A Preliminary Study on Interpretability of Machine Learning: Diagnosis of Internal Leakage in Feedwater Heaters

Kibeom Son, Gibeom Kim, Gyunyoung Heo*
Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Yongin-si, Gyeonggi-do, 17104, Korea
**Corresponding author: gheo@khu.ac.kr*

1. Introduction

In order for machine learning or deep learning methods to be reliably used in safety critical fields, not just model accuracy but interpretability for the process in which results are derived is likely to play a major role [1].

In this paper, three machine learning methods were demonstrated with an example data to show their interpretability: Decision tree, one of the most inherently explainable and supervised method, Principal Component Analysis (PCA), a unsupervised feature extraction method which enables the visualization of data through dimension reduction, Artificial Neural Network (ANN) which is the basis of deep learning.

Focusing on the characteristics of feature extraction of each method, their results were compared and analyzed in terms of interpretability. The example data is the internal leakage scenarios of feedwater heaters (FWHs) in nuclear power plants (NPPs) obtained by thermal performance simulations [2].

2. Methods

2.1 Decision Tree

The decision tree model is a kind of machine learning methods, and solves classification and regression problems through step-wise binary logics. This model is well known as a machine learning method with good interpretability. Comparing to others, its concept is intuitive, so the details are left out.

2.2 Singular Value Decomposition (SVD)

The PCA is a kind of unsupervised feature extraction method, and can adequately characterize high-dimensional forms of correlated data by reducing dimensions. It is a way to find a new axis that preserves the variance of existing data as much as possible, and to project the data on that axis.

Basically, analysts use a covariance matrix to perform the PCA, but we are able to use the SVD when a covariance matrix is not available.

Unlike the eigenvalue decomposition, the SVD has the advantage of being able to decompose a non-square matrix numerically stable [3, 4].

There is a large data matrix $A \in \mathbb{C}^{n\times m}$ such as *Eq.(1)*.

$$A = \begin{bmatrix} | & & | & & | \\ a_1 & \cdots & a_k & \dots & a_m \\ | & & | & & | \end{bmatrix} \quad (1)$$

Where, *n* is a number of variables, *m* is a number of datasets, $\mathbb{C}^{n\times m}$ denotes the dimension of the corresponding matrix, and vector a_k is the *k*-th measurement taken from simulation or experiment.

This matrix is decomposed through the SVD as *Eq.(2)*.

$$A = U\Sigma V^T \quad (2)$$

Where, $U \in \mathbb{C}^{n\times n}$ and $V \in \mathbb{C}^{m\times m}$ are unitary matrices with orthonormal columns, $\Sigma \in \mathbb{C}^{n\times m}$ is a real matrix and a diagonal matrix with a positive elements.

The final form of the SVD can be listed as an elemental combination of each matrix as *Eq.(3)*.

$$A_k = \sigma_1 u_1 v_1^T + \cdots + \sigma_k u_k v_k^T \quad (3)$$

Where, A_k is the approximation with the value closest to the base matrix *A*, σ_k is the *k*-th Singular value of the matrix *A*, u_k is the *k*-th element of matrix *U*, v_k is the *k*-th element of matrix *V*, and *k* is the rank.

$\sigma_1 u_1 v_1^T$ is the principal component with the largest value and since the value decreases as the order of the principal component increases, several low-order principal components can provide an approximation of matrix *A*.

The foregoing is demonstrated by the Eckart-Young theorem, as shown in *Eq.(4)*.

$$\|A - B\| \geq \|A - A_k\| \quad (4)$$

When the rank of matrix *A* is greater than *k*, and the rank of matrix *B* is *k*, the distance between *A* and *B* is greater than the distance between *A* and A_k.

By using PCA/SVD to reduce the dimensions of data, analyst can perform machine learning more efficiently.

2.3 Artificial Neural Network

An ANN is based on perceptron, the most basic form of neural network, and consists of input layer equal to the number of features of the input data, hidden layer processing the input data, and output layer representing the calculation results.

