Hindawi
Security and Communication Networks
Volume 2017, Article ID 1869787, 10 pages
https://doi.org/10.1155/2017/1869787

Research Article

Locality-Based Visual Outlier Detection Algorithm for Time Series

Zhihua Li,[1,2] Ziyuan Li,[1] Ning Yu,[2] and Steven Wen[2]

[1]*Department of Computer Science, School of Internet of Things Engineering, Jiangnan University, Jiangsu, Wuxi 214122, China*
[2]*Department of Computer Science, Georgia State University, Atlanta, GA 30303, USA*

Correspondence should be addressed to Zhihua Li; zhli@jiangnan.edu.cn

Received 22 August 2016; Revised 8 June 2017; Accepted 6 July 2017; Published 22 August 2017

Academic Editor: Emanuele Maiorana

Physiological theories indicate that the deepest impression for time series data with respect to the human visual system is its extreme value. Based on this principle, by researching the strategies of extreme-point-based hierarchy segmentation, the hierarchy-segmentation-based data extraction method for time series, and the ideas of locality outlier, a novel outlier detection model and method for time series are proposed. The presented algorithm intuitively labels an outlier factor to each subsequence in time series such that the visual outlier detection gets relatively direct. The experimental results demonstrate the average advantage of the developed method over the compared methods and the efficient data reduction capability for time series, which indicates the promising performance of the proposed method and its practical application value.

1. Introduction

Time series, widely existing in various applications [1] such as sensor network data collection [2–4], credit card fraud data [1], and environment monitoring data [2, 5–8], is one of the major types of big data. In fact, time series is an ordered sequence of observed data with respect to time; highly intuitive and usually most of the desired key information can be directly obtained from the different variations or distributions via the human visual system. On the other hand, physiological experiments have demonstrated that the deepest impression for sequence data with respect to the human visual system is its extreme value [9, 10], so it intuitively inspired us to study the visual outlier detection method with respect to the outlier events based on this principle.

Generally, there are three types of outliers: collective outliers, point outliers, and contextual outliers [8]. Identification of outliers can lead to the discovery of significant clues and has practical applications in various fields, such as financial risk management [1, 10], anomaly detection [5], and disaster alarm in environment monitoring [2, 5–7]. In the past few decades, this issue has been addressed in academia and attracted an increasing amount of attention. Some of the outlier detection approaches are based on notably different assumptions, intuitions, and models and also differ substantially in the scaling, range, and even meaning of values [11]. Furthermore, some other methods are developed on the basis of the technologies themselves such as the cluster-based detection method [4], the immunology-based detection method [12], and the SVM-based detection method [8]. Regardless of any type of time series, there always exist many valuable characteristics in most locations, such as the locality features neighboring the real outlier, the locality characteristic maybe more meaningful than the global information. For example, when a doctor diagnoses a disease based on the electrocardiogram, the ECG's local information is enough for finding the lesion. However, most of the aforementioned methods are unable to detect the outliers in time series locally and visually.

Although most of the previous researches [1–8] have addressed the outlier detection in time series, there still exist some challenges to undertake; for example, different time series appear out of synchronism, results of the traditional

similarity calculation method are no longer available, the periodical outlier in time series is hard to detect, the determination of the outlier threshold is unreasonable, and so on. In this paper, a hierarchy-segmentation-data-extraction-based outlier detection method is proposed. Our scheme integrates the investigation on the following to achieve relatively high effectiveness and efficiency: (a) studying the extreme-point discriminating strategy based on hierarchy segmentation; (b) the hierarchy-segmentation-based data extraction (HSDE) method for time series; (c) the outlier detection model; and (d) the locality outlier detection algorithm. Specific to the outlier identification, here, unlike all previous attempts to solve this problem, the proposed method depends on the departure from the location of the objects from its expected hierarchy rather than its global structure. Additionally, being labeled as an "outlier" here is not an either/or proposition. Instead, the proposed method assigns a local outlier factor to each detected subsequence, and the factor is the level to whether the object is outlying. Our major contributions are detailed as follows.

(1) The relation between the distribution characteristic in time series and the recognition mechanism associated with the human visual system is addressed, and the HSDE-based visual outliers detection method distinguishes the outliers directly without requiring previously observed training data.

(2) The locality-based outlier detection idea is successfully transferred into the realization for data mining of time series; in contrast, the previous LOF algorithms are only applicable to numerical data.

(3) A novel hierarchy-segmentation-based data extraction method for time series and its associated outlier detection model are presented.

The remainder of this paper is organized as follows. The related works are introduced in Section 2. In Section 3, we describe the new hierarchy-segmentation-based strategy and the related data extraction method. In Section 4, we improve the key ideas in LOF algorithm and derive the framework of the HSDE-based outlier detection model and algorithm. Promising experimental results on benchmarking datasets are presented in Section 5, which are followed by the concluding remarks in Section 6.

2. Related Works

A wide variety of studies investigating outlier detection have been examined; various outlier detection methods, such as global versus local, scoring versus labeling, and supervised versus unsupervised, were proposed [13]. Most of them are developed from different identification ideas of outliers, respectively, such as similarity measurement or dissimilarity measurement. Due to the specificity of time series, only a small part of detection methods are able to detect the outliers in time series.

