pISSN : 2508-3384 eISSN : 2508-3392
https://doi.org/10.22640/lxsiri.2021.51.2.141
Journal of Cadastre & Land InformatiX Vol.51 No.2 (2021)
한국국토정보공사「지적과 국토정보」제51권 제2호 2021년 12월 p. 141-150

기계학습 Adaboost에 기초한 미세먼지 등급 지도

Particulate Matter Rating Map based on Machine Learning with Adaboost Algorithm

정 종 철*
Jeong, Jong-Chul

Abstract

Fine dust is a substance that greatly affects human health, and various studies have been conducted in this regard. Due to the human influence of particulate matter, various studies are being conducted to predict particulate matter grade using past data measured in the monitoring network of Seoul city.

In this paper, predictive model have focused on particulate matter concentration in May, 2019, Seoul. The air pollutant variables were used to training such as SO_2, CO, NO_2, O_3. The predictive model based on Adaboost, and training model was dividing PM_{10} and $PM_{2.5}$. As a result of the prediction performance comparison through confusion matrix, the Adaboost model was more conformable for predicting the particulate matter concentration grade. Although air pollutant variables have a higher correlation with PM2.5, training model need to train a lot of data and to use additional variables such as traffic volume to predict more effective PM_{10} and $PM_{2.5}$ distribution grade.

Keywords: Particulate Matter(PM), Machine Learning, Adaboost, Air Pollutants

1. 서론

미세먼지, 즉 작은 입자의 먼지들은 기후변화와 환경문제의 주요 인자로 시민들에게 악영향을 미친다. 특히, 노약자 및 청소년 등과 같은 사회적 약자가 대상일 경우에는 보다 큰 위험에 노출될 수 있다. 입자가 작은 미세먼지에 장시간 노출될 경우, 호흡기 질환뿐만 아니라 심혈관, 내분비, 신경정신, 피부에도 영향을 미치게 된다(이승복 2019). 미세먼지는 가스 형태 대기오염물질인 오존, 이산화황, 일산화탄소 보다 노출시에 인체에 영향력이 큰 것으로 보고되고 있다(강민성 등, 2016). 해외 연구에서는 대기오염물질 중 미세먼지의 노출과 질병률, 사망률이 밀접한 연관성을 가지고 있으며, 노출 정도를 줄임으로써 기대수명을 연장할 수 있다고 제시하였다(Lary et al. 2015). 따라서 다양한 연구를 통해 미세먼지 저감대책을 제시하거나(김운수 2014), 세밀한 측정망 구축을 위한 측정소 위치선정(유재환 등 2011; 최임조 등 2016) 방법과 미세먼지 분포 경향을 분석하는 등(성선용 2019; 정종철 2014)의 다양한 연구 결과가 발표되었다.

* 남서울대학교 드론공간정보공학과 교수 Professor, Dept. of Drone Spatial-Information Eng. Namseoul Univ. (jjc1017@daum.net))

미세먼지의 경우 자연적인 발생과 인위적인 발생으로 구분되는데, PM_{10}과 같은 미세입자의 경우 자동차 운행 중의 연료 연소나 다양한 시설물의 화학물질 처리 과정에서 생성된다. 이 과정에서 미세먼지와 같은 물질 뿐 아니라 황산화물, 일산화탄소 등의 추가적인 오염물질 또한 발생한다. 이와 같은 발생원을 기반으로 미세먼지와 그 외의 대기오염물질은 상관성이 존재하는 관계임을 확인할 수 있다. 미세먼지를 비롯한 대기오염물질의 피해는 시민들에게 공간적인 분포정보를 제공해야 할 필요성이 있으며, 국가에서는 미세먼지를 포함한 대기오염 물질 현황을 환경부 에어코리아 웹사이트와 서울시를 비롯한 지방자치단체에서 홈페이지를 통해 시민들에게 실시간 농도를 제공하고 있다.

미세먼지와 초미세먼지와 같은 경우, 도시대기 측정소와 도로변 측정소를 기반으로 측정되고, 실시간으로 시민들에게 가장 가까운 측정소의 관측정보를 제공하고 있다. 하지만 실시간 측정되는 미세먼지는 시민에게 제공되는 정보가 좋음과 나쁨과 같은 인식정보의 효과적 전달을 위한 등급평가로 정보 제공이 이루어지고 있다. 미세먼지를 PM_{10}과 $PM_{2.5}$로 구분하여 측정값의 등급을 구분하고 이에 대한 결과로 시민들은 미세먼지 등급정보를 제공받는다.

