기상 인자와 대기오염 인자를 활용한 LSTM 기반의 미세먼지 농도 예측

유지훈*, 신동일*, 신동규*
*세종대학교 컴퓨터공학과
yoojihoon@sju.ac.kr, dshin@sejong.ac.kr, shindk@sejong.ac.kr

LSTM-based Fine Dust Concentration Prediction using Meteorogical factors and Air Pollution factors

Jihoon Yoo*, Dongil Shin*, Dongkyoo Shin*
*Dept. of Computer Engineering, Sejong University

요 약

미세먼지(PM10, PM2.5)는 배출가스 증가와 함께 빠르게 악화되어 왔으며, 다양한 화학성분 뿐만 아니라 금속 성분이 포함되어 있어 인체에 큰 유해성을 발생한다. 이에 정부는 미세먼지 저감 정책 및 법률을 통해 개선하고자 했지만, 2013년부터 그 효력을 잃기 시작하였다. 이에 본 연구에서는 미세먼지 저감 정책 및 법률을 수립하는데 있어 가장 중요한 요소인 미세먼지 농도를 예측하는 연구를 진행한다. 이전 연구들에서 미세먼지 영향 요소들이 시계열 기반의 데이터(기상인자와 대기오염 인자)인 것을 확인하였기에, 시계열 데이터에 좋은 성능을 보이는 LSTM 알고리즘을 사용하여 학습 후, 서울시 '구별' '시간단위' 미세먼지 농도 예측에 대한 예측 오차(RMSE, MAE)성능을 비교하였다. 실험 결과 PM10의 경우 (7.2, 4.78), PM2.5의 경우 (4.7, 3.2)의 예측 오차를 보였으며, 금천구의 경우 PM10이 (5.3, 3.71), PM2.5에서 (3.5, 2.5)로 가장 좋은 성능을 보였다.

1. 서론

대기오염 문제는 자동차 매연, 화석 연료 등과 같은 배출 가스 증가와 함께 빠르게 악화되기 시작했으며, 이중 입자의 직경이 10μm 이하의 일반 미세먼지(PM10)와 2.5μm 이하의 초미세먼지(PM2.5)는 1급 발암물질로 지정될 만큼 심각한 문제로 야기되었다.[1] 이러한 미세먼지는 다양한 화학성분 뿐만 아니라, 금속 성분이 포함되어 있기 때문에 인체의 호흡기나 심장 질환에 큰 유해성을 발생시킨다.[2, 3] 이에 정부는 미세먼지 저감 정책 및 법률과 같은 다양한 방법을 통해 미세먼지 오염도를 개선하였으나, 2013년부터 개선 추세가 정체되었고, 2016년에는 고농도 미세먼지가 자주 발생하여 현재까지 부정적인 영향을 끼치고 있다. 미세먼지가 다시금 사회적 관심과 문제를 야기하기 시작하면서 정부에서도 미세먼지 오염도 개선을 위한 정책 및 법률을 재수립하기 위해 준비하고 있다.[4] 이전과 같이 미세먼지 오염도의 개선 추세가 정체되는 것을 방지하기 위해 미세먼지 발생에 연관된 영향 인자와 농도에 대한 예측이 매우 중요해 지고 있다.

이에 본 연구에서는 국립환경과학원과 아테네 지역의 연구를 바탕으로 미세먼지 발생에 기상 인자와 대기오염 인자가 연관된 인자라는 것을 확인할 수 있었다.[5, 6] 두 가지 인자들은 시간에 따른 패턴을 가지는 시계열 기반의 데이터로 구성되기에, 시계열 기반의 데이터에 좋은 성능을 보이는 LSTM(Long Short Term Memory) 알고리즘을 통해서 미세먼지 예측을 모델링한다. 최종적으로 서울시 '구별' '시간단위'의 기상 및 대기오염 인자 데이터를 모델링된 LSTM을 통해 미세먼지 농도 예측에 대한 실험을 진행한다.

