Interpretability of Deep Neural Networks for Abnormal Situations Diagnosis in Nuclear Power Plants: A Rule Extraction Technique

Ji Hun Park, Sang Won Oh, Min Seon Kim, Man Gyun Na*
Department of Nuclear Engr., Chosun Univ., 10 Chosundae 1-gil, Dong-gu, Gwangju, Republic of Korea, 61452
**Corresponding author: magyna@chosun.ac.kr*

1. Introduction

Nuclear Power Plants (NPPs) have Abnormal Operating Procedures (AOPs) to prepare for abnormal operating situations. When abnormal operating situations occur, the operators select the prepared AOPs and perform the mitigation actions to return to normal operating conditions. However, there are more than 200 different AOPs, and the operators have to check the symptom requirements listed in each AOP in order to select one of them. This means selecting the optimal AOP that satisfies the symptom requirements while checking for rapidly changing NPP variables in abnormal operating situations. In addition, the operators have to perform this process quickly and accurately in order to quickly return to normal operating conditions. This process can also increase the possibility of human error due to the psychological stress on the operators.

To reduce the possibility of the human errors, studies are being conducted using Artificial Intelligence (AI). For example, if the AI can recommend the optimal AOP to the operators, it can reduce the possibility of human error. However, it is unclear when this AI-based optimal AOP recommendation system will be applied to NPPs. This is related to the problem known as the black box characteristic of AI. This refers to situations where the AI does not provide any reason for judgment about the results it derives. The stakes are too high for the safety-critical NPPs to accept the results of the AI without any reason for judgment.

In this paper, we apply the Efficient CLAuse-wIse Rule Extraction (ECLAIRE) method, a rule extraction technique, to address the black box characteristic of AI. The result of the rule extraction technique comes in the form of "IF… THEN". It is a rule-based system, which is one of the easier forms for humans to understand. To apply the rule extraction technique, we utilize the Deep Neural Network (DNN) method to develop a model that can recommend the optimal AOP. In addition, the data required to develop the DNN model were collected using the Compact Nuclear Simulator (CNS). The developed DNN model is derived in the form of a rule-based system by applying the rule extraction technique. To quantitatively evaluate the derived rule-based systems, metrics called accuracy and fidelity are utilized. As a result, the DNN method with low interpretability can be made to have high interpretability by utilizing the rule extraction technique.

2. Methods

This section describes the ECLAIRE method, a rule extraction technique that can extract ruleset from the DNN model, and the combination of grid search and early stopping techniques used to optimize the DNN model.

2.1 Efficient Clause-wise Rule Extraction

The ECLAIRE method is one of the rule extraction methods and is applied to increase the interpretability of the DNN method [1]. In general, the DNN method is intended for multiple hidden layers between the input and output layers and is known to have low interpretability [2]. It is also categorized as the deep learning methods, which means it has high performance compared to machine learning methods (i.e., decision tree, random forest, etc.). In other words, it is a method with high performance but low interpretability.

Fig. 1 shows the result of applying the ECLAIRE method to a DNN model that recommends AOPs. Here, the meanings for term, clause, rule, and conclusion are as follows:

1. Term: the minimal set of a ruleset (i.e., $x_3 > v_3$)
2. Clause: the conjunction of terms (utilizing "and"; i.e., $x_3 > v_3$ *and* $x_{12} < v_{12}$ *and* … $x_{35} > v_{35}$)
3. Rule: the conjunction of clauses (utilizing "or"; i.e., Rule 3)
4. Ruleset: the conjunction of rules
5. Conclusion: the result when the clause is satisfied (i.e., AOP 3)

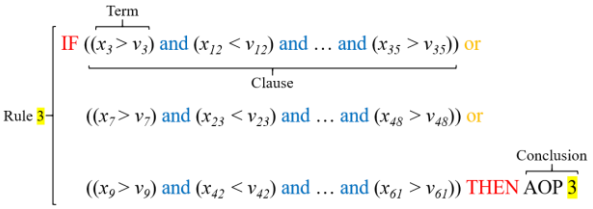

Fig. 1. Example of extracted rule.

