A Comparison of Four Methods with two ways for Missing Data Imputation

임소현*

*** 이화여자대학교 통계학과**

요 약

결측치를 처리하는 문제는 모든 연구의 일부분이며 데이터 분석에 중요한 요소이다. 우리는 효율적이고 유효한 분석을 제공하기 위해 결측치를 제거하거나 적절한 값으로 대체해야 한다. 또한 결측치를 처리함에 있어서 train data 와 test data 가 각각 주어졌을 때, 나누어 처리할지, 합쳐서 처리할지는 중요한 문제이다. 따라서, 이 연구에서는 네 가지의 대입 방법 : Mean imputation, KNN imputation, MICE, MissForest 을 이용하여 비교해보고자 한다. 실제 분류 문제를 다루는 데이터를 이용하여 MCAR, MAR, MNAR 를 가정하고, 결측치를 잘 채워 넣는지와 잘 예측하는지에 대해 RMSE, Miscalssification error 로 비교해 보고자 한다.

키워드 : missing values, multiple imputation

1. 서 론

결측치는 대부분의 연구 분야에서 일어나는 공통적인 문제이다. Rubin 에 따르면 결측치는 세 가지 메커니즘에 의해 발생한다. MCAR(Missing Completely at Rnadom) 은 변수의 종류와 변수의 값과 상관없이 완전히 랜덤하게 발생한다. 이 경우에는 분석에 영향을 주지 않지만, 실제로 MCAR 인 경우는 거의 없다고 알려져 있다. MAR(Missing at Random)은 한 변수에 대하여 관측된 값에는 의존하지만, 관측되지 않은 값에는 의존하지 않는 경우이다. Dropout case 가 이 경우에 해당한다. MNAR(Missing Not at Random)은 누락된 변수의 관측된 값 및 관측되지 않은 값에도 의존하는 경우이다. 누락된 변수의 값과 누락된 이유가 관련이 있다.

Complete-Case Analysis(CC)는 결측치가 존재하는 관측치를 제거하는 방법으로 결측값이 MCAR 일 때, 주로 사용한다. 결측치를 쉽게 처리할 수 있다는 장점이 있지만, 데이터 수가 적거나, 결측치의 비율이 높다면 불필요한 정보의 손실이 일어날 수 있다. 따라서 결측치를 제거하지 않고, 적절한 값으로 대체하는 것이 중요하다. 본 연구에서는 네 가지 Missing Imputation 방법을 이용하여, 세 종류의 결측치(MCAR, MAR, MNAR)에 대해서 비교해 보고자 한다.

네 가지 방법 비교 시, train data, test data 를 나누어 처리하는 방법과, 두 데이터를 합쳐서 처리하는 방법으로 나누어서 비교해보고자 한다. 대부분의 분류 문제들은 train data 와 test data 를 나누어 분석을 진행한다. 기존 연구에서는 나누어져 있는 데이터에 대해서 missing imputation 을 할 때, 어떻게 처리하는지에 대해서는 다루지 않는다. 본 연구에서는 이 두 가지 방법에 대해서 중점적으로 다룸으로써 다른 연구들과의 차별점을 갖는다. 결측치를 채울 때, 해당 값을 실제 값과 비교해서 얼마나 잘 채웠는지(RMSE), 채운 후, 분류 문제에 대해서 얼마나 좋은 예측력을 보이는지(Misclassfication error)로 비교한다.

2. 모형 및 Imputation Methods

2.1 로지스틱 회귀모형

반응변수가 범주형 변수인 자료에 흔히 쓰이는 모형은 로지스틱 회귀모형(logistic regression model)이다. 반응변수(Y)가 두 개의 범주로 이루어진 경우, 이 변수는 성공에 대한 확률 $P(Y=1)=\pi$와 실패에 대한 확률 $P(Y=0)=1-\pi$를 갖는 이항분포로 표현할 수 있다. 로지스틱 모형은 이러한 반응변수 Y 와 여러 독립변수와의 관계를 설명하는데 사용된다.

$$\text{logit}(\pi_i) = log\frac{\pi_i}{1-\pi_i} = \beta \cdot X_i$$

$$\pi_i = \frac{e^{\beta \cdot X_i}}{1+e^{\beta \cdot X_i}}$$

2.2 평균 대치법(Mean Imputation)

이 방법은 관측 또는 실험되어 얻어진 자료의 평균값으로 결측값을 대치하는 방법이다. 평균 대치법은 사용하기가 간단하고 Complete Analysis 에 비해 효율성이 향상된다. 그러나 관측된 자료를 토대로 한 추정값으로 결측값을 대치함으로써 통계량의 표준오차가 과소 추정되는 문제가 있다.

