The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 21, No. 1, pp.181-187, Feb. 28, 2021. pISSN 2289-0238, eISSN 2289-0246

https://doi.org/10.7236/JIIBC.2021.21.1.181
JIIBC 2021-1-23

기상 데이터에서 대기 오염도 요소의 결측치 보완 기법 제안

Proposal to Supplement the Missing Values of Air Pollution Levels in Meteorological Dataset

조동철*, 한희일**

Dong-Chol Jo*, Hee-Il Hahn**

요 약 최근 들어 대기오염으로 인한 피해를 줄이기 위하여 다양한 대기오염 요소를 측정, 분석하고 있다. 이 과정에서 다양한 원인으로 인하여 적지 않은 결측치가 발생한다. 이를 보완하기 위해서는 방대한 크기의 학습 데이터를 필요로 한다. 본 논문에서는 적은 양의 학습 데이터를 이용하여 오존, 이산화탄소, 초미세먼지 등을 측정하는 과정에서 발생하는 결측치를 효과적으로 보완하는 통계적 기법을 제안한다. 제안 알고리즘은 우선 기상데이터와 대기오염도 요소 간의 상관관계, p-값 등의 통계정보 분석을 통해 결측치 보완에 긍정적인 영향을 줄 것이라 예상되는 기상데이터 그룹을 추출한 다음, 이들을 분석하여 효율적이고 효과적으로 결측치를 보완하는 기법이다. 제안 알고리즘의 성능을 확인하기 위하여 다양한 실험을 통하여 널리 알려진 대표적인 알고리즘들과 그 특성을 비교분석한다.

Abstract Recently, various air pollution factors have been measured and analyzed to reduce damages caused by it. In this process, many missing values occur due to various causes. To compensate for this, basically a vast amount of training data is required. This paper proposes a statistical techniques that effectively compensates for missing values generated in the process of measuring ozone, carbon dioxide, and ultra-fine dust using a small amount of learning data. The proposed algorithm first extracts a group of meteorological data that is expected to have positive effects on the correction of missing values through statistical information analysis such as the correlation between meteorological data and air pollution level factors, p-value, etc. It is a technique that efficiently and effectively compensates for missing values by analyzing them. In order to confirm the performance of the proposed algorithm, we analyze its characteristics through various experiments and compare the performance of the well-known representative algorithms with ours.

Key Words : Air Pollution, Meteorological Data, Missing Value, Statistical Inference, Statistical Hypothesis

*학생회원, 한국외국어대학교 공과대학 정보통신공학과
**정회원, 한국외국어대학교 공과대학 정보통신공학과
접수일자 2021년 1월 12일, 수정완료 2021년 2월 3일
게재확정일자 2021년 2월 5일

Received: 12 January, 2020 / Revised: 3 February, 2020 / Accepted: 5 February, 2021
*Corresponding Author: hihahn@hufs.ac.kr
Dept. Information and Communications Eng., College of Engineering, Hankuk University of Foreign Studies, Korea

Ⅰ. 서 론

최근 들어 대기오염으로 인한 피해가 증가함에 따라 대기오염도 예측을 위한 다양한 연구[1, 2, 3]가 진행되고 있다. 그 중에서 오존(O_3), 이산화탄소 (CO_2), 초미세먼지 등과 같은 대기오염도 요소의 예보 시스템에서 결측치 처리는 주요 관심사 중 하나이다. 일반적으로 데이터에서 특정 정보가 누락된 것을 결측치라고 부르며 대기오염도 요소에서의 결측치는 센서의 오작동, 고장, 사람의 실수 등 다양한 원인에 의하여 발생한다. 이 때 발생한 결측치로 인한 불완전한 자료는 분석, 예측 모델에서의 편향된 모수 추정과 같은 문제로 성능이 떨어지거나 예측 자체가 안 되는 등의 문제가 발생할 수 있다. 이러한 이유로 결측치 처리는 데이터 분석 및 예측 등에 있어서 반드시 이루어져야 하는 전처리 과정이다.

