Handling anomaly in residential energy consumption data

Youhee Choi
Energy ICT Research Section
ETRI
Daejeon, Korea
yhchoi@etri.re.kr

Tai Yeon Ku
Energy ICT Research Section
ETRI
Daejeon, Korea
kutai@etri.re.kr

Wan-Ki Park
Energy ICT Research Section
ETRI
Daejeon, Korea
wkpark@etri.re.kr

***Abstract*— The operation of HVAC (heating, ventilation, and air-conditioning) accounts for a large proportion of energy consumption in buildings. Accurate estimation of the energy demand for efficient operation of HVAC is important. In this respect, many researches have been conducted to collect energy consumption data for building energy management and to use the collected data for analysis and prediction. Recently, with the advancement of AI technology, there are many studies to apply AI technology to energy management. In this regard, since performance of AI models depend on the quality of training data, a method for effectively handling missing values and outliers in training data should be considered. This study proposes a method for handling missing values and outliers considering the semantics of data required for energy consumption prediction.**

Keywords—energy, outlier, missing value, energy consumption

I. INTRODUCTION

Demand for efficient cooling and heating management is increasing due to extreme weather changes due to global warming. Recently, efforts in related studies have been made for the application of AI technology for efficient heating and cooling management is also being conducted. Basically, in order to apply AI technology, training data for learning is required, and the quality of the data affects the performance of AI model. For efficient cooling and heating management, data such as thermal energy consumption data and sensor values for measuring temperature and pressure on the thermal energy supply/consumption side can be required. Such data is meaningful only when it is normally collected and accumulated over a long period of time, but missing values or outliers may occur due to unexpected circumstances. In this respect, various data imputation studies have been made to deal with these missing values or outliers, and recently, studies using AI have been made[1-5]. However, AI is a black box type method in which rules are found and applied through data learning inside the system without human intervention. For this reason, it is difficult to know exactly which rules are applied to deal with missing values and outliers. In addition, when handling missing value or outlier data, the method for handling the outlier or missing value may differ not only from the past history of the data, but also from the hidden meaning of the data and its relationship with other data. Therefore, in this paper, we propose a method for handling missing values and outliers considering the semantics of energy data in order to prevent energy consumption prediction models from taking a long learning time or degrading accuracy performance due to outliers and missing values.

II. RELATED RESEARCHES

A. Machine learning methods for handling missing values and outliers

Various missing data imputation researches have been conducted applying deep learning methods such as k-nearest neighbor (K-NN) method[1], support vector machines (SVM)[2], and convolutional neural network-long short term memory (CNN-LSTM) [3]. Kim et al presented a learning-based adaptive imputation method (LAI) for handling missing power data in an energy system by applying k-NN algorithm[1]. Kasim Zor et al. compared a linear interpolation method and a marginal mean imputation method to handle missing data for energy prediction applying multilayer perceptron neural networks (MLPNN) and SVM[2]. Hussain et al. presented a framework for handling large missing values gaps observed in electricity consumption timeseries data of IoT based home appliances[3].

Also, existing researches about outlier handling methods have been conducted applying supervised learning methods such as SVM, linear regression, decision tree, and artificial neural network(ANN), etc. and unsupervised learning methods such as clustering algorithms and association rules[4]. Lei et al presented a dynamic anomaly detection algorithm for building energy consumption data for dynamic detection of point anomalies and collective anomalies[5].

III. HANDLING ANOMALY IN RESIDENTIAL ENERGY CONSUMPTION DATA

Fig. 1 shows the energy consumption data collected from 0:00 on June 1, 2023 to 23:00 on June 19, 2023 in descending order, for one apartment complex (468 households), for all households at one-hour intervals.

