JOURNAL OF THE ENVIRONMENT

Original article

Vol. 15, No. 1: 41-48 (2022)

기계학습을 이용한 초미세먼지($PM_{2.5}$) 예측 연구 동향

정하윤[1], 곽경환[1,2,*]

[1]강원대학교 환경학과, [2]강원대학교 환경융합학부

Review on the Prediction Studies of $PM_{2.5}$ Using Machine Learning

Ha-Yoon Jeong[1], Kyung-Hwan Kwak[1,2,*]

[1]Department of Environmental Science, Kangwon National University, Chuncheon 24341, Korea
[2]School of Natural Resources and Environmental Science, Kangwon National University, Chuncheon 24341, Korea

Abstract Machine learning, which has been recently used to predict the concentration of fine particulate matter ($PM_{2.5}$), can learn a large amount of data and perform classification or regression analysis. Among well-known machine learning algorithms, random forest and XGBoost have less prediction errors and do not have over-fitting problem. It is possible to check which variables have a great influence on the learning process by calculating the importance index. Because it is difficult to understand the classification and prediction process due to the characteristics of machine learning technology, it is necessary to select an appropriate set of input variables suitable for characteristics of the target variable (e.g., $PM_{2.5}$). The $PM_{2.5}$ concentration prediction model developed based on random forest and XGBoost reported in previous domestic and international studies showed similar or better prediction performance compared with other prediction techniques. There were important input variables related to the occurrence factors of $PM_{2.5}$ or high concentration cases such as AOD (aerosol optical depth), PM_{10}, relative humidity, and maximum wind speed. In particular, in domestic studies which include the influence of upwind countries, the input variables related to the upwind countries have a great influence on the model performance. The best model for predicting $PM_{2.5}$ concentration depends on the type, period, and dataset of the input variable, so an algorithm suitable for the data should be used through testing experiments. To select an appropriate input variables, we need to check and compare the prediction accuracy of various sets of input variables, as documented by previous studies.

Key words: Machine Learning, Fine Particulate Matter, Random forest, XGBoost, Prediction model

1. 서 론

급격한 산업화 및 기후변화로 인해 발생한 대기오염은 인체에 부정적인 영향을 미치며 전 세계적으로 문제가 되고 있다(Jacob and Winner, 2009; Choubin et al., 2020). 그 중 미세먼지(Particulate Matter; PM)는 호흡기 질환, 심혈관 질환, 뇌혈관 및 폐암 등 다양한 질병의 발병률 및 사망률과 상관관계가 있음이 밝혀졌다(Atkinson et al., 2010; Correia et al., 2013; Fang et al., 2013; Cadelis et al., 2014). 경제협력개발기구(Organisation for Economic Co-operation and Development; OECD)는 한국의 PM_{10}과 $PM_{2.5}$로 인한 대기오염이 심각한 수준이므로, 앞으로의 대책을 마련하지 않을 경우 OECD 회원국 중 2060년 조기 사망률 및 경제적 피해가 가장 클 것이라고 발표했다(OECD, 2017). 이에 우리나라는 효과적인 대기질 관리를 위해 국가적인 차원에서 대기오염측정망을 설치·운영하고, 노출 피해를 최소화하기 위해 2013년부터 미세먼지 예보를 시행하고 있다.

미세먼지 농도를 예측하는 방법은 크게 과정 기반 모델

***Corresponding author.** Kyung-Hwan Kwak
Tel. +82-33-250-8575 Fax. +82-33-259-5563
E-mail. khkwak@kangwon.ac.kr

링과 통계 기반 모델링으로 나뉜다. 국내에서 수치 예보 모델을 활용한 연구를 살펴보면 Kim et al. (2011)은 ADAM (Asian Dust Aerosol Model)과 CFORS (Chemical Weather Forecasting System) 모델을 활용해 황사 발생지의 PM_{10} 농도를 예측하였다. 이후, 손 등(2016)은 CMAQ (Community Multiscale Air Quality Model)을 사용해 서울지역의 PM_{10} 농도를 예측하였고, 정과 문(2015)은 WRF-Chem (Weather Research and Forecasting-Chemistry) 모델로 한반도의 황사, PM_{10} 농도를 모의하여 예측하였다. 하지만 화학 수송 모델과 같은 수치모델은 입력 자료의 불확실성과 정확한 미세먼지 발생원 조사 등에 한계가 있다. 따라서 이런 어려움을 피하기 위한 방법으로 측정망 데이터와 기상관측자료의 상관관계를 분석한 통계 모델링도 다수 수행되었으며 (구 등, 2010), 적절하게 변수 간 관계가 구현될 경우 수치모델보다 더 높은 예측 정확도를 보이기도 한다 (Hrust et al., 2009; 조 등, 2019).

