기상데이터와 머신러닝을 활용한 미세먼지농도 예측 모델

임준묵, 고선호, 김제완*
*한밭대학교 창의융합학과

An estimation model of fine dust concentration using weather data and machine learning

Lim, Joon-Mook, Ko, Sunho, and Kim, Jewan*
Hanbat National University
E-mail : jmlim@hanbat.ac.kr

요 약

본 연구에서는 에어코리아의 미세먼지 농도 측정치와 기상자료개방포털에서 실시간으로 제공하는 기상관련 다양한 정보를 활용하여 미세먼지의 농도를 예측할 수 있는 수리적 모델을 개발하였다. 수리모델에서는 다양한 국내 계절성 변수들과 대기상태 변수들을 다중회귀분석을 통해 미세먼지농도에 유의한 영향을 미치는 변수를 추출하고 그 변수들을 토대로 머신러닝기법인 ANN(Artificial Neural Network)과 SVM(Support Vector Machine)을 사용하여 미세먼지농도를 예측할 수 있는 모형을 제안하였다. 제안모형은 과거의 미세먼지 및 기상데이터를 활용하여 그 효과성을 검증할 수 있었다.

1. 서론

최근 미세먼지 수치가 급격히 상승함에 따라 사람들의 관심이 나날이 높아지고 있다. 미세먼지의 노출은 호흡기만이 아니라 심혈관계 질병의 발생에도 영향을 끼치며, 심하게는 사망률도 증가하는 것으로 연구되고 있다.[1] 미세먼지를 자체를 없애는 것은 사실상 불가능한 일이다. 하지만 미세먼지 농도를 예측하여, 이에 대한 노출을 최소화하는 것이 가장 좋은 방법일 것이다.

기상자료개방포털(https://data.kma.go.kr/)에서 제공하는 기상자료 데이터는 종관기상관측장비(ASOS : Automated Synoptic Observing System)에 의한 자동관측과 목측에 의한 수동관측으로 실시된다. 현재 전국에 94소를 운영하고 있고 지상 부근의 대기상태를 실시간으로 관측하기 위한 기본 장비로서 기온, 습도, 풍향, 풍속, 기압, 강수량, 일조, 일사, 지면온도, 초상온도, 지중온도를 매분 관측한다. 수동관측요소는 적설, 구름, 기타 일기현상 등이며, 실시간, 매 정시 또는 3시간 간격으로 관측한다.

한편 한국환경공단의 에어코리아(http://www.airkorea.or.kr/)는 공기오염상태를 확인할 수 있는 전국 실시간 대기오염도 공개 홈페이지로써, 전국의 대기오염 측정망에서 측정되는 아황산가스, 일산화탄소, 이산화질소, 오존, 미세먼지(PM10), 초미세먼지(PM2.5)의 데이터를 실시간으로 제공하고 있다.

에어코리아는 전국 97개 시,군에 설치된 323개의 도시대기 측정망, 도로변대기 측정망, 국가배경 측정망, 교외대기 측정망에서 측정된 대기환경기준물질의 측정자료를 다양한 형태로 표출하여 국민들에

게 실시간으로 제공하고 있다.

국립환경과학원(2016)에 따르면 수도권 고농도 미세먼지 발생 원인을 분석해 본 결과 중국 등의 국외 요인보다는 국내의 기상상태에 따른 2차 미세먼지의 생성 등이 미세먼지의 농도에 크게 영향을 미치고 있는 것으로 밝혀졌다.[2]

본 연구에서는 에어코리아의 미세먼지 농도 측정치와 기상자료개방포털에서 실시간으로 제공하는 기상관련 다양한 정보를 활용하여 국내 계절성 변수들과 대기상태 변수들을 다중회귀 분석(Multiple regression analysis)을 통해 미세먼지농도에 유의한 영향을 미치는 변수를 추출하고, 그 변수들을 토대로 머신러닝기법인 ANN(Artificial Neural Network)과 SVM(Support Vector Machine)을 사용하여 미세먼지 농도를 예측할 수 있는 모형을 제안한다.

