By combining the hidden Markov model and Fourier transform, a normal behavior and abnormal behavior model was established to realize the judgment of anomaly.

Because data points in the time series have a strict order, the traditional outlier detection method is mainly used to solve the anomaly detection problem of disordered data. However, this method is difficult to apply to the anomaly detection problem of ordered data. Additionally, in previous studies, the anomalies in time series are directly defined as sequence points or patterns that deviate from a certain training model. The detection efficiency of anomalies is not comprehensive, and the ability to detect and process massive data is insufficient. To solve the above two problems, an anomaly detection method based on time series feature representation is proposed in this study. First, piecewise linear representation of time series is realized by introducing the temporal edge operator (TEO). Then, data characteristics of each subsegment are extracted to realize the transformation of monitoring data information from order to disorder. The defect that traditional methods cannot handle ordered data can be resolved. Finally, the local outlier factor algorithm is used to detect the anomaly degree of each feature data point in a feature space. Based on the set threshold, abnormal points can be identified, which can improve the efficiency and accuracy of detecting anomaly points. Through three sets of data experiments, the effectiveness and accuracy of the algorithm are verified.

2. Time Series Feature Extraction

2.1 Basic Concept

To mine the information in time series, it is necessary to clarify the concept of time series first. Time series is an ordered set of elements composed of recorded values and recorded time, denoted as

$$X = \{x_1 = (v_1, t_1), x_2 = (v_2, t_2), \ldots, x_n = (v_n, t_n)\} \quad \text{(Eq. 1)}$$

The element $x_i = (v_i, t_i)$ represents the recorded value $v_i$ at the moment $t_i$; and the recorded time $t_i$ is strictly increased.

For the time series, $X = \{x_1, x_2, \ldots, x_n\}$, assuming that the time coordinate set of its piecewise points is $(t_1, t_2, \ldots, t_n)$, its piecewise linear representation can be expressed as
\[ X(t) = \begin{cases} 
  f_t(t, w_t) + e_t(t), & t \in [1, t_1] \\
  f_t(t, w_2) + e_2(t), & t \in (t_1, t_2) \\
  \vdots \\
  f_t(t, w_k) + e_k(t), & t \in (t_{k-1}, n] 
\end{cases} \tag{Eq. 2} \]

The function \( f_t(t,w_t) \) represents the linear function of two endpoints in a connection interval; \( w_t \) represents the coordinates of two endpoints in the time interval \([w_{t-1}, w_t]\); \( e_k(t) \) represents the error between original data and the linear representation in the time interval. The essence of piecewise linear representation of time series is to represent the original time series with multiple linear functions.

### 2.2 Piecewise linear representation based on TEO

The selection of piecewise points is the key to the piecewise linear representation algorithm. An edge operator is used to detect the edge of the image. Based on the original image, the grayscale step change of each image pixel in a certain area is investigated. The Sobel operator detects the edge of the image according to the grayscale-weighted difference in the adjacent points around the pixel and reaches the extremum at the edge. In this study, the Sobel operator is used as the prototype and the TEO conforming to the characteristics of the time series is used to identify the time edge data in the time series. These data are connected in turn to obtain the piecewise linear representation of the time series, which is referred to as the TEO representation of the time series. For the time series, the TEO is defined as follows:

\[ \text{Teo}(t,u) = \{ w(i)*(x_{u_i} - x_i) | i = -u, ..., -2, -1, 0, 1, 2, ..., u \} \tag{Eq. 3} \]

In which \( w(i) \) is the weight of the detection window. In this study, it is set as \( w(i) = abs(i), i = \{-1, ..., -u, 0, u, ..., 1\} \). When a point is close to the center of the detection window, the corresponding weight is higher. This point can be an edge point.

The edge amplitude of each point in the time series can be obtained using the convolution operation between the TEO and the time series. The edge amplitude represents the variation of time series trend in this point field. The higher the edge amplitude is, the greater the change trend of this point is compared with other points. The lower edge amplitude indicates that the change trend of the point and the surrounding point are closer. The points with a large margin are selected as the segmentation points of time series, and the original data are segmented. Then, the time series can be piecewise linearly represented by linear interpolation.

### 2.3 Feature extraction

Feature extraction can be conducted after linear piecewise representation of monitoring time series data. The more characteristic data are extracted, the more complete the original data information can be reflected. However, such data extraction will also increase the extent of calculation. Although fewer characteristic extraction can reduce the overall calculation, it may also cause lack of original data information, which reduces the accuracy and reliability of the whole algorithm. Therefore, based on the piecewise linear representation of the time series, this paper extracts three data characteristics of slope, length and mean of each line segment mode, and then the characteristic representation of each segment is denoted. The characteristics of time series are as follows:

\[ X_c \rightarrow \begin{pmatrix} 
  s_1 & s_2 & \cdots & s_p \\
  l_1 & l_2 & \cdots & l_p \\
  m_1 & m_2 & \cdots & m_p 
\end{pmatrix} \tag{Eq. 4} \]

In which \( s \) is the slope of subsequence, \( l \) is the length of the subsequence, and \( m \) is the mean of the data points contained in the subsequence.

