The Korean Journal of Applied Statistics
2021, Vol. 34, No. 2, 177–190

DOI: https://doi.org/10.5351/KJAS.2021.34.2.177
Print ISSN1225-066X / Online ISSN 2383-5818

Correlated variable importance for random forests

Seung Beom Shin[a], Hyung Jun Cho[1,a]

[a]Department of Statistics, Korea University

Abstract

Random forests is a popular method that improves the instability and accuracy of decision trees by ensembles. In contrast to increasing the accuracy, the ease of interpretation is sacrificed; hence, to compensate for this, variable importance is provided. The variable importance indicates which variable plays a role more importantly in constructing the random forests. However, when a predictor is correlated with other predictors, the variable importance of the existing importance algorithm may be distorted. The downward bias of correlated predictors may reduce the importance of truly important predictors. We propose a new algorithm remedying the downward bias of correlated predictors. The performance of the proposed algorithm is demonstrated by the simulated data and illustrated by the real data.

Keywords: random forests, variable importance, correlation

1. 서론

랜덤포레스트 (Breiman, 2001)는 여러 개의 의사결정나무의 결과를 결합하여 예측 정확도를 높이는 방법으로, 분류와 회귀에서 높은 성능을 보여 데이터 분석 시에 많이 사용되고 있다. 이러한 랜덤포레스트는 결과에 대한 해석을 위해 변수중요도를 사용해왔다. 하지만 변수중요도가 편향된 정보를 전달할 때가 있다. 랜덤포레스트 모형에 사용되는 예측변수 중에 2개 이상의 변수가 선형 또는 비선형 상관관계에 있을 때 변수 중요도는 낮게 편향된 수치를 보인다. 이 현상에 대해 Breiman (2001, p.24)은 다음과 같이 설명했다. “비슷한 변수들이 함께 입력된 경우, 각 변수는 하나의 랜덤포레스트에서 비슷한 빈도로 선택된다. 이러한 과정에서 예측변수들이 서로에게 혼동을 주어, 서로의 변수중요도를 낮추는 상황이 발생할 수 있다.” 그 외에 Strobl 등 (2007), Archer 과 Kimes (2008), Nicodemus 등 (2010) 등 여러 연구자가 예측변수 간에 상관관계가 있을 때 랜덤포레스트 변수중요도가 하향 편향되는 문제를 연구했다. 더 나아가 변수중요도의 하향 편향이 발생하는 상황에서 적합한 변수 선택을 하기 위한 연구도 여러 번 진행됐다 (Genuer 등, 2010; Gregorutti 등, 2017; Ruamo, 2019). 또한, 편향 문제를 해결하기 위해 Strobl 등 (2008)이 조건부 변수중요도를 제안했다. 하지만 여전히 변수중요도의 편향 문제는 남아있다.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2018R1D1A1B07044479), by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (2019R1A4A1028134) and by a Korea University Grant (K2019631).

[1] Corresponding author: Department of Statistics, Korea University, 145 Anam-ro 145, Seongbuk-Gu, Seoul 02841, Korea. E-mail: hj4cho@korea.ac.kr

Published 30 April 2021 / journal homepage: http://kjas.or.kr

본 논문은 기존의 랜덤포레스트 변수중요도를 수정하여 상관예측변수로 인한 변수중요도 하향 편향을 개선한 방법을 제안한다. 본 논문의 구성은 다음과 같다. 2장은 분류모형에서 계산할 수 있는 기존의 변수중요도 2가지와 수정한 변수중요도 2가지의 알고리즘에 관해 설명한다. 3장은 모의실험을 통해 기존의 변수중요도에서 발생하는 변수 간에 선형 또는 비선형 상관관계가 있을 경우의 하향 편향에 대해 논의하고, 수정한 변수중요도에서 편향이 개선되는 것을 확인한다. 4장은 실제 자료의 변수중요도를 계산하여 기존의 변수중요도와 수정한 변수중요도가 어떤 차이를 보이며, 그에 따라 수정한 변수중요도가 어떤 유용성을 가졌는지 논의한다.

2. 분류 변수중요도

본 논문에서 제안하는 변수중요도는 기존의 변수중요도 2가지의 알고리즘을 예측변수 간에 선형 또는 비선형 상관관계가 있더라도 변수중요도가 하향 편향되지 않도록 수정한 것이다. 단, 본 논문의 2장과 3장, 4장에서는 분류 문제에 관한 내용만 다룬다. 회귀 문제의 수정한 변수중요도 알고리즘과 그에 따른 결과는 분류 문제의 경우와 크게 다르지 않다.

하나의 랜덤포레스트 생성과정은 다음과 같다. $p\,(j = 1, \ldots, p)$개의 예측변수 $X \in R^p$를 이용하여 C개의 범주 중에 하나의 값을 갖는 $Y \in \{1, \ldots, C\}$를 예측하기 위해 n개의 학습 자료인 $D = \{\mathbf{X}_i,\ Y_i\}_{i=1}^n$가 주어졌을 때, $M(m = 1, \ldots, M)$개의 의사결정나무를 생성한다. 이때, m번째 의사결정나무를 만들기 위해 D를 확률변수 Θ_m으로 추출한 $D(\Theta_m)$을 사용하며, 이렇게 생성된 m번째 분류기를 $h(\mathbf{X}, \Theta_m)$라고 하자. 랜덤포레스트는 M개의 분류기로 분류된 결과를 다수결의 원칙에 따라 가장 많이 분류된 범주로 Y를 예측한다. 이렇게 생성한 랜덤포레스트로부터 다음의 변수중요도를 구할 수 있다.

2.1. MDI 변수중요도

분류모형에서 사용할 수 있는 첫 번째 변수중요도는 mean decrease impurity (MDI)이다 (Breiman 등, 1984; Biau과 Scornet, 2016). 이 척도는 의사결정나무에서 예측변수가 마디를 분리함으로써 발생하는 지니(Gini) 불순도(impurity)의 감소량을 누적시켜 M개의 의사결정나무에 대해 평균을 구한 값이다. MDI 변수중요도를 계산하는 알고리즘은 아래와 같다.