최근에는 빅데이터를 활용하여 인공 신경망이나 기계학습법을 이용한 미세먼지 예측 연구가 다수 수행되고 있다. 과거 기계학습은 데이터 수집의 어려움으로 실용성이 낮았지만, 현재는 과학 기술의 발전으로 다양한 분야에 적용되어 우수한 성능을 보인다 (김과 조, 2013; 김, 2014). 최근 미세먼지의 농도 예측을 위한 방법으로 각광받고 있는 기계학습 기법은 많은 양의 과거 데이터에서 학습한 독립변수와 종속 변수 간 관계를 바탕으로 통계적인 모델을 만드는 방법이다. 예측하고자 하는 대상 기간과 공간 범위에 적합한 과거 데이터를 선정한 후 그 특징을 구현하는 데 적합한 알고리즘을 선택하는 것이 중요하다. 이러한 점 때문에 기계학습법을 활용하여 $PM_{2.5}$를 예측한 국내외 많은 연구에서 가장 정확한 기계학습법에 관하여 제각각 다른 결론이 도출되기도 한다 (이 등, 2006; Ho et al., 2021; Gupta et al., 2021; 박 등, 2021; 길과 이, 2021). 정확한 $PM_{2.5}$ 예측에는 최적의 기계학습법 선정 외에도 적절한 입력변수의 조합, 충분한 학습 데이터 확보 등이 전제되어야 한다. 따라서 절대적으로 정확성을 담보하는 기계학습법이 존재하기 어려운 만큼, 주어진 학습 데이터와 예측 변수에 따라 사전 연구 결과와 경험에 근거하여 연구자가 독자적으로 방법론을 선택하는 것이 중요하다.

본 논문에서는 1장에서 본 연구의 배경과 필요성에 대해 설명하고 2장에서 대표적인 트리 기반 기계학습 기법인 랜덤 포레스트 (Random Forest)와 XGBoost (eXtreme Gradient Boosting)에 관해 설명하였다. 이어서 3장과 4장에서 해외와 국내 연구 사례를 통해 $PM_{2.5}$ 예측 연구와 중요 입력 변수를 정리하였다. 5장에서 선행 연구의 결론과 기계학습 모델 사용 시 유의 사항에 대해 언급하였다. 이를 통해 추후 빅데이터를 활용한 기계학습 기반의 국내 $PM_{2.5}$ 농도 예측 연구를 활성화하기 위한 방향성을 제시하고자 한다.

2. 랜덤 포레스트와 XGBoost

랜덤 포레스트, XGBoost 모형은 여러 분야에서 안정성과 정확도가 입증되어 다수 사용되고 있다 (김과 이, 2020; 신 등, 2021). 이러한 트리 기반 앙상블 학습 (ensemble learning) 모형은 다른 기계학습법 혹은 신경망 구성 기법과 유사한 예측 능력을 보여주면서도 (Reid et al., 2015; Di et al., 2019) 상대적으로 사용자 친화적이라는 장점을 가지고 있다 (Chen et al., 2018). 또한 수치 예보 모델이 입력 데이터의 특성보다 모형 설정과 구축 방법에 더 의존하는 경향이 있는 데 반해, 기계학습 기법은 입력 데이터를 기반으로 예측값을 산출하므로 예측 변동성이 크지 않다 (Singh et al., 2017; Joharestani et al., 2019). 랜덤 포레스트는 다수의 의사결정 트리 (Decision tree)를 학습하는 앙상블 학습 모형이다 (Fig. 1). Breiman (2001)에 의해 제시되었으며 Ho (1995)와 Amit and Geman (1997)의 연구에 영향을 받았다. 의사결정 트리는 정해진 규칙에 따라 데이터를 분류하며 목표에 가장 가까운 결과에 도달하도록 하는데, 간단하고 이해하기 쉽지만 과적합 (Overfitting)의 문제가 있어 실제적인 사용이 어렵다. 랜덤 포레스트는 이러한 의사결정 트리를 무작위로 생성하여 나온 예측값 중 최빈값 혹은 평균값으로 결과를 도출한다. 이러한 과정을 통해 구성된 모델은 예측 오차가 적고 과적합 문제가 발생하지 않는다는 장점이 있으며 (Breiman, 2001), 설명변수가 다수일 때 예측력이 뛰어나고 매우 안정적인 모형을 제공한다 (Siroky, 2009). 다만 데이터 크기에 비례하여 수백~수천 개의 트리를 형성하므로 메모리 소모가 크고 속도가 느리다는 단점이 있다.

XGBoost는 Chen and Guestrin (2016)에 의해 제시되었으며 기존 GBM (Gradiant Boosting Machine) 모형을 개선한 알고리즘이다 (Fig. 2). GBM은 랜덤 포레스트와 마찬가지로 다수의 의사결정 트리로 구성된 앙상블 학습 모형이지만, 무작위적으로 트리를 생성하는 배깅 (Bagging)이 아

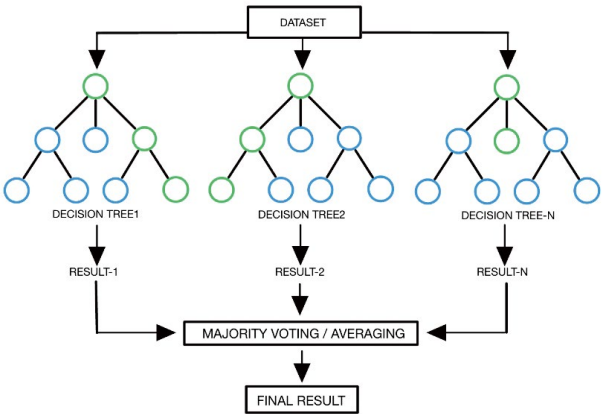

Fig. 1. Diagram of a random forest model. The final result is derived by majority voting or averaging the results from multiple decision trees.

