기계학습 군집 알고리즘을 이용한 미세먼지 비선형성 완화방안

이상권 · 조경우 · 오창헌*

한국기술교육대학교

Non-linearity Mitigation Method of Particulate Matter using Machine Learning Clustering Algorithms

Sang-gwon Lee · Kyoung-woo Cho · Chang-heon Oh*

[1]Department of Electrical, Electronics & Communication Engineering, Korea University of Technology and Education(KOREATECH)

E-mail : tkdrnjs507@koreatech.ac.kr / pinokio622@koreatech.ac.kr / choh@koreatech.ac.kr

요 약

고농도 미세먼지 발생이 증가함에 따라 미세먼지 예측에 많은 관심이 집중되고 있다. 미세먼지는 대기 중에 있는 직경 $10\mu m$ 이하의 밀입자 물질을 말하며, 온도, 상대습도, 풍속 등의 기상 변화에 영향을 받는다. 따라서 미세먼지 예측을 위해 기상 정보와의 상관관계를 분석하는 다양한 연구가 진행되었다. 하지만 미세먼지의 비선형적 시계열 분포는 예측 모델의 복잡도를 증가시키고, 부정확한 예측값을 초래할 수 있다. 본 연구에서는 기계학습의 군집 알고리즘 및 분류알고리즘을 이용하여 미세먼지의 비선형적 특성을 완화하고자 한다. 사용된 기계학습 알고리즘은 병합군집, 밀도기반군집이며, 각 알고리즘을 통한 군집결과를 비교, 분석하였다.

ABSTRACT

As the generation of high concentration particulate matter increases, much attention is focused on the prediction of particulate matter. Particulate matter refers to particulate matter less than $10\mu m$ diameter in the atmosphere and is affected by weather changes such as temperature, relative humidity and wind speed. Therefore, various studies have been conducted to analyze the correlation with weather information for particulate matter prediction. However, the nonlinear time series distribution of particulate matter increases the complexity of the prediction model and can lead to inaccurate predictions. In this paper, we try to mitigate the nonlinear characteristics of particulate matter by using cluster algorithm and classification algorithm of machine learning. The machine learning algorithms used are agglomerative clustering, density-based spatial clustering of applications with noise(DBSCAN).

키워드

Particulate matter, Machine learning, Agglomerative clustering, DBSCAN

Ⅰ. 서 론

대기 오염 물질 중 하나인 미세먼지는 인체에 영향이 가장 큰 것으로 알려져 있으며, 호흡기 및 심혈관계 질환 발생과 연관이 있는 것으로 보고되고 있다[1]. 미세머지는 대기 중에 있는 직경 $10\mu m$ 이하의 밀입자 물질을 말하며, 산불 또는 황사로 인한 자연

* corresponding author

적 배출, 석탄 연소 및 경유 연소를 통한 인위적 배출 등을 통해 발생한다[2].

미세먼지는 다양한 기상 인자와 상관관계를 형성하며, 예측을 위한 파라미터로 활용 할 수 있다[3],[4]. 하지만 미세먼지 농도의 지역 및 일일 대기 환경에 따른 비선형적 특성은 예측 모델의 복잡도를 증가시키고, 정확한 예측을 어렵게 만든다[5].

이에 본 논문에서는 미세먼지 예측에 앞서, 미세먼지 농도의 비선형적 시계열 분포를 기계학습의 군집 알고리즘을 통해 완화하고자 한다. 사용된 알고리즘은 병합군집, DBSCAN이며, 성능을 평가하기 위해 미세먼지 농도의 군집 결과를 확인하였다.

II. 기계학습 기반 군집 알고리즘

대표적인 기계학습 기반 군집 알고리즘은 비지도 학습 알고리즘으로써 데이터가 지닌 유사한 특성을 파악하여 분류하는 기법이다. 대표적인 알고리즘은 k-means, 병합군집, DBSCAN이며, 선행된 연구를 통해 k-means의 군집 특성을 파악하였다[6]. 따라서 본 논문에서는 병합군집, DBSCAN의 군집 특성을 파악하고자 한다.

병합군집은 계층적 군집으로써 초기 모든 d-차원 상의 벡터가 단일 군집을 이루고, 인접한 군집과 차례대로 새로운 군집을 형성하는 알고리즘이다. DBSCAN의 경우, 데이터의 밀집도를 분석하여 군집하는 방법으로써, 잡음(noise)을 포함한 공간 데이터와 다양한 모양 및 크기를 가진 데이터 집합에 적합하다.

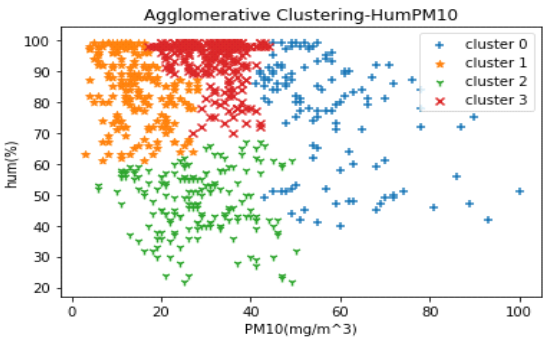

그림 1. 미세먼지-습도 데이터를 이용한 병합군집

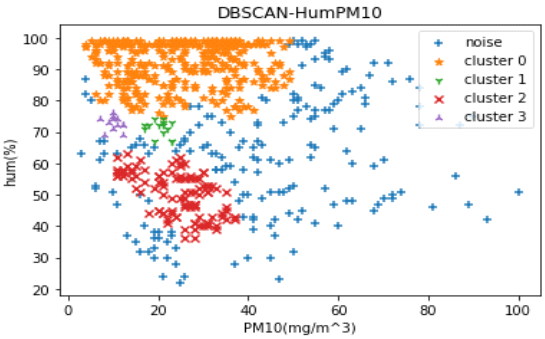

그림 2. 미세먼지-습도 데이터를 이용한 DBSCAN

그림 1은 2017년 9월 한 달 동안 측정된 723개의 미세먼지 농도와 습도 데이터를 병합군집 한 결과이다. 그래프 좌표상의 밀집 지역으로부터 우선적으로 군집을 이루기 때문에 가장 밀집도가 높은 cluster 3에서 224개를 분류했다. 다음으로는 cluster 1에서 204개, cluster 2에서 180개, cluster 0에서 116개로 분류하였다.

