미세먼지 수집 · 분석 · 예측 Modeling 구축을 위한 위치선정 및 알고리즘 적합성 검증 방안 연구

정종진[O], 심흥섭[*]
[O]한국방송통신대학교대학원 정보과학과,
[*]동양대학교 컴퓨터 군사학과
e-mail: erunking@naver.com[O], enforce1023@nate.com[*]

For the establishment of fine dust collection, analysis, and prediction modeling A Study on the Location Selection and Algorithm Conformance Verification Method

Jung Jong Jin[O], Sim Heung Sup[*]
[O]Dept. of Information Science, Korea National Open University,
[*]Dept. of Computer and Military Affairs, Dongyang University

● 요 약 ●

미세먼지 수집을 위하여 필요한 위치 선정 방안과 위치 선정시 중요한 바람길분석, 수요조사, 유동인구, 교통량 등의 중요 기준을 반영하여 최종 선정하여야 하며, 이에 따라 설치된 측정기로부터 데이터 수집을 위해 지역적, 환경적, 지형적 요소를 감안하여 수집 항목을 결정하여야 한다.

데이터 수집시 실시간 또는 배치(Batch)로 할 것인지 여부를 결정하여야 하며, 이 보고서에서는 실시간으로 데이터 수집하는 경우를 설명하였다. 데이터 수집시 정확도를 높이기 위해 결측값, 이상값인 전처리 단계를 거쳐서 분석과 Modeling 구축을 통하여 정확도가 높은 알고리즘을 선정하여야한다.

정확도가 높은 알고리즘은 검증용 데이터 셋으로 적합성을 검증하여, 측정기 설치 위치의 적합성, 데이터 수집의 적합성, Modeling 구축 및 평가가 적합함을 지표로서 제시하여 적합성 검증을 하고자 한다.

키워드: 바람길분석, 정확도(Accuracy), 데이터 셋(Data Set), 결측값, 이상값

I. 서론

환경부는 지름 2.5㎛ 이하인 미세먼지[1](PM2.5) 환경기준을 일평균 35㎍/㎥ 및 연평균 15㎍/㎥로 강화하는 정책을 발표하였다.[1]

Table 1. 미세먼지(PM2.5) 환경기준 강화내용
(단위 : ㎍/㎥)

구 분	한 국		주요 선진국		기 타		
	현행	개정	미국	일본	WHO	EU	중국
연평균	25	**15**	15	15	10	25	35
일평균	50	**35**	35	35	25	(없음)	75

이에 따라 미세먼지 환경기준 및 예보기준 강화와 별도로 주의보, 경보 기준도 강화를 추진하여 '대기환경 보전법 시행규칙'을 개정하였으며, 대기오염으로 우리나라 조기 사망자수는 연간 1만 8천여 명으로 환경부는 발표하였다.[2022.05.03.자 환경부 보고자료]

지자체별로 미세먼지를 관리하기 위하여 현황조사에 따라 측정기를 설치하고 있다.

미세먼지 간이측정기 설치를 위해 현황진단과 여건분석 자료는 다양한 환경요소를 반영하여 분석이 필요하다. 본 논문에서는 환경적 요소로는 지리적, 지역적, 산업적 관점을 기준으로 선정하였으며, 경제적 요소는 거시적, 미시적 요소를 선정하였다. 기술적 요소로는

1) 먼지는 입자의 크기에 따라 50㎛ 이하인 총먼지(TSP, Total Suspended Particles)와 10㎛ 이하의 미세먼지(PM, Particulate Matter)로 구분되며 PM10과 PM2.5의 기준은 다음과 같다.
- PM10 : 지름이 10㎛보다 작은 미세먼지 (사람 머리카락 지름보다 약 1/5~1/7 정도 작은 크기)
- PM2.5 : 지름이 2.5㎛보다 작은 미세먼지 (사람 머리카락 지름보다 약 1/20~1/30 정도 작은 크기)

정보통신 기술의 발전, 측정기의 가격 다변화 등의 요소를 선정하여 분석하였다.