As to the distance-based outlier detection methods in time series, there are four main dissimilarity measurements and their related evolution works, such as Euclidean distance (ED), dynamic time warping (DTW), symbolic aggregate approximation (SAX), and extended symbolic aggregate approximation (Extended-SAX) and their derived outlier detection schemes. The associated outlier detection methods that are developed from the four types of distance all inherit their own advantages or disadvantages without exception. ED is well known for its simple computation and sound universality, but it can only carry out the time series of equal length and cannot recognize the variation trend of time series [13, 14]. DTW can well overcome the first disadvantage of ED and can support the time warping of time series. However, its computing complexity and time complexity are high, which limits its application range. Chiu et al. [15] proposed the symbolic aggregate approximation (SAX) approach. SAX firstly symbolizes the time series and then carries out data similarity measure of the symbolic data. This method was easy to use and independent of specific experimental data. With relatively strong universality, the approach has been widely used [16–18]. However, the essence of similarity measure in SAX is based on ED or DTW, so it is inevitable to inherit their disadvantages.

Naess and Gaidai [9] developed a feature space-based outlier detection method based on SAX. The feature space-based outlier detection method can reduce the number of features effectively and compress the scale of time series. It was easy to miss some important features in the process of reduction. And also, it was unable to detect the outliers in time series visually. Extended symbolic aggregate approximation (Extended-SAX) [19] was developed from SAX, and an outlier detection method was also presented. Extended-SAX needed to depend on the piecewise aggregate approximation (PAA) representation for dimensionality reduction that minimizes dimensionality by the mean values of equal sized subsequences. Undoubtedly, the final distance measurement in Extended-SAX also depended on ED or DTW. Furthermore, the PAA still needed more time to strengthen the computation complexity. The outlier detection method based on Extended-SAX is unable to detect the outliers in time series visually. More so, all of the above methods realized the outlier detection through the so-called "distance measurement" rather than the locality distribution characteristic of time series.

This paper also uses DTW as the dissimilarity measurement. The HSDE-based outlier detection scheme is also inspired by the strategy of the local outlier factor LOF [19] and its incremental LOF algorithm [20], whereby we address the collective outlier detection by DTW-based methods and aim to enumerate the desired outliers in time series visually via the locality distribution characteristics of data points. Particularly, the outliers are visually enumerated to detect by the human visual system. Finally, comparison studies are also performed with the feature space-based outlier detection method [9] and the Extended-SAX-based outlier detection method [19], and the analysis results are also presented.

3. Hierarchy-Segmentation-Based Time Series Extraction

3.1. Extreme-Point-Based Hierarchy Segmentation. According to the physiological theories [9, 10], the extreme value in time series (i.e., either the maximum value or the minimum value) usually gives people the deepest impression. Based on this

```
For i = 2 : (n − 1)
    ∀pre ∈ [max{i − p, 1}, i),  ∀fol ∈ (i, min{i + p, n}]
   IF x_j ≥ x_pre AND x_j ≥ x_fol
    Flag = 1
    Else IF x_j < x_pre AND x_j < x_fol
    Flag = −1
    Else Flag = 0
  END
 END
Return Flag
```

Pseudocode 1: Function EPD(X, n, p).

```
Initial HM
For p = 1 : Max_p
For j = 1 : n
   If HM[j] == p − 1 AND EPD(X, j, p) == 1
     HM[j]++;
   Else If HM[j] == 1 − p && EPD(X, j, p) == −1
        HM[j]−−;
   End
End
HM[1] = HM[n] = Max_p + 1;
Return HM
```

Pseudocode 2: Function HM(X, Max_p).

principle, this paper presents a new concept: "*hierarchy* of time series."

Definition 1. Given a time series $X = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}(1 \le i \le n)$, before and after x_i, wherein the interval of x_i is $[-p, p]$, if $x_j \in \langle x_{i-p}, \ldots, x_i, \ldots, x_j, \ldots, x_{i+p}\rangle$ and x_j is the maximum value or minimum value in $\langle x_{i-p}, \ldots, x_i, \ldots, x_j, \ldots, x_{i+p}\rangle$, then it is called the hierarchy of $\langle x_{i-p}, \ldots, x_i, \ldots, x_j, \ldots, x_{i+p}\rangle$ and $2p$ is the size of its corresponding marked window.

Definition 2 ($\langle x_{i-p}, \ldots, x_i, \ldots, x_j, \ldots, x_{i+p}\rangle \in X$). The absolute value of p is called its "hierarchy value."

In the following, x_j is used to represent the corresponding subsequence $\langle x_{i-p}, \ldots, x_i, \ldots, x_j, \ldots, x_{i+p}\rangle$ and its hierarchy value is $|p|$.

In this, the "hierarchy value" describes the importance level of x_j in time series. The larger the hierarchy value, the higher the importance of x_j in time series. Therefore, the hierarchy value is also entirely used to represent the importance level of x_j in time series.

Based on the characteristic of the hierarchy of different data points in time series, the hierarchy-segmentation-based data extraction (HSDE) for time series is proposed, which includes stages such as extreme-pointed discriminating (EPD), hierarchy marking of time series (HM) and hierarchy segmentation series accessing (HSSA).

(1) Extreme-Pointed Discriminating. In this section, extreme-pointed discriminating (EPD) function is discussed. In a time series X, $\langle x_{i-p}, \ldots, x_i, \ldots, x_j, \ldots, x_{i+p}\rangle$ is a subsequence of X. If x_j is "$\arg\max\langle x_{i-p}, \ldots, x_j, \ldots, x_j, \ldots, x_{i+p}\rangle$," then the returned value of EPD is noted as Flag = 1; if x_j is "$\arg\min\langle x_{i-p}, \ldots, x_j, \ldots, x_j, \ldots, x_{i+p}\rangle$," then the returned value of EPD is noted as Flag = −1; otherwise, Flag = 0. The pseudocode of EPD is expressed in Pseudocode 1.

