Journal of the Korean Data Analysis Society (August 2020), 22(4), 1661-1671.
https://doi.org/10.37727/jkdas.2020.22.4.1661

XGBoost 앙상블에 의한 서울시 초미세먼지 예측*

김혁[1]

요 약

미세먼지는 사람의 건강에 많은 영향을 미치는 물질로서 환경과 건강에 대한 관심이 높아지면서 이와 관련하여 다양한 연구가 이루어지고 있다. 대기환경 분야에서는 전통적으로 수치모형을 이용하여 미세먼지와 관련된 연구를 수행하였으나 기계학습 분야가 발달함에 따라 최근 들어 기계학습을 이용한 연구도 수행하고 있다. 본 연구는 초미세먼지 농도의 예측에 관한 연구로서 기존의 연구들이 대부분 단기(1시간~2일 후) 예측에 그치는데 반해 중기(1주일 후) 예측을 시도하고 있다. 서울지역의 일평균 초미세먼지 농도를 보통(36μm미만)과 나쁨(36μm이상)의 이 단계로 나누어 이를 예측하는 분류모형을 만들었다. 분류모형으로는 XGBoost를 사용하였으며 모형의 성능 극대화를 위하여 각기 다른 평가지표를 기준으로 초모수 최적화(hyperparameter optimization)를 시도한 개별 모형들을 만든 후 앙상블 모형을 구축하였다. 2016년부터 2018년까지 3년의 자료를 이용하여 모형을 훈련시킨 후 2019년 1월부터 6월까지의 서울시 일 평균 미세먼지 농도를 예측한 결과 단일 모형보다 본 연구에서 제시한 앙상블 모형의 성능이 좋음을 확인하였다. 또한 미세먼지의 등급이 나쁜 경우보다 보통인 경우가 훨씬 많은 것을 고려하여 범주의 불균형 상태를 보완하는 방법을 적용하였으나 환경물질 자료만 설명변수로 사용할 때는 효과적이나 환경물질과 기상 정보에 관한 변수들을 모두 사용할 때는 오히려 모형의 성능에 악영향을 가져오는 것을 확인하였다.

주요용어 : 미세먼지, 예측, 기계학습, 앙상블, XGBoost.

1. 서론

미세먼지는 아주 작은 크기의 오염 물질을 말하며 먼지의 크기에 따라 미세먼지(직경 10μm이하)와 초미세먼지(직경 2.5μm이하)로 구분할 수 있다. 환경에 대한 국민들의 관심이 커짐에 따라 미세먼지에 대한 국가적 종합계획이 수립되었으며 이에 따라 다양한 연구 및 정책 등이 진행 중이다. 기상 예보와 마찬가지로 미세먼지 예보에 대해서도 많은 연구가 이루어지고 있으며 최근에는 기계학습 방법의 발달에 따라 기존의 대기과학 분야에 사용하던 수치해석모형이 아닌 기계학습 모형을 이용한 연구도 이루어지고 있다.

본 연구에서는 서울시의 일평균 초미세먼지 농도를 예측하는 모형을 생성하였다. 대부분의 미세먼지 예측 모형들이 단기 예측(1시간 후 ~ 2일)에 그치고 있는데 반해 본 연구에서는 중기 예측(1

*이 논문은 2018년도 호서대학교의 재원으로 학술연구비 지원을 받아 수행된 연구임(20180306).
[1]31499 충청남도 아산시 배방읍 호서로79번길 20, 호서대학교 과학기술융합대학 빅데이터경영공학부 부교수.
E-mail : hkim@hoseo.edu
[접수 2020년 8월 5일; 2020년 8월 17일; 게재확정 2020년 8월 20일]

주일)을 시도하였다. 미세먼지 농도에 영향을 주는 다양한 대기, 환경 변수들의 시공간적 특성 때문에 예측 기간이 길어짐에 따라 충분한 예측력을 갖춘 모형을 만드는 것이 쉽지 않으나 중기 예측 결과는 지자체에서 미세먼지 대책을 적용하거나 일반 국민이 여행을 계획하는 등 모형의 활용도가 많은 주제이다.

미세먼지보다는 초미세먼지에 대한 관심이 커지고 있고, 대부분의 사람들은 미세먼지의 정확한 수치보다는 미세먼지 등급에 관심을 가지는 점에 착안하여 예측하고자 하는 반응변수를 다음과 같이 정의하였다. 초미세먼지 예보 등급 좋음(0~15μm)과 보통(16~36μm)을 보통으로 분류하고, 나쁨(36~75μm)과 매우 나쁨(76μm 이상)을 나쁨으로 분류하여 4단계로 이루어진 예보 등급에 대한 예측을 이진분류모형으로 단순화하였다. 본 연구는 다음과 같이 구성되었다. 2장에서는 미세먼지와 관련된 선행 연구들을 살펴보며, 3장에서는 본 연구에서 활용하는 다양한 대기, 환경 자료와 모형의 설계 과정을 소개한다. 4장에서는 3장에서 소개한 자료를 이용하여 단일 모형과 앙상블 모형을 구축하며 생성된 모형들의 성능을 평가한다. 그리고 마지막 장에서는 결론과 함께 추가적인 연구 방향을 제시한다.