meter_id	date_time	gas	hotwater
S4200154000883468	2023-06-19 23:00:00	64.8	820.7
S4200154044227383	2023-06-19 23:00:00	213.9	742.8
S4200154454387764	2023-06-19 23:00:00	138.8	505.2
S4200154784042063	2023-06-19 23:00:00	73.2	214.4
S4200154938737187	2023-06-19 23:00:00	98.9	473.2
S4200118361119339	2023-06-01 00:00:00	199.5	494.1
S4200119377555359	2023-06-01 00:00:00	158.5	468.4
S4200115440405068	2023-06-01 00:00:00	123.0	419.7
S4200115790025849	2023-06-01 00:00:00	45.6	508.9
S4200115941737782	2023-06-01 00:00:00	94.7	845.8

201240 rows × 4 columns
[Energy usage data collected from 2023-06-01 0:00:00~ to 2023-06-19 23:00:00]

Column name	Description
meter_id	Individual household ID
date_time	Date and time that cumulative usage value are collected for each household
gas	Cumulative usage values for gas
hotwater	Cumulative usage values for hot water

Fig. 1. The example of dataset of residential energy consumption

In Fig. 2, (a) shows the number of data collected for each hour of the day, and (b) shows the number of data collected by day for the entire period.

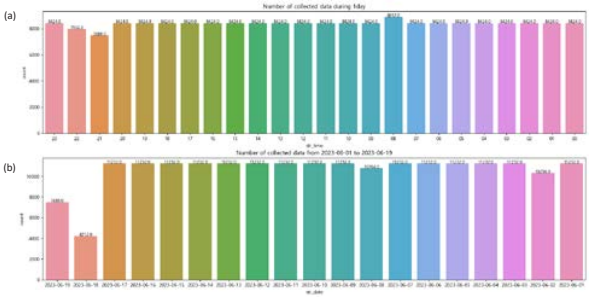


Fig. 2. The statistics of the number of example data collected

As shown in Fig. 2, there should be 12168 data (468 households * 26 days) for each time zone in (a), but it can be seen that the number is insufficient, and in (b) for all dates there should be 11232 (468 households * 24 hours) data should exist, but it can be seen that the number of data is insufficient on some dates. Cases in which anomaly can occur based on the example data can be classified into two cases that the first case is where data for a certain period of time are missing and the second case is where temporary error in measuring value has been occurred.

The first case that can be classified as an outlier is when the difference between the accumulated consumption data of the previous time zone and the accumulated consumption data after one hour for each household is a negative value. The following sample data are cases where the accumulated consumption difference values are negative values, and can be divided into two cases: normal case (a) or abnormal case (b) as shown in Fig. 3.

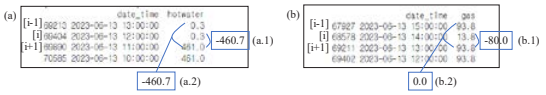


Fig. 3. Examples of candidates of outliers

Case (a) can be seen as a case where the accumulated consumption data is reset. In (a), the accumulated consumption data was 461.0 until 11:00 on June 13, but it was measured as 0.3 at 12:00 on the same day. And the value measured at 13:00 is also 0.3, so it is difficult to regard it as a measurement error that has occurred temporarily. Case (b) can be regarded as a case where the measured value is abnormal due to a temporary measurement error. In (b), the accumulated consumption data was 93.8 at 13:00 on June 13, but it was measured as 13.8 at 14:00. At 15:00, it was recorded again as 93.8. Since it was measured as 93.8 even at 13:00, it can be regarded as a temporary error.

In the case of (a), the difference between the value of the (i+1)-th time zone that is an hour before the i-th time zone and the value of the i-th time zone is a negative value as shown in (a.1). Also, the difference between the value of the (i-1)-th time zone that is an hour after the i-th time zone and the (i+1)-th time zone value is still a negative value as in (a.2). In the case of (b.2), the difference between the value of the (i-1)-th time zone and the value of the (i+1)-th time zone is a non-negative value, so it can be regarded as the accumulated consumption data became similar to the previous one. That is, it can be regarded that an error has occurred temporarily.

In order to distinguish these cases (a) and (b), the value of the time zone one hour later than the i-th time zone is compared with the value of the time zone one hour before the i-th time zone as shown in Fig. 4. Even in this case, if the difference is a negative number, it means that the accumulated consumption data is still being measured lower than the previous accumulated consumption data, so this value can be determined as a normal value rather than a temporary error.

