The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 1, pp.603-608, January 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2024.10.1.603

JCCT 2024-1-74

결측값 대체를 위한 데이터 재현 기법 비교

Comparison of Data Reconstruction Methods for Missing Value Imputation

김청호*, 강기훈**

Cheongho Kim*, Kee-Hoon Kang**

요 약 무응답 및 결측값은 표본 탈락, 설문조사에 대한 답변 회피 등으로 발생하며 정보의 손실 및 편향된 추론의 가능성이 있는 문제가 발생하게 되며, 이 경우 결측값을 적절한 값으로 바꾸는 대체가 필요하게 된다. 본 논문에서는 결측값에 대한 대체 방법으로 제안되었던 평균 대체, 다중회귀 대체, 랜덤 포레스트 대체, K-최근접 이웃 대체, 그리고 딥러닝을 기본으로 한 오토인코더 대체와 잡음제거 오토인코더 대체 방법을 비교한다. 결측값을 대체하는 이러한 방법들에 대해 설명하고, 연속형의 모의실험 데이터와 실제 데이터에 접목시켜 각 방법들을 비교하였다. 비교 결과 대부분의 경우에서 다중 대체 방법인 랜덤 포레스트 대체 방법과 잡음제거 오토인코더 대체 방법의 성능이 좋았음을 확인하였다.

주요어 : 평균 대체, 랜덤 포레스트 대체, K-최근접 이웃 대체, 오토인코더 대체, 잡음제거 오토인코더 대체

Abstract Nonresponse and missing values are caused by sample dropouts and avoidance of answers to surveys. In this case, problems with the possibility of information loss and biased reasoning arise, and a replacement of missing values with appropriate values is required. In this paper, as an alternative to missing values imputation, we compare several replacement methods, which use mean, linear regression, random forest, K-nearest neighbor, autoencoder and denoising autoencoder based on deep learning. These methods of imputing missing values are explained, and each method is compared by using continuous simulation data and real data. The comparison results confirm that in most cases, the performance of the random forest imputation method and the denoising autoencoder imputation method are better than the others.

Key words : Mean Imputation, Random Forest Imputation, K-nearest Neighbor Imputation, Autoencoder Imputation, Denoising Autoencoder Imputation

Ⅰ. 서 론

설문조사와 종단연구에서 보편적으로 발생하게 되는 결측값은 어떻게 처리하느냐에 따라 분석 결과에 심각한 영향을 줄 수 있는 중요한 문제이다. 결측값이 존재할 때 처리하는 방법은 두 가지로 나뉘는데, 결측값을 분석에서 제외시키는 결측값 제거 방법과 결측값을 적절한 값으로 대체하여 완전한 데이터를 구성한 뒤 분석에 사용하는 방법이 있다. 결측값을 제거하는 방법은 결측값이 다수 존재하는 경우나, 제거한 결측값이 전체 데이터에

*준회원, 한국외국어대학교 통계학과 석사 (제1저자)
**정회원, 한국외국어대학교 통계학과 교수 (교신저자)
접수일: 2023년 10월 20일, 수정완료일: 2023년 11월 14일
게재확정일: 2023년 12월 10일

Received: October 20, 2023 / Revised: November 14, 2023
Accepted: December 10, 2023
**Corresponding Author: khkang@hufs.ac.kr
Dept. of Statistics, Hankuk Univ. of Foreign Studies, Korea

유의미한 정보를 지니고 있을 경우의 정보 손실은 분석 결과의 편의(Bias)를 일으키게 된다. 이러한 이유로 본 논문에서는 결측값을 단순히 제거하지 않고 대체하는 방법을 살펴보고자 한다.

결측값은 생성된 이유에 따라 세 가지 분류로 나뉜다. 결측이 자료 안의 어떤 변수와도 관련이 없으며 결측값의 분포는 다른 모든 관측 값들과 무관한 경우에는 완전임의결측(Missing Completely At Random, MCAR), 결측이 일어난 관측값은 결측에 영향을 주지 않으나 다른 관측 값들이 결측에 영향을 끼치는 경우에는 임의결측(Missing At Random, MAR), 결측값과 결측 여부 간에 연관이 있을 경우에는 비임의결측(Missing Not At Random, MNAR)이라고 한다.

