한국역신악회지 제15권 제4호
Volume 15, Number 4, Nov. 2020
https://doi.org/10.46251/INNOS.2020.11.15.4.339

미세먼지 예측을 위한 기계학습 모델 간 성능 비교 연구: 국내 발생 데이터를 중심으로

성상하* · 김상진** · 류민호***

초 록

미세먼지는 대기오염의 주요인으로 각종 호흡계 질병을 초래해 건강을 위협하고 있다. 이러한 상황에서 미세먼지를 줄이기 위해 많은 정책들이 발표되었지만 효과는 크지 않은 상황이다. 본 연구는 국내 미세먼지 수치에 영향을 미치는 다양한 변수들을 분석하고, 미세먼지 수치를 예측하는 데 있어서 최적의 알고리즘을 제안한다. 미세먼지에 영향을 미치는 주요 변수는 기상, 대기오염물질 유입 등의 해외 유입 요인과 국내에서 발생되는 대기오염 요인으로 구분된다. 국내의 미세먼지 수치를 예측하기 위해 교통량, 화력발전량 등의 내부적 요인의 데이터를 활용했다. 분석에는 엑스지부스트(XGBoost), 랜덤포레스트(Random Forest), 서포트벡터머신(SVM), 인공신경망(ANN)의 알고리즘이 적용되었고, 제시된 모형을 통해 미세먼지 수치 예측을 진행했다. 분석 결과, 교통량, 화력발전 등 국내 관련 변수가 추가됨에 따라 기계학습 모형의 예측 정확도가 증가되는 것을 확인할 수 있었다. 또한, Random Forest와 XGBoost를 활용해 변수중요도를 추출한 결과 대기오염물질이 미세먼지를 예측하는데 가장 중요한 변수로 평가되었다. 제시된 알고리즘 중에서는 인공신경망 기법이 가장 예측 정확도가 높은 것을 확인할 수 있었다. 본 연구를 통해 미세먼지 수치 예측에 필요한 국내 변수들의 영향력을 파악하고, 적합한 예측모형을 선정할 수 있다. 또한 이를 활용해 향후 미세먼지 수치를 예측할 수 있다.

주제어; 미세먼지, 국내 요인, 예측, 기계학습, 인공신경망

논문접수일 2020. 9. 26 논문수정일 2020. 11. 22 논문게재확정일 2020. 11. 23

* 제1저자, 동아대학교 경영정보학과 박사과정, sangha@donga.ac.kr
** 공동저자, 동아대학교 경영정보학과 교수, skim10@dau.ac.kr
*** 교신저자, 동아대학교 경영정보학과 교수, ryumh12@dau.ac.kr
본 연구는 동아대학교 연구지원을 받아서 작성되었습니다.

Ⅰ. 서론

최근 대기 오염으로 인한 다양한 경제 사회적 문제가 야기되면서 이에 대한 세계적인 관심이 증가하고 있다. 세계보건기구는 전 세계 92%가 대기 오염으로 인해 영향을 받고 있다고 보고했으며, 해마다 많은 사람들이 직접적인 피해를 입고 있다고 밝혔다. 국내의 경우 대기 오염의 주요 원인 중 하나로 먼지를 꼽을 수 있다. 먼지란 대기 중에 떠다니거나 흩날려오는 작은 물질을 의미하는데, 이 중 지름이 10μm미만의 아주 작은 물질을 미세먼지라고 한다. 즉, 미세먼지란 대기 중에 떠다니며, 눈에 보이지 않을 정도로 작은 먼지를 일컫는다. 이처럼 미세먼지는 매우 작기 때문에 체내에 침투하여 건강에 악영향을 미친다. 미세먼지 입자에는 금속, 질산염, 황산염 등 호흡기 질환을 일으킬 수 있는 이물질을 포함한다. 따라서 호흡기 질환뿐만 아니라 다양한 질환을 유발할 수 있다. 이러한 미세먼지의 원인은 산불, 황사와 같은 자연적인 원인과 매연, 공장 등과 같은 인위적인 원인이 있다.

중위도 지역에 위치한 우리나라는 편서풍의 영향으로 인해 중국에서 배출된 오염물질에 직접적으로 노출되어 있다. 이러한 오염물질은 국내 대기 오염에 큰 영향을 끼치며, 국내 미세먼지의 주요 발생 원인은 중국발 미세먼지라고 할 수 있다. 하지만 국내에서도 공장이나 발전소, 자동차, 선박 등에서 발생하는 대기 오염물질로 인해 미세먼지가 발생하고 있다. 특히, 석유, 석탄 등과 같은 화석연료가 타면서 나오는 대기오염물질이 국내 미세먼지의 주 유발원인으로 알려져 있기 때문에 국내에서도 미세먼지 오염도를 저감시키기 위해 다양한 노력이 필요하다(박순애, 신현재, 2017). 국내에서는 차량 운행 제한, 수소차 보급 장려 등의 노력을 하였으나, 효과는 미비하였다. 또한 미세먼지를 줄이기 위한 다양한 정책이 논의되고 있으나, 실질적인 효과에 대한 분석이 부족하다(김형건, 2018). 이러한 무분별한 미세먼지 저감 정책이 실행될 경우 불필요한 자원을 낭비하게 된다. 국립기상과학원에서는 많은 국가예산을 투입해 여러 기상예측모델을 개선하려고하고 있지만 그 결과물에 대한 가치는 정확히 파악되고 있지 않다(김혜민 et al, 2020). 이렇듯, 미세먼지 문제에 대한 사람들의 관심이 늘어남에 따라 환경적 요인이 개인 삶의 만족도에 큰 영향을 미치고 있다(이미숙 & 진형익, 2018).

본 연구에서는 기계학습 기법을 활용해 국내 기상 데이터를 분석하고, 미세먼지 농도를 예측할 수 있는 방법을 찾고자 한다. 기계학습 기법을 활용할 경우 일

반적인 선형 예측모형에서 필요한 선형 가정의 제약에서 벗어날 수 있으며, 예측력이 높다는 장점이 있다(이태형, 전명진, 2018). 이에 본 연구에서는 기계학습 기법을 활용해 미세먼지 데이터를 분석하고자 한다.

