# 기상 데이터를 활용한 LSTM 기반 미세먼지 농도 예측 방법 비교

Comparison of LSTM-based Fine Dust Concentration Prediction Method using Meteorology Data

저자
(Authors)

서양모, 염재홍
Seo, Yang-Mo, Yom, Jae-Hong

출처
(Source)

한국측량학회 학술대회자료집 , 2019.4, 117-120(4 pages)

발행처
(Publisher)

한국측량학회
Korea Society of Surveying, Geodesy, Photogrammetry, and Cartography

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09305058

APA Style

서양모, 염재홍 (2019). 기상 데이터를 활용한 LSTM 기반 미세먼지 농도 예측 방법 비교. 한국측량학회 학술대회 자료집, 117-120

*Korean Society of Surveying, Geodesy, Photogrammetry, and Cartography 2019 / Seoul / Korea*

# 기상 데이터를 활용한 LSTM 기반 미세먼지 농도 예측 방법 비교

# Comparison of LSTM-based Fine Dust Concentration Prediction Method using Meteorology Data

서양모[1] · 염재홍[2]
Seo, Yang-Mo · Yom, Jae-Hong

1) 세종대학교 대학원 지구정보공학과 석사과정 (E-mail:yangmoseo@gmail.com)
2) 교신저자 · 정회원 · 세종대학교 공과대학 환경에너지공간융합학과 교수 (E-mail:jhyom@sejong.ac.kr)

**초 록**

미세먼지로 인한 대기오염이 심각해짐에 따라 정책의 수립 및 시행을 위해 정확한 미세먼지 농도를 예측하는 것이 중요해지고 있으며, 다양한 방법을 이용한 미세먼지 오염도 예측 연구들이 수행 되고 있다. 본 연구에서는 에어 코리아의 대기오염측정망에서 측정한 PM10의 이력 데이터와 기상청에서 제공하는 지역별상세관측자료(AWS) 데이터를 활용하여 LSTM(Long Short Term Memory)모델의 입력 시퀀스 형태에 따른 다양한 미세먼지 농도 예측 모델을 구성하고, 각 모델 별 학습을 통하여 최소의 오차(RMSE, MAPE)를 갖는 파라미터를 선정, 각 모델의 예측 결과를 비교하였다.

핵심어 : LSTM, 미세먼지, 기상데이터

## 1. 서 론

연중 미세먼지로 인한 주의보 및 경보 발령 횟수와 지속 기간 그리고 그 강도가 증가함에 따라 미세먼지가 개인의 건강에 대한 위협을 넘어 일상생활에 지장을 주기 시작 하였다. 이에 미세먼지에 대한 사회 구성원의 관심도가 증가하고 미세먼지로 인한 대기오염을 해소가 현대사회의 주요한 문제로 대두되고 있다. 이로 인해 대기 중 미세먼지를 줄이기 위한 정책의 수립 및 시행을 위해 미세먼지 농도를 정확하게 예측하는 것이 중요시 되고 있으며 다양한 분야에서 연구가 수행되고 있다. 전송완(2017)은 대구지역의 대기질 측정정보(NO2, SO2, CO, O3, PM10)와 기상정보(평균기온, 최고기온, 최저기온, 일 강수량, 평균풍속, 최대풍속)의 일 단위 평균을 구하고, MLR(Multi-linear Regression), SVR(Support Vector Regression), ARIMA(Auto Regressive Integrated Moving Average), ARIMAX(Auto Regressive Integrated Moving Average with Exogeneous Input)을 사용하여 예측 성능을 RMSE(Root Mean Square Error)로 비교 하였다. 임준묵(2018)은 기상관측 데이터 중 다중 회귀분석을 통해 미세먼지 농도에 영향을 미치는 항목을 추출하고, ANN(Artificial Neural Network)과 SVM(Support Vector Machine)을 사용하여 미세먼지 농도를 매우나쁨, 나쁨, 보통, 좋음 의 4가지 범주로 예측하였다. Park et al.(2017)은 2005년부터 2016년 까지의 PM10 관측데이터를 수집하여 노이즈 제거를 위해 전 처리 하고 이를 30일 단위의 시퀀스(sequence)데이터로 변환하여 LSTM(Long Short Term Memory)모델을 학습시

키고 미세먼지 농도를 예측하여 기존의 Linear Regression, RNN과의 성능을 MSE(Mean Square Error), RMSE(Root Mean Square Error)를 통해 비교 하였다.

