Outlier detection method based on standard scores

Ya-Nan Wang[1,2,3,4,a, †]

[1]*Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing, 100097, China*
[2]*National Engineering Research Center for Information Technology in Agriculture, Beijing 100097, China*
[3]*Key Laboratory of Agri-informatics, Ministry of Agriculture, Beijing 100097, China*
[4]*Beijing Engineering Research Center of Agricultural Internet on Things, Beijing 100097, China*
awangyn@nercita.org.cn
[†] *Ya-Nan WANG*

Agricultural monitoring data is the basis of environmental early warning, however, abnormal data is inevitable in the monitoring process. Aiming at the problem that the data of wireless sensor network is abnormal, an outlier detection method is proposed, in order to achieve the purpose of accurate calibration data. The method is optimized by using the theory of MovingRange and standard scores, which greatly reduces the time complexity and space complexity of the algorithm. Real application results show that the abnormal data detected by this method are basically consistent with the actual situation and the accuracy is over 90%. The experimental results suggest that this method can effectively complete the detection of abnormal data which is mass and the deviation of the characteristics is not obvious.

Keywords: Wireless sensor networks； Abnormal; Standard score; Moving range.

1. Introduction

At present, the Iot technology is rapidly development and has been widely used in the actual production [1]. Internet of things technology are more relying on the sensor technology, upload the information automatically through the sensor has great randomness and which leads to abnormal data [2].These abnormal data can lead to the abnormal operation of the Internet of things, and even provide experts and farmers with wrong decision-making basis. Therefore, the abnormal data detection has become a problem that cannot be ignored [3-4].

In order to avoid the harm caused by the above problems, the experts and scholars have conducted a lot of research on abnormal data detection. At present,

Copyright © 2017, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

the methods of detecting abnormal data are mainly as follows [5]: Statistical methods based on statistics, statistical methods based on distance, and statistical methods based on clustering.

However, in the actual production process, the response speed of the system seriously affects the user experience and system performance, which puts forward higher requirements for the time complexity and space complexity of the abnormal data detection methods. The method can detect more than 90% of the abnormal data, and has a very low time complexity and space complexity which achieve the purpose of abnormal data detection, and canensure the operating efficiency.

2. Abnormal Data Detection Model

2.1 Abnormal data detection principle

The standard score is also called the Z score which can reflect the relative position of the detection points in the overall distribution. Normalized values reflect the relative distance between the first I variable value and the average value of the sample [6]. Let z for the relative standard distance of the mean value and the detection point, x as the detection point, μ as the population mean[7], σ as the standard deviation, then z scores can be expressed as

$$z = (x - \mu)/\sigma \tag{1}$$

Among the formula, $\sigma \neq 0$ and$\mu = E(x)$, so the arithmetic mean can be expressed as

$$M = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

The variance can be expressed as$\sigma^2 = Var(x)$, and so the standard deviation can be expressed as

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{3}$$

If the standard value is 1, it represents that the numerical value of the i-th sample point is equal to that of the standard deviation. If the normalized value is 2, it represents that the numerical value of the i-th sample point is twice as large as the standard deviation value, that is, the relative distance between the i-th sample and that of the standard deviation of the sample is 2.According to the

"Pauta criterion", the abnormal data is that the distance between the mean and the averageis three times bigger thanthe standard deviation of the sample.The height anomaly data is that the distance between the average value and the standard deviation of the sample is greater than three times.However,in practical application, whether the data should be removed or not, it depands on the practical circumstances [8-9].

2.2 Abnormal data detection algorithm based on moving range

In the calculation formula of the standard score, the probability distribution of the sample is used, But when the sample data amount is larger, when computing the probability distribution of variance will consume a large amount of resources, which has a greater impact on the monitoring data storage speed and system response time. At the same time, the data collected every day is more than 30 thousand, the data is massive, and so the efficiency of the algorithm must be taken into account when we do abnormal data detection. In order to solve the problem of system efficiency, a method for detecting abnormal data is proposed.

The algorithm uses the historical data of the same period of time, due to the seasonal, geographical and other factors, the numerical value of the detection index is greatly influenced, and therefore, the historical data of the last month in the same area were selected as the sample data. When there is no abnormal data in the sample data, and the volatility of the data is not large, the discrete degree of the sample can be replaced by the moving range.

Moving Range is the difference between the maximum and the minimum in the sample, when an additional data point is obtained, we add the data to the sample and at the same time we remove the oldest point of samples and according to such theory, a new sample set is got.Then we calculated the range based on the new sample [10-11].

According to the above principle, the calculation formula of the standard fraction can be optimized as:

$$z = (x - \mu)/R \tag{4}$$

Among the formula, let R for the range of the new sample, Max as the Maximum value of the new sample and Min as the minimum value of the new sample, then the Moving range can be expressed as

$$R = Max - Min \tag{5}$$

Thealgorithm for detecting outliers is summarized in Algorithm 1, see Fig. 1 for details.