Step 1. 특정 중간 마디를 분리할 때 지니 불순도 감소를 최대로 하는 변수를 선택한다.

$$\begin{aligned}\underset{j}{\operatorname{argmin}}\,\Delta g(X_j, t) = \underset{j}{\operatorname{argmin}}\Bigg[&\left(1 - \sum_{j=1}^{C} \Pr(Y = y|t)^2\right) \\ &- \frac{N(t_L)}{N(t)}\left(1 - \sum_{j=1}^{C} \Pr(Y = y|t_L)^2\right) - \frac{N(t_R)}{N(t)}\left(1 - \sum_{j=1}^{C} \Pr(Y = y|t_R)^2\right)\Bigg]. \qquad (2.1)\end{aligned}$$

이때, $t_L = \{x \in t;\ X_j < z\}$, $t_R = \{x \in t;\ X_j \geq z\}$이고, $N(t)$, $N(t_L)$, $N(t_R)$은 각각 중간 마디 t, t_L, t_R에 속하는 표본의 수를 의미한다.

Step 2. m번째 의사결정나무 T_m에서 j번째 예측변수 X_j가 감소시킨 불순도를 누적한다.

$$G(X_j, T_m) = \sum_{t \in T_m} \Delta g(X_j,\ t) I\left(\underset{j}{\operatorname{argmin}}\,\Delta g(X_j,\ t) = j\right). \qquad (2.2)$$

이때, $m = 1, 2, \ldots, M$

Step 3. M개의 의사결정나무에서 j번째 예측변수 X_j가 감소시킨 불순도의 평균을 구한다.

$$\mathrm{MDI}\left(X_j\right) = \frac{1}{M}\sum_{m=1}^{M} G\left(X_j,\ T_m\right). \tag{2.3}$$

Step 4. Steps 2~3의 과정을 모든 예측변수에 대해 반복하여 각 변수의 변수중요도를 계산한다.

2.2. Max MDI 변수중요도

이 변수중요도는 상관예측변수의 중요도 하향 편향 문제를 해결하기 위해 MDI를 수정한 척도이다. 이것은 예측변수 간에 상관관계에 있는 변수가 없으면 변수중요도가 편향되지 않는 특징을 활용한다. 즉, 변수중요도를 알고 싶은 예측변수를 제외한 모든 예측변수의 값들을 무작위로 재배열하여 예측변수 간의 상관관계를 제거하는 것이다. 자세한 알고리즘은 아래와 같다.

Step 1. 특정 중간 마디를 분리할 때 지니 불순도 감소를 최대로 하는 변수를 선택한다.

$$\begin{aligned}\underset{j}{\operatorname{argmin}}\,\Delta g(X_j,\ t) = \underset{j}{\operatorname{argmin}}\Bigg[&\left(1-\sum_{j=1}^{C}\Pr(Y=y|t)^2\right)\\ &-\frac{N(t_L)}{N(t)}\left(1-\sum_{j=1}^{C}\Pr(Y=y|t_L)^2\right)-\frac{N(t_R)}{N(t)}\left(1-\sum_{j=1}^{C}\Pr(Y=y|t_R)^2\right)\Bigg].\end{aligned} \tag{2.4}$$

이때, $t_L = \{x \in t;\ X_j < z\}$, $t_R = \{x \in t;\ X_j \geq z\}$이고, $N(t)$, $N(t_L)$, $N(t_R)$은 각각 중간 마디 t, t_L, t_R에 속하는 표본의 수를 의미한다.

Step 2. m번째 의사결정나무 T_m에서 j번째 예측변수 X_j가 감소시킨 불순도를 누적한다.

$$G\left(X_j,\ T_m\right) = \sum_{t\in T_m}\Delta g(X_j,\ t)I\left(\underset{j}{\operatorname{argmin}}\,\Delta g(X_j,\ t) = j\right). \tag{2.5}$$

이때, $m = 1, 2, \ldots, M$.

Step 3. M개의 의사결정나무에서 j번째 예측변수 X_j가 감소시킨 불순도의 평균을 구한다.

$$\mathrm{MDI}\left(X_j\right) = \frac{1}{M}\sum_{m=1}^{M} G\left(X_j,\ T_m\right). \tag{2.6}$$

Step 4. Steps 1~3의 과정에 X_j를 제외한 모든 예측변수의 값을 각 예측변수 내에서 무작위로 재배열한 자료 $D^{(-j)*}$를 사용해 $\mathrm{MDI}^{(-j)*}(X_j)$를 구한다.

Step 5. X_j의 두 개의 수치 중에 큰 값을 수정한 변수중요도로 지정한다.

$$\text{Max MDI}\left(X_j\right) = \max\left\{\mathrm{MDI}\left(X_j\right),\ \mathrm{MDI}^{(-j)*}\left(X_j\right)\right\}. \tag{2.7}$$

Step 6. Steps 2~5의 과정을 모든 변수에 대해 반복하여 각 변수의 변수중요도를 계산한다.

2.3. MDA 변수중요도

분류문제에서 사용할 수 있는 두 번째 변수중요도는 mean decrease accuracy (MDA)이다 (Breiman, 2001; Biau과 Scornet, 2016; RColorBrewer과 Liaw, 2018). 이 변수중요도는 특정 변수가 존재함으로써 랜덤포레스트의 예측력이 얼마나 향상되는가를 나타내는 척도로, 원래의 예측변수를 사용했을 때와 예측변수의 값들을 무작위로 재배열 시켜 예측변수가 가진 정보를 제거한 후 사용했을 때 랜덤포레스트의 성능을 비교한다. MDA 변수중요도를 계산하는 알고리즘은 아래와 같다.

Step 1. m번째 의사결정나무 T_m 구축 시 추출되지 않은 out-of-bag (OOB)자료 O_m을 사용하여 예측 정확도를 계산한다.

$$A_m = \frac{1}{N(O_m)} \sum_{(x,\ Y)\in O_m} I[Y = h(x,\ \Theta_m)]. \tag{2.8}$$

이때, $N(O_m)$은 O_m의 표본 수이고, $h(x,\ \Theta_m)$은 m번째 의사결정나무 분류기에 예측변수 x를 입력한 것이다.

