Reliability Evaluation Method of Predictive Results in Deep Learning

YUSUKE YAMASHINA[*1] KENJI TAKAO[*2]
NOBUHARU IWASHITA[*2] TETSUYA KIZU[*3]
NOZOMU HAMASAKI[*4]

As one of the techniques for anomaly detection and future forecasts utilizing time-series operating data, there have been a growing number of applications of long short-term memory (LSTM), one form of deep learning. LSTM is expected to demonstrate high prediction accuracy when actual operating conditions are close to those simulated in training. If they are not, however, there is a possibility that the prediction accuracy could decrease significantly. Meanwhile, since there has been no means to know whether the predictive results are reliable in conventional LSTM, it has not been applied much to fields where a high level of reliability would be required such as plant operation control.

This report will describe the predictive results produced by a deep learning model utilizing LSTM developed jointly with Kyushu University, as well as a technique to evaluate its reliability.

1. Introduction

Machine learning technology, specifically deep learning, has advanced to the extent that prediction and judgement can be made even in the kinds of things that “humans cannot make out as a pattern with clarity”, if there is a large enough amount of training data. In terms of the application of deep learning, however, unless a large set of “appropriate” training data is available, it ends up producing erroneous results. There have been several cases reported such as accidents due to false recognition in autonomous driving(1).

LSTM, a deep learning model, is a Recurrent Neural Network (RNN) with a recurrent mechanism, which as shown in **Figure 1** consists of LSTM cells including neurons and their links, where the product of each cell’s output and the weight of its links is added in the order of the cell’s distance to the input to obtain the final value. The weights of individual links and those in LSTM cells are optimized at the time of training.

In conventional LSTM training, however, the weights are optimized so that the difference between the correct value and predictive one should be minimized, and not all LSTM cells are used. In other words, there is a possibility that some of the weights are left unoptimized at the time of training. Therefore, depending on the state of neural firing, it is possible that neurons which never fired during training could fire in actual operation, resulting in various adverse effects such as veering far from the predictive results (See **Figure 2** which describes LSTM cells as neurons).

On the other hand, in terms of deep learning in the image processing field, there have been several reports concerning the reliability evaluation of models utilizing the neuron coverage (the number of neurons activated by a set of test inputs) as an index, as well as some research findings where the prediction accuracy increased by generating a set of test data which would maximize the coverage and using it to retrain the model.

This report features a form of deep learning utilizing time-series data (LSTM) and describes the verification results of a boiler’s validity as an example by (1) developing a new evaluation index (coverage) suitable for LSTM, (2) creating a highly-robust model based on the development

*1 CIS Department, ICT Solution Headquarters, Mitsubishi Heavy Industries, Ltd.
*2 Chief Staff Manager, CIS Department, ICT Solution Headquarters, Mitsubishi Heavy Industries, Ltd.
*3 Chief Staff Manager, Boiler Engineering Department, Boiler Technology Integration Division, Mitsubishi Power. Ltd.
*4 Boiler Engineering Department, Boiler Technology Integration Division, Mitsubishi Power. Ltd.

of an evaluation function utilizing the coverage index, and (3) developing a method of evaluating the reliability of the predictive results at the time of actual operation.