따라서 본 연구에서는 미세먼지에 영향을 미칠 수 있는 변수 중 국내 변수들을 분석하여 각 변수의 영향력을 파악하고, 미세먼지 수치에 가장 큰 영향을 미치는 변수가 무엇인지 알아보고자 한다. 또한 다양한 기계학습 기법을 활용해 미세먼지 예측모델을 구축하고, 예측모델 간 예측 정확도 비교를 통해 미세먼지 수치 예측에 가장 적합한 기계학습 모델을 선정한다. 모델에 대한 일반적인 성능 측정 및 비교 결과뿐만 아니라 이전 변수 값을 활용하여 향후 발생할 미세먼지 수치에 대한 예측 결과를 제시하고자 한다.

본 논문의 구성은 다음과 같다. 제1장 서론에서는 본 연구의 배경과 필요성에 대해 설명하고 연구의 범위 및 연구 방법을 정의했다. 제2장에서는 이론적배경과 미세먼지 예측 관련 선행연구에 대해 서술했다. 미세먼지에 영향을 미치는 변수와 예측 알고리즘에 대한 내용이 포함되어 있다. 제3장에서는 본 연구에서 활용된 예측 관련 방법론에 대해 서술한다. 제4장에서는 연구에서 활용한 데이터에 대해 설명하고, 알고리즘의 성능을 확인한다. 마지막 제5장에서는 본 연구의 결론과 향후 연구방향에 대해 제시한다.

Ⅱ. 이론적 배경

2.1 미세먼지 수치에 미치는 변수

미세먼지의 원인을 찾기 위한 다양한 연구들이 진행되고 있다. 김형건(2018)은 미세먼지의 원인 중 하나로 우리나라의 연료 소비량에 대해 연구를 진행하였으며, 간접적으로 영향을 미칠 수 있는 것으로 나타났다(김형건, 2018).

강연욱 등(2019)은 국내 미세먼지 원인 중 하나로 화력발전소를 꼽았으며, 미세먼지를 저감하기 위한 모니터링 기술의 중요성에 대해 연구를 진행했다(강연욱 et al, 2019). 유태종 등(2019)은 석탄 화력 발전소에서 배출되는 미세먼지의 거동에 대한 연구를 수행하였으며, 여러 기상상황에 따라 미세먼지 수치가 변화함을 확인했다. 풍속이 약하거나 풍향이 일정하지 않을 경우 미세먼지 정체 현상

이 발생하면서 일부 지역에 미세먼지 수치가 높게 나타남을 확인할 수 있었다(유태종, 유동현, 2019). 또한 일부 연구에서는 시멘트와 석회 공장의 미세먼지 배출량이 주요 원인으로 꼽고 있다(조성환 et al, 2016). 이밖에도 김다빈 등(2018)은 고농도 미세먼지의 발생 원인을 연료연소, 산업생산, 자동차 배기가스 등을 원인으로 꼽고 있다(김다빈, 문윤섭, 2018).

또 다른 선행연구에서는 기후적 요인과 대기성분 요인을 미세먼지의 원인으로 꼽고 있다. 박기형 등(2019)은 미세먼지 농도의 원인으로 계절적 요인을 제시했다. 연구결과 지역 배출 및 국지 기상이 미세먼지 농도에 영향을 끼치는 것으로 나타났다(박기형 et al, 2019). 박윤서 등(2018)은 대기성분과 풍향이 미세먼지 수치에 영향을 미치는 것으로 주장했다. 강원도 지역의 대기오염물질 배출량을 풍향에 따라 측정하여 농도변화율을 산출했다. 연구결과 서풍계열이 불 경우 농도변화율이 크게 나타남을 확인할 수 있었다(박윤서 et al, 2018).

요약하면, 미세먼지에 대한 원인으로 풍향 풍속 등 대기의 흐름에 의한 자연유입 요인과 자동차배기가스, 연료 및 산업생산 등의 산업 발생요인으로 구분해 볼 수 있다. 이는 〈표-1〉에 정리되어 있다. 〈표-1〉에서 확인할 수 있듯이 많은 연구들이 저마다 중요한 변수를 활용해 미세먼지를 예측하고자 하였다. 하지만 국내 미세먼지 발생의 다양한 요인을 다룬 연구는 거의 없었다. 또한 기계학습 기법을 통해 미세먼지 수치를 예측하고자 할 때, 모델 성능의 단순 비교에 그쳤다. 본 연구는 이 자연유입 요인과 산업 발생요인의 변수들을 모두 활용해, 미세먼지의 주원인을 파악한다.

〈표-1〉 참고문헌의 주요 활용 변수

참고문헌	주요 활용 변수
김형건, 2018	국내 연료 소비량
김다빈, 문윤섭, 2018	산업 및 자동차 배기가스
강연욱 et al., 2019	석탄 화력 발전소 미세먼지 배출량
조성환 et al., 2016	석회 공장 미세먼지 배출량
유태종, 유동현, 2019	풍속과 풍향
박윤서 et al., 2018	풍향
박기형 et al., 2019	계절적 요인

2.2 미세먼지 수치 예측

최근 기계학습에 대한 다양한 기법들이 개발되면서, 기계학습을 통해 미세먼지의 수치를 예측하고자 하는 연구가 많이 진행되었다. 전송완 등(2017)은 다양한 예측 모델을 활용해 미세먼지 농도를 예측하고자 했다. 일반적인 예측 모델인 MLR(Multi-Linear Regression; MLR), ARIMA(Auto Regressive Integrated Moving Average; ARIMA)와 기계학습 기법 중 하나인 SVR(Support Vector Regression; SVR)을 사용해 미세먼지 농도를 예측하고자 하였다. 활용된 데이터는 대구지역의 대기질 정보(NO2, SO2, CO, O3, PM10)과 기상정보(기온, 강수량, 풍속)을 활용해 예측을 실시했다. 예측결과 많은 변수를 고려했을 경우는 SVR 모델이 뛰어난 예측력을 보였으나, 시간변수만을 고려했을 때는 ARIMA 모형의 정확도가 더 높은 것으로 나타났다(전송완 et al, 2017).

