The Korean Journal of Applied Statistics (2018) DOI: https://doi.org/10.5351/KJAS.2018.31.2.265
31(2), 265–285

Prediction of fine dust PM_{10} using a deep neural network model

Seonghyeon Jeon[a] · Young Sook Son[a,1]

[a]Department of Statistics, Chonnam National University

(Received February 5, 2018; Revised March 1, 2018; Accepted March 14, 2018)

Abstract

In this study, we applied a deep neural network model to predict four grades of fine dust PM_{10}, 'Good, Moderate, Bad, Very Bad' and two grades, 'Good or Moderate and Bad or Very Bad'. The deep neural network model and existing classification techniques (such as neural network model, multinomial logistic regression model, support vector machine, and random forest) were applied to fine dust daily data observed from 2010 to 2015 in six major metropolitan areas of Korea. Data analysis shows that the deep neural network model outperforms others in the sense of accuracy.

Keywords: fine dust PM_{10}, neural network, multinomial logistic regression, support vector machine, random forest, deep neural network, accuracy

1. 서론

National Institute of Environmental Research (2016)에 의하면 미세먼지(particulate matter; PM)는 입자의 크기에 따라 두 가지로 분류가 되는데 직경이 1000분의 10mm보다 작은 먼지를 PM_{10}이라 하고, 직경이 1000분의 2.5mm보다 작은 먼지를 $PM_{2.5}$라고 한다. PM_{10}은 대기 중으로 직접 방출되거나 생성되는 1차 오염물질에 속하고 $PM_{2.5}$는 오염물질이 대기 중에서 2차 반응하여 생성되는 2차 오염물질에 속한다. 미세먼지는 입자가 매우 작아 인체의 기관에서 걸러지지 못하고 몸속까지 스며들게 되는데 이때 몸속에서 세포와 염증반응이 발생하여 여러 질환 등이 유발될 수 있다. 2013년 10월 세계보건기구 산하 국제암연구소(International Agency for Research on Cancer)는 미세먼지를 인간에게 암을 일으키는 것으로 확인된 1군 발암물질로 분류할 정도로 미세먼지는 인체에 치명적인 대기오염물질이다.

Korean Ministry of Environment (2016)에 의하면 우리나라의 미세먼지 PM_{10} 오염도는 2001년부터 2006년까지는 보통 수준인 51–61$\mu g/m^3$ 사이를 오르내렸지만 수도권 대기환경관리 기본계획(2005-2014년)의 시행과 더불어 2007년부터 약간의 감소 추세로 돌아섰다가 2016년 들어 대기질 개선이 다소 정체되는 모습을 보이고 있다고 한다. 또한 2014년도의 시도별 미세먼지 현황을 확인해보면 연평균 환경기준인 50$\mu g/m^3$을 초과하는 지역이 17개 지역 중 4개로 나타났다. 2012년도부터 2014년도까지 세계 주요도시의 미세먼지 농도를 비교한 결과를 보면 서울의 미세먼지 농도는 미국 LA의 1.5배, 프랑스

[1]Corresponding author: Department of Statistics, Chonnam National University, 77, Yongbong-ro, Buk-Gu, Gwangju 61186, Korea. E-mail: ysson@jnu.ac.kr

파리의 2.1배, 영국 런던의 2.3배로 높게 나타났는데 이는 높은 인구밀도, 고도의 도시화 진행, 지리적 위치, 그리고 기상여건 등이 유리하지 않은 원인으로 파악되었다.

최근 중국의 황사와 미세먼지로 인한 대기오염에 관련된 뉴스와 기사가 많이 등장하고 있는데 구글 트렌드(Google Trends)의 '미세먼지' 검색어의 관심도 변화를 보면 2013년도 후반부터 급격하게 관심이 늘어나는 것을 볼 수 있고 2017년에 최고치를 기록하였다. 이는 중국의 황사와 미세먼지가 건강에 미치는 위험성이 크게 부각되면서 나타난 결과라고 할 수 있다. 중국 대부분 지역에서의 미세먼지 농도는 2012년도에 약 80μg/m^{3}에 육박하였고 2012년 상반기 베이징의 평균 $PM_{2.5}$ 농도는 124μg/m^{3}로 서울의 4배를 보였다. 이러한 높은 수치의 미세먼지 농도를 보이는 중국의 대기가 북서풍이 부는 겨울철과 황사가 잦은 봄철에 주로 한국의 대기에 유입되어 영향을 미치게 된다.

2013년도 이후로 수도권을 중심으로 미세먼지의 발생이 현저해짐에 따라 국민들의 관심과 우려가 늘어나게 되었고 이에 따라 환경부에서는 2015년 11월부터 전국을 18개 권역으로 세분화하여 내일의 미세먼지 예보결과를 '좋음, 보통, 나쁨, 매우 나쁨'의 4개 범주로 매일 4회(오전 5시/11시, 오후 5시/11시) 국민들에게 제공하고 있다. 이제 국민들은 매일 발표되는 미세먼지의 예보에 따라 외출, 환기, 실외활동 등과 같은 여러 일상사를 결정하게 되었다. 따라서 미세먼지의 예보는 매우 중요하게 되어 이에 관련된 연구들도 늘어나고 있다.

Koo 등 (2010)은 2006년 1월 1일부터 2009년 5월 31일까지 서울 남동지역을 대상으로 신경망, 회귀, 의사결정모형을 사용하여 미세먼지 통계예보모형을 개발하였다. 전일 17시에 익일 미세먼지 평균 농도를 예보하는 내일 예보모형과 당일 오전 9시에 확정 예보하는 당일 예보모형으로 나누어 분석하였다. 예보변수로는 대기질 측정자료, 기상자료, Mesoscale Model 5 (MM5)로 계산한 예보자료를 사용하였고 미세먼지의 농도를 수치에 따라 5개의 범주로 구분하여 반응변수로 사용하였다. 위의 데이터로 모형을 학습시키고 2010년 4월 1일부터 7월 31일까지 실제 예보시스템을 운영하여 예보정확도를 평가하여 당일 예보모형의 지수 일치도는 80.8%, 거짓 경보율은 12.5%, 감지 확률은 77.8%를 보였고 내일 예보모형은 지수 일치도는 72.4%, 거짓 경보율은 0.0%, 감지 확률은 42.9%를 보였다.

Lee (2011)는 2003년부터 2009년까지 평택시 비전동 관측소의 시간별 PM_{10} 데이터에 대해 자기회귀오차모형을 적합시켜 대기 오염물질인 오존(O_3), 이산화질소(NO_2), 일산화탄소(CO), 아황산가스(SO_2)의 농도와 기상 요소인 일 최고온도, 풍속, 상대습도, 강수량, 일사량, 운량 등과 같은 예측변수들이 미세먼지 PM_{10}의 농도에 어떤 영향을 주는지를 해석하였다.

Lee 등 (2014)은 2008년부터 2011년까지 전국 16개 시도 지역에서 관측한 일평균 미세먼지 데이터에 대하여 시간 및 공간에 대한 상관관계를 동시에 고려한 희박 벡터자기상관회귀 모형을 사용하여 적합의 타당성을 검증하였다. 일별 평균 미세먼지 농도에 대해서 장기간 종속성을 나타내는 것을 보였고 벡터자기회귀모형에 비해 더 간결하면서 좋은 예측력을 보였다.

Kwon 등 (2015)은 2011년 1월부터 2014년 6월까지 서울시 종로구에서 관측한 미세먼지 데이터에 대하여 분위수 부스팅에 의한 예측 모형 성능을 평가하였다. 익일의 미세먼지 PM_{10}을 100μg/m^{3}보다 크면 1로, 작으면 0으로 설정하여 반응변수로 사용하였고 예측변수로는 대기오염 물질에 속하는 O_3, NO_2, CO, SO_2의 농도, 기상요소에 속하는 평균기온, 최고기온, 최저기온, 평균풍속, 최대풍속, 순간최대풍속, 기압, 수증기압, 일조량, 운량, 강우, 습도, 황사 여부 등을 사용하였다. 현업에서 사용하는 의사결정나무 모형과 비교하여 $\tau = 0.60$일 때의 분위수 부스팅 모형에서 예측정확도, 감지확률이 각각 2.6%p, 68.9%p 높일 수 있음을 보였고 적절한 τ 값을 설정해 효율적인 예보가 가능하다는 것을 밝혔다.

Lee 등 (2017)은 2015년 8월 24일부터 2016년 8월 23일까지 결측 일을 제외한 360일의 서울시 25개 구 데이터에 대하여 공간패널모형을 사용하여 $PM_{2.5}$ 농도에 영향을 끼치는 유의미한 변수들을 알아보

고 호흡기 환자수를 파악하는데 $PM_{2.5}$ 농도 예측이 중요함을 보였다. 공간패널모형에 사용된 예측변수는 기상 요소에 속하는 일별 평균기온, 최저기온, 최고기온, 평균습도, 최저습도, 최고습도, 평균풍속, 최고풍속, 평균 강수량, 풍향과 대기오염 물질에 속하는 PM_{10}, NO_2, O_3, CO, SO_2이다. 분석 결과, 일반화 선형모형에서 SO_2, NO_2, O_3, 최고습도, 최고풍속, 평균풍속, 평균 강수량, 평균풍속, 계절성 변수들이 유의하였고 공간패널모형에서 또한 모두 유의함을 보였다. 그리고 공간패널모형에서 지역의 인접성에 의해 주변 지역의 $PM_{2.5}$ 농도에 서로 영향을 받는다는 것을 밝혔다.

이상과 같이 기존의 미세먼지 농도 예측 연구에서는 주로 대기오염 물질 자료와 기상 자료 등을 예측변수로 사용하여 분석을 하였는데, 이는 국외에서 발생한 대기오염 물질의 영향력을 반영하지 못한다. Korean Ministry of Environment (2016)에 의하면 2014년도 고농도 미세먼지 발생사례를 분석한 결과, 국외 영향이 가장 컸던 날은 강한 북서풍에 황사가 남부지역을 중심으로 유입되어 전체 PM_{10} 중 80% 이상이 국외 영향인 것으로 나타났고, 황사가 없는 날로서 외부 영향이 가장 높았던 날은 중국 북동지방에 위치한 고기압으로 인해 국외 대기오염물질 영향이 74% 수준까지 나타났다. 이와 같이 국외 영향에 의한 미세먼지 농도의 비율이 높다는 분석 결과가 존재하고 언론 매체에서도 중국에서 발생하여 우리나라로 넘어오는 황사의 영향에 대해 심각하게 다루는 추세이다. 따라서 우리나라 미세먼지 농도에 직접적인 영향을 주는 중국의 미세먼지 농도를 예측변수로 고려하는 것이 바람직할 것이다.

