DBpia

딥러닝 분석을 이용한 미세먼지 농도 예측에 관한 연구

A Study on The Prediction of The Fine-Dust Concentration Using RNN/LSTM

저자
(Authors)

송재철, 나경필, 김문철, 김명하, 임승택, 심용기
Jai-Chul Song, Kyoung-Pil Ra, Moon-Chel Kim, Myoung-Ha Kim, Seung-Taek Lim, Yong-Gi Sim

출처
(Source)

대한전자공학회 학술대회 , 2019.6, 1400-1405(6 pages)

발행처
(Publisher)

대한전자공학회
The Institute of Electronics and Information Engineers

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE08762275

APA Style

송재철, 나경필, 김문철, 김명하, 임승택, 심용기 (2019). 딥러닝 분석을 이용한 미세먼지 농도 예측에 관한 연구. 대한전자공학회 학술대회, 1400-1405

이용정보
(Accessed)

부산도서관
210.103.83.***
2021/09/24 13:52 (KST)

저작권 안내

DBpia에서 제공되는 모든 저작물의 저작권은 원저작자에게 있으며, 누리미디어는 각 저작물의 내용을 보증하거나 책임을 지지 않습니다. 그리고 DBpia에서 제공되는 저작물은 DBpia와 구독계약을 체결한 기관소속 이용자 혹은 해당 저작물의 개별 구매자가 비영리적으로만 이용할 수 있습니다. 그러므로 이에 위반하여 DBpia에서 제공되는 저작물을 복제, 전송 등의 방법으로 무단 이용하는 경우 관련 법령에 따라 민, 형사상의 책임을 질 수 있습니다.

Copyright Information

Copyright of all literary works provided by DBpia belongs to the copyright holder(s)and Nurimedia does not guarantee contents of the literary work or assume responsibility for the same. In addition, the literary works provided by DBpia may only be used by the users affiliated to the institutions which executed a subscription agreement with DBpia or the individual purchasers of the literary work(s)for non-commercial purposes. Therefore, any person who illegally uses the literary works provided by DBpia by means of reproduction or transmission shall assume civil and criminal responsibility according to applicable laws and regulations.

딥러닝 분석을 이용한 미세먼지 농도 예측에 관한 연구

*송재철
인덕대학교 정보통신공학과
e-mail : jcsong@induk.ac.kr
**나경필, 김문철, 김명하, 임승택, 심용기
에코스솔루션(주)
e-mail : kp.ra, mc.kim, mh.kim,st.lim,yg.sim@ecoss.co.kr

A Study on The Prediction of The Fine-Dust Concentration Using RNN/LSTM

*Jai-Chul Song
Department of Information and Communication Engineering
Induk University
**Kyoung-Pil Ra, Moon-Chel Kim, Myoung-Ha Kim, Seung-Taek Lim, Yong-Gi Sim
ECOSSOLUTION

Abstract

In this paper, we try to predict the fine dust concentration and analyze the fine dust trend using the deep running. In this paper, a prediction model of fine dust concentration was developed and analyzed considering that the weather such as wind direction, wind speed, temperature, humidity, air pressure affects the fine dust concentration. Since the fine dust concentration has a time series characteristic, it is modeled using RNN. This study was conducted using tensorflow.

I. 서론

미세먼지 농도는 산업화가 진행됨에 따라 큰 폭으로 증가하고 있으며, 건강을 크게 위협하는 대기오염 문제이다. 본 논문에서는 미세먼지 예측 모델을 만들어서 미세먼지 농도를 예측 및 분석하고자 한다.

미세먼지 농도를 예측하는 모델을 만드는데 있어서 미세먼지 농도는 그 지역에서의 풍향, 풍속, 기온, 습도, 기압 등의 날씨 정보와 어느정도 상관관계가 있을 수 있다고 고려하였고, 이를 이용하여 예측모델을 만들어 보고자 한다.

미세먼지 농도는 시간의 흐름에 따른 패턴을 가지는 시계열 데이터이므로 RNN/LSTM(Long Short Term Memory) 알고리즘을 이용해 모델링 하였고, 텐서플로우를 이용하여 본 연구를 진행하였다. '일' 단위 및 '시' 단위로 미세먼지 농도를 학습 및 예측해보고, 학습기간 또한 다르게 하여 각각의 결과를 비교해 보고자 한다. 그리고 각각의 날씨요소를 고려하여 RNN/LSTM 모델을 구성 후 각각의 모델의 예측 결과를 비교하여 보고자 한다.

