



Detection and Correction Method of Erroneous Data Using Quantile Pattern and LSTM

Chulhyun Hwang¹, Hosung Kim², and Hoekyung Jung^{3*}, *Member, KIICE*

¹Data Maroo Co., Daejeon, Korea

²K-Water, Daejeon, Korea

³Department of Computer Engineering, Pai Chai University, Daejeon 35345, Korea

Abstract

The data of K-Water waterworks is collected from various sensors and used as basic data for the operation and analysis of various devices. In this way, the importance of the sensor data is very high, but it contains misleading data due to the characteristics of the sensor in the external environment. However, the cleansing method for the missing data is concentrated on the prediction of the missing data, so the research on the detection and prediction method of the missing data is poor. This is a study to detect wrong data by converting collected data into quintiles and patterning them. It is confirmed that the accuracy of detecting false data intentionally generated from real data is higher than that of the conventional method in all cases. Future research we will prove the proposed system's efficiency and accuracy in various environments.

Index Terms: Data cleansing, Data quality, IoT, Water supply information

I. INTRODUCTION

Globally, the severity of water problems is intensifying, and there are also conflicts between countries. Therefore, attempts to combine big data analysis and intelligent technology, which have recently come to the fore in water management, are constantly being tried as part of the preparations for the fourth industrial revolution. In Korea, K-Water and others are leading in this way [1, 2].

In order to introduce big data analysis and intelligent technology into water information management, a high level of data quality must be provided. However, efforts to secure data quality in water management and related research have relied on simple rule-based algorithms or empirical methods due to the complexity of the collection system and the variety of sensors [3-5].

Many existing studies have focused on predicting missing

data. Long short-term memory (LSTM), a deep-running technology, it is known to have high accuracy in time series data [6]. However, the deep learning technique is a method of performing prediction on the detected data, and it is not a detection method. Unless the detection performance is guaranteed, the prediction performance is bound to have an effect. Detection of missing data is an area that is not desired, and detection performance is a serious problem because it has a very serious influence. In particular, error data compared to missing data is a much more complex problem because it occurs within the normal data range of the device and is a domain that needs the help of a domain expert.

This paper aims at improving the detection performance of the erroneous data presented above by using learning data without depending on expert judgment. To do this, we proposed the use of quantiles. The composition of the paper is

Received 22 October 2018, Revised 19 November 2018, Accepted 20 November 2018

*Corresponding Author Hoekyung Jung (E-mail: hkjung@pcu.ac.kr, Tel: +82-42-520-5640)

Department of Computer Engineering, Pai Chai University, 155-40 Baejae-ro, Seo-gu, Daejeon 35345, Korea.

Open Access <https://doi.org/10.6109/jicce.2018.16.4.242>

print ISSN: 2234-8255 online ISSN: 2234-8883

[©] This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © The Korea Institute of Information and Communication Engineering

as follows: in Section I, the necessity and outline of the study are explained. In Section II, related researches on data used in water information and prediction methods of erroneous and missing data are presented. In Section III, we propose a method of detecting false data based on quantiles. In Section IV, we present experimental methods and results. Finally, Section V presents conclusions and future research areas.

II. EXISTING RESEARCHES ABOUT DATA REFINEMENT

Data refinement is a process of discovering and classifying meaningless or erroneous data types using machine learning methods. It is a process of retrieving and correcting errors, inconsistencies, and missing data to improve the quality of data.

Data refinement is an essential process for data processing and analysis and is an infrastructure that provides confidence in processing results. Therefore, K-Water has also been in need of quality control for many erroneous and missing data generated during the process of data processing on water quality and quantity of water treatment plant.

In order to predict the missing data, various statistical techniques have been mainly used and researches on machine learning such as artificial neural networks have been carried out [7]. In particular, statistical methods for predicting missing data due to natural conditions, such as weather deterioration, have been studied [8, 9].

Recent research has found that recurrent neural network (RNN), which is a deep learning technology, can achieve higher performance than existing statistical techniques for missing data imputation using time-series characteristics of IoT data [10].

Despite this interest and research on missing data, research has not been conducted on erroneous data found within normal ranges. Erroneous data is much more domain knowledge than missing data and its detection process is difficult. This paper presents the detection process of erroneous data in order to overcome the limitation of this study.

