Journal of Korea Multimedia Society Vol. 24, No. 2, February 2021(pp. 245-254)
https://doi.org/10.9717/kmms.2020.24.2.245

디지털 데이터에서 데이터 전처리를 위한 자동화된 결측 구간 대치 방법에 관한 연구

김종찬†, 심춘보††, 정세훈†††

A Study on Automatic Missing Value Imputation Replacement Method for Data Processing in Digital Data

Jong-Chan, Kim†, Chun-Bo, Sim††, Se-Hoon, Jung†††

ABSTRACT

We proposed the research on an analysis and prediction model that allows the identification of outliers or abnormality in the data followed by effective and rapid imputation of missing values was conducted. This model is expected to analyze efficiently the problems in the data based on the calibrated raw data. As a result, a system that can adequately utilize the data was constructed by using the introduced KNN + MLE algorithm. With this algorithm, the problems in some of the existing KNN-based missing data imputation algorithms such as ignoring the missing values in some data sections or discarding normal observations were effectively addressed. A comparative evaluation was performed between the existing imputation approaches such as K-means, KNN, MEI, and MI as well as the data missing mechanisms including MCAR, MAR, and NI to check the effectiveness/efficiency of the proposed algorithm, and its superiority in all aspects was confirmed.

Key words: KNN, Maximum Likelihood Estimation, Missing Value Imputation, Processing

1. 서 론

광범위한 데이터가 생성되고 이를 활용하는 시스템에서 로우데이터에 대한 중요성은 매우 높다. 그러나 데이터의 범주가 너무 광범위하고 결정론적 분석을 위한 고정된 패턴 형태의 데이터가 확보되지 않는 상황에서는 데이터의 활용도가 떨어지는 문제점이 발생하고 있다. 특히 IoT 데이터는 수집과 통신 과정에서 이상치와 결측치 값이 다수 발생할 수 있다. 이러한 데이터는 분석 시스템에서 편향된 결과나 분석 품질 및 분석 정확도가 낮아지는 문제점으로 귀결되고 있다[1-3]. 분석 시스템에서 이상치와 결측치에 대한 전처리기법의 기존 연구는 국내외에서 활발하게 진행되고 있다[4-7]. 일반적으로 분석 시스템을 활용하기 데이터의 전처리 과정에서 발생하는 데이터 결측처리 방법은 크게 2가지로 구분된다. 결측치

※ Corresponding Author : Se-Hoon, Jung, Address: (36729) 1375 Gyeongdong-ro, Andong-si, Gyeongsnagbuk-do, Korea, TEL : +82-54-820-6894, FAX : +82-54-820-6825, E-mail : jungsh@anu.ac.kr
Receipt date : Jan. 19, 2021, Approval date : Jan. 21, 2021
† Department of Computer Eng., Sunchun National University
(E-mail : seaghost@sunchon.ac.kr)
†† School of Information Commucation·Multimedia Eng., Sunchun National University
(E-mail : cbsim@sunchon.ac.kr)
††† School of Creative Convergence, Andong National University
※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020 R1I1A3054843). and this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2019 R1G1A1002205).

를 제거하는 방법과 결측치를 보정하는 방법으로 구분할 수 있다.

데이터 결측치를 제거하는 방법은 다음 두 가지로 구분된다. 첫 번째 기법은 결측치 제거 방법인 완전 제거법이 있다. 완전 제거법은 수집된 모든 변수에서 유효한 값을 가진 관측치만이 분석 및 예측 모델의 대상이 된다는 필요충분조건을 가진다. 두 번째 기법은 결측치 대체 방법으로 단순 대치법과 다중 대치법으로 연구되고 있다. 이러한 결측치 대체 방법은 중심 경향값 대체기법, 랜덤 추출기법, 회귀 삽입 기법, 다중 대체기법이 있다. 중심 경향값 대체기법은 결측치를 데이터의 평균 등 중심 경향 값으로 대체하는 방식이다. 랜덤 추출 기법은 결측치를 랜덤에 의해 자주 나타나는 값으로 대체하는 방식이다. 회귀 삽입 기법은 각 관측치의 특성을 고려하여 회귀계수에 의한 값들을 삽입하는 방식이다. 다중 대체 기법은 다양한 모델을 여러번 반복하여 통합된 결과로 대체하는 방식이다.