결측값을 대체하는 방법으로는 단일대체 방법과 다중대체 방법이 있다. 단일대체 방법은 한 개의 결측값에 대해 한 개의 값으로 대체하는 방법으로써, 자료를 수집한 사람이 가용한 모든 자료를 바탕으로 대체 값을 추정할 수 있으므로 활용성이 높아지고 분석이 쉽다는 장점이 있다. 하지만, 결측 때문에 발생하게 되는 불확실성을 고려하지는 않으므로 추정량의 분산이 과소 추정되는 문제점을 가지게 된다. 대표적으로는 사용하기에 간단한 평균대체(Mean Imputation) 방법이 있다.

다중대체 방법은 단일대체 방법에서의 불확실성을 고려하여 한 개의 결측값에 대하여 최소 두 번 이상 대체를 시행하여 여러 개의 대체된 자료를 이용하는 방법이다.[1] 이 방법은 결측값이 완전임의결측에 의한 경우뿐만 아니라 임의결측인 경우에도 사용할 수 있다. 본 논문에서는 [2]에 의해 제안된 베이지안 통계에서 모수적으로 사후분포를 추정하는 방법 중 한 변수씩 순차적으로 표본을 생성하는 방법인 순차회귀 다중대체(Sequential Regression Multiple Imputation)를 고려하였다.

순차회귀 다중대체는 한 번에 한 변수씩 표본을 생성하기 때문에 여러 변수를 고려하더라도 각 변수의 특성에 맞는 분포로부터 표본을 생성할 수 있다. 이러한 다중대체 방법을 기반으로, 다중회귀 대체 방법과 회귀 모형의 사후분포에서 추출한 예측값과 최소제곱법을 이용하여 구한 예측값들 사이의 예측 거리를 이용하여 평균으로 대체하는 예측 평균매칭(Predictive Mean Matching, PMM) 대체 방법을 사용하였다. 또한 대체를 위한 비모수적인 방법 중 [3]에 제안된 자료의 유사성 측도를 이용하는 결측값을 대체하는 방법인 K-최근접이웃(K-nearest Neighbors, KNN) 대체 방법을 사용하였고, [4]에 제안된 비모수적인 방법인 랜덤 포레스트(Random Forest) 대체 방법을 사용하였다.

다양한 통계적 방법으로 결측값을 대체하는 방법 이외에 딥러닝 기반의 방법으로 결측값을 대체하는 오토인코더(Autoencoder) 대체도 소개한다. 오토인코더란 입력변수를 가능한 한 그대로 복원해 출력하도록 학습된 신경망을 이르는 말이다. 원래는 차원축소 또는 특징변수 추출에 사용되었으나, 최근에는 생성모델을 이용하여 결측값을 대체하는 중요한 도구로 사용되고 있다. 또한, 입력벡터 변수에 잡음을 적절한 정도로 추가해 변수를 입력으로 사용한 오토인코더 방법인 잡음제거 오토인코더(Denoising Autoencoder) 대체 방법도 사용한다.

본 논문은 다음과 같이 구성된다. 2장에서는 결측값을 대체하는 다양한 방법을 소개하고, 3장에서는 여러 모형을 적용한 대체방법의 성능을 모의실험을 통해 비교한다. 4장에서는 실제 데이터를 이용하여 결측값 대체의 효과를 비교한다. 마지막으로 결론에서 비교 결과를 요약 정리한다.

II. 결측값 대체 방법

본 장에서는 서론에서 언급한 결측값을 대체하는 방법들을 소개한다. 결측값을 대체하는 방법 중 다중회귀 대체와 예측 평균매칭 대체는 베이지안 통계에서 사후분포를 추정하는 방법 중 조건부모형을 사용하는 [5]에 제안된 MICE(Multiple Imputation by Chained Equations) 알고리즘을 사용한다.

MICE란 순차적으로 표본을 생성하는 방법으로써, 한 번에 한 변수씩 표본을 생성하기 때문에 각 변수의 특성에 맞는 분포로부터 표본을 생성할 수 있다. 설사 변수간의 상관성이 순차적이지 않더라도, 해당 변수를 제외하고 나머지 변수들에 대한 조건부모형을 통해 표본을 생성할 수 있는데, 자세한 것은 [2]를 참고하면 된다.