(2) Hierarchy Marking of Time Series. Hierarchy marking of time series (HM) function is discussed in this section. EPD function is utilized for discrimination of extreme points. The pseudocode of HM can be expressed in the diagram below. Because p is always a positive integer, here, a predetermined parameter Max_p is defined as the upper value of p, which is an experiential parameter. Namely, $p \in [1, \text{Max\_}p]$. The obtained hierarchy mark of each data point x_j in X is noted as $\text{HM} = \{\text{hm}_1, \text{hm}_2, \ldots, \text{hm}_j, \ldots, \text{hm}_n\}$ and hm_j ($j = 1, 2, \ldots, n$) represents the hierarchy value of each corresponding x_j as shown in Pseudocode 2.

After HM processing is done, the hierarchy values of the obtained HM and x_j correspond, respectively.

```
Initial HSS
For j = 1 : n
   If |HM(j)| = Max_p
      Add x_j to HSS
   End
END
Return HSS
```

Pseudocode 3: Function HSSA (X, HM, Max_p).

(3) Hierarchy Segmentation Series Accessing. The process of hierarchy segmentation series accessing (HSSA) function, along with the original time order in X, selects the data points that satisfy $|\text{HM}(j)| = \text{Max\_}p$ in terms of the HM. The selected data points are reconstructed as a new hierarchy segmentation series (HSS). The pseudocode of HSSA function is expressedin Pseudocode 3.

In fact, after HSSA processing is done, the HSS corresponds to X after data *reduction*, while attempting to maintain as much key information as possible.

3.2. Hierarchy-Segmentation-Based Data Extraction. In fact, the number of the new obtained time series HSS is far less than that of the original time series X. However, before and after the HSSA processing, the information is likely to remain similar without further changes. Therefore, data compression has been conducted simultaneously. What received more attention is that the new time series reduction HSS can successfully represent the original time series X only if the hierarchy value in HM is properly selected. As a result, we call it the hierarchy-segmentation-based data extraction (HSDE) method.

4. HSDE-Based Outlier Detection Scheme

4.1. The Local Outlier Factor and Detection Principle. In this section, our goal is to evaluate the practical applications value of HSDE-based methods. Inspired by the method developed in [19, 20], we extend the main idea of the local outlier factor (LOF) into data mining of time series, wherein the LOF is a local level that depends on how isolated the object is with respect to the surrounding neighborhoods. Moreover, our final goal is to assign an outlier factor (the level to which the object is outlying) to each subsequence in time series. Undoubtedly, this paper implements some key improvements of the steps in the previous algorithms [19, 20] and maintains some of the same locating outliers detection principles, such as the k-distance of an object x, k-distance neighborhood of an object, and reachability distance of an object x with respect to object o[19, 20]. The distance in k-distance of an object x is redefined as DTW(x, o) between x and an object such that (1) for at least k objects it holds that DTW$(x, o') <$ DTW(x, o) and (2) for at most $k - 1$ objects it holds that DTW$(x, o') <$ DTW(x, o). k is a positive integer which always represents the number of objects and must be predetermined by experimentation. Additionally, the k-distance neighborhood of x contains each object whose distance from x is not greater than the k-distance; that is,

$$N_{k-\text{distance}(x)}(x) = \{o \in X \setminus \{x\} \mid \text{DTW}(x, o) \leq k - \text{distance}(x)\}. \tag{1}$$

These objects o are called the k-nearest neighbors of x. The reachability distance of object x with respect to object o is defined as follows. Namely, it is defined as the following formula:

$$\text{reach-dist}_k(x, o) = \max\{k\text{-distance}(o), \text{DTW}(x, o)\}. \tag{2}$$

The set of the reachability distances of an object x is denoted as $S_{\text{reach-dist}_k}(x)$. The smaller the value of $|S_{\text{reach-dist}_k}(x)|$ is, the lower the number of the objects x in reachability distance of an object x with respect to object o is. In contrast, the larger value indicates that the object x has more neighborhoods and also falls inside more locations of the reachability distance of other objects. For example, given a temporary time series, which is illustrated in Figure 1, it is clear that the object x_1 is not located inside any other 2-distance neighborhoods and is far away from the others. Therefore, the object x_4 falls inside the reachability distance of the others.

Further, to improve the main principles developed in the algorithms [19, 20] to be suitable for handling time series, we continue to define two additionally important notions, the local reachability density of an object x and the local outlier factor of an object x, as shown in formulae (3) and (4), respectively.

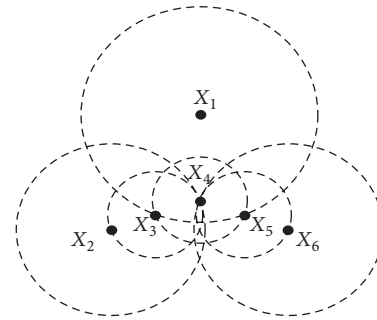


Figure 1: 2-distance neighborhoods.

Definition 3. The local reachability density $\rho_k(x)$ of an object x is defined as the following formula:

$$\rho_k(x) = \frac{\left|S_{\text{reach-dist}_k}(x)\right|}{\sum_{y \in \text{reach-dist}_k(x)} \left|S_{N_{k\text{-distance}(y)}}(y)\right|}, \tag{3}$$

where $\rho_k(x)$ is a level of the local density of the object x. $\rho_k(x)$ is the ratio of the number of the reachability distances of the object x and the total sum of the reachability distance of the objects. Obviously, the definition is subject to $0 \leq \rho_k(x) \leq 1$.

