Hindawi
Wireless Communications and Mobile Computing
Volume 2017, Article ID 6817627, 13 pages
https://doi.org/10.1155/2017/6817627

Research Article

A Variable Impacts Measurement in Random Forest for Mobile Cloud Computing

Jae-Hee Hur,[1] Sun-Young Ihm,[2] and Young-Ho Park[1]

[1] *Department of IT Engineering, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul 04310, Republic of Korea*
[2] *Big Data Using Research Center, Sookmyung Women's University, Cheongpa-ro 47-gil 100, Yongsan-gu, Seoul 04310, Republic of Korea*

Correspondence should be addressed to Young-Ho Park; yhpark@sm.ac.kr

Received 13 April 2017; Accepted 13 June 2017; Published 7 September 2017

Academic Editor: B. B. Gupta

Recently, the importance of mobile cloud computing has increased. Mobile devices can collect personal data from various sensors within a shorter period of time and sensor-based data consists of valuable information from users. Advanced computation power and data analysis technology based on cloud computing provide an opportunity to classify massive sensor data into given labels. Random forest algorithm is known as black box model which is hardly able to interpret the hidden process inside. In this paper, we propose a method that analyzes the variable impact in random forest algorithm to clarify which variable affects classification accuracy the most. We apply Shapley Value with random forest to analyze the variable impact. Under the assumption that every variable cooperates as players in the cooperative game situation, Shapley Value fairly distributes the payoff of variables. Our proposed method calculates the relative contributions of the variables within its classification process. In this paper, we analyze the influence of variables and list the priority of variables that affect classification accuracy result. Our proposed method proves its suitability for data interpretation in black box model like a random forest so that the algorithm is applicable in mobile cloud computing environment.

1. Introduction

Mobile cloud computing becomes a significant issue for data mining. Since multimodal sensor data is gathered from mobile devices, data mining in a mobile cloud environment is an important research area. Multidimensional data from mobile devices such as health information and GPS increases exponentially so that it becomes difficult to handle manually.

There are some researches on the progress that measures variable impact in classification and regression from the big data with multidimensional attributes by using data mining algorithms. As data becomes more complex, the importance of research in interpreting the meaning of data classification and regression results is increasing. The main problem of the multidimensional data analysis is the curse of dimensionality. Since high-dimensional data streams in real time, which is so-called "small n large p" problem, dimension reduction is a critical issue for efficient data analysis. The following examples illustrate the increasing need for research to identify important variables that have affected classification as well as increasing classification accuracy.

Example 1. Assume the situation that the doctor who diagnosed patient P used a data mining algorithm to determine whether the patient had cancer or not. The algorithm that completes its training process based on the patient data that was judged as cancer-positive in previous data judged patient P as cancer-positive. Before the doctor makes a definitive diagnosis to patient P, the doctor wants to know the specific reason that learning algorithm gave the cancer-positive diagnosis to patient P.

Example 2. We assume two people B as the banker and C as the customer. C wants to borrow money from the bank. When C visited the bank and asked B for his loan approval, B would like to know about the transaction history of C. Before making a confirmation, B wants to predict whether C has the ability to repay the loan or not. Since transaction data is composed of multidimensional attributes, it is impossible for B to investigate all data. Therefore, a data mining algorithm can support the decision based on the database by querying historical data of C. When the algorithm makes a suggestion to allow loan towards C, B may want to inspect the decision that algorithm made and which variables gave major impact to the result.

As examples suggested above, the needs for variable impact measurements research is increasing. However, even if the prediction accuracy of the learning algorithm is high, there is a danger that the reliability of the doctor's diagnosis may deteriorate if the physician cannot directly confirm the cause of the algorithm result. Also, in the second case, it is very important for the banking industry to determine what data from the customer has affected the classification results before deciding whether to approve the customer's loan or not.

Recently in bioinformatics field, as personal medical data becomes more complicated and accumulated in real time, the related work was proposed [1–3]. There is an increasing demand for research algorithms that can accurately predict patient's disease name in the multidimensional property [4]. Therefore, it is important to measure which variable among the various attributes contained in the individual's medical data has affected the prediction results of the algorithm. Random forest algorithm performs reliable classification in this area. Statnikov et al. [5] applied binary and multicategory classification towards cancer diagnosis. The paper investigates that random forests are outperformed by SVM. Díaz-Uriarte and Alvarez de Andrés [6] prove that random forest algorithm is well suited for a large number of datasets and solve the classification problem on gene selection issue. Wu et al. [7] compare five machine learning algorithms, linear discriminant analysis, kNN classifier, bagging and boosting classification trees, SVM, and random forest.

However, random forest algorithm has a critical problem. Since it is a black box model, we cannot see which variable is affected in classification result. It is important to interpret the result of the classification with variable importance measurement. Hapfelmeier et al. [8] investigate the variable importance measurement when the data contains missing values. The research proposed allocating variables randomly instead of permuting value to overcome the drawback of previous approaches which do not consider the missing data. Also, Gregorutti et al. [9] proposed new algorithm to eliminate variables recursively to predict with a smaller number of data. The algorithm is efficient when the high-dimensional regression or classification is required.

In this paper, we propose a new method that accurately grasps the influence of relative classification among variables in measuring the influence of classification of variables using random forest algorithm in an attempt to solve the problems. To solve this problem, this paper proposes a method to incorporate the economics theory called Shapley Value into the MDA index.

1.1. Random Forest.

The random forest algorithm, which is a kind of ensemble learning technique, generates several decision trees by bootstrapping the learning data and arbitrarily learns them. We then combine the learning results of all the trees to obtain the average in the case of regression and the prediction accuracy in the case of classification by the majority. By learning random decision trees and then averaging them, random forests solve the over sum problem by reducing the variance compared to the single decision trees. In particular, random forests are more suitable for the field of bioinformatics through the study that they have a good performance when sorting data with multidimensional data attributes but small number of data for each "small n large p." However, the random forest algorithm corresponding to the black box model has a high prediction accuracy, but it has a disadvantage in that it cannot intuitively interpret data in which the classification is performed directly in the internal process.

The principle of random forest operation is as follows. First, various subsets are arbitrarily generated from existing learning data for random forest learning. The most important characteristic of the random forest is the bagging. Bagging was proposed by Breiman [2] in 1996 as a shorthand for bootstrap aggregation. The decision tree was originally good in classification, but, due to overloading, random forests use bootstrap to perturb data. According to Breiman, bagging predictors are a method of generating multiple versions of a predictor and using an aggregated predictor. The bagging can improve the accuracy rate of the algorithm because the perturbation in learning set could cause a change in predictor construction. Research on the stability of variable impact measurements based on random forest algorithm received high attention in these days [10]. In a recent study, the variable impact measurement is divided into two categories: Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA).

1.2. Variable Impact Measurement Index.

Linear regression analysis and decision trees are the most frequently used algorithms for verifying the influence of classification results [11]. However, as the data age becomes more complex as the age of big data grows, linear regression algorithms do not show effective classification results. It is easy to intuitively interpret the learning result, and a decision tree with good performance has emerged as an alternative to multidimensional property classification of data. However, the decision trees are overly compliant with the training data, and there is a problem of overarching consensus that the accuracy of the test data prediction is relatively low. A random forest method has been proposed to solve the problem of prediction accuracy of decision tree.

