효율적인 대기오염 관리를 위한 저가형 대기오염물질 측정 센서 보정 방안 제안 - 머신러닝(HybridRNN) 모델

서울대학교 SEOUL NATIONAL UNIVERSITY

건설환경공학부 2014-14724 강지우
지도교수 : 김재영

1. Introduction

최근 전국민의 관심사가 된 미세먼지를 비롯한 대기오염 물질은 국민들의 건강과 생명을 위협한다. 산업화와 도시화로 인해 도시의 대기오염은 갈수록 심해질 것으로 예상되기 때문에 모든 국가들이 대기오염 관리에 큰 관심을 보이고 있다. 한국환경공단에서는 대기오염으로 인한 피해예방을 위해 미세먼지(PM2.5, PM10), 아황산가스, 일산화탄소, 이산화질소, 오존을 대기환경기준물질 6개 항목으로 지정한 뒤, 이에 대한 대기오염도를 측정 및 공개하고 있다. 현재 대한민국에서는 정부기관에서 고정된 측정소에 (경기 109, 서울 25개소 등) 분석 장비를 설치해 대기오염물질의 농도를 측정하고 있다[1]. 하지만 이러한 고정식 대기오염 측정시스템은 설치간격이 넓기 때문에 공간변화에 따른 대기오염도 변화를 세밀하게 분석하는데 한계가 있다. 따라서, 최근에는 이동식 대기오염 측정소, 촘촘히 설치한 센서들을 이용한 대기오염 측정 시스템 등 도시의 대기오염 관리를 세밀하게 할 수 있는 방법들이 검토되고 있다[2].

하지만 센서들을 촘촘히 설치하기 위해서는 정부기관에서 운영하는 고정 측정소에 비해 저성능, 저가의 센서를 사용해야하기 때문에, 측정값에 대한 신뢰도 문제가 있다. 따라서 본 연구에서는 저가의 센서를 통해 아황산가스, 일산화탄소, 이산화질소, 오존의 농도를 측정한 결과의 정확도를 올릴 수 있는 보정 방법에 대해 제안했다. 센서 측정값을 온도, 습도, 풍속을 이용해 RNN과 DNN을 결합한 HybridRNN 모델을 통해 보정했으며, 다중선형회귀, DNN, RNN 3가지 방법으로 보정한 결과를 기준으로 그 성능을 검증했다.

2. Materials & Methods

2.1 Data

Table.1 Dataset Information

	X				Y
	Sensor	온도	풍속	습도	Station
t	X_S^t	X_T^t	X_W^t	X_H^t	Y^t
…	…	…	…	…	…
t'	$X_S^{t'}$	$X_T^{t'}$	$X_W^{t'}$	$X_H^{t'}$	$Y^{t'}$

본 연구에서는 대한민국 경기도 시흥시 정왕동에 위치한 정왕보건지소에서 측정한 대기오염자료와 기상자료를 이용했다[1,3]. 아황산가스, 일산화탄소, 이산화질소, 오존 네가지 농도에 대해 각각 Table1과 같은 Data를 사용했다. 센서 측정 결과(Sensor), 온도, 풍속, 습도 데이터를 통해 고정 측정소 측정 결과(Station)에 가깝게 보정했다.

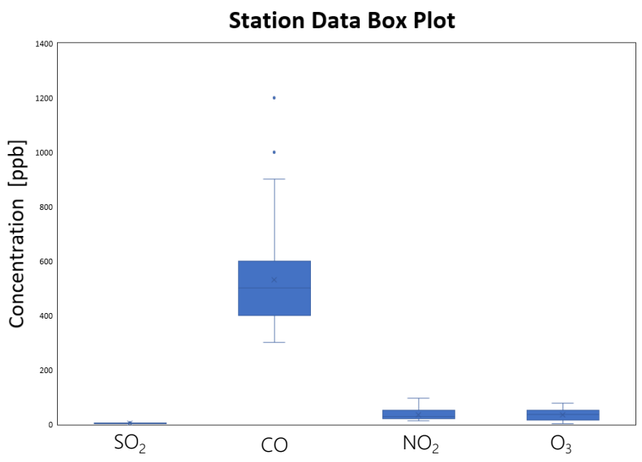

Figure.1 Box Plot of Station Data

Figure 1은 본 연구에서 사용한 Y값(고정 측정소 측정 자료, Station)의 데이터 분포를 나타낸 Box plot이다.