결측치 보완과 관련한 연구는 크게 통계적 기법[4, 5, 6]과 딥러닝 기법[7, 8, 9] 등으로 양분된다. 최근들어 딥러닝 기술의 급속한 발전으로 결측치 보완을 위한 많은 연구가 딥러닝 기법을 사용하는데 그 이유는 변수 간 상관관계를 파악하여 높은 성능을 보이기 때문이다. 이 기법은 딥러닝의 특징 추출 능력을 통해 시간상의 상관관계와 변수 간의 상관관계를 동시에 고려하여 결측치 문제를 해결하며, 오랜 시간 상의 상관관계를 고려할수록 다변량 시계열 데이터의 결측치 처리 성능이 향상되는 것으로 알려져 있다. 하지만 딥러닝을 이용한 결측치 보완 기법은 다음과 같은 고질적인 문제가 내재한다. 우선 데이터가 부족한 환경에서 정상적으로 작동하기 어렵고 성능향상을 위해서는 딥러닝 알고리즘을 학습하기 위한 방대한 학습 데이터가 구축되어야 한다. 두 번째로는 모델들의 학습시간이 오래 걸리므로 분석 모델의 결과를 빠르게 반영하여야 하는 스트림 환경에서는 사용이 비효율적이다는 한계점을 보인다. 특히, 본 논문에서 이용하는 대기 오염도 요소는 시계열 데이터이면서 레이블이 존재하지 않는다. 따라서 딥러닝의 분류모델을 사용하기 어려운 문제가 있다. 이러한 이유로 본 논문에서는 통계적 기법을 채택한다.

통계적 기법 중에서 가장 간단한 단순 대치법은 발생한 결측치를 단순히 제거하거나 평균값으로 보완하는 방식인데, 통계적 추론에 사용된 통계량의 효율성 및 일치성 등을 부분적으로 보완해준다는 특징이 있다. 이 때 발생하는 추정량 표준오차의 과소추정 또는 계산의 난해성 문제를 보완한 다중 대치법[10, 11, 12]은 단순 대치법을 여러 번 수행하여 여러 개의 결측치 보완 자료를 얻어 분석하는 방법이다. 다중 대치법의 과정에는 표준적 통계분석이 존재하는데 이는 딥러닝에서의 변수 간 상관관계와 기본적으로 다르다. 딥러닝 기법이 변수 간 상관관계를 파악하는 과정을 통해 높은 성능을 보이기 때문에 통계적 분석[13, 14]에서도 변수 간 상관관계를 파악하는 과정을 거친다면 성능을 향상시킬 수 있을 것이다. 그리고 통계적 기법을 사용하는데 있어서 단순히 변수 간 상관관계를 따지는 것 이외에 p-값, 자기상관관계 (Autocorrelation Function: ACF), 편자기상관관계 (Partial Autocorrelation Function: PACF) 분석 등을 통해 그룹을 나누어 이용하면 딥러닝의 잇점을 챙기면서 위에서 설명한 딥러닝의 고질적인 문제를 피할 수 있다.

본 논문에서는 대기오염도 요소에서 발생하는 결측치를 보완하기 위해 기상 데이터를 이용한다. 제안하는 방식은 우선 대기 오염도 요소와 기상 데이터의 상관도, 계절성 등을 파악하여 기상데이터를 여러 그룹으로 나눈 후 관련성이 높은 그룹과 낮은 그룹을 선정한다. 그런 다음 선정된 그룹의 자료를 이용하여 대기오염도 요소의 결측치를 보완한다. 이 때, 선정된 그룹에 따라 관련성이 높고 신뢰도가 높은 그룹에서는 결측치 보완의 성능이 높고 관련성이 낮고 신뢰도가 낮은 그룹에서는 결측치 보완의 성능이 낮을 것이라는 가설을 세운다. 이후 제시한 기법을 적용하여 대기 오염도 데이터에서 발생한 결측치를 추정하고 널리 알려진 알고리즘들과 성능을 비교분석한다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 관련연구를 리뷰하고, 제안기법을 Ⅲ장에서 설명한다. Ⅳ장에서는 기존의 대표적인 알고리즘들인 Amelia, Mice, MissForest 등과 제안기법의 성능에 대하여 실험결과를 기반으로 비교분석한다. 마지막으로 Ⅴ장에서는 결론을 맺고 향후 연구 진행방향에 대하여 논의한다.

