International Journal of Control and Automation
Vol. 11, No. 1 (2018), pp.91-98
http//dx.doi.org/10.14257/ijca.2018.11.1.08

Time Series Abnormal Data Detection for Smart Factory

Won-chang Lee[1], Jae-Han Cho[2] and LeeSub Lee[3]

[1,2,3]Kumoh National Institute of Technology
61, YangHo-Ro, Gumi, South Korea
{ [1]lwczzang, [2]jaehanfs, [3]eesub }@kumoh.ac.kr

Abstract

Recently, new facilities are being built as smart factories and the transition from existing facilities to smart factories has been increasing. Therefore, demands for processing data generated from a large number of sensors are rapidly increasing In the smart factory system, existing error processing is concentrated on simple methods such as out of range errors. However, there are various types of errors. Accurate detection and handling of errors are important parts of the productivity and quality of manufacturing.

This research proposes a real–time analysis system based on Complex Event Processing(CEP) which detects abnormal data in terms of time-series using Least Square Method(LSM). The proposed method will provide high performance real time detection of error data and the way of figure out the window size which generates optimum error the detection rates.

Keywords*: Big data, Least Square Method, Error detection, Apache Storm, Mutation Testing*

1. Introduction

Nowadays, the smart factory is a hot issue. New and existing facilities are also increasingly being converted to smart factories. Since Hadoop is a framework of using total collection analysis, Big data processing using a system such as Hadoop is not suitable to the smart factory that requires real time processing [1]. Therefore, it is necessary to develop a Complex Event Processing (CEP) [2] based real-time analysis system for increasing data processing requirements. CEP is a recent emerging technology, mainly developed by Apache and Microsoft. This technique is suitable for real-time big data processing with a structure that is stored after analyzing unlike the existing big data framework.

Huge number of data is generated from multiple sensors in smart factory environment. Therefore, it is necessary to process a large amount of data in real time. As shown in the Figure 1, since Hadoop stores the received data in the database and then processes the data, it is not suitable for real-time processing because it is stored on disks. As shown in Figure 2, on the other hand, the CEP handles sensor data directly from memory without storing it on disk.

Received (October 15, 2017), Review Result (December 22, 2017), Accepted (January 12, 2018)

ISSN: 2005-4297 IJCA

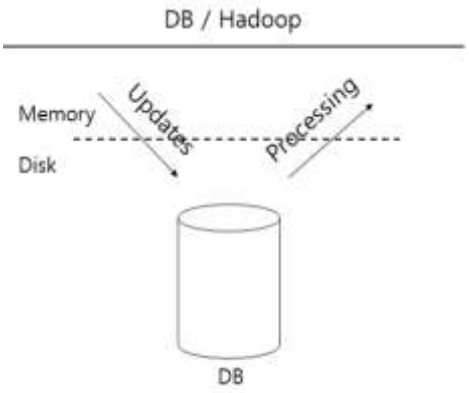

Figure 1. Using Hadoop in Data Analysis

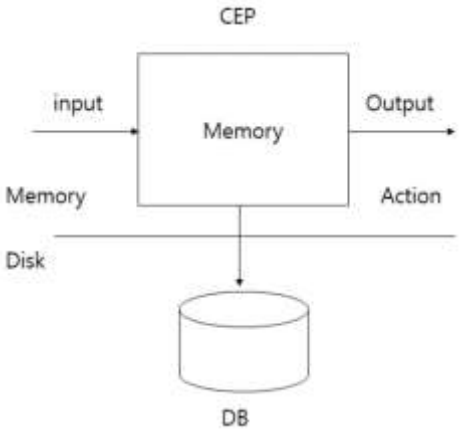

Figure 2. Using CEP in Data Analysis

In the smart factory system, existing error processing is concentrated only on out-of-range error detection [3]. However, there are various kinds of errors. Accurate detection and handling of various errors is also an important part of the efficiency for manufacturing. This study suggests the utilizing LSM (Least Square Method) to detect abnormal data through time series.