2. 선행연구

미세먼지는 환경과학에서 다루는 물질이나 미세먼지에 대한 관심이 커지면서 여러 분야에서 다양한 연구들이 수행되고 있다. 먼저 미세먼지의 공간적인 특성을 고려한 지역 간 연관성에 대한 Park(2019)의 연구와 동태적 조건부 상관구조를 활용한 Kim, Park(2019)한 연구가 있다. 또한 Choi, Park(2016)은 축소된 상자그림을 활용하여 미세먼지 데이터의 군집분석을 시도하였다.

본 연구의 주제인 미세먼지의 예측과 관련해서는 적용된 방법론에 따라 선행연구를 다음과 같이 분류할 수 있다. 회귀분석 모형을 적용한 연구로는, 일반선형회귀모형보다 주성분회귀모형의 성능이 좋음을 보인 Nazif et al.(2017)와 단계별 회귀분석의 적용을 제안한 Ahani, Salari, Shadman(2019)의 연구가 있다. 시계열 분석을 활용한 연구로는 서울시 미세먼지 예측을 시도한 Oh, Shin, Shin, Jeong(2017), 중국 타이위안시의 미세먼지 예측을 시도한 Zhang et al.(2017), 터키의 앙카라시 미세먼지 예측을 시도한 Akdi et al.(2020) 등이 있다. 전통적인 기계학습 모형을 이용한 연구로는, Random Forests(Breiman, 2001)를 활용한 Brokamp et al.(2018)와 support vector machine(Cortes, Vapnik, 1995)을 활용한 Zhou et al.(2019) 등의 연구가 있다. 2010년 이후 Deep Learning(Hinton, Osindero, Teh, 2006) 방법이 여러 분야에서 널리 활용되었다. 미세먼지 예측 모형에서도 딥러닝을 활용한 여러 연구가 시도되었다. Li et al.(2016)의 연구는 DNN(Deep Neural Networks)을 이용한 미세먼지 예측 모형을 구축하는데 있어서 모형의 기본 구조를 어떤 식으로 구성해야 하는지에 대한 연구이며 Soh, Chang, Huang(2018)는 DNN을 활용하여 대만과 중국의 베이징의 미세먼지 예측을 시도한 연구이다. RNN(Recurrent Neural Networks)을 적용한 Ong, Sugiura, Zettsu(2016)와 LSTM(Long Short-Term Memory)을 활용한 Li et al.(2017)는 모두 미세먼지의 시계열적 특성을 고려하였다. 한편 시공간적인 특성을 모두 고려하여 딥러닝의 CNN(Convolutional Neural Networks)과 LSTM의 융합 모형을 적용한 Pak et al.(2020)의 연구도 있다. 한편 단일 모형이 아니라 여러 모형들의 앙상블 모형을 제안한 연구들도 있다. Kuang et al.(2020)는 랜덤 포레스트, Gradient Boosting, K-nearest nearbor 모형을 앙상블하였으며 Joharestani et al.(2019)는 랜덤 포레스트, XGBoost, 딥러닝을 앙상블

로 하여 최종모형을 제안하였다. 그 외에 대부분의 미세먼지 예측 모형이 지역 단위로 이루어지는 데 반해 개별 집 단위의 예측을 시도한 Tong et al.(2020)의 연구와 Cho, An(2018)과 같이 천식, 당뇨, 고혈압, 아토피, 이상지질형증의 유병률과 미세먼지 농도간의 관계를 베타회귀분석을 이용하여 분석한 연구 등도 있다.

