한국빅데이터학회지
제6권 제2호, 2021, pp. 39-49 https://doi.org/10.36498/kbigdt.2021.6.2.39

실외공기측정기 자료를 이용한 도심 기상 예측 기계학습 모형 비교*

Comparison of Machine Learning Techniques in Urban Weather Prediction using Air Quality Sensor Data

박종찬 · 박헌진†

인하대학교 통계학과

요 약

최근 국가 관측망, 기업 공기 측정기 등을 통해 많고 다양한 기상 데이터가 수집되고 있다. 기계학습 기법을 통해 기상 예측하려는 노력이 곳곳에서 이루어지고 있으며, 국내 미세먼지는 농도가 증가해오고 사람들의 관심이 높아 가장 관심있는 예측 대상 중 하나이다. 본 연구에서는 서울시 전역에 설치된 840여 개 실외공기측정기 데이터를 사용하여 PM10 · PM2.5 예측 모형을 비교하고자 한다. 5분 뒤 미세먼지 농도 예측을 통해 실시간으로 정보를 제공할 수 있으며, 이는 10분 · 30분 · 1시간 뒤 예측 모형 개발에 기반이 될 수 있다. 잡음 제거, 결측치 대체 등의 데이터 전처리를 진행하였고, 시 · 공간 변수를 고려할 수 있는 파생 변수를 생성하였다. 모형의 매개변수는 반응 표면 방법을 통해 선택하였다. XGBoost, 랜덤포레스트, 딥러닝(Multilayer Perceptron)을 예측 모형으로 사용하여, 미세먼지 농도와 예측값의 차이를 확인하고, 모형 간 성능을 비교하고자 한다.

■ 중심어 : 미세먼지 농도, 기계학습, 시공간 모형, 기상 자료

Abstract

Recently, large and diverse weather data are being collected by sensors from various sources. Efforts to predict the concentration of fine dust through machine learning are being made everywhere, and this study intends to compare PM10 and PM2.5 prediction models using data from 840 outdoor air meters installed throughout the city. Information can be provided in real time by predicting the concentration of fine dust after 5 minutes, and can be the basis for model development after 10 minutes, 30 minutes, and 1 hour. Data preprocessing was performed, such as noise removal and missing value replacement, and a derived variable that considers temporal and spatial variables was created. The parameters of the model were selected through the response surface method. XGBoost, Random Forest, and Deep Learning (Multilayer Perceptron) are used as predictive models to check the difference between fine dust concentration and predicted values, and to compare the performance between models.

■ Keyword : PM10 · PM2.5, machine learning, spatio-temporal model, weather data

2021년 11월 26일 접수; 2021년 12월 11일 수정본 접수; 2021년 12월 17일 게재 확정.

* 본 연구는 국토교통부의 '빅데이터 기반 항공안전관리 기술 개발 및 플랫폼 구축(21BDAS-B158275-02)' 연구의 지원에 의하여 이루어진 연구로서, 관계부처에 감사드립니다. This work is supported by the Korea Agency for Infrastructure Technology Advancement(KAIA) grand funded by the Ministry of Land, Infrastrucrue and Transport (Grant 21BDAS-B158275-02).

† 교신저자 (hjpark@inha.ac.kr)

Ⅰ. 서 론

기상 현상은 인간의 삶에 많은 영향을 끼친다. 강수・바람・구름 등 대기 중에서 일어나는 물리적인 현상들은 인간뿐만 아니라 동・식물까지 큰 관여를 한다. 최근에는 인간이 인위적으로 만들어낸 미세먼지가 큰 관심 속에 놓여 있다.

농경사회부터 현재에 이르기까지 기상 예측에 대한 노력은 끊임없이 이어져 왔으며, 오래 전부터 기상 예측의 시작은 데이터의 수집으로부터 시작되었다. 눈으로 관측할 수 있는 태양의 위치나 구름의 양, 기압 차로 인해 발생하는 신체의 변화 등을 통해 기상 예측을 수행할 수 있었으며, 현대 사회에서는 기상을 측정할 수 있는 기계, 센서 등을 만들어 데이터를 자동으로 수집하고 인과관계를 이용하여 기상 예측 및 예보를 시행하였다.

최근 몇 년간 CPU, GPU 등의 하드웨어 발전으로 컴퓨터 계산 능력이 급격하게 향상되고, 대량의 데이터셋을 수집하고 배포할 수 있는 인터넷 시장의 성장으로 빅데이터 시대가 도래했다. 현재는 금융・제조・유통뿐만 아니라 기상・게임・공공 등 대부분 분야에서 빅데이터를 적용하고 활용하고자 하는 노력이 활발하다.