### 3. Abnormality Calculation

#### 3.1 Local outlier factor detection algorithm
The local outlier factor detection algorithm measures the anomaly degree of the corresponding points by calculating the density of each characteristic object in the characteristic space. If the density of the characteristic object is large, it is less likely to be abnormal. Conversely, if the density of the characteristic object is small, it is more likely to be an outlier. After mapping the original time series to an object set \( D \) in the characteristic space, the characteristics of each subline segment are represented by \( p \), and then the anomaly degree of each feature object in the object combination can be calculated from the local outlier factor detection algorithm. The following is the basic concept in the algorithm calculation process.

1. **Distance function:** the distance between the object \( p(l_1,s_1,m_1) \) and the object \( q(l_2,s_2,m_2) \) is defined as:
   \[
   d(p,q) = \sqrt{(l_1 - l_2)^2 + (s_1 - s_2)^2 + (m_1 - m_2)^2}
   \]  
   (Eq. 5)

2. **\( k \)-nearest neighbor distance (\( k\text{-dist}(p) \)):** for a given positive integer \( k \) and data point set \( D \), the \( K\)-dist of point \( P \) in \( D \) satisfies the following two points: (1) at least \( k \) points satisfy \( o \in D, d(p,o) \leq k - dist(p) \) (2) at most \( k-1 \) points satisfy: \( o \in D, d(p,o) < k - dist(p) \).

3. **\( K \) nearest neighbor reachable distance from point \( q \) to point \( p \):**
   \[
   r - dist_k(q,p) = \max (d(q,p),k - dist(p))
   \]

4. **\( K \) local reachable density of point \( q \):**
   \[
   lrd(q) = \frac{k}{\sum_{p \in k(q)} r - dist_k(q,p)}
   \]  
   (Eq. 6)

5. **Local outlier factor of point \( p \):**
   \[
   LOF(q) = \frac{\sum_{p \in k(q)} lrd(p)}{lrd(p)}
   \]  
   (Eq. 7)

If the local anomaly factor of point \( q \) is large, it means that the local range of the point contains sparse points. This point is more likely to be abnormal.

6. **Abnormal pattern:** An abnormal pattern is a pattern that is significantly different from other patterns or has an abnormal behavior in the time series piecewise linear pattern representation. In this study, the line pattern whose local anomaly factor exceeds the abnormal threshold is the abnormal pattern and the degree of pattern abnormality is measured by calculating the local anomaly factor.

### 3.2 Abnormality calculation method

After calculating the anomaly degree of each characteristic point in the characteristic space, the characteristic anomaly degree can be assigned to each data point on the corresponding linear segment. Thereby, the anomaly degree of each point of the original data is obtained. The steps of calculating anomaly degree through the local anomaly factor detection algorithm are as follows: (1) For the input time series, piecewise linear representation is first performed. The slope \( s \), length \( l \), and mean \( m \) of each subsegment are extracted and the ordered time series data into the characteristic object set \( D \) in the disordered characteristic space are transformed. (2) Calculate the \( K \) nearest neighbor distance of each object \( P \) in the object set \( D \) and its \( K \) nearest neighbor reachable distance with other objects \( q \). (3) Calculate the local reachability density and local outlier factor of each data point \( P \) in object set \( D \). (4) The local anomaly factors of each feature object in the characteristic space are assigned to the original time series data to obtain the local anomaly factors of each point in original data.

### 4. Abnormality Threshold Calculation

The threshold selection is essential for the anomaly detection of time series data. An exception is the data that exceed the threshold. At present, the commonly used threshold-setting rule is to manually set various thresholds and analyze the effect. Obviously, this will increase the amount of calculation, and the results of the threshold are not universal. Based on statistical laws, define the variance of the data in a certain sample space is \( \sigma \) and the mean is \( \mu \). Then when the data distribution in this space obeys the normal distribution, if certain sample data are outside the \( (\mu - 3\sigma, \mu + 3\sigma) \) interval, the data can be regarded as abnormal data, which are called 3\( \sigma \) criterion. In this study, the distance between time series feature points and other feature points in the feature space is calculated to express the magnitude of its abnormality. In a long-term stable monitoring environment, when the amount of monitoring data
is large enough, the time characteristic distribution of the collected data points obeys the normal distribution\(^{[12]}\). If data characteristics of a data point significantly deviate from the data characteristics of other points, the data can be considered abnormal data. Therefore, based on the 3σ criterion, the threshold for all outliers is set as follows\(^{[13]}\):
\[
\lambda = \mu + 3\sigma
\]
(Eq. 8)

where \(\mu\) is the mean value of the abnormality of all data points; \(\lambda\) is the abnormality threshold; \(\sigma\) is the standard deviation of the abnormality of all data points. If the abnormality of each point of the original time series data is greater than the abnormal threshold, the data point is considered abnormal data. If the abnormality of the data point is less than the abnormal threshold, the data point is deemed normal data.