본 연구에서는 환경부에서 발표하는 미세먼지 PM_{10} 농도의 4가지 범주인 '좋음, 보통, 나쁨, 매우 나쁨'의 익일 예측 뿐만 아니라 2가지 범주인 '좋음 혹은 보통, 나쁨 혹은 매우 나쁨'의 익일 예측에 목표를 둔다. 예측변수로는 기존의 연구들에서 주로 사용하였던 대기오염 물질 및 기상 자료와 관련된 예측변수 외에 전일 및 이틀 전 중국의 미세먼지($PM_{2.5}$) 농도와 계절변수를 예측변수에 추가하였다. 예측모형으로 사용된 심층 신경망모형(deep neural network model)에 대한 다양한 실험을 통하여 기존의 분류기법들인 신경망모형(neural network model; NN), support vector machine (SVM), 다항 로지스틱 회귀모형, 그리고 random forest (RF) 기법과 비교하여 정확도 측면에서 보다 우수한 결과를 얻을 수 있었다.

2. 기술통계분석

Table 2.1은 자료분석에서 사용한 변수들을 정의한 표로서 각 변수들은 2010년 1월 1일부터 2015년 12월 31일까지 6년동안 수집한 일별 데이터이다. 대기오염 물질 데이터는 전국의 시군구별 177개 대기오염 물질 측정소들 중 6개 대도시 지역인 서울특별시 강남구, 부산광역시 해운대구, 인천광역시 부평구, 대전광역시 서구, 광주광역시 북구, 대구광역시 달서구를 선정하여 시간별 미세먼지 PM_{10}, SO_2, O_3, NO_2, CO를 일별로 평균하여 일별 데이터를 만들었다. PM_{10}의 경우 과거부터 관측되어져 왔지만 $PM_{2.5}$는 2015년 1월 1일부터 대기환경기준이 시행되었기 때문에 2015년도 이전에는 관측되지 않았다. 따라서 본 논문에서는 $PM_{2.5}$ 대신 PM_{10}을 사용하였으며, 연속형 변수인 PM_{10}을 Table 2.2와 같이 환경부에서 예보하는 4개의 순서형 범주인 좋음(Good), 보통(Moderate), 나쁨(Bad), 매우 나쁨(Very Bad) 범주로 나눈 다항 범주형 반응변수로 변환하여 사용하였다. 대기오염 물질 데이터는 한국환경공단(www.airkorea.or.kr)의 대기환경 측정자료를 활용하였다.

기상 관련 데이터는 기상자료개방포털(data.kma.go.kr)에서 방재기상관측장비로 관측한 일별 평균기온(meanTemp), 최저기온(minTemp), 최고기온(maxTemp), 최대풍속(maxWind), 평균풍속(meanWind), 일강수량(Rain)을 사용하였다.

서로 다른 측정소에서 측정된 대기 관련 변수와 기상 관련 변수를 데이터로 활용하기 위해서 대기오염 물질 측정소 및 기상관측 측정소의 경도 및 위도를 이용해 서로 가장 가까운 지역의 거리를 계산하여 매

Table 2.1. Definition of variables

Variable type	Variable name	Variable description
Air pollutants	PM_{10}	Particulate matter($< 10\mu m$) of the day
	PM_{10y}	Particulate matter($< 10\mu m$) of the previous 1 day
	SO_2	Sulfur dioxide of the previous 1 day
	O_3	Ozone of the previous 1 day
	NO_2	Nitrogen dioxide of the previous 1 day
	CO	Carbon monoxide of the previous 1 day
	ChinaPM1	China's Particulate matter($< 2.5\mu m$) of the previous 1 day
	ChinaPM2	China's Particulate matter($< 2.5\mu m$) of the previous 2 day
Meteorological elements	meanTemp	The average temperature of the previous 1 day
	minTemp	The lowest temperature of the previous 1 day
	maxTemp	The highest temperature of the previous 1 day
	meanWind	The average wind velocity of the previous 1 day
	maxWind	The highest wind velocity of the previous 1 day
	Rain	no rain(0) or rain(1) of the previous 1 day
Other	Season	Spring(3–5), Summer(6–8), Autumn(9–11), Winter(12–2)

Table 2.2. Grades of PM_{10} used as a categorical response variable (uint: $\mu g/m^3$)

Grade	Good	Moderate	Bad	Very Bad
PM_{10}	0–30	31–80	81–150	>151

칭시켰다. 따라서 대기오염 물질 데이터의 측정소 지역을 기준으로 가장 가까운 기상 측정소에서 측정된 데이터가 기상 변수로써 활용되었다.

익일의 미세먼지를 예측하기 위해서 대기오염 물질 변수와 기상 변수 모두 전일 변수 값들을 예측변수로 사용하였다. 중국 미세먼지(ChinaPM1, ChinaPM2)의 경우는 미국 국무부 대기 질 모니터링 프로그램 사이트(www.stateair.net)에서 제공하는 베이징 관측소의 시간별 $PM_{2.5}$ 데이터를 일자별로 평균하여 사용하였다. 베이징 지역의 $PM_{2.5}$를 사용한 이유는 중국의 관측소들 중 우리나라와 비교적 가까운 위치에 있는 지역이면서 계절풍에 의한 영향을 가장 많이 주는 중국 북동지역에 속하기 때문이다. 또한 당일의 중국 미세먼지가 당일 바로 한국에 영향을 주기보다는 중국의 대기가 한국으로 넘어오는 시간이 존재하기 때문에 전일 및 이틀전의 중국 미세먼지 농도를 사용하였다.

기상 요소들 중 강수량 변수의 경우는 대부분의 값이 0으로 비가 오지 않은 날이 압도적으로 많으며, 비가 온 날의 경우도 강수량 값의 범위가 커서 분산이 커질 수 있기 때문에 강수 여부(Rain)를 나타내는 이항 범주형 변수를 사용하였다.

계절에 따라 우리나라에 불어오는 바람 방향이 바뀌고 대기의 순환이 달라질 수 있다. 따라서 대기오염 물질 관측 데이터의 각 시점에 따라 봄(3월부터 5월까지), 여름(6월에서 8월까지), 가을(9월에서 11월까지), 겨울(12월에서 2월까지)을 나타내는 범주형 계절변수(Season)를 예측변수로 사용하였다.

Table 2.3은 각 지역별 기술통계 분석 결과이다. 각 지역별 미세먼지 PM_{10} 농도의 평균을 보면 부산이 약 35.7로 가장 낮은 농도를 보이고 인천이 약 50.7로 가장 높은 농도를 보인다. 전반적으로 중국과 가장 가까운 인천 및 서울 지역의 미세먼지 농도가 높게 나타나고 중국과 가장 먼 지역인 부산에서는 미세먼지 농도가 낮게 나타나는 것을 알 수 있다. 중국 베이징의 경우 우리나라 6개 대도시에 비해 미세먼지 농도의 평균 및 표준편차가 대략 2배 이상으로 높다.

평균의 관점에서 보자면 우리나라 6개 대도시의 미세먼지 농도는 환경부가 정의한 보통(Moderate)과

Table 2.3. Descriptive statistics by area

Variable		Area						
		Total	Seoul	Busan	Incheon	Daejeon	Gwangju	Daegu
PM_{10}	Mean	44.5525	47.3599	35.6818	50.6964	41.9032	46.0341	45.5397
	SD	26.4438	30.4616	21.3227	28.4316	24.7637	26.9613	22.3526
PM_{10y}	Mean	44.5411	47.3451	35.6659	50.6934	41.8940	46.0216	45.5268
	SD	26.4437	30.4590	21.3187	28.4324	24.7655	26.9624	22.3514
SO_2	Mean	0.0049	0.0053	0.0064	0.0076	0.0030	0.0036	0.0034
	SD	0.0027	0.0019	0.0029	0.0024	0.0014	0.0013	0.0023
O_3	Mean	0.0240	0.0185	0.0324	0.0182	0.0225	0.0264	0.0260
	SD	0.0130	0.0121	0.0122	0.0100	0.0125	0.0112	0.0137
NO_2	Mean	0.0266	0.0382	0.0230	0.0368	0.0220	0.0222	0.0164
	SD	0.0135	0.0139	0.0114	0.0124	0.0090	0.0078	0.0093
CO	Mean	0.5387	0.4742	0.6000	0.6255	0.5290	0.5733	0.4141
	SD	0.2214	0.1846	0.2137	0.2447	0.2053	0.1720	0.2297
ChinaPM1	Mean				95.9066			
	SD				77.6848			
ChinaPM2	Mean				95.9045			
	SD				77.7022			
meanTemp	Mean	12.9913	12.2643	15.1776	12.3879	12.3156	13.4022	12.8787
	SD	9.8708	10.7301	7.4860	9.8469	10.4493	9.6546	9.9976
minTemp	Mean	8.5447	7.4713	13.0156	8.8322	6.9737	8.7032	7.2619
	SD	10.5560	11.1102	8.0904	9.9862	11.2160	10.2507	10.8748
maxTemp	Mean	18.0480	17.7312	17.3753	16.4953	18.3474	18.7640	19.4189
	SD	9.9134	11.0127	7.1387	10.3066	10.2649	9.7690	9.7384
meanWind	Mean	2.2237	1.5905	4.6037	1.9728	1.9304	2.0728	1.7326
	SD	1.4925	0.8405	2.1467	0.9454	1.0167	1.0032	0.7363
maxWind	Mean	5.3500	4.2718	9.3223	4.4470	5.0997	4.9578	4.9357
	SD	2.5450	1.5024	3.3676	1.5540	2.0122	1.5806	1.74659
Rain	Freq(0)	9,766	1,621	1,794	1,684	1,589	1,456	1,622
	Freq(1)	3,373	569	396	506	601	734	567

나쁨(Bad)의 경계 값인 $80\mu g/m^3$ 보다는 훨씬 낮지만 표준편차가 매우 큰 편이다. 각 지역별 미세먼지 PM_{10} 농도의 상자그림을 나타낸 Figure 2.1을 보면 나쁨(Bad) 수준인 $80\mu g/m^3$ 이상인 날들이 꽤 존재한다. 각 지역 및 미세먼지 등급별 빈도표를 나타내는 Table 2.4를 보면 나쁨(Bad) 및 매우 나쁨(Very Bad)은 전체 7.5%이고 인천이 11.6%로 가장 높고 부산이 2.6%로 가장 낮은 편이다.

미세먼지의 예측에서 좋음(Good) 혹은 보통(Moderate)의 예측 오류보다 나쁨(Bad) 혹은 매우 나쁨(Very Bad)의 예측 오류가 발생했을 때 위험이 보다 크기 때문에 이 등급들의 정확한 예측이 필수적이다. 그러나 전체 자료 중 7.5%를 점유하는 나쁨(Bad) 및 매우 나쁨(Very Bad)의 분포는 매우 불균형적인 데이터를 구성하고 있음을 알 수 있다.