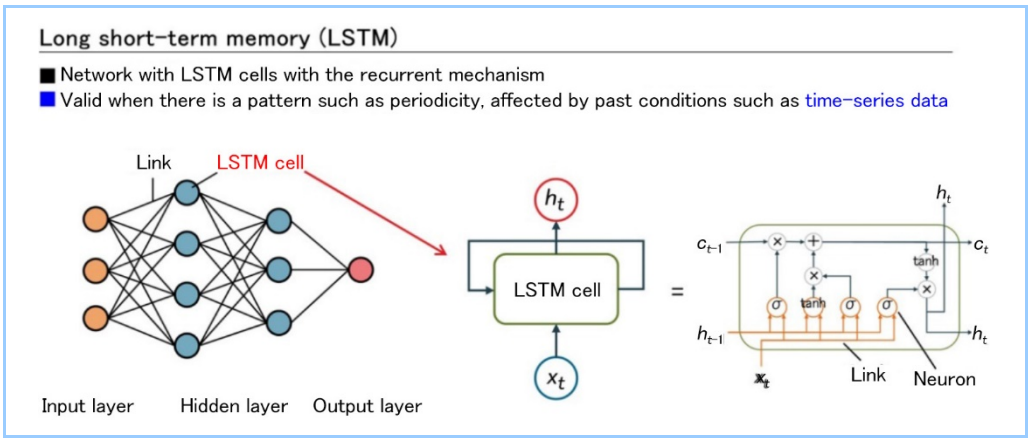

Figure 1 LSTM model overview

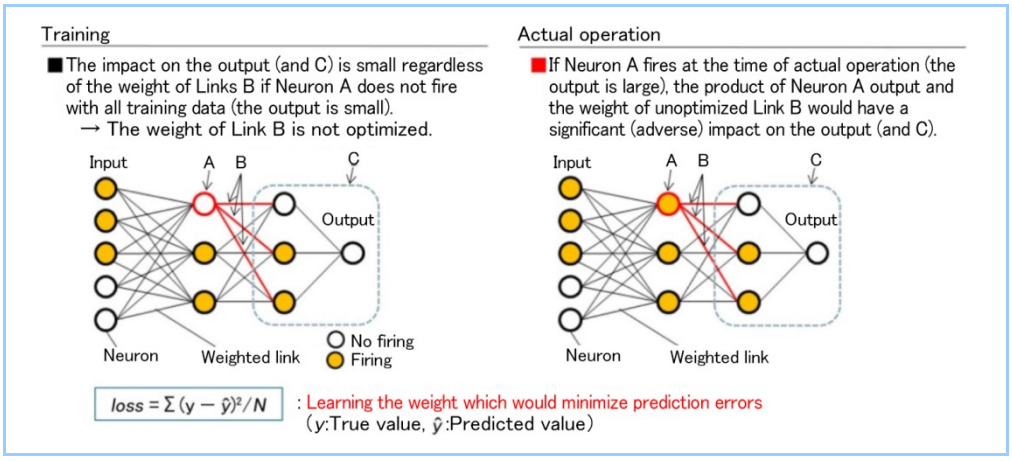

Figure 2 Challenges in the training in deep learning methods

2. Reliability evaluation method in deep learning

2.1 Evaluation index for a deep learning model with time-series data

LSTM has, unlike a regular neural network, a recurrent layer which is a recurrent structure. In this recurrent structure, there are 2 internal states (Neuron coverage h_t and Cell state coverage c_t as shown in **Figure 3**. h_t is a value propagated to the next layer whereas c_t is a value propagated in the direction of time on the same layer). Since these 2 types of values affect the neurons firing, we used them as the indices for the model's reliability (to see whether the individual weights are optimized) by evaluating them quantitatively as coverages.

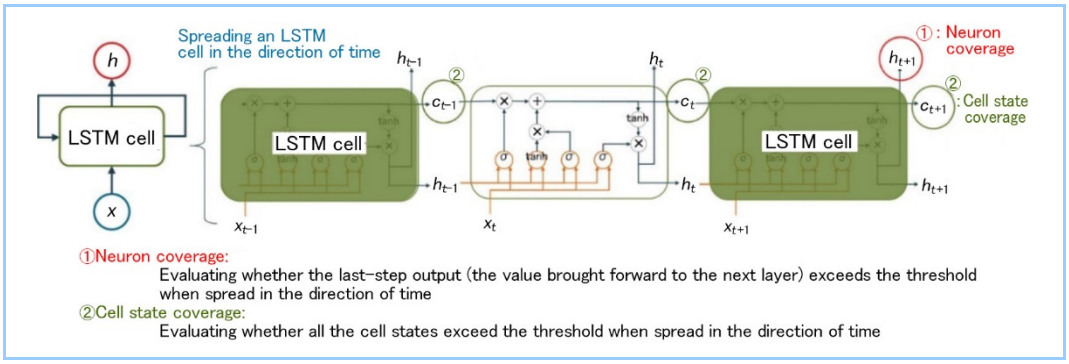

Figure 3 Evaluation index for a deep learning model with time-series data

2.2 Development of an evaluation function considering the coverage index

When creating a neural network model, we normally make the model learn weights so that it would minimize the evaluation function consisting of the error between actual measurement value