III. ERRONEOUS DATA DETECTION ALGORITHM USING QUANTILES

In this paper, we propose an algorithm that detects specific error data and predicts normal data values when data input is simultaneously in multiple IoT sensor environments.

The most widely used prediction algorithm in the IoT environment is mean imputation method. The average replacement method is a method of predicting the average value of data generated recently (T_p-n to T_p-1) before the

prediction time T_p . The mean imputation method is mainly used in the prediction of sensor data because it is simple to implement and exhibits good performance when certain values are continuously generated.

The error detection method using the mean imputation method determines an error when the difference between the predicted value and the input value exceeds the threshold value and replaces the actual data with the predicted value. This method has a problem of detecting normal data as error data when the prediction accuracy is low. In addition, detection and correction accuracy vary greatly depending on how the threshold is adjusted when predicting high-precision data. In addition, there is a problem in that the interaction between data input at the same time cannot be considered.

The proposed algorithm uses two methods to detect and predict error data.

First, we utilize quantiles to convert precision IoT sensor data into a simplified pattern. The quantile pattern verifies whether the data of currently input sensors have occurred in the past.

The second uses LSTM, a deep learning method to predict time series data. If the quantile is simply patterned by categorizing the sensor data, the LSTM algorithm generates the predicted value while maintaining the accuracy of the sensor.

The difference between the existing algorithm and the proposed algorithm is as follows.

Existing algorithms use only prediction algorithms to detect error data. However, the proposed algorithm is a consensus model considering both the results of the quantile pattern and the LSTM prediction algorithm. By performing quantile pattern and LSTM prediction at the same time, it is possible to easily retrieve past data and maintain accuracy. Finally, since the data input at the same time is converted into a quantile pattern, the correlation between the sensors can be considered.

Fig. 1 shows the difference between the existing algorithm and the proposed algorithm.

The proposed algorithm consists of 4 steps.

The first step is to build a historical data pattern into a DB. The proposed algorithm utilizes n-quantiles to prevent too many patterns from being generated. And it saves pure data pattern by removing time information and redundancy.

The second step is to convert a new data pattern into a quantile to see if there is a matching pattern in the constructed DB. If there is no matching pattern, you may suspect that one or more of the entered data contains error data.

The third step is to identify the error data among the data patterns entered at the same time. Remove data one by one to see if there is a matching pattern in the DB. If there is a matching pattern, the removed data is judged as error data. Fig. 2 shows the overall flow of the proposed algorithm.

If the quantile pattern does not exist in the existing DB, the quantile of one of the patterns is excluded and then

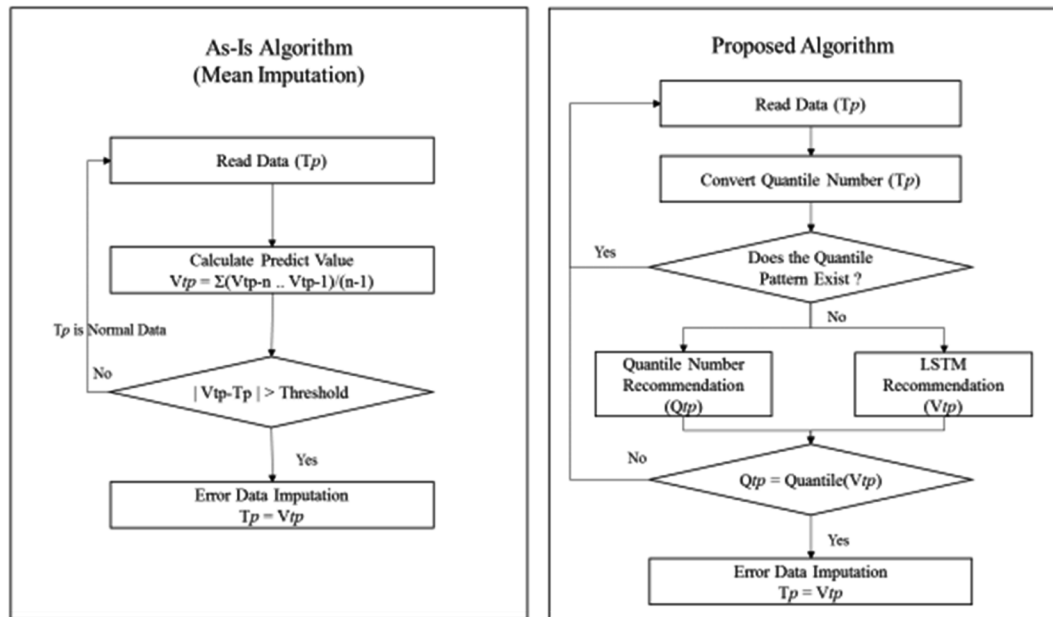


Fig. 1. As-Is algorithm versus proposed algorithm.