서양모 등(2019)은 딥러닝 기법 중 하나인 LSTM(Long-Short Term Memory; LSTM)을 활용해 미세먼지 오염정도를 분류하고자 했다. LSTM은 일반적인 딥러닝 모델에 비해 시간적 특성을 보존할 수 있다는 특징을 활용해, 2015년 1월 1일부터 2018년 9월 30일까지의 서울지역의 미세먼지 농도를 예측했다. 대기질 데이터를 활용해 미세먼지 농도를 예측하였으며, 이를 일정 기준에 의해 분류했다. 실험결과, Hidden Node의 수준을 늘릴 경우 더 높은 정확도를 얻을 수 있다는 결론을 얻었다(서양모 et al, 2019). LSTM의 높은 정확도로 인해 많은 연구에서 이를 추가적으로 활용했다. 조경우 등은 천안 지역의 기상 데이터와 대기오염 물질 데이터를 활용해 미세먼지 농도를 예측했으며, 그 결과 평균 87.59%의 결과값을 확보했다(조경우, 2019). 김종수 등(2019)은 일반적인 기상 데이터와 대기오염 물질 데이터뿐만 아니라 강설량과 압력, 습도 등의 추가적인 변수를 활용해 분석을 진행하였다. 그 결과, 단기적인 예측에서는 높은 성능을 보였지만, 장기예측에서는 예측의 한계점이 나타났다(김종수, 이창훈, 2019). 정용진 등은 LSTM에 비해 간소화된 GRU(Gated Recurrent Unit) 모델을 활용해 미세먼지 농도를 예측했다. GRU는 LSTM의 구조를 단순화한 모형으로 적은 컴퓨팅 성능을 요구함과 동시에 빠른 학습 속도를 가지고 있다. 기상데이터와 대기오염 물질 데이터를 활용해 미세먼지 농도를 범주화하여 분류하였으며, 분류 정확도는 PM10 기준 약 87%, PM2.5 기준 약 89%의 정확도를 나타냈다(정용진 et al., 2019).

본 연구에서는 선행연구에서 활용된 변수를 참고하여 변수를 선정한 뒤, 미세먼지 수치를 예측하고자 한다. 예측의 정확성을 높이기 위해 미세먼지의 수치 영향을 주는 다양한 변수들이 활용되었다. 특히, 국내 미세먼지 농도 증가의 요인 중 하나인 석탄 화석 연료 사용량을 반영하기 위해 교통 요인와 발전소 요인를 추가적으로 활용했다(김형건, 2018; 강연욱 et al, 2019). 발전소 요인에 대한 변수로는 월 평균 화력 발전량을 활용했다. 교통량 변수는 주요 항의 선적 운행에 대한 데이터를 활용했다. 선적 운행 데이터의 경우 기존 선행연구에서 거의 다뤄지지 않은 변수이다. 하지만 최근 국내 주요항의 미세먼지 농도가 선적 운행에 의해 발생한다는 연구결과가 나타나고 있다(육근형, 2017). 따라서 본 연구에서는 선적 운행에 대한 데이터를 추가적으로 활용하여 예측 정확도를 높이고자 한다. 본 연구의 목적은 국내 요인들이 미세먼지에 대한 예측력이 얼마나 되는지 확인하는데 있다. 또한, 본 연구는 미세먼지 농도 예측을 위한 최적의 모델을 선정하기 위해 다양한 모델 간 성능 비교를 실시하고 있으며, 미세먼지 예측에 가장 큰 영향을 주는 변수가 무엇인지를 탐색한다. 뿐만 아니라 기존에 성능 비교로 그쳤던 연구를 확장시켜 향후 미세먼지 수치를 예측할 수 있는 모형을 제시한다. 실제 황사, 연무 등의 대기의 예측정확도가 향상 될 경우, 정확도 1%당 한계지불의사액(Marginal Willingness to Pay)이 48원 증가한다(김혜민 et al, 2020). 즉, 개선된 미세먼지 예측 모형을 활용해 실제 미세먼지 예측 분야에 적용시킬 경우 보다 경제적인 미세먼지 정책을 수립할 수 있을 것으로 기대된다.

Ⅲ. 연구방법론

1950년대 인공신경망이라는 개념이 소개되면서 시작된 인공지능은 이후 2000년대 초까지 침체를 겪었으나, 최근 다양한 방법론과 컴퓨터 성능의 향상으로 다시 주목받고 있다. 기계학습은 인간이 과거의 경험과 다양한 실험을 기반으로 새로운 것을 학습하는 모형을 컴퓨터 알고리즘에 적용한 것이다. 특정 데이터를 기계학습 모델에 입력해주면 컴퓨터가 스스로 문제를 해결해가며, 기존에 찾아내기 어려웠던 해답을 제시한다. 이를 기계학습이라고 할 수 있으며, 기계학습을 통해 가치 있는 정보를 찾아낼 수 있다. 기계학습의 장점은 과거의 경험이나

연구자의 주관이 개입이 되는 것이 아니라 객관적인 데이터 자체를 설계된 알고리즘에 학습하고 구축하는 것이다(오지훈, 김정섭, 2017). 기계학습은 지도 학습과 비지도 학습으로 구분된다. 지도학습은 입력과 출력 값이 존재하는 데이터를 기반으로 데이터의 특성들과 관련된 레이블 사이의 관계를 학습시켜 새로운 모형을 만드는 것이다. 학습을 통해 만들어진 모형에 새로운 데이터를 적용하면 미래에 대한 예측 또는 새로운 데이터를 분류하는 것이 가능하다. 비지도 학습은 데이터의 특성을 알 수 없을 때 컴퓨터가 일정한 기준으로 알려지지 않은 데이터의 패턴이나 데이터 사이의 관계를 찾아내는 것이다.

본 연구에서는 지도학습 기법 중 XGBoost(Extreme Gradient Boosting; XGBoost), Random Forest, SVM, 인공신경망(Artificial Neural Network; ANN) 기법을 활용해 미세먼지 농도를 예측하고자 한다.

3.1 XGBoost

XGBoost 모델은 의사결정 나무 알고리즘의 한 종류이다. XGBoost는 의사결정 나무 방법의 장점인 데이터 전처리 과정의 단순함 가지고 있으면서, 다른 의사결정 나무 방법들에 비해 비교적 빠른 계산 속도를 보여준다(황혜진 et al, 2018). XGBoost는 여러 개의 회귀나무(Classfication and Regression Tress: CART)를 이용해서 오차 값을 줄여나가면서 최적의 트리를 찾아나간다. XGBoost는 훈련과정에서의 데이터 손실을 최소화하면서, 결과의 과적합을 피하기 위해 트리의 복잡도를 조절한다. 트리의 깊이가 0에서부터 시작해서, 가지치기 과정을 통해 최적의 모델을 찾는다. 가지치기 과정에서 정보획득을 순서대로 연산하고, 정보획득 점수가 음의 값을 가질 때까지 가지를 제거한다. XGBoost는 설정한 횟수만큼 무작위로 트리를 생성하고 계산을 반복한다. 최종적으로 점수가 높은 트리들을 조합하여 모델을 생성한다(Tianqi Chen, Carlos Guestrin, 2016). 최근 다양한 데이터분석 경진 대회에서 많이 활용되고 있으며, 속도 및 정확성 부분에서 인정받고 있다.

