미세먼지 머신러닝 예측모델에 대한 고찰

서동규*, 주재걸**
*고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
**고려대학교 정보대학 컴퓨터학과
e-mail:bluexs@naver.com

Insight of fine dust using Bigdata

Dong-Kyu Seo*, Jae-Gul Choo**
*Dept of Big Data Science, Korea University
**Dept of Computer Science and Engineering, Korea University

요 약

본 연구에서는 온 국민의 관심사이자 국가적 재난의 문제로 대두되고 있는 미세먼지의 최근 제시된 예측모델에 대해 고찰해 본다. 이를 통해 미세먼지와 머신러닝에 대한 이해를 넓히고자 한다.

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 데이터부족으로 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 각계의 여러 노력에도 불구하고, 단기 내에 획기적인 개선이 어려운 것이 사실인 바, 미세먼지농도의 예측을 통해 이에 대한 노출을 최소화하는 것이 국민건강을 지키기 위한 최선의 방법이라고 사료된다.

4차 산업혁명시대에 이른 지금, 빅데이터에 대한 분석 및 머신러닝을 통한 예측기법이 발달하고, 이를 활용한 다양한 주제의 연구가 이뤄지고 있는 바, 본고에서는, 기존 논문에서 실행된 머신러닝을 통한 미세먼지농도 예측모델을 고찰해보고, 이를 통해 해당연구에 대한 지식을 넓히고자 한다.

2. 기존연구에 대한 고찰

2.1 수치예측모델

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.

확보한 변수의 종류는 총 23개로, 한국환경공단제공데이터 5개(미세먼지(PM10), 이산화질소(NO2), 일산화탄소(CO), 아황산가스(SO2)), 기상청제공데이터 16개(기온, 강수량, 풍속, 풍향, 습도, 증기압, 이슬점온도, 현지기압, 해면기압, 일조, 일사, 전운량, 중하층운량, 최저운고, 시정, 지면온도), 날짜 2개(월, 시)루 구성되어 있으며, 이 중 미세먼지(PM10)를 종속변수로, 나머지 변수들을 독립변수로 설정하였다.

우선, 다중회귀분석 및 다중공선성 분석을 통해 불필요한 변수들을 제거 후 남은 13개의 독립변수(O3, CO, SO2, 강수량, 풍속, 풍향, 일조, 일사, 전운량, 중하층운량, 최저운고, 시정, 달(월))를 대상으로 종속변수(PM10)에 대한 다중회귀분석, ANN, SVM을 수행하였다.

각 예측모델에 시험자료를 투입하여 수행 결과, 다중회귀분석 : SVM : ANN = 74.35% : 80.35% : 85.1%로 비교적 높은 정확도를 도출하였다. 2차적인 미세먼지 발생

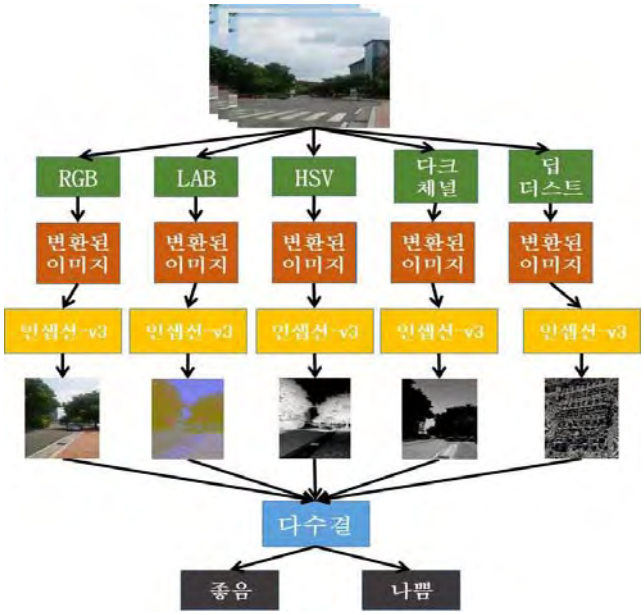
의 원인이 되는 O3, SO2[4]를 독립변수에 포함시킨 것이 주 원인으로 사료된다.

2.2 시각데이터를 활용한 머신러닝모델[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 가진다.

추출된 비디오 시퀀스를 다양한 기법을 이용해 변환시키고 변환된 이미지들 각각을 인셉션을 이용해 분석 후 나온 예측값을 다수결(majority vote)을 통해 결합하여 예측하는 앙상블 모형을 사용하여 미세먼지의 좋음, 나쁨을 예측하였다.



2.3 GBM(Gradient Boosting Machine)예측모델[6]

GBM은 여러 개의 의사결정모델을 연결하여 강력한 모델을 만드는 부스팅 방식의 앙상블 기법으로, 이전에 만들어진 의사결정모델의 오차를 보완하는 방식으로 순차적으로 모델이 연결되어 뒤로 갈수록 오차가 작아지게 된다.

행정안전부는 최근 1월 보도자료를 통해 NASA Aqua 위성 MODIS 센서데이터 및 인천지역 관측데이터를 전처리하여 GBM모델을 학습하여 최적 모델을 구성했다고 밝혔다.

'15.1월부터 '18.3월까지의 인천 지역 미세먼지·대기오염 데이터(환경부, 28,464건), 미국항공우주국(NASA)에서 제공하는 동북아 지역의 위성 센서 데이터* 및 에어로넷(AERONET)**의 지상관측 센서 데이터를 활용하였다.

* NASA Aqua 위성의 MODIS(중간해상도 영상 분광계) 센서 데이터로 미세먼지와 같이 공기 중에 떠 있는 작은 입자인 에어로졸을 관측

** NASA가 운영하는 국제 공동 에어로졸 관측 네트워크로 지상에서 관측

이를 통해 '18년 1분기를 예측한 결과, 미세먼지(PM_{10}) 84.4%, 초미세먼지($PM_{2.5}$) 77.8%의 정확도를 보였다고 한다. 주요 예측변수로는, 미세먼지의 경우 풍향, 강우량, 서해안 및 중국 산둥성 지역의 에어로졸 농도로, 초미세먼지의 경우 풍속, 풍향 및 중국 내몽골, 베이징·허베이성 지역의 에어로졸 농도로 나타났다.

앞선 모델과 달리 머신러닝을 활용하여 내일의 미세먼지 예측을 위한 미세먼지 예측모델을 개발하고, 미세먼지에 영향을 미치는 주요 요인을 파악한 것이 차이점이며, 국외요인의 영향을 도출한 것이 유의미한 특성이다.

3. 결론

미세먼지에 대한 국민의 불안은 계속 높아지고 있으나, 정부의 미세먼지 측정장비는 전국 390여개 지점에 설치되어 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 판단되며, 본고에서는 머신러닝을 활용한 다양한 미세먼지 예측모델을 고찰해보았다.

데이터의 양과 질이 예측모델의 정확도에 크게 기여하는 바, 더 다양하고 많은 데이터 확보를 통해 예측모델 개선이 가능할 것으로 예상된다. 기존 예측모델에서 활용된 기상데이터, 시각데이터 외에, 향후 굴뚝자동측정기기(TMS) 부착 사업장의 대기오염물질 배출량이 실시간으로 공개될 예정인 바, 이를 활용한 미세먼지 예측모델 개선과 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)
[6] 미세먼지, 빅데이터로 예측한다 (행정안전부 보도자료, 2019.01)