KIPS Trans. Softw. and Data Eng.
Vol.9, No.4 pp.129~136 pISSN: 2287-5905

https://doi.org/10.3745/KTSDE.2020.9.4.129

Hourly Prediction of Particulate Matter (PM2.5) Concentration Using Time Series Data and Random Forest

Deukwoo Lee† · Soowon Lee††

ABSTRACT

PM2.5 which is a very tiny air particulate matter even smaller than PM10 has been issued in the environmental problem. Since PM2.5 can cause eye diseases or respiratory problems and infiltrate even deep blood vessels in the brain, it is important to predict PM2.5. However, it is difficult to predict PM2.5 because there is no clear explanation yet regarding the creation and the movement of PM2.5. Thus, prediction methods which not only predict PM2.5 accurately but also have the interpretability of the result are needed. To predict hourly PM2.5 of Seoul city, we propose a method using random forest with the adjusted bootstrap number from the time series ground data preprocessed on different sources. With this method, the prediction model can be trained uniformly on hourly information and the result has the interpretability. To evaluate the prediction performance, we conducted comparative experiments. As a result, the performance of the proposed method was superior against other models in all labels. Also, the proposed method showed the importance of the variables regarding the creation of PM2.5 and the effect of China.

Keywords : Particulate Matter, PM2.5, Time Series Data, Machine Learning, Random Forest

시계열 데이터와 랜덤 포레스트를 활용한 시간당 초미세먼지 농도 예측

이 득 우† · 이 수 원††

요 약

최근 환경 문제에서 중요한 화두로 떠오른 초미세먼지(PM2.5)는 미세먼지(PM10)보다도 작은 부유물질이다. PM2.5는 안구나 호흡기 질환을 일으키며 뇌혈관에까지 침투할 수 있어서 시간별로 수치를 예측하여 대비하는 것이 중요하다. 그러나 PM2.5의 생성과 이동에 관한 명확한 설명이 아직까지는 제시되지 않고 있어서 예측에 어려움이 따른다. 따라서 PM2.5 예측뿐만 아니라 예측 결과에 대한 설명력을 갖는 예측 방법이 제시될 필요가 있다. 본 연구에서는 서울시의 시간당 PM2.5를 예측하고자 하며, 이를 위해 각기 다른 지상관측 데이터를 시계열로 전처리하고 부트스트랩 수를 조정한 랜덤 포레스트(Random Forest)를 데이터 학습 및 예측에 사용하는 방법을 제안한다. 이 방법은 예측 모델이 입력 데이터의 시각별 정보를 균형 있게 학습하게 하며 예측 결과에 대한 설명이 가능하다는 장점을 갖는다. 예측 정확도 평가를 위해 기존 모델과의 비교실험을 수행한 결과 제안 방법은 모든 레이블에서 가장 뛰어난 예측 성능을 보였으며, PM2.5의 생성과 관련된 변수와 중국의 영향과 관련된 변수가 예측 결과에 중요한 영향을 미치는 것을 보여주었다.

키워드 : 초미세먼지, PM2.5, 시계열 데이터, 기계학습, 랜덤 포레스트

1. 서 론

최근 환경 문제에서 중요한 화두로 떠오른 초미세먼지(PM2.5)는 안구나 호흡기 질환을 일으키며 산업시설에도 피해를 줄 수 있다[1]. PM2.5는 입자의 지름이 2.5㎛ 이하인 초미세먼지로, 입자의 지름이 10㎛ 이하인 PM10보다도 작아서 뇌혈관에까지 침투할 수 있다. 따라서 시간별로 이 수치를 예측하여 대비하는 것이 중요하다.

PM2.5를 포함한 미세먼지는 생성 원인에 따라 다시 1차, 2차 미세먼지로 나뉜다. 1차 미세먼지는 고체 상태의 먼지인데 흙먼지, 꽃가루, 산업 시설의 분진 등이 이에 해당한다. 2차 미세먼지는 가스 상태의 먼지인데 화석연료가 연소될 때 나오는 CO, NO2 같은 가스와 다른 가스 또는 수증기 사이의 화학반응으로 생긴다. 수도권 PM2.5는 2/3가 2차 미세먼

※ 이 논문은 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터 지원사업의 연구결과로 수행되었음(IITP-2020-2018-0-01419).
† 준 회 원 : 숭실대학교 융합소프트웨어학과 석사과정
†† 정 회 원 : 숭실대학교 소프트웨어학부 교수
Manuscript Received : August 23, 2019
First Revision : November 5, 2019
Accepted : December 19, 2019
* Corresponding Author : Soowon Lee(swlee@ssu.ac.kr)

지, 나머지가 1차 미세먼지이다[2]. 따라서 PM2.5 예측에 사용되는 변수는 1차 미세먼지와 2차 미세먼지의 특성을 모두 고려할 수 있어야 한다.

PM2.5는 대기가 정체되었을 때 많이 생성된다. 따라서 기상 상태가 어떠한지가 중요한데 이를 반영하기 위해 기온, 풍속, 강수량 등의 기상 정보가 변수로 필요하다. 또한 PM2.5는 화석 연료 연소로 생긴 가스 간의 화학반응으로 생긴다. 이를 반영하기 위해 CO, NO2, O3와 같은 대기오염물질 정보가 변수로 필요하다[1, 2]. 그밖에 PM2.5 생성의 재료가 되는 대기오염물질이나 미세먼지가 국외로부터 국내에 많이 유입될 경우 PM2.5가 많이 생성된다. 이를 반영하기 위해 국경 근처의 기상 정보, 대기오염물질 정보 및 미세먼지 수치가 변수로 필요하다. PM2.5 생성에는 1시간 이상의 시간이 소요되는 것으로 알려져 있기 때문에 입력 변수들을 시계열로 묶어서 대기의 추세를 반영할 필요도 있다. 하지만 아직까지 지상관측 기상 정보와 대기오염물질 정보를 시계열 데이터로 만들어 PM2.5 예측을 수행한 연구는 없다.

