International Journal of Internet, Broadcasting and Communication Vol.15 No.2 240-247 (2023)
http://dx.doi.org/10.7236/IJIBC.2023.15.2.240

IJIBC 23-2-27

A Study on Abnormal Data Processing Process of LSTM AE - With applying Data based Intelligent Factory

Youn-A Min

Professor, Applied Software Engineering, Hanyang Cyber University, Korea
yah0612@hycu.ac.kr

Abstract

In this paper, effective data management in industrial sites such as intelligent factories using time series data was studied. For effective management of time series data, variables considering the significance of the data were used, and hyper parameters calculated through LSTM AE were applied. We propose an optimized modeling considering the importance of each data section, and through this, outlier data of time series data can be efficiently processed. In the case of applying data significance and applying hyper parameters to which the research in this paper was applied, it was confirmed that the error rate was measured at 5.4%/4.8%/3.3%, and the significance of each data section and the significance of applying hyper parameters to optimize modeling were confirmed.

Keywords: *LSTM, AE, intelligent factory, AI, abnormal-data*

1. Introduction

Data-based intelligent factory is a technology that integrates and automates the entire process from product planning to sales using cutting-edge information technology, and refers to an intelligent factory that implements digital information [1, 2]. As another meaning of data-based intelligent factory, it also refers to production automation that improves productivity by linking and integrating data in real time and pursues the production of customized products through energy saving [2].

Recently, due to environmental factors such as Corona 19, the importance of non-face-to-face and remote has emerged, and interest in the application of data-based intelligent factory technology has increased [1]

According to Statista[1-3], a global consulting firm, the global data-based intelligent factory market is reported to grow by 9.6% annually from $153.7 billion in 2019 to $244 billion in 2024 [1-3] The United States, the leader in the field, has an average annual average of 8.8 % or more, and China has an average annual growth rate of 12.2% [1-3].

In the case of domestic data-based intelligent factories, they are operated mainly in the manufacturing industry, and the market size, which was 10.42 trillion won in 2018, is expected to increase by 11.4% or more per year on average, increasing to 19.75 trillion won in 2024 [2].

Manuscript Received: April. 16, 2023 / Revised: April. 20, 2023 / Accepted: April. 22, 2023
Corresponding Author: yah0612@hycu.ac.kr
Tel: +82-2-2290-0872, Fax: +82-2-2290-0872
Professor, Applied Software Engineering, Hanyang Cyber University, Korea

Table1. Smart factory market size by country (Unit: USD billion, %)

NAT	year: 2022	year: 2024(exp)	growth rate
America	29	41	9
China	26	47	12
Japan	16	26	10
German	11	17	10
Kores	10	15	11

The reason for the growth of domestic data-based intelligent factories is that it is possible to secure various data and analyze big data generated during factory operation [3-5]. However, the artificial intelligence technology currently applied to data-based intelligent factories does not deviate significantly from big data analysis. In this situation, overlooking abnormal data in many data-driven intelligent factories can lead to transformation of results. If the scale and application range of the data is wide, the damage can be greater [3-5].

In this study, as a method to efficiently process abnormal data of data-based intelligent factories, the modified LSTM AE processing is studied for time-series data-based data, and the experiments and results of the research are verified by applying the data generated in practice.

2. Related work

2.1 Artificial Intelligence Technology of Data-based Intelligent Factory

The data-based intelligent factory applies technology that enables automation by securing, analyzing, and utilizing data from various machines. Currently, in Korea, data-based intelligent factories are the most used in the manufacturing sector [2, 3]. Compared to general factories managing factories with production facilities and control systems, data-based intelligent factories enable intelligence and data analysis through data collection [4-6]. Figure 1 is the data processing process of the intelligent factory.

Figure 1. Data process of the Intelligent Factory

Currently, artificial intelligence technologies that can be used in data-based intelligent factories include generative design and smart work for product design and design [1 ,4-6]. Generative design is a technology that can create and recommend customized product designs [3]. The smart workbench enables safe work environment and customized work plan assignment for workers [1, 4-6]. The intelligent sensory interaction guide introduces and utilizes systems such as augmented reality and virtual reality so that workers can handle tasks skillfully. Table 2 shows the application procedure of the data-based intelligent factory applied by POSCO ICT. For product quality control, information is acquired using various camera systems, and data learning and modeling are possible based on the data collected through image classification. Also, quality evaluation is possible through predictive models [6].