2. Abnormal Data Detection with CEP for Smart Factory

The most popular CEP engines are Microsoft StreamInsight, Apache Spark, and Apache Storm. Microsoft StreamInsight has a problem that is dependent on Microsoft's environment [4]. Apache Spark uses a batch-oriented approach with Hadoop [5]. Apache Storm, an open-source software produced by Twitter, is a technology that allows large-scale data to be analyzed in real time. Hadoop is a large-scale distributed processing system specialized for batch analysis, Storm is a distributed processing system specialized for real-time analysis. That is why we applied the Apache Storm [6].

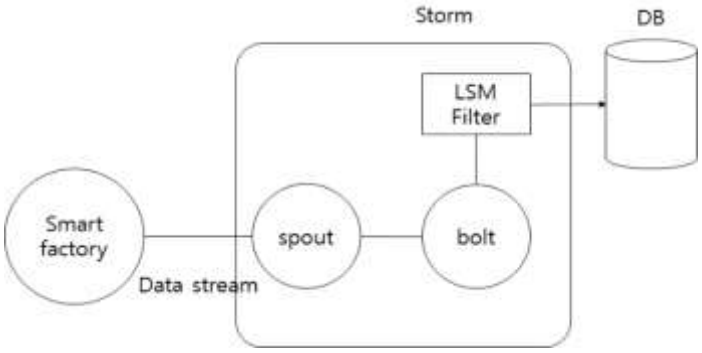

Figure 3. Structure of Apache Storm Based CEP

Figure 3 shows the structure of the CEP based on Apache Storm. The data generated by the sensors of the smart factories are transferred to the spout module where they are converted into tag value format data [7]. These data are consumed by the bolt and filtered by the filter attached to the bolt. This filter is made by applying LSM [8][9]. Consequently, the filtered sensor data is stored in the database and the subsequent processing proceeds.

Figure 4 shows the internal structure of the CEP system and its interfaces with external systems. When the system receives sensor data as an event from sensors of smart factory, the system extracts abnormal data and transfers it to external event handlers. The sensor data is stored in the event list in a tuple format (tag, value) by the event adapter implemented in the form of Spout. The filter is implemented in the form of Bolt, which provides various types of abnormal data extraction techniques such as LSM filters, boundary value filters, and so on. With this architecture new detecting method can be easily embedded by adding a new filter. This paper focuses on LSM filters. When the publisher detects abnormal data, it sends to the corresponding event handler.

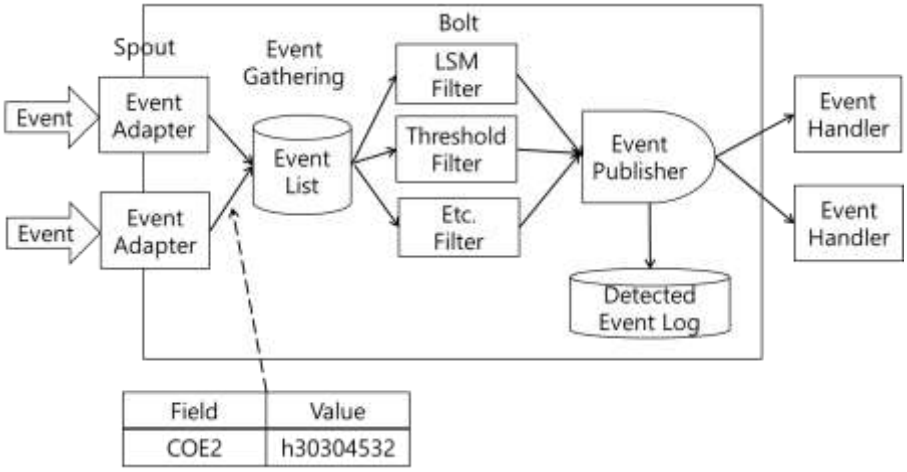

Figure 4. Architecture of CEP

3. Time Series Analysis for Abnormal Data Detection

There are various types of errors in the data generated in the manufacturing process. Previous studies have focused only on detection of data that is outside of the error tolerance range, but data on a pattern of time series error cannot be detected. As shown in Figure 4, the datum indicated by the arrow cannot be physically generated, so it can only be seen as the noise of the sensor. If this data is not filtered, the accuracy of the data may be a problem.

