JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 29 No. 4, pp. 39-46, April 2024
https://doi.org/10.9708/jksci.2024.29.04.039

Missing Value Imputation Technique for Water Quality Dataset

Jin-Young Jun*, Youn-A Min**

*Student, Graduate School of Engineering, Hanyang Cyber University, Seoul, Korea
**Professor, Dept. of Applied Software Engineering, Hanyang Cyber University, Seoul, Korea

[Abstract]

Many researchers make efforts to evaluate water quality using various models. Such models require a dataset without missing values, but in real world, most datasets include missing values for various reasons. Simple deletion of samples having missing value(s) could distort distribution of the underlying data and pose a significant risk of biasing the model's inference when the missing mechanism is not MCAR. In this study, to explore the most appropriate technique for handing missing values in water quality data, several imputation techniques were experimented based on existing KNN and MICE imputation with/without the generative neural network model, Autoencoder(AE) and Denoising Autoencoder(DAE). The results shows that KNN and MICE combined imputation without generative networks provides the closest estimated values to the true values. When evaluating binary classification models based on support vector machine and ensemble algorithms after applying the combined imputation technique to the observed water quality dataset with missing values, it shows better performance in terms of Accuracy, F1 score, RoC-AuC score and MCC compared to those evaluated after deleting samples having missing values.

▸ **Key words**: Water Quality Data, Missing Value, MCAR, MICE Imputation, Combined Imputation

[요 약]

많은 연구자들이 다양한 모델을 이용하여 물의 수질을 평가하기 위해 노력하고 있다. 평가 모델에는 결측값이 없는 데이터셋이 필요하지만, 관측 데이터셋에는 결측값이 다수 포함되는 것이 현실이다. 단순히 결측값을 삭제하는 방법은 경우에 따라 기저 데이터의 분포를 왜곡시키고 모델의 예측성능에도 편의(bias)를 불러올 위험성이 있다. 본 연구에서는 수질 데이터의 결측값 처리에 적합한 기법을 탐색하기 위해, 기존의 KNN과 MICE Imputation, 그리고 생성형 신경망 모델인 Autoencoder와 Denoising Autoencoder를 기반으로 몇 가지 대치 기법을 실험하였다. 실험 결과, KNN과 MICE Imputation의 결과를 평균한 Combined Imputation이 실측치에 가장 가깝게 값을 추정하였으며, 이 기법을 적용하여 결측값을 처리한 관측 데이터셋을 support vector machine과 ensemble 기반의 분류 모델로 평가한 결과, 결측값을 삭제했을 때에 비해 Accuracy, F1 score, ROC-AUC score, 그리고 MCC(Mathews Correlation Coefficient) 지표가 향상되었다.

▸ **주제어**: 수질 데이터, 결측값, MCAR, MICE 대치 기법, Combined 대치 기법

• First Author: Jin-Young Jun, Corresponding Author: Youn-A Min
*Jin-Young Jun (2022201681@hycu.ac.kr), Graduate School of Engineering, Hanyang Cyber University
**Youn-A Min (yah0612@hycu.ac.kr), Dept. of Applied Software Engineering, Hanyang Cyber University
• Received: 2024. 02. 23, Revised: 2024. 03. 15, Accepted: 2024. 03. 21.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

2020년 UN에 따르면, 전 세계 폐수 중 80%가 적절히 처리되지 못한 채 생태계로 다시 유입되고 있으며, 약 22억 명의 사람들이 안전한 식수에 접근하지 못하고 있다[1]. 안전한 식수는 인간을 포함한 지구상 모든 생물 종의 생명 유지를 위한 필수 자원일 뿐만 아니라, 경제 성장 및 사회의 발전과도 밀접한 관련이 있으므로, 많은 물환경 연구자들이 수질을 정확하게 평가하기 위해 노력하고 있다.

수질평가를 위해서는 다양한 기계학습 기반 모델이 사용될 수 있지만, 이러한 모델들은 모두 학습과 예측을 위해 결측값이 없는 완전한 데이터셋을 요구한다. 그러나, 현실에서는 다양한 이유로 결측값을 포함하는 불완전한 데이터셋이 되는 경우가 많다. 데이터셋으로부터 결측값이 포함된 표본을 '삭제'하는 것(listwise or case deletion)은 가장 기본적인 처리방법이지만 전체 표본의 크기를 줄어들게 하고, 결측 메커니즘이 완전 무작위(MCAR)가 아닌 경우에는 관측 데이터의 분포를 왜곡시킬 수 있다[2]. 이러한 문제를 해결하기 위해 단순히 '삭제'하는 전략 대신 특정한 알고리즘에 의해 추정된 값으로 결측값을 교체하는 '대치(Imputation)' 전략을 사용하기도 한다. 기계학습 기술에 초점을 맞춘 결측값 대치 기법 관련 문헌들을 조사한 기존 연구에서는 서로 다른 두 가지 종류의 데이터셋에 KNN(k-nearest neighbor) 기반의 대치 기법과 missForest 대치 기법을 적용한 실험을 통해, 결측값에 대한 대치 기법은 분석 대상 데이터의 특성과 결측률을 고려하여 선택되어야 함을 보여주었다[3].

