Is the Surrogate Model Interpretable?

Sangwon Kim **Mira Jeong** **Byoung Chul Ko** [*]

Department of Computer Engineering
Keimyung University
Daegu, South Korea
{eddiesangwonkim, mystroll24}@gmail.com, niceko@kmu.ac.kr

Abstract

In this study, we propose a lightweight surrogate random forest (SRF) algorithm that can be interpreted through a new rule removal method. First, we create a lightweight student SRF that can replace the existing heavy, deep, but high-performance teacher model using teacher–student learning framework. However, the student SRF obtained in this way must maintain the performance of the teacher model, and thus the degree of weight reduction is extremely limited. Therefore, in this paper, we propose a cross-entropy Shapley (CES) value to achieve an optimized model weight reduction. The rule contribution of the decision trees constituting the random forest is calculated based on the CES value to further lighten the SRF and measure the importance of the SHapley Additive exPlanations (SHAP) feature for the features at the same time. To demonstrate the superior performance of the proposed lightweight method, several comparative experiments were conducted on various datasets. We evaluate experimentally that the proposed method is superior to existing deep neural networks and machine learning methods in terms of the number of parameters and the amount of memory use, and demonstrates the importance of the features.

1 Introduction

In recent years, an ability to observe the cause and effect of reasoning in the machine learning model is required, and as a result, the necessity of research into explainable artificial intelligence AI (XAI) or interpretable machine learning (IML) has emerged. XAI technology can be largely divided into transparent and post hoc explanations. The latter case is a method of explaining the cause of inference by learning an explainable subsystem while leaving the black box model as is. However, because this method does not change the heavy deep neural network (DNN) structure itself, it still requires a high-end system, and the model itself cannot be interpreted. This study focuses on a transparent design that makes the deep model itself transparent such that the model's inference decision process can be understood. Through a lightweight system, it is possible to analyze features that affect the inference, and to understand how model structures operate in the inference process.

2 Surrogate random forest

DNN and machine learning models achieve a higher performance as the model becomes deeper and heavier, but have a disadvantage in that memory usage increases and the speed is slowed down owing to many different parameters. To solve this problem, a model compression method based on

[*] Correspond author

1st NeurIPS workshop on Interpretable Inductive Biases and Physically Structured Learning (2020), virtual.

the teacher–student (T-S) framework was introduced that can construct a shallow network model by reducing the size of the model while maintaining a similar performance as an existing deep and wide network model [1]. The T-S framework constructs a large-scale teacher model with numerous parameters, and a small-scale student model is therefore trained to derive similar results as the teacher model but with fewer parameters, while maintaining the excellent classification performance.

In this paper, we propose a surrogate random forest (SRF) model that can maintain the performance while applying fewer parameters using the T-S framework. The process of creating the SRF is as follows: The training dataset is divided into dataset A for training the teacher model and dataset B for training the student model. First, the teacher DNN model is obtained by applying data set A, labeled 0 and 1 (hard target), to AutoML [2]. Then, by inputting the unlabeled dataset B into the trained teacher DNN, a soft target (label), which is a class-specific probability value output from softmax, is assigned as a label to dataset B. Now, we construct a student RF model by reducing the number of trees using dataset B and N-fold cross-validation. The student RF model can be trained to consider the inter-class relationships of the teacher DNN model by using training data labeled as a soft target. The student RF model created through this learning process is called the SRF model.

3 Lightness of SRF

The SRF model has a much smaller scale than the teacher model. However, the size of the student model that is initially defined must have more than a certain parameter to achieve the same performance as the teacher model. Therefore, even with the surrogate model based on the T-S framework, the size that can be reduced is limited, and it still has the disadvantage of low transparency. To solve this limitation, in this paper, an additional SRF model reduction method is proposed that applies rule removal based on the CES value to further lighten the SRF model obtained from the T-S framework.

3.1 Decision of rule contribution

The RF is an ensemble model of decision trees [3], and each decision tree is a set of rules that are paths from a root to intermediate and leaf nodes. We can reduce the weight of the RF by evaluating the contribution of all rules constituting the decision trees and removing the rules with a low contribution. In this study, the Shapley value [4] was used to determine the contribution of the rule. The existing Shapley value was used to measure the contribution of feature vectors input from a machine learning model. This value measures the difference in accuracy according to the presence or absence of a specific feature, and the greater the difference, the higher the degree of contribution given to the corresponding feature. In this study, instead of determining the contribution of the input feature, the Shapley value is used to determine the contribution to the rule in the RF, and a new cross-entropy Shapley (CES) value is proposed to evaluate the contribution to each rule in more detail. Whereas the existing rule removal method [5, 6] evaluates the prediction accuracy according to the rule of the tree, the proposed CES value can be evaluated in more detail by considering the prediction probability value for each class according to the rule. The rule contribution evaluation method using the CES value is as follows:

$$\Phi_j(T_{CE}) = -\sum_{R\subseteq\{v_1,\ldots,v_n\}\setminus v_j} \frac{|R|!(n-|R|-1)!}{n!}(T_{CE}(R\cup v_j) - T_{CE}(R)) \tag{1}$$

where v_j means the j-th rule constituting RF, and R is a subset of the n rules constituting the learned RF, excluding the v_j-th rule. $T_{CE}(\bullet)$ denotes the CES value for $(\bullet)$. By measuring the CES value for each rule of the previously learned SRF, the model is lightened by removing the rule with low contribution.