데이터 결측치를 보정하는 방법은 다음 2가지로 구분된다. 첫 번째 기법은 결측치 발생 구간에 인접한 로우데이터를 선형 보간법 알고리즘을 통해 보정하는 방법이다. 이 기법은 매우 간단하며 측정 데이터에 일관된 패턴이 있거나 이상치 및 결측치 데이터의 간격 폭이 짧은 경우 결정론적 분석 모델에 우수한 결과를 나타낸다. 그러나 수집되 데이터에 일관된 패턴이 부족하거나 이상치 및 결측치 로우데이터의 간격 폭이 넓을 경우에는 분석 모델의 정확성은 떨어질 수 있다는 문제점이 존재한다. 두 번째 기법은 결측 시점에서 과거의 데이터를 기반으로 결측치가 발생한 구간과 동일한 패턴의 구간 데이터를 활용하여 결측치를 대체하는 K-최근접 이웃 알고리즘이다. 해당 알고리즘 역시 동일한 과거 시점의 패턴을 활용하기 때문에 패턴의 형태가 근접할 경우 더 우수한 결과를 도출한다. 그러나 기존에 연구된 결측치 대체 기법은 패턴이 존재할 경우 우수한 결과를 확인할 수 있지만 로우데이터는 예상치 못한 특수한 요인으로 인하여 일정한 패턴과 주기적인 패턴의 데이터 수집이 불가능할 수 있다. 주기적인 패턴을 지니고 있는 특정 데이터 셋에서는 기존 연구 방식의 결과가 우수하지만 여러 종류의 패턴을 가지고 있는 데이터 셋에서는 결측 구간에 대한 패턴을 확보하는 방법이 적합하지 않다.

본 연구에서는 디지털 환경에서 데이터 전처리를 위한 데이터 결측 및 보정 처리 알고리즘을 제안한다. 첫 번째는 결측치 데이터의 보정 알고리즘을 통해 결측치 데이터를 대체하며, 두 번째는 보정된 로우데이터를 기준으로 데이터 이상 여부와 특이점을 확인할 수 있도록 변형된 K-means와 주성분 분석 알고리즘을 결합한 데이터 분석 모델을 제안한다.

본 연구에서는 결측치 보정 연구를 위해 데이터의 불규칙한 시계열 데이터 패턴 및 결측 처리 방안을 머신러닝 기법을 사용하여 데이터 결측이 발생한 지점의 상황과 가장 유사한 과거 시점의 상황을 찾아 유사 과거 시점의 데이터로 결측을 대체하는 방안을 제안하고자 한다. 기존 최근접 이웃 알고리즘은 결측값 대치 시 결측값을 포함하지 않는 완전한 패턴만을 대상으로 하기 때문에 결측값을 포함하는 인스턴스의 관측값들을 활용하지 못하는 단점이 있었지만, 본 연구에서는 이러한 불규칙 패턴을 구분하여 관측정보로 활용하고 패턴의 후보로 선택하고자 한다. 이를 위해 최대 우도 추정법 알고리즘과 최근접 이웃 알고리즘을 결합한 기법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 데이터 보정 및 제시한 방법인 최대 우도 추정법과 관련된 기존 연구 내용을 제시하며, 3장에서는 제안하는 데이터 결측 및 보정 알고리즘 제시하고, 4장에서는 제안한 알고리즘에 대한 성능평가에 대해 제시한다. 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

2.1 결측 데이터 대치법

데이터 분석 및 예측 모델에 사용되는 로우데이터들은 다양한 형태로 수집되게 된다. 결측값은 이러한 관측 또는 실험되어 얻어지는 로우데이터에 종종 나타나는 현상으로 데이터에 결측값이 포함되어 실험에 사용되지 않고 버려지는데, 이러한 부분은 데이터 손실로 야기되어 일부 데이터에 편중된 학습 모델을 생성할 수 있다. 결측치를 가진 데이터 분석을 위한 통계적 분석 모형들은 현재에도 수 많은 연구들에 의해 문제를 해결하고 있다.

결측/무응답을 가진 데이터를 분석에 활용할 수 있는 수치적 방법론의 하나인 결측 대치법은 현재에도 연구가 진행중이다[3-4]. 결측값이 포함된 데이터

Table 1. A Processing Method of Missing Data Mechanism.

Item		Instance 1	Instance 2	Instance 3	Instance 4	Instance 5	Instance 6	Instance 7	Instance 8
Valuable 1		A	A	B	B	C	C	D	D
Valuable 2	Value	85	101	111	80	115	95	87	116
	MCAR	85	101	111	80	115	95	87	116
	MAR	85	101	Missing	Missing	115	95	87	116
	NI	85	Missing	Missing	80	Missing	95	87	Missing

에 대한 결측값 처리를 결정하기 위해서는 결측치 처리 알고리즘을 이해하는 것이 중요하다. 결측 매커니즘은 결측값과 데이터 변수들과의 연관 관계를 의미한다. 결측 매커니즘은 변수들에 대한 결측의 의존 여부에 따라 Missing Completely At Random(MCAR), Missing At Random(MAR), Non Ignorable(NI)로 구분된다. Missing Completely At Random은 결측이 변수의 결측값 포함 여부와 상관없이 어떠한 변수들과도 무관한 경우를 의미하는 것으로 결측이 랜덤하게 분포한 경우를 뜻한다[7].

Missing At Random은 결측이 오직 결측값을 포함하지 않는 변수들과 연관이 있는 경우이며, Non Ignorable은 결측이 오직 결측값을 포함하는 변수들과 관련이 있는 경우에 포함된다. 표 1은 두 개의 변수로 구성된 8개의 데이터 인스턴스들을 통해 3가지 결측 메커니즘을 구분하였다. Missing Completely At Random은 변수 2의 결측이 결측값을 포함하지 않는 변수 1과 결측값을 포함하는 변수 2에 무관함을 보이고 있다. Missing At Random는 변수 2의 실제값에 관계없이 변수 1이 B일때만 결측이 발생하여 오직 결측값을 포함하지 않는 변수 1의 값에 의존함을 알 수 있다. Non Ignorable은 변수 1에 관계없이 변수 2의 값이 100이상일 경우에 결측이 발생하여 결측값이 포함된 변수 2와 관련이 있음을 보이고 있다[7].