3. 연구 방법

3.1. 자료

최근 몇 년간 공공데이터에 대한 개방이 적극적으로 이루어짐에 따라 본 연구에서 사용한 기상 및 환경물질에 대한 자료들은 관련 인터넷 사이트에서 모두 자유롭게 이용할 수 있다. 기상 자료는 기상자료개방포털(datat.kma.go.kr)의 ASOS(Automated Synoptic Observing System) 관측자료들을 활용하고, 환경물질에 대한 자료는 에어코리아(www.airkorea.or.kr)에서 제공하는 자료들을 사용하였다. ASOS는 종관기상관측장비로 전국 500여 지역에 설치된 방재기상관측장비와는 달리 96개 지점에만 설치되어 활용되고 있다. 개수가 적은 만큼 더 정교한 관측이 가능하며 일부 관측 요소에 대해서는 사람에 의한 관측도 실시되고 있다. 환경물질 자료로는 SO2, CO, O3, NO2, PM10, PM2.5을 활용하며 기상 자료로는 기온, 강수, 풍속, 풍향, 이슬점온도, 습도, 증기압, 현지기압, 해면기압, 일조, 일사, 적설, 3시간 신적설, 전운량, 중하층운량(10분위), 운형, 최고운고, 시정, 지면상태, 현상번호, 지면온도, 5cm 지중온도, 10cm 지중온도, 20cm 지중온도, 30cm 지중온도를 수집하였다. 서울 전역의 관측소로부터 2016년 1월 1일부터 2019년 6월 30일까지 관측된 자료들로부터 각각의 변수들에 대한 일 평균값을 계산하였다. 평균 계산을 하는데 있어서 결측이 발생한 값에 대해서는 R의 missForest 패키지에서 제공하는 MissForest 기법(Stekhoven, Buehlmann, 2012)에 의해서 결측값을 대체하였다. MissForest 기법은 랜덤 포레스트를 이용한 비모수적 결측값 대체 방법으로 빠른 실행속도와 수치형 변수 외에 범주형 변수에도 적용 가능한 장점 때문에 널리 사용되고 있다. 그리고 본 연구에서 예측하고자 하는 것은 중기 예측이기 때문에 1주일 뒤의 서울시 일 평균 미세먼지 농도를 보통(36μm미만)과 나쁨(36μm이상)으로 분류하여 반응변수로 사용하였다. 본 연구에서 훈련용 자료로 활용하는 기간은 2016년 1월 1일부터 2018년 12월 31일까지이며 연도별 서울시 일 평균 초미세먼지 농도와 농도 등급이 나쁨인 일수는 다음과 같다.

Table 1. Daily average PM2.5 concentration in Seoul between 2016 and 2018

Year	Average PM2.5 concentration	The number of days whose PM2.5 status is bad (PM2.5 concentration is greater than 36μm.)
2016	26.3	73
2017	25.6	64
2018	23.8	61

또한, 평가용 자료로 활용하는 2019년 1월 1일부터 2019년 6월 30일까지의 서울시 일 평균 초미세먼지 농도에 대해서는 월별로 구분하면 다음과 같다.

Table 2. Daily average PM2.5 concentration in Seoul between January 2019 and June 2019.

Month	Average PM2.5 concentration(μm)	The number of days whose PM2.5 status is bad (PM2.5 concentration is greater than 36μm.)
January	37.6	11
February	35.4	12
March	44.6	13
April	20.7	3
May	20.9	11
June	19.8	2
Average	31.2	8.7

우리 나라의 (초)미세먼지는 계절적 영향으로 인하여 겨울과 봄에 심해졌다가 여름, 가을에 약해지는 특성을 보이고 있다. 이는 2019년에도 동일하나 2019년 5월에는 2019년 5월 11일부터 2019년 5월 17일 사이의 이상 현상(7일 중 6일간 초미세먼지 농도가 나쁨)으로 인하여 초미세먼지 농도 등급이 나쁜 일수가 다른 해의 5월과는 다르게 비정상적으로 많았다.

3.2. 분류 모형 설정

본 연구에서 사용하는 분류 모형은 Chen, Guestrin(2016)에 의해 개발된 XGBoost(eXtreme Graditent Boosting)이다. XGBoost의 기본형은 그래디언트 부스팅(gradient boosting)으로, 이는 20년 전에 개발되었으나(Friedman, 2001; Friedman, 2002) 기존의 그래디언트 부스팅 방법의 단점인 느린 수행 시간 및 과적합 규제 부재 등의 문제를 해결함으로서 빠른 학습 속도와 뛰어난 성능을 보이는 분류 모형이다. XGBoost는 그래디언트 부스팅과 마찬가지로 여러 초모수(hyperparameter)들을 설정할 수 있다. 본 연구에서 탐색한 초모수들과 각 모수별 특징은 다음과 같다(Kim, 2019).

1. 반복 회수: 부스팅(boosting)의 시행 회수이다. 반복 회수가 증가함에 따라 모형의 성능이 좋아지나 너무 큰 반복 회수는 오히려 과적합이 발생할 수 있다.

2. 학습률: 값이 너무 작으면 학습하는데 많은 시간이 걸리며, 값이 너무 크면 반복 회수가 많더라도 모형의 최적화에 실패할 수 있다.

3. gamma(또는 min split loss): 가지를 분할하기 위해서 요구되는 손실함수의 축소값을 의미한다. 모형의 정규화에 영향을 미치며 큰 값을 가질수록 보수적인 모형이 만들어진다.

4. 최대 허용 깊이: XGBoost 모형은 여러 개의 가지를 만들게 되는데 이렇게 만들어지는 모형의 최대 깊이를 의미한다. 깊이가 깊을수록 복잡한 모형이 만들어지나 과적합의 가능성도 올라간다.

5. min child weight: 자식 노드에 있어야 하는 최소한의 자료 개수. 큰 값을 가질수록 과적합을 방지하는 역할을 한다.

