한국정보전자통신기술학회논문지(jkiiect)'24-6, Vol.17 No.3
논문 24-17-03-176 http://dx.doi.org/10.17661/jkiiect.2024.17.3.176

결측치 비율이 높은 시계열 데이터 분석 및 예측을 위한 머신러닝 모델 구축

고방원*, 한용희**

Development of a Machine Learning Model for Imputing Time Series Data with Massive Missing Values

Bangwon Ko*, Yong Hee Han**

요 약 본 연구는 결측치 비율이 높은 시계열 데이터를 효과적으로 분석하고 예측할 수 있는 머신러닝 모델을 구축하기 위해 다양한 결측치 처리 방법을 비교 분석하였다. 이를 위해 PSMF(Predictive State Model Filtering), MissForest, IBFI(Imputation By Feature Importance) 방법을 적용하였으며, 이후 LightGBM, XGBoost, EBM(Explainable Boosting Machines) 머신러닝 모델을 사용하여 예측 성능을 평가하였다. 연구 결과, 결측치 처리 방법 중에서는 MissForest와 IBFI가 비선형적 데이터 패턴을 잘 반영하여 가장 높은 성능을 나타냈으며, 머신러닝 모델 중에서는 XGBoost와 EBM 모델이 LightGBM 모델보다 더 높은 성능을 보였다. 본 연구는 결측치 비율이 높은 시계열 데이터의 분석 및 예측에 있어 비선형적 결측치 처리 방법과 머신러닝 모델의 조합이 중요함을 강조하며, 실무적으로 유용한 방법론을 제시하였다.

Abstract In this study, we compared and analyzed various methods of missing data handling to build a machine learning model that can effectively analyze and predict time series data with a high percentage of missing values. For this purpose, Predictive State Model Filtering (PSMF), MissForest, and Imputation By Feature Importance (IBFI) methods were applied, and their prediction performance was evaluated using LightGBM, XGBoost, and Explainable Boosting Machines (EBM) machine learning models. The results of the study showed that MissForest and IBFI performed the best among the methods for handling missing values, reflecting the nonlinear data patterns, and that XGBoost and EBM models performed better than LightGBM. This study emphasizes the importance of combining nonlinear imputation methods and machine learning models in the analysis and prediction of time series data with a high percentage of missing values, and provides a practical methodology.

Key Words : Imputing, Machine learning, Missing values, Missingness, Time series data

1. 서론

시계열 데이터는 시간에 따라 관측된 값을 기록한 데이터로, 금융, 기후, 의료, 에너지 등 다양한 분야에서 중요한 역할을 한다. 이러한 데이터는 시간의 흐름에 따라 변동하는 패턴을 분석하고 미래를 예측하는 데 유용하다. 그러나 실세계 시계열 데이터는 종종 결측치가 발생하며, 이는 데이터 분석 및 예측의 정확성을 저해하는 주요 요인이 된다[1]. 결측치는 데이터 수집 과정에서의 오류, 장비 고장, 기록 누락 등 다양한 이유로 발생할 수 있다. 결측치가 높은 데이터는 특히 더 많은 도전과제를 제시하며, 이를 효과적으로 처리하

This work was supported by the Soongsil University Research Fund(Convergence Research) of 2020.
*Department of Statistics and Actuarial Science, Soongsil University
**Corresponding Author: Department of Entrepreneurship and Small Business, Soongsil University (amade@ssu.ac.kr)
Received June 10, 2024 Revised June 21, 2024 Accepted June 24, 2024

는 것은 예측 모델 구축의 핵심이다.

본 연구의 목적은 결측치 비율이 높은 시계열 데이터를 효과적으로 분석하고 예측할 수 있는 머신러닝 모델을 구축하는 데 있다. 이를 위해 본 연구에서는 다양한 머신러닝 모델을 비교 분석하여 결측치 처리 및 예측 성능이 가장 뛰어난 모델을 도출하였다. 결측치 비율이 높은 시계열 데이터는 분석 및 예측 과정에서 다음과 같은 여러 문제를 발생시킨다. 첫째, 결측치는 데이터의 패턴을 왜곡시켜 모델의 학습을 방해할 수 있다. 둘째, 결측치로 인해 데이터의 일관성이 떨어져 예측의 신뢰성을 낮출 수 있다. 셋째, 결측치를 적절히 처리하지 않으면 모델이 잘못된 결론을 도출할 위험이 존재한다. 따라서 결측치를 효과적으로 처리하는 방법을 모색하는 것은 매우 중요하다[2].