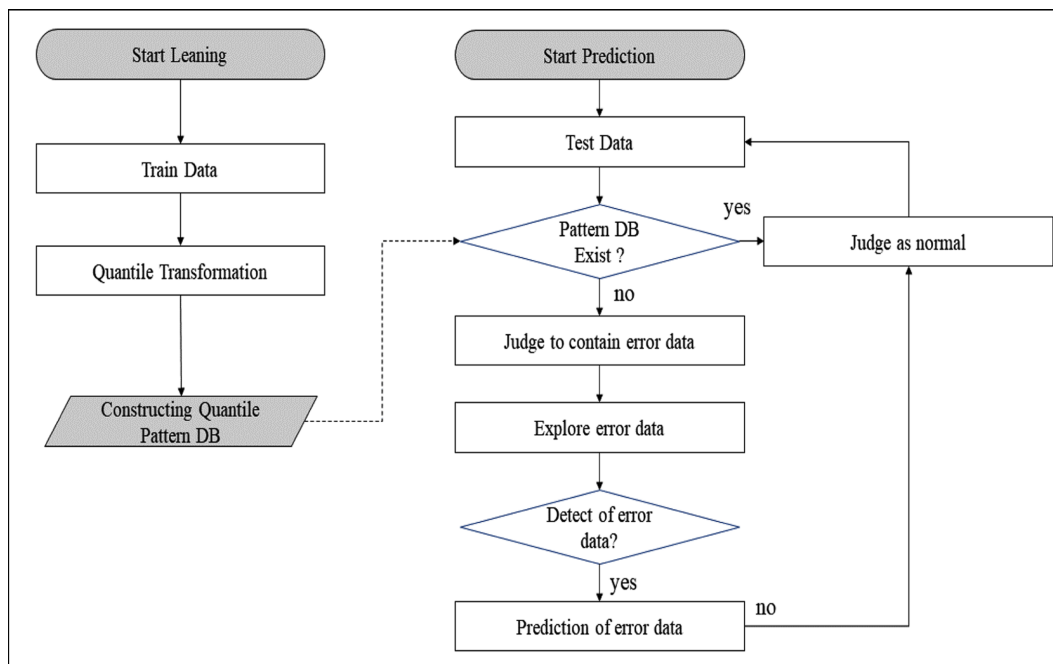


Fig. 2. Process flow of proposed algorithm.

retrieved again. If a pattern excluding one quartile of the sensor is present in the existing DB, we assume that the sensor except the quartile contains the error value. Fig. 3 shows the process of determining error data among data patterns.

The final step 4 is to correct the error determination data when it is judged to be error data. If the existing quantiles recorded in the DB match the quantiles of the data predicted by the LSTM, the prediction values of the LSTM are determined as the correction values of the corresponding data.

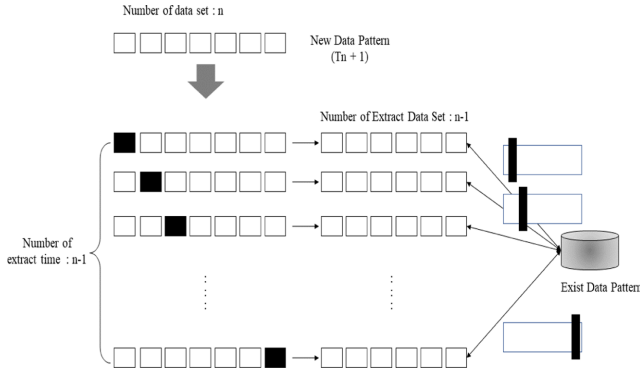


Fig. 3. Procedure for determining the error data.

IV. EXPERIMENT

The experimental method uses the proposed algorithm and the existing algorithm to compute and compare the detection accuracy and the detection error rate, which detect any erroneous data, for the experimental data.