6. subsample rate: 나무 생성을 할 때 전체 관측치가 아닌 일부 관측치를 활용한다. 적절한 subsample rate을 설정하면 모형의 훈련 시간이 단축되며 과적합을 방지할 수 있다.

7. column sample rate: 나무 생성시 사용하는 입력 변수의 비율이다. 이를 통해 일부 변수만 활용할 수 있으며 subsample rate와 같은 효과를 가진다.

초모수 최적화를 위해서는 훈련용 자료(training set)를 다시 훈련용 자료와 검증용 자료(validation set)로 나눌 필요가 있다. 두 자료를 나누는 일반적인 방법은 k-교차검증(cross validation)이다. 그러나 본 연구에서 활용하는 자료는 시계열자료이기 때문에 훈련용 자료와 검증용 자료에 속하는 개체들을 임의로 정하게 되면 올바른 모형의 평가를 할 수 없다. 그렇기 때문에 일반적인 교차검증과는 다르게 년도별로 구분하여 훈련용 자료와 검증용 자료를 구분하였다. 훈련용 자료와 검증용 자료를 구분한 후에는 다양한 초모수들의 조합을 이용하여 모형들을 만들고, 검증용 자료를 대상으로 모형의 성능을 평가하였다. 최적의 초모수를 찾는 방법으로는 grid search, random search, Bayesian optimization 등이 있는데 본 연구에서는 random search를 500번 적용하여 최적의 조합을 찾았다.

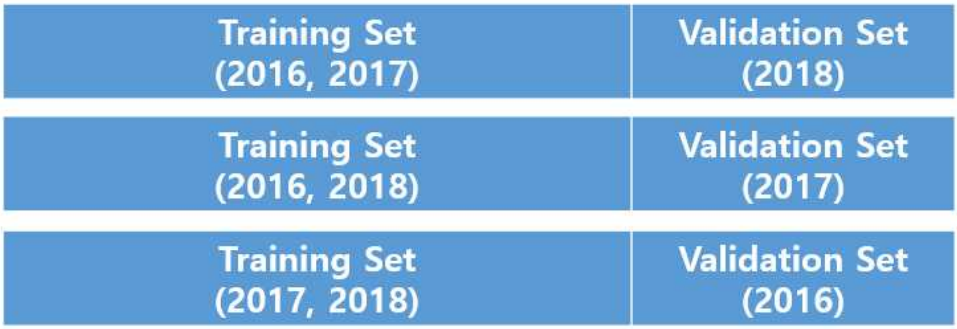

Figure 1. 3-cross validation for the hyperparameter tuning

본 연구에서 사용한 모든 코드는 오픈 소스(open source) 언어인 64비트 R 3.6.1을 활용하였으며 R에서 추가적으로 설치해서 사용한 패키지는 다음과 같다. 대용량의 자료를 빠르게 저장하고 불러오기 위해서 data.table 패키지를 사용했으며 XGBoost 모형의 사용을 위해서 xgboost 패키지를 이용하였다. 결측값 대체를 위해서는 missForest 패키지, 그리고 자료의 불균형 문제를 해소하기 위해서는 ROSE 패키지를 사용하였다. 그 외 교차검증(cross validation), 초모수 최적화들을 위한 과정은 별도의 패키지를 이용하지 않고 직접 코딩하여 구현하였다.

4. 모형 적용 결과

4.1. 단일 분류 모형

활용 가능한 설명변수들을 크게 환경물질에 관련된 변수와 기상 정보에 관련된 변수로 구분하였다. 그리고 환경물질에 관련된 변수들만 이용해서 모형을 만들고, 추가적으로 기상 정보들을 이용한 모형도 만들었다. Table 1에서 초미세먼지가 나쁜 날의 비율은 18.1%이다. 그러므로 반응변수의 소수 범주와 다수 범주 간의 비율을 조정하는 것이 모형의 성능을 향상시킬 수 있다. 반응변수의 범주들의 비율 차이가 있는 것을 불균형문제(imbalance problem)라 하며 이를 해결하는 여러 방법들이 존재한다. 본 연구에서는 소수 범주의 분포를 추정하여 소수 범주에 속하는 가상의 개체들을 생성해내는 ROSE(Random Over-Sampling Examples)(Menardi, Torelli, 2014) 기법을 적용하였다. ROSE 기법은 성능의 우수성과 함께 소수 범주에 대한 가상의 개체를 생성하더라도 분포의 모양이

그대로 유지되는 장점 때문에 널리 사용되는 불균형 해결 기법 중의 하나이다. 그리고 초미세먼지 농도 예측을 위해 고려하는 과거 자료의 기간을 달리해서 모형들을 만들어 보았다. 5일, 10일, 15일, 20일의 네 가지 경우로 구분하여 모형을 만들었다. 이처럼 각기 다른 설명변수들의 사용(2가지), ROSE 기법의 적용 유무(2가지), 각기 다른 과거 자료 사용 기간(4가지) 등 모두 16가지 조합에 대해 모형을 만들었으며 각각의 모형에 대해서는 3.2에서 설명한 초모수 최적화 과정을 적용하였다.