Step 2. j번째 예측변수 값을 무작위로 재배열한 $O_m^{(j)*}$자료를 사용하여 Step 1과 같은 방법으로 예측 정확도 $A_m^{(j)*}(X_j)$를 계산하고 $A_m(X_j)$과의 차이를 계산한다.

$$\Delta A_m\left(X_j\right) = A_m - A_m^{(j)*}\left(X_j\right). \tag{2.9}$$

이때, $A_m^{(j)*} = (1/N(O_m^{(j)*}))\sum_{(x,\ Y)\in O_m^{(j)*}} I[Y = h(x,\ \Theta_m^{(j)*})]$.

Step 3. ΔA_m를 M개의 의사결정나무에 대해 평균을 낸다.

$$\mathrm{MDA}\left(X_j\right) = \frac{1}{M}\sum_{m=1}^{M}\Delta A_m\left(X_j\right). \tag{2.10}$$

Step 4. 만약, ΔA_m의 표준편차가 0이 아닌 경우, ΔA_m의 평균을 표준편차로 나누어 정규화한다. ΔA_m의 표준편차가 0인 경우엔 Step 3의 $MDA(X_j)$를 j번째 예측변수의 변수중요도로 한다.

$$\mathrm{MDA}\left(X_j\right) = \frac{1}{M}\frac{1}{\mathrm{sd}\left(\Delta A_1\left(X_j\right),\ \Delta A_2\left(X_j\right),\ldots,\ \Delta A_M\left(X_j\right)\right)}\sum_{l=1}^{M}\Delta A_l\left(X_j\right). \tag{2.11}$$

이때, $sd(\Delta A_1(X_j),\ \Delta A_2(X_j),\ldots,\ \Delta A_M(X_j)) \neq 0$이다.

Step 5. Steps 2~4의 과정을 모든 예측변수에 대해 반복하여 각 변수의 변수중요도를 계산한다.

2.4. Max MDA 변수중요도

이 변수중요도는 본 논문에서 제안하는 두 번째 변수중요도이다. 이것은 상관관계에 있는 변수가 없으면 변수중요도가 편향되지 않는 특징을 활용한다. 즉, 변수중요도를 알고 싶은 예측변수를 제외한 모든 예측변수의 값들을 무작위 재배열하여 예측변수 간의 상관관계를 제거하는 것이다. 자세한 알고리즘은 아래와 같다.

Step 1. m번째 의사결정나무 T_m 구축 시 추출되지 않은 OOB자료 O_m을 사용하여 예측 정확도를 계산한다.

$$A_m = \frac{1}{N(O_m)} \sum_{(x,Y)\in O_m} I[Y = h(x,\ \Theta_m)]. \tag{2.12}$$

이때, $N(O_m)$은 O_m의 표본 수이고, $h(x,\ \Theta_m)$은 m번째 의사결정나무 분류기에 예측변수 x를 입력한 것이다.

Step 2. j번째 예측변수 값을 무작위로 재배열한 $O_m^{(j)*}$자료를 사용하여 Step 1과 같은 방법으로 예측 정확도 $A_m^{(j)*}(X_j)$를 계산하고 $A_m(X_j)$과의 차이를 계산한다.

$$\Delta A_m\left(X_j\right) = A_m - A_m^{(j)*}\left(X_j\right). \tag{2.13}$$

이때, $A_m^{(j)*} = (1/N(O_m^{(j)*}))\sum_{(x,Y)\in O_m^{(j)*}} I[Y = h(x, \Theta_m^{(j)*})]$.

Step 3. ΔA_m를 M개의 의사결정나무에 대해 평균을 낸다.

$$\mathrm{MDA}\left(X_j\right) = \frac{1}{M}\sum_{m=1}^{M}\Delta A_m\left(X_j\right). \tag{2.14}$$

Step 4. 만약, ΔA_m의 표준편차가 0이 아닌 경우, ΔA_m의 평균을 표준편차로 나누어 정규화한다. ΔA_m의 표준편차가 0인 경우엔 Step 3의 값을 $\mathrm{MDA}(X_j)$로 한다.

$$\mathrm{MDA}\left(X_j\right) = \frac{1}{M}\frac{1}{\mathrm{sd}\left(\Delta A_1\left(X_j\right), \Delta A_2\left(X_j\right), \ldots, \Delta A_M\left(X_j\right)\right)}\sum_{l=1}^{M}\Delta A_l\left(X_j\right). \tag{2.15}$$

이때, $\mathrm{sd}(\Delta A_1(X_j), \Delta A_2(X_j), \ldots, \Delta A_M(X_j)) \neq 0$이다.

Step 5. Steps 1~4의 과정에 j번째 예측변수를 제외한 모든 변수값을 무작위로 재배열한 $D^{(-j)*}$를 사용하여 $\mathrm{MDA}^{(-j)*}(X_j)$를 구한다.

Step 6. X_j의 두 개의 수치 중에 큰 값을 수정한 변수중요도로 지정한다.

$$\text{Max MDA}\left(X_j\right) = \max\left\{\mathrm{MDA}\left(X_j\right),\ \mathrm{MDA}^{(-j)*}\left(X_j\right)\right\}. \tag{2.16}$$

Step 7. Steps 2~6의 과정을 모든 변수에 대해 반복하여 각 변수의 변수중요도를 계산한다.

3. 모의실험

본 논문의 2장에서 제시하는 Max MDI와 Max MDA 변수중요도는 기존에 있던 MDI와 MDA 변수중요도에서 나타나는 예측변수 간에 선형 또는 비선형 상관관계가 존재할 때의 하향 편향을 개선하기 위해 수정한 것이다. 3장에서는 수정한 변수중요도가 기존의 두 변수중요도에서 나타나는 하향 편향을 실제로 개선하는지 확인하기 위해 모의실험을 진행했다. 모의실험은 총 3가지 상황이 있고, 모두 200개의 관측치를 사용했다. 랜덤포레스트 내의 $M = 500$번 무작위 추출로 인해 변수중요도에 변동이 있어서 상황마다 100번씩 반복하여 결과를 얻었다. 또한, 랜덤포레스트를 적합할 때 마디를 분리하기 위한 무작위 추출변수는 모든 변수가 같은 확률로 사용될 수 있게 1개로 지정했다. 2개 이상으로 지정했을 때는 중요한 변수와 중요하지 않은 변수와의 격차가 커지지만, 편향문제에 대해서는 비슷한 결과가 나오기에 생략하도록 한다.

