Journal of Korea Multimedia Society Vol. 25, No. 2, February 2022(pp. 257-268)
https://doi.org/10.9717/kmms.2022.25.2.257

머신러닝기반의 데이터 결측 구간의 자동 보정 및 분석 예측 모델에 대한 연구

정세훈[†], 이한성[††], 김준영[†††], 심춘보[††††]

A Novel on Auto Imputation and Analysis Prediction Model of Data Missing Scope based on Machine Learning

Se-Hoon, Jung[†], Han-Sung, Lee[††], Jun-Yeong, Kim[†††], Chun-Bo, Sim[††††]

ABSTRACT

When there is a missing value in the raw data, if ignore the missing values and proceed with the analysis, the accuracy decrease due to the decrease in the number of sample. The method of imputation and analyzing patterns and significant values can compensate for the problem of lower analysis quality and analysis accuracy as a result of bias rather than simply removing missing values. In this study, we proposed to study irregular data patterns and missing processing methods of data using machine learning techniques for the study of correction of missing values. we would like to propose a plan to replace the missing with data from a similar past point in time by finding the situation at the time when the missing data occurred. Unlike previous studies, data correction techniques present new algorithms using DNN and KNN-MLE techniques. As a result of the performance evaluation, the ANAE measurement value compared to the existing missing section correction algorithm confirmed a performance improvement of about 0.041 to 0.321.

Key words: Data Imputation, Deep Neural Network, KNN, Missing Value, MLE

1. 서 론

구글 딥마인드의 알파고 도전 이후, 인공지능에 대한 관심이 폭발적으로 증가하였고 인공지능 관련 연구와 예산이 매년 늘어남에 따라 인공지능 기술은 눈부신 발전을 이루고 있다. 특히 컴퓨터 비전, 자연어 처리 분야에서 엄청난 성과를 거두고 일상생활에 인공지능이 적용되면서 인공지능은 이제 선택이 아닌 필수가 되었다. 빅데이터 시대가 진행되는 있는 현재에 데이터의 과잉 생산은 이슈가 되지 않고 있는 실정이다. 어느 분야에서든 새롭게 생성되는 데이터의 규모는 현재 데이터베이스 관리 시스템을 초과하

※ Corresponding Author : Chun-Bo, Sim, Address: (57922) 255, Jungang-ro, Suncheon-si, Jeollanam-do, Republic of Korea, TEL : +82-61-750-3834, FAX : +82-61-750-3830, E-mail : cbsim@scnu.ac.kr
Receipt date : Jan. 17, 2022, Approval date : Jan. 24, 2022
[†] School of Creative Convergence, Andong National University(E-mail : jungsh@anu.ac.kr)
[††] School of Creative Convergence, Andong National University(E-mail : mohan@anu.ac.kr)
[†††] School of ICT Convergence Engineering, Sunchon National University(E-mail : kjoone3k@naver.com)
[††††] School of ICT Convergence Engineering, Sunchon National University

※ This paper was supported by(in part) Sunchon National University Research Fund in 2020.(Grant number: 2020-0208) and this research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education (2020R1I1A3054843) and this paper is an extended version of a conference paper published in S. H. Jung, J. C. Kim, and C.B. Sim, A Novel on Outlier Detection Algorithm using MLE Method based on Machine Learning, Proceedings of the Fall Conference of the Korea Multimedia Society, Vol. 24, No. 2, Poster Session B-9, 2021

는 이슈가 지속적으로 발생되고 있다. 이러한 데이터 생성은 이전과는 다르게 데이터 분석과 예측에 많은 기본 베이스가 되고 있다. AlphaGo와 Naver Papago와 같은 인공지능 모델은 학습 데이터가 많은 수록 데이터의 예측과 분석 정확도는 높아지는 상관관계를 유지하고 있다. 그러나 이러한 데이터 홍수시대에도 특정 분야를 분석하기에는 데이터가 부족하거나 수집된 데이터의 일부 구간이 오류로 인한 손실 데이터가 발생하는 데이터 결핍 현상이 꾸준하게 증가하고 있다. 데이터 예측을 위한 로우 데이터의 분석에 있어 로우 데이터 부족 현상은 분석 모델에 편향된 예측 모형이나 예측 품질 및 예측 정확도의 신뢰성에 대한 문제점으로 지적되고 있다. 그리고 인공지능 기술을 이끌고 있는 딥러닝이 성능에 비례하여 엄청난 양의 빅데이터로 인해 발생한 문제로써 빅데이터를 구축하기에 너무 많은 비용이 소모되거나, 구축할 수 없는 분야로 확장되면서 데이터의 중요도 문제는 점차 심화되고 있다. 이와 같이 향후 인공지능 기술이 필요한 대다수의 환경은 빅데이터를 축적하기 어려운 환경이기 때문에 빅데이터 의존성이 낮은 기술과 새로운 인공지능 데이터 수집 및 보정 기술이 필요한 실정이다. 데이터 보정 기술은 결측치 대체 기술로 지속적으로 연구되고 있다. 수집된 로우 데이터의 결측 구간을 보정하는 방법은 2가지로 정의할 수 있다. 첫번째 방법은 로우 데이터의 결측 구간에 대한 결측 구간 완전 제거 기법이 있다. 결측 구간 완전 제거 방법은 입력되는 로우 데이터의 결측 구간에서 유효한 값을 가진 로우 데이터 수집 구간만 대상이 되는 방법이다. 두번째 방법은 수집된 로우 데이터의 결측 구간에 대한 단순 변경과 다중 변경으로 분류할 수 있다. 단순 보정 기법으로는 평균 보정(Mean/Median Imputation, 이하 MMI)과 특징을 분석한 최대 빈도수 보정(Most Frequent Imputation, 이하 MFI), 로우 데이터의 특징 분석 후 회귀계수를 이용한 지정값 보정(Constant Imputation, 이하 CI), 불특정 보정(Random Imputation, 이하 RI), KNN 보정(KNN Imputation)등이 연구되고 있다. 그리고 다중 보정 기법으로는 체인 부등식기반의 다중 보정(Multi Imputation by Chained Equation, 이하 MI)이 연구되고 있다[1-5]. 기존 데이터 결측 구간의 보정 연구는 수집된 로우 데이터에 일부 특징을 통해 결측 구간을 보정하는 연구가 주로 연구되었다. 특히 평균 보정 기법은 결측 구간의 특징에 대한 상관관계가 분석되지 않았으며, 범주형 특징에는 부적절하다. 최대 빈도수 보정은 데이터의 편향을 만들 수 있으며, 결측 구간간 상관관계가 고려되지 않는 문제점이 있다[6-8]. k-최근접 이웃(k-Nearst Neighbors, 이하 KNN) 보정은 메모리가 많이 필요하며, 이상점에 대한 민감도가 상대적으로 높다. 다중 보정 기법은 메모리 소모 문제가 발생한다. 이러한 문제점들을 보완하기 위한 딥러닝(Deep Learning)기법의 Datawig 알고리즘이 연구되었지만 대규모 데이터셋에 대한 보정 비용과 보정해야 할 특정 구간에 대한 비교 특징 구간을 직접 설정해야 하는 문제점이 발생하고 있다. 기존 결측 구간 보정 방법은 메모리 사용량, 데이터 민감도, 특징 구간에 대한 직접적인 연결의 문제점과 데이터 보정 기법에서 많이 활용된 KNN 알고리즘의 결측 구간 보정 시 결측 구간을 포함하지 않고 특징이 서로 상이한 완전 분석 패턴만을 활용하기 때문에 결측 구간을 포함하는 과거 관측값의 특징을 적용하지 못하는 문제점이 있었다. 본 연구에서는 로우 데이터에 포함된 불규칙한 결측 구간의 결측값들을 분석하고 보정된 결과를 활용하고자 한다.