모형의 평가를 위해서는 정확도(accuracy), 특이도(specificity), 민감도(sensitivity), 정밀도(precision)와 함께 민감도와 정밀도의 조화평균인 F1을 사용한다. 초미세먼지 농도의 등급에 대한 자료는 불균형 자료이기 때문에 정확도만을 이용해서 모형의 결과를 평가하는 경우에는 왜곡된 결과를 가져올 수 있기 때문이다. 그러므로 본 연구에서는 민감도와 정밀도를 동시에 고려하는 F1을 이용하여 모형의 성능을 평가한다.

Table 3. Result of the prediction of PM2.5 in Seoul for the single XGBoost

description of the variables used for the prediction	ROSE algorithm	the past days used for the prediction	accuracy	specificity	sensitivity	precision	F1
environmental variables	No	5	0.6188	0.7674	0.2500	0.3023	0.2737
		10	0.6464	0.7829	0.3077	0.3636	0.3333
		15	0.6298	0.7597	0.3077	0.3404	0.3232
		20	0.6354	0.7752	0.2885	0.3409	0.3125
	Yes	5	0.6630	0.7674	0.4038	0.4118	0.4078
		10	0.6519	0.7364	0.4423	0.4035	0.4220
		15	0.6519	0.7287	0.4615	0.4068	0.4324
		20	0.6464	0.7364	0.4231	0.3929	0.4074
environmental variables + atmospheric variables	No	5	0.6740	0.7674	0.4423	0.4340	0.4381
		10	0.7017	0.7829	0.5000	0.4815	0.4906
		15	0.6685	0.7519	0.4615	0.4286	0.4444
		20	0.6740	0.7519	0.4808	0.4386	0.4587
	Yes	5	0.6298	0.6977	0.4615	0.3810	0.4174
		10	0.6851	0.7674	0.4808	0.4545	0.4673
		15	0.6575	0.7209	0.5000	0.4194	0.4561
		20	0.6409	0.6977	0.5000	0.4000	0.4444

Table 3에서 단일 XGBoost 모형으로 가장 좋은 성능을 보이는 조합은 과거 10일치의 환경물질 변수와 기상변수를 사용하고, 반응변수의 불균형상태에 대해서는 보정을 하지 않은 모형이 가장 좋은 성능을 보이는 것을 알 수 있다. ROSE 기법이 환경물질만 적용하는 경우에는 모형의 성능을 더 향상시키나 환경물질변수와 기상변수를 모두 사용할 때는 오히려 모형의 성능에 악영향을 가져오는 이유는 다음과 같다. 반응변수에 대한 불균형 해결 기법을 사용하는 이유는, 소수 범주의 개체수가 부족한 경우에는 소수 범주의 개체에 대한 특징 파악이 쉽지 않기 때문에 해당 범주에 대한 예측력이 많이 낮아지기 때문이다. 그러나 외형적으로 보이는 개수가 적더라도 소수 범주에 대한 특성 파악이 충분히 이루어진다면 불균형 해결 기법의 사용은 오히려 악영향을 가져올 수 있다(He, Garcia, 2009). 환경물질 변수만 사용하는 경우에는 초미세먼지 등급이 나쁜 경우에 대한 충분한 해석이 부족하기 때문에 ROSE 기법을 이용해서 소수 범주에 대한 가상의 자료를 생성하는 것이 도움이 된다. 그러나 환경물질 변수와 함께 기상변수도 사용하는 경우에는 소수 범주의 특성을

파악하기에 충분한 변수들이 확보되었기 때문에 오히려 가상의 자료가 추가되는 것이 모형의 성능에 악영향을 미치는 것이다.

