R기계학습 소프트웨어를 이용한 미세먼지 예측

장재호*
*일산대진고등학교
e-mail:pooh1210jjh@naver.com

Korean Dust Prediction using R Machine Learning Software

Jae-Ho Jang*
*Il-San Dae-Jin High School

요 약

최근, 한국에서는 사람들이 미세먼지로 많은 고통을 받고 있으며, 특히, 초미세먼지(PM2.5)의 경우에는 생성될 때, 화학적인 2차 반응에 의하여 생성되는 것으로 여겨지고 있다. 본 논문에서는 R에서 제공하는 기계학습 프로그램을 이용하여 초미세먼지를 예측하기 위한 실험을 진행하였다. R소프트웨어는 빅데이터 및 통계 분석을 위해서 많이 사용되고 있는 프로그램이다. 최근에는 인공지능의 기계학습을 위한 기능도 제공하고 있는데, 데이터 예측을 할 때, 사용하면 매우 유용하다.

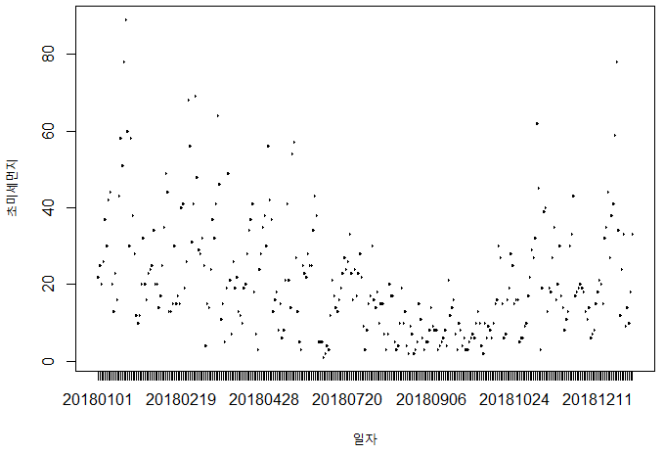
1. 서론

초 미세먼지는 사람들에게 매우 안 좋은 영향을 끼치고 있며, 최근 들어, 국내에서는 사람들이 미세먼지로 많은 고통을 받고 있다. 초미세먼지(PM2.5)의 경우에는 생성될 때, 화학적인 2차 반응에 의하여 생성되는 것으로 여겨지고 있다. 국내의 미세 먼지 예측은 온도, 습도, 바람, 구름과 같은 기상 환경과 중국의 미세 먼지 수치를 고려하여 기상청에서 매일 예보를 하고 있다. 초 미세먼지 수치가 높을 것으로 예상될 경우, 사람들은 외부 활동을 자제하거나 차량의 운행을 자제하는 등의 활동을 통해, 건강에 미치는 영향을 최소화 하게 된다. 따라서, 정확한 미세 먼지 예측은 아주 중요하다고 할 수 있다 [1].

최근 들어, 이러한 미세 먼지 예측을 위해, 기계 학습을 이용하는 연구가 진행 되고 있다. 기계 학습을 위해서는 인공 신경망을 정의하고 이를 훈련시킨 후, 예측에 사용해야 한다. 인공 신경망 훈련을 위해서는 과거의 데이터가 필요 하다. 초 미세 먼지의 경우, 10년 전부터 시간 단위로 측정하여, 자료를 제공하고 있기 때문에, 신경망 훈련을 위한 많은 데이터가 축적되어 있다. 현재, 기계 학습 기능을 제공하는 도구는 매우 많이 나와 있다. R은 통계 분석과 데이터 분석을 위한 도구로서 많이 사용되고 있으며, 기계학습을 위한 패키지도 제공학고 있다[2]. 본 논문에서는 R이 제공하는 기계 학습 패키지는 이용하여, 초미세먼지 예측 실험을 진행 하였다.

2. 초미세먼지 데이터 수집 및 신경망 훈련

본 연구에서는 기계학습에 사용될 초미세 먼지 데이터를 일별로 1년간의 데이터를 수집하였다. 물론 PM10도 수집하면, 좋겠지만, 시간이 너무 오래 걸려서 초미세먼지 PM2.5데이터만 일별로 수집하여, 엑셀로 저장하였다. 미세먼지데이터는 환경부 홈페이지 또는 공공 데이터 정부 제공 사이트를 방문하여 다운로드 하였다. 데이터 수집 기간은 2018년 1월1일부터 2018년 12월31일까지 이다.. 일별 데이터의 분포는 <그림 1> 과 같다.



<그림 1> 인공 신경망 훈련을 위한 일별초미세먼지 데이터 분포

수집된 데이터를 입력 데이터로 사용하여 신경망을 훈련시키고 최대 5일간의 초미세먼지를 예측하는 실험을 진행하였다. 신경망 기반 예측의 정확성을 높이기 위해서는 최적의 신경망 모형을 정의해야 한다. 인공 신경망 모형 정의는 입력층의 노드 개수, 은닉층의 노드 개수, 출력층의