부산도서관 | IP:210.103.83.*** | Accessed 2021/09/24 13:52(KST)

Ⅱ. 본론

2.1 RNN

RNN에 대한 기본적인 아이디어는 순차적인 정보를 처리한다는 데 있다. 기존의 신경망 구조에서는 모든 입력(과 출력)이 각각 독립적이라고 가정했지만, 많은 경우에 이는 옳지 않은 방법이다. 한 예로, 문장에서 다음에 나올 단어를 추측하고 싶다면 이전에 나온 단어들을 아는 것이 큰 도움이 될 것이다. RNN이 recurrent 하다고 불리는 이유는 동일한 태스크를 한 시퀀스의 모든 요소마다 적용하고, 출력 결과는 이전의 계산 결과에 영향을 받기 때문이다. 다른 방식으로 생각해 보자면, RNN은 현재지 계산된 결과에 대한 "메모리" 정보를 갖고 있다고 볼 수도 있다. 이론적으로 RNN은 임의의 길이의 시퀀스 정보를 처리할 수 있지만, 실제로는 비교적 짧은 시퀀스만 효과적으로 처리할 수 있다.

2.2 LSTM

RNN은 시간의 흐름에 따른 데이터의 변동 추이를 계산해 연속해서 입력될 데이터의 값을 예측할 수 있게 해준다. 단순 RNN 모델에서는 시간이 흐름에 따라서 장기의존성(Long-Term Dependency) 문제가 생기지만, LSTM은 이러한 문제를 forget gate, input gate, output gate를 활용함으로써 해결한다[3].

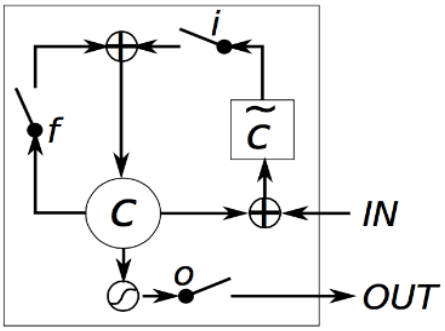

그림 1. LSTM Structure

cell state는 일종의 컨베이어 벨트 역할을 한다. 덕분에 state가 꽤 오래 경과하더라도 gradient가 비교적 전파가 잘 되게 된다. LSTM cell의 수식은 아래와 같다. ⊙는 요소별 곱셈을 뜻하는 Hadamard product 연산자이다.

$$f_t = \sigma\left(W_{xh_f} x_t + W_{hh_f} h_{t-1} + b_{h_f} \right)$$
$$i_t = \sigma\left(W_{xh_i} x_t + W_{hh_i} h_{t-1} + b_{h_i} \right)$$
$$o_t = \sigma\left(W_{xh_o} x_t + W_{hh_o} h_{t-1} + b_{h_o} \right)$$
$$g_t = \tanh\left(W_{xh_g} x_t + W_{hh_g} h_{t-1} + h_{h_g} \right)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$
$$h_t = o_t \odot \tanh(c_t)$$

forget gate f_t는 '과거 정보를 잊기'를 위한 게이트이다. h_{t-1}과 x_t를 받아 sigmoid를 취해준 값이 바로 forget gate가 내보내는 값이 된다. 시그모이드 함수의 출력 범위는 0에서 1 사이이기 때문에 그 값이 0이라면 이전 상태의 정보는 잊고, 1이라면 이전 상태의 정보를 온전히 기억하게 된다.

input gate i_t는 '현재 정보를 기억하기' 위한 게이트이다. h_{t-1}과 x_t를 받아 sigmoid를 취하고, 또 같은 입력으로 hyperbolic tangent를 취해준 다음 Hadamard product 연산을 한 값이 바로 input gate가 내보내는 값이 된다.

Ⅲ. 실험

본 연구에서는 기상청에서 제공하고 있는 풍속, 풍향, 습도, 현지기압, 기온, 미세먼지 농도를 사용했다. 본 연구에서는 여러 지역에서의 미세먼지 농도를 '일' 단위 그리고 '시' 단위로 각각 학습 및 예측하였다. 또한 학습 기간 별 성능 차이도 테스트 해보았다.