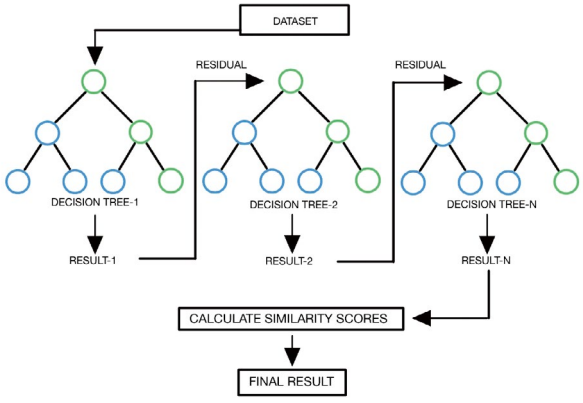

Fig. 2. Diagram of a XGBoost model. XGBoost uses a boosting method that learns by weighting the errors of the previous tree.

닌 이전 트리의 약점을 반영하는 부스팅(Boosting) 기법을 사용한다. 이로 인해 의사결정 트리의 쉽고 직관적인 점이 잘 반영된다는 장점이 있지만 병렬 처리가 불가해 학습 속도가 느리고 자체적으로 과적합 규제가 없다는 단점이 있었다. XGBoost는 이러한 부분을 개선하여 병렬 학습 지원, 과적합 규제 기능을 지원하며 범용성이 좋다는 장점이 있다.

3. 해외 연구 사례

Hu et al. (2017)은 미국 대륙의 일평균 지상 $PM_{2.5}$ 농도를 추정하는 랜덤 포레스트 모델을 개발하여 국가적 규모에서의 기계학습 알고리즘을 구현하고자 했다. 12 km × 12 km 격자로 나타냈으며 입력 변수로는 10 km 해상도의 AOD (Aerosol Optical Depth) 데이터와 기상장, 토지 이용 변수를 통합하여 사용했다. 예측값에 대해 교차 검증을 수행한 결과 R^2 값은 0.80, 일일 예측에 대한 MPE, RMSE는 각각 1.78 μg m^{-3}, 2.83 μg m^{-3}으로 국가 및 지역 규모에서 모두 랜덤 포레스트를 활용한 기계학습법이 높은 예측 정확도를 나타냈다. 또한, 측정소와 요일별로 학습 데이터를 분할하여 공간적, 시간적 변동에 대해 교차 검증을 수행한 결과 R^2 값이 각각 0.70, 0.79이므로, $PM_{2.5}$와 예측 변수 간의 관계에서 시간적 변동이 더 잘 설명됨을 나타낸다. 다만 저자는 10 km 해상도를 사용할 경우 시간 단축 면에서는 좋지만, 예측 정확도를 높이기 위해서는 더 상세한 격자를 사용해야 한다고 밝혔다. 이후 Di et al. (2019)는 미국 전역을 대상으로 더 높은 시공간 해상도를 가진 예측을 하고자 했고, 랜덤 포레스트를 포함한 세 가지 기계학습 알고리즘을 통합하여 각각의 모델이 서로를 보완하고자 했다. 예측 변수로는 AOD 및 위성 데이터, 화학 수송 모델 예측자료, 토지 이용 변수, 기상 자료 등을 사용했다. 그리고 2000년부터 2015년까지 1 km × 1 km 격자에서 100 m × 100 m 격자로 축소하여 공간 해상도를 높였다. 일농도 예측값은 R^2이 0.86, 연간 농도 예측값은 R^2이 0.89로 모의 성능이 향상되었음을 확인했다.

Chen et al. (2018)은 2005~2016년 중국의 $PM_{2.5}$ 일농도를 추정하기 위해 2014~2016년 동안 중국 전역에 있는 1,479개 관측소의 $PM_{2.5}$ 데이터를 사용했다. AOD, 기상 조건 등의 변수도 사용했으며 랜덤 포레스트 기법과 기존에 사용되던 2개의 회귀 모델을 구성하고 비교했다. 교차검증 결과 랜덤 포레스트 모델의 R^2값은 0.83, RMSE는 18 μg m^{-3}로 다른 회귀 모델보다 훨씬 유의한 예측값을 보였다. 이는 다른 모델을 사용하여 중국의 $PM_{2.5}$ 농도를 추정하려고 시도했던 이전 연구들보다 더 높은 정확도를 보였다(Fang et al., 2016; Ma et al., 2016; You et al., 2016).

Wong et al. (2021)은 대만의 $PM_{2.5}$ 일평균 농도 예측을 위해 대만 전역의 73개 측정소 자료를 활용했다. 사용 기간은 2006~2016년이며 약 280,000개의 관측값을 분석에 사용했다. 그 외에 EPA 환경 자원 데이터, 기상 자료, 토지 사용 자료, 도로 네트워크 지도, MODIS 식생 지수를 포함한 데이터 세트를 구성했다. 이후 기존에 사용되던 토지 사용 회귀 모델을 활용해 중요 변수를 추출하고, 농도 예측 모델로 심층 신경망, 랜덤 포레스트, XGBoost를 사용했다. 예측 결과 XGBoost의 RMSE가 4.41 μg m^{-3}, R^2은 0.94로 가장

Table 1. Input variables and important variables of machine learning for $PM_{2.5}$ concentration prediction model development studies.

