Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 979035, 6 pages
http://dx.doi.org/10.1155/2013/979035

Research Article

The Outlier Interval Detection Algorithms on Astronautical Time Series Data

Wei Hu and Junpeng Bao

School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Correspondence should be addressed to Junpeng Bao; baojp@mail.xjtu.edu.cn

Received 6 November 2012; Accepted 11 February 2013

Academic Editor: Tsung-Chih Lin

The Outlier Interval Detection is a crucial technique to analyze spacecraft fault, locate exception, and implement intelligent fault diagnosis system. The paper proposes two OID algorithms on astronautical Time Series Data, that is, variance based OID (VOID) and FFT and k nearest Neighbour based OID (FKOID). The VOID algorithm divides TSD into many intervals and measures each interval's outlier score according to its variance. This algorithm can detect the outlier intervals with great fluctuation in the time domain. It is a simple and fast algorithm with less time complexity, but it ignores the frequency information. The FKOID algorithm extracts the frequency information of each interval by means of Fast Fourier Transform, so as to calculate the distances between frequency features, and adopts the KNN method to measure the outlier score according to the sum of distances between the interval's frequency vector and the K nearest frequency vectors. It detects the outlier intervals in a refined way at an appropriate expense of the time and is valid to detect the outlier intervals in both frequency and time domains.

1. Introduction

The Time Series Data is a sequence of values observed in some periods. Usually, it takes a long time to continuously observe an object and record its data so the accumulated TSD is often in a very large amount. A significant issue is that how to mine latent or interesting knowledge from the huge amount of TSD and apply what is mined to the future. The Astronautical data (AD) is a typical big TSD, which is gathered by frequently and continuously observing spacecrafts. The AD mining techniques have many significant applications, including but not limited to analyzing satellites' working status, judging faults/errors, and forecasting its next state in the near future. Since it is hard to touch or measure spacecrafts in a shouting distance. The technique to detect outlier intervals from AD is important. The outlier intervals are the exceptional or unusual parts in a long period of the TSD, which cover abundant information about spacecraft faults or special events. The Outlier Interval Detection technique can provide foundation proofs for intelligent astronautical fault analysis, diagnosis and prediction.

Although, most spacecrafts work well in most of their lives, the faults and abnormal situations take place occasionally. Namely, the outlier intervals are always drowned in long pieces of regular data. Obviously, it is a very tough mission for a person to find out outliers and make out regulations from the huge AD. Human experts will spend not only plenty of efforts and time, but also some luck. This paper proposes two algorithms to automatically detect the outlier intervals in AD by means of data mining. These algorithms not only are able to promote analysis efficiency, raise diagnosis system's response performance, but also can be used to support spacecraft fault prediction and diagnosis system as well as astronautical intelligent monitoring system. Additionally, the OID technique has a great perspective in the area of industry production process monitoring, financial administration, medical data analysis, disaster alarm, network intrusion detection, credit card fraud detection, and so forth.

2. Related Work

Generally, there are four ways to implement Outlier Interval Detection, that is, distance based approach, statistic based approach, deviation based approach, and clustering based approach.

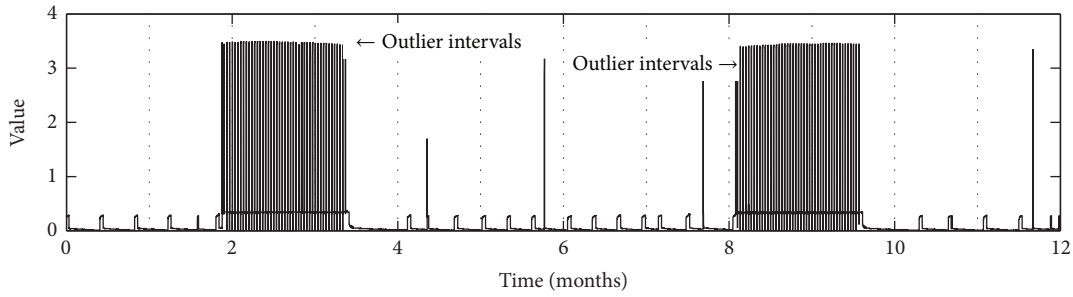

Figure 1: The examples of the outlier interval.

