韓國IT서비스學會誌
第18卷 第1號
2019年 3月, pp.173-186

Journal of Information Technology Services
https://doi.org/10.9716/KITS.2019.18.1.173

기상환경데이터와 머신러닝을 활용한 미세먼지농도 예측 모델*

임 준 묵**

An Estimation Model of Fine Dust Concentration Using Meteorological Environment Data and Machine Learning*

Joon-Mook Lim**

■ Abstract ■

Recently, as the amount of fine dust has risen rapidly, our interest is increasing day by day. It is virtually impossible to remove fine dust. However, it is best to predict the concentration of fine dust and minimize exposure to it. In this study, we developed a mathematical model that can predict the concentration of fine dust using various information related to the weather and air quality, which is provided in real time in 'Air Korea (http://www.airkorea.or.kr/)' and 'Weather Data Open Portal (https://data.kma.go.kr/).' In the mathematical model, various domestic seasonal variables and atmospheric state variables are extracted by multiple regression analysis. The parameters that have significant influence on the fine dust concentration are extracted, and using ANN (Artificial Neural Network) and SVM (Support Vector Machine), which are machine learning techniques, we proposed a prediction model. The proposed model can verify its effectiveness by using past dust and weather big data.

Keyword : Fine Dust, Machine Learning, Weather Data, Bigdata

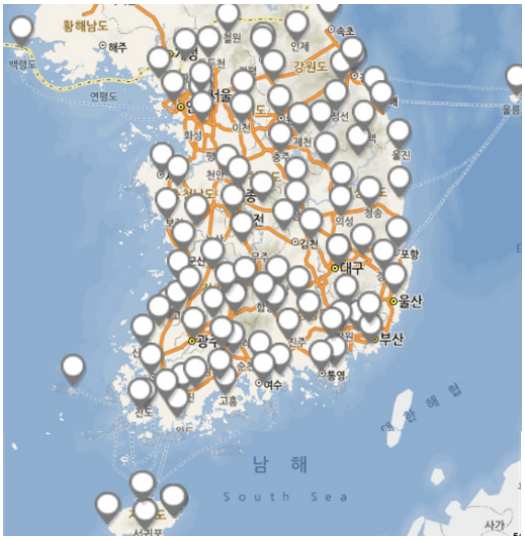
Submitted : November 8, 2018 1st Revision : December 17, 2018 Accepted : January 26, 2019

* 본 연구는 2015년 한밭대학교의 대학회계 연구비 지원을 받아 연구되었음.

** 한밭대학교 공과대학 창의융합학과 교수

1. 서 론

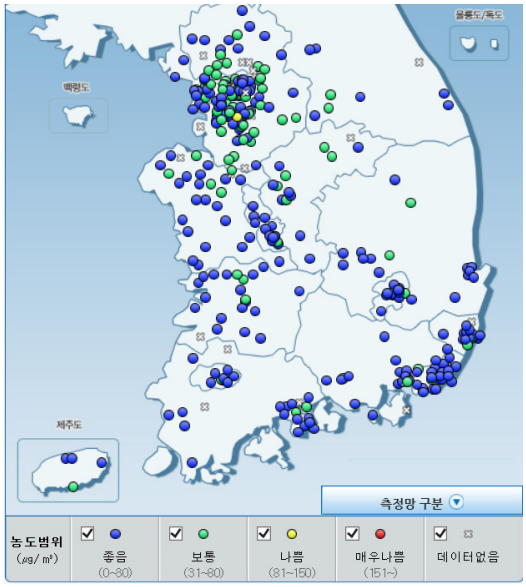
최근 미세먼지 수치가 급격히 상승함에 따라 사람들의 관심이 나날이 높아지고 있다. 미세먼지의 노출은 호흡기만이 아니라 심혈관계 질병의 발생에도 영향을 끼치며, 심하게는 사망률도 증가하는 것으로 연구되고 있다(Shin et al., 2018). 건강 문제에 더해 산업현장 및 농업현장에 대해도 상당한 피해를 입히고 있다. 미세먼지 자체를 없애는 것은 사실상 불가능한 일이다. 하지만 미세먼지 농도를 예측하여 이에 대한 노출을 최소화하는 것이 가장 좋은 방법일 것이다.



〈Figure 1〉 Weather Data Station Location Map (https://data.kma.go.kr/)

기상자료개방포털(https://data.kma.go.kr/)에서 제공하는 기상자료 데이터는 종관기상관측장비(ASOS : Automated Synoptic Observing System)에 의한 자동관측과 목측에 의한 수동관측으로 실시된다. <Figure 1>과 같이 현재 전국에 94소를 운영하고 있고 지상 부근의 대기상태를 실시간으로 관측하기 위한 기본 장비로서 기온, 습도, 풍향, 풍속, 기압, 강수량, 일조, 일사, 지면온도, 초상온도, 지중온도를 매분 관측한다. 수동관측요소는 적설, 구름, 기타 일기현상 등이며, 실시간, 매 정시 또는 3시간 간격으로 관측한다.

한편 한국환경공단의 에어코리아(http://www.airkorea.or.kr/)는 공기오염상태를 확인할 수 있는 전국 실시간 대기오염도를 공개하는 홈페이지로써, 전국의 대기오염 측정망에서 측정되는 아황산가스(SO_2), 일산화탄소(CO), 이산화질소(NO_2), 오존(O_3), 미세먼지(PM10), 초미세먼지(PM2.5) 등의 대기의 질에 관련된 데이터를 실시간으로 제공하고 있다.



〈Figure 2〉 Location of fine Dust Meter Station (http://www.airkorea.or.kr/)

<Figure 2>에서 보는 바와 같이 에어코리아는 전국 97개 시, 군에 설치된 323개의 도시대기 측정망, 도로변대기 측정망, 국가배경 측정망, 교외대기 측정망에서 측정된 대기환경기준물질의 측정자료를 다양한 형태로 표출하여 국민들에게 실시간으로 제공하고 있다.

<Figure 1>과 <Figure 2>에서 알 수 있듯이 기상자료 측정소의 위치는 전국에 골고루 분포해 있는 반면에 미세먼지 측정소 위치는 특정 지역에 집중되어 있는 것을 확인할 수 있다. 미세먼지 측정소의 위치가 골고루 분포하고 있지 않기 때문에

측정 센서가 위치한 장소의 데이터를 기준으로 주변 장소의 미세먼지 농도까지 일반화 시키는 문제점이 있다.