PM2.5 예측을 어렵게 만드는 이유 중 하나는 입력 변수 사이의 상호작용 때문이다. 일반적으로 대기오염물질 중 NO2와 O3는 비선형적인 상호작용을 하는 것으로 알려져 있으며, 계절의 영향을 많이 받는 한국의 기상 정보와 대기오염물질 사이의 관계도 여전히 많은 연구를 필요로 한다[3]. 이러한 이유로 PM2.5 예측의 경우 변수에 대한 설명력을 확보하는 것이 중요하다.

기존의 연구들은 위성의 레이더를 사용하여 부유물질을 간접 측정한 '에어로솔 광학두께(AOD)' 데이터를 사용하거나, 기상 모델로 추정하여 만든 데이터를 사용하여 실제 미세먼지를 추정하였다. 하지만 이 경우 실제 지상에서 관측기로 측정한 지상관측 데이터에 비해 데이터의 정확도가 떨어지며, 입력 변수에 대한 설명력이 부족하다는 한계가 있다. 또한 국내의 미세먼지 수치 예측 연구는 아직까지는 대부분 월별 또는 일별[4] 예측 위주이고, PM2.5보다는 PM10을 예측 대상으로 삼기 때문에 입력 변수에 대한 설명력을 갖는 시간당 PM2.5 예측 연구가 부족하다.

본 연구는 입력 변수에 대한 설명력을 갖는 서울시의 시간당 PM2.5 예측 연구를 위해, 각기 다른 지상관측 데이터를 시계열로 전처리하고, 부트스트랩 수를 조정한 랜덤 포레스트(Random Forest)를 데이터 학습 및 예측에 사용하는 방법을 제안한다. 또한 본 연구는 중국과 인접한 서울시의 지리적 특성을 반영하기 위해 시계열로 묶은 백령도의 기상 정보, 대기오염물질 정보 및 미세먼지 수치를 입력 변수로 사용한다.

2. 연구 배경지식

2.1 에어로솔 광학두께

'에어로솔'이란 먼지, 황사, 안개 등 대기부유물질을 의미하며 주로 위성의 레이더 등으로 대기부유물질을 관측할 때 이를 총칭하는 용어로 쓰인다. '에어로솔 광학두께(Aerosol Optical Depth; AOD)'란 '에어로솔'로 인해 태양복사가 지표 도달 전까지 얼마나 감쇄되는지를 위성의 레이더 센서 등으로 측정한 데이터이다. 해당 측정 값은 위성 사진 위에 색상으로 표시된다. 에어로솔 광학두께는 미세먼지 수치와는 다른 값이기 때문에, 에어로솔 광학두께 측정 데이터로부터 실제 미세먼지 수치를 추정해야 한다.

에어로솔 광학두께는 넓은 지역의 미세먼지 흐름을 추정하기 좋다는 장점이 있지만, 과대 측정될 수 있다는 단점도 존재한다. 왜냐하면 에어로솔 광학두께는 인간의 생활 고도가 아닌 상층부 대기에서까지 값이 측정되며, 습도가 높은 경우 에어로솔 입자가 크게 파악되기 때문이다[3]. 또한 태양복사를 측정할 수 없는 저녁 및 밤 시간 대에는 측정이 불가능하다. 따라서 실제 미세먼지 수치 추정을 하기 위해서, 에어로솔 광학두께 측정 값을 모델을 통해 추가로 보정하는 과정이 별도로 필요하다.

2.2 지상 관측

지상관측이란 지상에 설치한 측정기를 통해 미세먼지 등의 수치를 직접 측정하는 방법인데, 미세먼지의 경우 자동측정법과 수동측정법이 있고 이 중에서 자동측정법 중 하나인 '베타선(β-ray) 흡수 방식'[5]이 주로 사용된다. 베타선 흡수 방식이란 측정기 속 여과지에 수집한 먼지에 베타선을 통과시켜서 흡수되는 베타선 양을 측정해 미세먼지의 농도를 측정하는 방식이다. 지상관측은 에어로솔 광학두께로부터의 추정보다 더 정확하다는 장점이 있다. 지상관측은 에어로솔 광학두께에 비해 생활 고도에 가까운 대기 중 미세먼지를 직접 측정하기 때문이다. 국내에서는 미세먼지 이외에도 CO2나 O3 등의 대기오염물질이 각각 다른 방식으로 지상관측 되고 기온, 풍속, 강수량, 일조량 등의 기상 정보 또한 기상관측소의 센서를 통해 지상관측 된다.

2.3 관련 연구

특정 변수를 사용하여 미세먼지를 예측한 연구들이 있다. Choi et al.[6]은 국내에 유입된 황사를 입력 변수로 사용하여 황사와 강릉 지역 시간당 미세먼지 간의 상관관계를 보여주었다. 하지만 이 연구는 2차 미세먼지의 원인인 대기오염물질 정보는 반영하지 않았다. Seo et al.[7]은 바이오매스 연소 시 발생하는 유기화합물인 Levoglucosan과 미세먼지의 상관관계를 보여주었지만 이 연구 수행을 위해서는 Levoglucosan을 미세먼지에서 분류하는 과정이 필요하다. Huang et al.[8]은 에어로솔 광학두께 데이터와 지상관측 데이터로부터 중국 북부의 월별 PM2.5 농도를 랜덤 포레스트 모델을 사용해 예측하였다. 하지만 미세먼지는 시간 마다 변하기 때문에 시간 단위로 예측하는 것이 중요하다.