3.2 Random Forest

랜덤 포레스트(Random Forest)는 전통적인 의사결정나무 기법을 발전시킨

기법이다. 여러 개의 의사 결정 트리를 이용해서 최적의 모델을 찾아내는 기계학습 기법으로 분류 문제와 회귀 문제에 주로 사용된다. 학습 과정에서 생성된 여러 개의 트리로부터 평균 예측률을 출력한다. 입력 변수의 노이즈에 취약한 의사결정나무의 단점을 보완한 모형이라고 할 수 있다(문재욱; 2019). 랜덤 포레스트는 편향과 분산의 균형을 맞추는 것이 중요한 기법이다. 편향은 모델이 예측한 값과 실제 값의 차이를 나타낸다. 분산은 학습 데이터의 결과를 얼마만큼 일반화할 수 있는가를 나타내는 것이다. 편향과 분산은 서로 반비례 관계에 있는데, 예측률을 높이기 위해서 학습 데이터에 과적합할 경우 분산이 높아져 모델을 일반화하기 어렵고, 분산을 줄여 일반화 가능성을 높인다면 학습 오차율이 높아진다. 랜덤 포레스트는 전체 데이터를 학습에 사용 가능하고, 다른 분석 방법에 비해 전처리 과정에서의 복잡도가 낮은 장점이 있다. 뿐만 아니라 높은 정확도를 나타내기 때문에 다양한 분야에서 사용되고 있다(Leo Breiman, 2001).

3.3 SVM

SVM은 일반적으로 서로 다른 집단으로 분류할 수 있도록 하는 기계학습 기법이다. 훈련 데이터들을 학습시켜 최적의 초평면을 찾는다. 최적의 초평면에 의해 분리된 집단 사이의 마진을 최대화 시키는 것이 학습 목적이며, 이때 최적의 초평면에서 가장 가까운 데이터를 서포터 벡터라고 한다. 즉, 서로 다른 범주에 속한 관측치 사이에 간격이 최대가 되는 선을 찾는다. 즉, 집단으로 분류할 수 있는 기준을 바탕으로 새로운 데이터가 주어졌을 때, 어느 집단으로 분류하는지를 판단하는 알고리즘이라고 할 수 있다(정용진 et al, 2020).

SVM은 다양한 커널이 존재하는데 본 연구에서는 rbf 커널을 사용하였다. rbf 커널은 복잡한 비선형적인 문제를 해결하는데 적합한 커널이다. 이 경우 종속변수가 연속형 변수일경우도 활용할 수 있다. 일반적으로 예측 모델에서 활용되며, SVR로 불리게 된다. 분석방법은 SVM과 동일하다(H. Drucker et al, 1996). 앞서 소개된 XGBoost와 RandomForest와 함께 꾸준히 활용되고 있는 알고리즘이다. 하지만 세부 파라미터 정의가 복잡할 경우 연산에 많은 시간이 소요된다는 단점이 있다.

3.4 ANN

인공신경망은 신경 세포인 뉴런의 형태를 참고하여 제안된 신경망 모델이며, 일반적으로 신경망은 입력층(Input layer)과 은닉층(Hidden layer), 출력층(Output layer)으로 구성되어 있다. 입력된 정보의 출력이 다음 층에 입력되는 형태로 이루어지며, 이러한 결합을 통해 오른쪽에서 왼쪽으로 정보가 전달된다. 인공신경망은 학습 데이터와 학습 결과의 오차를 최소화 할 수 있는 방향으로 가중치를 업데이트하며 학습을 진행한다. 학습은 손실 값이 최소화되는 지점까지 이루어진다. 학습된 모델을 활용해 회귀 분석, 분류, 패턴 인식 등 다양한 부분에 활용할 수 있다(A.K. Jain et al, 1996). 최근에 인공신경망은 주가 예측 및 자연재난 예측 등과 같이 다양한 분야에서 활용되고 있다(이상원, 2002; 최민희 et al, 2019).

Ⅳ. 실험결과

4.1 데이터

본 연구에서는 대기 오염 데이터, 기상 데이터, 교통 데이터, 발전량 데이터를 활용하여 미세먼지 수치를 예측했다. 예측을 위한 미세먼지 데이터는 pm10에 대한 월별 평균 측정치이다. 환경부에서 제공하는 월별 도시별 대기오염도 데이터 중 전국 평균 데이터를 활용했다. 대기오염물질 데이터는 아황산가스, 오존, 이산화질소, 일산화탄소에 대한 월별 평균 측정치로 구성되어 있다. 2010년 1월부터 2019년 11월까지 데이터를 수집했으며, 환경부에서 제공하는 대기오염도 현황 데이터의 전국 평균치를 활용했다. 기상 데이터는 기상청에서 제공하는 기온, 강수량, 풍속에 대한 월별 평균 측정치로 구성되어 있다. 2010년 1월부터 2020년 6월까지의 데이터를 수집했으며, 전국 평균 데이터를 활용했다.

교통량 데이터는 항만의 화물입항현황에 대한 데이터를 활용했다. 해양수산부에서 제공하고 있으며, 화물, 유류, 화물환적에 대한 전국 항구의 물건 입하량을 나타내고 있다. 월별 전국 항구의 입하량 총계 데이터를 활용했다. 수집된 데이터는 2010년 6월부터 2020년 4월까지의 데이터로 구성되어 있다.