4.2. 앙상블 모형

앙상블 기법은 주어진 자료를 이용하여 여러 개의 서로 다른 단일 모형(single classifier)을 생성한 후, 이들의 예측 결과를 종합하여 하나의 최종 결과를 도출하는 방법이다(Kuncheva, Rodriguez, 2012). 앙상블 모형 내에 속한 모형들 간의 다양성(diversity)의 성질에 의해서 단일 모형보다 좋은 성능을 보이게 되고 간단한 결합으로도 모형의 성능을 향상시킬 수 있기 때문에 예측 모형을 만들 때 널리 사용되는 방법이다(Kuncheva, Whitaker, 2003). 4.1에서는 XGBoost 기법을 활용하여 단일 모형을 구축하였다. 과거 10일치의 환경물질 및 기상 자료를 활용하고, 소수 범주에 대해서는 보정을 하지 않은 채 모형을 만들 때 가장 좋은 성능을 가져왔다. 이렇게 만들어진 여러 개의 XGBoost 모형을 결합하여 앙상블 모형을 만들려고 한다. 앙상블 모형을 설계할 때 중요한 것은 속한 단일 모형들 간의 다양성이다. 아무리 성능이 좋더라도 단일 모형들이 모두 같은 결과를 예측한다면 모형들 간의 결과를 결합하는 것이 의미가 없기 때문이다. 본 연구에서는 XGBoost 기법으로 활용된 모형들만을 가지고 앙상블 모형을 구성한다. 모형들이 다양성을 가져야 하므로 모형들 간에 다른 기준에 의한 초모수 최적화를 시도하였다. 3개의 단일 모형을 구축하였으며 첫 번째 모형은 최대 정확도를 가지도록 초모수 최적화를 시행하였다. 두 번 째 모형은 최대 민감도를 가지도록 초모수 최적화를 시도했으며 마지막 모형은 F1값이 최대가 되도록 초모수 최적화를 시도하였다. 이를 통해 생성된 세 개의 모형들은 각각 다른 평가 기준에 대해 최적의 성능을 보이는 모형들이 된다. 이렇게 생성된 모형에 대해 단순 다수결(simple majority voting)에 의하여 최종 결과를 도출하였으며 위 과정을 10회 반복한 평가 지표들의 평균값은 다음과 같다.

Table 4. Comparison of the model performance between the single models and the ensemble model for the prediction of PM2.5 in Seoul

model	accuracy	specificity	sensitivity	precision	F1
ensemble model	0.7293	0.8140	0.5192	0.5294	0.5243
single model whose hyperparameters are optimized to maximize F1	0.7017	0.7829	0.5000	0.4815	0.4906
single model whose hyperparameters are optimized to maximize the sensitivity	0.6961	0.7752	0.5000	0.4727	0.4860
single model whose hyperparameters are optimized to maximize the accuracy	0.7127	0.8140	0.4615	0.5000	0.4800

Table 4로부터 앙상블 모형이 각각의 단일 모형들보다 더 성능이 좋은 것을 알수 있으며 , 모형들 간의 평균 F1 점수에 대해 통계적 유의성을 검정하기 위해 분산분석을 적용하였다.

Table 5. ANOVA test for comparison of the model performance by F1 score

	DF	Sum Sq	Mean Sq	F-value	p-value
group	3	0.01972	0.00657	4.275	0.0111*
residuals	36	0.05534	0.00154		

또한 모형들 간에 유의미한 성능 차이가 있는지를 쌍으로 비교하기 위해 Tukey의 HSD(honestly significant difference) test를 적용하였다. Table 6으로부터 앙상블 모형은 민감도에 기반한 초모수 최적화를 적용한 단일 모형과 정확도에 기반한 초모수 최적화를 적용한 단일 모형보다 성능이 더 좋은 것을 확인할 수 있다.

Table 6. Tukey's HSD test for the pairwise comparisons

pairwise comparison	diff	lower bound	upper bound	p-value
ensemble model - single model optimized by F1	0.0337	-0.0023	0.0697	0.3124
ensemble model - single model optimized by sensitivity	0.0383	0.0023	0.0743	0.0124*
ensemble model - single model optimized by accuracy	0.0443	0.0083	0.0803	0.0326*
single model by F1 - single model by sensitivity	0.0046	-0.0314	0.0406	0.4436
single model by F1 - single model by accuracy	0.0106	-0.0254	0.0466	0.6784
single model by sensitivity - single model by accuracy	0.0060	-0.0300	0.0420	0.9702

5. 결론

미세먼지는 국민 건강에 많은 영향을 미치는 환경물질로서 2010년대 들어 사람들의 많은 관심을 받고 있다. 국가에서도 정부 차원의 미세먼지 대책을 수립하는 등 많은 연구가 이어지고 있으며 그에 따라 기상 예보와 같은 미세먼지 농도 예보도 많이 활용되고 있다.

본 논문에서는 2016년부터 2018년까지의 환경물질 및 기상 자료를 수집하여 일주일 뒤의 서울시 일 평균 초미세먼지 농도를 예측하는 모형을 만들었다. 기계학습의 여러 분류 모형 중 좋은 성능과 함께 빠른 훈련 속도로 인하여 많이 활용되고 있는 XGBoost를 이용하였으며 500번의 random search를 통해 초모수를 최적화하였다. 다양한 설정에서 최적의 모형을 찾기 위해서 시도하였으며 최종적으로 과거 10일치의 환경물질과 기상 변수들을 모두 사용하고, 소수 범주의 개수를 조정하지 않는 설정에서 가장 좋은 성능을 보였다. 한편 예측력의 극대화를 위해서 앙상블 모형을 구성하였다. 앙상블 모형에 있어서 가장 중요한 점은 앙상블을 구성하는 모형들 간의 다양성을 확보하는 것이다. 세 가지의 다른 모형을 생성하기 위해서 초모수를 최적화할 때 다른 평가지표에 따라 최적화 과정을 거쳤다. 각각의 모형들은 F1을 최대화하도록 초모수가 정의된 모형, 정확도를 최대화하도록 초모수가 정의된 모형, 민감도를 최대화하도록 초모수가 정의된 모형으로 이루어졌으며 이들 모형을 결합한 앙상블 모형이 각기 다른 세 개의 단일 모형들보다 더 좋은 성능을 보이는 것을 확인하였다.