결측 매커니즘을 통해 발생되는 결측값을 대치하기 위한 기존 결측값의 대치법은 2가지 형태인 단순 대치법과 다중 대치법으로 구분된다. 단순 대치법은 Completes Analysis, Mean Imputation, Single Stochastic Imputation 다중 대치법은 Multiple Imputation으로 구성된다. Completes Analysis는 완전 제거법으로 불완전 자료를 모두 무시하고 관측된 자료만으로 표준적 통계 추정 기법을 적용한다. Completes Analysis는 분석이 쉽다는 특징이 있지만 관측된 자료를 삭제함으로써 편향된 결과 또는 통계적 추론의 타당성 측면에서는 문제점이 존재한다. Mean Imputation는 관측된 자료의 적절한 평균값으로 대치하여 불완전한 자료를 완전한 자료로 변경하는 기법이다. 대표적으로 조건부 평균 대치법과 비조건부 평균 대치법이 있다. 조건부 평균 대치법의 확장으로써 Buck's 방법이 대표적이다.

Single Stochastic Imputation(SSI)은 평균대치법에서 오차의 과소 추정 문제점을 상쇄하기 위해 제안된 알고리즘으로 통계량을 결측값에 치환할 때 적합한 확률값을 치환한 후 대치하는 기법이다. Single Stochastic Imputation의 알고리즘은 지적된 과소 추정 문제점을 일부분 보완하였지만 일반적인 문제를 제외한 경우에 추정량의 표준 오차 계산 자체가 어려운 문제점이 존재한다. Multiple Imputation 방법은 SSI의 문제점을 보완하기 위해 단순 대치법처럼 m번의 대치 기법을 통해 m개의 가상적 데이터를 생성하는 분석하는 기법이다. Multiple Imputation(MI)은 3단계로써, 대치, 분석, 결합단계로 분류된다. [8]에서는 디지털 데이터의 결측 구간이 발생하는 시점을 기준 n개의 데이터 시점(시간+값)과 데이터 결측 구간이 사라지는 지점에서 측정된 데이터 위치값을 결합하여 하나의 벡터로 설정하는 연구를 제시하였다. 전력사용량 데이터로 이루어진 데이터 벡터의 차분값을 추출하여 기준 벡터에 특징값을 결합한 벡터를 활용하고 벡터와 이전 벡터간의 거리 유사도를 LWED(Linear Weighted Euclidean Distance)으로 적용하여 수치화하였다. [7]의 연구에서는 SW 품질 측정 비용 추정을 위해 수집된 데이터의 결측에 대한 수치값을 해결하는 방법을 제시하였으며, 기존의 결측값 처리 방법들을 비교 및 분석하였다. 비교 대상으로는 결측값을 포함한 데이터를 삭제하는 기법인 완전 제거법과 결측값 대치법인 평균 대치법, K-NN이다. [8]의 연구에서는 SW 품질 측정 비용 추정 모델을 위해 결측값 대치법이 완전 제거법보다 효율적

이며, 비용 추정 모델을 위한 결측값 대치 방법으로 KNN을 제안하였다. [9]의 연구에서는 데이터에 결측값을 포함하여 4가지 결측값 처리 기법들을 측정하고 그 결과를 평가하였다. 평가 대상으로는 완전 제거법, 평균 대치법, 유사 유형 대치 기법, 최대 우도 추정법이었고, 실험 평가에서는 실험 데이터 셋의 규모가 클 경우 최대 우도 추정법의 적용이 적절하는 결과를 제시하였다. [10]의 연구에서는 소규모 데이터의 결측값 처리를 위해 KNN 알고리즘을 활용한 클래스기반의 대치법을 통해 결측값을 처리하는 알고리즘을 제안하였다. [10]의 연구에서는 데이터값 및 데이터 범주형 결측값에 적용할 수 있으나 데이터의 규모가 작을 경우에만 알고리즘 적용이 가능한 단점이 있다.