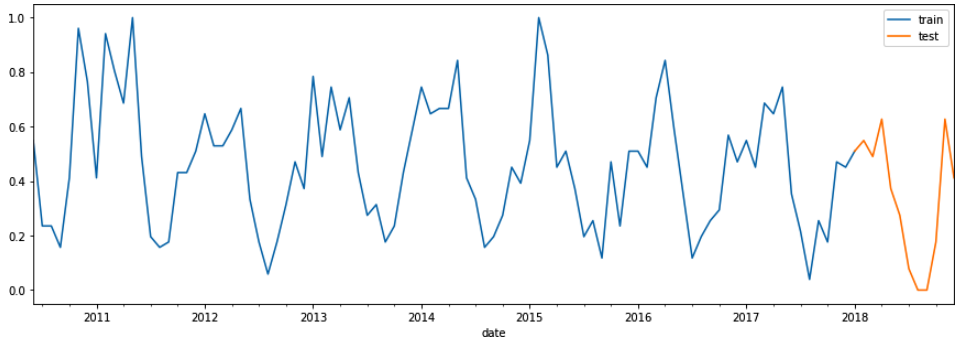
발전량 데이터는 한국전력거래소에서 제공한 데이터이며, 무연탄, 유연탄, 중

유, LNG 등의 화석연료를 사용해 발전한 발전량에 대한 데이터이다. 수집 데이터는 2010년 1월부터 2018년 12월까지로 구성되어 있다. 아래 〈표-2〉에서는 각 데이터에 대한 기술통계량이 제시되어 있다.

〈표-2〉 기술통계량

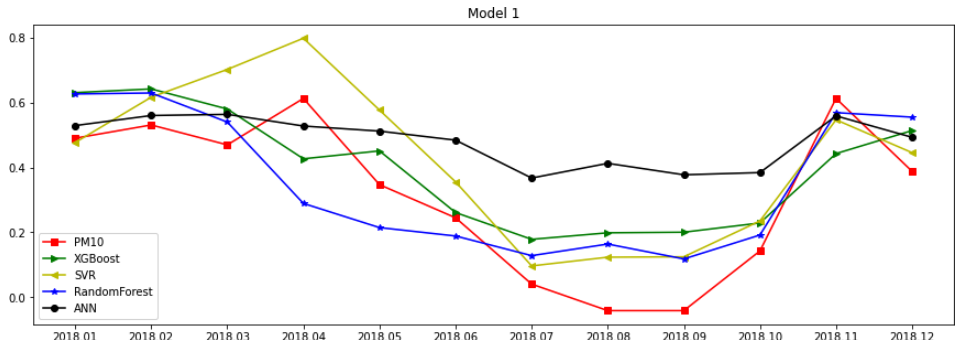
구분	PM10 (㎍/㎥)	SO2 (ppm)	O3 (ppm)	N2O4 (ppm)	CO (ppm)	기온 (°C)	강수 (mm)	풍속 (m/s)	입하량 (R/T)	발전량 (GWh)
Count	103	103	103	103	103	103	103	103	103	103
Mean	46.78	0.005	0.03	0.02	0.50	13.25	107.23	1.98	10369966	18264961
Std.	11.89	0.001	0.008	0.005	0.10	9.5	99	0.28	1222110	1721628
Min.	24	0.003	0.013	0.012	0.4	-4.8	5.6	1.3	8020408	15234639
25%	37	0.004	0.019	0.019	0.4	5.2	37	1.8	9695542	16912636
50%	47	0.005	0.026	0.024	0.5	13.9	67	1.9	10252227	17977931
75%	54	0.006	0.032	0.026	0.6	21.8	131	2.2	10973679	19530009
Max.	75	0.008	0.045	0.035	0.8	27.3	491	2.7	15872211	23083099

수집된 데이터의 기간이 상이해 최종 데이터 셋은 2010년 6월부터 2018년 12월까지로 구성했다. 각 변수의 값의 크기를 맞춰주기 위해 0~1 사이로 정규화를 실시한 뒤, 분석을 진행했다. 미세먼지 농도 예측을 위해 2010년 6월부터 2017년 12월까지의 데이터를 학습 데이터로 활용했으며, 2018년 1월부터 2018년 12월까지의 데이터를 검증 데이터로 활용했다. 〈그림 1〉에서는 분석에 활용한 미세먼지 수치를 각 시점에 따라 그래프로 나타내고 있다.



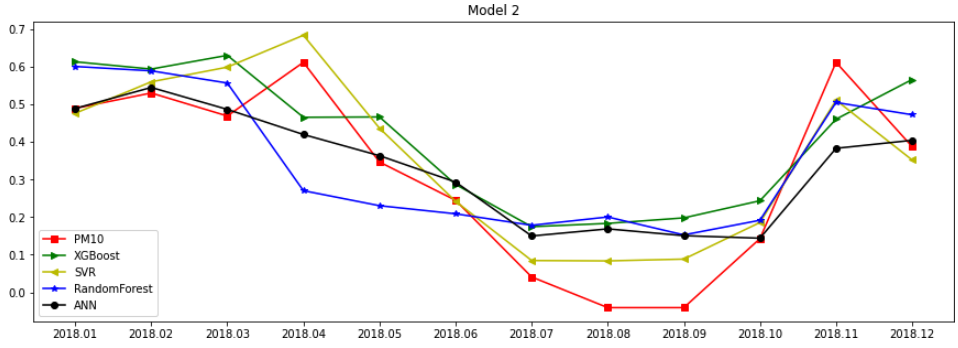
〈그림 1〉 연도별 미세먼지 수치

4.2 알고리즘 별 결과 비교



〈그림 2〉 기상 + 대기질 데이터를 활용한 미세먼지 예측

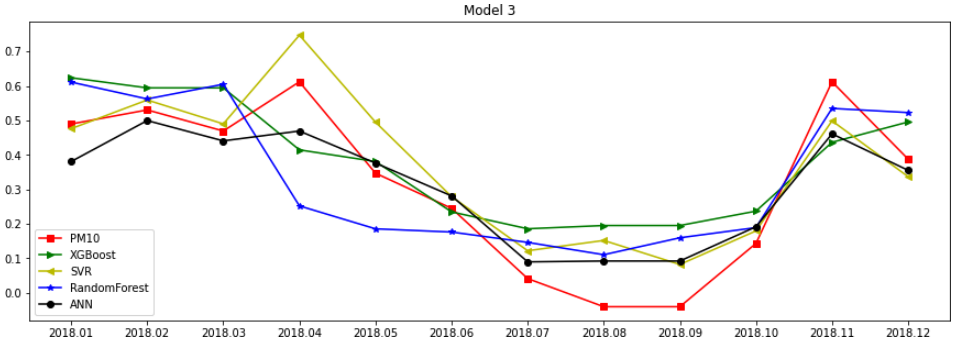
〈그림 2〉는 대기질 데이터와 기후 데이터를 활용해 미세먼지 농도를 예측한 결과를 보여준다. PM10은 실제 미세먼지 관측치를 나타내며, 다른 그래프는 각 모형의 예측 결과를 나타낸다. 분석 결과, 예측 정확도는 MAE 기준 SVR - RandomForest - XGBoost - 인공신경망 순으로 우수하게 나타났다. 성능에 대한 자세한 결과 값은 〈표-1〉에 명시되어 있다. 가장 높은 예측정확도를 보인 SVR의 경우 실제 미세먼지 수치와 다소 큰 차이가 있지만 전체적인 추세 흐름을 잘 찾아가는 것을 확인할 수 있다.