그림 2는 병합군집과 동일한 데이터를 사용하여 DBSCAN을 수행한 결과이다. 좌표상의 모든 데이터의 거리를 측정하여 군집하기 때문에 데이터가 밀집된 군집과 비교적 적은 데이터를 포함한 군집이 모두 보였다. 일정 거리를 벗어난 데이터는 모두 noise로 취급하기 때문에 밀집 지역 이외의 데이터는 모두 무효한 데이터로 취급된다.

III. 성능 평가

그림 3은 병합군집 수행 결과를 Boxplot으로 보여주고 있다. base는 군집 수행 이전의 데이터이며, $3\mu g/m^3$부터 $70\mu g/m^3$까지 고르게 분포하였다. 특히, $70\mu g/m^3$ 이상부터 이상치 데이터가 다수 보였다. cluster 0은 고농도 미세먼지로 분류되며, 몇몇 이상치 데이터가 관찰된다. cluster 1부터 cluster 3은 저 농도와 중간 농도의 미세먼지 구간이 구분됨을 보였다.

DBSCAN의 경우, 각 좌표상의 데이터 사이의 거리를 측정하여 근접한 데이터를 군집하기 때문에 특성이 확실하게 구분된 데이터에 적합하다. 하지만 미세먼지 및 습도 데이터가 나타내는 부정확한 분포는 알고리즘 수행 후 많은 noise와 함께 편향된 군집 결과를 보였다.

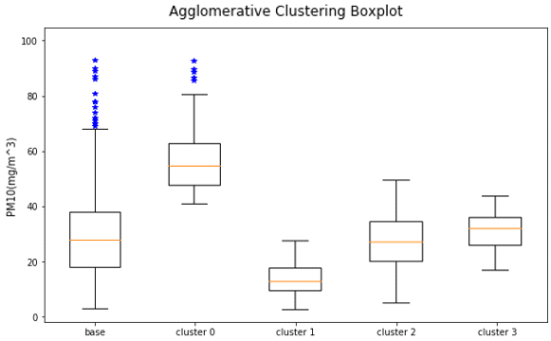

그림 3. 병합군집 결과를 이용한 BoxPlot

IV. 결 론

미세먼지의 정확한 예측을 위해 다양한 기상 정보가 활용될 수 있다. 하지만 미세먼지의 비선형적 시계열 분포는 예측 모델의 복잡도를 증가 시키고, 부정확한 예측 결과를 도출할 수 있다. 따라서 본 연구에서는 기계학습 기반의 군집 알고리즘인 병합군집과 DBSCAN을 이용하여 미세먼지의 비선형적 특성을 완화하고자 한다. 미세먼지와 습도 데이터를 사용하여 군집 알고리즘 수행하였으며, Boxplot으로 평가하였다. 병합군집의 경우, 고농도 미세먼지 구간에서 몇몇 이상치가 발생하였으며, 저 농도 및 중간 농도를 분별할 수 있음을 보였다. 하지만 DBSCAN은 많은 noise와 편향된 군집 결과로 인해 미세먼지 군집에 적합하지 않음을 나타내었다. 병합군집을 이용한 미세먼지 예측 시나리오를 설계할 경우, 보다 정확한 예측 결과를 도출할 것으로 판단된다. 향후 연구로는 농도별로 구분된 미세먼지 데이터를 통해 기계학습 및 시계열분석을 이용한 구간별 미세먼지 예측모델을 설계할 예정이다.

References

[1] K .Katsouyanni, G. Touloumi, C. Spix and J. Schwartz, “Short term effects of ambient sulphur dioxide and particulate matter on mortality in 12 European cities: results from time series data from the APHEA project,” *Bmj*, pp. 1658-1663, Jun. 1997.

[2] Y. P. Kim, “Air pollution in seoul caused by aerosols,” *The Journal of Korean Society for Atmospheric Environment*, Vol. 22, No. 5, pp. 535-553, Oct. 2006.

[3] H. Y. Son and C. H. Kim, “Interpretating the spectral characteristics of measured particle concentrations in busan,” *The Journal of Korean Society for Atmospheric Environment*, Vol. 25, No. 2, pp. 133-140, Apr. 2009.

[4] M. K. Shin, C. D. Lee, H. S. Ha, C. S. Choe, and Y. H. Kim, “The influence of meteorological factors on PM10 concentration in incheon,” T*he Journal of Korean Society for Atmospheric Environment*, Vol. 23, No. 3, pp. 322-331, Jun. 2007.

[5] A study on the meteorological characterization of high concentration particulate matte, ChungNam Institute Seohaean Research Institute, Research Report, pp. 1-50, Mar. 2017.

[6] S. G. Lee, K. W. Cho, C. G. Kang and C. H. Oh, “Nonlinear mitigation method of particulate matter using k-means clustering,” in *Proceeding of Conference on Korea Navigation Institute*, Seoul, pp. 126-128, 2018.