5. Anomaly Detection Data Experiment

5.1 Time series data with abnormal series

To verify the applicability of the algorithm in this study, anomaly detection is conducted on the dataset containing abnormal sequences. Keogh et al.\(^{[14]}\) used the simulation dataset to study the anomaly detection of singular modes in time series. Time series are generated by the following random process:

\[
\begin{align*}
X_1(t) &= \sin\left(\frac{50\pi}{N} t\right) \\
X_2(t) &= \sin\left(\frac{50\pi}{N} t\right) + n(t) + e_1(t) + e_2(t)
\end{align*}
\]

(Eq. 9)

where \(n(t) (t = 1, 2, \ldots, N, N = 800)\) is a Gaussian noise with a mean value of 0 and a standard deviation of 0.1. \(e_1(t)\) and \(e_2(t)\) two abnormal subsequences, defined as follows:

\[
\begin{align*}
e_1(t) &= \sin\left(\frac{25\pi}{N} t\right), \quad t \in [400, 432] \\
e_2(t) &= 0.5 \times \sin\left(\frac{50\pi}{N} t\right), \quad t \in [300, 350]
\end{align*}
\]

(Eq. 10)

The images of \(X_1(t)\) and \(X_2(t)\) are shown in Figure 1. Time series \(X_2(t)\) contains two abnormal sequences, which are distributed in the two intervals \([400, 432]\) and \([300, 350]\). The anomaly-processing results detected by the proposed time series anomaly detection algorithm based on the characteristic space are shown in Figure 2.
Figure 2 intuitively reflects the abnormality of the time series, in which the red dashed line is the calculated abnormality threshold. It can be seen from the figure that when the abnormal sequence is added to the original sequence, the local abnormality factor of the abnormal subsequence and the abnormality factor of the surrounding subsequences have an obvious change. The data distributed in the intervals [300, 350] and [400, 432] have a high degree of abnormality, and this subsequence can be regarded as an abnormal sequence. This abnormal sequence can also be explained by theoretical analysis. The curve amplitude and frequency in the intervals [300, 350] and [400, 432] are significantly different from other intervals. The data features extracted by the time series feature extraction algorithm in this study are far away from other feature points in the feature space. Therefore, the data in these two intervals have a high degree of abnormality. Additionally, the abnormal factors of other data points change around 1, which agrees with the theoretical values of abnormal factors that normal data should have. The anomaly detection algorithm based on characteristic space proposed by this research is correct and effective for the identification and detection of anomalous events in time series.

5.2 Time series data with outliers
Experimental data in this section come from the open dataset provided by the Berkeley-Inter Laboratory in the United States. From this dataset, 1,000 temperature monitoring data of a sensor are selected within the day of February 28. Because the distribution of abnormal data in the original data is not obvious, the abnormal data points are manually added at eight points of 100, 200, 300, 400, 500, 600, 700, and 800, which are taken as the anomaly detection objects. The original data curve and the data curve with an exception are shown in Figure 3.

In the data curve distribution image, it is obvious that the overall fluctuation of the original data is small, and there is no obvious abnormal point. After exception handling, the manually added
exception points are significantly different from other data points. The results of the anomaly detection using the proposed algorithm are shown in Figure 4.

![Abnormality of temperature data with abnormal points](image)

**Figure 4.** Abnormality of temperature data with abnormal points

Three conclusions can be drawn from the calculation results: (1) The abnormality threshold calculated by the proposed algorithm is 9.151, which can well identify eight anomaly points. The effectiveness of the threshold-setting rules in this study has been proven. (2) It can be seen from the numerical size distribution of local abnormal factors in the result diagram that the abnormal degree of most data points is approximately equal to 1 except for some points whose abnormal degree is greater than 1. This observation not only proves the theoretical correctness of the algorithm in this paper, but also conforms to the abnormal distribution of actual monitoring data. (3) In addition to the high abnormality of the eight manually added abnormal points, the abnormality of some data is also relatively high in the interval [400,500]. Compared with other intervals, the original data in this interval exist larger fluctuations, which indicate that there are abnormal points in this interval.