미세먼지를 제외한 대기오염 물질 변수 중에서 아황산가스(SO_2) 및 일산화탄소(CO) 농도는 인천, 오존(O_3) 농도는 부산, 이산화질소(NO_2) 농도는 서울이 가장 높았다. 기상 관련 변수들을 살펴보면 부산을 제외하고는 대체로 기상적인 특성이 비슷하다. 부산의 경우는 다른 지역에 비해 평균기온과 최저기온이 높고 비가 내리는 날이 적으며 평균풍속, 최대풍속의 값이 매우 높은 편이다.

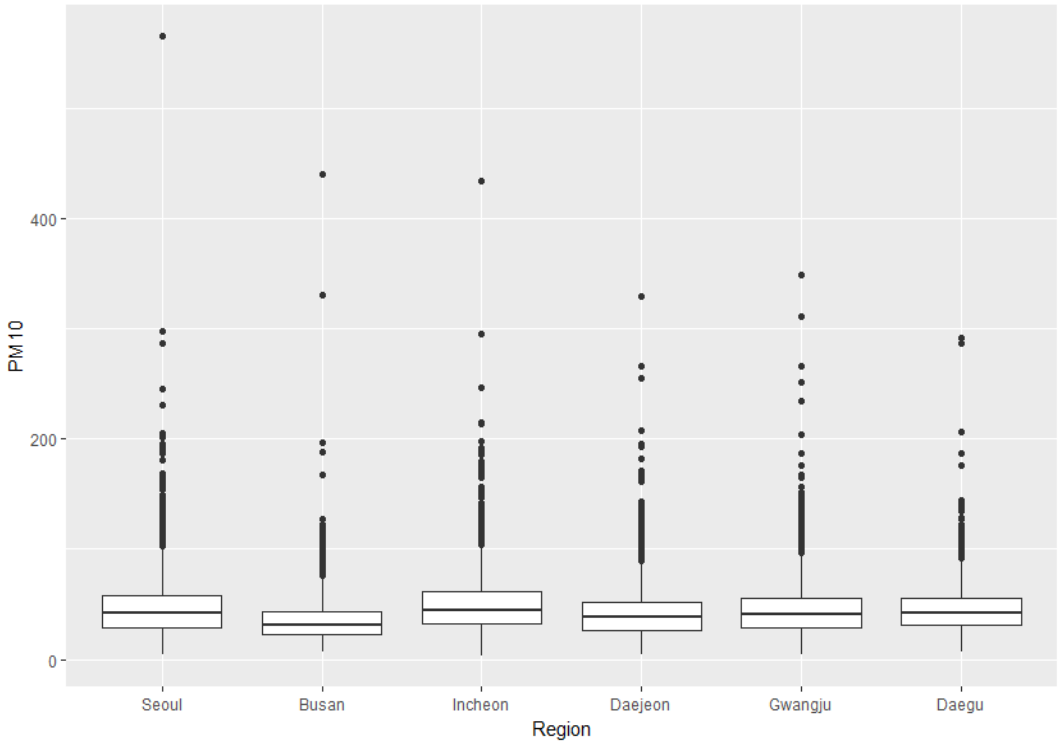


Figure 2.1. Box plots of PM_{10} by area.

Table 2.4. Frequency table by PM_{10} grade and area

Grade	Area						
	Total	Seoul	Busan	Incheon	Daejeon	Gwangju	Daegu
Good	3811(30.0%)	594(27.3%)	1019(47.8%)	445(20.4%)	715(33.3%)	607(27.7%)	431(23.0%)
Moderate	7947(62.5%)	1372(63.1%)	1058(49.6%)	1488(68.1%)	1310(60.9%)	1391(63.6%)	1328(70.8%)
Bad	867(6.8%)	181(8.3%)	52(2.4%)	230(10.5%)	114(5.3%)	177(8.1%)	113(6.0%)
Very Bad	83(0.7%)	26(1.2%)	5(0.2%)	23(1.1%)	11(0.5%)	13(0.6%)	5(0.3%)

3. 미세먼지와 예측변수들 간의 관계

제3절에서는 예측변수들과 반응변수인 미세먼지 농도 변수(PM_{10}) 와의 관계를 알아보고자 한다. Table 3.1은 PM_{10}과 각 예측변수들 간의 상관계수, 그리고 PM_{10} 변수의 4개의 범주에서 각 예측변수들의 평균값을 보여준다. 상관분석에서는 유의수준 5% 하에서 모든 상관계수들은 유의한 결과를 보였고, PM_{10} 변수의 4개 범주별 각 예측변수의 차이에 대한 ANOVA 검정에서도 유의수준 5% 하에서 모두 유의한 결과를 보였다.

상관계수들을 보면 전날 미세먼지 농도 변수(PM_{10y})와의 상관계수가 0.6179로 PM_{10}과 가장 선형 관련성이 높은 변수로 나타났다. 그 다음으로 약 0.3 정도의 상관계수을 갖는 NO_2, CO가 PM_{10}과 양의 선형 관련성이 약하게 존재함을 보였고 O_3의 경우는 매우 선형성이 낮은 0에 가까운 값으로 나타났다. 1일 전 중국 미세먼지와 2일 전 중국 미세먼지 변수는 각각 0.1705, 0.1368로 매우 약한 양의 상관관계

Table 3.1. Relationship between PM_{10} and each predictor

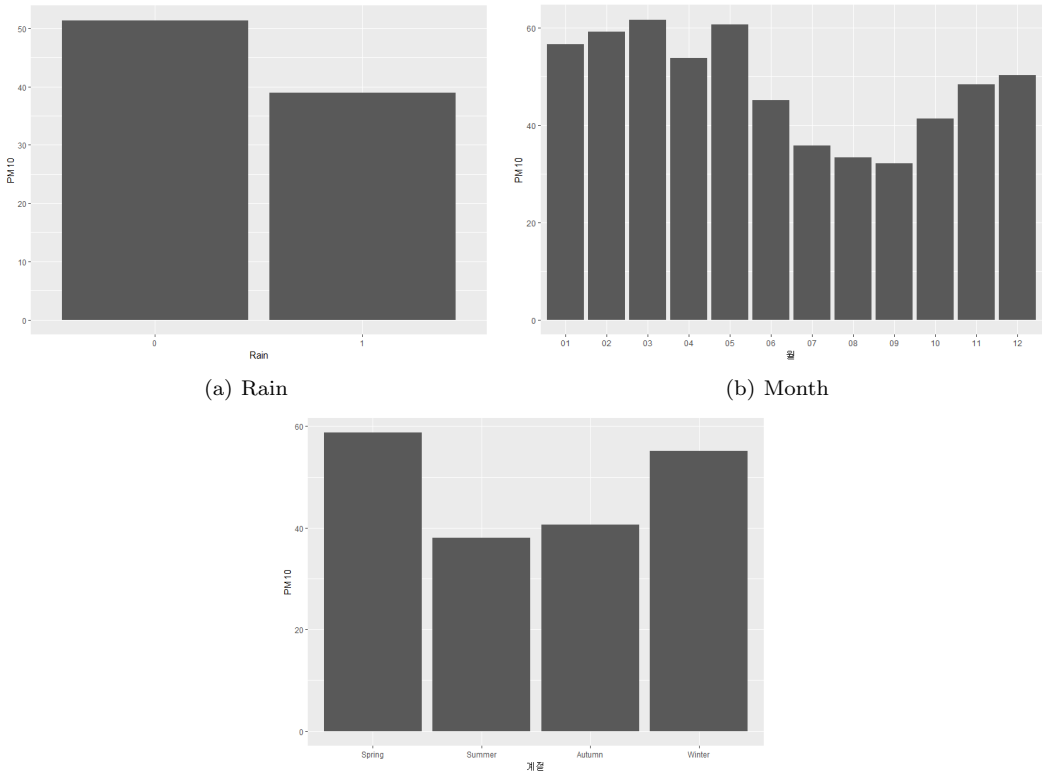
Variable	Correlation coefficient	Mean of variable by grade			
		Good	Moderate	Bad	Very Bad
PM_{10y}	0.6179	29.0504	46.9625	81.9071	135.5268
SO_2	0.2485	0.0043	0.0050	0.0070	0.0068
O_3	0.0314	0.0224	0.0230	0.0247	0.0240
NO_2	0.3080	0.0184	0.0256	0.0340	0.0345
CO	0.2934	0.4737	0.5484	0.7353	0.7043
ChinaPM1	0.1705	84.3274	97.6374	129.6883	133.7645
ChinaPM2	0.1368	90.7450	94.8335	125.2360	145.0907
meanTemp	−0.2257	16.6527	11.6965	8.4972	9.5840
minTemp	−0.2645	13.1387	6.9745	3.0120	4.3296
maxTemp	−0.1615	20.6479	17.0726	14.6927	15.2144
meanWind	−0.1433	2.6248	2.0930	1.8551	2.1021
maxWind	−0.1191	5.3860	5.1422	4.8160	5.3860

를 보였다. 기상 변수들은 모두 음의 상관관계를 보이는데 온도와 관련된 평균기온, 최저기온, 최고기온이 약한 음의 상관관계를 가지고 바람과 관련된 평균풍속, 최대풍속은 낮은 음의 상관관계를 보였다.

PM_{10} 변수의 각 범주별로 예측변수들의 평균값을 비교해보면 모든 대기 오염물질 변수들과 중국 미세먼지 변수는 범주가 나쁨(Bad) 쪽으로 나빠짐에 따라 평균값이 증가하는 경향을 보이며 반대로 기상 변수들은 평균값이 감소한다. 그러나 매우 나쁨(Very Bad)에서는 이러한 규칙이 성립하지 않는다. 전반적으로 연속형 변수인 PM_{10} 변수와 예측변수들은 높은 선형 상관성을 보이지는 않지만 범주형 반응변수로 변환하였을 때는 나쁨(Bad)까지의 순서형 반응 범주와 각 예측변수가 선형성을 나타냄을 알 수 있다. 이상과 같은 상관분석에 의하면 일반적으로 대기오염물질 농도가 작을수록, 온도는 높을수록, 바람이 불수록 미세먼지 농도는 낮아진다고 볼 수 있다.