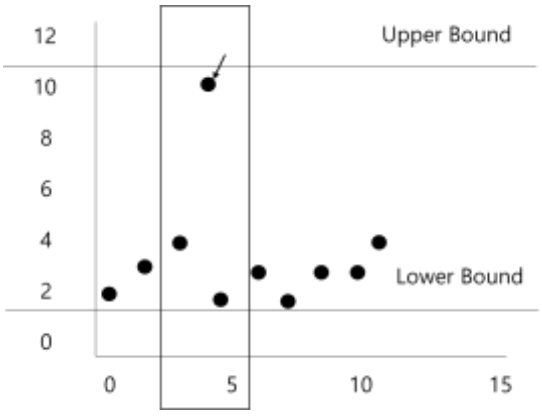

Figure 5. An Example of Undetected Sensor Data

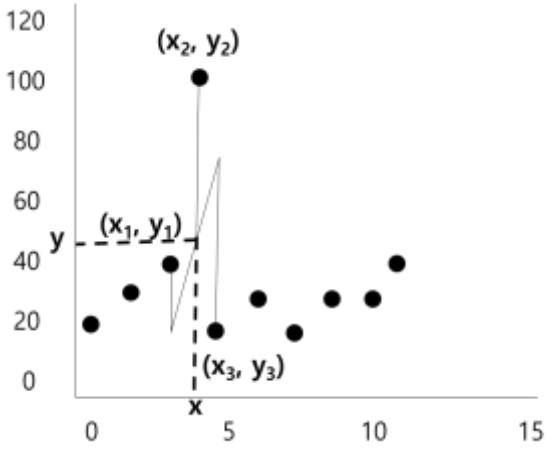

Figure 6. An Example of Sensor Data Detection using LSM

Figure 5 shows an example of the sensor data of the boiler where the X-axis represents time and the Y-axis represents temperature. The data indicated by the arrow is within the acceptable range of the boiler temperature but it was corrupted by noises and cannot be physically present. If such data is input, there is a high possibility that errors will occur in subsequent processing, so it should be detected and removed in real time.

Figure 6 shows how to filter these abnormal data using LSM. Providing that the size of the window is 3, the LSM can calculate the coordinates (x, y) of the expected value by letting the average value of the X axis be x and the average value of the Y-axis be y. After calculating the distance value between the expected value and the actual value of the fourth data, if the value exceeds the predetermined threshold, it is found that there is a problem with the data.

The purpose of using LSM is founding out the following linear model of data which minimize residual values.

$$f(x_i) = ax_i + b \tag{1}$$

The first step is finding out the a and the b values which minimize the left side of the equation (1) where w is the window size.

$$\sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} r_i^2 = \sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} (y_j - (ax_j - b))^2 \tag{2}$$

Two equations which are partial differentiations of a and b are shown in the following equations (3) and (4) respectively.

$$\sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} y_j x_j - a\sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} {x_j}^2 - b\sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} x_j = 0 \tag{3}$$

$$\sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} y_j - a\sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} x_j - nb = 0 \tag{4}$$

From the above equations the slope value a that can be derived as the following equation (5).

$$\mathrm{a} = \frac{\sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} (y_j - \bar{y})(x_j - \bar{x})}{\sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} (x_j - \bar{x})^2} \tag{5}$$

The second step is to find the window size that will most accurately detect abnormal data. Since the optimal window size may vary depending on the type and characteristics of the data, the window size should be determine using the mutation test method. The mutation test method is a method of deliberately inserting error data into normal data and testing how many errors a particular test method finds.

In the last step, the LSM equation derived from the previous step is used to check that whether the input data is normal or abnormal data. Figure 7 shows the process of detecting abnormal data detection. From the input stream, LSM method analyses the data with the slope value and window size from the above steps, then it can be decide the input data is normal or not with ease.