1. 평균 대체

평균 대체(Mean Imputation)는 각 변수에서 응답자들의 평균을 구하여 그 층의 무응답을 이 평균값으로 대체하는 방법이다. 자료를 수집한 사람이 가용한 모든 자료들을 바탕으로 대체 값을 추정할 수 있기 때문에 활용성이 높아지고 분석이 쉽다는 장점이 있지만, 결측 때문에

발생하게 되는 불확실성을 고려하지는 않으므로 추정량의 분산이 과소 추정되는 문제점을 가지게 된다. 그럼에도, 사회과학조사에서 실제적으로 가장 많이 사용되고 있는 방법이기도 하다.

2. 다중회귀 대체

다중회귀 대체는 무응답이 있는 변수를 종속변수로 하고 응답된 보조변수들을 독립변수로 하는 선형회귀모형(Linear Regression)을 적용하는 방법이다. 다중 회귀 대체는 한 변수씩 순차적으로 표본을 생성하는 조건부분포모형을 사용한 방법으로서, [6]에서 확인할 수 있듯이 대체하고자 하는 결측이 있는 변수(X_k)를 반응변수로, 나머지 변수를 독립변수로 하여 다음의 회귀 모형을 적합시켜 대체한다.

$$X_k = X_{-k}^T \beta + \epsilon_{-k}$$

여기서, X_{-k}는 k번째 변수를 제외한 변수들의 자료이고, β는 이에 대응하는 회귀 계수 벡터, ϵ_{-k}은 정규분포를 따르는 오차이다.

3. 예측 평균매칭 대체

예측 평균매칭 (Predictive Mean Matching; PMM) 대체는 다중회귀 모형에 의해 대체된 값을 가장 가까운 관측값에 일치시키는 방법으로, 선형 회귀분석의 변형된 형태이다. 즉, 회귀모형의 사후분포에서 추출한 예측값과 최소제곱법을 이용하여 구한 예측값들 사이의 예측 거리를 이용하여 평균으로 대체하는 방법이다.[7]

다중회귀 대체 방법의 사후분포에서 추출한 β^*를 이용하여 결측된 자료들의 벡터인 $\hat{X}_k^{mis} = X_{-k}^T \beta^*$를 계산한다. 그리고 최소제곱법에 의해 구한 $\hat{\beta}$을 이용하여 대체하고자 하는 변수의 예측 벡터인 $\hat{X}_k^{obs} = X_{-k}^T \hat{\beta}$을 계산하고 다음의 예측거리 Δ를 최소화하는 X_k^{obs} 를 구한다. 본 논문에서는 [8]에서 사용된 알고리즘을 적용한다.

$$\Delta = \|\hat{X}_k^{obs} - \hat{X}_k^{mis}\|$$

4. 랜덤 포레스트 대체

랜덤 포레스트 (Random Forest; RF) 대체는 [4]에서 제안된 방법으로 MICE 알고리즘과 마찬가지로 종속변수를 결측값이 있는 변수 중 하나를 선정하고, 이를 제외한 나머지 독립변수로 하여 종속변수에 대한 예측값을 결측값의 대체 값으로 사용하는 것이다. 대부분의 단일 대체방법들과 달리, 수치형 및 범주형 변수가 혼합된 데이터에 적용할 수 있어 서로 다른 변수들로 구성된 각기 다른 나무에 서로 다른 데이터셋을 적용하므로, 단일 대체방법의 큰 약점인 변동성 측면과 다양성을 보완할 수 있다.[4]

5. K-최근접 이웃 대체

K-최근접 이웃(K-nearest Neighbors; KNN) 대체는 비모수적 대체 방법으로써,[3] 결측값을 갖는 개체와 가장 근접한 K개의 관측된 개체를 이용하여 결측값을 대체하는 방법이다. 근접 정도를 측정하는 기준으로는 유클리드 거리를 사용하여 결측값과 관측값들과의 유사도를 정의하였으며, 유사도가 높은 K개의 선택된 관측값들의 평균으로 결측값을 대체한다.