2.3 K-Nearest Neighbor imputation

KNN 은 다차원 공간에서 가장 가까운 k 개의 이웃과 점을 연결하는 데 유용한 알고리즘이다. 연속형, 이산형, 범주형 등 데이터의 종류에 상관없이 데이터의 결측치를 처리하는 데 유용하다. 결측치를 처리할 때, KNN 을 사용하여 관측치가 있는 다른 변수들을 기준으로 값이 가장 가까운 점의 값으로 근사치를 구할 수 있다.

2.4 MICE

MICE 는 연쇄방정식을 이용한 대중결측대치 알고리즘으로, 단순 대치법에 비해 여러 대치를 생성하여 결측값의 불확실성을 관리한다. 변수별로 대체값을 만드는 모델을 지정하여 변수별로 결측치를 대체한다

1. 결측치들을 적당한 값으로 초기값을 정한다.
2. 결측치의 양에 따라 변수의 인덱스를 정렬하여 벡터를 k 로 둔다.
3. 종료 조건 γ 을 충족하지 않는다면 다음을 반복한다.:
4. X_{old}^{imp}에 이전에 결측 대치한 자료 행렬을 할당한다.
5. k 의 각 원소 s 에 대해 다음을 반복한다.
6. 랜덤 포레스트 모형 $y_{obs}^{(s)} \sim x_{obs}^{(s)}$을 적합시킨다.
7. 6 의 모형을 이용하여 $y_{miss}^{(s)} \sim x_{miss}^{(s)}$ 을 적합시킨다.
8. 추정된 $y_{miss}^{(s)}$를 이용해 결측 대치하여 x_{new}^{imp}행렬을 만든다.
9. γ 을 갱신한다.
10. 최종 결측 대치된 행렬 x^{imp}을 얻는다.

<표 1> Mice 알고리즘

2.5 MissForest

missForest 는 기계학습 알고리즘인 랜덤 포레스트(random forest)를 이용한 결측 대치 알고리즘이다. 알고리즘 실행 과정은 <표 2>과 같다.

1. 결측치들을 적당한 값으로 대체한다.
2. 결측치 양에 따라 변수의 index 를 정렬한 벡터를 k 로 둔다.
3. γ를 충족하지 않는 동안 다음을 반복한다.
4. X_{old}^{imp}에 이전에 결측 대치한 자료 행렬을 할당한다.
5. k 의 각 원소 s 에 대해 다음을 반복한다.
6. 랜덤 포레스트 모형 $y_{obs}^{(s)} \sim x_{obs}^{(s)}$을 적합시킨다.
7. 6 의 모형을 이용하여 $x_{mis}^{(s)}$로 $y_{mis}^{(s)}$를 추정한다.
8. 추정된 $y_{mis}^{(s)}$를 이용해 결측 대치하여 X_{new}^{imp}행렬을 만든다.
9. γ 를 업데이트 시킨다.
10. 최종 결측 대치된 행렬 X^{imp}를 얻는다.

<표 2> MissForest 알고리즘

종료조건 γ는 양의 실수이고, 알고리즘 실행 전에 미리 설정되는 값이다. γ는 이전 단계의 대치 결과와 현재 단계의 대치 결과 간의 차이가 충분히 작을 때 알고리즘을 종료한다. missForest 가 랜덤 포레스트 모형을 이용하기 때문에 알고리즘 운용을 조절하는 매개변수는 랜덤 포레스트 알고리즘의 매개변수를 따른다.

