International Journal of Multimedia and Ubiquitous Engineering
Vol. 10, No. 2 (2015), pp. 51-60
http://dx.doi.org/10.14257/ijmue.2015.10.2.05

Application Research of Linear Programming on Mining Outliers of Time Series

Yingying Min[1,2] and Changlin Ao [1,*]

[1]*School of computer and information engineering*
Harbin University of Commerce
[2]*College of Computer and Information Engineering*
Harbin University of Commerce
Harbin, China
myy80@126.com

Abstract

Based on data mining of outliers of time series, current studies have developed many methods, but there exist certain disadvantages for each method. This paper developed the method of mining outliers of time series based on linear programming model, verified high efficiency through an actual case and make up for the shortcoming of other methods.

Keywords: *Time series, Outliers, Data mining, Linear programming*

At present, in many scientific fields have accumulated large amounts of data. Data often contains some unknown information. In such circumstances gradually developed a new discipline that is data mining. The data mining has some data with time obviously. This is the time series data mining. In time series data mining work people often ignore the abnormal data. However, some abnormal data is able to respond to some unexpected situations, especially in the electronic commerce environment. The time series data mining technology can help the electronic commerce enterprise learn from data mining, potential rules. Then these findings shown by visualization method can help enterprises to deepen the understanding of knowledge acquisition, optimize enterprise decision management, customer relationship management, collaborative business management, marketing management, website maintenance management and risk control management and confirm the target market, to realize the personalized marketing, and obtain greater competitive advantage.

Time series outlier data mining has many method, such as: the statistical method, distance based method, density based method, biological method, based on model method etc. But each method all has different limitations. This paper, based on time series outlier data mining is a mining method of linear programming based on time series outlier data. And a good effect is obtained by the empirical analysis of the 2011 stock data. Hopefully, this method can increase the accuracy of time series the anomaly detection.

1. Research Status

Outlier detection has 20 years of history, and a good many methods has developed, such as statistical method: in the selection of a suitable distribution model. The model data selected is not consistency test to find out abnormal. The problem with this approach is that most of the

* Corresponding Author

ISSN: 1975-0080 IJMUE

tests are only suitable for the single attribute. The distribution of data is necessary before the use of statistical methods, which relates to the model parameters of the problem. It is difficult to determine the general.

Outlier detection method based on distance: the idea of this method is that if the distance from point P to the distance data set S is greater than a certain fixed value d, this point is determined to be an abnormal point. This method does not need to know the distribution model of data set. It is valid for arbitrary distribution model, but it can only detect the global outliers, not detect the local outliers.

Outlier detection method based on the density: the concept of density here is resembled to that of in physics, which refers to the number in the data in some area of existing data points.

Abnomaly detection algorithm based on distance calculation data set: the average distance of each point to its close point of this method, the average distance is smaller the points in the region is more, and the density is bigger. The algorithm detects the local outlier factor data centralized, local outlier factor is greater. The abnormal data can become large and the smaller the possibility.

Detection algorithm based on Clustering: the remaining after clustering data as abnormal data. Data set by clustering algorithm which has similar properties to the data together as a class. As a class and not aggregation or very small abnormal data. The algorithm is mainly used to cluster data set. The research direction is to improve the validity of clustering for outlier detection, optimization research and so on.

At present, the main detection methods of outliers. Every method is not comprehensive for outlier detection. So I want to find a more accurate detection method of abnormal.

2. To Establish the Linear Programming Time Series Outlier Data Mining Model based on

2.1The Model Building Process

According to the traditional time series outlier data mining method to establish a mining model of linear programming based on time series outlier data. We make the following assumptions:

Given a set of time series $X = (x_1, x_2, ..., x_n)$. Point $x_{t_i} = (v_{t_i}, t_1)$ Representation of time series at t_i Moments of the observation value y_{t_i}. Use $(N_1, N_2, ..., N_k)$ Indicates a point x_{t_i}'s k the neighbor point set. The observations recorded as collections is $(y_1, y_2, ..., y_n)$.The given threshold T, a_i is Coefficient. A phenomenon of programming model is as follows:

$$\min f(x) = \sum_{i=1}^{n} a_i x_i$$

$$s.t\begin{cases} d(x,y)=[\sum_{i=n}^{n}|x_i-y_i|]^{1/r} \\ |x_i-\bar{x}|>3\sigma \\ \dfrac{\sum_{i=1}^{k}x_i|y_i-y_{N_i}|}{\sum_{i=1}^{k}x_i}>T \end{cases} \quad (1)$$

r=1 and 2 this distance is the Euclidean distance and absolute distance.