Table 2. Smart Factory System in POSCO

Sensing	Analyze	Control
with IoT, Connected	Big Data, Data-driven	AI, Intelligent

Many of the automation devices used in data-based intelligent factories rely on data abnormalities to determine whether accidents and errors occur, and control through AR/VR is also used. It is very important to detect anomalies in a data-based intelligent factory and perform predictive maintenance before equipment failure [6]. Anomaly detection can be applied as a method for preliminary maintenance, and many companies are implementing outlier detection based on data. Data outlier detection is to find anomalies by recognizing anomalies in the data flow. Especially when time series data are significant [2-6]. As for existing domestic research for data outlier detection, S University and Company T have jointly developed a monitoring system for real-time preprocessing of production process data and detection of outliers. In order to do this, a preliminary analysis is being conducted by analyzing the data collected by the bearing vibration sensor [2-6].

2.2. LSTM AE

RNN is an algorithm to compensate for the disadvantage that existing deep neural network algorithms cannot think continuously, and it repeats itself so that previous information is reflected in the current information prediction [8]. LSTM is an algorithm for solving the shortcomings of long-term memory retention in RNNs, and learning that requires a long dependency period is possible. LSTM has the same chain structure as RNN, but each iteration module has multiple layers and exchanges information with each other to preserve and utilize memory [7-9]. LSTM (Long Short-Term Memory Network) is one of RNN (Recurrent Neural Network) types that deal with time series data among deep learning algorithms [7].

Figure 2 shows the process of LSTM. LSTM changes the data flow by giving various values to the cell state through several gates. Each gate has a sigmoid layer like RNN, and this layer is called a forget gate layer [8-10]. In this step, information about previous data is received and processed. After that, through the input gate, using the sigmoid layer like the previous gates, it determines whether to update through new

information coming in the future. A new cell state is formed through the input gate. Finally, the filter value is set through the cell state through the forget gate and the output result is determined [9,10-12,15].

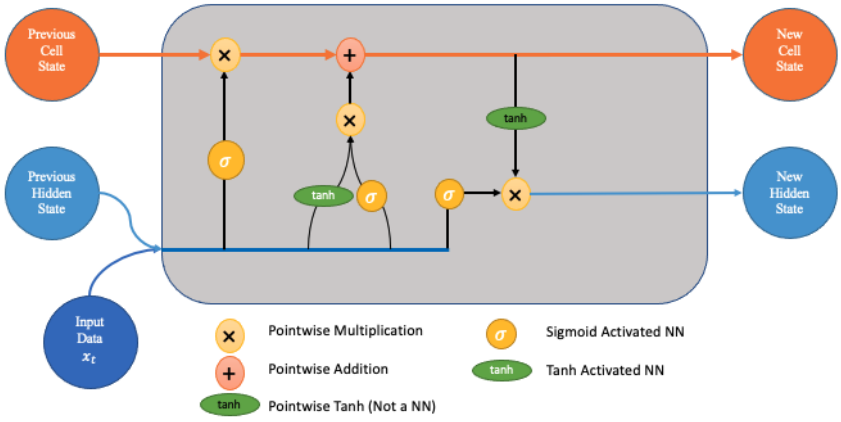

Figure 2. LSTM Process Diagram

3. Abnormal data processing process with LSTM AE

3.1 Modeling & Research Method

In this paper, anomaly data pre-processing and processing fixation of time series data mainly used in data-based intelligent factories were studied through LSTM AE.