(1) The distance based OID calculates the distance between objects firstly, and then the outliers are defined as the objects whose distances to others exceed the given threshold [1]. This is a popular approach because it is simple and easy to implement and does require the knowledge of the data distribution in advance.

Angiulli et al. have done a lot of research on this way [2–5]. Angiulli and Pizzuti [2] defined an object's outlier score as the sum of distances between the object and its K nearest neighbors. They also introduced a notion of outlier detection solving set [3], a subset of the data set, to improve the distance based outlier detection performance. Angiulli and Fassetti [4] proposed a DOLPHIN algorithm to carry out the distance based outlier detection in the large database. Later, they proposed a new model [5] that considers only the sum of the distances between the object and the others in the sliding window.

Chen et al. [6] proposed a rough set and K Nearest Neighbor (KNN) based method to detect the outliers on the mixed continuous and discrete dataset. Bhaduri and Matthews [7] introduced two distributed algorithms and a novel indexing scheme to speed up the distance based outlier detection. Li et al. [8] transformed interval values into real values and then employed the KNN approach to find outliers.

(2) The statistic based OID approach requires the knowledge the probability distribution of the data in advance, which is the foundation of the approach. But for a specific sample space, it is usually hard to know the exact distribution of the data. The key work of the approach is to perform plenty of tests in order to get the most proper distribution model. For example, Takeuchi and Yamanishi [9] proposed a unifying framework that employed a Gaussian mixture model for statistical outlier detection.

(3) The deviation based OID extracts the main features of the center objects, and then the outliers are considered as the objects whose features deviate from the centers remarkably. For example, Oliveira and Meira [10] proposed a neural network method to forecast the thresholds for detecting outliers.

(4) The clustering based OID considers the outliers as the by-products of clustering [11–13], that is, the objects that do not belong to any normal cluster.

Some OID methods exploit the Fourier Transform and the Wavelet Transform to extract data features and mine the outliers in a special domain. For example, Rasheed et al. [14] proposed a Fast Fourier Transform and Inverse Fast Fourier Transform based OID method. Grané and Veiga [15] propose a Wavelet Transform based method to detect outliers in financial data.

3. The Outlier Interval Detection

3.1. The Solo Variant TSD and the Outlier Interval. The solo variant TSD is the observed value of the single object or attribute in a period. Generally, a long period of TSD is divided into many intervals according to the time scale. Thus an interval is a piece of time and the values in the range. The time span of each interval can be equal to each other, such as a day, a week, or 10 days. Some applications may employ unequal interval. But in this paper, a TSD is divided into identical length of intervals.

The OID of the solo variant aims to find out the oddest K outlier intervals from a long period of solo variant TSD, such as the changes of a satellite battery's voltage in a year. An outlier interval indicates that the variant varies abnormally in the time span so that it deserves a special care. It often reflects some events happened or the symptoms of a fault. The reason of the abnormal data may be diverse, such as device fault, external interference, changes of temperature, and so forth. Figure 1 illustrates some examples of the outlier interval where the 3rd, 4th, 9th, and 10th intervals are apparently different from others.

3.2. The Variance Based Outlier Interval Detection. In a real astronautical dataset, the abnormally varying data often result in a great fluctuation amplitude. The degree of the amplitude can be directly reflected by the variance in the interval. Namely, the outlier score can be measured by the variance. The higher the value of the variance is, the odder the interval. It is easy to mine the top K oddest intervals by means of sorting their variances in descending order.

