Keras를 이용한 대기오염이 유해질환에 미치는 위험 예측 시스템

이지수[O], 이유정[*], 윤수한[*], 문유진[**]
[O]한국외국어대학교 Global Business & Technology학부,
[**]한국외국어대학교 Global Business & Technology학부,
[*]한국외국어대학교
e-mail: {gksm1215[O], sh119112[*], ysuhani98[*]}@gmail.com, yjmoon@hufs.ac.kr[**]

A Risk Prediction System of Air Pollution Influencing Diseases Utilzing Keras

Jisu Lee[O], Yu-jeong Lee[*], Soo-han Yoon[*], Yoo-Jin Moon[**]
[O]Division of Global Business & Technology, Hankuk University of Foreign Studies,
[**]Division of Global Business & Technology, Hankuk University of Foreign Studies,
[*]Hankuk University of Foreign Studies

● 요 약 ●

이 연구는 대기오염과 미세먼지의 각 성분이 질환에 미치는 영향에 대한 데이터만 존재한다면 어떠한 질환이든 위험도 예측 결과를 알 수 있는 것에 의미가 있다. 또한 기존의 대기정보에 따른 정보를 예상하는데 필요한 데이터 종류와 수가 많았으며 계산의 복잡성이 높았고 정보의 제공 범위가 넓었다. 하지만 이 연구는 과거 대기 데이터와 딥러닝을 통해서 낮은 비용으로 더욱 자세하게 유해질환 위험도를 예측하는 시스템을 구축하였다. 이 연구에서 구축한 시스템은 예측 결과 88.9%의 정확도를 보였다. 이 시스템은 입력되는 데이터의 정보에 따라 세분화된 지역의 대기환경 정보 또한 파악 가능하며 그 과정이 매우 간편하고 유용하다. 이 시스템은 공기질 예측을 위해 유용하게 사용될 수 있을 것이라고 사료된다.

키워드: 신경망(Neural Network), 인공지능(AI), 대기오염(Air Pollution), 케라스(Keras)

I. Introduction

전 세계의 대기오염 위성지도에서 한국과 중국이 가장 오염이 높은 것으로 드러났다. 이러한 상황을 바탕으로 이 시스템은 서울의 대기정보를 이용하여 공기의 각 성분마다 유해질환에 미치는 영향도에 대한 예측 데이터를 제공하는데 목적이 있으며 이를 통해 대기오염의 위험성에 대한 경각심을 제공하고자 한다.

이 프로젝트는 대기오염과 미세먼지의 각 성분이 유해질환에 미치는 영향에 대한 데이터만 존재한다면 어떠한 질환이든 위험도 예측 결과를 알 수 있는 것에 의미가 있다. 또한, 기존의 대기정보에 따른 정보를 예상하는데 필요한 데이터 종류/수가 많았으며 계산의 복잡성이 높았고 정보의 제공 범위가 넓었다. 하지만 이 시스템은 과거 대기 데이터와 딥러닝을 통해서 낮은 비용으로 더욱 자세하게 유해질환 위험도를 예측하는 시스템을 구축해 활용 가능하다. 이와 더불어 기존의 미세먼지 예측 위주 시스템이 담지 못한 수치 또한 종합적으로 예측할 수 있다.

II. Preliminaries

1. Related works

최근 미세먼지에 대해 다양한 연구가 진행되고 있다. 한양대학교 측에서 진행한 기존 프로젝트는 국내 미세먼지 수준을 예측하는 프로그램으로 데이터 예측을 위해 국내외 미세먼지, 풍향, 풍속 데이터를 사용하여 CRNN모델을 설계했다.

기존 프로젝트는 이 연구 프로젝트와 목적이 유사하지만, 결론적으로 얻고자 하는 결과가 다른 것을 알 수 있었다. 기존 프로젝트는 국내외 미세먼지, 풍향, 풍속 데이터 등 다양한 요소를 분석해 결론적으로 미세먼지 예측, 그 수준을 통해 위험도를 예측하고자 했으며 이 연구에서 진행한 프로젝트는 대기오염 물질의 각 성분 데이터를 이용하여 각 성분들이 질병에 미치는 영향을 파악하고 질병 발생률에 대한 위험도를 얻고자 했다. 따라서 이 연구의 인공지능 시스템은 이전에 없던 유용한 정보를 제공할 수 있을 것으로 예상된다.

