Journal of the Korean Data & Information Science Society
2023, **34**(5), 763–773

http://dx.doi.org/10.7465/jkdi.2023.34.5.763
한국데이터정보과학회지

머신러닝을 이용한 미세먼지 예측 연구[†]

김삼용[1]

[1]중앙대학교 응용통계학과

접수 2023년 7월 24일, 수정 2023년 8월 7일, 게재확정 2023년 8월 11일

요 약

미세먼지란, 대기 중에 떠다니거나 흩날려 내려오는 입자상물질인 먼지 중 입자의 지름이 10㎍ 이하인 먼지를 말하며, 으로 표기하기도 한다. 이러한, 미세먼지는 매우 작은 크기로, 코나 기관지에서 걸러지지 않고 몸속에 스며들어 천식과 폐질환 또는 면역세포의 작용을 통해 염증을 일으키기도 한다. 최근 한국이 세계적으로 미세먼지 농도가 가장 높은 국가인 것이 밝혀졌는데, 미세먼지는 건강뿐 아니라 생태계 및 농작물에도 직접적인 영향을 미치기 때문에 정확한 예보 시스템을 통한 대책마련을 강구하는 것이 중요하다. 따라서, 본 논문에서는 기상청이 제공하는 기상 데이터와 에어코리아에서 제공하는 대기오염물질 데이터를 이용하여 미세먼지 농도의 머신러닝 예측 성능 비교를 하고자 하였다. 지역으로는 황사의 유입 경로인 산둥반도와 가장 인접한 인천광역시의 데이터를 추출하였고, 인천시의 다양한 기상 요인 및 대기오염물질들의 상관관계 확인 후, 모형을 구축하였다. 모형으로는 MLP, RNN, LSTM, GRU 그리고 CNN을 사용하였고, 기본적인 하이퍼파라미터와 단일층으로 구성하여 예측 성능을 비교하였다. 그 후, GRU1 (단일층) 모형에 층을 추가한 GRU2 모형을 새롭게 구성하여 가장 예측 성능이 좋았던 GRU1 모형과 비교해보았다. 예측 성능은 테스트 데이터에서 MAE와 RMSE로 평가하였다. 대부분 비슷한 예측 성능을 보였지만, GRU1 모형이 MAE 8.80, RMSE 14.61로 다른 모형들에 비해 가장 성능이 우수하다는 것을 확인할 수 있었다. 가장 예측 성능이 낮은 모형은 MLP 모형이며, 그 뒤로는 RNN, LSTM, GRU2, CNN 순으로 예측 성능이 우수하였다.

주요용어: 머신러닝, 미세먼지 예측, CNN, GRU, LSTM.

1. 머리말

미세먼지란, 대기 중에 떠다니거나 흩날려 내려오는 입자상물질인 먼지 중 입자의 지름이 10㎍ 이하인 먼지를 말하며, PM_{10}으로 표기한다. 이러한, 미세먼지는 매우 작은 크기로, 코나 기관지에서 걸러지지 않고 몸속에 스며들어 천식과 폐질환 또는 면역세포의 작용을 통해 염증을 일으키기도 한다. 그런데 최근 한국이 전 세계에서 미세먼지 농도가 가장 높은 국가라는 연구 결과가 밝혀져 논란이 일고 있다. 실제 OECD 자료에 따르면, 2019년 기준 한국의 미세먼지 (PM_{10}및 $PM_{2.5}$) 농도는 27.4㎍/㎥으로 비교대상 국가들 중 매우 높은 수준이다. OECD 평균치 (13㎍/㎥)보다 2배 정도 높으며, 핀란드 (5.6㎍/㎥)에 비하면 4배 이상 심한 수치이다. 이러한 미세먼지는 인체, 생태계 및 농작물에 직접적인 영향을 미칠 수 있기 때문에 최근 국내외에서 중요한 이슈로 떠오르고 있다.

한국에서의 미세먼지에 대한 환경기준은 2015년부터 시행되고 있으며, 해당 연도에 정부에서는 미세먼지를 대기환경기준물질로 지정하여 기준을 설정 및 예보시스템을 개발할 예정이라고 발표하였다

[†] 이 논문은 2022년도 중앙대학교 연구년 결과물로 제출됨.
[1] (06974) 서울특별시 동작구 흑석로 84, 중앙대학교 응용통계학과, 교수. E-mail: sahm@cau.ac.kr