3. Datasets

Imputation 결과를 비교하기 위해서 UCI machine learning repository 에서 제공하는 *Blood Transfusion Service Center Data Set* 을 사용하였다. *Transfusion* 데이터는 2007 년 5 월에 헌혈 여부에 대한 분류 문제로, 총 499 개의 관측치와 5 개의 변수(Recency – 마지막 기증 이후 개월 / Frequency – 여태까지의 기증 횟수 / Monetary – 총 기증한 헌혈의 양(c.c). / Time – 첫 기증 이후 개월 / Y)를 가지고 있다. 전체 데이터의 종속변수(Y)는 0 이 271 개(54.3%), 1 이 178 개(35.7%)로 이루어져 있다.

3.1 분석 방법

Transfusion 데이터를 이용하여, U =$\alpha_1 I^* + \alpha_2$ Y + Z 라는 새로운 변수를 생성하였다. 여기서 I^*는 결측치를 생성할 변수를, Y 은 반응변수, Z 는 정규분포로부터 독립적으로 생성한 값이다. U 가 양수이면, I^*로 사용된 설명변수의 값을 NA 로 만들어 결측치를 생성하였다. MCAR, MAR, MNAR 은 α_1과 α_2의 값을 각각 다르게 설정하였다. 자세한 값은 다음과 같다.

1. MCAR Selection : α_1=$\alpha_2 = 0$
2. MAR Selection : α_1=0, α_2=1
3. MNAR Selection : α_1=1, α_2=0

위의 방법으로 세 종류의 결측치를 생성하였고, 결측치를 생성한 변수(I^*)에는 Monetary, Time 두 변수가 사용되었다. 결측치를 만든 전체 데이터의 70%을 training data, 나머지 30%을 test data 로 임의로 나누어 분석에 사용하였다. 결측치 종류별 데이터에 존재하는 결측치 수와 비율은 <표 3>과 같다.

		Monetary	Time	
MCAR	train	92 (29%)	102 (32%)	194 (15.3%)
	test	37 (28%)	48 (35%)	85 (15.3%)
	Total 279 (15.5%)			

MAR	train	92 (29%)	106 (33%)	198 (15.7%)
	test	40 (29%)	43 (32%)	83 (15.4%)
	Total		281 (15.6%)	
MNAR	train	93 (29%)	101 (32%)	194 (15.3%)
	test	36 (27%)	49 (36%)	85 (15.3%)
	Total		279 (15.5%)	

<표 3> 데이터의 결측치 수 (비율)

각각의 경우에 대하여 split 방법과 combine 방법을 이용하여 분석을 진행하였다. split 방법은 train data 와 test data 를 나누어 놓고, 각각의 데이터를 가지고 결측치 처리를 하여 분석하는 방법이다. combine 방법은 train data 와 test data 를 하나로 합쳐서 결측치 처리를 하여 분석하였다.

3.2 평가 기준

모델 성능은 두 가지 척도로 비교해보고자 한다.

평균 제곱근 오차(RMSE)는 일반적으로 회귀의 평가를 위한 지표로 오차를 제곱해 평균한 값에 루트를 씌운 값이다. 본 연구에서는 실제 값과 대입한 값의 차이를 사용하였고, 각 변수들의 가중치를 동일하게 해주기 위해 전체 데이터를 표준화 후 계산하였다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i^{obs} - X_i^{imputed})^2}{n}}$$

오분류율(Error rate, Misclassification rate)이란 전체 데이터에서 잘못 분류한 관측치의 비율이며, 직관적으로 모델 예측 성능을 나타내는 평가 지표이다.