기상 예측 문제는 다양한 변수가 작용하기 때문에 먼 미래를 예측할수록 예측력이 떨어진다. 하지만, 국가 관측망이나 기업에서 설치하고 있는 공기측정기를 통해 대량의 다양한 데이터가 쌓이고 있어, 기계학습을 통한 기상 예측력 향상을 위한 노력이 이어지고 있다. 특히 국내 수도권은 미세먼지 농도가 높아지고 관심도가 높아짐에 따라, 미세먼지 농도 예측에 큰 관심이 쏠려지고 있다.

따라서, 본 연구에서는 서울시에 위치한 840여 개의 실외공기측정기로부터 수집된 기상 데이터를 이용하여, 5분 뒤 미세먼지 농도 예측 모형을 생성하고자 한다. 예측 모형에는 대표적으로 성능이 좋은 기계학습 모형인 XGBoost, 랜덤포레스트, Deep Learning(Multilayer Perceptron)을 사용한다. 5분 뒤 예측 모형을 구축함으로써, 나아가 10분 뒤, 30분 뒤, 1시간 뒤 예측 모형 생성에 기반을 마련할 수 있고, 골고루 분포된 실외공기측정기로부터 생성됐기 때문에 특정 지역에서의 미세먼지 농도를 확인할 수 있다. 많은 변수의 영향을 받고 급격하게 변화하는 미세먼지 농도를 예측함으로써, 사람들의 생활 방식을 개선할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 시공간 데이터 및 기계학습 모형에 관해 설명하고, 3장에서는 데이터 전처리부터 매개변수 선택, 모형 적합 과정을 소개한다. 4장은 연구에 활용된 데이터 셋으로 예측 모형 적합 결과와 성능 비교 결과를 보여준다. 5장은 분석 결과를 종합하고 결론으로 마무리 짓는다.

Ⅱ. 관련 연구

본 장에서는 관련 연구를 살펴보고자 한다. 시공간 데이터 분석을 소개하고, 연구에 사용될 기계학습 모형인 XGBoost, 랜덤포레스트, 딥러닝{Multilayer Perceptron}을 설명하고자 한다.

2.1 시공간 데이터 분석

소셜 미디어, 보건 의료, 농업, 교통, 기후 과학 분야 등에서 시공간 데이터는 많은 양이 수집되고 응용되고 있다. 시공간 데이터는 일반적으로 두 가지의 특징을 갖는다. 관측치가 주변 위치와 시간에 영향을 받아 독립이 아니고 상관관계가 있는 자기상관성과 다양한 방법과 수준에서 나타나는 이질성(heterogeneity)이 두 특징이다. 기존 데이터 마이닝 알고리즘에 적용하기 위해선 위 두 가지 성질을 고려해야 한다[1].

시공간 데이터는 다양한 유형으로 분류된다. 일반적으로 4가지 범주로 설명되며, 어떠한 지점이나 시간에서 이산적으로 발생하는 사건 데이터(예, 도시 내 범죄 사건 발생)와 이동 물체의 궤적을 추적하는 궤적 데이터(예, 경찰의 감시 차량 순찰 경로), 움직이는 참조물로부터 연속적인 시공간 필드가 측정되는 포인트 참조(point referenc) 데이터(예, 풍선을 이용한 표면 온도 측정), 고정된 셀로부터 수집되는 래스터(raster) 데이터(예, 뇌 fMRI 데이터)가 있다[1].

시공간 데이터 또한 클러스터링, 이상 탐지, 예측, 패턴 마이닝, 변화 탐지 등 다양한 방법 및 문제들이 적용된다. 본 논문에서는 고정된 센서로부터 수집되는 기상 데이터를 사용하고, 시공간 변수를 고려한 파생 변수를 생성하여 세 가지 기계학습 모형을 적용하고자 한다. 적합 된 모형을 기반으로 기상 데이터에서 모형 간 차이를 확인하고 추후 기상 데이터 병합 시 모형 선택에 기반을 제시한다.

2.2 XGBoost

XGBoost 모형은 트리 기반의 앙상블 모형이다. 의사결정나무를 차례대로 학습해 가면서 각 트리는 앞서 생성된 트리의 잔차를 학습하여 개선해 나간다. Gradient tree boosting 방법에서 트리 분할 지점을 찾는 알고리즘과 시스템 디자인 부분을 변경한 모형으로 Chen and Guestrin(2016)에 의해 제안되었다[5]. XGBoost 이전의 몇몇 트리 모형들은 변수별 모든 속성값을 고려하여 분할 지점을 탐색하고 최적값을 찾아내었다. 모든 경우의 수를 고려하기 때문에 정확도는 높지만 계산 비용이 크고 데이터 크기가 클 때 메모리 할당 문제가 있었다. XGBoost는 근사 알고리즘으로 분할 지점의 최적해를 찾아내고, 병렬 계산 처리를 통해 계산 비용을 줄였다. 근사 알고리즘은 데이터를 분위 수에 따라 후보 분할 지점을 제시하고 집계된 통계량을 통해 최적해를 찾는다. 트리 모형 학습에서 가장 시간을 많이 소비하는 부분 중 하나는 정렬된 순서의 데이터를 얻는 것이다. 해당 비용을 감소하기 위해 XGBoost는 *block*으로 불리는 인-메모리 단위로 저장하는 방식을 제안하였다. 그 외에도 하드웨어적인 최적화를 구현하여 빠른 속도의 학습과 높은 정확도를 얻어내었다. 회귀・분류 모형 모두 적용 가능하며, 크기가 큰 데이터로의 확장성을 갖추고 최소한의 자원으로 문제를 해결할 수 있는 모형으로 평가받는다.