본 연구는 수질 데이터셋에서 결측값을 삭제하는 대신, 전통적인 결측값 대치 알고리즘과 딥러닝 알고리즘에 기반한 몇 가지 생성형 대치 기법을 적용하여 결측값을 추정한 후, 실측치에 가장 가까운 값으로 추정하는 대치 기법은 무엇인지 알아보고, 대치 기법별 물의 음용성을 예측하는 이진 분류 예측모델의 성능을 비교하여 수질 데이터셋에 적절한 결측값 대치 기법은 무엇인지 알아보고자 한다.

II. Preliminaries

1. Related works

1.1 Water Quality Prediction

물의 음용성(water potability)을 예측하는 우수한 기계학습 알고리즘 모델을 탐색하기 위해 9개의 수치형 독립변수와 하나의 이진형 종속 변수로 구성된 3,276개의 표본으로 구성된 데이터셋을 기반으로, 결측값을 삭제한 후 다양한 기계학습 모델의 예측성능을 비교한 연구에서는 SVC가 가장 우수한 예측성능을 보였다고 하였다[4]. 동일한 수질 데이터셋을 기반으로 한 또 다른 연구에서는 결측값을 삭제하는 대신 평균으로 대치하고, 데이터 균형 맞춤을 위한 SMOTE 기반 oversampling을 추가하여 여러 기계학습 모델의 예측성능을 비교한 결과, SVC보다 Random Forest가 더 우수한 성능을 보였다고 하였다[5]. 동일한 수질 데이터셋을 기반으로 했던 두 선행연구의 결과는, 물의 음용성(water potability) 예측을 위한 수질 데이터에서 적절한 결측값 대치 기법의 적용이 예측모델의 성능에 영향을 주었음을 짐작하게 한다. 이상의 두 선행연구의 목적이 예측성능이 우수한 기계학습 모델을 개발하는 것에 있었던 반면, 본 연구의 목적은 예측성능 개선에 영향을 미칠 수 있는 적절한 결측치 대치 기법을 탐색하는 것이므로, 다음으로는 전통적인 결측값 대치 기법과 최근 많이 연구되고 있는 생성형 대치 기법에 관한 관련 연구를 중심으로 살펴보고자 한다.

1.2 Imputation Techniques

결측값 대치 기법은 추정값의 후보가 하나인지 여러 개인지에 따라 단일 대치(single imputation)와 다중 대치(multiple imputation)로 구분할 수 있다. 다중 대치는 실제 데이터의 분포를 제대로 반영하지 못하는 단일 대치의 문제를 해결하기 위한 것으로, 실제 값에 대한 불확실성을 반영하여 값을 추정할 수 있도록 여러 개의 single imputation 기법을 조합한 후 집계하여 추정한다[3,6].

대표적인 다중 대치 기법에는 선형회귀 모델로 다변량 대치(multivariate imputation)를 독립변수마다 반복하는 MICE(Multiple Imputation of Chained Equation)가 있다[7]. 다변량 대치란 결측값이 있는 독립변수별로 나머지 다른 독립변수들의 함수로 모델링한 후, 독립변수별 모델을 사용하여 추정하는 방법을 말한다. MICE와 유사하지만, 선형회귀 모델 대신 RandomForest를 사용하는 missForest는 생성된 모델 중 가장 우수한 모델 하나만을 이용하여 추정하므로 single imputation에 속한다고 볼 수 있다[8]. 한편 missForest와 유사하게 RandomForest를 적용하지만 가장 우수한 하나의 모델을 선택하는 대신 생성된 여러 개의 모델을 집계하여 최종값을 추정하는 MICE with RandomForest도 있으며, RandomForest와 달리 트리의 수직 확장 구조로 동작하는 LightGBM 기반의 MICE도 있다.