Definition 4. The local outlier factor is defined as the following formula:

$$\text{LOF}(x) = 1 - \rho_k(x). \tag{4}$$

The local outlier factor (LOF) of each $x_i \in X$ is computed by formula (4) and is ordered in either an ascending or a descending order. As a result, the range of the outlier factor for each subsequence, $x_i \in X$, is clear.

4.2. The Outlier Detection Model. Based on the above studies, this paper presents an outlier detection model for time series that is shown in Figure 2. The outlier detection process mainly includes the following stages: the hierarchy-segmentation-based data extraction (HSDE) method for time series, the computation of k-distance, the computation of reachability k-distance neighborhoods and the local reachability density, the computation of the outlier factor, and labeling of the outlier sequence. Here, each stage is strictly conducted in terms of the aforementioned details.

4.3. The Proposed Method. Based on the proposed model in Section 4.2, the HSDE-based outlier detection method is summarized as shown in Pseudocode 4.

In this, the computation of the hierarchy-segmentation-based data extraction (HSDE) for time series requires the most time, while the main cost of time complexity is the double loop in the HM function, and the time complexity is $O(\text{lognMax\_}p)$. It is clear that the time complexity is similar in the other stages of the HSDE-based outlier detection algorithm and is no more than $O(\text{lognMax\_}p)$, in which each stage is conducted sequentially. Therefore, the total time complexity of the HSDE-based outlier detection algorithm is $O(\text{lognMax\_}p)$.

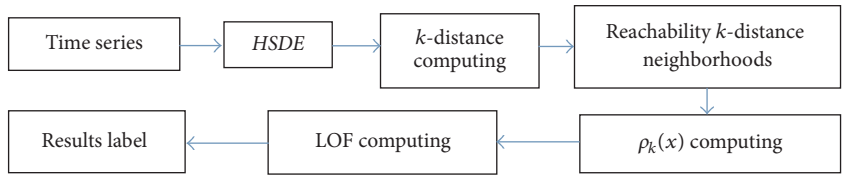


Figure 2: HSDE-based outlier detection model.

```
Input: X //the time series
 Max_p //the maximum of the hierarchy
 k //the k in k-distance
 λ //the threshold of the local outlier factor LOF
 Output: LOF(X)
 (1) HSS = HSDE(X)
 (2) for each x_j in HSS do {
       Computing the k-distance neighborhood of x_j;
       Computing the reachability distance of object x_j;
       According to Formula (3), computing the ρ_k(x)
       According to Formula (4), computing the LOF(x_j)
       if (LOF(x_j) ≥ λ)
          Label x_j as outlier
          }
 (3) return LOF(X)
```

Pseudocode 4: HSDE-based outlier detection method.

5. Experiment Result and Analysis

5.1. Experiment Arrangement. We arrange several experiments on three datasets: including Keogh_Data [21], ECG_Data [22], and Ma_Data [23], respectively. The experiments aim to validate both the detection capability and its effectiveness and efficiency. All of the experiments are realized using Matlab R2010b.

5.2. Evaluation Indices. This study also inherits the traditional indices [24], including false negative rate and false positive rate, and they are redefined as the following formulae, respectively:

$$R_{\text{FalseNegative}} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{5}$$

$$R_{\text{FalsePositive}} = \frac{\text{FN}}{\text{TN} + \text{FN}} \tag{6}$$

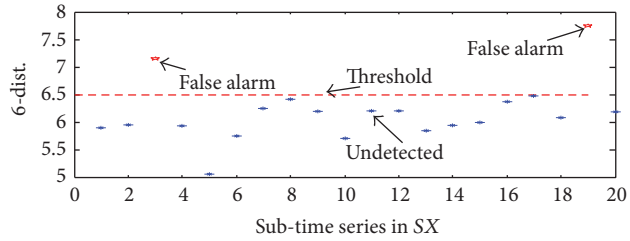
where TP, FP, FN, and TN are expressed in Table 1, where false negative rate ($R_{\text{FalsePositive}}$) denotes the ratio between the number of normal items wrongly recognized as outliers and the total number of the detected outliers, which is defined and formalized as formulation (5); a smaller false negative rate also indicates a higher outlier detection performance; false positive rate ($R_{\text{FalseNegative}}$) is expressed as the ratio between the misdetection outliers and the total number of the real outliers, which is shown as that formalized in formula (6); a lower rate implies a higher detection accuracy and prominent efficiency.

Table 1: The detected results.

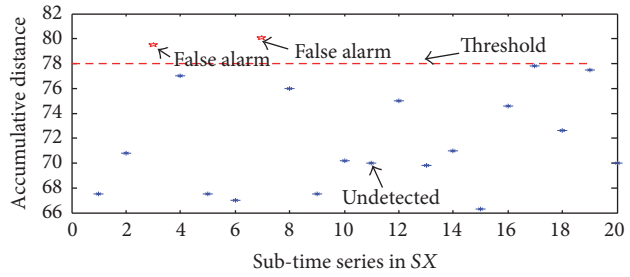
Reality	Detected normal	Detected outliers
Normal	True positive (TP)	False negative (FN)
Outliers	False positive (FP)	True negative (TN)

5.3. Result Analysis. Three benchmarking time series datasets, Keogh_Data [21], ECG_Data [22], and Ma_Data [23], are employed to the experiment. Experimental comparisons between different detection methods, including the feature space-based method [9], the Extended-SAX-based method [19], and the proposed method in this paper, are also done in terms of the evaluation indices with the best parameters in each method. We compared all three approaches on the same tasks: (1) the first is the training data, with several slightly noisy data points; (2) the second is a time series containing a synthetic "outlier," which was created with the same parameters as the training subsequence [25]; (3) to guarantee the fairness of comparison results, the time series datasets are user-partitioned into equal subsequences to highlight the outliers and degrade the complexity of data processing; and (4) the best parameters in each method are selected through several training experiments.