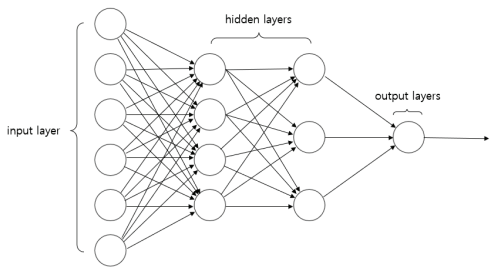
2.3 랜덤포레스트

랜덤포레스트는 Brieman(2001)에 의해 제안된 방법으로 트리 기반의 앙상블 모형이다[4]. 붓스트랩(bootstrap) 표본을 생성하여 다수의 의사결정나무를 만들고 각 결과를 종합하는 방법이다. 의사결정나무 모형은 회귀 분석이나 딥러닝 등과 같은 다른 방법에 비해 분석이 비교적 간단하고, 데이터가 분할되어 나무의 뿌리로 분할되므로 이해하기 쉽고 설명하기 쉬운 장점이 있다. 하지만, 예측력이 낮고 변수가 증가함에 따라 과적합(overfitting) 문제를 지닌다. 랜덤포레스트는 이러한 단점을 극복하기 위해 다수의 의사결정나무 결과를 종합하는 배깅(bagging) 방법을 사용하고, 나무 모형을 분할할 때 변수를 임의로 추출하는 방식을 사용한다. 분산을 감소시켜 예측오차를 줄이는 효과를 보여주며, 나무 모형 간 상관관계를 줄여주어 예측오차를 더 줄여주는 효과를 보여준다. 과적합 문제를 극복하는 장점이 있고, 소수의 변수에 영향을 크게 받지 않는 장점이 있다. 또한 bootstrap에 사용되지 않은 Out-of-bag(OOB) 표본을 평가할 때 사용할 수 있고, 변수 중요도를 계산해주어 변수들의 중요도를 판별할 수 있다.

2.4 딥러닝 (Multilayer Perceptron)

딥러닝은 기계학습 기법의 하나로 사람의 뇌 구조를 모방한 형태의 모형이다. 기본적으로 입력층(Input layer)과 출력층(Output layer)이 존재하며, 층 사이에는 2개 이상의 여러 개의 은닉층(Hidden layer)이 존재한다. 이미지 분야에서 많이 사용되는 합성곱 신경망(Convolution Neural Network, CNN)과 언어나 음성 분야에서 많이 사용되는 순환 신경망(Recurrent Neural network, RNN) 또한 딥러닝의 일부에 해당하는 모형이다.

컴퓨팅 능력이 향상되고, 빅데이터 시대가 도래함에 따라 데이터가 다양하고 커져 딥러닝의 많은 문제가 해결되고 활발히 연구되고 있다. 하지만, 대표적인 문제점으로 과적합과 매개변수 선택이 있다. 딥러닝의 매개변수는 활성화 함수, 손실 함수, 초기 가중치 분포, 은닉층 및 뉴런 개수, Epoch 등 많은 종류의 가짓수가 존재한다. 어떤 매개변수를 선택하느냐에 따라 모형의 구조 및 성능이 달라지며, 데이터에 따라 최적값을 찾는 매개변수 종류는 달라진다.



〈그림 1〉 2개의 hidden layer를 가진 딥러닝 모형 구조

Ⅲ. 분석 방법

실외공기측정기로부터 수집된 데이터를 사용하여 5분 뒤의 PM10・PM2.5 수치를 예측하는 모형을 각각 만들고자 한다. 모형은 기계학습 기법 중 빅데이터에서 성능이 뛰어난 XGBoost, 랜덤포레스트와 딥러닝(Multilayer Perceptron)을 사용한다.

실외공기측정기 데이터는 센서 데이터로 시간 측면에서의 잡음 제거와 공간 측면에서의 잡음 제거를 진행하였다. 시간 측면에서의 잡음 제거는 RLWS(Robust Locally Weighted Regression)을 사용하여 이상치를 제거한 후, LOESS(Local Regression)을 사용하여 제거하였다. 공간 측면에서의 잡음 제거는 GAM (Generalized Addictivd Model)을 사용하였고, 동시에 결측값 대체를 진행하였다. 실외공기측정기로부터 수집된 변수를 이용하여 파생 변수를 생성하였고, 모형의 매개변수 선택은 Golden Section Search와 Response Surface Model을 사용하여 진행하였다.