Target area	Target period	Input variables	Important variables	References
USA	2011	$PM_{2.5}$, AOD (10 km), weather, land use variables	$PM_{2.5}$ convolution layer, AOD	Hu *et al.* (2017)
USA	2000~2015	$PM_{2.5}$, EOS AOD, Satellite-based measurements, chemical transport model prediction, land-use variables, weather	$PM_{2.5}$, CMAQ $PM_{2.5}$, $PM_{2.5}$ elemental carbon	Di *et al.* (2019)
USA 7 states	2000~2015	Terra AOD, $PM_{2.5}$, Meteorology, land use, roads, green spaces, spatiotemporal elements, convolution layer	–	Brokamp *et al.* (2018)
China	2014~2016	$PM_{2.5}$, MODIS AOD, weather (average temperature, relative humidity, atmospheric pressure, wind speed), land cover data	Day of year, AOD, daily temperature	Chen *et al.* (2018)
Taiwan	2006~2016	EPA environmental resources dataset, meteorological data, land-use inventory, landmark dataset, digital road network map, digital terrain model, MODIS (NDVI), power plant distribution dataset	Season (Winter), SO_2, O_3, NO_2	Wong *et al.* (2021)
Yangtze River Delta, China	2002~2018	TOA, $PM_{2.5}$, remote sensing data, meteorological field, land use data	PBLH, Relative humidity	Yang *et al.* (2020)
Beijing-Tianjin-Hebei, China	2010~2016	$PM_{2.5}$, AOD, weather (45), land use variables, altitude, population, road, chemical transport model data	Land data, AOD	Zhao *et al.* (2020)
Beijing, China	2017~2019	PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, O_3, weather, AQI	AQI, CO, PM_{10}	Zhang *et al.* (2020)
South Korea	2015~2016	PM_{10}, $PM_{2.5}$, GOCI AOD and satellite data, combined model results (weather, emissions, solar radiation), population density by region and year	AOD, maximum wind speed, solar radiation, dew point temperature, population density, SO_2, NH_3	Park *et al.* (2019)
Seoul, South Korea	2015~2019	PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, O_3, weather	NO_2, Average temperature	Park (2021)
Seoul, South Korea	2015~2018	Seoul and Baengnyeongdo PM_{10}, $PM_{2.5}$, weather	SO_2, NO_2, CO, O_3, Baengnyeongdo PM_{10}, $PM_{2.5}$	Lee and Lee (2020)
Gyeonggi, South Korea	2017	PM_{10}, $PM_{2.5}$, SO_2, NO_2, CO, O_3, weather (temperature, wind direction, wind speed, precipitation, humidity), traffic	–	Won (2020)

우수했다.

Brokamp et al. (2018)은 미국 7개 카운티에서 1 km × 1 km로 시공간 해상도가 높은 $PM_{2.5}$ 농도 예측을 위해 AOD와 11개의 입력 변수를 활용해 랜덤 포레스트 모델을 훈련했다. 연구 기간은 2000~2015년이고 사용된 데이터는 AOD 위성 자료(Terra, Aqua), $PM_{2.5}$ (AQS), 기상, 토지 이용, 도로, 녹지, 시공간 요소, 합성곱 레이어(convolution layer)를 사용했다. AOD 데이터에서 누락된 값을 예측 혹은 다른 데이터로 대체하거나 아예 제외하고 사용하던 다른 연구들과 달리, 본 연구에서는 AOD 결측도 지상의 $PM_{2.5}$와 연관이 있을 것이라 가정하고 앙상블 학습 모델을 구성했다. 이 모델은 교차 검증된 MAE가 0.95 μg m^{-3}, R^2가 0.91로 양호한 성능을 보였다. 이는 공간 및 시간적으로 교차 검증했을 때에도 마찬가지였다. 전반적으로 시공간 랜덤 포레스트 $PM_{2.5}$ 모델은 우수한 정확도를 보여주었으며 관련 건강 결과와의 연관성을 정량화하기 위해 장기간 및 급성 $PM_{2.5}$ 노출에 대한 고해상도 평가에 유용할 것이라고 했다.

Yang et al. (2020)은 $PM_{2.5}$ 추정을 위해 흔히 사용되는 AOD의 결측값이 많아 올바른 추정이 어렵다고 판단하

여 대신 TOA (Top of Atmosphere) 반사율을 사용했다. 연구 대상 지역은 대기오염이 심각한 양쯔강 삼각주 지역(Yangtze River Delta; YRD)이다. 랜덤 포레스트 모델의 경우 $PM_{2.5}$ 관측값과 예측값의 R^2가 0.96으로 높은 상관관계를 보였고, 이는 선형 회귀 모델보다 62.5% 높게 나타났으며 RMSE 및 MAE도 각각 4.21 μg m^{-3}, 2.85 μg m^{-3}로 감소했다. 이는 같은 지역에 대해 개발되었던 고급 통계 모델 LME 및 GWR (R^2은 각각 0.81, 0.79)에 비해 좋은 상관성을 보였다. 총 10개의 기상 변수 중 이전 연구와 동일하게 PBLH와 상대습도가 중요한 예측 변수였다(Yang et al., 2019). 반면 시공간 요소 교차검증을 수행했을 때 Ma et al. (2016)과는 다르게 $PM_{2.5}$에 대해 시간보다 공간적 변동성을 더 잘 나타냈음을 시사했다.

Zhao et al. (2020)의 연구는 중국의 대표적인 오염 지역인 3개 도시(베이징-톈진-허베이)에서 높은 시공간 해상도(0.01° × 0.01°)의 $PM_{2.5}$ 일농도를 추정하기 위해 랜덤 포레스트 모델을 개발했다. 입력 변수는 $PM_{2.5}$ 측정 데이터, AOD, 기상 자료(재분석 데이터), 토지 이용 변수, 고도, 인구수, 도로망, CMAQ 모델로 모의한 배출량 자료이다. 이렇게 만들어진 모델은 교차 검증 R^2값이 0.83~0.86으로 적절한 성능을 보였으며, 다른 연구(Lin et al., 2015; Lin et al., 2016; You et al., 2016; Pang et al., 2018)와 유사하게 지형 요인과 AOD 데이터가 중요한 기여 요인으로 작용했다.

