풍향풍속과 미세먼지의 상관관계 분석과 LSTM을 이용한 미세먼지 예측

Analysis of Correlation of Wind Direction/Speed and Particulate Matter(PM10) and Prediction of Particulate Matter Using LSTM

저자 (Authors)
조수현, 정미리, 이진향, 오일석, 한영태
Jo Soohyun, Miri Jeong, Jinhyang Lee, Ilseok Oh, Yeongtae Han

출처 (Source)
한국정보과학회 학술발표논문집 , 2020.7, 1649-1651 (3 pages)

발행처 (Publisher)
한국정보과학회
The Korean Institute of Information Scientists and Engineers

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE09874876

APA Style
조수현, 정미리, 이진향, 오일석, 한영태 (2020). 풍향풍속과 미세먼지의 상관관계 분석과 LSTM을 이용한 미세먼지 예측. 한국정보과학회 학술발표논문집, 1649-1651.

풍향풍속과 미세먼지의 상관관계 분석과 LSTM을 이용한 미세먼지 예측

조수현°, 정미리°, 이진향°, 오일석, 한영태
{eunyong5675, 1889543, wlsgid2378}@naver.com, {isoh, hanyt}@jbnu.ac.kr
전북대학교 컴퓨터공학부

Analysis of Correlation of Wind Direction/Speed and Particulate Matter(PM10) and Prediction of Particulate Matter Using LSTM

Jo Soohyun, Miri Jeong, Jinhyang Lee, Ilseok Oh, Yeongtae Han
Division of Computer Science and Engineering, Jeonbuk National University

요 약

동아시아 지역의 급격한 산업화로 인한 미세먼지 농도의 증가는 주요한 사회적 이슈가 되었고 정확한 미세먼지 농도 예측에 대한 필요성이 증가하고 있다. 바람이 미세먼지를 실어 나르기 때문에 풍향과 풍속이 미세먼지 농도에 상당한 영향을 미칠 것이라는 전제하에 기상 데이터를 활용하여 풍향/풍속과 미세먼지 농도의 상관관계를 정량적으로 분석한다. 또한 LSTM(long short-term memory)으로 기상 데이터를 모델링하고 미세먼지를 예측한다. 기상청이 수집한 대한민국 17개 지역의 2017년 1월부터 2019년 6월까지의 일간 최다풍향, 평균풍속을 사용하였고 이를 7:3 비율로 훈련 집합과 테스트 집합으로 나누었다. 훈련 집합을 사용하여 서울시 미세먼지 농도를 예측하는 LSTM 모델을 학습하였고 테스트 집합으로 모델의 정확도를 평가하였다. 실험 결과 지난 최근 2일 데이터를 보고 내일을 예측하는 경우의 정확률이 최대였으며 이때 Mean Absolute Error(MAE)는 14.34이다.

1. 서 론

20세기 중반에 발생하고 있는 다양한 대기오염 문제는 인간에게 많은 영향을 미치고 있다. 특히 대기오염으로 인한 사망률과 발병률의 증가는 심각성을 야기하고 있다. 과거에는 무분별한 배출에 규제를 강화하면 대기오염의 문제가 건강에 미치는 위험요소로서의 문제가 적어질 것이라고 생각했지만, 최근 유럽, 미국과 같은 대기오염 수준이 상대적으로 낮은 국가에서도 사망률 및 발병률이 증가하고 있다고 보고된다. 이러한 인체에 미치는 영향은 대부분 화석연료를 연소할 때 발생하는 미세먼지에 의해 나타난 것으로 최근 다양한 연구가 밝히고 있다. 세계보건기구에서는 대기오염으로 인해 매년 3백만 명 가량이 사망에 이르는 것으로 조사하였으며, 불확실한 사망자수를 고려했을 때 실제 사망자수는 140만 명에서 최대 600만 명까지 이를 것으로 예측하였다. 이러한 수치는 전 세계 매년 5천 5백만 명 사망자의 약 5%를 차지하는 비율이다.