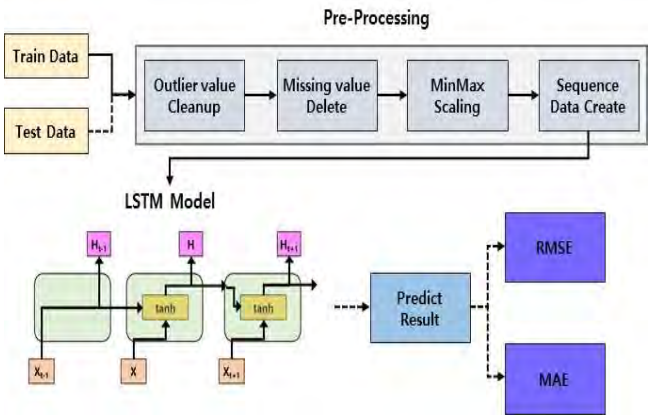
2. 최근 연구 동향

미세먼지 농도를 정확하게 예측하는 것이 중요해진 만큼 다양한 연구 방법을 통해 미세먼지를 예측하는 연구가 진행되어왔다. 이전에는 연구에서는 McKendry와 Zhao의 경우 다층신경망(Multi Layer Perceptron) 및 단일 회귀(Regression) 알고리즘들을 사용하여 단일 회귀 알고리즘이 더 좋은 성능을 보인다고 발표하였지만, 좋은 예측 성능을 보이지 못

하였다.[7, 8] 차진욱의 연구에서는 기상 인자와 대기 오염 데이터를 사용하여 ANN(Artificial Neural Network)과 KNN(K-Nearest Neighbor) 알고리즘을 응용하여 PM10에 대한 예측을 시도 하였으며, Yadav는 PCA(Principal Component Analysis)를 활용하여 최상의 예측 결과를 제공하는 입력 변수 조합을 ANN 알고리즘 학습에 사용하여 예측을 진행했다.[9, 10] 하지만 위의 모든 연구들은 시계열 데이터의 특성을 고려하지 않고 학습을 진행하여 좋은 예측 성능이 나오지 않았다. 이에 반해 데이터의 시계열 특성을 고려하는 LSTM 및 RNN과 같은 알고리즘을 통해 학습을 진행한 이홍석과 Di Antonio연구에서는 미세먼지 예측에 대한 좋은 성능을 보인 것을 확인할 수 있었다.[11, 12]

3. LSTM 알고리즘 기반의 미세먼지 예측 모델

그림 1은 연구에서 제안하는 LSTM 모델의 학습 및 테스트에 대한 흐름을 보여준다.



(그림 1) LSTM 기반의 미세먼지 예측 모델

LSTM은 순환신경망(Recurrent Neural Network)의 한 종류로, 기존 순환신경망과 달리 신경망 내부에 input, forget output gate cell을 포함하고 있는 모델이다. 신경망 내부 메모리를 대체하는 cell을 통해, 긴 시퀀스 학습에 발생되는 타임스태프를 학습하지 못하는 기울기 소실 문제(Vanishing Gradient Problem) 또는 처음 입력된 데이터 학습이 제대로 반영되지 않는 장기 의존성 문제(Long-Term Dependencies)를 해결할 수 있는 시계열 데이터에 강한 모델이다.

3.1 데이터 세트

데이터는 '2015년 ~ 2019년 11월'까지의 서울시 '구별' '시간 단위'의 기상인자와 대기오염 인자로 구성된다. 기상 인자는 온도, 풍향과 같은 기상의 원인이 되는 5개의 속성을 사용하며, 빠른 시간 변화를 나타내는 요소에 대해 측정 순간의 값 혹은 측정 시간 평균값을 사용한다. 기상 인자 데이터는 "기상자료개발포털(https://data.kma.go.kr/cmmn/main.do)"에서 제공된다.