Zhang et al. (2020)은 베이징의 $PM_{2.5}$ 일평균 농도 예측을 위해 랜덤 포레스트와 XGBoost를 사용하고 LSTM과 비교했다. 랜덤 포레스트를 사용하여 중요 변수를 추출하고, 해당 변수들을 XGBoost에 입력하여 사용하였다. 예측 결과 XGBoost과 LSTM의 RMSE는 각각 8.630 μg m^{-3}, 23.519 μg m^{-3}이며 R^2은 각각 0.95, 0.54로 XGBoost의 예측 성능이 높았다. 예측 시간에서도 XGBoost가 LSTM보다 빨랐다.

4. 국내 연구 사례

Park et al. (2019)는 한국의 대기오염측정망이 주로 도시에 집중되어 있으므로 이를 보완하기 위해 AOD와 대기 오염물질 측정 자료를 사용하여 랜덤 포레스트 모델을 개발했다. 추정된 PM_{10}와 $PM_{2.5}$의 R^2값은 0.78 및 0.73이고 RMSE가 17.08 μg m^{-3}, 8.25 μg m^{-3}로 우수한 성능이었다. 특히 이 모델은 높은 PM 농도를 잘 예측했다. 가장 중요한 변수는 AOD이었고 그 외 최대 풍속, 태양복사와 이슬점 온도와 같은 기상 변수도 기여 변수로 밝혀졌다. 또한 인구 밀도, SO_2 및 NH_3 배출량과 같은 인위적 요인이 $PM_{2.5}$ 농도 예측에 중요 지수로 밝혀졌다. 예측 결과를 과정 기반 모델인 GEOS-Chem, CMAQ 예측값과 비교했을 때 랜덤 포레스트 모델이 더욱 나은 성능을 보였다.

박(2021)은 서울시의 2015~2019년 기상 및 대기오염물질 데이터로 월평균 PM_{10}과 $PM_{2.5}$를 예측하고자 했다. 예측 모델로는 선형 회귀(Linear Regression), 서포트 벡터 머신(Support Vector Machine; SVM), 보팅(Voting), 랜덤 포레스트, 엑스트라 랜덤 트리(Extra Random Tree), GBM, XGBoost를 사용하여 총 7개의 기계학습 모델을 비교 평가했다. 개별 모델을 결합하여 예측하는 보팅에는 선형 회귀와 의사결정 회귀(Decision Tree Regression), K-최근접 이웃 회귀(K-Nearest Neighbor; K-NN)를 사용했다. 모델 사용 전 입력 변수 값 범위를 맞추기 위해 정규화 방법으로 데이터 스케일링(Feature scaling)을 진행했다. 모델 평가를 위한 성능 지표는 MAE, MSE, RMSE, MSLE를 사용했다. 결과적으로 $PM_{2.5}$ 농도 예측 시 입력 변수로 기상 데이터만 사용한 경우 XGBoost가 가장 우수했으며, 대기오염물질 데이터까지 사용한 경우 선형 회귀가 다른 모델보다 우수했고 랜덤 포레스트보다 XGBoost가 좋은 성능을 보였다. 다만 선형 회귀로는 입력 변수의 비선형성을 구분하기 어려우므로 랜덤 포레스트, XGBoost와 같은 비선형 모형에서 각 입력 변수의 중요도를 우선 확인하는 과정이 필요하다는 단점이 있다.

이와 이(2020)는 미세먼지를 시간 단위로 예측하는 것이 중요하다고 판단하여 $PM_{2.5}$ 생성과 연관된 입력 변수들을 시계열로 묶어 서울시의 시간당 $PM_{2.5}$를 예측하고자 했다. 입력 변수로는 2015년부터 2018년까지의 서울시 기상 및 대기오염물질 데이터와 중국과 인접한 지리적 특성을 반영하여 백령도의 기상, 대기오염물질 수치를 사용했다. 그리고 시간당 예측을 위해 입력 변수와 목적 변수를 타입 스텝 n으로 묶어 시계열 전처리를 수행했다. 예측 등급 분류를 위해 한국, 미국, 일본 등이 사용하는 $PM_{2.5}$ 분류 기준을 참고했는데, '매우 나쁨' 등급의 경우 전체에서 차지하는 데이터의 비율이 낮아 '나쁨'과 함께 묶어 학습을 진행했다. 상위 10개의 변수 중요도를 확인한 결과, $PM_{2.5}$의 생성과 관련된 NO_2, CO, O_3, SO_2가 높은 순위를 차지했고 백령도의 미세먼지(PM_{10}, $PM_{2.5}$) 데이터 또한 중요하게 나타났다. 본 연구에서 제안한 랜덤 포레스트 방법과 기존에 주로 사용되던 모델들(LSTM, Logistic Regression)의 비

교 실험을 진행한 결과, F-1 score가 0.775~0.8인 랜덤 포레스트가 모든 레이블 및 타임 스텝에서 가장 뛰어난 성능을 보였다.