6. 오토인코더 대체

[8]에 제안된 오토인코더 (AutoEncoder; AE) 방법은 인코더로 입력 데이터를 압축하고 디코더로 압축한 데이터를 입력 형태의 데이터로 복원시키는 신경망을 활용한 딥러닝 방법이다. 딥러닝은 많은 경우 텍스트 자료 분석에 이용되었으며,[9] 이 신경망은 입력 계층에서 불완전한 데이터의 표현을 학습하고 출력 계층에서 새로운 값을 재현한다. 간단한 신경망처럼 보이지만 네트워크에 여러 가지 방법으로 제약을 줌으로써 어려운 신경망으로 만든다. 예를 들면, 은닉층의 뉴런 수를 입력층보다 작게 하여 데이터를 축소하거나 여러 은닉층을 쌓아서, 입력 데이터와 출력 데이터 간의 차이를 보다 줄여 줄 수 있고, 또한 입력 데이터에 노이즈를 추가한 후 원본 입력을 복원할 수 있도록 네트워크를 학습시키는 등 다양한 오토인코더가 있다.[10] 오토인코더에 대한 이론적인 배경과 절차는 [11]을 참고할 수 있고, 이를 이용한 결측값 대체를 다룬 최근 리뷰 논문으로는 [12]가 있다.

7. 잡음제거 오토인코더 대체

[13]에서 제안된 잡음제거 오토인코더 (Denoising AutoEncoder, DAE)는 데이터에 잡음이 추가되었을 때, 이러한 잡음을 제거하여 잡음을 붙이기 전 데이터와의 차이를 최소화하는 목적을 가지는 오토인코더이다. DAE

는 AE와 유사하지만 고의로 입력 데이터에 잡음을 추가하고, 추가된 잡음을 토대로 학습된 데이터에서 나오는 결과값이 잡음을 삽입하기 전의 순수 입력값인지를 확인하는 알고리즘이다. 정규분포를 이용하여 입력 데이터에 잡음을 추가할 수도 있으나 본 연구에서는 데이터를 20%의 비율로 드롭아웃[14] 하여 일부를 0으로 설정하여 잡음을 추가하는 방법을 적용한 후에 오토인코더를 실행한다. 손실 함수로는 기존 오토인코더와 같은 입력과 출력의 오차 제곱이 아닌 잡음으로 손상되기 이전의 초기 데이터와 출력의 오차 제곱으로 한다.

III. 모의실험

1. 개요

본 논문에서의 모의실험은 결측값 대체 방법에 따라 모형의 성능을 확인하기 위해 연속형 속성의 변수들을 가지며 각 변수의 평균과 분산의 크기를 다르게 하고, 또한 결측값을 MCAR과 MAR의 가정하에 결측을 발생시켜 자료를 생성한다. 결측값을 MCAR으로 했을 경우, 생성된 자료는 $X_1, X_2, \cdots, X_{10}$ 총 10개의 연속형 변수로 만들었고, 각각 변수들의 결측률을 5%, 10%, 15% 20%인 4가지 경우의 결측률을 고려하여 각각의 결측 자료를 생성하였다. 표본의 크기는 500, 1000, 2000으로 총 3가지 경우의 표본을 고려하여 각 100번의 모의실험을 반복하였다. 본 논문에서는 이 중에 결측률이 5%. 15%이고 표본크기가 1000인 경우의 결과만을 제시하며, 나머지 경우도 비슷한 결과를 얻을 수 있다.

결측 대체 방법으로는 II절에서 다룬 것들을 사용하였으며, k-최근접 이웃 대체 방법의 경우 k값을 1부터 10까지 변화시키며 구해진 값들의 평균을 이용하여 대체하였다. 성능 평가 방법으로는 실제 관측된 값과 추정된 값의 차이에 대한 측도인 평균 제곱근 오차(Root Mean Squared Error; RMSE)를 이용하였고, 총 100번의 반복을 통해 평균과 그에 따른 표준 오차를 계산하였다.

2. 완전임의결측 자료

다음에 주어진 바와 같이 변수 $X_1, X_2, \cdots, X_5$는 연속형 속성을 가지며 다변량 정규분포를 따르도록 설정하였고, 나머지 변수들은 지수분포와 균일분포에서 생성했다.

