석 사 학 위 논 문

머신러닝을 활용한 미세먼지농도예측모델에 대한 연구

A Study on Prediction Model of Fine Dust Using Machine Learning

고려대학교 컴퓨터정보통신대학원
빅데이터융합학 전공
서동규

2019년 6월

주 재 걸 교 수 지 도
석 사 학 위 논 문

머신러닝을 활용한 미세먼지농도예측모델에 대한 연구

A Study on Prediction Model of Fine Dust Using Machine Learning

이 논문을 공학 석사학위 논문으로 제출함.

2019년 6월

고려대학교 컴퓨터정보통신대학원
빅데이터융합학 전공
서동규

서동규의 공학 석사학위논문 심사를 완료함.

2019년 6월

위원장 주재걸 (인)

위 원 강재우 (인)

위 원 김현철 (인)

목 차

그 림 목 차

표 목 차

1. 서론

세계보건기구(WHO)에서 1급 발암물질로 분류한 미세먼지의 고농도 현상이 잦아지는 등 국민건강과 생명을 직접적으로 위협하는 미세먼지는 이제 온 국민의 관심사이자 국가적 재난의 문제로 대두되고 있다. 정부에서도 최근 2019년 3월 26일 미세먼지 저감 및 관리에 관한 특별법(약칭 : 미세먼지법)을 제정 및 시행하는 등 이의 해결을 위한 다각도의 노력을 진행하고 있다.

미세먼지 저감을 위해 주요요인 파악이 요구되며, 이는 크게 국외요인과 국내요인으로 분류할 수 있다. 주요 국외요인으로는 중국 대기오염의 국내유입을 들 수 있고, 주요 국내요인으로는 산업활동에 따른 대기오염물질 배출을 들 수 있다.

국외요인에 해당하는 중국 대기오염은 외교적 이해관계 및 국외 기상 데이터 부족으로 인해 실증이 어려운 상황이며, 최근에서야 한중 양국의 공동연구 및 공동대응방안 협의가 계획되고 있다[1].

국내요인 중 산업활동에 따른 대기오염물질 배출량은 최근 환경부에서 공개한 굴뚝자동측정기기(TMS) 부착 사업장의 2018년도 대기오염물질 연간배출량 자료를 통해 확인할 수 있다[2]. 이에 따르면, 발전업이 14만 5,467톤(44%), 시멘트 제조업이 6만 7,104톤(20%), 제철 제강업이 6만 3,384톤(19%), 석유화학제품업이 3만 5,299톤(11%), 기타 업종이 1만 8,791톤(6%)으로, 발전업이 제일 큰 비중을 차지하고 있음을 알 수 있다.

외교를 통한 대응과 국내 산업활동에서의 대기오염물질 저감 강제 등 다각도의 노력에도 불구하고, 단기 내에 획기적인 개선이 어려운 것이 사실인 바, 미세

먼지농도에 대한 노출을 최소화하는 것이 국민건강을 지키기 위한 최선의 방법이라고 사료된다.

미세먼지에 대한 국민의 불안은 계속 높아지고 있으나, 정부의 미세먼지 측정장비는 전국 390여개 지점 설치에 그치고 있어, 측정장비의 절대수가 부족한 바, 일반에 공개되는 실시간대기오염도와 실제생활에서 체감하는 공기질이 차이가 나는 경우가 빈번하다. 현실적으로, 미세먼지에 대한 예측을 통해 이에 대한 노출을 최소화하는 것이 중요하다고 판단된다.

4차 산업혁명시대에 이른 지금, 빅데이터에 대한 분석 및 머신러닝을 통한 예측기법이 발달하고, 이를 활용한 다양한 주제의 연구가 이뤄지고 있는 바, 본 연구에서는 머신러닝을 활용한 미세먼지 예측모델을 제안하고자 한다.

2. 기존연구에 대한 고찰

2.1 수치예측모델[3]

기준지를 서울 종로인근으로, 기간을 '14.1월부터 '17.9월까지로 설정하고, 한국환경공단에서 제공하는 미세먼지 농도측정치, 대기환경기준물질 측정치와, 기상청에서 제공하는 기상데이터를 활용하여 다중회귀분석, 인공신경망(ANN), 서포트벡터머신(SVM) 기법을 통해 미세먼지 예측모델을 생성하였다.

