https://doi.org/10.21289/KSIC.2021.24.3.323

유해가스 배출량에 대한 시계열 예측 모형의 비교연구

A Comparison Study of Forecasting Time Series Models for the Harmful Gas Emission

장문수[1], 허요섭[2], 정현상[2], 박소영[1*]

Moonsoo Jang[1], Yoseob Heo[2], Hyunsang Chung[2], Soyoung Park[1*]

〈Abstract〉

With global warming and pollution problems, accurate forecasting of the harmful gases would be an essential alarm in our life. In this paper, we forecast the emission of the five gases(SOx, NO_2, NH_3, H_2S, CH_4) using the time series model of ARIMA, the learning algorithms of Random forest, and LSTM. We find that the gas emission data depends on the short-term memory and behaves like a random walk. As a result, we compare the RMSE, MAE, and MAPE as the measure of the prediction performance under the same conditions given to three models. We find that ARIMA forecasts the gas emissions more precisely than the other two learning-based methods. Besides, the ARIMA model is more suitable for the real-time forecasts of gas emissions because it is faster for modeling than the two learning algorithms.

Keywords : Time-series forecasting, ARIMA, Random Forest, LSTM, Harmful gas emission

1 정회원, 부산대학교(Pusan National University), 통계학과
2 정회원, 한국과학기술정보연구원(KISTI), 부산울산경남지원
1* 교신저자, 부산대학교 통계학과, 조교수
E-mail: soyoung@pusan.ac.kr

1 Department of Statistics, Pusan National University
2 Busan · Ulsan · Gyeongnam Branch, Korea Institute of Science and Technology Information(KISTI)
1* Corresponding Author, Assistant Professor, Department of Statistics, Pusan National University,
E-mail: soyoung@pusan.ac.kr

1. 서 론

유해가스 배출량 예측의 정확도는 공장 주변에서의 공기의 질을 제어하고, 환경오염에 관한 정책을 결정하는 데 있어서 중요한 문제이다. 공장 주변에서 발생하는 유해 가스들은 그 종류가 다양할 뿐만 아니라, 실시간으로 각 가스량이 변화한다. 따라서, 실시간 상황에 대처가 가능한 시계열 예측모형을 통해 단기간의 미래 유해 가스량을 예측함으로써, 위험수위를 넘는 양에 대하여 사전에 대응할 수 있는 시스템을 갖추고자 하는 수요가 지역 제조산업을 중심으로 증가하고 있는 추세이다[1].

기존 연구에 따르면, Sen et al.[2]은 ARIMA (Autoregressive integrated Moving Average) 모형을 통한 온실가스 배출량의 예측을 시도하였고, Fang et al.[3]은 가우시안 프로세스 회귀(Gaussian process regression)를 이용하여 이산화탄소 배출량에 대한 예측을 시도하였다. 최근에는 머신러닝(machine learning), 딥러닝(deep learning)과 같이 방대한 데이터를 학습시켜 예측을 시도하는 알고리즘들을 이용한 연구가 주목을 받고 있다. 일례로, Radojević et al.[4]의 연구에서는 인공신경망(ANN, Artificial Neural Network)을 이용하여 온실가스 배출량의 예측을 시도하였다. 그러나 기존의 연구들은 온실가스 배출량이라는 세부적인 연구 주제에 집중되어 있어, 공장 주변의 유해가스 배출량에 대한 예측과는 그 모형의 적합도와 예측 성능이 다를 가능성이 있다.

따라서 본 연구에서는 유해가스 배출량에 대한 실증적인 자료를 바탕으로, 통계적인 예측 모델과 학습 기반의 알고리즘을 사용하여 예측 결과를 비교해 각 모델의 성능을 비교하는 연구를 수행하고자 한다.

2. 시계열 예측모형

본 연구에서는 세 가지의 시계열 예측모형 사용하며, 첫 번째로 대표적인 시계열 예측모형인 ARIMA, 두 번째로 기계학습알고리즘 중 하나인 랜덤 포레스트(RF, Random forest), 세 번째로는 딥러닝 알고리즘인 LSTM(Long Short-Term Memory) 모델을 비교하고자 한다.

2.1 ARIMA

시계열 데이터는 일반적인 데이터와는 다르게 자기 상관성(Auto Correlation)이 존재하는 데이터로, 현재의 값은 이전의 값에 영향을 받는 데이터를 말한다. 이러한 시계열 자료의 분석에 사용되는 대표적인 통계 방법으로는 ARMA (Autoregressive Moving Average), ARIMA (Autoregressive integrated Moving Average), SARIMA(Seasonal ARIMA) 등이 있다.

