Journal of Health Informatics and Statistics

Original Article

J Health Info Stat 2022;47(3):217-221 • https://doi.org/10.21032/jhis.2022.47.3.217

pISSN 2465-8014 | eISSN 2465-8022

기계학습방법을 이용한 순서형 결측자료 대체의 성능비교

손세림[1], 안형진[2]

[1]고려대학교 의과대학 의학통계학교실 박사과정생, [2]고려대학교 의과대학 의학통계학교실 교수

Performance Comparison of Imputation Methods Using Machine Learning Techniques for Ordinal Missing Data

Serhim Son[1], Hyonggin An[2]

[1]Doctoral Student, Department of Biostatistics, Korea University College of Medicine, Seoul; [2]Professor, Department of Biostatistics, Korea University College of Medicine, Seoul, Korea

Objectives: When missing values occur, complete case analysis can cause biased results. In this paper, we discuss imputation methods using machine learning techniques when missing values occurred in ordinal variables. **Methods:** We consider two machine learning techniques, the ordinal decision tree and the random forest, for the imputation of missing values. We use the ordinal decision tree treating variables as ordinal, and the random forest as nominal. In addition, we apply the cumulative logistic model. The results are compared with complete case analysis using empirical bias, empirical mean squared error and accuracy. The same methods are applied using the Korea National Health and Nutrition Examination Survey. **Results:** In the case of five ordinal categories, machine learning techniques yield better performance than the cumulative logistic and complete case. The ordinal decision shows lower bias while random forest shows higher accuracy. In the case of 3 categories, random forest produces better performance in all respects. In the case study, biased results are also identified if we use complete case analysis. Random forest shows the best performance, and the parametric method shows similar performance to the ordinal decision tree. **Conclusions:** Missing imputation using machine learning techniques can reduce bias and improve performance. If possible, it is recommended to use the ordinal decision tree to impute missing values that reflects the meaning of order. If it is not possible, it is recommended to treat them at least as nominal variables and then impute.

Key words: Machine learning, Regression analysis, Decision tree, Big data, Health survey

서 론

자료수집 과정에서 결측은 다양한 원인으로 발생한다. 결측이 발생한 자료에서 가장 흔하게 사용되는 응답한 개체만을 이용한 분석(complete case analysis)이다. 하지만 이 방법은 자료의 일부분만을 사용하기 때문에 자료의 특성과 분포가 모집단과 상이해져 대표성에 문제가 발생할 수 있다. 또한, 표본수 감소에 따른 정보의 손실로 정밀도(precision)가 낮아져 통계적인 검정력(power)이 낮아질 수 있다[1]. 결측으로 인한 문제들을 해결할 수 있는 방법 중 하나인 대체(imputation)는 결측값들을 그럴듯한 값으로 채우는 방법이다. Jonathan et al. [2]은 완전히 응답한 개체만을 이용한 분석이 아닌 결측 대체에 대한 필요성을 제시하였고, 최근 발표되는 논문들에서는 결측이 발생한 변수를 그대로 분석에 사용하지 않고 대체한 후 분석에 사용한다[3,4].

조사자료에서 많이 사용되는 문항은 리커트 척도와 같은 순서형 문

Corresponding author: Hyonggin An
73 Goryeodae-ro, Seongbuk-gu, Seoul 02841, Korea
Tel: +82-2-2286-1437, E-mail: hyonggin@gmail.com

Received: June 13, 2022 **Accepted:** August 28, 2022 **Published:** August 31, 2022

*This paper is modified Serhim Son's master's degree paper at Korea University.

No potential conflict of interest relevant to this article was reported.

How to cite this article:
Son S, An H. Performance comparison of imputation methods using machine learning techniques for ordinal missing data. J Health Info Stat 2022;47(3):217-221. Doi: https://doi.org/10.21032/jhis.2022.47.3.217