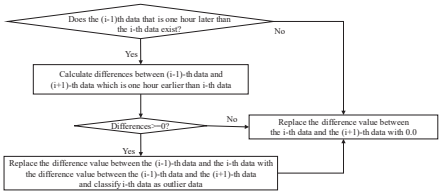


Fig. 4. The classification logic of outliers

The second case that can be classified as an outlier is when the difference between the accumulated consumption data of the previous time zone and the accumulated consumption data after one hour is above average value. The following sample data are cases where the accumulated consumption difference values are above average values, and can be divided into two cases: a normal case (a) or an abnormal case (b) as shown in Fig. 5. The case (a) is a case where the accumulated amount of gas used is continuously increasing, and a large amount of consumption occurred at a specific point in time (a.1). The case (b) is a case after the data of the previous time zone has missed, a large amount of used amount occurred at a specific point in time (b.1).

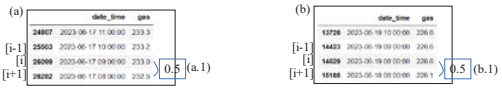


Fig. 5. Examples of candidates of outliers

In order to distinguish (a) and (b) in Fig. 5, if the difference is greater than average, it is necessary to examine whether the data of the previous time zone has missed as shown in Fig. 6. If the time interval from the measurement time of the (i+1)-th data to the i-th data is greater than an hour interval, it can be considered as meaning that a missing data exists. If the time interval is greater than 1 hour, the i-th value is divided by the missing time interval since there may be an influence by the value accumulated during the missed period. If the difference divided by the corresponding time interval is greater than the average, the i-th difference can be replaced by the divided difference, and if it is smaller than the average, the i-th difference can be replaced by average value. In addition, the i-th data can be classified as an outlier data to identify that the previous time zone data is missing.

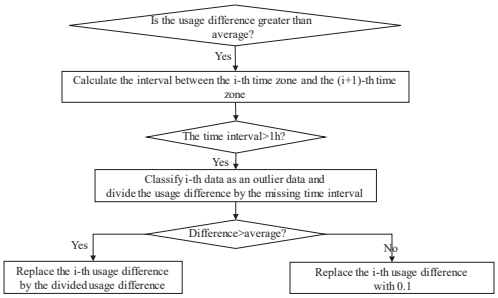


Fig. 6. The classification logic of outliers

IV. EXAMPLES

Fig. 7 shows the number of gas consumption difference values for the entire dataset and the data list corresponding to candidates of outliers.

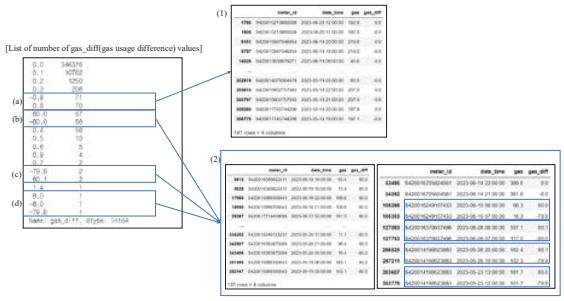


Fig. 7. Examples of candidates of outliers

In (1), the data list corresponding to (a) in Fig. 7, except for one data difference, it can be seen that the negative values of accumulated consumption difference for each household became positive values equal to the same amount after 1 hour. In the case of (1), it can be regarded as the accumulated gas consumption was measured as a value smaller than the previous hour due to a temporary value error and then it was measured as a normal value with no change in gas consumption. In (2), which is the data list corresponding to (b), (c), and (d) in Fig. 7, in the case of (b), there are 66 values that accumulated consumption differences are -80.0, while there are 67 values that accumulated consumption differences are 80.0. This can be seen as the case in which some change in gas consumption occurred when the time of occurrence of a temporary measurement error. Similarly, cases (c) and (d) also include cases in which a positive value of the same amount is not measured after a negative value is measured. Comparing with the actual data list in (2), the values calculated as -79.9 and -79.8 were calculated as 80.0 in the next time period, respectively, indicating that the change in gas consumption occurred when the time of occurrence of a temporary measurement error.