ARMA 모형은 AR(Autocorrelation)모형과 MA (Moving average)모형을 합친 것으로, 현재 시계열 값은 과거의 데이터와 오차에 의해 설명되는 모형이다. ARIMA모형은 차분(difference)과 함께 ARMA모형을 사용하며, 비정상(non-Stationary) 시계열에도 사용할 수 있는 장점이 있다. SARIMA 모형은 계절성 데이터에 대해 효과적인 모형이다. 시계열자료를 Z_t라고 할 때, ARIMA(p,d,q)모형의 식(1)과 같이 나타난다.

$$\begin{aligned} & \phi_p(B)(1-B)^d(Z_t-\mu)=\theta_q(B)\epsilon_t \\ where \quad & \theta_q(B)=1-\theta_1B-\ldots-\theta_qB^q \\ & \phi_q(B)=1-\phi_1B-\ldots-\phi_pB^p \end{aligned} \tag{1}$$

B는 후행 연산자(Back shift operator), μ는 평균, d는 몇 번 차분했는지를 을 의미한다. $\phi_p(B)$와 $\theta_q(B)$는 각각 p차, q차 다항식으로 p와 q는 AR(p), MA(q) 모형의 모수를 말한다.

2.2 랜덤 포레스트(Random forest)

랜덤 포레스트[5]는 앙상블 모델 기반의 기계학습 알고리즘으로, 종속변수의 분류(classification) 또는 예측(regression) 시에 사용된다. 주어진 설명변수와 변수를 무작위 표본 추출하여 각 시뮬레이션 상황별로 의사 결정 나무(decision tree)를 형성한다. 이를 통해 최대한 많은 비상관화된 의사결정 나무(uncorrelated decision tree)를 구성하고, 새로운 예측을 위한 설명변수가 입력되었을 때, 모든 의사결정 나무의 결정으로부터 다수의 결정을 최종 결정 값으로 선택한다.

기존 연구로부터, 랜덤 포레스트 알고리즘이 이미지 분류, 변수선택 등에 많이 사용되나 시계열 자료의 예측에는 많이 사용되지 않는다는 것을 확인하였다. 이에 본 연구에서 랜덤 포레스트 알고리즘을 시계열 예측모형으로 사용하기 위하여 종속변수 한 개를 현재값, 설명변수를 여러 개의 직전 과거값으로 구성하여 학습시키는 방식을 사용하였다.

2.3 LSTM

Hochreiter et al.[6]에 의해 개발된 LSTM (Long Short-Term Memory) 모델은 순차적인 데이터를 학습시키는 딥러닝 모델 중 하나로, RNN (Recurrent Neural Network)의 단점을 보완한 모델이다. RNN은 순서가 있는 데이터를 학습시키고 처리하는데 적합한 모델이지만, 기울기의 소실(gradient vanishing) 문제나 기울기 폭발(gradient exploding)문제 때문에 장기적인 데이터는 기억하지 못하는 장기기억 의존성(Long-term dependency)라는 치명적인 단점이 있다. LSTM은 RNN에 메모리 셀(memory cell)이라는 새로운 노드를 추가하여 기울기의 소실이나 기울기 폭발 문제를 해결해 장기기억 능력을 향상시킨 모형이다.

이러한 딥러닝 기반의 순차적인 데이터에 대한 모형들은 자연어 처리(NLP, Natural Language Processing)[7], 음성 인식(speech recognition)[8], 시계열 예측(time-series forecasting)[9] 등 광범위하게 사용되고 있다. RNN과 LSTM에 대한 자세한 내용은 Sherstinsky[10]에서 확인할 수 있다.

3. 유해가스 배출량 예측모형의 성능 비교

3.1 데이터 소개

본 연구에서 사용한 데이터는 2021년 2월 1일부터 2021년 2월 20일까지 수집한 데이터이며, 수집 장소는 석유화학 및 종합 화학공장들이 다수 위치한 울산테크노파크(울산광역시 두왕동 소재) 건물 옥상에서, 유해가스 배출량을 검출하는 센서를 이용해 측정한 데이터다. 설치된 센서는 약 10초 간격으로 여러 종류의 유해가스 배출량이 측정되고 있으며, 황산화물(SOx), 이산화질소(NO_2), 암모니아(NH_3), 황화수소(H_2S), 메탄(CH_4) 등의 총 다섯 종의 유해가스의 배출량을 측정하였다. 공단에서 배출되는 가스는 공기 중으로 확산이 되고, 옥외에 설치된 센서함에는 팬이 돌아가며 공기를

빨아들이게 된다. Fig. 1에는 실제 설치된 센서의 모습을 나타내었고, Table 1에는 사용한 센서에 대한 상세 정보를 수록하였다.