(Kim과 Yeo, 2019). 또한 2013년 이후, 언론과 세계보건기구(World Health Organization; WHO)에서 미세먼지를 발암물질로 분류 및 보도에 따라 현재 미세먼지 농도에 대한 국민들의 불안이 높은 상황이다. 하지만 미세먼지는 산업 등 여러요인들과 복잡하게 연관되어 있어 완전한 제거는 불가능하다. 따라서 정확한 미세먼지 농도 예보 시스템을 이용하여 예측한다면 다양한 대안법 마련이 가능할 것이다. 이를 위하여 국내외에서는 미세먼지 농도 예측 성능 개선 연구가 활발하게 진행되고 있다 (HanJoo Lee 등, 2023). 특히, 인공지능 알고리즘의 발전에 따라 머신러닝 및 딥러닝 모형의 연구가 계속되고 있다. 예를 들면, 국외에서는 Sharma 등 (2022)은 호주의 시간별 미세먼지 농도 데이터와 12개 위성과 대기오염 물질 데이터를 이용하여 CNN (Convolutional neural networks), GRU (Gated recurrent unit), LSTM (Long-short term memory) 등 머신러닝 모형 중 CNN-GRU 모형이 PM_{10} 농도를 예측하는 데에 가장 효과적이며, 기존의 다른 딥러닝 모형에 비해 더 낮은 예측 오차를 보이는 것을 확인하였다. 또한, Kujawska 등 (2022)는 폴란드의 2017년부터 2019년까지의 시간별 대기오염물질과 기상변수들에 SVM (Support vector machine), ANN (Artifical neural network), LSTM 등의 모형을 예측에 적용하였고, ANN 모형이 다양한 시간대에서 PM_{10} 농도를 예측하는 데에 효과적인 것을 밝혔다. Plocoste 등 (2023)은 2005년부터 2012년까지의 일별 평균 대기오염 물질 데이터로 SVR (Suport vector regresson), KNN (K-nearest neighbors), GBR (Gradient boosting regression) 등 모형 중, GBR 모형이 가장 우수한 예측 성능을 보이는 것을 확인하였다. 이 연구는 특히 카리브 해역에서의 PM_{10} 분포 특성을 강조하였고, 향후 연구에서는 추가 매개 변수를 사용하여 모델의 성능을 더욱 개선할 수 있을 것이라 주장했다. Adnane 등 (2022)는 아가디르, 모로코 지역의 미세먼지 농도 데이터와 온도, 습도, 풍속, 풍향 및 대기오염물질인 SO2 (아황산가스 농도), PM10 (미세먼지 농도), NO2 (이산화질소 농도), NOx (질소산화물), CO (일산화탄소 농도), O3 (오존 농도), C6H6 (벤젠)과 같은 변수 데이터를 사용하여 NARX-ANN (Non-linear autoregressive neural network with multiple exogenous variables) 모형에 적용하였다. Bouakline 등 (2022)는 10년간의 미세먼지 데이터를 사용하였고, PM_{10} 의 하루 예측을 진행하였다. 그 결과, LSTM, RNN (Recurrent neural network)및 GRU에 유전 알고리즘 GA (Genetic Algorithm)기법을 사용하여 튜닝한 후 EFS (Exhaustive feature selection) 방법을 통해 예측하였을 때 세 가지 예측 결과가 비슷한 성능을 보였다. 하지만 각 모형의 성능을 향상시키기 위해 최적화 기법을 탐색하는 등 추가적인 개선이 필요하다고 주장했다. Ramli 등 (2023)은 말레이시아 반도 내 대기질 모니터링 데이터를 사용하여 8가지 대기질 매개 변수와 함께 BMA (Bayesian model averaging)을 사용하여 다음날의 미세먼지 농도를 예측하였다. 이 모형은 불확실성을 고려한 통계 모형 적합 방법으로 다양한 모형의 조합을 고려할 수 있는데, 결과적으로 해당 모형에 상대습도, 풍속 그리고 PM_{10} 을 적용하였을 때, 예측성능이 가장 우수한 결과를 확인하였다. Isikdag Umit (2022)는 선형회귀, NARnets (Non-linear autoregressive neural network), LSTM 모형 중, NARnets 모형이 단기적인 시간 의존성을 가진 미세먼지 농도 시계열 예측에서 가장 정확한 예측 결과를 보이는 것을 확인하였고, 해당 모형이 효과적인 예측 모형이라고 추천하였다. Ivanov 등 (2022)은 RF (Random forest) 머신러닝 모형을 사용하여 8개의 기상변수 등을 적용하여 PM_{10} 수준을 7일까지 예측하였고, 해당 모형은 93% 정도 설명력을 지닌 것을 확인하였다. Veleva 등 (2022)는 불가리아의 수자원부가 2010년부터 2021년까지 수집한 미세먼지 농도 측정 데이터에 Box-Jenkins ARIMA (Autoregressive integrated moving average) 모형을 적용하였고, 추가적으로 GARCH (Generalized autoRegressive conditional heteroskedasticity) 모형을 결합하여 잔차의 변동성을 고려하였다. 해당 모형은 연간 미세먼지 농도의 변동성을 효과적으로 예측하는 것을 입증하였다. Fernando 등 (2012)는 대기오염 데이터와 시간적 요인인 주말/평일 데이터를 사용하여 PM_{10}를 예측하고자 하였다. 결론적으로, EnviNNet (prototype stochastic NN model based on neural networks) 신경망 기반의 모형이 가장 빠른 연산속도와 MAE 기준 19.01로 우수한 예측 성능을 보였다. Cekim (2020)은 9년동안의 월별 미세먼지 농도 데이터에