모의실험 결과를 보기에 앞서, 본 논문에서 사용하는 상자 그림은 변수중요도의 분위수를 나타내지 않음을 주의해야 한다. 상자 그림에서 중심선은 평균을 나타내고 25분위수, 75분위수 선은 평균에서 변수중요도의 표준편차만큼 떨어진 지점을 나타내며, 양 끝 선은 평균에서 2 표준편차만큼 떨어진 지점을 나타낸다.

Table 1: Case 1. Variable importance

	X_1	X_2	X_3	X_4	X_5	X_6
MDI	18.0(0.2)	17.1(0.1)	19.7(0.2)	13.9(0.1)	13.9(0.1)	14.0(0.1)
Max MDI	20.6(0.2)	19.9(0.2)	20.7(0.2)	16.6(0.1)	16.7(0.1)	16.6(0.1)
MDA	10.6(0.5)	7.7(0.4)	13.5(0.5)	1.3(0.3)	0.9(0.2)	−0.4(0.3)
Max MDA	13.1(0.5)	10.9(0.4)	14.4(0.5)	3.7(0.2)	3.5(0.2)	2.9(0.2)

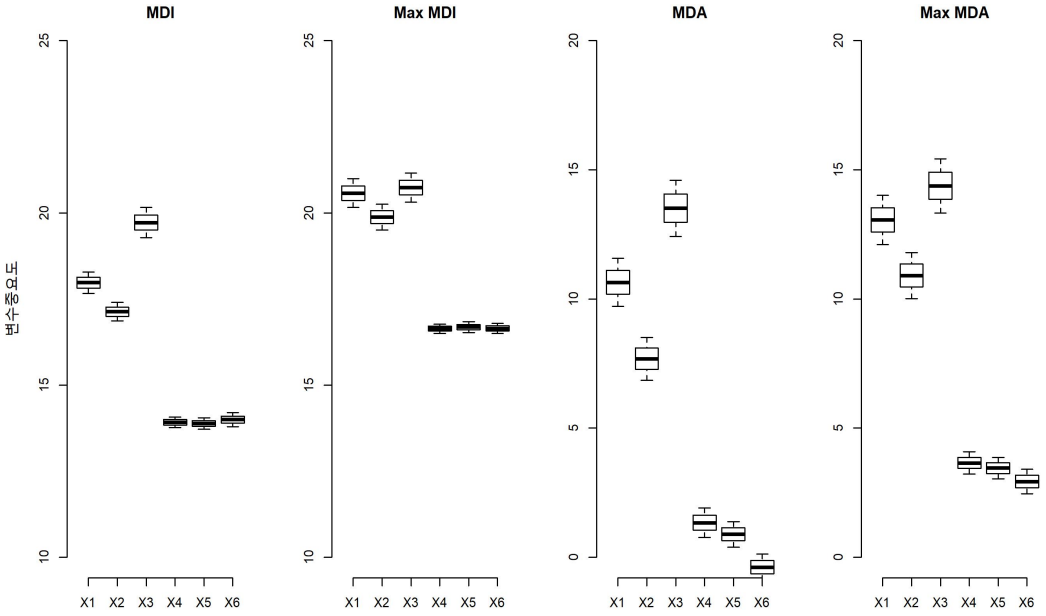

Figure 1: *Case 1. Distribution of variable importance.*

3.1. Case 1

첫 번째 모의실험은 두 예측변수 간에 강한 선형 상관관계가 있을 때의 상황이다. X_1과 X_2는 평균이 0이고 분산이 1이며, 공분산이 0.9인 이변량 정규분포를 따른다. X_3는 표준정규분포를 따른다. X_4와 X_5는 X_1, X_2처럼 이변량 정규분포를 따르고 X_6는 표준정규분포를 따르지만, X_2, X_4, X_5, X_6는 반응변수 생성에 영향을 주지 않는다. 반응변수 Y는 X_1과 X_3에 의해 다음과 같이 생성된다.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 0.5 + X_{i1} + X_{i3}, \quad Y_i \sim B(1,\ \pi_i). \tag{3.1}$$

이렇게 생성된 Y를 예측하기 위해 $X_1, \ldots, X_6$를 사용하여 Figure 1 (Table 1)과 같은 변수중요도를 얻었다. Figure 1을 보면 MDI와 MDA 변수중요도에서 X_2와 상관계수가 0.9인 X_1의 변수중요도가 X_3에 비해 낮아지는 것을 볼 수 있다. 하지만 Max MDI와 Max MDA에서 이러한 편향은 개선된다. MDA와 Max MDA의 경우, 반응변수에 영향을 주지 않지만 서로 상관관계가 있는 X_4, X_5의 변수중요도가 X_6보다 높게 나타나지만 중요한 변수에 비해 낮은 수치를 보여 중요하지 않은 변수임을 알 수 있다.

Table 2: Case 2. Variable importance

	X_1	X_2	X_3	X_4	X_5	X_6
MDI	16.2(0.1)	15.6(0.1)	15.3(0.1)	15.6(0.1)	19.6(0.2)	13.8(0.1)
Max MDI	20.9(0.2)	19.9(0.2)	19.8(0.2)	20.0(0.2)	20.9(0.2)	16.1(0.1)
MDA	10.2(0.4)	6.2(0.4)	5.9(0.3)	7.1(0.4)	14.3(0.5)	0.3(0.3)
Max MDA	13.6(0.4)	11.2(0.5)	10.7(0.5)	11.7(0.4)	15.1(0.5)	2.9(0.2)