딥러닝 모델은 복잡한 패턴을 학습하는 데 강력한 도구이지만, 다음과 같은 이유로 본 연구에서는 딥러닝 모델이 아닌 머신러닝 모델을 사용하여 결측치 비율이 높은 시계열 데이터를 분석하였다.

1. 연산 자원 요구: 딥러닝 모델의 실행을 위해 대규모 데이터셋과 높은 연산 자원이 필요하며, 실시간 처리 시스템이나 임베디드 시스템과 같이 빠른 실행 속도가 필요하거나 시스템 자원이 낮은 경우 문제가 발생한다. 또한 결측치가 많은 데이터 처리에 사용되는 딥러닝 모델을 훈련시키기 위한 데이터 전처리, 모델 튜닝 과정이 복잡하고 비용이 많이 소요된다.

2. 과적합 문제: 딥러닝 모델은 데이터의 패턴을 과도하게 학습하여 과적합될 위험이 있다. 이는 특히 데이터의 결측치 비율이 높을 때 더욱 심각해질 수 있다.

3. 해석 가능성: 딥러닝 모델은 일반적으로 블랙박스 모델로 간주되며, 결과를 해석하는 데 어려움이 있다. 반면, 머신러닝 모델은 상대적으로 해석 가능성이 높아 결과를 이해하고 설명하는 데 유리하다.

본 연구에서는 결측치 비율이 높은 시계열 데이터에 먼저 PSMF(Predictive State Model Filtering), Miss Forest, IBFI(Imputation By Feature Importance)를 적용하였다. 이 알고리즘들은 효과적으로 결측치를 대체하며, 상대적으로 적은 연산 자원으로도 높은 성능을 발휘할 수 있다. 이후 LightGBM, XGBoost, EBM (Explainable Boosting Machines) 머신러닝 모델을 사용하여 모델의 정확도를 분석하였다. 또한, 각 모델의 예측 성능을 비교 분석하여, 시계열 데이터의 결측치 처리 및 예측에 가장 적합한 모델을 도출하였다.

본 논문의 구성은 다음과 같다. 2장에서는 시계열 데이터의 결측치 처리 방법과 딥러닝 및 머신러닝 모델의 적용 사례를 검토한다. 특히, LightGBM, XGBoost, EBM 모델의 특성과 장단점을 분석하였다. 3장에서는 본 연구에서 사용한 데이터셋과 모델 학습 및 평가 방법을 설명하였다. 4장에서는 각 모델의 예측 성능을 비교 분석하고, 실험 결과를 바탕으로 가장 적합한 모델을 도출하였다. 5장에서는 연구 결과를 요약하고, 본 연구의 한계 및 향후 연구 방향을 제시하였다. 본 연구는 결측치 비율이 높은 시계열 데이터의 분석 및 예측에 있어 실용적이고 효과적인 방법을 제시함으로써, 다양한 분야에서의 데이터 분석 및 예측 정확성을 향상시키는 데 기여하고자 한다.

2. 선행 연구

결측치 처리는 데이터 분석에서 매우 중요한 문제로, 여러 연구에서 다양한 접근 방법이 제시되었다. 결측치 처리 방법은 크게 통계적 기법, 행렬 기반 기법, 회귀분석 기법, 딥러닝 기반 기법으로 나눌 수 있다. 이러한 방법들은 각각의 장단점이 있으며, 시계열 데이터의 특성에 따라 선택적으로 적용될 수 있다.

통계적 기법은 결측치 처리를 위한 가장 기본적인 접근 방법으로, 주로 단순 대입법과 다중 대입법이 사용된다. 단순 대입법에는 결측치 주변 값의 평균, 중앙값, 최빈값 등을 대체하는 방법이 포함된다. 이러한 방법들은 간단하고 빠르지만, 데이터의 패턴을 제대로 반영하지 못하는 단점이 있다. 이를 보완하기 위해 다중 대입법(multiple imputation)이 사용되며, 이는 결측값을 여러 번 대체하여 대체값의 불확실성을 반영한다.