(a) Installed sensor box

(b) View of sensor box and Ulsan industrial complex

Fig. 1 Actual installed sensor box and view of industrial complex

Fig. 2에는 다섯 가지 유해가스 배출량을 2월 1일부터 20일간 측정된 값에 대한 시계열도표를 나타내었다. 붉은색 점선을 기준으로 각각 10일씩 데이터를 나누었으며, 2월 1일부터 10일까지의 데이터를 학습용(training), 11일부터 20일까지의 데이터를 테스트용(testing)으로 사용한다.

Table 2는 5가지 유해가스 배출량 데이터의 전체 기간에 대한 평균과 표준편차, 정상성(stationary)검정인 KPSS(Kwiatkowski-Phillips-Schmidt-Shin) 검정결과[11]를 나타낸다. 5가지 유해가스 배출량의 원자료에 대하여 정상성 검정결과 모두 정상성을 갖는다는 귀무가설이 기각되었으나, 1차 차분($Z_t - Z_{t-1}$)을 시도하자 모두 정상성을 갖는다는 검정결과를 나타내었다.

Table 1. Specifications of sensors

Gas type	Manufacturer	Output range	Resolution
SOx	Tesla ENG (Korea, Republic of)	0~100 ppm	≤0.1 ppm
NO_2		0~100 ppm	≤0.1 ppm
NH_3		0~200 ppm	≤0.1 ppm
H_2S		0~200 ppm	≤0.1 ppm
CH_4		0~60,000 ppb	≤30 ppb

Table 2. Basic statistics for each harmful gas

Gas type	Mean	Std.	KPSS (original)	KPSS (1-lag diff.)
SOx	16.60	4.95	18.96 **	0.004
NO_2	37.64	8.34	6.42 **	0.001
NH_3	47.57	22.78	12.30 **	0.002
H_2S	45.30	18.39	9.32 **	0.002
CH_4	1592	423.4	7.08 **	0.004

Std. : Standard deviation
Significance codes : **, * for 1% and 5% levels;
n = 132,487 (the number of data for 02/01~02/20)

Fig. 3에는 황산화물(SOx)에 대한 자기 상관함수(Autocorrelation function)를 나타내었으며, Fig. 3(a)와 Fig. 3(b)는 각각 원 데이터와 1차 차분 된 데이터에 대한 자기상관계수(ACF, Autocorrelation coefficient)를 나타낸 그래프이다. 왼쪽의 그림을 보면 자기 상관 함수가 시차가