측정소를 모든 곳에 설치하기에는 어려운 것이 현실이다. 하지만 기상자료개방포털에서 제공하는 기상자료데이터는 전국에 골고루 위치하고 있기 때문에 기상자료를 기반으로 한 미세먼지농도를 예측하는 수리모델을 개발하게 된다면 관측재원(측정소)이 없는 곳에서도 미세먼지 농도를 예측 할 수 있기 때문에 미세먼지에 의한 피해를 예방하는데 도움이 될 것이다.

국립환경과학원에 따르면 수도권 고농도 미세먼지 발생 원인을 분석해 본 결과 중국 등의 국외 요인보다는 국내의 기상상태에 따른 2차 미세먼지의 생성 등이 미세먼지의 농도에 크게 영향을 미치고 있는 것으로 밝혀졌다(Lee, 2016).

본 연구에서는 에어코리아의 대기질 및 미세먼지 농도 측정치와 기상자료개방포털에서 실시간으로 제공하는 기상관련 다양한 정보를 활용하여 국내 계절성 변수들과 대기상태 변수들을 다중회귀분석(multiple regression analysis)을 통해 미세먼지농도에 유의한 영향을 미치는 변수를 추출하고, 그 변수들을 토대로 머신러닝(machine learning) 기법인 ANN(Artificial Neural Network)과 SVM (Support Vector Machine)을 사용하여 미세먼지 농도를 예측할 수 있는 모델을 제안한다.

제안 모델은 과거의 미세먼지 및 기상데이터를 활용하여 그 효과성을 검증할 것이다.

본 연구에서 제안된 모델을 사용하면 관측재원이 없는 곳에서도 쉽게 획득이 가능한 기상데이터를 활용하여 미세먼지 농도를 예측할 수 있고, 이를 통해서 미세먼지의 피해를 사전에 예방할 수 있는 기회를 가질 수 있을 것이다

제 2장에서는 미세먼지의 예측에 관련된 기존의 다양한 연구결과를 살펴본 후, 본 연구에서 제안한 예측모델의 특성과 차이점에 대해서 기술한다.

제 3장에서는 기상자료개방포털과 에어코리아로부터 수집한 미세먼지를 포함한 대기질 및 기상관련 빅데이터에 대한 설명을 하였다. 제 4장에서는 머신러닝과 관련된 미세먼지 예측모델로 다중회귀분석, SVM(Support Vector Machine) 및 ANN (Artificial Neural Network)을 활용한 미세먼지 예측모델을 도출하였다. 마지막으로 제시한 3가지 모델의 결과를 비교하고 결론을 도출하였다.

2. 미세먼지농도 예측 기존연구 및 예측모델 제안

2.1 미세먼지농도 예측 기존연구

그 동안 기상환경자료는 미세먼지의 농도에 가장 큰 영향을 미칠 것으로 예상하고 많은 연구들이 수행되어 왔다. 신문기등(Shin et al., 2007)은 주요 기상인자로 풍향, 풍속, 습도, 일기유형, 해륙풍의 5가지에 대해서 미세먼지와의 연관성을 분석하고 있지만 그 밖의 다양하고 방대한 기상자료를 고려하지 못한 한계점을 가진다.

또한 기상자료와 대기질 환경자료를 중심으로 회귀분석모형이나 인공신경망모형을 활용한 미세먼지 예보시스템을 개발한 연구 사례가 있다(Koo et al., 2010). 하지만 주어진 예보시스템에서는 오늘의 미세먼지측정치와 기상자료를 바탕으로 내일의 미세먼지를 예측하는 것으로서, 미세먼지의 측정치가 없는 지역에 대한 예측은 원천적으로 불가능하여, 본 연구에서 지향하는 모든 지역에서의 미세먼지농도 예측을 위한 모형에 직접적으로 활용하는 데는 한계가 있다.

최근에는 미세먼지농도 예측을 위해서 대기질정보와 기상정보를 사용하고, 선형예측기법, 머신러닝을 활용한 비선형기법, 시계열모형 등의 다양한 예측기법들이 활용되고 있다. 전송완 등(Joun et al., 2017)의 연구에 따르면, SVR(Support Vector Regression)의 비선형기법이 선형회귀모형기법이나 시계열분석모형에 비해서 상대적으로 좋은 성과를 나타냄을 보이고 있다.

또한 오병두 등(Joun et al., 2016)은 5개의 기상

인자(일기유형, 기온, 상대습도, 풍속, 풍향)만을 고려하고 머신러닝 기법을 활용하여 미세먼지를 예측하는 모형을 제시하였는데, Multilayer Perceptron (MLP) 알고리즘이 가장 좋은 성능을 나타냄을 보였다. 하지만 사용한 기상인자의 정보가 한정되어 있고, 미세먼지 농도와 직접적으로 관련이 있는 대기질의 자료를 배제하여 정확성에 한계를 보였다. 강태천 등(Kang et al., 2017)은 Domain Adaptation 방법을 활용한 기계학습기반의 미세먼지 농도예측모형을 제시하고 있는데, 기상데이터와 교통데이터를 조합해서 사용하는 것이 특징이다. 하지만 구역별 미세먼지의 농도를 예측하는 데는 우수한 성능을 보였으나 시간이나 구체적인 위치에 따른 변화를 반영하지 못하는 단점을 가지고 있다. 차진욱 등(Cha et al., 2018)은 데이터마이닝 기법인 ANN(Artificial Neural Network) 알고리즘과 K-NN 알고리즘을 상호 응용하여 미세먼지를 예측하는 모형을 제안하여 ANN 및 K-NN보다 향상된 예측률을 보이고 있으나, 미세먼지의 예측은 일반적으로 범주형(매우나쁨, 나쁨, 보통, 좋음)으로 이루어지므로 범주형에 좋은 성능을 보이는 다른 예측기법들과의 직접적인 비교는 되지 못하고 있다.

2.2 미세먼지농도 예측모델 제안

위의 미세먼지농도 예측에 대한 기존 연구결과를 바탕으로 본 연구에서는 다음과 같은 모델을 제안한다.

일반적으로 미세먼지의 농도는 측정장소의 위치, 측정환경에 영향을 많이 받는다. 하지만 현재까지 우리나라의 미세먼지 측정소는 기상자료 측정소에 비해서 <Figure 2>에서 살펴본 바와 같이 매우 불규칙하게 설치되어 있고, 측정소가 설치되지 않은 장소가 많아서 예측에 어려움이 많다. 따라서 본 연구에서는 기상자료개방포털로부터 실시간으로 얻을 수 있는 16개의 다양한 기상자료와 에어코리아로부터 얻을 수 있는 4개의 대기질 관련 자료를 모두 수집하여, 모델의 입력변수로 활용한다.