Figure 1 shows a basic structure of an ANN and expanding this concept gives some insight into the principle of deep learning.

Assuming a linear relationship between layers, it will have the following relationship as *Eq.(5)*.

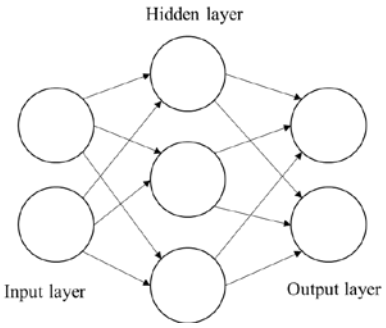

Fig. 1. The conceptual structure of NN

$$x^{(1)} = A_1 x$$
$$x^{(2)} = A_2 x^{(1)} \quad (5)$$
$$\boldsymbol{y = A_3 x^{(2)}}$$

Where $\boldsymbol{A_1}$ means the first weight, $\boldsymbol{x^{(1)}}$ means the first layer, and when there are *M* layers, it can be expressed in the following as *Eq.(6)*.

$$y = A_M A_{M-1} \cdots A_1 x \quad (6)$$

If the relationship between layers is nonlinear, the right side can be expressed as a function, which is called an active function, for instance, Logistic, TanH, and Rectified Linear Unit (ReLU).

3. Case study

3.1 Internal leakage in feedwater heaters

In this paper, thermal performance analysis prepared by the simulation of Performance Evaluation of Power System Efficiencies (PEPSE), and the modelled plant is a turbine cycle of OPR-1000 type reactor.

It was assumed that leakage occurred at the FWH #5, 6, 7, and the leakage locations were modeled in four types: pass partition plates, inlet side tube-sheets, U-tubes section, and outlet side tube-sheets.

The input data is the leakage data of the FWHs used for machine learning in this paper, and consists of 14 features, 4 leakage locations, and 52 datasets.

Table I: Meaning of the features

Features	Meanings
X1	Electric output
X2	Flow rate
X3	Temperature difference at tube side (FWH # 7)
X4	Temperature difference at shell side (FWH # 7)
X5	TTD (FWH #7)
X6	DCA (FWH #7)
X7	Temperature difference at tube side (FWH # 6)
X8	Temperature difference at shell side (FWH # 6)
X9	TTD (FWH #6)
X10	DCA (FWH #6)
X11	Temperature difference at tube side (FWH # 5)
X12	Temperature difference at shell side (FWH # 5)
X13	TTD (FWH #5)
X14	DCA (FWH #5)

Table 1 contains the detailed meanings of the input data. The features were selected as electric output, flow rate, temperature difference at shell side, temperature difference at tube side, Terminal Temperature Diffusion (TTD), and Drain Cooler Approach (DCA). In other words, data can be defined as classification problems with 14 dimensions, 4 labels, and 52 sets.

This section discusses the results of performing decision tree on input data. For visualization, the maximum depth is limited to 4, and Figure 2 shows the results of the decision tree model.