하지만 등급 산정의 방법이 단순한 농도 기준의 단계이고 PM_{10}과 $PM_{2.5}$로 구분하여 측정된 측정값만을 기준으로 하고 있어서 본 연구에서는 미세먼지와 상관성이 높은 다른 오염물질과의 상관성을 기반으로 미세먼지 등급예측 모델을 생성하였다.

등급산정은 측정값의 공간적 구분으로 지방자치단체의 행정구역이나 서울시 구별 행정구역을 기준으로 분석하고 이를 제공해야 하는데 공간분석의 도구가 적용된 연구는 부족한 실정이다. 이러한 문제를 해결하기 위해 기계학습 예측모델을 접합하여 미세먼지 결측값에 대한 예측을 제시하는 연구가 진행되었으며(정종철 2017; 이기혁 2020), 미세먼지를 예측하기 위한 학습 변수로 기상인자, 교통량과 같은 상관성을 가지는 데이터를 활용한 공간분석 연구가 진행되었다.

공간분석과 함께 기계학습을 기반으로 한 미세먼지 예측모델을 제시하는 해외 연구가 최근 발표되고 있다(Rochelle et al. 2020; Jan et al. 2017; Hamed Karimian et al. 2019). 하지만 대기오염 측정망과 기상인자 및 교통량과의 상관성을 분석하는 과정에서 동일한 지점에서 측정이 가능한 자료의 수집이 어렵고, 미세먼지와 동일한 지역에서 측정되는 다른 대기오염물질만을 활용하여 학습을 진행될 경우, 보다 정확한 예측 모델을 제시할 수 있을 것으로 판단되었다.

본 연구는 측정 미세먼지 값을 기반으로 피해등급 예측을 동일 시간대에 측정된 다른 대기오염물질을 기반으로 미세먼지 등급을 추정하고자 하였다. 또한 분류된 결과를 기반으로 서울시 행정구역에 따른 미세먼지 등급 지도를 제시하는 연구를 진행하였다.

2. 선행이론

머신러닝은 Frank Rosenblatt가 개발한 대표적 지도학습 알고리즘인 퍼셉트론으로부터 발전되었다. 하지만 정확도 및 학습지역에 대한 여러 가지 제약으로 인해, 최신의 머신러닝 알고리즘으로는 활용하기 힘들다. 하지만 기계학습 알고리즘을 이용한 주택 모기지 금리에 대한 시민들의 감정예측(김윤기 2019)과 같이 기계학습 알고리즘은 다양한 분야에 활발하게 적용되고 있고, 현재 사용되는 기계학습 알고리즘에 경우, 예측된 결과 값의 오차에 대한 가중치를 갱신함으로써 다중 신경망에 최적의 기능을 구현할 수 있다는 것을 증명하며 알고리즘을 발전시켰다.

기계학습을 위한 감독분류 알고리즘은 기존의 선행연구에서도 다양하게 존재하지만, 본 연구에서는 Adaboost기법을 활용하였다. Adaboost는 Yoav와 Robert가 제안한 기계학습 메타 알고리즘이다. Adaboost는 이전에 분류된 약분류기(weak classifier)를

기준으로 상호 보완하여 순차적으로 재학습시키는 방식의 학습 알고리즘이다. 학습과정에서 이전의 학습 데이터의 학습 결과와 다음 학습데이터의 학습에서 문제점이 존재할 경우 이전의 학습 결과를 보완하는 형식으로 진행되는 Adaboost는 최종적으로 학습데이터를 기준으로 예측 성능을 향상 시킬 수 있는 장점이 존재한다.

Adaboost에서 최종 분류식은 다음과 같이 T개의 나무 모형 예측값 $f(x)$의 가중합으로 계산된다.

$$F(x) = sign(\sum_{t=1}^{T} \alpha_t f_t(x))$$

여기서, $f_t(x)$는 t번째 약한 학습기(weak learner)로 예측한 y값을 의미하며, α_t은 t번째 학습 결과에 곱해지는 가중치를 의미한다. $F(x)$의 값이 양수일 경우 1, 음수일 경우 -1을 출력해 예측 결과를 낸다. 만약 데이터가 n개의 샘플이며, $x_i \in R^d, y_i \in -1,1$일 때 Adaboost의 가중치 계산은 다음 순서와 같다. 첫 번째, 모든 샘플에 똑같은 가중치를 주어 가중치를 초기화 한다.

$$D_1(x_i, y_i) = \frac{1}{n}, i = 1,...,n$$

$t = 1,..., T$ 번째 반복마다 오차 계산 후 t번째 약한 학습기에 곱해지는 가중치 α_t을 다음과 같이 계산한다.