Ⅱ. 관련 연구

이 절에서는 결측치 보완을 위한 통계적 기법을 중심으로 설명한다. 단순 대치법은 결측치를 가진 자료분석에 적용하기 용이하고, 통계적 추론에 사용된 통계량의 효율성과 일치성 문제를 부분적으로 보완해준다. 하지만 추정량 표준오차의 과소추정 또는 계산의 난해성 등의 문제가 있기 때문에 이러한 단순 대치법의 문제를 보완한 다중 대치법을 이용한다.

1. 다중 대치법

다중 대치법은 대치, 분석, 결합의 세 단계로 구성된다. 대치 단계에서 결측 메카니즘을 가정하고 대치 모형을 결정한 후 m 개의 데이터를 생성한다. 분석 단계에서는 대치 단계에서 생성된 m 개의 데이터를 각각 표준적 통계 분석을 통해 추정량과 분산을 계산한다. 마지막 결합 단계에서 분석단계에서 생성된 m 개의 추정량과 분산을 결합하여 통계적 추론을 한다. 대표적으로 MICE (Multiple Imputation by Chained Equations)[10], MissForest[11], Amelia[12] 등의 기법이 이에 속한다.

MICE[10]는 연쇄 방정식에 의한 다변량 대치를 뜻한다. 이 패키지는 다변량 결측 데이터에 대하여 여러 대치(값)을 작성한다. 연속, 이진, 정렬되지 않은 범주 형과 정렬된 범주형 데이터의 혼합을 대치할 수 있고 연속 2단계 데이터를 대치하고 수동 대치를 통해 대치 간에 일관성을 유지할 수 있다. MissForest[11]는 랜덤한 트리를 얻기 위하여 데이터를 부트스트랩(주어진 데이터 셋에서 무작위로 중복을 허용하여 n 개의 데이터를 선택)하여 포레스트(forest)를 구성하며 전체 데이터를 학습시키는 대신 샘플의 결과물을 각 트리의 입력 값으로 이용하여 학습한다. 또한 파티션을 나눌 때 변수에 무작위성을 부여하여 변수 중 일부만 선택하고 그 중에서 최적의 변수를 선택하는 앙상블 기법(ensemble learning)을 이용한다. Missforest는 결측 여부가 해당 변수의 값에 의해 결정되는 결측치를 보완할 때 특히 적합하다[15].

Amelia[12]는 EM(Expectation Maximization) 알고리즘을 부트스트랩하여 결측치를 보완하기 위한 데이터를 생성한다. 이후 분석과정을 거치고 결측치가 존재하는 원본 데이터와 병합한다. EM 알고리즘은 관측되지 않은 잠재변수(결측치)가 있는 자료에서 최대우도(Maximum Likelihood)나 최대사후확률(Maximum a Posteriori)을 갖는 매개변수를 찾는 반복적인 알고리즘이다. 이 기법은 로그-가능도 함수를 결측치에 대한 가중평균으로 기댓값을 구하는 E 단계와, 그렇게 가중평균이 된 로그-가능도 함수를 최대화하는 모수치의 값을 찾는 M 단계로 구성된다. 두 단계를 로그 가능도 함수의 값이 유의미하게 변화하지 않을 때까지 반복 수행한다.

2. 통계적 분석

통계분석의 방법 중 본 논문에서는 추론통계 기법을 이용한다. 추론통계란 단순히 숫자를 요약하는 것이 아닌 어떤 값이 발생할 확률을 계산하는 통계기법이다. 이는 수집한 데이터 간의 관계가 우연히 발생했는지에 대한 확률을 알아내는 방법으로 우연히 나타날 확률이 낮다면 통계적으로 유의하다(statistically significant)라고 말하고 반대로 우연히 나타날 확률이 크다면 통계적으로 유의하지 않다라고 결론을 내린다.