3.2 Rule removal using mini-grouping

In general, the number of rules of a RF ranges from tens of thousands to millions of rules, depending on the tree and node depth. The individual contribution check and rule removal process for the entire rule not only takes a significant amount of processing time, it can also cause an over-fitting of the model. Therefore, in this study, a random mini-grouping method is proposed to perform optimal rule removal while minimizing the overfitting problem caused by rule removal based on checking the individual contributions. In the random mini-grouping, the rules of the RF are randomly clustered into K groups, and the degree of contribution is evaluated by estimating the CES value for each group. This process is repeated H times, and each rule has $|H|$ contributions. The average value is then calculated to determine the final contribution of the rule. Finally, by removing the rule with a low contribution according to the measured contribution, the model can significantly reduce the model size while maintaining the existing prediction performance. Algorithm 1 introduces the weight reduction process through random mini-grouping with the CES value of the SRF model.

Algorithm 1 Rule removal using random mini-grouping

```
Input: Trained $SRF$, Number of group $K$, rule set $R = r_1, r_2, ..., r_M$, Number of rules $M$, Number of
    iteration $H$, rule elimination rate $\delta$
Output: A lighten random forest model $iRF$
for $h = 1\ to\ H$ do                                              ▷ randomly mini-grouping
    for $k = 1\ to\ K$ do
        $g_k \leftarrow randomly\ extract\ rule - subset\ from\ R$
        $G^h \leftarrow G_h \cup g_k$
    end for
    for $k = 1\ to\ K$ do                                          ▷ calculate the CES Value $\phi$
      $\Phi_k^h(T_{CE}) \leftarrow -\sum_{G^h \subseteq \{g_1,\ldots,g_K\} \setminus g_k} \frac{|G^h|!(K-|G^h|-1)!}{K!}(T_{CE}(G^h \cup g_i) - T_{CE}(G^h))$
    end for
end for
for $h = 1\ to\ H$ do                                              ▷ calculate the average $\phi$ of each rule
    for $m = 1\ to\ M$ do
      $\phi'_n \leftarrow \frac{1}{H}\sum^K \phi_k^h$
    end for
end for
      $R' \leftarrow sort\ the\ rules\ by\ \phi'$
for $m = 1\ to\ M$ do
    if $m < (M \cdot \delta)$ then
      $then\ iRF \leftarrow RF - r_m,\ r_m \leftarrow \{x|x \subseteq R'\}$
    end if
end for
```

4 Experimental results

In this study, the AutoGluon AutoML [2] toolkit was used to generate the teacher model, and Scikit-learn and Python were used to implement the SRF. We conducted comparative experiments and a performance evaluation of the proposed method and the latest algorithms using UCI datasets [7] (Adult Income, MONK's problem, Breast Cancer Wisconsin, and Statlog). To implement the proposed algorithm experimentally, AutoML [2] was used to generate the teacher model, and the initial student SRF model was constructed using 100 decision trees.

When AutoML was applied as the teacher model as shown in Table 1, it was confirmed that the accuracy according to the rule removal rate of the SRF showed a mostly similar prediction accuracy as that of the teacher model even when 80% (RF 0.8) of the total rule was removed. In addition, for the evaluation of the mini-grouping, the number of groups was evaluated for 5 cases (20, 10, 5, 3, and 1), and when rule removal was performed based on 10 or more groups, the performance improved by up to 2.83%.

To evaluate the superior performance of the proposed method, a comparative experiment was conducted using the complete RF, ExtRa [8], k-NN [9], SVMs [10], BestRF [11], and a teacher DNN based on Auto Gluon applying another UCI dataset (MONK's problem, Breast Cancer Wisconsin, or Statlog). As shown in Fig. 1 (a), the proposed SRF method showed a much better performance than the complete RF, ExtRa, k-NN, SVMs, and BRF algorithms. In particular, the SRF model obtained through a weight reduction showed a better performance than the teacher DNN generated by AutoML for the MONK's problem and Statlog data. These results show that the proposed CES value and random mini-grouping method effectively remove unimportant rules that degrade the performance, thereby increasing the performance.