행렬 기반 기법은 데이터를 저차원 행렬로 분해하여 결측치를 대체하는 방법이다. PSMF 등의 행렬 분해 기반 접근법은 데이터를 두 개의 저차원 행렬로 분해하고 이를 통해 원래 데이터를 복원하는 과정을 포함한다.

회귀분석 기법은 과거 데이터를 기반으로 결측치를

예측하는 방법이다. 가장 간단한 방법으로는 선형 회귀 분석이 있으며, 이는 빠르고 간단하게 결측치를 처리할 수 있다. 그러나 선형 회귀분석은 전체적인 시계열 특성을 반영하지 못한다는 단점이 있으며, 이를 보완하기 위해 AR(AutoRegressive) 모델이 도입되었다. AR 모델은 과거의 값을 이용하여 현재의 값을 예측하는 방법으로, 시계열 데이터의 시간적 종속성을 반영한다. 딥러닝 기법은 결측치 처리에 있어서 높은 성능을 발휘하며, RNN(Recurrent Neural Network), GAN(Generative Adversarial Network), TCN(Temporal Convolutional Networks)은 복잡한 데이터 패턴과 구조를 학습하여 결측치를 예측하는 데 효과적이다. RNN은 시계열 데이터의 시간적 특성을 모델링하여 결측치를 처리하며, GRU-D(Gated Recurrent Unit for missing Data)와 같은 변형 모델들은 결측치의 시간적 간격을 고려하여 더욱 정밀한 예측을 가능하게 한다. GAN은 생성자, 판별자의 적대적 학습을 통해 결측치를 예측하며, GAIN(Generative Adversarial Imputation Networks)과 같은 모델은 결측 데이터의 분포를 학습하여 실제와 유사한 데이터를 생성한다. TCN은 CNN(Convolutional Neural Network)을 기반으로 한 모델로, 시계열 데이터의 패턴 인식을 강화한다. TCN은 고정된 크기의 필터를 사용하여 시간의 흐름에 따른 패턴을 학습하며, 병렬 처리가 가능하다는 장점이 존재한다. 그러나 TCN 또한 대규모 데이터셋과 높은 연산 자원을 필요로 하며, 결측치 비율이 높은 데이터에서의 정확도가 높지 않다.

본 연구에서는 결측치 비율이 높은 시계열 데이터 분석에 적합한 머신러닝 모델을 검토하였다. 머신러닝 모델은 딥러닝 모델보다 연산 자원 요구가 적고, 해석 가능성이 높은 장점이 존재한다. LightGBM은 마이크로소프트가 개발한 그레이디언트 부스팅 프레임워크로, 빠른 학습 속도 및 높은 효율성을 제공한다. 이 모델은 결측치 처리를 내장하고 있으며, 대규모 데이터셋에서도 높은 성능을 발휘한다. LightGBM은 트리 기반의 학습 알고리즘으로, 데이터의 분포, 특성을 효과적으로 반영할 수 있다[3]. XGBoost는 그레이디언트 부스팅 알고리즘을 기반으로 하며, 높은 예측 정확도, 효율성을 제공한다. XGBoost는 결측치 처리 및 과적합 방지를 위한 다양한 기능을 제공하여, 시계열 데이터 분석에 유리하다. 또한, 이 모델은 다양한 하이퍼파라미터 튜닝 옵션을 제공하여 모델 성능을 최적화할 수 있다[4]. EBM은 모델의 해석 가능성을 높이기 위해 개발된 알고리즘으로, 변수 간 상호작용을 직관적으로 이해할 수 있다. EBM은 부스팅 알고리즘의 일종으로, 각 변수의 효과를 독립적으로 추정하며, 모델의 투명성과 해석 가능성을 제공한다[5].

선행 연구에서는 결측치 처리 및 시계열 데이터 분석을 위한 다양한 방법들이 제안되었다. 딥러닝 모델은 복잡한 패턴을 학습하는 데 강력한 도구이지만, 높은 연산 자원과 과적합 문제로 인해 결측치가 많은 데이터에서는 한계를 가진다. 반면, 머신러닝 모델은 결측치를 효과적으로 처리하면서도 높은 예측 성능을 발휘할 수 있다. 이와 같은 다양한 접근 방법 중에서, 본 연구에서는 결측치 비율이 높은 시계열 데이터 분석에 적합한 머신러닝 모델을 검토하기 위해 PSMF, MissForest, IBFI를 선택하였다. 이들 방법을 선택한 이유는 다음과 같다.