유의확률을 이용해 가설을 검정하는 방법을 통계적 가설 검정(statistical hypothesis test)라 한다. 유의확률(significance probability, p-value)이란 집단 또는 데이터 간 우연히 차이가 있는 데이터가 발생할 확률이다. 즉, 유의확률이 작다는 것은 우연이라고 보기 힘들다는 뜻이고 유의확률이 크다는 것은 우연일 가능성이 높다는 뜻이다. p-값(p-value)으로 통계적 유의도를 판단하는데, 일반적으로 p-값이 0.05미만일 경우 통계적으로 유의하다고 판단한다.

기상 데이터 간 상관관계와 계절성 등을 파악하여 기상 데이터를 그룹으로 나누고, 나눈 그룹을 이산화탄소, 오존, 초미세먼지 등의 대기오염도 요소의 결측치를 통계기법을 적용하여 추정한다. 나눈 그룹을 대기오염도 요소에 적용할 때 상관도와 계절성을 파악하여 적용하여야 하는데, 상관도를 파악하기 위해 피어슨 상관계수를 사용하고, 피어슨 상관계수의 신뢰도를 측정하기 위해 다음과 같은 p-값을 이용한다.

$$p-value = 1-erf\left(\frac{|z|}{\sqrt{2}}\right) \tag{1}$$

여기서 $z=\frac{\overline{X}-\mu}{\sigma}$ 인데, $\overline{X}$는 표본평균이고 μ와 σ는 모집단 분포의 평균과 표준편차를 의미한다. 그리고 $erf(x)$는 다음과 같이 정의된다.

$$erf(x)=\frac{2}{\sqrt{\pi}}exp\left(\int_0^x(-t^2)dt\right) \tag{2}$$

위 식에서 x는 관측된 값이고 p-값은 x보다 더 큰 값이 관측될 확률을 의미한다. p-값은 검정 통계량에 대한 유의 확률을 나타내는데, 귀무가설의 신뢰구간을 벗어나는 확률로 기각역보다 유의 확률이 작아야 귀무가설을 기각할 수 있다. 본 논문에서는 단측 검정을 기준으로 유의수준은 0.05로 정한다.

변수 간의 관계성을 분석하기 위하여 상관분석(Correlation Analysis)를 이용한다. 상관분석이란 연속변수 간 관련이 있는지 검정하는 통계기법이다. 이 때 상관계수(Correlation Coefficient)를 통해 변수 간 관련도를 파악하는데[16], 본 논문에서는 연속형 자료와 연속

형 자료 간의 선형관계를 나타내는 다음과 같은 피어슨 상관계수를 채택한다.

$$p = \frac{\sum_i (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2 (y_i - \mu_y)^2}} \quad (3)$$

위 식에서 x_i와 y_i는 변수 x, y의 i번째 표본을, μ_x와 μ_y는 x와 y의 평균을 각각 나타낸다. 계절성, 정상성 등을 파악하는 과정은 통계적 분석방법 중 ACF와 PACF 기법을 이용한다. ACF는 자기상관성을 파악하기 위한 함수로, 동일한 변수라도 어떤 시차를 가지고 스스로와 유사한 지 알아내는데 목적이 있다. 시차 k의 ACF는 다음 식으로 정의된다.

$$\rho_k = \frac{\gamma_{t,t-k}}{\sqrt{\gamma_{t,t}\gamma_{t-k,t-k}}} \quad (4)$$

여기서 $\gamma_{t,s} = E[(Y_t - \mu_t)(Y_s - \mu_s)]$이고 μ_t와 μ_s는 각각 랜덤 프로세스 Y_t와 Y_s의 평균을 나타낸다.