The time complexity of the standard variance definition is $O(N^2)$, where N is the number of data because it traverses twice the whole sequence. For a huge amount of AD, it is better to execute a faster algorithm with a smaller time com plexity. We transform the standard variance defination to the form given in (1). Both results are identical, but (1) needs only

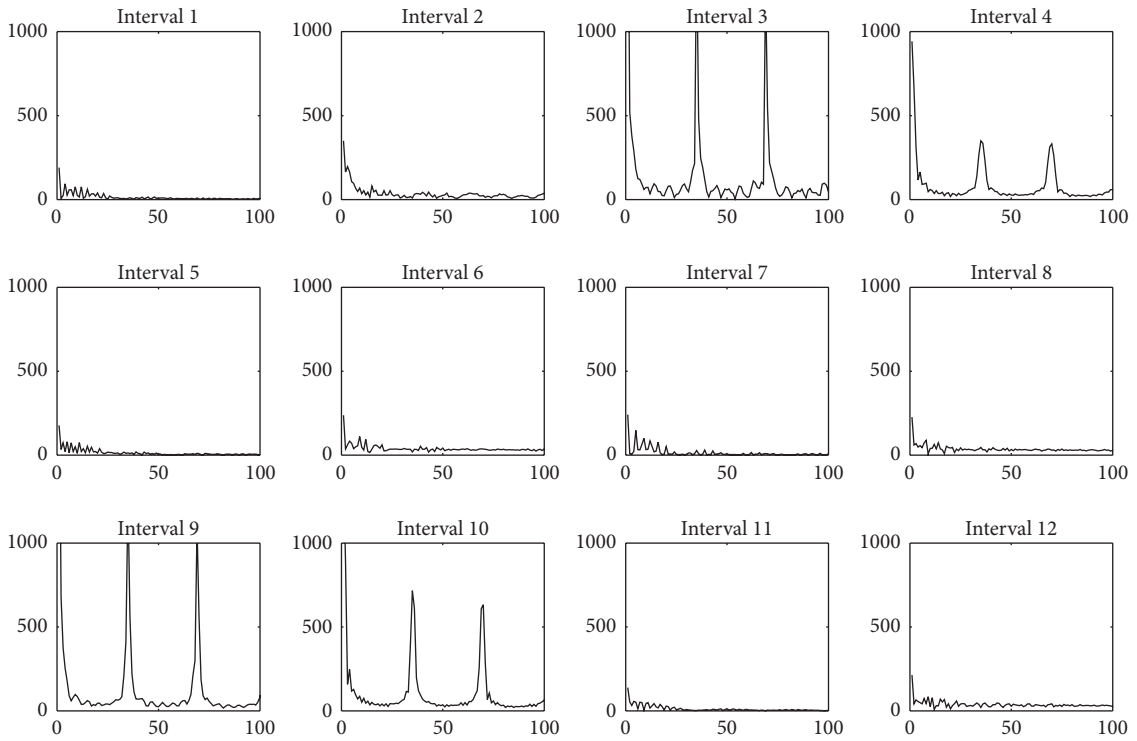

Figure 2: The frequency spectra of 12 intervals of the data shown in the Figure 1.

Input: X is the TSD, and c is the count of the intervals.
Output: The top K outlier intervals in time domain and their scores.
(1) T is the set of c intervals divided from X, that is, $T = \{t_i \mid \bigcup_{i=1}^{c} t_i = X,\ t_i \cap t_j = \phi (i \neq j)\}$
(2) for each t_i in T
(3) $a = \sum_{x_j \in t_i} x_j^2,\ b = \sum_{x_j \in t_i} x_j$
(4) $\text{var}(t_i) = \dfrac{a}{|t_i|} - \left(\dfrac{b}{|t_i|}\right)^2$
(5) endfor
(6) $\text{Outlier}(T, K) = \text{argmax}_{t_i \in T}^{K}\{\text{var}(t_i)\}$
The time complexity of the VOID is $O(N)$ where N is the number of data in the TSD.