Figure 3.1은 강수 여부, 월별, 계절별로 PM_{10}의 평균값을 나타낸 막대 그래프이다. 강수 여부에 따라 구분한 그래프 (a)에서 비가 왔을 경우(1)가 비가 오지 않았을 경우(0)에 비해 미세먼지의 농도가 낮은 것을 알 수 있다. 월별 그래프(b)의 경우 미세먼지의 농도가 3월에 가장 높게, 그 다음으로 5월, 2월, 1월 순으로 나타났고 9월, 8월, 7월 순으로 매우 낮은 값을 보이는데 이러한 결과를 통해 미세먼지의 농도는 시간에 따라 확연한 차이를 나타낸다고 할 수 있다. 월별 그래프에 근거해서 봄(3–5월), 여름(6–8월), 가을(9–11월), 겨울(12–2월)의 계절변수(Season)를 만들어 표현한 계절별 그래프 (c)에서는 봄 및 겨울철의 미세먼지의 농도가 여름 및 가을에 비해 높다. Korean Ministry of Environment (2016)에 의하면 일반적으로 봄에는 황사를 동반한 고농도 미세먼지가 발생할 가능성이 크다. 반면 비가 많은 여름철에는 미세먼지와 같은 대기오염물질이 빗방울에 씻겨 제거됨으로써 미세먼지의 농도가 낮으며 천고마비의 계절인 가을에는 기압계의 흐름이 빠르고 지역적인 대기의 순환이 원활하여 미세먼지의 농도가 낮아진다. 난방 등 연료사용이 증가하는 겨울이 되면 다시 미세먼지 농도가 높아진다고 한다.

Table 3.2는 범주형 예측변수인 강수(Rain) 및 계절(Season)에 대하여 범주형 반응변수인 PM_{10}과의 독립성 검정을 수행하기 위한 교차표로서 각 셀 안의 수치는 해당 범주 조합의 빈도를 나타내며 괄호 안의 수치는 행 퍼센트이다. 강수(Rain) 변수를 보면 비가 오지 않았을 때에 비해 비가 왔을 때 좋음(Good)의 비율은 높아지고 나쁨(Bad)의 비율은 낮아짐을 알 수 있다. 또한 계절(Season) 변수에서는 여름과 가을은 봄과 겨울에 비해 좋음(Good)의 비율은 높아지고 나쁨(Bad)의 비율은 낮아짐을 알 수 있다. 강수 및 계절 변수에 대하여 카이제곱 독립성 검정을 수행한 결과, 5% 유의수준 하에서 강수 유무 및 계절에 따라 PM_{10}의 농도 차이가 존재한다.



(a) Rain (b) Month

(c) Season

Figure 3.1. Bar graphs of categorical predictors.

Table 3.2. Cross table between PM_{10} grade and categorical predictors

Variable	Levels of variable	PM_{10} grade			
		Good	Moderate	Bad	Very Bad
Rain	0 (no rain)	2204(23.3%)	6444(68.1%)	759(8.0%)	57(0.6%)
	1 (rain)	1607(49.5%)	1503(46.3%)	108(3.3%)	26(0.8%)
Season	Spring	467(14.5%)	2328(72.4%)	370(11.5%)	51(1.6%)
	Summer	1382(44.2%)	1684(53.9%)	60(1.9%)	0(0.0%)
	Autumn	1231(38.4%)	1859(58.0%)	104(3.2%)	10(0.3%)
	Winter	731(23.1%)	2076(65.7%)	333(10.5%)	22(0.7%)

이상과 같이 Table 2.1의 각 예측변수들은 범주형 반응변수 PM_{10}을 예측하는데 유의한 변수들이라 판단하여 이후 자료분석에서 예측변수로 사용하였다.

4. 기존 분류 기법에 의한 미세먼지 예측

본 연구에서는 미세먼지 농도를 예측하기 위해 기존의 주요 분류 기법에 해당하는 신경망모형, SVM, 다항 로지스틱 회귀모형, Random Forest 기법을 사용하였다. 모든 분석 기법은 R Project를 사용하였고 R 프로그램 내에서 사용된 패키지는 nnet, e1071, randomForest이다.

4.1. 신경망모형

신경망모형은 비선형함수에 의한 예측변수들의 결합을 각 은닉마디에 전달하고 각 은닉마디들의 결합을 출력마디에 전달함으로서 반응변수 값을 예측하거나 반응변수의 범주를 분류하는 모형이다. 신경망모형의 구조는 예측변수들로 구성되는 입력층, 은닉마디들로 구성되는 은닉층, 그리고 반응변수의 범주들로 구성되는 출력층으로 이루어진다.

본 연구에서 사용한 nnet 패키지의 신경망모형은 활성함수로 선형(linear) 함수 혹은 시그모이드(sigmoid) 함수를 선택할 수 있는데 시뮬레이션 결과 시그모이드 함수를 사용한 것이 예측 정확도가 더 높았기 때문에 시그모이드 함수를 사용하기로 한다.

은닉층이 1개인 경우 신경망모형은 다음과 같이 구성된다. $X_1, X_2, \ldots, X_p$를 예측변수라 하면 $j\,(j = 1, 2, \ldots, J)$번째 은닉마디 H_j는 식 (4.1)과 같은 시그모이드 함수 $\sigma(\cdot)$에 의해서 계산된다.

$$H_j = \sigma(\zeta_j) = \frac{1}{1 + \exp(-\zeta_j)}, \tag{4.1}$$

여기서

$$\zeta_j = u_{0j} + u_{1j}X_1 + u_{2j}X_2 + \cdots + u_{pj}X_p.$$

최종적으로 관측치 Y가 범주 $k\,(k = 1(\text{Good}), 2(\text{Moderate}), 3(\text{Bad}), 4(\text{Very Bad}))$일 확률 $P(Y = k)$를 식 (4.2)와 같이 계산하여 가장 높은 확률 값을 주는 범주로서 미세먼지 농도에 대한 등급 k를 결정한다.

$$P(Y = k) = \frac{\exp(\eta_k)}{\sum_{k=1}^{4}\exp(\eta_k)}, \tag{4.2}$$

여기서

$$\eta_k = v_{0k} + v_{1k}H_1 + v_{2k}H_2 + \cdots + v_{Jk}H_J.$$

4.2. 다항 로지스틱 회귀모형

다항 로지스틱 회귀모형은 범주형 반응변수가 갖는 범주가 세개 이상일 때 반응변수의 분류에 사용하는 모형으로 관측치가 범주 $k\,(k = 1, 2, 3, 4)$일 확률 $P(Y = k)$을 식 (4.3)과 같이 계산하여 가장 높은 확률 값을 주는 범주로 미세먼지 농도에 대한 등급을 결정한다.

$$P(Y = k) = \frac{\exp(\delta_k)}{\sum_{k=1}^{4}\exp(\delta_k)}, \tag{4.3}$$

여기서

$$\delta_k = \begin{cases} \beta_{0k} + \beta_{1k}X_1 + \beta_{2k}X_2 + \cdots + \beta_{pk}X_p, & k = 1, 2, 3, \\ 0, & k = 4 \end{cases}$$

4.3. Support vector machine

SVM은 두 범주 사이의 거리를 최대로 해주는 초평면(hyperplane)을 분류 함수로 사용하여 반응변수 값을 분류하는 기계학습법으로 변수들 간의 관계를 파악하기는 힘들지만 많은 응용문제에서 우수한 성능을 보여주는 분류기법이다. SVM은 기본적으로 이항 분류 문제를 푸는 알고리즘이며 다항 범주의 분류는 이항 분류 규칙을 따른다. 즉, 모든 다항 범주에 대해서 1 대 1 이항 분류를 대응시킨 후 투표에 의해 각 관측치가 속할 최종 범주를 찾는다.

Table 4.1. Accuracy in the 4 way classification

Data	Model	Area					
		Seoul	Busan	Incheon	Daejeon	Gwangju	Daegu
Training	NN	72.30%	77.88%	73.24%	74.04%	74.22%	78.48%
	Logistic	73.26%	73.73%	72.96%	75.29%	73.94%	77.80%
	SVM	74.81%	79.09%	76.47%	76.92%	76.39%	78.94%
	RF	71.62%	73.86%	72.20%	72.89%	73.56%	77.12%
Validation	NN	73.48%	74.45%	74.96%	73.99%	74.22%	77.42%
	Logistic	71.88%	71.83%	74.54%	74.14%	73.94%	77.80%
	SVM	71.74%	73.04%	74.54%	72.70%	74.79%	77.59%
	RF	71.88%	74.04%	75.39%	73.28%	75.50%	76.06%
Test	NN	75.07%	71.15%	75.62%	68.98%	70.14%	71.39%
	Logistic	73.67%	73.77%	74.52%	68.07%	66.58%	72.89%
	SVM	75.07%	73.77%	71.19%	71.39%	69.04%	71.99%
	RF	72.55%	72.13%	72.85%	71.99%	71.78%	70.48%

NN = neural network model; SVM = support vector machine; RF = random forest.

4.4. Random forest

Random forest는 의사결정나무 기법을 기반으로 하여 의사결정나무의 불안정성과 낮은 예측력을 보완한 기계학습법으로 다수의 의사결정나무를 결합하여 반응변수를 예측한다. 주어진 하나의 데이터에서 bootstrap을 사용하여 여러 데이터를 생성하고 모델링을 하여 결합한 후에 최종 모형을 만들어 내는 앙상블 기법인 bagging과 변수에도 임의성을 반영한 노드 최적화 방법을 적용하여 기존 bagging 모형보다도 많은 초평면을 갖고 안정적인 모형을 얻게 된다.

4.5. 기존 분류기법들에 의한 미세먼지 예측

수치분석에 사용된 자료는 2010년부터 2015년까지의 결측값을 제외한 11,750개 자료이다. 이들 자료 중에서 2010년 1월 1일부터 2014년 12월 31일까지 5년간의 일별 데이터를 훈련용(training) 및 평가용 데이터(validation data)에 각각 60% 및 40%의 비율로 랜덤하게 할당하여 모형화를 위해서 사용하였다. 그리고 2015년 1월 1일 부터 12월 31일까지 1년간의 일별 데이터를 검증용 데이터(test data)로 사용하여 각 모형의 예측력을 평가하는데 사용하였다.

Table 4.1은 신경망모형(NN), SVM, 다항 로지스틱 회귀모형(Logistic), Random Forest(RF) 분류기법을 적용하였을 때의 각 지역별 미세먼지 예측의 정확도를 나타낸 표이다. 3개 데이타별 정확도는 몇 케이스를 제외하고 대부분 70% 이상이다. 신경망모형의 경우 nnet 패키지의 nnet 함수를 사용하였고 함수 내의 옵션으로는 은닉층의 수를 3개, 은닉마디의 수를 200으로 고정하여 적합하였다. 다항 로지스틱 회귀모형은 nnet 패키지의 multinom 함수를 사용하였다. SVM은 e1071 패키지의 svm 함수를 사용하였고 커널함수로 linear, polynomial, radial을 선택할 수 있는데 평가용 데이터에 대하여 가장 좋은 결과를 보이는 커널함수를 적용시켰다. 모수들에 대한 값은 모두 디폴트 값으로 사용하였다. Random Forest는 randomForest 패키지의 randomForest 함수를 사용하여 적합시켰다.