Algorithm 1

Input: True data x

1、get new sample.

2、Compute $R = Max - Min$ by equation (5).

3、Compute $\mu = E(x)$.

4、Compute $z = (x - \mu)/R$ by equation (4).

Output: Estimated result z

Fig. 1. Algorithm of data rectification.

3. Experiments

3.1 Experimental data

In order to verify the validity of the proposed anomaly detection method, the simulation test is carried out based on the real data. The real air temperature data forexperimental is fromBeijing Xiaotangshan modern agriculture demonstration base. The air temperature monitoring data of the monitoring station was used as test data that from June 30, 2016 02:00 to July 30, 2016 02:00 and these numbers are up to 43200 records. And then based on the sample data, 100 records are ready to be detected, and 23 abnormal data is added into the sample waiting for detection. The air temperature data is shown in table 1 below.

Table 1. Experimental data

No.	True data	No.	True data	No.	True data
1	28.6	**11**	27.8	**21**	29
2	28.5	**12**	28	**22**	28.9
3	28.5	**13**	29.8	**23**	28.7
4	28.4	**14**	29.6	**24**	28.6
5	28.2	**15**	29.1	**25**	28.21
6	28	**16**	28.9	**26**	29.21
7	27.9	**17**	28.6	**27**	28.50
8	27.8	**18**	29.3	**28**	28.74
9	27.8	**19**	29.2	**29**	27.77
10	27.7	**20**	29.1	**30**	27.46
			……		

3.2 Experimental result

According to the method of “Pauta criterion”, the exception data is the value that the distance between the mean and the average is more than two times of the

standard deviation. These values are called highly abnormal and should be directly removed. We compare the monitoring results with the abnormal data for intervention and the result is shown in Table 2 below.

Table 2. Results comparison

sample records	detected records	abnormal records	monitoring results	Detection rate
43200	100	23	21	91.3%

Through the data shown in the Table2, we can find that the abnormal data detected by this method are basically consistent with the actual situation and the accuracy is over 90%.Meanwhile, the complexityof the method proposed above is O(1) which greatly reduces the time complexity and space complexity of the algorithm. The test strategy has been successful applied to practical system of outlier detection algorithm, the detection strategy for dealing with data quality has been improved significantly, and the method can be directly carries on the statistical analysis.

4. Summary

Based on the research of abnormal data detection theory and combined with the actual situation, ananomaly detection method is proposed that combined the standard score test method and the moving range theory.The test strategy has been successful applied to practical system of outlier detection algorithm and the abnormal data detected by this method are basically consistent with the actual situation and the accuracy is over 90%.

Acknowledgement

This work was supported by Special Fund for Agro-scientific Research in the Public Interest (201303107).

References

[1]. Xing-Huai ZHAGN, Realization of the Hospital Database Abnom1al Automatic Monitoring Based on the SMS MODEM,J, China Digital Medicine. 7(2012): 112-114.
[2]. Hua HUANG, Data Anomaly Detection Method of Sensor Nodes in Internet of Things, J, Computer Simulation.29 (2012):159-162.
[3]. Hong LIU, Design of anomaly detection model of database, J, Economic vision. 29(2012):159-162.

[4]. Demyana, Nathan, The Relationship Among E-service Quality Dimensions, Overall Internet Banking Service Quality, and Customer Satisfaction in the USA, J, Journal of Modern Accounting and Auditing, 4(2014):479-493.
[5]. Ying LEI, Abnormal data detection in agriculture search engine, D, University of Science and Technology of China, 2010.
[6]. Xiang-Fen KONG, Zhen HE, Jian-GuoCHE, A Comparativen Study on Con¯dence Interval of Cp in Terms of Standard Deviation Estimated by Different Methods, J, Chinese Journal of Applied Probability and Statistics.25(2009):164-170.
[7]. An-Yuan ZHU, Error analysis and comparison of different methods of estimating total standard deviation, J, china market. 26(2013):23-33.
[8]. Ai-Wu ZHOU, Xian-Hui WANG, Hui-Ting LIU, Vocabulary Correlation Text Clustering Based on How Ne, J, MICROELECTRONICS & COMPUTER, 04(2015):90-93.
[9]. Jin-Qi SU, Hui-Feng XUE, Hai-Liang ZHAN,K-means Initial Clustering Center Optimal Algorithm Based on Partitioning, J, Microellectronics & Computer. 26(2009):8-11.
[10]. Qi-Fa XU, Jin-Xiu ZHANG, Cui-Xia JIANG, Evaluating Multiperiod VaR via Nonlinear Quantile Regression Model, J, Chinese Journal of Management Scienc. 23(2015):56-65.
[11]. Bian-Xia SUN, Ming-Jin WANG, A New Class GARCH Model based on Price Range, J, Journal of Applied Statistics and Management. 32(2013):259-267.