환경적 요소는 지리적 관점으로 미세먼지 측정기를 설치를 위한 대기환경을 조사와 지역의 특징을 파악과 더불어 지역적 기후 및 환경변화와 어린이 등 민감 계층 보호를 위한 대책이 필요하다.

경제적 요소인 정보통신 기술의 발전으로 대기환경 현장 정보를 실시간 양방향으로 전달 가능하여 미세먼지 행동요령 전파를 통한 가시적인 경제적 이득을 이루고 있다. 간이측정기의 다변화로 인하여 환경부 1등급 제품에 비용 측면에서 예산확보가 용이 해졌다.

미세먼지 측정기 설치 지역인 양주시는 지역적으로 산업의 중심지로서 산업단지가 11개소 있으며, 양주시 미세먼지 발생 기여율 측면에서 45%정도로 분석하였다. 산업단지 대기 배출 시설 사업장의 92.6%를 차지하는 4종, 5종 사업장의 발생기여도가 전체 38%를 차지하여 미세먼지 배출원인 인벤토리 관리가 필요하다.

기후 환경적으로는 미세먼지 고농도 시기인 10월에서 4월까지 미세먼지 농도가 높게 분포되었다.

대내외적인 요인으로는 지역 주민과 취약계층의 지속적인 민원 발생과 도시개발 사업으로 미세먼지 및 비산먼지 등 대기질 관련 환경문제 해결책이 시급하다.

대기환경 관리를 체계적이고 지속적으로 관리하기 위하여 미세먼지 측정기를 바람길을 분석하여 미세먼지 농도가 높은 위치에 설치하여 실시간으로 데이터를 수집하여 지자체의 미세먼지 저감 대책에 반영하여야 한다.

이 논문에서는 측정기를 설치하는 위치의 적합성과 데이터를 수집하여 분석, 모델, 예측 시스템 알고리즘 선정방안의 적합성을 검증하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 미세먼지 측정기 위치 선정의 적합성에 대하여 관련된 연구 동향을 알아보고, 3장에서는 데이터 수집을 통한 분석, Modeling, 예측 알고리즘의 선정방안 적합성을 알아보고 4장에서는 결론 및 향후 연구 과제를 제시하며 끝을 맺는다.

2. 관련 연구

2.1 미세먼지 위치 선정 방안

지자체별로 대기 오염 물질인 미세먼지 측정기를 설치하여 미세먼지 데이터 수집과 분석을 통하여 미세먼지 저감대책을 마련하기 위하여 미세먼지 유발요인을 억제하고 미세먼지 수집을 위하여 측정기를 설치하기 위해선 많은 사전 절차가 필요하다.[2]

Table 2. 측정기 위치 선정을 위한 절차

1	미세먼지 조사 및 과거 자료 수집 방안
	↓
2	미세먼지 분포 및 활동인구 분석
	↓
3	지역별, 지형별 바람형성 분석
	↓
4	미세먼지 민감 지역 분석
	↓
5	최종 위치 선정

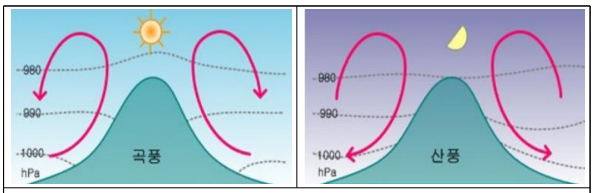
절차 기준에 따라 미세먼지 측정기 위치를 선정 계획을 준비하고 추가적으로 각 행정동 별 민원 발생 조사와 사전 의견수렴을 통하여 최종 위치를 선정하여야 한다.

위치를 선정하였다면 바람길 영향도에 따른 위치 조정이 필요하다.

2.2 바람길 분석

바람길 분석의 필요성은 각 도시별 지형이 차이가 남으로 지역별 건물과 바람길을 분석하여 위치를 선정하여야 한다.