PSMF: 상태 모델(state model)과 관찰 모델(observation model)을 사용하여 이전 상태와 현재 관찰값을 바탕으로 다음 상태를 계산하며, 이를 통해 시간에 따른 시스템의 동작을 예측하고 잡음을 포함한 데이터를 필터링한다. PSMF는 과거 상태와 관찰을 기반으로 하는 선형 모델로 미래 상태를 예측하므로, 선형적인 데이터 패턴에서의 예측 정확도가 높다[6].

MissForest: 랜덤 포레스트(Random Forest)를 기반으로 하여, 각 특성의 결측치를 반복적으로 대체하는 과정에서 비선형적인 관계를 학습한다. 랜덤 포레스트는 여러 개의 결정 트리를 결합한 앙상블 학습 방법으로, 비선형적 관계와 상호작용을 잘 포착할 수 있다. MissForest의 절차는 다음과 같다.

1. 초기 대체: 모든 결측값을 초기값(일반적으로 평균 또는 중앙값)으로 대체한다.

2. 모델 학습: 각 변수에 대해, 결측치가 없는 데이터를 사용하여 랜덤 포레스트 모델을 학습한다.

3. 결측치 대체: 학습된 모델을 사용하여 해당 변수의 결측치를 예측하고, 예측값으로 대체한다.

4. 반복: 2번과 3번 과정을 반복하며, 모델이 수렴

할 때까지 진행한다.

이 과정에서 랜덤 포레스트는 변수 간 복잡한 상호작용과 비선형적 패턴을 학습하여, 결측값을 더욱 정확하게 대체한다[7].

IBFI: 특성 중요도를 기반으로 결측치를 대체하는 방법으로, 각 특성이 종속 변수에 미치는 영향을 고려하여 결측값을 대체하는 데 주안점을 둔다. IBFI의 절차는 다음과 같다.

1. 특성 중요도 계산: 랜덤 포레스트를 사용하여 각 특성의 중요도를 계산한다. 랜덤 포레스트는 특성의 비선형적 관계와 상호작용을 효과적으로 포착할 수 있다.

2. 결측치 대체 순서 결정: 계산된 특성 중요도를 기반으로, 결측치가 있는 특성들을 대체할 순서를 정한다. 중요한 특성부터 결측치를 대체함으로써, 모델의 예측 성능을 극대화할 수 있다.

3. 대체: 각 특성에 대해, 결측치가 없는 데이터를 사용하여 모델을 학습하고, 이를 통해 결측치를 예측하여 대체한다.

IBFI는 랜덤 포레스트의 특성 중요도를 활용함으로써 데이터 내의 비선형적 패턴과 상호작용을 잘 반영하며, 중요한 특성부터 결측치를 대체하는 접근법은 모델의 예측 정확도를 향상시킨다[8].

이러한 결측치 대체 방법들은 각기 다른 특성을 가지며, 결측치 비율이 높은 시계열 데이터에서 발생할 수 있는 다양한 문제를 효과적으로 처리할 수 있다. 선행 연구 모두 공통적으로 다양한 결측치 처리 방법을 적용하여 데이터의 패턴을 보존하고 머신러닝 모델을 활용하여 예측 성능을 극대화하려고 시도하였으며, 본 연구도 그러한 시도의 일환이다. 하지만 통계적 기법, 행렬 기반 기법, 회귀분석 기법은 기법의 원리상 가설 공간(hypothesis space)이 한정되어 복잡한 데이터 패턴과 구조를 학습하여 결측치를 예측하는 데 한계가 존재한다. 딥러닝 기법은 매우 풍부한 가설 공간을 가지고 있으므로 복잡한 데이터 패턴과 구조 학습에 강점을 보이나, 1장에서 설명한 것처럼 필요한 연산 자원 수준이 높아서 적용에 제약이 존재한다. 본 연구는 결측치 비율이 높은 데이터를 대상으로 더욱 정교한 비선형적 결측치 처리 방법(MissForest, IBFI)을 적용하여 기존 기법인 PSMF와 비교하고, 예측 모델로서 적용에 제약이 거의 존재하지 않고 빠른 실행 속도를 가진 머신러닝 기법인 XGBoost, LightGBM, EBM의 성능을 분석한 점에서 기존 연구와 차별화된다.