기존에는 미세먼지를 시간 단위로 예측하기 위해 대상 변수의 이전 관측값과 이동 평균을 고려하는 ARMA(Durbin, 1960) 또는 ARIMA(Box and Jenkins, 1976)와 같은 시계열 모델이 많이 사용되었다. 그러나 이 방법들은 미세먼지의

변화만을 계산하기 때문에 미세먼지의 급격한 비선형적 변화에 취약하며[9], 미세먼지 이외의 여러 입력 변수를 함께 학습한 딥러닝 또는 하이브리드 모델 등에 비해 좋지 못한 성능을 보였다[10-13]. 또한 ARIMA에 외부 변수의 영향을 추가한 ARIMAX(Box and Tiao, 1981) 역시 좋지 못한 성능을 보였다[14]. 최근 연구들에서는 시간 단위 예측이라고 하더라도 ARIMA나 ARIMAX보다는 딥러닝이나 하이브리드 방식의 모델이 제안되고 있다[15, 16].

그 외에 Kang et al. [17]은 부천시를 그리드로 나눈 뒤 Domain adversarial network를 적용하여 지상관측 데이터로부터 측정소가 없는 곳의 미세먼지를 예측했다. OH et al. [18]는 K-nearest neighbor를 사용하여 지상관측 데이터로부터 서울시 인접구 단위로 미세먼지를 예측했다. Cha et al. [19]는 지상관측 데이터로부터 인공신경망과 K-nearest neighbor를 사용하여 미세먼지를 예측했다. 그 외에 지상관측 데이터로부터 기계학습 모델을 적용하여 미세먼지를 예측한 연구들이 있지만 이 방법들은 모두 PM10만을 예측하였고, 미세먼지 예측에서 중요한 고농도 구간에서의 예측 성능 결과가 낮거나 성능 평가가 부재하다[20, 21].

3. 제안 방법

3.1 입력 변수

본 연구는 서울시 시간당 PM2.5 예측을 위해 시간당 지상관측 데이터를 사용하는데, 지상관측 데이터 중에서 서울시의 기상 정보 및 서울시 25개구의 대기오염물질 정보가 입력 변수로 사용된다. 또한 중국과 인접한 서울시의 지리적 특성을 반영하기 위해 백령도의 기상 정보, 대기오염물질 정보 및 미세먼지 수치가 입력 변수로 사용된다. 공장이 없고 인구가 적은 백령도의 지상관측 데이터는 중국으로부터 유입된 대기오염의 영향을 반영하기 좋기 때문이다. 본 연구는 백령도와 서울의 지상관측 데이터를 동시에 학습시켜서 서울 PM2.5예측을 위한 넓은 범위의 기상 흐름을 반영하고자 한다. 본 연구의 입력 변수는 총 34개이며 그 구성은 아래 Table 1과 같다.

3.2 데이터 시계열 전처리

본 연구에서는 서울시 시간당 PM2.5 예측을 위해 입력 변수 및 예측 대상을 타임스텝 n으로 묶는 시계열 전처리가 수행된다. 이를 통해 일정 시간단위의 입력 변수 모두를 예측모델에서 학습시킬 수 있다.

시각 t_i의 입력 변수 34개 전체를 $X(t_i)$라고 했을 때, t_1부터 t_n까지 n시간의 전처리된 데이터를 모델에서 학습하여 t_{n+1}시각의 서울시 PM2.5를 예측하는 수식은 다음과 같다.

$$Predicted\_PM\ 2.5(t_{(n+1)}) = Model(X(t_1), X(t_2), \cdots, X(t_n)) \quad (1)$$

Table 1. Description of Input Variables

	Variable Type	Description
Seoul City (Numeric)	Weather	Wind Speed, Wind Degree, Rain, Vapor, Sunlight, etc (12 Variables)
	Air Pollutant	CO2, NO2, O3, SO2 (4 Variables)
Baekryung Island (Numeric)	Weather	Wind Speed, Wind Degree, Rain, Vapor, Sunlight, etc (12 Variables)
	Air Pollutant	CO2, NO2, O3, SO2 (4 Variables)
	Particulate Matter	PM10, PM2.5 (2 Variables)
Target (Category)	Particulate Matter	PM2.5 of Seoul City

예측 대상인 서울시 PM2.5를 제외한 입력 값은 0에서 1 사이의 값으로 정규화되었다. 일부 기존 연구들에서는 변수에 결측값이 있을 경우 해당 변수의 평균값 등으로 결측값을 보정하였으나, 본 연구에서는 타임스텝 전처리 시 값이 누락된 변수가 하나라도 있는 경우 데이터 집합에서 제외시켰다.

3.3 학습 모델

본 연구에서는 랜덤 포레스트 분류기[22]를 학습 모델로 사용한다. 최근 미세먼지 예측 연구에서 사용되는 LSTM의 경우 입력 데이터의 시퀀스 또는 시계열이 길어질수록 앞의 입력 정보가 소실되고[23] 결과 설명력이 부족하다는 문제점이 있다. 또한 하이브리드 모델의 경우에는 딥러닝 모델처럼 설명력이 떨어질 수 있다. 반면 랜덤 포레스트는 모든 입력 변수를 균형 있게 학습할 수 있기 때문에 시계열 전처리를 수행한 데이터가 입력될 경우 모든 입력 변수의 시각별 정보를 소실하지 않고 학습할 수 있다. 또한 결과에 대한 설명력이 앞의 모델들에 비해 상대적으로 좋다는 장점이 있다. 이러한 랜덤 포레스트의 특성은 본 연구의 입력데이터(3.2 참조)와 함께 시너지 효과를 가져올 수 있다.