수집한 13개의 변수(O3, CO, SO2, 강수량, 풍속, 풍향, 일조, 일사, 전운량, 중하층운량, 최저운고, 시정, 달(월))를 독립변수로 설정하고, 미세먼지(PM10)를 종속변수로 설정하여, 다중회귀분석, ANN, SVM을 통해 예측모델을 생성하였다.

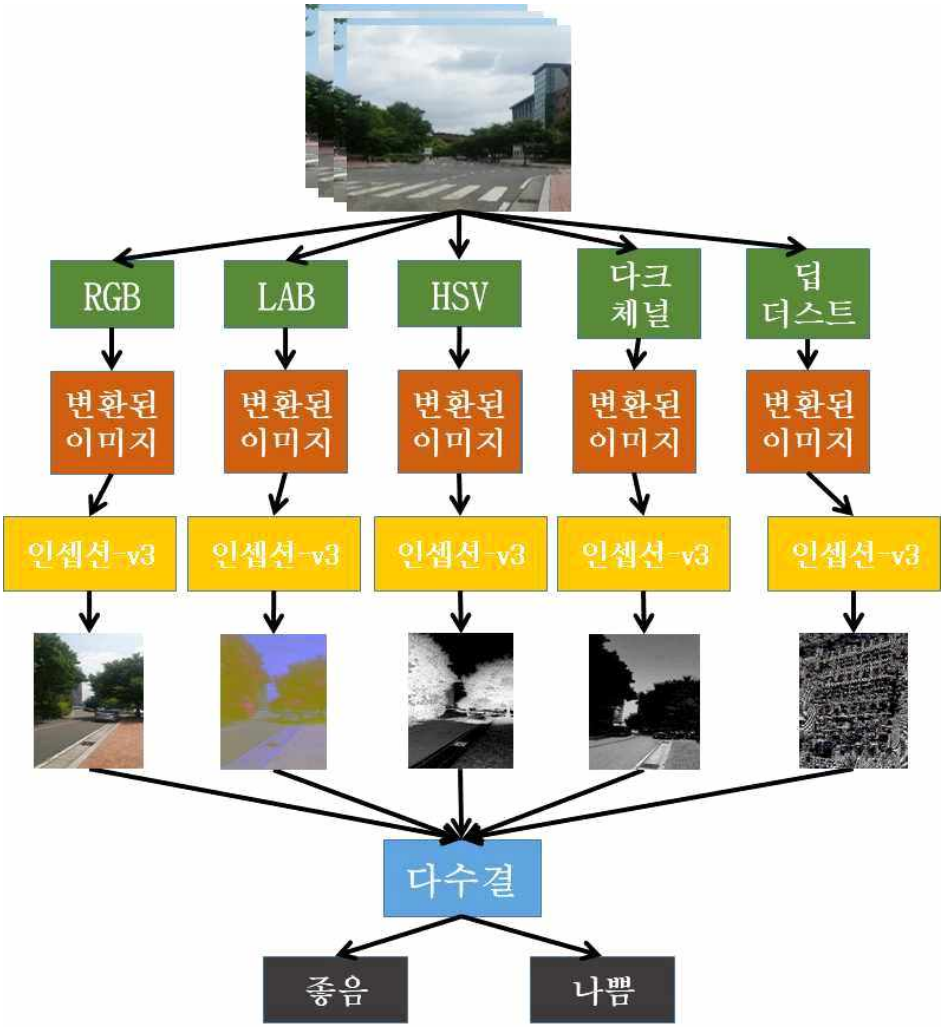
그 결과, 각 예측모델의 정확도는 다중회귀분석 : SVM : ANN = 74.35% : 80.35% : 85.1%로 비교적 높은 정확도를 보였다. 2차적인 미세먼지 발생의 원인이 되는 O3, SO2를 독립변수에 포함시킨 것이 주 원인[4]으로 사료된다.

