2015 International Conference on Computer Science and Applications

The Financial Data of Anomaly Detection Research based on Time Series

Chen-Ming Guo
School of Computer Engineering and Science, Shanghai University,
Shanghai, China
e-mail: 15821081538@163.com

Ling-Yu Xu
School of Computer Engineering and Science, Shanghai University,
Shanghai, China
e-mail: xly@shu.edu.cn

Hui-Fang Liu
School of Computer Engineering and Science, Shanghai University,
Shanghai, China
e-mail: 925937109@qq.com

Lei Wang
School of Computer Engineering and Science, Shanghai University,
Shanghai, China
e-mail: 12716328@qq.com

Xiang Yu
School of Computer Engineering and Science, Shanghai University,
Shanghai, China
e-mail: 260442576@qq.com

Bo Han
School of Computer Engineering and Science, Shanghai University,
Shanghai, China
e-mail: 289757418@qq.com

***Abstract*—The rapid development of financial markets make the financial data variability and unpredictability, the abnormal fluctuations in financial data often contain important information. Financial data is generated over time, so the time series mining method widely used in the financial data of anomaly detection. The traditional time series of anomaly detection method is to find out one of the biggest point of outliers from a series of randomly generated numbers. It often do not consider the time sequence of time series, and the concern is not only a point of unusual data, but the abnormal sub-sequences in the time series anomaly detection. This paper presents a method that combines the activity and density of time series. It uses the time sequence of time series and sub-sequences' features effectively to discover the anomalies. Experimental results show that this method can be more effective and accurate to find the anomalies from the time series of financial data.**

Keywords-time series; data mining; anomaly detection; density

I. INTRODUCTION

Time series anomaly detection is an important aspect of data mining, widespread in various fields, such as the financial field, the Internet industry and so on. Hawkins [1] gives a definition of abnormal essential: abnormity is the different data in a data set, make people suspect that the different data is not a random deviation but in a completely different mechanism. Anomaly detection of research on algorithm can be divided into five categories: the method based on distribution [2]; Based on the depth [3]. Based on the clustering method [4]; the method based on distance and the method based on density [5].

Based on the distribution and the method based on the depth have developed in the field of statistics. The method based on the distribution assume the known data set of distribution. According to the distribution of the data set for each object conformance testing, if the object and the distribution does not conform to, then think it is abnormal. In fact, the distribution of actual data sets are often unknown, and are often difficult to obtain in estimates; The method based on the depth is only suitable for 2 d or 3 d data, for more than four dimensional data set it is inefficient; The method based on distance and based on density are more popular algorithm in anomaly detection at this stage. The common denominator is to find the deviation from the public data as an abnormal point. The method based on distance usually does not consider the time sequence of data set, so it can bring great influence to the result in the time series of anomaly detection. Method based on density is developed on the basis of method based on distance, thus bound by the same constraint. Many optimization clustering method was used to find the public data and dig out the abnormal information from the vast amounts of data. Data clustering is one of the most commonly used and the most effective means in data mining.

The above-mentioned methods are given to find out the Markedly different data points from the data set. They often do not consider the time sequence of time series data set, and time series of anomaly detection is not always focus on the points on the time series but its sub-sequence of abnormal fluctuations. In this paper, the time series will be divided into different successive time sub-sequences. The method in this paper combines the successive time sub-sequences' activity and the relative density of time sub-sequence to determine the time sequence of abnormal points and sub-sequences.

II. RELATED WORK

When dealing with the anomaly detection, different clustering strategy[6] of obtaining abnormal points set has been put forward. With continuous renewal of clustering methods, in

978-1-4799-9961-3/15 $31.00 © 2015 IEEE
DOI 10.1109/CSA.2015.42

recent years, there are some methods which were combined by distance and density [7] has been proposed. These methods use K-medoids to calculate distance and include the idea of DBSCAN, which solve the problem that the K-medoids method has a bad result on processing aspheric data. However, time series data in time order has not been considered. The theme of K-medoids and methods K-means algorithm [8-9] is to gathering those points close to clustering center. The definition of clustering center is the average value of all date in each category. The result depends on the choice of the initial cluster centers, which is bad effect on dealing with aspheric data [10].