There are two main indicators to measure the influence of classification of a variable through the random forest. One is the Mean Decrease Impurity (MDI) index, which measures the classification impact of variables by totaling the amount of decrease in impurity as the classification is performed, and the other is the sum of the amount of decrease in accuracy depending on the presence or absence of specific variables (Mean Decrease Accuracy). However, since both indicators adapt biasedly to the order of variables in the tree structure, there is a disadvantage in that the influence of classification is provided at a larger value than the actual value. According to [12], there is a disadvantage that two indicators cannot accurately determine the classification influence because they cannot distinguish false correlation due to data characteristics. The paper [12] has therefore proposed a technique to measure the influence of conditional variable classification to solve this problem. However, this technique has the limitation that it cannot accurately grasp the influence of relative classification between variables and inconsistently provides priority of classification influence.

This paper has the following contributions:

(1) We propose a measuring technique of variable impacts based on Shapley Value method on random forest regression. The proposed method attempts to solve the problem that highly correlated variables gain relatively high contribution no matter what their real contribution in prediction is.

(2) We proposed a method that demonstrates the impact of variable coalitions. Considering that not only individual variables are important but also the variable impact of variable sets is, our proposed method is able to inspect the interaction between variables. It will increase the overall accuracy of a variable when a high priority of classification influence is improved when it is used as a partitioning variable in the tree.

(3) Finally, we propose a coherent ranking of variable impacts based on the marginal contribution of each variable.

The rest of this paper is organized as follows. In Section 2, we describe related work about variable impact measurement in random forest regression algorithm. In Section 3, we explain the economics theoretical method Shapley value with its basic structure. In Section 4, we propose a Shapley Value-based variable impact measurement method. In Section 5, we show the experiments with previous methods and our proposed method. In Section 6, we summarize our research and conclude the paper.

2. Related Work

In this section, we discuss the previous research for measuring variable impact index. In Section 2.1, we introduce the previous research about variable impact measuring technique in a random forest. In Section 2.2, we describe several data mining algorithms that applied Shapley Value.

2.1. Variable Impact Measurement Index. We explain the related research of variable impact measurement index in a random forest. The representative methods of the variable impact measurement index are Mean Decrease Impurity (MDI) and Mean Decrease Accuracy (MDA) proposed by Breiman [2]. Also, to improve its performance, Strobl et al. [12] proposed conditional variable impact measuring technique for random forests.

2.1.1. Mean Decrease Impurity. Breiman [2] proposed the variable impact measurement index called MDI based on impurity. Data impurity index was used to decide where we want to make a split and variables that are often made to make a split. Therefore, the MDI assumes that the amount of impurity reduction when the individual variable is selected as the partition node is the contribution in the random forests. Therefore, the sum of the impurity reductions in all the trees is calculated as the importance of the variable. For impurity reduction, classification trees use Gini coefficient index or information gain and regression trees use mean value of variables.

The equation of variable importance (VI) for variable x_j is as follows. To calculate variable importance for MDI method, it adds up the decrease of Gini index of each of the variables from 1 to n_{tree}, which means the number of trees, and gets the average of all.

The Formula of Mean Decrease Impurity [12]

$$\mathrm{VI}\left(x_j\right) = \frac{1}{n_{\text{tree}}}\left[1 - \sum_{k=1}^{n_{\text{tree}}} \mathrm{Gini}\left(j\right)^k\right]. \tag{1}$$

MDI has the advantage of being easy to compute, but it has the disadvantage that it can be biased only for categorical variables that contain multidimensional attributes. For example, if there are continuous variables and categorical variables that contain several classes, this means that the variables are more likely to be biased because they can be judged to be more superficially partitioned when categorical variables are selected under the same conditions. When attempting to split a tree into a specific variable, the most effective partitioning is the moment when the impurity is lowest. If the degree of impurity is reduced to a maximum by a single partition, this partition is considered to be an efficient partition, which means a high contribution to tree partitioning.

On the contrary, when attempting to divide into a specific variable, if the amount of decrease in impurity before and after the division is 0, it is meaningless to perform the division because the data is not classified through the variable. Therefore, in this case, the importance of the variable is judged to be zero.

2.1.2. Mean Decrease Accuracy. MDA is also called permutation importance. This is because when a decision tree is created based on a set of learning datasets divided through subsampling, the intuition behind permutation has an importance that is not a useful feature for predicting an outcome. OOB (Out-Of-Bag) is one of the subsampling techniques to calculate prediction error of each of the training samples utilizing bootstrap aggregation. MDA is the method that calculates variable importance by permutation and the method uses OOB to divide its sample data. In other words, OOB estimates more accurate prediction value by computing OOB accuracy before and after the permutation of variable x_j and compute the difference.

Since $t \in \{1, 2, 3, \ldots, n\text{tree}\}$, the variable importance of x_j in tree t is the averaged value of the difference between predicted class before permuting x_j, which is $y_i = f(x_i)$, and

Table 1: Regression coefficient simulation design [12].

X_j	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
β_j	5	5	2	0	−5	−5	−2	0	0	0	0	0

after permuting variable x_j, which is $y_i = f(x_i^j)$, in certain observation i.

The Formula of Mean Decrease Accuracy [12]

$$\mathrm{VI}\left(x_j\right) = \frac{1}{n_{\mathrm{tree}}} \cdot \sum_{t=1}^{n_{\mathrm{tree}}} \frac{\sum_{i\in\mathrm{OOB}} I\left(y_i = f\left(x_i\right)\right) - \sum_{i\in\mathrm{OOB}} I\left(y_i = f\left(x_i^j\right)\right)}{|\mathrm{OOB}|}. \quad (2)$$

2.1.3. Conditional Variable Importance. Strobl et al. [12] identified the bias selection problem in MDI and MDA. Both methods are sensitive when it comes to selecting split variables so that the selected variables are biased. In the case of predictive variables with a false correlation, the influence of variables is overestimated. This suggests a way to conditionally replace the variable X_j within the range of the specified variable $Z = X_1, X_2, X_3, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$ through splitting by random permutation in which the variable X_j of the input data is replaced with the independent variable Z. The research shows a simulation to figure out the problem in Table 1. The variables above refer to the following meaning. The first row of the figure is input variable and the second row is its weights towards predictor y. In this simulation, $X_1 \sim X_4$ are correlated.

2.2. Data Mining Algorithm with Shapley Value. In this section, we examine related studies on data mining techniques applying Shapley Value. Most of the studies show that the reason for applying the Shapley Value is to grasp objectively important indicators of the variable or feature in various algorithms.