랜덤 포레스트는 분류 또는 회귀 모두에 사용될 수 있는 방법으로, 편향이 적은 여러 의사결정트리(Decision Tree)의 예측결과 중 다수의 결과를 선택(Majority Vote)하여 분산과 편향이 적은 단일한 학습 결과를 얻는 앙상블 트리 기법이다. 이때 편향이 적은 각각의 의사결정트리를 만들기 위해 학습 데이터를 전체 데이터 크기만큼 복원 추출하여 여러 개의 학습 데이터를 샘플링하는 방법인 부트스트랩(Bootstrap)이 함께 사용된다. 또한 특정 변수가 집중 학습되는 것을 막기 위해, 의사결정트리의 노드를 분할할 때 총 q 개의 입력 변수 중 매번 새로 선택된 m개의 입력 변수가 노드 분할 시 고려 대상이 된다. 이때 변수의 개수 m은 일반적으로 $m \approx \sqrt{q}$ 또는 $m \approx \log_2 q$로 결정된다. 랜덤 포레스트 학습에 사용되는 전체 부트 스트랩 샘플링 수(의사결정트리 수)가 B,

각 부트 스트랩 샘플 집합을 b_j라고 할 때, 각 단일 트리의 예측 결과 $Pred\ (b_j)$를 활용한 랜덤 포레스트 분류기의 최종 예측 결과는 다음과 같이 결정된다.

$$\begin{aligned} &Predicted\_PM\ 2.5(t_{(n+1)}) \\ &= Majority\_Vote\ \left[Pred\ (b_1), Pred\ (b_2), \cdots, Pred\ (b_B)\right] \end{aligned} \tag{2}$$

랜덤 포레스트 분류기는 특정 변수가 각 단일 의사결정트리에서 데이터를 분할할 때 얻을 수 있는 정보량의 평균으로 해당 변수의 중요도를 계산한다. 부트스트랩 샘플 b_j의 데이터들이 시각 t_i의 특정 변수 x_q기준에서 레이블 k일 확률을 $\hat{p}(b_j, k, x_q(t_i))$라고 하고, 정보량 기준을 Entropy로 잡을 때, 변수 $x_q(t_i)$의 샘플 b_j에서의 정보량인 I_{b_j}의 기대값은 다음과 같다.

$$\begin{aligned} &I_{b_j}(x_q(t_i)) = \\ &-\sum_{k=1}^{K} \hat{p}(b_j, k, x_q(t_i)) \times \log\ \hat{p}(b_j, k, x_q(t_i)) \end{aligned} \tag{3}$$

이때 총 B 개의 의사결정트리를 사용한 랜덤 포레스트 분류기에서 변수 $x_q(t_i)$의 중요도는 다음과 같다.

$$I_B(x_q(t_i)) = \frac{1}{B}\sum_{b=1}^{B} I_{b_j}(x_q(t_i)) \tag{4}$$

본 연구에서 전처리 한 데이터에는 시각 마다 입력 변수가 존재하기 때문에, 전체 시각에서 특정 변수의 중요도를 측정하기 위해 다음과 같이 특정 변수의 시각별 중요도를 합산하였다.

$$Total\_I_B(x_q) = \sum_{i=1}^{n} I_B(x_q(t_i)) \tag{5}$$

Equation (5)의 결과로 얻은 변수별 중요도는 전체 변수의 백분위 상대점수로 환산된다. 본 연구의 제안 모델 구조도는 Fig. 1과 같다.

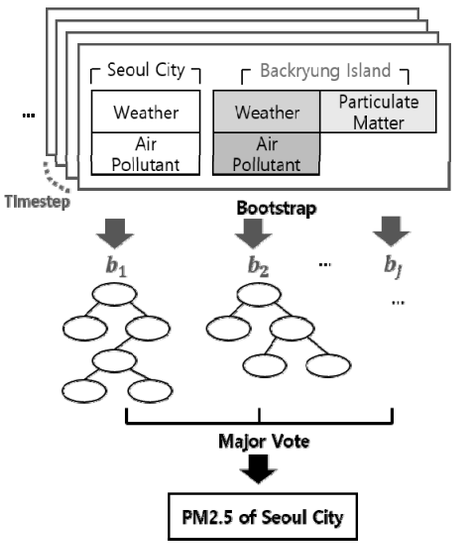

Fig. 1. Structure of the Proposed Method

4. 실 험

4.1 데이터

본 연구에서는 한국 기상자료 개방포털(https://data.kma.go.kr)에서 제공하는 2015.1.1-2018.12.31 기간의 시간당 기상정보 데이터와, 에어코리아(https://www.airkorea.or.kr)에서 제공하는 같은 기간의 시간당 대기오염물질 및 미세먼지 데이터를 학습, 검증 및 테스트에 사용한다. 전체 데이터 수는 타임스텝에 따른 전처리마다 조금씩 달라지며 학습, 검증, 테스트 데이터는 6:2:2의 비율로 Random 분할한다.