〈그림 3〉 교통 데이터를 추가로 활용한 미세먼지 예측

〈그림 3〉은 대기질 데이터와 기후 데이터, 교통량 데이터를 활용해 미세먼지 농도를 예측한 예측 그래프를 나타내고 있다. MAE를 기준으로 랜덤 포레스트를

제외한 모든 모델의 성능이 향상되었다. 특히, 인공신경망의 경우 그 성능 향상 폭이 큰 것을 확인할 수 있다. 예측정확도는 MAE 기준 SVR - 인공신경망 - XGBoost - RandomForest 순으로 우수하게 나타났다. 교통량 데이터가 포함된 모형의 경우 〈그림 2〉에 비해 전체적인 추세를 잘 따라가고 있음을 확인할 수 있다. 특히 SVR의 경우 〈그림 2〉에서 나타난 그래프에 비해 오차가 눈에 띄게 줄었음을 확인할 수 있다.



〈그림 4〉 발전량 데이터를 추가로 활용한 미세먼지 예측

〈그림 4〉는 기존 변수에 발전량 데이터를 추가적으로 활용한 모형에 대한 미세먼지 농도 예측 그래프이다. MAE를 기준으로 모든 모델의 성능이 향상되었다. 특히, SVR과 인공신경망의 경우 다른 두 모형에 비해 높은 예측 정확도를 보여주고 있다. 예측정확도는 MAE 기준 인공신경망 - SVR - XGBoost - RandomForest 순으로 우수하게 나타났다. 〈그림 4〉에서는 인공신경망의 적합도가 높음을 확인할 수 있다. 이전 〈그림 2〉과 〈그림 3〉에서는 다소 평평한 모습을 보였으나 〈그림 4〉에서는 전체적인 추세를 잘 따라감을 확인할 수 있다.

〈표-3〉은 각 측정지표별 미세먼지 예측에 대한 정확도를 나타내고 있다. 변수가 추가적으로 적용될 경우 대체적으로 예측의 오차율이 줄어듦을 확인할 수 있다.

기상 데이터와 대기질 데이터를 활용한 모형의 경우 SVR이 가장 좋은 결과를 보였다. 결과 값이 0~1사이로 정규화되었음을 감안했을 때, MAE 기준 약 0.1의 오차율을 보여주고 있다. XGBoost와 RandomForest는 각각 0.1391, 0.1269의 오차를 보였으며, 인공신경망은 0.1762로 오차가 가장 크다고 할 수 있다.

〈표-3〉 미세먼지 예측 모델의 정확도 측정 값

구분		XGBoost	RandomForest	SVR	Neural Network
기상 + 대기질	MAE	0.1391	0.1269	**0.1036**	0.1762
	RMSE	0.1591	0.1583	**0.1137**	0.2089
	MAPE	283.38	258.05	289.46	**234.35**
교통 데이터 추가	MAE	0.1375	0.1422	**0.0989**	0.1086
	RMSE	0.1517	0.1695	**0.1105**	0.1416
	MAPE	290.13	258.03	276.07	**74.09**
발전량 데이터 추가	MAE	0.1290	0.1305	0.0789	**0.0652**
	RMSE	0.1592	0.1489	**0.1138**	0.2046
	MAPE	290.13	258.03	276.07	**55.45**

교통 데이터가 추가된 모형의 경우에도 SVR의 예측력이 가장 좋게 나타났다. MAE 기준 약 0.05의 오차율 감소가 있었다. 인공신경망 모형의 경우 오차율이 약 0.1로 이전 예측과 비교해서 상당히 높은 개선률을 보였다. 이와 달리 XGBoost와 RandomForest의 경우 크게 개선이 되지 않았다.

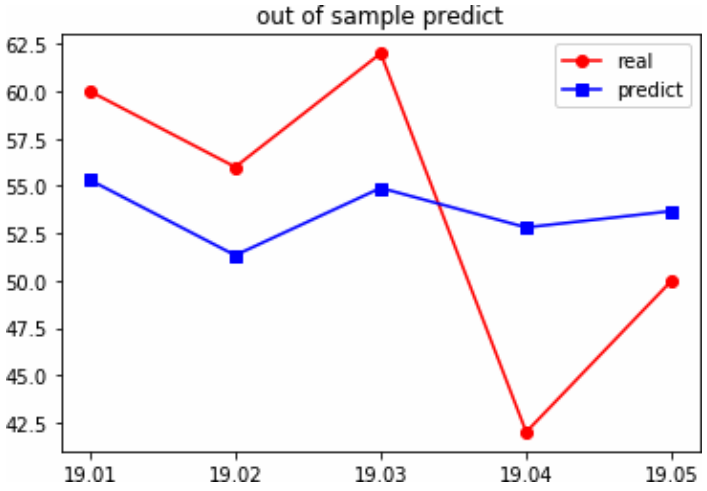
발전량 데이터가 추가된 모형의 경우 인공신경망의 예측력이 가장 좋게 나타났다. MAE 기준 약 0.065의 오차율을 보여주고 있다. 또한 SVR의 예측력도 크게 향상되었다. XGBoost와 RandomForest도 예측력이 어느정도 개선되었으나 그 효과는 미비하다.