본 연구에서는 LSTM을 이용한 기상데이터 및 PM10 이력 데이터 기반의 미세먼지 농도 예측 모델을 제안한다. 일단위의 PM10 이력 데이터와 기상데이터를 시퀀스 데이터로 변환하고 시퀀스 데이터의 형태에 따라 싱글 시퀀스, 멀티 시퀀스로 구분하여 이를 LSTM 모델에 입력하여 모델을 학습시킨다. 학습된 모델을 바탕으로 PM10의 농도를 예측하고 각 모델별 예측 성능을 평가하기 위해 RMSE, MAPE값을 비교하였다.

## 2. 연구방법

### 2.1 데이터 수집 및 전처리

미세먼지 데이터는 에어코리아의 대기오염측정망을 통하여 측정된 도심 및 거주 지역의 데이터를 2015년 1월 1일부터 2018년 9월 30일까지 서울지역의 25개 측정소를 대상으로 수집하였다. 수집한 데이터는 미세먼지(PM10) 초미세먼지(PM2.5), 오존(O3), 일산화질소(NO2), 일산화탄소(CO), 아황산가스(SO2)의 측정값으로 구성되어 있으며, 본 연구에서는 미세먼지(PM10)만 필터링 하여 결측치(-999, NaN)를 주변값을 이용하여 보정하고, 일단위 데이터로 통합하여 사용하였다.

기상데이터는 기상자료개방포털에서 제공하는 자동기상관측장비(AWS, Automatic Weather System)를 통하여 측정된 일별 평균기온, 최저기온, 최고기온, 일강수량, 최대순간풍속, 평균풍속, 최대순간풍속풍향 데이터를 2015년 1월 1일부터 2018년 9월 30일까지 서울지역의 29개 측정소를 대상으로 수집하였다. 수집한 데이터는 결측치(-999, NaN)를 주변값을 이용하여 보정하고 일단위로 통합된 미세먼지 데이터와 측정소 설치주소 및 측정날짜를 기준으로 결합하였다.