2.2 최대 우도 추정법

최대 우도 추정법(Maximum Likehood Estimation)은 영국의 통계학자 피셔(Fisher)가 20세기에 고안한 파라미터 추정법이다. 우도함수(Likehood Function)를 최대화하는 파라미터 Θ를 결정하는 방법이다. 결합밀도함수로 구해진 우도함수 $p(D|\theta)$이다. 여기서 D는 데이터 집합을 의미하며, 학습 데이터가 많을수록 함수의 폭이 좁아진다. 우도를 최대로 하는 파라미터 값이 Θ로 표시되어 있다. 우도를 최대로 하는 파라미터 Θ의 위치가 우도와 로그 우도 함수일 경우가 동일함을 알 수 있다. 최대 우도 추정법을 이용하여 Θ을 구하는 과정은 다음과 같다. 파라미터 $\theta = (\theta_1, \theta_2, ..., \theta_i)$ 집합으로 구성된 특정 확률밀도함수 $p(x|\theta)$로부터 포함된 데이터를 $x = (x_1, x_2, ..., x_i)$로 정의할 때, 이러한 x로부터 파라미터를 예측하기 위해 $p(x|\theta)$는 식 (1)와 같이 정의된다. 식 (1)은 확률함수이므로 가장 큰 확률값을 발생시키는 Θ값을 추정값으로 결정한다. $p(x|\theta)$는 파라미터 Θ에 의한 데이터 집합의 우도 함수이다. 우도함수를 합으로 변경하기 위해서 로그를 취하며, 식 (2)는 로그 우도 함수를 정의한 식이다. 단조증가 함수이므로 우도를 최대로 하는 파라미터 Θ를 찾기 위해 Θ에 관한 편미분식을 0으로 식 (3)과 같이 정리된다. 결측값 대치 기법의 기존 연구에서 최대 우도 추정법은 결측 매커니즘 중 Missing At Random를 전제로 하며, 다변량 정규 분포가 아닌 데이터에 대해서도 활용 가능 하지만 다른 결측값 대치법에 비하여 데이터 크기가 큰 규모의 데이터가 필요하다는 단점을 보완해야 한다.[11-14].

$$p(x|\theta) = \prod_{i=1}^{n} p(x_i|\theta) \tag{1}$$

$$\log p(x|\theta) = \prod_{k=1}^{n} \log p(x_k|\theta) \tag{2}$$

$$\frac{\partial}{\partial\theta}\log p(x|\theta) = -\prod_{k=1}^{n}\frac{\partial}{\partial\theta}\log p(x_k|\theta) = 0 \tag{3}$$

3. 제안하는 데이터 결측 및 보정 알고리즘

Fig. 1은 데이터 결측 및 보정 알고리즘의 구성도이다. 데이터에 결측치가 있을 때 이를 무시하고 분석을 진행할 경우 표본 수의 감소로 인해 검정력이 낮아지게 된다. 결측치들을 단순히 제거하기 보다는 관련 패턴 및 유의미한 값으로 대치하여 분석하는 방법은 편향된 결과나 분석 품질 및 분석 정확도가 낮아지는 문제점을 보완할 수 있다. 결측치 대치에 관한 기존 연구는 KNN 알고리즘을 적용한다, 이러한 알고리즘은 통계적 접근과 다르게 데이터 분포에 대한 가설 검증이 불필요하다는 장점이 있다. 일정한 패턴이나 동일한 패턴이 존재하지 않을 경우 결측치 대치에 대한 성능을 낮아지는 문제점이 있다. 일반적으로 데이터는 대규모일 수 있고 또는 소규모일 수 있고, 패턴이 일정하지 않는 경우도 존재한다. 데이터의 이러한 특성을 반영한 결측 대치 연구가 요구되고 있다.

본 논문에서는 이러한 문제점을 해결하고자 K-최근접 이웃 알고리즘과 최대 우도 추정법을 결합한 알고리즘을 제안한다. 최대 우도 추정법은 통계적 패

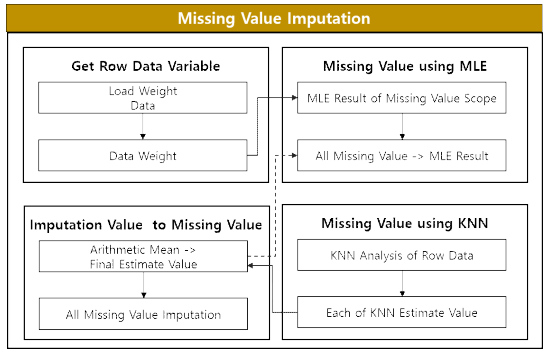

Fig. 1. Data Missing Value Imputation and Replacement Algorithm Configuration.

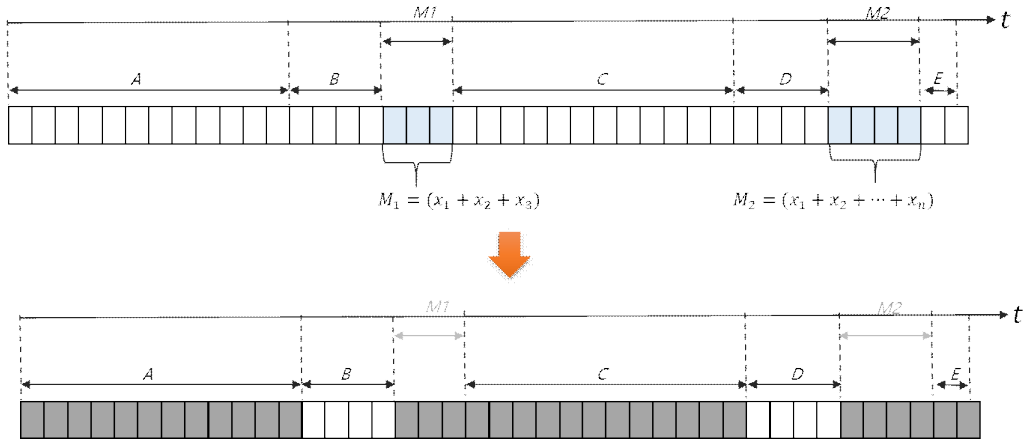

Fig. 2. Example of Data in Missing Value (K-Nearest Neighbor Algorithm).