Pseudocode 1

once traverse so its time complexity is $O(N)$ one has

$$\text{var} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\overline{x})^2. \tag{1}$$

Pseudocode 1 is the pseudo code of the VOID.

3.3. FFT and K Nearest Neighbour Based Outlier Interval Detection. The VOID algorithm can quickly detect the outlier intervals in time domain. However, many real AD are periodical in some extent. The violent frequency fluctuations also imply something happened. On the other hand, the violent fluctuations in time domain lead to changes in frequency domain. Figure 2 illustrates the frequency spectra of the twelve intervals of the data shown in Figure 1. It is clear that the frequency fluctuations in the 3rd, 4th, 9th, and 10th intervals are distinctly greater than others.

In order to detect the outliers in a fine granularity, more frequencies have to be taken into account. So a feature vector of an interval is made of the whole frequency band from the lowest frequency to the highest. The outlier score of an interval is measured according to the distance between feature vectors instead of variance. Moreover, an amplitude threshold is set to decline noises, namely; the value is assigned to 0 if it is not higher than the threshold, otherwise it keeps its value.

The FKOID algorithm firstly divides the whole TSD into c intervals equally. Secondly, it executes FFT on each interval to get the frequency spectrum of the interval, which builds a feature vector after the low energy frequencies are set to 0 by the amplitude threshold. Thirdly, the Euclidean distances between every pair of feature vectors are calculated. It should

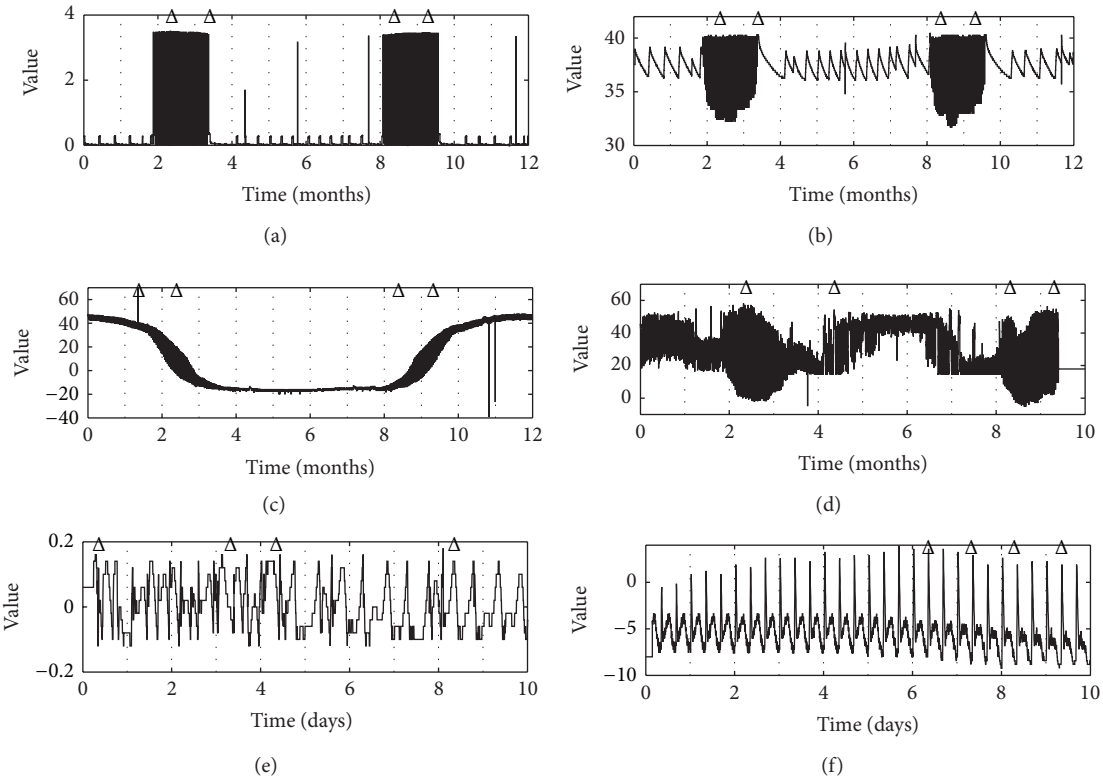

Figure 3: The VOID results on the 6 real astronautical data where the symbol Δ marks the outlier intervals.