The anomaly detection based on time series is to find out a largest sub sequence which is most different with others in time sequence T, such the sequence is called abnormal sequences. This paper is to identify the most "disharmony" sequence from a given time sequence. Although there are many experts and scholars research on anomaly detection currently, but there is no universally accepted definition of anomaly detection until now. In recent years, as a branch of data mining, the abnormality detecting is being more and more attention and study. Researchers have proposed many methods about anomaly detection, for example, Ma et al. [11] used the vector regression model to train historical events sequence, when new time series data deviate from the model, it is considered to be a strange sequence. Shahabi et al [12] proposed an improved TSA-Tree algorithm, which through the local maxima of the wavelet coefficients to search the activity of different sequences and locate abnormal patterns, but the drawback is that the definition of abnormal patterns in the article is based on wavelet coefficients, some singular model cannot find. Moradi, who proposed an anomaly detection method based on time series RF and SVM classifier. Richard et al. [12] proposed a new anomaly time series detection algorithms based on phase space.

III. BASED ON TIME SERIES ACTIVITY AND DENSITY OF ABNORMAL FINDINGS

To find abnormal sub-sequences of time series, the whole time series is divided into different sub-sequences by using the sliding window in this paper. Then we calculate the activity between adjacent sequences using the area method. If the sub-sequences' activity is bigger than the entire sequence's average activity , we combined it with its k distance to calculate the sub-sequences' local density. Given the density threshold a, if the density of the sub-sequence is bigger than a, the sub-sequences is abnormal. Related definitions are given following:

Definition 1 Time series: $X = <x_1 = (v_1, t_1), x_2 = (v_2, t_2), .., x_T = (v_T, t_T)>$ is a time sequence. T which is strictly increasing belongs to the set of real numbers, representing the length of time series.

Typically, the abnormal time digging time series, is no longer just concerned about the unusual circumstances of a point on the entire time series but the abnormal fluctuations in the sequence, and therefore, the need for time-series sequence is described, which is defined as follow:

Definition 2 Sub-sequences: Given a time series X whose length is T, the sub-sequence is defined as: $S = X_p, X_{p+1}, .., X_{p+o-1}$, where P is the starting point in the sequence of the original data point and o is the length of S.

Definition 3 Sliding window: Given a time series X whose length is T, a sub-sequence S whose length is n, then we define n as the size of the sliding window. The sliding window moves over the time series X, and intercepts X into a collection G which constituted of all Sub-sequences.

Definition 4 Activity between sub-sequences: Given the collection G, $A = \{a_1, a_2 ... a_n\}$ and $B = \{b_1, b_2 ... b_n\}$ are two adjacent sub-sequence in G.AS represents the area which consists of the sub-sequence A and the x-axis.AS:

$$\text{AS}=\frac{1}{2}*\sum_{i=1}^{n-1}(a_i + a_{i+1}) \qquad (1)$$

Where a_i is the ith point's value in A.BS is in the same way , the sub-sequence activity of sequences A and B is Si(A, B)

$$\text{Si(A,B)}=\frac{1}{2}*|\text{AS} - BS| \qquad (2)$$

The bigger the value of Si(A,B) ,the lower activity between the two sub-sequence. The smaller the value of Si(A,B), the higher activity between the two sub-sequence .

Definition 5 The average activity of time series: given the length of T, time series X , the sliding window whose length is n ,and the sub-sequences set denoted by $G = \{G_1, G_2, .. G_n\}$, then the average activity of time series is defined as following:

$$\text{Z(X)}=\frac{1}{\text{n}}*\sum_{i=1}^{n}\text{Si}\left(G_i, G_{i+1}\right) \qquad (3)$$

Where Gi is the i-th sub-sequence in G and n denotes the number of sub-sequences.