바람길 분석은 도시의 복합한 토지피복, 지형, 건물을 직접적으로 고려하여 공기의 흐름을 분석하여 도시 규모의 바람생성, 추적, 흐름을 분석하여 반영하여야 한다.[3]



자료출처 : 케이웨더

Fig. 1. 공기 흐름도

2.3 측정기 위치 최종 선정 검증

미세먼지 측정기 위치를 최종 선정하기 위해서는 우선적으로 측정기를 설치할 위치를 조사한 다음 바람길 분석을 통한 위치 조정하여 최종 위치를 선정하여야 적합한 데이터 수집이 가능하다.

또한, 측정기의 성능에 따라 수집할 항목을 결정하여 환경기반 대기질 수집이 적정하다고 할 수 있다.

3. 미세먼지 데이터 수집 · 분석 · 모델 · 예측 검증

가. 데이터 수집 방안

미세먼지 데이터 수집은 실시간(5분) 단위로 데이터를 수신받아서 처리하여야 현장의 대기질 상황을 정확히 파악할 수 있다. 측정기는 국가측정망과 간이측정망으로 분류되며, 지자체에서 운영하는 것은 간이측정망을 구축하여 데이터를 수집하게 된다.[4]

데이터 수집에서 기초 비교 측정데이터는 국가측정망 데이터를 확보하고, 간이측정망 자료는 실시간 자료를 확보하여 두 데이터를 수집하여 결측치와 이상치를 전처리 후 기초통계량을 조사하여 정규분포 신뢰도 구간에 들어오면 데이터 수집은 적정한 것으로 판단하면 된다.

나. 데이터 분석 방안

분석 모델 설계·구축 시에는 해당 모델의 학습데이터·평가데이터·검증데이터를 통해 최적의 모델을 선정 및 적용하기 위해 하나 이상의 모델을 준비하여 최소한의 시간에 탐색적 분석을 완료하는 것이 성공적인 분석의 관건으로 단위 분석에 대한 예상 소요 시간을 추정해 필요 시 샘플링할 수 있다.[5]

표본을 추출하는 방법에는 다양한 추출방법을 이용하며 보통 실무에서는 단순, 층화 등 혼합해 사용하여 추출한다. 분석에서는 학습데이터 80%, 평가데이터 20%로 분류하여 데이터 셋을 준비하는 과정이다.

다. 데이터 Modeling 구축 평가 방안

Modeling에서는 통계기반 데이터 분석을 위하여 정형/비정형 데이터를 분석 목적에 따라 가설을 설정하고 분석 모델을 만들어 평가하는 요소로서 이산형 종속변수 평가항목 1종 오류, 2종 오류, 정밀도, 정확도, 특이도, 민감도, 오류율, FP Rate, TP Rate 등이고, 연속형 종속변수 평가항목은 MAPE(Mean Absolute Percent Error), RMSE(Root Mean Square Error) 등이다.

분류(Classification) 모델은 의사결정나무, 로지스틱회귀분석, KNN 등을 이용하며, 클러스터링(Clustering) 모델은 K-Means을 이용하여 분류한다.

Modeling의 정확성을 높이기 위해 앙상블 모델은 배깅(Bagging)과 부스팅(Boosting) 등을 적용한다.[6]

데이터 셋 분할 시 훈련데이터 셋(Training Data Set)은 머신러닝 기법이나 알고리즘 학습을 적용하여 결과 파라미터를 도출하는 데이터 셋을 의미한다.

데이터 셋 분할 시 테스트 데이터 셋 (Test Data Set)은 훈련데이터 셋을 통해 도출된 머신러닝 학습 결과를 실제 적용하여 결과 예측(Prediction)을 수행할 새로운 데이터 셋을 말한다.

데이터 셋 분할 시 검증용 데이터 셋 (Validation Data Set)은 모델의 성능을 평가하거나 개선하기 위해 사용되는 데이터 셋으로서, 모델 학습 자체 혹은 결과 개선에 활용 되거나, 모델 평가 검증에 사용되는 데이터 셋을 의미하며, Modeling 자체에는 사용되지 않고, 결과 예측에만 사용되는 테스트 데이터 셋 (Test DataSet)과는 다르다.