σ is error. $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, $\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$.

Parameters using the 2010 Shenzhen stock data for model fitting do the following estimation:

Table 1. The Parameter Estimation Value Table

Model No.	Parameters 1	Parameters 2	Parameters 3	Correlation coefficient
1	0. 027001	0. 024666	0. 021582	0. 9763
2	0. 022002	0. 035787	0. 021472	0. 9768
3	0. 028001	0. 064784	0. 024666	0. 9862
4	0. 025001	0. 067692	0. 024615	0. 9722
5	0. 030001	0. 057732	0. 024742	0. 9786
6	0. 019001	0. 008145	0. 011213	0. 9798
7	0. 007001	−0. 035246	0. 014098	0. 9789
8	0. 044001	−0. 032128	0. 003012	0. 9788
9	0. 003001	−0. 050151	−0. 003009	0. 9743
10	0. 021001	−0. 019408	0. 014301	0. 9786
11	0. 011001	−0. 037412	0. 004044	0. 9823
12	0. 007001	−0. 035246	0. 014098	0. 9845
13	0. 003001	−0. 050151	−0. 003009	0. 9825
14	0. 027001	0. 024666	0. 021582	0. 9837
15	0. 024001	0. 031762	0. 021516	0. 9787
16	0. 005001	−0. 003015	0. 001005	0. 9781
17	0. 030001	0. 057732	0. 024742	0. 9791
18	0. 028001	0. 064748	0. 024666	0. 9759
19	0. 025001	0. 067692	0. 024615	0. 9783
20	0. 008001	0. 048487	0. 008064	0. 9792
21	0. 006791	0. 0432406	0. 007609	0. 9765
22	0. 007654	0. 0543208	0. 087901	0. 9775
23	0. 008765	0. 0234591	0. 012353	0. 9734
24	0. 002437	0. 0342187	0. 054128	0. 9765
25	0. 002345	0. 0542897	0. 056723	0. 9754

26	0. 005432	0. 0564328	0. 039871	0. 9768
27	0. 005648	0. 3420982	0. 020897	0. 9775
28	0. 004531	0. 2043512	0. 030987	0. 9723
29	0. 004609	0. 3210981	0. 080976	0. 9745
30	0. 004608	0. 3210881	0. 050987	0. 9781
31	0. 002786	0. 3241091	0. 005637	0. 9745
32	0. 030291	0. 0892012	0. 004762	0. 9762
33	0. 030965	0. 0891102	0. 029845	0. 9763
34	0. 030987	0. 0028761	0. 054309	0. 9781
35	0. 040598	0. 0067539	0. 0565349	0. 9678
36	0. 030456	0. 0045321	0. 0345211	0. 9782
37	0. 028765	0. 0049832	0. 0398767	0. 9756
38	0. 050621	0. 00243109	0. 0034219	0. 9765
39	0. 045312	0. 0045321	0. 0054209	0. 9735
40	0. 065431	0. 0052987	0. 0034521	0. 9762

From the above data we can see that fit best is the third group model. So the selected third groups of data as the parameter values of the model. This model was established to complete.

2.2. The Model Empirical

In order to validate the models to forecast the abnormal points of time series is accurate. We will use the 2011 data of Shenzhen stock.