이에 본 연구에서는 결측 구간 보정 연구를 위해 수집되는 로우 데이터의 특징 패턴 분석 및 결측 보정 방안을 심층 신경망(Deep Neural Network, 이하 DNN) 및 보정된 KNN-MLE기반의 머신러닝 알고리즘 연구를 제안한다. 이를 위해 수집된 로우 데이터의 결측 구간에 대한 특징 분석과 유사 구간에 대한 자동 보정 방법을 제안하고자 한다. 이를 위해 심층 신경망과 최대 우도 추정법 기법과 KNN 알고리즘을 결합한 방법을 제안한다. 본 논문에서는 DNN, 보정화된 KNN-MLE 방법을 통해 수집되는 로우 데이터에서 발생하는 결측 구간의 특징을 분석하고 결측 구간에 보정이 가능하다.

본 논문의 구성은 다음과 같다. 2장에서는 로우 데이터 보정 기법에 해당하는 KNN 알고리즘과 결측 보정 기존 연구 내용을 제시하며, 3장에서는 제안하는 DNN 및 보정화된 KNN-MLE기반의 로우 데이터 결측 구간에 포함된 결측 보정 알고리즘 모델과 데이터 예측 모델을 제시한다. 4장에서는 제안한 모델에 대한 연구평가를 제시한다. 5장에서 결론과 향후 연구에 대한 내용을 제시한다.

2. 관련 연구

이번 장에서는 로우 데이터의 특징 분석 및 결측 구간의 보정 기법에 대한 기본 개념을 소개하고 결측 데이터에 대한 보정 기법과 관련된 기존 연구를 소개하며, 결측 구간 데이터가 포함된 분석 및 예측 모델의 기존 문제점을 파악하고 본 연구에서 목표로 하는 결측치 자동 보정과 예측 모델에 대한 연구를 확인한다.

2.1 K-최근접 이웃 알고리즘

S. Shiliang 등의 연구[9]에서는 제시한 KNN 알고리즘은 가치 함수(Value Function)를 설정하고 분류하는 머신러닝 기법과 달리, 분류하려는 데이터를 학습대상 데이터로 구분하고 특징이 유사한 k개를 선택하여 데이터의 분류 기준을 결정하는 기법으로써, 학습 및 테스트 대상 데이터의 특징을 모두 백터화한다. 그리고 고유값들에 대한 유클리디안(Euclidean Distance) 측정값을 평가하여 인스턴스간 유사도를 결정하는 방식이다. KNN 알고리즘은 학습 데이터의 볼륨과 신뢰도, 테이터별 특징 선택 기법 및 특징 k에 따라 데이터 분류기의 효율성이 제시된다. 미리 부류가 결정되어 있는 데이터 분류 기법으로는 서포트 벡터 머신(Support Vector Machine, 이하 SVM), 나이브베이즈(NaiveBayes), KNN 등이 분류 머신러닝 알고리즘으로 분류된다. 그러나 학습 이전에 분류가 결정되지 않거나 여러 분류가 서로 중복으로 분류되어 있을 경우에는 KNN 알고리즘을 활용하는 것이 더 효율적이다. 그러나 기존 연구된 최근접 이웃 알고리즘은 결측구간에 포함된 데이터와 최근접 위치에 배치된 결측치가 존재하지 않는 데이터들을 활용하여 결측값을 보정하는 기법이다. 결측 구간에 포함된 결측값의 보정시 결측값을 갖고 있지 않는 완전한 데이터만을 활용하기 때문에 결측값을 포함하는 데이터 구간의 보정값들을 알고리즘에 활용하지 못하는 문제점이 포함하고 있다.

2.2 최대 우도 추정 기법

통계학자 피셔(Fisher)는 20세기에 제안한 데이터의 파라미터 추정 기법으로 최대 우도 추정 기법(Maximum Likehood Estimation, 이하 MLE)을 제안하였다. 우도함수(Likehood Function, 이하 LF)내의 범위 중 함수를 최대화하는 매개변수 Θ를 측정하는 방법이다. Fig. 1의 위쪽 부분은 MLE의 측정에 포함되는 LF와 Log-LF이다. 그래프의 맨 위쪽 부분은 1차원 데이터이며, MLE 측정을 위한 후보 확률분포로 분석될 수 있다. Fig. 1의 중앙 부분은 결합밀도함수로 분석된 LF이며, 수식은 $p(D|\theta)$이다. D는 로우 데이터들의 집합을 의미한다. 그리고 로우 데이터가 많으면 많을수록 데이터 분포 간격이 좁아지게 된다. 우도를 최대로 하는 매개변수 값이 Θ로 표시된다. Fig. 1의 마지막 부분은 LF에 대한 Log-LF를

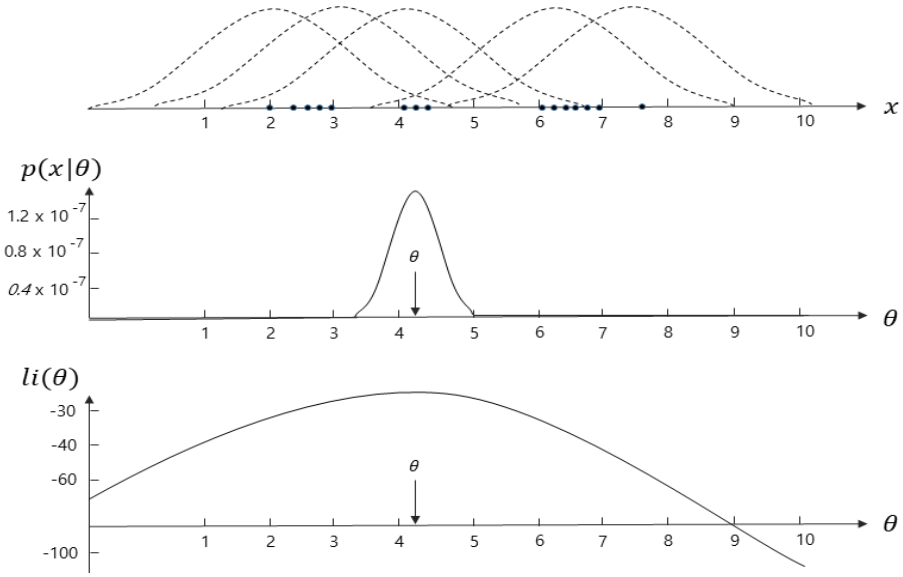


Fig. 1. The likelihood function and log-likelihood function for MLE.

도식화한 이미지이다. 그리고 우도를 최대로 하는 매개변수 Θ의 절대적 위치는 우도와 Log-LF가 동일함을 확인할 수 있다[10].