본 연구와 관련한 향후 과제는 다음과 같다. 첫째, 본 연구에서는 XGBoost 모형만을 활용하여 모형들을 구성하였다. 그러나 분류 모형에는 XGBoost 외에 랜덤 포레스트, LightGBM, CatBoost, 딥러닝 등 다양한 기법들이 있다. 이들 기법들을 종합적으로 활용하여 앙상블 모형을 구축하게 되면 더 풍부한 다양성에 의해서 앙상블 모형의 추가적인 성능 향상을 기대할 수 있다. 둘째, 본 연구에서는 미세먼지 농도의 추정을 위해서 과거 10일치의 자료를 이용하여 모형을 훈련시키고 있다. 변수들 간의 시계열적 특성에 대해서는 특별한 조작 없이 입력 변수로 활용하고 있는데, 연속적으로 존재하는 자료 값들의 특성을 반영하여 정교한 변수 변환을 수행할 때 성능 향상을 꾀할 수 있다. 마지막으로, 본 모형에서는 서울의 초미세먼지 예측을 위해서 서울에 소재한 관측소의 정보만을 활용하고 있다. 그러나 미세먼지는 시간에 따라 이동을 하는 물질이고 특히 일주일 뒤의 초미세먼

지를 예측하게 되었을 때는 중국의 영향을 무시할 수 없다(Lee, Ho, Lee, Choi, Song, 2013). 중국 관측소의 공식적인 미세먼지 자료는 쉽게 구할 수 없는 자료로서 본 연구에서는 사용하지 못했으나 향후 미세먼지를 관리하는 국가연구기관인 국립환경과학원의 협조를 얻어 해당 자료를 얻어서 연구에 활용하고자 한다.

References

Ahani, I. K., Salari, M., Shadman, A. (2019). Statistical models for multi-step-ahead forecasting of fine particulate matter in urban areas, *Atmospheric Pollution Research*, 10(3), 689-700.

Akdi, Y., Okkaoglu, Y., Goveren, E., Yucel, M. E. (2020). Estimation and forecasting of PM10 air pollution in Ankara via time series and harmonic regression, *International Journal of Environmental Science and Technology*, 17, 3677-3690.

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

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

Chen, T., Guestrin, C. (2016). XGBoost: A scalable tree boosting system, *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Cho, E. Y., An, H. (2018). Analysis of the relationship between regional prevalence and the average concentration of particulate matter using beta regression, *Journal of the Korean Data Analysis Society*, 20(4), 1791-1800. (in Korean).

Choi, H., Park, C. (2016). Clustering analysis of particulate matter data using shrinkage boxplot, *Journal of the Korean Data Analysis Society*, 18(5B), 2435-2443. (in Korean).

Cortes, C., Vapnik, V. N. (1995). Support-vector networks, *Machine Learning*, 20(3), 273-297.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine, *The Annals of Statistics*, 29(5), 1189-1232.

Friedman, J. H. (2002). Stochastic gradient boosting, *Computational Statistics and Data Analysis*, 38(4), 367-378.

Joharestani, M. Z., Cao, C., Ni, X., Bashir, B., Talebiesfandarani, S. (2019). PM2.5 prediction based on random forests, XGBoost, and deep learning using multisource remote sensing data, *Atmosphere*, 10(7), 373.

Kim, H. (2019). Study on the prediction of the number of spectators and it's factors in pro sports by machine learning method, *Journal of the Korean Data Analysis Society*, 21(4), 1867-1880. (in Korean).

Kim, S. T., Park. M. S. (2019). Particulate matter time-series data analysis based on the dynamic conditional correlation structure, *Journal of the Korean Data Analysis Society*, 21(6), 2859-2871. (in Korean).

Kuncheva, L. I., Rodriguez, J. J. (2012). A weighted voting framework for classifiers ensembles, Knowledge and Information Systems, 38, 259-275.

Kuncheva, L. I., Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, *Machine Learning*, 51(2), 181-207.

He, H., Garcia, E. A. (2009). Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263-1284.

Hinton, G. E., Osindero, S., Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets, *Neural Computation*, 18(7), 1527-1554.

Lee, S., Ho, C. H., Lee, Y. G., Choi, H. J., Song, C. K. (2013). Influence of transboundary air pollutants from China on the high-PM10 episode in Seoul, Korea for the period October 16-20, 2008, *Atmospheric Environment*, 77, 430-439.