ARIMA, ETS (Error trend and seasonality), SSA (Singular spectrum analysis) 시계열 예측 방법을 적용하여 SSA 모형이 가장 좋은 예측 성능을 보이는 것을 확인하였다. 또한, Ceylan과 Bulkan (2018)은 다중 선형 회귀 모형과 ANN 모형을 사용하였고, Francsci 등 (2018)은 K-means 군집화방법과 ANN과 MLP (Multi layer perceptron)를 결합한 모형을 사용하여 오염이 심한 지역에서의 군집화 방법히 효과적이라는 것을 주장하였다. Wang 등 (2022)는 고전적인 시계열 분석 방법과 회귀분석을 결합한 R-A&R-M (Residual analysis and regression-based autoregressive and moving average) 모델을 제안하였다. 국내에서는 Hoonja Lee (2010)는 경기도 수원시의 2003년부터 2009년 미세먼지 농도 자료와 이산화황, 이산화질소, 일산화탄소, 오존, 기상변수로는 최고온도, 풍속, 상대습도, 강수량, 일사량, 운량을 사용하여 자기회귀오차모형을 통해 월별 미세먼지 농도를 예측하고자 하였으며, Yoon과 Lee (2023)가 안동지역 내 3년동안의 시간별 데이터를 수집하고, SVM, RF, GBM (Gradient boosting machine) 그리고 XGBoost (eXtrme Gradient boosting) 모형을 적용하여 $PM_{2.5}$ (초미세먼지 농도)를 예측하고자 하였다. 그 결과, XGBoost 모형이 MAE (mean absolute error) 기준 4.02로 가장 우수한 예측성능을 보였다. 해당 연구의 기상변수로는 기온, 풍속, 풍향, 습도, 증기압, 현지 기압, 해면 기압, 지면 온도 및 대기 오염 물질이 사용되었다. Jung 등 (2021)은 천안시의 기상 및 대기오염 데이터에 RNN과 LSTM 모형의 예측 성능을 비교하였고, 결론적으로 LSTM에 Dropout 층을 추가한 모형이 RMSE (root mean square error) 10.25로 정확도가 우수하다고 주장하였다. Kim과 Jeong (2022)는 서울 지점의 미세먼지 농도를 DNN, RF, SVR, LSTM 모형을 이용하여 실시간으로 예측하였고, SVR 모형의 예측 성능이 미세먼지 농도 등급별로 비교하였을 때 가장 높은 것을 확인하였다.

이렇듯 국내외에서 미세먼지 농도 예측 알고리즘 연구가 활발히 진행되고 있지만, 국내에서는 다양한 딥러닝 모형을 적용한 미세먼지 농도 장기 예측 연구가 부족한 실정이다. 따라서, 본 연구에서는 시계열에 적합하고 최근 이미지 분류에 가장 많이 쓰이지만 예측에도 다양하게 적용되는 머신러닝 모형들의 예측 성능을 비교해보고자 한다.

다음 2장에서는 미세먼지 농도 예측을 위해 사용한 MLP, RNN, LSTM, GRU, CNN 모형에 대하여 소개한다. 3장에서는 본 논문에 사용한 기상 및 대기오염 데이터와 전처리 방법에 대하여 설명하고, 4장에서는 2장에서 언급한 머신러닝 모형들을 데이터에 적용한 후, 예측 성능을 비교한다. 마지막으로 5장에서는 결론 및 향후 연구 방향에 대하여 논의한다.

2. 예측 모형

본 논문에서는 미세먼지 농도 예측을 위하여 머신러닝 알고리즘을 사용하였다. MLP, RNN, LSTM, GRU 그리고 최근 다양하게 적용되는 CNN 모형을 예측에 적용하였고, 기본적인 모형의 예측성능을 확인하기 위하여 단일 은닉층의 입력노드 수 (unit)를 32로 설정하였다.

2.1. MLP (Multilayer percaptron) 모형

MLP (Multilayer percaptron) 모형은 지도학습 신경망의 한 종류로, 입력 데이터와 출력 데이터 사이의 관계를 학습하는 함수를 최적화하며 진행된다. 이 모형은 아래 Figure 2.1과 같이 Input layer, Hidden layers, Output layer으로 정의된다. Input layer은 이 모형이 처리할 데이터를 받아들인 후 다음 단계로 데이터를 전달하는 역할을 한다. 이후 Hidden layers에서 이전 계층에서 입력 받은 데이터를 활성화 함수와 가중치를 통해 처리한다. Output layer은 처리된 데이터를 바탕으로 예측 데이터를 결과로 제공한다. 이러한 방식으로 출력된 값이 실제 값과 가능한 비슷하도록 가중치를 조정한다 (Gardner와 Dorling, 1998).

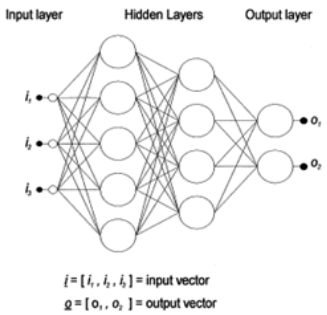

Figure 2.1 A multilayer perceptron with two hidden layers.