Table 4.1이 좋음(Good), 보통(Moderate), 나쁨(Bad), 매우 나쁨(Very Bad)의 4개 등급에 대한 정확도를 나타낸 반면 Table 4.2는 분류된 4개 등급 중 나쁨(Bad) 혹은 매우 나쁨(Very Bad) 등급을 합하여 정확도를 나타낸 표이다. 대부분의 지역에서는 30% 내외의 낮은 정확도를 보이며 최소 0%에서 최대 35%의 정확도에 이를 정도로 편차가 매우 크다. 특히 2015년 자료인 검증용 데이터에서 정확도

Table 4.2. Accuracy of Bad or Very Bad in the 4 way classification

Data	Model	Area					
		Seoul	Busan	Incheon	Daejeon	Gwangju	Daegu
Training	NN	16.16%	0.00%	25.00%	15.52%	17.98%	26.67%
	Logistic	28.28%	16.67%	25.71%	29.31%	24.72%	33.33%
	SVM	26.26%	12.50%	35.00%	18.97%	19.10%	28.33%
	RF	19.19%	8.33%	27.86%	18.97%	23.60%	21.67%
Validation	NN	18.31%	0.00%	26.58%	18.60%	23.19%	21.05%
	Logistic	22.54%	33.33%	21.52%	20.93%	26.09%	21.05%
	SVM	16.90%	0.00%	12.66%	4.65%	17.39%	2.63%
	RF	16.90%	11.11%	24.05%	4.65%	31.88%	7.89%
Test	NN	16.00%	0.00%	17.24%	15.79%	10.34%	5.90%
	Logistic	20.00%	0.00%	17.24%	10.53%	3.45%	0.00%
	SVM	12.00%	0.00%	6.90%	0.00%	3.45%	0.00%
	RF	20.00%	0.00%	10.34%	5.26%	3.45%	0.00%

NN = neural network model; SVM = support vector machine; RF = random forest.

가 상대적으로 낮아지는데 그 이유는 2015년 자료가 그 이전의 자료들과 약간 다른 구조를 가지고 있었다. 예를 들면 2015년 이전 자료들인 훈련용 및 평가용 데이터에서 PM_{10}과 PM_{10y}의 상관계수는 각각 0.6541 및 0.6369인 반면에 2015년 자료에서는 0.5047로 낮아지며, 중국미세먼지 수치도 2015년의 경우 그 이전의 자료에 비해서 작을 뿐 만 아니라 매우 나쁨(Very Bad) 등급에서 역으로 가장 작은 수치를 보인다. 또한 부산의 경우 0%의 정확도를 많이 보이는 이유는 나쁨(Bad) 혹은 매우 나쁨(Very Bad)으로 분류되는 자료에 하나 이상의 결측값이 존재하여 분석 대상에서 제외된 비율이 다른 지역에 비해 높기 때문이다. 즉 나쁨(Bad) 혹은 매우 나쁨(Very Bad)에 해당하는 자료 수가 타 지역에 비해 매우 작은데 분류까지 정확히 못한 결과이며 이러한 결과는 이후 분석에서도 계속되는 현상이다.

Table 2.4의 미세먼지 농도 등급별 비율을 보면 나쁨(Bad)이 최저 2.4%에서 최고 10.5%, 매우 나쁨(Very Bad)이 최저 0.2%에서 최고 1.2%로 매우 낮은 비율을 가진다. 따라서 Table 4.1에서 약 70%이상의 정확도는 대부분 좋음(Good) 혹은 보통(Moderate) 등급의 정확한 분류 돗수에 의해 예측된 것임을 알 수 있다.

그러나 실제 미세먼지 예보 시에는 좋음(Good) 혹은 보통(Moderate)보다 나쁨(Bad) 혹은 매우 나쁨(Very Bad)을 정확히 예측하는 것이 더 중요하다. 즉, 좋음(Good)을 나쁨(Bad)으로 잘못 예측할 때보다 나쁨(Bad)을 좋음(Good)으로 잘못 예측할 때의 오류가 더 위험하다. 따라서 좋음(Good) 혹은 보통(Moderate)을 좋음 혹은 보통(Good|Moderate) 등급으로 통합하고 나쁨(Bad) 혹은 매우 나쁨(Very Bad)을 나쁨 혹은 매우 나쁨(Bad|Very Bad) 등급으로 통합하여 이항 분류를 하였을 때 민감도 및 특이도 측면에서의 비율을 확인해보고자 한다.

Table 4.3은 나쁨 혹은 매우 나쁨(Bad|Very Bad) 및 좋음 혹은 보통(Good|Moderate)의 이항 분류시 정오분류표이다. 여기서 실제 데이터가 나쁨 혹은 매우 나쁨(Bad|Very Bad) 등급인 경우에, true positive (TP)는 나쁨 혹은 매우 나쁨(Bad|Very Bad)등급으로, false negative (FN)은 좋음 혹은 보통(Good|Moderate) 등급으로 예측한 일수이다. 반면에 실제 데이터가 좋음 혹은 보통(Good|Moderate) 인 경우에, true negative (TN)은 좋음 혹은 보통(Good|Moderate) 등급으로, false positive (FP)는 나쁨 혹은 매우 나쁨(Bad|Very Bad)등급으로 예측한 일수이다.

이제 좋음 혹은 보통(Good|Moderate) 및 나쁨 혹은 매우 나쁨(Bad|Very Bad)의 이항 분류시 민

Table 4.3. Confusion matrix

Data	Prediction	
	Bad|Very Bad	Good|Moderate
Bad|Very Bad	TP	FN
Good|Moderate	FP	TN

TP = true positive; FN = false negative; FP = false positive; TN = true negative.

감도(Sensitivity 혹은 True Positive Rate), 특이도(Specificity 혹은 True Negative Rate) 및 정확도(Accuracy)는 식 (4.4)와 같이 정의된다.

$$\begin{aligned} \text{Sensitivity} &= \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{Specificity} &= \frac{\text{TN}}{\text{FP}+\text{TN}}, \\ \text{Accuracy} &= \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}}. \end{aligned} \tag{4.4}$$

신경망모형과 로지스틱 회귀모형에 의한 이항분류에서는 관심 범주인 나쁨 혹은 매우 나쁨(Bad|Very Bad)의 확률을 추정해 준다. 따라서 사용자는 데이터 특성에 따라 적절한 확률을 threshold로 사용할 수 있다. 미세먼지의 예측 문제에서는 정확도를 적당한 선에서 유지하면서 나쁨 혹은 매우 나쁨(Bad|Very Bad) 등급에 대한 예측력이 높아야 한다. 만약 민감도를 최대로 하는 threshold를 설정하게 되면 나쁨 혹은 매우 나쁨(Bad|Very Bad) 등급에 대한 예측의 정확도는 100%가 되지만 전체 정확도는 약 10% 정도로 떨어지게 되어 바람직한 결과를 얻을 수 없게 된다. 따라서 본 실험에서는 threshold 값으로 첫째, 정확도를 최대로 하는 threshold(Max Accuracy)를 택하거나 둘째, 불균형의 데이터에서 Matthew's correlation coefficient (MCC)를 사용하여 MCC의 절댓값을 최대로 하는 threshold(Max Absolute MCC)를 택하여 균형잡힌 예측을 제공해주는 두 가지 방법으로 나누어 실험하였다.

MCC의 식은 다음과 같이 정의된다.

$$\text{MCC} = \frac{\text{TP}/N - S \times P}{\sqrt{PS(1-S)(1-P)}}, \tag{4.5}$$

여기서

$$N = \text{TN} + \text{TP} + \text{FN} + \text{FP}, \quad S = \frac{\text{TP}+\text{FN}}{N}, \quad P = \frac{\text{TP}+\text{FP}}{N}.$$

Table 4.4는 미세먼지의 좋음 혹은 보통(Good|Moderate)과 나쁨 혹은 매우 나쁨(Bad|Very Bad)의 2개 등급으로 구분하였을 때 각 지역별 검증용 데이터에 대한 정확도, 민감도, 특이도를 계산한 표이다. Table 2.4에서 보듯이 실제로 좋음 혹은 보통(Good|Moderate)의 일수가 나쁨 혹은 매우 나쁨(Bad|Very Bad)의 일수에 비해 압도적으로 많기 때문에 정확도 및 특이도는 높으나 상대적으로 민감도 즉, 미세먼지가 나쁨 혹은 매우 나쁨(Bad|Very Bad) 등급에 대한 정확도가 대체로 낮은 편이다. 그러나 MCC의 절댓값을 최대로 하는 threshold를 사용한 신경망모형은 가장 민감도를 개선시킨 결과를 보여주고 있다.

이제 제5장에서는 심층 신경망모형의 다양한 실험을 통하여 앞서 소개했던 기존의 분류 모형들보다 정확도를 높여줄 수 있는 지를 모색해보기로 한다.

Table 4.4. Accuracy of Bad|Very Bad for test data in the binary classification

Threshold	Model	Rate	Area					
			Seoul	Busan	Incheon	Daejeon	Gwangju	Daegu
Max accuracy	NN	Accuracy	93.28%	96.39%	90.58%	94.28%	92.33%	91.87%
		Sensitivity	44.00%	0.00%	37.93%	15.79%	33.33%	11.76%
		Specificity	96.99%	99.32%	95.18%	99.04%	92.70%	96.19%
	Logistic	Accuracy	94.12%	96.72%	92.24%	93.98%	91.51%	94.88%
		Sensitivity	20.00%	0.00%	3.45%	5.26%	6.90%	5.88%
		Specificity	99.70%	99.66%	100.00%	99.36%	98.81%	99.68%
Max absolute MCC	NN	Accuracy	93.28%	96.39%	89.75%	92.47%	92.33%	89.46%
		Sensitivity	44.00%	0.00%	41.38%	26.32%	47.62%	29.41%
		Specificity	96.99%	99.32%	93.98%	96.49%	94.48%	92.70%
	Logistic	Accuracy	94.40%	96.07%	92.24%	93.98%	92.33%	94.88%
		Sensitivity	32.00%	0.00%	13.79%	5.26%	6.90%	5.88%
		Specificity	99.10%	98.99%	100.00%	99.36%	99.70%	99.68%
None	SVM	Accuracy	94.12%	97.05%	91.69%	92.17%	93.98%	94.28%
		Sensitivity	24.00%	0.00%	13.79%	0.00%	6.90%	5.88%
		Specificity	99.40%	100.00%	98.50%	99.68%	98.81%	99.05%
	RF	Accuracy	93.84%	97.05%	92.52%	94.28%	90.68%	93.98%
		Sensitivity	24.00%	0.00%	10.35%	5.26%	6.70%	0.00%
		Specificity	99.10%	100.00%	99.70%	99.68%	97.92%	99.05%

NN = neural network model; SVM = support vector machine; RF = random forest.