제안 모형은 과거의 미세먼지 및 기상데이터를 활용하여 그 효과성을 검증할 것이다. 본 연구에서 제안된 모델을 사용하면 관측재원이 없는 곳에서도 쉽게 획득이 가능한 기상데이터를 활용하여 미세먼지 농도를 예측할 수 있고, 이를 통해서 미세먼지의 피해를 사전에 예방할 수 있는 기회를 가질 수 있을 것이다

2. 기상과 미세먼지의 빅데이터 수집 및 처리

2.1 데이터수집

(1) 미세먼지관련 빅데이터 수집

미세먼지 관련 빅데이터는 에어코리아 홈페이지에서 수집하였으며, 우리나라의 대표적인 지역 한 군데를 지정하여 수집하였다. 지정한 장소는 서울시 종로구 종로 169 (종묘주차장 앞)이며 수집한 자료는 미세먼지(PM10), 오존(O3), 이산화질소(NO2), 일산화탄소(CO), 아황산가스(SO2)의 다섯가지 측정값이다. 여기서 미세먼지(PM10)는 (㎍/㎥)으로 측정되지만 일반적으로 다음과 같이 4가지 범주(좋음(15이하), 보통(15-35이하), 나쁨(35-75이하), 매우나쁨(75초과))로 나누어 사용한다. 수집기간은 2014-01-01 00시부터 2017년 9월 30일 00시까지의 3년 9개월간의 자료이다. [표 1]은 에어코리아에서 수집한 자료의 종류와 일시에 따른 측정값 및 단위를 보여준다.

[표 1] 미세먼지 관련 자료(에어코리아)

날짜 (년-월-일-시)	PM10 (㎍/㎥)	오 존 (ppm)	이산화질소 (ppm)	일산화탄소 (ppm)	아황산가스 (ppm)
14-01-01-00	163	0.004	0.04	0.8	0.015
14-01-01-01	152	0.004	0.04	0.9	0.012
14-01-01-02	153	0.003	0.042	0.9	0.011
14-01-01-03	159	0.003	0.043	1.0	0.012
…	…	…	…	…	…
17-09-29-21	15	0.005	0.4	0.017	0.039
17-09-29-22	17	0.006	0.4	0.016	0.041
17-09-29-23	18	0.005	0.4	0.018	0.037
17-09-30-00	15	0.00	0.4	0.017	0.036

(2) 기상관련 빅데이터 수집

[표 1]에서 미세먼지 관련 측정치가 수집된 장소와 가장 가까운 기상관측소는 서울기상관측소로 서울특별시 종로구 신문로2가 1-43에 위치해 있다. 따라서 서울기상관측소로부터 측정된 기상관련자료를 '기상자료개방포털'로부터 얻을 수 있다. 수집된 기상관측자료는 기온, 강수량, 풍속, 풍향, 습도, 증기압, 이슬점온도, 현지기압, 일조, 일사, 전운량, 중하층운량, 최저운고, 시정, 지면온도의 15가지 자료이다. 미세먼지 관련자료와 수집기간은 2014-01-01 00시부터 2017년 9월 30일 00시까지로 일치하도록 하였다. [표 2]는 기상자료개방포털에서 일시별로 수집한 자료의 종류와 측정단위를 보여준다.

