CaRFE: Candidates Recursive Feature Elimination for Black-Box Model

Song Youngjae, Kim Kwangsu*

Sungkyunkwan University, *Sungkyunkwan University

thddudwo1313@gmail.com, *kim.kwangsu@skku.edu

Abstract

RFE(Recursive Feature Elimination) is a widely-used greedy algorithm for a feature selection. However, the greedy feature selection does not guarantee obtaining an optimal feature subset as there is no method to predict perfectly precise feature importance for a black-box model such as a deep neural network. Thus, this paper proposes a novel variation of RFE for the black-box model, namely CaRFE(Candidates RFE). Through two experiments in the solar power forecasting dataset, we empirically show that CaRFE outperforms RFE in model performance and can reach a nearly optimal feature subset.

Ⅰ. Introduction

RFE[1] is a greedy feature selection technique eliminating unimportant features through evaluating the feature importance. Although there have been proposed several variation of the RFE[4, 5] so far, they usually focus on the traditional machine learning model rather than a black-box model such as deep neural network. As there is no method to predict perfectly precise feature importance for deep neural network, the greedy manner of the RFE misleads it to reach an optimal feature subset for deep neural network. Thus, we need an improved version of the RFE which is applicable to a black box model such as deep neural network.

Motivated by this, we propose a practical feature elimination framework, CaRFE(Candidates Recursive Feature Elimination). The CaRFE employs a candidates search to maintain the algorithm much less greedy and exploits both the feature importance and model performance metric. This paper computes a feature importance and model performance with KernelSHAP[2] and customized NMAE(Normalized Mean Absolute Error), respectively.

Through two experiments in the solar power forecasting dataset, we show that the CaRFE outperforms the RFE in model performance. Additionally, we identify that the CaRFE can reach a nearly optimal feature subset by comparing it with an exhaustive search. The contribution of this work is that we introduce an improved and practical version of the RFE appliable to a black-box model.

Ⅱ. CaRFE: Candidates Recursive Feature Elimination

CaRFE consists of 2 stages per one iteration to eliminate one unimportant feature. The first stage identifies candidate feature subsets C_i, $i \in 1,2,...,p$, that the most unimportant feature is eliminated from the initial feature set F. Given the $C_1, C_2, ..., C_p$, the second stage stores a feature subset with the best model performance among them. Then the feature subset is assigned for the following initial feature set. The procedure above is recursively iterated until the number of remained features becomes k. Finally, among the stored candidates, CaRFE select a feature subset with the best performance. Algorithm 1 presents the details about the CaRFE algorithm. The CaRFE becomes RFE if we assign p to 1 and do not care about the model performance. The CaRFE is much less greedy than RFE due to the candidate search and able to reach a nearly optimal feature set as the user searches for the appropriate hyper-parameter p.

Algorithm 1: CaRFE

Input: F = Initial feature set
$M(F)$ = trained model with feature set F
$S(M)$ = list of feature importances for M
p = number of candidates
k = number of final features

Initialize the empty list *result*
while $|F| > k$ **do**
  Sort $S(M(F))$ in ascending order
  /* stage 1 starts */
  for $i = 1, 2, ..., p$ **do**
    $C_i \leftarrow F$
    $s_i \in S(M(C_i))$
    Remove the feature with s_i from C_i
  end
  /* stage 2 starts */
  Let C_{best} a feature subset with the best performance among $C_1, C_2, ..., C_p$
  $F \leftarrow C_{best}$
  Append C_{best} to *result*
end
Let *final* a feature subset with the best performance in *result*
return *final*

Ⅲ. Experiment

We apply RFE and exhaustive search to compare with the CaRFE. We work with the solar power forecasting dataset of the Ulsan solar plant in South Korea. The initial feature set F consists of 10 features: 'season',

'cos(time)', 'GHI', 'DNI', 'Temperature', 'Humidity', 'Wind_x', 'Wind_y', 'clearness' and 'Visibility'. In addition, we set 3 for both k and p. We implement an MLP architecture constructed by 1 dense layer with 256 neurons, 3 dense layers with 128 neurons, 1 dense layer with 64 neurons and 1 dense layer with 32 neurons sequentially. Dropout layers are placed with a 0.2 dropout rate between each layer. The activation function is ReLU for all dense layers. The optimizer is the Adam[3] with a 0.001 learning rate and the epochs are 100. The model performance metric used here is defined below.

$$NMAE = \frac{100}{c} \times \frac{1}{N} \sum_{i=1}^{N} |y_i - f(x_i)|$$
$$i \in \{i \mid y_i \geq c \times 0.1\}$$

The y_i and $f(x_i)$ indicate an actual power generation and the predicted one, respectively. And the c stands for the capacity of the solar panel in the Ulsan solar plant: 500. N is the number of all instances for the dataset.

