Anomaly Detection of Environmental Sensor Data using Recurrent Neural Network at the Edge Device

JaeMyoung KIM
Intelligent Robotics Research Division
ETRI
Daejeon, Rep. of Korea
jaemkim@etri.re.kr

Young Wook Cho
CEO
Chironsoft Co., Ltd.
Daejeon, Rep. of Korea
ywcho@chironsoft.com

Do-hyun KIM
Intelligent Robotics Research Division
ETRI
Daejeon, Rep. of Korea
dohyun@etri.re.kr

***Abstract*—Advance in sensor technology brings numerous challenges with it in the context of data collection, storage and processing. Edge-enabled AI processing of sensor data is a large part of the sensor data processing. Sensor data about environments have natural errors and incompleteness in the collection process and need to be processed in real-time. Due to large volumes of data, monitoring and reporting of analyzed results need to be processed at the edge side of generated data. This paper proposes a time series data anomaly detection method that is based on neural network. Our models are evaluated using synthesis data generated from time series with trend, seasonal and noise component. In the case of zero-based dataset, GRU model with hidden cells (240 cells) and using only input values without additional features, applied with 0.3 dropout, produced 100% recall and 99.7% accuracy. In the case of non-zero-based dataset, LTSM model with hidden cells (240 cells) and using only input values without additional features, produced 86.7% recall and 99.5% accuracy. We also suggest the edge monitoring system with anomaly detection function of each environmental sensor using our pretrained detection model. Users can recognize an environmental status of the workplace using the prediction method with previous sensor outputs in real-time.**

Keywords*—*anomaly detection, environment sensor data, RNN based model, edge computing, environment monitoring system

I. Introduction

Due to the development of ICT technology, especially 5G communication technology, IoT technology is also applied to the industry, and it is developing into an application that automatically and accurately monitors the workplace environment in combination with artificial intelligence technology.

Advance in sensor technology brings numerous challenges with it in the context of data collection, storing and processing. Also, Edge-enabled AI processing of sensor data is one of the industry-oriented technology. Sensor data about environments have natural errors and incompleteness in the collection process and need to be processed in real-time. Due to large volumes of data, monitoring and reporting of analyzed results need to be processed at the edge side of generated data[1].

In HSE (Health, Safety and Environment) management, it is necessary to provide a method for checking and reporting the environmental conditions in real time about the sensor data installed in the workplace. Especially we proposed a workplace risk prediction method based on neural networks using environment sensor data in real-time and implemented monitoring and reporting system of workplace environmental status to users promptly[2].

When the trained model with the previously tested algorithm was applied in a real system, but there is no visible change of sensor values beyond the range of normal levels during testing periods. In order to solve the problems, there is a need for an anomaly detection to handle work area risks immediately with the change in sensor values[3].

We check the normality of sensor data in the edge system which collects sensor data. If the checked data is normal, the workplace environmental risk prediction is executed, and if it is abnormal, the situation is notified to the user. By predicting workplace environmental risks using normal sensor data, more robust workplace environmental monitoring and notification will be possible.

In this paper, we provide a method to check and notify the abnormal condition by machine learning in real-time of sensor data installed in the workplace about the abnormal condition of the worker environment. We create some synthetic dataset for learning the real-time data anomaly detection algorithms and test seven models based on GRU(Gated Recurrent Units) and LSTM(Long Short-Term Memory) for predicting anomalies in time series data.

In hydrogen sulfide(H_2S) gas sensor data of the zero-based dataset, which has less than 0.1 ppm normal value, the model with 240 hidden cells and 0.3 dropout based on GRU showed the best performance with 100% reproducibility and 99.7% accuracy. In oxygen(O_2) gas sensor data of the non-zero-based dataset, which has the normal range of 19.5~23.5%, the model with 240 hidden cells based on LSTM showed the best performance with a reproducibility of 86.7% and an accuracy of 99.5%.

We explain the basic concept and method of anomaly detection and describe how to create dataset, how to construct models for experiment, how to detect anomalies using neural networks, and how to evaluate suggested models, and finally summarize the experimental results. As our future work, we suggest an environmental gas detection monitoring system using the edge system.