결측값 대치 기법은 관측된 데이터의 관계로부터 추정

하는 Discriminative imputation 방법과 관측된 데이터의 특성을 학습하여 재구성해 낸 데이터에서 추정값을 가져오는 Generative imputation으로 구분하기도 한다. MICE와 missForest는 Discriminative imputation이라 할 수 있고, 입력 데이터의 숨은 특성을 학습하여 입력과 유사한 데이터를 생성해 내는 Autoencoder를 사용한 결측값 대치 기법은 Generative imputation이라 할 수 있다. Generative imputation 관련 연구로는 결측값의 존재를 특별한 형태의 노이즈로 인식하고, Denoising Autoencoder(DAE)를 이용하여 대치할 값을 생성하는 모델을 연구한 MIDA와 kNN과 Denoising Autoencoder를 이용한 NAA(Neighborhood Aware Autoencoder), 그리고 적대적 생성망인 GAN을 이용하여 대치값을 생성하는 GAIN 등 많은 연구가 있다[9,10,11]. 본 연구에서는 수질 데이터셋의 구조가 복잡하지 않은 점과 선행연구인 MIDA의 내용을 고려하여 AE와 DAE를 기반 모델로 하는 생성형 imputation 기법에 관심을 두고자 한다.

1.3 AE and DAE

Autoencoder(AE)는 입력 데이터로부터 숨어 있는 특성(Hidden feature)을 찾는 Encoder와, Encoder가 생성한 Hidden feature를 이용하여 다시 원래의 입력 데이터를 재구성해 내는 Decoder로 구성된 비지도 학습 알고리즘이다[12]. Encoder가 입력 데이터의 차원보다 더 작은 크기의 Hidden feature를 생성하면 Undercomplete AE라 하고, 더 입력 데이터의 차원보다 큰 Hidden feature를 생성하면 Overcomplete AE라 한다.

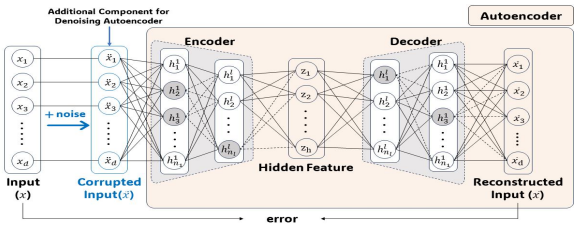

Fig. 1. AE and DAE Overview

Denoising Autoencoder(DAE)는 AE와 유사하지만, 입력 데이터가 어느 정도 손상된 상태라도 데이터가 존재하는 가상의 공간인 manifold 상에서는 같은 곳에 분포한다는 가정하에 입력 데이터에 약간의 노이즈를 추가한다는 차이가 있다[13]. AE와 DAE의 전반적인 구조는 Fig. 1과 같다.

III. The Proposed Scheme

본 장에서는 대표적 Discriminative Imputation technique인 KNN, MICE 그리고 이 둘을 조합한 Combined Imputation으로 구성된 세 가지 전통적인 대치 기법과 AE 및 DAE 기반의 몇 가지 생성형 대치 기법을 실험한다. AE와 DAE를 기반으로 결측값 대치 데이터셋을 각각 생성하고, 생성된 값과 실측치 간의 오차를 비교함으로써, 수질 데이터셋에 적합한 결측값 대치 기법은 무엇인지를 탐색한다. 또한, 실험 대치 기법들을 하나씩 관측 데이터셋에 적용하여 실험용 데이터셋들을 생성하고, 이들을 기반으로 여러 가지 이진 분류 모델을 학습한 후, 대치 기법별 모델들의 예측성능을 비교한다.

본 연구에서 사용하는 데이터셋은 물의 음용성을 예측하는 모델을 비교한 선행연구에서 사용한 수질 데이터셋과 동일한 독립변수와 종속 변수를 가진다. 이 수질 데이터셋의 특성 변수명과 설명 및 자료형, 결측의 크기 및 비율은 Table 1과 같다.

1. Experimental Setup

1.1 Dataset

Kaggle에 공개된 수질 데이터셋으로 Table 1과 같이 9개의 수치형 독립변수와 1개의 이진형 종속 변수 구조를 가진다[14]. 종속 변수인 Potability에는 결측값이 존재하지 않지만, 전체 표본 크기의 약 38.6%인 1,265건의 표본에서 적어도 하나 이상의 독립변수에 결측이 존재한다. 종속 변수의 분포 상태는, 음용이 가능함을 나타내는 “1”이 1,998건(61%)이고, 음용 불가능을 나타내는 “0”이 1,278건(39%)이다. Little의 통계적 가설 검정법에 따른 검정과 결측 상태의 시각화를 검토한 결과, 수질 데이터셋의 결측값은 완전 무작위적으로 발생한 것(MCAR, Missing Completely At Random)으로 판단되었다. 결측값 대치 기법의 효과를 확인하는 것이 목적이므로 독립변수를 선택하는 처리 과정인 Feature Selection을 거치지 않고 9개의 독립변수를 모두 사용한다.

1.2 Experimental Imputation Techniques

k개의 최근접 이웃의 평균을 구하여 추정하는 KNN Imputation과 LightGBM 기반의 MICE Imputation, 그리고, 딥러닝 모델인 AE와 DAE를 기반으로 다음과 같이 9가지의 결측값 대치 기법을 실험하였다.