If the sub-sequence's activity is greater than the average activity, we calculate the density . Following is the method, first define the distance between time series sequences and the k Distance:

Definition 6 The distance of time sub-sequences: given the sub-sequence's collection G ,the sub-sequence A = {a1, a2 ... an} and B = {b1, b2 ... bn} are two sub-sequences in G, each sequence was seen as a point of n-dimensional space, the distance of time sub-sequences(A and B) (Euclidean distance):

$$\text{D(A,B)}=\sqrt{\sum_{i=1}^{n}\left(A_i - B_i\right)^2} \qquad (4)$$

Definition 7 The K distance of sub-sequence p: For positive integer k, the K distance of sub-sequence p can be referred as the k-distance (p). In the sample space, there is a sub-sequence o, p and o's distance referred as D (p, o). If the following two conditions are met, we think that k-distance (p) = D (p, o).

(1) In the sample space, there are at least k time sequence q, such that D (p, q) <= D (p, o);

(2) In the sample space, there are at most k-1 time sequence q, such that D (p, q) <D (p, o);

Definition 8 The k-distance neighborhood of time sub-sequence p: Given the k-distance of the time sequence p, then the distance between the time sequence p less than or equal to k-distance (p) is called the neighborhood, denoted as: Nkdisk (p) .The neighborhood's center is p, and the radius of the area is k-distance (p) .Since there may exist many k-distance, the set includes at least k time sequence.

Definition 9 Time density sub-sequence p: Given time sub-sequence of the p with k-distance (p), and the time sub-sequence p with the neighbourhood Nkdisk (p), the density is defined as:

$$\mathrm{M(p)}=\frac{1}{\mathrm{k}}*\sum_{q\in \mathrm{Nkdisk(p)}}\frac{\mathrm{k-distance(q)}}{\mathrm{k-distance(p)}} \tag{5}$$

Definition 10 Abnormal time sequence: Given time series X, the sub-sequence p, we can get the activity of sequences p. If the activity of p is greater than the whole time series' activity and obtained the number of the neighbourhood by mapping p into multidimensional space. We can calculate the density .With the threshold u, if and M (p) <u, p will be the abnormal time sequence.

A. Algorithm for pseudo-code

```
Input:Time series $X = <x_1 = (v_1, t_1), x_2 = (v_2, t_2)..,x_T = (v_T, t_T)>$
Best_ anomaly_ sub-sequence=0;
K_distance=k; Z=0;Sliding_window=S; X_length=T;
G：={}//the set of sub-sequences in sequence X
M=u;// the threshold of time series
For p=1 to T
    G.add(Xp-Xp+s-1);// get the sub- sequences set $G = \{G_1, G_2, ..G_n\}$;
For j=1 to n
    Using formula (1)、(2)、(3) get Z  // The average activity of time series
End
For j=1 to n //Scan the G of sub-sequences
    If(Si[Gj,Gj+1])>Z// Using formula (2)
        Continue;
    Else
        Using formula (4)、(5)get M(Gj)
        If(M(Gj)<u)
            Best_ anomaly_ sub- sequence= Gj;
End
Return  Best_ anomaly_ sub- sequence;
```

IV. EXPERIMENTAL RESULTS

This section firstly describes the data sources and evaluation methods of experiments, then we verify the theory proposed in this paper from two angles, in section 4.2, we apply our approach and the other method on the benchmark for the purpose of comparison. We use precision, recall and F1 as parameters for evaluation. The standard criterion is from the professional news and post content.

A. Data sources and evaluation methods

In order to verify the accuracy and applicability of this method, the experimental data sets are from two information spaces, one space from the online financial space in the stock forum, and the other from reality financial space in the stock market.

Cyberspace data is the number of posts that are from four stock areas in the Eastern wealth network. Post time is from January 2012 to March 2013, their stock codes are 600559,600199,600779,600600. The time series A is divided according to the time granularity $O = \{day\}$. This paper selected the real space stock volume data which has the same areas with the posts, the time is the same as the stocks in posts and divided by time granularity $O = \{day\}$ into time series B.