2. Data sources

이 연구에서 사용하는 데이터는 서울열린데이터광장에서 제공하는 서울시 가간별 시간평균 대기환경 정보이다 [1]. 데이터는 측정일시, 지역코드, 권역명, 측정소코드, 측정소명, 미세먼지 1시간, 미세먼지 24시간, 초미세먼지, 오존, 이산화질소, 일산화탄소, 아황산가스 농도로 구성된 18,000개 데이터 셋이다.

III. The Proposed Scheme

1. Variables

미세먼지가 유해 질병에 미치는 영향에 대한 데이터를 얻기 위해 논문 [2]를 참고하였다. Table 1 자료를 참고하였으며 CO(일산화탄소), NO2(이산화질소), SO2(아황산가스), O3(오존), PM(미세먼지)가 대기오염으로 인한 심혈관 질환과의 통계적으로 유의한 양의 연관성을 나타내는 것을 파악할 수 있었다. 이를 기반으로 위험도= CO*1.039 + NO2*1.027 + SO2*1.028 + O3*1.012+ PM*1.002의 방법을 통해 미세먼지의 위험도 변수를 제작하였다. 따라서 이 연구에서는 5가지 요소를 독립변수로 설정하였고, 위험도 데이터를 종속변수로 설정하였다.

Table 1. Air Pollution Data

Pollutant [lag]	Relative risk (95% CI)	
	With Asian Dust Days	Without Asian Dust Days
(All-aged)		
CO [lag 1]	1.039 (1.027-1.051)	1.044 (1.031-1.057)
O_3 [lag 1]	1.012 (0.999-1.026)	1.011 (0.998-1.025)
PM_{10} [lag 3]	1.002 (0.995-1.009)	1.008 (0.997-1.018)
NO_2 [lag 2]	1.027 (1.014-1.039)	1.028 (1.015-1.040)
SO_2 [lag 1]	1.028 (1.017-1.040)	1.034 (1.022-1.046)

2. Data normalization

더 정확한 결과값을 추출하기 위해 기존의 각 성분의 데이터들을 -1과 1사이의 수로 정규화 변환하였다. 이후 각 성분과 영향도를 계산하여 위험 수준별 데이터를 제작하였다.

IV. Experiment

이 연구에서는 TensorFlow와 Keras를 활용하여 대기오염이 유해 질환에 미치는 위험을 예측하는 딥러닝 시스템을 구축하였다. 이 시스템 실행 결과, sigmoid 함수를 사용했을 때 정확도가 0.8892임을 확인하였다. 이에 노드와 은닉층, Loss Function, Optimizer 등 여러 값을 수정하고 비교해가며 원하는 가중치와 정확도 결과를 얻을 수 있었다.

결과적으로 테스트 정확도는 0.8892, 각 변수의 가중치는 실제 참고한 데이터와 매우 유사하게 나오는 것을 확인할 수 있었다.

V. Conclusions

본 시스템은 대기오염으로 인한 심혈관 질환 데이터 이외에도 대기오염으로 인한 사망률, 대기오염으로 인한 호흡기 질환 사망률 등 다양한 결론을 도출할 수 있다. 이를 통해 대기오염이 인체에 미치는 유해한 영향에 대한 정보와 위험도 등 대기오염에 대해 자세한 정보를 제공하고, 사람들로 하여금 대기오염과 미세먼지의 심각한 유해성에 대해 경각심을 주는 데 기여한다.

REFERENCES

[1] Seoul Open Data Plaza. 2021. https://data.seoul.go.kr

[2] Jiyoung Son, etc., “Analysis of Yellow Dust in Evaluating City Air Pollution - Focusing on Death and Death Rate in Seoul”, Journal of Korea Environment Health Society, Vol.35, No.4, 2009.