2.3 데이터 결측 구간 대치법

로우 데이터들은 수집되는 형태에 따라 다양한 타입을 포함하고 있다. 실험을 통해 얻어지는 로우 데이터에는 외부 또는 내부적인 방법으로 인하여 결측값이 포함된 결측 구간이 포함되며, 이러한 결측값은 활용되지 않고 전처리 과정에서 손실되는 경우가 발생하고 있다. 이러한 손실은 데이터 분석 및 예측 모델에 있어 편향(Bias)를 발생시켜 모델의 성능을 하락시키는 주요 원인으로 분석되고 있다. 결측치가 포함된 결측 구간을 활용한 데이터 분석 모델에 있어 이를 해결하기 위한 통계적 방법들은 지속적으로 연구가 진행되고 있다. 로우 데이터에 포함된 결측 구간의 결측치에 대한 보정 기법을 활용하기 위해서는 결측 보정을 처리하는 방법을 이해하는 것이 필수적이다. 결측 분포 방법은 결측값을 포함하는 결측 구간과 데이터에 포함된 변수들의 상관 및 연관 분석 관계를 해석하는 것이 필요하다. 결측 분포 방법은 변수들에 대한 결측 구간 존재 여부에 따라 결측 완전(Missing value Completely At Random, 이하 MCAR), 랜덤 결측(Missing value At Random, 이하 MAR), 연관 결측(value Non Ignorable, 이하 NI)로 구분된다[11]. 결측 완전(MCAR)은 결측구간이 포함된 변수와 상관없이 어떠한 변수들과도 상관 및 연관 관계가 없다는 것으로 결측구간에 포함된 결측값이 랜덤하게 배치되는 것을 뜻한다. 랜덤 결측(MAR)은 결측 구간에 포함된 결측값이 결측값을 포함하지 않는 데이터들과 연관 관계가 존재하는 경우를 뜻한다. 연관 결측(NI)은 결측 구간에 포함된 결측값이 오직 결측값을 포함하는 데이터들과 연관 관계가 존재하는 경우를 뜻한다.결측 분포 방법을 통해 발생되는 결측값을 보정하기 위한 결측 구간 보정 방법은 2가지 형태인 단순 보정 방법과 다중 보정 방법으로 분리된다. 단순 보정 방법은 완전 제거 기법, 평균 보정 기법, 지정값 보정 기법, 불특정 보정 기법, KNN 보정 기법, Datawig 보정 기법, 단일 확률 보정 기법, 다중 보정 방법은 체인 부등식 보정 기반의 다중 보정으로 구성된다. 완전 제거 보정 기법은 결측구간에 포함된 결측값들을 모두 삭제하고 결측구간이 포함되지 않는 데이터로 분석을 진행하는 표준적 통계 보정 기법을 적용한다. 완전 제거 보정 기법은 결측 구간 분석이 쉽다는 장점을 가지고 있으며, 결측구간이 포함된 로우 데이터를 모두 삭제함으로써 편향된 보정 결과 또는 통계적 추론의 적정성 측면에서 여러 문제점이 제시되고 있다. 평균 보정 기법은 결측 구간이 포함되지 않는 데이터의 확률적 통계의 평균값으로 보정하여 불완전한 데이터를 완전한 데이터로 보정하는 기법이다. 평균 보정 기법은 비조건부 확률적 평균 보정 기법과 조건부 확률적 평균 보정 기법이 있다. 조건부 확률적 평균 보정 기법의 확장으로써 Buck's 보정 기법이 연구되고 있다. 단일 확률적 보정 기법은 평균 보정 기법에서 결측 구간의 보정값의 표준 오차에 대한 과소적합 문제점를 보완하기 위해 연구된 기법으로 보정된 결측 구간의 추정 통계량을 결측값으로 보정할 때 단일 확률값을 보정값으로 처리하는 기법이다. 단일 확률적 보정 기법은 결측 구간의 예측 보정값의 표준 오차가 과소적합 되는 문제점을 개선하였지만 비교적 데이터 크기가 크지 않는 데이터를 제외한 데이터에서는 결측 구간의 예측 보정값의 표준 오차 측정이 어렵다는 문제점이 부각되고 있다. 체인 부등식기반의 다중 보정 기법은 이러한 문제점을 보완하기 위하여 단순 보정 기법처럼 한번만 수행하지 않고 m번의 보정을 통해 m개의 가상적 데이터로 보정하는 기법이다. 체인 부등식기반의 다중 보정 기법은 3단계로 구분되며, 보정-분석-통합 단계로 활용된다[11].

M. Ingunn 등의 연구[12]에서는 결측값이 포함된 실제 데이터를 기준으로 4가지 결측값 보정 기법들을 비교하였다. 비교 대상 기법들은 완전 제거 보정 기법, 확률적 평균 보정 기법, 유사 응답 유형 보정 기법(similar response pattern imputation), MLE기법을 제안하였고, 성능평가로는 데이터 규모가 클 경우 MLE기법이 결측 구간 보정에 적합하다는 것을 성능평가로 제시하였다. S. Qinbao 등의 연구[13]에서는 소규모 데이터의 결측 구간 결측값 보정을 위해 K-NN에 활용한 클래스 확률적 평균 보정 기법(class mean imputation)이라는 새로운 결측 구간 보정 기법을 제안하였다. 제안된 기법은 수치 및 범주형 결측 구간에 포함된 결측값에 적용할 수 있으나 로우 데이터가 소규모(평균 로우 데이터 100개 이하)인 경우에만 알고리즘 활용과 성능이 우수하다는 제약 사

항이 문제점으로 지적되고 있다.

3. 데이터 결측 구간 대치 및 분석 예측 모델설계

3.1 전체 시스템 구성도

Fig. 2는 본 논문에서 제안하는 데이터 결측 대치 및 분석 예측 모델의 구성도이다. 모델은 크게 2가지 형태로 데이터 결측 구간 대치 및 분석 모델을 제안한다. 제안하는 모델은 Analysis and Prediction Model과 Missing Value Imputation로 구성된다. Analysis and Prediction Model Part는 보정된 로우 데이터를 기준으로 전처리 과정을 수행한다. 전처리 모듈에는 분석 모델에 적용되는 데이터의 이상점에 대한 데이터 제거와 정규화 과정을 포함하고 있다. 데이터 이상점과 특이점을 분석할 수 있도록 강화된 K-means 알고리즘과 주성분 분석(Principal Component Analysis, 이하 PCA) 알고리즘을 포함한 보정 데이터의 분석 및 예측 기능을 수행한다. Missing Value Imputation Part는 결측 구간의 추출을 위하여 DNN 모델과 KNN 알고리즘과 최대 우도 추정 알고리즘을 결합한 KNN-MLE 보정 알고리즘을 통해 결측 구간 데이터를 대치하는 기능을 수행한다.