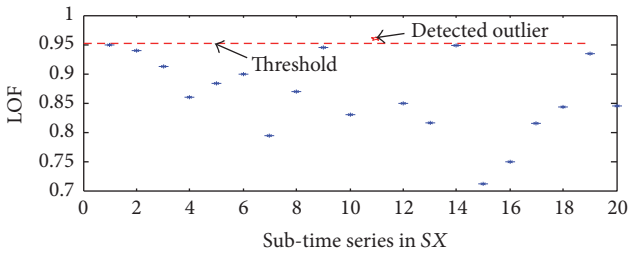
Experiment 1 (Keogh_Data). Keogh_Data [21] is the experiment time series by Keogh, which is generated by several randomized procedures and whose length is 800, in which an additive Gaussian noise with an average value of "0" and a standard deviation of "1" is added. In addition, there exist



(a) The results using the feature space-based outlier detection method



(b) The results using the extended *SAX*-based outlier detection method



(c) The results using the proposed outlier detection method

FIGURE 3: Experimental results on Keogh_Data.

outlier events in the range between the 400th and the 432nd data points in order to concentrate the outlier data points. Here, the 800 time series data points are separated into 20 subsequences, and each subsequence is 40 data points long. These 20 subsequences are reconstructed as a new time series dataset that is denoted as $SX = \{X_1, \ldots, X_{11}, \ldots, X_{20}\}$, which is implemented for the experiment. In the range between the 400th and 432nd data points, the corresponding 11th (e.g., X_{11}) subsequence is the real outlier. In the feature space-based outlier detection method, the number of subsequences is 20; k in k-distance is 6; the parameter α in Extended-SAX-based outlier detection method is 4; and Max_p in the proposed outlier detection method is 2. The experimental results are shown in Figure 3 and Table 2, wherein the threshold value is a user-predefined parameter based on several experienced observations.

Figure 3(a) shows the generalization 6-distance neighborhoods of each subsequence in SX by the feature space-based outlier detection method. It is clear that the value of X_{19} is the maximum, the value of X_{11} is relatively smaller, and the value of X_5 is the minimum. Figure 3(b) shows the generalization outlier factor of each subsequence in SX by the Extended-SAX-based outlier detection method. It is clear that the values of X_3 and X_7 are the maximum and their values are nearly equal. This case also indicates that the outlier factors of X_3 and X_7 are the maximum. In contrast, the accumulated distance of X_{11} is relatively smaller and less prominent. The accumulative distance of X_{11} is neither the maximum nor the minimum one. Figure 3(c) shows the LOF of each subsequence in SX by the proposed outlier detection method. It is clear that the value of X_{11} is the maximum. In this study, this case indicates that the LOF of X_{11} is the highest one, and it is consistent with the real time series.

On the other hand, an experimental comparison is shown in Table 2. The comprehensive performance of the proposed method is superior to the other compared ones. In Table 2, the total number of the real outliers is small, regardless of whether they are detected or not, which causes the evaluation indices of $R_{\text{FalseNegative}}$ and $R_{\text{FalsePositive}}$ to be extremely high or low according to the definitions. Comparatively, the proposed method is prominent.

According to the above findings, the generalizations of the 6-distance neighborhoods method and the Extended-SAX-based outlier detection method are unable to find the outlier subsequence. The generalization of 6-distance neighborhoods method introduced X_3 and X_{19} false alarms of approximatively equal magnitude, and the Extended-SAX-based outlier detection method introduced X_3 and X_7 false alarms of approximatively equal magnitude. Unlike the other two compared approaches, the proposed outlier detection method shows a strong peak for the range of the outlier subsequence, as it successfully detected the outlier X_{11}. Although X_3, X_7, and X_{19} are not real outliers, the proposed outlier detection method also shows X_3 and X_{19} at a relatively high outlier "level," but no more than that of the real outlier X_{11}. This situation indicates that the proposed outlier detection algorithm might have a practical application value. Although Figure 3 just shows the results at "2-distance," similar results may be observed at other hierarchies, and some outlier patterns might exist at different "hierarchy."

Experiment 2 (ECG_Data). ECG_Data [22] is a time series dataset with 3570 data points, in which there exist outlier events in the range between the 2300th and 2500th data points. Here, the ECG data are separated into 25 subsequences in order to highlight the outlier data points, and each subsequence is 150 data points long. These 25 subsequences are created as a new time series dataset $SY = \{Y_1, \ldots, Y_{16}, Y_{17}, \ldots, Y_{25}\}$, which is implemented in the experiment. In terms of the real outliers in ECG_Data, the 16th subsequence (e.g., Y_{16}) and the 17th subsequence (e.g., Y_{17}) are the outliers. We compared all three methods under consideration. In the feature space-based method, the segmentation number is 50; k in k-distance is 8; the parameter α in Extended-SAX-based outlier detection method is 4; and Max_p in the proposed method is 4. The experimental results are shown in Figure 4 and Table 3. In Figure 4, the threshold value is a user-predefined parameter based on the experienced observation.

Figure 4(a) shows the generalization 8-distance neighborhoods of time series by the feature space-based outlier detection method. It is clear that the 8-distance values of Y_{16}

Table 2: The experimental comparison using Keogh_Data.