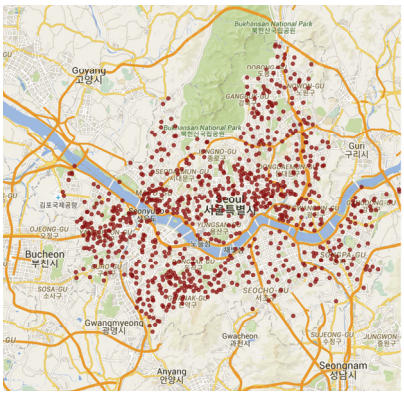
3.1 데이터 셋

데이터는 서울 지역에 위치한 845개의 실외공기측정기로부터 수집된 데이터를 사용한다. 측정기는 서울 지역에 골고루 분포되어 있으며,

〈표 1〉 실외공기측정기 데이터

Serial	Time	PM 10	PM 2.5	Temperature	Humidity
V02Q1940043	202005010000	12	4	22.9	46
V02Q1940043	202005010002	14	4	22.9	46
…	…	…	…	…	…
V02Q1940955	202103312346	153	84	24.2	76
V02Q1940955	202103312348	76	46	24.3	75

CCTV 폴대에 2~3m 높이로 설치되어 있다. 광산란 방식으로 수집되며, 기온・습도・PM10・PM2.5의 수치가 분 단위로 수집된다. 수집되는 시간 간격은 일정하지 않지만 대부분 2분 간격으로 수집된다. 원본 데이터는 2020년 5월부터 2021년 3월까지 수집된 데이터를 사용하였으며, 총 175,387,922개의 행으로 구성된다.



〈그림 2〉 서울시 실외공기측정기 분포

3.2 데이터 전처리

3.2.1 이상치 제거

빅데이터 분석 업무에서는 많은 변수가 기록되고 추출된다. 일관된 분석을 얻기 위해, 첫 단계의 업무 중 하나는 관측치 중 바깥에 위치한 이상치를 탐지하는 것이다. 탐지된 이상치는 잘못된 데이터의 후보들이며, 이는 적절하지 못한 모형의 구축으로 이어질 수 있다. 따라서, 모형 적합 전이나 분석을 하기 전, 이상치를 식별하고 제거하는 것은 좋은 결과를 얻기 위해 매우 중요하다.

이상치는 데이터 구조나 적용된 탐지 방법의 가정에 따라 다르게 정의될 수 있다. Hawkins (1980)은 이상치를 다른 메커니즘으로 생성되었다고 의심을 불러일으킬 정도로 다른 데이터로부터 많이 벗어난 값으로 정의하였다[6]. Barnett and Lewis(1994)은 이상치가 발생하는 다른 표본들과 비교했을 때, 현저하게 벗어난 것으로 정의하였다[2]. 비슷하게 Johnson(1992)은 데이터셋의 나머지 부분과 비교했을 때 일치하지 않는 것을 이상치로 정의하였다[7].

본 연구에서는 정규분포를 가정하여 RLWS (Robust Locally Weighted Regression)을 적합 시키고, 잔차의 $\overline{X} \pm 3s$를 벗어나는 값을 이상치로 판단하고 제거하였다. RLWS의 매개변수인 이웃 점 개수는 10-fold Cross Validation을 사용하여 선택하였다.

3.2.2 잡음 제거

센서로부터 수집된 원본 데이터는 원하는 정보를 얻기 위해 다양한 과정이 요구된다. 원본 데이터는 임의의 잡음을 포함하기 때문에 데이터의 품질을 개선하기 위해 전처리가 필요하다. 그중 잘 활용되는 절차 중 하나는 잡음 제거이다. 일반적으로 잡음 제거 알고리즘은 세 가지로 분류된다. 이전 시점 값과 현재 시점 값을 사용하여 현재 시점 값을 추정하는 Filtering과 이전 시점 값과 다음 시점 값을 사용하여 현재 시점 값을 추정하는 Smoothing, 이전 시점 값으로 이후 시점을 예측하는 Prediction이 일반적으로 분류되는 잡음 제거 알고리즘이다. 데이터의 이용 가능성이나 처리 목적에 따라 선택하여 사용할 수 있다[8].

Smoothing 기법은 과거 데이터와 미래 데이터가 있다면 잡음 제거에서 좋은 결과를 보여준다. 그중 LOESS는 함수를 규명하는 대신 평활 매개변수 값만 제공하면 되는 장점을 갖고 있으며, 매우 유연하여 복잡한 프로세스를 모형화하는 데 적합하다. 본 연구에서는 11개월간 수집된 과거 및 미래 데이터가 있고, 데이터 간격이 조밀하여 Smoothing 기법의 하나인 LOESS (Local Regression)을 사용하여 잡음을 제거하고자 한다. 매개변수인 이웃 점 개수는 10-fold Cross Validation을 사용하여 선택하였고, 모형

을 적합 후 5분 단위로 예측값을 추출하여, 지역이 다른 센서 간 시간 단위를 맞추고 잡음 제거를 진행하였다.