3.2 데이터 결측 구간 대치 설계

로우 데이터상에 결측 구간이 포함된 결측치가 존재할 경우 보정하지 않고 모델 분석을 진행되면 로우 데이터의 표본 수 감소로 신뢰성 및 검정력이 현저하게 낮아지거나 편향된 결과를 도출할 수 있다. 결측 구간이 포함된 로우 데이터의 결측값은 결측 보정 기법에 해당하는 단순 제거 기법 보다는 결측값이 포함되거나 포함되지 않는 데이터 패턴과 로우 데이터의 유의미한 특징을 포함한 데이터 값으로 보정하여 데이터를 분석하는 방법이 데이터 분석 정확도의 편향된 결과나 품질 및 정확도가 상대적으로 감소되는 단점을 보완할 수 있다. 결측값 보정에 관한 기존 연구에는 KNN 알고리즘이 주로 활용하였다, 해당 알고리즘은 전통적인 확률적 기법과 달리 로우 데이터 분포에 대한 가설이 불필요하다는 장점이 있다. 그러나 일정한 결측 구간의 패턴이나 동일한 결측 구간의 패턴이 존재하지 않을 때, 결측치 보정에 대한 정확성이 감소되는 단점이 있다. 로우 데이터는 규모가 클 수도 있고 규모가 작을 수도 있다. 또한 패턴이 일정하지 않는 경우도 존재한다. 로우 데이터의 불규칙한 패턴 특성을 반영한 결측 보정 연구가 필요한 실정이다. 본 연구에서는 기존에 제기된 문제점과 논의사항을 보완하고자 KNN 알고리즘과 MLE 알고리즘을 혼합한 결측구간 보정 모델을 제안한다. MLE는 전통적인 통계기반의 패턴 추론으로 최적의 매개변수를 추정하는 기법으로 로우 데이터 크기가 커질수록 변수의 가능도가 추정값을 정확하게 찾을 수 있다는 장점을 가지고 있다. KNN 알고리즘은 기존에 연구된 결측 구간 보정 알고리즘에 많이 활용되었으며, 알고리즘의 구현은 비교적 간단하다. KNN을 적용한 결측 보정 알고리즘은 관련 연구에서 제시한 결측 구간 구성 방법 3가지에 활용될 수 있으며, 결측 구간의 결측 보정의 결과도 높은 수준이다. Fig. 3과 같이 수집된 로우 데이터는 결측 구간 구성 3가지 패턴을 기준으로 5개의 결측 구간 패턴을 적용하

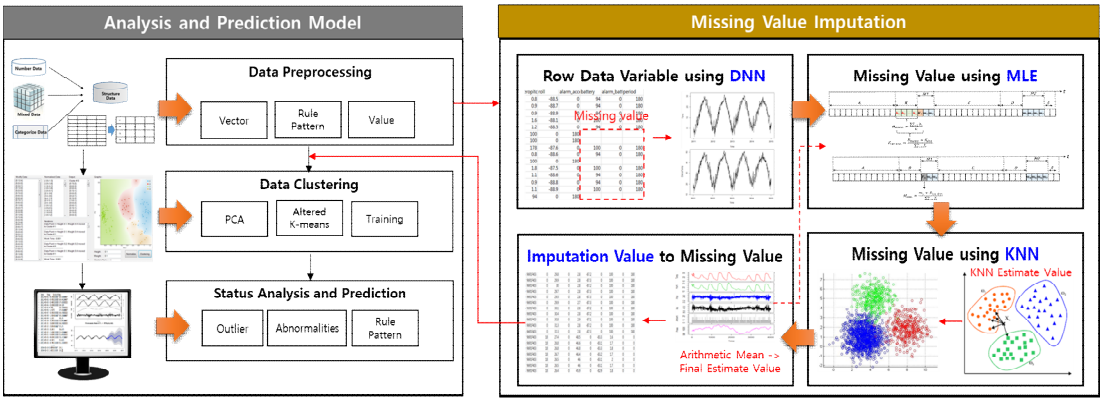


Fig. 2. Structure Diagram of Proposed Model.

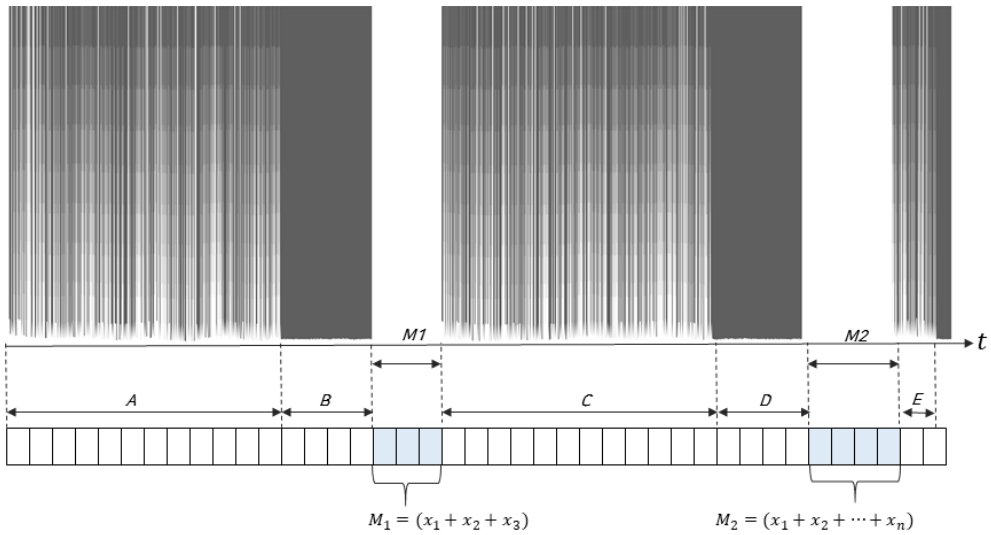


Fig. 3. Example of Data in Missing Value.

였다. A, B, C 구간은 이상치와 일부 결측 구간을 적용하였으며, B와 D 구간은 안정적인 로우 데이터가 존재하는 구간이다. 파란색으로 표시된 M_1 구간과 M_2 구간에서는 구간 전체에서 데이터 결측이 발생하였다. 기존 연구는 결측이 발생한 일부 구간의 보정 연구로 결측 패턴이 발생하지 않는 데이터 구간의 패턴 정보를 적용하는 기법이다. 기존 연구 기법은 2장에서 언급한 부분처럼 비조건부 확률적 평균 보정 기법, 조건부 확률적 평균 보정 기법, 단일 확률적 보정 기법, 체인 부등식기반의 다중 보정 기법이 존재한다. 그러나 KNN 알고리즘을 적용할 경우 Fig. 4와 같이 결측값 대치시 결측값을 포함하는 패턴이나 유효한 데이터가 존재하는 A, C, E 구간에 대해서는 결측치 발생 구간이 아님에도 불구하고 결측 대치 측정 구간에 유효한 관측정보를 제공할 수 없다는 문제점이 존재한다. 결국 B 구간과 D 구간만 활용하여 M_1과 M_2의 결측 구간에 대한 대치법을 처리하게 된다.

본 논문에서는 기존 연구 방식에서 활용되는 B, D 구간 뿐만 아니라 결측 구간이 포함된 A, B, C 구간에 대한 결측 패턴을 같이 적용하여 M_1 구간과 M_2구간의 결측치를 보정하는 방법으로 Fig. 4와 같

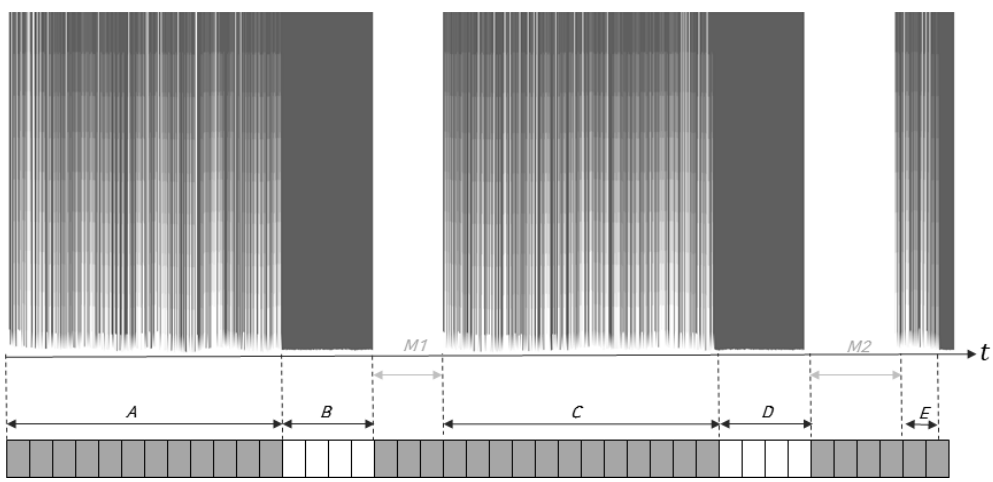


Fig. 4. Example of Data in Missing Value (K–Nearest Neighbor Algorithm).