Table 1: The top 4 outlier scores of both algorithms on the 6 real astronautical datasets where the number in the brackets is the ID of the interval.

Algorithm	(a)	(b)	(c)	(d)	(e)	(f)
VOID Figure 3	1.2001(3), 1.0300(9), 0.6976(10), 0.3677(4)	1.7812(4), 1.6907(9), 1.4850(3), 1.3747(10)	131.66(3), 77.06(10), 72.81(9), 21.37(2)	135.83(5), 128.09(10), 127.37(3), 110.52(9)	0.0076(5), 0.0062(1), 0.0055(4), 0.0055(9)	5.4093(10), 5.3548(7), 5.3369(8), 5.1904(9)
FKOID Figure 4	7211(3), 6515(9), 6027(10), 4862(4)	9013(3), 8856(9), 7145(10), 5746(4)	28163(3), 28083(9), 26842(10), 18941(2)	39757(5), 39619(2), 39182(1), 38768(4),	60.94(4), 55.02(1), 53.10(5), 53.03(9)	1479(1), 1306(10), 1149(7), 1096(2)

be noted that some frequencies may have an extremely large amplitude so that they are overwhelming in the vector. That will cover the other frequencies' effect in terms of Euclidean distance. So we set a top threshold to limit the maximum amplitude value of those overwhelming frequencies. At last, the outlier score of an interval is the sum of distances between the interval's feature vector and its K nearest neighbors.

The idea of the FKOID is inspired by the method of Grané and Veiga [15]. The longer the vector's distance is, the odder the interval is. Pseudocode 2 is the pseudo code of the FKOID.

The FKOID algorithm has to maintain a distance matrix and fetches the k nearest neighbors of each interval. The time complexity of that is $O(kwc^2)$, where k is the number of nearest neighbors, w is the number of data in an interval, and c is the count of intervals. Since the first stage is the Fast Fourier Transform of each interval, the entire time complexity of the FKOID is $O(N \times \lg(w) + kwc^2)$.

4. Experimental Results

We use six real astronautical datasets to test the algorithms. As shown in Figure 3, these AD represent some popular tendencies in the real world. The data of (a), (b), and (c) have distinct violent fluctuations in time domain. The data of (e) and (f) have no great fluctuation in time domain. The data of (d) is the most complicated, which contains diverse variation tendency. Figure 3 illustrate the top 4 outlier intervals detected by the VOID algorithm, and Figure 4 is