3.2.2.1 결측치 대체

데이터 분석은 결측치가 존재할 경우, 종종 결과에 방해받는다. 결측치를 처리하지 않고 모형 생성을 하게 되면 모형의 성능을 감소시키고 편향된 모형을 생성할 수 있다. 결측치는 데이터 추출, 데이터 수집 단계에서 발생할 수 있다. 또한 결측치는 크게 세 가지 가정인 MCAR (Missing completely at random), MAR(Missing at random), NMAR(Not missing at random)로 분류된다.

앞서 LOESS를 통해 실외 공기 측정기의 잡음을 제거하고 5분 단위로 추출된 데이터셋은 기존에 시간 간격이 일정하지 않고 미수신 구간이 존재하기 때문에, 결측치가 존재한다. GAM (Generalized addictive model)을 통해 시점(5분 단위)마다 공간 모형을 적합한 후, 모든 센서의 위치에(결측치를 보유한 센서 위치 포함) GAM 적합 값으로 대체해주었다. 설명 변수는 위도와 경도를 서울을 중심으로 km 단위로 환산한 값을 사용하였고, 교호작용 항을 추가해 적합하였다.

$$g(\mu) = \alpha + s_1(latitude) + s_2(longitude) + s_3(latitude, longitude) \quad (1)$$

〈그림 3〉 GAM 모형 적합 결과 시각화

GAM 모형의 항은 Radial Basis Function을 사용하였고, penalty term은 thin plate regression spline 모형을 사용하였다. 매개변수인 smoothing parameter는 GCV(Genralized Cross Validation)으로 선택하였고, basis function 개수는 10-fold Cross Validation을 통해 선택하였다.

3.3 모형 적합

3.3.1 파생 변수

실외공기측정기로부터 수집된 데이터는 serial · time · PM10 · PM2.5 · 습도 · 기온의 변수를 가진다. 5분 뒤의 PM10 · PM2.5 값을 예측하기 위해선 시공간 데이터의 자기상관성을 고려해야 한다. 한 센서로부터 수집된 가까운 시간 간격의 데이터는 관찰은 독립적으로 이루어지지만 서로 상관관계가 있다. 마찬가지로 가까운 거리에 위치한 센서들은 서로 공간자기상관성을 갖는다.

PM10 · PM2.5 · 습도 · 기온 변수의 이전 시점 값을 사용하여 이전 시점 변수를 생성하고, 센서별 가장 가까운 5개의 센서를 선정하여 해당 센서의 이전 시점 변수를 생성해주었다. 파생 변수 생성 결과는 <표 2>과 같다.

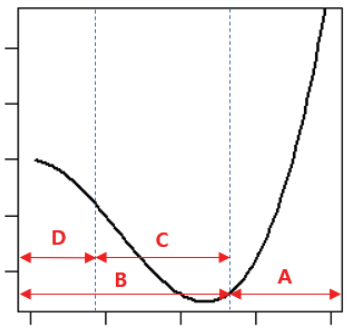
〈표 2〉 파생 변수

PM10 예측 모형	PM2.5 예측 모형	변수설명
month	month	월
PM10_t1～5	PM2.5_t1～5	미세먼지 농도 이전시점 5개 값(5분～25분 전)
temp_t1～5	temp_t1～5	기온 이전시점 5개 값(5분～25분 전)
humi_t1～5	humi_t1～5	습도 이전시점 5개 값(5분～25분 전)
nb1_pm10_t1～5, nb2_pm10_t1～5, nb3_pm10_t1～5, nb4_pm10_t1～5, nb5_pm10_t1～5	nb1_pm25_t1～5, nb2_pm25_t1～5, nb3_pm25_t1～5, nb4_pm25_t1～5, nb5_pm25_t1～5	1～5번째로 근접한 센서의 이전시점 5개 값(5분～25분 전)

3.3.2 매개변수 선택

3.3.2.1 황금 분할 탐색(Golden Section Search)

황금 분할 탐색(Golden Section Search)은 두 점으로 생성된 구간 내에 존재하는 극값을 찾는 방법이다. 본 연구에서는 각 매개변수에 따른 RMSE 값이 극솟값이 되는 구간을 찾기 위해 사용한다. 그림 0과 같이 전체 구간 중 황금 비율로 두 구간 A, B를 나누고 A와 B 구간 중 극솟값이 존재하는 구간을 선택한다. 극솟값이 존재하지 않는 구간은 제외하며 극솟값이 존재하는 구간에서 다시 황금 비율로 구간을 C, D로 나눈 후 일정 조건을 만족할 때까지 반복한다. 해당 방법을 XGBoost, 랜덤포레스트, 딥러닝(Multilayer Perceptron)의 매개변수에 적용하여 각 매개변수 별 초기 구간을 설정한다.