우리나라에서 2000년대를 기점으로 대기오염문제를 인지하고 이로 인한 사망률 또는 발병률에 대한 연구가 계속해서 이루어지고 있다. 특히 다른 나라에 비해 우리나라의 미세먼지의 수치는 비교적 높기 때문에 미세먼지로 인한 위해성은 매우 높을 것으로 추정되고 있다[1].

본 논문에서는 기상청에서 수집한 풍향/풍속과 미세먼지 데이터를 LSTM모델로 학습하여 다음날의 서울시 미세먼지 농도를 예측하는 모델을 제안한다. LSTM은 시계열데이터를 분석할 때 사용하는 머신러닝 모델중 하나로 시계열 데이터인 미세먼지 농도, 풍향/풍속을 예측 및 분석하기에 적합하다. 사용된 풍향/풍속은 훈련집합과 테스트집합을 7:3 비율로 나누었고 Min-Max Scaling을 통해 정규화 하여 LSTM을 학습시켰다. LSTM 모델의 성능을 3가지 정량적 평가지표 Mean Absolute Error(MAE), Mean Squared Error(MSE), Root Mean Squared Error(RMSE)를 사용하여 평가하였다. 실험결과 모델의 성능은 MAE, MSE, RMSE는 각각 14.34, 472.48, 21.73 이다. 또한 LSTM의 기상데이터 입력 시 윈도우 크기에 따른 성능 비교도 하였다. 실험결과 2일간의 기상데이터로 다음날을 예측할 때 성능이 가장 좋았다. 풍향/풍속과 미세먼지 사이의 피어슨 상관관계를 분석하여 예측 모델을 적용하기에 적합한 데이터라는 사실을 입증하였다.

2. 관련연구

미세먼지 농도를 예측하기 위해 다양한 모델을 사용한 연구들이 진행되고 있다. 특히 신경망 모델을 활용해 시계열 자료를 분석하는 연구가 최근에 많이 이루어지고 있다.

[2]는 기질정보와 기상정보를 수집하여 MLR, ARIMA, SVR, ARIMAX 모델을 사용하여 미세먼지 예측 성능을 비교분석 하였다. 실험결과 ARIMAX 모델이 성능이 가장 좋았다. [3]은 PM10, PM2.5 이산화질소, 오존, 일산화탄소, 아황산가스를 측정한 데이터를 통해 SVM(Support Vector Machine) 모델을 사용하여 미세먼지 농도를 예측하였다. [4]는 미세먼지 수치를 신경망과 KNN을 사용하여 예측하였다. 예측 모델의 출력값은 매우 나쁨, 나쁨, 보통, 좋음의 4단계로 구분하여 예측하였다. [5]는 ANN(Artificial Neural Network)과 SVM 모델을 활용해 미세먼지를 예측하는 연구를 수행했다. 다중회귀분석을 사용해 미세먼지 관련하여 영향을 미치는 풍향/풍속 변수를 선택했으며 ANN 모델이 가장 좋은 성능을 나타냈다. [6]의 연구에서는 5개의 기상 데이터를 머신러닝을 사용하여 미세먼지를 예측하는 연구를 진행했다. MLP(Multilayer Perceptron) 모델이 가장 높은 정확도를 나타냈다. [7]의 연구에서는 기상환경 변수와

계절에 의해 달라지는 기상 변수를 활용해 SVM, DNN(Deep Neural Network), 랜덤포레스트(Random Forest), 다중로지스틱 회귀분석(Multinomial Logistic Regression)로 분석하여 미세먼지 농도를 예측하였다. [8]은 AWS를 활용하여 IoT센서로부터 생성되는 스트리밍 데이터를 수집해 20초 단위의 평균값으로 구성된 30개의 시퀀스 데이터로 변환하여 실내 미세먼지 농도를 예측하는 연구를 진행했다. RNN(Recurrent Neural Network)과 LSTM 모델을 사용해 기상환경 데이터와 대기오염 데이터를 변수로 예측하였고 LSTM 모델이 RNN 모델보다 더 좋은 성능을 나타냈다. [9]는 Logistic regression, CART, Boosting, 랜덤 포레스트를 사용해 홍콩의 미세먼지 농도를 예측하였다. 실험결과 Logistic regression과 랜덤 포레스트가 가장 좋은 성능을 기록하였다.