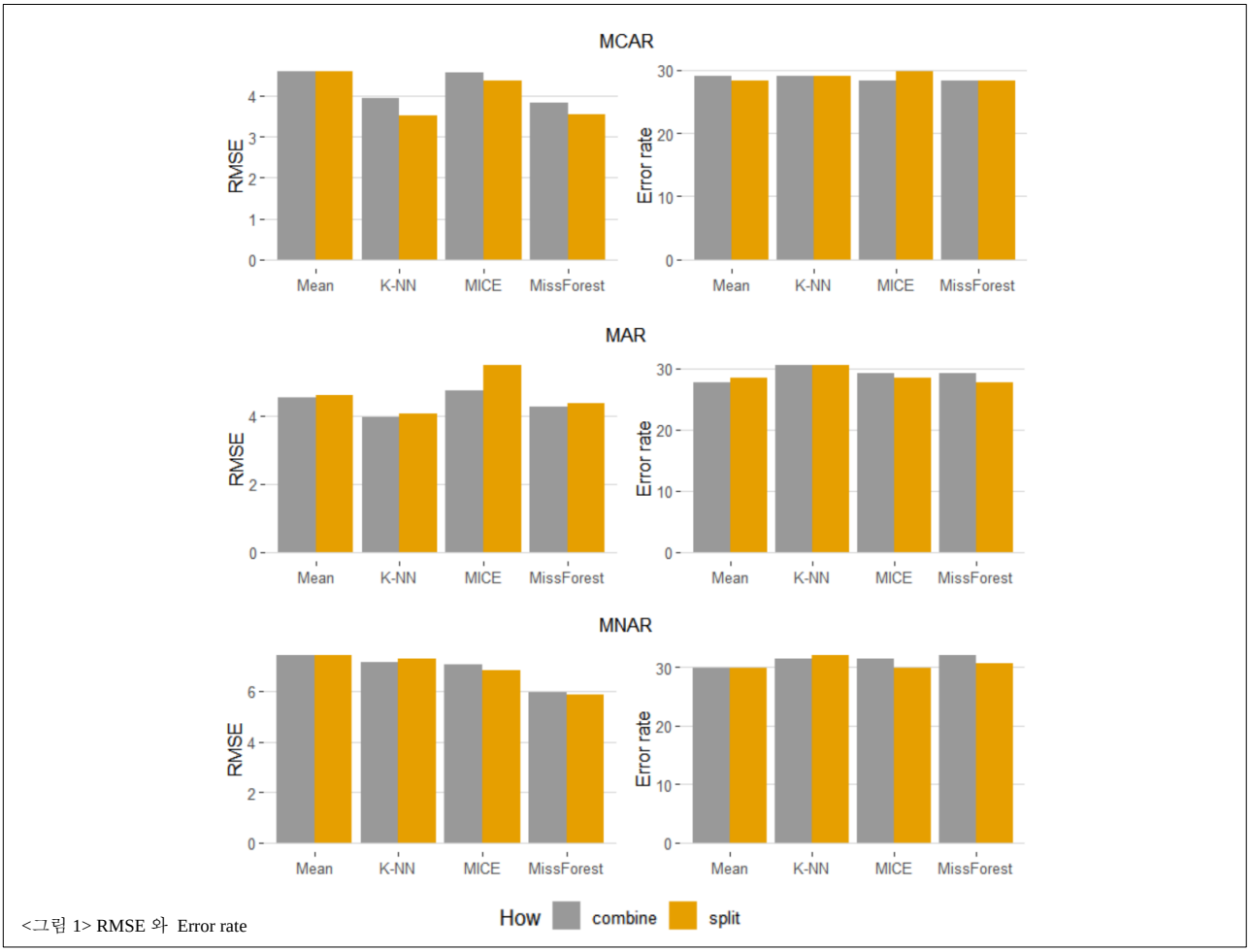
$$Error\ rate = \%\ of\ misclassified\ samples$$

결론적으로 대입된 값을 이용해 RMSE 를 구하고, 로지스틱 회귀로 모델링 후 예측된 값으로 오분류율을 구하여 여러 방법 별로 비교해 보았다.

4. 결 과

먼저 MCAR 에서 RMSE 의 경우, 평균값 대체를 제외하면 combine 보다 split 하여 결측치를 채웠을 때 낮은 오차를 보였다. Split 하였을 때는 K-NN, MissForest, Mice, Mean 순으로, combine 을 하였을 때는 MissForest, K-NN, Mice, Mean 순으로 좋았다.

오분류율의 경우 K-NN 과 MissForest 는 두 방법의 값



<그림 1> RMSE 와 Error rate

이 같았고, Mean 은 split 이, Mice 는 combine 이 좀 더 좋았다.

