DNN 과 LSTM 기반의 대기질 예측 모델 성능 비교 연구

조성재*, 김준석**, 김성희**, 윤주상**
*동의대학교 응용소프트웨어공학과
**동의대학교 산업ICT기술공학과
xc11161@gmail.com, junsuk.kim@deu.ac.kr, sh.kim@deu.ac.kr, jsyoun@deu.ac.kr

A Comparative Study on the Performance of Air Quality Prediction Model Based on DNN and LSTM

Sung-Jae Jo*, Junsuk Kim**, Sung-Hee Kim**, Joosang Youn**
*Dept. of Application Software Engineering, Dong-eui University
**Dept. of Industrial ICT Engineering, Dong-Eui University

요 약

최근 인공지능을 활용한 대기질 예측 모델 개발 연구가 활발히 진행 중이다. 특히 시계열 데이터 기반 예측 시스템 개발에 장점을 가진 DNN, LSTM 알고리즘을 활용한 다양한 예측 시스템이 제안되고 있다. 본 논문에서는 LSTM을 활용한 모델과 Fully-Connected 기반의 DNN 모델을 활용한 대기질 예측 시스템을 개발하고 두 모델의 예측 정확도를 비교한다. 성능 평가 결과를 보면 LSTM 모델이 DNN 모델보다 모든 면에서 좋은 결과를 보여줬다. 그리고 이산화황(SO_2), 이산화질소(NO_2), 초미세먼지($PM_{2.5}$)에 대해서는 그 차이가 두드러지게 나타났다.

1. 서론

최근 인공지능 기술을 활용한 대기질 예측 연구가 활발히 진행 중이다. 특히, DNN(Deep Neural Network)을 활용한 연구와 LSTM(Long Short-Term Memory)을 활용한 연구가 제안되었다[1, 2]. 이 외에도 많은 연구에서 다양한 알고리즘을 적용한 대기질 예측 모델이 제안되고 있지만 사용된 알고리즘에 따른 예측 정확도를 비교하는 연구 결과는 아직 제안되어 있지 않다. 본 논문은 기존 시계열 데이터 학습을 위해 사용되는 대표적 알고리즘인 DNN과 LSTM 알고리즘을 활용한 대기질 예측 모델을 개발하고 두 예측 모델의 대기질 예측 정확도를 비교 분석할 것이다.

이어지는 2장에서는 관련 연구를 기술하고, 3장에서는 [1]에서 제안된 DNN 모델과 LSTM 기반 대기질 예측 시스템을 제시할 것이다. 4장에서 각 모델의 성능을 비교 분석할 것이고, 마지막으로 5장에서 논문의 결론을 내린다.

2. 관련 연구

[1]에서는 완전연결신경망을 기반으로 하는 대기질 예측 모델을 제안했다. 학습을 위한 데이터는 대기오염물질의 시간대별 농도자료 6종(SO_2, CO, O_3, NO_2, PM_{10}, $PM_{2.5}$)과 이와 상응하는 시간대의 기상정보(온도, 습도, 풍향, 풍속) 4종을 활용한 10종이 활용됐으며, 총 3년간의 연속 시간별 데이터가 학습에 사용되었다. 그 결과 각각의 대기오염물질 농도에 대한 예측 정확도는 75% ~ 88%를 보였다.

3. 대기질 예측 모델

3.1 학습 데이터

대기질 예측 모델에 사용된 학습데이터는 한국환경공단 에어코리아에서 제공하는 대기오염물질의 시간대별 농도자료 6종(SO_2, CO, O_3, NO_2, PM_{10}, $PM_{2.5}$) 및 기상청 기상자료개방포털에서 제공하는 기상자료 4종(온도, 습도, 풍향, 풍속)으로 서울 중구 지역의 3년간(2016.01.01. ~ 2018.12.31.)의 데이터를 사용했다.

3.2 DNN 기반 대기질 예측 모델

DNN모델은 [1]논문에서 사용된 구현 방법을 기반으로 했으며, 일부 수정을 거쳤다. 모델은 입력층

과 3개의 은닉층 그리고 출력층으로 설계되고, Keras의 Sequential Model과 Dense Layer로 구현했다. 각 은닉층에 할당되는 노드의 개수는 각각의 대기오염물질에 따라 100 ~ 400 개가 할당된다. 모델의 입력으로는 대기오염물질 6종의 농도자료와 기상자료 4종이 사용되고 예측 결과는 입력 시간으로부터 한 시간 후의 대기오염물질 농도를 나타내도록 했다.

3.3 LSTM 기반 대기질 예측 모델

LSTM 모델은 기본적으로 입력층, 은닉층, 출력층 3계층 구조를 갖지만, 미세먼지(PM_{10})와 초미세먼지($PM_{2.5}$)에 대해서는 예측 정확도를 높이고자 은닉층을 한 층 늘린 4계층으로 설계했다. 그리고 3계층 모델의 은닉층에는 4개의 LSTM 유닛을 할당했고, 4계층 모델의 은닉층에는 각각 256, 128개의 LSTM 유닛을 할당했다. 모델의 구현은 Keras의 Sequential Model과 LSTM layer를 활용해 구현했다. 모델의 입력으로는 각각의 물질에 대해 직전 100시간 동안의 시간대별 연속 관측 자료가 적용되고 예측 결과는 입력 데이터로부터 한 시간 후의 대기오염물질 농도를 나타내도록 했다.