(표 1) 기상 인자 예시

날짜	기온	풍향	풍속	강수량	습도
15-01-01-01	-4.3	289.9	3.4	0.0	0.0
15-01-01-02	-10.1	341.8	5.3	0.0	48.2
...	...	...	...	...	...
19-11-30-24	3.8	78.4	0.7	0.0	71.0

대기오염 인자는 대기 오염 현상에 영향을 주는 물질들로, 발생원에 의해서 1차 오염물질과 상호 반응을 통해 생성되는 2차 대기오염 물질로 분류된다. 대기오염 인자 데이터는 6개의 속성으로 구성되며, 이중 PM10과 PM2.5는 예측 값으로 사용된다. 해당 데이터는 "에어코리아(www.airkorea.or.kr)"에서 제공된다.

(표 2) 대기오염 인자 예시

날짜	SO2	NO2	O3	CO	PM 10	PM 2.5
15-01-01-01	0.005	0.7	0.025	0.015	31	8
15-01-01-02	0.005	0.2	0.019	0.008	70	3
...	...	...	...	...	...	...
19-11-30-24	0.005	0.4	0.02	0.027	29	24

3.2 전처리 과정 (Pre-Processing)

전처리 과정은 그림 1의 Pre-Processing에서 제시되어 있으며, 모델 예측 성능에 부정적인 영향을 주는 요소를 제거 해주는 과정이다.

1. Outlier Cleanup : 2015, 2016년 몇몇 지역의 PM10 및 PM2.5의 수치가 900, 1000과 같이 특이 값을 해당 지역의 정상적인 가장 큰 PM10, PM2.5의 수치로 변경한다.
2. Missing Value Deletes : 데이터 구조에서 각각의 속성에 대한 Null Value를 제거하는 단계이다. 2019년 데이터 세트의 '습도' 속성은 대다수의 지역에서 데이터가 Null Value로 되어 있어 속성 자체를 제거하였다. 또한 PM10 및 PM2.5의 Null Value의 경우 예측 값(Target Value) 이므로, Null Value를 다른 값으로 채우기 보다는 row를

제거하는 방법을 사용하였다.

3. MinMax Scaling : 데이터들의 간의 분포 차이가 큰 경우 학습에 부하가 가는 것을 방지하기 위해서 예측 값(다음 시간 PM10, PM2.5)을 제외한 모든 데이터에 대해 MinMax Scaler를 사용해 0~1 사이 값으로 데이터의 범위를 일치시킨다.
4. Sequence Data Create : LSTM 알고리즘은 시퀀스 데이터를 모델의 입력으로 사용하기 때문에, 사용되는 데이터 세트를 LSTM 입력에 가능한 포맷으로 재구성해야한다. 본 연구에서는 Pandas Library의 Shift() 함수를 사용하여, 미세먼지 농도가 예측 되는 시점(t+1)을 기준으로 모델의 시퀀스 길이만큼 이전 시간(t-i)과 현재 시간(t)의 미세먼지 농도, 대기오염 인자 및 기상 인자가 시퀀스로 연속되는 형태로 구성된다.

$$X_{Seq} = \{(X_{1(t-i)}, Y_{1(t-i)}, X_{1(t_i)}, Y_{1(t_i)}), \\ \ldots (X_{i(t-1)}, Y_{i(t-1)}, X_{i(t)}, Y_{i(t)})\}$$
$$Y_{pm10} = \{Y_{1(t+1)}, Y_{2(t+1)}, \ldots Y_{i(t+1)}\}$$
$$Y_{pm2.5} = \{Y_{1(t+1)}, Y_{2(t+1)}, \ldots Y_{i(t+1)}\}$$

$X = \{$기상인자, 대기오염인자$\}$, $t =$ 현재시간
$i =$ 시퀀스길이 $t-i$: 시퀀스길이 만큼 이전시간
$Y = \{pm10 \text{ or } pm2.5\}$

(그림 2) 시퀀스 데이터 구성

4. 실험 및 평가

제안된 모델의 실험 환경은 표 4와 같으며, 전체 데이터 중에서 2015년 1월 1일 1시부터 2018년 12월 31일 24시까지 데이터는 학습을 위한 데이터로 2019년 1월 1일 1시부터 2019년 11월 30일 24시까지 데이터는 모델의 예측 성능을 테스트하기 위한 데이터로 사용한다. 모델 학습에 사용되는 파라미터는 표 5에 제시되어있으며, 모델의 성능 평가는 다음 시간(t+1)에 대한 두 개의 미세먼지 농도(PM10, PM2.5)에 대한 예측 오차를 통해 확인한다.