항이다. 대표적으로 국민건강영양조사자료 7기 자료를 보면, 개인의 주관적 건강상태, 혹은 개인의 주관적 체형인식 상태를 묻는다. 순서형 자료에서 결측이 발생했을 때 적용할 수 있는 방법에는 순서형 로지스틱 회귀분석이 있다. 이 방법은 회귀 대체법으로 완전히 관측된 자료를 이용하여 모형을 만들고, 확률을 추정하는 방법이다. 하지만 이 방법은 안정적인 결과를 도출하기 위해서는 많은 수의 연구대상자가 필요하고[5], 외삽으로 인해 데이터의 실제 범위에 벗어난 값으로 대체가 될 수도 있다[1]. 이때 대안적으로 사용할 수 있는 방법은 기계학습방법으로, 학습을 통해 모형을 설정하고, 이 모형을 이용하여 결측값을 예측하는 방법이다. 모델의 해석력과 복잡성은 서로 상충하기 때문에 만약 연구 목적이 추론을 하는 것이라면, 복잡한 기계학습법을 사용해서 해석이 어려워지는 것을 선호하지 않은 수 있다[6]. 반면 성능이 좋은 예측을 하는 것이 목적이라면, 기계학습기법의 사용이 추천될 수 있다. 따라서, 본 연구에서는 순서형 자료에서 발생한 결측자료에 대해 기계학습방법을 적용하여 대체 성능을 비교하였다. 순서형 자료는 종종 명목형 자료로 취급되어 분석에 사용되기 때문에, 명목형으로도 취급하여 결과를 비교하였다.

Kevin et al. [7]은 순서형 자료에서 무응답이 발생했을 때 마지막 값으로 결측 대체(last observation carried forward, LOCF)방법과 모형기반 방법을 비교하였다. 또한, Jonsson et al. [8]은 리커트 척도에 대하여 k-최근접 이웃 알고리즘을 적용하여 성능을 확인하였고, Podolsky and Sho [9]은 소득자료를 구간별로 나눈 순서형 자료에 기계학습법을 적용하여 성능을 비교하였다. 하지만 좀 더 다양하고 일반적인 상황에 대한 연구는 제한적이다. 따라서 본 연구에서는 국내자료에서 많이 확인할 수 있는 3점 척도와 5점 척도의 상황을 가정하여 기계학습기법과 모수적 방법을 모의실험을 통해 결측값 대체 성능을 비교하였다.

연구 방법

순서형 의사결정나무

순서형 의사결정나무는 classification and regression trees (CART) 알고리즘에 근거하여 작동한다. CART 알고리즘은 적절한 분리기준과 정지기준의 반복을 통해 하나의 의사결정나무를 얻는 것이 목적이다. CART 알고리즘의 분리 기준은 예측변수가 범주형 변수인 경우에는 지니 불순도(gini impurity)를 이용하여 분리한다. 지니 불순도는 하나의 집합에 얼마나 많은 클래스를 갖는지 측정하는 지표로, 하나의 집합에 모든 같은 클래스가 속한다면 지니 불순도의 값은 0이 된다. CART 알고리즘에서, 하나의 마디를 분리할 수 있는 가장 적절한 기준은 지니 불순도의 감소를 최대화할 수 있는 기준을 찾는 것이다.

N개의 관측치에 p개의 공변량 $X_1, X_2, \ldots, X_p$과, $\{w_1 < w_2 < \cdots < w_J\}$를 만족하고 J개의 카테고리를 갖는 예측변수 Y가 있다고 가정했을 때, 순서형 변수의 특징을 반영한 지니 불순도 식은 (1)이다. 이때, $p(w_k|t)$는 마디 t에 속한 k번째 카테고리의 비율을 의미하며, $p(w_1|t) + p(w_2|t) + \cdots + p(w_J|t) = 1$이 된다.

$$i_{GG}(t) = \sum_{k=1}^{J}\sum_{l=1}^{J} C(w_k|w_l)p(w_k|t)p(w_l|t) \quad (1)$$

이때, $C(w_k|w_l)$는 클래스 l에 속하는 관측치를 클래스 k에 잘못 분류했을 때 발생하는 오분류 비용함수로, 순서형 반응변수의 예측을 위해 추가된 항이다[10]. 만약 $k = l$이면, 제대로 분류가 되었기 때문에 오분류 비용은 0이 된다.