턴 추론 기법으로 최적의 Θ를 분석하는 것으로 데이터 규모가 클수록 사건의 가능도가 최대가 되는 추정값을 정확하게 찾는 것이 가능하다는 특징이 있다. 기존 연구에서 활용되는 K-최근접 이웃 알고리즘은 비교적 구현이 간단하며, 2장에서 언급한 결측 매커니즘 3가지 분류에 대해 강건하며 결측값 대치의 결과도 우수한 편이다. Fig. 2와 같이 수집된 데이터에는 5개의 패턴으로 구분된다. A(일부 이상치 및 결측치 존재), B, C(일부 이상치 및 결측치 존재), D, E(일부 이상치 및 결측치 존재) 구간은 정상적인 데이터가 존재하는 구간이며, 파란색으로 표시된 M_1과 M_2인 2개의 구간에서 데이터 결측치가 발생하였다. 결측치가 발생한 구간에 대해서는 정상적인 데이터 구간을 활용하는 방법이 기존 연구 방법인 평균 대치법과 조건부 평균 대치법, 다중 대치법이다.

본 논문에서는 Fig. 2와 같이 활용되지 않는 구간인 A, C, E 구간을 활용하여 결측치를 대치하는 방법을 제안한다. 이를 위해 최대 우도 추정법을 활용하여 이 문제점을 해결한다. 최대 우도 추정법은 전체 데이터를 활용하여 최적의 Θ를 찾는 매커니즘을 활용한다. 데이터의 결측 발생 구간을 제외한 모든 구간을 활용하고 결측값에 대한 추정값을 계산하여 결측값 대치를 진행한다. K-최근접 이웃 알고리즘과 최대 우도 추정법 기반의 데이터 결측 추정 및 보정을 위해 제안하는 결측값 대치 알고리즘은 Fig. 3과 같다.

제안하는 알고리즘은 먼저 데이터에 포함된 결측값을 수집된 전체 데이터에 대한 최대 우도 추정법의 결과값으로 1차 대치한다. 이는 추후 K-최근접 이웃 알고리즘 적용 시 결측값을 포함하는 인스턴스들의 관찰 데이터 활용을 가능하게 하는 것을 목적으로 한다. 두 번째로 결측 구간의 전체 결측값을 최대 우도 추정 결과값으로 초기화한 인스턴스들 중 하나의 인스턴스에 대해 최대 우도 추정 결과값을 결측값으로 변경한다. 이는 하나의 결측값에 대한 추정값을

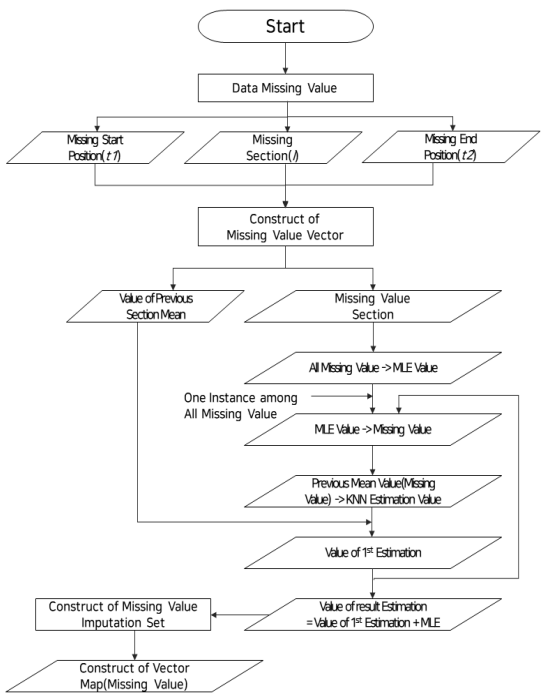

Fig. 3. Flow chart of Proposed KNN-MLE Method.

계산하기 위하여 결측값으로 변경된 인스턴스를 제외한 전체 데이터를 K-최근접 이웃 알고리즘 적용하기 위함이다. 세 번째로는 결측값을 가진 하나의 인스턴스에 대해 K-최근접 이웃 알고리즘을 적용하여 결측값 대치의 추정값을 계산한다.