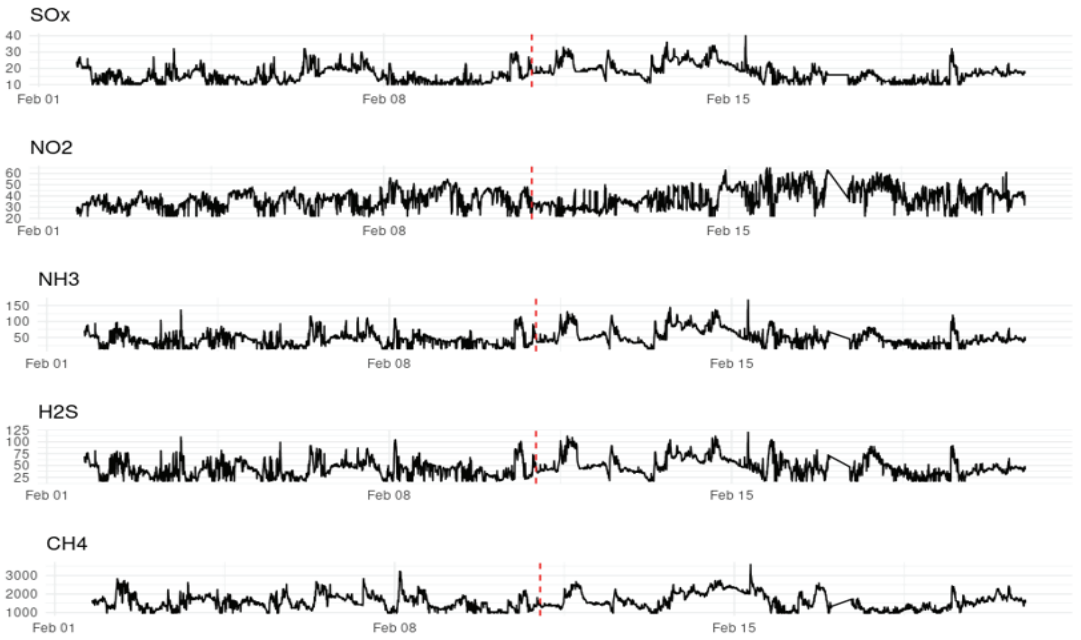


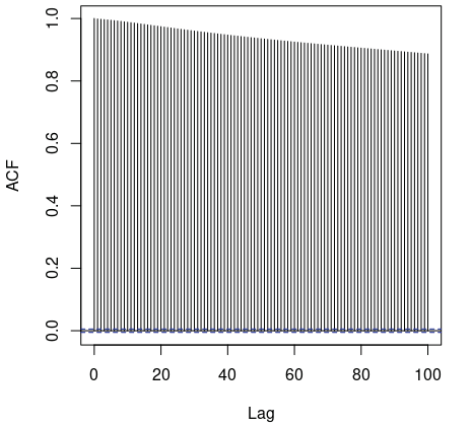
Fig. 2 Time-series plots for five harmful gases(SOx, NO_2, NH_3, H_2S, CH_4)

증가해도 매우 천천히 감소하고 있지만, 1차 차분 후에는 다음 시차 간의 자기 상관성이 확연히 감소하는 것이 확인되었다. 이는 유해가스 배출량의 현 시차와 지난 시차 간의 종속성이 시간 차이가 증가하더라도 지속되고 있으나, 차분을 통해 데이터의 추세를 제거해주면 시차 간 종속성이 확연히 감소함을 의미한다. 따라서 데이터가 랜덤워크(random walk)의 특징을 보이고 있음이 확인되었으며, 다른 4개 가스의 배출량에 대해서도 ACF 결과가 비슷하게 나타나, 유해가스 배출량은 모두 랜덤워크의 특징을 보임이 확인되었다.

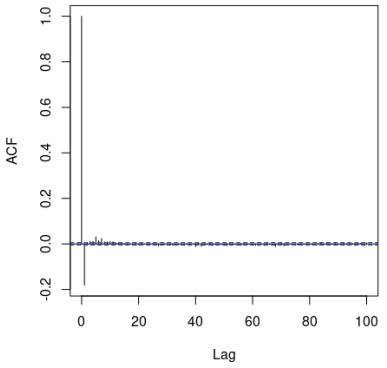
3.2 모형 적합 결과

본 연구에서는 ARIMA, 랜덤 포레스트, LSTM 모형이 유해가스 배출량을 예측하는 것에 대해 같은 기간에 대한 예측을 실시하며 성능을 비교하고자 한다. ARIMA모형은 다른 두 모형과는 달리 학습 과정이 없고, 주어진 기간 내에 반복적으로 적합(fitting)을 실시하여 최적의 모수($\hat{p},\hat{d},\hat{q}$)를 추정한 뒤, 생성된 모델을 사용해 다음 예측값을 계산하는 과정을 거친다.

기계학습알고리즘인 랜덤 포레스트와 LSTM 알고리즘은 ARIMA와는 달리 학습 과정이 필요하다. 따라서 데이터의 학습을 위해 2월 1일부터 10일간의 총 66,126개의 데이터를 Fig. 4와 같이 분할한 뒤 뒤 모형에 학습시켰다. Fig. 4는 시차가 10일 때의 학습용 데이터의 형태를 나타내고 있다. 예를 들어 현재 시점 X_{11}을 종속변수로, 과거 10개의 가스량 값인 $X_1, \cdots, X_{10}$을 학습용 설명변수로 설정하는 방식이다. 만약 시차가 30인 경우, Fig. 4에서 X train에 들어가는 데이터의 개수가 30개가 된다. ARIMA모형 적합시에는 가장 가까운 과거 10개 또는 30개를 사용하였으며, 주로



(a) ACF plots of SOx



(b) ACF plots of differenced SOx

Fig. 3 ACF plots of two types of SOx

ARIMA(0,0,0) 또는 ARIMA(0,1,0)으로 적합된 결과를 얻었다.