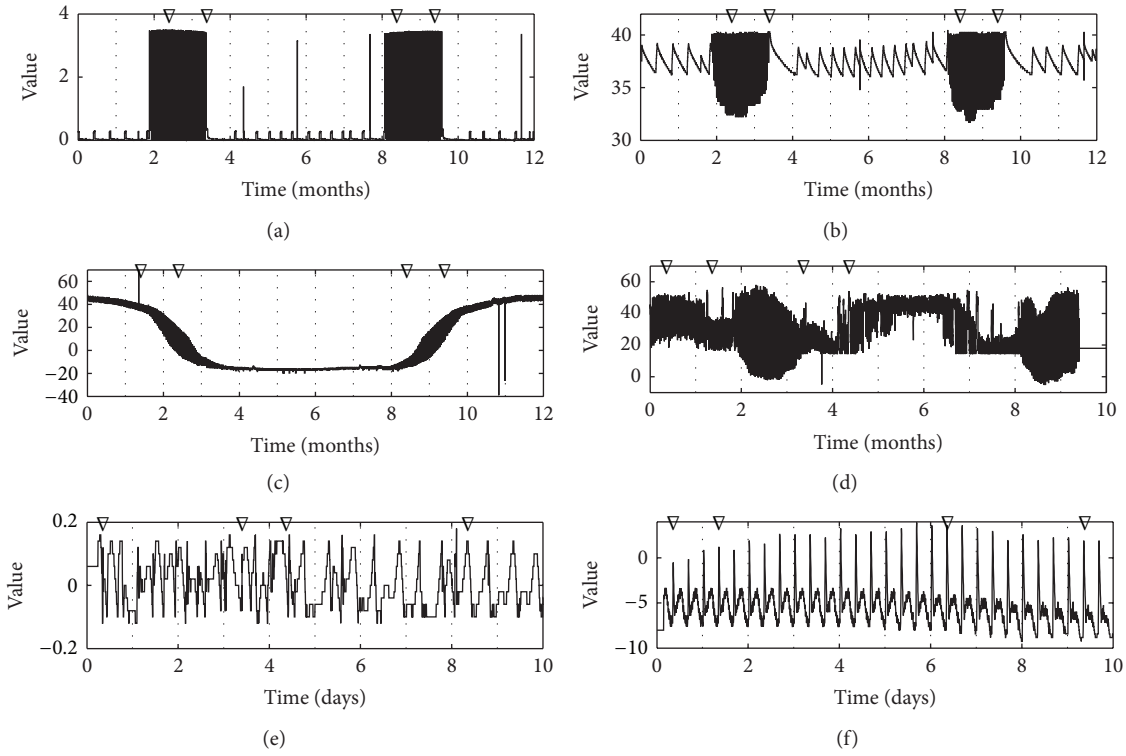

Figure 4: The FKOID results on the 6 real astronautical data where the symbol ∇ marks the outlier intervals.

Input: X is the TSD, and c is the count of the intervals, $[\tau_1, \tau_2]$ are the amplitude thresholds.
Output: The top K outlier intervals and their scores.

(1) T is the set of c intervals divided from X, that is, $T = \{t_i \mid \bigcup_{i=1}^{c} t_i = X, t_i \cap t_j = \phi(i \neq j)\}$

(2) for each t_i in T

(3) $f_{t_i} = \mathrm{FFT}(t_i) \otimes [\tau_1, \tau_2]$, where $a \otimes [\tau_1, \tau_2] = \begin{cases} \tau_2, & a \geq \tau_2 \\ a, & \tau_1 < a < \tau_2 \\ 0, & a \leq \tau_1 \end{cases}$

(4) endfor

(5) for each t_i in T

(6) $d(t_i) = \sum_{f_{t_j} \in knn(f_{t_i})} |f_{t_i} - f_{t_j}|^2$

(7) endfor

(8) $\mathrm{Outlier}(T, K) = \mathrm{argmax}_{t_i \in T}^{K} \{d(t_i)\}$

Pseudocode 2

the results of FKOID, respectively. The Table 1 lists the detailed outlier scores on the 6 AD.

Based on the experimental results, the VOID algorithm is fit for the case that data varies slightly in the regular situation whereas it becomes violent in the irregular situation. If the normal data waves frequently and varies widely, then the VOID algorithm will have a great error. Additionally, the VOID algorithm is suitable to detect OID in time domain, but failed in the case that data varies peacefully in time domain but violently in frequency domain.

The FKOID algorithm can solve the problem of OID in frequency domain. It is shown in Figure 4 that the violent fluctuations in time domain have corresponding abnormal changes in the frequency domain. As a result, the FKOID algorithm can detect outlier intervals in time domain as well. However, the FKOID algorithm can detect the very subtle outliers at the expense of long running time because it has a high time complexity.