고려하는 예측모델로는 앞의 문헌 연구(Joun et al., 2017; Oh et al., 2016)에서 살펴본 바와 같이 SVM(Support Vector Machine) 및 ANN(Artificial Neural Network)의 머신러닝기법이 다른 기법들에 비해서 상대적으로 우수한 성능을 보이는 것으로 실험적으로 밝혀졌으므로 같은 기법을 사용하기로 한다. 다만, 기상자료와 대기질 자료 등의 매우 많은 인자를 사용하므로 모델의 유의 인자를 선별하는 방법으로 다중선형회귀분석의 방법을 추가적으로 활용하기로 한다.

미세먼지는 계절성을 보이므로 시계열분석 모형이 성과를 낼 수 있는 것으로 평가(Oh et al., 2017)되기는 하지만 시계열분석 모형을 직접 활용하기 위해서는 미세먼지의 농도에 대한 과거의 자료가 축적되어 있어야 하므로 미세먼지 측정장치가 없는 지역에 대한 예측을 고려하는 본 연구에서는 시계열 예측모형은 사용하지 않기로 한다.

하지만 예측에 있어서 시간적인 요소는 매우 중요하므로 제안하는 모델에 측정한 데이터의 해당 월과 시를 입력변수로 포함시켜 예측에 있어서 계절성과 시간적인 요인을 포함하도록 한다.

3. 기상환경데이터와 미세먼지의 빅데이터 수집 및 처리

3.1 데이터수집

3.1.1 미세먼지 및 대기질 관련 빅데이터 수집

미세먼지 및 대기질 관련 빅데이터는 에어코리아(Air Korea) 홈페이지에서 수집하였으며, 우리나라의 대표적인 지역 중 한 군데를 지정하여 수집하였다. 데이터의 수집은 다양한 곳의 데이터를 수집하여 활용하는 것이 모델의 정확도를 높이는 데는 효과가 있을 것이지만, 그 양이 워낙 방대하여 예측모델 알고리즘의 학습을 수행하는 데 시간상의 물리적인 한계가 있을 것으로 판단되고, 수집된 자료가 미세먼지농도의 예측을 위해 제안하는 모델의 학습과 성과평가를 위해서 사용하는 것이 목적

이므로 한 군데만의 자료를 가지고 충분한 성과가 도출되는지를 알아보는 것에 한정하기로 한다.

지정한 장소는 서울시 종로구 종로 169(종묘주차장 앞)이며 수집한 자료는 미세먼지(PM10), 오존(O_3), 이산화질소(NO_2), 일산화탄소(CO), 아황산가스(SO_2)의 다섯 가지 측정값이다. 여기서 미세먼지(PM10)는 ($\mu g/m^3$)으로 측정되지만 일반적으로 다음과 같이 4가지 범주((좋음(0~30 이하), 보통(31~80 이하), 나쁨(81~150 이하), 매우나쁨(151 초과))로 나누어 사용한다. 수집기간은 2014년 1월 1일 00시부터 2017년 9월 30일 00시까지의 1시간 단위로 측정한 3년 9개월간의 자료이다.

<Table 1>은 에어코리아에서 수집한 자료의 종류와 일시에 따른 측정값 및 단위를 보여준다.

〈Table 1〉 Fine Dust Related Data(Air Korea)

Date (Y-M-D-H)	PM10 ($\mu g/m^3$)	O_3 (ppm)	NO_2 (ppm)	CO (ppm)	SO_2 (ppm)
14-01-01-00	163	0.004	0.04	0.8	0.015
14-01-01-01	152	0.004	0.04	0.9	0.012
14-01-01-02	153	0.003	0.042	0.9	0.011
14-01-01-03	159	0.003	0.043	1.0	0.012
…	…	…	…	…	…
17-09-29-21	15	0.005	0.4	0.017	0.039
17-09-29-22	17	0.006	0.4	0.016	0.041
17-09-29-23	18	0.005	0.4	0.018	0.037
17-09-30-00	15	0.00	0.4	0.017	0.036

3.1.2 기상관련 빅데이터 수집

<Table 1>에서 미세먼지 관련 측정치가 수집된 장소와 가장 가까운 기상관측소는 서울기상관측소(종로구)로 서울특별시 종로구 신문로2가 1-43에 위치해 있다. 따라서 서울기상관측소로부터 측정된 기상관련 자료를 '기상자료개방포털'로부터 얻을 수 있다. 수집된 기상관측자료는 기온, 강수량, 풍속, 풍향, 습도, 증기압, 이슬점온도, 현지기압, 해면기압, 일조, 일사, 전운량, 중하층운량, 최저운고, 시정, 지면온도의 16가지 자료이다. 미세먼지 관련 자료와 수집기간은 2014년 1월 1일 00시부터 2017년 9월 30일 00시까지로 미세먼지관련 자료와 연월일시가 일치하도록 하였다. <Table 2>는 기상자료개방포털에서 일시별로 수집한 자료의 종류와 측정단위를 보여준다.

〈Table 2〉 Weather Related Data(https://data.kma.go.kr/)

Date (Y-M-D-H)	14-01-01-00	14-01-01-01	14-01-01-02	14-01-01-03	…	17-09-29-21	17-09-29-22	17-09-29-23	17-09-30-00
Temperatures(°C)	3.3	2.6	1.7	1.4	…	3.3	2.6	1.7	1.4
Precipitation(mm)					…				
Wind velocity(m/s)	3,8	2.3	1.7	1.4	…	0.7	0.9	1	1.1
Wind direction(16Directions)	250	250	250	250	…	360	270	290	290
Humidity(%)	65	66	67	60	…	41	42	44	53
Vapor pressure(hpa)	5	4.9	4.6	4.1	…	2	2	2	2.4
Dew point temperature(°C)	-2.6	-3.1	-3.7	-5.5	…	4.5	5.2	5.2	6.4
Local Pressure(hpa)	1001.9	1002.2	1002.4	1002.5	…	1008.9	1009.5	1009.8	1009.6
Sea surface pressure(hpa)	1012.5	1012.9	1013.1	1012.9	…	1019	1019.6	1019.9	1019.8
sunshine(hr)	0.528	0.528	0.528	0.528	…	0.528	0.528	0.528	0.528
Solar radiation(MJ/m^2)	0.977	0.977	0.977	0.977	…	0.977	0.977	0.977	0.977
Total Cloudiness(10percentiles)	6	4.976	4.976	0	…	4	4.976	4.976	0
Middle and low Cloudiness(10percentiles)	6	3.086	3.086	0	…	0	0	0	0
Lowest Cloud height(100m)	10	13.182	13.182	13.182	…	13.182	13.182	13.182	13.182
Visibility(10m)	600	1400.137	1400.137	600	…	2000	2000	2000	2000
Ground temperature(°C)	0	-0.1	-0.3	-0.4	…	15.1	14.6	14.1	13.5