세 모형을 이용하여 각 유해가스 배출량의 예측에 사용되는 학습용 데이터의 기간은 같은 기간으로 고정하였고, 표본 외 예측(out-of sample forecasting)을 통해 예측력을 비교한다.

5개의 유해가스 중 황산화물(SOx)가스 배출량에 대한 예측 결과를 Fig. 5에 나타내었다. 검은색 선은 예측해야 할 황산화물의 실제 값(original)을 의미하며, 빨간색 선은 ARIMA 모형의 예측결과, 연두색 선은 LSTM 모형의 예측결과, 파란색 선은 랜덤 포레스트 모형의 예측결과

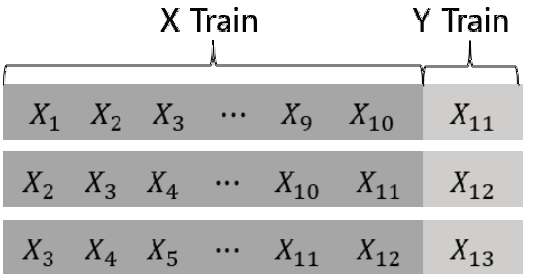


Fig. 4 One step ahead out-of-sample forecasting for LSTM and RF

를 의미한다. Fig. 5의 결과는 과거 30개의 자료를 사용하여 모형 적합과 예측에 사용하였다.

3.3 모형의 성능 비교

최종적으로 66,361개의 테스트 데이터를 사용하여 세 모형에 적합하여 예측결과를 비교하였으며, 일반적으로 예측모델의 성능 비교에 사용되는 지표들인 RMSE(Root Mean Squared Error), MAE(Mean Absolute Error), MAPE(Mean Absolute Percentage Error)의 세 가지 값을 모형의 예측 성능 비교에 사용한다. 식(2)-(4)에 각 지표의 계산식을 나타내었으며, 세 지표 모두 값이 낮을수록 좋은 모형이라고 판단할 수 있다.

Table 3에 각 모델에 대한 세 가지 지표들의

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \quad (2)$$

$$MAE = \frac{\sum_{i=1}^{n}|Y_i - \hat{Y}_i|}{n} \quad (3)$$

$$MAPE = \frac{\sum_{i=1}^{n}\frac{|Y_i - \hat{Y}_i|}{Y_i} \times 100}{n} \quad (4)$$