Some experimental data are given in the following table:

Table 2. Stock Data Table

Serial number	The transaction date	The index name	The total rate of return	The overall level of index
1	20110301	Shenzhen A shares in circulation market value weighted market index	0. 008912	6558. 486
2	20110306	Shenzhen A shares in circulation market value weighted market index	0. 058023	6873. 354
3	20110311	Shenzhen A shares in circulation market value weighted market index	0. 005674	7026. 657
4	20110320	Shenzhen A shares in circulation market value weighted market index	0. 00534	7064. 324
5	20110401	Shenzhen A shares in circulation market value weighted market index	0. 05876	7502. 989
6	20110407	Shenzhen A shares in circulation market value weighted market index	0. 003384	7818. 115
7	20110413	Shenzhen A shares in circulation market value weighted market index	0. 041121	7555. 21
8	20110417	Shenzhen A shares in circulation market value weighted market index	-0. 03356	7564. 544
9	20110422	Shenzhen A shares in circulation market value weighted market index	-0. 00905	7568. 297
10	20110429	Shenzhen A shares in circulation market value weighted market index	0. 007912	7785. 513

11	20110503	Shenzhen A shares in circulation market value weighted market index	0.02765	7719.34
12	20110509	Shenzhen A shares in circulation market value weighted market index	0.000765	7952.56
13	20110517	Shenzhen A shares in circulation market value weighted market index	0.02178	7896.435
14	20110524	Shenzhen A shares in circulation market value weighted market index	-0.04532	7658.987
15	20110601	Shenzhen A shares in circulation market value weighted market index	-0.03454	7377.654
16	20110602	Shenzhen A shares in circulation market value weighted market index	-0.00975	7323.985
17	20110603	Shenzhen A shares in circulation market value weighted market index	-0.02747	7934.905
18	20110610	Shenzhen A shares in circulation market value weighted market index	0.006935	7787.075
19	20110615	Shenzhen A shares in circulation market value weighted market index	0.0069264	7908.028
20	20110618	Shenzhen A shares in circulation market value weighted market index	0.0038762	7783.098
21	20110620	Shenzhen A shares in circulation market value weighted market index	0.0029874	7782.039
22	20110625	Shenzhen A shares in circulation market value weighted market index	0.0039867	7781.329
23	20110628	Shenzhen A shares in circulation market value weighted market index	0.0043892	7791.423
24	20110630	Shenzhen A shares in circulation market value weighted market index	0.0039875	7618.291
25	20110702	Shenzhen A shares in circulation market value weighted market index	0.0028942	7683.372
26	20110705	Shenzhen A shares in circulation market value weighted market index	0.0028947	7638.293
27	20110708	Shenzhen A shares in circulation market value weighted market index	0.0038921	7632.291
28	20110710	Shenzhen A shares in circulation market value weighted market index	0.0029863	7702.902
29	20110715	Shenzhen A shares in circulation market value weighted market index	0.0019872	7701.896
30	20110719	Shenzhen A shares in circulation market value weighted market index	0.0298304	7698.329
31	20110721	Shenzhen A shares in circulation market value weighted market index	0.0109821	7649.312
32	20110725	Shenzhen A shares in circulation market value weighted market index	0.1028632	7639.209
33	20110729	Shenzhen A shares in circulation market value weighted market index	0.1098703	7689.306
34	20110730	Shenzhen A shares in circulation market value weighted market index	0.0209831	7791.783
35	20110805	Shenzhen A shares in circulation market value weighted market index	0.0090803	7782.307

In order to make the data more intuitive we draw the stock about graph:

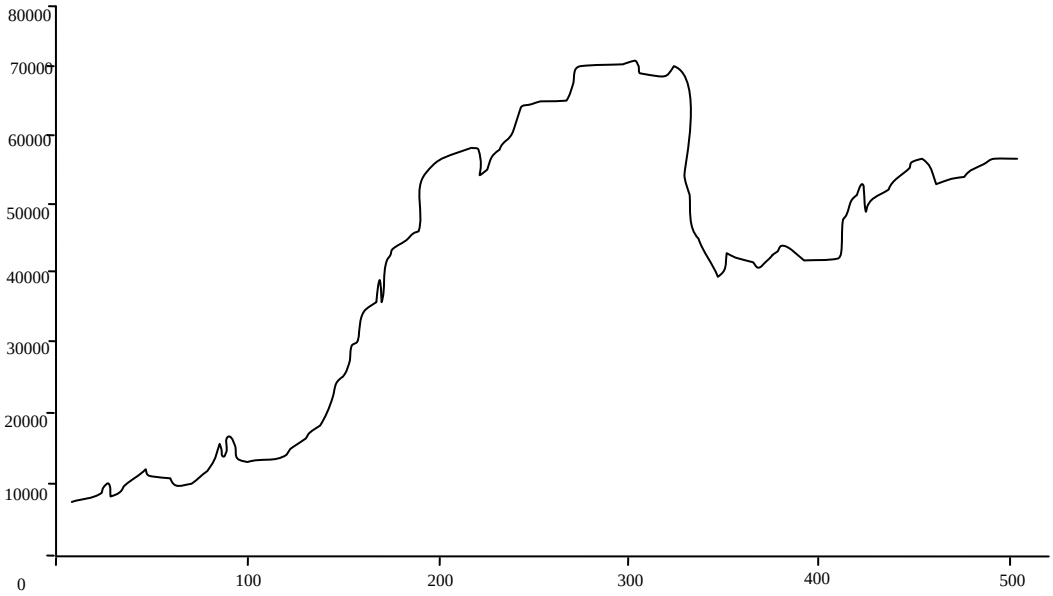

Figure 1. Stock Data Distribution Map

Detection of the use of these data for anomaly detection. The results are compared with the results based on the distance. The comparison results show outliers in linear programming model based on more detailed testing results.

The following is a more detailed result:

Table 3. Outlier Detection Table

Experiment	Algorithm	Parameter setting	Test results	Time consumption
Experiment one	Based on the distance	distance=0.25,n=30	Abnormal point number 53	40 Minutes
		distance=0.38,n=26	Abnormal point number 4	26 Minutes
		distance=0.35,n=20	Abnormal point number 45	37 Minutes
	Based on linear programming	n=15	Abnormal point number 9	5 Minutes
Experiment two	Based on the distance	distance=0.25,n=30	Abnormal point number 23	30 Minutes
		distance=0.35,n=20	Abnormal point number 17	36 Minutes
		distance=0.30,n=25	Abnormal point number 15	18 Minutes
	Based on linear programming	n=15	Abnormal point number 6	3 Minutes

See from the above comparison of the test results can be (1) detection time is shortened to a lot. Detection results are more precise. It indicates that the model has good effect. But the model also has the disadvantage of this model, for abnormal points of inspection are not obvious. The improvement on this issue model.

2.3. Model Improvement

Model (1) in the detection of outliers in the performance is very good. But the abnormal point's detection is not obvious. So to improve it. In the improvement of a small variables ε in the model. The variables need to be determined according to the actual conditions by our self.

This is the model into the following form:

$$\min f(x) = \sum_{i=1}^{n} a_i x_i$$

$$s.t \begin{cases} d(x,y) = [\sum_{i=n}^{n} |x_i - y_i|]^{1/r} \\ |x_i - \overline{x}| > 3\sigma \\ \dfrac{\sum_{i=1}^{k} (x_i |y_i - y_{N_i}| + \varepsilon)}{\sum_{i=1}^{k} x_i} > T \end{cases} \quad (2)$$

Below we verify this model, first used in model (1) in 2011 August to September stock data were detected. The detection results are as follows:

Table 4. Results a Model of Anomaly Detection

The transaction date	The total rate of return	The overall level of index	Abnormal / normal
20110801	0. 028102	4929. 495	Normal
20110804	0. 037432	5114. 004	Normal
20110809	0. 053064	5346. 78	Normal
20110811	0. 030807	15551. 7	Normal
20110819	0. 0000766	15674. 34	Abnormal
20110823	0. 015076	15624. 64	Abnormal
20110825	0. 026598	15987. 25	Abnormal
20110826	0. 005286	15684. 4	Abnormal
20110829	0. 029541	15696. 46	Normal
20110901	−0. 05058	15660. 14	Normal
20110904	0. 002123	15703. 66	Normal
20110906	−0. 00638	15374. 98	Normal
20110910	0. 00876	15447. 26	Normal
20110916	−0. 05763	15387. 79	Normal
20110918	0. 013433	5549. 578	Normal
20110920	−0. 01147	5523. 834	Abnormal
20110923	0. 025054	5398. 768	Abnormal
20110926	0. 000784	5467. 235	Abnormal
20110930	−0. 02336	5558. 967	Normal

