국가 측정소(AirKorea) 공기질 분석을 통한 예측 모델 개발

저자 : 김규민, 정미윤

소속 : 경상국립대학교

일자 : 2023년 5월 20일

초록(Abstract)

**I. 서론(Introduction)**

미세먼지는 우리 눈에 보이지 않을 정도로 작은 먼지 입자로, 입자 크기에 따라 지름 10 이하의 먼지를 미세먼지()라고 하며, 2.5) 이하의 먼지를 초미세먼지()라고 한다. 미세먼지 조성은 매우 다양하나, 주로 탄소성분(유기탄소, 원소탄소), 이온성분(황산염, 질산염, 암모늄), 광물성분 등으로 구성되어 있다. 이 미세먼지들은 대부분 사람의 코털이나 기관지 점막에 의해 걸려져 배출되지만, 미세먼지(PM-10)는 입자의 지름이 사람 머리카락 굴기의 1/5 ~1/7 정도인 10 *이하로 매우 작아 코, 구강, 기관지에서 걸러지지 않고 몸속으로 스며든다고 한다. 따라서 기관지에 미세먼지가 쌓이면 가래가 생기고 기침이 잦아지며 기관지 점막이 건조해지며 세균이 쉽게 체내에 침투할 수 있는 극심한 문제가 생길 수 있어 어떠한 조치가 필요한 때이다. 예를 들면 우리나라에서 발생하는 미세먼지는 60%정도가 중국이나 몽골의 사막지역에서 발생한 흙먼지가 편서풍을 타고 봄철에 날아오는 황사인데 황사 입자의 크기는 대부분 5~8 정도나 된다.[1]*

미세먼지는 호흡기의 흉부부위를 통과할 수 있을 만큼 작은 입자이기 때문에 지상의 기타 오존 및 대기 오염 물질(, , )[2]등의 물질과 같이 인체 건강에 더 큰 위험을 초래한다. 이러한 문제 때문에 국제보건기구(World Health Organization, WHO)에서 이 두가지 종류에 대한 측정을 통해 공기오염의 정도를 평가하고 있다. 그리고 보고된 바에 따르면, 2019년 전세계에서420만명의 미성숙아동이 대기오염으로 인해 사망을 했다고 한다.[3] 이렇게 미세먼지로 인해 발생하는 심각한 문제들로 인해 국민들의 안전을 보장하기 위해 미세먼지를 측정하고 국민들에게 위험수치를 알려주기 위해 우리나라는 환경부와 한국환경공단에서 전 국토에505개의 도시대기 측정소를 운영하여 실시간 운영하고 있는 중이다.[4]

미세먼지 예측을 위해 본 논문에서는 국내 기업인 ㈜아림사이언스 측에서 제공해준 OA200 (sensirion SPS30이 내재된) 광학 미세먼지 측정기를 경남 진주시 가좌동에 있는 경상국립대 351동(OA200\_1)[그림1.1, 1.2]과 경남 진주시 남강동에 있는 진주성의 창렬사(OA200\_2)[그림 2.1, 2.2]에 설치를 하여 2022년 12월 01일부터 측정을 실시해 현재 2023년 6월 측정을 진행중인 상태이다. 따라서 본 논문에서는 측정된 12월부터 4월간의 데이터와 AirKorea에서 측정하고 있는 경남 진주시 상대동과 상봉동의 미세먼지 측정 데이터를 가지고 AirKorea에서 측정하고 있는 경남 진주시 대안동의 미세먼지 평가지표(좋음, 보통, 나쁨, 매우나쁨)[그림3]를 예측해보기 위해 분류에 사용되는 머신러닝 알고리즘인 KNN(K-Nearest Neighbor), SVM(Support Vector Machine), DCT(Decision Tree), RF(Random Forest), RC(Rositic Classifier), MLP(Multi Layer Perceptron)등을 사용해 예측을 수행해보고 알고리즘의 적합성 평가를 하고자 한다.