2.2.1. Feature Selection Method. Cohen et al. [3] proposed Shapley Value-based feature selection method. To deal with the curse of dimensionality to improve the accuracy of prediction, Contribution-Selection algorithm (CSA) ranks each feature by its contribution value using Shapley Value. According to the rank of feature contribution, the algorithm performs either forward selection or backward elimination. When it comes to performing the forward selection, the algorithm selects certain number of features from the highest contribution values. Otherwise, it selects features from the lowest contribution values to eliminate.

2.2.2. Multiple Regression Analysis. Lipovetsky and Conklin [13] used the Shapley Value to analyze the relative importance of predictive variables in the multiple regression models. Multiple regression analysis is a statistical analysis technique that estimates the causal relationship between two or more independent variables. Shapley Value compares the average of all possible subsets within the prediction model to improve the prediction accuracy by calculating the importance of individual variables.

2.2.3. Multiagent Reinforcement Learning. In a dynamic environment where multiple agents communicate with each other, each agent looks for a single equilibrium point to determine its behavior. In this study, Bowling and Manuela [14] combine the Shapley Value with a probabilistic model that combines the Markov Decision Processes and matrices for efficient reinforcement learning.

3. Shapley Value Model

In this section, we explain about Shapley Value model which corresponds to the game theory of economics area. We explain Shapley Value for each step in Sections 4.1, 4.2, and 4.3.

3.1. What Is the Shapley Value? Shapley Value was proposed by Lloyd Shapley in 1953, the theory about fair distribution with players in a mutual interest relationship in a cooperative game situation. In game theory, the game can be divided into two types. One is a cooperative game in which players form certain coalitions by mutual agreement to maximize their communal payoff and the other is a noncooperative game where players maximize the interests by acting individually rather than from any collaboration with each other. According to the Shapley Value, players form coalitions and create certain common payoff. Players in each coalition receive differentiated payoff based on the fair distribution of their contributions using Shapley Value.

3.2. Basic Structure. The following concept is used to describe the Shapley Value [15]. First, there is a player who wants to participate in the game.

According to [15], the theorem is defined with a given coalitional game (N, v). There is a unique payoff division $x(v) = \varphi(N, v)$ that divides the full payoff of the grand coalition and that satisfies the *symmetry, dummy* player, and *additivity* axioms. According to the theorem, Shapley Value follows the axiom to make fair distribution towards players in the coalition.

First, Shapley Value follows the *symmetry* axiom that distributes benefit to the player who made the same amount of contributions.

Axiom 1 (see [15]). For each π in $\prod(U)$, $\Phi_{\pi i}[\pi v] = \Phi_i[v]$.

Second, Shapley Value follows the *efficiency* axiom that distributes the collective payoff generated within the coalition to the payoff with any remainder.

Axiom 2 (see [15]). For each carrier N of v, $\sum_n \Phi_i[v] = v(N)$.

Third, Shapley Value follows the *additivity* axiom, which is also called the law of aggregation. This axiom describes two games X and Y, and the sum of $v(X)$ and $v(Y)$ should be the same as $v(X+Y)$.

Axiom 3 (see [15]). For any two games v and w,

$$\Phi[v+w] = \Phi[v] + \Phi[w]. \tag{3}$$

The Shapley Value is a theory that equitably and reasonably distributes the collective payoff generated from the coalition to its players. Therefore, the following formula is used to calculate the Shapley Value for the player i, assuming the probability that each player will be placed in any order in each coalition will be the same. The Shapley Value can be obtained by averaging the marginal contributions that can be obtained when players are placed in any order within the coalition through the following formula. The equation is defined by Shapley [15] as follows.

Given a coalition game N, the Shapley Value of player i is given by the following.

The Formula of Shapley Value [15]

$$\Phi_i(v) = \sum_{S \subseteq N\{i\}} \frac{|S|!\,(n-|S|-1)!}{n!}\left[v(S\cup\{i\}) - v(S)\right]. \tag{4}$$

A set of players is formed $N = \{1, 2, 3, \ldots, n\}$. Thus, n players of the total set N are constructed. Secondly, there is a coalition formed by the players in the cooperative game situation to maximize their payoff. In this case, S means all subsets of the total set N, and the coalition that included all of n players is called grand coalition. Thirdly, there is a payoff that players are willing to gain from the coalition. The value of the subset S for the entire set N is represented by the characteristic function $v(S)$.

4. Measuring Variable Impacts

In this section, we explain the proposed method. In this study, we propose a method to apply Shapley Value from the game theory to solve the problems of the previous research for the variable impact of the random forests algorithm. Our research follows five steps' process. The details are as follows.

4.1. Contribution Calculation Step. First of all, we calculate each contribution of variables. When we generate various regression trees in random forests algorithm, we traverse each of the tree paths to assign each value of variables used in regression trees. We can assign a single path per a single coalition. We perform this contribution calculation step based on the MDA method, which permutes random variables to calculate the prediction accuracy of variables so that we are able to calculate the marginal contribution of each variable.

4.2. Construction Step. Secondly, we construct coalitions of all variables used in random forests. We consider coalitions for individual variables as players of the cooperative game situation by connecting the contributions of specific variables. Each variable has its own payoff according to the joint contribution with each coalition. Figure 1 describes the step.

4.3. Assignment Step. Thirdly, we assign each coalition with their contribution values. We assign value in every coalition.

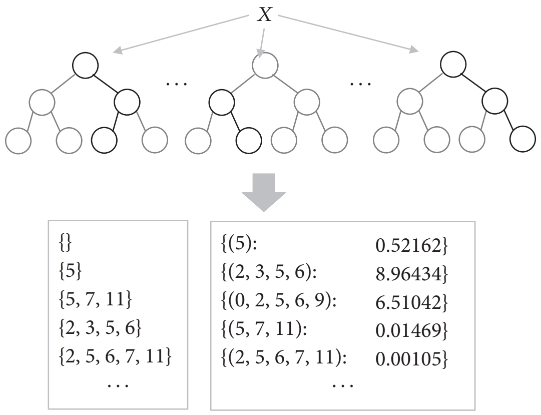


Figure 1: Construction step in Shapley Value applied method.

In this case, the number of coalitions is the same as the power set for all the variables used in the regression tree. We compare the coalition formed in step 5.2 with the power set of the variables used in random forests. If a variable does not belong to the same tree path and a value for all power sets is not assigned, the value of this coalition is assumed to be zero. This is because the coalition determined that there is no contribution to prediction accuracy since it is a coalition that did not contribute to the regression tree.

4.4. Calculation Step. Fourth, we calculate variable impact using Shapley Value method. We combine variables and their contribution as {(key): value} structure. Based on step 5.3, the assignment of both the variable and the contribution leads to obtaining the Shapley Value to figure out the variable impact of the individual variable.

4.5. Ranking Step. Finally, we provide a coherent ranking based on the variable impact. Shapley Value is calculated for the impact of individual variables as well as the priority value of the variable impact based on the value assigned to the contribution of the coalition. In this case, the ranking can be considered not only for rankings for individual variables but also for impact on the value of coalition. It is possible to line up the highest ranking of variable impact or the lowest ranking. In future work, we can use this ranking as dimension reduction method to improve prediction accuracy rate.