3.2 데이터 전처리

미세먼지관련 빅데이터와 기상관련 빅데이터는 서로 다른 테이블에 존재하는데, 우선 공통되는 날짜(년-월-일-시)를 키(key)로 하여 하나로 결합하였고, 미세먼지(PM10)를 종속변수로 하고 나머지 미세먼지관련자료(4), 기상자료(16), 날짜(월-시)(2)의 총 22개의 측정값을 독립변수로 설정하였다.

수집한 데이터에는 결측값이 존재하는데, 강수량의 결측값은 0으로 일조, 일사를 비롯한 변수들의 결측값은 평균값으로 대치했다.

년-월-일-시에 따라 수집된 총 자료의 수는 32,784개의 자료이다. 이 중에서 임의로 30,000개의 자료를 추출하여 훈련을 위한 자료로 사용하였고, 나머지 2,784개의 자료는 예측모형의 정확도를 평가하기 위한 시험자료로 사용하였다.

4. 머신러닝기법을 활용한 미세먼지의 예측모델

4.1 다중회귀분석에 의한 미세먼지 예측

다중회귀분석은 변수 간의 인과 관계를 통계적 방법에 의해 추정하는 회귀분석의 일종이다. 회귀분석에는 원인이 되는 독립변수와 결과가 되는 종속 변수가 존재하는데, 이때 종속 변수는 하나이고 독립변수가 2개 이상인 회귀 모델에 대한 분석을 수행하는 방법이 다중회귀분석(multiple regression analysis)이다. 다중회귀분석의 기본적인 목표는 다음과 같은 다중 회귀식에서 상수 및 계수를 추정하는 것이다.

$$Y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3 \cdots \beta_i x_i+\varepsilon_i$$

(x : 독립변수, Y : 종속변수, β : 회귀계수, β_0 : Y 절편, $\beta_0, ..., i$: 독립변수의 기울기, ε_i : 오차)

주어진 자료를 가지고 R프로그램을 활용하여, 모든 독립변수를 포함하여 다중회귀분석을 수행한 결과를 요약하면 <Table 3>과 같다.

<Table 3>으로부터 유의수준 5%에서 유의하지 않은 현지기압, 해면기압, 시의 독립변수 3개를 1차적으로 제거하였다. 제거 후의 독립변수들 간의 상관관계를 파악하기 위해서 다중공선성(multicollinearity)을 측정한 결과는 <Table 4>와 같다.

〈Table 3〉 Multiple Regression Analysis

	Estimate	Std. Error	t value	Pr (> |t|)
(Intercept)	347.436	34.214	10.155	0.000
O_3	118.158	11.753	10.054	0.000
NO_2	7.352	1.481	4.963	0.000
CO	27.593	0.668	41.325	0.000
SO_2	399.908	21.470	18.626	0.000
Temperatures	-2.197	0.171	-12.847	0.000
Precipitation	-0.914	0.149	-6.132	0.000
Wind velocity	0.771	0.129	6.000	0.000
Wind direction	0.031	0.001	20.899	0.000
Humidity	-1.129	0.039	-28.797	0.000
Vapor pressure	-1.073	0.060	-17.908	0.000
Local Pressure	**3.779**	**2.156**	**1.753**	**0.080**
Sea surface pressure	**-3.914**	**2.135**	**-1.834**	**0.067**
Dew point temperature	3.305	0.149	22.222	0.000
Sunshine	-7.327	0.757	-9.674	0.000
Solar radiation	2.625	0.468	5.613	0.000
Total Cloudiness	-0.873	0.068	-12.857	0.000
Middle and low Cloudiness	-0.458	0.073	-6.267	0.000
Lowest Cloud height	0.154	0.022	6.873	0.000
Visibility	-0.029	0.000	-82.226	0.000
Ground temperature	-0.367	0.042	-8.735	0.000
Month	-1.330	0.051	-26.001	0.000
Hour	**-0.020**	**0.024**	**-0.836**	**0.403**

〈Table 4〉 Multi-Collinearity Analysis Result

Variable	O_3	NO_2	CO	SO_2	Temperatures
Multi-collinearity	1.62	6.73	2.32	5.87	111.36
Variable	Precipitation	Wind velocity	Wind direction	Humidity	Vapor pressure
Multi-collinearity	1.02	1.32	1.18	28.29	11.15
Variable	Dew point temperature	sunshine	Solar radiation	Total Cloudiness	Middle and low Cloudiness
Multi-collinearity	150.61	3.02	3.85	2.65	2.42
Variable	Lowest Cloud height	Visibility	Ground temperature	Month	
Multi-collinearity	1.14	1.95	14.60	1.30	

일반적으로, 다중공선성 값이 5 이상이면 해당 변수가 다른 변수와 상관관계가 높아 변수의 회귀계수 추정을 어렵게 하여 위험하며, 10 이상이면 '매우위험'이라고 알려져 있다(Tofallis, 2016). 따라서 다중공선성 결과값이 10을 넘는 기온, 습도, 증기압, 이슬점온도, 지면온도의 5개변수를 2차적으로 제거하였고, 5를 넘는 NO_2와 SO_2는 피어슨의 상관계수값이 0.9를 넘는 것으로 조사되어 다중공선성값이 더 높은 NO_2를 제거하였다.