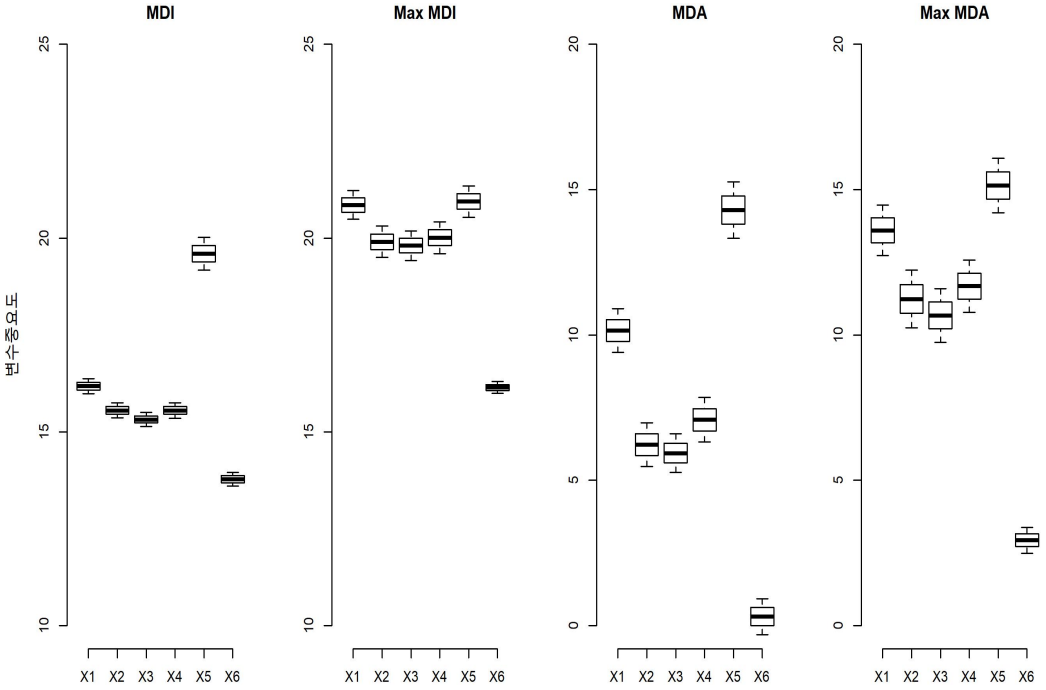

Figure 2: *Case 2. Distribution of variable importance.*

3.2. Case 2

두 번째 모의실험은 4개의 예측변수 간에 상관관계가 있을 때의 상황이다. $X_1, \ldots, X_4$는 평균이 0이고 분산이 1이며, 두 변수 간의 공분산이 0.9인 다변량 정규분포를 따른다. X_5와 X_6는 표준정규분포를 따른다. 반응변수 Y는 X_1과 X_5에 의해 다음과 같이 생성된다.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = 0.5 + X_{i1} + X_{i5}, \quad Y_i \sim B(1,\ \pi_i). \tag{3.2}$$

이렇게 생성된 Y를 예측하기 위해 $X_1, \ldots, X_6$를 사용하여 Figure 2 (Table 2)와 같은 변수중요도를 얻었다. Figure 2를 보면 MDI와 MDA 변수중요도에서 X_1의 변수중요도가 X_5에 비해 낮아지는 것을 볼 수 있다. 하지만 Max MDI와 Max MDA에서 이러한 편향은 개선된다.

Table 3: Case 3. Variable importance

	X_1	X_2	X_3	X_4	X_5	X_6
MDI	21.7(0.2)	13.5(0.1)	13.4(0.1)	13.3(0.1)	25.7(0.3)	11.3(0.1)
Max MDI	27.0(0.3)	18.8(0.1)	18.6(0.1)	18.7(0.1)	27.3(0.3)	16.7(0.1)
MDA	23.9(0.4)	5.9(0.3)	5.4(0.3)	5.2(0.3)	30.5(0.4)	−0.5(0.2)
Max MDA	26.3(0.5)	8.2(0.3)	7.7(0.4)	7.6(0.3)	30.5(0.5)	2.9(0.2)

3.3. Case 3

세 번째 모의실험은 X_2, X_3, X_4가 표준정규분포를 따르고 X_1을 다음과 같은 식을 통해 생성했을 때의 상황이다.

$$X_1 = X_2^2 + X_3^2 + X_4^2. \tag{3.3}$$

X_5는 X_1과 같은 분포를 만들기 위해 자유도가 3인 카이제곱분포를 따른다.

$$X_5 \sim \chi^2(3). \tag{3.4}$$

X_6는 표준정규분포를 따르고 반응변수 Y는 X_1과 X_5에 의해 다음과 같이 생성된다. 상수항을 –5.5로 설정한 것은 X_1과 X_5가 양수 값을 갖기 때문에 π_i가 1에 가까운 값이 되는 것을 방지하기 위함이다.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -5.5 + X_{i1} + X_{i5}, \quad Y_i \sim B(1,\ \pi_i). \tag{3.5}$$

Figure 3 (Table 3)을 보면 MDI와 MDA 변수중요도에서 X_1의 변수중요도가 X_5에 비해 낮아지는 것을 볼 수 있다. 하지만 Max MDI와 Max MDA에서 이러한 편향은 개선된다.

총 3가지의 상황에 대해 변수중요도를 계산한 결과, 모든 상황에서 MDI, MDA 변수중요도가 선형 또는 비선형 상관관계가 있는 예측변수에 대해 하향 편향된 수치를 보였다. 하지만 Max MDI 변수중요도는 예측변수 간에 선형 또는 비선형 상관관계가 있음에도 불구하고 편향되지 않은 수치를 보였다. Max MDA 변수중요도는 Max MDI만큼 개선되지 않았지만, MDA보다 개선된 결과를 보임을 알 수 있었다.

4. 사례분석

본 논문에서 제안하는 수정한 변수중요도가 실제 자료에서 어떤 결과를 만들어내는지 확인하기 위해 예측변수 간에 상관관계가 높은 실제 자료를 사용하여 기존의 변수중요도와 수정한 변수중요도 간에 어떤 결과의 차이가 생기는지 확인한다.

4.1. 자료 소개

이 장에서 사용할 Vehicle Silhouette 자료는 Turing Institute, Glasgow, Scotland에서 제공되어 UCI Machine Learning Repository에서 확인할 수 있는 공식 자료이다. 이 자료는 오펠(Opel)과 사브(Saab), 버스(Bus), 밴(Van) 4종류의 자동차 윤곽을 여러 회전 각도에서 촬영하여 특징 추출기를 통해 특징을 수치화한 것으로, 자동차 윤곽에서 나타나는 축의 길이로부터 여러 변수를 생성했기 때문에 예측변수 간에 높은 상관관계가 있다. Table 4에서 변수들을 확인 할 수 있으며, Table 5에서 변수들의 상관계수가 높은 것을 볼 수 있다.

