지역별 미세먼지 발생 데이터 클러스터링 메소드 설계 및 구현

Designing and Implementing Clustering Method of Particulate Matter Data by Region

문 주 환* • 윤 홍 식**
Moon, Ju-Hwan • Yoon, Hong-Sik

요 약

본 연구는 우리나라의 지역별 미세먼지 발생 데이터에 대한 수집과 그에 대한 분석, 처리 방법에 대한 연구로 수집된 미세먼지 데이터에 대한 클러스터링 메소드를 설계하고 구현하는 것을 목표로한다. 본 연구에서는 기상청 산하의 30여개의 관측소에서 측정된 미세먼지 데이터를 기반으로 클러스터링 작업에 대한 전처리를 실시한다. 이러한 전 처리에는 각 관측소의 미세먼지 데이터의 시계열 그래프의 유사도를 비교하기 위하여 Dynamic Time Warping알고리즘을 활용하였으며 이를 통해 산출되는 DTW값을 통하여 유사도가 높은 미세먼지 측정 지역별 클러스터링을 수행해 클러스터링 군별 미세먼지 발생 원인에 대한 분석과 대비, 피해저감 방안등의 대책 마련을 위한 자료로서 활용됨을 목적으로 한다.

keywords : 미세먼지, Dynamic Time Warping, 데이터 클러스터링

1. 서 론

본 연구는 우리나라의 대표적 재해중 하나인 미세먼지에 대하여 미세먼지의 발생 원인에 대한 분석과 그에 대한 예방 및 피해저감 방안 마련을 위한 자료를 제시하기 위하여 진행 되었다. 우리나라는 중국에서 날아오는 분진을 비롯하여 자동차와 생활분진, 공장과 발전소로부터의 공해 등 다양한 오염을 통해서 미세먼지 및 초 미세먼지가 생성되고 이렇게 생성된 미세먼지로 인하여 전반적으로 다른 국가들에 비하여 높은 미세먼지 농도를 보이고 있으며, 이러한 미세먼지에 대한 관심 및 우려는 날로 증가되고 있는 추세이다. 본 연구에서는 이러한 미세먼지로 인한 재해에 대한 분석과 대비, 피해저감 방안 마련을 위한 방법으로서 DTW 기법을 통한 지역별 클러스터링 메소드를 설계하고 이를 구현함으로서 미세먼지 발생 및 농도에 대하여 클러스터링 군을 분류해 내고 분류된 지역에 대한 미세먼지 발생 및 농도 데이터의 상관관계 및 인과관계에 대한 분석을 통해 보다 효율적인 미세먼지 대책 마련을 위한 자료를 구축한다.

2. 본론

2.1. 분석 데이터

* 정회원 • 성균관대학교 방재안전공학협동과정 석사과정 moonjuhwan@skku.edu
** 정회원 • 성균관대학교 건설환경시스템공학과 교수

본 연구에서 분석된 데이터는 한국환경공단의 실시간 대기정보 PM10 미세먼지 데이터를 바탕으로 진행되었다. 그중 메소드의 설계와 구현에는 수집된 데이터 중 특이성을 띈 패턴을 갖고 있는 것으로 파악되는 2015년 11월 29개 관측소에서의 측정 값을 사용하였다.

2.2. 클러스터링을 위한 데이터 전처리

클러스터링을 위한 시계열 그래프에 대한 분석에 사용된 Dynamic Time Warping(이하 DTW) 알고리즘은 지역별 미세먼지 데이터의 시계열 그래프에 대한 유사도를 측정하여 이를 정량적 데이터로 나타내기 위하여 사용되었다. 본 연구에서 쓰인 DTW알고리즘은 Euclidean 거리 측정 방식을 통한 두 그래프의 각 시간 지점별 데이터의 유사성분석 방법이 측정 지역 간의 시간별 딜레이가 발생하는 자료에 대한 유사도를 측정함에 있어서 적합하지 않다는 판단에 따라 각 시간 지점별이 아닌 동적인 시간 개념을 반영한다. 또한 본 연구에서는 시간의 경과에 따른 가중치를 부여하기 위하여 동적인 시간 개념 하에서 지점에 대한 거리 측정을 Pythagorean방식을 적용하여 계산함으로써 측정 지역 간 발생하는 시간별 딜레이를 고려한 미세먼지 농도의 변화와 지역별 상관관계에 대한 데이터 분석이 가능하였다. 본 연구에서는 이를 위한 가중치로 시간별 미세먼지 데이터 변화 값의 표준 편차 평균값을 기반으로 추출한 임의의 값으로 가중치를 지정하였으며 이는 차후 연구를 통해 변경 및 보완한다.

2.3. 지역간 패턴 유사도 수치화

위 과정에서 도출된 Pythagorean DTW 알고리즘을 각 지역 간의 네트워크에 적용시켜 관측소에서 측정된 미세먼지 데이터 간의 유사도를 비교하여 DTW값을 산출해 낸다. 이렇게 산출된 값을 통하여 유사도를 정형화된 수치로 표현할 수 있다. 수치화 되어 표현된 각 지역 측정소간의 유사도를 통하여 각 측정소간 유사도를 비교 및 분석할 수 있고 또한 일반적인 클러스터링 메소드에 적용시켜 측정 지역별 클러스터링 그룹화를 구현할 수 있다.

3. 결론

위의 연구 과정을 통하여 국내 지역별 미세먼지 발생 데이터에 클러스터링을 수행하였고 이렇게 수행된 클러스터링의 결과로 도출된 미세먼지 유사 발생 지역 그룹을 통하여 그룹별 미세먼지 발생 원인에 대한 분석과 대비, 피해 저감 방안등의 대책 마련을 위한 연구에 활용될 자료를 구축하였다. 다만 본 연구는 짧은 기간 내에 전반적인 연구 개요에 대한 작성에 미치치 못한 것이 사실이다. 따라사 이어지는 차후 연구를 통하여 전반적인 내용을 보완하고 상기한 과정에 대한 서술 및 구체화를 통하여 보다 더 효과적인 연구 자료로서 활용될 수 있는 방안을 마련키로 한다.

감사의 글

본 연구는 국민안전처장관의 방재안전분야 전문인력 양성사업으로 지원되었습니다.

참고문헌

Carmelo Cassisi, Placido Montalto. (2012) Similarity Measyres and Dimensionality Reduction Techniques for Time Series Data Mining, Advances in Data Mining Knowledge Discovery and Applications

Wes McKinny (2012) Python for Data Analysis, *OREIILLY*