$$\alpha_t = \frac{1}{2} ln(\frac{1-\epsilon_t}{\epsilon_t})$$

여기서 ϵ이 0일 경우 α_t는 기하학적으로 커지게 되며 오류가 적으면 해당 모델은 더 많은 가중치를 가지게 된다. ϵ이 0.5면 α_t는 0이 되기 때문에 50%의 정확도를 가진 분류기는 반반으로 분류하게 된다. 따라서 이 경우는 분류기에서 자체적으로 무시하게 된다. ϵ이 1이면 α_t는 음수 값이 되기 때문에 50%의 보다 오차가 크면 잘못 분류하는 방향을 조정하기 위해 분류기는 반대 결과를 도출하도록 가중치를 조정한다.

3. 연구방법

3.1. 연구 대상지

본 연구에서는 서울시와 환경부에서 운영하는 대기오염 측정망을 중심으로 수집된 미세먼지 관련 데이터를 AirKorea에서 수집하고 데이터를 분류하고 PM_{10}과 $PM_{2.5}$로 구분하여 기계학습을 진행하였다.

서울시 미세먼지를 측정하는 측정소는 대표적인 40곳을 적용하였다. 대기측정소는 도시대기 측정소와 도로변 측정소로 구분하는데, 도시대기 측정소의 경우 행정구별 1개씩 도심 내에 위치하며 총 25개의 측정소를 선정하였다. 도로변 측정소는 서울시 내에 존재하는 교통량이 많은 대로변 주변에 설치되어 있으며, 15개의 측정소를 선정하였다. 본 연구에서 수집한 미세먼지 농도는 40개 측정소의 위치로 Figure 1과 같다.

환경부(2020)에서 제공한 '2020 환경통계연감'에 따르면 미세먼지는 봄과 겨울철에 높은 경향을 보여주는 것으로 나타났다(정종철 2019). 이는 계절적으로 나타나는 기상조건과 중국으로부터 유입되는 미세먼지의 영향이 크게 나타난 시기를 고려하여 본 연구에서는 2018년 11월부터 2019년 3월까지의 5개월 동안의 서울시 대기측정소 값을 수집하였으며 수집된 데이터는 미세먼지(PM_{10})과 초미세먼지($PM_{2.5}$) 그리고 학습변수로 활용할 SO_2, NO_2, CO, O_3를 수집하였다.

3.2. 데이터 처리

본 연구는 Figure 2와 같은 흐름으로 진행되었다. 미세먼지 데이터는 에어코리아에서 제공하는 시간별 데이터를 수집하였으며, 이중 도시대기 측정소 25개

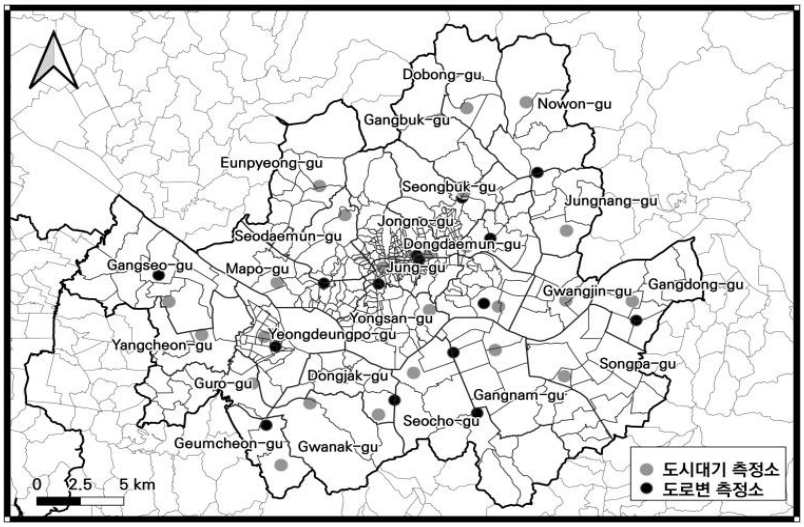


Figure 1. Current status of Seoul Metropolitan Government's measuring station.

Table 1. Standard for fine dust levels
Source: Ministry of Environment.