4.2. 제안 알고리즘 적합

이 자료로부터 변수중요도를 계산하기 위해 랜덤포레스트를 사용할 때 각 마디를 분리하는 무작위 추출 변수는 모의실험과 마찬가지로 1개를 시도했다.

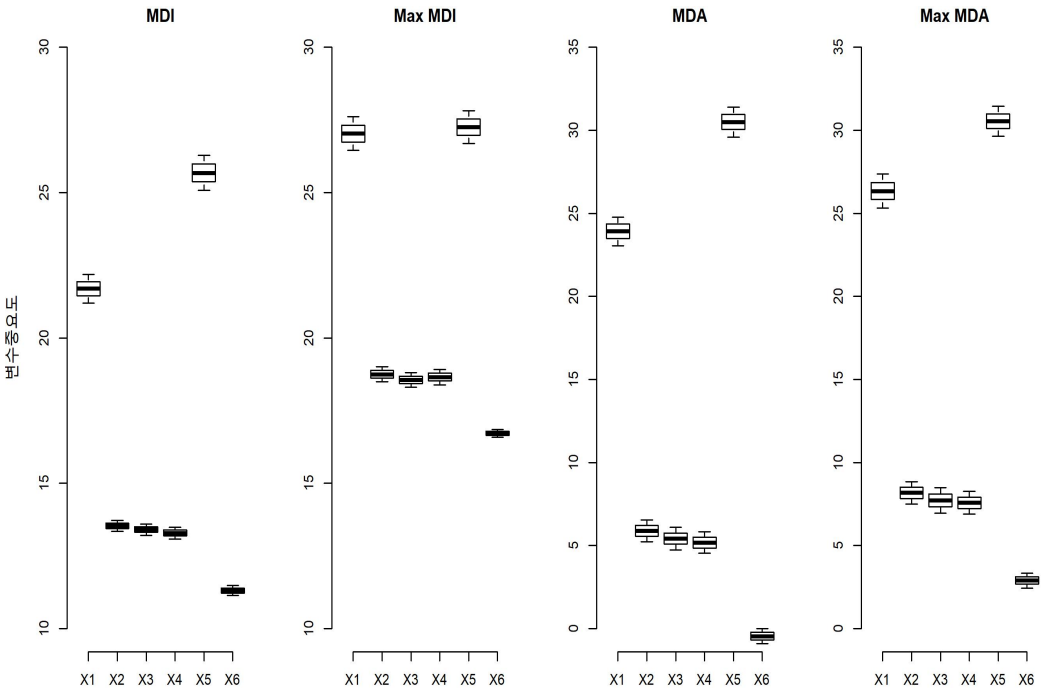

Figure 3: *Case 3. Distribution of variable importance.*

Table 4: Vehicle silhouette data set description

Variable	Variable name
X_1	Compactness
X_2	Circularity
X_3	Distance Circularity
X_4	Radius ratio
X_5	Pr.axis aspect ratio
X_6	Max length aspect ratio
X_7	Scatter ratio
X_8	Elongatedness
X_9	Pr.axis rectangularity
X_{10}	Max length rectangularity
X_{11}	Scaled variance along major axis
X_{12}	Scaled variance along minor axis
X_{13}	Scaled radius of gyration
X_{14}	Skewness about major axis
X_{15}	Skewness about minor axis
X_{16}	Kurtosis about major axis
X_{17}	Kurtosis about minor axis
X_{18}	Hollows ratio
Y	Car type

Table 5: Pearson correlation between variables

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
X_1	1.00	0.69	0.79	0.69	0.09	0.15	0.81	−0.79	0.81	0.68	0.76	0.82	0.59	−0.25	0.23	0.16	0.30	0.37
X_2	0.69	1.00	0.80	0.62	0.15	0.25	0.86	−0.83	0.86	0.97	0.81	0.85	0.94	0.06	0.15	−0.02	−0.11	0.04
X_3	0.79	0.80	1.00	0.77	0.16	0.26	0.91	−0.91	0.90	0.77	0.86	0.89	0.71	−0.23	0.12	0.26	0.15	0.34
X_4	0.69	0.62	0.77	1.00	0.67	0.45	0.74	−0.79	0.71	0.57	0.80	0.73	0.54	−0.18	0.05	0.17	0.38	0.47
X_5	0.09	0.15	0.16	0.67	1.00	0.65	0.11	−0.19	0.08	0.13	0.27	0.09	0.12	0.15	−0.06	−0.03	0.24	0.27
X_6	0.15	0.25	0.26	0.45	0.65	1.00	0.17	−0.18	0.16	0.31	0.32	0.14	0.19	0.29	0.02	0.04	−0.03	0.14
X_7	0.81	0.86	0.91	0.74	0.11	0.17	1.00	−0.97	0.99	0.81	0.95	1.00	0.80	−0.03	0.07	0.21	0.00	0.12
X_8	−0.79	−0.83	−0.91	−0.79	−0.19	−0.18	−0.97	1.00	−0.95	−0.78	−0.94	−0.96	−0.77	0.10	−0.05	−0.19	−0.11	−0.22
X_9	0.81	0.86	0.90	0.71	0.08	0.16	0.99	−0.95	1.00	0.81	0.94	0.99	0.80	−0.02	0.08	0.21	−0.02	0.10
X_{10}	0.68	0.97	0.77	0.57	0.13	0.31	0.81	−0.78	0.81	1.00	0.75	0.80	0.87	0.04	0.14	0.00	−0.11	0.08
X_{11}	0.76	0.81	0.86	0.80	0.27	0.32	0.95	−0.94	0.94	0.75	1.00	0.95	0.78	0.11	0.04	0.19	0.01	0.09
X_{12}	0.82	0.85	0.89	0.73	0.09	0.14	1.00	−0.96	0.99	0.80	0.95	1.00	0.80	−0.02	0.08	0.20	0.01	0.11
X_{13}	0.59	0.94	0.71	0.54	0.12	0.19	0.80	−0.77	0.80	0.87	0.78	0.80	1.00	0.19	0.17	−0.06	−0.23	−0.12
X_{14}	−0.25	0.06	−0.23	−0.18	0.15	0.29	−0.03	0.10	−0.02	0.04	0.11	−0.02	0.19	1.00	−0.09	−0.13	−0.75	−0.81
X_{15}	0.23	0.15	0.12	0.05	−0.06	0.02	0.07	−0.05	0.08	0.14	0.04	0.08	0.17	−0.09	1.00	−0.04	0.12	0.10
X_{16}	0.16	−0.02	0.26	0.17	−0.03	0.04	0.21	−0.19	0.21	0.00	0.19	0.20	−0.06	−0.13	−0.04	1.00	0.08	0.20
X_{17}	0.30	−0.11	0.15	0.38	0.24	−0.03	0.00	−0.11	−0.02	−0.11	0.01	0.01	−0.23	−0.75	0.12	0.08	1.00	0.89
X_{18}	0.37	0.04	0.34	0.47	0.27	0.14	0.12	−0.22	0.10	0.08	0.09	0.11	−0.12	−0.81	0.10	0.20	0.89	1.00