(표 4) 실험 환경

구분	이름
Language	Python
Library	Tensorflow 2.1.0 Keras 2.2.4-tf Scikitlearn 0.22.1
GPU	Nvidia Geforce RTX 2070 Super
Memory	64GB

(표 5) 예측 모델 Hyperparameter

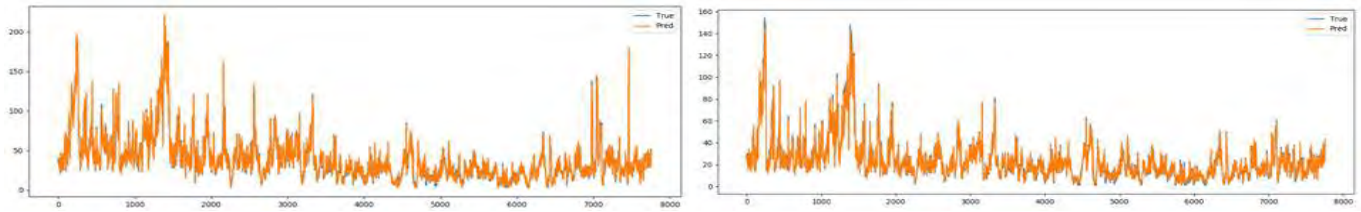
Parameter	Value
epoch	50
batch_size	32
Hidden Layer	256
Optimizer	Adam
learning rate	0.001
Sequence Length	1

예측 오차를 구하기 위해서, RMSE(Root Mean Square Error) 및 MAE(Mean Absoult Error)를 통해 실제 값과 예측 값의 차이를 계산한다. 실험 결과는 표 6에 제시되어 있으며, 대부분의 지역에서 낮은 수치의 예측 오차(RMSE, MAE)가 계산된 것을 확인할 수 있었다. 하지만 지역별 예측 오차의 값의 차이가 생각보다 높은 것을 볼 수 있는데, 전처리의 Missing Value Delete 단계에서 지역별로 삭제된 데이터양이 다른 것이 원인이라 생각된다.

(표 6) 구별 RMSE, MAE 성능 지표

서울시 구	RMSE		MAE	
	PM10	PM2.5	PM10	PM2.5
중구	5.7	3.75	3.7	2.5
용산구	5.9	4.4	3.9	2.9
광진구	7.9	5.4	5.2	3.7
성동구	8.5	5.5	5.9	3.8
중랑구	6.4	4.3	4.3	2.9
동대문구	6.5	4.4	4.2	2.8
성북구	8.0	5.2	5.4	3.7
도봉구	7.3	4.9	5.0	3.3
은평구	7.9	4.7	5.2	3.2
서대문구	7.7	5.07	5.0	3.3
마포구	8.3	5.59	5.4	3.7
강서구	7.4	4.51	4.9	3.08
구로구	8.2	5.63	5.3	3.7
영등포구	8.07	6.04	5.3	4.08
동작구	6.4	4.25	4.0	2.8
관악구	8.6	6.04	5.8	4.2
강남구	6.4	4.63	4.2	3.1
서초구	8.8	5.26	5.6	3.4
송파구	7.4	4.93	5.2	3.4
강동구	6.6	4.12	4.2	2.8
금천구	**5.3**	**3.71**	**3.5**	**2.5**
강북구	7.6	4.3	4.9	2.9
양천구	6.9	4.1	4.4	2.8
노원구	5.7	4.2	3.7	2.9
평균	7.2	4.78	4.7	3.2

가장 예측 성능이 좋은 곳은 금천구로 PM10(5.3, 3,71) 및 PM2.5(3.5, 2.5) 속성 모두 가장 낮은 예측 오차를 보였으며, 이에 대한 예측 그래프 그림 3을 통해 실제 미세먼지 농도와 예측 값이 매우 근사한



(그림 3) 금천구 미세먼지 예측 그래프 PM10(좌), PM2.5(우)

것을 확인할 수 있다. 하지만 PM10의 경우 미세먼지 농도가 200μm이상인 경우, 과대 예측되는 성향을 보인다.