다음으로 MAR 일 때는 전체적으로 split 보다 combine 인 경우의 RMSE 가 더 낮았다. 특히 MICE 의 경우 차이가 0.75 정도로 다른 방법의 차는 0.1 인 것에 비해 큰 폭으로 차이가 났다. Split 과 combine 모두, k-nn, MissForest, Mean, Mice 순으로 좋았다. 예측력의 관점에서는 Mean 의 경우는 combine 일 때가 더 좋았지만, MICE 와 MissForest 는 split 에서 더 좋았다. K-NN 의 두 결과는 동일하였다.

마지막으로 MNAR 에서는 RMSE 값이 MissForest, Mice, K-NN, Mean 순으로 좋았고 K-NN 은 combine 인 경우에, MICE, MissForest 는 split 인 경우에 더 좋았다. 오분류율로 비교해 보았을 때는 Mean, Mice, MissForest, K-NN 순으로 좋았다.

5. 결 론

결측치가 있는 자료는 분석의 과정과 결과 전반을 왜곡시킬 수 있다. 따라서 결측치가 있는 자료에서 올바르게 데이터를 채우는 방법이 필요하다. 본 연구는 결측치를 대치할 때, MCAR, MAR, MNAR 이라는 가정하에 올바르게 대치하기 위해 split 방법과 combine 방법을 두 가지 측면(RMSE, 오분류율)으로 비교하였다. 표준화된 RMSE 를 비교함으로써 참값과 비교해 결측치를 잘 처리했는지를 검증했고, 오분류율 를 비교함으로써 예측력에는 어떤 차이가 있는지 검증하였다.

MCAR 의 경우 train 과 test 를 나누어 MissForest 로 결측치를 채웠을 때, RMSE 도 낮고 예측력 또한 좋았다.

MAR 의 경우에는 train 과 test 를 합쳐서 결측치를 채우되, 결측값이 잘 채워졌는지를 기준으로 두면 K-NN 을, 예측력 기준이면 평균값 대체가 좋았다. MissForest 는 RMSE 와 오분류율이 네 방법 중 제일 좋지는 않았지만 전반적으로 나쁘지 않았다.

MNAR 에서는 MissForest 가 두 데이터를 나누어 결측치를 처리하였을 때 다른 방법들에 비해 RMSE 가 현저히 낮았고, 예측력 기준에 있어서도 다른 방법과 비슷하였다.

다만, 한 가지의 데이터를 이용하여 분석을 진행하였기 때문에 본 연구의 결과가 다른 자료에도 똑같이 적용될 수 있는가에 대해서 확인이 필요하다. 또한 실제 데이터에는 MCAR, MAR, MNAR 이 섞여 존재하는데 한 가지로만 특정 지었기 때문에 추가 연구의 필요성도 제기된다.

몇 가지 한계점에도 불구하고 연구 의의는 크게 두가지의 기여점이 있다. 첫째, 결측치 대체값을 비교하는 기존 연구는 대부분 MCAR 을 가정하고 있지만, 우리는 MCAR, MAR, MNAR 세 가지 가정하에 분석을 진행하였다. 세 경우, 좋은 성능을 보인 방법론이 차이가 있었기 때문에 좀 더 자세한 결과를 얻을 수 있었다. 둘째, 결측값을 잘 대치했는지 뿐만 아니라 예측률의 측면에서도 평가를 하였다. 또한, 전체 데이터로만 대치를 진행하지 않고 train data 와 test data 를 나눈 split 방법으로도 분석을 진행하였기 때문에 다양한 시도를 한 것에 의의가 있다.

참고문헌

[1] Rubin DB (1976) Inference and missing data. Biometrika 63: 581-592.B

[2] Schmitt et al.(2015) A comparison of Six Methods for Missing Data Imputation, Biomet Biostat 6:1

[3] White et al.(2011) Multiple imputation using chained equations: issues and guidance for practice, AM Wood - Statistics in medicine.

[4] Therese D. Pigott (2001) A Review of Methods for Missing Data, Educational Research and Evaluation Vol. 7, No. 4, pp. 353-383

[5] Stekhoven et al.(2012) MissForest—non-parametric missing value imputation for mixed-type data, BIOINFORMATICS Vol. 28 no. 1, pages 112–118

[6] Stef et al (2010) Journal of Statistical Software

[7] Margarida G. M. S. Cardoso, UCI machine learning repository