수치 모델링 및 머신러닝을 이용한 대기 오염 예측

수학과 오서영, 신영민

미세먼지는 아황산가스, 질소 산화물, 납, 오존, 일산화탄소 등을 포함하는 대기오염물질이다. 이러한 미세먼지의 농도가 증가할수록 전체 사망률과 심장 혈관 및 호흡기계 질환으로 인한 사망률이 증가한다. 그렇기에 미세먼지 농도에 대한 정보는 국민의 주요 관심사이며 더불어 장기적인 미세먼지 예측이 크게 중요해짐에 따라 대기오염 예측 모델 개발 추진이 필요하게 되었다.

대기오염 예측 모델을 만들기 위해 크게 두 가지 방법론을 사용하고자 한다. 첫 번째는 '수학적 모델링'을 통한 방법이다. 수학적 모델링은 주어진 상황에 맞게 변수를 설정하여 함수를 제작하는 것이다. 다시 말해, 주어진 문제를 수학적인 구조로 정리하게 되는 것이다. 우리가 베이스로 사용할 모델은 '대류확산 방정식' 인데, 이는 말 그대로 대류방정식과 확산방정식의 결합으로 이루어진 식으로, 입자나 에너지 등과 같은 물리량의 대류와 확산으로 인한 움직임을 표현하는 방정식이다. 여기서 대류란 바람 등 외부 힘이나 움직임에 의해 물리량이 움직이는 것을 말하며, 확산이란 밀도나 농도 차이 등에 의하여 물질을 이루는 입자 등이 농도가 높은 곳에서 낮은 쪽으로 퍼져나가는 것을 의미한다. 이 대류확산 방정식이 실질적으로 공기질 모델링에 많이 쓰인다고 하여, 미세먼지 예측 모델링에 활용하고자 한다. 이 방정식은 해석적인 방법으로 정확한 해를 구하기 어렵기 때문에, '수치적 해법'을 사용해야한다. 중앙 차분(centered difference) 과 명시적 유한차분법 (explicit finite difference method) 을 노이만 경계조건과 함께 사용하여 기본적인 모델을 만들어 냈고, 대한민국 풍속, 풍향 데이터와 미세먼지 데이터를 모델에 반영했다.

두 번째로 사용한 방법론은 머신러닝이다. 기계학습이라고도 불리는 머신러닝은 데이터를 입력 받아 인간의 도움 없이 컴퓨터가 스스로 새로운 규칙을 학습하는 것이다. 사람이 직접 함수를 제작하는 수학적 모델링과 달리 머신러닝은 데이터를 통한 학습으로 최적화된 함수를 만들어낸다. 수학적 모델링에서 사용한 풍속, 풍향 데이터와 미세먼지 데이터를 활용한 회귀분석 (Regression), 또는 순환 신경망 (Recurrent Neural Network) 등을 활용할 것이다. 순환 신경망은 유닛간의 연결이 순환적 구조를 갖는 특징을 갖고 있는데, 이러한 구조는 시변적 동적 특징을 모델링 할 수 있도록 신경망 내부에 상태를 저장할 수 있게 해준다. 우리는 2019년 4월 5일 하루 24시간에 대한 미세먼지 농도를 예측할 것이기 때문에 순환 구조를 가진 모델이 적합하다고 판단했다.

방법론과 상관없이 우리는 데이터를 활용한다. 공간에 대한 통계자료가 필요할 때 가장 좋은 방법은 모든 지점에서 필요로 하는 자료를 직접 획득하는 것이다. 그러나 비용과 시간적인 한계가 있어 모든 지점에서 원하는 값을 얻는다는 것은 현실적으로 불가능하다. 이러한 문제를 해결하기 위해 우리는 '공간 보간법'을 사용했다. 보간법은 실제로 많은 특정지점을 선정하여 관측 값을 얻은 후, 이 데이터를 이용하여 알고자하는 지점의 값을 예측하기 위해 많이 사용되고 있으며, 우리는 Inverse Distance Weighted 보간법을 사용하는 것이 미세먼지 데이터에 적합하다고 판단했다.

여러 모델을 통한 예측결과를 분석하고 비교하여 더 나은 모델링을 수행하고, 더 나아가 많은 사람들이 이러한 예측을 활용할 수 있도록 어플리케이션 또는 웹에 모델을 연동시키는 방안도 고려해볼 예정이다. 이러한 예측을 잘 활용하면 데이터를 측정하는데 필요한 비용을 줄일 수 있으며, 실시간으로 미세먼지 농도가 변화하는 양상을 누구나 쉽게 확인 할 수 있다.

[참고자료]

[1] 김영록, 『산업응용수학의 기본』 , 경문사(2017), p111-119

[2] "보간법 (Interpolation)", 네이버 블로그, https://m.blog.naver.com/ceoyangsj/100169531489