이때 처음 결측값을 최대 우도 추정법의 결과값으로 1차 대치한 인스턴스들도 K-최근접 이웃 알고리즘 결과 계산에 모두 활용되어 결측값을 가진 인스턴스들의 관측 데이터를 모두 이용할 수 있게 된다. 기존 K-최근접 이웃 알고리즘에서는 결측값을 가진 M_1가 M_2의 결측 대치값을 찾는 과정에서 제외되었지만 본 논문에서 제안하는 알고리즘에서는 최대 우도 추정법을 통해 1차 대치된 값을 M_1가 가지고 있으므로 M_1의 결측값을 임시 벡터 테이블에 저장하고 첫 번째 단계에서 적용된 최대 우도 추정법 결과값을 관측 데이터로 활용하여 M_2의 결측 대치값을 찾는 과정에 포함시키는 것이 특징이다. 네 번째 단계로는 해당하는 결측 구간의 이전 데이터 구간의 산술평균과 세 번째 단계에서 계산된 K-최근접 이웃 알고리즘 추정값을 산술평균하여 결측값에 대한 결측값 인스턴스의 1차 추정값을 산출한다. 이 과정은 데이터의 특성을 반영하여 기존 수집된 데이터의 영속성과 지속성을 반영하기 위하여 추가된 과정이다. 다섯 번째 단계로는 첫 번째 단계에서 변경한 결측값에 대한 각 인스턴스의 최대 우도 추정 결과값과 네 번째 단계에서 계산된 1차 추정값을 산술평균하여 결측값에 대한 결측값 인스턴스의 최종 추정값을 산출한다. 이러한 일련의 과정을 결측값의 결과값을 1차 대치한 모든 인스턴스들이 만족할 때까지 반복하는 과정을 적용한다. 마지막으로 최종 추정값들을 모두 저장하여 결측값이 추정값으로 대치된 완전한 결측 대치 벡터가 완성되면 결측 대치가 완료된다. 인스턴스에 대한 대치되는 추정값이 인스턴스 추정에 반영되지 않도록 최종값을 데이터에 즉시 반영하지 하고 임시로 저장 후 최종 결과에 반영한다.

4. 데이터 결측 대치 및 보정 알고리즘 성능평가

제안된 결측 대치 알고리즘은 기존 연구 기법과 비교하여 최대 우도 측정법을 활용한 K-최근접 이웃 알고리즘을 결합하였다. 이를 통해 소규모 데이터 및 일부 데이터를 기준으로 결측 대치값을 결정하였던 기존 K-최근접 이웃 알고리즘과 비교하여 대규모 및 전체 데이터를 활용할 수 있다는 특징이 있다. 이러한 결측 대치 기법의 정확도 비교 연구에 활용되는 평가 척도는 다음과 같다. 평가 척도는 Average Absolute Error(AAE)와 Average Normalized Absolute Error(ANAE)로 구분할 수 있다. Average Absolute Error(AAE)는 특정 변수에 포함된 결측값에 대한 결측값 대치 정확도를 확인하기 위해서 평가에 반영한다. 본 논문에서는 데이터를 최대 우도 추정법을 기반으로 특정 결측 구간에 해당되는 ANAE 수치를 기존 연구와 비교평가를 진행하여 데이터의 결측 대치 및 보정의 우수성을 평가한다.

4.1 데이터 형태

본 논문에서는 IoT 데이터 분석 및 결측치 대치 보정의 알고리즘을 확인하기 위하여 센서 데이터는 Table 2와 같다. 알고리즘 성능평가에 활용하기 위한 데이터 수집 기간은 2018년 1월 1일부터 2018년 12월 31일까지 12개월의 데이터이며, 환경 센서 데이터이다. 측정 주기는 1시간 단위이며 수집된 데이터 타입는 Temperature, Acceropitch, Roll이다. 전체 데이터셋은 크기는 100MB이며, 로우데이터당 51,312개의 데이터가 수집되었다.

4.2 데이터의 결측 구간 추출 및 보정

IoT 데이터는 Temperature, Acceropitch, Roll의 로우데이터이며, 수집된 데이터는 로우데이터당 51,312개의 데이터를 실험에 반영하여 결측구간에 대한 결측 대치 및 보정 수행을 진행하였다. 결측 대치 및 데이터 전처리를 통해 필터링된 IoT 데이터는

Table 2. Components of the dataset.

Name	Variable name	Explanation
Temperature (℃)	temp.	Total of temperature by 1 hour
Acceropitch	Acc.	Total of acceropitch by 1 hour
Roll	Roll	Total of roll by 1 hour

다음과 같다. Temperature는 IoT 센서의 이상 여부를 확인할 수 있는 데이터로 활용되며, Acceropitch, Roll은 IoT 센서의 움직임을 확인할 수 있는 데이터로 활용된다. Fig. 4는 수집된 로우데이터 원본이며, Temperature, Acceropitch, Roll의 51,312개 데이터의 분포도이다. Temperature는 0도에서 32.4도까지의 범위를 나타내고 있다. Acceropitch는 23.0도에서 25.6도 사이에 분포되어 있으며, Roll은 -65.4도에서 -64.0도 사이에 분포도를 나타내고 있다. 수집된 데이터에는 결측치 및 이상치 데이터가 포함되어 있어 제안하는 결측 대치 알고리즘을 통해 보정하였다. Fig. 5는 결측 대치를 통해 보정된 분포도이다. Temperature는 15.4도에서 32.4도까지의 범위를 나타내고 있다. Acceropitch는 23.2도에서 25.5도 사이에 분포되어 있으며, Roll은 -65.5도에서 -63.0도 사이에 분포도로 보정되었다.

