다층 퍼셉트론 신경망을 이용한 미세먼지 AQI 지수 예측

조경우 · 이종성 · 오창헌*

한국기술교육대학교(KOREATECH)

Particulate Matter AQI Index Prediction using Multi-Layer Perceptron Network

Kyoung-woo Cho · Jong-sung Lee · Chang-heon Oh*

Korea University of Technology and Education(KOREATECH)

E-mail : pinokio622@koreatech.ac.kr

요 약

미세먼지로 인한 대기오염 및 인체 영향에 대한 많은 발표로 인해 미세먼지 예보는 많은 대중의 관심을 받고 있다. 이로 인해 통계 모델링 기법과 함께 기계학습 기법을 사용하여 미세먼지 예보 정확도를 올리기 위한 다양한 노력이 수행되고 있다. 본 논문에서는 미세먼지 예측을 위해 다층 퍼셉트론 신경망을 활용한 미세먼지 AQI 지수 예측을 수행한다. 이를 위해 다수의 연구에서 공통적으로 사용된 기상 인자와 미세먼지 농도값을 이용하여 예측 모델을 설계하고 4단계의 미세먼지 AQI 예측 정확도를 비교한다.

ABSTRACT

With many announcements on air pollution and human effects from particulate matters, particulate matter forecasts are attracting a lot of public attention. As a result, various efforts have been made to increase the accuracy of particulate matter forecasting by using statistical modeling and machine learning technique. In this paper, the particulate matter AQI index prediction is performed using the multilayer perceptron neural network for particulate matter prediction. For this purpose, a prediction model is designed by using the meteorological factors and particulate matter concentration values commonly used in a number of studies, and the accuracy of the particulate matter AQI prediction is compared.

키워드

Particulate matter, Multi-layer perceptron, Neural network, Deep learning

Ⅰ. 서 론

미세먼지로 인한 대기오염 및 인체 영향에 대한 많은 발표로 인해 미세먼지 예보는 많은 대중의 관심을 받고 있다. 특히, 우리나라의 경우 미세먼지로 인한 대기오염이 선진 주요 도시 대비 높은 수준을 나타내고 있어 미세먼지 예보 확인은 일상생활을 결정짓는 중요한 요인으로 작용하고 있다 [1]. 이로 인해 높은 미세먼지 예보 정확도를 요구하고 있으나, 공간과 시간에 따라 크기와 화학 성분이 지속적으로 변하는 미세먼지 비선형 성질로 인하여 예보 정확도 향상에 어려움을 겪고 있다. 이에 기존 통계적 선형 방법보다 좋은 결과를 제공하는 기계학습 예측 모델 적용에 대한 다양한 시도가 이루어지고 있다 [2-4].

본 논문에서는 인공 신경망 알고리즘 중 하나인 다층 퍼셉트론 신경망을 사용하여 미세먼지 AQI(Air Quality Index) 지수를 예측하는 예측 모델

* corresponding author

을 설계한다. 이를 위해 다수의 연구에서 공통적으로 사용된 기상 인자와 미세먼지 농도값을 이용하여 예측 모델을 설계하고 4단계의 미세먼지 AQI 지수를 사용하여 예측 정확도를 비교한다.

II. 예측 모델 설계

예측 모델에 사용된 데이터의 경우, 과거 10년간(2009년~2018년 9월) 천안 지역의 온도, 평균 풍속, 최대 풍향, 습도와 같은 일 평균 기상 데이터와 O_3, NO_2, CO, SO_2, PM_{10}과 같은 대기오염 물질 데이터를 활용하였다. 대기오염 물질 데이터의 경우, 측정소 장비 유지 보수로 인한 결측치를 최대한 제거하기 위하여 천안 지역 측정소 3곳의 데이터를 평균을 취하여 활용하였다. 또한 최대 풍향 데이터의 경우 16방위로 표현된 범주형 변수임을 고려하여 one hot encoding을 통한 데이터 전처리를 수행 후 총 24개의 노드로 구성된 input layer를 구성하였다. Hidden layer의 경우 3개의 은닉층으로 구성하였으며, 활성화 함수는 ReLu, 최적화 함수는 adam을 사용하였다. Output layer의 경우 AQI 지수의 '좋음, 보통, 나쁨, 매우 나쁨'을 분류하기 위해 4개의 노드로 구성하였으며, softmax를 활성화 함수로 사용하였다. 이후, 예측 모델의 하이퍼 파라미터 최적화를 위해 랜덤 탐색 방법을 사용하여 표 1과 같이 모델의 최종 파라미터를 설정하였다.