Name	$R_{\text{FalseNegative}}$ (%)	$R_{\text{FalsePositive}}$ (%)
The feature space-based method	100	100
The extended SAX-based method	100	100
The proposed method	0	0

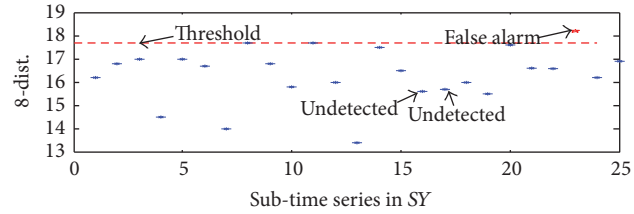
Table 3: The experimental comparison using ECG_Data.

Name	$R_{\text{FalseNegative}}$ (%)	$R_{\text{FalsePositive}}$ (%)
The feature space-based method	100	100
The extended SAX-based method	50	0
The proposed method	0	0

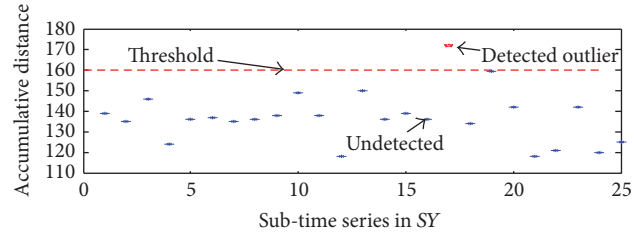
and Y_{17} are neither the maximum nor the minimum ones. In contrast, the 8-distance value of Y_{23} is relatively larger, but in fact Y_{23} is not a real outlier. Figure 4(b) shows the generalization outlier factor of each subsequence in SY by the Extended-SAX-based outlier detection method. It is clear that the accumulated distance value of Y_{17} is the maximum one and that of Y_{19} is the second maximum, whereas the accumulated distance value of Y_{16} is relatively smaller and less prominent. Namely, the other outlier Y_{16} has not been found. Figure 4(c) shows the generalization LOF of each subsequence in SY by the proposed outlier detection method. It is clear that the LOFs of Y_{16} and Y_{17} are larger than those of the others. Here, this case indicates that the LOF of Y_{16} and Y_{17} is the largest one, and it is consistent with the real time series.

According to the above discussion, the feature space-based outlier detection method is unable to find the outliers entirely, while introducing several subsequences false alarms of approximatively equal magnitude. The Extended-SAX-based outlier detection method found only one of the real outlier series instead of the two. It is clear that the Extended-SAX-based outlier detection method introduced Y_{19} false alarm. Unlike the other two compared approaches, the proposed outlier detection method shows a strong peak for the range of the real outlier data points by successfully detecting the outliers Y_{16} and Y_{17}. Although the LOF of the normal Y_{24} is no more than that of the outliers, it is regretful that the proposed outlier detection method also shows Y_{24} a relatively higher LOF value. In essence, as seen from Figure 4(c), the corresponding level of Y_{24} is of equal magnitude to the other normal data points without any extreme performance. Through analysis, we found that the reason this was caused is because of the experienced parameter Max_p in the proposed method. The length of subsequence is separated and marked by the parameter of Max_p. This situation results in the locality outlier instead of the global one, which is only outlying in its neighborhoods rather than in the global time series.

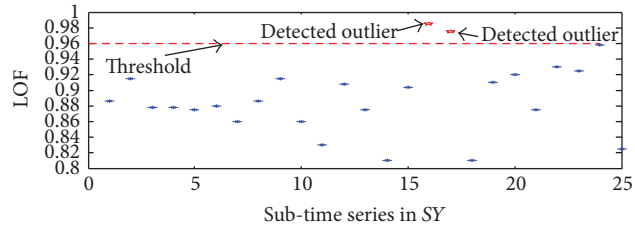
Additionally, the experimental comparison is shown in Table 3. The comprehensive performance of the proposed method is superior to the other compared ones. In Table 3, because of similar reasons, the number of real outliers is small; this results in extremely high or low evaluation indices



(a) The results using the feature space-based outlier detection method



(b) The results using the extended SAX-based outlier detection method

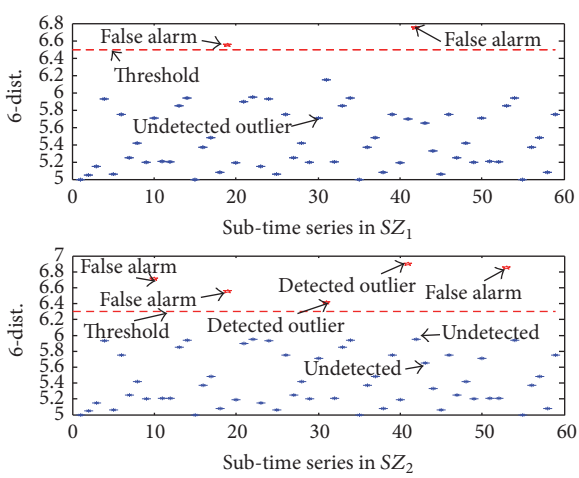


(c) The results using the proposed outlier detection method

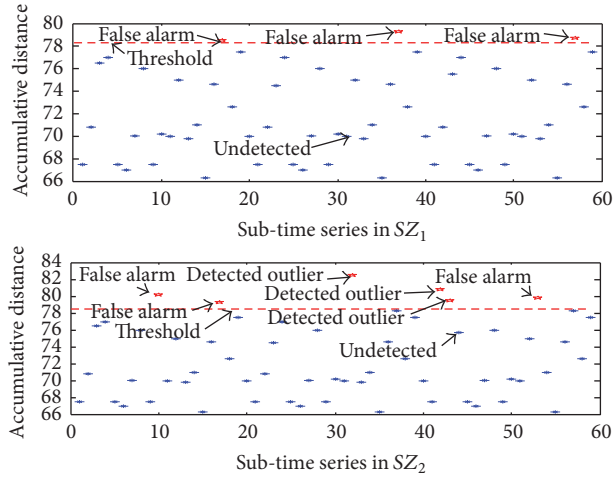
Figure 4: Experimental results on ECG.

of $R_{\text{FalseNegative}}$ and $R_{\text{FalsePositive}}$. Fortunately, it has no influence on the proposed method with a relatively stronger outlier detection capability.