PACF는 자기상관성을 파악하되, Y_t와 Y_{t-s} 사이에 있는 $Y_{t-1}\cdots, Y_{t-(k-1)}$의 영향을 제거하고 오직 둘 사이의 관계를 파악하는데 목적이 있다. PACF는 다음과 같이 정의되는데,

$$\phi_k = \frac{E[(e_t - \bar{e}_t)(e_{t-k} - \bar{e}_{t-k})]}{\sqrt{var(e_t)var(e_{t-k})}} \quad (5)$$

여기서 e_t와 e_{t-k}는 각각 다음 식으로 주어진다.

$$e_t = Y_t - \alpha_1 Y_{t-1} - \cdots - \alpha_{k-1} Y_{t-(k-1)} \quad (6)$$
$$e_{t-k} = Y_t - \beta_1 Y_{t-1} - \cdots - \beta_{k-1} Y_{t-(k-1)}$$

Ⅲ. 제안 기법

기상데이터의 요소 간 상관계수, p-값, ACF, PACF 등의 분석을 통해 상관성, 계절성, 정상성 등에 따라 데이터 그룹을 나눈다. 표 1은 본 논문에서 이용한 기상 데이터를 이용하여 구한 데이터 그룹을 보여 준다. 각 그룹 내의 요소들은 상호 간에 강한 양의 상관관계를 보여준다.

표 1. 기상데이터 요소 간 상관관계, p-값에 따른 데이터 그룹 분류

Table 1. Division into groups of the elements in weather dataset according to the correlation and the p-values among their elements.

그룹 A	그룹 B	그룹 C	그룹 D
평균기온	평균 현지기압	합계 일조시간	일 강수량
평균 이슬점 온도	평균 해면기압	합계 일사량	평균풍속
평균 증기압			평균 상대습도
평균 지면온도			

기상 데이터를 구성하는 평균기온, 일 강수량, 평균 풍속, 평균 이슬점 온도 등, 11 가지 요소에 대해 상관관계와 p-value 값을 분석하는데 피어슨 상관계수와 단측검정을 기준으로 유의수준을 0.05로 정하여 분석한다. 표 2는 각 그룹 간의 상관관계를 보여준다.

표 2. 기상데이터 그룹 간 상관관계

Table 2. Correlation among the elements in weather dataset.

	그룹 A	그룹 B	그룹 C
그룹 A	strong positive	strong negative	weak positive
그룹 B	strong negative	strong positive	weak negative
그룹 C	weak positive	weak negative	strong positive

표 2에 의하면 그룹 A는 그룹 B와 강한 음의 상관관계를 가지나 그룹 C와는 약한 양의 상관관계를 갖는다. 표에는 제시되어 있지 않으나, 그룹 D는 다른 세 그룹과의 상관관계가 일정하지 않다. 또한, 그룹 A와 그룹 B는 계절성의 강도가 높지만 그룹 D는 낮고 그룹 C는 중간 정도의 강도를 갖는다. 다음 단계로, 대기오염도 요소와 기상데이터 그룹 간의 관계를 파악하여 어떤 그룹과 상관도가 높은지를 파악한다. 예를 들어, 오존은 그룹 A, B, C에서 모두 낮은 p-값과 높은 상관계수를 갖는 양상을 보이므로, 이들 세 그룹을 이용하여 결측치를 보완한다. 이산화탄소와 초미세먼지 등에도 동일한 방법을 적용함으로써 높은 성능을 기대할 수 있다.

그리고 상관성, 계절성, 정상성 등에 따라 나눈 그룹과 대기오염도 요소의 관계를 파악하여 특정관계가 결측치 추정과정에 긍정적인 영향을 주는지 여부를 파악한다. 제

안 알고리즘은 다음과 같이 네 개의 과정에 따라 동작한다.

과정 1: 오존, 이산화탄소, 초미세먼지 등, 각 요소에 높은 성능을 보이는 그룹과 요소들 간의 선형회귀모델을 구한다.

과정 2: 선형회귀모델의 그래프와 오존, 이산화탄소, 초미세먼지의 잔차와 기울기를 비교하여 잔차가 가장 낮고, 기울기가 큰 선형회귀 모델의 그래프를 추출한다. 추출한 그래프가 정규분포를 따르는지 확인한다.

과정 3: 스케일-위치 그래프를 통해 표준화 잔차를 구하여 이상치(outlier)를 파악한다. 여기서 과정 2에서 추출한 그래프가 정규분포 요건을 충족하지 못하거나 이상치 수치가 높으면 제외한다.