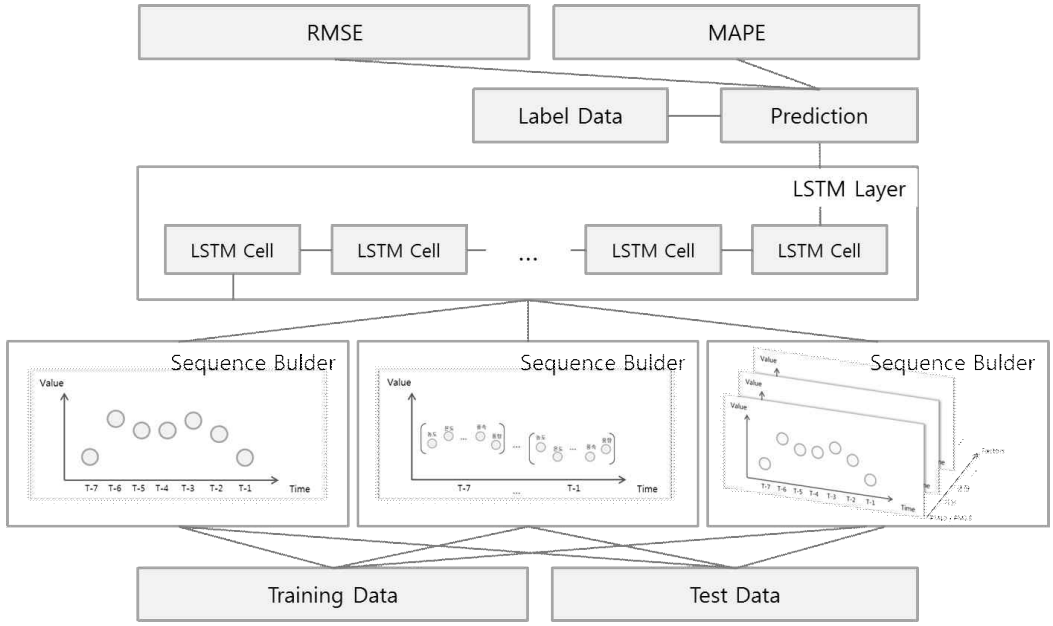
### 2.2 LSTM 모델 구성

RNN(Recurrent Neural Network)은 연속되는 데이터의 모델링에 강점이 있으며, 각 입력 단계에서 학습된 정보가 다음 입력 단계를 거치면서 조금씩 수정되고 데이터 입력이 끝나면 데이터 전체에 대한 정보를 학습하게 된다. LSTM은 RNN의 한 종류로 데이터의 길이가 길어져 학습 및 수정의 단계가 길어지면 처음 입력된 데이터의 정보의 학습 내용이 제대로 반영되지 못하는 장기 의존성문제(Long-Term Dependencies)를 해결한 모델이다 Hochreiter(1997). 본 연구에서는 미세먼지(PM10) 농도 데이터와 같은 시간에 관측한 기상데이터를 이용하여 미세먼지(PM10)를 예측하기 위해 LSTM기반의 모델을 구성 하였으며, 입력되는 미세먼지(PM10) 농도와 기상데이터를 시퀀스 데이터로 변환하는 방법에 따라 그림1과 같이 구분하였다. 시퀀스 데이터의 형태는 미세먼지 농도 예측 대상이 되는 시점을 t로 볼 때 모델의 시퀀스 길이만큼 과거의 미세먼지 농도가 연속되는 형태(1), 미세먼지 농도와 기상정보가 시간 순서대로 연속되는 형태(2), 미세먼지 농도와, 기온, 풍속 등이 각각의 별도의 시퀀스로 연속되는 형태(3)로 구분 하였다.

(1) $X_{sequence} = \{j_1, j_2, \cdots j_{t-1}, j_t\}$, $j_t$=t일전의 PM10 농도

(2) $X_{sequence} = \{j_1, k_1, \cdots j_t, k_t\}$, $j_t$=t일전의 PM10 농도, $k_t$=t일전의 기상정보

(3) $X_{sequence} = \{x_1, x_2, \cdots x_{t-1}, x_t\}$, $x_t = \{j_t, k_t\}$ t일전의 PM10 농도와 기상정보의 시퀀스




[그림 1] 시퀀스 형태에 따른 LSTM 기반 미세먼지 예측 모델

## 3. 연구내용

결합된 미세먼지 및 기상데이터는 측정 항목마다 서로 다른 스케일을 갖고 있으므로 LSTM을 이용한 학습 과정에서 특정 데이터의 큰 숫자에 과도한 영향을 받아 과적합(Over Fitting)하는 현상을 방지하기 위해 0에서 1사이의 값으로 정규화를 수행하였다. 전체 데이터 중 대기오염 측정망과 자동기상관측장비의 설치 위치가 가장 근접한 강남구 데이터를 대상으로 2015년 1월 1일부터 2017년 12월 31일 까지 데이터는 학습을 위한 데이터로 2018년 1월 1일 부터 2018년 9월 30일 까지 데이터는 모델의 정확성을 테스트하기 위한 데이터로 구분 하였다.

학습 데이터를 대상으로 반복 실험을 통하여 급격한 미세먼지 농도 변화 및 차이를 모델링하기에 적합한 시퀀스의 길이를  7일로 고정하였으며 다음날의 미세먼지 농도를 예측하여 학습 오차가 최소화 되는 파라미터를 결정하기 위해 학습 모델의 Epoch과 Hidden Node를 변경하며 RMSE와 MAPE를 표1과 표2와 같이 측정 하였다.

| Sequence Length | Hidden Node | Epoch | RMSE | MAPE |
|---|---|---|---|---|
| 7 | 30 | 40 | 6.442 | 25.477 |
| 7 | 30 | 50 | 3.693 | 15.019 |
| **7** | **30** | **60** | **2.508** | **8.273** |
| 7 | 30 | 70 | 10.097 | 40.725 |

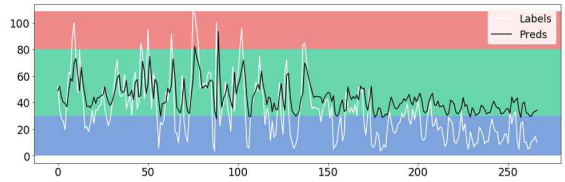
[표 1] Epoch에 따른 모델 성능평가

| Sequence Length | Hidden Node | Epoch | RMSE | MAPE |
|---|---|---|---|---|
| **7** | **30** | **60** | 2.508 | **8.273** |
| 7 | 50 | 60 | 5.772 | 28.237 |
| 7 | **70** | **60** | **2.248** | **11.480** |
| 7 | 100 | 60 | 2.901 | 10.821 |