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \end{pmatrix} \sim N\left(\begin{pmatrix} 40 \\ 10 \\ 30 \\ 50 \\ 60 \end{pmatrix}, \begin{pmatrix} 5 & 30 & 30 & 20 & 15 \\ 30 & 10 & 25 & 15 & 5 \\ 30 & 25 & 15 & 20 & 15 \\ 20 & 15 & 10 & 20 & 15 \\ 15 & 5 & 15 & 15 & 25 \end{pmatrix} \right)$$

$$X_6 \sim \mathrm{Exp}(10), X_7 \sim U(10,50), X_8 \sim U(0,60),$$
$$X_9 \sim U(-10,40), X_{10} \sim \mathrm{Exp}(5)$$

[그림 1]은 데이터가 각 변수마다 데이터가 1000개 일 경우의 대체 방법에 대한 RMSE의 Boxplot 결과를 보여준다. 결측값 비율이 증가 할 때 마다 전체적으로 RMSE가 증가하는 것을 볼 수 있다. 결측 비율이 5%일 때는 랜덤 포레스트 대체 방법의 RMSE가 가장 좋았으며, 15%의 경우에는 잡음제거 오토인코더 대체 방법의 RMSE가 가장 좋았다. 제시하지는 않았지만 표준오차를 보면 랜덤 포레스트 대체 방법이 가장 작았으며, 평균값에 대한 변동성이 가장 작다고 할 수 있다.

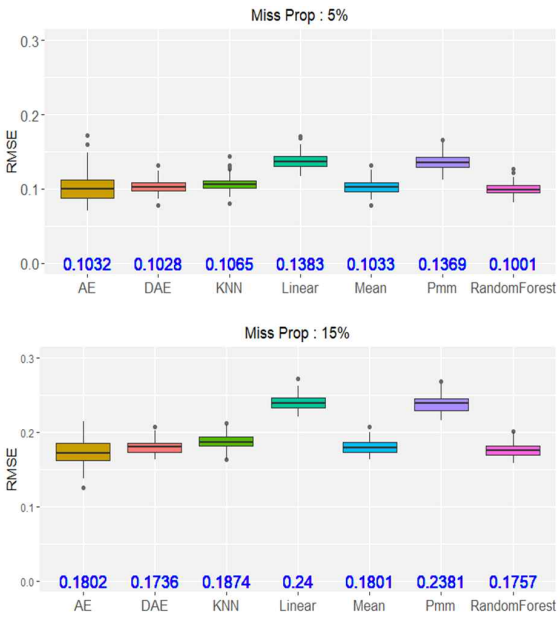

그림 1. 데이터 1000개의 각 대체 방법에 대한 RMSE 결과
Figure 1. RMSE results for each imputation for 1000 data

3. 임의결측 자료

데이터 구성에 결측을 발생시킬 변수를 첫 번째로는 $U(0,1)$의 분포에서 랜덤하게 하나의 변수를 생성한 뒤, 이 변수에 해당되는 데이터들의 값이 0.1보다 작을 경우 그에 대응되는 나머지 변수들의 값들 중 절반을 결측값으로 만든다. 두 번째의 가정은 첫 번째 가정에서 데이터의 값이 0.2보다 작을 경우에 나머지 변수들의 값들 중

절반을 결측값으로 만든다. 표본의 크기는 1000개로 설정 하였고, 반복횟수는 앞서 완전임의결측 자료와 동일하게 100번을 진행하여 대체 방법들에 대한 비교를 한다.

표 1. 임계값이 0.1, 0.2일 때 각 대체 방법에 대한 RMSE 결과
Table 1. RMSE results when the thresholds are 0.1, 0.2

MAR	AE	DAE	KNN	Linear	Mean	RF
0.1	0.1061 (0.0021)	0.1042 (0.0026)	0.1057 (0.0022)	0.1446 (0.0027)	0.1060 (0.0020)	0.1037 (0.0021)
0.2	0.1511 (0.0027)	0.1495 (0.0031)	0.1503 (0.0029)	0.2054 (0.0036)	0.1510 (0.0027)	0.1479 (0.0028)