과정 4: 위 과정이 끝난 후 선형회귀모델의 그룹을 선정하여 오존, 이산화탄소, 초미세먼지의 결측치 보완에 사용하고 다중선형회귀분석을 통해 결측치를 보완한다.

과정 4에서는 선형회귀와 다중 선형회귀를 함께 이용한다. 회귀모형에 포함되는 예측변수는, 종속변수와는 상관관계가 높지만 예측 변수들 간에는 낮은 상관관계를 갖도록 선정된다. 선정된 예측변수와 종속변수들 간에 선형회귀를 각각 적용한 후, 상관계수와와 변화율이 크고 잔차가 작은 순서로 두 개의 예측변수를 최종 선정하여 다중선형회귀에 적용함으로써 결측치를 보완한다.

Ⅳ. 실험 결과

본 논문에서는 2016년 1월 1일부터 2018년 12월 31일까지 3 년간 서울시 강남구의 평균 대기오염도 정보를 이용하여 다음과 같은 데이터셋을 구축한다. 대기오염도 요소 중 오존, 이산화탄소, 초미세먼지를 선정하여 사용하는데, 그 데이터는 총 3,288개로 구성된다. 그리고 동일한 장소와 동일한 기간 동안 기상청에서 공개한 종관기상관측 자료로부터 다음과 같은 데이터셋을 추출하여 구성한다. 이 데이터셋은 평균기온, 일 강수량, 평균 풍속, 평균 이슬점 온도, 평균 상대습도, 평균 증기압, 평균 현지 기압, 평균 해면 기압, 합계 일조시간, 합계 일사량, 평균 지면온도 등을 포함하며 총 12,056개로 구성된다. 대기오염도 요소와 기상 데이터의 결측치 수와 그 비율은 표 3과 표 4에 각각 제시한다.

표 3. 대기오염도 요소의 결측치 수 및 비율
Table 3. The numbers of missing values and their rates in the elements of the air pollution levels.

	오존	이산화탄소	미세먼지
결측치 갯수	58	62	57
결측치 비율(%)	5.2	5.6	5.2

실험을 위하여 오존, 이산화탄소, 초미세먼지 데이터에서 각각 3%, 10%, 20%의 결측치를 무작위로 발생시킨다. 표 1에 제시한 네 개의 그룹 중에서 그룹 D는 다른 세 그룹과의 상관관계가 불규칙적이므로 그룹 D를 제외한 나머지 그룹으로 실험군을 구성한다. 실험군은 오존, 이산화탄소, 초미세먼지와 비교적 높은 상관계수와 낮은 p-값을 보인다.

표 4. 기상데이터 요소의 결측치 수 및 비율
Table 4. The numbers of missing values and their rates in the elements of the weather dataset.

	일 강수량	평균 풍속	합계 일조시간	합계 일사량
결측치 갯수	682	3	9	9
결측치 비율(%)	62.2	0.27	0.82	0.82

표 5. 결측치 비율에 따른 오존에 대한 결측치 보완 성능 비교 (단위 : ppm)
Table 5. Performances of supplementing missing values for the ozon data when the MICE, MissForest, Amelia and our proposed model are applied. (Unit : ppm)

결측치비율	MICE	MissForest	Amelia	제안 알고리즘
3%	0.001392	0.000098	0.003873	0.001454
10%	0.005433	0.006335	0.012824	0.006059
20%	0.012890	0.011750	0.025368	0.015699

제안 기법의 성능을 확인하기 위하여 Amelia, Mice, MissForest 알고리즘의 결과와 비교한다. 이 때, 평균제곱근오차(Root Mean Square Error : RMSE)를 평가지표로 이용한다. 표 5는 오존 데이터에 대하여 결측치 보완을 수행한 다음 실측 정보(ground-truth)와의 RMSE 계산 결과를 보여준다. 마찬가지로 이산화탄소, 초미세먼

지에 대하여 동일한 실험을 수행하였을 때 RMSE 성능은 각각 표 6과 표 7에 제시한다. 결측치 비율에 따라 성능 차이가 다르게 나타나지만 널리 알려진 대표적인 세 알고리즘들에 비하여 전체적으로 대등한 성능을 보여주고 있음을 알 수 있다.