2.2 Sensors

Table.2 Sensor Informations & Statistics

	SO_2	CO	NO_2	O_3
Sensor	SO2-B4	CO-B4	NO2-B43F	OX-B431
Noise [ppb]	5	4	15	15
Max-min [ppb]	3	600	84	75
Average [ppb]	3.52	530.43	34.73	33.32

본 연구에 사용된 센서들은 모두 *Alphasense* 사의 전기화학식 센서이다. Table 2는 본 연구에서 사용한 센서 정보(이름, Noise)와 Figure 1에 대한 통계량이다[4-7]. SO_2 의 경우 평균이 3.52ppb, 최대값-최소값이 3ppb인 자료를 Noise가 5ppb인 센서로 측정했기 때문에 센서 측정값에 대한 신뢰도가 매우 낮다고 할 수 있다.

2.3 DNN

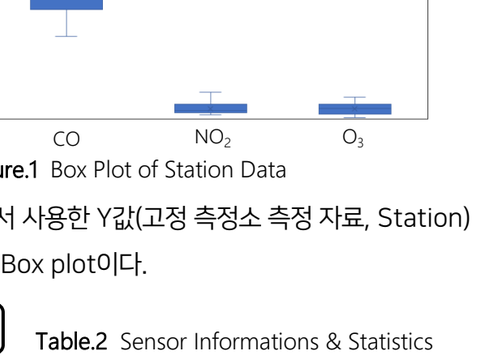

Figure.2 Sample DNN structure

본 연구에서는 순방향 완전 연결 신경망으로 DNN을 구현해 사용하였다. Figure 2와 유사한 구조이며, Sensor 측정값, 기온, 풍속, 습도 4가지 값을 이용해 Station 측정값에 가깝도록 학습을 진행했다. 활성함수로는 정류 선형 유닛(ReLU)을 사용했고, Hidden Layer수, Node 개수 등은 관련 선행 연구를 참고해 선정하였다[8].

2.4 RNN

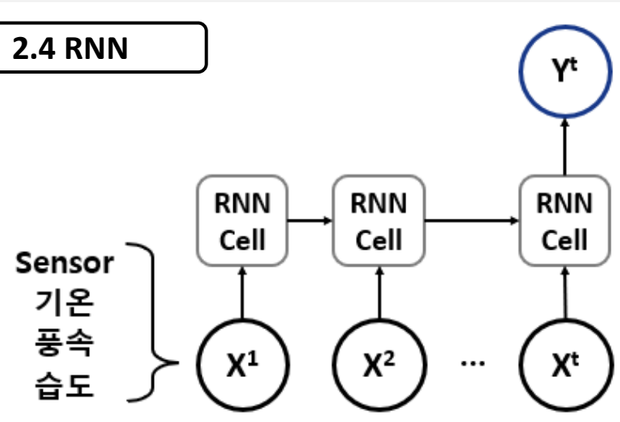

Figure.3 Sample RNN structure

본 연구에서는 Figure 3과 같은 다대일 순환 신경망을 구현해 사용하였다. 해당 순환 신경망은 설정한 길이의 X값 시퀀스를 이용해 Y값을 구한다. 이때, 시퀀스 안에서 각 RNN Cell이 다음 Cell로 정보를 전달하기 때문에, 시계열 데이터 학습에 유리한 것으로 알려져 있으며, 이는 데이터가 시간 종속성을 가지고 있는 본 연구에도 적합한 모델이다[9].