[표 2] 기상관련 자료(기상자료개방포털)

날짜 (월-일-시)	기온 (° C)	강수량 (mm)	풍속 (m/s)	풍향 (16방위)	습도 (%)	증기압 (hpa)
14-01-01-00	3.3		3.8	250	65	5
14-01-01-01	2.6		2.3	250	66	4.9
14-01-01-02	1.7		1.7	250	67	4.6
14-01-01-03	1.4		1.4	250	60	4.1
…	…	…	…	…	…	…
17-09-29-21	3.3		0.7	360	41	2
17-09-29-22	2.6		0.9	270	42	2
17-09-29-23	1.7		1	290	44	2
17-09-30-00	1.4		1.1	290	53	2.4

날짜 (월-일-시)	이슬점온도 (° C)	현지기압 (hpa)	해면기압 (hpa)	일조 (hr)	일사 (MJ/m2)
14-01-01-00	-2.6	1001.9	1012.5	0.528	0.977
14-01-01-01	-3.1	1002.2	1012.9	0.528	0.977
14-01-01-02	-3.7	1002.4	1013.1	0.528	0.977
14-01-01-03	-5.5	1002.5	1012.9	0.528	0.977
…	…	…	…	…	…
17-09-29-21	4.5	1008.9	1019	0.528	0.977
17-09-29-22	5.2	1009.5	1019.6	0.528	0.977
17-09-29-23	5.2	1009.8	1019.9	0.528	0.977
17-09-30-00	6.4	1009.6	1019.8	0.528	0.977

날짜 (월-일-시)	전운량 (10분위)	중하층운량 (10분위)	최저운고 (100m)	시정 (10m)	지면온도 (° C)
14-01-01-00	6	6	10	600	0
14-01-01-01	4.976	3.086	13.182	1400.137	-0.1
14-01-01-02	4.976	3.086	13.182	1400.137	-0.3
14-01-01-03	0	0	13.182	600	-0.4
…	…	…	…	…	…
17-09-29-21	4	0	13.182	2000	15.1
17-09-29-22	4.976	0	13.182	2000	14.6
17-09-29-23	4.976	0	13.182	2000	14.1
17-09-30-00	0	0	13.182	2000	13.5

2.2 데이터 전처리

미세먼지관련 빅데이터와 기상관련 빅데이터는

서로 다른 테이블에 존재하는데, 우선 공통되는 날짜(월-일-시)를 키(key)로하여 하나로 결합하였고, 미세먼지(PM10)을 종속변수로 하고 나머지 미세먼지관련자료(4), 기상자료(16), 날짜(월-시)(2)의 총 22개의 측정값을 독립변수로 설정하였다. 수집한 데이터에는 결측값이 존재하는데, 강수량의 결측값은 0으로 일조, 일사를 비롯한 변수들의 결측값은 평균값으로 대치했다. 년-월-일-시에 따라 수집된 총 자료의 수는 32,784개의 자료이다. 이 중에서 임의로 30,000개의 자료를 추출하여 훈련을 위한 자료로 사용하였고, 나머지 2,784개의 자료는 예측모형의 정확도를 평가하기 위한 시험자료로 사용하였다.

3. 머신러닝기법을 활용한 미세먼지의 예측모형

3.1 다중회귀분석

다중 회귀 분석은 변수 간의 인과 관계를 통계적 방법에 의해 추정하는 회귀 분석의 일종이다. 다중 회귀 분석의 기본적인 목표는 다음과 같은 다중 회귀식에서 상수 및 계수를 구하는 것이다.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \cdots \beta_i x_i + \varepsilon_i$$

(x: 독립변수, y: 종속변수, β: 회귀계수, β_0: Y절편, $\beta_1 \sim \beta_i$: 독립변수의 기울기)

주어진 자료를 사용하여 다중회귀분석을 수행한 결과를 요약하면 [표 3]과 같다.