Secondly by using the model (2) in 2011 August to September stock data were detected, the detection results are as follows:

Table 5. Results a Model of Anomaly Detection

The transaction date	The total rate of return	The overall level of index	Abnormal / normal
20110801	0. 028102	4929. 495	normal
20110804	0. 037432	5114. 004	normal
20110809	0. 053064	5346. 78	Abnormal
20110811	0. 030807	15551. 7	normal
20110819	0. 0000766	15674. 34	Abnormal
20110823	0. 015076	15624. 64	Abnormal
20110825	0. 026598	15987. 25	Abnormal
20110826	0. 005286	15684. 4	Abnormal
20110829	0. 029541	15696. 46	normal
20110901	−0. 05058	15660. 14	normal
20110904	0. 002123	15703. 66	Abnormal
20110906	−0. 00638	15374. 98	Abnormal
20110910	0. 00876	15447. 26	normal
20110916	−0. 05763	15387. 79	normal
20110918	0. 013433	5549. 578	normal
20110920	−0. 01147	5523. 834	Abnormal
20110923	0. 025054	5398. 768	Abnormal
20110926	0. 000784	5467. 235	Abnormal
20110930	−0. 02336	5558. 967	normal

Controls from Table 5 and table 4 can be seen: 4 showed a normal point in the Table. In Table 5 are considered to be outliers. Visible model (2) detection more accurate.

3. Summary

This paper has given two linear programming based time series outlier detection model. Model test results achieve the expected aim. However, to determine the model one k and model two ε values are difficult. This in future research can be studied further.

Reference

[1] M. S. Chen, J. Han and P. S. Yu, “Data mining, an overview from database perspective”, IEEE Transaction on Knowledge and Data Engineering, vol. 8, no. 6, **(1996)**, pp. 866-883
[2] J. L. Hua, M. Kamber, “Data Mining, Concepts and Techiiiques”, Morgan Kaufmann Publisher, Inc, **(2001)**, pp. 384-396.
[3] E. M. Knorr and T. N. Raymond, “Algorithms for Mining Distance-Based Outliers in Large Datasets”, Proceedings of the 24th VLDB Conference, New York, USA, **(1998)**, pp. 392-403.
[4] M. M. Breuning, H. P. Kriegel, T. N. Raymond and J. Sander, “LOF, Identifying Density Based Local Outliers”, ACM SIGMOD 2000 Int, Conf, On Management of Data, Dallas, TX, **(2003)**, pp. 93-104.
[5] E. Keogh, S. Lonardi and B. Yuan-Chi, “Finding Surpring Patterns in a Time Series Database in Linear Time and space”, SIGKDD , 02.July, Edmonton, Alberta, Canada, **(2002)**, pp. 23-26.
[6] X. Kui and J. Jingping, Pattern recognition method of symbolic time series of J and artificial intelligence, vol. 20, no. 2, **(2007)**, pp.154-161.

[7] L. Aiguo, Q. Zheng and H. Shengping, "Extracting similar patterns in time series data", Journal of Xi'an Jiao Tong University, 36(12), **(2002),** pp. 275-1278.
[8] W. Yongjun, "The mining of master's degree thesis, Hefei", Anhui University, (**2010)**, pp. 9-22.
[9] Linsen, "Anomaly detection in time series research and application Master thesis, Nanjing", Hohai University, **(2008)**, pp. 10-32.
[10] K. Yamanishi and J. Takeuchi, "A unifying Framework for Detecting Outliters and Change Points From Non-stationary Time Series Data SIGKDD'02", **(2002)** July, pp. 676-681.

Authors

Yingying Min, Associate Professor, teacher of School of the Computer and Information Engineering, Harbin University of Commerce. Her main research fields are Data mining and mathematical modeling.

Changlin Ao, doctor professor, his research fields are Evaluation theory and method, system reliability modeling and simulation.