Acta Technica 62 (2017), No. 7A, 1013–1022 © 2017 Institute of Thermomechanics CAS, v.v.i.

Analysis of ATM transaction status characteristics and anomaly detection

Duzhen Zhang[1], Liuqian Sun[2], Xiaotong Ye[3], Haochi Zhang[4], Yinggang Li[5]

Abstract. After detailed analysis of topic and data, a set of early warning system based on analysis of transaction status characteristics and anomaly detection has been designed. Discrimination of characteristics for anomaly situation has been verified. The feasibility and possibility of further improvement of model have been evaluated comprehensively. Based on the extracted characteristic parameters, mechanism model has been constructed and the rules that can be discovered have been recorded as rule bank, with which the normal or abnormal status of ATM transaction has been described. The first perspective starts from single index, adopts wavelet analysis to make anomaly detection for transaction volume, success rate and response time data and combines the characteristics of each index to propose two anomaly detection programs based on wavelet analysis; the second perspective starts from joint index and adopts isolation-forest algorithm to make joint detection for transaction volume, success rate and response time data. Through data backtesting and other methods, the accuracy and timeliness of detection program have been tested. The results have shown that this program can detect the collected abnormal situations in rule bank timely and accurately; at the same time, it can also discover the unknown abnormal situation timely and the false alarm rate is low.

Key words. Correlation analysis, Variance analysis, Anomaly detection, Wavelet analysis, Isolation-forest algorithm.

1. Introduction

The data in this paper is time series data; at the same time, there are three variables, including transaction volume, success rate and response time. Its distribution

[1] College of Qilu Software, ShandongUniversity, Jinan City, China

[2] Department of Information Engineering, Shandong university of science and technology University, Taian City, China

[3] College of International Education, Henan University, Kaifeng City, China

[4] Department of Electronic Engineering, Institute of Information Technology, Guilin University of Electric Technology, Guilin City, China

[5] School of Shandong University of Science and Techology, Taian City, China

has regularity in time, which the time can be divided based on different cycles of minute, hour, day, week and month etc. For multivariate variable data, it firstly divides it to three one variable time series, explores the distribution and regularity of different characteristics in time cycle; secondly, explore the relationship between variables at time period after division. The initially selected feature characteristics here include:

(1) Describing concentration degree: mean and median;

(2) Describing degree of dispersion: standard deviation;

(3) Describing deviation degree: Z-score and fractile;

(4) Describing multivariate correlation degree: covariance and correlation coefficient;

(5)Through analyzing the descriptions of above characteristic parameters for variables at different time periods, finally it will extract the requested information which can provide anomaly detection at maximum as characteristics parameter.

In a further way, apply the formerly extracted characteristic parameters to depict the deviation degree of normal situation. The system monitoring staffs will set the threshold value, and then carry out the alarm for abnormal situations. The seriousness degree of abnormal situation can be measured from two aspects, including:

(1) Deviation between measurement value of every minute and expected value

(2) Duration time and happening times for deviation

If increase the data that can be collected, for ATM transaction system, in addition to the materials in attachment as well as the data attained from processing of materials, it can also increase the statistics for transaction types and transaction scale.

2. Basic hypotheses

Table 1. Main symbols, marks and their explanations

Mean value	μ
Median	Med
Standard deviation	σ
Z-score	$z-score$
Fractile	α
Covariance	Cov
Correlation coefficient	$Corr$

1) Exclude the influence of human operation error and failure happened to the front end of ATM application system on data offered in the topic.

2) Assume the four possible situations offered in the topic as the main influencing factors.

3) Assume that only consider about the data offered in the topic is 2017, working days and holidays are divided based on 2017.

4) Assume various kinds of possible anomalous independent distribution and

believe that the probability for the abnormality is very small.

5) Assume the added data can be collected effectively and scientifically.

3. Extraction of characteristics parameters

3.1. Analysis of transaction volume

Data analysis in one day

Transaction volumes in different times of the day have obvious fluctuations, which can be divided into transaction peak period and transaction trough period. The following has selected data from Mar. 8^{th} to 9^{th} for display.