〈그림 4〉 황금 분할 탐색 구간 선택 과정

3.3.2.2 반응 표면 방법 (Response Surface Method)

실험계획법은 최소 비용으로 최대 정보를 얻는 데에 의의가 있다. 실험계획법 중 반응 표면 방법(Response Surface Method)은 다양한 요인들과 요인들 간의 교호작용(Interaction)이 반응(Response)에 미치는 영향을 분석하고 최적화된 조건 값을 얻기 위하여 Box와 Wilson이 고안해 낸 방법이다[3]. 반응 표면 방법은 최소한의 실험 횟수로 최적의 조건을 얻어내도록 실험을 설계해야 한다.

일반적으로 사용되는 일원, 이원, 다원 배치법에서는 각 인자의 수준(Level)을 정하고 여러 인자들의 수준 조합 중 하나가 최적의 조건으로 선택된다. 반응 표면 방법은 설정되지 않은 수준의 값이 최적값으로 선택될 수 있으며, 인자의 수준 내에서 직선 관계가 아닌 곡선의 관계를 맺을 때, 유용하게 사용된다.

반응 표면 방법은 회귀 분석을 적용하여 인자들 간의 교호작용을 고려하고 정상점(Stationary Point)를 찾는다. 분석 모형은 일차항과 교차항, 이차항에 포함 여부에 따라 3가지로 분류된다. 일반적으로 일차항과 교차항을 포함한 모형인 Screening Response(1) 모형으로 초깃값을 탐색한다. 정상점 근처에 도달할 때까지 Steepest Ascent 모형을 사용하며, 설정한 범위에서 정상점이 존재할 때 Second-order 모형(2)을 사용한다. 본 연구에서는 황금 분할 탐색으로 얻어진 범위 내에 정상점이 존재한다는 가정하에 Optimization 모형만을 사용한다. 인자의 개수가 2일 때, 모형의 수식 표현은 다음과 같다.

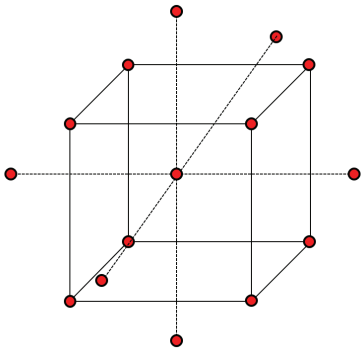
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \quad (2)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_{11}^2 + \beta_{22} x_{22}^2 + \beta_{12} x_1 x_2 + \epsilon \quad (3)$$

3.3.2.3 Box-Behnken Design(BBD)

반응 표면 방법을 사용하기 위해선 함수를 추정할 수 있는 데이터가 필요하다. 모든 조합으로 실험을 설계한다면 표면을 잘 추정할 수 있지만, 실험 수가 많을수록 큰 비용이 소모된다. 반응 표면 방법에서 가장 많이 사용되는 설계법은 Box-Behnken Design(BBD)과 Central Composite Design(CCD)이 있다.

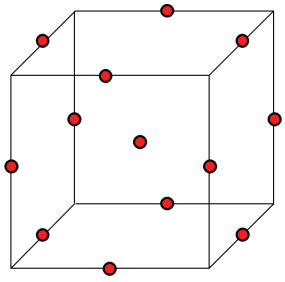
Central Composite Design(CCD)은 그림 0과 같이 2-수준 요인 설계에서 실험한 점과 중앙점, 축점(axial point)으로 구성된다.



〈그림 5〉 Central Composite Design의 실험점

Central Composite Design의 장점은 추가적인 실험을 통해 회전 가능성을 가진 설계를 한 것이다. 회전 가능성은 설계의 중심으로부터 모든 점에서 예측 분산이 일정해 예측의 질을 향상시키는 장점이 있다. 하지만, 요인의 수준을 벗어나기 때문에, 정의되지 않은 수준이 포함되는 경우가 발생할 수 있고, 실험 횟수가 증가하는 단점을 갖는다.

Box와 Behnken에 의해 소개된 Box-Behnken Design(BBD)은 Central Composite Design과 다르게 극단 값을 사용하지 않는다.



〈그림 6〉 Box-Behnken Design의 실험점

Box-Behnken Design은 실험 횟수가 비교적 적고, 요인 수에 따라 회전 가능성을 갖는다. 또한, Central Composite Design과 달리 설계 요인 수준을 벗어나는 경우가 발생하지 않는다. 본 연구에서는 이러한 장점으로 인하여 Box-Behnken design을 사용하여 각 모형의 매개변수를 선택하였다.