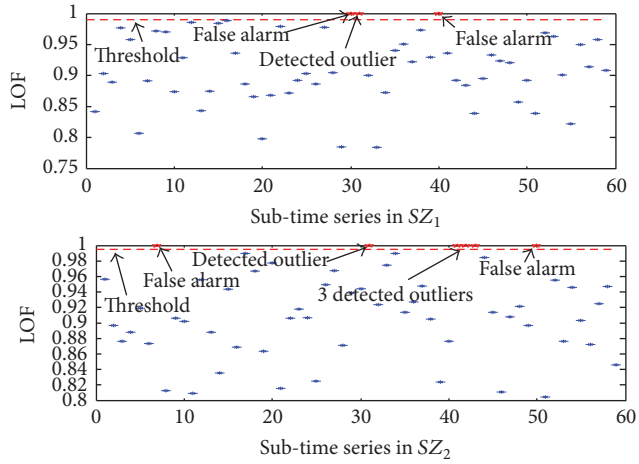
Experiment 3 (Ma_Data). Ma_Data [23] includes three pieces of synthetic time series that are generated from a user-predefined stochastic process, respectively; each time series has 1200 data points, wherein $X_1(t)$ is the normal distribution without outliers and the others of $X_2(t)$ and $X_3(t)$ are with an additive Gaussian noise with zero mean and a SDT of 0.1. The outlier event is between the ranges of [600–620] in $X_2(t)$, and the outlier events are in the



(a) The results using the feature space-based outlier detection method



(b) The results using the Extended-SAX-based outlier detection method



(c) The results using the proposed outlier detection method

Figure 5: Experimental results on Ma_Data.

range of [600–620] and [820–870] in $X_3(t)$. Here, the 1200 data points are separated into 60 subsequences, and each subsequence is 20 data points long. These 60 subsequences are reconstructed as two pieces of new time series datasets which are denoted as $SZ_1 = \{Z_1^1, Z_2^1, \ldots, Z_{31}^1, \ldots, Z_{60}^1\}$ and $SZ_2 = \{Z_1^2, \ldots, Z_{31}^2, \ldots, Z_{41}^2, Z_{42}^2, Z_{43}^2, \ldots, Z_{60}^2\}$, and they are correspond to $X_2(t)$ and $X_3(t)$, respectively. Namely, Z_{31}^1 is the outlier in SZ_1 of $X_2(t)$; Z_{31}^2, Z_{41}^2, Z_{42}^2, and Z_{43}^2 are the outliers in SZ_2 of $X_3(t)$; in particular, Z_{43}^2 is partially an outlier. In the feature space-based outlier detection method, the number of subsequences is 60; k in k-distance is 6; the parameter α in Extended-SAX-based outlier detection method is 4; and Max_p in the proposed outlier detection method is 2. The experimental comparison and results are shown in Figure 5 and Tables 4 and 5.

In summary, from Figure 5 and Tables 4 and 5, the proposed outlier detection method successfully detects the outliers. In this respect, the proposed method outperforms the other compared approaches. However, the proposed method also introduces several higher LOFs of the normal data points and gives out some false alarms; this situation also emerges in both Keogh and ECG datasets; namely, $R_{\text{FalseNegative}}$ and $R_{\text{FalsePositive}}$ are relatively high in the three experimental datasets. It seems to be insufficient. However, just depending on this specificity, some unknown outliers hidden inside the time series might be found by configuring different hierarchy without any prior knowledge or expert opinion. Namely, some new interpretations can be presented with the help of the expert opinions or other domain knowledge. Therefore, these experimental results completely indicate that the proposed method has not only a relatively perfect outlier detection capability for time series but also a potential ability for outlier detection in some unknown fields.

6. Conclusions

In this paper, the HSDE method, HSDE-based outlier detection model, and outlier detection scheme are proposed. The

Table 4: The experimental comparison using $x_2(t)$ of Ma_Data.

Name	$R_{\text{FalseNegative}}$ (%)	$R_{\text{FalsePositive}}$ (%)
The feature space-based method	100	100
The extended SAX-based method	100	100
The proposed method	0	66.7

Table 5: The experimental comparison using $x_3(t)$ of Ma_Data.

Name	$R_{\text{FalseNegative}}$ (%)	$R_{\text{FalsePositive}}$ (%)
The feature space-based method	50	60
The extended SAX-based method	25	50
The proposed method	33.3	33.3

advantages of the proposed method can be summarized as follows:

(1) By the studies, the HSDE-based visual outlier detection method does not require previously observed normal data.

(2) The HSDE-based outlier detection visual method can find outliers by enumerating all of the outlier subsequences and even determine the final outliers in terms of intuition.

(3) It is more practical to assign a factor of being an outlier to each hierarchy of the different subsequences in time series, so that the outlier can be detected directly.

(4) The proposed method visually enumerates the outlier subsequence in time series based on its outlier factor.

(5) The results directly present strong visual evidence for monitoring outliers without any data converting.