2.2. RNN (Recurrent neural network) 모형

RNN은 시퀀스 데이터를 처리하는 데 적합한 인공신경망의 한 종류로, 각 뉴런이 방향성을 가지고 연결되어 순환 구조를 이루며 내부의 hidden state가 이전 정보를 기억하면서 시간적인 정보를 처리한다. RNN은 주어진 시퀀스 $x = (x_1, x_2, \cdots, x_t$ 에 대해 hidden state, h_t를 아래와 같은 식을 활용하여 업데이트한다.

$$h_{t-1} = \begin{cases} 0, & t = 0 \\ \phi(h_t, x_t), & otherwise \end{cases} \tag{2.1}$$

위 식에서 phi 는 비선형 함수이다. RNN은 시퀀스의 다음 요소에 대한 확률 분포를 출력한다. 이는 hidden state, h_t를 고려하여 수행되며 이러한 시퀀스의 확률은 아래의 식과 같이 분해할 수 있다.

$$p(x_1, \cdots, x_t) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\cdots p(x_t|x_1, \cdots, x_{t-1}) \tag{2.2}$$

각 조건부 확률 분포는 다음과 같이 표현된다.

$$p(x_t|x_1, \cdots, x_{t-1}) = g(h_t) \tag{2.3}$$

이는 시퀀스 내의 x_t가 이전 요소들에 의해 의존하고 변화한다는 사실을 반영한다. 이를 활용하여 복잡한 패턴을 학습하고 예측하는 것이 가능해진다 (Chung 등, 2014).

2.3. LSTM (Long short-term memory) 모형

LSTM(Long Short-Term Memory)은 반복 신경망(RNN)의 한 종류로, 긴 시퀀스 데이터에서 이전 정보가 나중에 발생한 데이터에 영향을 주지 않는 문제를 해결하기 위해 고안되었다. LSTM은 메모리 셀과 세 가지 종류의 게이트인 input gate, forget gate, output gate를 통해 진행되며 이 과정을 거치며 시퀀스 데이터 간 의존 관계를 기억할 수 있게 된다. 각 게이트는 다음과 같은 식으로 표현될 수 있다 (Dey 와 Salem, 2017).

$$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \end{aligned} \tag{2.4}$$

위의 식에서 Input gate인 i_t는 cell state에 저장할 정보를 결정하고 forget gate인 f_t는 cell state에서 삭제할 정보를 결정하며, output gate인 o_t는 cell state에서 출력할 정보를 결정한다. h_t는 hidden state이며, W와 U는 각 층을 위한 가중치, 각 층의 b는 편향값을 의미한다. 게이트들은 sigmoid 활성화 함수(σ)를 통해 0과 1 사이의 값으로 조정한다. sigmoid 함수를 통해 출력된 값은 0에 가까운 값을 가지면 정보를 많이 잃어버린 상태로, 1에 가까운 값을 가지면 정보를 많이 기억하고 있는 상태로 해석할 수 있다. output gate는 앞서 업데이트된 cell state를 바탕으로 결과를 출력하며 cell state의 어떤 정보를 전달할지 결정하는 역할을 한다. 이렇게 각 게이트의 연산과 cell gate의 업데이트를 통해 데이터간 의존 관계를 효율적으로 활용하는 것이 가능해진다 (Gers 등, 2000).

2.4. GRU (Gated Recurrent Unit) 모형

Gated Recurrent Unit은 RNN의 변형 형태로, LSTM의 복잡성을 줄이면서 시퀀스 데이터에서 이전 정보가 나중에 발생한 데이터에 영향을 주는 특성을 처리할 수 있는 모델이다. GRU는 LSTM의 세 개의 게이트를 두 개로 축소하여 update gate, 와 reset gate만을 가진다. 이 두 가지 게이트는 GRU의 진행 과정에 중요한 역할을 한다. GRU와 update gate, reset gate 수식은 다음과 같이 표현된다.

$$\begin{aligned} h_t &= (1 - z_t) \bigodot h_{t-1} + z_t \bigodot \widetilde{h}_{t-1} \\ \widetilde{h}_{t-1} &= g(W_t x_t + U_h(r_t \bigodot h_{t-1}) + b_h) \\ z_t &= \sigma(W_t x_t + U_z h_{t-1} + b_h) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \end{aligned} \tag{2.5}$$

r_t는 reset gate로 과거의 정보를 리셋시키며 현 시점의 정보 x_t와 직전 시점의 은닉층의 값 h_{t-1}에 출력가중치와 입력가중치인 W와 U를 곱하고 편향값 b를 더하여 얻을 수 있다. z_t는 update gate, $\widetilde{h}_t$는 h_t이전의 cell state를 의미한다. σ는 sigmoid 함수를, $\bigodot$은 Hadamard product를 의미한다. 이러한 구조를 가진 GRU는 LSTM과 비교하여 매개 변수가 적으며, 그로 인해 학습 과정에서 계산 효율성이 높아진다. 그럼에도 불구하고 여러 연구에서 GRU는 대부분 LSTM과 비교해 동등하거나, 심지어 더 우수한 성능을 보인다. 게이트의 개수를 줄여 복잡성이 낮아진 점이 GRU의 성능에 크게 영향을 미치지 않았다는 것을 보여준다 (Dey 와 Salem, 2017).