We can verify our concept and models in real environment edge monitoring system with anomaly detection function of each environmental sensor using our pretrained detection model. Through this system, we will be able to confirm that the proposed anomaly detection method works well and easily determines the abnormality of the sensor data and accurately predicts the risk of the workplace environment. Users can recognize an environmental status of the workplace using the prediction method with previous sensor outputs in real-time.

II. Anomaly Detection Methods

A. Definition of anomaly detection

Anomalies are also referred to as abnormalities, deviants, or outliers in the data mining and statistics literature. Anomaly detection is the identification of rare items, events or observations which raise suspicions by differing significantly

978-1-7281-6758-9/20/$31.00 ©2020 IEEE

from the majority of the data. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions[4,5].

B. *Methods of anomaly detection*

The anomaly detection algorithm for sensor time series data can be classified as follows [6,7].

- Statistical methods

These methods use past measurement data to approximate a model of the correct behavior of a sensor. Whenever a new measurement data is registered, it is compared to the model data. If the results are not statistically significant, the new measured data is marked as an anomaly. An example of a very common statistical anomaly detection method is the low-high pass filter.

- Probabilistic method

Classification of anomalies is performed by measuring the probability of measured data with respect to the model. If the probability falls below a predefined threshold, then it is labelled as an anomalous event. Methods based on probability distribution are typically very computationally expensive and do not scale well. So anomaly detection with streaming data rarely uses these methods.

- Proximity-based method

These methods rely on distances between measured data to distinguish between anomalous and correct data. A very famous proximity-based algorithm is the Local Outlier Factor (LOF). These method include the cluster-based methods. The measured data are first used to create clusters and then are labelled as anomalous if new measurement that are assigned to small and isolated clusters that are far from their cluster's centroid.

- Prediction-based method

These methods use past measured data to train a model that can predict the value of the next measurement in the sensors' time series data. If the actual measured data are too different from the predicted one, then it is labelled as anomalous. There are many prediction-based algorithms for anomaly detection. There are some methods based on very simple machine learning models, such as 1-class SVMs, while there are more complex neural networks. such as LSTM or GRU cells.

Anomaly detection methods are used in various application such as petroleum industry applications[8] and are tried in various ways about streams of sensor data such as unsupervised online detection and prediction, supervised and unsupervised combined hybrid detection[9,10].

Our environmental sensor data have time series and real-time properties and anomaly events of these data will rarely happen. Other methods except prediction-based methods can analyze these data deeply, but are difficult in real-time processing. In this paper, we are using prediction-based anomaly detection methods and prefer the recurrent neural network(RNN) based prediction methods to other algorithm based methods due to above data properties. By predicting whether the measured data is anomaly data to occur in the future, up-coming status can be checked in real time and it can be provided to the user who is in charge of the event.

III. Anomaly Detection Models and Evaluations

A. *Creating datasets*

We synthesize the time series data using random seasonal, trend and noise components in order to generate a value as close as possible to the actual sensor data. Datasets have two types of data according to the safe state value of gases which is generated using gas detect devices.

The first type is data having a normal range of 19.5 to 23.5%, such as oxygen(O_2) and the other type is data having a normal range of 0.1 ppm or less, such as hydrogen sulfide (H_2S). The former is called a non-zero valued dataset(Non-Zero_AD), and the latter is called a zero valued dataset(Zero_ AD).

Each dataset has 14,400 training data, which is measured in minutes and corresponds to 10 days. In addition, the test data in each dataset has 6,171, which is 30% of the training data. This is to inject abnormal data to measure performance. The number of abnormal data is 30, which is 0.5% of the number of test data.

1) *Non-zero valued dataset (Non-Zero_AD)*

Non-Zero_AD is a data set with the following element values for training data and test data. Table I summarizes the characteristics.

- Training Data:
  - seasonal factor : level = 0.2, length = 720
  - trend factor : level = 0.3, number = 30
  - noise factor : normal distribution with mean = 20 and standard deviation = 0.3
- Test Data:
  - seasonal factor : level = 0.2, length = 360
  - trend factor : level = 0.3, number = 15
  - noise factor : normal distribution with mean = 20 and standard deviation = 0.3

TABLE I. Characteristics of Dataset

Parameter	Training data	Test data
No of data	14,400	6,171
mean	20.860	2.065
Standard deviation	0.589	0.408
Min value	18.929	18.680
25%	20.424	19.916
50%	20.835	20.180
75%	21.259	20.466
Max value	22.937	22.050
No of anomaly data	0	30

2) *Zero valued dataset (Zero_AD)*

Zero_AD is a data set with the following element values for training data and test data. Table II summarizes the characteristics.