The operation of the ECLAIRE method consists of 2 steps. The first step extracts a ruleset from each hidden layer of the DNN model. This is illustrated in Fig. 2. Specifically, a surrogate model is trained by combining the weight parameters from the hidden layer with the predicted values from the output layer. The surrogate model utilizes the C5.0 algorithm [3, 4], a refinement of

the decision tree, which is a highly interpretable method to extract ruleset. In other words, the strategy is to increase the interpretability of the DNN model by utilizing the weight parameters and predicted values of the less interpretable DNN model to train the more interpretable C5.0 algorithm. This process is repeated for the number of hidden layers in the DNN model. This means that a ruleset is generated from each hidden layer.

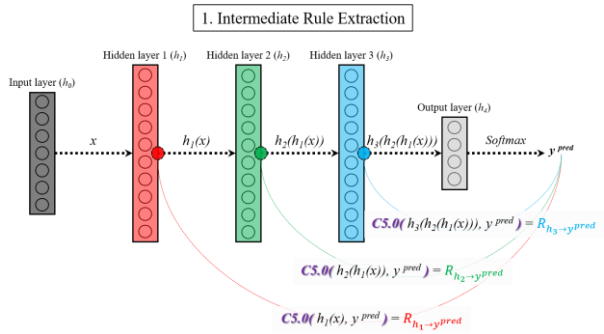

Fig. 2. Example of extracting ruleset from each hidden layer of the DNN [1].

However, the ruleset extracted from each hidden layer is not associated with any input variables (x). This means that a ruleset consisting of weight parameters (i.e., $h_1(x)$, $h_2(h_1(x))$, and $h_3(h_2(h_1(x)))$ has been created. It is necessary to combine the ruleset extracted from each hidden layer to construct a ruleset that associates the input variables with the predicted values (y^{pred}). This process is performed in the second step. Fig. 3 shows the second step schematically.

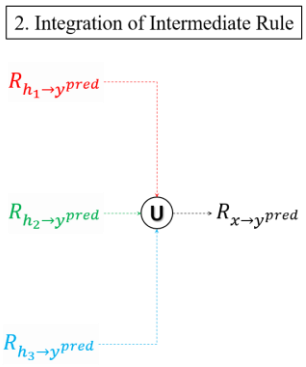

Fig. 3. Example of combining rulesets extracted from each hidden layer [1].

Specifically, the ruleset extracted from each hidden layer is decomposed in the following order, utilizing the training data that was used to train the DNN model. First, the ruleset is decomposed into clauses. Second, a conclusion is derived by sequentially injecting weight parameters to determine whether a clause is satisfied or not. Third, a surrogate model (i.e., the C5.0 algorithm) is trained by combining the input variables and the conclusions; since the weight parameters are constructed based on the input variables, it is assumed that the conclusion is also derived based on the input variables. Fourth, a ruleset is constructed to associate the input variables with the conclusion via the trained surrogate model. Fifth, the new ruleset is decomposed into clauses. Sixth, unsatisfied clauses are removed from the ruleset by sequentially injecting the input variables. Seventh, the above process is repeated on the ruleset extracted from each hidden layer. Finally, the final ruleset is completed by integrating the satisfied clauses.

2.2 Grid Search and Early Stopping

The DNN model that performs the AOP suggestions is the base for the application of the ECLAIRE method. Therefore, the performance of the DNN model is also an important factor. For the optimization of the DNN model, we used a method that combines grid search and early stopping techniques.

The grid search technique sets arbitrary hyper-parameters and sequentially trains the DNN model by utilizing the set hyper-parameter combinations; in this paper, we set the learning rate, activation function, number of layers, and number of nodes as hyper-parameters. In general, an evaluation metric is utilized to select the optimal DNN model. This process can be exhaustively checked for all hyper-parameters, but it has the disadvantage of being very time-consuming.