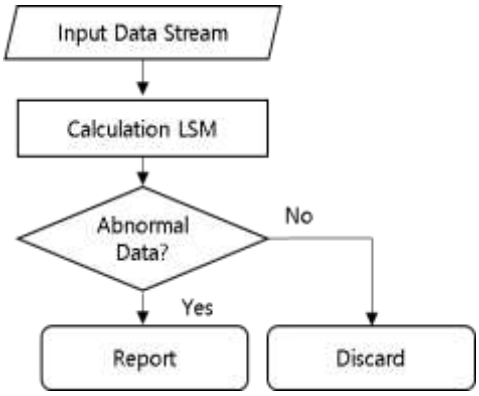

Figure 7. The Process of Detecting Abnormal Data

4. Experiments and Analysis

This experiment assumes a temperature control system for the virtual boiler. Experimental data was generated by inserting error data into the temperature data for using the mutation test technique. The goal of the test is to find the optimal window size that can find the most errors inserted.

The result of the experiment is shown in Table 1, if the window size is 10, all the inserted abnormal data are detected. If the window size is 5 or 15, it does not detect all abnormal data. This means that it is important to calculate the appropriate window size for practical application.

Table 1. Table Label

Time / Size	10	20	30	40	50	60	70	80	90
Apache Storm	0	0	0	0	0	0	0	0	0
5	1	2	1	1	1	2	1	0	0
10	2	4	3	5	3	2	2	4	3
15	2	2	1	2	1	1	0	2	1
Error count	2	4	3	5	3	2	2	4	3

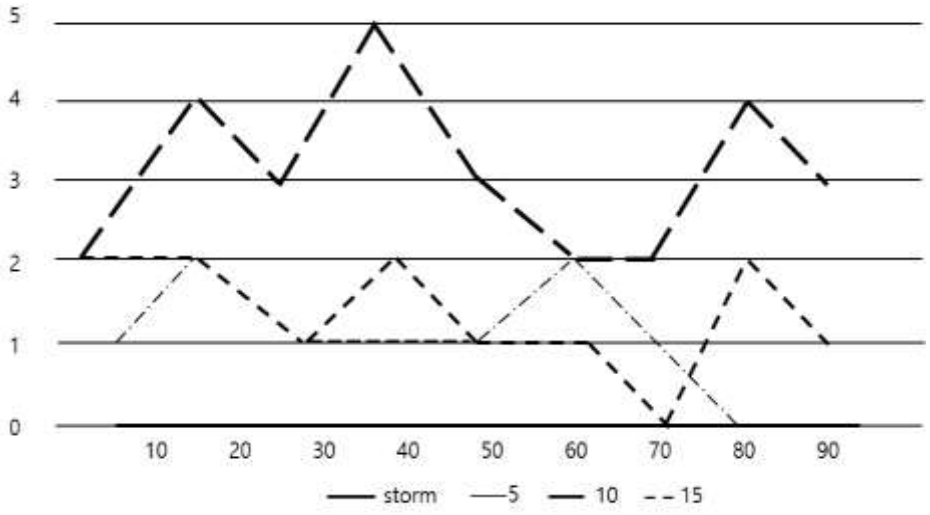

Figure 8. The Number of Abnormal Data Detected

Figure 8 shows the experimental results as a graph. X-axis shows the time series data that is input and Y-axis shows the number of detected errors.

LSM computes the slope to represent the values of the time series data within the window. Excrement shows that if the size of the window is large, the slope value for a large number of data is calculated. Therefore, a somewhat gentle slope appears and an accurate value cannot be measured. On the other hand, if the size of the window is small, it may be difficult to measure the accurate value since a rather steep slope can be measured. Therefore, it is the most important issue to specify the appropriate window size according to the characteristics of data and determining window size step should be done before the analysis.

5. Conclusion

Big data processing using systems such as Hadoop is not suitable for the smart factory which requires real-time processing. Therefore, it is mandatory to develop a CEP based real-time analysis system for increasing data processing requirements. In the smart factory system, existing error processing is concentrated on simple methods such as out of range errors. However, there are various types of errors. Accurate detection and handling of errors are important parts of the productivity and quality of manufacturing.

Recently, new facilities are being built as smart factories and the transition from existing facilities to smart factories has been increasing. Therefore, demands for

processing data generated from a large number of sensors are rapidly increasing. In this paper, we propose real-time big data processing of smart factories and detection of abnormal data using LSM with CEP engine suitable for smart factory. The paper applies CEP based on Apache Storm which processes input data in memory.