Table 1. Water Quality Dataset

Variable Type	Columns	Description	Missing (rate)	Type
Independent variable	pH	A measure of how acidic/basic water is (1 ~ 14)	491 (14.98%)	Numerical (float)
	Hardness	Capacity of water to precipitate soap in mg/L	0	
	Solids	Total dissolved solids in ppm	0	
	Chloramines	Amount of Chloramines in ppm	0	
	Sulfate	Amount of Sulfates dissolved in mg/L	781 (23.84%)	
	Conductivity	Electrical conductivity of water in μS/cm	0	
	Organic_carbon	Amount of organic carbon in ppm	0	
	Trihalomethanes	Amount of Trihalomethanes in μg/L	162 (4.95%)	
	Turbidity	Measure of light emiting property of water in NTU	0	
Dependent variable	Potability	Indicates if water is safe for human consumption. Potable -1 and Not potable -0	0	Categorical
Total number of samples having any missing value at least one feature variable			1,265 (38.6%)	-

Case 1: KNN Imputation
Case 2: MICE Imputation(based on lightGBM)
Case 3: Combined Imputation(mean of Case 1 and 2)
Case 4: KNN + AE Imputation
Case 5: MICE + AE Imputation
Case 6: Combine + AE Imputation
Case 7: KNN + DAE Imputation
Case 8: MICE + DAE Imputation
Case 9: Combined + DAE Imputation

실험할 대치 기법 중 Case 4 ~ Case 9에서 AE와 DAE를 선정한 이유는 다음과 같다. 첫째, AE를 선택한 이유는 사용하는 수질 데이터셋은 독립변수의 크기가 9로 그리 크지 않으며, 연속형 독립변수로만 이루어진 비교적 단순한 구조라는 점과 imputation value를 생성하는 정확한 모델을 설계하는 것은 또 하나의 어려운 연구 문제라는 점을 고려하여 생성형 모델 중 가장 기본적이라 할 수 있는 모델부터 실험을 시작하기 위함이며, 두 번째로 DAE를 선택한 이유는 imputation 모델로는 노이즈를 처리할 수 있는 생성형 모델인 DAE가 이상적이라고 밝힌 선행연구인 MIDA를 고려한 것이다.

실험에서 AE와 DAE 모델의 학습 횟수는 모델의 학습 오차 수준이 더 낮아지지 않는 상태가 충분히 유지되는 경향을 보일 때까지 여러 값의 실험을 통해 최종 5,000회로 설정하였으며, Optimizer는 파라미터 최적화 과정에서 스텝 크기를 자동으로 조절하는 Adam을, 비선형 활성화 함수는 ReLU로 하였다. AE의 경우, 모델의 복잡성은 피하되, 하나의 은닉층을 가진 AE에 비해 중요한 특성만 더 심도 있게 학습할 수 있도록 은닉층에 4개의 층을 쌓아 Stacked AE 구조를 만들었으며, Encoder가 최종 출력하는 Hidden feature의 차원은 불필요한 특성이 제거될 수 있도록 입력 데이터의 차원보다 작은 4차원의 Undercomplete 구조로 하였다. 한편, DAE 모델은 AE와 동일하게 4개의 은닉층, Adam optimizer, ReLU 활성화 함수를 적용하며, 가우시안 분포를 따르는 임의의 노이즈를 관측 데이터에 추가하여 학습하게 하였다. 또한, Encoder가 최종 출력하는 Hidden feature의 차원은 독립변수 사이의 숨은 관계를 학습할 수 있도록 입력 데이터의 차원보다 큰 Overcomplete 구조로 하였다. 본 연구에서 사용한 DAE 모델은 Overcomplete 구조라는 점을 제외하면 은닉층의 수와 최적화 함수 및 활성화 함수, 그리고 Hidden feature의 차원과 노이즈 추가 처리 부분에서 선행연구인 MIDA와는 차이가 있다.

1.3 Experiment for Imputation Techniques

3,276개의 표본 크기를 갖는 관측 데이터셋에서 결측값을 모두 제거하고 남은 2,011개의 표본을 갖는 데이터셋(complete_dataset)을 추출하고 이 데이터셋의 복사본 3개를 생성한다. 각 복사본에 10%, 20%, 30%의 결측값을 완전 무작위(MCAR)로 발생시켜 3개의 결측 데이터셋을 만든다. 3개의 결측 데이터셋에 대한 실측치 데이터셋은 결측값을 모두 제거한 데이터셋인 complete_dataset으로부터 생성한다. 3개의 결측 데이터셋에 Case 1 ~ Case 9의 대치 기법을 각각 적용하여 결측값이 대치된 데이터셋들을 생성한 후, 결측률별 실측치 데이터셋과의 차이를 계산하여 오차를 구한다.