원 등(2021)은 경기도 반월시화국가산업단지의 대기오염물질과 기상, 교통자료를 활용하여 각 변수 간 유의성을 파악하고 다양한 기계학습 모형 및 딥러닝으로 산업단지의 시간별 $PM_{2.5}$를 예측하고자 했다. 랜덤 포레스트, XGBoost 외에도 LightGBM (Light Gradient Boosting Model), DNN (Deep Neural Network), 보팅 모델을 사용했으며 비교 분석을 위해 에어코리아의 24시간 예측 이동평균 방법을 활용했다. 예측 전 회귀 분석을 통해 해당 변수들이 $PM_{2.5}$ 농도와 유의미한 관계($R^2=0.804$)이고 모든 변수의 P-value가 0.05 미만으로 유의하며, VIF (Variance Inflation Factor)값이 10 미만으로 변수 간의 관계가 독립적임을 확인했다. 모델별 $PM_{2.5}$ 농도를 RMSE 기준으로 예측한 결과 모두 6.70 μg m^{-3} 미만으로 에어코리아 예측 모델인 10.77 μg m^{-3}에 비해 매우 높은 성능을 보였다. 랜덤 포레스트와 XGBoost는 각각 6.27 μg m^{-3}, 6.42 μg m^{-3}로 랜덤 포레스트가 조금 더 나은 예측 결과를 보였다.

5. 결론 및 유의사항

기계학습 기법을 사용한 국가적 혹은 지역적 규모의 $PM_{2.5}$ 농도 예측 연구에서는 다음과 같은 결론을 확인할 수 있었다. 랜덤 포레스트 혹은 XGBoost 기반으로 개발된 $PM_{2.5}$ 농도 예측 모델은 수치 예보 모델이나 다른 기계학습, 딥러닝 기법과 비교하였을 때 대부분 유사하거나 더 좋은 예측 성능을 보였다. 연구마다 사용된 입력 변수가 달랐지만 언급된 중요 입력 변수로는 AOD, PM_{10}, 상대습도, 최대풍속 등 PM의 생성 혹은 고농도 사례 발생 요인과 관련된 변수들이 있었다. 특히 해외 사례와 달리 국내에서는 지리적 특성과 편서풍으로 인해 중국의 영향을 받으므로, 이를 고려하는 것이 중요한 것으로 판단한 연구 결과들이 눈에 띄었다. 이를 위해 중국의 대기질 데이터 혹은 백령도의 기상과 대기오염물질 데이터를 활용하여 영향을 파악하고자 했고, 중요도 지수 확인 결과에서도 중국과 관련된 입력 변수들이 PM 농도 예측에 주요한 영향을 끼치는 것으로 나타났다.

연구 논문을 살펴본 결과 각 연구마다 $PM_{2.5}$ 예측에 대해 가장 우수했다고 하는 모델이 서로 상이했다. 사용하는 입력 변수와 데이터의 구성에 따라 모델의 예측 결과가 달라지므로 연구자는 기계학습 기법 사용 시 다양한 알고리즘을 대입해보고, 연구에 가장 알맞은 모델을 사용해야 한다. 추가적으로 랜덤 포레스트 모델의 개선 방법에 대해서 입력 변수의 개수를 두고 상반되는 의견이 있었다. Park et al. (2019)은 입력 변수의 개수와 기간을 늘리면 모델 성능이 개선될 것이라고 언급했다. 하지만 Hu et al. (2017)은 중요하지 않은 입력 변수를 제외해야 성능이 향상되었다고 밝혔다. 이는 랜덤 포레스트 모델이 보이는 현상으로, 모델에 사용되는 예측 변수 중 2개 이상의 변수가 선형 또는 비선형 상관관계에 있을 때 변수 중요도가 낮게 편향된 수치를 보인다. 이는 비슷한 변수들이 입력된 경우에 각 변수가 하나의 랜덤 포레스트에서 비슷한 빈도로 선택되므로 예측 변수들이 서로에게 혼동을 주기 때문에 서로의 변수 중요도를 낮추는 상황이 발생할 수 있다(Breiman, 2001). 따라서 이러한 편향 문제가 남아있기 때문에 연구자는 적합한 변수 선택을 해야 하며(Genuer et al., 2010; Gregorutti et al., 2017), 다양한 변수 조합의 예측 정확도를 비교하여 식별하거나 여러 선행 연구를 살펴 어떤 입력 변수가 중요하게 작용하는지 확인해야 함을 제시하고자 한다.

사 사

이 연구는 2020년도 과학기술정보통신부의 재원으로 수행된 한국연구재단 신진연구지원사업(NRF-2020R1C1C1012354)과 2021년도 환경부 "미세먼지관리 전문인력 양성사업"의 지원을 받아 수행되었습니다.

참고문헌

구윤서, 윤희영, 권희용, 유숙현. 2010. 미세먼지 예보시스템 개발. 한국대기환경학회지 26(6): 666-682.

길준수, 이미혜. 2021. 인공신경망 모델과 배경대기 측정자료를 활용한 서울시 $PM_{2.5}$ 농도 단기예측 및 입력변수의 기여도 분석. 한국대기환경학회지 37(6): 862-870.

김용대, 조광현. 2013. 빅데이터와 통계학. 한국데이터정보과학회지 24(5): 959-974.

김인중. 2014. Deep Learning: 기계학습의 새로운 트랜드. 정보와 통신 31(11): 52-57.

김인호, 이경섭. 2020. 트리 기반 앙상블 방법을 활용한 자동 평가 모형 개발 및 평가: 서울특별시 주거용 아파트를 사례로. 한국데이터정보과학회지 31(2): 375-389.

박서희, 김미애, 임정호. 2021. 부스팅 기반 기계학습기법을 이용한 지상 미세먼지 농도 산출. 대한원격탐사학회지 37(2): 321-335.