Table 6: Variable importance

	MDI	max MDI	MDA	max MDA
X_1	34.15	49.02	21.86	21.86
X_2	31.35	47.44	18.46	18.67
X_3	38.71	56.42	24.44	25.54
X_4	34.43	56.55	19.05	23.52
X_5	32.24	40.90	24.77	24.77
X_6	43.91	49.63	26.51	26.51
X_7	38.73	64.73	21.51	27.24
X_8	37.45	61.23	21.66	27.85
X_9	26.40	54.52	20.66	27.82
X_{10}	37.61	45.11	23.00	23.00
X_{11}	39.34	62.52	19.52	25.96
X_{12}	41.33	65.95	21.68	26.96
X_{13}	30.78	48.21	18.46	19.01
X_{14}	33.99	50.02	21.77	21.77
X_{15}	24.81	37.60	13.68	13.68
X_{16}	26.24	43.35	11.56	12.85
X_{17}	29.39	43.04	20.68	20.68
X_{18}	33.46	46.99	23.61	23.61

Figure 4와 Figure 5 (Table 6)를 보면 기존의 변수중요도와 수정한 변수중요도의 차이가 큰 변수가 있다. 그중에 유심히 볼만한 것으로 자동차의 형태가 사각형에 가까운 정도를 나타내는 9번째 변수인 사각형 변수가 있다. 4종류의 자동차를 분류할 때 그 축이 직각인 정도가 중요한 역할을 할 수 있음을 고려하면 이 변수는 높은 변수중요도를 가져야 할 것이다. 그러한 기대와 반대로, 9번째 변수는 Table 5에서 확인할 수 있는 높은 상관관계에 있는 예측변수들의 간섭으로 Figure 4와 Figure 5에서 낮게 편향된 변수중요도를 가진다. 하지만

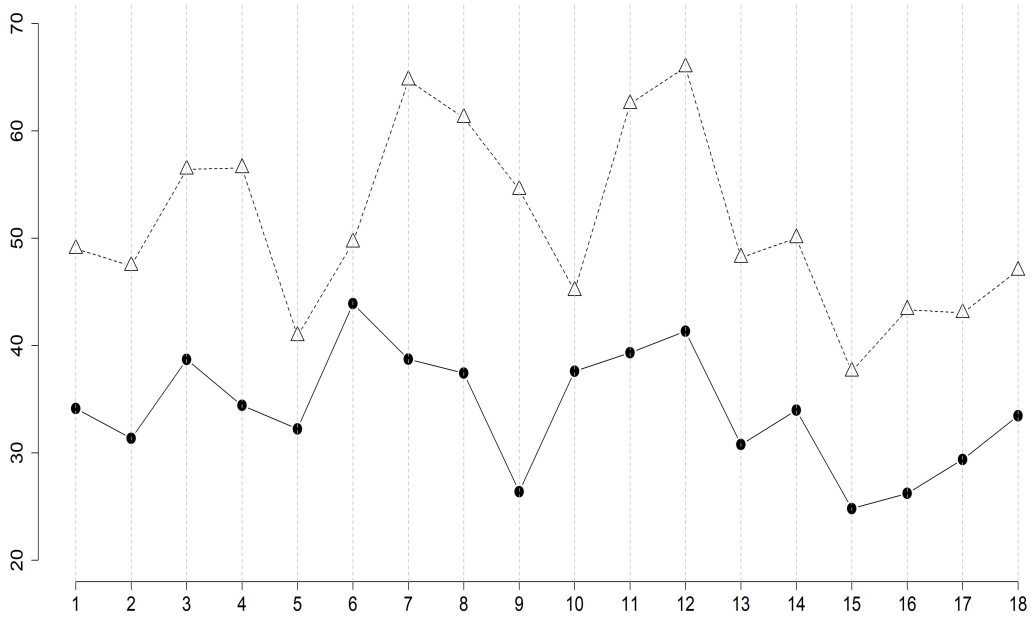

Figure 4: *MDI variable importance.*

Max MDI에선 9번째 변수가 적당한 수준의 변수중요도를 가지며, Max MDA에선 가장 높은 변수중요도를 가진다. 또한, MDI와 MDA 결과에서 얇고 긴 정도를 나타내는 6번째 변수 Max length aspect ratio가 가장 높은 변수중요도를 보였는데, Max MDI와 Max MDA에서 가장 높은 변수중요도를 가진 변수가 바뀌게 된다. 6번째 변수보다 7번째와 8번째, 9번째, 12번째 변수가 자동차 분류에 더 큰 역할을 할 수 있음을 알려준다. 즉, 수정한 변수중요도가 기존 변수중요도의 하향 편향이 개선된 변수중요도를 보여 모형에 대해 다른 해석을 할 수 있다.