모든 데이터를 복합적으로 사용한 경우, SVR과 인공신경망의 오차율이 낮게 나왔음을 확인할 수 있다. 두 모형의 오차율은 MAE 기준 0.0789와 0.0652로 실제 값에 근접함을 확인할 수 있다. 또한 인공신경망 모형의 경우 변수가 추가됨에 따라 오차율이 크게 감소하는 것을 확인할 수 있다. 이와 같은 추세를 확인했을 때, 인공신경망의 모델 개선의 여지가 크다고 볼 수 있다. 즉, 인공신경망 모형의 경우 미세먼지에 영향을 미칠 수 있는 유의미한 변수가 추가된다면 보다 정확한 예측을 할 수 있을 것으로 판단된다.

4.3 결과예측

다수의 변수를 활용해 미세먼지 수치를 예측한 결과 인공신경망 모형의 예측

에 대한 적합성이 우수한 것으로 나타났다. 따라서 본 연구에서는 추가적인 예측 결과를 제시하기 위해 인공신경망을 활용하여 미래의 미세먼지 결과 값을 도출하였다. 향후 5개월의 미세먼지 예측 값을 도출하였으며, 그 결과는 〈그림 5〉와 같다. 〈그림 5〉에서 나타난 인공신경망의 적합성의 경우 이전 결과에 비해 좋지 않은 결과가 나타났다. 특히, 2019년 4월 달의 평균 미세먼지 수치는 급격하게 감소하였지만 이를 인공신경망은 반영하지 못했다. 하지만 인공신경망을 사용한 예측 모형의 경우 전체적인 추세를 잘 반영하고 있다고 할 수 있다. 정확한 값을 예측하지는 못하였지만 등락의 추세는 어느정도 파악하고 있다고 볼 수 있다.



〈그림 5〉 인공신경망 모형을 활용한 미래 결과 예측

V. 결론

본 연구는 다양한 미세먼지 관련 변수를 활용해 미세먼지 농도 예측에 관한 연구를 수행했다. 실험결과 국내 관련 변수들이 추가됨에 따라 예측 모형의 정확도가 향상되는 것을 확인할 수 있었다. 한편, RandomForest와 XGBoost 기법을 통해 변수 중요도를 측정한 결과 국내 대기질이 예측에 가장 큰 영향을 주는 변수로 나타났다. 본 연구의 결과를 통해서 미세먼지를 예측함에 있어 국내 요인을 고려하는 것이 예측의 정확도를 점진적으로 증가할 수 있다는 점을 확인했고,

미세먼지의 저감을 위해서 국내 발생 요인에 대한 추가적인 조치가 필요하다는 사실도 확인할 수 있었다.

모델간 비교 결과에서는 인공신경망의 성능이 가장 우수한 것으로 파악되었다. 관련 선행연구에서는 RandomForest가 대체적으로 우수한 모형으로 선정되었지만, 본 연구에서는 동일한 연구결과가 나타나지 않았다. 비슷한 방법으로 예측을 진행하는 XGBoost 또한 낮은 정확도를 나타내었다. SVM가 기상 데이터와 대기질 데이터만을 활용해 미세먼지 수치를 예측할 경우, 가장 좋은 모델이라고 할 수 있다. 또한 교통량과 발전량 데이터가 추가되면서 SVR 모형의 예측력이 개선되었다. 변수가 추가되면서 꾸준히 예측력이 상승하였으며, 최종 모형에서도 좋은 결과를 보여주었다.

그러나 국내 관련 변수들이 추가되면서 예측력의 개선은 SVM 보다 인공신경망 모델에서 더 뚜렷하게 나타남을 확인할 수 있었다. 즉, 인공신경망을 활용한 예측 모형에서는 변수가 추가됨에 따라 모형의 정확도가 크게 개선되는 것을 확인할 수 있었다. 특히, 모든 변수가 고려되었을 때, 0.065의 오차율을 보이며 가장 우수한 모형으로 선정되었다. 따라서 본 연구에서는 미세먼지 예측 모델로 인공신경망이 가장 적합하였다. 인공신경망 기반 미세먼지 예측 모형의 경우 정확한 값을 나타내지는 못했지만, 추세를 벗어나지 않는 수준에서 결과를 제시한다. 향후 모형이 보완될 경우 추세 뿐만 아니라 정확한 결과 값을 예측할 수 있을 것이라고 판단된다. 본 연구를 통해 미세먼지 예측 모델의 정확도를 향상시킬 수 있는 방법론이 제시되고, 이를 통해 미세먼지로 인한 사회, 경제적인 피해를 줄일 수 있는 정책을 수립할 수 있을 것으로 기대된다. 특히 분석결과에 따라 미세먼지에 직접적인 영향을 미치는 대기오염물질(SO2, O3, N2O4, CO)을 줄일 수 있는 정책을 수립할 경우 미세먼지 수치에 직접적인 영향을 미칠 수 있을 것으로 기대된다.

본 연구에서는 미세먼지 예측을 위해 국내외의 다양한 변수들을 활용했다. 그럼에도 불구하고, 미세먼지에 영향을 줄 수 있는 다양한 요인들을 추가로 발굴하는 것이 중요하다. 한편, 국내 미세먼지의 가장 큰 요인은 중국에서 날아오는 대기오염물질이라고 할 수 있다. 그러나 본 연구에서는 국내에서 발생한 주요 환경오염 변수만을 고려하여 중국에서 미치는 영향력을 설명하지 못했다. 즉, 중국의 대기오염과 관련된 변수와 중국과 연관된 계절풍에 대한 영향력에 대한 설명이

부족하다. 향후 중국내 대기오염의 원인을 규명할 수 있는 다양한 변수들도 같이 고려한다면 예측의 정확도를 더 크게 향상시킬 수 있을 것으로 기대된다. 본 연구에서 확인한 바와 같이 인공신경망의 경우 변수가 추가됨에 따라 모형의 개선 정도가 크기 때문에, 유의미한 변수를 추가적으로 추출하여 미세먼지 수치 예측의 정확도를 개선할 수 있을 것이다. 뿐만 아니라 인공신경망 기법을 통해 미래의 미세먼지 수치도 예측하여 적절한 대비를 할 수 있을 것이다. 이를 통해 보다 경제적인 정책개발 및 투자가 가능할 것으로 기대된다.

[참고문헌]

강연욱 · 천성남 · 한경남 · 김태옥 (2019), “국내 화력발전소 배출 미세먼지 저감 및 모니터링 기술 분석,” 『대한전기학회 학술대회 논문집』, 1932-1933.

김다빈 · 문윤섭 (2018), “청주시 미세먼지 PM2.5 고농도 발생원인 분석 및 교육적 방안,” 『한국환경교육학회 학술대회 자료집』, 93-94.

