Research Article

Locality-Based Visual Outlier Detection Algorithm for Time Series

Zhihua Li,1,2 Ziyuan Li,1 Ning Yu,2 and Steven Wen2

1Department of Computer Science, School of Internet of Things Engineering, Jiangnan University, Wuxi, China
2Department of Computer Science, Georgia State University, Atlanta, GA, 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

Copyright © 2017 Zhihua Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

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

In the following, \( x_j \) is used to represent the corresponding subsequence \( \langle x_{i-p}, \ldots, x_i, \ldots, x_{i+p}\rangle \) 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 (HSDSE) 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 \( X, \langle x_{i-p}, \ldots, x_i, \ldots, x_{i+p}\rangle \) is a subsequence of \( X \). \( x_j \) is \( \text{arg max}(x_{i-p}, \ldots, x_j, \ldots, x_{i+p}) \); then the returned value of EPD is noted as \( \text{Flag} = 1 \); if \( x_j \) is \( \text{arg min}(x_{i-p}, \ldots, x_j, \ldots, x_{i+p}) \); then the returned value of EPD is noted as \( \text{Flag} = -1 \); otherwise, \( \text{Flag} = 0 \). 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 \( \text{Max.}p \) 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 \( \text{HM} = \{h_{m_1}, h_{m_2}, \ldots, h_{m_j}, \ldots, h_{m_p}\} \) and \( h_{m_j} (j = 1, 2, \ldots, n) \) 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.

(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 expressed in 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 (HSDSE) 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 of the 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(676,680),(909,995) that 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 $\text{DTW}(x, o') < \text{DTW}(x, o)$ and (2) for at most $k - 1$ objects it holds that $\text{DTW}(x, o') > \text{DTW}(x, o)$. $k$ is a positive integer which always represents the number of objects and must be predetermined by experimentation. Additionally, the $k$-distance neighborhood of an object $x$ contains each object whose distance from $x$ is not greater than the $k$-distance; that is,

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

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:

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

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 $x_4$ 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.

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

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 $0 \leq \rho_k(x) \leq 1$.

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

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 $O(\log n \text{Max}_p)$. It is clear that the time complexity is similar to the other stages of the HSDE-based outlier detection algorithm and is no more than $O(\log n \text{Max}_p)$, in which each stage is conducted sequentially. Therefore, the total time complexity of the HSDE-based outlier detection algorithm is $O(\log n \text{Max}_p)$. 
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:

\[ R_{\text{FalseNegative}} = \frac{FP}{TN + FP} \]  \hspace{1cm} (5)  
\[ R_{\text{FalsePositive}} = \frac{FN}{TN + FN} \]  \hspace{1cm} (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>
<thead>
<tr>
<th>Reality</th>
<th>Detected normal</th>
<th>Detected outliers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>True positive (TP)</td>
<td>False negative (FN)</td>
</tr>
<tr>
<td>Outliers</td>
<td>False positive (FP)</td>
<td>True negative (TN)</td>
</tr>
</tbody>
</table>

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
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 $SX = \{X_1, ..., X_{11}, ..., X_{20}\}$, 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 $\alpha$ in Extended-SAX-based outlier detection method is 4; and $\text{Max}_p$ 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-defined 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 $X_{11}$ is the maximum, the value of $X_7$ 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 $X_5$, $X_7$, 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 $SY = \{Y_1, ..., Y_{16}, Y_{17}, ..., Y_{25}\}$, 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 $\alpha$ in Extended-SAX-based outlier detection method is 4; and $\text{Max}_p$ 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-defined 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.

<table>
<thead>
<tr>
<th>Name</th>
<th>$R_{\text{FalseNegative}}$ (%)</th>
<th>$R_{\text{FalsePositive}}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The feature space-based method</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>The extended SAX-based method</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>The proposed method</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3: The experimental comparison using ECG Data.

<table>
<thead>
<tr>
<th>Name</th>
<th>$R_{\text{FalseNegative}}$ (%)</th>
<th>$R_{\text{FalsePositive}}$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>The feature space-based method</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>The extended SAX-based method</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>The proposed method</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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 $\text{Max}_p$ in the proposed method. The length of subsequence is separated and marked by the parameter of $\text{Max}_p$. 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 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
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 $SZ_1 = \{Z_1^1, Z_1^2, \ldots, Z_1^{31}, \ldots, Z_1^{60}\}$ and $SZ_2 = \{Z_2^1, Z_2^2, \ldots, Z_2^{31}, \ldots, Z_2^{41}, Z_2^{42}, Z_2^{43}, \ldots, Z_2^{60}\}$, 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 $\alpha$ in Extended-SAX-based outlier detection method is 4; and $\text{Max}_p$ 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, $R_{\text{False Negative}}$ and $R_{\text{False Positive}}$ 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
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