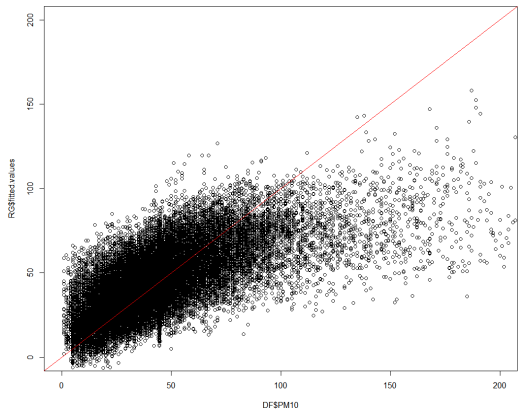
이제 최종적으로 남은 13개의 독립변수(O_3, CO, SO_2, 강수량, 풍속, 풍향, 일조, 일사, 전운량, 중하층운량, 최저운고, 시정, 달(월))을 대상으로 종속변수(PM10)에 대한 다중회귀분석을 수행하였다.

최종적으로 얻어진 회귀식은 다음과 같다.

$$\begin{aligned} PM10 = {} & 5.386 + 200.13O_3 + 31.21CO + 642.5SO_2 \\ & - 2.25\text{강수량} + 1.513\text{풍속} + 0.034\text{풍향} \\ & - 3.508\text{일조} + 2.426\text{일사} - 0.73\text{전운량} \\ & - 0.98\text{중하층운량} + 0.333\text{최저운고} \\ & - 0.023\text{시정} - 1.5\text{달} \end{aligned}$$

회귀식에 의한 추정값과 실제값의 분포도를 그래프로 나타내면 <Figure 3>과 같다.

회귀식의 정확도를 평가하기 위해 MAPE(mean absolute percentage error) 값을 계산한 결과 47.25를 얻었다. MAPE 값이 50 이하로 매우 합리적인 예측이 되고 있음을 알 수 있다(Tofallis, 2016).



〈Figure 3〉 Estimation and Actual Value Distribution According to Regression Equation

또한 회귀식의 예측 정확도를 측정하기 위해서 사전에 준비한 2,784개의 시험자료(test data)를 바탕으로 예측을 수행한 결과 <Table 5>와 같은 결과를 얻었다.

<Table 5>에서 보는 바와 같이, 다중회귀분석을 미세먼지 예측모델로 사용하여 예측하였을 경우, 예측과 실제의 범주값이 일치하는 비율은 2,784의 자료 중에서 1,886개가 정확히 일치하여 67.7%의 일치도를 보여준다. 하지만 2,784개의 자료 중에서 범주별로 주어지는 자료의 개수가 상이하므로 Confusion Matrix 형태로 정확도를 나타내면 다음과 같다. 여기서 '정밀도'는 예를 들어 예측을 '좋음'이라고 했을 때, 실제로도 '좋음'의 결과가 나온

비율이 70.4%임을 의미한다. 이와 대응해서 민감도는 예를 들어 미세먼지가 실제 '보통'이었는데 맞춘(예측) 비율이 84.3%라는 의미이다. F-measure는 정밀도와 민감도를 조합해서 조화평균으로 나타낸 수치이다.

〈Table 5〉 Multi-Regression Analysis Result of Fine Dust Estimation

real \ prediction	very bad	bad	average	good
very bad	0	19	16	0
bad	0	55	155	0
average	0	48	1348	203
good	0	1	456	483

〈Table 6〉 Accuracy of Multiple Regression Analysis Prediction Model

category	Precision	Recall	F-measure
very bad	0.000	0.000	0.000
bad	0.447	0.262	0.330
average	0.683	0.843	0.754
good	0.704	0.514	0.594

<Table 6>에서 살펴보면 다중회귀분석으로는 '매우나쁨'의 예측은 수행하지 못하는 것으로 보인다. 미세먼지가 '보통'인 경우가 84.3%로 가장 높은 민감도 예측율을 보이고 있음을 알 수 있다.

4.2 SVM에 의한 미세먼지 예측

본 절에서는 미세먼지의 예측을 위해서 SVM(Support Vector Machine)기법을 사용하였다.

4.2.1 SVM 이론

SVM은 Vapnik이 제안한 머신러닝 기법으로, 경험적 위험 최소화 원칙을 기반으로 하는 다른 통상적인 기계학습 기법과는 달리 구조적 위험 최소화를 기반으로 하여 일반화 오류의 상한을 최소화할 수 있는 머신러닝(machine learning) 기법이다(Cortes and Vapnik, 1995). 이 중에서 SVM은 ANN 기법의 문제점으로 지적되는 과적합(overfitting) 문제를 벌칙(penalty)항을 이용하여 피할 수 있으며, 또한 함수 근사에 있어서 이상데이터(outlier)에 둔감하기 때문에 높은 일반화 성능을 가진다. 따라서, 만약 동일한 데이터를 활용할 경우, 데이터의 특성에 따라 인공신경망 기법에 비해 상대적으로 예측력이 우수한 모델의 구현이 가능한 장점이 있다.

4.2.2 SVM을 활용한 미세먼지 예측모델

위에서 살펴본 바와 같이 범주형 데이터의 분류 능력이 뛰어난 SVM을 활용하여 미세먼지의 범주형 예측을 수행하고자 한다.

학습을 위한 자료는 다중회귀분석에서 사용하였던 자료와 동일한 자료를 사용하였다. 단, 자료의 값들 간의 차이가 크므로 값을 0~1사이의 값으로 표준화((x-min(x))/(max(x)-min(x)), 여기서 x = 측정값)하여 사용하였다. 미세먼지에 대한 측정값(PM10)은 앞에서 언급한 기준에 따라 '매우나쁨', '나쁨', '보통', '좋음'의 네 가지 범주로 구분하여 사용하였다.

SVM의 실행을 위해서 R 프로그램의 e1071 패키기를 사용하였다. Soft margin SVM 방법을 선택하였고, 커널은 비선형 함수인 다양한 형태의 데이터에서 가장 잘 수행되는 Gaussian RBF(Radial Basis Function) 커널인 'radial'을 선택했다(Liu et al., 2015).