- Training Data:
  - seasonal factor : level = 0.03, length = 720
  - trend factor : level = 0.05, number = 30
  - noise factor : normal distribution with mean = 2.0 and standard deviation = 0.05

- Test Data:
  - seasonal factor : level = 0.03, length = 360
  - trend factor : level = 0.05, number = 15
  - noise factor : normal distribution with mean = 2.0 and standard deviation = 0.05

TABLE II. CHARACTERISTICS OF DATASET

Parameter	Training data	Test data
No of data	14,400	6,171
mean	2.131	2.065
Standard deviation	0.074	0.063
Min value	1.861	1.825
25%	2.080	2.023
50%	2.129	2.062
75%	2.182	2.106
Max value	2.411	2.447
No of anomaly data	0	30

B. Creating Models

We evaluate 14 models. We use RNNs(recurrent neural network) which have GRU(Gated Recurrent Unit) and LSTM(Long-Short Term Memory) hidden cells. Those cells are more effective than vanilla RNN hidden cells. We also use various hyperparameters which are the number of hidden cells, additional features, autoencoder, and dropout. The additional features are moving averages and data with lag time of 1.

We set time steps to 30 because we think the window size would be enough to predict next value. The dropout was processed by a dropout layer without using the dropout option given to the RNN hidden cell.

TABLE III. GRU BASED MODELS

Model	parameters
Model 1	GRU with 50 hidden cells
Model 2	GRU with 240 hidden cells
Model 3	GRU with 240 hidden cells + Dropout(p=0.3)
Model 4	GRU with 240 hidden cells + additional features
Model 5	GRU with 240 hidden cells + additional features + Dropout(p=0.3)
Model 6	AutoEncoder(32-16-1) + GRU with 50 hidden cells + additional features
Model 7	AutoEncoder(32-16-1) + GRU with 50 hidden cells + additional features + dropout(p=0.3)

TABLE IV. LSTM BASED MODELS

Model	parameters
Model 1	LSTM with 50 hidden cells
Model 2	LSTM with 240 hidden cells
Model 3	LSTM with 240 hidden cells + Dropout(p=0.3)
Model 4	LSTM with 240 hidden cells + additional features
Model 5	LSTM with 240 hidden cells + additional features + Dropout(p=0.3)
Model 6	AutoEncoder(32-16-1) + LSTM with 50 hidden cells + additional features
Model 7	AutoEncoder(32-16-1) + LSTM with 50 hidden cells + additional features + dropout(p=0.3)

C. How to evaluate Models

An outlier is a data point which differs significantly from other sensor data. Statistically, we define outliers which any data point less than the Lower Bound or more than the Upper Bound is considered as an outlier as given below:

Lower Bound: (Q1 - 1.5 * IQR)
Upper Bound: (Q3 + 1.5 * IQR)

where Q1 is the first quartile of the data, i.e., to say 25% of the data lies between minimum and Q1, Q3 is the third quartile of the data, i.e., to say 75% of the data lies between minimum and Q3, and IQR = Q3 - Q1[11].

We used the following method to detect anomaly data. First, a variety of defined models are trained using the training data created above. Predictions are made on the test data using the trained model. Next, when the actual sensor data value exceeds the above range of value predicted by the model, it is determined as anomaly.

We will choose the best model from confusion matrix which is calculated as follows: Find the confusion matrix for the given test data which is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known and then calculate accuracy, True Positive rate which called recall, and precision with the value of the obtained confusion matrix[12].

In the case of a hazardous gas measurement, a false positive, which is predicted to be dangerous when the actual data value is not dangerous, is accepted as a warning and no action is required. However, a false negative, which is predicted to be dangerous when the actual data value is not dangerous, is serious. Therefore, a model that minimizes the FN value is a good model.