1.4 Evaluation Metrics for Imputation Techniques

결측값이 대치된 데이터셋을 실측치 데이터셋과 비교하여 그 오차를 MSE(Mean Square Error), RMSE(Root Mean Square Error), MAE(Mean Absolute Error) 기준으로 비교한다. RMSE와 MAE는 지표 값을 통해 오차의 크기를 직관적으로 이해할 수 있는 장점이 있지만, 모델 학습 과정에서 이상치(Outliers)에 더 큰 오차를 부여하지

는 못한다. 반면, MSE는 직관적 이해도는 낮으나 비교하는 두 대상의 차이가 클수록 더 큰 값의 오차로 계산되므로 모델이 학습할 때 이상치에 더 강인하게 대처하도록 유도할 수 있다는 장점이 있다.

1.5 Experiment for Binary Classification Models

SVC(kernel='poly')를 포함하여 Bagging 계열의 앙상블 기법인 RandomForest, Boosting 계열의 알고리즘인 LightGBM과 GradientBoosting 기법을 적용하여 4가지 이진 분류 모델을 생성하였다. 본 연구의 목적은 분류 기법을 탐색하는 것이 아니라 결측값의 대치 기법이 분류 성능에 미치는 영향을 관찰함으로써 수질 데이터셋의 결측값을 처리하기 위한 적절한 결측값 대치 기법을 탐색하기 위한 것이므로, 분류 예측모델은 모두 기본 파라미터 설정을 유지한다. 결측값이 있는 표본을 포함하여 총 3,276개의 관측 표본에 9가지 실험 대치 기법을 적용하여 생성한 각 데이터셋을 80:20의 비율로 나누어 예측모델의 학습과 평가에 사용하였다. 학습용 데이터셋의 표본 크기는 2,620개이며 평가용 데이터셋의 크기는 656개이다. Case 4 ~ Case 9에서 사용할 대치값 생성 모델은 30%의 결측률 데이터셋의 80%로 학습한 AE와 DAE로 하였으며, 모든 독립변수의 값은 0 ~ 1의 범위로 표준화하였다. 수질 데이터셋의 균형도는 61%:39%로 심각한 불균형은 아니므로, 수질 데이터에 Oversampling 기법인 SMOTE를 적용했던 선행연구와는 달리, 본 연구에서는 모델 학습 시 학습 데이터의 균형을 맞추기 위한 처리는 하지 않는다. 하지만, 평가에서는 불균형 상태를 고려할 수 있도록 다양한 평가지표를 사용하였다. 마지막으로, 3,276개의 표본으로 이루어진 관측 데이터셋에서 결측값을 모두 제거하고 남은 2,011개의 표본 데이터셋으로 학습하고 평가한 예측모델은 예측성능 비교를 위한 기준 모델(Case 0, baseline)로 삼는다.

1.6 Evaluation Metrics for Classification Models

물의 음용성 예측모델을 연구한 기존의 두 연구에서는 정확도(Accuracy)만을 예측모델의 성능 지표로 하였으나, 본 연구에서는 정확도와 함께 ROC_AUC score, F1 score, 그리고 Matthew Correlation Coefficient(MCC)를 성능 지표로 설정하였다. ROC_AUC score와 F1 score는 불균형 데이터셋에 대한 대표적 이진 분류 성능 지표이다. ROC_AUC score는 '가능'과 '불가능'으로 분류되는 두 가지 관점을 모두 중요하게 여기는 것으로, 0.5를 기준으로 이보다 크고 1에 가까울수록 우수한 성능을 나타낸다. 한편, 조화평균인 F1 score는 0 ~ 1의 범위를 가지며, 1에 가까울수록 성능이 우수하다. 물의 음용성 분류는 음용 가능한 물을 음용 가능하다고 예측하는 True Positive도 중요하지만, 음용 불가능한 물을 음용 가능하지 않다고 예측하는 True Negative 관점 또한 중요하다. 그러므로, True Positive 관점 중심의 F1 score 외에, True Negative 관점도 함께 고려하는 메튜 상관계수(Matthews Correlation Coefficient, MCC)를 평가 지표에 포함한다. MCC는 -1 ~ 1의 범위를 가지며, 1에 가까울수록 실제값과의 상관관계가 높고, 0에 가까울수록 상관관계가 낮으며, -1에 가까울수록 반대의 상관관계가 있다고 해석한다.

2. Results of Imputation Experiment

Case 1 ~ Case 9의 실험을 위해 준비된 데이터셋의 표본 크기는 Table 3과 같다.