Li, X., Peng, L., Hu, Y., Shao, J., Chi, T. (2016). Deep learning architecture for air quality predictions, *Environmental Science and Pollution Research*23, 22408-22417.

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation, *Environmental Pollution*, 231, 997-1004.

Menardi, G., Torelli, N. (2014). Training and assessing classification rules with imbalanced data, *Data Mining and Knowledge Discovery*, 28(1), 92-122.

Nazif, A., Mohammed, N. I., Malakahmad, A., Abualqumboz, M. S. (2017). Regression and multivariate models for predicting particulate matter concentration level, *Environmental Science and Pollution Research*, 25(1), 283-289.

Oh, J., Shin, H., Shin, Y., Jeong, H. C. (2017). Forecasting the particulate matter in Seoul using a univariate time series approach, *Journal of the Korean Data Analysis Society*, 19(5B), 2457-2468. (in Korean).

Ong, B. T., Sugiura, K., Zettsu, K. (2016). Dynamically pre-trained deep recurrent neural networks using environmental monitoring data for predicting PM2.5, *Neural Computing and Applications*, 27, 1553-1566.

Pak, U., Ma, J., Ryu, U., Ryom, K., Juhyok, U., Pak, K., Pak, C. (2020). Deep learning-based PM2.5 prediction considering the spatiotemporal correlations: A case study of Beijing, China, *Science of the Total Environment*, 699(10), 133561.

Park, M. S. (2019). Regional assoficatino of the particulate matters, *Journal of the Korean Data Analysis Society*, 21(3), 1169-1181. (in Korean).

Soh, P.-W., Chang, J.-W., Huang, J.-W. (2018). Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations, *IEEE Access*, 6, 38186-38199.

Stekhoven, D. J., Buehlmann, P. (2012). MissForest – nonparametric missing value imputation for mixed-type data, *Bioinformatics*, 28(1), 112-118.

Tong, X., Ho, J. M. W., Li, Z., Lui, K. H., Kwok, T. C. Y., Tsoi, K. K. F., Ho, K. F. (2020). Prediction model for airparticulate matter levels in the households of elderly individuals in Hong Kong, *Science of The Total Environment*, 717, 135323.

Yazdi, M., Kuang, Z., Dimakopoulou, K., Barratt, B., Suel, E., Amini, H., Lyapustin, A., Katsouyanni, K., Schwartz, J. (2020). Predicting fine particulate matter (PM2.5) in the Greater London area: An ensemble approach using machine learning methods, Remote Sensing, 12(6), 914.

Zhang, H., Zhang, S., Wang, P., Qin, Y., Wang, H. (2017). Forecasting of particulate matter time series using wavelet analysis and wavelet-ARMA/ARIMA model in Taiyuan, China. *Journal of the Air & Waste Management Associaion*, 67(7), 776-788.

Zhou, Y., Chang, F. J., Chang, L. C., Kao, I. F., Wang, Y. S., Kang, C. C. (2019). Multi-ouput support vector machine for regional multi-step-ahead PM2.5 forecasting, *Science of The Total Environment*, 651(1), 230-240.

The Prediction of PM2.5 in Seoul through XGBoost Ensemble*

Hyeuk Kim[1]

Abstract

Fine dust is a substance that greatly affects human health, and as interest in the environment and health increases, various studies have been conducted in this regard. In the field of atmospheric environment, studies related to fine dust have been traditionally performed using numerical models, but the research based on the machine learning has been conducted recently as the field of machine learning has developed. This study is about the prediction of ultrafine dust concentration, and it attempts to predict the middle (after 1 week) prediction while most of the existing studies are only short-term (after 1 hour to 2 days) prediction. A classification model was develeped to predict the average daily ultrafine dust concentration in Seoul by dividing it into two stages: normal (less than 36) and bad (more than 36). XGBoost was used as the classification model and an ensemble model was constructed after creating individual models through the hyperparameter optimization based on different evaluation indicators in order to maximize the performance of the model. The average daily fine dust concentration in Seoul from January to June 2019 was predicted after training the model using data from 2016 to 2018, and the performance of the ensemble model presented in this study was higher than that of a single model. In addition, the method of imbalance problem was applied since the number of cases where the level of fine dust is bad is much less than the number of cases where the level of fine dust is normal. It works effectively when only environmental material data is used as explanatory variables, but it adversely affects the performance of the model when both variables related to environmental materials and weather information are used.

Keywords : particulate matter, prediction, machine learning, ensemble, XGBoost.

*This Research was supported by the Academic Research Fund of Hoseo University in 2018(20180306).
[1]31499 20, Associate Professor, Division of Big Data and Management Engineering, College of Convergence Science & Technology, Hoseo University, 20, Hoseo-ro 79beon-gil, Baebang-eup, Asan-si, Chungcheongnam-do, 31499, Korea. E-mail: hkim@hoseo.edu
[Received 5 August 2020; Revised 17 August 2020; Accepted 20 August 2020]