and predicted value (the Mean Squared Error (MSE) is commonly used for time-series data). As opposed to that, we have developed a new evaluation index which can factor in not only the error, but also the coverage by adding the evaluation index developed in the previous section as a coverage term to the loss function. This would allow no neurons to be left unoptimized at the time of training—preventing an adverse impact from firing during actual operation—and thereby stabilizing the level of accuracy.

The training is illustrated in **Figure 4**.

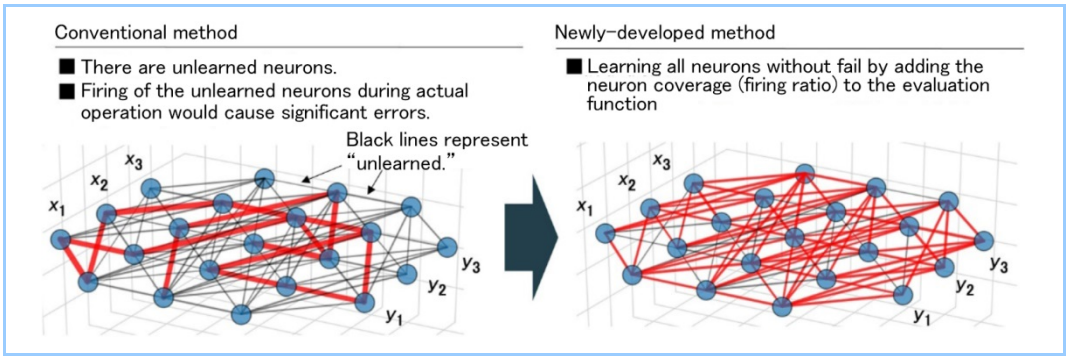

Figure 4 Training method considering the coverage

2.3 Reliability evaluation of predictive results based on the coverage index

Generally, machine-learning models including deep learning have a high possibility of the prediction accuracy decreasing significantly when the input is outside of the range the model has learned. Therefore, it would be too risky to utilize the predictive results for control purposes. Accordingly, we have developed a method for evaluating the reliability of predicted values as well. Specifically, since the coverage index allows us to determine whether the individual neurons have fired, as shown in **Figure 5**, we evaluate the data distribution at the time of actual operation quantitatively to see if it is the same as that at the time of training, in accordance with the difference in the firing patterns at the time of training and actual operation.

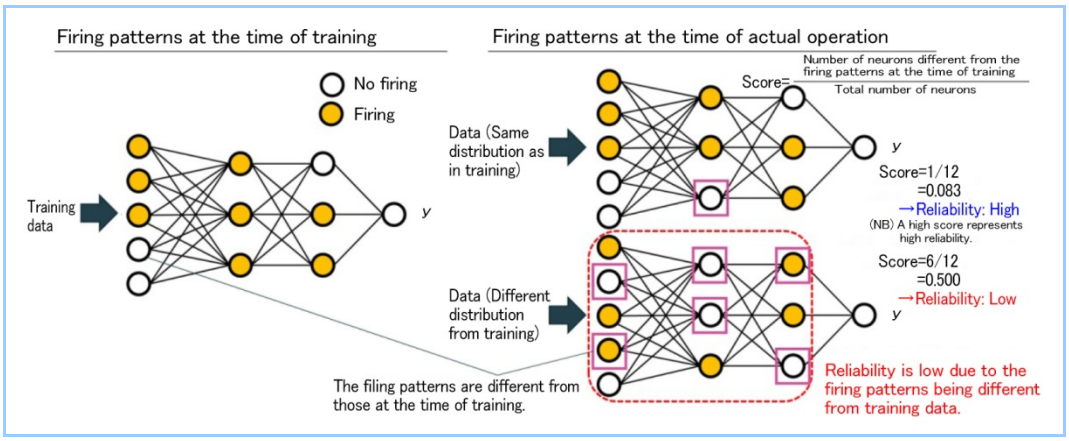

Figure 5 Reliability evaluation of the predictive results utilizing the coverage index