However, improvements on the proposed method require further study, for example, how to determine the threshold value of outliers by the proposed algorithm itself and lower the higher false alarm ratio as well as handle "each point" in time series and how to utilize the sliding window technology to separate the time series instead of user-conducted separation, which will be investigated in succeeding studies.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the Future Research Projects Funds for the Science and Technology Department of Jiangsu Province (Grant no. BY2013015-23) and the Fundamental Research Funds for the Ministry of Education (Grant no. JUSRP211A 41).

References

[1] V. Barnett and T. Lewis, *Outliers in Statistical Data*, John Wiley, New York, NY, USA, 1995.

[2] M. A. Osborne, S. J. Roberts, A. Rogers, and N. R. Jennings, "Real-time information processing of environmental sensor network data using Bayesian Gaussian processes," *ACM Transactions on Sensor Networks*, vol. 9, 2012.

[3] Y. Diao, D. Ganesan, and G. Mathur, "Rethinking data management for storage-centric sensor networks," *CIDR*, vol. 7, pp. 22–31, 2007.

[4] J. Yin and M. M. Gaber, "Clustering distributed time series in sensor networks," in *Proceedings of the 8th IEEE International Conference on Data Mining, (ICDM '08)*, pp. 678–687, Pisa, Italy, December 2008.

[5] M. Gupta, A. Sharma B, and H. Chen, "Context-aware time series anomaly detection for complex systems," *Workshop Notes*, 2013.

[6] I. Vasilescu, K. Kotay, D. Rus, M. Dunbabin, and P. Corke, "Data collection, storage, and retrieval with an underwater sensor network," in *Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems (SenSys '05)*, pp. 154–165, ACM, San Diego, Calif, USA, 2005.

[7] G. A. Hollinger, S. Choudhary, P. Qarabaqi et al., "Communication protocols for underwater data collection using a robotic sensor network," in *Proceedings of the 2011 IEEE GLOBECOM Workshops, (GC Wkshps '11)*, pp. 1308–1313, IEEE, Houston, TX, USA, December 2011.

[8] R. Isermann, "Process fault detection based on modeling and estimation methods—a survey," *Automatica*, vol. 20, no. 4, pp. 387–404, 1984.

[9] A. Naess and O. Gaidai, "Estimation of extreme values from sampled time series," *Structural Safety*, vol. 31, no. 4, pp. 325–334, 2009.

[10] Y. Liu and J. A. Tawn, "Volatility model selection for extremes of financial time series," *Journal of Statistical Planning and Inference*, vol. 143, no. 3, pp. 520–530, 2013.

[11] E. Achtert, H.-P. Kriegel, L. Reichert, E. Schubert, R. Wojdanowski, and A. Zimek, "Visual evaluation of outlier detection models," *Lecture Notes in Computer Science*, pp. 396–399, 2010.

[12] D. Dasgupta and S. Forrest, "Novelty detection in time series data using ideas from immunology," in *Proceedings of The International Conference on Intelligent System*, pp. 82–87, 1996.

[13] J. Ma and S. Perkins, "Time-series novelty detection using one-class support vector machines," in *Proceedings of the International Joint Conference on Neural Networks (IEEE '03)*, vol. 3, pp. 1741–1745, Portland, OR, USA, July 2003.

[14] E. Achtert, H.-P. Kriegel, L. Reichert, E. Schubert, R. Wojdanowski, and A. Zimek, "Visual evaluation of outlier detection models," *Lecture Notes in Computer Science*, no. 2, pp. 396–399, 2010.

[15] B. Chiu, E. Keogh, and S. Lonardi, "Probabilistic discovery of time series motifs," in *proceeding of the 9th ACM SIGKDD International Conference on nowledge Discovery and Data Mining (KDD '03)*, pp. 493–498, Washington, DC, 2003.

[16] J. Lin, E. Keogh, S. Lonardi, and etal., "A symbolic representation of time series," in *proceedings of the 8th ACM SIGMOD workshop on research issues in data mining and knowledge discovery*, pp. 2–11, 2003.

[17] B. H. Jun, "Fault detection using dynamic time warping (DTW) algorithm and discriminant analysis for swine wastewater treatment," *Journal of Hazardous Materials*, vol. 185, no. 1, pp. 262–268, 2011.

[18] S. Adwan and H. Arof, "On improving dynamic time warping for pattern matching," *Measurement: Journal of the International Measurement Confederation*, vol. 45, no. 6, pp. 1609–1620, 2012.

[19] B. Lkhagva, Y. Suzuki, and K. Kawagoe, "New time series data representation ESAX for financial applications," in *Proceedings of the 22nd International Conference on Data Engineering Workshops, (ICDEW '06)*, Atlanta, GA, USA, April 2006.

[20] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," in *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD '00)*, vol. 29, pp. 93–104, June 2000.

[21] E. Keogh, S. Lonardi, and Y. Chiu B, "Finding surprising patterns in a time series database in linear time and space," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 550–556, Edmonton, Alberta, Canada, 2002.

[22] A. L. Goldberger, L. A. N. Amaral, L. Glass et al., "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[23] J. Ma and S. Perkins, "Online novelty detection on temporal sequences," in *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Dta Mining (KDD '03)*, pp. 613–618, ACM, Wash, USA, August 2003.

[24] S. Smaha, "Haystack: an intrusion detection system," in *Proceedings of the Fourth Aerospace Computer Security Applications*, pp. 37–44, Orlando, FL, USA, 1988.

[25] D. Pokrajac, A. Lazarevic, and L. J. Latecki, "Incremental local outlier detection for data streams," in *Proceedings of the 1st IEEE Symposium on Computational Intelligence and Data Mining (CIDM '07)*, pp. 504–515, IEEE Press, Honolulu, Hawaii, USA, April 2007.