5. 심층 신경망모형에 의한 미세먼지 예측

최근에 예측분야에서 각광 받고 있는 심층 신경망 모형은 기존 신경망모형의 느린 속도와 과적합 문제를 해결하면서도 높은 예측력을 보이는 기계학습법들 중 하나로 알려져 있다. 이제 제4장에서 사용했던 기존의 주요한 분류 기법들에 비해 심층 신경망모형을 사용하였을 때 보다 개선된 예측 결과를 보이는지 살펴보기로 한다. 본 논문에서는 심층 신경망모형에 의한 미세먼지 예측을 위해서 R Project 내의 h2o 패키지를 사용하여 분석하였다 (Arno, 2015; The H2O.ai team, 2017).

5.1. 심층 신경망모형의 개요

심층 신경망모형은 4.1절에서 소개한 신경망모형에 기초하여 입력층과 출력층 사이에 2개 이상의 은닉층을 가지고 있고 다수의 은닉노드를 포함하는 신경망모형이라고 할 수 있다. 심층 신경망모형에서 다층 퍼셉트론은 Backpropagation으로 학습되는데 가중치들은 Stochastic gradient descent에 의해 갱신된다.

h2o 패키지를 사용하여 심층 신경망모형의 분석에서 사용할 수 있는 활성함수로는 Hyperbolic tangent, Rectified Linear Unit (RELU), Maxout이 있다. 심층 신경망모형의 경우 다수의 은닉층과 은닉노드를 필요로 하기 때문에 기존의 Sigmoid와 같은 활성함수보다 학습이 빠른 함수를 사용한다. 이제 예측변수를 $X_1, X_2, \ldots, X_p$라 하고 각 예측변수들에 대한 가중치를 $w_1, w_2, \ldots, w_p$라 하면 예측변수들의 가중결합 α를 다음과 같이 표현할 수 있다.

$$\alpha = \sum_{i}^{p} w_j x_j + b. \tag{5.1}$$

Hyperbolic tangent 함수는 4.1절의 Sigmoid 함수보다 빠른 학습을 위해 Sigmoid 함수를 변형하여 값의 범위를 −1부터 1로 늘린 형태이다. 가중결합 α에 대한 활성함수 Hyperbolic tangent의 식은 다음과 같다.

$$f(\alpha) = \frac{\exp(\alpha) - \exp(-\alpha)}{\exp(\alpha) + \exp(-\alpha)}. \tag{5.2}$$

RELU 함수는 최근 가장 많이 쓰이고 있는 활성함수로 Hyperbolic tangent 함수처럼 기존의 Sigmoid 함수를 개선시킨 함수이다. Sigmoid 함수는 Gradient descent를 사용한 Backpropagation 과정에서 여러 층들을 거쳐가면서 Gradient가 0으로 수렴하게 되어 다층의 신경망 구조에서는 좋은 결과를 내지 못한다. 반면에 RELU 함수는 0보다 작은 값에 대해 0을 출력하여 부분 활성화가 가능하고 선형함수이기 때문에 계산도 빠르게 수행되는 장점이 있다. 가중결합 α에 대한 활성함수 RELU의 식은 다음과 같다.

$$f(\alpha) = \max(0, \alpha). \tag{5.3}$$

Maxout 함수는 RELU 함수와 Leaky RELU 함수를 일반화한 형태이다. Leaky RELU 함수란 기존의 RELU 함수에서 입력값이 0보다 작을 경우 미분값이 항상 0으로 되어 해당 신경망이 활성화되지 않게 되는 문제를 보완한 것이다. Maxout은 위의 두 함수에 의한 가중결합 α_1, α_2를 계산하여 더 큰 값을 가지는 형태이다. RELU의 장점을 보유하면서 RELU의 단점을 보완하지만 그만큼 추정해야할 모수가 두 배가 되어 계산 시간이 길어지는 단점이 있다. 가중결합 α_1, α_2에 대한 활성함수 Maxout의 식은 다음과 같다.

$$f(\alpha) = \max(\alpha_1, \alpha_2). \tag{5.4}$$

5.2. 심층 신경망모형의 최적화

심층 신경망모형을 적합시킬 때 은닉층과 은닉노드의 수가 많아지면서 과적합의 문제가 발생할 수 있다. 즉, 훈련용 데이터에서는 예측을 잘 하지만 평가용 데이터에서 예측력이 떨어지는 결과가 나올 수 있다. 따라서 편의-분산 상쇄(bias-variance trade off)측면에서 분산을 줄여 새로운 데이터에 대하여 예측을 잘 수행하도록 하는 방법이 필요하다. 심층 신경망분석에서 과적합을 방지할 수 있는 방법으로 정규화와 Dropout이 있다.

정규화의 방법으로는 L1, L2 정규화가 있으며 이중에서 주로 L2 정규화를 사용하는데 이는 Gradient descent에 의해 가중치가 갱신되는 과정에서 loss 항에 가중치 값이 너무 커지지 않도록 제한한다. 가중치가 클 경우 그만큼 더 큰 페널티를 주어 과적합을 방지하는 효과를 갖는다. j번째 가중치를 w_j라고 하고 페널티를 조정하는 모수를 λ라고 하면 기존의 loss 함수에서 다음의 $l2$ 식을 더하게 된다. λ의 값으로는 주로 0.01 또는 0.001을 사용하는데 가중치가 클 경우 더 큰 λ 값을 주어 페널티를 부여한다.

$$l2 = \frac{1}{2}\lambda \sum_j {w_j}^2. \tag{5.5}$$

L1 정규화는 절댓값 기호를 사용하여 가중치에 주는 페널티가 크기에 비례하지 않고 항상 상수 형태의 값을 가지게 되어 L2 정규화에 비해 더 자유로운 값을 가지게 된다. j번째 가중치를 w_j라고 하고 페널티를 조정하는 모수를 λ라고 하면 기존의 loss 함수에서 다음의 $l1$ 식을 더하게 된다.

$$l1 = \lambda \sum_j |w_j|. \tag{5.6}$$

Table 5.1. Definition of model number

Regularization	Value	Epoch(90)			Epoch(100)		
		Tangent	RELU	Maxout	Tangent	RELU	Maxout
L1	0.01	1	11	21	31	41	51
	0.001	2	12	22	32	42	52
	0.0001	3	13	23	33	43	53
L2	0.01	4	14	24	34	44	54
	0.001	5	15	25	35	45	55
	0.0001	6	16	26	36	46	56
Dropout	0.3	7	17	27	37	47	57
	0.5	8	18	28	38	48	58
	0.4, 0.7	9	19	29	39	49	59
	0.6, 0.3	10	20	30	40	50	60

RELU = Rectified Linear Unit.

Dropout은 각 은닉층에서 일정 비율의 은닉노드를 제외하여 나머지의 은닉노드로만 모형화에 사용하는 방법이다. 따라서 훈련용 데이터에 의해 적합이 이루어지는 과정에서 과적합을 방지하고 새로운 데이터에 대한 예측력이 향상될 수 있다. 또한 심층 신경망모형 적합 과정에서 각 반복마다 Dropout이 적용되는 은닉노드가 변화하면서 랜덤한 신경망을 사용하게 되어 앙상블의 기능을 가진다.

5.3. 심층 신경망모형의 적합

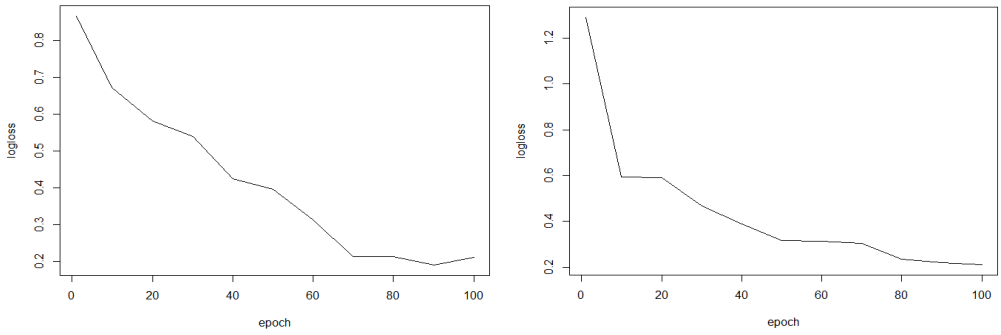
심층 신경망모형을 적합할 때 사용자가 활성함수를 결정하고 은닉층, 은닉노드의 수, Epoch, 정규화 등의 옵션을 결정해야 한다. 본 연구에서는 이러한 옵션 값을 결정할 때 다양한 시뮬레이션을 통하여 결정하였다. 모형화는 h2o 패키지의 h2o.deeplearning 함수를 사용하였다. Epoch 옵션은 모형훈련을 수행하는 과정에서 데이터에 대한 한 번의 학습 과정을 의미하는데 loss를 작게 하는 적당한 Epoch를 정해주는 것이 좋다. Epoch가 너무 크면 수행 시간도 오래 걸릴 뿐만 아니라 과적합의 문제가 생길 수 있기 때문에 적절한 크기의 Epoch를 갖는 활성함수의 사용이 필요하다.

먼저 각 지역별 데이터를 사용하여 Epoch를 조절해가면서 Hyperbolic tangent, RELU, Maxout 3가지의 활성함수들에 대해 적합시킨 모형의 loss 값을 최소로 하는 적절한 Epoch를 찾아보았다. Epoch은 $1, 10, 20, \ldots, 100$의 값을 사용하였고 loss 값은 h2o.logloss 함수에 의해 계산된 로그 형태의 loss 값인 logloss를 사용하였다.

Figure 5.1은 서울시 데이터에 대한 Epoch와 logloss의 그래프인데 Hyperbolic tangent, RELU, Maxout 활성함수들에 대해 각각 Epoch 값이 90 또는 100일 때 최소의 loss를 갖게 된다. 다른 지역들의 데이터도 마찬가지로 Epoch 값이 90 또는 100일 때 최소의 loss를 보였다.

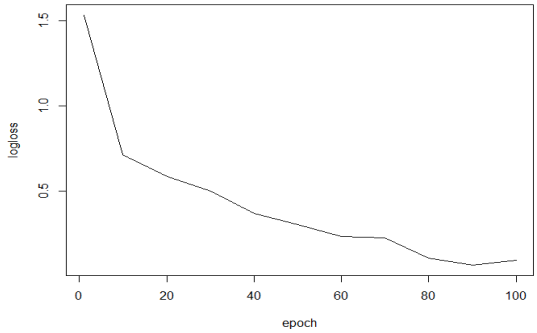
은닉층과 은닉마디의 수에 대해서는 너무 복잡해지지 않도록 적당한 개수인 은닉층 2개 혹은 3개, 은닉마디 각각 100개 혹은 200개로 시뮬레이션 해본 결과 은닉층이 3개이고 은닉마디가 각각 200개인 경우가 가장 적합도가 높게 나타났다. 따라서 은닉층과 은닉마디의 수는 3개, 200개로 모든 활성함수에 대하여 고정하도록 한다.