4.2 예측 레이블

본 연구에서는 현재 한국, 미국 일본 등이 사용하는 PM2.5 분류기준을 참고하여 예측 레이블을 분류하였다. 해당 분류기준에 따르면 PM2.5의 수치(시간당 평균 ㎍/㎥)가 0~15인 경우 '좋음', 16~35인 경우 '보통', 36~75인 경우 '나쁨', 76 이상인 경우 '매우나쁨'으로 분류된다. 하지만 '매우나쁨' 레이블의 개수가 전체에서 차지하는 비율이 적고, '나쁨' 이상인 경우 위험성을 예측하기에 충분하기 때문에, 본 연구에서는 '나쁨'과 '매우나쁨'을 '나쁨'으로 묶어 학습하였다. 예를 들어 타임스텝을 2로 묶었을 때 전처리 후 레이블별 데이터 수는 Table 2와 같다.

Table 2. Number of Data for Each Target Label (Timestep=2)

	Good	Normal	Bad	Total
Train	127,374	182,620	71,316	381,310
Validation	42,671	60,540	23,898	127,104
Test	42,265	61,131	23,708	127,104

4.3 학습 모델 파라미터 설정

본 연구에서는 랜덤 포레스트에서 사용할 부트스트랩 샘플링 횟수(의사결정트리 수) B를 정하기 위해 해당 파라미터를 변화시키며 검증 데이터 상에서 모델 성능의 변화를 관찰하였다. 랜덤 포레스트 구현에는 Ubuntu 16.04, Python 3.5, sklearn 라이브러리가 사용되었다. 타임스텝을 2시간으로 묶어서 학습한 뒤 1시간 후를 예측 할 때 결과는 Table 3 및 Fig. 2와 같다.

Table 3. F1-Scores by Tree Number on Validation Data

	PM2.5 Label		
B	Good	Normal	Bad
1	0.741	0.710	0.676
10	0.817	0.788	0.756
30	0.824	0.801	0.773
100	0.828	0.806	0.780
200	0.828	0.807	0.781
300	0.829	0.808	0.782

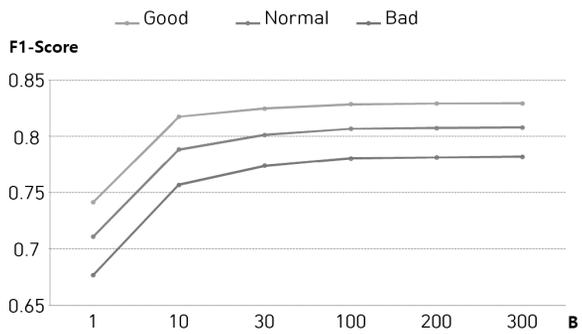

Fig. 2. F1-Scores by Tree Number on Validation Data

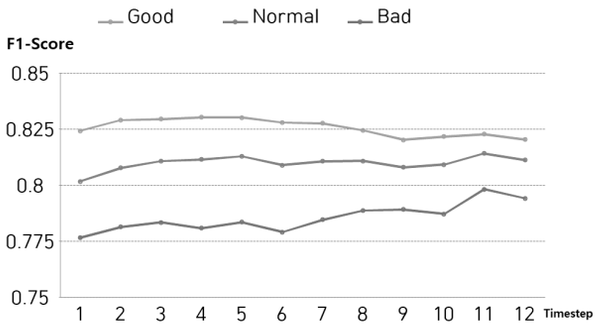

Fig. 3. F1-Scores by Timestep on Validation Data

Table 3 및 Fig. 2에서 확인할 수 있듯이, B가 200을 넘어서자 예측 정확도의 상승폭이 매우 작아진 것을 알 수 있다. B가 커질수록 학습에 필요한 시간이 길어지기 때문에 본 연구에서는 B의 값을 200으로 설정하였다.

트리 기법의 경우 가지치기를 통해 모델이 학습데이터에 과적합 되는 것을 막을 수 있는데, 이를 위해 가지치기를 실험한 결과 본 연구의 학습 모델에 가지치기를 하지 않고 트리의 노드를 학습 데이터 상에서 끝까지 학습시킬 때 오히려 검증 데이터 상에서의 성능이 좋게 나왔다. 이것은 비슷한 패턴을 반복 및 순환하는 기상 정보의 특성이 학습 및 검증 데이터에서 드러난 것으로 해석 가능하다.

이 외에 모델의 노드 분할 기준인 정보량 지수는 Gini가 아닌 Entropy로, 각 노드 분할 시 Random으로 고려할 입력 변수 개수 m은 가 아닌 근사값으로 정했을 때 미세하게 더 높은 성능을 보였다. 검증데이터를 통해 결정된 학습 모델의 하이퍼파라미터는 Table 4와 같다.

Table 4. Hyperparameters for the Proposed Random Forest

Name	Value
n_estimators (Number of Tree (Bootstrap))	200
Criterion	'Entropy'
max_depth	None
max_features	'log2'
min_impurity_decrease	0
bootstrap	True

4.4 타임스텝별 다음 시간 예측 결과

본 연구는 모델의 시계열 학습 성능 비교를 위해 입력 데이터의 타임스텝을 1시간에서 12시간까지 조정하며 1시간 후 PM2.5 농도를 예측하는 성능을 비교했다. 그 결과는 Fig. 3과 같다.