The result of handling outliers by applying the logic described in the previous chapter is shown in Fig. 8.

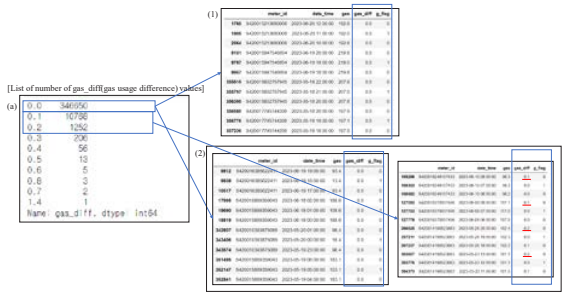


Fig. 8. Example datasets after handling outliers

In Fig. 8(a), it can be seen that the values of 0.0, 0.1, and 0.2 are increased instead of negative values in Fig. 7. In addition, the negative values that existed in (1) and (2) of Fig. 7 are replaced with 0.0 or 0.1 and 0.2 that reflect accumulated values of actual consumption in (1) and (2) of Fig. 8, and 'g_flag's of the replaced values are set to 1 showing them as outliers.

Fig. 9 shows the number of hot water consumption difference values for the entire dataset and the data list corresponding to candidate of outliers.

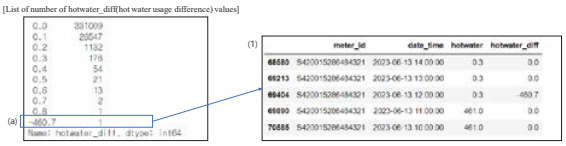


Fig. 9. Examples of candidates of outliers

In (1), the measured values before the time period in which a negative value(a) was measured for the household and the measured values after the time period when the negative value was measured are shown. As shown in Fig. 9 (1), Until 11:00 on June 13, 2023, 461.0 was measured as accumulated hot water consumption, but from an hour later, it is measured as 0.3. Accordingly, although the negative value was calculated at 12:00, since accumulated consumption data are maintained as same values thereafter, it is difficult to regard it as a temporary error. In this case, since the difference in accumulated consumption between 11:00 and 12:00 is meaningless, the difference in consumption at 12:00 is set to 0.0 according to the logic described in the previous chapter as shown in Fig. 10. And to show it as the outlier, the 'hw_flag' value of accumulated hot water consumption data at 12:00 was set to 1.

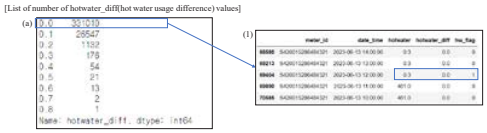


Fig. 10. Example datasets after preprocessing

Fig. 11 shows the number of gas consumption difference values for the entire dataset and the data list corresponding to candidates of outliers that are above average.

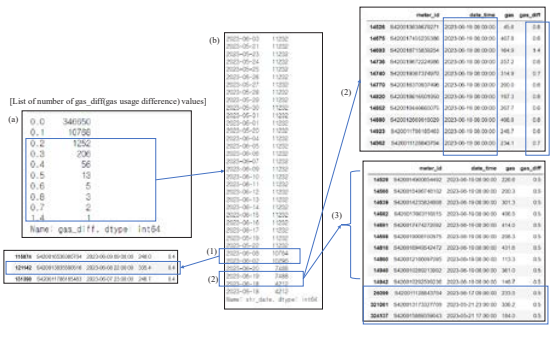
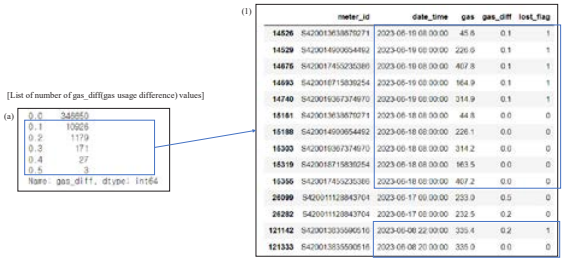