4.3 데이터의 결측 보정 비교평가

본 논문에서 제안하는 IoT 데이터의 결측 대치 성능평가를 위해 1,241개의 결측치 데이터를 5%에서부터 100%까지 비율을 조정하며, 결측 매커니즘이 MCAR, MAR, NI를 생성하여 K-means, KNN, MEI, MI의 알고리즘과 성능평가를 진행하였다. ANAE 수치는 작을수록 결측 대치의 정확성이 우수한 것으로 평가된다.

Fig. 6과 같이 본 논문에서 제안하는 KNN-MLE 알고리즘이 기존 연구인 K-means, KNN, MEI, MI 기법보다 우수한 것을 확인할 수 있다. K-means, KNN, MEI, MII는 확률적인 통계기법의 알고리즘으로 결측 비율이 높아질수록 성능이 개선되는 효과를 보였다. 특히 KNN 알고리즘은 기존 다른 연구 방식과 다르게 결측 비율이 높아질수록 정확성이 확연하게 개선되는 것을 확인할 수 있다. 이러한 현상은 KNN 알고리즘의 특징으로 데이터 규모 및 결측 비율이 향상할수록 성능이 우수하다는 가설을 확인할 수 있었다. 그러나 MI 알고리즘은 소규모 데이터이거나 결측 비율이 낮을 경우 우수한 결과를 보였지만 결측 비율이 높아질수록 결측 구간이 포함된 정상 데이터 구간을 반영하지 못하여 성능이 떨어지는 결과를 보였다. 그리고 제안하는 알고리즘은 KNN을 적용할 때 MLE로 변경된 데이터들을 활용하기 때문에 KNN의 성능 저하에 크게 영향을 받지 않다는 것을 확인하였다. 결측값 대치 기법별 ANAE 측정 값들에 대한 평균의 비교가 유의미한지 평가하기 위한 t Test와 w Test 결과이다.

각 비율별로 t Test와 w Test를 각각 확인하였으며, KNN, MI의 결측 비율이 5%일 때, 비교 평가를 제외하고 모든 구간과 영역에서 제안하는 알고리즘의 p-value가 유의수준인 0.05보다 작은 것을 확인하였다. 이러한 결과는 KNN과 MI의 결측 비율이 5%

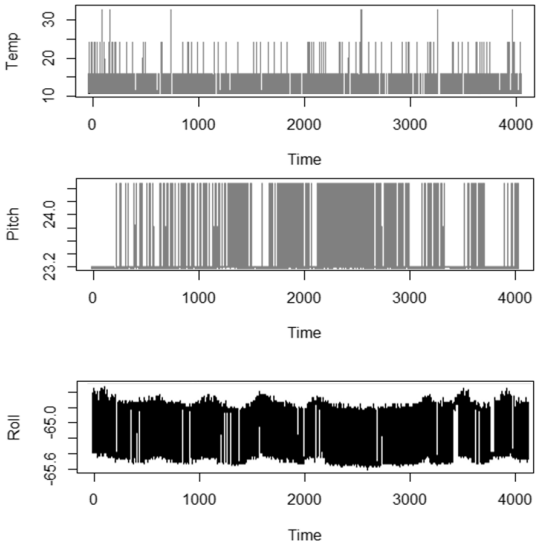

Fig. 4. IoT Original Data.

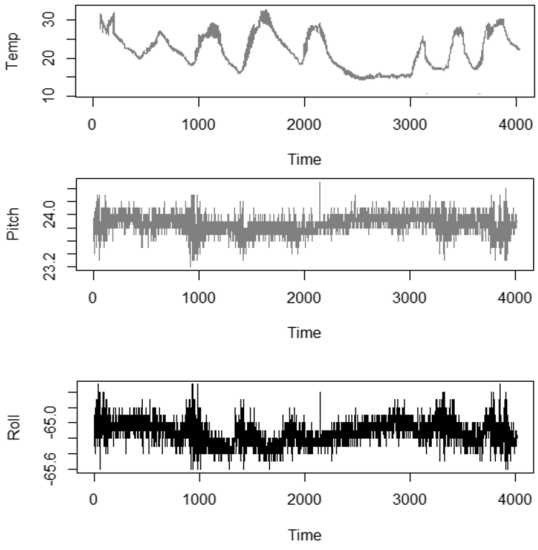

Fig. 5. Data of Missing Value Imputation Processing.