Table 3. Sample size for imputation experiment

missing rate	ms_ds	N_df	T_df	M_df
10%	2011	1811	200	200
20%		1623	388	388
30%		1406	605	605

Table 3에서 "ms_ds"는 관측 데이터셋에서 결측값을 포함하는 표본을 모두 삭제한 데이터셋이고, M_df는 ms_ds에서 임의로 추가된 결측값을 포함하는 데이터셋의 크기이다. T_df는 M_df와 기본적으로 같지만, M_df에 결측으로 표시된 셀의 실측치를 가지고 있는 데이터셋이다. N_df는 ms_ds에서 결측값을 하나도 포함하지 않는 데이터셋으로서, AE와 DAE 모델의 학습을 위해 사용된 학습 데이터셋이다. Case 1 ~ Case 9의 대치 기법을 결측률별 M_df에 적용하여 생성한 결측값 대치 데이터셋(I_df)과 실측치 데이터셋(T_df) 간의 평균 오차는 Table 4와 같다.

Table 4. Imputation Errors

Case ID	Mean of 3 missingness cases		
	MSE	RMSE	MAE
Case 1	1.13e+06	**1057.18**	**91.87**
Case 2	1.86e+06	1364.43	120.08
Case 3	1.15e+06	**1062.3**	**94.37**
Case 4	2.47e+06	1556.37	136.8
Case 5	2.55e+06	1591.37	140.85
Case 6	2.36e+06	1523.80	134.38
Case 7	2.54e+06	1573.45	138.73
Case 8	2.63e+06	1610.84	142.63
Case 9	2.44e+06	1543.79	136.39

실측치와 가장 작은 오차를 보인 대치 기법은 Case 1인 KNN Imputation이었으며, 두 번째로 작은 오차를 보인 기법은 Case 3인 Combined Imputation이었다. Case 1과 Case 3을 제외한 나머지 기법들에서는 이 두 기법의 결과에 비해 비교적 큰 폭의 오차가 관찰되어, 수질 데이터셋의 결측값 대치 기법으로는 KNN Imputation이나 KNN과 MICE Imputation 결과의 평균으로 추정한 Combined Imputation이 적절한 것으로 보였다.

3. Experiment Results of Classification Model

결측값 처리 기법별 4가지 분류 모델(RandomForest, SVC, lightGBM, GradientBoosting)들의 평균 성능을 표로 정리한 것은 Table 5와 같다. Table 6은 True Positive와 True Negative의 관점을 모두 고려하는 지표인 MCC를 기준으로 모델별 결측값 처리 기법에 따른 예측성능을 비교한 것이고, Table 7은 True Positive 관점인 Precision과 Recall 성능을 동시에 고려하는 조화평균, F1 score 기준의 비교이며, Fig. 2는 Table 5의 결과를 시각화한 것이다.

Table 5. Mean Performances

Case ID	Acc	ROC_ AUC	F1 Score	MCC
0	0.612	0.667	0.507	0.250
1	0.604	0.658	0.452	0.231
2	0.582	0.626	0.422	0.180
3	**0.638**	**0.690**	**0.521**	**0.305**
4	0.588	0.635	0.417	0.195
5	0.586	0.630	0.423	0.188
6	0.587	0.633	0.416	0.192
7	0.601	0.630	0.447	0.228
8	0.590	0.625	0.435	0.199
9	0.598	0.627	0.448	0.218

IV. Conclusions

독립변수 사이의 상관성이 거의 없는 9개의 연속형 변수로 구성되었으며, 3,276건의 표본 중 38.6%에 해당하는 1,265건의 표본에 결측값이 존재하는 수질 데이터셋을 기반으로 결측값을 처리하기 위한 적절한 결측값 대치 기법을 탐색하고자 9가지의 실험 Case를 설정하였다. 실험 Case 별로 결측값을 대치한 데이터셋을 생성하여 실측값과 사이의 오차를 점검하였다. 또한, 실험 Case 별로 결측값을 대치한 데이터셋을 80:20으로 각각 학습용과 검증용 데이터셋으로 분할한 후, 4가지의 기계학습 모델에 학습용 데이터셋을 학습시켰다. 이후, 검증용 데이터셋을 이용하여 실험 Case별로 4개의 예측모델 성능 평균을 비교하였다.

실험 결과, 9가지 실험 기법(Case) 중 어떤 기법이 실측값과 가장 근사한 값을 생성해 내는지를 보기 위한 실험에서는 KNN(Case 1) 또는 KNN과 MICE 기법으로 생성된 데이터의 값을 평균한 Combined Imputation(Case 3)을 적용했을 때 실측값과의 오차가 가장 작았다. 그리고, 실험 기법별로 생성한 데이터셋을 이용하여 4가지의 기계학습 알고리즘 기반 예측모델을 학습시키고, 학습에서 사용하지 않은 검증 데이터셋으로 예측성능을 평가하여, 그 평균을 비교한 결과에서는 4개의 실험 평가 지표에서 모두 Combined Imputation을 적용한 경우가 가장 우수한 성능을 보였다.