3. Validation of the newly-developed method with a boiler

There are multiple high-pressure soot blowers inside a boiler and there is a need to optimize their operating intervals. The restriction to the optimal interval is the heat exchanger's outlet temperature, which needs to be within the threshold range. Therefore, we aim to determine the optimal operating interval by predicting the future outlet temperature utilizing our new deep learning method. From an operational planning aspect, the outlet gas temperature needs to be predicted 2 days ahead, and the activation of soot blowers is scheduled based on the predicted value. Therefore, it is necessary to evaluate the reliability of the predicted value before the true value is available. Furthermore, as soot blowers are greatly affected by the fuel properties, the prediction needs to be both highly accurate and highly reliable.

We validated this method utilizing operating data from 3 different periods as shown in

Figure 6. The validating conditions include a total of 11 input variables consisting of the boiler inlet temperature, outlet temperature and accumulated down-time (reset to 0 when restarting) of 9 high-pressure soot blowers (SB). The objective variable is set as the outlet temperature 2 days later and the data sampling time is 10 minutes. This deep learning model is structured to have one LSTM layer as the hidden layer (50 units). One of the 3 data sets in Figure 6 is designated as training data and other 2 are validation data.

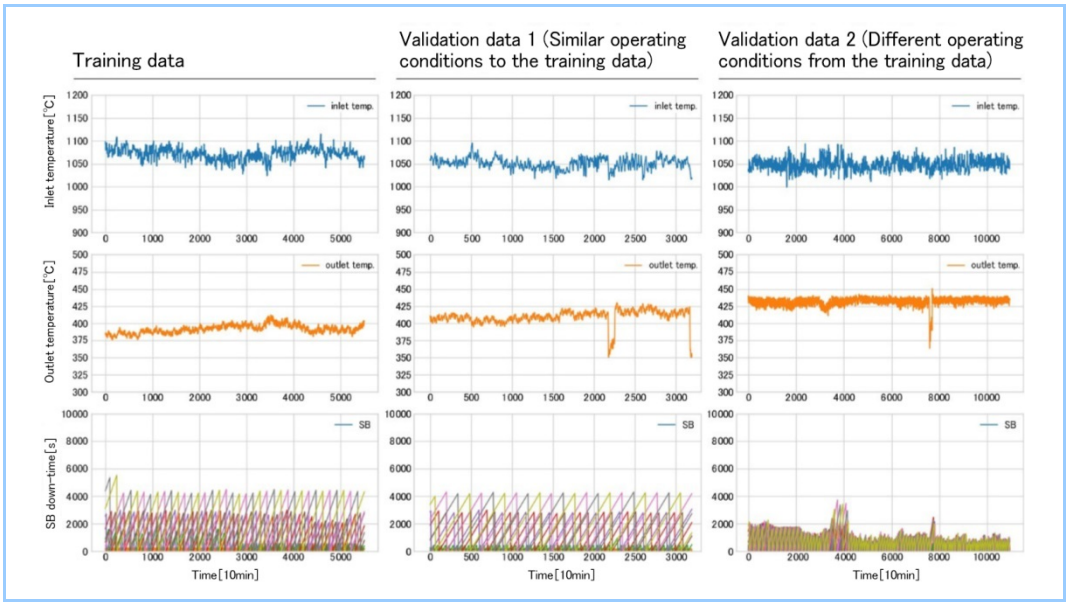

Figure 6 Actual measurement data of boiler gas temperature

Firstly, **Figure 7** shows the training results in accordance with the conventional method. The model is constructed to match the training data with accuracy. The neuron coverage is 0.54, however, and from this it can be seen that many neurons do not fire. Furthermore, in terms of the operation data close to the training data (Validation data 1), the coverage increases, which indicates that the neurons that did not fire at the time of training do fire. Therefore, we can confirm that the Root Mean Square Error (RMSE) increases in some parts. With respect to the operation data outside the training range (Validation data 2 where the outlet temperature is higher than that at the time of training), the coverage increases even further, meaning a significant decrease in the prediction accuracy. Accordingly, we have confirmed that, due to the weights unoptimized at the time of training, the accuracy decreases at the time of actual operation.