그림1.1 그림1.2

야외, 하늘, 건물, 나무이(가) 표시된 사진

자동 생성된 설명텍스트, 벽, 빛, 실내이(가) 표시된 사진

자동 생성된 설명

그림2.1 그림2.2

지도, 텍스트이(가) 표시된 사진

자동 생성된 설명

그림3

**Ⅱ. 이론적 배경**

**Ⅱ-1. Class1 미세먼지 측정기**

**Ⅱ-2. 광학 미세먼지 측정기**

이러한 이유 때문에 본 논문에서는 PM2.5를 예측하기로 결정하였다.

**Ⅱ-3. KNN(K-nearest neigbor)**

**Ⅱ-4. SVM(Support Vector Machine)**

**Ⅱ-5. DCT(Decision Tree)**

**Ⅱ-6. RF(Random Forest)**

**Ⅱ-7. RC(Rogistic Classifier)**

**Ⅱ-8. MLP(Multi Layer Perceptron)**

**Ⅲ. 분석 내용 및 과정**

**Ⅲ-1. 전 처리 과정**

**Ⅲ-1.1. 사분위수로 이상치 제거**

현재까지 모은 데이터인 2022년 12월1일부터 2023년 4월30일까지의 모든 데이터를 그래프로 그려보면[그림 4.1], OA200\_1과 OA200\_2에서 다른 데이터에 비해 특정한 날에 엄청나게 높은 수치를 나타나고 있는 것을 알 수 있다. 따라서 각각의 데이터의 사분위수를 구해 IQR(Interquartile Range)의 10배 이상 벗어난 값들을 이상치로 판단하고 제거해 주었다. [그림 4.2]

**Ⅲ-1.2. 각각의 데이터에 풍향, 풍속, 온 습도 데이터 삽입**

현재 데이터를 보면[표1.1, 1.2, 1.3, 1.4, 1.5], OA200 데이터들은 풍향, 풍속 데이터가 존재하지 않고, airKorea 데이터들은 풍향, 풍속, 온 습도 데이터가 존재하지 않는다는 것을 알 수 있다. 따라서 OA200 데이터 들엔 기상청 기상자료 개방 포털에서 2022년 12월부터 ‘2023년 4월까지의 풍향, 풍속, 온 습도 데이터를 받아와 OA200\_1은 ‘경남 진주시 정촌면’에서 받아온 데이터를, OA200\_2와 airKorea\_상봉동은 ‘경남 진주시 상봉동’에서 받아온 데이터, airKorea\_상대동은 ‘경남 진주시 상대동’에서 받아온 데이터, airKorea\_대안동은 ‘경남 진주시 대안동’에서 받아온 데이터를 각각 삽입해주었다. [표2.1, 2.2, 2.3, 2.4, 2.5]

**Ⅲ-1.3. 모든 데이터들의 속성값이 NaN 데이터 삭제 후, 모든 데이터 인덱스 병합**

표2.3을 보면 ‘PM2.5’, ‘temp’, ‘humi’, ‘wind\_dir’, ‘wind\_speed’의 값이 NaN인 값이 존재한다. 모든 데이터를 pandas 라이브러리의 info함수를 사용해 확인해보면, airKorea데이터들에서 많은 NaN값이 존재한다는 것을 알 수 있다. [그림5.1] 따라서 모든 속성값에서 NaN 값을 pandas 라이브러리의 dropna함수를 사용해 제거해주었으며 set함수의 intersection함수 인덱스를 병합해주었다. [그림 5.2]

**Ⅲ-1.4. OA200 데이터의 값 보정**

현재 간접 미세먼지 측정 센서라고 불리는 광학 미세먼지 측정기는 습도에 따라 민감하게 변화한다고 알려져 있다. 즉 상대습도가 증가하거나 감소함에 따라 미세먼지 농도 또한 비슷한 비율로 증가하거나 감소한다는 의미이다. 특히 상대습도가 50.8~74.5%인 구간에서 상대습도와 미세먼지의 상관성이 높다.[5] 이러한 문제점 때문에 처음엔 이 습도와 상관관계를 분석해 측정된 데이터를 통해 미세먼지가 특이하게 높은 지점이 있으면 습도가 얼마나 나오고 이 값이 얼만큼 존재하는지 확인하여 살짝 보정을 하려 했다. 그림3을 보면 상봉동과 진주성의 데이터가 거리가 가장 가까우므로 기상청 기상자료 개방 포털에서 상봉동 2022년 12월부터 2023년 4월까지의 온 습도, 풍향, 풍속 데이터를 받아와 기존 데이터(OA200\_2, AirKorea 상봉동)에 삽입을 하고 비교를 해보았다. 우선 OA200\_2의 데이터 셋에는 온습도가 나타나 있기에 풍속과 풍향의 데이터만을 삽입하였고, AirKorea 데이터 셋에는 온,습도,풍속,풍향의 4가지 데이터를 다 삽입해 나타내었다.[표1,2]