지도학습 머신러닝 기법은 예측하고자 하는 목적변수의 형태에 따라 명목형/이산형으로 나누며 분류모델 지도학습 머신러닝 기법으로 구분하고, 목적변수가 연속형 변수의 경우 수치 예측모델 지도학습 머신러닝 기법으로 구분한다.

분류 모델을 위한 지도학습 머신러닝 기법은 적용할 알고리즘을 다음 중에서 선정하여 k-NN, Naive Bayes, Bayesian Network, Logistic Regression, Decision Tree, Random Forest, SVM, HMM, FFT, Neural Network, Deep Learning 학습한다.

수치 예측 모델을 위한 지도학습 머신러닝 기법은 다음 알고리즘을 적용해야 하며 종류로는 Regression, Neural Network, SVM, Decision Tree, Time Series : AR, MA, ARIMA 등 사용 여부를 결정해야 한다.

Table 3. 교차검증 결과

Model	AUC	CA	F1	Precision	Recall
kNN	0.614	0.096	0.082	0.080	0.096
SVM	0.581	0.108	0.084	0.090	0.108
Neural Network	0.654	0.120	0.115	0.112	0.120
Logistic Regression	0.783	0.163	0.143	0.131	0.163

* 측정도구 : Orange 3.32.0

표3에서는 다양한 모델로 교차검증을 수행한 결과이며, Logistic 회귀분석의 AUC가 가장 높은 수치를 나타내고 있으며 이 모델을 선정하여 Modeling을 평가하는 것이 적정하다고 판단하였다.

머신러닝 기반 데이터 분석 모델 성능 평가 도구 및 지표는 예측값과 실제값 비교를 위한 혼동 행렬 (Confusion Matrix)이며 주요 모델 평가 지표 (Accuracy, Precision, Sensitivity, Specificity, FP Rate, Kappa, ROC 커브 등)을 활용하여 Modeling을 평가한다.[7]

라. 모델 평가 검증 방안

데이터 셋 분할 시 시험 데이터 셋(Test Data Set)은 훈련 데이터 셋을 통해 도출된 머신러닝 학습 결과를 실제적용하여 결과 예측(Prediction)을 수행할 새로운 데이터 셋을 이용하여 검증한다.

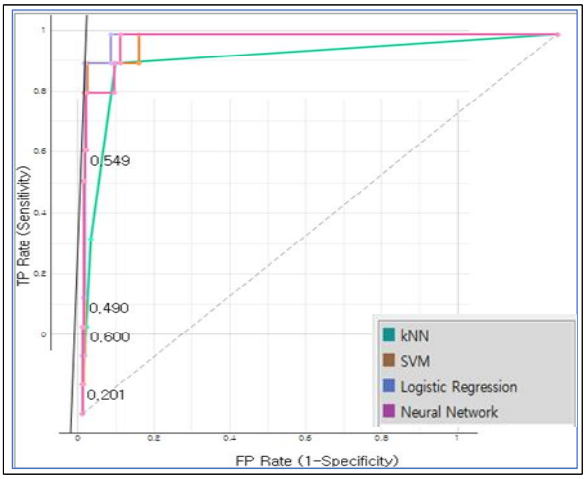
머신러닝 기반 빅데이터 분석 모델 성능 평가 도구 및 지표는 다음을 포함한다.

지도학습(분류)의 평가로는 예측분류와 실제분류 결과 대조를 통한 분류정확도와 빅데이터 분석의 성과 기준은 분석 결과를 활용하여 얻을 수 있는 성과에 대한 지표를 활용하여 평가한다.

- 비용편익분석(Cost Benefit Analysis) 방법
- IT 투자 성과평가 방법론
- BSC(Balanced Score Card)

분석 성과지표는 반드시 정의하여 평가해야하는 것은 아니며, 분석의 성격에 따라 정의하기 어려운 경우도 있다.

