DNN 기반의 미세먼지 농도별 이진 분류 모델

이종성 · 정용진 · 오창헌*

한국기술교육대학교

DNN based Binary Classification Model by Particular Matter Concentration

Jong-sung Lee · Yong-jin Jung · Chang-heon Oh*

Korea University of Technology and Education

E-mail : lee8611@koreatech.ac.kr

요 약

미세먼지 예측의 경우 농도에 따른 특성으로 인해 예측 모델의 학습이 잘 이루어지지 않는 문제가 있다. 이러한 문제를 해결하기 위해 저농도와 고농도에 대한 개별 예측 모델을 구분하여 설계할 필요가 있다. 따라서 미세먼지 농도를 저농도와 고농도로 구분하기 위한 분류 모델이 필요하다. 본 논문은 미세먼지 농도 $80\mu g/m^3$을 기준으로 저농도와 고농도를 구분하기 위한 분류 모델을 제안한다. 분류 모델의 알고리즘은 DNN을 사용하였으며, 하이퍼 파라미터 탐색 후 최적의 파라미터를 적용하여 분류 모델을 설계하였다. 모델의 성능 평가 결과, 저농도 분류의 경우 97.54%, 고농도 분류의 경우 85.51%의 분류 성능을 확인하였다.

ABSTRACT

There is a problem that learning of a prediction model is not well performed depending on the characteristics of each particular matter concentration. To solve this problem, it is necessary to design a prediction model for low concentration and high concentration separately. Therefore, a classification model is needed to classify the concentration of particular matter into low and high concentrations. This paper proposes a classification model to classify low and high concentrations based on the concentration of particular matter. DNN was used as the classification model algorithm, and the classification model was designed by applying the optimal parameters after searching for hyper parameters. As for the result of evaluating the performance of the model, 97.54% of the low concentration classification was measured. And in the case of high concentration classification, 85.51% was measured.

키워드

DNN, Particular matter, Neural Network, Classification

Ⅰ. 서 론

미세먼지 예측의 정확도 향상을 위해 다양한 연구가 이루어지고 있으나 미세먼지 농도에 따른 특성으로 인해 학습이 잘 이루어지지 않는 문제가 있다[1].

이러한 문제를 해결하기 위해 저농도와 고농도를 구분하여 각 농도에 대한 특성이 반영된 예측 모델의 설계가 필요하며, 해당하는 모델들을 설계하기 위해 농도별 분류 모델이 필요하다. 본 논문에서는 미세먼지 농도를 $80\mu g/m^3$의 수치를 기준으로 저농도와 고농도로 구분하기 위한 분류 모델을 제안한다. 분류 알고리즘은 DNN(deep neural network)을 사용하며 분류 모델의 파라미터 최적화 후 분류에 대한 성능 평가를 진행한다.

Ⅱ. 분류 모델 설계

DNN 알고리즘을 이용하여 미세먼지 농도를 저

* corresponding author

농도와 고농도로 구분하기 위해 모델 학습이 필요하다. 미세먼지 농도에 영향을 주는 데이터로 기상 데이터와 대기오염 물질 데이터가 일반적으로 사용되며, 모델의 학습을 위한 데이터로 사용하였다. 기상 데이터와 대기오염 물질 데이터는 천안시에서 수집한 데이터를 사용하였으며, 많은 표본의 학습을 위해 데이터의 누락이 적은 2009년부터 2018년 구간의 데이터를 사용하였다[2][3]. 기상 데이터는 온도, 습도, 풍속, 풍향을 포함하며, 대기오염 물질 데이터는 PM_{10}, O_3, CO, NO_2, SO_2으로 구성된다. 저농도와 고농도 분류를 위한 학습 결과의 기준을 설정하기 위해 수집된 PM_{10}을 $80\mu g/m^3$를 기준으로 두 개의 클래스로 구분하였다. 다른 데이터들의 경우, one hot encoding과 min max scaling을 적용하여 각 데이터들의 다양한 특성을 동일한 특성으로 전처리를 진행하였다. 모델의 학습과 성능 평가를 위한 데이터의 구성은 75%의 훈련데이터와 25%의 실험데이터로 구분하였다. 학습 결과의 검증을 위한 데이터의 경우, 모델의 최적 파라미터를 구하기 위한 grid search 교차 검증을 진행함에 따라 훈련데이터 중 20%로 구성하였다.