Fig. 3에서 확인할 수 있듯이, '좋음' 레이블은 타임스텝이 4에서 5 사이일 때까지 예측 정확도가 상승하다가 이후 하락한다. '보통'과 '나쁨' 레이블은 등락이 있지만 타임스텝이 11일 때 예측 정확도가 가장 높다. 공통적으로 모든 레이블에서 타임스텝이 작은 것보다는 어느 정도의 값을 가질 때 예측 정확도가 상승하였는데, 이것은 생성에 시간이 걸리는 PM2.5의 특성이 반영된 것으로 보인다. 또한 '나쁨' 레이블의 정확도가 타임스텝 11일 때 가장 높은 것으로 보아, 고농도 PM2.5를 예측하려면 긴 시간의 경과를 학습해야 한다는 것과 '나쁨' 레이블의 결정 경계가 가장 복잡하다는 것을 알 수 있다.

4.5 변수 중요도

본 연구에서는 입력 변수에 대한 설명력을 얻기 위해, 특정 변수 기준으로 데이터를 분할할 때 얻을 수 있는 정보량에 기반한 변수별 중요도를 백분위 상대점수로 만들어 상위 10개를 추출하였다(3.3. Equation (3)(4)(5) 참조). 그 결과는 Table 5와 같다.

Table 5. Top10 Input Variables by Importance

	Timestep=1		Timestep=12	
Rank	Variable	Importance	Variable	Importance
1	NO2	11.11	NO2	10.68
2	CO	10.39	CO	8.09
3	O3	8.42	O3	7.77
4	bPM2.5	5.68	bPM2.5	6.86
5	SO2	5.04	SO2	5.35
6	bPM10	3.39	bPM10	4.14
7	dew	3.38	bNO2	3.70
8	bTemp	3.09	bO3	3.20
9	vapor	2.93	humid	3.18
10	humid	2.90	dew	3.11

Table 5에서 확인할 수 있듯이, PM2.5의 생성과 관련된 NO2, CO, O3, SO2가 높은 순위에 있으며, 백령도의 미세먼지(bPM2.5, bPM10) 역시 상위권에 있다. 이 경향은 타임스텝에 상관없이 드러난다. 즉, PM2.5 예측에서는 PM2.5의 생성과 관련된 변수들이 주요하며, 백령도의 미세먼지와 관련된 변수가 중요한 것으로 보아 중국의 영향을 고려하는 것 역시 중요하다고 볼 수 있다. 하지만 34개의 변수 중 미세먼지의 이동과 관련된 기상 정보 관련 변수는 상위 10위 안에는 들지 못하였다.

4.6 비교 실험

본 연구에서는 제안 방법과 기존 모델들의 비교실험을 진행했다. 비교 실험에는 일반적으로 분류 예측에 많이 사용되는 Logistic Regression, 도메인과 상관없이 좋은 성능을 보여주는 순환신경망 중 하나인 LSTM(Long Short-Term Memory)이 사용되었다. 구현에는 각각 sklearn, tensorflow1.14-gpu가 사용되었으며 그래픽카드 사양은 GeForce GTX 1060 6GB이다. 두 모델 모두 검증 데이터에서 Grid Search 방식을 통해 '나쁨' 레이블 예측 성능이 가장 좋았던 하이퍼파라미터를 결정하였다.

Logistic Regression의 경우 대용량 데이터 학습 시 효율적인 것으로 알려진 SAGA[24]를 최적화 방법으로 사용하였으며 허용오차(tol) 및 정규화 값(C)과 penalty 파라미터로 예측 성능을 검증하였다. Logistic Regression의 최적 하이퍼파라미터는 Table 6과 같으며 검증 성능은 Table 7과 같다.

Table 6. Hyperparameters for Logistic Regression

Name	Hyperparameters	Optimal
penalty	['l1', 'l2', 'elasticnet']	**'l1'**
tol	[0.00001, 0.0001, 0.001, 0.01, 0.1]	**0.1**
C	[5.0, 4.0, 3.0, 2.0, 1.0, 0.1]	**1.0**

Table 7. F1-Score of Hyperparameters for Logistic Regression

penalty	tol	C	F1-Score for 'Bad' Label
L1	**0.1**	**1.0**	**0.5864**
L1	0.1	2.0	0.5863
elasticnet	0.1	5.0	0.5859
...	...	...	...
L2	0.1	1.0	0.5811
...	...	..	..
L1	0.00001	2.0	0.5459
L1	0.0001	2.0	0.5459
...	...	...	...

LSTM의 경우 Optimizer는 최근 많은 연구에서 사용되는 Adam[25]을 사용하였으며 권장 학습률인 0.001과 다른 학습률로 예측 성능을 검증하였다. 학습률 이외에는 배치 크기(batch_size) 및 셀의 출력 노드 수(num_units)와 셀 개수(number of cell)로 예측 성능을 검증하였다. Epoch 수는 최대 200으로 설정하되 검증 데이터에서 모델의 손실 값이 증가할 때 조기 종료(Early Stopping)하여 자동으로 결정하였다. LSTM의 최적 하이퍼파라미터는 Table 8과 같으며 검증 성능은 Table 9과 같다.

LSTM 하이퍼파라미터 검증 과정에서 학습률이 0.01~0.0001 사이일 경우에는 epoch 200 내에서 손실 값이 증가하며 학습이 자동으로 종료되었다. 그러나 학습률이 0.00001일 경우에는 epoch 200 내에서 학습이 종료되지 못하였다. 이후 학습률을 0.00001로 두고 추가로 몇 쌍의 하이퍼파라미터를 랜덤 선택하여 epoch 제한이 없이 학습시켜 보았는데 하이퍼파라미터 조합당 epoch 1000 이상 및 11시간 이상이 소요되었지만 더 높은 성능이 나타나진 못했다.