이 도식화할 수 있다. 이를 위해 MLE을 적용하여 결측 구간의 보정 문제를 보완한다. MLE는 전체 로우 데이터를 활용하여 최적의 매개변수를 찾아가는 매커니즘에 적용한다. 로우 데이터의 결측치가 발생한 구간을 제외하고 나머지 전체 구간의 패턴을 적용하여, 결측 구간이 포함된 결측치에 대한 임시 추정값을 계산하고 결측값 보정를 수행한다.

Table 1과 같이 결측 구간 대치법은 총 6 Step으로 구성된다. Step 1은 데이터에 발생한 모든 결측값들을 최대 우도 추정법 결과값으로 임시 변경한다. Step 2는 최대 우도 추정법 결과값으로 변경한 인스턴스들 중 하나의 인스턴스만 최대 우도 추정법 결과값을 결측값으로 변경(나머지는 모두 최대 우도 추정법 결과값으로 대치) 한다. Step 3은 결측 구간에 포함된 결측값을 가진 하나의 데이터에 대해 KNN을 적용하여 K-최근접 이웃 추정값 계산한다. Step 4는 이전 데이터 구간의 평균 수치와 K-최근접 이웃 추정값을 결합한 산술평균 수치를 1차 추정값으로 산출한다. Step 5는 최대 우도 추정법 결과값과 1차 추정값에 대한 산술평균을 최종 추정값으로 산출하며 모든 결측값의 추정값을 계산하기까지 Step 1부터 Step 5단계를 반복적으로 수행한다. Step 6은 모든 결측값들에 대한 최종 추정값을 저장하여 완전한 결측치 데이터 벡터 생성하게 된다.

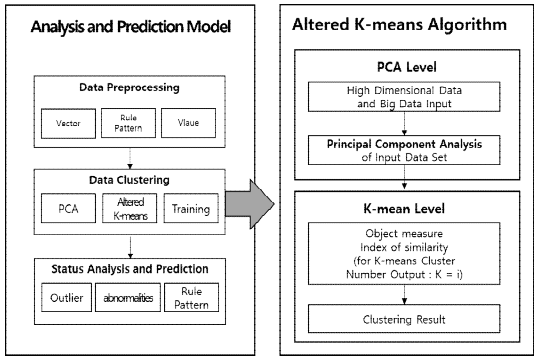
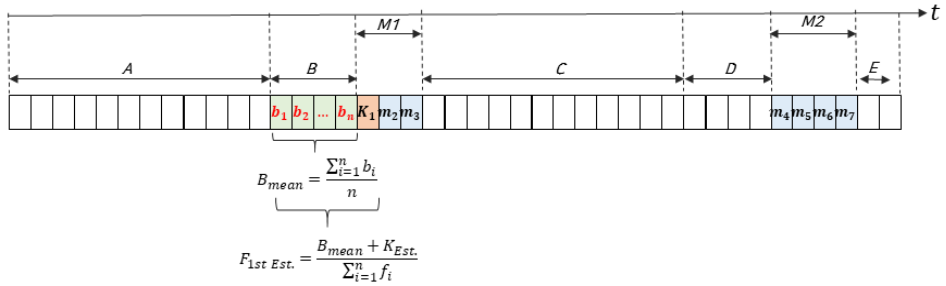
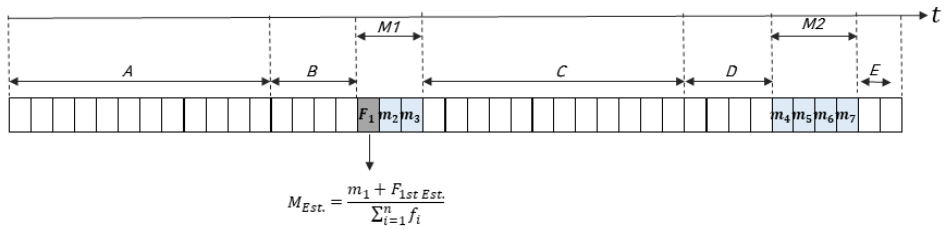


Fig. 5. Flow Chart of Analysis and Prediction Model.

3.3 데이터 분석 및 예측 모델 설계

Fig. 5는 본 논문에서 제안하는 결측치가 보정된 로우 데이터 예측 및 분석 흐름도이다. 제안하는 분석 모델은 결측치가 보정된 로우 데이터를 클러스터링할 수 있는 강화된 K-means알고리즘을 적용하여 데이터를 분류 및 분석한다. 강화화된 K-means의 자동화된 K값 선택은 입력되는 시계열 로우 데이터

Table 1. Construct of KNN-MLE Algorithm.

Step	Description
Step 4	
Step 5	

x_n : Missing Value Instance, m_n : MLE Estimated Value, K_n : KNN Estimated Value, f_n : 1st Missing Estimated Value, F_n : Final Missing Imputation Value

에서 PCA의 데이터 벡터 공간 차원 감소를 통해 값을 결정한다. 보정된 데이터를 기반으로 강화된 K-means 알고리즘을 제안하기 위해 2가지 고려 사항을 반영해야 한다. 첫째, 결측치가 보정된 로우 데이터들의 분류를 위한 연구기법은 데이터의 벡터 공간 및 산술적 거리 측정 방법이 필요하다. 특히 군집화를 위한 K값 선택의 로우 데이터 공간분할 방법과 클러스터간 산술적 공간 거리 측정 기법이 필요하다. 둘째, 로우 데이터에 대한 최적화된 중심값의 초기 공간 또는 위치 확보가 필요하다. 다차원 로우 데이터의 최적화된 클러스터링 특징으로 차원 축소 기법을 포함해야 한다. 차원 축소를 통해 로우 데이터에 대한 클러스터간의 상관성을 최소화하면 클러스터의 신뢰성 있는 초기 공간 벡터 확보와 정확도를 향상시킬 수 있다는 장점이 존재한다. 마지막으로, 주어진 로우 데이터에 포함된 이상점을 최소화해야 한다. 군집화를 위한 초기 공간 벡터 중심값 선택 에서 공간 데이터 이상점을 중심값으로 선택하게 되면 클러스터 결과의 정확성이 상대적으로 낮아진다는 문제점이 있다. 벡터 공간분할과 산술적 거리 측정 기법의 개선된 K-means 알고리즘을 통해 로우 데이터에 대한 이상점의 오류율을 최소화해야 한다.