Concentration (㎍/㎥, Daily average.)	good	normal	bad	very bad
PM_{10}	0-30	31-81	81-150	150 over
$PM_{2.5}$	0-15	16-35	36-75	76 over

소의 데이터를 취득하였다. 미세먼지농도의 공간적 분포를 구분하는 방법으로 환경부에서 제공하는 피해등급을 활용하였다. 환경부는 일평균 미세먼지를 기준으로 피해등급을 구분하여 시민들에게 제공하고 있으며, Table 1과 같다. 따라서 본 연구에선 취득한 모든 미세먼지 데이터를 Table 1과 같은 4가지 등급으로 구분하여 학습을 진행하였다.

대기오염 물질은 오존(O_3), 이산화질소(NO_2), 일산화탄소(CO), 아황산가스(SO_2)이며, 이는 조경우 등(2019)과 Guang Yang et al.(2020)이 딥러닝을 기반으로 학습한 변수로, 미세먼지와의 상관성을 가지고 있음을 제시한 바 있어서 본 연구의 적용기법에 응용하였다.

Table 2. As a result of analyzing the correlation between fine dust(PM_{10}) and air pollutants.

fine dust	correlation coefficient			
	SO_2	NO_2	CO	O_3
(p < .01)	0.193	0.538	-0.065	0.325

4. 연구결과

4.1. 미세먼지와 대기물질의 상관성 분석

모델 학습에 앞서 본 연구에서 사용한 미세먼지 데이터와 학습변수인 대기오염물질과의 상관성을 확인하는 과정은 본 연구의 학습변수에 대한 선정의 유의미를 평가하는 단계로 수행하였다. 본 연구에서는 미세먼지(PM_{10})과 초미세먼지($PM_{2.5}$)를 구분하고 측정값에 대하여 유의확률과 Pearson 상관계수 값을 통해 변수간의 상관성을 분석하였다. Table 2는 미세먼지와 대기오염물질 간의 상관성 분석 결과이다.

상관성 분석 결과에서 미세먼지(PM_{10})와 가장 높은 상관성을 가지는 대기오염물질 변수는 NO_2로 나타났

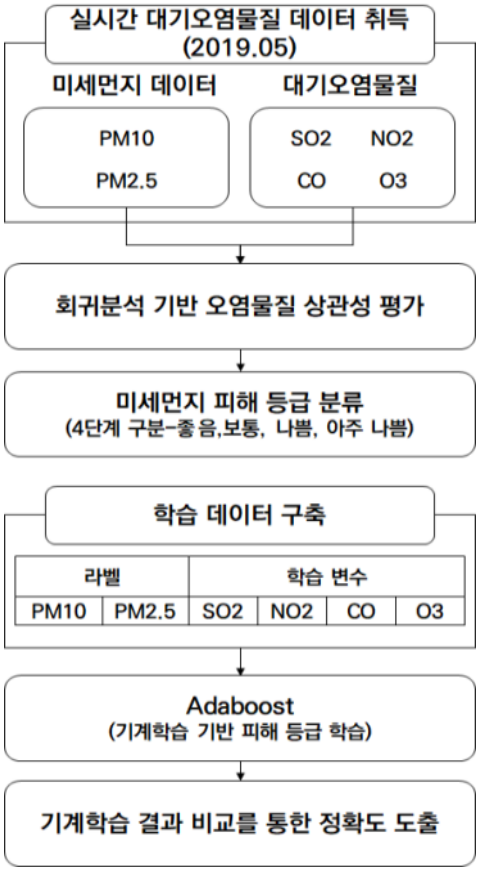


Figure 2. Research flow chart

다. 미세먼지와 NO_2 사이에서는 0.538로 비교적 높은 정적 상관계수($p<0.01$)로 나타낸다. 이와 반대로 CO의 경우 미세먼지와 음의 상관관계임을 확인할 수 있었다. Table 3은 초미세먼지($PM_{2.5}$)와 대기오염물질 간의 상관성 분석 결과이다. 초미세먼지의 경우에도 미세먼지와 마찬가지로 NO_2와 가장 높은 상관성을 확인할 수 있었으며, 학습에 활용할 4가지 변수에 대하여 높은 상관성을 가지는 것으로 나타났다.

초미세먼지와 NO_2 사이의 pearson 상관계수는 0.644로 미세먼지보다 높은 정적 상관계수를 보여주는 것을 확인하였다($p<0.01$). CO의 경우에도 초미세먼지와 음의 상관관계로 나타났다. 이와 같은 미세먼지와 초미세먼지에 대한 상관성 분석을 통해 대기오염물질을 학습변수로 지정하게 될 경우 미세먼지에 대한 등급 예측보다 초미세먼지에 대한 등급 예측이 보다 높은 정확도를 도출할 수 있다고 판단되었다.