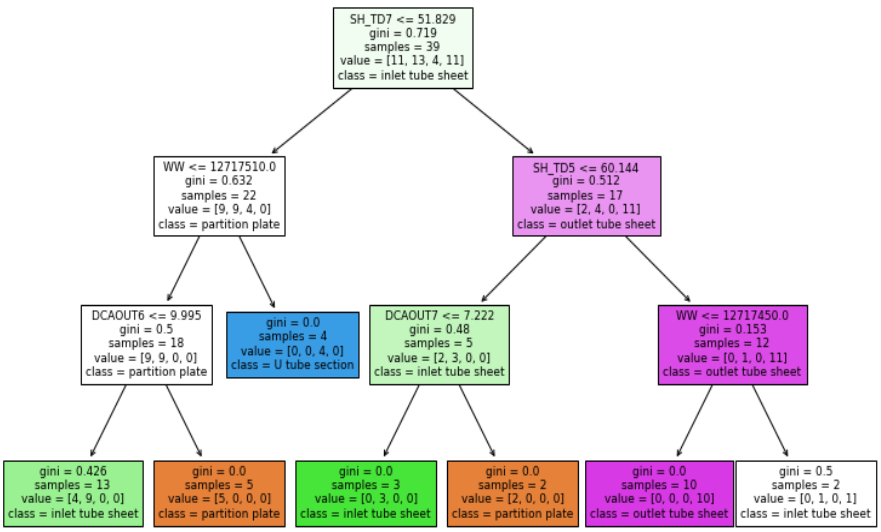

Fig. 2. The result of decision tree learning

3.2 Interpretability of Decision Tree

In case of U tube section leakage, it can be distinguished based on the main stream flow rate. So it is possible to interpret that the leakage of the partition plate is closely related to the DCA of FWH #6, 7.

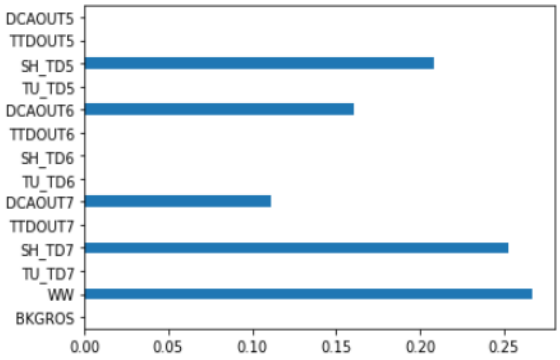

Fig. 3. The feature importance of decision tree

Figure 3 shows a bar graph of the analysis results of the importance of the decision tree model. Through this graph, it can be inferred that main stream flow rate, temperature difference at shell side and DCA are contributing greatly.

3.3 Interpretability of SVD

This section describes the process of applying SVD/PCA for input data.

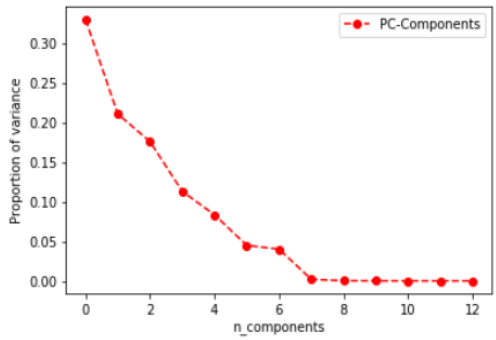

Fig. 4. The scree curve of PCA

Figure 4 shows the accuracy of the reduced model according to the number of principal components, and most of the existing 14 dimensions of data can be expressed with just 6 principal components.

Figure 5 shows the relationship between principal components (PCs) and features, and each principal component is extracted in a combination of all features.

An equal sign between features in one principal component means that there is a common correlation between those features.

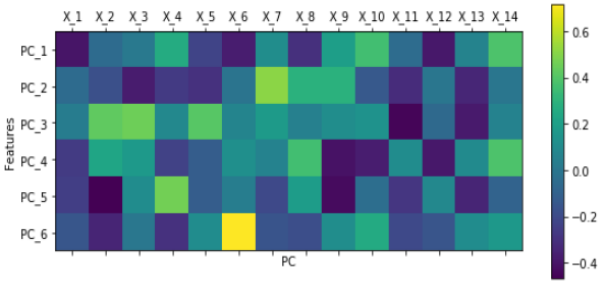

Fig. 5. The heatmap of PCs

As noted in *Eq.(3)*, it is necessary to interpret the results in consideration of the fact that the smaller the order of principal components, the greater the contribution to the model.