The processing process proposed in this paper is as follows. First, in order to collect normal data used in the general manufacturing industry, normal temperature data for a certain period of time used by the Y institution, which is involved in the combined heat treatment business, was collected. Thereafter, data pre-processing is performed to detect and process abnormal data. In the process, hyper parameters are created to be used for data modeling and training, and optimization. By adding intermittent abnormal data to 20% of normal data, missing values are treated. Through time series analysis when processing missing value, if ite noise occurs more frequently than the allowable value within a significant period, an error is checked from the first occurrence to the y-t period. At this time, hyper parameter weight (weight for trust) and lim (allowable value) are generated. The weight (weight for trust) is the tolerance range between normal and outlier using Auto-Encoder and is set to 1 if the reliability is 95% or more, 2 if it is 90%, and 3 otherwise. It can be changed according to the size of the data and the importance of the task. For lim (allowable value), the data update period is calculated as 1 considering the characteristics of the data, and the |re-update period (value between 0 and 1)-1]*weight is calculated. After that, normal data and abnormal data are split into the test, train, and valid data, and algorithms through LSTM are applied to normal data, and then new allowable values and hyper parameter changes are saved.

After applying the algorithm through LSTM to abnormal data, the new allowable value is saved and the hyper parameter changes are saved. After that, when the LSTM cell state is updated, the detection noise

data and DC data (don't care data) for detecting new outliers are separated and removed and taken care of.

In LSTM, the weight of the previous section is used and the following algorithm is used to determine the cell state of the forget layer.

Table 3. Data Preprocessing

c(t) = f(t) * c(t-1) + tanh(f(t)) + weight * c(t-1) + b // b:bias
if c(t) > Threshold : lim=rand()/
else : lim=1

Finally, for the accuracy evaluation, the accuracy and harmonic average of the accuracy after application of the AE and LSTM algorithms are performed and measured.

The conceptual diagram of the proposed processing process is as follow Figure 3.

- Data collection and abnormal data generation
- Data preprocessing and missing value processing
- Hyper parameter creation through the process of 2)
- Weight : Normal value, outlier value, and tolerance range using Auto-Encoer.
- lim (permissible value): Considering the characteristics of the data, the data update period is calculated as 1 and | Renewal period (value between 0 and 1) -1] * weight
- Split into the test, train, and valid data for normal and abnormal data
- Algorithm application through LSTM for normal data > Saving new allowable values and updating hyper parameter values
- Algorithm application through LSTM for abnormal data > Save new allowable values and update s hyper parameter values - Detect whether there are new outliers Noise data and DC data (Don't Care Data) are separated and removed and taken action.
- Accuracy after accuracy evaluation, Harmonized averaging

Figure 3. Conceptual Diagram of Processing

The detailed process for hyper parameter creation among the above process is shown in below, Figure 4

Pre-process : {Normal data (a) -> Abnormal abnormal data (b) -> Compression -> Auto encoder -> Create new hyper parameter -> Restore}

↓

processing

↓

Hyper parameter and initial value-threshold settings to be used for modeling and optimization

Figure 4. Detailed Process for Hyper Parameter Creation

● Experiment and performance evaluation

In order to test the contents proposed in this study, 432,000 temperature-related normal data used in the general manufacturing industry (company Y) are collected. For the experiment, noise was applied to transform 5% of the data into abnormal data.

The environment for modeling and learning is as follows Figure 5.

- GeForce RTX 2080 GPU
- Python Keras 2.3
- Algorithm: LSTM AE
- Optimizer: Adam
- Activation Function: Relu
- Loss Function: MSE
- Batch Size: 32
- Epochs: Early Stopping Callback by Keras

Figure 5.Eenvironment for Modeling

As a normalization process for data pre-processing, data normalization (Robust) and Z-score are applied as processing processes to select high-efficiency data.

Table 4 is part of the code for normalization.

Table 4. Part of the Code for Normalization

```
…
scaler = preprocessing.RobustScaler()
robust_df = scaler.fit_transform(x)
robust_df = pd.DataFrame(robust_df, columns =['x1', 'x2'])
scaler = preprocessing.StandardScaler()
standard_df = scaler.fit_transform(x)
standard_df = pd.DataFrame(standard_df, columns =['x1', 'x2'])
…
```

The epoch was set to 1000, and a threshold was set and applied through PR_Curve for each epoch.