5. 결론

본 연구에서는 미세먼지 농도를 예측을 위해 시계열 데이터에 좋은 성능을 보이는 LSTM 알고리즘을 사용하여 미세먼지 농도 예측을 평가 했다. 데이터는 이전 연구를 분석을 통해 기상 인자와 대기오염 인자를 활용하였으며, 4단계의 전치리 과정을 통해 학습과 테스트에 사용할 데이터를 정리하였다. 학습된 모델에 RMSE 및 MAE를 사용하여 미세먼지 농도의 예측 오차를 계산한 결과 평균적으로 PM10의 경우 '7.2, 4.78', PM2.5의 경우 '4.7, 3.2'의 예측 오차를 보였으며, 금천구의 경우 PM10이 '5.3, 3.71' PM2.5이 '3.5, 2.5'로 가장 좋은 성능을 보였다. 하지만 PM10의 경우 200이상일 경우 과대 예측하는 현상을 보였으며, '시간 단위'로 구성된 데이터의 Null Value가 많아 지역별로 예측 오차 값의 차이가 발생되었다. 추후 시간별 데이터와 일치하는 일별 평균값으로 Null Value를 대체하여 실험해보는 방법과 통계 모델인 ARIMA, VAR과 같은 방법의 예측 성능을 비교해보는 연구를 계획하고 있다.

Acknowledgement

This work was supported by the National Research Foundationof Korea (NRF) grant funded by the Korea government (MSIT) (2018R1D1A1B07050633)

참고문헌

[1] 우정헌. "수도권 미세먼지 환경 개선을 위한 미국의 대기환경정책 사례 조사 연구." 한국대기환경학회지 (국문) 25.6 (2009): 579-593.

[2] 최종규, et al. "미세먼지의 질병에 미치는 유해성." 생명과학회지 30.2 (2020): 191-201.

[3] Jaafari, Jalil, et al. "Characterization, risk assessment and potential source identification of PM10 in Tehran." Microchemical Journal 154 (2020): 104533.

[4] 김봉균, et al. "미세먼지 저감을 위한 정책 선정 연구." 한국전자거래학회지 25.1 (2020): 109-121.

[5] 공부주, 한진석, 이민도, 이정영, 박진수, "기상인자가 미세먼지 농도에 미치는 영향 연구", 국립환경과학원, pp. 1-137, 2006.

[6] Abatzoglou, G., Chaloulakou, A., Assimacopoulos, D., Lekkas, T, "Prediction of air pollution episodes: Extreme value theory applied in Athens," Environmental technology, vol.17, NO.4, pp349-359, 1996.

[7] Pires, J. C. M., Martins, F. G., Sousa, S. I. V., Ferraz, M. C. M. A., & Pereira, M. C., "Prediction of the daily mean PM10 concentrations using linear models," American Journal of Environmental Sciences, vol.4, no.5, pp.445-453, 2008.

[8] Zhao, Yin, "Machine learning algorithms for predicting roadside fine particulate matter concentration level in Hong Kong Central,"

[9] 차진욱, and 김장영. "미세먼지 수치 예측 모델 구현을 위한 데이터마이닝 알고리즘 개발." 한국정보통신학회논문지 22.4 (2018): 595-601.

[10] Yadav, V., and S. Nath. "Novel hybrid model for daily prediction of PM 10 using principal component analysis and artificial neural network." International Journal of Environmental Science and Technology 16.6 (2019): 2839-2848.

[11] 이홍석, et al. "도심지 교통흐름 및 미세먼지 예측을 위한 딥러닝 LSTM 프레임워크." 정보과학회논문지 47.3 (2020): 292-297.

[12] Di Antonio, Ludovico, et al. "Multivariate Prediction of PM 10 Concentration by LSTM Neural Networks." 2019 Photonics & Electromagnetics Research Symposium-Fall (PIERS-Fall). IEEE, 2019.