오분류 비용함수를 계산하기 위해서는 반응변수 Y의 각 카테고리에 해당하는 스코어 $\{S_1 < S_2 < \cdots < S_J\}$가 있다고 가정한다. 반응변수 Y에 대한 카테고리를 증가하는 값으로 가정했기 때문에, 그에 대응하는 스코어도 증가하는 값으로 가정한다. 오분류 비용함수의 값은 스코어의 적절한 변형을 통해서 구할 수 있다. 오분류 비용함수를 선형함수와 이차함수로 나눌 수 있고, 각각 식 (2), (3)과 같다.

$$i_{GG1}(t) = \sum_{k=1}^{J}\sum_{l=1}^{J} |s_k - s_l|p(w_k|t)p(w_l|t) \quad (2)$$

$$i_{GG2}(t) = \sum_{k=1}^{J}\sum_{l=1}^{J} (s_k - s_l)^2 p(w_k|t)p(w_l|t) \quad (3)$$

CART 알고리즘에서 지니 불순도의 감소를 위해 계속해서 마디를 분리하면 과적합의 문제가 발생할 수 있다. 이를 해결하기 위해서는 가지치기(pruning)의 과정이 필요하다[11]. 가지치기의 종류에는 하나의 큰 나무를 만들기 전에 몇 가지의 값들을 미리 지정하여 제한된 크기의 나무를 만드는 사전 가지치기와 나무가 만들어진 후에 적절한 크기로 마디를 자르는 사후 가지치기가 있다. 본 논문에서는 사후 가지치기를 이용하여 과적합을 예방하였다.

랜덤 포레스트

랜덤 포레스트는 의사결정나무를 기반으로 작동하는 앙상블기법 중 하나이다. 앙상블기법은 다수의 약한 알고리즘들을 여러 번 작동하여 하나의 강한 알고리즘을 만드는 방법으로, 여러 번의 의사결정나무를 독립적으로 학습하여 강한 하나의 숲을 만드는 방법이다[6]. 자세한 알고리즘은 다음과 같다[12].

(Step 1) B = 1, 2, ⋯, B에 대하여,

(1) 훈련 자료에서 크기가 N인 부트스트랩 표본 Y^*을 생성한다.

(2) 부트스트랩 표본 Y^*에서 의사결정나무의 각 끝단노드(terminal node)가 최소 마디 크기인 $node_{min}$에 도달할 때까지 아래의 세 단계를 반복적으로 수행하여 하나의 랜덤 포레스트 $T_b(x)$를 생성한다.

Table 1. Parameters for generating simulation data

	5 Categories	3 Categories
$\beta_{0,1}$	-4	-2
$\beta_{0,2}$	-2	2
$\beta_{0,3}$	1	-
$\beta_{0,4}$	3	-
β_1	0.02	0.02
β_2	0.02	0.02
β_3	0.02	0.02
β_4	0.02	0.02
β_5	10	-7.5
β_6	-7	-4.5
β_7	-7	8.5

(a) p개의 공변량 중에서 m개를 무작위로 선택한다(단, m<p).
(b) 선택된 m개의 공변량 중에서 가장 좋은 변수 또는 분할점을 찾는다.
(c) 노드를 두 개의 자식노드(daughter node)로 분할한다.

(Step 2) B개의 랜덤 포레스트를 결합하여 최종 예측식을 생성한다. 이때, 의사결정나무가 범주형이면 가장 많이 예측된 값을 도출한다. 최종 예측식은 식 (4)와 같다.

$$T^*(x) = majority\ vote\{T_b(x)_1^B\} \quad (4)$$

랜덤 포레스트는 p개의 공변량 중에서 가장 강력한 m개의 공변량을 선택해서 나무를 만들기 때문에, 분산을 감소시키며[6], 과적합을 예방할 수 있다는 장점이 있다[12].

모의실험의 설계

순서형 자료에 대한 기계학습기법의 성능비교를 위한 자료생성은 순서형 로지스틱회귀분석을 기반으로 범주가 3개인 경우와 5개인 경우를 각각 생성하였다. 자료생성을 위한 모수값은 Table 1과 같다. 카테고리가 5개인 경우, i=1,⋯,4에 대하여 자료생성을 위한 식은 (5)와 같다.

$$X_1 \sim N(0,1),\ X_2 \sim bin(1,0.5),\ X_3 \sim exp(1),\ X_4 \sim N(0,1)$$
$$logit(Y \le i) = \beta_{0,i} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1^2 + \beta_6 X_3^2 + \beta_7 X_4^2 \quad (5)$$