'일' 단위 학습은 2016년~2018년도의 데이터를, '시' 단위 학습은 2018년도의 데이터를 사용하여 training 및 test를 진행하였다.

기상정보와 미세먼지농도의 데이터는 기상자료개방포털(https://data.kma.go.kr/cmmn/main.do)과 에어코리아 (https://www.airkorea.or.kr/web)에서 수집하였으며 '시' 그리고 '일' 단위로 데이터를 습득하였다.

본 실험에서의 LSTM 모델의 구현을 위해 다음 요소들을 고려하였다.

1) LSTM 네트워크를 위한 데이터 준비에는 시간 단계가 포함되어 있다. 사건이 될 관측치가 시간 단계가 되는 표본이 될 것이며 관측 된 변수는 특징이 된다. 다음 시간 단계에서 출력을 예측하기 위해 입력으로 time series에서 이전 시간 단계를 취할 수 있다. 과거의 관측치를 별도의 입력 피쳐로 표현하는 대신, 이를 한 입력 피처의 time step로 사용할 수 있다. 이는 실

부산도서관 | IP:210.103.83.*** | Accessed 2021/09/24 13:52(KST)

제로 문제보다 정확한 프레이밍이다. 데이터를 재구성할 때를 제외하고 열을 시간 단계 차원으로 설정하고 형상 차원을 다시 1로 변경한다. 즉, t가 10일 경우 10개월 이전의 미세먼지 농도, 기상정보로부터 현 시점 t까지의 데이터를 가지고 예측 모델을 학습시켜 다음 값을 예측한다.

2) LSTM 네트워크에는 긴 시퀀스에서 기억할 수있는 메모리가 있다. 일반적으로 model을 수용 할 때 각 training batch 후에 model.predict() 또는 model.evaluate()를 호출 할 때마다 네트워크 내의 상태가 재설정된다. LSTM 계층을 “stateful” 함으로써 Keras에서 LSTM 네트워크의 내부 상태가 지워지는 경우를 보다 세밀하게 제어 할 수 있다. 즉, 전체 교육 과정에 걸쳐 상태를 구축 할 수 있으며 예측이 필요한 경우 상태를 유지할 수도 있다. 네트워크에 연결될 때 교육 데이터가 섞이지 않도록 조심해야한다. 또한 model.reset_states()에 대한 호출을 통해 train 데이터(epoch)에 노출 될 때마다 네트워크 상태를 명시적으로 재설정해야 한다.

3) LSTM은 deep network 아키텍처를 활용하여 trainning을 받을 수 있다. LSTM 네트워크는 다른 레이어 유형을 쌓을 수 있는 것과 같은 방법으로 Keras에 쌓을 수 있다. 각 후속 LSTM 레이어 이전의 LSTM 레이어는 시퀀스를 반환해야 한다. 이 작업은 레이어의 return_sequences 매개 변수를 True로 설정하여 수행 할 수 있다.

Ⅳ. 결과분석

본 연구에서 모델 제작을 위한 데이터를 훈련용(training)과 검증용(test) 데이터로 구분한 뒤 진행했다. 검증용 데이터는 예측 모델로부터 계산된 결과로써 실제 값 미세먼지 농도 값과 비교가 가능하다.

학습 파라미터들은 다음과 같이 하였다.

seq_length = 5
data_dim = 8
hidden_dim = 20
output_dim = 1
learning_rate = 0.01
iterations = 600

4.1 일 단위 학습 및 예측 결과

학습은 2016~2018년간 ‘일’ 단위로 학습을 진행하였고, 그 뒤 90일간의 미세먼지 농도 예측값을 LSTM을 이용하여 나타내었다. <그림 2~4>는 실제값과 예측값 그래프를 나타낸다.