한편, Table 6과 Table 7을 보면, 비록 SVC 모델의 예측성능에서 결측값을 삭제하고 예측모델을 학습했던 Case 0의 경우가 Case 3에 비해, MCC와 F1 score에서 조금 더 높게 (MCC: 0.31 vs 0.292, F1 score: 0.62 vs 0.605) 나타났으나, 이는 Case 0에 적용된 테스트 데이터셋의 크기(403개)가 Case 3에 적용된 테스트 데이터셋의 크기

Table 6. MCC of Classification Models

Model	Case 0	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9
SVC	**0.310**	0.225	0.159	**0.292**	0.232	0.213	0.213	0.216	0.223	0.216
RFC	0.285	0.244	0.177	**0.341**	0.240	0.211	0.218	0.294	0.239	0.259
LGBM	0.212	0.241	0.200	**0.270**	0.170	0.167	0.170	0.179	0.134	0.175
GNB	0.193	0.214	0.185	**0.316**	0.136	0.162	0.168	0.224	0.198	0.223
mean	0.250	0.231	0.180	**0.305**	0.195	0.188	0.192	0.228	0.199	0.218

Table 7. F1 scores of Classification Models

Model	Case 0	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9
SVC	**0.620**	0.556	0.530	**0.605**	0.548	0.538	0.536	0.540	0.542	0.539
RFC	0.447	0.365	0.324	**0.495**	0.361	0.366	0.341	0.420	0.394	0.397
LGBM	0.553	0.561	0.530	**0.568**	0.502	0.495	0.496	0.508	0.492	0.516
GNB	0.410	0.326	0.303	**0.414**	0.255	0.293	0.290	0.320	0.313	0.341
mean	0.507	0.452	0.422	**0.521**	0.417	0.423	0.416	0.447	0.435	0.448

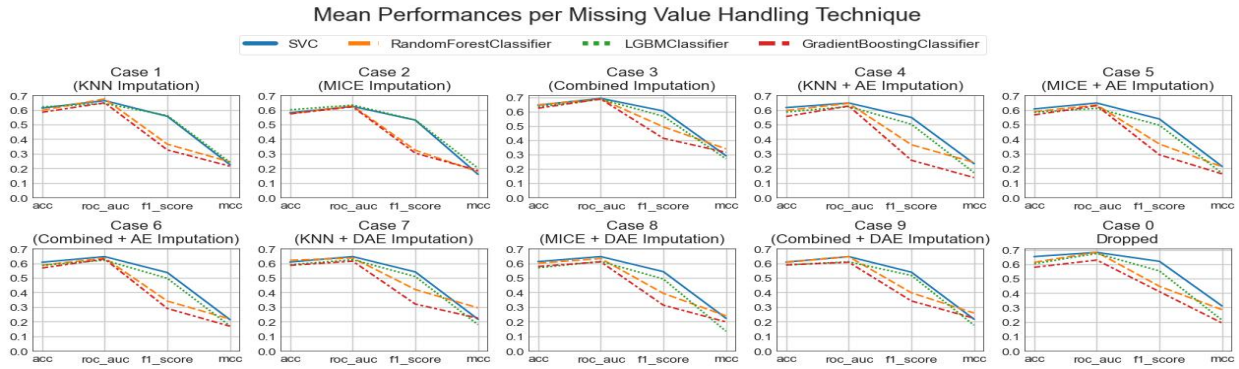

Fig. 2. Mean Performances per Missing Value Handling Technique

(656개)보다 38.6%나 작았음을 고려해 볼 때, 의미를 두기는 어려운 결과이다.

Table 8. Mean Effect of Combined Imputation

Perform. Metric	Deletion (Case 0)	Combined (Case 3)	Effect
Accuracy	0.612	0.638	0.026 ↑
ROC_AUC	0.667	0.690	0.023 ↑
F1 score	0.507	0.521	0.014 ↑
MCC	0.25	0.305	0.055 ↑

Table 8은 결측값이 삭제된 데이터셋과 Combined Imputation을 적용하여 결측값을 대치한 데이터셋으로 학습한 4개 예측모델의 평균 성능을 지표별로 정리한 것이다. 이상의 결과를 종합해 볼 때 다음과 같은 결론을 내릴 수 있다. 첫째, 수질 데이터셋에 대해서는 결측값을 삭제하는 것보다 Combined Imputation 기법으로 결측값을 대치하는 것이 다양한 기계학습 모델의 성능에 긍정적인 영향을 미칠 수 있다. 둘째, Combined Imputation 기법을 적용하면 SVC를 제외한 RandomForest, LightGBM, GradientBoosting 모델의 성능이 모두 개선된 것으로 보아, 적어도 트리 기반의 예측모델에서는 Combined Imputation 기법이 예측성능 향상에 긍정적인 영향을 줄 수 있을 것으로 보인다. 각 모델의 예측성능 중 MCC에 미친 Combined Imputation의 영향은 Table 9와 같다. 셋째, Combined Imputation을 적용했을 때 실험한 평가지표 중 MCC의 성능개선 폭이 가장 크게 나타난 것을 보아, True Positive뿐만 아니라 True Negative도 중요하게 평가되는 분야라면 Combined Imputation을 적용해 결측값을 대치하려는 노력을 해 볼 가치가 있을 것이다.