In this paper, F-1, precision and recall will be as the experimental evaluation index, which is based upon abnormal content of the post and the news is given as a criterion. LetC.day data is the time set corresponding to standard outliers set in set A, and the M.day is the calculated outliers time set corresponding to our method, the evaluation index is calculated as follows:

$$\text{Precision}\ \mathrm{P} = | M.day \cap C.day \,|\, / \,|\, M.day \,|$$

$$\text{Recall}\ \mathrm{R} = | M.day \cap C.day \,|\, / \,|\, C.day \,|$$

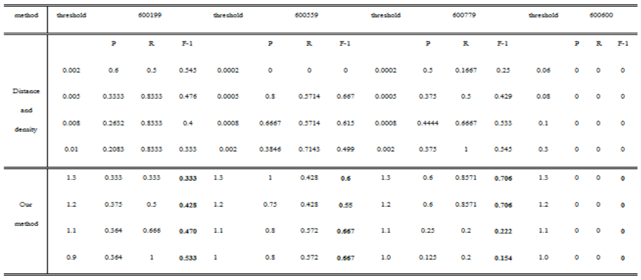
$$F\text{-}1 = 2 * \mathrm{P} * \mathrm{R} / (\mathrm{P} + \mathrm{R})$$

B. Comparison with experimental results

This article selects the method proposed in literature [11] as the comparison method to analysis, and selected four different levels of thresholds to compute the precision, recall and F-1, the method proposed in this paper select the same levels of threshold. The following table 1 is the experimental results in cyberspace, table 2 is realistic space experiment results.

Table 1 is the result of the four different areas in cyberspace, we can see from table 1, the change of the results of four cases in different thresholds, when the F-1 get the optimal value the thresholds in three areas are not the same, this shows that in order to get a better result must be identified better thresholds value, which in 600600 did not find any right of abnormal data shows that this method is more sensitive to the data source area. The method of four different levels of the results is stable, the best result is better than the results in the literature [11], it shows that the method proposed in this paper has better accuracy. Table 2 is the result of the four areas in real space, the situation is similar to table 1, the result is better than that of the comparison method.

TABLE I. R、P AND F-1OF TWO METHODS COMPARISON IN CYBERSPACE



method	threshold	600199			threshold	600559			threshold	600779			threshold	600600		
		P	R	F-1		P	R	F-1		P	R	F-1		P	R	F-1
Distance and density	0.002	0.6	0.5	0.545	0.0002	0	0	0	0.0002	0.5	0.1667	0.25	0.06	0	0	0
	0.005	0.3333	0.8333	0.476	0.0005	0.8	0.5714	0.667	0.0005	0.375	0.5	0.429	0.08	0	0	0
	0.008	0.2632	0.8333	0.4	0.0008	0.6667	0.5714	0.615	0.0008	0.4444	0.6667	0.533	0.1	0	0	0
	0.01	0.2083	0.8333	0.333	0.002	0.3846	0.7143	0.499	0.002	0.375	1	0.545	0.3	0	0	0
Our method	1.3	0.333	0.333	**0.333**	1.3	1	0.428	**0.6**	1.3	0.6	0.8571	**0.706**	1.3	0	0	**0**
	1.2	0.375	0.5	**0.428**	1.2	0.75	0.428	**0.55**	1.2	0.6	0.8571	**0.706**	1.2	0	0	**0**
	1.1	0.364	0.666	**0.470**	1.1	0.8	0.572	**0.667**	1.1	0.25	0.2	**0.222**	1.1	0	0	**0**
	0.9	0.364	1	**0.533**	1	0.8	0.572	**0.667**	1.0	0.125	0.2	**0.154**	1.0	0	0	**0**

TABLE II. R、P ANDF-1OF TWO METHODS COMPARISON IN REAL SPACE

method	threshold	600199			threshold	600559			threshold	600779			threshold	600600		
		P	R	F-1		P	R	F-1		P	R	F-1		P	R	F-1
Distance and density	0.002	0.6	0.5	0.545	0.0002	0	0	0	0.0002	0.5	0.1667	0.25	0.06	0	0	0
	0.005	0.3333	0.8333	0.476	0.0005	0.8	0.5714	0.667	0.0005	0.375	0.5	0.429	0.08	0	0	0
	0.008	0.2632	0.8333	0.4	0.0008	0.6667	0.5714	0.615	0.0008	0.4444	0.6667	0.533	0.1	0	0	0
	0.01	0.2083	0.8333	0.333	0.002	0.3846	0.7143	0.499	0.002	0.375	1	0.545	0.3	0	0	0
Our method	1.9	1	0.428	0.6	6	1	0.428	**0.6**	6	0.7	0.6	**0.65**	1.7	0	0	**0**
	1.6	0.666	0.428	**0.521**	3	0.75	0.428	**0.545**	5	0.583	0.6	**0.591**	1.6	0	0	**0**
	1.5	0.5	0.428	**0.461**	2	0.7	0.572	**0.635**	4	0.2632	0.8333	**0.4**	1.5	0	0	**0**
	1.4	0.4	0.428	**0.413**	1.5	0.7	0.572	**0.635**	3	0.2632	0.8333	**0.4**	1.4	0	0	**0**