[표 1]은 $U(0,1)$에서 샘플링한 임계값을 0.1과 0.2로 하여 이 값보다 작을 경우 결측값을 생성한 뒤에 각 각 대체 방법을 100번 반복했을 때 RMSE의 평균값과 표준오차에 대한 결과를 보여준다. 임계값을 0.1로 했을 때 RMSE의 평균값은 랜덤 포레스트가 가장 작게 나왔으며 그 뒤로 잡음제거 오토인코더와 K-최근접 이웃 대체 방법 순으로 결과가 나왔으며, 다중 회귀 대체 방법이 가장 크게 나왔다. RMSE를 추정값으로 하여 표준오차를 보면, 오토인코더 대체 방법과 평균 대체 방법이 가장 작고, 다중 회귀 대체방법이 가장 표준 오차가 크게 나왔다. 임계값을 0.2로 했을 때의 결과를 확인 해보면 0.1의 경우보다 전체적으로 RMSE의 평균값이 커지는 것을 확인할 수 있고, 방법들을 비교해보면 임계값이 0.1일 때와 마찬가지로 랜덤 포레스트 대체 방법이 평균값에서 가장 RMSE가 작게 나왔다.

Ⅳ. 실제 데이터를 이용한 비교

자료는 공공데이터포털[15]에서 제공하는 2019년 1월-7월 서울교통공사의 각 역들의 호선, 역 번호, 역명, 일평균 사용인원(명)에 대한 자료이다. 자료 중 본 논문에서 사용한 변수로는 1월-7월까지 각 역들의 호선 사용인원(명)이다. 결측 변수는 1월-7월까지 각 역들의 호선 사용인원(명)인 7가지 변수이다. 결측값 대체 효과를 비교하기 위하여 완전한 자료에서 임의로 결측값을 완전임의결측 방법으로 생성하였다. 그 후 Ⅱ장에서 소개한 방법들을 이용하여 결측값을 대체하였다. MCAR에 대한 결측값은 랜덤으로 각 각 5%, 10%을 형성 되도록 했고, 결과에 대한 비교는 대체 방법들을 각 변수마다의 RMSE를 계산하고 그 값들의 평균을 [표 2]에 제시하였다.

결측률을 5%로 했을때의 결측값을 MCAR으로 하여 생성했을 때, 각 각의 대체 방법 결과로는 잡음제거 오토인코더 대체 방법의 RMSE가 값이 가장 작게 나왔으며, 그 뒤로 랜덤 포레스트 대체 방법이 좋았다. 결측률 10%의 결과를 보면 5%와 마찬가지로 잡음제거 오토인코더 대체 방법이 가장 좋았으며, 평균 대체 방법의 RMSE가 가장 크게 나왔다.

표 2. 사용인원 데이터의 각 대체 방법에 대한 RMSE 결과
Table 2. RMSE results of each imputation for user data

MCAR	AE	DAE	KNN	Linear	Mean	RF
5%	0.0372	0.0151	0.0196	0.0699	0.1578	0.0171
10%	0.0858	0.0179	0.0559	0.0720	0.2537	0.0286

Ⅴ. 결론

본 연구에서는 결측값을 대체하는 방법들을 소개하고 비교하는 내용을 주로 다루었는데, 단일 대체 방법인 평균 대체 방법과 다중 대체 방법인 다중 회귀 대체, 예측평균매칭 대체, 랜덤 포레스트 대체, K-최근접 이웃 대체, 오토인코더 대체, 잡음제거 오토인코더 대체 방법까지 총 7가지 결측값을 대체 하는 방법을 사용하였다.

모의실험에 대한 전체적인 결과를 봤을 때, 잡음제거 오토인코더 대체 방법과 랜덤 포레스트 대체 방법의 결과가 대부분의 경우에 가장 RMSE의 평균값이 작았다. 또한, 대체 방법에 대한 성능 비교를 위해 RMSE를 계산하는 시스템 타임을 계산했는데, 그 결과를 보면 랜덤 포레스트 대체 방법이 시간 측면에서는 다른 대체 방법들에 비해 오래 걸리는 것을 확인할 수 있었다.

실제 데이터를 통한 실험의 결과에 앞서, 실제 데이터를 대체 했을 때와 실험 데이터를 대체 했을 때의 차이점으로는 실험 데이터는 분포를 가정했기 때문에 평균으로 대체했을 경우에 편향성이 크게 나타나지 않았지만, 실제 데이터의 결과에서는 다른 대체방법보다 평균 대체 방법의 RMSE값이 눈에 띄게 큰 것을 확인할 수 있었고, 잡음제거 오토인코더 대체 방법의 값이 가장 작은것을 확인했다.