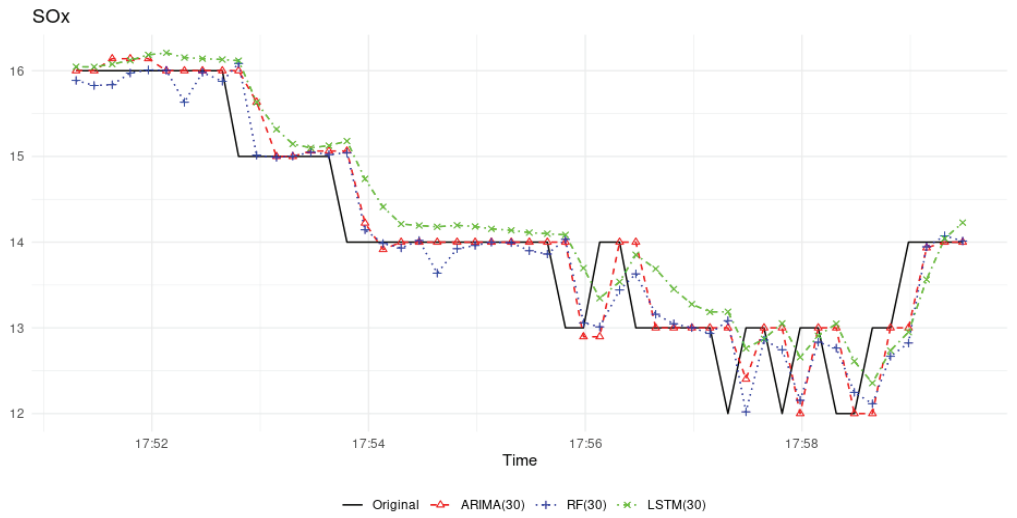


Fig. 5 Results of the predicted values of SOx gas for 50 time steps

값을 나타내었다. 유해가스의 종류와 시차별로 가장 낮은 값을 굵게 표시했으며, 소수점 넷째 자리까지 나타냈다. 전체적으로 LSTM이 가장 높은 RMSE, MAE, MAPE 값을 가지고 있으며, 랜덤 포레스트가 ARIMA보다 약간 더 높은 값을 가지는 것을 확인할 수 있다. LSTM모형의 경우, 가스 배출량의 변동성을 잘 이해하지 못하여 급격한 감소 또는 증가에 약하다는 특징을 보여주고 있다. ARIMA 모형이 대부분의 경우에서 가장 좋은 모델로 선택된 것을 볼 수 있으며, 시차가 10일 때보다 30일 때 모든 경우에서 RMSE, MAE, MAPE 값이 작아지는 것을 볼 수 있다. 이와 반대로 LSTM과 RF는 시차가 커질수록 각 지표의 값들이 상승하는 경향을 보인다.

4. 결 론

본 연구에서 다섯 가지의 유해가스 배출량을 예측하는 모형으로 ARIMA, 랜덤 포레스트, LSTM을 사용해 예측 성능을 비교하였다. 본 연구에서 사용한 데이터에 대해서는 ARIMA 모형이 기계학습을 기반으로 하는 랜덤 포레스트나 LSTM보다 더 좋은 성능을 나타내었다. 실제 모델 적합을 수행했던 대부분의 기체에서 ARIAM 모형이 다른 두 모델에 비해 더 좋은 성능을 나타냈고, 성능지표로도 가장 좋은 값을 보였다. 일례로 SOx의 경우, Time lag가 30일 때, ARIMA는 RMSE 값이 0.2888(LSTM 0.3548, RF 0.3420), MAE 값은 0.1016(LSTM 0.2283, RF 0.1251), MAPE 값은 0.0062(LSTM 0.0136, RF 0.0072)로 가장 낮았다. 이는 랜덤 포레스트와 LSTM 알고리즘에 지난 10일간의 유해가스 데이터를 가지고 학습을 시켰으나, 유해가스 데이터의 특성이 10일이라는 상대적으로 오래된 과거 데이터의 패턴에 의존하기보다는 가까운 과거 시점의 데이터에 더 의존하는 특성이 있었기 때문으로 풀이된다.

Time lag가 10일 경우에는 ARIMA 보다 랜덤 포레스트가 더 좋은 성능지표를 나타내는 경우도 있었다. 예를 들어, SOx의 경우 MAPE 값이 RF

Table 3. Results of each model for every harmful gas with both time lag

Gas type	Model	Time lag = 10			Time lag = 30		
		RMSE	MAE	MAPE	RMSE	MAE	MAPE
SOx	LSTM	0.9447	0.8888	0.0537	0.3548	0.2283	0.0136
	RF	0.3336	0.1070	**0.0062**	0.3420	0.1251	0.0072
	ARIMA	**0.3019**	**0.1056**	0.0065	**0.2888**	**0.1016**	**0.0062**
NO_2	LSTM	1.6249	0.9382	0.0272	2.1338	1.6566	0.0423
	RF	1.4636	0.4640	0.0112	1.5084	0.5326	0.0132
	ARIMA	**1.4179**	**0.3334**	**0.0091**	**1.4007**	**0.3290**	**0.0089**
NH_3	LSTM	1.7183	1.2006	0.0336	2.0590	1.5574	0.0333
	RF	1.4031	0.6271	0.1448	1.7086	0.7137	0.0155
	ARIMA	**1.2711**	**0.5924**	**0.0149**	**1.2005**	**0.5688**	**0.0142**
H_2S	LSTM	1.2386	0.9007	0.0235	1.4301	1.0481	0.0266
	RF	**0.7768**	**0.3738**	**0.0088**	0.8557	0.4223	0.0096
	ARIMA	0.8422	0.3766	0.0092	**0.7974**	**0.3624**	**0.0089**
CH_4	LSTM	43.3419	36.6836	0.0267	29.8359	23.0421	0.0166
	RF	**19.8263**	11.7781	0.0079	19.5671	11.1294	0.0075
	ARIMA	19.8845	**9.7910**	**0.0067**	**18.7459**	**9.3614**	**0.0064**

RMSE = root mean squared error; MAE = mean absolute error; MAPE = mean absolute percentage error
RF = Random Forest

가 다른 두 모델에 비해 0.0062로 가장 낮았으며, H_2S의 경우는 RF가 RMSE 값은 0.7688(LSTM 1.2386, ARIMA 0.8422), MAE 값은 0.3738(LSTM 0.9007, ARIMA 0.3766), MAPE 값은 0.0088(LSTM 0.0235, ARIMA 0.0092)로 가장 좋은 성능을 나타냈다. 그러나 그때에도 랜덤 포레스트와 ARIMA 모형의 성능지표 차이는 크게 나타나지 않은 것을 확인할 수 있다.