따라서 본 연구에서는 이들 방법을 사용하여 결측치를 대체한 후, 머신러닝 모델을 적용하여 예측 성능을 평가하였다. 본 연구는 상대적으로 적은 연산 자원으로도 높은 성능을 발휘하는 LightGBM, XGBoost, EBM과 같은 머신러닝 모델을 사용하여 결측치 비율이 높은 시계열 데이터를 분석하고 예측하는 데 주력한다. 본 연구는 다양한 머신러닝 모델의 성능을 비교 분석하여, 시계열 데이터의 결측치 처리 및 예측에 가장 적합한 모델을 도출하였다.

3. 연구 방법 및 연구 결과

본 연구에서는 결측치 대체 방법을 적용한 데이터를 사용하여 각 머신러닝 모델(XGBoost, LightGBM, EBM)을 학습시켰다. 머신러닝 모델을 사용하여 시계열 데이터를 분석하기 위해 다음과 같은 과정을 수행하였다.

1. 모델 학습: 결측치가 대체된 데이터를 사용하여 머신러닝 모델을 학습시킨다.

2. 예측 수행: 학습된 모델을 사용하여 테스트 데이터에 대한 예측을 수행한다.

3. 성능 평가: 예측 결과를 기반으로 성능 지표를 계산하여 모델의 성능을 평가한다.

결측치 비율이 높은 데이터 분석 및 예측 모델 평가의 경우 두 가지 평가 방법이 존재한다. 결측치 비율이 매우 낮은 데이터셋의 경우 데이터를 일부 삭제를 통해 결측치 비율을 높인 후 결측치 처리 모델로 삭제한 값을 대체하고 대체된 값과 원 값 간 차이를 통해 모델의 성능을 직접적으로 비교한다. 결측치 비율이 매우 높은 데이터셋의 경우 결측치 대체를 비교할 정답이 없으므로, 결측치 대체 모델을 통해 결측치가 채워진 데이터에 분류나 회귀를 추가로 수행하여 전체 모델의 성능을 간접적으로 비교한다. 본 연구에서는 두 평가 방법을 모두 적용하였으며, 평가에 사용된 첫 번째 데이터셋은 한국환경공단 산하 대기환경 정보 실시간 공개 시스템인 에어코리아[9]에서 제공하는 서울시의 미세먼지(PM10), 초미세먼지(PM2.5) 시계열 데이터이

다. 데이터 수집 기간은 2023년 1월 1일부터 12월 31일까지이며, 수집 주기는 1시간이다. 데이터셋에는 표 1과 같이 시간별로 측정된 미세먼지, 초미세먼지 농도가 기록되어 있으며, 결측치 비율이 4.6%로 너무 낮으므로 본 연구에서는 높은 결측치 발생 상황을 시뮬레이션하기 위해 무작위로 30%의 데이터를 추가로 제거하였다.

표 1. 서울시 미세먼지, 초미세먼지 시계열 데이터
Table 1. Time series data on fine and ultra-fine particulate matter in Seoul

Location	Time	PM10	PM2.5
Jung-gu	2023010101	57	52
Jung-gu	2023010102	65	60
Jung-gu	2023010103	73	63
Jung-gu	2023010104	76	71
Jung-gu	2023010105	78	73
Jung-gu	2023010106	68	61
Jung-gu	2023010107	67	58
Jung-gu	2023010108	73	65
Jung-gu	2023010109	77	65

평가에 사용된 두 번째 데이터셋은 MIMIC(Multiparameter Intelligent Monitoring in Intensive Care) II 버전 2.6 데이터셋[10]으로 2001년부터 2008년까지 특정 병원의 중환자실 환자 방문 12,000건을 대상으로 수집된 종합적인 임상 데이터셋이며, 표 2에서 예시된 것처럼 결측치 비율이 80.7%로 매우 높으므로 추가 데이터 제거를 수행하지 않았다.