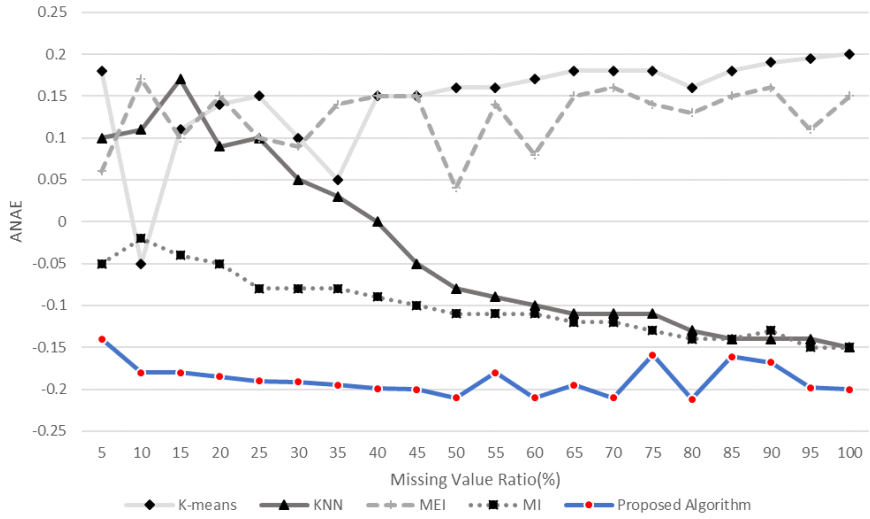

Fig. 6. The Missing Value Imputation and Replacement of Compared with the Existing Algorithm.

일 때 정확도 측면에서는 제안하는 알고리즘과 유사하다는 의미이며, 해당 구간을 제외한 모든 비율과 구간에서는 제안하는 알고리즘의 결측 비율의 정확도가 더 우수하다는 성능평가 결과이다.

5. 결 론

본 논문에서는 K-최근접 이웃 알고리즘과 최대 우도 추정법을 통해 데이터에서 발생하는 결측치를 빠르게 대치하고, 데이터 이상 여부와 특이점 확인에 대한 연구를 제안하였다. 이를 통해 보정된 로우데이터를 기준으로 데이터의 이상유무 및 외력을 통해 발생되는 문제점을 빠르게 분석할 수 있는 모델을 구현하였다. 특히 IoT 센서 데이터에서 수집되는 결측 데이터를 보정하고 이를 활용하기 위해 K-최근접 이웃 알고리즘과 최대 우도 추정법을 결합한 KNN+MLE 알고리즘을 적용하여 데이터 전처리 및 데이터를 적절하게 활용할 수 있는 시스템을 제시하였다. 제안된 알고리즘을 통해 기존 결측 대치 알고리즘인 KNN 알고리즘의 단점으로 확인된 일부 구간의 결측치와 정상적인 관측 데이터의 버려지는 문제점을 보완하였다. 이를 확인하기 위해 기존 결측 대치법(K-means, KNN, MEI, MI)과 결측 매커니즘(MCAR, MAR, NI)과의 비교평가를 확인하였다.

향후 연구로는 제안한 알고리즘을 기반으로 추가적인 데이터 범위 확대와 데이터 규모를 확대하여 데이터의 불균형 현상을 해결하여 데이터간의 의미론적인 상관관계를 파악하고 다양한 예측 알고리즘을 연구하여 비교 및 평가하고자 한다. 이러한 연구결과를 바탕으로 결측치가 보정된 데이터를 기준으로 강화학습 모델을 적용한 데이터 분석 시스템을 구축할 것이다.

REFERENCE

[1] 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.

[2] S. Zahriah, R. Yusof, and J. Watada. “FINNIM: Iterative Imputation of Missing Values in Dissolved Gas Analysis Dataset,” *IEEE Transactions on Industrial Informatics*, Vol. 10, No. 4, pp. 2093-2102, 2014.

[3] 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.

[4] C. C. Turrado, F. S. Lasheras, J. L. C-R, A.-J. P-P, M.G. Melero, and F. J. C. Juez, “A Hybrid Algorithm for Missing Data Imputation and Its Application to Electrical Data Loggers,”

Sensors, Vol. 16, No. 9, pp. 1-13, 2016.

[5] 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.

[6] 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.

[7] D.H. Lee, K.A. Yoon, and D.H. Bae, "A Missing Data Imputation by Combining K nearest Neighbor with Maximum Likelihood Estimation for Numerical Software Project Data," *Journal of KIISE: Software and Applications*, Vol. 36, No. 4, pp. 273-282, 2009.

[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] I. Myrtveit, 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.

[10] Q. Song 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.

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

[12] 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.

[13] 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.

[14] 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.

김 종 찬

2000년 순천대학교 컴퓨터과학과 졸업(이학사)
2002년 순천대학교 대학원 컴퓨터과학과 졸업(이학석사)
2007년 순천대학교 대학원 컴퓨터과학과 졸업(이학박사)
2013년 서울대학교 자동화 시스템 연구소 선임연구원
관심분야 : 영상 처리, HCI, 컴퓨터그래픽스, 기계학습, 데이터 분석 및 예측

심 춘 보

1996년 2월 전북대학교 컴퓨터공학과(공학사)
1998년 2월 전북대학교 컴퓨터공학과(공학석사)
2003년 2월 전북대학교 컴퓨터공학과(공학박사)
2005년 3월~현재 순천대학교 정보통신·멀티미디어공학부 교수
관심분야 : 빅데이터 시스템, 머신러닝, IoT/IoE 플랫폼, 멀티미디어

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 안동대학교 창의융합학부 조교수
관심분야 : 최적화 알고리즘, 강화학습, 블록체인