이와 더불어, 유해가스의 예측은 그 특성상 실시간으로 이루어져야 하며, 유해가스가 기준 농도보다 높아지게 되는 경우, 즉각적인 후속 조치가 필요하다. 따라서 모형 학습에 시간이 많이 필요한 랜덤 포레스트나 LSTM은 실시간 가스량 예측에 적절하지 않을 수 있다. 따라서, 유해가스 배출량은 오래된 과거의 패턴보다 가까운 과거에 더 큰 영향을 받기 때문에 최근 과거 값들을 이용하여 최적의 적합을 하는 ARIMA 모형이 기계학습 방법인 랜덤 포레스트나 LSTM보다 더 좋은 성능과 빠른 예측을 보인다는 점에서 유해가스 배출량 예측에 도움이 될 수 있다고 판단된다.

그러나 각 모델에 대해서 세부적인 하이퍼 파라미터 튜닝(hyper-parameter tuning)이나 모델 최적화 과정을 거치지 않은 채, 전반적인 예측 성능 비교만을 했기 때문에 본 연구결과만으로 단정적인 모델 간 비교는 어렵다고 볼 수 있다. 향후 연구를 통해 이러한 부분에 대한 연구를 통해 유해가스 배출 예측에 대한 고도화된 모델 구축이 가능해진다면 제조산업 및 화학 공단에서 유해가스 누출 사고로 인한 재산상의 손해와 인명 피해를 사전에 차단하거나 최소화하는 데 기여할 것으로 사료된다.

사 사

본 연구는 2020학년도 부산대학교 신임교수연구정착금 지원사업과, 과학기술정보통신부 과학기술기반 지역수요맞춤형 R&D지원 사업으로부터 한국과학기술정보연구원을 통해 지원받아 수행되었습니다(CN20120US001).

참고문헌

[1] Marshall, J. Park, S. Park, “Transfer Learning 기법을 이용한 가스 누출 영역 분할 성능 비교,” 한국산업융합학회 논문집, 23(3), pp.481-489, (2020).

[2] P. Sen, M. Roy, and P. Pal, “Application of ARIMA for forecasting energy consumption and GHG emission: A case study of an Indian pig iron manufacturing organization.” Energy, 116, pp. 1031-1038, (2016).

[3] D. Fang, X. Zhang, Q. Yu, T. C. Jin, and L. Tian, “A novel method for carbon dioxide emission forecasting based on improved Gaussian processes regression,” Journal of cleaner production, 173, pp. 143-150, (2018).

[4] D. Radojević, V. Pocajt, L. Popović, A. Perić-Grujić, and M. Ristić, “Forecasting of greenhouse gas emissions in Serbia using artificial neural networks,” Energy Sources, Part A: Recovery, Utilization, and Environmental Effects, 35, pp. 733-740, (2013).

[5] L. Breiman. “Random forests,” Machine learning 45, pp. 5-32, (2001)

[6] S. Hochreiter and J. Schmidhuber. “Long short-term memory,” Neural computation 9.8 1735-1780, (1997).

[7] Z. Huang, W. Xu and K. Yu. “Bidirectional LSTM-CRF models for sequence tagging,” arXiv preprint arXiv:1508.01991, (2015).

[8] A. Graves, N. Jaitly, and A.R. Mohamed. “Hybrid speech recognition with deep bidirectional LSTM,” IEEE workshop on automatic speech recognition and understanding, (2013).

[9] A. Sagheer and M. Kotb. “Time series forecasting of petroleum production using deep LSTM recurrent networks,” Neurocomputing, 323, pp. 203-213, (2019).

[10] A. Sherstinsky “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network,” Physica D: Nonlinear Phenomena, 404, (2020).

[11] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, “Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?,” Journal of econometrics, 54(1-3), pp. 159-178, (1992).

(접수: 2021.05.04. 수정: 2021.05.24. 게재확정: 2021.06.04.)