SVM의 예측률을 더욱 높이기 위해 SVM 모델의 결정 경계선의 폭을 변경 할 수 있는 제약 매개변수(cost)의 비용을 결정하여 soft 마진을 택할 수 있다. 다시 말해, 초평면의 기울기인 gamma 값과 과적합에 따른 비용 cost의 최적조합을 결정하여 과적합 되지 않는 SVM 모델을 만들 수 있다. 따라서 과적합과 과소적합을 일으키지 않는 (cost, gamma) 값의 최적 조합이 중요한데, 본 연구에서는 R의 튜닝함수(tune.svm())와 사전실험을 통해서 결정하였다. 튜닝을 위해 gamma값은 $2^{-1} \sim 2^{1}$을 cost 값은 $2^{-2} \sim 2^{5}$을 사용하였고, 실험결과 (gamma, cost) = (0.5, 1.0)에서 최적의 결과를 보여줌을 확인하고

SVM 예측에서 매개변수로 사용하였다. SVM을 활용한 미세먼지 예측 R코드의 일부는 다음과 같다.

dust_svm ← svm(PM10~O_3+CO+SO_2+강수량+풍속+풍향+일조+일사+전운량+중하층운량+최저운고+시정+달, data = dust_train, scaled = TRUE, kernel = "radial", gamma = 0.5, C = 1)

위와 같은 방법을 통해 얻어진 SVM 예측모델을 활용하여 다중회귀분석에서 사용한 자료와 같은 2,784개의 시험자료에 대해서 미세먼지 예측을 수행한 결과를 요약하면 <Table 7>과 같다.

〈Table 7〉 Result of Fine Dust Prediction by SVM

real \ prediction	very bad	bad	average	good
very bad	2	12	21	0
bad	0	41	165	4
average	0	10	1391	198
good	0	1	333	606

<Table 7>에서 보는 바와 같이, SVM을 미세먼지 예측모델로 사용하여 예측하였을 경우, 예측과 실제의 범주값이 정확히 일치하는 비율은 2,784의 자료 중에서 2,040개가 정확히 일치하여 73.3%의 일치도를 보여준다. 하지만 다중회귀분석에서와 마찬가지로 Confusion Matrix 형태로 정확도를 나타내면 다음과 같다.

〈Table 8〉 Accuracy of SVM Prediction Model

category	Precision	Recall	F-measure
very bad	1.000	0.057	0.108
bad	0.641	0.195	0.299
average	0.728	0.870	0.793
good	0.750	0.645	0.693

<Table 8>의 결과를 살펴보면, SVM을 사용한 예측모델은 다중회귀모델보다 높은 정밀도를 보임을 알 수 있다. '좋음'의 경우에 75.0%로 가장 높은 정밀도를 보였고, '보통' 범주에서 가장 높은 민감도 87.0%를 보인다.

4.3 ANN에 의한 미세먼지 예측

본 절에서는 미세먼지의 예측을 위해서 ANN 기법을 사용하였다.

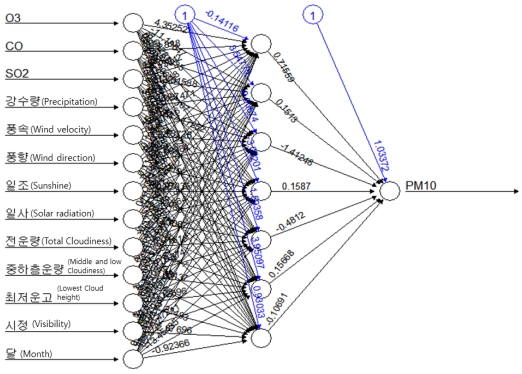
ANN은 머신러닝과 인지과학에서 동물(특히 인간)의 신경망에서 영감을 얻은 통계학적 학습 알고리즘이다. 다른 층의 뉴런(노드)들 사이의 연결 패턴, 연결의 가중치를 갱신하는 학습과정, 마지막으로 뉴런의 가중입력을 활성화도 출력으로 바꿔주는 활성화 함수, 이 세 가지의 인자를 이용해 정의된다(Hagan et al., 2014).

일반적으로 ANN은 [입력층(input layer)-은닉층(hidden layer)-출력층(output layer)]으로 구성되는데 본 연구에서도 3층의 다층모형을 사용하였다. 입력층의 각 노드에는 다중회귀분석에서 독립변수로 결정되었던 13개의 변수(O_3, CO, SO_2, 강수량, 풍속, 풍향, 일조, 일사, 전운량, 중하층운량, 최저운고, 시정, 달(월))를 사용하였고, 각 입력 노드(node)에 각 기상자료 값을 할당하였다. 입력층으로 입력되는 학습자료는 다중회귀분석에서 사용하였던 자료와 같은 자료를 사용하였으며, 또한 SVM에서와 마찬가지로 0~1사이의 표준화된 값을 사용하였다.

ANN에서 은닉층을 어떻게 구성하느냐에 따라서 출력값의 정확도에 많은 영향을 끼치게 된다. 따라서 본 연구에서도 가장 적합한 은닉층의 노드 수를 정하기 위해서 사전실험을 수행하였다. 사전실험결과로부터 은닉층의 노드 수가 7인 경우에 가장 높은 정확도를 보여, ANN의 본 모델에서도 은닉층의 노드 수를 7로 하였다.

출력층은 최종 예측변수인 미세먼지(PM10)을 고려하여 1개의 노드로 구성하였다. 에어코리아에서 처음에 주어지는 미세먼지(PM10)값은 단위가 ($\mu g/m^3$)인 실수 값이므로, 그대로 실수 값으로 사용하였으며 다만 ANN의 출력 값 특성을 고려하여 0~1사이의 표준화된 값으로 변환하여 사용하였다.

최종적으로 구성된 ANN 모델은 <Figure 4>와 같다.



〈Figure 4〉 ANN Model for Prediction of Fine Dust Concentration

구성된 ANN에 대해서 30,000개의 학습자료를 사용하여 ANN의 weight 값들을 훈련시켰다. 학습에서는 R프로그램의 neuralnet() 패키지를 사용하였으며, 은닉층의 노드 수는 7로 하였다. 학습 후 최종적으로 얻어진 ANN 미세먼지 예측모델과 입력층, 은닉층, 출력층간의 각 아크(arc)에 대한 가중치 최종값을 <Figure 4>에서 확인할 수 있다. 위에서 얻어진 ANN 예측모델을 활용하여 2,784개의 시험자료에 대해서 예측을 수행하였고 요약하면 <Table 9>와 같다.

〈Table 9〉 Result of Fine Dust Prediction by ANN

real \ prediction	very bad	bad	average	good
very bad	6	20	9	0
bad	0	98	110	2
average	0	73	1346	180
good	0	2	357	581

<Table 9>로부터 ANN의 경우 예측과 실제의 범주값이 정확히 일치하는 비율은 2,784의 자료 중에서 2,031개가 정확히 일치하여 73.0%의 일치도를 보여준다. 하지만 앞의 경우와 동일하게 Confusion Matrix로 정확도를 요약하면 다음과 같다.