Table 3. As a result of analyzing the correlation between ultra-fine dust($PM_{2.5}$) and air pollutants.

ultra-fine dust	Air pollutant variables pearson correlation coefficient			
	SO2	NO2	CO	O3
(p < .01)	0.259	0.644	-0.170	0.433

Table 4. The results of Fine dust levels

Adaboost Classification results	Observation data grade			
	good	normal	bad	very bad
good	41	284	41	0
normal	1,498	10,469	2,411	92
bad	31	629	1,145	314
very bad	6	97	277	210

4.2. 모델 학습 결과

4.2.1. 취득 데이터 학습 결과

본 연구는 미세먼지와 초미세먼지를 구분하여 두 개의 분포 등급에 대한 모델을 학습하였다. 서울시 도시대기 측정소에서 추출한 17,545개의 데이터를 기준으로 Adaboost를 통해 학습된 미세먼지 등급 결과는 Table 4와 같다. 전체 데이터 중 보통 등급의 데이터가 가장 많은 양의 분류 결과를 나타내었으며, 이에 대한 결과로 가장 많은 양의 데이터가 보통 등급으로 모델 상에서 재분류 되는 것을 확인할 수 있었다.

초미세먼지와 관측데이터간의 학습 분류 결과는 Table 5와 같이 나타난다. 초미세먼지의 분류 결과 또한 관측 데이터 상에서 가장 많은 학습 데이터를 가지

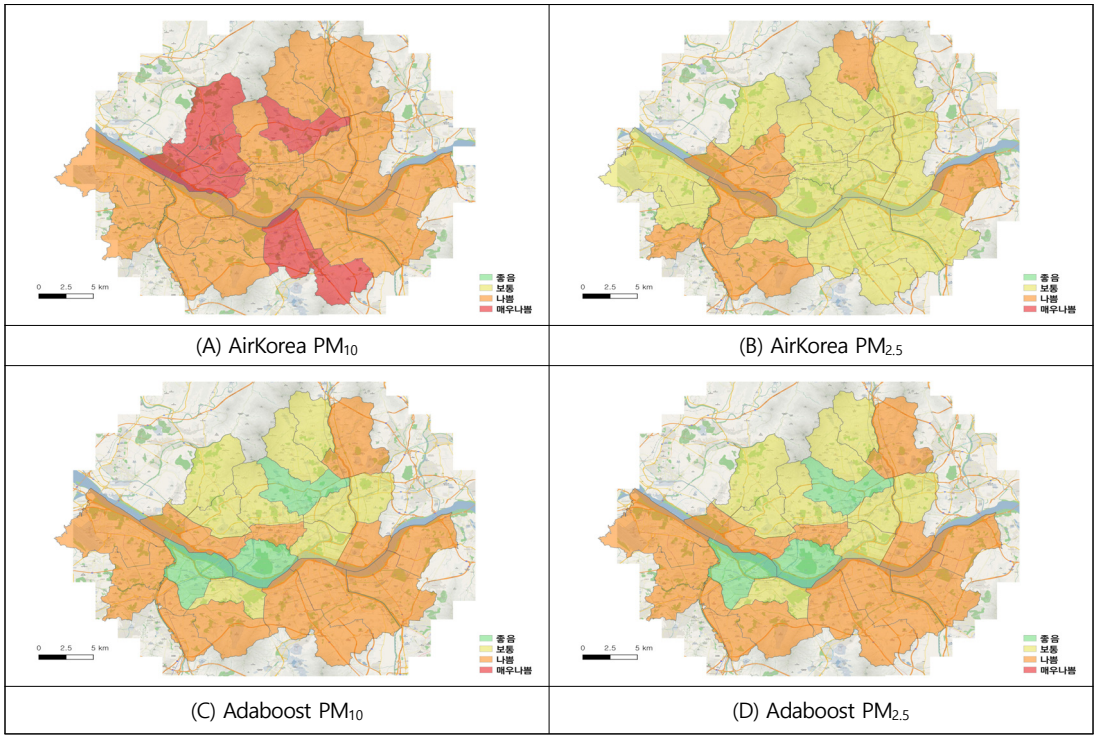
Table 5. The results of the ultra-fine dust level

Adaboost Classification results	Observation data grade			
	good	normal	bad	very bad
good	1358	1785	15	9
normal	1352	8265	599	202
bad	16	1665	1033	319
very bad	1	174	396	356

고 있는 보통 등급의 분류가 나타나는 것으로 확인되었다. 또한 전반적으로 미세먼지보다는 초미세먼지와 대기오염 상의 상관성을 결과를 통해 확인할 수 있었다.