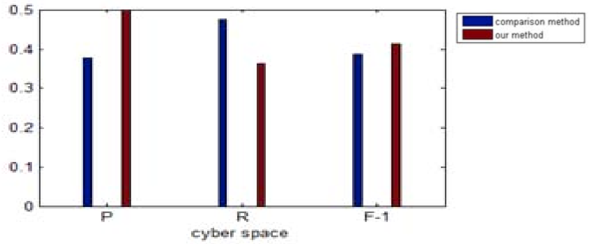


Figure 1. R、P andF-1of two methods in cyberspace

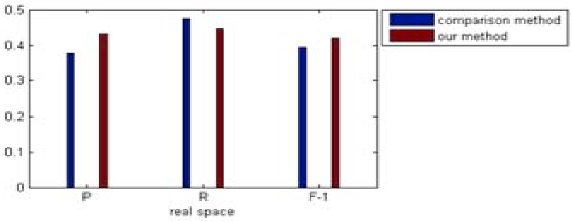


Figure 2. R、P andF-1of two methods in realspace

Fig 1 is the best average results of R、P and F-1 in cyber space of two methods, and Fig 2 is the best average results of R、P and F-1 in real space of two methods, in these two fighters we can see that the results of F-1 that are computed by our method is better.

V. SUMMARY

In this paper, the time series of anomaly detection and its algorithm is carried on the analysis and research in depth. we propose a method of time series outlier detection based in activity and density on the basis of time series outlier detection, the experimental results show that the proposed time series anomaly detection algorithms in this paper is more stable and accurate to discover the outlier sets. Experiments on dataset is further evidence in the stock market, we can effectively detect anomaly change of time series by this algorithm.

REFERENCES

[1] Hawkins D. Identification of outliers. London : Chapman and Hall,1980,304-315.

[2] Eskin E, Arnold A, Prerau M, Portnoy L and Stolfo S. A geometric frame-work for unsupervised anomaly detection. In Proceedings of Applications of Data Mining in Computer Security. 2002. Pages: 78-100.

[3] Z He, X Xu and S Deng. Discovering cluster-based local outliers [J]. Pattern Recognition Letters, 2003, 24 (9/10) :1641-1650.

[4] D Pokrajac, A LAzarevic, LJ Latechi. Incremental local outlier detection for data streams. IEEE Symposium on Computational Intelligence and Data Mining (CIDM). Apr, 2007.

[5] Sebyala A, Olukemi, T and Sacks L. Active platform security through intrusion detection using naive bayesian network for anomaly detection. The 2002 London Communications Symposium.2002.

[6] KNORR E.GtL Algorithms for mining distance-based outliers in large datasets[A]. In Pro VLDB'98[C]. NY,1998:69-85.

[7] Alex R,Alessandro L.Clustering by Fast Search and Find of Density Peaks. Science. 2014, 344(96191): 1492-1496.

[8] R. Xu, D. Wunsch 2nd, IEEE Trans. Neural Netw. 16 , 645–678(2005).

[9] L. Kaufman, P. J. Rousseeuw, Finding Groups in Data:An Introduction to Cluster Analysis , vol. 344(Wiley-Interscience, New York, 2009).

[10] A. K. Jain, Pattern Recognit. Lett. 31 , 651 –666 (2010).

[11] Mei-yu sun. Time series of anomaly detection method based on distance and density study [J]. Computer engineering and application. 2012 (20)

[12] Yong-Yeol Ahn, James P. Bagrow & Sune Lehmann.Link communities reveal multiscale complexity in networks[J]. nature 2010, 466(7307): 761-764.