IV. RESULTS OF EXPERIMENTS

A. Non-Zero_AD case

TABLE V. GRU BASED MODELS CONFUSION MATRIX

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Hidden cells	50	8X30	8X30	8X30	8X30	50	50
AutoEncoder	0	0	0	0	0	32x16	32x16
dropout	0	0	0.3	0	0.3	0	0.3
Feature #	1	1	1	3	3	3	3
TP	25	25	25	24	22	25	23
TN	6,057	6,057	6,054	6,055	6,030	6,058	6,048
FP	53	53	56	55	80	52	62
FN	5	5	5	6	8	5	7
recall	0.833	0.833	0.833	0.800	0.733	0.833	0.767
precision	0.331	0.321	0.309	0.304	0.216	0.325	0.271
accuracy	0.995	0.991	0.994	0.994	0.989	0.995	0.992

In the case of oxygen saturation synthesis data, Model 1 and 6 showed the almost same performance in terms of recall and accuracy in the GRU-based model, while Model 1 showed the highest performance in precision. This is the model learned from simple base GRU model.

TABLE VI. LSTM BASED MODELS CONFUSION MATRIX

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Hidden cells	50	8X30	8X30	8X30	8X30	50	50
AutoEncoder	0	0	0	0	0	32x16	32x16
dropout	0	0	0.3	0	0.3	0	0.3
Feature #	1	1	1	3	3	3	3
TP	25	26	25	25	19	23	26
TN	6057	6,057	6,058	6,056	6,017	6,047	6,054
FP	53	53	52	54	93	63	56
FN	5	4	5	5	11	7	4
recall	0.833	0.867	0.833	0.833	0.633	0.767	0.867
precision	0.321	0.329	0.325	0.316	0.170	0.267	0.317
accuracy	0.995	0.995	0.995	0.994	0.986	0.992	0.994

In LSTM-based models, Model 2 and 7 were the best in terms of recall, but Model 2 showed better performance than Model 7 in terms of accuracy and precision. In comparison between the GRU-based model and the LSTM-based model, the LSTM Model 2 showed high performance, and in the case of the LSTM-based model, it can be seen that a simple input with sufficient hidden cells shows good performance.

B. *Zero_AD case*

TABLE VII. GRU BASED MODELS CONFUSION MATRIX

Model	1	2	3	4	5	6	7
Hidden cells	50	8X30	8X30	8X30	8X30	50	50
AutoEncoder	0	0	0	0	0	32x16	32x16
dropout	0	0	0.3	0	0.3	0	0.3
Feature #	1	1	1	3	3	3	3
TP	30	30	30	30	29	30	30
TN	6057	6,057	6,059	6,044	5,979	6,052	6,024
FP	53	53	51	66	131	58	86
FN	0	0	0	0	1	0	0
recall	1.000	1.000	1.000	1.000	0.967	1.000	1.000
precision	0.361	0.361	0.370	0.313	0.181	0.341	0.259
accuracy	0.991	0.991	0.992	0.989	0.979	0.991	0.986

The anomaly detection result of the hydrogen sulfide synthesis dataset showed high recall (1.0) and high accuracy (0.99) in all models, but Model 3 had the highest precision in GRU-based models.

TABLE VIII. LSTM BASED MODELS CONFUSION MATRIX

Model	1	2	3	4	5	6	7
Hidden cells	50	8X30	8X30	8X30	8X30	50	50
AutoEncoder	0	0	0	0	0	32x16	32x16
dropout	0	0	0.3	0	0.3	0	0.3
Feature #	1	1	1	3	3	3	3
TP	30	30	30	30	29	30	30
TN	6057	6,057	6,057	6,044	5,948	6,053	6,054
FP	53	53	53	66	162	57	56
FN	0	0	0	0	1	0	0
recall	1.000	1.000	1.000	1.000	0.967	1.000	1.000
precision	0.361	0.361	0.361	0.313	0.152	0.345	0.349
accuracy	0.996	0.996	0.996	0.994	0.978	0.996	0.996

In LSTM-based models, Model 1, 2 and 3 showed the same and best recall, precision and accuracy. These models are almost same performance of GRU based models. But FP value in GRU based model is lower than one of LSTM based model. So, Model 3 in GRI based model is the best.

C. *Choosing Models*

In this apaper, several neural network models and synthetic data sets similar to the real data patterns for anomaly were generated and analyzed into a confusion matrix for performance measures.

As a result, in the case of Non-Zero_AD, which has an oxygen saturation range of 19.5 to 23.5%, the model with 240 hidden cells and additional features based on LSTM showed the best performance with 99.5% recall and 86.7% accuracy.