표 2. MIMIC II 버전 2.6 데이터셋
Table 2. MIMIC II version 2.6 dataset

Time	GCS	HR	NIDiasABP	NIMAP	NISysABP	Temp	pH	PaCO2	HCT
0:07	15	73	65	33	147	35.1			
0:37	15	77	58	91	157	35.6			
1:11							7.45	34	24.7
1:26	3	88				35.1			
1:31		88				34.8	7.44	33	
1:38							7.44	33	
1:46		88				34.5			
1:56		88				34.8			
2:11		88				35.1			

연구에서 사용된 머신러닝 모델과 결측치 대체 방법의 학습에 설정된 하이퍼파라미터는 다음과 같다: XGBoost(학습률: 0.1, 최대 깊이: 6, n_estimator: 100), LightGBM(학습률: 0.1, 최대 깊이: -1, num_leaves: 31), EBM(학습률: 0.01, 교차 검증 횟수: 5)

성능 평가에 사용된 지표는 다음과 같다.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y_i}\right)^2}$$

RMSE(Root Mean Squared Error)는 예측 오차의 크기를 측정하며, 값이 작을수록 모델의 예측이 정확함을 의미한다.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y_i}\right|$$

MAE(Mean Absolute Error) 또한 예측 오차의 크기를 측정하며, RMSE 대비 이상치에 덜 민감하다.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y_i}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y_i}\right)^2}$$

R^2는 예측값이 실제값을 얼마나 잘 설명하는지를 나타내는 지표로, 1에 가까울수록 모델이 데이터를 잘 설명함을 의미한다.

각 결측치 대체 방법(PSMF, MissForest, IBFI)을 각 머신러닝 모델(XGBoost, LightGBM, EBM)에 적용한 성능 평가 결과가 표 3과 같으며, 데이터셋 1은 미세먼지/초미세먼지 데이터셋, 데이터셋 2는 MIMIC II 데이터셋을 의미한다.

표 3. 성능 평가 결과
Table 3. Performance evaluation results

Dataset	Imputation method	Model	RMSE	MAE	R^2
Dataset 1	PSMF	XGBoost	16.32	12.45	0.72
		LightGBM	16.76	12.89	0.71
		EBM	16.40	12.60	0.72
	MissForest	XGBoost	14.90	11.10	0.83
		LightGBM	15.20	11.35	0.82
		EBM	15.05	11.22	0.83
	IBFI	XGBoost	15.00	11.20	0.83
		LightGBM	15.30	11.45	0.82
		EBM	15.10	11.25	0.83
Dataset 2	PSMF	XGBoost	13.52	9.83	0.64
		LightGBM	14.10	10.76	0.67
		EBM	14.68	10.90	0.63
	MissForest	XGBoost	12.67	9.73	0.66
		LightGBM	13.25	9.93	0.67
		EBM	11.97	8.91	0.64
	IBFI	XGBoost	12.46	8.65	0.71
		LightGBM	13.41	9.54	0.71
		EBM	12.86	9.35	0.66

표 3에 의하면, 결측치 대체 방법 중 PSMF는 Miss Forest, IBFI 대비 낮은 성능을 보였으며, 이는 PSMF가 선형적인 데이터 패턴에 적합하며 비선형적인 패턴을 충분히 반영하지 못하기 때문으로 판단된다. 또한 PSMF는 XGBoost 모델과의 조합이 가장 높은 성능을 나타냈다. MissForest를 사용한 결측치 대체는 모든 머신러닝 모델과의 조합에서 안정적이고 높은 성능을 나타냈으며, 이는 MissForest가 랜덤 포레스트 알고리즘을 기반으로 하여 비선형적인 데이터 패턴을 효과적으로 반영할 수 있기 때문으로 판단된다. MissForest는 특히 XGBoost 모델과의 조합에서 가장 높은 성능을 보였다. IBFI는 MissForest와 유사한 높은 성능을 보였으며, 특히 XGBoost, EBM 모델과의 조합에서 높은 예측 성능을 나타냈다.

머신러닝 모델 중 XGBoost는 전반적으로 가장 높은 성능을 보인 모델로, MissForest, IBFI 결측치 대체 방법과의 조합에서 우수한 예측 결과를 나타냈다. LightGBM은 XGBoost와 비교했을 때 성능이 다소 낮았지만, 여전히 우수한 예측 성능을 보였으며, MissForest, IBFI 결측치 대체 방법과의 조합에서 안정적인 성능을 나타냈다. EBM은 XGBoost보다는 다소 낮지만 LightGBM보다는 높은 성능을 나타냈으며, XGBoost와 유사하게 MissForest, IBFI와의 조합에서 높은 성능을 보였다 (특히 MIMIC II 데이터셋에서 우수한 결과를 보임).