4. 모델 성능 평가

4.1 모델 성능 평가 방법

모델의 예측 결과를 실제 관측값과 비교하기 위해 각각의 대기오염물질에 대해 IOA(index of agreement), ME(bias) NRMSE (Normalized RMSE) 값을 산출하여 표로 정리하여 나타냈다. 비교에 사용된 수식은 논문[1]의 모델 검증 과정에서 사용된 수식이다.

$$IOA = 1 - \frac{\sum_{i=1}^{n}(|O_i - M_i|)^2}{\sum_{i=1}^{n}(|M_i - \overline{M}| + |O_i - \overline{O}_i|)^2}$$

$$ME(bias) = \frac{1}{n}\sum_{i=1}^{n}(O_i - M_i)$$

$$NRMSE = \frac{100}{\overline{O}} \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - M_i)^2}$$

여기서 O_i는 관측 값, M_i는 모델의 예측 값을 나타낸다. $\overline{O}$ 와 $\overline{M}$는 각각 관측 값과 모델 예측 값의 평균이다. n은 검증에 사용된 데이터의 개수다.

4.2 모델 성능 평가 결과

모델의 성능을 평가하기 위해 사용된 검증데이터는 학습 과정에서 사용되지 않은 데이터로 서울 중구 지역의 3개월간(2019.01.01. ~ 2019.03.31.) 데이터다. 생성된 모델이 일정한 성능을 보이는지 확인하기 위해 입력데이터에 대해 총 10회의 예측을 진행했고, 그 결과 산출된 IOA, ME(bias), NRMSE 값의 평균을 <표 1>에 제시했다.

대기오염물질	IOA		ME(bias)		NRMSE	
	LSTM	DNN	LSTM	DNN	LSTM	DNN
SO_2	0.915	0.669	-0.277	0.461	14.653	29.145
CO	0.967	0.953	-14.719	29.422	13.373	15.312
O_3	0.969	0.944	0.530	0.857	24.859	32.048
NO_2	0.966	0.898	-0.594	4.961	16.454	25.312
PM_{10}	0.987	0.977	-0.264	5.121	12.417	16.106
$PM_{2.5}$	0.987	0.960	-0.296	5.738	16.354	25.739

<표 1> 검증 결과

IOA는 관측값과 예측값의 시계열 유사성을 나타내는 척도로 0 ~ 1 사이의 값을 가진다. IOA 값이 1에 근접할수록 관측값과 모델의 예측값이 시계열에 대해 일치함을 뜻한다. 모든 경우에서 LSTM의 IOA 값이 DNN에 비해 높게 나타났고, 특히 이산화황(SO_2)에서 그 차이가 두드러지게 나타났다.

ME(bias)는 관측값과 모델의 예측값 간의 평균 편향을 나타내는 지표로 0에 근접할수록 편향이 적음을 의미한다. 단위는 각 대기오염물질의 농도와 동일하다. 두 모델에서 ME(bias)의 값은 매우 낮은 수치를 보였다. 하지만 일산화탄소(CO), 이산화질소(NO_2), 미세먼지(PM_{10}), 초미세먼지($PM_{2.5}$)에서 LSTM이 비교적 적은 편향을 나타냈다.

NRMSE는 RMSE 값을 관측 평균으로 나누어 백분율로 나타낸 것이다. 모든 경우에서 LSTM이 DNN에 비해 낮았고 이산화황(SO_2), 오존(O_3), 이산화질소(NO_2), 초미세먼지($PM_{2.5}$) 큰 차이를 보였다.

5. 결론

모델의 성능 평가 결과 모든 지표에서 LSTM 기반 모델이 DNN 기반 모델 보다 좋은 예측 결과를 보였다. 특히 이산화황(SO_2), 이산화질소(NO_2), 초미

세먼지($PM_{2.5}$)에 대해서는 유의미한 차이를 보였다.

본 결과는 대기질 데이터가 시계열에 의존적임과 동시에 LSTM 모델이 시계열 데이터 기반 예측 시스템에 강점을 가지고 있기 때문이다[3]. 하지만 학습하는 과정에서 에포크(epoch)횟수, 배치(batch)사이즈 그리고 입력 데이터의 형태 등 모든 조건이 동일하게 적용되지 않아 절대적인 비교는 불가능했다는 한계점이 있다.

ACKNOWLEDGMENT

본 연구는 ICT R&D 혁신 바우처 지원 사업의 지원을 받아 수행된 결과임. (20190019150 012002_104)

참고문헌

[1] 조경학, 이병영, 권명흠, 김석철. (2019). 심층 신경망을 이용한 대기질 예측. 한국대기환경학회지, 35(2), 214-225.

[2] İ. Kök, M. U. Şimşek and S. Özdemir, "A deep learning model for air quality prediction in smart cities," 2017 IEEE International Conference on Big Data (Big Data), Boston, MA, 2017, pp. 1983-1990.

[3] 백창룡. (2013). 한국의 미세먼지 시계열 분석: 장기종속 시계열 혹은 비정상 평균변화모형. 응용통계연구, 26(6), 987-998.