2.5 HybridRNN

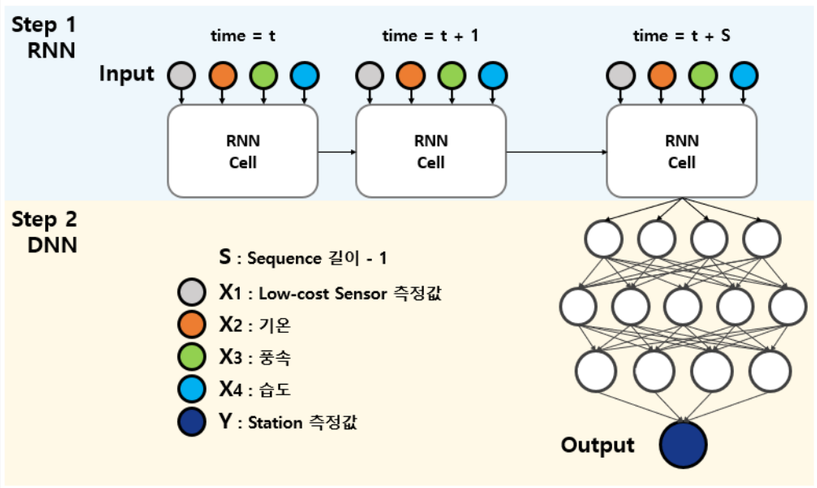

Figure.4 HybridRNN structure

본 연구에서 제시하는 저가 센서 보정 방안은 Figure 4와 같은 HybridRNN 모델이다. 기존의 전기화학식 센서 보정 관련 연구에서는 다중선형회귀, DNN 등 순방향 신경망, RNN 및 LSTM 등 순환 신경망을 이용해 센서 측정값을 보정했다. 하지만 미세먼지를 측정하는 광학식 센서 보정 관련 연구에서는 순환신경망과 순방향 신경망을 융합한 형태의 모델에 대한 연구가 있다. 이 경우, 측정값이 시계열의 데이터라는 특성을 살리며, 입력과 출력 사이 비선형의 관계를 고려해 보정을 할 수 있기 때문에 더 유리하다[8].

HybridRNN 모델의 진행은 다음과 같다.

Step 1) Time sequence의 길이만큼의 Input Data를 RNN에 입력해 시간 종속 변수들을 구한다.

Step 2) 시간 종속 변수들을 DNN에 입력해 Output을 구한다.

2.6 R^2 & RMSE

본 연구에서는 Raw 센서값, 각각 다중선형회귀, RNN, DNN, HybridRNN을 통한 보정값 5가지에 대해 그 결과를 비교 및 평가했다. 다양한 지표들이 회귀 모델 평가를 위해 쓰이는데, 센서 측정값 보정 방법을 평가할 때는 R^2, RMSE 두 가지 지표로 충분하다[10]. R^2과 RMSE 는 다음 식과 같다.

$$R^2 = 1 - \sum_{i=1}^{n} \frac{(t_i - y_i)^2}{(t_i - \bar{t})^2} \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{t}_i)^2}$$

$t_i, y_i, \bar{t}, n$ 은 각각 i번째 이론값, i번째 예측값, 이론값의 평균, 샘플 수를 의미한다.

3. Results & Discussion

3.1 Calibration

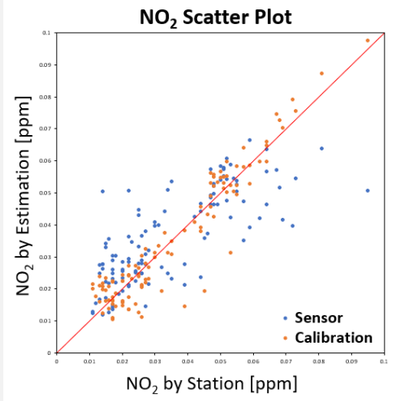

Figure.5 NO_2 Scatter Plot for Raw Sensor, Calibration result

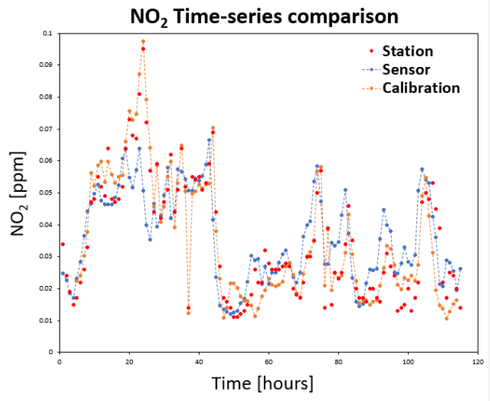

Figure.6 NO_2 Time-series comparison for Raw Sensor, Calibration result

Figure 5와 6은 NO_2 를 측정한 Raw 센서값과 보정 결과를 각각 Scatter Plot과 Time-series 로 비교한 결과이다. 이 때, Raw 센서값의 R^2 값과 RMSE 값은 각각 0.59, 11.90ppb 이며 보정 결과의 R^2 값과 RMSE 값은 각각 0.90, 6.48ppb 이다. R^2 값이 클 수록, RMSE 값이 작을 수록 이론값(Station)에 더 가까운 것을 확인할 수 있다.