We also sort the features by decreasing the feature importance obtained in the process of calculating CES values of the rule and plotting them as shown in Fig. 1 (b). The following figure shows the

Table 1: Comparison of accuracy performance according to rule removal of AutoML-based teacher model and SRF learned through teacher model using UCI Adult Income

Methods		Accuracy (%)				
Teacher	AutoML	**85.40**				
Methods (Rule Rate)		Number of Groups				
		20	10	5	3	1
Surrogate	RF (1.0)	85.24	85.24	85.24	85.24	85.24
	RF (0.9)	85.27	85.25	85.19	85.26	85.24
	RF (0.8)	**85.35**	85.25	85.20	85.21	85.24
	RF (0.7)	85.25	85.27	85.20	85.31	85.24
	RF (0.6)	85.26	85.25	85.19	85.27	85.19
	RF (0.5)	85.25	85.22	85.25	85.22	85.12
	RF (0.4)	85.17	85.12	85.19	85.14	85.05
	RF (0.3)	85.06	85.08	84.06	85.03	84.48
	RF (0.2)	84.06	84.40	83.51	84.30	82.52

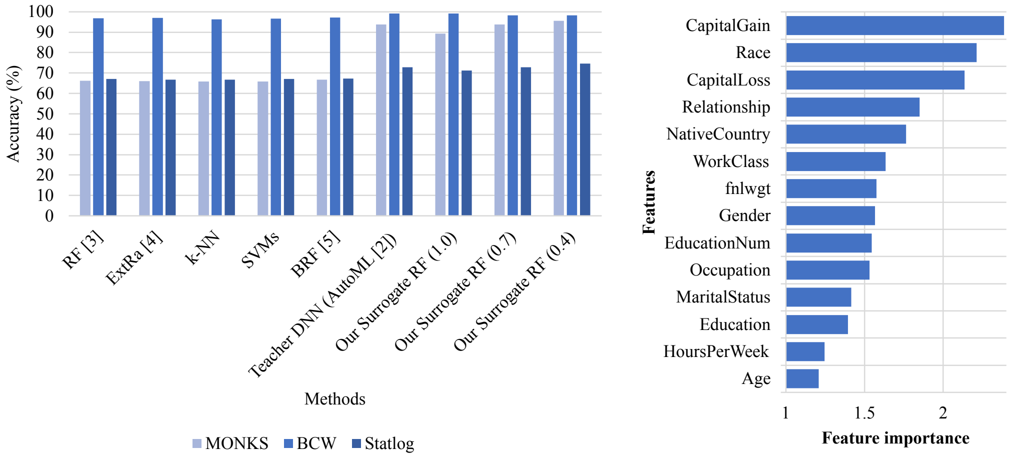

Figure 1: (a) Comparison of machine learning model performance evaluation using UCI data sets (MONK's problem (monks), Breast Cancer Wisconsin (bcw), and Statlog), (b) Feature importance measured as the mean of the CES values using the UCI Adult Income dataset. The marital status was shown to be the most important feature, changing the predicted income probability by an average of 2.5%.

feature importance for the SRF trained in advance to predict the UCI Adult Income dataset. The dataset consists of 14 features plotted along the vertical axis of the figure. To gather each feature importance, we stacked the calculated cross-Shapley values corresponding to each feature during the training phase. Finally, we averaged a number of CES values in a feature by feature manner.

5 Conclusion

The proposed method for reducing the weight of the SRF guarantees a predictive performance equivalent to that of the existing model, while confirming the ability to compress the model on a small scale. In conclusion, by using the model weight reduction method proposed in this paper, not only can the model processing speed be improved, the analysis power of the model can also increase.

In future research, we will develop an SRF model that can be applied to a wider variety of data, including images, and expand it to a model that can visualize the feature importance of and the correlation among the different features.

Acknowledgments and Disclosure of Funding

This work was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education through the Basic Science Research Program under Grant 2019R1I1A3A01042506.

References

[1] Sangjun Kim, Sooyeong Kwak, Byoung Chul Ko. Fast Pedestrian Detection in Surveillance Video Based on Soft Target Training of Shallow Random Forest. *IEEE ACCESS*, 7:124185-12426, 2019.

[2] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, Alexander Smola. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. *arXiv preprint arXiv:2003.06505*, 2020.

[3] Leo Breiman. Random forest. *Machine Learning*, 45:5-32, 2001.

[4] Lloyd Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307-317, 1953.

[5] Hongyu Yang, Cynthia Rudin, Margo Seltzer. Scalable Bayesian rule lists. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3921-3930, 2017.

[6] Sangwon Kim, Mira Jeong, Byoung Chul Ko. Energy Efficient Pupil Tracking Based on Rule Distillation of Cascade Regression Forest. *Sensors*, 20(18):1-17, 2020.

[7] Census income dataset. *University of California, Irvine*. https://archive.ics.uci.edu/ml/datasets/census+income

[8] Pierre Geurts, Damien Ernst, Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3-42, 2006.

[9] Dennis Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, 2(3):408-421, 1972.

[10] Corinna Cortes, Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273-297, 1995.

[11] Hanyuan Hang, Xiaoyu Liu, Ingo Steinwart. Best-scored Random Forest Classification. *arXiv preprint arXiv:1905.11028*, 2019.