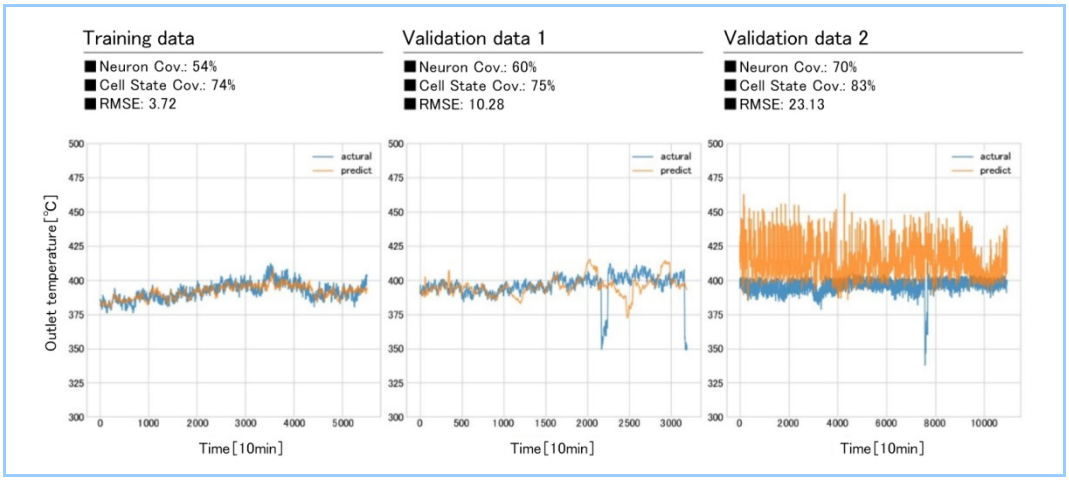

Figure 7 Coverage evaluation result

Next, the results of training and prediction with the evaluation function in light of the coverage are shown in **Figure 8**, from which it can be seen that the neuron coverage at the time of training is 1, meaning the weights of all the links are optimized.

Furthermore, the prediction accuracy increases both in Validation data 1 and 2. Especially

with Validation data 2, we have successfully obtained results showing the RMSE reduced by 28% from 23.13°C to 6.55°C, which means that the level of accuracy is stabilized.

Lastly, we will discuss the reliability of the predictive results. **Figure 9** shows the reliability of the predictive results with Validation data 2, where it can be confirmed that when the prediction accuracy decreases, so does the reliability. Accordingly, we have confirmed that the reliability evaluation would inform us beforehand of the situation where the RMSE increases due to the input of data outside the range covered by the pre-trained model.