There are various kinds of errors in sensor data generated in smart factories. However, existing researches focus only on detection of data outside the error range of data, so that abnormal data that can be detected in a time series cannot be filtered. The detection and processing of the error are directly related to the accuracy of the data. This will be very helpful for the quality for the subsequent processing. For this purpose, this study proposes a method to detect abnormal data using LSM.

There are data generated from various kinds of sensors in smart factory. For example, there are various data such as the temperature, the illuminance, the oxygen concentration, the voltage, the current, and the distance measurement. In addition, the judgment of a suitable window size may vary depending on the application field. This model can handle this diversity only by specifying parameters.

Future research will include implementing the proposed system. This model provides out-of-bound error detection and detection of time-series errors using LSM. More intelligent analysis methods than LSM for time series data analysis will be included in future studies.

Acknowledgements

This paper was supported by the Kumoh National Institute of Technology Research Grant. This article is a revised and expanded version of a paper entitled “Abnormal Data Detection with CEP Engine for Smart Factory” presented at SERSC NGCIT 2017 on August 16-18, 2017 at Liberty Central Saigon Riverside Hotel, Ho Chi Minh City, Vietnam.

References

[1] X. Zhang and G. Wang, “Hadoop-Based System Design for Website Intrusion Detection and Analysis”, pp.1171-1174, 2015 IEEE international Conference on Smart city, **(2015)**.
[2] W. Yang, “Computing data quality indicators on Big Data streams using a CEP”, computational intelligence for Multimedia Understanding, 2015 Internation workshop, **(2015)**, pp.29-30
[3] M. Kurth, C. Schleyer and D. Feuser, “Smart factory and education: An integrated automation concept”, pp.1057-1061, 2016 IEEE 11th conference of Industrial Electronics and Applications, **(2016)**.
[4] M. Ali, B. Chandramouli, J. Goldstein and R. Schindlauer, “The Extensibility Framework in Microsoft StreamInsight”, IEEE, ICDE Conference, **(2011)**.
[5] 5. Seyoon Ko and J.-H. Won, “Processing large-scale data with Apache Spark”, The Korean Jaurnal of Applied Statistice, **(2016)**, pp.1077-1094.
[6] S.H. Kwon, D. Park, H. Bang and Y. Park, “Real-time and Parallel Semantic Translation Technique for Large-Scale Streaming Sensor Data in an IoT Environment”, korea information science society, 1th, **(2007)**, pp.1-10.
[7] Q. Guo and J. Huang, “A complex event processing based approach of multi-sensor data fusion in IoT sensing systems”, 2015 4th International Conference on computer science and network Technology, **(2015)**, pp.548-551.
[8] B. Rehman, C. Liu and L. Wang, “Least Square Method : A Novel Approach to Determine Symmetrical Components of Power System”, J Electr Eng Technol., **(2017)**, pp. 39-44
[9] H. Chi, “A Discussions on the Least-Square Method in the Course of Error theory and Data Processing”, 2015 International Conference on computational intelligence and Communication Networks, **(2015)**, pp.486-489.

Authors

Won-Chang Lee, he is a M.S. student in department of computer engineering at Kumoh National Institute of Technology, Gyeongsangbuk-do, Korea. He received his B.S. degree in department of computer engineering from Kumoh National Institute of Technology in 2015. His research interests include software engineering. Bigdata

Jae-Han Cho, he is a Ph.D. student in department of computer engineering at Kumoh National Institute of Technology, Gyeongsangbuk-do, Korea. He received B.S and M.S. degree in Computer Engineering from Kumoh National Institute of Technology, His research interests include software engineering, database design, and GPGPU based parallel processing. Bigdata

Lee-Sub Lee, he is an Associate professor of Department of Computer Engineering at the Kumoh National Institute of Technology Gyeongsangbuk-do, Korea. He received B.S. in Mathematics and M.S. degree in Computer Engineering from Sogang University, Seoul, Korea. He received his Ph.D. in Computer Engineering from Korea University, Seoul, Korea. He has worked as a senior researcher at Samsung SDS 1990 to 2004. His research work has been on the Software Engineering and Database System. His recent interest focuses on Software Testing, Bigdata