5. Experiments

In this section, we measure the variable impact by using Shapley Value-based method in random forest regression. On the experiments, we compare variable impacts with other measuring techniques which are previously researched with our proposed method: MDI and MDA.

The experimental environment is Intel(r) Core(TM) i7-6700HQ CPU @ 2.60 GHz/2.592 GHz, RAM 16.0 GB, x64 Windows OS. We use R and Python 3.5 as programming languages: we mainly use Python for the experiments in variable impact measure techniques and we use R for the data visualization and application towards previous works for random forest algorithm with the randomForest package.

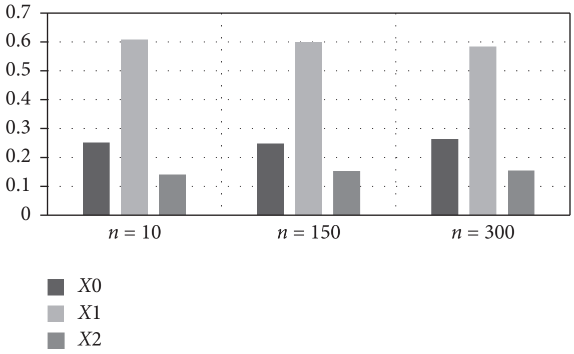


Figure 2: The variable impact of MDI.

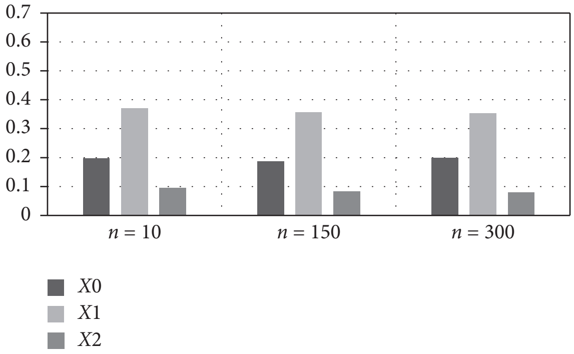


Figure 3: The variable impact of MDA.

Table 2: Variable impact of three variables ($n$tree = 10).

	MDI	MDA	SVC
x_0	0.2509	0.1980	0.0185
x_1	0.6085	0.3709	0.0229
x_2	0.1407	0.0952	0.0201

Table 3: Variable impact of three variables ($n$tree = 150).

	MDI	MDA	SVC
x_0	0.2479	0.1865	0.0276
x_1	0.6003	0.3561	0.0267
x_2	0.1518	0.0833	0.0314

Table 4: Variable impact of three variables ($n$tree = 300).

	MDI	MDA	SVC
x_0	0.2629	0.1996	0.0339
x_1	0.5841	0.3536	0.0260
x_2	0.154	0.0801	0.0137

Experiment 1. In the previous experiment, we figure out the bias selection problem in previous variable impact measuring techniques: MDI and MDA. We set certain formula to simplify the problem. Assuming that there is a formula $y = x_0 + x_1 + x_2$, we have predictor y and three variables x_0, x_1, and x_2. The variables equally contribute to predictor value because predictor y is a sum of three variables. Therefore when we measure the variable impact of variables, each of the three variables has to be equal to the same variable impact.

However, MDI and MDA show certain bias in variable selection stage. Variable impact measurement results of x_0, x_1, and x_2 do not have the same impact. For the brief description of the bias selection, we compare the experiments with 10 regression trees of MDI and MDA in random forest performance. The parameter $n$tree means a number of split variables. That is, when $n$tree = 1, we select one variable as a split point of regression tree. The number of randomly generated input variables are 10000 and its mean value is 0 and its standard deviation is 0.1.

Figures 2 and 3 show the experiment result about bias selection in MDI and MDA. Even if the weights of all variables are the same as one, the result shows that x_1 has the highest variable impact of those variables.

When $n$tree = 1, the probability that one of the three variables is selected as a split variable is equal to one-third, so that the variable impact is distributed relatively the same. However, we can see there is a bias selection when it comes to choosing more than one split variable. Nevertheless, the measurement of MDI, which measures variable impact through data impurity reduction, shows that x_2 has a higher variable impact than x_0 or x_1.

To solve this bias selection problem, we applied the Shapley Value-based technique. We generated 10, 150, and 300 regression trees, respectively, to measure the variable impact. Table 2 shows the classification impact measure based on the generated tree.

As shown in Tables 2, 3, and 4 which describe the variable impacts of three techniques, the influence of x_1 is about twice that of the other two variables. Since all variables have the same weight for the predictive variable y, this measure of influence is biased. Comparing the performance when $n = 10$, MDI measured that the influence difference between the variables was large, and, in particular, x_1 and x_2 showed a difference of two times or more. MDA measurement also showed that x_2 had the highest influence as MDI. Also, these results were not affected by the number of trees. Therefore, even if arbitrary trees are generated through random forests, there is no significant difference between the biased variable influences provided by MDI and MDA.

However, the proposed method based on Shapley Value (SVC) reduces the difference in influence between these variables. As Table 2 shows, there is almost no difference in variable impact between variables in the calculated performance through Shapley Value-based method when $n$tree = 10. Table 3 shows the experiment result of three variable impact measurement techniques with $n$tree = 150. Table 4 shows the result when n is 300. We show that the value of SVC seems relatively similar, especially when n is 10 and 150.

Unlike the previous method, which had a large difference in variable impact between variables, the Shapley Value-based method suggested a solution to this bias problem by reducing the difference in influence between variables.

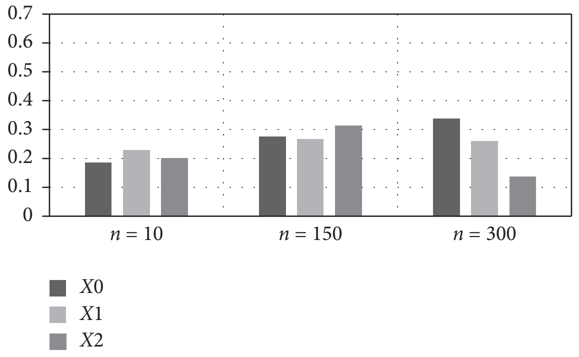
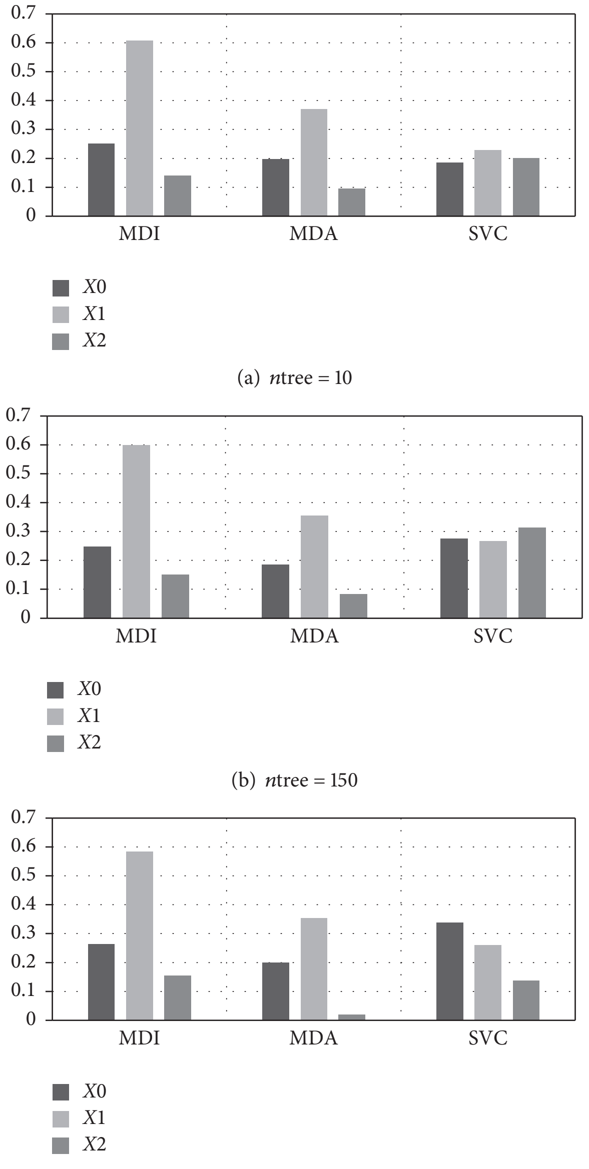