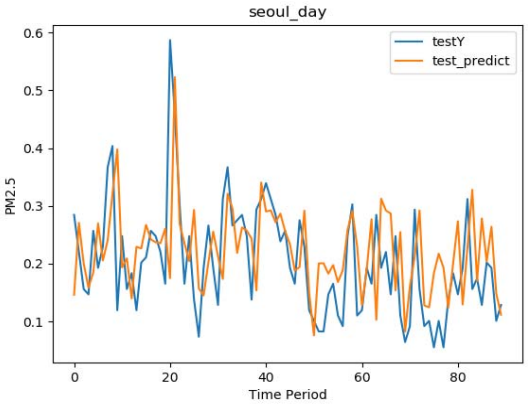

그림 2. 서울 미세먼지 예측 그래프

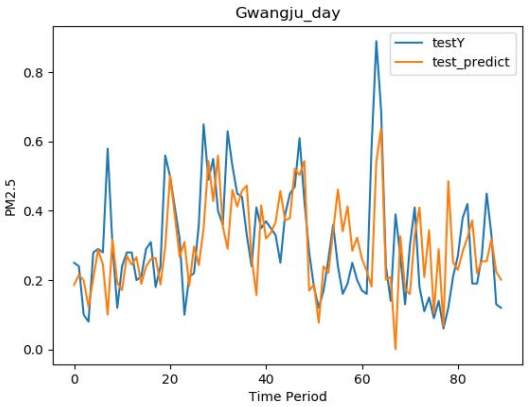

그림 3. 광주 미세먼지 예측 그래프

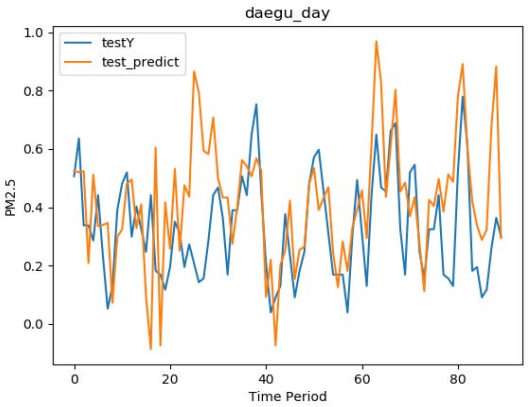

그림 4. 대구 미세먼지 예측 그래프

결과를 보면 큼직 큼직 하게는 어느정도 실제 데이터를 예측 하는 부분들도 있지만, 맞지 않는 부분들과 결과와 차이나는 부분들이 군데 군데 있는 것을 확인

부산도서관 | IP:210.103.83.*** | Accessed 2021/09/24 13:52(KST)

할 수 있다.