4. 제안하는 모델의 성능평가

4.1 데이터 집합 및 성능평가 환경

본 연구에서는 로우 데이터 예측 분석 및 결측 구간이 포함된 결측값 보정의 알고리즘의 성능을 확인하기 위하여 전력 관련 로우 데이터를 활용한다. 이를 확인하기 위한 성능평가 반영 데이터 수집 기간은 2007년 1월 1일부터 2015년 12월 31일까지 108개월의 데이터이다. 측정 주기는 1분 단위이며 수집된 데이터 타입는 온도, 습도, 가속도, 빛 감지, 압력, 대기환경, 소리감지, 초음파, 지자기, 전류이다. 전체 데이터 셋은 크기는 1.5GB이며, 로우 데이터는 33,864,675개의 데이터가 수집되었다. 수집된 데이터 중 본 논문의 성능평가에 반영될 타입은 데이터는 온도, 습도, 가속도, 압력의 로우 데이터이다. 성능평가에 반영될 데이터는 로우 데이터는 13,312,920개의 데이터를 반영하였다. 데이터의 크기는 586MB이며, 이 중 결측치로 구분된 로우 데이터는 약 10%에 해당되는 1,431,208개의 데이터가 분류되었다.

본 연구에서 진행되는 성능평가 환경 구성은 다음과 같다. OS는 Windows 10이며, CPU는 Intel Core i9-11900K이며, RAM은 64GB이다. GPU는 Geforce RTX 3090 24GB이며, 개발 언어는 Python 3.6, IDE는 Anaconda의 Spyder 4.2.5를 활용하였다. Library는 Tensorflow 1.14.0, Keras 2.3.1을 활용하였으며, DNN은 GRU 모델을 활용하였으며, 학습모델 구성은 AlexNet을 적용하였다.

4.2 결측 보정 알고리즘 평가 기준

제안된 결측 구간 자동 보정 및 예측 분석 모델의 결측 구간 자동 보정 알고리즘은 기존 연구 기법과 비교하여 DNN 및 KNN-MLE 알고리즘을 결합하였다. 이를 통해 빅데이터의 일부 데이터를 기준으로 결측 구간 대치를 결정하였던 기존 결측치 보정 알고리즘과 비교하여 전체 데이터를 활용할 수 있다는 특징을 포함하고 있다. 본 연구에서는 결측치 자동 보정 방법의 성능평가에 활용될 평가 기준을 다음과 같이 정의한다. 알고리즘 평가 기준는 평균절대오차(Average Absolute Error, 이하 AAE)와 평균표준절대오차(Average Normalized Absolute Error, 이하 ANAE)로 분류할 수 있다. AAE는 데이터에 포함된 변수 중 결측구간에 대한 결측값 보정의 정확도를 평가하기 위해서 적용된다. 입력되는 데이터를 DNN과 KNN-MLE를 기반으로 특정 결측치가 포함된 결측 구간에 AAE 평가값을 기존 연구와 비교 수치를 평가한다. 그러나 데이터 결측 구간에 대한 AAE 성능평가는 비교할 수 없다는 문제점이 존재한다. 본 연구에서는 ANAE로 제안하는 모델의 성능평가를 진행한다. ANAE평가 척도는 AAE의 평가 척도의 문제점인 데이터 결측 구간마다 발생하는 값의 범위가 다를 경우 성능비교를 진행할 수 없는 문제점을 보완한 평가 기준이다. ANAE는 데이터 변수의 결측 구간에 대한 결측 보정 추정값이 반영된 추정값과 실제값의 절대오차를 표준값으로 평가하여 표준 절대오차를 구하고 오차값을 평균으로 결측 구간 대치 측정값으로 대입할 수 있다. 관련 수식은 식 (1)과 같다.

4.3 데이터의 결측 구간 자동 보정 기법 성능평가

본 논문에서 제안하는 데이터의 결측 구간 보정 기법에 대한 성능평가를 위해 1,431,280개의 결측 구간의 결측치가 포함된 로우 데이터의 비중을 5%에

$$Average\ Normalized Absolute\ Error = |Data_{missing\ value}|^{-1} \sum_{i=1}^{|Data_{missing\ value}|} \frac{|x_i - \hat{x_i}| - \mu}{\sigma} x_i \quad (1)$$

$|Data_{missing\ value}|$ = *Missing Value Variable*
x_i = *Actual of Missing Value*
$\hat{x_i}$ = *Estimation of Missing Value*
μ = *Error Mean*
σ = *Error Standard Deviation*

서부터 5%단위로 구분하여 100%까지 비율을 조정하였다. 결측 보정 기법인 KNN, CI, MI, Datawig를 적용하여 총 30회에 걸쳐 ANAE 측정 평가를 진행하였다. ANAE 측정값은 작을수록 결측 구간 보정의 정확성이 우수한 것으로 분류할 수 있다. Fig. 6과 같이 본 논문에서 제안하는 DNN 및 KNN-MLE 알고리즘이 기존 연구인 KNN, MEI, MI, Datawig 기법보다 약 0.041에서 0.321까지 상대적으로 우수함을 확인할 수 있다. CI, MI는 확률적인 전통적 통계기법의 알고리즘으로 결측 구간의 결측치 포함 비율이 높아질수록 결측치 보정에 대한 성능이 개선되는 효과를 확인였다. 특히 제안하는 알고리즘은 기존 연구 방식과 다르게 로우 데이터의 결측 구간의 결측치 포함 비율과 결측치의 보정 성능이 높아질수록 보정 정확도가 급격하게 개선되는 부분을 확인할 수 있다. 이와 같은 부분은 제안하는 알고리즘의 특징인 로우 데이터 크기 및 결측 구간에 포함된 결측치 비율이 높을수록 데이터 보정 정확도가 높은 가설을 검증할 수 있었다. KNN 알고리즘은 로우 데이터의 크기가 작거나 결측 구간에 포함된 결측치 비율이 상대적으로 낮은 경우에 기존 결측 보정 기법보다 우수한 결과를 제시되었다. 그러나 결측 구간에 포함된 결측치 비율이 상대적으로 높을 때, 결측 구간이 일부 포함된 데이터 구간을 제대로 반영하지 못해 보정에 대한 성능이 낮아지는 문제점이 있다. 제안하는 결측 구간 자동 보정 알고리즘은 KNN을 적용하는 보정 구간에 MLE을 적용한 결측 구간 보정 기법을 적용하므로 KNN 알고리즘의 결측 데이터 보정 성능 저하에는 영향을 받지 않는다. Table 2는 결측값 보정 기법에 대한 기존연구와 제안하는 모델의 ANAE값에 대한 확률적 평균의 성능평가 항목이 유의미한지 측정하기위해서 t 검증과 w 검증을 측정한 결과이다. 전체 결측 t Test와 w Test를 각각 확인하였으며, KNN, Datawig의 알고리즘에 있어 전체 데이터 중 결측 포함을 5%일 경우를 제외하고 전체 결측율 포함 구간에서 제안하는 DNN, KNN_MLE기반 보정 모델의 p값을 기준으로 유의수준할 수 있는 0.05보다 이하인 것을 확인하였다. 성능평가 결과는 KNN과 Datawig가 전체 데이터 중 결측 구간의 포함이 5%일 경우 결측 보정에 대한 정확성 부분에서는 제안하는 모델과 유사하다는 의미이다. 5% 구간을 제외한 전체 평가 범위에서는 제안하는 모델의 결측율 설정 구간에