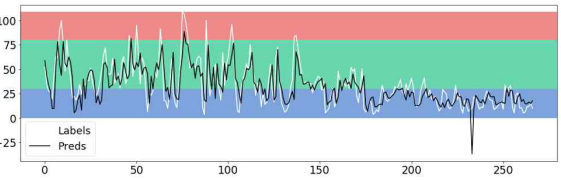
[표 2] Hidden Node에 따른 모델 성능평가

측정결과 모델의 학습 파리미터가 Epoch 60, Hidden Node 30일 때 MAPE가 가정 작게 나타났으며, Epoch 60, Hidden Node 70일 때 RMSE가 가장 작게 나타났다. RMSE가 가장 작은 파라미터 설정으로 시퀀스 형태에 따른 예측 성능을 비교 하였다.

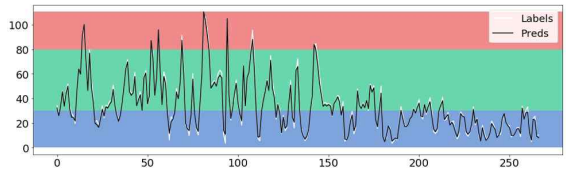
## 4. 결과 분석




[그림 2] LSTM(PM10 Only)




[그림 3] LSTM(PM10+AWS)




[그림 4] LSTM(PM10, AWS Multi sequence)

| Model | RMSE | MAPE |
|---|---|---|
| LSTM(PM10 Only) | 19.604 | 103.080 |
| LSTM(PM10+AWS) | 16.300 | 54.635 |
| LSTM(Multi) | 2.248 | 11.480 |

[표 3] 시퀀스 형태에 따른 예측성능 비교

그림2는 시퀀스 데이터가 길이만큼 과거의 미세먼지 농도가 연속되는 형태일 경우, 그림3은 시퀀스 데이터가 미세먼지 농도와 기상정보가 시간 순서대로 연속되는 형태일 경우, 그림4은 시퀀스 데이터가 미세먼지 농도와, 기온, 풍속 등이 각각의 별도의 시퀀스로 연속되는 형태일 경우로, 실제 대기오염 측정망을 통하여 측정된 PM10 농도와 학습된 모델을 통하여 예측된 다음날의 PM10 농도를 좋음(파란색), 보통(초록색), 나쁨(노란색), 매우나쁨(빨간색)의 범주에서 비교하였다. 표3의 RMSE값을 통하여 입력 데이터를 별도의 시퀀스로 구성하는 방법이 실제값과 예측값의 오차가 작으면서 MAPE값을 통해 오차의 변동폭이 적게 나타남을 확인 할 수 있다. 그림2와 그림4를 비교해보면 그림2는 예측값이 실제 값과 유사한 패턴을 보이지만 고점과 저점에서 차이가 발생하고 차이의 폭이 다양하게 나타난다.

## 5. 결 론

본 논문에서는 자동기상관측장비에서 관측한 기상데이터를 이용하여 PM10 농도를 예측하는 LSTM 기반의 모델을 시퀀스의 형태에 따라 구성하고 비교 하였다. 비교 결과 LSTM 모델에 입력되는 데이터를 다차원의 시퀀스로 구성하여 학습하는 경우 예측성능이 가장 정확하였다. 향후에는 예측 대상 시간을 늘리고 급격한 농도 변화의 예측 정확성을 향상시키기 위하여 주변 측정소의 이력 데이터를 활용하는 연구를 수행 할 예정이다.

## 참고문헌


전송완, et al. (2017), 미세먼지 농도 예측 알고리즘 성능 비교, 2017 한국소프트웨어종합학술대회 논문집, 한국정보과학회, pp. 775-777.

임준묵, et al. (2018), 기상데이터와 머신러닝을 활용한 미세먼지농도 예측 모델, 한국IT서비스학회학술대회논문집 2018, pp. 691-694.

Park, J.-H., et al. (2017), PM10 density forecast model using long short term memory, 2017 Ninth International Conference on Ubiquitous and Future Networks (ICUFN), IEEE, pp. 25-30.

Hochreiter, S. and J. J. N. c. Schmidhuber (1997), Long short-term memory., 9(8): 1735-1780.