2.2 시각데이터활용모델[5]

대구 계명대학교의 캠퍼스와 인근지역에서 '17.6월부터 8월까지, 총 3개월간 수집된 681개의 비디오 데이터를 생성했다. N초동안 촬영된 1080x1920픽셀을 가지는 단일비디오에서 시간 순으로 비디오 시퀀스를 K개(K=평균 초당 25프레임) 추출하였고, 추출된 시퀀스는 3개의 RGB값을 가지기 때문에 하나의 시퀀스는 1080x1920x3의 데이터배열을 가지게 된다.

추출된 비디오 시퀀스를 다수의 기법을 이용해 변환시키고, 변환된 이미지들 각각을 인셉션을 이용해 분석 후 도출된 예측값을 다수결(majority vote)을 통해

결합하여 예측하는 앙상블 모형을 사용하여 미세먼지의 좋음, 나쁨을 예측하였다.



(그림 1) 시각데이터활용모델

3. 연구계획

3.1 연구개요

기간을 2016년 1월부터 2019년 3월까지로 설정하고, 한국환경공단에서 제공하는 초미세먼지 농도측정치와, 기상청에서 제공하는 기상데이터를 활용하여 다항로지스틱회귀분석, 최근접이웃(KNN), 서포트벡터머신(SVM) 기법을 통해 미세먼지 예측모델을 생성하고자 한다.

초미세먼지 농도측정치 데이터수집 대상장소는 서울특별시 중구 덕수궁길 15 시청서소문별관 3동(이하 서울로 기재)과, 인천광역시 옹진군 백령면 연화리 산 241-2(이하 백령도로 기재)을 선정하였으며, 기상데이터수집 대상 장소는 각 지역별 최인접 장소인 서울특별시 종로구 신문로2가 1-43 서울기상관측소(이하 서울로 기재)와 인천광역시 옹진군 백령면 진촌리 1031(이하 백령도로 기재)로 선정하였다.

3.2 백령도 선정 의의

백령도는 우리나라 최북단에 위치한 섬으로, 산업시설이 전무하여 미세먼지 발생인자가 없음에도 불구하고 미세먼지가 측정되고 있는데, 이는 외부 지역에서의 미세먼지 유입이 원인인 바, 중국 동부 산업도시 옌타이, 칭다오 등에서의 서풍을 통한 유입이 큰 원인으로 보여진다. 본 연구에서는 서울 초미세먼지 예측모델에 있어, 백령도의 초미세먼지와 풍향이 주는 영향을 알아보고, 국내 초미세먼지 발생원인 중 국외요인의 유의성을 확인하고자 한다.

(그림 2) 백령도 지리적 위치

3.3 연구계획

첫 번째 연구계획은, 독립변수로 서울 기상데이터를, 종속변수로 서울 초미세먼지 농도측정치데이터로 설정하여 머신러닝기법을 활용하여 예측모델을 생성한 후, 해당 예측모델의 정확도를 도출하는 것이다.

두 번째 연구계획은, 독립변수로 서울 기상데이터, 백령도 기상데이터 및 백령도 초미세먼지 농도측정치데이터를, 종속변수로 서울 초미세먼지 농도측정치데이터를 설정하여 머신러닝기법을 활용하여 예측모델을 생성한 후, 해당 예측모델의 정확도를 도출하는 것이다.

첫 번째와 두 번째 방법을 통해 생성된 모델 간 정확도를 비교해보고, 백령도의 기상데이터 및 초미세먼지 농도측정치데이터가 예측모델에 주는 영향을 알아보기로 한다.

	예측모델1	예측모델2
독립변수	서울 기상데이터	서울 기상데이터 백령도 기상데이터 백령도 초미세먼지농도
종속변수	서울 초미세먼지농도	서울 초미세먼지농도
머신러닝기법 (R프로그램)	다항 로지스틱회귀분석, 최근접이웃(KNN), 서포트벡터머신(SVM)	

[표 1] 연구계획

4. 연구수행

4.1 데이터수집 및 가공

한국환경공단에서 제공하는 미세먼지 농도측정치데이터를 통해 2016년 1월부터 2019년 3월까지 3년 3개월 간 서울과 백령도의 일별 초미세먼지(PM2.5) 데이터를 수집하였다.