많은 기존연구들이 머신러닝 알고리즘의 시계열 예측모델을 사용하여 미세먼지 농도를 예측하였다. 그러나 예측모델의 입력 값인 기상정보의 윈도우 크기 최적 값을 실험적으로 제시하지 않았다. 본 논문에서는 LSTM 입력의 윈도우 크기를 실험을 통하여 최적 값을 제시한다.

3. 풍향/풍속 데이터 수집 및 미세먼지와 상관관계 분석

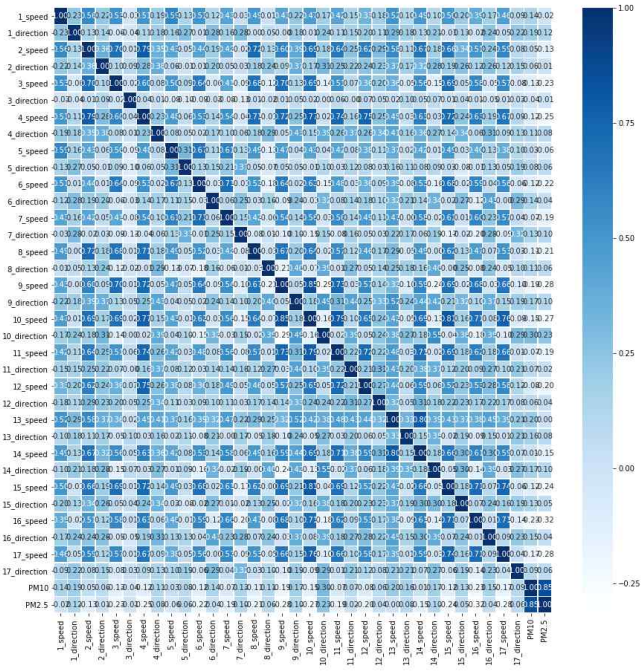

그림 1. 17개 지역의 기상 데이터와 미세먼지 간의 피어슨 상관관계 분석

본 연구에서는 한국환경공단에서 제공하는 기상 데이터를 활용하였다. 임의로 17개 지역을 선정하였고 2017년 1월부터 2019년 6월까지의 최다풍향(16방위, 360도) 및 평균풍속(m/s) 데이터와 기상청의 서울시 미세먼지 및 초미세먼지 데이터를 수집하였다. 이 데이터를 7:3 비율로 분할하여 훈련 집합과 테스트 집합으로 나누었다. 또한 수집한 기상 데이터가 미세먼지를 예측하는 LSTM의 입력으로 적합한지 판단하기 위하여 기상 데이터와 미세먼지 간의 상관 분석을 하였다.

상관분석은 피어슨 상관분석 방법을 통하여 미세먼지와 풍향/풍속간에 선형적 또는 비선형적 관계가 형성되는지 파악하였다. 상관분석의 결과값이 완전히 동일하다면 +1, 완전히 다르다면 0을 나타내며 반대방향으로 동일하다면 -1을 나타낸다. 그림 1은 17개 지역의 풍향/풍속과 미세먼지 간의 피어슨 상관관계 분석 결과이다. 상관분석 결과로 17개 지역의 평균풍속과 PM10농도는 평균 0.1이며 최다풍향과 PM10농도는 0.13으로 약한 양적 선형관계를 이루고 있음을 확인하였고 평균풍속보다 최다풍향이 미세먼지 농도에 더 영향을 미친다는 것을 파악하였다. 이 분석을 통하여 수집한 풍향/풍속이 예측 모델을 학습시킬 때 적합한 데이터임을 확인하였다.