Ⅳ. 분석 결과

전처리가 완료된 데이터의 관측치 수는 65,821,548개이고, 변수는 42개이다. 각 모형의 매개변수를 선택하기에 앞서 전체 데이터셋의 30%를 평가용 데이터로 분류한 후, 70%의 데이터만을 사용하여 모형 적합을 진행했다. 훈련용 데이터셋의 관측치 개수는 총 45,870,864개이고, 평가용 데이터셋의 관측치 개수는 19,950,684개이다.

모형별 설정할 매개변수는 <표 3>과 같다.

〈표 3〉 모형별 조정 매개변수

모형	매개변수	변수설명
XGBoost	eta	학습률 조정 변수
	gamma	분할을 위한 최소 손실 감소량
	max depth	최대 트리 깊이
	colsample bytree	훈련 관측치 샘플링 비율
	subsample	변수 샘플링 비율
랜덤포레스트	number of trees	트리 개수
	mtry	분할 후보 변수 개수
	max depth	최대 트리 깊이
딥러닝	hidden nodes	은닉층 노드 개수
	hidden layers	은닉층 개수
	epochs	역전파 학습 횟수
	dropout ratio	드롭아웃 노드 비율
	l1	L1 노름(norm) 가중치 규제
	activation	활성 함수
	initial weight distribution	초기 가중치 분포

4.1 황금 분할 탐색을 통한 매개변수 구간 설정

황금 분할 탐색을 사용하여 매개변수의 실험 범위를 설정한다. 설정 방법은 실험할 매개변수를 제외한 나머지 매개변수는 고정하고, 실험할 매개변수의 값을 변화시켜가며 최적의 구간을 찾는다. 값의 변화가 없는 범주형 변수인 활성함수와 초기 가중치 분포는 황금 분할 탐색을 수행하지 않았다. 황금 분할 탐색으로 찾은 구간은 Box-Behnken Design에 사용될 것이므로 매개변수별 수준 간격의 변화는 균일하게 지정하였다. 모형별 최종적으로 선정한 구간은 <표 4>와 같다.

〈표 4〉 황금 분할 탐색 결과

모형	매개변수	구간(PM10)	구간(PM2.5)
XGBoost	eta	[0.03.0.07]	[0.05,0.09]
	gamma	[9,13]	[9,13]
	max depth	[4,8]	[4,8]
	colsample bytree	[0.5,0.9]	[0.5,0.9]
	subsample	[0.5,0.9]	[0.5,0.9]
랜덤 포레스트	number of trees	[2000,3000]	[2000,3000]
	mtry	[10,20]	[10,20]
	max depth	[40,60]	[40,60]
딥러닝	hidden nodes	[50,150]	[50,150]
	hidden layers	[8,12]	[8,12]
	epochs	[150,250]	[200,300]
	dropout ratio	[0.7,0.9]	[0.7,0.9]
	l1	[0,2]	[0,2]
	activation	Rectifier, Tanh	
	initial weight distribution	Uniform Adaptive, Normal	

4.2 반응 표면 방법 적용

본 연구에서는 반응 표면 방법을 위해 Box-Behnken Design을 사용하였다. 황금 분할 탐색으로 얻어진 구간을 요인 수준 설계에 적용하였다. 모형별 요인 개수에 따라 XGBoost는 46번, 랜덤포레스트는 15번, 딥러닝(Multilayer Perceptron)은 46×4=184번의 실험 횟수를 갖는다[3]. 실험 점에서의 RMSE 값을 계산한 후, RMSE를 반응변수로 반응 표면 분석을 수행하였다. Optimization 모형을 사용하여 각 모형의 매개변수 최적값을 선정하였다. 최종 선정 결과는 <표 5>와 같다.

〈표 5〉 반응 표면 방법 매개변수 선정 결과

모형	매개변수	매개변수 최적값(PM10)	매개변수 최적값(PM2.5)
XGBoost	eta	0.03	0.05
	gamma	9	9
	max depth	6	6
	colsample bytree	0.9	0.9
	subsample	0.5	0.9
랜덤 포레스트	number of trees	3000	2500
	mtry	15	15
	max depth	40	50
딥러닝	hidden nodes	50	50
	hidden layers	10	10
	epochs	200	250
	dropout ratio	0.7	0.9
	l1	0	0
	activation	Tanh	Tanh
	initial weight distribution	Normal	Normal

4.3 모형 비교

반응 표면 방법으로 선정된 매개변수 최적값을 사용하여 각 모형을 적합하였다. 모형 비교는 평가용 데이터를 사용하였고, RMSE 값을 확인하였다. 최종 적합 결과<표 6>, PM10 예측 모형의 경우 랜덤포레스트가 RMSE가 가장 낮아 PM10을 정확하게 예측하는 모형으로 드러났다.