표 6. 결측치 비율에 따른 이산화탄소에 대한 결측치 보완 성능 비교 (단위 : ppm)

Table 6. Performances of supplementing missing values for the dioxide data when the MICE, MissForest, Amelia and our proposed model are applied. (Unit : ppm)

결측치비율	MICE	MissForest	Amelia	제안 알고리즘
3%	0.3472	0.5636	0.8371	0.6734
10%	0.9908	0.9923	1.6668	1.2658
20%	2.6636	3.3707	5.9335	3.4583

표 7. 결측치 비율에 따른 초미세먼지에 대한 결측치 보완 성능 비교 (단위 : ㎍/㎥)

Table 7. Performances of supplementing missing values for the ultra fine dust data when the MICE, MissForest, Amelia and our proposed model are applied. (Unit : ppm)

결측치비율	MICE	MissForest	Amelia	제안 알고리즘
3%	7663.64	4796.168	8678.834	7838.325
10%	22245.64	16128.5	33865.63	30322.48
20%	58121.48	42019.25	98630.61	78347.91

V. 결 론

본 논문에서는 기상데이터들을 분석하여 상관도, 신뢰도, 계절성 등에 따라 네 개의 그룹으로 나누고 대기 오염도 요소와 기상 데이터 그룹 간의 관계를 파악하여 통계적으로 결측치를 구하는 알고리즘을 제안하였다. 제안 알고리즘은 딥러닝 기법을 이용하는 알고리즘들에 비하여 월씬 적은 양의 학습 데이터로 모수를 추정할 수 있는 장점이 있다. 제안 알고리즘의 성능을 확인하기 위하여 R에서 제공하는 대표적인 결측치 보완기법인 MICE, MisForest, Amelia 등과 RMSE 관점에서 성능을 비교하였다. 결측치 비율에 따라 성능차이가 다르게 나타나지만 기존의 대표적인 세 알고리즘들에 비하여 전체적으로 대등한 성능을 보여주고 있음을 확인하였다.

향 후에는 결측치 추정의 정확도를 향상시키기 위하여 기상 데이터와 대기 오염도 간의 상호 관련성을 보다 세분화하여 구하는 알고리즘 개발을 위한 연구를 계속 진행할 계획이다.

References

[1] Jihoon Yoo, Dongil Shin, Dongkyoo Shin, "Prediction System of Fine Particulate Matter Concentration Index by Meteorological and Air Pollution Material Factors Based on Apache Spark and Machine Learning", SoICT 2019: Proceedings of the Tenth International Symposium on Information and Communication Technology, pp 479-485, December 2019.
DOI: https://doi.org/10.1145/3368926.3369684

[2] Kyunghak Cho, Byoung-young Lee, Myeongheum Kwon, and Seogcheol Kim, "Air Quality Prediction Using a Deep Neural Network Model", Journal of Korean Society for Atmospheric Environment, Vol.35, No.2, pp 214-225, April 2019.
DOI: https://doi.org/10.5572/KOSAE.2019.35.2.214

[3] Taehee Kim, Jisu Myoung, Yonghee Lee, Insuk Suh, and Limsuk Jang, "A Study on Influence of Meteorological Patterns on Data Assimilation Effect Using the Air Quality Prediction Model", Journal of Korean Society for Atmospheric Environment, Vol. 35, No. 1, pp 49-59, Feb 2019.
DOI: https://doi.org/10.5572/KOSAE.2019.35.1.049

[4] Myungsik Jeon and Seungjun Shin, "Comparison Study for Missing Imputation Methods: A Focus on Canonical Discriminant Analysis", Journal of the Korean Data Analysis Society, Vol. 9, No. 2, pp. 673-685, Sep. 2007.
UCI: G704-000930.2007.9.2.023

[5] Sung Cheol Yun, "Imputation of Missing Values", Journal of Preventive Medicine and Public Health, Vol. 37, Issue 3, pp 209-211, 2004.
URI:https://www.jpmph.org/journal/view.php?number=1772