4. 실 험

4.1 학습 모델 및 학습 방법

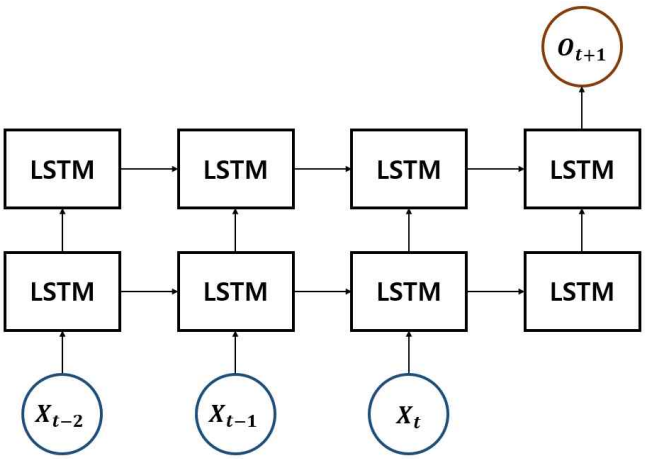

그림 2. 사용한 LSTM 구조
(x=풍향/풍속, 미세먼지농도, o=예측된 미세먼지 농도)

본 논문에서 사용한 LSTM 모델은 Tensorflow를 사용하여 구현하였다. 설계된 학습 모델은 그림 2에서 윈도우 크기만큼 x 입력을 받고 다음 날의 o 출력을 갖는다. 여기서 x는 풍향/풍속과 미세먼지 농도이며 o는 예측된 미세먼지 농도이다. LSTM 모델은 2층으로 구성된 LSTM을 사용하였고 각각의 노드의 수는 256으로 구성하였다. LSTM의 활성화 함수로 hyperbolic tangent 함수를 사용하였다. 또한 각 층 간에 과잉적합(overfitting)을 방지하기 위해 Dropout을 사용하였다. 가중치는 Xavier 방법으로 초기화 하였다. 학습에서 사용한 손실함수는 Mean Absolute Error(MAE)를 사용하였고 최적화 알고리즘으로 Adam 알고리즘을 사용하였고 학습률로 0.1을 사용하였다. 수집한 풍향/풍속 중 시스템 오류로 수집되지 않은 날의 데이터는 선형보간법을 이용하여 보간 하였고 수집된 풍향/풍속과 PM10 수치를 Min max scaling으로 전처리하여 사용하였다.

4.2 평가방법 및 평가 결과

예측 모델의 정량적 평가를 위해 Mean Absolute Error(MAE), Mean Squared Error(MSE), Root Mean Square Error(RMSE)를 사용하였다. 이때, MAE, MSE, RMSE는 아래와 같다.

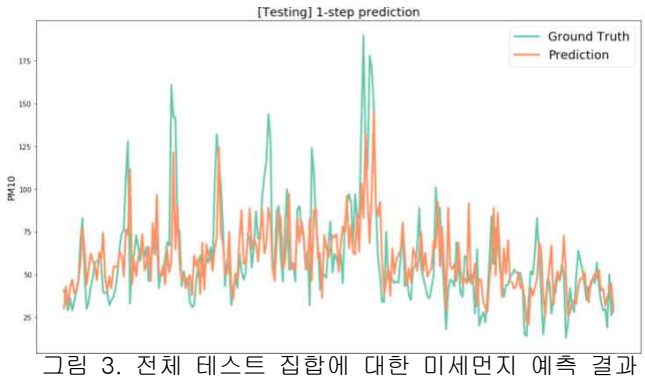

그림 3. 전체 테스트 집합에 대한 미세먼지 예측 결과

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

$$RMSE = \sqrt{MSE}$$

y_i는 예측값, x_i는 실제값, n는 샘플 개수이다. 윈도우 크기가 2일 때 학습된 모델을 테스트 집합으로 평가 했을 때 MAE는 14.3, MSE는 472.48, RMSE는 21.73이 나왔다. 실제 예측 결과는 그림 3와 같으며 전반적인 예측 결과가 실제 결과와 차이가 크지 않다는 것을 확인 할 수 있다.