1차적으로 활용된 데이터의 경우 수집된 데이터의 수가 차이가 나고, 이에 대한 모델의 과적합 문제가 발생 할 수 있다고 판단하여, 두 개의 데이터에 대한 2차 학습을 진행하였다. 수집된 데이터를 기반으로 미세먼지와 초미세먼지의 등급이 같은 데이터 중 각 등급별로 500개씩 총 2,000개의 데이터 세트를 재 추출하였다. 그 후 1,600개(80%)의 데이터를 훈련을 위한 데이터로 선정하였으며, 나머지 400개(20%)의 데이터를 테스트 데이터로 활용하였다. 따라서 미세먼지와 초미세먼지는 같은 등급으로 통일된 결과를 가지고 있으며, 미세먼지를 기반으로 분류된 결과 학습 정확도는 교차검증을 통해 89.64%까지 상승하였다. 이러한 학습 모델을 기반으로 400개의 테스트 데이터를 분류할 경우 Table 6과 같은 결과로 나타났다. 미세먼지의 측정정보를 시민에게 제공하는 미세먼지 등급정보로 제공하는 지도의 생성은 Figure 3과 같다. 이는 측정망에서 제공하는 미세먼지 등급의 공간적 범위를

Figure 3. Fine dust level map of 12:00 May 4, 2019, (A) AirKorea PM_{10}, (B) AirKorea $PM_{2.5}$, (C) Adaboost PM_{10}, (D) Adaboost $PM_{2.5}$



(A) AirKorea PM_{10} (B) AirKorea $PM_{2.5}$

(C) Adaboost PM_{10} (D) Adaboost $PM_{2.5}$

Table 6. The result of the 2nd round of fine dust level

Adaboost Classification results	Observation data grade			
	good	normal	bad	very bad
good	83	5	12	0
normal	41	20	29	10
bad	9	6	49	36
very bad	1	10	32	57

미세먼지가 심각한 5월의 샘플 자료를 기초로 본 연구의 Adaboost알고리즘으로 구분하여 비교할 수 있었다.

2차 학습을 통한 결과에서는 이전의 미세먼지를 분류하는 과정에서 보통에서만 많은 데이터가 분류되는 과정이 제거되어 모든 등급에서 일치 데이터가 가장 많은 수를 보여주는 것을 확인할 수 있었다. 각각의 등급에 대한 일치율은 Table 7과 같다.

4.2.2. 미세먼지 등급지도

본 연구에서는 생성된 모델을 기반으로 미세먼지 등급지도를 제작하여 실제 미세먼지 등급지도와 어느 정도 차이를 나타내는지에 대한 경향을 파악하였으며, 수집된 데이터 중 2019년 4월 5일 12시의 데이터를 활용하여 Figure 3과 같이 등급지도를 제작하였다. PM_{10}의 경우 해당 분포 경향이 나쁨과 매우 나쁨으로만 나타나는 것에 비해 실제 모델 분류 결과에서는 매우 나쁨 등급의 분류가 이루어지지 않았으며 서울 내부 중심 지역에 한해서 피해등급이 감소된 것으로 나타났다. $PM_{2.5}$의 경우에도 보통과 나쁨으로 등급이 구분되어 있으나 모델 분류 결과에는 PM_{10}과 같은 등급 분류 결과가 나타났다. 예측된 미세먼지는 PM_{10}과 $PM_{2.5}$를 동일한 등급으로 구분한 모델을 통해 결과가 제시되었으며, 향후 이러한 구분 모델은 대기오염 특성물질을 기반으로 PM_{10}과 $PM_{2.5}$ 데이터를 분류할 수 있을 것으로 판단된다.

Table 7. Match rate by test data grade.

concordance rate(%)		good	normal	bad	very bad
1 fine	PM_{10}	2.60	91.20	29.57	34.09
	$PM_{2.5}$	49.80	69.52	50.56	40.18
2 fine dust		61.94	48.78	40.16	55.34

5. 결론

5.1. 결론 및 제언

본 연구는 시민들에게 제공되고 있는 미세먼지 데이터를 기초로 시-공간적인 미세먼지 등급지도를 작성하고, 시민들에게 예측된 미세먼지 등급존(Zoon)을 제시하였다. 또한 미세먼지와 같은 지점에서 측정되는 대기오염물질을 기반으로 상관성을 분석 후 이러한 상관성을 기반으로 한 미세먼지 등급분류를 기계학습을 통해 진행하였다.