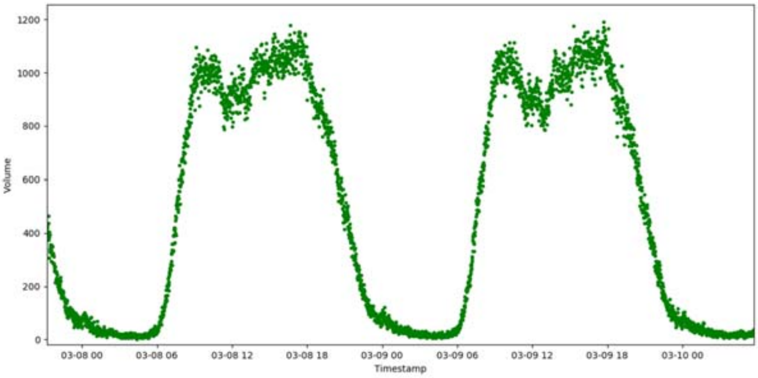

Fig. 1. Transaction Volumes on Mar. 8^{th} and 9^{th}

The daily transaction volume is similar to the figure. It is trough period from midnight to early morning. From 6:00am to 9:00am, it is rising interval; from 9:00am to 18:00pm, it is peak period, (in which between 12:00 and 14:00, there is small trough period within peak period), from 18:00 to 22:00, it is lower interval. The above conclusions can be explained reasonably through regularities of human economic activities. Their specific regularities will be taken as important standards for judging abnormal situations.

Draw the data into histogram and fit its probability distribution function. Adopt Kernel-Density-Estimation, which is a non-parameter method for estimating probability density function. $(x_1, x_2, ...x_n)$ is n sample points of independent and identical distribution F. Set its probability density function is f, kernel density estimation is as following:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x_- x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})\,. \tag{1}$$

$K(.)$ is kernel function (non-negative and integral is 1, meet the property of probability density and mean value is 0), $h > 0$ is a smoothing parameter and called as bandwidth. $K_h(x) = x/hK(x/h)$ is scaled-kernel function.

Firstly, take minute as the unit to divide data set, select transaction volumes at

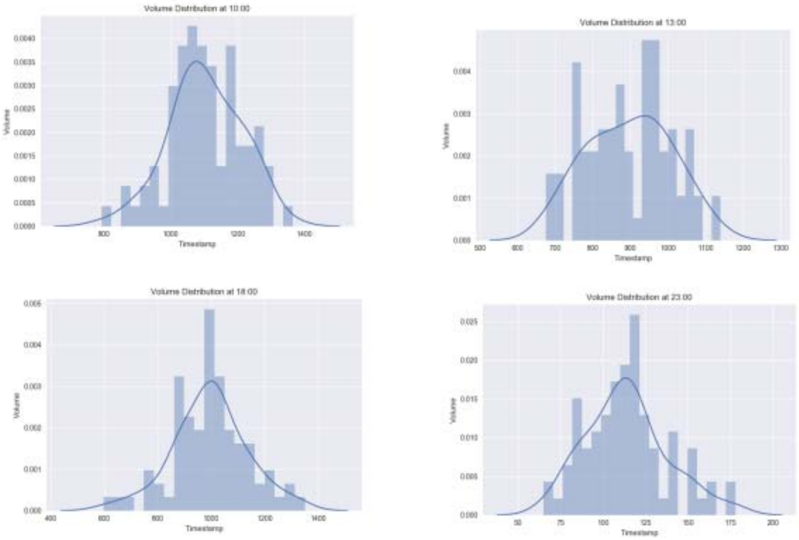

Fig. 2. Transaction volumes at 10:00, 13:00, 18:00 and 23:00

10:00, 13:00, 18:00 and 23:00 to draw histogram, adopt formula to fit the optimal probability distribution (kernel function), discover that it meets normal distribution generally. In the following, try to use bigger time interval to divide data set.