The existing algorithms apply ‘mean imputation’ which is a method to apply past average data for a certain period of time when judging or correcting error data. Mean imputation is the most widely used algorithm in industrial field.

The data used in the experiments are divided into learning data and test data, all of which are extracted from the actual operating environment. Both data use the data set of consecutive times for the safety of the experiment and generate the wrong data by applying arbitrary operation to the test data.

The test objective is to measure how accurately the false data generated randomly is detected, to generate a predicted value for the detected false data, and then to compare it with the initial input value to calculate the predictive rate.

The data used in the experiment is based on data extracted from the IoT sensor collected at the site managed by K-Water. The experimental data was constructed as follows.

The learning data is for 1 year from 00:00 on October 1, 2016 to 23:59 on September 30, 2017. Data was collected from every 5 IoT sensors every minute. The total number of data is 525,600. The test data consists of a total of 500 data collected every minute from five IoT sensors on October 1, 2017. The learning and test data are produced by the same sensors.

Experimental procedure generates experimental data and performs learning for each algorithm. Then, the detection accuracy and the prediction rate of the erroneous data and the missing data are calculated. Table 1 compares the experimental method of the existing algorithm and the proposed algorithm.

Experimental results are analyzed by comparing the experimental results with the conventional methods and the experimental results with the proposed algorithms. The results of

Table 1. Percentage of erroneous data detected

Division	Exist algorithm	Proposal algorithm
Learning	Unnecessary	Quantile pattern DB LSTM learning
Detection	Difference between average and input value	Comparison of quantile pattern DB and quantile number
Prediction	Average value	LSTM predict number

Table 2. Percent of erroneous data detected as erroneous data

Division	#001	#002	#003	#004	#005	Avg.
Exist						
20%	98	80	82	80	76	83.2
40%	40	48	26	44	40	39.6
60%	0	0	14	0	4	3.6
80%	0	0	0	0	0	0
100%	0	0	0	0	0	0
Proposal	82	38	72	60	26	55.6



Fig. 4. Percent of erroneous data detected as erroneous data.

the comparison between the existing method and the proposed method are shown in Table 2 and Fig. 4, respectively.

The experimental method uses the proposed algorithm and the existing algorithm to compute and compare the detection accuracy and the detection error rate, which detect any erroneous data, for the experimental data.

As shown in Fig. 5 and Table 3, the result of detecting normal data as erroneous data is presented. Especially, in case of the existing method 20% error level, the detection rate of erroneous data is high, but at the same time, the rate of detecting normal data as erroneous data also becomes high.

This problem cannot be used as a detection method because it adversely affects data quality. Therefore, it shows the highest detection rate in all areas except 20%. This can adversely affect data quality and cannot be used as a detec-



Fig. 5. Judgment error for normal data.

Table 3. Number of normal data detected as erroneous data

Division	#001	#002	#003	#004	#005	Avg.
Exist						
20%	0	67	69	0	48	36.8
40%	0	0	31	0	0	6.2
60%	0	0	31	0	0	6.2
80%	0	0	0	0	0	0
100%	0	0	0	0	0	0
Proposal	0	0	0	0	0	0

tion method because it ignores actual data. Therefore, an error level of 20% is excluded from the analysis of the experimental results. In all cases except the 20% error level, the proposed detection method shows the highest detection rate.

V. CONCLUSION

Finding erroneous data is a more difficult area than detecting or predicting missing data. In general, erroneous data detection is a field that has not been studied, especially since normal data can be used to determine erroneous data and complex business rules must be considered. In this study, we proposed a method to increase the detection rate of erroneous data by using quantile patterns and confirmed that the performance is higher than previous studies.

Among the valid cases, the existing method showed a detection performance improvement of 34% including the detection error rate. If the detection error does not occur, the detection rate of the proposed method is 55.6% while the error detection rate is 0 in the conventional method.

In conclusion, it can be seen that the proposed method has a higher detection performance than the conventional method.

Future research should be devoted to the solving of the prediction and long-term prediction problems of data identified as erroneous data.

ACKNOWLEDGMENTS

This research was supported by The Leading Human Resource Training Program of Regional Neo industry through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (No. 2016H1D5A1911091). And this study was supported by the research grant of Pai Chai University in 2018.