Figure 4: The variable impact of SVC.



(a) $n$tree = 10

(b) $n$tree = 150

(c) $n$tree = 300

Figure 5: The variable impact of three techniques.

Figures 4 and 5 show the variable impacts of each technique based on Tables 2, 3, and 4. MDI and MDA represent clear importance of x_1 that does not have any significant importance. The variable importance maintains its ranking no matter how many trees are constructed.

On the other hand, the result of SVC are shown in Figure 5, and the graphs indicate that the importance of x_1 is estimated to be similar to other variables rather than other techniques like MDI and MDA. Since the impacts of variables are the same in the formula above, the smaller difference of x_0, x_1, and x_2 means more accurate importance has been estimated.

However, our proposed method has a limitation that the variation of the variable impact range greatly occurs according to the number of tree parameter. The drawback comes out when $n$tree is 300. In this experiment, the difference between variables impact in the case of $n = 10$ and $n = 150$ was significantly reduced, but the difference between x_0 and x_2 in the case of $n = 300$ is more than twice. Although SVC seems to be biased in calculating variable impacts when $n$tree = 300, the method still performs better interpretation than MDI or MDA.

Figure 6 is a boxplot graph that shows the range of variable impacts based on MDI technique. The axis x is variable impact of variables and the axis y is three variables. As the result above already showed, the technique provides biased impacts towards the values x_0, x_1, and x_2. Figure 7 shows the variable impact of MDA technique. The variables seem to have similar importance to the result of MDI technique.

However, Figure 8 shows the experiment with SVC. The importance of three variables seems more average than other techniques, even if the range seems unstable. Since the previous research demonstrates that Shapley Value selects entirely different variable with other classifiers in the case of feature selection technique [15], the contribution values calculated from the candidate variables are modified within iterations.

Therefore if there is a significant change in the combination of candidate variables, the range of variable impact can be fluctuated. In terms of the influence of x_0 and x_2, the proposed method SVC is limited to provide a complete solution yet there are still biases which were presented in MDI and MDA. In SVC technique, the variable impacts show huge fluctuation. The reason for the fluctuation is that Shapley Value calculates every coalition to evaluate the impact of each variable x_j on the classification. Since random forest algorithm constructs the coalitions randomly and not every coalition affects the result, the payoff of the coalitions which are not initially constructed from random forest are considered to be 0. It is because the coalition without any contribution towards result has no payoff. Therefore, there is a fluctuation when we average all marginal contribution of those coalitions. The issue causes the instability of variable impacts from SVC.

Yet, the contribution of this research is that even if the fluctuation of a range of variable impact is larger than MDI or MDA, SVC can be judged more reliable on relative

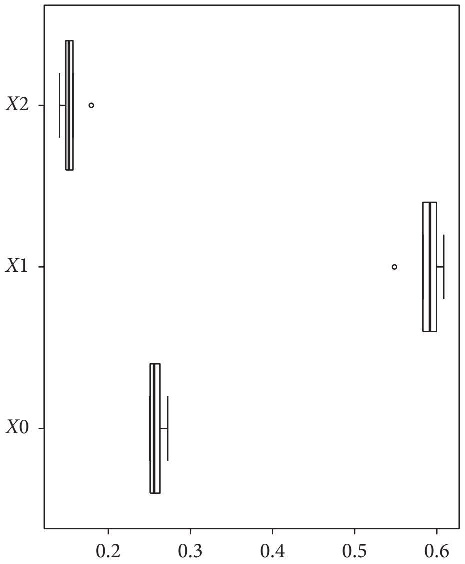


Figure 6: A boxplot graph of variable impacts with three variables on MDI technique.

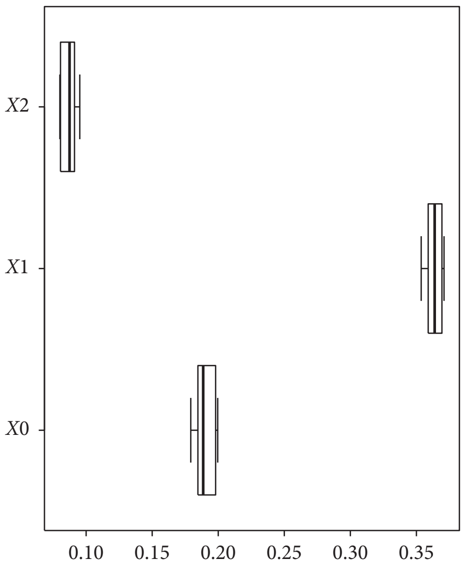


Figure 7: A boxplot graph of variable impacts with three variables on MDA technique.

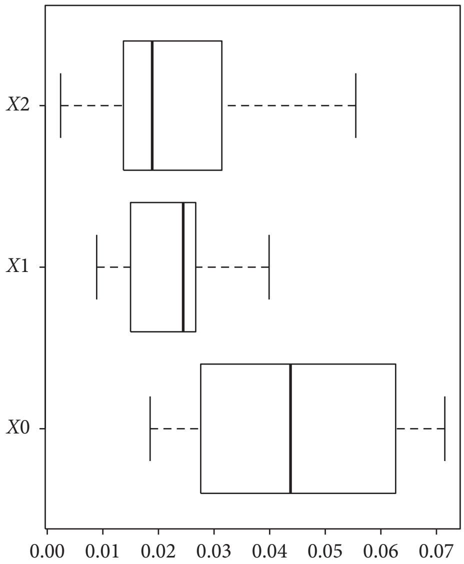
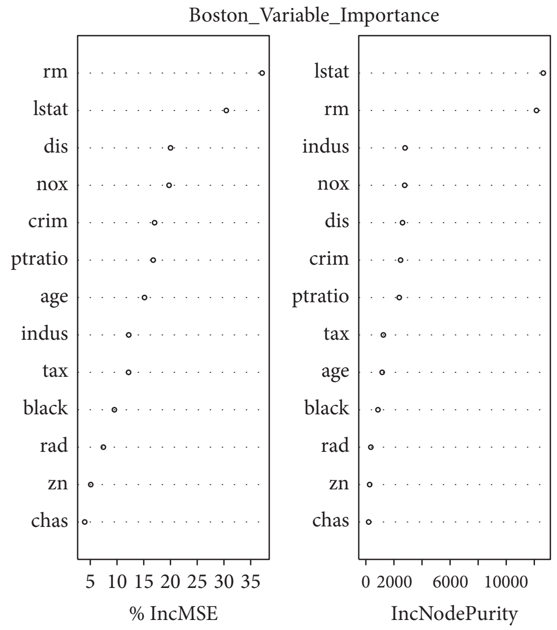


Figure 8: A boxplot graph of variable impacts with three variables on SVC technique.