Table 5.1은 모형 번호를 정의한 표로 은닉층 3개, 은닉마디의 수를 각각 200개로 고정한 상태에서 Epoch 값과 Hyperbolic tangent, RELU, Maxout 3가지의 활성함수와 L1 정규화, L2 정규화, Dropout 기법들에 대한 λ와 Dropout 비율 값을 다르게 하여 각각에 대한 모형 번호를 정의하였다. 정규화 옵션을 사용했을 때 λ 값을 0.01, 0.001, 0.0001로 각각 사용하였고 Dropout 기법을 사용했을 때에는 전체



(a) Hyperbolic tangent

(b) RELU



(c) Maxout

Figure 5.1. Epoch and logloss graph by activation functions. RELU = Rectified Linear Unit.

Table 5.2. Accuracy in the 4 way classification using deep neural network models

Value	Area					
	Seoul	Busan	Incheon	Daejeon	Gwangju	Daegu
Model Number	4	34	2	48	27	51
Epoch	90	100	90	100	90	100
Activation function	Tangent	Tangent	Tangent	RELU	Maxout	Maxout
Regularization	L2	L2	L1	Dropout	Dropout	L1
Lambda or Ratio	0.01	0.01	0.001	0.5	0.3	0.01
Training	71.43%	74.66%	75.62%	73.08%	75.92%	79.28%
Validation	73.95%	75.08%	76.73%	73.49%	74.52%	74.40%
Test	75.35%	74.43%	75.68%	74.43%	73.80%	77.93%

노드의 Dropout 비율을 0.3, 0.5와 같이 고정한 방법과 입력노드, 은닉노드의 Dropout 비율을 다르게 하는 방법을 사용하였다. 표에서 제시한 1번부터 30번까지의 모형은 Epoch 값을 90으로 고정한 모형이고 31번부터 60번까지의 모형은 Epoch 값을 100으로 고정한 모형으로 정의한다.

Figure 5.2는 Table 5.1에서 정의한 1번부터 60번까지의 모형에 대하여 각 지역별로 그래프 상단에는 평가용 데이터의 정확도를 실선으로 나타내고 하단에는 훈련용 데이터와 평가용 데이터의 정확도 차이를 점선으로 표시하였다. 각 지역별 그래프에서 실선의 최대값 및 점선의 최소값을 나타내는 최적모형을 잠정적으로 각 2개씩 선택하여 점선 형태의 수직선으로 표시하였다. 최적의 모형 선택 기준은 예측

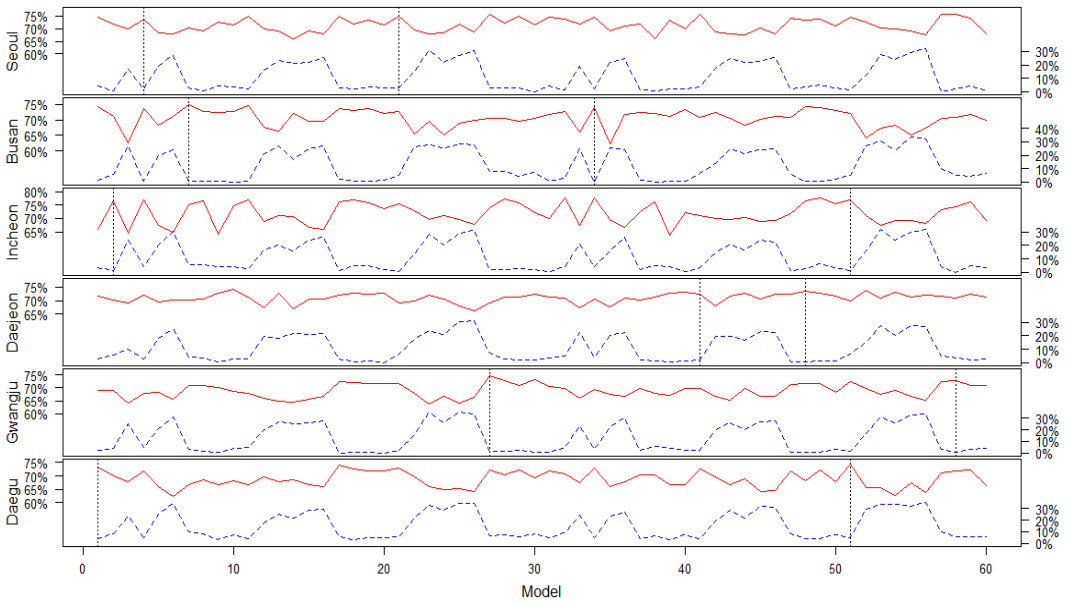


Figure 5.2. Accuracy for validation data(left axis) and difference of accuracy for training and validation data(right axis).

Table 5.3. Accuracy of Bad or Very Bad in the 4 way classification using deep neural network models

Data	Area					
	Seoul	Busan	Incheon	Daejeon	Gwangju	Daegu
Training	34.34%	41.67%	35.00%	24.14%	61.80%	28.33%
Validation	33.80%	33.33%	31.65%	23.26%	46.38%	18.42%
Test	24.00%	11.11%	24.14%	15.79%	31.03%	11.76%

의 안정성을 위해서 훈련용 데이터와 평가용 데이터의 정확도 차이가 약 5% 내외로 작은 값이면서, 예측의 정확성을 위해서 평가용 데이터의 정확도를 최대로 하는 모형으로 판단하였다.

Table 5.2는 Figure 5.2에서 선택된 각 지역별 2개의 모형들 중 보다 더 우수한 모형을 최종적으로 선택하여 나타낸 표이다. 선택된 각 모형 번호에 대하여 적용된 모형화 방법을 요약하고 각 데이터별 정확도를 나타내었다. Table 5.2가 좋음(Good), 보통(Moderate), 나쁨(Bad), 매우 나쁨(Very Bad)의 4개 등급에 대한 정확도를 나타낸 반면 Table 5.3은 분류된 4개 등급 중 나쁨(Bad) 혹은 매우 나쁨(Very Bad) 등급을 합하여 이 2개 등급의 정확도를 나타낸 표이다. 4원 분류에서 기존 분류기법에 의한 4개 등급 분류 결과인 Table 4.1 및 2개 등급 분류 결과인 Table 4.2와 비교하였을 때 심층 신경망모형에 의한 예측 결과는 검증용 데이터에서 보다 정확도가 높아졌다.

Table 5.4는 심층 신경망모형에 의한 검증용 데이터에 대한 이항 분류의 예측 결과를 보여준다. 각 지역에 대한 심층 신경망모형의 옵션으로는 Figure 5.2에서와 같은 방법으로 선택된 모형을 사용하였다. Table 5.4에서 첫째 방법인 최대 정확도를 갖는 threshold를 선택했을 경우는 threshold(Max Accuracy)를 지정하지 않은 경우와 유사한 결과를 보이는데, 즉 정확도와 특이도는 높지만 민감도는 여전히 낮은 편이다. 따라서 낮은 민감도를 보완하기 위해서 MCC의 절댓값을 최대로 하는 threshold(Max absolute MCC)를 설정하여 예측하였을 때 정확도와 특이도는 대체로 90% 내외 값으로 유지되면서 민

Table 5.4. Accuracy of Bad|Very Bad for test data in the binary classification using deep neural network models

Threshold	Rate	Area					
		Seoul	Busan	Incheon	Daejeon	Gwangju	Daegu
Max accuracy	Accuracy	94.68%	97.05%	93.63%	93.67%	91.78%	94.28%
	Sensitivity	36.00%	11.11%	24.14%	5.26%	20.69%	5.88%
	Specificity	99.10%	99.66%	99.70%	99.04%	97.92%	99.05%
Max absolute MCC	Accuracy	92.44%	97.05%	89.47%	86.75%	86.85%	91.87%
	Sensitivity	44.00%	11.11%	44.83%	47.37%	48.28%	47.06%
	Specificity	96.08%	99.66%	93.37%	89.14%	90.18%	94.29%

MCC = Matthew's correlation coefficient.

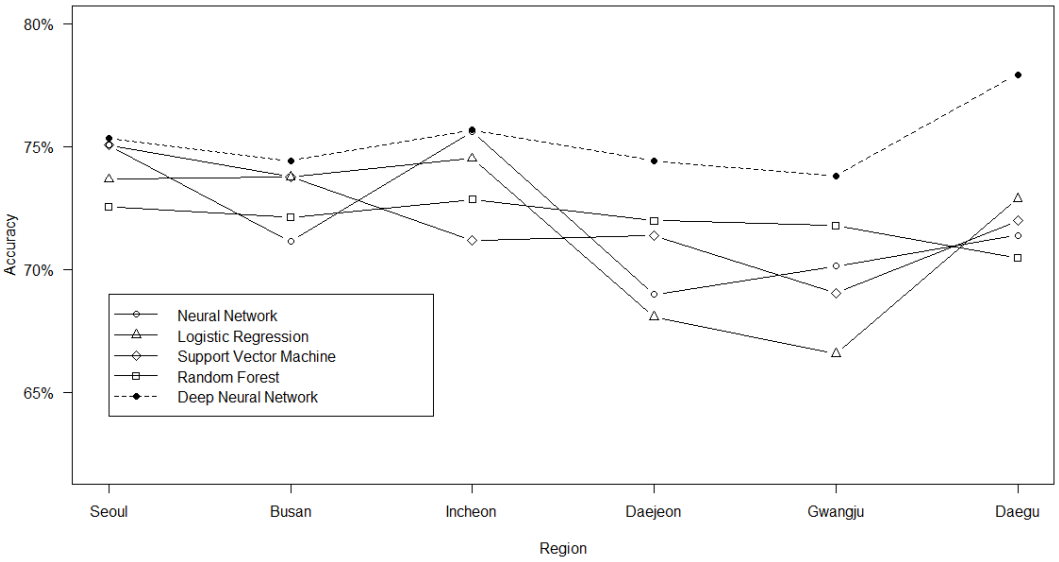


Figure 5.3. Accuracy for test data in the 4 way classification.

감도는 높아졌다.