노드 갯수를 정의하여 이루어진다. 구조가 다른 여러 개의 인공 신경망 모형을 정의 하고 각 모형에 대한 훈련, 예측 실험을 통해 가장 정확성이 높고 훈련 시간이 적게 걸리는 모형을 최적 모형으로 선정한다. 최적의 신경망 모형을 찾아내는 과정은 총 5단계로 이루어 진다. 1단계에서는 경험에 기반 하여, 인공 신경망 모형을 정의한다. 2단계에서는 수집된 훈련용 입력 데이터를 읽는다. 3단계에서는 수집된 데이터에서 정답 데이터를 선정 하여, 읽는다. 4단계에서는 신경망을 훈련 시킨다. 5단계에서는 훈련된 신경망을 이용하여 값을 예측한다. 6단계 예측된 값과 해당 기간의 정답 데이터를 비교하여, 정의된 인공 신경망에 대한 에러값을 구한다. 본 실험에서는 8개 정도의 신경망 모델을 구성한 후, 각 신경망을 입력 데이터로 훈련을 시켰다. 이후, 훈련된 신경망을 이용하여, 가장 최근 5일 이전의 데이터를 입력 데이터로 사용하여, 최근 5일간의 초미세먼지 양을 예측 하도록 하였다. 예측된 수치와 가장 최근 5일간의 실제 데이터와 비교하여, 각 모델의 정확성을 평가하였다. 예측의 정확도를 평가하기 위해서 MAPE(Mean Absolute Percentage Error)값을 사용하였다. MAPE는 아래의 공식을 이용하여 계산한다.

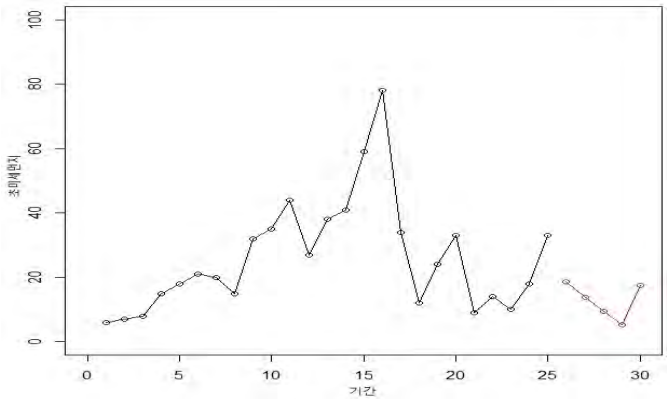
MAPE = [(|예측값(n)-측정값(n)|/측정값(n))/측정갯수] * 100 (%)

실험에서 구성된 모델과 각 모델의 MAPE 계산 결과값은 <표 1>과 같다.

〈표 1〉 실험에서 시험한 신경망 모델 및 에러값

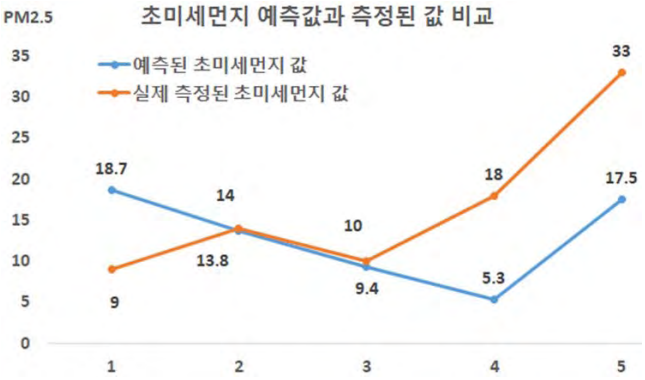
	입력층 노드 갯수	은닉층 노드갯수	출력층 노드갯수	훈련에 걸리는 시간	에러 (MAPE)
모형1	5	10	5	1초	73.56
모형2	5	20	5	1초	93.79
모형3	10	20	5	1.2초	81.16
모형4	10	30	5	1.5초	78.96
모형5	15	30	5	12초	46.29
모형6	15	50	5	17초	43.27
모형7	20	40	5	훈련불가능	측정불가
모형8	20	60	5	훈련불가능	측정불가

위의 결과를 분석해 보면, 입력 노드 수를 20으로 설정할 경우, R에서 nnet함수를 수행하는 도중에 에러가 발생하여, 훈련작업을 수행할 수 없었다. 따라서, 모형 1에서 모형 6중에 하나를 선택하는 수밖에 없었다. 본 실험에서는 모형 5을 선택하였다. 모형 중에서 에러값이 두 번째로 작고 훈련시간도 비교적 적게 걸렸기 때문이다. 훈련된 모형 5를 이용하여, 5일 간의 미세먼지 예측을 수행하였다. 예측 결과는 <그림 2>와 같다. 그림에서 붉은색으로 표시된 부분이 예측된 초미세 먼지 수치 값이다. 예측 이전에 검은색으로 된 미세먼지 값은 5일 이전 25일 동안의 실제 미세 먼지 값을 의미 한다.



<그림 2> 훈련된 인공 신경망을 이용한 미세먼지 예측 결과

<그림 3>은 5일간에 대한 초미세 먼지 예측값과 실제 미세 먼지 값과 비교한 결과이다. 5일중 이틀, 삼일은 예측값과 실제값이 거의 비슷하게 나왔으나, 1,3,5일은 차이가 많이 나는 것을 알 수 있다.



<그림 3> 초미세 먼지 예측값과 실제값 비교

3. 결 론

본 연구에서는 R이 제공하는 신경망 학습 및 예측 기능을 사용하여, 초미세먼지 예측 실험을 수행해 보았다. 실험에서는 서울시 강남구에서 365일 동안 측정한 일일 평균 미세먼지 데이터를 사용 하였다. 신경망을 훈련시킨 후, 최근 5일 동안의 미세먼지 측정을 해본 결과, 5일중 이틀간의 미세 먼지 예측결과는 실제와 유사하게 나왔으며, 3일간의 예측결과는 많은 차이가 났다.

참고문헌

[1] 김선영, "미세먼지 대기오염과 건강영향", 한국환경독성학회 2014 춘계국제학술대회, 2014, pp.154-175.

[2] R을 활용한 기계 학습, 브레트 란츠 지음, 전철욱 옮김, 에이콘출판사, 2014년.