To compensate for these shortcomings, an early stopping technique was combined. The early stopping technique utilizes validation data that is independent of the training and testing data to confirm the learning of the DNN model. If the degree of learning does not improve, the learning is terminated. This process not only avoids the risk of potential overfitting of the DNN model, but also reduces the time required for the grid search technique.

3. Data Acquisition and Pre-processing

Data were collected to develop a DNN model for recommending AOPs. Here, the data were collected using the CNS. The CNS is an NPP simulator tool that simulates a Westinghouse-993 MWe 3-loop pressurized water reactor; it has the same parameters as the Kori-1 and Kori-2 NPPs in the Republic of Korea [5]. It can simulate normal, abnormal, and emergency operating conditions, and various faults can be injected. In this paper, we utilized the CNS to collect 15 scenarios, which are shown in Table I.

Table I: List of Simulated Scenarios

No.	Scenario name
1	Normal operating condition
2	PRZ pressure channel failure (high)
3	PRZ pressure channel failure (low)
4	PRZ water level channel failure (low)

5	SG water level channel failure (low)
6	SG water level channel failure (high)
7	Continuous insertion of control rod
8	PRZ PORV open due to failure
9	PRZ safety valve open due to failure
10	PRZ spray valve open due to failure
11	RHX front section rupture
12	CVCS to CCWS leakage
13	Leakage at back section of charging flow control valve
14	RCS to CCWS leakage
15	SG tube rupture
PRZ: Pressurizer, SG: Steam generator, PORV: Power operated relief valve, RHX: Regenerative heat exchanger, CVCS: Chemical and volume control system, CCWS: Component cooling water system, RCS: Reactor coolant system.	

The data obtained from the CNS consists of 2,222 variables, and it is necessary to select only the optimal input variables to train the DNN model. This is because unnecessary input variables disturb the learning of the AI model. In this paper, 2 DNN models are developed for various applications of the rule extraction technique. Specifically, DNN models are developed based on 8 and 15 scenarios. Furthermore, 27 and 137 input variables are selected for each DNN model by analyzing the symptom requirements of AOPs.

Normalization is performed based on the selected input variables. It is usually done to improve the learning speed. Among the various normalization techniques, the min-max normalization is used in this paper. This converts the values between 0 and 1 based on the minimum and maximum values of each variable.

4. Result

4.1 Optimization of the DNN models

In this paper, 2 DNN models were trained and optimized; each DNN model was trained with 8 and 15 scenarios, respectively. A combined method of grid search and early stopping technique was used for optimization. The main hyper-parameters for optimization were the learning rate, activation function, number of hidden layers, and number of nodes. The optimal hyper-parameters of the DNN model with 8 scenarios trained are as follows:

1. Learning rate: 0.01
2. Activation function: Sigmoid
3. Number of hidden layers: 3
4. Number of nodes: 128, 64, 32

The optimal hyper-parameters for the DNN model with 15 scenarios trained are as follows:

1. Learning rate: 0.001
2. Activation function: Sigmoid
3. Number of hidden layers: 2
4. Number of nodes: 64, 32

To evaluate the performance of the 2 DNN models, accuracy was used as the evaluation measure. The accuracy is a measure of how closely a DNN model's predicted result matches the actual correct answer, with a value closer to 100% indicating better performance.

The performance evaluation of the DNN model with 8 scenarios trained achieved 100% accuracy on both training and testing data. The DNN model with 15 scenarios trained shows 99% accuracy on the training data and 95% accuracy on the testing data.