〈Table 10〉 Accuracy of ANN Prediction Model

category	Precision	Recall	F-measure
very bad	1.000	0.171	0.293
bad	0.508	0.467	0.486
average	0.739	0.842	0.787
good	0.761	0.618	0.682

<Table 10>으로부터 ANN 예측모델의 경우 '매우나쁨'에서 정밀도가 100%로 '매우나쁨'으로 예측한 경우는 실제로도 모두 '매우나쁨' 결과가 나왔음을 알 수 있다. 또한 민감도는 '보통'의 범주에서 84.2%로 가장 높게 나타났다.

4.4 예측모델의 비교

앞에서 기상자료를 활용하여 미세먼지농도를 예측하기 위한 모델로 다중회귀분석, SVM, ANN의 세 가지를 활용하였고, 2,784개의 시험자료에 대해서 각각 예측을 수행하고 정확도를 평가하였다.

하지만, 예측값이 '매우나쁨' 또는 '나쁨'을 예측하였을 때, 실제값이 '매우나쁨' 또는 '나쁨'으로 실제 관측되는 경우나 '보통' 또는 '좋음'으로 예측하였을 때 '보통' 또는 '좋음'의 실제값을 보일 경우 일반적으로 용인되는 수준이므로, 이것을 고려하면 기상환경자료와 다중회귀분석, SVM 및 ANN을 이용한 미세먼지 예측모델의 정확도는 <Table 11>과 같다.

4.4.1 정밀도(Precision) 비교

<Table 11>에서 보는 바와 같이 예측한 범주 내에서 실제로 예측이 맞은 비율을 나타내는 정밀도 측면에서 살펴보면,

경우 1) '나쁨/매우나쁨' 범주 :
다중회귀 < ANN < SVM
경우 2) '보통/좋음' 범주 :
SVM < 다중회귀 < ANN

와 같다. 따라서, 정밀도 측면에서는 종합적으로 ANN을 사용한 예측모델이 가장 정밀도를 보임을 알 수 있다.

〈Table 11〉 Accuracy Comparison of Multiple Regression, SVM, and ANN Prediction Models

prediction models	Category	Precision	Recall	F-measure
Multiple Regression	bad/very bad	0.602	0.302	0.402
	average/good	0.936	0.981	0.958
SVM	bad/very bad	0.833	0.224	0.354
	average/good	0.930	0.996	0.962
ANN	bad/very bad	0.623	0.506	0.559
	average/good	0.953	0.970	0.962

4.4.2 민감도(Recall) 비교

민감도는 실제의 어는 범주에 속하는 미세먼지의 농도를 정확히 예측하였을 비율을 나타낸다. <Table 11>로부터 두 가지 경우로 나눠서 살펴보면 다음과 같다.

경우 1) '나쁨/매우나쁨' 범주 :
SVM < 다중회귀 < ANN
경우 2) '보통/좋음' 범주 :
ANN < 다중회귀 < SVM

민감도 측면에서 '나쁨/매우나쁨' 범주에서 SVM과 다중회귀모델의 경우가 22.4%, 30.2%로 매우 낮은 반면, ANN의 경우는 50.6%로 상대적으로 높게 나타나서 민감도는 ANN의 경우가 좋은 것으로 평가된다.

4.4.3 종합비교

정밀도와 민감도를 종합(조화평균)하여 보여주는 F-measure를 중심으로 다중회귀, SVM, ANN 모델의 정확도를 비교하기로 한다.

'보통/좋음'의 범주에서는

다중회귀(95.8%) < SVM, ANN(96.2%)

로 세 모델 모두 매우 높은 정확도를 보일 뿐만아니라 그 차이도 0.4%point 이내로 매우 작다.

반면에, '나쁨/매우나쁨'의 범주에서는

SVM(35.4%) < 다중회귀(40.2%) < ANN(55.9%)

로 나타나 모델간의 차이도 상당히 크고, 정확도 측면에서도 SVM과 다중회귀는 매우 낮은 반면, ANN은 55.9%로 상대적으로 높게 나타났다.

종합적으로 볼 때, 미세먼지와 대기질관련 자료를 사용하고 예측 알고리즘으로 ANN을 사용한 미세먼지 예측모델이 가장 높은 정확도를 보임을 실험의 결과를 통해서 알 수 있다.

이것은 제시한 예측모형과 비교 데이터의 불일치로 직접적인 비교는 어려울 수 있지만, 가장 최근에 발표된 미세먼지농도 예측관련 연구결과(Cha et al., 2018)에서 제시한 예측정확도 83.4%와 비교해서도 우수한 결과로 판단된다.

5. 결 론

본 연구에서는 미세먼지 농도의 예측을 위해서 미세먼지 농도와 상관관계가 매우 높을 것으로 밝혀진 기상자료와 대기질 관련 환경자료를 활용하였다.

예측을 위한 모델로는 다중회귀분석, SVM, ANN을 사용하였다. 각 예측모델에 대해서 시험자료를 바탕으로 예측을 수행한 결과 예측 정확도는 '보통/좋음'의 범주에서 다중회귀 : SVM : ANN = 95.8% : 96.2% : 96.2%를 얻어 세 모델 모두 95% 이상의 매우 높은 정확도를 보임을 확인할 수 있었다. 반면에 '나쁨/매우나쁨'의 범주에서는 다중회귀 : SVM : ANN = 40.2% : 35.4% : 55.9%를 보여 세 모델 모두 높지 않은 정확도를 보였으나 ANN 모델의 경우가 상대적으로 높은(55.9%) 정확도를 보였다. 종합적으로 볼 때, 기상환경자료와 머신러닝기법 중의 하나인 ANN을 활용한 미세먼지 예측모델은 매우 효과가 높은 것으로 판단된다.

본 연구에서 다룬 예측모델은 학술적으로 미세먼지농도의 예측에 머신러닝기법을 활용하여 예측정확도를 높였다는 데에 의의가 있다. 일반적으로 미세먼지농도에 큰 영향을 미치는 요소는 기상데이터 및 대기환경데이터를 들 수 있는데, 이 데이터들은

방대한 양의 빅데이터로 기존의 일반적인 통계기법으로 예측할 경우 그 정확도가 낮아질 수 있는 단점이 있었다. 또한 본 연구에서 제안한 미세먼지농도 예측모델은 국외적인 요인이나 기타 환경적인 요인을 고려하지 않고, 손쉽게 획득이 가능한 기상 및 대기환경자료만을 가지고도 미세먼지의 예측을 효과적으로 수행할 수 있음을 보이고 있어 실무적 활용도가 매우 높을 것으로 판단된다.