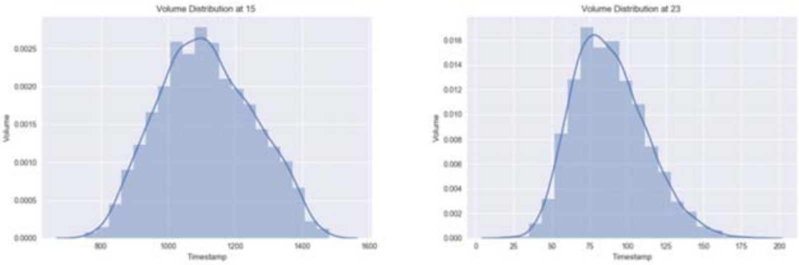

Fig. 3. Transaction volumes at 15pm and 23pm

Take hour as the unit to divide data set, it can be seen that its kernel function fitting is good and it nears to normal distribution at the same time. However, the accuracy is not high by taking hour as the unit and it will lose some information; at the same time, it can only be used in the situation where changes of variables are not big for time.

To explain the universality of this regularity, take 10min between 1min and 1h as the unit to divide data set, it can also be found that it is very close to normal distribution.

At this moment, it can draw a conclusion that transaction volume data at the same time of different days meets normal distribution and the μ and σ of normal function can be estimated based on sample.

However, there is another problem whether the mean value and variance at this

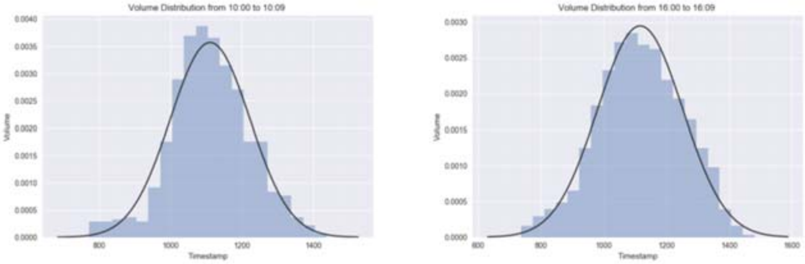

Fig. 4. Transaction volumes distribution at 10:00-10:09, 16:00-16:09

period need to be calculated with all the known data. If there is trend in the data, using more data will not bring better effect, because the data which is far away from current time can't well represent current trend. Therefore, adopt HP filtering method to decompose transaction volume into trend part and cycle part, and then continue to decompose trend part and cycle part.

HP filtering method is a decomposition method of time series in state space, which equals to minimize the linear filtering of fluctuation variance and is usually used to decompose the trend factors and cycle fluctuation factors in time series.

For time series $y_t, t = 1, 2, ..., T$, HP filtering decomposition meets the trend component T_{y_t} in following formula:

$$\min_{T_t}\{\sum_{t=1}^{T}(y_t - T_{y_t})^2 + \lambda\sum_{t=2}^{T}[(T y_{t+1} - T_{y_t}) - (T_{y_t} - T_{y_{t-1}})]\}^2 . \tag{2}$$

In which λ is the weight of various fluctuation degrees in the trend. For annual data, it usually selects experience value $\lambda = 100$. HP filtering can attain cycle component $C_{y_t} = y_t - T_{y_t}$ after decomposition $C_{y_t} = y_t - T_{y_t}$.

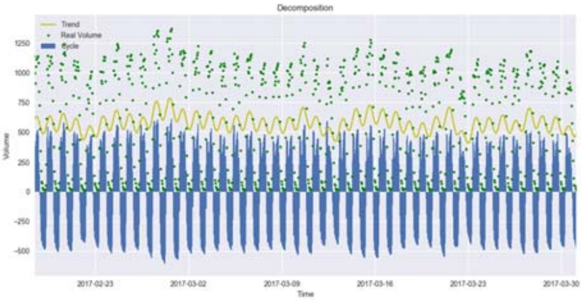

Fig. 5. Transaction volume is decomposed into trend part and cycle part

It can be seen that in the trend, there is basically no additional information, while the originally decomposed trend fluctuates around one line and it indicates that there is no obvious trend. Therefore, at the same time of different dates (taking 1min as the time interval), it can use the data of three months to calculate μ and σ and take them as the characteristic parameters of transaction volume. Z-Score:

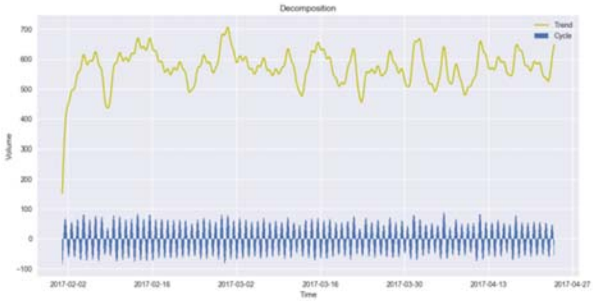

Fig. 6. Re-decomposition for trend part

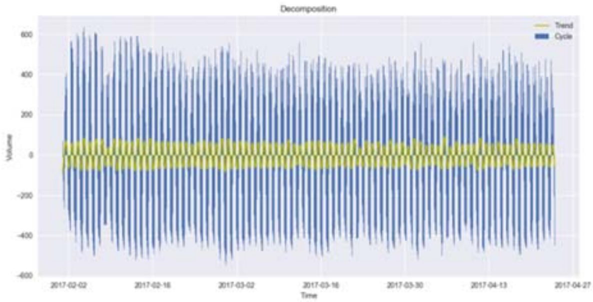

Fig. 7. Re-decomposition for cycle part

standard score is also called as Z score or Z value. As the statistical measurement, it examines the relationship between one score and the average value of a set of scores. Here, it also takes Z score as the characteristic parameter of transaction volume. It can be both positive and negative. The calculation formula is as following:

$$z = \frac{x - \mu}{\sigma}\,. \tag{3}$$

x is standardized random variable, μ is mean value of sample, σ is standard deviation of sample.

3.2. Success rate analysis

Figure 10 and figure 11 are success rate scatter diagrams from January to April. It can be seen obviously that the success rate generally distributes between 0.8 and 1.0. Abnormal situations happen in late March and later April obviously.

After detailed observation of daily success rate, it is found out that it presents a changing trend on daily cycle basis. It is relatively stable between 8:00 and 21:00, the mean value and variance distribute along straight line approximately. In comparison, the distributions of success rates from midnight to early morning are different, which can be divided into two categories, including success rate is 1 and success rate is not 1. The latter is with higher degree of dispersion and presents the shape of trailing.

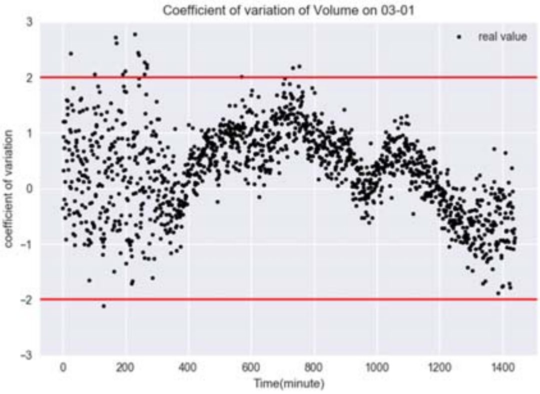

Fig. 8. Z score of transaction volume per minute on Mar. 1^{st}

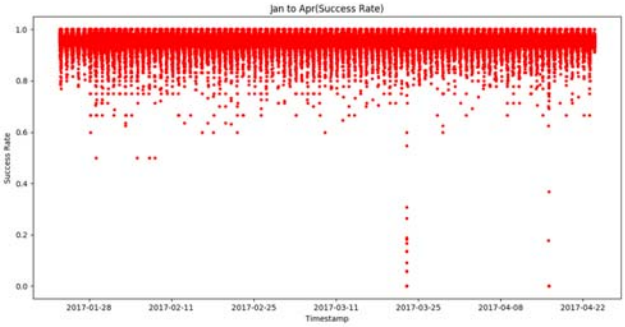

Fig. 9. Success rates from January to April

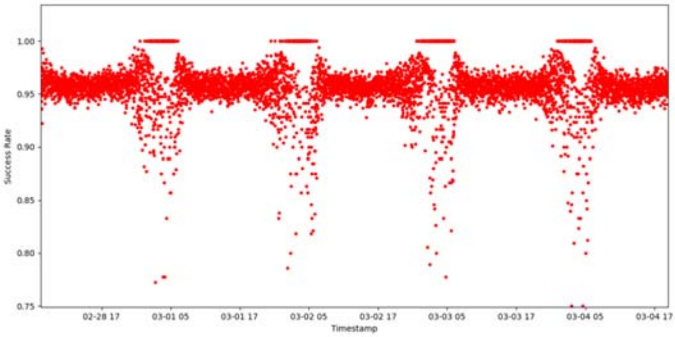

Fig. 10. Success rate scatter diagram of four continuous days

At the same time, through drawing fold line graph, it is found out the situation with success rate at 1 does not happen continuously; therefore, the success rate will appear in these two situations based on certain ratios.