[표 3] 다중회귀 분석결과

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	347.436	34.214	10.155	0.000
O3	118.158	11.753	10.054	0.000
NO2	7.352	1.481	4.963	0.000
CO	27.593	0.668	41.325	0.000
SO2	399.908	21.470	18.626	0.000
기온	-2.197	0.171	-12.847	0.000
강수량	-0.914	0.149	-6.132	0.000
풍속	0.771	0.129	6.000	0.000
풍향	0.031	0.001	20.899	0.000
습도	-1.129	0.039	-28.797	0.000
증기압	-1.073	0.060	-17.908	0.000
현지기압	**3.779**	**2.156**	**1.753**	**0.080**
해면기압	**-3.914**	**2.135**	**-1.834**	**0.067**
이슬점온도	3.305	0.149	22.222	0.000
일조	-7.327	0.757	-9.674	0.000
일사	2.625	0.468	5.613	0.000
전운량	-0.873	0.068	-12.857	0.000
중하층운량	-0.458	0.073	-6.267	0.000
최저운고	0.154	0.022	6.873	0.000
시정	-0.029	0.000	-82.226	0.000
지면온도	-0.367	0.042	-8.735	0.000
달(월)	-1.330	0.051	-26.001	0.000
시	**-0.020**	**0.024**	**-0.836**	**0.403**

[표 3]으로부터 유의수준 5%에서 유의하지 않은 현지기압, 해면기압, 시의 독립변수 3개를 1차적으로 제거하였다. 제거 후의 독립변수들 간의 다중공선성을 측정한 결과는 [표 4]와 같다.

[표 4] 다중공선성 분석결과

O3	NO2	CO	SO2	기온
1.62	6.73	2.32	5.87	111.36
강수량	**풍속**	**풍향**	**습도**	**증기압**
1.02	1.32	1.18	28.29	11.15
이슬점온도	**일조**	**일사**	**전운량**	**중하층운량**
150.61	3.02	3.85	2.65	2.42
최저운고	**시정**	**지면온도**	**달**	
1.14	1.95	14.60	1.30	

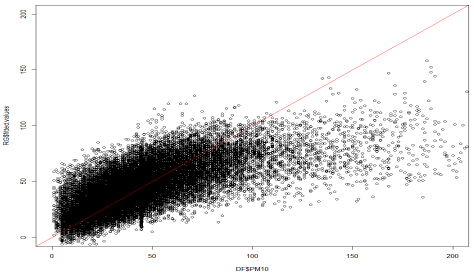
일반적으로, 다중공선성 값이 5이상이면 위험, 10이상이면 매우위험이다. 따라서 다중공선성 결과 값이 10을 넘는 기온, 습도, 증기압온도, 이슬점온도, 지면온도의 5개변수를 2차적으로 제거하였고, 5를 넘는 NO2와 SO2는 피어슨의 상관계수값이 0.9를 넘는 것으로 조사되어 다중공선성값이 상대적으로 높은 NO2를 제거하였다.

이제 최종적으로 남은 13개의 독립변수(O3, CO, SO2, 강수량, 풍속, 풍향, 일조, 일사, 전운량, 중하층운량, 최저운고, 시정, 달(월))를 대상으로 종속변수(PM10)에 대한 다중회귀분석을 수행하였다.

최종적으로 얻어진 회귀식은 다음과 같다.

$$\begin{aligned} PM10 = {} & 5.386 + 200.13O_3 + 31.21CO + 642.5SO_2 \\ & - 2.25\text{강수량} + 1.513\text{풍속} + 0.034\text{풍향} \\ & - 3.508\text{일조} + 2.426\text{일사} - 0.73\text{전운량} \\ & - 0.98\text{중하층운량} + 0.333\text{최저운고} \\ & - 0.023\text{시정} - 1.5\text{달} \end{aligned}$$

회귀식에 의한 추정값과 실제값의 분포도를 그래프로 나타내면 [그림 1]과 같다.



[그림 1] 회귀식에 따른 추정값과 실제값 분포도

회귀식의 정확도를 평가하기 위해 MAPE 값을 계산한 결과 47.24593를 얻었다. MAPE값이 50이하로 매우 합리적인 예측이 되고 있음을 알 수 있다. 또한 회귀식의 예측 정확도를 측정하기 위해서 2,784개의 시험자료를

바탕으로 예측을 수행한 결과 [표 5]와 같은 결과를 얻었다.