4. 결론

본 연구에서는 결측치 비율이 높은 시계열 데이터를 분석하고 예측하기 위해 다양한 결측치 대체 방법과 머신러닝 모델을 적용하였다. 성능 평가 결과, PSMF, MissForest, IBFI를 사용한 결측치 대체 방법 중 MissForest, IBFI가 가장 높은 성능을 보였으며, XGBoost, EBM 모델이 LightGBM 모델보다 높은 성능을 보였다. 또한, 결측치 비율이 높은 시계열 데이터를 분석할 때는 MissForest, IBFI와 같은 비선형적 방법을 사용하여 결측치를 대체한 후, XGBoost, EBM 머신러닝 모델을 적용하는 것이 최적임을 확인하였다.

본 연구에는 다음과 같은 한계가 존재한다. 첫째, 본 연구에서는 두 가지 데이터셋(서울시 미세먼지 데이터셋, MIMIC II 데이터셋)만을 사용하여 평가를 진행하였으므로 향후 연구에서는 다양한 도메인의 데이터셋을 사용하여 본 연구의 결과를 검증할 필요가 존재한다. 둘째, 본 연구에서는 RMSE, MAE, R^2와 같은 전통적인 성능 지표를 사용하였으며, 향후 연구에서는 더 다양한 성능 지표를 도입하여 모델의 성능을 다각도로 평가할 필요가 존재한다. 예를 들어, 결측치 대체의 정확성 외에도 데이터의 시계열 패턴 보존 여부, 예측값의 안정성 등을 평가할 수 있는 지표를 고려할 수 있다. 또한 향후 연구에서는 더 다양한 결측치 대체 방법과 머신러닝 모델을 적용한 성능 비교 분석이 필요하다.

REFERENCES

[1] Little, Roderick. J., Donald B. Rubin. "Statistical analysis with missing data." 793). 793, John Wiley & Sons, 2019.

[2] Van Buuren, Stef, Karin Groothuis-Oudshoorn. "mice: Multivariate imputation by chained equations in R." Journal of Statistical Software, 45, 1-67, 2011.

[3] Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. "LightGBM: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems, 3146 -3154, 2017.

[4] Chen, Tiangi, Carlos Guestrin. "XGBoost: A scalable tree boosting system." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794, 2016.

[5] Lou, Yin, Rich Caruana, Johannes Gehrke. "Intelligible models for classification and regression." Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 150-158, 2012.

[6] Akyildiz, Omer Deniz, Gerrit van den Burg, Theodoros Damoulas, Mark Steel. "Probabilistic sequential matrix factorization." arXiv preprin

t, arXiv:1910.03906, 2019.
[7] Stekhoven, Daniel J., Peter Buhlmann. "MissForest: non-parametric missing value imputation for mixed-type data." Bioinformatics, 28-1, 112-118, 2012.
[8] Mir, Adil Aslam, Kimberlee Jane Kearfott, Fatih Vehbi Celebi, Muhammad Rafique. "Imputation by feature importance (IBFI): A methodology to envelop machine learning method for imputing missing patterns in time series data." PloS one, 17-1, e0262131, 2022.
[9] Air Korea, https://www.airkorea.or.kr/web/last_amb_hour_data?pMENU_NO=123
[10] PhysioNet, https://archive.physionet.org/mimic2

저자약력

고 방 원 (Bangwon Ko) **[정회원]**

- 서울대학교 수학과 (학사)
- 서울대학교 통계학과 (석사)
- 아이오와 주립대 통계학과 (박사)
- (현)숭실대학교 정보통계·보험수리학과 교수

〈관심분야〉 보험수리, 금융공학, 머신러닝

한 용 희 (Yong Hee Han) **[정회원]**

- 한양대학교 산업공학과 (학사)
- 조지아공대 산업공학과 (석사)
- 조지아공대 산업공학과 (박사)
- 삼성전자 메모리사업부 (책임연구원)
- (현)숭실대학교 벤처중소기업학과 교수

〈관심분야〉 머신러닝, 딥러닝, 스마트팩토리