4. Summary

Exploration methods are comprehensive and diversified and it tries to discover the connections of characteristics from different perspectives. Basic method of data science has been applied to explore data from different perspectives comprehensively, including the regularity of data changing with time, connections between characteristics. At the same time, it is not satisfied with qualitative analysis, adopts correlation measurement, stability test of time series and function relation fitting between variables, reveals the connections between characteristics specifically in a quantitative way and adopts knowledge reasoning to judge the corresponding reasons for some special phenomena. Wavelet analysis is a useful tool for anomaly detection; its application is very innovative and has been verified with prominent detection effect by the industry. As the wavelet analysis principle is difficult to understand study and practice from the simplest knowledge and finally attain good result to make up for the shortages of statistical model method. The application of machine learning method brings a brand new thinking. Adopt the principle of clustering and classification first and training later to attain the model and then compare with the former results.

References

[1] S. L. Fernandes, V. P. Gurupur, N. R. Sunder, N. Arunkumar, S. Kadry: *A novel nonintrusive decision support approach for heart rate measurement* (2017) Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2017.07.002

[2] N. Arunkumar, K. Ramkumar, V. Venkatraman, E. Abdulhay, S. L. Fernandes, S. Kadry, S. Segal: *Classification of focal and non focal EEG using entropies.* Pattern Recognition Letters, *94* (2017), 112–117

[3] W. S. Pan, S. Z. Chen, Z. Y. Feng.: *Investigating the Collaborative Intention and Semantic Structure among Co-occurring Tags using Graph Theory.* International Enterprise Distributed Object Computing Conference, IEEE, Beijing, (2012), 190–195.

[4] Y. Y. Zhang, Q. Li, W. J. Welsh, P. V. Moghe, and K. E. Uhrich: *Micellar and Structural Stability of Nanoscale Amphiphilic Polymers: Implications for Anti-atherosclerotic Bioactivity*, Biomaterials, *84* (2016), 230–240.

[5] L. R. Stephygraph, N. Arunkumar, V. Venkatraman: *Wireless mobile robot control through human machine interface using brain signals*, 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2015 - Proceedings, (2015), art. No. 7225484, 596–603.

[6] N. Arunkumar, V. S. Balaji, S. Ramesh, S. Natarajan, V. R. Likhita, S. Sundari: *Automatic detection of epileptic seizures using independent component analysis algorithm*, IEEE-International Conference on Advances in Engineering, Science and Management, ICAESM-2012, (2012), art. No. 6215903, 542–544.

[7] Y. Du, Y. Z. Chen, Y. Y. Zhuang, C. Zhu, F. J. Tang, J. Huang: *Probing Nanostrain via a Mechanically Designed Optical Fiber Interferometer.* IEEE Photonics Technology Letters, *29* (2017), 1348–1351.

[8] W. S. Pan, S. Z. Chen, Z. Y. Feng: *Automatic Clustering of Social Tag using Community Detection.* Applied Mathematics & Information Sciences, *7* (2013), 2, 675–681.

[9] Y. Y. Zhang, E. Mintzer, and K. E. Uhrich: *Synthesis and Characterization of PEGylated Bolaamphiphiles with Enhanced Retention in Liposomes*, Journal of Colloid and Interface Science, *482* (2016), 19–26.

[10] N. Arunkumar, K. M. Mohamed Sirajudeen: *Approximate Entropy based ayurvedic pulse diagnosis for diabetics - A case study*, TISC 2011 - Proceedings of the 3rd International Conference on Trendz in Information Sciences and Computing, (2011), art. No. 6169099, 133–135.

Received May 7, 2017