카테고리가 3개인 경우 j=1,2에 대하여 자료생성을 위한 식은 (6)과 같다.

$$logit(Y \le j) = \beta_{0,j} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1^2 + \beta_6 X_3^2 + \beta_7 X_4^2 \quad (6)$$

각 모형에서 절편값을 조정하여 범주별 비율을 조정하였다. 표본의 크기는 1,000으로 생성하였고, 1,000번 반복을 시행하였다. 결측치는

Table 2. Parameters for generating response model

	5 Categories	3 Categories
α_0	-1.5	-1.5
α_1	3	3
α_2	3	3
α_3	3	3
α_4	3	3
Actual missing rate (%)	30.4	30.4

임의결측 가정하에서 30%을 목표로 결측을 발생시켰다. 응답모형은 식 (7)과 같으며, 모수값은 Table 2에 제시되어 있다.

$$logit[(P(R=1)|X_1,X_2,X_3,X_4)] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \alpha_4 X_4,$$
$$R = \{\begin{matrix}1, y = observed \\ 0, y = missing\end{matrix}\} \quad (7)$$

결측치를 대체하는 방법으로 순서형 의사결정나무, 명목형으로 취급하는 랜덤 포레스트, 순서형 로지스틱 회귀분석을 적용하였다. 본 모의실험의 결과는 결측이 발생하기 전 평균값을 모수로 하여 추정치의 경험적 편향과 경험적 평균제곱오차를 비교하였다. $\bar{\mu}_i$가 대체 전 원자료의 평균이고, $\hat{\mu}_i$가 결측값이 대체된 자료의 평균일 때, 경험적 편향과 경험적 평균제곱오차의 식은 (8), (9)과 같다.

$$Empirical\ bias = 100 * |E(\hat{\mu}_i - \bar{\mu}_i)| \quad (8)$$
$$Empirical\ mean\ squared\ error = 100 * E(\hat{\mu}_i - \bar{\mu}_i)^2 \quad (9)$$

또한, 전체 결측치 중에서 정확하게 예측된 비율을 의미하는 정확도를 제시하였다. 모의실험 및 사례연구를 위해서 R 프로그램을 사용하였다.

사례연구

국민건강영양조사는 매년 질병관리청의 감독하에 우리나라 국민 1만 명에 대한 건강수준, 건강관련 의식 및 행태, 식품 실태조사에 대한 국가단위 통계를 산출하기 위한 조사이다. 본 연구에서는 국민건강영양조사의 2014-2018년 자료의 '주관적 체형 인식 수준' 변수에 대한 결측 대체를 적용하였다. 무응답 대체를 위한 데이터를 생성하기 위해 19세 이상 28,819명의 성인을 대상자로 선정하였으며, 30%의 결측을 목표로 1,000번의 모의실험을 시행하였다. 해당 변수를 대체하기 위한 보조변수로는 나이, 성별, 키, 몸무게, 허리둘레를 선정하였다. 결측 데이터를 만들기 위한 모수값은 Table 3에 제시되어 있다.

Table 3. Parameters for generating case study data

Parameters	
Intercept	-5.50
Sex	1.50
Age	0.70
Height	0.11
Weight	0.45
Waist circumference	-0.89
Actual missing rate (%)	29.6

Table 4. Results of simulation

Methods	5 Categories			3 Categories		
	Bias	MSE	Accuracy (%)	Bias	MSE	Accuracy (%)
DT (Lin)	0.61	0.97	61.2	1.63	0.31	70.9
DT (Quad)	0.70	1.14	60.6	1.31	0.31	70.4
RF	1.94	0.23	72.0	1.36	0.09	78.9
Cum	12.37	1.93	25.1	5.49	0.40	31.8
CC	19.67	4.06		8.84	0.82	

DT, decision tree; Lin, linear; Quad, quadratic; RF, random forest; Cum, cumulative logistic; CC, complete case; MSE, mean squared error.