Table 8. Hyperparameters for LSTM

Name	Hyperparameters	Optimal
batch_size	[512, 256, 128]	**256**
num_units (cell output)	[512, 256, 128]	**512**
number of cell	[3, 2, 1]	**2**
learning_rate	[0.01, 0.001, 0.0001, 0.00001]	**0.0001**
epochs	-	Early Stop

Table 9. F1-Score of Hyperparameters for LSTM

number of cell	batch size	num units	learning rate	F1-Score for 'Bad' Label
2	**256**	**512**	**0.0001**	**0.7771**
3	256	256	0.0001	0.7754
1	256	512	0.0001	0.7696
...	...	...	...	...
1	128	256	0.001	0.7641
3	512	256	0.001	0.7640
...	...	...	...	...
1	512	256	0.01	0.7354
...	...	...	...	...

타임스텝이 2일 때 본 연구의 제안 방법과 최적 하이퍼파라미터의 Logistic Regression 및 LSTM의 테스트 데이터에서의 예측 성능 비교실험 결과는 Table 10과 같다.

Table 10. F1-Scores of the Proposed Method and Other Models (Timestep=2)

Model	PM2.5 Label		
	Good	Normal	Bad
Logistic Regression	0.684	0.700	0.588
LSTM	0.822	0.792	0.775
Proposed Method	**0.828**	**0.808**	**0.779**

Table 10에서 확인할 수 있듯이, 제안 방법의 예측 성능이 모든 레이블에서 가장 뛰어나다. 특히 제안 방법은 시계열 학습에서 강점을 갖는 LSTM보다도 높은 성능을 보여주었다. 나아가 제안 방법은 딥러닝 모델에서는 알기 어려운 변수의 중요도 또한 계산할 수 있었다(4.5 참조).

다음으로 PM2.5 예측에서 중요한 '나쁨' 레이블을 대상으로 타임스텝을 변화시키며 비교 실험을 진행하였다. 결과는 Fig. 4와 같다.

Fig. 4에서 확인할 수 있듯이, 제안 방법의 예측 성능이 모든 타임스텝에서 가장 높은 성능과 고른 추세를 보였다.

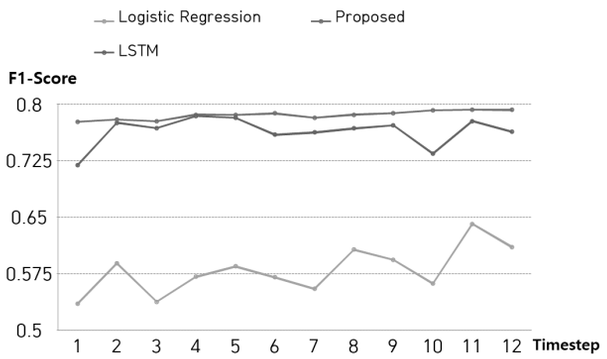

Fig. 4. F1-Scores for 'Bad' Label of Proposed Method and Other Models by Timestep

LSTM의 경우 타임스텝에 따라 등락을 보였으며, 타임스텝 2, 4, 5 구간에서는 제안방법에 근접한 성능을 보였다. Logistic Regression의 경우 타임스텝이 증가할수록 예측 정확도가 향상되는 구간이 존재하지만 등락폭이 컸으며 전체적인 성능은 가장 낮은 것으로 평가되었다. 결과적으로 제안 방법은 모든 레이블 및 타임스텝에서 가장 뛰어난 성능을 보였다.

5. 결 론

본 연구에서는 각기 다른 지상 관측 데이터를 시계열로 전처리 한 뒤 부트 스트랩 수를 조정한 랜덤 포레스트를 활용하여 서울시의 시간당 PM2.5 농도를 예측하는 방법을 제안했다. 이 방법은 시계열 처리된 입력 변수의 모든 시각별 정보를 균형 있게 학습할 수 있었으며, 비교 실험 결과 본 연구의 제안 방법이 기존 모델 보다 모든 레이블에서 더 뛰어난 성능을 보였다. 또한 본 연구에서는 입력 변수에 대한 설명력을 높이기 위해 각 변수의 타임스텝별 중요도를 합산하여 변수의 중요도를 계산하였고 그 결과 PM2.5의 생성과 관련된 변수와 중국의 영향과 관련된 변수들이 중요했음을 확인하였다.

향후에는 본 연구를 확장하여 연구의 범위를 서울시뿐만 아니라 다른 국내 도시로 확장하고 국가 전체로 확대할 수 있을 것이다. 또한 제안 방법을 심층신경망과 결합하여 추가적인 성능 향상을 가져올 수 있을 것으로도 기대된다.

References

[1] H. J. Lee, Y. Jeong, S. T. Kim, and W. S. Lee, "Atmospheric Circulation Patterns Associated with Particulate Matter over South Korea and Their Future Projection," *Journal of Climate Change Research*, Vol.9, No.4, pp.423-433, 2018.

[2] Ministry of Environment, "What is Fine Dust?," Republic of Korea's Ministry of Environment, Apr. 2016.

[3] National Institute of Environmental Research & NASA, "KORUS-AQ: An International Cooperative Air Quality Field Study in Korea," KORUS-AQ, 2016.

[4] D. Lee and S. Lee, "Prediction of fine Dust(PM2.5) Concentration Based on RBF Kernel SVM," *Proceedings of the ISSAT international Conference on Data Science in Business, Finance and Industry*, pp.114-117, Jul. 2019.