(a) MDA

(b) MDI

Figure 9: The graph of variable impact measurement of random forest regression.

relationship towards variables. Our proposed method gave a significant reduction of bias that was provided from MDI and MDA in variable impact calculation.

Experiment 2. In this experiment, we use a real dataset which is named Boston Housing Data [11]. Boston Housing Data provides 506 instances and 13 attributes that affect the housing price of Boston. The description of Boston Housing Data is shown in (ii) in the Notations.

Figure 9 shows variable impacts measured by previous methods. A graph on the left shows variable impacts measured by MDA method, and another graph on the right shows variable impacts measured by MDI method from *randomForest* package in R library [16]. The highest top two variables are RM and LSTAT in those variable impacts measurement index.

However, this ranking is not considerable on variable impacts between correlated variables. For example, there is a phenomenon called multicollinearity issue. Multicollinearity means that more than two input variables are highly correlated so that the impact of those variables is overestimated. Since the phenomenon spoils the relevant importance

Table 5: A comparison of variable impact measuring technique on Boston Housing Data (*n*tree = 20).

	MDI	MDA	SVC
CRIM	0.0164	0.012	0.0701
ZN	0.0015	0.0009	0.0240
INDUS	0.052	0.0206	0.007
CHAS	0.0009	0.0001	0
NOX	0.0119	0.0065	0.057
RM	0.5842	0.803	0.1705
AGE	0.0307	0.0118	0.0718
DIS	0.0177	0.0108	0.0733
RAD	0.003	0.0013	0.0678
TAX	0.017	0.0118	0
PTRATIO	0.0387	0.0203	0.0889
B	0.0123	0.0021	0.0648
LSTAT	0.2138	0.2404	0.3555

between input variable and predictor, we need to minimize the possibility of multicollinearity.

In this data, there is a high correlation among NOX, INDUS, and TAX. INDUS means the proportion of nonretail business acres per town and NOX means nitric oxides concentration. It is inferable that INDUS and NOX have a positive correlation: as the proportion of industrial area increases, the ratio of nitric oxides concentration also increases. The tax ratio increases when INDUS is increased. Therefore, INDUS, NOX, and TAX are highly correlated.

However, the impact of those correlated variables is relatively high. Even though those variables gained rather smaller contribution than LSTAT or RM, the ranking should be reliable in order to make a reliable decision. It is more efficient if we use only one of those variables which are correlated to each other. Eliminating unnecessary variables due to the variable impact ranking reduces dimensionality. In order to resolve this problem, we use our proposed method. The experiment steps are followed.

First, we compare the prediction accuracy of the random forest regression tree when we permute a certain variable randomly so that the marginal contribution of the specific variable to be calculated. Second, we construct coalitions for individual variables as players of the cooperative game situation by connecting the contributions of specific variables. Each variable has its own payoff according to the joint contribution with each coalition. Third, we assigned a power set as a coalition of N with contribution values that we calculated by MDA. Finally, based on the {(key): value} structure, we calculated Shapley Value.

We implemented MDI and MDA in Python to compare the variable impact measurement results with our Shapley Value-based method proposed in this research. In MDA, we shuffle dataset 10 times for permutation towards random variables. We used a cross-validation technique that permutes the variables randomly for comparison.

Tables 5, 6, and 7 show the result of our experiments. Table 5 shows the result of variable impacts of measuring techniques of random forest regression by generating 20 trees. Table 6 generates 50 trees and Table 7 generates 100 trees to measure variable impact. In this experiment, we unify *mtry* parameter as 6. It means that regression trees in random forests select 6 variables as a split criterion to perform prediction. Instead, each experiment has a different *n*tree parameter, which is the number of trees in random forests. It means that tree generation process iterates 20, 50, and 100 times.

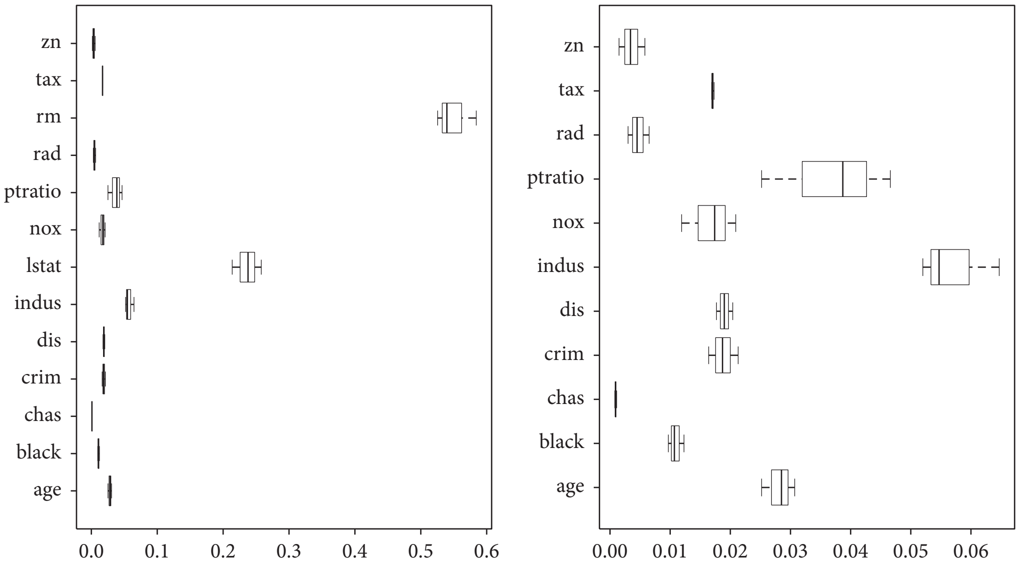
Table 6: A comparison of variable impact measuring technique on Boston Housing Data (*n*tree = 50).