연구 결과

모의실험 결과

Table 4는 카테고리가 5개인 경우 완전히 응답한 개체를 이용한 분석, 순서형 의사결정 나무, 랜덤 포레스트, 순서형 로지스틱 회귀분석을 이용한 대체자료의 평균, 평균제곱오차, 정확도를 제시하였다. 완전히 응답한 개체만을 이용한 분석은 편향이 19.67으로 가장 큰 값을 보였고, 이것은 완전히 응답한 개체만을 이용해서 분석한다면 편향된 결과를 제시할 수 있음을 의미한다. 순서형 로지스틱 회귀분석의 경우도 편향이 12.37로 모수적 모형을 이용하여 대체를 하는 경우 통계적 모형설정이 중요함을 확인할 수 있었다. 순서형 의사결정나무를 사용한 대체 결과 편향이 각각 0.61, 0.70로 제일 작았고, 랜덤 포레스트의 경우 편향이 1.94로 순서형 의사결정나무보다 약간 큰 값이었다. 반면 평균제곱오차의 경우 랜덤 포레스트가 0.23으로 가장 안정적이었다. 정확도의 경우 랜덤 포레스트가 약 72%로 제일 높았고, 순서형 의사결정나무가 각각 60% 내외였다.

범주가 3개인 경우도 완전히 응답한 개체만을 이용한 경우만 이용한 경우 편향이 8.84로 제일 컸으며 대체의 필요성을 확인하였다. 순서형 로지스틱 회귀분석도 편향이 5.49로 완전히 응답한 개체만을 이용한 방법과 유사한 결과를 확인할 수 있었다. 순서형 의사결정나무의 경우 오분류 함수를 이차함수로 지정한다면 편향이 1.31로 제일 작았다. 랜덤 포레스트의 편향은 1.36이었고, 평균제곱오차가 0.09로 여전히 제일 안정적인 결과를 확인할 수 있었다. 정확도의 경우 기계학습기법이 모두 70% 이상이었다.

Table 5. Results of case study

Methods	Bias	MSE	Accuracy (%)
DT (Lin)	4.87	0.25	54.8
DT (Quad)	5.00	0.26	53.7
RF	0.22	0.00	64.9
Cum	4.73	0.22	63.0
CC	11.23	1.26	

DT, decision tree; Lin, linear; Quad, quadratic; RF, random forest; Cum, cumulative logistic; CC, complete case; MSE, mean squared error.

사례연구 결과

사례연구 결과는 Table 5에 제시되어 있다. 사례연구 결과, 모의실험 결과와 비슷한 것을 확인할 수 있다. 무응답 대체를 하지 않고 응답한 개체만을 이용한 분석시 편향된 결과를 확인할 수 있었고 실제 자료에서도 무응답 대체의 필요성을 확인하였다. 모의실험결과와 유사하게 랜덤 포레스트가 편향이 0.22로 제일 작았으며 정확도가 약 64%로 제일 성능이 좋았다. 다음으로 순서형 로지스틱 회귀분석의 편향이 4.73이었고 순서형 의사결정나무가 4.87로 유사한 값을 보였다. 실제 자료에서는 대체 모형이 잘 만들어진다면 모수적 모형을 사용하여 대체를 하여도 편향이 크지 않은 대체 자료를 얻을 수 있음을 확인하였다.

고 찰

본 연구는 명목형 변수 중 순서형 변수에 대해서 기계학습기법을 사용해 대체 성능을 비교하였다. 변수의 성질을 반영하기 위해 순서형 의사결정나무, 내포된 순서를 무시하고 명목형으로 취급한 후 랜덤 포레스트, 모수적 방법인 순서형 로지스틱 회귀분석을 적용하였고, 완전히 응답한 개체의 자료와 비교하여 결과를 제시하였다. 모의실험과 사례연구 결과, 완전히 응답한 개체만을 이용해서 비교를 했을 때 원자료와 비교해서 편향이 발생해 결측 대체의 필요성을 확인하였다. 자료의 형태가 복잡한 경우 모수적 모형인 순서형 로지스틱 회귀분석은 완전히 응답한 개체만을 이용한 분석과 유사한 크기의 편향을 보였다. 하지만 결측대체를 위한 보조변수가 적절하게 설정되고 결측이 발생한 자료의 구조가 복잡하지 않다면 모수적 모형도 우수한 성능을 보였다. 랜덤 포레스트는 모의실험과 사례연구에서 모두 좋은 성능을 보여, 순서형 변수를 명목형 변수로 취급한 후 대체를 하여도 편향이 낮은 결과를 얻을 수 있었다. 다만, 변수의 카테고리가 많아지고 자료의

구조가 복잡할수록 순서형 변수의 순서를 반영하여 대체를 하는 것이 더 작은 편향을 얻을 수 있기 때문에 데이터의 구조와 변수의 카테고리에 맞게 대체 방법을 선택할 수 있다. 하지만, 의사결정나무는 편향을 줄이기 위해서 아주 많은 수의 가지를 생성하려고 하기 때문에 [5], 시간소모가 클 수 있다. 따라서, 연구자의 상황에 맞는 기계학습기법와 모수적 방법의 선택을 추천한다.