In the case of Zero_AD, which synthesized hydrogen sulfide sensor values with less than 0.1 ppm, the model with 240 hidden cells and 0.3 dropout based on GRU showed the best performance with 100% call and 99.7% accuracy.

The GRU-based and LSTM-based models have similar performance in the simple parameters. Therefore, it is better to select a model with hidden cells adequately, apply the trained model through the generated training data, and improve the model by reflecting the actual operating data.

The anomalies happen rarely in the real environmental gas detect situation. So, it is possible to predict the risk condition of the workplace more realistically by introducing an anomaly detection algorithm.

V. FUTURE WORKS

We propose the environmental gas detection monitoring system in below figure. It will be implemented in edge devices. The sensing data enters the Anomaly Detection module at the edge device. The Anomaly Detection module determines whether the given data is an anomaly based on the trained model. If the given data is in anomaly state, the user is notified to take action. If the given data is normal, it enters the Environment Status Prediction module and sends the results of the workplace environment risk prediction based on the trained model to the user.

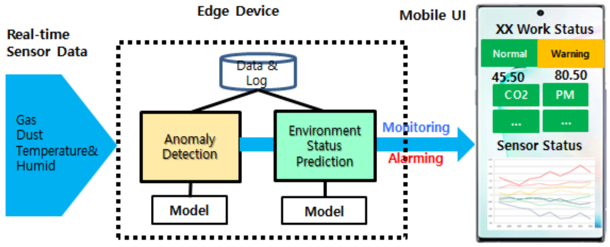

Fig. 1. Configuration of the monitoring system

The reason for deploying the system on the edge device is not only that it can quickly detect and take action on dangerous gas conditions, but also that it provides more robust prediction results by using the Environment Status Prediction module with verified data.

Also the reason why such a system is necessary may be that it is not suitable in a real environment because the data used for learning is synthetic data. Therefore, it is necessary to continuously improve the prediction model by reflecting the current sensor data situation.

It is considered that the prediction results can be inferred on the edge device due to the relatively small trained model. The predicted monitoring result of the workplace risk condition can be provided to the user promptly and easily.

ACKNOWLEDGMENT

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [20ZS1200, Fundamental Technology Research for Human-Centric Autonomous Intelligent Systems]

REFERENCES

[1] Charu C. Aggarwal (editor), Managing and Mining Sensor Data, Springer, 2013.

[2] JaeMyoung Kim, Young Wook Cho and Byung-Tae Jang, "Sensor Data based Work Area Environment Risk Prediction and Monitoring in the Shipbuilding," Proceedings of The 10th Int'l Conference on ICT Convergence, ICTC 2019, pp.906-909.

[3] Young Wook Cho(2019), "A Study on Real-time Data based Risk Prediction and Anomaly Detection Methods in Large Scale Work Areas," Doctoral dissertation, Sunmoon University, Asan, Rep. of Korea.

[4] Wikipedia, "Anomaly detection," https://en.wikipedia.org/wiki/Anomaly_detection.

[5] Raghavendra Chalapathy and Sanjay Chawla, "Deep Learning for Anomaly Detection: A Survey," arXiv:1901.03407, January 24, 2019.

[6] Federico Giannoni, Marco Mancini, and Federico Marinelli, "Anomaly Detection Models for IoT Time Series Data," arXiv:1812.00890, 30 Nov 2018.

[7] Varun Chandola, Arindam Banerjee and Vipin Kumar, "Anomaly Detection : A Survey," ACM Computing Surveys, Volume 41 Issue 3, July 2009.

[8] Luis Martí et al, "Anomaly Detection Based on Sensor Data in Petroleum Industry Applications," Sensors 2015, 15, 2774-2797.

[9] Niko Reunanen et al, "Unsupervised online detection and prediction of outliers in streams of sensor data," International Journal of Data Science and Analytics volume 9, pages 285–314(2020).

[10] Mohsin Munir et al, "FuseAD: Unsupervised Anomaly Detection in Streaming Sensors Data by Fusing Statistical and Deep Learning Models," Sensors 2019, 19(11).

[11] Shivam Chaudhary, "Why "1.5" in IQR Method of Outlier Detection?" Sep 28, 2019, https://medium.com/.

[12] Data school, Simple guide to confusion matrix terminology, https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/, March 25, 2014.