박홍진. 2021. 기상 데이터와 대기 환경 데이터 기반 (초) 미세먼지 분석과 예측. 한국정보전자통신기술학회 논문지 14(4): 328-337.

손건태, 하미향, 이순환. 2016. CMAQ 예측치를 이용한 서울지역 PM_{10} 농도 예측모형. Journal of The Korean Data Analysis Society (JKDAS) 18: 3001-3009.

신지안, 문지훈, 노승민. 2021. 설명 가능한 정기예금 가입 여부 예측을 위한 앙상블 학습 기반 분류 모델들의 비교 분석. 한국전자거래학회지 26(3): 97-117.

원동준, 김선겸, 김영훈, 송규원. 2021. 기계학습을 활용한 경기도 산업단지 미세먼지 예측. 정보과학회논문지 48(7): 764-773.

이득우, 이수원. 2020. 시계열 데이터와 랜덤 포레스트를 활용한 시간당 초미세먼지 농도 예측. 정보처리학회논문지 9(4): 129-136.

이영섭, 김현구, 박종석, 김희경. 2006. 변수변환을 통한 포항지역 미세먼지의 통계적 예보모형에 관한 연구. 한국대기환경학회지 22(5): 614-626.

정옥진, 문윤섭. 2015. WRF-Chem 모델을 이용한 2010년 한반도의 황사 예측에 관한 연구. 한국지구과학회지 36(1): 90-108.

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

Amit, Y. and D. Geman. 1997. Shape quantization and recognition with randomized trees. Neural Computation 9(7): 1545-1588.

Atkinson, C. J., J. D. Fitzgerald and N. A. Hipps. 2010. Potential mechanisms for achieving agricultural benefits from biochar application to temperate soils: a review. Plant and Soil, 337(1): 1-18.

Breiman, L. 2001. Random forests. Machine Learning 45(1): 5-32.

Brokamp, C., R. Jandarov, M. Hossain and P. Ryan. 2018. Predicting daily urban fine particulate matter concentrations using a random forest model. Environmental Science & Technology 52(7): 4173-4179.

Cadelis, G., R. Tourres and J. Molinie. 2014. Short-term effects of the particulate pollutants contained in Saharan dust on the visits of children to the emergency department due to asthmatic conditions in Guadeloupe (French Archipelago of the Caribbean). Plos One 9(3): e91136.

Chen, G., S. Li, L. D. Knibbs, N. A. Hamm, W. Cao, T. Li, J. Guo, H. Ren, M. J. Abramson and Y. Guo. 2018. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information. Science of the Total Environment 636: 52-60.

Chen, T. and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785-794.

Choubin, B., M. Abdolshahnejad, E. Moradi, X. Querol, A. Mosavi, S. Shamshirband and P. Ghamisi. 2020. Spatial hazard assessment of the PM_{10} using machine learning models in Barcelona, Spain. Science of The Total Environment 701: 134474.

Correia, A. W., C. A. Pope III, D. W. Dockery, Y. Wang, M. Ezzati and F. Dominici. 2013. The effect of air pollution control on life expectancy in the United States: an analysis of 545 US counties for the period 2000 to 2007. Epidemiology 24(1): 23.

Di, Q., H. Amini, L. Shi, I. Kloog, R. Silvern, J. Kelly, M. B. Sabath, C. Choirat, P. Koutrakis, A. Lyapustin, Y. Wang, L. J. Mickley and J. Schwartz. 2019. An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. Environment International 130: 104909.

Díaz-Robles, L. A., J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson and J. A. Moncada-Herrera. 2008. A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile. Atmospheric Environment 42(35): 8331-8340.

Fang, J. K., M. A. Mello Athayde, C. H. Schonberg, D. I. Kline, O. Hoegh Guldberg and S. Dove. 2013. Sponge biomass and bioerosion rates increase under ocean warming and acidification. Global Change Biology 19(12): 3581-3591.

Fang, X., B. Zou, X. Liu, T. Sternberg and L. Zhai. 2016. Satellite-based ground $PM_{2.5}$ estimation using timely structure adaptive modeling. Remote Sensing of Environment 186: 152-163.

Genuer, R., J. M. Poggi and C. Tuleau-Malot. 2010. Variable selection using random forests. Pattern Recognition Letters 31(14): 2225-2236.

Gregorutti, B., B. Michel and P. Saint-Pierre. 2017. Correlation and variable importance in random forests. Statistics and Computing 27(3): 659-678.

Gupta, P., S. Zhan, V. Mishra, A. Aekakkararungroj, A. Markert, S. Paibong and F. Chishtie. 2021. Machine learning algorithm for estimating surface $PM_{2.5}$ in Thailand. Aerosol and Air Quality Research 21(11): 210105.

Ho, T. K. 1995. Random decision forests, In Proceedings of 3rd international conference on document analysis and recognition, Vol. 1, pp. 278-282.

Ho, C.-H., I. Park, H.-R. Oh, H.-J. Gim, S.-K. Hur and J. Kim. 2021. Development of a $PM_{2.5}$ prediction model using a recurrent neural network algorithm for the Seoul metropolitan area, Republic of Korea. Atmospheric Environment 245: 118021.

Hrust, L., Z. B. Klaić, J. Križan, O. Antonić and P. Hercog. 2009. Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. Atmospheric Environment 43(35): 5588-5596.

Hu, X., J. H. Belle, X. Meng, A. Wildani, L. A. Waller, M. J. Strickland and Y. Liu. 2017. Estimating $PM_{2.5}$ concentrations in the conterminous United States using the random forest approach. Environmental Science & Technology 51(12): 6936-6944.