Table 4. ROC Analysis 검증



* 측정도구 : Orange 3.32.0

표4에서는 AUC-ROC Curvesms 다양한 임계값에서 모델의 분류 성능을 나타내는 결과이며, AUC가 가장 높은 모델은 Logistic Regression으로 나타났다.

마. 최종 알고리즘 검증 방안

지도학습 머신러닝 분석 모델 결과는 모델 훈련에 사용한 훈련데이터에 편향된 결과값을 내는 경향이 많으므로, 머신러닝 알고리즘이 데이터로부터 얼마나 잘 학습했는지를 평가하기 위해 평가 데이터 셋을 이용하여 모델의 정확도를 평가하게 된다. 여기서 수치예측 목적의 지도학습 머신러닝일 경우, 평가 데이터 세트에서 계산된 예측값과 실제 목표 변수값 간의 평균제곱오차(MSE) 등을 사용하여 측정하게 되며 분류 예측 목적의 지도학습 머신러닝일 경우, 목표변수의 예측된 분류와 실제 분류가 얼마나 일치하는지 분류정확도 등을 계산하여 측정하게 된다.

이 과정에서 예측분류와 실제 분류 결과 대조를 위해 혼돈행렬 등을 작성하여 분류정확도 등을 확인하는 절차를 거치게 된다. 이 과정에서 민감도(Sensitivity, 혹은 Recall, Hit Ratio 등으로도 불림)나 정밀도(Precision) 등의 평가 지표를 계산하여 Modeling 성능을 판단하여 적합성을 판단한다.

4. 결론 및 향후 연구

본 연구는 미세먼지 측정기 설치를 위한 적합성 검증 반안에 대하여 연구하였으며, 측정기를 설치 시 중요한 요소는 지역의 교통량, 유동인구, 산업단지 밀집지역 등을 사전 조사하고 지자체의 비산 먼지 발생지역을 추가보완 하면 좋은 위치 선정 방안 자료가 될 것이다.

미세먼지 데이터 수집을 통하여 분석 시 중요한 알고리즘 선정과 Modeling 구축을 위하여 SVM과 RandomForest를 이용하여 정확도를 파악하였다.

본 연구를 바탕으로 Modeling 구축의 정확도와 예측 분석을 위한 Modeling 구축을 위하여 앙상블(Xboot, Bagging 등)을 이용하여 추가적으로 신뢰도 높은 예측 Modeling을 구축을 위하여 추가 연구가 진행 중이다.

REFERENCES

[1] 환경부, 미세먼지 환경기준 미국, 일본 수준으로 강화. 2018. 03.21

[2] 페이저 측정 시스템의 측정기 최적 배치, 김재훈, 1999

[3] 배민기, 정용일, 최은희, 이채연, 양호진, 이광진, 김보은, 이승욱.(2019).충청북도 고농도 미세먼지 발생 원인규명을 위한 바람길 분석.연구보고서,(),1-100.

[4] 김남호(Nam Ho Kim). (2018). 드론을 이용한 대기환경정보 수집장치 개발 및 응용 연구. 스마트미디어저널, 7(4), 40-47.

[5] 임준묵(Lim, Joon-Mook);고선호(Ko, Sunho);김제완(Kim, Jewan). (2018). 기상데이터와 머신러닝을 활용한 미세먼지 농도 예측 모델. 한국IT서비스학회 학술대회 논문집, 2018(3), 691-694.

[6] 김성태, 구윤서. (2015). 미세먼지 앙상블 예보기법 개발. 한국도시환경학회지, 15(3), 251-260.

[7] C.-H. Kim et al., “A Study on Statistical Parameters for the Evaluation of Regional Air Quality Modeling Results - Focused on Fine Dust Modeling -,” Journal of Environmental Impact Assessment, vol. 29, no. 4, pp. 272–285, Aug. 2020.