종합하여 Figure 5.3-5.5는 본 연구에서 논의된 모든 분류기법에 의한 미세먼지 등급의 정확도를 검증용 데이터에 대해 요약하여 나타낸 그림이다. 즉, Figure 5.3은 4원 분류에서 미세먼지 4개 등급의 정확도를 나타낸 Table 4.1 및 Table 5.2로 부터 각 모형별로 비교한 그래프이다. Figure 5.4는 4원 분류에서 미세먼지 4개 등급 중 나쁨(Bad) 혹은 매우 나쁨(Very Bad) 등급을 합하여 정확도를 나타낸 Table 4.2 및 Table 5.3로 부터 각 모형별로 비교한 그래프이다. Figure 5.5는 이항 분류에서 나쁨 혹은 매우 나쁨(Bad|Very Bad) 등급의 정확도(민감도)를 나타낸 Table 4.4 및 Table 5.4로 부터 각 모형별로 비교한 그래프이다. 이때, 신경망모형, 로지스틱 회귀모형, 그리고 심층 신경망모형의 경우는 MCC의 절댓값을 최대로 하는 threshold를 사용한 경우를 나타내었다. 전체적으로 심층 신경망모형은 검증용 데이터 기준 하에서 다른 분류기법에 비해 우위인 정확도를 보이고 있음을 알 수 있다.

6. 결론

본 연구에서는 2010년부터 2015년까지 6년 동안 서울특별시 강남구, 부산광역시 해운대구, 인천광역시

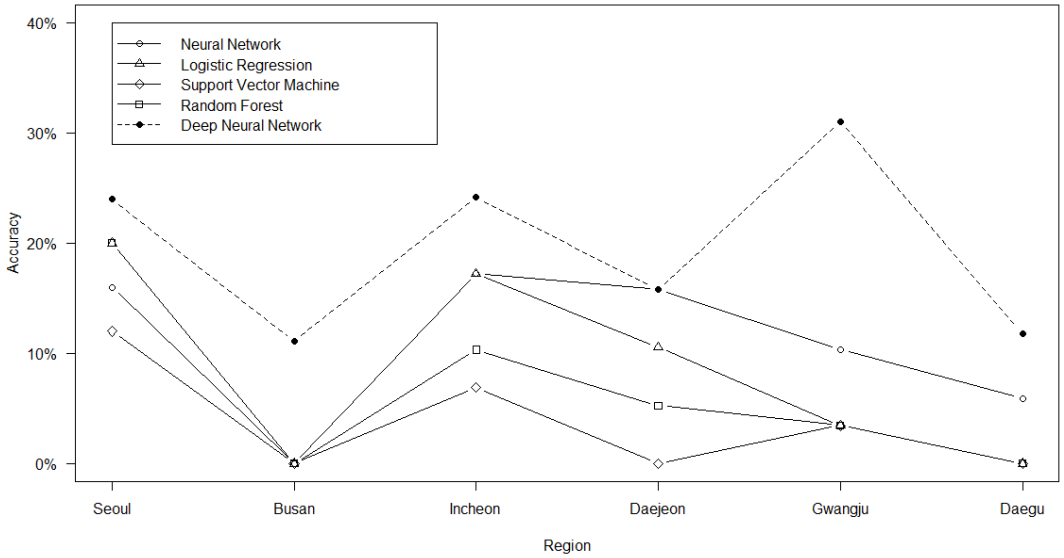


Figure 5.4. Accuracy of Bad or Very Bad for test data in the 4 way classification.

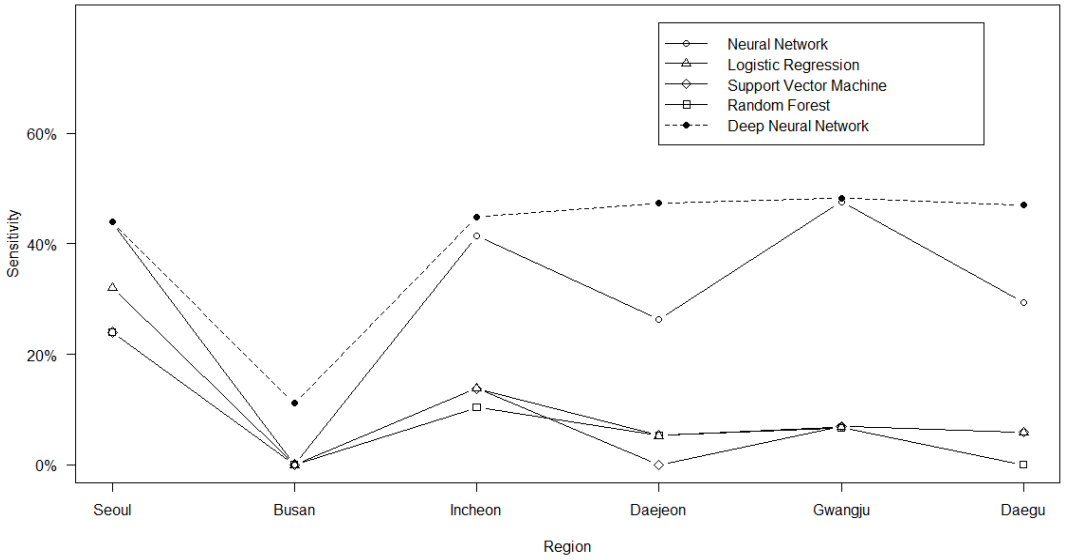


Figure 5.5. Accuracy(Sensitivity) of Bad|Very Bad for test data in the binary classification.

부평구, 대전광역시 서구, 광주광역시 북구, 대구광역시 달서구 지역을 대상으로 환경부에서 실제 예보하는 미세먼지의 4가지 등급인 '좋음, 보통, 나쁨, 매우 나쁨'의 예측 그리고 2가지 등급인 '좋음 혹은 보통, 나쁨 혹은 매우 나쁨'의 예측을 수행하였다. 미세먼지의 등급 예측을 위해서 예측변수로서 기존에 사용되던 대기오염 물질, 기상 요소 및 전날 미세먼지의 농도에 추가하여 중국 미세먼지의 농도를 사용하였다.

미세먼지의 예측을 위하여 신경망모형, Support Vector Machine, 다항 로지스틱 회귀모형, Random

Forest와 같은 기존의 주요 분류 기법 그리고 딥러닝기법에 속하는 심층 신경망모형을 사용하여 시뮬레이션을 통해 가장 성능이 좋은 활성함수와 모수를 찾고 최적화 기법을 통해 미세먼지의 4가지 등급을 예측하였을 때 비교적 높은 정확도를 보였지만 미세먼지가 나쁨(Bad) 혹은 매우 나쁨(Very Bad)인 등급에 대해서는 상대적으로 정확도가 낮았다. 그러나 신경망모형과 심층 신경망모형에 의한 이항 분류에서는 MCC의 절댓값을 최대로 하는 threshold 값을 사용하여 분류하게 되면 정확도를 적당히 높게 유지하면서 낮은 비율로 구성된 불균형 데이터에 대해서도 균형있는 예측 결과를 보여주었다. 이는 실제 미세먼지의 예보를 했을 때 나쁨 등급을 좋음 등급으로 분류할 때의 위험성을 줄이고 미세먼지 예방 차원에서 더 좋은 영향을 줄 수 있을 것이라 생각한다. 자료분석 결과를 종합하면 미세먼지의 4가지 등급 혹은 2가지 등급의 예측 정확도는 심층 신경망모형이 기존의 분류기법에 비해서 더 우수함을 보여주었다.

본 연구의 자료분석 과정에서는 각 지역별로 미세먼지, 대기오염 물질, 기상 요소의 특성이 다르다는 것을 파악하여 국내 6개 대표 대도시 지역의 자료에 대한 분석을 실시하였다. 비록 지면 관계상 생략하였지만 각 지역이 아닌 각 계절별로 나누어 자료분석을 하였을때도 미세먼지의 4가지 등급 혹은 2가지 등급의 예측 정확도 측면에서 심층 신경망모형이 기존의 분류기법에 비해서 대체로 더 우수하였다. 일반적으로 심층 신경망모형은 데이터가 더 많이 축적될수록 예측력이 높아지게 되는데, 본 연구에서 보다 더 오랜 기간의 데이터를 수집하고 특성이 유사한 지역들을 묶어서 분석하게 된다면 예측력은 더 높아질 것으로 생각된다. 또한 전날 얻을 수 있는 변수만을 활용하여 분석하였는데 이를 실시간으로 얻게 되어 분석할 수 있다면 시간별 예보 모형도 가능해질 것이고 기존의 모형보다 더 정확한 예측 성능을 보일 것이라고 기대된다.

References

Arno, C., Jessica, L., Erin, L., Viraj, P., and Anisha, A. (2015). *Deep Learning with H2O* (3rd ed), H2O.ai, Inc. California.

Koo, Y. S., Yun, H. Y., Kwon, H. Y., and Yu, S. H. (2010). A development of PM_{10} forecasting system, *Journal of Korean Society for Atmospheric Environment*, **26**, 666–682.

Korean Ministry of Environment (2016). If you know right now, it is seen. What on the earth is the fine dust? *Korean Ministry of Environment*.

Kwon, J. H., Lim, Y. J., and Oh, H. S. (2015). Particulate Matter prediction using Quantile boosting, *The Korean Journal of Applied Statistics*, **28**, 83–92.

Lee, H. J. (2011). Analysis of PM_{10} concentration using auto-regressive error model at Pyeongtaek City in Korea, *Journal of Korean Society for Atmospheric Environment*, **27**, 358–366.

Lee, J. H., Kim, Y. M., and Kim, Y. K. (2017). Spatial panel analysis for PM2.5 concentrations in Korea, *Journal of Korean Society for Atmospheric Environment*, **27**, 358–366.

Lee, W. S. and Baek, C. R. (2014). The sparse vector autoregressive model for PM_{10} in Korea, *Journal of the Korean Data and Information Science Society*, **25**, 807–817.

National Institute of Environmental Research (2016). Annual report of air quality in Korea 2015, *Korean Ministry of Environment*.

The H2O.ai team (2017). R Interface for H2O, *CRAN*.

심층 신경망모형을 사용한 미세먼지 PM_{10}의 예측

전성현[a] · 손영숙[a,1]

[a]전남대학교 통계학과

(2018년 2월 5일 접수, 2018년 3월 1일 수정, 2018년 3월 14일 채택)

요 약

본 연구에서는 미세먼지 PM_{10}의 4가지 분류 등급인 '좋음, 보통, 나쁨, 매우 나쁨' 그리고 2가지 분류 등급인 '좋음 혹은 보통, 나쁨 혹은 매우 나쁨'을 예측하기 위해서 심층 신경망모형을 사용하였다. 2010년부터 2015년까지 국내 6개 대도시 지역에서 관측한 일별 미세먼지 데이터에 대하여 기존 분류기법인 신경망모형, 다항 로지스틱 회귀모형, Support Vector Machine, Random Forest을 적용했을 때에 비해서 심층 신경망모형의 정확도는 더 높아졌다.

주요용어: 미세먼지 PM_{10}, 신경망, 다항 로지스틱 회귀, support vector machine, random forest, 심층 신경망, 정확도

[1]교신저자: (61186) 광주광역시 북구 용봉로 77(용봉동), 전남대학교 통계학과. E-mail: ysson@jnu.ac.kr