Jacob, D. J. and D. A. Winner. 2009. Effect of climate change on air quality. Atmospheric Environment 43(1): 51-63.

Jeong, Y., Y. Youn, S. Cho, S. Kim, M. Huh and Y. Lee. 2020. Prediction of Daily PM_{10} Concentration for Air Korea Stations Using Artificial Intelligence with LDAPS Weather Data, MODIS AOD, and Chinese Air Quality Data. Korean Journal of Remote Sensing 36(4): 573-586.

Joharestani, M. Z., C. Cao, X. Ni, B. Bashir and S. Talebiesfandarani. 2019. $PM_{2.5}$ prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere 10(7): 373.

Kim, S. B., K. Yumimoto, I. Uno and Y. Chun. 2011. Dust model intercomparison between ADAM and CFORS/Dust for Asian dust case in 2007 (March 28-April 3), Sola, 7 (Special_Edition): 25-28.

Lin, C., Y. Li, A. K. Lau, X. Deng, K. T. Tim, J. C. Fung, C. Li, Z. Li, X. Lu, X. Zhang and Q. Yu. 2016. Estimation of long-term population exposure to $PM_{2.5}$ for dense urban areas using 1-km MODIS data. Remote Sensing of Environment 179: 13-22.

Lin, C., Y. Li, Z. Yuan, A. K. Lau, C. Li and J. C. Fung. 2015. Using satellite remote sensing data to estimate the high-resolution distribution of ground-level $PM_{2.5}$. Remote Sensing of Environment 156: 117-128.

Ma, R., J. Ban, Q. Wang, Y. Zhang, Y. Yang, M. Z. He, S. Li, W. Shi and T. Li. 2021. Random forest model based fine scale spatiotemporal O_3 trends in the Beijing-Tianjin-Hebei region in China, 2010 to 2017. Environmental Pollution 276: 116635.

Ma, Z., Y. Liu, Q. Zhao, M. Liu, Y. Zhou and J. Bi. 2016. Satellite-derived high resolution $PM_{2.5}$ concentrations in Yangtze River Delta Region of China using improved linear mixed effects model. Atmospheric Environment 133: 156-164.

OECD. 2017. OECD Environmental Performance Reviews, OECD Publishing.

Özdemir, U. and S. Taner. 2014. Impacts of meteorological factors on PM_{10}: Artificial neural networks (ANN) and multiple linear regression (MLR) approaches. Environmental Forensics 15(4): 329-336.

Pang, J., Z. Liu, X. Wang, J. Bresch, J. Ban, D. Chen and J. Kim. 2018. Assimilating AOD retrievals from GOCI and VIIRS to forecast surface $PM_{2.5}$ episodes over Eastern China. Atmospheric Environment 179: 288-304.

Park, S., M. Shin, J. Im, C. K. Song, M. Choi, J. Kim, S. Lee, R. Park, J. Kim, D. W. Lee and S. K. Kim. 2019. Estimation of ground-level particulate matter concentrations through the synergistic use of satellite observations and process-based models over South Korea. Atmospheric Chemistry and Physics 19(2): 1097-1113.

Reid, C. E., M. Jerrett, M. L. Petersen, G. G. Pfister, P. E. Morefield, I. B. Tager, S. M. Raffuse and J. R. Balmes. 2015. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. Environmental Science & Technology 49(6): 3887-3896.

Singh, B., P. Sihag and K. Singh. 2017. Modelling of impact of water quality on infiltration rate of soil by random forest regression. Modeling Earth Systems and Environment 3(3): 999-1004.

Siroky, D. S. 2009. Navigating random forests and related advances in algorithmic modeling. Statistics Surveys 3: 147-163.

Strobl, C., J. Malley and G. Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods 14(4): 323-348.

Wong, P. Y., H. Y. Lee, Y. C. Chen, Y. T. Zeng, Y. R. Chern, N. T. Chen, S. C. Lung, H. Su and C. D. Wu. 2021. Using a land use regression model with machine learning to estimate ground level $PM_{2.5}$. Environmental Pollution 277: 116846.

Yang, L., H. Xu and Z. Jin. 2019. Estimating ground-level $PM_{2.5}$ over a coastal region of China using satellite AOD and a combined model. Journal of Cleaner Production 227: 472-482.

Yang, L., H. Xu and S. Yu. 2020. Estimating $PM_{2.5}$ concentrations in Yangtze River Delta region of China using random forest model and the Top-of-Atmosphere reflectance. Journal of Environmental Management 272: 111061.

You, W., Z. Zang, L. Zhang, Y. Li, X. Pan and W. Wang. 2016. National-scale estimates of ground-level $PM_{2.5}$ concentration in China using geographically weighted regression based on 3 km resolution MODIS AOD. Remote Sensing 8(3): 184.

Zhang, L., Y. Ji, T. Liu and J. Li. 2020. $PM_{2.5}$ Prediction Based on XGBoost. In 2020 7th International Conference on Information Science and Control Engineering (ICISCE), pp. 1011-1014. IEEE.

Zhao, C., Q. Wang, J. Ban, Z. Liu, Y. Zhang, R. Ma, S. Li and T. Li. 2020. Estimating the daily $PM_{2.5}$ concentration in the Beijing-Tianjin-Hebei region using a random forest model with a $0.01° \times 0.01°$ spatial resolution. Environment International 134: 105297.