2.5. CNN (Convolutional neural network) 모형

CNN은 이미지와 같은 데이터의 패턴을 인식하는 데 효과적인 딥러닝 모델로, 데이터 내의 위치 정보 즉 이미지 내 픽셀들 사이의 연관성과 구조를 보존하며 합성곱 연산을 통해 특징을 추출한다. CNN은 Convolutional layer, Pooling layer, Fully connected layer 등 여러 계층을 통해 구성된다. 먼저 Convolutional layer에서는 입력 데이터에 필터를 적용해 피처맵을 생성한다. 입력 데이터에 필터를 적용하여 이미지의 특정 패턴이 어디에 존재하는 지를 확인한 정보가 피처맵이다. 필터의 크기는 일반적으로 매개변수로 결정되며, 가중치 학습을 통해 최적화된다. 활성화 함수는 Convolutional layer에서 출력된 값에 적용된다. 대표적으로 Relu 함수가 활용되며, 비선형성이 적용된다. Pooling layer 계층은 Convolutional layer를 통과하고 활성화 함수가 적용된 피처맵의 차원을 축소하는 역할을 한다. 차원을 줄임으로써 계산 효율성을 높이고 모델의 학습 속도를 향상시키며 과적합을 방지한다. 마지막으로 Fully connected layer에서는 피처맵을 1차원 벡터로 평탄화한 후 분류기를 통해 최종적인 예측을 수행한다. CNN의 학습 과정은 역전파 알고리즘을 통해 이루어지며 손실함수를 최소화하는 방향으로 가중치를 갱신한다 (LeCun 등, 1998).

3. 데이터 및 자료 분석

3.1. 데이터 및 분석방법

본 논문의 연구 주제인 미세먼지 농도(PM_{10}) 자료는 에어코리아 (`https://www.airkorea.or.kr`)에서 수집하였다. 기상 자료는 기상청에서 수집하였고, 아래 Figure 3.1과 같이 2016년 1월부터 2019년 12월까지의 시간별로 인천광역시의 데이터를 추출하였다. 인천광역시는 실제로, 2022년 미세먼지 및 초미세먼지 농도가 전국 7대 도시 중 가장 나쁜 수준을 기록하였고, 보건환경연구원에 따르면 인천의 대기질이 악화된 이유는 고비사막이나 내몽골 등에서 온 황사의 영향으로 추정하였다. 따라서, 국내에서 미세먼지 농도 유입 경로인 산둥반도와 가장 인접한 인천광역시의 미세먼지 농도를 정확히 예측할 수 있다면 다른 지역의 미세먼지 농도 예측과 미세먼지 저감 대책 모색에도 큰 도움이 될 것으로 보인다 (Oh 등, 2019). 2020년 이후로는 2019년 12월에 발생한 코로나19의 여파로 대기오염 수치가 매우 낮아졌기 때문에 데이터에서 제외하였다 (Choi와 Cheong, 2021). PM_{10}을 예측하기 위한 기상변수

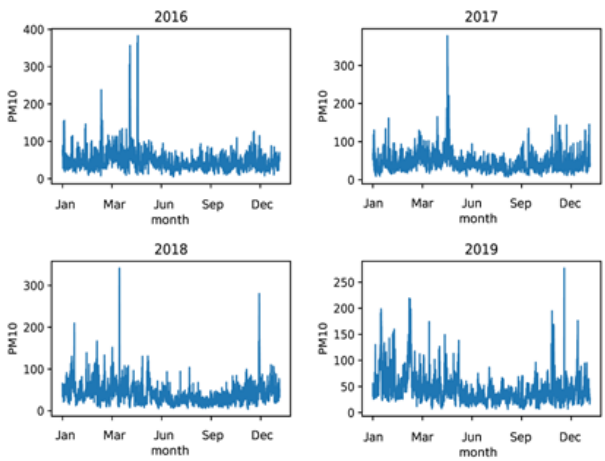

Figure 3.1 Graph of PM_{10} in Incheon by year

로는 기온, 강수량, 풍속, 풍향, 습도, 증기압, 이슬점 온도, 현지기압, 해면기압, 전운량, 지면온도를 사용하였으며, 대기오염 변수로는 SO2, CO, O3, NO2, PM2.5를 수집하였다. 추가적으로, 강수량과 전운량 같은 경우 오랫동안 데이터가 누락된 기간이 발생하여 선형보간법을 통해 보완하였다. 하지만, 선형보간법은 앞뒤 데이터가 반드시 존재해야하기 때문에 데이터 수집기간인 2016년 1월 1일부터 장기간 결측이 발생한 구간은 제외하였다. 데이터에서 2016년 1월부터 2018년 12월까지를 추출하여 훈련 및 검증자료 (training and validation data)로 사용했고, 나머지 2019년 1월부터 12월까지의 데이터는 테스트 데이터 (test data)로 각 예측 모형의 성능을 평가하였다. 본 논문에서 미세먼지 농도 예측을 위해 Python을 사용하였으며, 모형 적합 및 예측을 위해 sklearn과 tensorflow의 패키지를 주로 사용하였다. 또한 이 모형들은 초기값에 따른 결과가 달라질 수 있으므로 본 연구에서는 random_state 설정을 사용하여 초기값을 고정하였다.

3.2. 모형 성능 평가

본 논문에서 모형별 예측 성능 평가를 위해 MAE (mean absolute error)와 RMSE (root mean square error)를 사용하였다. 일반적으로 MAPE (mean absolute percentage error)가 모형 평가를 위해 널리 사용되지만, 미세먼지 농도 예측값이 0보다 작거나 0에 가까울 경우, MAPE 값이 매우커지거나 계산이 불가능하다는 단점이 발생한다. 따라서, 다음과 같이 정의되는 MAE와 RMSE 척도를 통하여 정확도를 평가하였다.

$$\begin{aligned} MAE &= \frac{\sum_{t=1}^{n} |Y_t - F_t|}{n} \\ RMSE &= \sqrt{\frac{\sum_{t=1}^{n} (Y_t - F_t)^2}{n}} \end{aligned} \tag{3.1}$$

여기서 n은 예측에 사용되는 데이터의 수이고, Y_t는 t시점에서 관측된 값이며, F_t는 시점에서 모형을 통해 예측된 값이다. MAE와 RMSE 값이 모두 작을수록 정확도가 우수하다는 것을 의미한다.