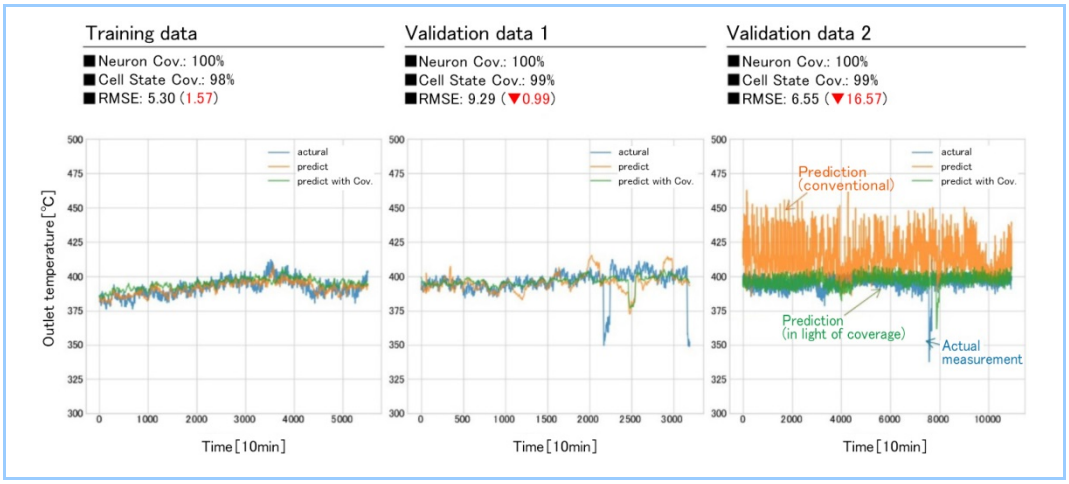

Figure 8 Training results in light of the coverage

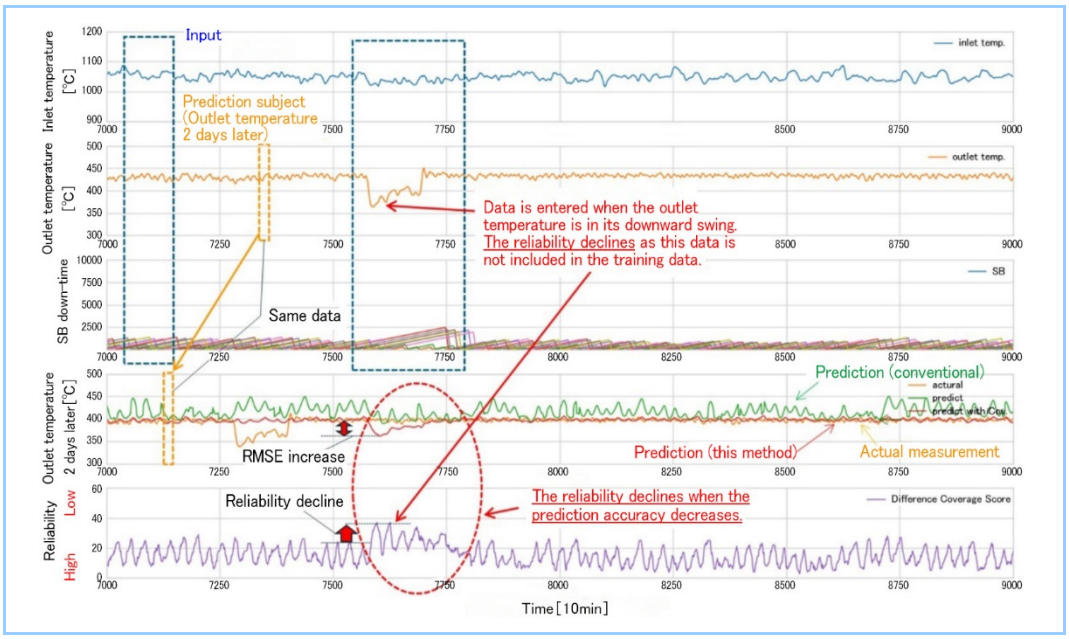

Figure 9 Reliability evaluation of the predictive results using the coverage index

4. Conclusion

This report has described a method of evaluating pre-trained models utilizing the coverage index, as well as a method of evaluating the reliability of the predictive results, with respect to the deep learning model covering time-series data. These methods would allow the reliability evaluation of pre-trained models prior to actual operation. Furthermore, the reliability of the predictive results according to the input data could be evaluated during actual operation as well, which would allow practical application of deep learning to fields such as control, where we can provide high-performance products and services utilizing deep learning as a business.

In the future, we intend to expand the application of this technology to validation in our customers' thermal power plants, remaining life prediction of the spare parts of our overall products, etc.

References

(1) Tian, et al., DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars, ICSE, 2018
(2) Hochreiter, et al., LONG SHORT-TERM MEMORY, Neural Computation, Vol.9, No.8, p.1735-1780, 1997
(3) L.Ma, et al,. DeepGauge: Multi-Granularity Testing Criteria for Deep Learning Systems, ASE, 2018
(4) L.Ma, et al., DeepMutation: Mutation Testing of Deep Learning Systems, ISSRE, 2018