오토인코더 대체 방법과 다양한 대체 방법들의 비교에 있어서 오토인코더 방법의 RMSE값이 가장 작은 것을 확인 할 수 있었다. 본 논문에서는 오토인코더 대체 방법을 발전시킨 잡음제거 오토인코더 대체 방법과 다른 다중 대체 방법을 추가하여 실험을 한 결과 잡음제거 오

토인코더 대체 방법이 RMSE 표준오차의 결과에서는 변동성이 크긴 했지만, 대부분은 좋은 결과를 얻었다. 사회과학 데이터에 대한 통계조사에서 결측값이 없는 자료는 드물고, 결측값이 있을 때 이를 무시하거나 단일 대체로 하여 대체를 하고 분석하면 편리할 수 있지만, 편향이 발생할 수 있으므로 다중 대체 방법들을 사용하는 것이 좋고, 잡음제거 오토인코더 대체 방법과 랜덤 포레스트 대체 방법이 성능면에서 가장 권장될만하다.

References

[1] Rubin, DB, *Multiple imputation for nonresponse in surveys,* John Wiley & Sons, New York, 1987

[2] Raghunathan TE, Lepkowski JM, Hoewyk JV, Solenberger P, “A multivariate technique for multiply imputing missing values using a sequence of regression models”, *Survey Methodology,* Vol. 27, pp. 85-95. 2001

[3] Dixon, JK, “Pattern recognition with partly missing data”, *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 9, pp. 617-621, 1979, DOI: 10.1109/TSMC.1979.4310090

[4] Stekhoven DJ, Bühlmann, P, “MissForest - non-parametric missing value imputation for mixed-type data”, *Bioinformatics,* Vol. 28, pp. 112-118. 2012, DOI: 10.1093/bioinformatics/btr597

[5] Van Buuren, S, Groothuis-Oudshoorn, K, “MICE: Multivariate imputation by chained equations in R”, *Journal of Statistical Software,* Vol. 45, pp. 1 - 67, 2011, DOI: 10.18637/jss.v045.i03

[6] Rubin, DB, “Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse”, *In proceedings of the survey research methods section of the American Statistical Association*, Vol. 1, pp. 20-28, 1978

[7] Little RJA, “A Test of Missing Completely at Random for Multivariate Data with Missing Values”, *Journal of the American Statistical Association*, Vol. 83, pp. 1198-1202, 1988, DOI: 10.1080/01621459

[8] LeCun Y, Bengio Y, Hinton GE, “Deep learning”, *Nature*, Vol. 521, pp. 436-444. 2015, DOI: 10.1038/nature14539

[9] Ko KH, “Study on Difference of Wordvectors Analysis Induced by Text Preprocessing for Deep Learning”, *The Journal of the Convergence on Culture Technology*, Vol. 8, No. 5, pp. 489-495, 2022, DOI: 10.17703/JCCT.2022.8.5.489

[10] Zhai J, Zhang S, Chen J, He Q, “Autoencoder and Its Various Variants”, 2018 *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Miyazaki, Japan, 2018, pp. 415-419, DOI: 10.1109/SMC.2018.00080.

[11] Bank D, Koenigstein N, Giryes, R. “Autoencoders”, available from arXiv:2003.05991v2, 2021, DOI: 10.48550/arXiv.2003.05991

[12] Pereira RC, Santos MS, Rodrigues PP, Abreu PH. “Reviewing autoencoders for missing data imputation: Technical trends, applications and outcomes”, *Journal of Artificial Intelligence Research*, Vol. 69, pp. 1255 - 1285, 2020, DOI: 10.1613/jair.1.12312

[13] Gondara L, Wang K, “MIDA : Multiple imputation using denoising autoencoders”, *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 260-272, 2018, DOI: 10.48550/arXiv.1705.02737

[14] Park JG, Choi ES, Kang MS, Jun YG, “Dropout Genetic Algorithm Analysis for Deep Learning Generalization Error Minimization”, *International Journal of Advanced Culture Technology*, Vol. 5, No. 2, pp. 74-81, 2017, DOI: 10.17703/IJACT.2017.5.2.74

[15] 공공데이터포털. https://www.data.go.kr/

※ 이 연구는 2023년도 한국외국어대학교 교원연구지원사업 지원에 의하여 이루어진 것임