Fig. 11. Examples of candidates of outliers

In (a) of Fig. 11, it can be seen that while most of the accumulated consumption difference values are 0.0 or 0.1, various values from 0.2 to 1.4 exist. In this case, it may be the case that the time interval with the previous data is more than one hour because the accumulated consumption data to be collected at an hourly interval is missing. In this regard, the data corresponding to the difference values from 1.4 to 0.6 in (a) correspond to the difference values when data of the previous time zone is missed, as in (2). In this case, the

'lost_flag' is set to 1 to show that is affected by missing data, and then set to the difference value calculated by the algorithm in Fig. 12. In addition, from 0.5 or less, as in case (3), it is necessary to distinguish between a normal case and a case due to missed data because the case of a large amount of actual consumption and the case of missing data are mixed.



	meter_id	date_time	gas	gas_diff	lost_flag
14526	S4200136338570271	2023-06-18 08:00:00	45.6	0.1	1
14529	S4200149006544162	2023-06-19 08:00:00	226.9	0.1	1
14676	S4200174552353B6	2023-06-19 08:00:00	407.8	0.1	1
14693	S4200187158382S4	2023-06-19 08:00:00	164.9	0.1	1
14740	S4200193673749?0	2023-06-19 08:00:00	314.9	0.1	1
15181	S4200136338570271	2023-06-18 08:00:00	44.8	0.0	0
15188	S4200149006544162	2023-06-18 08:00:00	226.1	0.0	0
15303	S4200193673749?0	2023-06-18 08:00:00	314.2	0.0	0
15319	S4200187158382S4	2023-06-18 08:00:00	163.5	0.0	0
15356	S4200174552353B6	2023-06-18 08:00:00	407.2	0.0	0
26099	S4200111288437O4	2023-06-17 08:00:00	233.0	0.5	0
26282	S4200111288437O4	2023-06-17 08:00:00	232.5	0.2	0
121142	S4200138355905I6	2023-06-06 22:00:00	335.4	0.2	1
121333	S4200138355905I6	2023-06-06 20:00:00	335.9	0.0	0

Fig. 12. Example datasets after preprocessing

V. CONCLUSIONS

In this paper, we proposed a method for handling missing values and outliers considering the semantics of energy consumption data in order to prevent energy consumption prediction models from taking a long learning time or degrading accuracy performance due to outliers and missing values. In the future, we will evaluate the proposed method's substantiality by applying to other sites.

ACKNOWLEDGMENT

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning(KETEP) and the Ministry of Trade, Industry & Energy(MOTIE) of the Republic of Korea (No. 2021202090028C)

REFERENCES

[1] M. Kim, S. Park, J. Lee, Y. Joo, and J. Choi, "Learning-based adaptive imputation method with kNN algorithm for missing power data", Energies, v.10, n.10, p.1668, Oct. 2017.

[2] K. Zor, O. Celik, O. Timur, H. B. Yildirim, and A. Teke, "Simple approaches to missing data for energy forecasting applications", 16th International Conference on Clean Energy (ICCE), Sep. 2018.

[3] S. Hussain, A. Aziz, Md. Hossen, N. Aziz, G. Murthy, and F. Mustakim, "A Novel Framework Based on CNN-LSTM Neural Network for Prediction of Missing Values in Electricity Consumption Time-Series Datasets", Journal of Information Processing Systems,18(1), p.115-129, Feb. 2022.

[4] C. Fan, X. Fu, Y. Zhao, and J. Wang, "Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data", Applied Energy, p.1123–35, 2018.

[5] L. Lei, B. Wu, X. Fang, L. Chen, H. Wu, and W. Liu, "A dynamic anomaly detection method of building energy consumption based on data mining technology", Energy, p.1-19, Oct. 2022.