또한, 기상청 제공자료를 통해 동일기간 서울의 평균기온, 최저기온, 최고기온, 일강수량, 평균풍속, 최다풍향, 평균이슬점온도, 평균상대습도, 평균현지기압, 가조시간, 합계일조시간, 평균전운량, 평균중하층운량, 평균지면온도 및 백령도의 최다풍향 데이터를 수집하였다.

초미세먼지데이터는 종속변수에 해당하므로, 결측치는 동일자의 전체 변수를 제거하였으며, 기상데이터 결측치는 전후 일자의 동일변수 해당 값의 평균값을 대입하였다. 그 결과, 수집된 자료는 1,143일간 17개 변수, 즉 19,431개의 데이터로 구성되었다.

초미세먼지 데이터를 농도별로 0~15(좋음) / 16~35(보통) / 36~75(나쁨) / 76~(매우 나쁨) (단위:㎍/㎥)으로 범주화하였다. 서울 기상데이터에 해당하는 14개 변수들을 대상으로 다중공선성분석을 수행하였으며, 그 결과값이 10 이상인 변수는 평균기온, 최저기온, 최고기온, 평균이슬점온도, 평균상대습도, 평균지면온도 인 것으로 확인되었다. 이 중 기온값을 가지는 평균기온, 최저기온, 최고기온, 평균이슬점온도, 평균지면온도 중 평균기온을 제외한 나머지 4개 변수들를 제거한 후, 남은 변수들을 대상으로 다시 다중공선성분석을 수행한 결과, 모든 변수가 10 이하의 결과값을 나타내었다.

평균기온	최저기온	최고기온	일강수량	평균풍속
748.49	116.90	158.11	1.50	1.45
최다풍향	평균이슬점온도	평균상대습도	평균현지기압	가조시간
1.33	480.55	51.33	4.37	6.39
합계일조시간	평균전운량	평균중하층운량	평균지면온도	
6.15	5.5	4.52	57.83	

[표 2] 변수제거 전 다중공선성 분석결과

평균기온	일강수량	평균풍속	최다풍향	평균상대습도
6.35	1.35	1.33	1.29	2.77
평균현지기압	가조시간	합계일조시간	평균전운량	평균중하층운량
4.35	5.18	4.95	5.45	4.41

[표 3] 변수제거 후 다중공선성 분석결과

최종적으로 1,143일간 서울 기상데이터 10개 변수, 백령도 기상데이터 1개 변수, 서울과 백령도 초미세먼지 2개 변수인 총 13개 변수, 즉 14,859개의 데이터를 연구수행에 활용하게 되었다. 연구수행을 위해 시계열 배열로 상위 915일간 데이터를 훈련자료로, 하위 228일간 데이터를 시험자료로 분할구성하였으며, 각 연구수행은 R프로그램을 활용하여 진행하였다.

4.2 다항 로지스틱회귀분석

종속변수가 4개이므로, 다항 로지스틱회귀분석을 수행하였다. 훈련자료를 통해 초미세먼지 예측모델을 생성 후, 시험자료를 통해 예측모델의 정확도를 도출하였다. 연구계획대로 두 가지의 모델을 생성하여 수행 결과, 예측모델1은 48.2%, 예측모델2는 62.7%의 정확도를 보여, 예측모델2가 우수한 정확도를 나타내었다.

실제 / 예측	좋음	보통	나쁨	매우 나쁨
좋음	3	0	0	0
보통	73	106	37	4
나쁨	0	4	1	0
매우 나쁨	0	0	0	0

[표 4] 다항로지스틱회귀분석 예측모델1

실제 / 예측	좋음	보통	나쁨	매우 나쁨
좋음	27	4	0	0
보통	48	94	17	0
나쁨	1	11	21	3
매우 나쁨	0	1	0	1

[표 5] 다항로지스틱회귀분석 예측모델2

4.3 최근접이웃(KNN)

최근접이웃(KNN) 기법을 활용하기 위해 각 독립변수의 데이터값을 0과 1 사이의 값으로 표준화하였으며, K값은 30을 설정하였다. 훈련자료를 통해 초미세먼지 예측모델을 생성 후, 시험자료를 통해 예측모델의 정확도를 도출하였다. 연구계획대로 두 가지의 모델을 생성하여 수행 결과, 예측모델1은 46.9%, 예측모델2는 51.7%의 정확도를 보여, 예측모델2가 우수한 정확도를 나타내었다.