3.2 Benchmarking

Table.3 R^2 (Left) and RMSE (Right) value for different methods

	R^2					RMSE [ppb]				
	Sensor	MLR	DNN	RNN	Hybrid	Sensor	MLR	DNN	RNN	Hybrid
SO_2	0.02	0.11	0.04	0.05	0.07	19.96	0.61	2.59	3.22	1.97
CO	0.62	0.63	0.74	0.90	0.93	115.43	100.47	85.77	53.07	45.03
NO_2	0.59	0.67	0.77	0.79	0.90	11.90	10.43	9.29	11.64	6.48
O_3	0.63	0.64	0.74	0.83	0.85	14.05	12.83	15.21	10.96	8.73

실험 결과 Table 3과 같이 SO_2 를 제외한 모든 대기오염 물질에서 본 연구에서 제안한 HybridRNN 모델을 통해 보정했을 때 R^2 값이 가장 크고, RMSE 값이 가장 작은 것을 확인할 수 있다.

SO_2 측정 센서의 경우, 센서의 정확도로 인해 학습이 되지 않을 정도로 데이터 신뢰도가 매우 낮다. SO_2에 대한 국내 대기환경 기준은 1시간 평균치 150ppb 이하이고, 본 연구에서 측정한 기간 동안 최대 6ppb였다[1]. 따라서 SO2 농도가 더 높은 곳에서 학습을 진행한다면 만족스러운 결과를 얻을 것으로 예상할 수 있다.

4. Conclusion

본 연구에서는 광학식 센서 보정 관련 연구에서 제안한 HybridLSTM 모델에서 아이디어를 얻어 HybridRNN 모델을 구현하여, 전기화학식 센서 보정에도 사용할 수 있다는 것을 검증하였다[8]. 센서값, 다중선형회귀, DNN, RNN 4가지 결과와 보정 성능을 비교했으며 본 연구에서 제안한 HybridRNN 모델이 가장 우수한 것을 확인했다. 이 모델은 기존 모델과 비교했을 때, (1) 대기오염물질 측정값 및 기온, 풍속, 습도의 시간 종속성, (2) Input 변수들과 Output 사이 비선형 관계를 고려할 수 있다는 점에서 더 우수한 성능을 낸다고 생각한다.

하지만 센서값을 보정하는 새로운 머신러닝 모델을 제시하는 기존 연구들의 대다수는 training set 1000개 이상, test set 100개 이상의 데이터를 사용한 반면, 본 연구에서는 training set 100개, test set 14개의 데이터를 사용했다. 따라서 데이터의 부족으로 모델 Hyper-Parameter의 Optimization이 잘 되었는지 확인하기 어렵다는 한계가 있다.

해당 모델을 통해 저가의 센서 측정값을 보정하여 정확한 대기오염 물질의 농도를 알 수 있게 된다면, 저가의 센서를 이용한 대기오염 감지 네트워크를 구축해 도시의 효율적인 대기오염 관리에 큰 도움이 될 것이다.

5. References

[1] Airkorea (한국환경공단), 국내 대기환경 기준 및 측정자료.

[2] Prashant Kumar, et al, "The rise of low-cost sensing for managing air pollution in cities", Environment International, 2015, 75, pp.199~205.

[3] 기상청, 공공기관 기상관측 자료.

[4] Alphasense, "SO2B4 Datasheet", 2017.

[5] Alphasense, "COB4 Datasheet", 2019.

[6] Alphasense, "NO2-B43F Datasheet", 2019.

[7] Alphasense, "OX-B431 Datasheet", 2019.

[8] Park, D. et al, "Assessment and Calibration of a Low-Cost PM2.5 Sensor Using Machine Learning (HybridLSTM Neural Network): Feasibility Study to Build an Air Quality Monitoring System", Atmosphere, 2021, 12, 1306.

[9] Han, P. et al, Calibrations of Low-Cost Air Pollution Monitoring Sensors for CO, NO2, O3, and SO2. Sensors 2021, 21, 256.

[10] Zusman, M. et al, "Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study", Environ. Int, 2020, 134, 105329.