PM2.5 예측 모형에서는 딥러닝이 가장 정확하게 예측하는 것으로 확인되었다.

〈표 6〉 최적 매개변수에 따른 모형 RMSE

모형	RMSE(PM10)	RMSE(PM2.5)
XGBoost	3.133886	1.947246
랜덤포레스트	1.031336	0.8491444
딥러닝	4.865213	0.642407

Ⅴ. 결론

본 연구에서는 서울시에 위치한 실외공기측정기로부터 데이터를 수집하여, 5분 뒤 미세먼지 농도를 예측하는 모형을 생성하였다. 잡음 제거, 이상치 제거 등의 데이터 전처리와 반응표면 방법을 통한 최적 매개변수 선택을 진행하였고, 예측 모형에는 XGBoost, 랜덤포레스트, 딥러닝(Multilayer Perceptron)을 사용하여 성능을 비교하였다. 랜덤포레스트가 5분뒤 PM10 농도 예측을 가장 정확하게 수행하였으며, 딥러닝은 5분뒤 PM2.5 농도 예측을 가장 정확하게 수행하였다.

본 연구에서 사용된 데이터는 실외공기측정망 센서로부터 수집된 데이터이다. 해당 센서는 지상으로부터 약 10m 높이에 설치되어 있다. 실제 신뢰성이 높은 국가관측망 센서는 건물의 옥상에 주로 설치되어 있고 수집 방식이 베타선 방식으로 본 연구에서 사용된 센서와 약간의 차이가 있다. 하지만, 서울시에 위치한 국가관측망 센서는 40여 개로 실외공기측정기 센서보다 적은 개수이며 한 시간 단위로 공표된다. 또한, 본 연구에 사용된 센서는 도심 곳곳에 위치해 있고 대부분 2~3분 간격으로 수집되어 실시간으로 미세먼지 맵을 이용자에게 제공하기에 유리하다.

본 연구에서 생성한 예측 모형은 5분 뒤를 예보하는 모형이다. 추후에는 반응 변숫값을 10분 뒤, 30분 뒤, 1시간 뒤 등으로 설정하여, 더 긴 시간 후를 예보하는 모형을 구축할 수 있다. 본 연구에 사용된 실외공기측정기 자료를 보았을 때, 미세먼지 농도 값은 10분, 30분, 1시간 단위로 값이 급변하지 않는다. 따라서 본 논문에 제시된 방법을 그대로 사용하여도 무방할 것이다. 단, 정확하고 실용성이 높은 시스템을 구축하기 위해선, 검증하는 절차가 필요할 것이며, 풍향·풍속·강수량 등의 기상 변수를 추가하거나, 200개 이상의 추가 설치된 센서를 모형에 포함하여 국가관측망 센서와의 차이를 규명할 수 있다. 또한, 본 연구에서 사용하지 않은 지역별 변수 및 공간적 요인에 관한 연구를 지속해서 수행할 필요가 있다.

참 고 문 헌

[1] Atluri, Gowtham, Anuj Karpatne, and Vipin Kumar. “Spatio-temporal data mining: A survey of problems and methods.” ACM Computing Surveys (CSUR) 51.4 (2018): 1-41.

[2] Barnett, Vic, and Toby Lewis. “Outliers in statistical data.” Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics (1984).

[3] Box, George EP, and Donald W. Behnken. “Some new three level designs for the study of quantitative variables.” Technometrics 2.4 (1960): 455-475.

[4] Breiman, Leo. “Random forests.” Machine learning 45.1 (2001): 5-32.

[5] Chen, Tianqi, and Carlos Guestrin. “Xgboost: A scalable tree boosting system.” Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[6] Hawkins, Douglas M. Identification of outliers. Vol. 11. London: Chapman and Hall, 1980.

[7] Johnson, Richard Arnold, and Dean W. Wichern.

Applied multivariate statistical analysis. Vol. 6. London, UK:: Pearson, 2014.

[8] Kowalski, Paweł, and Robert Smyk. “Review and comparison of smoothing algorithms for one-dimensional data noise reduction.” 2018 International Interdisciplinary PhD Workshop (IIPhDW). IEEE, 2018.

[9] 이범석. 반응 표면 방법을 이용한 딥러닝 매개변수 최적화 연구 . 인천: 인하대학교 대학원, 2017. Print.

저 자 소 개

박 헌 진(Heon Jin Park)

·1990년 9월～1994년 8월 : SAS Institute Inc. Senior Research Statistician

·1994년～현재 : 인하대학교 통계학과 교수, 데이터사이언학과 학과장

·관심분야 : 데이터마이닝, 시계열, 통계계산

박 종 찬(Jong-Chan Park)

·2020년 2월 : 인하대학교 통계학과 (학사)

·2020년 3월～현재 : 인하대학교 통계학과 석사과정

·관심분야 : 데이터 마이닝, 빅데이터, 머신러닝, 이상 탐지