	MDI	MDA	SVC
CRIM	0.0187	0.0125	0.0954
ZN	0.0034	0.0008	0.0620
INDUS	0.0547	0.02	0
CHAS	0.0008	0.0001	0.0069
NOX	0.0209	0.0077	0.0887
RM	0.5396	0.7758	0.4954
AGE	0.0252	0.0138	0.0913
DIS	0.019	0.0105	0.0104
RAD	0.0065	0.0013	0.0972
TAX	0.0173	0.0118	0.0996
PTRATIO	0.0252	0.0178	0.0286
B	0.0107	0.0024	0.0958
LSTAT	0.2579	0.2337	0.3652

Table 7: A comparison of variable impact measuring technique on Boston Housing Data (*n*tree = 100).

	MDI	MDA	SVC
CRIM	0.0213	0.012	0.0144
ZN	0.0058	0.001	0.0023
INDUS	0.0647	0.0173	0
CHAS	0.0011	0.0001	0.0202
NOX	0.0174	0.0062	0.0172
RM	0.5255	0.7853	0.483
AGE	0.0285	0.0126	0.021
DIS	0.0204	0.0097	0.0226
RAD	0.0045	0.001	0.02
TAX	0.0169	0.0119	0.0075
PTRATIO	0.0466	0.0162	0.0171
B	0.0097	0.0015	0.0192
LSTAT	0.2378	0.2258	0.982

The value of SVC with *n*tree = 100 in Table 7 changes rather than other values of SVC. Since random forest constructs trees randomly and uses only sampled trees to their classification, there are sparse coalitions that are not constructed from the model and have no contribution to the classification result. However, when it comes to calculating the contribution of each variable, we have to consider all cases of a coalition even though it has no contribution. Those coalitions are considered as dummy players, so we assign 0 payoff due to the axiom. When the number of trees got bigger, the number of dummy player coalitions which have no value



(a) The variable impacts with total variables

(b) The variable impacts with minor variables

Figure 10: The boxplot of variable impacts in MDI.

also increased. Finally, since we average all values of coalition to calculate SVC, the value of SVC is decreased.

As we mentioned before, RM and LSTAT are notably the most important variable in Boston house price prediction. Regardless of the number of the tree trained by random forests, RM and LSTAT are on the highest ranking. Also, we can see the result of our proposed method, which refers to Shapley Value-based Calculation, the highest ranking maintains the same as MDI and MDA.

We figure out that our proposed method solves the multicollinearity issue in biased variable impact measurement in MDI and MDA. Table 6 shows that the variable impact of INDUS, NOX, and TAX is almost similar in MDI. In the case of MDA, the variable impact of NOX is relatively low compared to INDUS or TAX. Still, the impact of those three variables is on average, which means three variables are all considered as a split criterion in random forests.

On the other hand, our proposed method reduces the possible multicollinearity problem. When $n$tree = 20, the impact of INDUS and TAX is close to zero. NOX has the variable impact as 0.057. The SVC only consider NOX as a split criterion in prediction and decide not to distribute an evident impact to INDUS and TAX since they are highly correlated to each other. Even when $n$tree = 50 or $n$tree = 100, the result of Shapley Value-based method shows that the variable impact of INDUS is zero.

The result of Shapley Value-based method is that the variable impact of CHAS is zero. CHAS refers to Charles River dummy variable. The variable seems relevant in both MDI and MDA for the lowest variable impact. So far it is possible to eliminate CHAS variable as a dummy variable which does not contribute to any prediction. For a null player with no contribution, it is the result of the axiom of the Shapley Value that the payoff is not distributed.

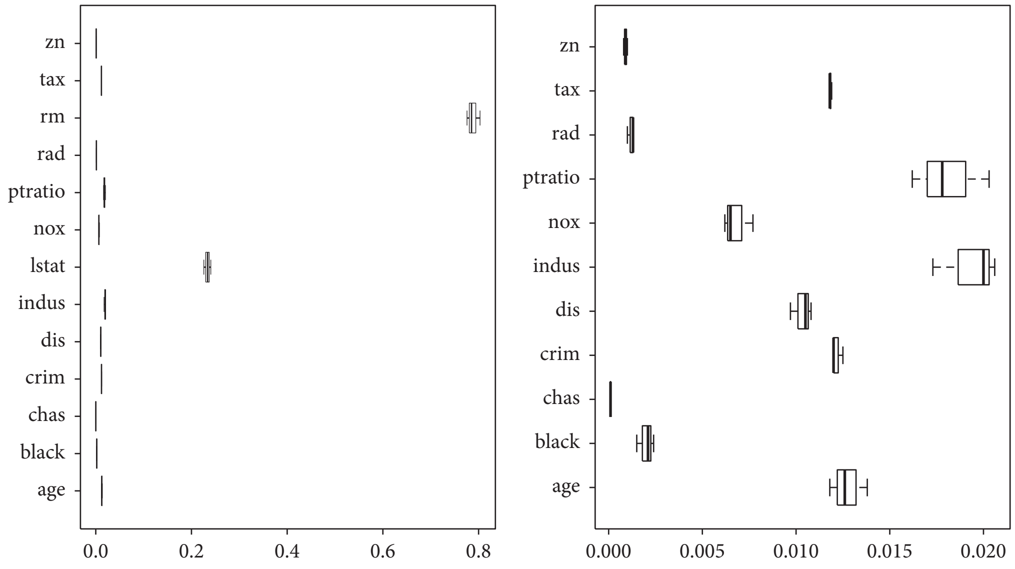
Table 8: Coherent ranking based on SVC.

Rank	Variable coalition	Value
1	{RM}	2.6902
2	{RM, TAX, LSTAT}	0.7613
3	{RM, CRIM, LSTAT}	0.5866
4	{RM, LSTAT}	0.5694
5	{LSTAT}	0.4391
6	{RM, DIS, LSTAT}	0.3147
7	{RM, INDUS}	0.2797
8	{RM, CRIM, AGE, LSTAT}	0.2589
9	{RM, INDUS, AGE, LSTAT}	0.2438
10	{RM, DIS}	0.2247

Figure 10 shows the boxplot graph of the variable impact of MDI. The graph (a) shows all the variable impacts of Boston Housing Data and graph (b) shows only the minor variables. Since RM and LSTAT have a major impact on prediction, the possible variance of minor variables can be omitted on graph (b). Figure 11 shows the boxplot graph of MDA and Figure 12 shows the boxplot graph of SVC.

However, our proposed method reveals the limitation in the experiment that SVC method provides highly unstable variable impacts rather than other techniques. The fluctuation of the range of variable impact seems to distract the experiment result. However, the average variable impact of SVC methods shows better importance than other techniques.