5. Conclusions

The OID technique can quickly deal with TSD to find the oddest objects, which often imply crucial exceptional events. It has a great perspective in the astronautical applications, such as the spacecraft fault prediction and diagnosis system,

astronautical intelligent monitoring system and other systems based on TSD.

This paper proposes two algorithms to detect the outlier intervals on astronautical data. The VOID algorithm directly exploits the variance of data to quickly detect the outlier intervals in time domain. The FKOID employ the full frequency band to build a feature vector of an interval and measure the outlier score by the distances sum of the K nearest neighbors. The FKOID is subtle enough to detect the outliers in a refined granularity in both frequency and time domains, but its time complexity is a little big.

However, the above algorithms are based on the identical length of interval. It is rather arbitrary in practice because the real outlier intervals may be varying in length. So it is our next work to study on the methods of the unequal length Outlier Interval Detection.

Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant 60903123) and the Baidu Theme Research Plan on Large Scale Machine Learning and Data Mining.

References

[1] E. Knorr and R. Ng, "Tucakov. Distance-based outliers: algorithms and applications," *VLDB Journal*, vol. 8, no. 3, pp. 237–253, 2000.

[2] F. Angiulli and C. Pizzuti, "Outlier mining in large high-dimensional data sets," *IEEE Transaction on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 203–215, 2005.

[3] F. Angiulli, S. Basta, and C. Pizzuti, "Distance-based detection and prediction of outliers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 2, pp. 145–160, 2006.

[4] F. Angiulli and F. Fassetti, "An efficient algorithm for mining distance-based outliers in very large datasets," *ACM Transactions on Knowledge Discovery from Data*, vol. 3, no. 1, pp. 1–57, 2009.

[5] F. Angiulli and F. Fassetti, "Distance-based outlier queries in data streams: the novel task and algorithms," *Data Mining and Knowledge Discovery*, vol. 20, no. 2, pp. 290–324, 2010.

[6] Y. Chen, D. Miao, and H. Zhang, "Neighborhood outlier detection," *Expert Systems with Applications*, vol. 37, no. 12, pp. 8745–8749, 2010.

[7] K. Bhaduri and B. L. Matthews, "Algorithms for speeding up distance-based outlier detection," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD '11)*, pp. 859–867, 2011.

[8] S. Li, R. Lee, and S. D. Lang, "Detecting outliers in interval data," in *Proceedings of the Southeast regional conference (ACM-SE '06)*, pp. 290–295, 2006.

[9] J. Takeuchi and K. Yamanishi, "A unifying framework for detecting outliers and change points from non-stationary time series data.," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 4, pp. 482–492, 2006.

[10] A. Oliveira and S. Meira, "Detecting novelties in time series through neural networks forecasting with robust confidence intervals," *Neurocomputing*, vol. 70, no. 1–3, pp. 79–92, 2006.

[11] C. Böhm, C. Faloutsos, and C. Plant, "Outlier-robust clustering using independent components," in *Proceedings of the 28th ACM SIGMOD International Conference on Management of Data (SIGMOD '08)*, pp. 185–198, 2008.

[12] N. Ade and B. Zadronzy, "Outlier detection by active learning," in *Proceedings of the 12th ACM SIGMOD International Conference on Management of Data (SIGMOD '06)*, pp. 504–509, 2006.

[13] C. Franke and M. Gertz, "ORDEN: outlier region detection and exploration in sensor networks," in *Proceedings of the 29th ACM SIGMOD International Conference on Management of Data (SIGMOD '09)*, pp. 1075–1078, 2009.

[14] F. Rasheed et al., "Fourier transform based spatial outlier mining," *Lecture Notes in Computer Science*, pp. 317–324, 2009.

[15] A. Grané and H. Veiga, "Wavelet-based detection of outliers in financial time series," *Computational Statistics & Data Analysis*, vol. 54, no. 11, pp. 2580–2593, 2010.