The experiments are consist of 2 parts: comparison to the RFE and comparison to the exhaustive search. We evaluate the performance of the CaRFE in both cases with the NMAE metric.

3.1 Comparison to the RFE

We examine how low NMAE the CaRFE and the RFE can reach. To guarantee the reliance for the result, we repeat the experiment 10 times with different random seeds and take the average of NMAE. Figure 1 presents the average NMAE plot by 10 repetitions of the CaRFE and the RFE. It shows that the CaRFE can reach way lower NMAE, 4.294, whereas RFE can only get 4.608 for the lowest NMAE.

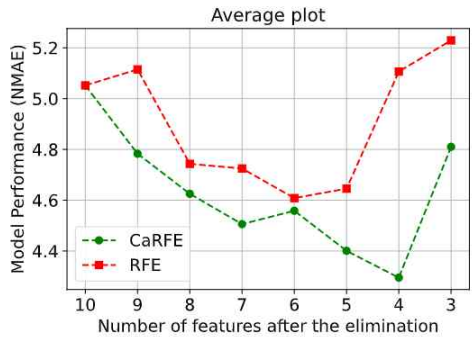

Fig 1. Average behaviors of NMAE during CaRFE and RFE

Additionally, since CaRFE is much less greedy in nature, the NMAE is lower than RFE over the whole interval. This indicates that CaRFE can select important features in practice regardless of the number of features a user wants to use.

3.2 Comparison to the exhaustive search

This experiment aims to ascertain whether the feature set obtained by CaRFE resembles the optimal feature set. We evaluate NMAE for all possible feature combinations with the exhaustive search for the given features: 968 cases in total, whereas CaRFE tries 21 cases for the given setting. If there is no significant difference between NMAE from the best feature set by CaRFE and exhaustive search, we can conclude that the CaRFE can find an almost optimal feature set in the given setting.

Table 1. CaRFE and exhaustive search comparison

	CaRFE	Exhaustive search
Best feature subset	season cos(time) GHI Wind_y	season cos(time) GHI Visibility
NMAE	4.212	4.191

Table 1 presents the best feature subset by the CaRFE and the exhaustive search. There is no significant gap between them with respect to the NMAE and the qualitative comparison. This implies that the CaRFE can find a nearly optimal feature set, unlike the RFE.

Ⅳ. Conclusion

In this work, we present the CaRFE, the improved version of RFE. The CaRFE exploits both the model performance metric(NMAE) and feature importance computed by KernelSHAP with candidates search. The CaRFE is more reliable than the RFE and applicable to a black box model. The two quantitative experiments show that the CaRFE outperforms the RFE and can find nearly optimal feature set in the solar power forecasting dataset.

ACKNOWLEDGMENT

This research was partly supported by the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program(IITP-2021-2015-0-00742) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00990, Platform Development and Proof of High Trust & Low Latency Processing for Heterogeneous·Atypical·Large Scaled Data in 5G-IoT Environment)

Reference

[1] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines" Machining Learning, vol. 46, no. 1-3, pp. 389-422, 2002.

[2] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions" In Advances in Neural Information Processing Systems, pp. 4768 - 4777, 2017.

[3] Diederik Kingma and Jimmy Ba. Adam, "A method for stochastic optimization" In ICLR, 2015.

[4] Chen and Jeong "Enhanced recursive feature elimination" In IEEE sixth international conference on machine learning and application. pp. 429-435, 2007

[5] H. Jeon and S. Oh "Hybrid-recursive feature elimination for efficient feature selection" Appl. Sci, 10. pp. 3211, 2020