REFERENCES

- [1] J. R. Kim, G. W. Shin, H. S. Kim, and S. T. Hong, "A study on cleansing algorithm for outlier data in water supply," in *Proceedings of the Korean Institute of Communications and Information Sciences Summer Conference*, pp. 19-20, 2017.
- [2] G. W. Choi, K. S. Song, and J. Kang, "Understanding and policy assignment of R&D of deep learning," Korea Institute of S&T Evaluation and Planning, 2016 [Internet], Available: <https://www.kistep.re.kr/c3/sub3.jsp?brdType=R&bbIdx=10484>.
- [3] S. M. Hong and A. Jang, "The development study on the integrated management system for water information based on ICT," *Journal of Korean Society of Environmental Engineers*, vol. 39, no. 12, pp. 723-732, 2017. DOI: 10.4491/KSEE.2017.39.12.723.
- [4] S. Baek, C. Seong, S. Choe, Y. Park, and M. Kim, "Mobile water quality monitoring system using ion-selective-electrodes," *Journal of the Institute of Electronics and Information Engineers*, vol. 55, no. 2, pp. 29-38, 2018. DOI: 10.5573/ieie.2018.55.2.29.
- [5] C. H. Kim, L. S. Kang, and H. J. Kim, "The development of information breakdown structure for integrated management of water filtration plants," *Journal of the Korean Society of Civil Engineers*, vol. 37, no. 5, pp. 863-869, 2017. DOI: 10.12652/Ksce.2017.37.5.0863.
- [6] V. Q. Nguyen, L. Van Ma, and J. Kim, "LSTM-based anomaly detection on big data for smart factory monitoring," *Journal of Digital Contents Society*, vol. 19, no. 4, pp. 789-799, 2018. DOI: 10.9728/dcs.2018.19.4.789.
- [7] J. M. Jerez, I. Molina, P. J. Garcia-Laencina, E. Alba, N. Ribelles, M. Martin, and L. Franco, "Missing data imputation using statistical and machine learning methods in a real breast cancer problem," *Artificial Intelligence in Medicine*, vol. 50, no. 2, pp. 105-115, 2010. DOI: 10.1016/j.artmed.2010.05.002.
- [8] F. Liu, Z. You, W. Shan, and J. Liu, "A grey system based missing sensor data estimation algorithm," in *Proceedings of 2012 2nd International Conference on Computer Science and Network Technology*, Changchun, China, pp. 482-486, 2012. DOI: 10.1109/ICCSNT.2012.6525982.
- [9] N. I. Nwulu, "Evaluation of machine learning classification algorithms & missing data imputation techniques," in *Proceedings of 2017 International Artificial Intelligence and Data Processing Symposium*, Malatya, Turkey, pp. 1-5, 2017. DOI: 10.1109/IDAP.2017.8090315.
- [10] Z. C. Lipton, D. C. Kale, and R. Wetzel, "Modeling missing data in clinical time series with RNNs," *Proceedings of Machine Learning Research*, vol. 56, pp. 253-270, 2016.



Cheolhyun Hwang

received the M.S. degree in 1995 and Ph.D. degree in 2017 from the Department of Computer Engineering of Pai chai University, Korea. From 1991 to 2000, he worked for Korea Navy as a Computer Officer. Since 2001, he has worked in the variety of IT professional companies, where he now works as a data consultant. His current research interests include deep learning, machine learning, IoT, big data, and artificial intelligence.



Hosung Kim

received the M.S. degree in 2008 from the Department of Electronic Information and Communication Engineering, Chungnam National University and Ph. D. degree in 2018 from the Department of Computer Engineering of Pai Chai University, Korea. Since 1995, he has worked in the Department of Management Information System at K-Water (Korea Water Resources Corporation), where he now works as a General Manager. His current research interests include data cleansing algorithm, deep learning, and big data.



Hoekyung Jung

received the M.S. degree in 1987 and Ph.D. degree in 1993 from the Department of Computer Engineering of Kwangwoon University, Korea. From 1994 to 1995, he worked for ETRI as a researcher. Since 1994, he has worked in the Department of Computer Engineering at Pai Chai University, where he now works as a professor. His current research interests include multimedia document architecture modeling, machine learning, IoT, big data, and artificial intelligence.