표 1. 모델 parameter

Parameter	Value
hidden layer nodes	20, 20, 20
batch size	60
dropout rate	0.2
l2 penalty	0.001

III. 실험 결과

본 논문에서 설계된 미세먼지 예측 모델의 성능 평가를 위해 모델 학습 후 실제 미세먼지와 예측치의 AQI 지수 비교를 수행하였다. 모델 학습을 위해 총 데이터의 75%를 학습 데이터로 구성하였으며, 예측 성능 비교를 위해 25%의 데이터를 검증 데이터로 활용하였다. 또한, 전체 데이터에서 작은 비율을 차지하는 고농도의 미세먼지 발생 비율을 고려하여 교차 검증을 통한 모델 학습을 진행하였다. 그림 1은 학습 횟수를 150으로 설정하였을 때 모델의 평균 loss 및 accuracy 그래프이며, loss는 0.51, accuracy는 0.81을 나타내었다.

그래프의 학습 횟수가 약 50~60회에 도달했을 때, 검증 데이터의 accuracy는 더 이상 증가되지 않음을 관찰할 수 있으며, loss 역시 더 이상 감소되지 않고 학습 횟수가 증가될수록 loss값이 상승하는 것을 볼 수 있다. 과도한 학습 횟수 설정은 학습 수행 시간의 증가뿐만 아니라 과대 적합 문제를 발생시킬 수 있어 학습 횟수를 50, 55, 60으로 설정하여 각 결과를 비교하였다. 표 2는 각 학습 횟수별 실제 미세먼지 값과 예측 결과의 AQI 지수를 비교한 결과이다.

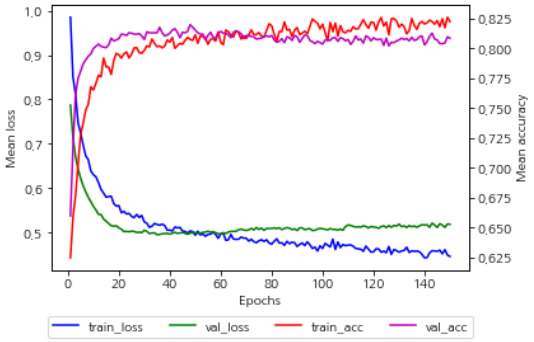

그림 1. 예측 모델 loss 및 accuracy(epoch=150)

표 2. 학습 횟수 별 예측 결과 비교

AQI 지수	빈도	학습 횟수		
		50	55	60
total	884	80.2%	80.7%	80.8%
좋음	198	63.6%	64.6%	62.1%
보통	606	93.1%	92.1%	93.4%
나쁨	74	25.7%	35.1%	32.4%
매우 나쁨	6	0%	16.6%	16.7%

실험 결과 각 학습 횟수에 따라 근소한 예측 성능 차이를 보였으며 학습 횟수가 60회 일 때 80.8%의 최고 정확도를 보였다. '매우 나쁨'의 AQI 지수의 경우, 발생 빈도가 6회로 매우 적어 정확한 예측 성능을 파악하기 어려우나 예측 오차 값을 확인하였을 경우, 대부분의 예측을 '나쁨'으로 과소 예측함을 확인하였다.

IV. 결 론

본 논문에서는 인공 신경망 알고리즘 중 하나인 다층 퍼셉트론 신경망을 사용한 미세먼지 AQI 지수 예측 모델을 설계하였다. 이를 위해 약 10년 치의 천안 지역 일평균 기상 및 대기오염 물질 데이터를 활용하여 설계한 모델을 통해 예측을 수행하였다. 실험 결과, 전체 예측 정확도는 약 80%를 나타내었으나 AQI 지수중 '나쁨'과 '매우 나쁨'의 예측 결과가 각각 약 33.7%, 16.7%로 낮은 성능을 보이는 것을 확인하였다. 실제 예측 결과를 확인한 결과 '나쁨' 이상의 데이터의 경우 전체적으로 한

단계 낮은 AQI 지수로 과소예측하는 경향을 보였다. 이는 전체 AQI 지수의 발생 빈도의 대부분을 '좋음'과 '보통'이 차지하고 있어 고농도 미세먼지에 해당하는 데이터 셋이 상대적으로 적어 발생하는 문제로 판단된다. 따라서 고농도 미세먼지와 상관도가 높은 설명 변수의 추가와 함께 계층별 교차 검증을 통한 학습 데이터의 클래스 비율을 고르게 할 경우 예측 성능을 향상시킬 수 있을 것으로 기대된다.

References

[1] Ministry of Science and ICT et. al., "Roadmap(plan) for the development of particulate matter technology," Policy Report, 2018.

[2] A. Chaloulakou, G. Grivas, and N. Spyrellis, "Neural Network and multiple regression models for PM10 prediction in Athens: A comparative assessment," *Journal of the Air & Waste Management Association*, vol. 53, no. 10, pp. 1183-1190, 2003.

[3] M. M. Dedovic, S. Avadakovic, I. Turkovic, N. Dautbasic, and T. Konjic, "Forecasting PM10 concentrations using neural networks and system for improving air quality," in *proceeding of 2016 XI International Symposium on Telecommunications (BIHTEL)*, Sarajevo, pp. 1-6, 2016.

[4] J. W. Cha, and J. Y. Kim, "Development of data mining algorithm for implementation of fine dust numerical prediction model," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 22, no. 4, pp. 595-601, 2018.