5.3 *Engineering time series data containing abnormal events*

During the construction of Shanghai Rail Transit Line 11, the Catholic Church of Xujiahui is traversed along the way [15]. To ensure safety of the church, 31 settlement observation points are arranged around it. The influence of the tunnel construction process on church can be divided into four stages: (1) MJS pile foundation construction stage; (2) Uplink crossing stage; (3) Downline crossing stage; (4) Long-term settlement stage. The collection of the monitoring data began on November 14, 2009 and ended on March 5, 2010. MJS pile foundation construction was completed on December 4, 2009. On January 11, 2010, the upper line was excavated below the church. On February 25, 2010, the downward line was excavated directly below the church. After March 1, 2010, the church entered the late settlement stage. To verify the effectiveness of the algorithm in this paper, the historical settlement monitoring data of four settlement observation points (F3, F4, F5, and F6) close to the tunnel upstream are selected for analysis. A single settlement observation point contains 90 sample data. The original monitoring data curve is shown in Figure 5.
Figure 5. Settlement monitoring curve of the Catholic Church

The calculation results of anomaly degree and anomaly threshold are shown in Fig. 6.

Figure 6. Abnormality of data at each measurement point of the Catholic Church

The results of the observation show that the four sensor monitoring curves have large anomalies at the thirty-seventh, forty-eighth, fifty-seventh, and seventy-fifth points, and the corresponding values exceed the threshold. It is observed that the monitoring time of these four points is exactly the four key construction nodes of the shield tunnel. The anomaly detection algorithm in this paper can well identify the anomaly changes in the four monitoring curves.
6. Conclusion

In this study, a time series anomaly detection algorithm based on feature space is proposed. Through experimental verification of three types of datasets, the following conclusions can be drawn:

(1) By introducing the TEO, the effective segmentation of the time series can be realized. This method can also compress and reduce noise points on the original data, and improve the efficiency of algorithm execution. Simultaneously, the influence of noise points on subsequent anomaly results is reduced.

(2) Combining the local anomaly factor detection algorithm with the time series piecewise linear representation algorithm can effectively identify the abnormal subsequences and points in the time series, and can give the abnormality of each point. The algorithm has high accuracy and applicability.
7. References

[1] Zheng X, Chen G, Zhou T. Cleaning of abnormal dam monitoring data considering the time series association. Water Power, 2020, 46(04):111-114+125.
[2] Hu M, Bai X, Xu W et al. Review of anomaly detection algorithms for multidimensional time series. Journal of Computer Applications, 2020, 40(06):1553-1564.
[3] Xiao H. Similarity search and outlier detection in time series. Fudan University, 2005.
[4] Zhou D, Liu Y, Ma W. Effective time series outlier detection algorithm based on segmentation. Computer Engineering and Applications, 2008, 44(35):145-147.
[5] Ren H, Liu M, Li Z, et al. A piecewise aggregate pattern representation approach for anomaly detection in time series. Knowledge-Based Systems, 2017, 135:29-39.
[6] Keogh E, Chakrabarti K, Pazzani M, et al. Dimensionality reduction for fast similarity search in large time series databases. Journal of Knowledge and Information Systems, 2013, 3(3):263-286.
[7] Keiichi T, Tatsuhiro S, Takumi I. Time series classification using MACD-histogram-based SAX and its performance evaluation//Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics. Piscataway: IEEE, 2016:2419-2424.
[8] Sun M. Research on discords detect on time series based on distance and density. Computer Engineering and Applications, 2012, 48(20):11-17.
[9] Li H P, Hu Z Y, Wu Y H, et al. Behavior modeling and anomaly detection based on semi-supervised learning. Journal of Software, 2017, 18(3):527-537.
[10] Sun M F, Huang F, Chen Y Z, et al. Markov chain anomaly detection model based on feature pattern. Computer Engineering, 2018, 32(24):155-157.
[11] Wang Q, Ni G Q, Pan Z S, et al. System call anomaly detection based on improved hidden Markov model. Journal of Data Acquisition and Processing, 2009, 24(4):508-513.
[12] Laptev N, Amizadeh S, Flint I. Generic and Scalable Framework for Automated Time-series Anomaly Detection. Knowledge Discovery and Data Mining, 2015, 1939-1947.
[13] LIU F L, YIN M, YUAN P. Research on Threshold Selection Algorithm for Anomaly Detection of Time Series Data Based on DBSCAN, Modern Computer, 2020(04):3-6+20.
[14] Keogh E, Lonardi S, Chiu W. Finding surprising patterns in a time series database in linear time and space//Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2002:550-556.
[15] Qi M, Huang A, Zhang H. An analysis & research of Shanghai metro L11 shield slanting across Xujiahui Catholic Church. China Municipal Engineering, 2011(06):44-45+48+83.

Acknowledgments
The work was supported by the National key research and development projects (Grant No. 2019YFC0605105).