4.3 윈도우 크기에 따른 평가 결과

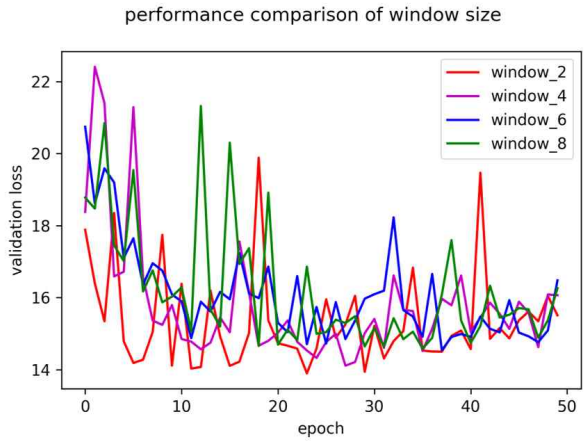

그림 4. 윈도우 크기에 따른 학습 결과

설계된 모델의 적절한 윈도우 크기를 찾기 위해서 윈도우를 2일, 4일, 6일, 8일로 설정하고 학습 집합으로 학습하였고 테스트 집합으로 평가하였다. 그림 4는 윈도우 크기에 따른 성능을 비교한 그래프이다. 실험 결과 윈도우 크기를 2일로 설정했을 때 MAE가 14.34로 가장 좋은 성능을 기록한 것을 볼 수 있었고 학습수렴 속도도 가장 빠른 것으로 나타났다.

5. 결론 및 향후 연구

본 논문에서는 미세먼지와 풍향/풍속의 상관관계를 분석하고 이에 따른 LSTM 모델을 사용해 PM10 농도를 예측하는 모델을 만들었다. 학습과 평가를 위해 대한민국 17개 지역의 기상정보를 수집하였다. 기상정보와 미세먼지의 상관분석 결과 양적 선형관계를 이루고 있음을 확인하였고 수집한 풍향/풍속이 예측 모델을 학습시킬 때 적합한 데이터임을 확인하였다. 수집된 데이터를 LSTM으로 학습하고 예측한 결과 예측값과 실제값의 농도차이가 크지 않음을 확인하였다. 또한 2일간의 데이터로 다음날을 예측했을 때 예측성능이 가장 높은 것을 확인하였다.

이번 연구는 일자별 데이터를 사용했지만 향후 연구에서는 시간별 데이터를 사용하고, 최다풍향, 평균 풍속과 같은 바람 데이터 이외에도 더 다양한 기상 데이터를 활용해 예측을 더 정확히 하고 17개 지역에 한정하지 않고 전국 지역의 데이터로 확장해 데이터 활용성을 높이고 오차율을 줄이는 과정이 앞으로의 연구 방향이다.

참 고 문 헌

[1] 신동천, “Health Effects of Ambient Particulate Matter”, 대한의사협회지, 50권 2호, pp.175-182, 2007.

[2] 전송완, 최제열, 배준현, “미세먼지 농도 예측 알고리즘 성능 비교”, 한국소프트웨어종합학술대회, pp.775-777, 2017.

[3] 오수진, 구자환, 김응모, “미세먼지 발생 locality를 기반으로 하는 농도 예측 기법”, 대한전자공학회학술대회, pp.1357-1360, 2017.

[4] 차진욱, 김장영, “미세먼지 수치 예측 모델 구현을 위한 데이터마이닝 알고리즘 개발”, 한국정보통신학회지 논문지, 22권 4호, pp.595-601, 2018.

[5] 임준묵, “기상환경데이터와 머신러닝을 활용한 미세먼지농도 예측 모델”, 한국IT서비스학회지, 18권 1호, pp.173-186, 2019.

[6] 오병두, 박지후, 김유섭, “Machine-Learning을 활용한 미세먼지(PM10) 농도 예측”, 한국정보과학회 학술발표논문집, pp.1674-1676, 2016.

[7] 전성현, 손영숙, “Prediction of fine dust PM10 using deep neural network model”, The Korean Journal of Applied Statics, 31권 2호, pp.265-285, 2018.

[8] 김삼근, 오택일, “IoT 스트리밍 센서 데이터에 기반한 실시간 PM10 농도 예측 LSTM 모델”, Journal of the Korea Academia-Industrail cooperation Society“, 19권 11호, pp.310-318, 2018.

[9] Y. Zhao, "Data Mining Algorithms for Predicting PM2.5 concentration level in HongKong“, Master thesis, University Sains Malaysia, 2014.