3.3. 모형 적합 결과

본 논문에서는 인천 지역의 2016년 1월부터 2018년 12월까지의 데이터를 각 머신러닝 모형에 적용해 2019년의 미세먼지 농도(PM_{10})를 예측하고자 하였다. 외생 변수로는 인천 지역의 기상 정보인 기온 (temp), 강수량 (mm), 풍속 (m/s), 풍향 (dir), 습도 (humid), 증기압 (hPa), 이슬점온도 (etemp), 현지기압 (hhPa), 해면기압 (shPa), 전운량 (cloud), 지면온도 (gtemp)를 사용하였고, 대기오염 정보로는 SO2, CO, O3, NO2, 농도를 사용하였다. 본 연구는 각 머신러닝 모형의 기본적인 미세먼지 농도 성능 평가를 위하여 가장 기본적인 은닉층 개수로 단일층과 초매개변수 (hyperparameter)를 설정하였다. 세부적인 초매개변수값으로 hidden state의 갯수인 n_unit을 32로 설정하였고, 각 모형층은 단일층으로 구성하였다. 또한 optimizer는'adam으로 통일하여 선정하였으며 모형 학습 시 epoch 수는 100, batch size는 기본값인 32, 입력데이터의 time step은 3으로 설정하였다. CNN 모형의 경우, kernel_size를 2로, filter 수는 모형별 n_unit과 같은 수인 32로 설정하였다. 이렇게 단일층만으로 구성된 모형별 예측 성능을 비교한 후 GRU 모형의 경우가 가장 예측 정확도가 우수하였고, 최종적으로는 은닉층 개수에 따른 GRU 모형의 예측 성능을 확인하기 위해, 은닉층 (unit = 64)을 추가한 모형 (GRU2)의 예측 성능 또한 확인해보았다. 각 머신러닝 모형의 성능을 향상시키기 위해 Figure 3.2과 같이 시계열 Corss-validation 방법을 적용하여 바로 전날의 정보를 포함하여 예측을 수행 및 실제 미세먼지 농도값과 예측값을 비교하고자 하였다.

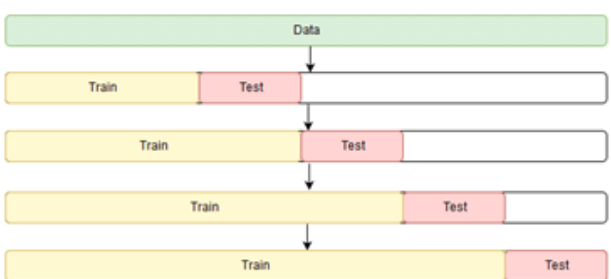

Figure 3.2 Time series cross-validation process

예측에 앞서, 피어슨 상관분석 방법을 통하여 예측하고자 하는 미세먼지 농도과 대기물질 및 기상변수 간의 상관관계를 확인하였다. 변수별 상관관계는 다음 Figure 3.3와 같다. 예측하고자하는 PM_{10}에 영향을 가장 많이 끼치는 변수는 $PM_{2.5}$변수임을 보였다. 그 다음으로는 CO, NO2, SO2, O3 순으로 대기오염물질 데이터가 뒤를 이었고, 기상변수는 증기압, 기온과 이슬점온도가 상관관계가 상대적으로 높은 수치를 보였다. 특히, 초미세먼지 농도 ($PM_{2.5}$)와의 상관관계는 0.68로 매우 높게 나타났고, 다른 대기오염물질들 또한 미세먼지 농도에 강한 양적 상관관계를 가지고 있다. 반면에, 기상 변수들은 상관관계 수치가 0.5 미만으로 상관관계가 약하다고 할 수 있다. 위 변수들을 통해 머신러닝 모형별 예측 정확도는 다음 Table 3.1와 같다.