[표 5] 다중회귀모형에 의한 미세먼지 예측 정확도

실제 / 예측	매우 나쁨	나쁨	보통	좋음
매우 나쁨	119	88	2	0
나쁨	178	936	430	61
보통	4	210	513	113
좋음	0	7	54	69

[표 5]에서 보는 바와 같이 실제값이 '매우나쁨' 또는 '나쁨'인 경우 예측값도 '매우나쁨' 또는 '나쁨'으로 예측한 경우는 제대로 된 예측이라고 판단하고, '보통' 또는 '좋음'의 경우도 같은 방법으로 고려할 때, 74.35%의 정확도를 보이는 것으로 판단된다.

3.2 SVM에 의한 미세먼지 예측

본 절에서는 미세먼지의 예측을 위해서 SVM기법을 사용하였다. 학습을 위한 자료는 다중회귀분석모형에서 사용하였던 자료를 그대로 사용하였다. 단, 자료의 값들 간에 차이가 크므로 값을 0-1사이의 값으로 표준화시켜 사용하였다.

또한 SVM 예측모형의 결과에 영향을 미치는 kernel은 radial을 사용하였으며, gamma와 cost의 값은 사전 튜닝(tune)을 수행하여 얻은 0.5와 2를 사용하였다. 2,784개의 시험자료에 대한 SVM에 의한 미세먼지 예측결과는 다음과 같다. [표 6]으로부터 예측 정확도는 80.35%이다.

[표 6] SVM에 의한 미세먼지 예측

실제 / 예측	매우 나쁨	나쁨	보통	좋음
매우 나쁨	100	20	4	0
나쁨	193	980	269	25
보통	7	235	695	148
좋음	1	6	31	70

3.3 ANN에 의한 미세먼지 예측

본 절에서는 미세먼지의 예측을 위해서 ANN기법을 사용하였다. 학습을 위한 자료는 다중회귀분석모형에서 사용하였던 자료를 그대로 사용하였으며, SVM에서와 마찬가지로 0-1사이의 값으로 표준화시켜 사용하였다. 단, 예측변수는 미세먼지(PM10)은 범주화값을 사용하였다.

예비 실험결과를 바탕으로 중간층(hidden)의 노드수는 7개를 사용하였다. ANN을 활용한 미세먼지 예측모형의 예측결과는 85.1%임이 확인되었다.

4. 결론

본 연구에서는 미세먼지의 예측을 위해서 미세먼지와 상관관계가 매우 높을 것으로 예상되는 기상자료를 활용하였다.

예측을 위한 모형으로는 다중회귀분석, 서포트벡터머신(SVM), 인공신경망(ANN)을 사용하였다. 각 예측모형에 대해서 시험자료를 바탕으로 예측을 수행한 결과 예측정확도로 다중회귀분석:SVM:ANN = 74.35%:80.35%:85.1%를 얻어, 기상데이터를 활용한 미세먼지의 예측은 ANN을 사용한 모델이 상대적으로 정확한 것으로 판명되었다.

본 연구에서 제시한 미세먼지 예측모형을 사용할 경우, 손쉽게 획득할 수 있는 기상예보 데이터를 활용하여 미래의 미세먼지의 정도를 비교적 정확히 예측할 수 있음을 알 수 있다.

[참고문헌]

[1] Shin, Eunkyung, Kim, Jaebum, and Choi, Yongrak, "A Study on the Data Model Design of Fine Dust Related Disease), " Jounal of The Korea Society of Information Technology Policy & Management, Vol.10, No.1, pp.655-659, 2018.

[2] 이미혜, "한 중 월경성 미세먼지 저감을 위한 공동연구(II)(Korea-China collaborative study to abate trans-boundary air pollution(II)),"국립환경과학원 연구용역보고서, 2016.6.

[3] 박진수 등, 미세먼지 2차 생성량 추정에 관한 연구(I). 국립환경과학원, 2017.