그 결과 첫째, 미세먼지와 초미세먼지 모두 대기오염물질 NO_2와 높은 상관성을 보여주는 것으로 나타났으며, 그중 초미세먼지와 더 높은 상관성을 나타냈다. 또한 일산화탄소는 미세먼지와 낮은 음의 상관성을 보여주었다.

둘째, Adaboost 기반 기계학습을 통해 미세먼지 데이터를 분류한 결과에서 보통 분류 등급의 데이터를 가장 잘 분류하는 것으로 나타났으며 초미세먼지의 경우 모든 데이터에 대한 분류가 다른 오분류 결과보다 많이 나타나는 것으로 확인되었다.

위와 같은 결과를 기반으로 미세먼지 등급화는 매우 중요한데 이는 지방자치단체에서 제공하는 미세먼지 등급에 대한 다양한 정보제공의 방법 개선과 등급 산정의 방법을 효과적으로 적용하는데 기여할 것으로 판단된다.

5.2. 연구의 한계 및 시사점

본 연구결과는 지방자치단체와 공공기관에서 제공하는 미세먼지등급을 분류하는 과정에서 기계학습을 접목할 수 있는 가능성을 확인하였다. 다만 기계학습을 통한 결과에서도 테스트 결과에서 정확도 자체가 높게 상승하지 않는 것을 확인하였으며, 학습된 모델 자체의 과적합이 발생할 가능성이 존재하였다. 또한 학습에 사용된 변수들이 같은 측정소에서 취득된 대기오염물질이지만, 미세먼지는 다양한 대기오염물질과의 상관성뿐만 아니라, 교통량, 측정지역의 고도, 건물의 밀집 정도 등 다양한 공간정보의 변수를 고려해야 할 필요성이 존재하지만, 연구 과정에서 사용된 변수는 대기오염물질에 한정되어 있어 분류된 결과가 미세먼지 등급에 대한 정확한 등급을 제시할 수 없는 것이 본 연구의 한계점으로 판단된다.

따라서 향후 연구 과제에서는 미세먼지와 상관성을 가지는 다양한 공간정보의 학습 변수를 활용해야 할 필요성이 파악되었으며, 다양한 공간정보의 지역정보를 포함한 모델적용 변수 데이터를 수집하고, 미세먼지와의 상관성 분석을 통해 높은 상관관계를 가지는 공간정보 데이터를 기반으로 새로운 모델을 구축해야 할 것을 향후 연구로 제시할 수 있으며, 대기오염물질 배출원에 기초한 입력자료의 생산 필요성을 파악하였다.

사사

이 논문은 2021년도 남서울대학교 학술연구비 지원에 의해 연구되었음.

참고문헌

References

환경부. 2020. 환경통계연감 .

Ministry of Environment. 2020. Environmental Statistics Yearbook.

강민성 외 4인. 2016. 한반도 주요 대도시의 PM10 농도 특성 및 배출량과의 상관성 분석, 한국환경과학학회. 25(8):1065－1076.

Kang MS et al. 2016. Analysis of the correlation between PM10 concentration characteristics and emissions in major metropolitan cities on the Korean Peninsula, Korean Society of Environmental Sciences. 25(8):1065-1076.

김윤기. 2019. 기계학습 알고리즘을 이용한 주택 모기지 금리에 대한 시민들의 감정예측. 지적과국토정보 49(1): 65-84.

Kim YK. 2019. Prediction of Citizens' emotions on home mortgage rates using machine learning algorithms. *Journal of cadastre & land InfomatiX*. 49(1): 65-84.

김운수. 2014. 서울시 초미세먼지($PM_{2.5}$) 관리방안, 서울연구원. 정책리포트 2014-182.

Kim WS. 2014. Seoul Metropolitan Government's ultra-fine dust ($PM_{2.5}$) management plan, Seoul Institute. 2014-182.

성선용. 2019. 미세먼지 농도의 시·공간적 분포 현황 및 잠재영향인자 고찰, 국토연구원. WP 19-04.

Sung SY . 2019. The status of temporal and spatial distribution of fine dust concentrations and the potential influencing factors are reviewed. National Research Institute of Human Settlements. WP 19-04.

유재환. 2011. 서울시 대기오염측정소의 중복성 분석, 서경대학교 대학원 학위논문.

Yoo JH. 2011. The redundancy analysis of the air pollution monitoring station in Seoul. Graduate School thesis of Seokyung University.

이기혁. 2020. 복합 신경망 구조를 이용한 미세먼지 위험단계 예측 모델 설계 및 분석, 한양대학교 대학원 학위논문.

Lee GH. 2020. Design and analysis of a model for predicting the risk level of fine dust using a complex neural network structure. Graduate degree thesis of Hanyang University.