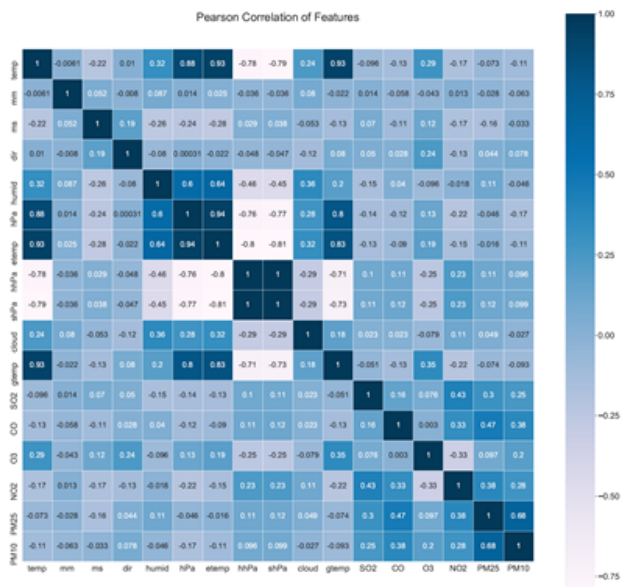

Figure 3.3 Pearson correlation of features in Incheon

Table 3.1 Test data fit performance results

Model	MAE	RMSE
MLP	10.23	14.87
RNN	9.17	15.40
LSTM	9.23	14.89
GRU1	**8.80**	**14.61**
GRU2	9.49	15.19
CNN	9.71	15.07

미세먼지 농도 예측 성능 비교 결과 은닉층을 추가한 모형 (GRU2)의 예측 성능이 더 우수할 것이라는 예상과 달리, 단일층으로 구성된 GRU 모형 (GRU1)이 MAE 8.80, RMSE 14.61로 다른 모형들에 비해 가장 성능이 우수하다는 것을 확인할 수 있었다. 가장 예측 성능이 낮은 모형은 MLP 모형이며, 그 뒤로는 RNN, LSTM, GRU2, CNN 순으로 예측 성능이 우수하였다.

4. 결론

본 논문에서는 최근 봄철마다 국민들의 관심이 커지고 있는 미세먼지 농도 (PM_{10})를 머신러닝 모형을 이용하여 예측 및 성능 비교를 하고자 하였다. 그 중, 인천광역시의 기상변수와 대기오염물질 데이터를 선형보간법으로 전처리하여 이용하고, 피어슨 상관분석을 통해 어떠한 변수가 미세먼지 농도에 큰 영향을 미치는지 확인하였다. 상관분석의 결과, 대기오염물질이 PM_{10}에 가장 주요한 영향을 미치는 것을 확인하였고, $PM_{2.5}$, CO, NO2, SO2, O3 순으로 상관관계가 높았다. 기상변수는 대기오염물질에 비해 미세먼지 농도에 상대적으로 작은 영향력을 가지고 있었으며, 그 중에서는 증기압, 기온과 이슬점 온도 변수의 영향이 있었다. 다양한 단일층 머신러닝 모형 (MLP, RNN, LSTM, GRU, CNN)을 적용하여 예측 결과를 비교해 보았다. 그 결과, GRU 모형의 MAE와 RMSE가 8.80과 14.61로 가장 낮으며 우수한 예측 성능을 보여주는 것을 확인하였다. 이에 따라, 가장 좋은 성능을 지닌 단일층으로 구성된 GRU 모형에 추가적인 층을 구축하여 GRU2 모형의 성능을 확인하였으나, 단일층으로 구성된 GRU1 모형이 더 좋은 성능을 보여주었다. 추후 해당 연구를 확장 및 예측 정확도 향상을 위해 미세먼지 예측에 중요한 다른 변수들을 탐색하는 다각적인 접근이 필요할 것으로 예상된다. 또한, 본 연구에서는 간단한 머신러닝 모형들을 적용하였으나 향후 CNN 모형을 활용한 머신러닝 앙상블 모형을 통해 미세먼지 농도 예측 알고리즘 발전에 보탬이 될 것으로 사료된다.

References

Adnane, A., Leghrib, R., Chaoufi, J., & Chirmata, A. (2022). Prediction of pm10 concentrations in the city of agadir (morocco) using non-linear autoregressive artificial neural networks with exogenous inputs (NARX). *Materials Today: Proceedings*, **52**, 146-151.

Bouakline, O., El Merabet, Y. and Khomsi, K. (2022). Deep-Learning models for daily pm10 forecasts using feature selection and genetic algorithm. *In 2022 8th International Conference on Optimization and Applications (ICOA) IEEE*, 1-4.

Cekim, H. O. (2020). Forecasting PM10 concentrations using time series models: A case of the most polluted cities in Turkey. *Environmental Science and Pollution Research*, **27**, 25612-25624.

Ceylan, Z. and Bulkan, S. E. R. O. L. (2018). Forecasting PM10 levels using ANN and MLR: A case study for Sakarya City. *Glob. Nest J*, **20**, 281-290.

Choi, W.-C. and Cheong, K.-S. (2021). Analysis of the factors affecting fine dust concentration before and after COVID-19. *Journal of the Korean Society of Hazard Mitigation*, **21**, 395-402.

Chung, J., Gulcehre, C., Cho, K. and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Dey, R. and Salem, F. M. (2017). Gate-variants of gated recurrent unit (GRU) neural networks. *In 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 1597-1600.

Franceschi, F., Cobo, M. and Figueredo, M. (2018). Discovering relationships and forecasting PM10 and PM2. 5 concentrations in Bogota, Colombia, using artificial neural networks, principal component analysis, and k-means clustering. *Atmospheric Pollution Research*, **9**, 912-922.

Fernando, H. J., Mammarella, M. C., Grandoni, G., Fedele, P., Di Marco, R., Dimitrova, R. and Hyde, P. (2012). Forecasting PM10 in metropolitan areas: Efficacy of neural networks. *Environmental Pollution*, **163**, 62-67.

Gardner, M. W. and Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) -A review of applications in the atmospheric sciences. *Atmospheric Environment*, **32**, 2627-2636.

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, **12**, 2451-2471.

Ivanov, A., Gocheva-Ilieva, S. and Stoimenova-Minova, M. (2022). Random forest regression for statistical modeling and forecasting of pm10. *In AIP Conference Proceedings*, *AIP Publishing*, **2522**.