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

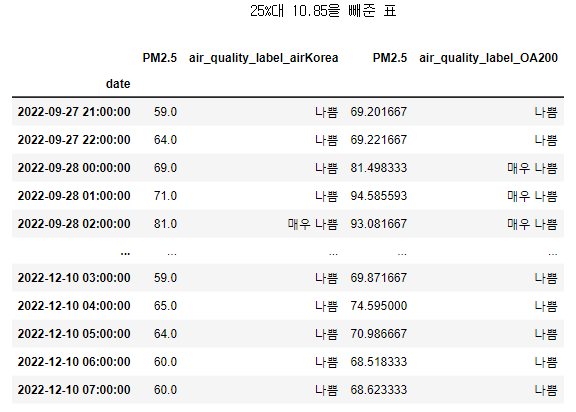
자동 생성된 설명 텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

그리고 전체 데이터를 확인하여 2022년 12월부터 2023년 4월까지의 데이터 중 평가지표가 ‘좋음’ 부터 ‘매우 나쁨’까지 골고루 분포되어 있는 1월달의 데이터를 가지고 OA200데이터셋의 특징을 살펴보기 위해 airKorea 데이터를 그래프로 그려보고, OA200의 데이터가 특히 높은 값을 나타내는 시간대를 찾아 비교를 해보았다. [그림4.1] 그림 4를 확인해보았을 때, OA200 데이터의 수치가 높게 나타나는 날은 (1/5~7일, 1/12~13일, 1/19일) 3개의 날로 간추릴 수 있었고, 이 데이터를 가지고 분석을 진행하였다. [그림4.2,4.3,4.4] 하지만 데이터를 보았을 때, 습도에 따라 데이터도 같이 올라가는 구간이 존재함을 확인했지만, 특정한 패턴을 찾을 수 가 없었다. 따라서 현재 쓰이는 간이 미세먼지 센서의 습도에 따른 보정을 적용하진 않았다.

그렇지만 현재 airKorea(상봉동) 데이터와 OA200\_2 데이터는 같은 위치에서 측정한 데이터가 아니므로 같은 위치에 있는 데이터를 ㈜아림 사이언스 측에서 제공해준 데이터를 가지고 비교를 해보기로 하였다. 2021년 8월10일부터 11월 4일까지 측정된 충북 음성군 금왕읍에 설치되어 있는 AirKorea데이터와 OA200센서의 측정데이터[5.1, 5.2]와 2022년 9월 01일부터 12월 29일 까지 측정된 충남 서천군 장항읍에 설치되어 있는 AirKorea 데이터와 OA200 센서의 측정데이터[그림6.1, 6.2]를 받아와 분석해보았다. 금왕 데이터는 pm2.5농도가 낮은 구간이 많고[그림 7] 장항항 데이터는 pm2.5농도가 높은 구간이 어느정도 존재한다는 것을 알 수 있다. [그림 8.1] 전체적인 데이터의 pm2.5 농도 분포가 높은 구간에서는 OA200 데이터가 높게 나타나고, pm2.5 농도 분포가 낮은 구간에서는 OA200 데이터가 낮게 나타나는 것을 알 수 있었다.