이승복. 2019. 미세먼지가 인체에 미치는 영향에 관한 연구동향, BRIC View 2019-T26

Lee SB.2019. Research trends on the effects of fine dust on the human body. BRIC View 2019-T26.

정종철. 2014. 서울시 PM10 공간분포 분석과 시계열 변화, 한국지리정보학회지. 17(1):61-69.

Jung JC. 2014. Seoul PM10 Spatial Distribution Analysis and Time Series Changes, *Journal of the Korean Geographic Information Society*. 17(1):61-69.

정종철. 2017. 서울시 미세먼지 관측망 위치 적정성 평가를 위한 공간정보 활용방안. 지적과국토정보 47(2): 175-184.

Jung JC. 2017. Spatial information application case for appropriate location assessment of PM10 observation network in Seoul city. *Journal of cadastre & land InfomatiX*. 47(2): 175-184.

정종철. 2019. 커널분석을 활용한 미세먼지 신규측정소 선정 -서울시 초등학교를 대상으로. 지적과국토정보 49(2): 83-92.

Jung JC. 2019. Selection of new particulate matter monitoring stations using Kernel analysis - Elementary schools, Seoul, Korea. J*ournal of cadastre & land InfomatiX*. 47(2): 175-184.

조경우. 2019. 미세먼지 예측을 위한 기계학습 알고리즘의 적합성 평가. 한국정보통신학회논문지. 23(1): 20-26.

Jo KW. 2019. Evaluation of the suitability of machine learning algorithms for predicting fine dust. *Paper of the Korean Society of Information and Communication*. 23(1):20-26.

최임조 외 2명. 2016. 다변량분석법을 활용한 수도권 지역의 대기오염측정망 평가, 한국환경과학회지. 25(5):673-681.

Choi IJ et al.2016.Evaluation of Air Pollution Measurement Network in the Seoul metropolitan area using multivariate analysis method. *Korean Society of Environmental Sciences*. 25(5):673-681.

Rochelle Schneider dos Santos et al. 2020. A satellite-based spatio-temporal machine learning model to reconstruct daily $PM_{2.5}$ concentrations across Great Britain, medRxiv 2020.07.19.20157396.

Jan Kleine Deters et al. 2017. Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters, Journal of Electrical and Computer Engineering, Volume 2017, Article ID 5106045.

Hamed Karimian et al. 2019, Evaluation of Different Machine Learning Approaches to Forecasting $PM_{2.5}$ Mass Concentrations. *Aerosol and Air Quality Research*, 19: 1400 – 1410.

Guang Yang et al. 2020, A Hybrid Deep Learning Model to Forecast Particulate Matter Concentration Levels in Seoul, South Korea, Atomsphere, 11, 348; doi:10.3390/atmos11040348.

Lary D.J. et al. 2015, Using Machine Learning to Estimate Global $PM_{2.5}$ for Environmental Health Studies. *Environmental Health Insights* 2015.9(S1):41-52.

2021년 09월 24일 원고접수(Received)
2021년 10월 28일 1차심사(1st Reviewed)
2021년 11월 05일 2차심사(2nd Reviewed)
2021년 11월 25일 게재확정(Accepted)

초 록

미세먼지는 사람의 건강에 많은 영향을 미치는 물질로서 이와 관련하여 다양한 연구가 이루어지고 있다. 미세먼지의 인체 영향으로 인해 서울시 모니터링 네트워크에서 측정된 과거 데이터를 활용하여 미세먼지를 예측하려는 다양한 연구가 진행되고 있다.

본 연구는 2019년 5월 서울시의 미세먼지를 중점으로 진행하였으며, 학습에 사용한 변수는 SO_2, CO, NO_2, O_3와 같은 대기오염물질 데이터를 활용하였다. 예측모델은 Adaboost에 기반하여 구축하였고, 훈련모델은 PM_{10}과 $PM_{2.5}$로 구분하였다. 에러 메트릭스를 통한 예측모델의 정확도 평가 결과로 Adaboost가 시도되었다. 대기오염물질은 초미세먼지와 더 높은 상관성을 보이는 것으로 나타났지만, 보다 효과적인 분포등급을 제시하기 위해서는 많은 양의 데이터를 학습하고, PM_{10}과 $PM_{2.5}$의 공간분포 등급을 효과적으로 예측하기 위해서 교통량 등의 추가적인 변수를 활용할 필요성이 있다고 판단된다.

주요어 : 미세먼지, 기계학습, Adaboost, 대기오염물질