4.2 Rule Extraction of DNN models

We apply the ECLAIRE method to the optimized DNN models to extract rulesets. The extracted rulesets are evaluated for accuracy and fidelity. The accuracy of the rulesets is the same as the accuracy of the previous DNN model evaluation, and the fidelity indicates how well the output of the DNN model matches the output of the ruleset. This fidelity is an important measure of how representative the derived ruleset is of the DNN model. The ruleset of the DNN model trained with 8 scenarios showed 100% accuracy and fidelity. The DNN model's ruleset with 15 scenarios trained showed 83% accuracy and 82% fidelity.

The rules extracted from the ruleset containing the 8 scenarios are shown in Fig. 4; it shows only the rules for some of the scenarios out of the total rules. The scenario numbers shown in Fig. 4 are listed in Table I.

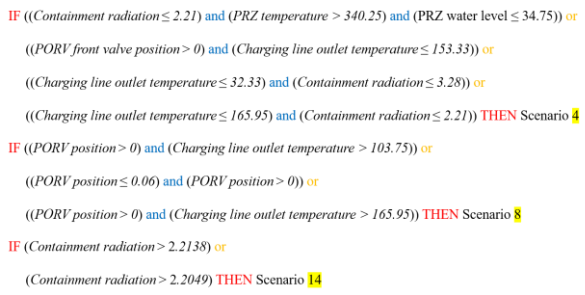

Fig. 4. The result of extracted rule.

The scenario 4 is a situation where the PRZ water level channel suddenly dictates low pressure, and the extracted rule provides containment radiation, PRZ temperature, PRZ water level, PORV front valve position, and charging line outlet temperature. The containment radiation is a significant variable in a Loss-Of-Coolant Accident (LOCA) situation, assumed to be derived because PRZ water level drops in LOCA situations as well. The remaining variables were derived under the influence of other scenarios.

The scenario 14 is a LOCA situation, and the extracted rule suggests containment radiation. This was extracted because it is the only scenario that raises the containment radiation among all the scenarios. In other

words, the containment radiation variable is the best basis for diagnosing this scenario.

5. Conclusion

In this study, we apply the ECLAIRE method, a rule extraction technique, to improve the low interpretability of DNN models. As an example, we develop a DNN model that recommends AOPs and apply the ECLARE method to this model. The application of the ECLAIRE method contributed to improving the interpretability of the DNN model. For the DNN model with 8 scenarios trained, the DNN accuracy, ruleset accuracy, and fidelity were confirmed to be 100%. This means that the DNN model and the rules in the ruleset match perfectly. However, the DNN model trained on 15 scenarios performed slightly worse, with 95% DNN accuracy, 83% ruleset accuracy, and 82% fidelity. In terms of the performance of the DNN model, 95% accuracy is also a high performance, but it showed low interpretability for the extracted rules. This means that in difficult problems (i.e., 15 scenarios), the interpretability of the extracted rules becomes lower.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (Ministry of Science and ICT) (Grant No. NRF-2018M2B2B1065651).

REFERENCES

[1] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural Computation, Vol. 1, No. 4, pp. 541-551, 1989.
[2] M. E. Zarlenga, Z. Shams, and M. Jamnik, Efficient decompositional rule extraction for deep neural networks, 1st Workshop on eXplainable AI approaches for debugging and diagnosis, arXiv preprint arXiv:2111, p. 12628, 2021.
[3] J. R. Quinlan, C4.5: programs for machine learning, pp. 17-55, Morgan Kaufmann, San Mateo, California, 2014.
[4] R. Pandya, and J. Pandya, C5.0 algorithm to improved decision tree with feature selection and reduced error pruning, International Journal of Computer Applications, Vol. 117, No. 16, pp. 18-21, 2015.
[5] J. C. Park, K. C. Kwon, B. S. Sim, J. T. Kim, D. Y. Lee, C. H. Kim, W. M. Park, S. J. Song, S. D. Yoo, and K. N. Yang, Performance and equipments upgrade of compact nuclear simulator, No. KAERI/RR—1794/97, KAERI, Daejeon and Republic of Korea, 1997.