Experiments were conducted in the same data and computing environment, a is the case of using the general classification (KNN) algorithm, b is the case of applying the general LSTM, and c is the result of applying the LSTM AE after processing missing values by applying hyper parameters. The graph shows the results using the accuracy rate (d) and error rate (e) of prediction through F_Score as the standards for performance evaluation. Figure 6 shows evaluation result.

Figure 6. Evaluation Result

4. Discussion

For effective management of time series data used in various industrial sites such as data-based intelligent factories, LSTM AE is applied in this paper. It was proposed to enable the application of hyper parameters for the significance of each data section and optimization of modeling, and through this, it was possible to efficiently process outliers for abnormal data of time series data. For the performance evaluation of this study, the epoch was set to 1000 and the threshold value was set and applied through PR_Curve for each epoch. In the case of data significance application and hyper parameter application, the error rates were measured to be 5.4%/4.8%/3.3%, respectively. In the case of this study, there are limitations in calculating hyper parameters and measuring data significance using a rather limited dataset. In the future, we plan to continue research on the applicability of various data sets and coping with various event situations.

References

[1] Smart contract processing, 2020, Available: https://www.lgcns.com/blog/cns-tech/30841/
[2] Bini, S.A et al., "Artificial intelligence, machine learning and cognitive computing", The Journal of Arthroplasty, Vol.33, No.8, pp.2358-2361, 2018. DOI: 10.1016/j.arth.2018.02.067
[3] KDI International Information Center, " Overseas Trend of Smart Factory for 2021-04", Available: https://eiec.kdi.re.kr/reviewCallDownliad
[4] Lindsay et al., "A Novel Stochastic LSTM Model Inspired by Quantum Machine Learning", 2023 24th International Symposium on Quality Electronic Design, pp. 05-07, 2023 DOI:10.1109/ISQED57927.2023.10129344

[5] Haruna et al., "CNN-LSTM Learning Approach for Classification of Foliar Disease of Apple", 2023 1st International Conference on Advanced Innovations in Smart Cities, pp. 23-25, 2023. DOI:10.1109/ICAISC56366.2023.10085039

[6] Chung.S, Jeon. JY et al., "Standardization strategy of smart factory for improving sme's global competitiveness", Journal of Korea Technology Innovation Society, Vol.19, No.3, pp.545-571, 2018. Available:https://koreascience.kr/article/JAKO201610364778724.pdf

[7] Van Quan Nguyen et al., "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.2018.19.4.789039

[8] Wonjin Jang et al., "RNN-LSTM Based Soil Moisture Estimation Using Terra MODIS NDVI and LST", Journal of the Korean Society of Agricultural Engineers, Vol.61, No.6, pp. 123 – 132, 2019. DOI:10.5389/KSAE.2019.61.6.123

[9] Tae-Won Jung et al., "Traffic-based reinforcement learning with neural network algorithm in fog computing environment", The International Journal of Internet, Broadcasting and Communication, Vol.12, No.1, pp. 144-150, 2020. DOI: 10.7236/IJIBC.2020.12.1.144

[10] Shen, Peng et al., "Pronunciation-Aware Unique Character Encoding for RNN Transducer-Based Mandarin Speech Recognition", J022 IEEE Spoken Language Technology Workshop, pp. 09-12, 2023. DOI:10.1109/SLT54892.2023.10022528

[11] Donkol, A.A.E. et al., "Optimization of Intrusion Detection Using Likely Point PSO and Enhanced LSTM-RNN Hybrid Technique in Communication Networks", IEEE Access, Vol. 11, pp. 9469 - 9482, 2023. DOI: 10.1109/ACCESS.2023.3240109

[12] N. Par et al., "Time-step interleaved weight reuse for LSTM neural network computing", IEEE Int. Symp. on Low Power Electron, pp. 13-18, 2020. DOI: 10.1145/3370748.3406561

[13] Y. Guan, Z. Yuan, G. Sun, J. Cong, “FPGA-based accelerator for long short-term memory recurrent neural networks”, ASP-DAC, pp. 629-634, 2017. DOI: 10.1109/ASPDAC.2017.7858394

[14] LSTM Process Diagram, Available: https://towardsdatascience.com/lstm-networks-a-detailed-explanation-8fae6aefc7f9