실제 / 예측	좋음	보통	나쁨	매우 나쁨
좋음	2	0	0	0
보통	74	100	33	3
나쁨	0	10	5	1
매우 나쁨	0	0	0	0

[표 6] 최근접이웃 예측모델 1

실제 / 예측	좋음	보통	나쁨	매우 나쁨
좋음	6	0	0	0
보통	70	108	34	0
나쁨	0	2	4	4
매우 나쁨	0	0	0	0

[표 7] 최근접이웃 예측모델 2

4.4 서포트벡터머신(SVM)

서포트벡터머신(SVM) 기법의 활용 결과에 영향을 미치는 kernel은 rbfdot를 사용하였다. 훈련자료를 통해 초미세먼지 예측모델을 생성 후, 시험자료를 통해 예측모델의 정확도를 도출하였다. 연구계획대로 두 가지의 모델을 생성하여 수행 결과, 예측모델1은 50.4%, 예측모델2는 60.5%의 정확도를 보여, 예측모델2가 우수한 정확도를 나타내었다.

실제 / 예측	좋음	보통	나쁨	매우 나쁨
좋음	5	0	0	0
보통	71	107	35	4
나쁨	0	3	3	0
매우 나쁨	0	0	0	0

[표 8] 서포트벡터머신 예측모델 1

실제 / 예측	좋음	보통	나쁨	매우 나쁨
좋음	16	1	1	0
보통	60	104	19	0
나쁨	0	5	18	4
매우 나쁨	0	0	0	0

[표 9] 서포트벡터머신 예측모델 2

5. 결론

세가지 머신러닝 기법을 통해 생성한 예측모델1과 예측모델2의 정확도를 비교해 본 결과, 서울 기상데이터와 백령도의 초미세먼지 및 풍향을 조합하여 생성한 예측모델2가, 서울 기상데이터만을 활용하여 생성한 예측모델1보다 더 높은 정확도를 보였다. 이를 통해, 백령도의 해당 데이터가 서울 초미세먼지를 예측함에 있어, 유의미한 변수임을 확인할 수 있었다.

또한, K값의 설정에 따라 예측모델 정확도 편차를 보이는 최근접이웃(KNN)기법에 비해, 다항 로지스틱회귀분석과 서포트벡터머신(SVM) 기법이 상대적으로 더 높은 정확도를 보였으며, 동시에 예측모델2에서 더 개선된 모습을 보였다.

머신러닝기법	예측모델1 정확도	예측모델2 정확도
다항 로지스틱회귀분석	48.2%	62.7%
최근접이웃(KNN)	46.9%	51.7%
서포트벡터머신(SVM)	50.4%	60.5%

[표 10] 연구수행 결과

일반적으로, 데이터의 양과 질이 예측모델의 정확도에 크게 기여하는 바, 더 다양하고 많은 데이터 확보를 통해 예측모델 개선이 가능할 것으로 예상된다. 향후 굴뚝자동측정기기(TMS) 부착 국내사업장의 대기오염물질 배출량이 실시간으로 공개될 예정인 바, 이를 활용한 미세먼지예측모델 개선과 국내 발생요인분석이 기대된다.

참고문헌

[1] 중국과의 공동대응 협력 및 고농도 미세먼지 긴급조치 강화(환경부 보도자료, 2019.03.07.)

[2] 2018년도 대기오염물질 연간배출량 조사자료(환경부)

[3] 기상데이터와 머신러닝을 활용한 미세먼지농도 예측모델(임준묵 등, 한국IT서비스학회 학술대회 논문집, 2018, pp 691-694)

[4] 바로 알면 보인다. 미세먼지, 도대체 뭘까?(환경부, 2016)

[5] 딥러닝에 기반한 미세먼지 예측 통합모델(김송이, 계명대학교 일반대학원, 2019)