본 연구는 순서형 변수의 오분류 함수로 선형함수와 이차함수 형태를 고려하였다. 하지만 순서형 변수의 경우 1점에서 결측이 발생했을 때, 2점으로 대체하는 것 보다 5점으로 대체하는 것이 더 오분류 비용이 크기 때문에 다양한 형태의 오분류 함수를 개발해 볼 수 있다. 또한, 본 연구는 임의결측가정하에서 모의실험 및 사례연구를 진행하였는데, 비임의결측가정인 경우에는 어떤 결과가 있는지를 확인해 볼 수 있다.

결 론

현재 보건학분야에서 사용되는 자료에는 순서형 변수들이 많다. 예를 들어, 국민건강영양조사에서는 '주관적 건강인식 상태', '주관적 체형인식 상태', '주관적 스트레스 인지 정도' 등의 변수가 있고, 건강보험공단 자료에는 '흡연량', '음주량' 등의 변수가 있다. 결측치 대체를 하지 않고 자료를 분석한다면 결과에 편향이 발생하기 때문에 적절한 방법을 통한 결측치 대체가 요구된다. 순서형 변수에 결측이 발생했을 때, 본 연구에서 적용한 기계학습기법처럼 순서의 의미를 배제하고 명목형으로 취급한 후 대체를 할 수도 있고, 순서의 의미를 반영하여 대체를 할 수도 있다. 본 연구를 통해서, 순서형 자료를 사용하는 연구자들이 통계적인 분석을 위한 양질의 데이터를 얻는 과정에 도움이 되는 것을 기대한다.

ORCID

Serhim Son *https://orcid.org/0000-0003-4539-4392*
Hyonggin An *https://orcid.org/0000-0002-0566-758X*

REFERENCES

1. Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken: John Wiley & Sons; 2019.
2. Sterne JAC, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338:b2393. DOI: 10.1136/bmj.b2393
3. Wang L, Bautista LE. Serum bilirubin and the risk of hypertension. Int J Epidemiol 2014;44(1):142-152. DOI: 10.1093/ije/dyu242
4. Aris IM, Rifas-Shiman SL, Li LJ, Kleinman KP, Coull BA, Gold DR, et al. Patterns of body mass index milestones in early life and cardiometabolic risk in early adolescence. Int J Epidemiol 2019;48(1):157-167. DOI: 10.1093/ije/dyy286
5. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. J Big Data 2021;8(1):140. DOI: 10.1186/s40537-021-00516-9
6. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer; 2013.
7. Kunzmann K, Wernisch L, Richardson S, Steyerberg EW, Lingsma H, Ercole A, et al. Imputation of ordinal outcomes: a comparison of approaches in traumatic brain injury. J Neurotrauma 2021;38(4):455-463. DOI: 10.1089/neu.2019.6858
8. Jonsson P, Wohlin C. An evaluation of k-nearest neighbour imputation using likert data. 10th International Symposium on Software Metrics, IEEE; 2004.
9. Podolsky S. Ordinal variable imputation for health survey data: a comparison between machine learning and non-machine learning methods. 2021.
10. Galimberti G, Soffritti G, Di Maso M. Classification trees for ordinal responses in R: the rpartScore package. J Stat Softw 2012;47:1-25. DOI: 10.18637/jss.v047.i10
11. Ture M, Kurt Omurlu I. Determining of complexity parameter for recursive partitioning trees by simulation of survival data and an application on breast cancer data. J Stat Manage Syst 2018;21(1):125-138. DOI: 10.1080/09720510.2017.1386878
12. Cutler A, Cutler DR, Stevens JR. Random Forests. In: Zhang C, Ma Y; editors. Ensemble machine learning: methods and applications. New York: Springer; 2012, p. 157-175.