김종수 · 이창훈 (2019), “머신러닝 기반의 미세먼지 장기 예측 모델 개발,” 『대한기계학회 춘추학술대회』, 424-425.

김형건 (2018), “미세먼지 원인 요소들의 영향력 변화 추정: 경유를 중심으로,” 『한국자료분석학회』, 20(2), 747-757.

김혜민 · 이승욱 · 김인겸 · 이대근 · 유승훈 (2020), “컨조인트 분석법을 이용한 기상관측장비 활용 및 예측모델 정확도 개선의 경제적 가치 추정,” 『한국혁신학회지』, 15(1), 301-320.

문재욱 · 정승원 · 김형준 · 황인준 (2019), “랜덤 포레스트를 활용한 지역별 하루 단위 감염병 발생 예측,” 『한국정보과학회』, 335-337.

박기형 · 장은화 · 정현철 · 유은철 · 조정구 (2019), “여름철 부산지역 초미세먼지(PM-2.5) 고농도 원인 분석,” 『한국대기환경학회 학술대회논문집』, 167-167.

박순애 · 신현재 (2017), “한국의 초미세먼지(PM2.5)의 영향요인 분석 : 풍향을 고려한 계절성 원인을 중심으로,” 『한국환경정책 · 평가연구원』, 25(1), 227-248.

박윤서 · 박지훈 · 강소영 · 이승하 · 손정석 · 유철 · 이상보 · 김정수 (2018), “미세먼지 오염 우심지역 고농도 원인분석 연구: 강원도에 대하여,” 『한국대기환경학회 학술대회논문집』, 146-146.

서양모 · 염재홍 (2019), “기상 데이터를 활용한 LSTM 기반 미세먼지 농도 예측 방법 비교,” 『한국측량학회 학술대회자료집』, 117-120.

오지훈 · 김정섭 (2017), “머신러닝을 적용한 주택가격 추정모형: MARS를 중심으로,” 『한국주택학회 하반기학술대회 자료집』, 153-171.

유태종 · 유동현 (2019), “화학 수송 모델과 기상 모델을 이용한 석탄 화력 발전소에서 배출되는 미세먼지의 거동 분석,” 『한국전산유체공학회지』, 24(4), 34-42.

육근형 (2017), “항만도시 미세먼지 배출현황과 대응 방향,” 『한국해양환경에너지학회 학술대회논문집』, 97.

이미숙・진형익 (2018), “환경적 요인이 삶의 만족도에 미치는 영향,” 『한국혁신학회지』, 13(4), 227-251.

이상원 (2002), “주가 예측을 위한 최적 인공신경망 모형 선택에 관한 연구,” 『인제대학교 교육대학원』.

이태형・전명진 (2018), “딥러닝 모형을 활용한 서울 주택가격지수 예측에 관한 연구 : 다변량 시계열 자료를 중심으로,” 『주택도시연구』, 8(2), 33-56.

전송완・최제열・배준현 (2017), “미세먼지 농도 예측 알고리즘 성능 비교,” 『한국정보과학회 학술발표논문집』, 775-777.

정용진・조경우・이종성・오창헌 (2019), “GRU를 이용한 미세먼지(PM10) 농도 예측 모델,” 『한국정보통신학회 종합학술대회 논문집』, 644-646.

조경우・정용진・이종성・오창헌 (2019), “LSTM을 이용한 PM10 미세먼지 농도 예측,” 『한국정보통신학회 종합학술대회 논문집』, 632-634.

조경우・정용진・이종성・오창헌 (2020), “SVM 알고리즘 기반의 PM10 이진 분류 모델,” 『한국정보통신학회』, 24(1), 308-310.

조성환・김현웅・한영지・김우진 (2016), “강원도 춘천과 영월에서 측정한 미세먼지 농도 특성 및 고농도 원인 분석,” 『한국대기환경학회지』, 100-113.

최민희・정남준・이규철・정재성・서인용 (2019), “자연재난에 의한 전력 설비 피해 예측을 위한 인공신경망(ANN) 알고리즘 개발,” 『전기학회논문지』, 68(9), 1085-1093.

황혜진・김수현・송규원 (2018), “XGBoost 모델 해석을 통한 노인의 인지능력 개선・악화 요인 탐구,” 『한국차세대컴퓨팅학회 논문지』, 14(3), 16-24.

A.K. Jain・Jianchang Mao・K.M. Mohiuddin (1996), “Artificial neural networks: a tutorial”, *Computer*, 29(3), 31-44.

Drucker, H., C.J. Burges, L. Kaufman, A. Smola and V. Vapnik (1997), “Supprot Vector Regression Machines,” *Neural Information Processing Systems 9*, 155-161.

Leo Breiman (2001), “Random Forests,” *Machine Learning*, 45(1), 5-32.

Tianqi Chen・Carlos Giestrin (2016), “XGBoost: A Scalable Tree Boosting System,” *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

A Comparative Study on the Performance of Machine Learning Models for the Prediction of Fine Dust: Focusing on Domestic and Overseas Factors

Sang Ha Sung* · Sangjin Kim** · Min Ho Ryu***

Abstract

Fine dust is a major cause of air pollution, causing various respiratory diseases and threatening health. This study analyzes various factors affecting fine dust levels in Korea, and proposes optimal algorithms in predicting fine dust levels. To do this, data from overseas inflow factors such as weather, air pollutants and internal factors of domestic traffic volume and thermal power generation etc are used. According to the analysis results, it was confirmed that the predicted accuracy of machine learning model increases as domestic related variables such as traffic volume and thermal power generation are added. In addition, the inflow of air pollutants from the overseas was evaluated as the most important variable for predicting fine dus. Among the various algorithms, artificial neural network techniques were found to have the highest predictive accuracy. Through this study, we will be able to grasp the influence of domestic/oversea-related variables needed to predict fine dust levels, and improve the accuracy of predictions through suitable models.

Key Words; fine dust, prediction, machine learning, foreign factors, domestic factors

* Ph.D Students, Department of Management Information Systems, Dong-A University, sangha@donga.ac.kr
** Professor, Department of Management Information Systems, Dong-A University, skim10@dau.ac.kr
*** Professor, Department of Management Information Systems, Dong-A University, Corresponding author, ryumh12@dau.ac.kr