다만, 본 연구의 결과비교에서 살펴본 바와 같이 예측모델의 한계로 나타난 '나쁨/매우나쁨'의 범주에서는 정확도가 높지 않은 편으로 해당범주의 예측율을 높이기 위한 모델의 보완이 추가적으로 요구된다. 또한 시간적 물리적 제약으로 인해 특정 장소에서만의 기상환경자료를 활용하여 모델을 학습하고 시험하였으나, 모델의 정확도를 보다 더 높이기 위해서는 다양한 곳의 기상환경데이터를 확보하여 학습할 수 있도록 하여야 할 것이다. 더불어 약 3년간의 기상환경데이터를 확보하여 예측에 활용하였으나 머신러닝과 빅데이터분석의 장점을 충분히 활용하기 위해서 확보가능한 과거의 충분한 데이터를 확보하여 학습에 활용하면 '나쁨/매우나쁨'의 범주예측에서 나타난 정확도의 한계를 극복할 수 있을 것으로 판단된다.

References

Cha, J.W. and J.Y. Kim, "Development of Data Mining Algorithm for Implementation of Fine Dust Numerical Prediction Model", *Journal of the Korea Institute of Information and Communication Engineering*, Vol.22, No.4, 2018, 595-601.

(차진욱, 김장명, "미세먼지 수치 예측 모델 구현을 위한 데이터마이닝 알고리즘 개발", *한국정보통신학회논문지*, 제22권, 제4호, 2018, 595-601.)

Cortes, C. and V. Vapnik, "Support-vector networks", *Machine Learning*, Vol.20, No.3, 1995, 273-297.

Hagan, M.T., H.B. Demuth, M.H. Beale, and O. Jesus, *Neural Network Design (2nd Edition)*, Martin Hagan, 2014.

Joun, S.W., J.Y. Choi, and J.H. Bae, "Performance Comparison of Algorithms for the Prediction of Fine Dust Concentration", *Proceedings of Korea Computer Congress, The Korean Institute of Information Scientists and Engineers*, No.12, 2017, 775-777.

(전송완, 최제열, 배준현, "미세먼지 농도 예측 알고리즘 성능 비교", 한국정보과학회 학술발표논문집, 제12호, 2017, 775-777.)

Kang, T.C. and H.B. Kang, "Machine Learning-based Estimation of the Concentration of Fine Particulate Matter Using Domain Adaptation Method", *Journal of Korea Multimedia Society*, Vol.20, No.8, 2017, 1208-1215.

(강태천, 강행봉, "Domain Adaptation 방법을 이용한 기계학습 기반의 미세먼지 농도 예측", *멀티미디어학회논문지*, 제20권, 제8호, 2017, 1208-1215.)

Koo, Y.S., H.Y. Yun, H.Y. Kwon, and S.H. Yu, "A Development of PM10 Forecasting System", *Journal of Korean Society for Atmospheric Environment*, Vol.26, No.6, 2010, 666-682.

(구윤서, 윤희영, 권희용, 유숙현, "미세먼지 예보시스템 개발", *한국대기환경학회지*, 제26권, 제6호, 2010, 666-682.)

Lee, J.G., *Multivariate Analysis and Data Mining with R*, Hwangso Academy, 2016.

(이재길, *다변량분석 및 데이터마이닝*, 황소걸음아카데미, 2016.)

Lee, M.H., Korea-China collaborative study to abate trans-boundary air pollution(II), *National Institute of Environmental Research, Research Report*, 2016. 6.

(이미혜, *한·중 월경성 미세먼지 저감을 위한 공동연구(II)*, 국립환경과학원, 최종보고서, 2016. 6.)

Liu, Z., M.J. Zuo, X. Zhao, and H. Xu, "An Analytical Approach to Fast Parameter Selection of Gaussian RBF Kernel for Support Vector Machine", *Journal of Information Science and Engineering*, Vol.31, No.2, 2015, 691-710.

Oh, B.D., J.H. Park, and Y.S. Kim, "Prediction of the concentration of PM10 using Machine-Learning", *Proceedings of The Korean Institute of Information Scientists and Engineers Winter Conferences*, No.12, 2016, 1674-1676.

(오병두, 박지후, 김유섭, "Machine-Learning을 활용한 미세먼지(PM10) 농도 예측", *한국정보과학회 학술발표논문집*, 제12호, 2016, 1674-1676.)

Oh, J.M., H.S. Shin, Y.S. Shin, and H.C. Jeong, "Forecasting the Particulate Matter in Seoul using a Univariate Time Series Approach", *Journal of the Korean Data Analysis Society*, Vol.19, No.5, 2017, 2457-2468.

(오종민, 신현수, 신예슬, 정형철, "시계열 분석을 활용한 서울시 미세먼지 예측", *한국자료분석학회지*, 제19권, 제5호, 2017, 2457-2468.)

Shin, E.K., J. Kim, and Y. Choi, "A Study on the Data Model Design of Fine Dust Related Disease", *Journal of The Korea Society of Information Technology Policy & Management*, Vol.10, No.1, 2018, 655-659.

(신경은, 김재범, 최용락, "미세먼지 관련 질병 데이터 모형 설계 연구", *한국IT정책경영학회논문지*, 제10권, 제1호, 2018, 655-659.)

Shin, M.K., C.D. Lee, H.S. Ha, C.S. Choe, and Y.H. Kim, "The Influence of Meteorological Factors on PM10 Concentration in Incheon", *Journal of Korean Society for Atmospheric Environment*, Vol.23, No.3, 2007, 322-331.

(신문기, 이충대, 하현섭, 최춘석, 김용희, "기상인자가 미세먼지 농도에 미치는 영향", *한국대기환경학회지*, 제23권, 제3호, 2007, 322-331.)

Tofallis, C., "A Better Measure of Relative Prediction Accuracy for Model Selection and Model Estimation", *Journal of the Operational Research Society*, Vol.66, No.8, 2016, 1352-1362.

About the Authors

Joon-Mook Lim (jmlim@hanbat.ac.kr)

Professor Joon-Mook Lim is currently a Professor of Bigdata Analytics at Department of Creative Convergence Engineering, Hanbat National University. He received his Ph.D. in Industrial Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1994. His current research interests include intelligent data analysis, optimization, data mining and simulation.