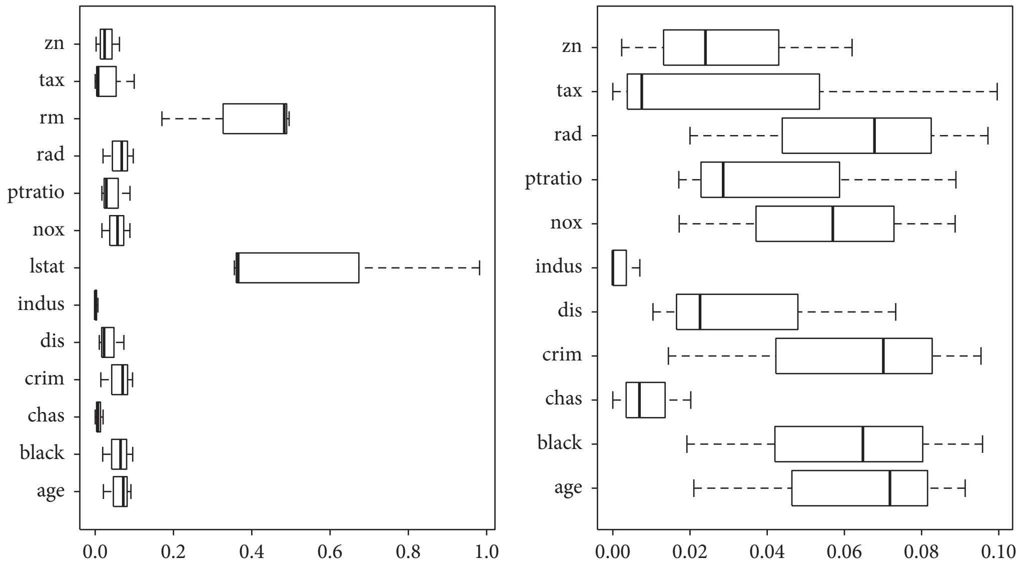
Table 8 shows the comparison of highly correlated variables TAX and INDUS whose impacts are calculated by each variable impact measuring technique. Even though the variance of Shapley Value-based method is the largest, the



(a) The variable impacts with total variables

(b) The variable impacts with minor variables

Figure 11: The boxplot of variable impacts in MDA.



(a) The variable impacts with total variables

(b) The variable impacts with minor variables

Figure 12: The boxplot of variable impacts in SVC.

mean value is meaningful. We find that multicollinearity issue is solved by the variable impacts of INDUS and TAX calculated with Shapley Value method.

6. Conclusion

In this paper, we proposed a method to measure the influence of variables using Shapley Value method in random forest algorithm. One of the existing methods for measuring the classification impact of variables is the Mean Decrease Impurity technique, which uses Gini coefficients to determine the influence of variables through data impurity reduction. The other is the Mean Decrease Accuracy method, which limits the influence of classification by calculating the difference in the prediction accuracy of the changing data by permitting the variable. Both indicators are commonly used to measure the classification impact of variables using real data.

Our proposed approach performs better than other approaches for two main reasons. First, our approach tries to solve the multicollinearity problem in other techniques. In the previous approach, the variable impact calculation was less accurate because of the correlated variables. In this paper, we proposed Shapley Value-based approach so that the payoff is fairly distributed by its contribution among variables.

Second, our approach considers not only the impact of the individual variable but also the impact of the group of variables. There are synergies between variables which perform effectively when those are combined. Previous approaches did not consider the impact of the group. However, in this paper, we would like to consider the impacts of the group so that we can inspect the synergies between variables.

Through this research, we have made the following three contributions. First, this paper presents the problems of existing techniques for finding the influence of variable classification using the random forest and tries to solve it by combining Shapley Value of economics theory. As Shapley Value is applied to a variety of machine learning or data mining algorithms, it is the first study to incorporate the Shapley Value of economics theory to measure the exact classification impact of random forests. Second, we can obtain the priority of the variables that affect the accuracy of the classification result through the proposed method. The proposed method improves the accuracy of random forest prediction based on this priority. Finally, this research improves the analytical power of the black box model. The interpretation of variable importance is critical in the classification problem. Our proposed method is suitable for measuring variable impact in black box model such as random forest. Furthermore, the algorithm is applicable in mobile cloud computing environment.

In future work, we will conduct the experiments with several different data. Moreover, we will research about reducing the complexity so that we could improve the performance of variable impacts measuring techniques based on Shapley Value.

Notations

(i) The Notation of Cooperative Game Theory [12]

n: Player willing to participate in cooperative game
N: Total set of n players
S: Coalition of the players who share common payoff
$v(S)$: A payoff that players gain from the coalition.

(ii) The Description of Boston Housing Data [11]

CRIM: Per capita crime rate by town
ZN: The proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS: Proportion of nonretail business acres per town
CHAS: Charles River dummy variable (=1 if tract bounds; 0 otherwise)
NOX: Nitric oxides concentration (parts per 10 million)
RM: Average number of rooms per dwelling
AGE: Proportion of owner-occupied units built prior to 1940
DIS: Weighted distances to five Boston employment centers
RAD: Index of accessibility to radial highways
TAX: Full-value property-tax rate per \$10,000
PTRATIO: Pupil-teacher ratio by town
BLACK: $1000(B_k - 0.63)^2$, where B_k is the proportion of blacks by town
LSTAT: % lower status of the population
MEDV: Median value of owner-occupied homes in \$1000.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (no. R0115-16-1009, Development of Smart Learning Interaction Contents for Acquiring Foreign Languages through Experiential Awareness).

References

[1] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.

[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] S. Cohen, E. Ruppin, and G. Dror, "Feature selection based on the Shapley value," *In Other Words*, vol. 1, 2005.

[4] O. Okun and H. Priisalu, "Random forest for gene expression based cancer classification: overlooked issues," in *Pattern Recognition and Image Analysis*, pp. 483–490, Springer, 2007.

[5] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," *BMC Bioinformatics*, vol. 9, article 319, pp. 1–10, 2008.

[6] R. Díaz-Uriarte and S. Alvarez de Andrés, "Gene selection and classification of microarray data using random forest," *BMC Bioinformatics*, vol. 7, no. 1, article 3, 2006.

[7] B. Wu, T. Abbott, D. Fishman et al., "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.

[8] A. Hapfelmeier, T. Hothorn, K. Ulm, and C. Strobl, "A new variable importance measure for random forests with missing data," *Statistics and Computing*, vol. 24, no. 1, pp. 21–34, 2014.

[9] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, vol. 27, no. 3, pp. 659–678, 2017.

[10] H. Wang, F. Yang, and Z. Luo, "An experimental study of the intrinsic stability of random forest variable importance measures," *BMC Bioinformatics*, vol. 17, no. 1, 2016.

[11] D. Harrison and D. L. Rubinfeld, "Hedonic prices and the demand for clean air," *Journal of Environmental Economics and Management*, vol. 5, pp. 81–102, 1978.

[12] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, article 307, 2008.

[13] S. Lipovetsky and M. Conklin, “Analysis of regression in game theory approach,” *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 319–330, 2001.

[14] M. Bowling and V. Manuela, *An Analysis of Stochastic Game Theory for Multiagent Reinforcement Learning*, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2000.

[15] L. S. Shapley, “A value for *n*-person games,” in *Contributions to the Theory of Games*, H. Kuhn and A. W. Tucker, Eds., vol. 2 of *Annals of Mathematics Studies*, pp. 307–317, Princeton University Press, Princeton, NJ, USA, 1953.

[16] S. RColorBrewer, A. Liaw, M. Wiener, and M. A. Liaw, “Package ‘randomForest’,” 2015.