이 차이를 해결해보기 위해 첫번째 단계로 농도지수가 높게 나타나는 장항항 데이터를 가지고 평가지표가 매우 나쁨 구간인 PM2.5가 76이상인 데이터의 OA200과 airKorea 데이터의 차이를 비교 해보기로 하였다. 이 방법으로는 OA200의 PM2.5와 airKorea의 PM2.5의 차이를 box plot을 그려 그 차이의 데이터 전체 개수 중 25%,20%,10%에 위치하는 값을 찾아 25%는 10.65, 20%는 9.81, 10%는 6.87이 나오는 것을 확인한 후, OA200의 PM2.5값이 76이상인 데이터에 나온 수치를 각각 빼 주고 airKorea의 평가지표와 바뀐 OA200의 평가지표를 비교해보았다. [표3.1, 3.2, 3.3]



데이터는 크게 변하지 않았지만, 평가지표는 확실히 차이가 존재하였다. 표3.1에선 기존 OA200이 ‘매우 나쁨’이었던 데이터가 ‘나쁨’으로 변한 데이터가 전체 87개의 데이터 중 33개인데 그 중airKorea데이터와 동일하게 맞춰진 데이터의 개수가 30개, airKorea가 ‘매우 나쁨’인데 OA200이‘나쁨’으로 변한 안 좋은 데이터가 3개 존재하였고 30/33=90.9% , 표3.2에선 ‘나쁨’으로 변한 43개 중 5개의 데이터가 안 좋은 데이터로, 표3.3에선 ‘나쁨’으로 변한 46개 중 5개의 데이터각 안 좋은 데이터로 나타났다. 이 수치를 확률로 나타내면 각각 로 10%의 값을 빼준 값이 가장 좋은 성능을 보이긴 했다. 하지만 확률의 차이가 크지 않기도 하고, 전체 데이터의 개수가 87개로 많지 않은 수였기 때문에 가장 많은 데이터의 개수가 맞춰진 25%대의 값인 10.65를 OA200의 PM2.5에서 빼는 보정을 하기로 결정하였다. 이렇게 해서 맞춰진 값을 확인해보았을 때, 어느 정도 높은 위치에서의 값이 잘 맞춰진 것을 확인할 수 있었다. [그림 8.2]

이와 같이 낮은 수치에서의 값을 보정하기 위해 똑같이 금왕 데이터에 적용시켜 평가지표가 ‘좋음’구간인 PM2.5가 15이하인 데이터의 OA200과 airKorea 데이터의 차이를 비교 해보기로 하였다. OA200의 PM2.5가 15이하인 데이터는 1362개의 데이터였고, 이 중 896개의 데이터가 OA200과 airKorea의 PM2.5평가지표가 같았고, 나머지 466개는 다른 평가지표의 값이 나왔다[표 4.1]. 그리고 데이터를 수정하기 위해 농도가 높은 구간에서 적용했던 방법과 같이 box plot을 그려 OA200의 PM2.5와 airKorea의 PM2.5의 차이를 그렸다. 그런데 데이터의 개수가 많아 낮은 구간에서 25%는 1.74, 20%는 1.0, 10%는 -0.1이 나와, 낮은 구간에서는 50%, 60%, 75%의 값을 찾아 50%는 4.81, 60%는 6.07, 75%는 8.48이 나와 OA200의 PM2.5값이 15이하인 데이터에 나온 수치를 각각 더해주고 airKorea의 평가지표와 바뀐 OA200의 평가지표를 비교해보았다. [표4.2, 4.3, 4.4]수정하지 않았을 때는 전체 1362개의 데이터 중 896개의 데이터의 평가지표가 같았지만, 수정 후에는 표 4.2에선 전체 1362개의 데이터 중 1075개의 데이터의 평가지표가 맞아졌고, 표 4.3에선 1362개 중 1081개, 표 4.4에선 1362개 중 896개가 맞아졌다. 여기서 60%와 75%사이에 더 좋은 데이터를 추출해낼 수 있는 값이 존재함을 예측하였고, 60%부터 75%까지의 값 1%대씩 다 확인한 후, 적당한 값으로 찾아보았다. 그런데 61%인 6.13을 대입한 결과 1362개 중 1077개가 맞아지는 것을 확인하고, 1081개보다 안 좋은 결과가 나오는 것을 확인한 후, 50%부터 60%의 값에 극점이 존재할 것이라 예상하고 값을 다시 찾아보았다. 찾아본 결과 51%는 1077개, 52%는 1075개, …, 56%는 1095개, 57%는 1095개, 58%는 1094개 59%는 1084개로 결론적으로 56%, 57%인 값인 5.48*,* 5.6이 적당한 값이라는 것을 확인하였다. [표4.5, 4.6] 따라서 가장 평가지표가 잘 맞춰진 56%,57% 중 56%인 5.48을 OA200의 PM2.5에서 더하는 보정을 하기로 결정하였다. 이렇게 해서 맞춰진 값을 확인해보았을 때, 어느 정도 낮은 위치에서의 값이 잘 맞춰진 것을 확인할 수 있었다. [그림 7.2] 그래서 결론적으로 ‘2022년 9.01~12.29일’ 측정된 데이터인 장항항 데이터는 높은 구간, 낮은 구간의 차이가 둘 다 존재하므로 앞선 두 가지 방법을 다 적용시켜 그래프를 그려본 결과 그림 8.1과 비교해보았을 때, 꽤나 유사한 그래프를 나타내었다. [그림 9]