4.2 시 단위 학습 및 예측 결과

학습은 2018년 1년을 '시' 단위로 학습 하였고, 그 뒤 90시간을 예측한 값과 실제 값을 그래프로 나타내었다 <그림 5~7>

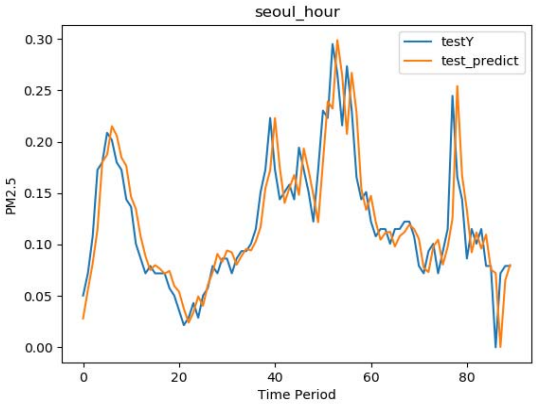

그림 5. 서울 미세먼지 예측 그래프

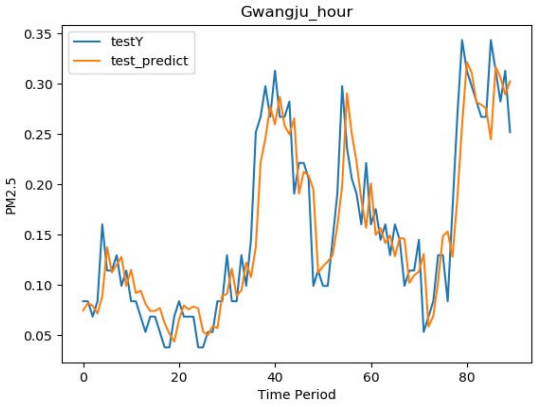

그림 6. 광주 미세먼지 예측 그래프

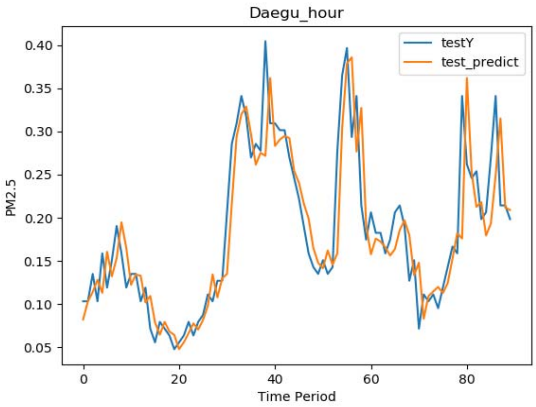

그림 7. 대구 미세먼지 예측 그래프

결과를 보면 실제 값을 잘 맞추고 있는 결과를 볼 수 있다. 군데군데 미세하게는 차이가 나는 부분들이 있지만 전체적으로는 예측값이 실제값과 비슷한 결과를 얻었다..

4.3 학습 기간과 예측 결과 비교분석

딥러닝 학습을 하게 될 때, 학습 데이터를 얼마냐 주느냐에 따라서도 예측 성능 값이 차이가 날 수 있다. 우리는 학습 기간을 차이를 둬서 결과가 어떻게 달라지는지 비교 분석 하였다.

학습 기간은 각각 40일, 365일을 '시' 단위로 서울에서의 데이터를 학습을 진행 하였고, 그 뒤 90시간의 예측값과 실제 값의 데이터를 그래프로 나타내었다<그림 8,9>.

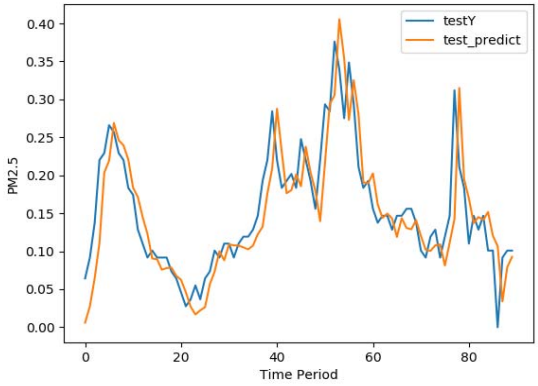

그림 8. 40일간 traning 데이터

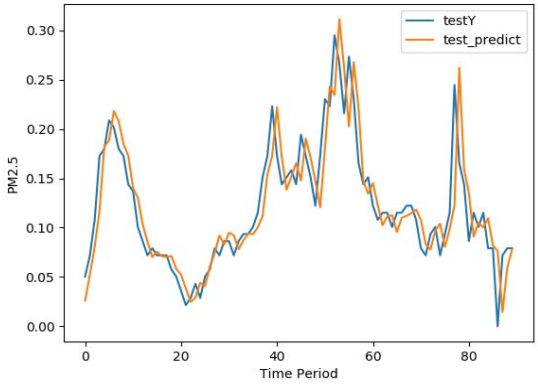

그림 9. 1년간 traning 데이터

두 경우의 예측 성능을 비교해 보면, 두 경우 모두 전체적으로는 예측를 잘 해내었지만, 세부적으로 보다 정확도 있게 예측을 해낸 것은 보다 길게 학습을 한 경우임을 확인할 수 있다.

부산도서관 | IP:210.103.83.*** | Accessed 2021/09/24 13:52(KST)

4.4 각 날씨요소를 적용한 미세먼지 농도 예측

각각의 날씨 요소별로 LSTM 예측 모델에 적용하여 예측 성능을 비교 분석 하였다. 결과는 <그림 10~16>에 나타내었고, 각각의 RMSE 값은 <표 1>에 나타내었다.