향후 Autoencoder 계열의 모델뿐 아니라 여러 가지 생성형 모델과 설정 조절에 기반한 후속 연구가 필요할 것으로 생각된다.

REFERENCES

[1] UN-Water, Water and Sanitation, *https://sdgs.un.org/topics/water-and-sanitation*

[2] Hyun Kang, "The prevention and handling of the missing data," Korean Journal of Anesthesiol, Vol. 64(5), 2013, pp. 402-406. DOI: 10.4097/kjae.2013.64.5.402

[3] Tlamelo Emmanuel, Thabiso Maupong, et al., "A survey on missing data in machine learning," Journal of Big Data, Vol. 8, Issue 1, Article no. 140, 2021, DOI:10.1186/s40537 -021-00516-9

[4] Jae-Ho Kim, Jang-Young Kim, "Performance Comparison of Classification performance Using Water Drinkability Data," Journal of the Korea Institute of Information and Communication Engineering, Vol. 27, No. 8, pp. 934-940, 2023, DOI: 0.6109/jkiice.2023.27.8.934

[5] Jinal Patel, Charmi Amipara, et al., "A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI," Computational Intelligence and Neuroscience, Vol. 2022, Article ID 9283293, 2022, DOI: 10.1155/2022/9283293

[6] Ratolojanahary Romy, Houé Ngouna Raymond, et al., "Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset," Expert Systems With Applications, 131 (2019) 299-307, DOI: 10.1016/j.eswa.2019.04.049

[7] L.L. Doove, S. Van Buuren, E. Dusseldorp, "Recursive partitioning for missing data imputation in the presence of interaction effects," Computational Statistics & Data Analysis, Vol. 72, April 2014, pp. 92-104, 2014, DOI: 10.1016/j.csda.2013.10.025

[8] Daniel J. Stekhoven, Peter Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," Bioinformatics, Vol. 28, Iss. 1, 2012, pp. 112-118, 2012, DOI: 10.1093/bioinformatics/btr597

[9] Gondara, L., Wang, K., "MIDA: Multiple Imputation Using Denoising Autoencoders," Advances in Knowledge Discovery and Data Mining. PAKDD 2018, DOI:10.1007/978-3-319-93040-4_21

[10] Konstantinos Psychogyios, Loukas Ilias, Dimitris Askounis, "Comparison of Missing Data Imputation Methods using the Framingham Heart study dataset," 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Ioannina, Greece, Sept. 2022, DOI:10.1109/BHI56158.2022.9926882

[11] Jinsung Yoon, James Jordon, Mihaela van der Schaar, "GAIN: Missing Data Imputation using Generative Adversarial Nets," Proceedings of the 35th International Conference on Machine Learning, Vol. 80, pp. 5689-5698, 2018

[12] Michelucci Umberto, “An Introduction to Autoencoders,” arXiv: 2201.03898 [cs.LG], 2022, DOI: 10.48550/arXiv.2201.03898

[13] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, Pierre-Antoine Manzagol, “Extracting and Composing Robust Features with Denoising Autoencoders,” Proceedings of the 25th International Conference on Machine Learning, pp. 1096-1103, Helsinki, Finland, July 2008

[14] Kaggle, Water Potability Prediction, *https://www.kaggle.com/code/ashfakyeafi/random-forest-with-water-quality/input*

Authors

Jin-Young Jun received the B.S. degree in Computer Science from Sungshin Women’s University, Korea, in 2003, and M.S. degree in e-Learning from Korea National Open University in 2020.

Jun is currently pursuing the second M.S. degree in Mechanical & IT Convergence Engineering from Graduate School of Engineering, Hanyang Cyber University, Seoul, Korea. She is interested in IoT, artificial intelligence and machine learning.

Youn-A Min Received a doctorate in computer engineering from Dongguk University. She served as a professor at Gachon University from 2016 to 2019, and has been working as a professor in the Department of Applied Software Engineering at Hanyang Cyber University since 2020. Dr. Min has been working as a professor in the Department of Applied Software Engineering at Hanyang Cyber University in Seoul since 2020, and is also a professor at the Graduate School of Mechanical and IT Convergence. She is interested in blockchain and blockchain-based artificial intelligence security.