**Ⅲ-2. 머신러닝 알고리즘 적용**

ⅣⅤⅥⅦⅧⅨⅩ

ㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇㅇ

실험 방법 및 분석 방법(Experimental)

결과(Result)

토의(Discussion)

결론 및 요약(Conclusion or Summary)

논문 주제 : 간이 미세먼지 센서 OA200과 중량농도법을 활용한 미세먼지 센서인 Class1의 5개월 측정 데이터로 특정 국가측정소 공기질 예측 분석 알고리즘 개발 및 모델링

설정 이유 : 현재 정확성이 높지만 설치 비용이 높고 이동성이 낮은 중량농도법을 활용한 Class1 미세먼지 측정기는 진주시 내 특정 지역에만 설치 되어 있다. 따라서 정확성이 높은 1등급의 간이 미세먼지 센서인 OA200을 진주시 특정 지역에 설치하고 센서들의 중앙의 센서가 설치되어 있지 않은 지역의 미세먼지 농도를 예측해보아 알고리즘으로 개발해보는 것을 목적으로 이 주제를 선정하게 되었음.

향후계획 : 현재 국가 측정소의 미세먼지 농도 측정 센서는 전 국토에 367개가 설치되어 있다. 이 측정 데이터들을 이용한 미세먼지 농도 예측 및 설계를 다른 간이 미세먼지 센서인 OA200의 pm2.5농도와 위치, 풍향, 풍속, 온도, 습도 같은 간단한 다수의 기상 인사들을 독립변수들로 사용하여 MC(다중 분류) 알고리즘이나 ANN(인공신경망)을 통해 예측 모델을 설계하여 F1 Score, AUC 같은 분류평가지표를 사용해 pm2.5의 4가지 분류인 좋음, 보통, 나쁨, 매우나쁨 수준을 예측해보는 알고리즘을 모델링하는 계획을 잡고 있다. 현재 예측하려는 위치는 진주시 대안동에 설치되어 있는 국가측정소(Airkorea) 데이터임. 하지만 분명 현 주제에서 추출하려는 국가측정소 2장소의 데이터는 소실된 데이터가 존재하므로 그 소실된 데이터들은 소거하는 작업과 국가측정소와 약간의 오차가 존재할 수 있는 OA200에서 측정된 데이터의 값을 수정하는 작업이 필요할 것임.

**참고문헌**

[1] AirKorea - 미세먼지 정의. Retrieved from https://airkorea.or.kr/portal/web/contents/contentView/?pMENU\_NO=141

[2] AirKorea - 대기환경기준물질. Retrieved from https://www.airkorea.or.kr/web/airMatter?pMENU\_NO=130

[3] World Health Organization (2022). Ambient (outdoor) air pollution. Retrieved from //www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health

[4] AirKorea - 에어코리아 소개. Retrieved from https://www.airkorea.or.kr/web/contents/contentView/?pMENU\_NO=91&cntnts\_no=1

[5] Lee, C. H., & Oh, S. N. (2020). Effect of Atmospheric Humidity on Fine Dust Measurement Using the Light Scattering Method. J. Korean Soc. Hazard Mitig., 20(1), 391-399. doi: 10.9798/KOSHAM.2020.20.1.391