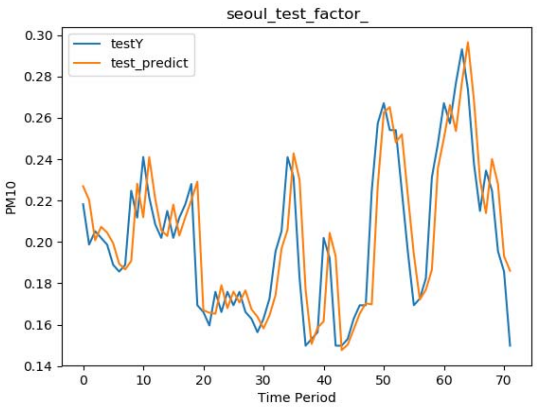

그림 10. 기상 요소 적용 안할 때의 예측 그래프

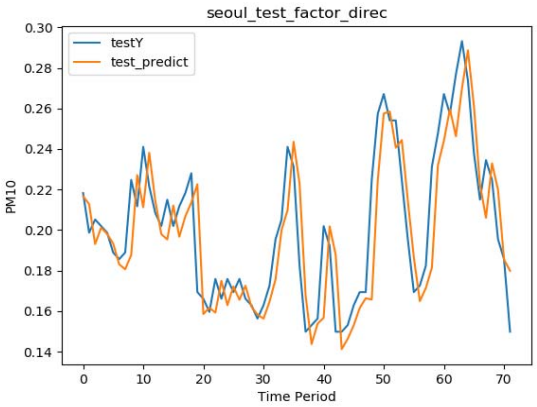

그림 11. 풍향 적용 예측 그래프

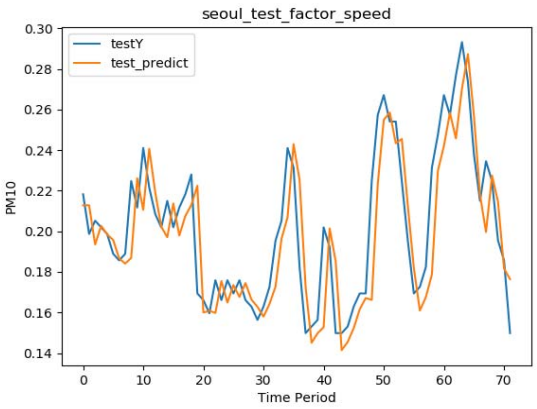

그림 12. 풍속 적용 예측 그래프

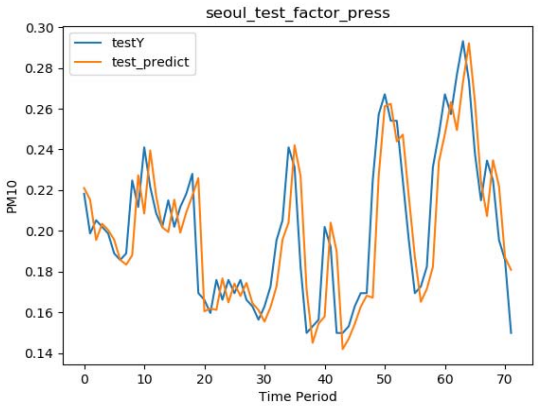

그림 13. 기압 적용 예측 그래프

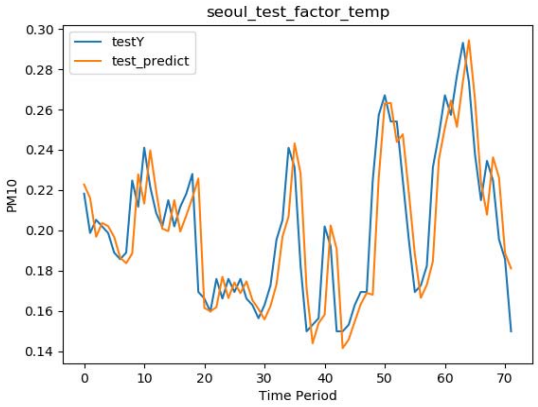

그림 14. 온도 적용 예측 그래프

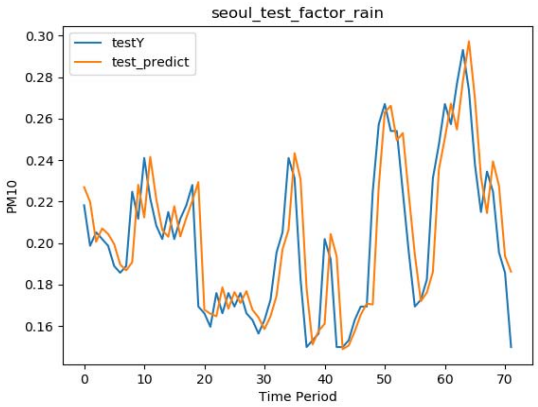

그림 15. 강수량 적용 예측 그래프

부산도서관 | IP:210.103.83.*** | Accessed 2021/09/24 13:52(KST)