[6] Yeonjin Kim, Heonjin Park, "Comparison of Missing Imputation Methods In Fine Dust Data", Journal of Big Data, Vol. 4, No. 2, pp. 105-110, 2019.
DOI: https://doi.org/10.36498/kbigdt.2019.4.2.105

[7] YoungJun Lee, Susik Yoon, and Jae-Gil Lee, "A Survey on Handling Missing Values in Multivariate Time Series Data Using Deep Learning", Communications of the Korean Institute of Information Scientists and Engineers, Vol. 35, No. 3, pp. 54-65, 2019.
URI: https://koasas.kaist.ac.kr/handle/10203/271398

[8] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, DWavid Sontag, and Yan Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values", Scientific Reports 8, 6085, 2018.

DOI: https://doi.org/10.1038/s41598-018-24271-9

[9] Joon-Mook Lim, "An Estimation Model of Fine Dust Concentration Using Meteorological Environment Data and Machine Learning", Journal of Information Technology Services, Vol. 18, Issue 1, pp. 173-186, Mar. 2019.
DOI: https://doi.org/10.9716/KITS.2019.18.1.173

[10] Patrick Royston and Ian R. White, "Multiple Imputation by Chained Equations (MICE): Implementation in Stata", Journal of Statistical Software, Vol. 45, Issue 4, Dec. 2011.
DOI: https://doi.org/10.18637/jss.v045.i04

[11] Daniel J.Stekhoven and Peter Bühlmann, "MissForest Nonparametric Missing Value Imputation for Mixed-Type Data", Bioinformatics, Vol. 28, Issue 1, pp. 112-118, 1 Jan. 2012.
DOI: https://doi.org/10.1093/bioinformatics/btr597

[12] James Honaker, Garay King, and Matthew Blackwell, "Amelia 2: A Program for Missing Data", Journal of Statistical Software, Vol. 45, Issue 7, Dec. 2011.
DOI: https://doi.org/10.18637/jss.v045.i07

[13] Yeonsu Lee and Dongwook Son, "An Analysis of the Relationships between the Characteristics of Urban Physical Environment and Air Pollution in Seoul", Journal of the Urban Design Institute of Korea, Vol. 16, No. 3, pp. 5-19, Jun. 2015.
UCI: G704-001620.2015.16.3.002

[14] SooHyun Eom, Chul-Ghi Lee, and Wookey Lee, "An Analysis of Relation between Climate and Air Pollution in Incheon", Journal of Information Technology and Architecture, Vol. 12. No. 2, pp. 271-276, Jun. 2015.
UCI: G704-SER000010357.2015.12.2.004

[15] Gyoo Seok Choi, Jong jin Park, Ha-Nam Nguyen, "A Study on Feature Selection Method based on Random Forest for Cancer Diagnosis System", Journal of Korean Institute of Information Technology, Vol. 6, No. 3, pp. 53-61, 2008.
UCI: G704-001947.2008.6.3.004

[16] Han Sang-Seok, "The Correlation Study of Factors which effect on Insulation Reliability of Inverter Motor Using Statistical Analysis", Journal of Korea Academia-Industrial cooperation Society, Vol. 11, No. 4, pp.1216-1221, 2010.
UCI: G704-001653.2010.11.4.044

저 자 소 개

조 동 철(학생회원)

- 2015년 한국외국어대학교 정보통신공학과 입학.
- 2021년 한국외국어대학교 정보통신공학과 학사 졸업예정.
- 주관심분야 : 신호처리, 컴퓨터비전, 머신러닝

한 희 일(정회원)

- 1984년 서울대학교 제어계측공학과 학사 졸업.
- 1988년 서울대학교 제어계측공학과 석사 졸업.
- 1995년 University of Arizona 전기및컴퓨터공학과 박사 졸업.
- 2019년 현재 한국외국어대학교 정보통신공학과 교수.
- 주관심분야 : 신호처리, 컴퓨터비전, 머신러닝, 기하 및 위상

※ 본 논문은 2020년도 한국외국어대학교 교내 학술연구지원에 의하여 연구되었음