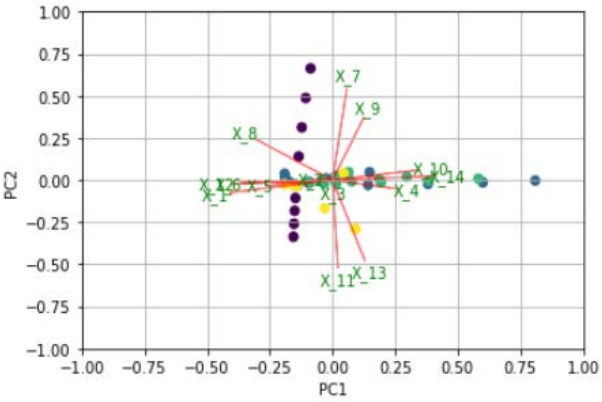

Fig. 6. The biplot for PC1 & PC2

Figure 6 illustrates the relationship between principal components and features on the PC1 and PC2 axes according to the aforementioned points. Point graphs are representations of the original data on axes of PC1 and PC2, and the length of the arrow graphs means the variance of each feature on that axis.

The closer the distance and direction of each feature is, the higher the correlation between the features. For example, X4, X10, and X14 on the right side are highly correlated with PC1 and can also be found in Figure 5.

In addition, in the case of partition plate leakage, which is dotted vertically and distributed, it can be seen that there is a significant association with X7 and X11, which was difficult to identify in the decision tree model.

On the other hand, it is difficult to derive physically significant results from the correlation between principal components and features.

When comparing the results of this with the decision tree model, feature X4, X6, X10, and X12 can be inferred that the interpretability of the two models are similar. The contribution of X4, X6, X10, and X12 are derived from the decision tree model, and in the PCA model, X4, X10 are distributed to the right and X6, X12 are distributed to the left in similar directions.

3.4 Interpretability of ANN

The deep learning methodology based on neural networks is known that interpretation is ambiguous or not enough due to the nature of learning structure or process.

The example problem conditions are 14 input layer nodes, 5 hidden layer nodes, 4 output layer nodes, and RELU active functions.

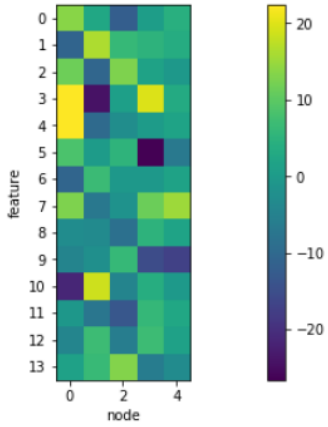

Fig. 7. The heatmap of weights between input layer and hidden layer

Figure 7 is a weighted heat map between the input layer and the first hidden layer, and a large weight can be inferred that it has a large contribution to the model or an appropriately entered feature.

The heatmap shows relatively large weights for the X4, X5, X6, X10, and X11 features, which are somewhat similar to those of the previous models.

The weight between the hidden layer and the output layer is also known to be more difficult to interpret, although visualization is possible. This is because the weight between the hidden layers and output layers alone cannot identify the whole model.

4. Conclusions

In this paper, we discussed the potential of interpretability with an example of FWH leakage diagnosis through machine learning methods: decision tree, SVD/PCA, ANN.

The decision tree method was able to clearly identify the classification process of the label. In the case of SVD/PCA, various visualizations were possible through the reduction of the dimension of the data.

While many artificial intelligence studies focus on accuracy, it is expected that the applicability will increase only when interpretability is taken into account.

This paper is a preliminary calculation for checking such characteristics, and will keep track of related research and technology.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP:Ministry of Science, ICT and Future Planning) (No. 2017M2B2B1072806)

REFERENCES

[1] G.Y. Heo, E. Zio, Statistics and Learning from Regulatory Viewpoint in Safety-Critical Systems, Korean Society for Prognostic and Health Management 2020, Seoul, Korea, July 21-23, 2020.

[2] G.Y. Heo, S.K. Lee, D.W. Kim, Diagnosis of Internal Leakage in Feedwater Heaters Using Neural Networks, The 7th International Topical Meeting on Nuclear Reactor Thermal Hydraulics, Operation and Safety, Seoul, Korea, October 5-9, 2008

[3] Steven L. Brunton, J.Nathan Kutz, Data-Driven Science and Engineering, CAMBRIDGE PRESS, 2019

[4] G. Strang, Linear Algebra and Learning from Data, WELLESLEY-CAMBRIDGE PRESS, 2019