[5] S. Choi, J. An, and Y. Jo, "Review of Analysis Principle of Fine Dust," *Prospectives of Industrial Chemistry*, Vol.21, No.2, pp.16-23, Apr. 2018.

[6] H. Choi and M. S. Lee, "Atmospheric Boundary Layer Influenced upon Hourly PM10, PM2.5, PM1 Concentrations and Their Correlations at Gangneung City before and after Yellow Dust Transportation from Gobi Desert," *Atmospheric Research*, Vol.7, No.1, pp.30-54, Feb. 2012.

[7] Y. H. Seo and J. Kweon, "Relation of Levoglucosan and the Outbreaks of High PM10 and PM2.5 Concentration Occurred in Seoul Air," *J. Korea Society of Environmental Administration*, Vol.19. No.1, pp.1-10, Mar. 2013.

[8] K. Huang, Q. Xiao, X. Meng, G. Geng, Y. Wang, A. Lyapustin, D. Gu, and Y. Liu, "Predicting monthly high-resolution PM2.5 concentrations with random forest model in the North China Plain," *Environmental Pollution*, Vol. 242, No.A, pp.675-683, 2018.

[9] Y. Lin, N. Mago, Y. Gao, Y. Li, Y. Y. Chiang, C. Shahabi, and J. L. Ambite, "Exploiting Spatiotemporal Patterns for Accurate Air Quality Forecasting using Deep Learning," *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL)*, pp.359-368, Nov. 2018.

[10] S. Jiaming. "PM 2 . 5 Concentration Prediction using Times Series Based Data Mining." 2015.

[11] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, "Forecasting Fine-Grained Air Quality Based on Big Data," *Proceedings of the 21th SIGKDD Conference on Knowledge Discovery and Data Mining*, pp.2267-2276, Aug. 2015.

[12] J. E. Choi, H. Lee, and J. Song, "Forecasting Daily PM10 Concentrations in Seoul using Various Data Mining Techniques," *Communications for Statistical Applications and Methods 2018*, Vol.25, No.2, 199-215, Mar. 2018.

[13] Y. Liang, S. Ke, J. Zhang, X. Yi, and Y. Zheng, "GeoMAN Multi-level Attention Networks for Geo-sensory Time Series Prediction," *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp.3428-3434, Jul. 2018.

[14] L. A. Diaz-Robles, J. C. Ortega, J. S. Fu, G. D. Reed, J. C. Chow, J. G. Watson, and J. A. Moncada-Herrera, "A Hybrid ARIMA and Artificial Neural Networks Model to Forecast Particulate Matter in Urban Areas: The Case of

Temuco, Chile," *Atmospheric Environment*, Vol.42, No.35, pp.8331-8340, Nov. 2008.

[15] J. Zhao, F. Deng, Y. Cai, and J. Chen, "Long Short-term Memory - Fully Connected (LSTM-FC) Neural Network for PM2.5 Concentration Prediction," *Chemosphere*, Vol.220, No.1, pp.486-492, 2019.

[16] Y. Cheng, H. Zhang, Z. Liu, L. Chen, and P. Wang, "Hybrid Algorithm for Short-Term Forecasting of PM2.5 in China," *Atmospheric Environment*, Vol.200, pp.264-279, 2019.

[17] T. C. Kang and H. B. Kang, "Machine Learning-based Estimation of the Concentration of Fine Particulate Matter Using Domain Adaptation Method," *Journal of Korea Multimedia Society*, Vol.20, No.8, pp.1208-1215, August. 2017.

[18] S. OH, J. Koo, and U. M. Kim, "Concentration Prediction Technique Based on Locality of Fine Dust Generation," *The Institute of Electronics Engineers of Korea 2017*, pp. 1357-1360, Jun. 2017.

[19] J. Cha and J. kim, "Development of Data Mining Algorithm for Implementation of Fine Dust Numerical Prediction Model," *Journal of the Korea Institute of Information and Communication Engineering*, Vol.22, No.4, pp.595-601, Apr. 2018.

[20] J. H. Kwon, Y. Lim, and H. S. Oh, "Particulate Matter Prediction using Quantile Boosting," *The Korean Journal of Applied Statistics*, Vol.28, No.1, pp.83-92, 2015.

[21] S. Joun, J. Choi, and J. Bae, "Performance Comparison of Algorithms for the Prediction of Fine Dust Concentration," *Korea Software Congress 2017*, pp.775-777, Dec. 2017.

[22] G. James, D. Witten, T. Hastie, and R. Tibshirani, "An Introduction to. Statistical Learning with Applications in R," Springer, 2017.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, pp.5998-6008, Dec. 2017.

[24] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives," *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, Jul. 2014.

[25] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, May. 2015.

이 득 우

https://orcid.org/0000-0001-7956-5570
e-mail : dugu@soongsil.ac.kr
2017년 ~ 현 재 숭실대학교 융합소프트웨어학과 석사과정
관심분야 : 기계학습, 데이터사이언스, 인공지능

이 수 원

https://orcid.org/0000-0001-5863-1188
e-mail : swlee@ssu.ac.kr
1982년 서울대학교 계산통계학과(학사)
1984년 한국과학기술원 전산학과(석사)
1994년 미국 University of Southern California 전산학과(박사)
1995년 ~ 현 재 숭실대학교 소프트웨어학부 교수
2003년 ~ 2004년 한국정보과학회 인공지능연구회 운영위원장
2008년 한국정보과학회 소프트웨어 및 응용 논문지 편집위원장
2008년 ~ 2012년 한국BI데이터마이닝학회 부회장
관심분야 : 데이터사이언스, 인공지능, 스포츠IT융합