DNN 알고리즘에 적용될 layer의 수는 2개로 정하였으며, 과대 적합을 완화 및 최적의 학습을 위해 hidden node, L2, dropout rate, batch size에 대한 하이퍼 파라미터 탐색을 진행하였다. 표 1은 하이퍼 파라미터 탐색 결과이며, 도출된 최적의 파라미터 값을 기반으로 모델의 설계를 진행하였다. 파라미터 탐색 결과, hidden node는 20, L2는 0.001, dropout rate는 0.3, batch size는 80으로 모델을 구성하였다.

표 1. 하이퍼 파라미터 탐색 결과

parameter	search region	value
hidden node	20 ~ 200 (interval 20)	20
L2	0, 0.1, 0.01, 0.001	0.001
dropout rate	0 ~ 0.5 (interval 0.1)	0.3
batch size	20 ~ 100 (interval 20)	80

III. 모델 성능 평가

하이퍼 파라미터를 이용하여 최적화가 진행 된 모델을 기반으로 미세먼지에 대한 이진 분류 성능 평가를 진행하였다. 표 2는 설계한 모델을 이용하여 미세먼지 농도를 저농도와 고농도로 분류한 결과이다. 학습 후 테스트에 사용된 실제 미세먼지 농도 데이터는 총 21,804개이며, $80\mu g/m^3$ 미만의 경우 19,713개, $80\mu g/m^3$ 이상의 경우 2,091개로 구성된다. 예측 결과, 저농도 예측의 경우 19,229개의 데이터가 일치하였으며 97.54%의 정확도를 보였다. 고농도 예측의 경우 1,788개의 데이터가 일치하였으며 85.51%의 정확도를 보였다. 전체 정확도의 경우 21,804개 중 21,017개의 예측 성공을 보였으며 96.39%의 정확도를 보였다.

표 2. 분류 모델 테스트 셋 분류 결과

구 분	실제 데이터 수	일치 데이터 수	정확도
전체	21,804	21,017	96.39%
저농도 ($80\mu g/m^3$ 미만)	19,713	19,229	97.54%
고농도 ($80\mu g/m^3$ 이상)	2,091	1,788	85.51%

IV. 결 론

본 논문에서는 농도별 예측 모델의 설계를 위한 DNN 알고리즘을 이용한 저농도, 고농도 분류 모델을 제안하였다. 분류 모델의 알고리즘은 DNN을 사용하였으며, 분류 모델의 학습을 위한 데이터는 천안시에서 수집한 기상데이터와 대기오염물질 데이터를 이용하였다. 각 데이터의 특성을 동일한 특성으로 사용하기 위해 전처리하였다. 훈련데이터와 실험데이터를 각 75%, 25%로 구성하여 진행하였다. 모델에 적용될 최적의 파라미터 도출을 위해 하이퍼 파라미터 탐색을 진행하여 도출된 결과를 기반으로 최종 모델을 설계하였다.

설계한 모델의 성능 평가를 위해 학습 후 테스트 셋을 이용하여 예측을 진행하였다. 전체 데이터 중 저농도에 대한 분류의 경우 97.54%, 고농도에 대한 분류는 85.51%의 정확도를 보였다. 전체 데이터 중 고농도에 비해 저농도 데이터의 비중이 높은 데이터 불균형의 문제로 고농도에 대한 분류 정확도가 상대적으로 낮았으며, 해당 분류 알고리즘을 이용하여 고농도에 대한 예측 진행시 성능하락의 문제를 야기할 수 있다. 따라서, 향후 고농도 분류 성능의 향상을 위한 연구를 진행할 계획이다.

Acknowledgement

이 논문은 2021년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (No. NRF-2019R1I1A3A01059038)

References

[1] K. W. Cho, Y. J. Jung, J. S. Lee, and C. H. Oh, "Separation Prediction Model by Concentration based on Deep Neural Network for Improving PM10 Forecast Accuracy," *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 24, No. 1. pp. 8-14, Jan. 2020.

[2] Air Korea. Inquiry of final confirmed measurement data [Internet]. Available : https://www.airkorea.or.kr/web/pastSearch?pMENU_NO=123.

[3] Korea Meteorological Administation. Weather Data Opening Portal Synthetic Weather Observation [Internet]. Available : https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36.