Jung, Y. J., Lee, J. S. and Oh, C. H. (2021). Performance comparison of pm10 prediction models based on RNN and LSTM. *The Korea Institute of Information and Commucation Engineering*, 280-282.

Kim, M. and Jeong, H.-S. (2022). Development of machine learning based prediction of particluate matter concentration in Seoul. *Journal of the Korean Data & Information Science Society*,**33**, 1095-1111.

Kujawska, J., Kulisz, M., Oleszczuk, P. and Cel, W. (2022). Machine learning methods to forecast the concentration of pm10 in lublin, poland. *Energies*, **15**, 6428.

LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**, 2278-2324.

Lee, H. (2010). Analysis of time series models for pm10 concentrations at the Suwon city in Korea. *Journal of the Korean Data & Information Science Society*, **21**, 1117-1124.

Lee, H., Jee, M., Kim, H., Jun, T. and Kim, C. (2023). Early prediction of fine dust concentration in seoul using weather and fine dust information. *Journal of Broadcast Engineering*, **28**, 285-292.

Oh, J., Kim, D., Kwon, H., Kim, S. and Lee, S. (2019). A study on the relationship between shandong peninsula and fine dust concentration in incheon metropolitan city. *Journal of the Conference of the Korean Society for Environmental Education*, 198-203.

Plocoste, T. and Laventure, S. (2023). Forecasting pm10 concentrations in the caribbean area Using machine learning models. *Atmosphere*, **14**, 134.

Ramli, N., Abdul Hamid, H., Yahaya, A. S., Ul-Saufie, A. Z., Mohamed Noor, N., Abu Seman, N. A. and Deak, G. (2023). Performance of bayesian model averaging (bma) for short-term prediction of pm10 concentration in the Peninsular Malaysia. *atmosphere*, **14**, 311.

Ramli, N. Norazrin, *et al*. (2023) Performance of Bayesian model averaging (BMA) for short-term prediction of PM10 concentration in the Peninsular Malaysia.*Atmosphere*, **14**, 311.

Sharma, E., Deo, R. C., Soar, J., Prasad, R., Parisi, A. V. and Raj, N. (2022). Novel hybrid deep learning model for satellite based pm10 forecasting in the most polluted Australian hotspots. *Atmospheric Environment*, **279**, 119111.

Veleva, E., Filipova, M. and Zheleva, I. (2022). Statistical study of particulate matter (pm10) air contamination in the city of Vidin, Bulgaria. *In AIP Conference Proceedings*, AIP Publishing, **2522**.

Wang, Z., Chen, H., Zhu, J. and Ding, Z. (2022). Daily pm2. 5 and pm10 forecasting using linear and nonlinear modeling framework based on robust local mean decomposition and moving window ensemble strategy. *Applied Soft Computing*, **114**, 108110.

Yeo, M. and Kim, Y. (2019). Trends of the pm10 concentrations and high pm10 concentration cases in korea. *Journal of Korean Society for Atmospheric Environment*, **35**, 249-264.

Yoon, J. and Lee, Y. (2023). Performance comparison of machine learning models for prediction of fine dust concentration. *Proceedings of Symposium of the Korean Institute of Communications and Information Sciences*, 1417-1418.

Journal of the Korean Data & Information Science Society
2023, **34**(5), 763–773

http://dx.doi.org/10.7465/jkdi.2023.34.5.763
한국데이터정보과학회지

A study on PM10 forecasting model using machine learning[†]

Sahm Kim[1]

[1]Department of Statistics, Chungang University

Received 24 July 2023, revised 7 August 2023, accepted 11 August 2023

Abstract

Fine dust refers to dust with a particle diameter of less than 10μg among dust, which is a particulate matter floating or flying down in the atmosphere, and is also referred to as PM_{10}. These fine dust is very small in size and permeates the body without being filtered from the nose or bronchial tubes, causing inflammation through asthma, lung disease, or the action of immune cells. Recently, it was found that Korea has the highest concentration of fine dust in the world, and it is important to take measures through an accurate forecast system because fine dust directly affects not only health but also ecosystems and crops. Therefore, this paper attempted to compare machine learning prediction performance of fine dust concentration using weather data provided by the Korea Meteorological Administration and air pollutant data provided by Air Korea. As for the region, data from Incheon Metropolitan City, which is the closest to the Shandong Peninsula, the inflow path of yellow dust, were extracted, and a model was built after confirming the correlation between various weather factors and air pollutants in Incheon. MLP, RNN, LSTM, GRU, and CNN were used as models, and predictive performance was compared by organizing basic hyperparameters and single layers. After that, the GRU2 model, which added layers to the GRU1 (single layer) model, was newly constructed and compared with the GRU1 model with the best prediction performance. Prediction performance was evaluated by MAE and RMSE in test data. Most of them showed similar predictive performance, but it was confirmed that the GRU1 model had the best performance compared to other models, with MAE 8.80 and RMSE 14.61. The model with the lowest prediction performance was the MLP model, followed by RNN, LSTM, GRU2, and CNN.

Keywords: CNN, GRU, LSTM, machine learning, PM_{10} forecasitng.

[†] This research was supported by the Chung-Ang University research grant in 2022.
[1] Professor, Department of Statistics, Chungang University, Seoul 06974, Korea.
E-mail: sahm@cau.ac.kr