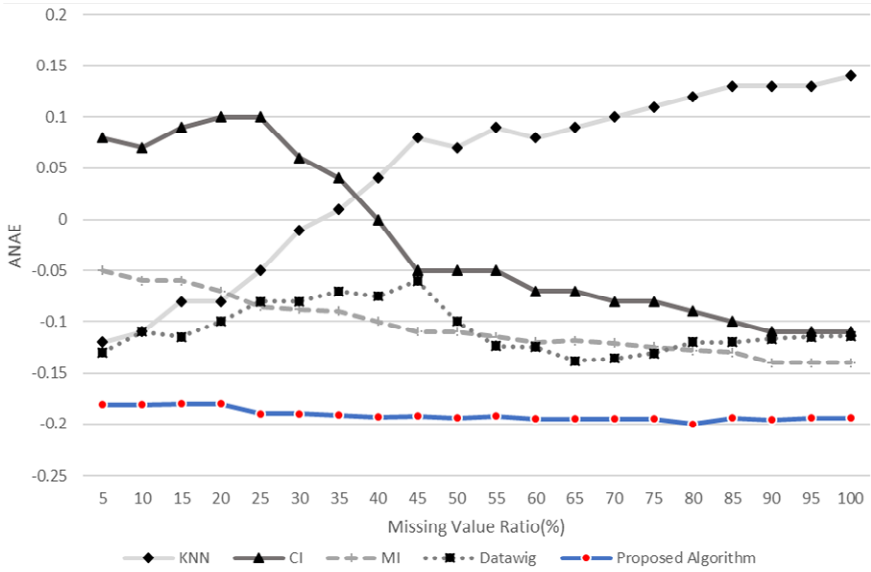


Fig. 6. Accuracy Result of existing Algorithm by Missing Value Imputation.

Table 2. t Test and w Test Result of ANAE Value through Proposed Algorithm and Missing Value.

Imputation Algorithm		KNN	CI	MI	Datawig
Ratio	Test				
5%	t Test	0.95	< 0.00	< 0.00	0.11
	w Test	0.17	< 0.00	< 0.00	0.22
90%	t Test	< 0.00	< 0.00	< 0.04	< 0.00
	w Test	< 0.00	< 0.00	< 0.00	< 0.00
95%	t Test	< 0.00	< 0.00	< 0.04	< 0.00
	w Test	< 0.00	< 0.00	< 0.03	< 0.00
100%	t Test	< 0.00	< 0.00	< 0.01	< 0.00
	w Test	< 0.00	< 0.00	< 0.01	< 0.00

서는 정확도 측면에서 더 우수하다는 성능결과를 확인하였다.

4.4 결측 데이터 보정을 통한 예측 모델

본 연구에서는 다수의 결측치가 포함된 로우 데이터(온도 데이터)를 제안하는 DNN 및 KNN-MLE 기법을 이용한 데이터 보정을 진행하였다. Fig. 7의 (a)는 로우 데이터로써 2,628,753개의 온도 데이터와 10%의 262,850개의 결측 데이터에 대한 결측치 보정을 진행하였으며, 보정된 데이터의 결과는 Fig. 7의 (b)와 같다. 보정된 데이터가 특징별로 군집화가 되도록 자동화된 클러스터 K-Value 알고리즘을 적용하여 데이터 모델을 예측하였다. 2007년부터 2016년까지 3개월 단위의 평균 기온에 대한 학습을 진행하였으며, 제안된 모델의 보정과 예측 모델이 적용된 결과는 Fig. 7의 (c)의 붉은색 예측선과 같다. Fig. 7의 (c)의 파란색 예측선은 기존 KNN 보정 기법과 DNN기반의 Datawig기반의 예측 모델이다. 측정 결과 오차는 최소 약 5도에서 최대 10까지의 오차가 발생하였다. 그리고 예측 모델의 정확성을 평가하기 위하여 로우 데이터의 학습모델에는 80%에 해당하는 2,103,003개의 데이터를 학습시켰으며, 나머지 20%에 해당하는 525,750개의 데이터는 테스트 모델에 적용하였다. 성능평가를 5회에 걸쳐 각각 평가를 진행하였으며, 평균 결과값을 확인하였다. 실험 결과 제안하는 분석 및 예측 모델을 활용할 경우 분석의 정확율은 평균 94,67%로 기존 타 연구에 비해 약 0366%~3.064%의 향상된 결과를 확인하였다.

5. 결론 및 향후 연구

인공지능 기술의 발전을 이끌고 있는 딥러닝의 성능에 비례하여 엄청난 양의 데이터가 필요한 실정이다. 그러나 빅데이터를 구축하기에 너무 많은 비용이 소모되거나, 구축할 수 없는 분야로 확장되면서 데이터의 중요도 문제는 점차 심화되고 있다. 이와 같이 향후 인공지능 기술이 필요한 대다수의 환경은 빅데이터를 축적하기 어려운 환경이기 때문에 빅데이터 의존성이 낮은 기술과 새로운 인공지능 데이터 수집하거나 손실되는 데이터에 대한 보정 기술이 필요한 실정이다. 본 논문에서는 로우 데이터의 결측 구간에 포함된 결측값을 보정하기 위한 연구를 제안하였다. 제안된 알고리즘을 통해 기존 결측 보정 알고리즘인 완전 제거, Datawig, KNN 알고리즘 등의 문제점으로 인식된 특정 구간의 결측값과 정상적인 로우 데이터의 손실 문제를 보완하였다. 데이터의 결측 구간 보정 기법에 대한 성능평가를 위해 1,431,280개의 결측 데이터를 기준으로 결측율을 5%에서부터 5% 단계로 조정하여 100%까지 성능평가를 진행하였다. 평가결과 기존 연구인 KNN, MEI, MI, Datawig 기법보다 약 0.041에서 0.321까지 상대적으로 우수하다는 점을 확인할 수 있다. 또한 데이터 예측 모델에 적용한 결과 온도 예측에 있어 약 5도에서 10도까지의 성능 차이를 확인할 수 있었다. 마지막으로 예측의 정확성 부분에 있어 평균 94.67%로 약 0.366%~3.064%의 우수성을 확인하였다. 향후 연구로는 연구결과를 바탕으로 결측 구간의 결측 데이터에 대한 보정된 데이터를 기준으로 메타학습 및 강화학습 모델을 적용한 데이터 예측 시스템을 연구할 것이다.