Max MDI와 Max MDA의 차이점이 있다면 Max MDA는 MDA와 값이 같은 때도 있다는 점이다. 이에 대해서는 두 가지 해석이 가능하다. 첫 번째로 해당 변수와 상관관계가 있는 변수가 없어 하향 편향이 일어나지 않았다고 볼 수 있으며, 두 번째로 해당 변수 하나만 사용할 때보다 여러 변수를 함께 사용할 때 그 변수가 더 중요해진다는 의미로, 다른 변수와의 상호작용이 있음을 암시한다.

5. 결론

이 논문의 모의실험에서 기존의 랜덤포레스트 변수중요도가 예측변수 간에 상관관계가 있을 때 하향 편향된 결과를 보임을 확인할 수 있었다. 그리고 실제 자료에 적용된 결과와 함께 이러한 편향이 수정한 변수중요도에서 개선됨을 확인할 수 있었다. 물론 이 논문의 모의실험에서 예측변수 간에 존재할 수 있는 모든 상관관계를 고려했다고 할 수는 없지만, 제시된 상황 내에서는 대부분의 하향 편향이 개선됐다고 할 수 있다.

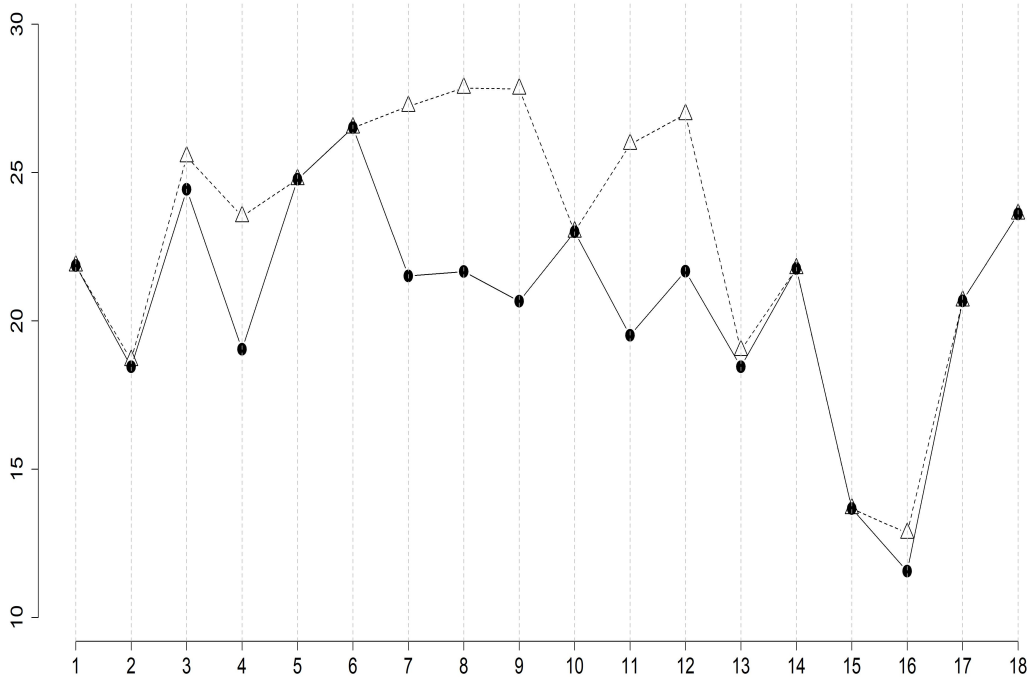

Figure 5: *MDA variable importance.*

References

Archer, K. J. and Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures, *Computational Statistics & Data Analysis*, **52**, 2249–2260.

Biau, G. and Scornet, E. (2016). A random forest guided tour, *Text*, **25**, 197–227.

Breiman, L. (2001). Random forests, *Machine Learning*, **45**, 5–32.

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3.1. Statistics Department University of California Berkeley, CA, USA.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*, Wadsworth, Belmont.

Genuer, R., Poggi, J. M., and Tuleau-Malot, C. (2010). Variable selection using random forests, *Pattern Recognition Letters*, **31**, 2225–2236.

Gregorutti, B., Michel, B., and Saint-Pierre, P. (2017). Correlation and variable importance in random forests, *Statistics and Computing*, **27**, 659–678.

Nicodemus, K. K., Malley, J. D., Strobl, C., and Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation, *Bioinformatics*, **11**, 1–13.

RColorBrewer, S. and Liaw, M. A. (2018). *Package 'randomForest'*, University of California Berkeley, CA, USA.

Rumao, S. (2019). Exploration of Variable Importance and Variable selection techniques in presence of correlated variables. Rochester Institute of Technology.

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for

random forests, *Bioinformatics*, **9**, 307.

Strobl, C., Boulesteix, A. L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution, *Bioinformatics*, **8**, 25.

Received December 8, 2020; Revised January 12, 2021; Accepted January 15, 2021

랜덤포레스트를 위한 상관예측변수 중요도

신승범[a], 조형준[1,a]

[a]고려대학교 통계학과

요 약

랜덤포레스트는 여러 의사결정나무 모형들을 융합하여 안정성과 예측력을 높여주기 때문에 종종 사용되는 방법이다. 예측력을 증가시키는 반면 해석의 용이성을 희생하기 때문에 이를 보상하기 위해 변수의 중요도를 제공한다. 변수의 중요도는 랜덤포레스트를 구축할 때 변수가 얼마나 중요한 역할을 하는지를 알려준다. 그러나 어떤 예측변수가 다른 예측변수들과 상관되어 있을 때 기존 알고리즘의 변수중요도는 왜곡될 수 있다. 상관된 예측변수들의 하향 편향은 예측변수의 중요도를 실제 중요도보다 낮게 측정하게 한다. 우리는 기존 알고리즘을 수정하여 상관 예측변수의 하향 편향을 회복하는 새로운 알고리즘을 제안한다. 제안된 알고리즘의 성능은 모의 자료에 의해 증명되고 실제 자료에 의해 설명된다.

주요용어: 랜덤포레스트, 변수중요도, 상관관계

이 논문은 제1저자 신승범의 석사학위 논문의 일부를 발췌한 것임.
이 논문은 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원(2018R1D1A1B07044479), 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(2019R1A4A1028134), 고려대학교 연구비 지원(K2019631)을 받아 수행된 연구이다.
[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: hj4cho@korea.ac.kr