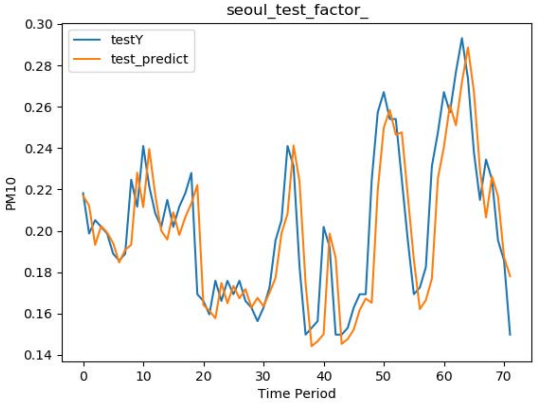

그림 16. 기상 요소 전부 적용 예측 그래프

적용날씨요소	RMSE
-	0.02089
풍향	0.02024
풍속	0.02065
기압	0.02056
온도	0.02045
강수량	0.02098
all	0.02088

표 1. 적용날씨요인별 RMSE값

<표 1>을 보면, 날씨 요소를 적용했을 때랑, 하지 않았을 때 비교해보면 RMSE값은 크게 차이가 나지 않는 것을 볼 수 있다.

그중에 RMSE값이 더 낮은 경우는 풍향, 온도, 기압 순으로 낮았다. 이를통해 풍향이 미세먼지 농도와 가장 연관이 높음을 추측해 볼 수 있다.

V. 결론

실험은 RNN/LSTM Algorithm을 사용하여 지역별, 기간별로 여러 지역에서의 미세먼지 농도를 예측하였고 성능을 분석하였다. 또한 날씨 요소를 적용하여 예측모델을 설계하였고 결과를 분석하였다.

실험을 통하여 미세먼지 농도를 예측 할 수 있었는데, 일 단위 분석 및 예측 시스템 보다는 시 단위로 분석 및 예측하는 경우 더 나은 예측 성능을 가진 결과를 보였다. 이를 통해 더욱 짧은 시간 단위로 분석 및 예측하는 경우 더욱 성능이 좋을 수 있을거라는 추축을 해 볼 수 있다.

딥러닝 모델을 통해 예측하기 위해서는 학습(training)을 충분히 시켜야 하는데, 학습 양이 많을수록 더욱 정확히 예측하는 시스템이 됨을 확인하였다. 하지만 너무 많은 양을 학습하게 된다면 계산에 너무 많은 자원이 소모가 될 수 있고, 계산의 양에 대비하여 효율성이 비교적 떨어질 수도 있을 것이다.

기상요소를 적용을 하였을때는 적용하지 않았을 때에 비해 성능은 크게 차이가 나지 않았다. 하지만 풍향을 요소를 적용하였을 때 비교적 나은 RMSE 값을 얻을 수 있었고, 그 다음이 온도, 기압 순이다. 이를 통해 풍향이 미세먼지 농도와 가장 관련이 깊다고 추측해 볼 수 있었다.

본 연구에서 더 나아갈 방향으로 데이터의 표본을 크게하는 것, 데이터에 대한 깊이 있는 이해를 하는 것이 되겠다. 향후에는 국내의 기상, 미세먼지 상황만 고려할 것이 아니라, 중국 등 주변국의 미세먼지 농도, 풍향(편서풍)등의 요소들도 고려해 볼 수 있겠다.

감사의 글

본 논문은 중소기업청에서 지원하는 2018년도 산학연협력 기술개발사업(No.S2613897)의 연구수행으로 인한 결과물임을 밝힙니다.

참고문헌

[1] LSTM Gating. Chung, Junyoung, et al. “Empirical evaluation of gated recurrent neural networks on sequence modeling.” (2014)

[2] Sepp Hochreiter, Jurgen Schmidhulber, “Long Short-Term Memory”, 1997

[3] Bo Liu Ying Wei, Yu Zhang, Qiang Yang, “Deep Neural Networks for High Dimension, Low Sample Size”, 2017

[4] Prajit Ramachandra, Barret Zoph, Quoc V.LE, “Searching for activation function”, Google Brain, 2017

[5] https://www.kma.go.kr

[6] https://www.airkorea.or.kr/web

부산도서관 | IP:210.103.83.*** | Accessed 2021/09/24 13:52(KST)