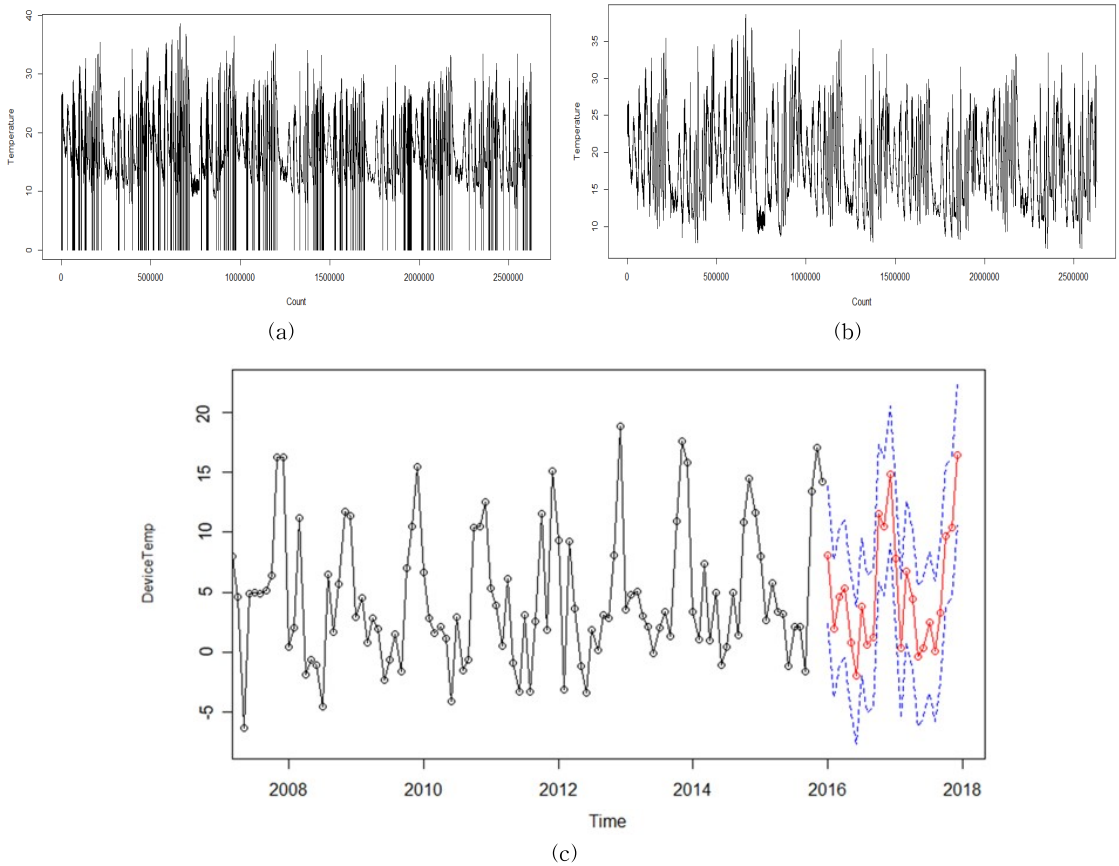


Fig. 7. Prediction Model Result of Raw Data using Missing Value Imputation. (a) Raw Data, (b) Imputation data, (c) Prediction model.

REFERENCE

[1] M.A. Kiasari, G.J. Jang, and M.H Lee, "Novel Iterative Approach using Generative and Discriminative Models for Classification with Missing Features," *Neurocomputing*, Vol. 225, pp. 23-30, 2017.

[2] V. Terauds and J. Sumner. "Maximum Likelihood Estimates of Rearrangement Distance: Implementing a Representation-Theoretic Approach," *Bulletin of Mathematical Biology*, Vol. 81, No. 2, pp. 535-567, 2019.

[3] M.P. Beugin, T. Gayet, D. Pontier, S. Devillard, and T. Jombart, "A Fast Likelihood Solution to the Genetic Clustering Problem," *Methods in Ecology and Evolution*, Vol. 9, No. 4, pp. 1006-1016, 2018.

[4] Q. Ding, J. Han, X. Zhao, and Y. Chen, "Missing-data Classification with the Extended Full-dimensional Gaussian Mixture Model: Applications to EMG-based Motion Recognition," *IEEE Transactions on Industrial Electronics*, Vol. 62, No. 8, pp. 4994-5005, 2015.

[5] M.K. Kim, S.D. Park, J.H. Lee, Y.J Joo, and J.K. Choi, "Learning-Based Adaptive Imputation Method with kNN Algorithm for Missing Power Data," *Energies*, Vol. 10, No. 10, pp. 1-20, 2017.

[6] C.C. Turrado, F.S. Lasheras, J.L. C-R, A.J. P-P, and F.J.C. Juez, "A New Missing Data Imputation Algorithm Applied to Electrical Data Loggers," *Sensors*, Vol. 15, No. 12 pp.

31069-31082, 2015.

[7] A. Chaudhry, W.Li, A. Basri, and F. Patenaude, "A Method for Improving Imputation and Prediction Accuracy of Highly Seasonal Univariate Data with Large Periods of Missingness," *Wireless Communications and Mobile Computing*, Vol. 2019, pp. 1-14, 2019.

[8] K. Strike, K.EI. Emam, and N. Madhavji, "Software Cost Estimation with Incomplete data," *IEEE Transactions on Software Engineering*, Vol. 27, No. 10, pp. 890-908, 2001.

[9] S. Shiliang and R. Huang, "An Adaptive k-Nearest Neighbor Algorithm," *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, Vol. 1, 2010.

[10] W.R. Parke, "Pseudo Maximum Likelihood Estimation: The Asymptotic Distribution," *The Annals of Statistics*, Vol. 14, No. 1, pp. 355-357, 1986.

[11] J.C. Kim, C.B. Sim, and S.H. Jung, "A Study on Automatic Missing Value Imputation Replacement Method for Data Processing in Digital Data," *Journal of Korea Multimedia Society*, Vol. 24, No. 2, pp. 245-254, 2021.

[12] M. Ingunn, E. Stensrud and U. H. Olsson, "Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods," *IEEE Transactions on Software Engineering*, Vol. 27, No. 11, pp. 999-1013, 2001.

[13] S. Qinbao and M. Shepperd, "A New Imputation Method for Small Software Project Data Sets," *Journal of Systems and Software*, Vol. 80, No. 1, pp. 51-62, 2007.

정 세 훈

2010년 2월 순천대학교 멀티미디어공학과 공학사
2012년 2월 순천대학교 멀티미디어공학과 공학석사
2017년 2월 순천대학교 멀티미디어공학과 공학박사

2018년 9월~2020년 2월 영산대학교 빅데이터융합전공 조교수
2020년 3월~현재 국립안동대학교 창의융합학부 조교수
관심분야 : 최적화 알고리즘, 강화학습, 블록체인

이 한 성

1996년 2월 고려대학교 전산학과 학사
2002년 2월 고려대학교 전산학과 석사
2008년 2월 고려대학교 전산학과 박사

2009년 11월~2014년 11월 한국전자통신연구원 (ETRI)
2014년 12월~2019년 8월 Samsung Research (삼성전자)
2019년 9월~2021년 2월 영산대학교 컴퓨터공학부 조교수
2021년 3월~현재 국립안동대학교 창의융합학부 조교수
관심분야 : 기계학습, 딥러닝, 소프트컴퓨팅, 컴퓨터비전, 멀티미디어 데이터마이닝, 네트워크보안, 침입탐지

김 준 영

2019년 2월 순천대학교 멀티미디어공학과 졸업(공학사)
2019년 3월 ~ 2021년 2월 순천대학교 멀티미디어공학과 공학석사
2021년 3월 ~ 현재 순천대학교 스마트융합학부 박사과정
관심분야 : IoT 상황인식, 빅데이터 처리 및 분석, 딥러닝

심 춘 보

1996년 2월 전북대학교 컴퓨터공학과 공학사
1998년 2월 전북대학교 컴퓨터공학과 공학석사
2003년 2월 전북대학교 컴퓨터공학과 공학박사

2005년 3월~현재 순천대학교 ICT융합학부 교수
관심분야 : 빅데이터 시스템, 머신러닝, IoT/IoE 플랫폼, 멀티미디어