Journal of the Korean Association of Geographic Information Studies, 21(4) 2018, pp. 64-80 ISSN 1226-9719(Print)
https://doi.org/10.11108/kagis.2018.21.4.064 ISSN 2287-6952(Online)

머신러닝기반 범죄발생 위험지역 예측*

허선영[1]·김주영[2※]·문태헌[3]

Predicting Crime Risky Area Using Machine Learning*

Sun-Young HEO[1]·Ju-Young KIM[2※]·Tae-Heon MOON[3]

요 약

우리나라의 시민들은 범죄에 대한 일반적인 사항만을 알 수 있을 뿐, 자신이 범죄위험에 얼마나 노출되어 있는지를 파악하기 어렵다. 경찰의 입장에서도 범죄발생 지역을 예측할 수 있다면 경찰력이 부족한 상황에서 효율성 있게 범죄에 대처 가능할 것이지만 아직 우리나라에서는 예측시스템이 없고, 관련 연구도 매우 부족한 실정이다. 이에 본 연구에서는 범죄발생 위험지역 예측 자동화 시스템 개발의 첫 번째 단계로 빅데이터로 구축 가능한 범죄정보와 도시지역 자료를 바탕으로 머신러닝 방식을 통해 한국형 범죄발생 위험지역 예측 모형을 개발하고자 한다. 또한 시나리오를 가정하여 범죄발생 확률을 지도로 시각화함으로써 사용자의 이해도를 높이도록 하였다. 선행 연구 및 사례에서 범죄발생에 영향을 미치는 요인 중 빅데이터로 구축 가능한 범죄정보, 날씨정보(기온, 강수량, 풍속, 습도, 일조, 일사, 적설, 전운량), 지역정보(평균 건폐율, 평균 용적율, 평균 높이, 총 건축물수, 평균 공시지가, 평균 주거용도면적, 평균 지상층수)를 머신러닝에 활용할 수 있도록 데이터를 사전 처리하였다. 머신러닝 알고리즘으로서 지도학습 모형 중 다양한 분야에서 활용되며 정확도가 높다고 알려진 의사결정나무모형, 랜덤포레스트모형, Support Vector Machine(SVM)모형을 활용하여 범죄 예측 모형을 구축하고 비교·분석하였다. 그 결과 평균 제곱근 오차(Root Mean Square Error, RMSE)가 낮아 예측력이 높은 의사결정나무모형을 최적모형으로 선정하였다. 이를 바탕으로 가장 빈번하게 발생하는 절도와 폭력범죄를 대상으로 시나리오를 작성하여 범죄 발생 위험지역을 예측한 결과, 사례도시 J시는 위험지역이 3가지 패턴으로 발생하는 것으로 나타났으며, 각각 발생확률을 3 등급으로 구분하여 250 x 250m 단위의 지도형태로 시각화할 수 있었다. 본 연구는 향후 자동화 시스템으로 개발하여 시시각각으로 변하는 도시 상황에 따라 실시간으로 예측 결과를 시각화하여 제공함으로써 보다 범죄로부터 안전한 도시환경 조성에 기여하고자 한다.

주요어 : 범죄예측, 머신러닝, 의사결정나무, 랜덤포레스트, 서포트벡터머신

2018년 10월 30일 접수 Received on October 30, 2018 / 2018년 11월 23일 수정 Revised on November 23, 2018 / 2018년 11월 26일 심사완료 Accepted on November 26, 2018

* 이 논문은 한국연구재단의 기초연구사업(2017R1A2B4012254)연구지원비에 의하여 수행되었음

1 경상대학교 공학연구원, Engineering Research Institute(ERI), Gyeongsang National University
2 (주)동명기술공단종합건축사사무소 도시사업본부 도시계획부, Dong Myeong Engineering Consultants & Architecture, Urban Development, Urban Planning
3 경상대학교 도시공학과, Dept. of Urban Engineering, Gyeongsang National University
※ Corresponding Author E-mail : jo0066@nate.com

ABSTRACT

In Korea, citizens can only know general information about crime. Thus it is difficult to know how much they are exposed to crime. If the police can predict the crime risky area, it will be possible to cope with the crime efficiently even though insufficient police and enforcement resources. However, there is no prediction system in Korea and the related researches are very much poor. From these backgrounds, the final goal of this study is to develop an automated crime prediction system. However, for the first step, we build a big data set which consists of local real crime information and urban physical or non-physical data. Then, we developed a crime prediction model through machine learning method. Finally, we assumed several possible scenarios and calculated the probability of crime and visualized the results in a map so as to increase the people's understanding. Among the factors affecting the crime occurrence revealed in previous and case studies, data was processed in the form of a big data for machine learning: real crime information, weather information (temperature, rainfall, wind speed, humidity, sunshine, insolation, snowfall, cloud cover) and local information (average building coverage, average floor area ratio, average building height, number of buildings, average appraised land value, average area of residential building, average number of ground floor). Among the supervised machine learning algorithms, the decision tree model, the random forest model, and the SVM model, which are known to be powerful and accurate in various fields were utilized to construct crime prevention model. As a result, decision tree model with the lowest RMSE was selected as an optimal prediction model. Based on this model, several scenarios were set for theft and violence cases which are the most frequent in the case city J, and the probability of crime was estimated by 250x250m grid. As a result, we could find that the high crime risky area is occurring in three patterns in case city J. The probability of crime was divided into three classes and visualized in map by 250 x 250m grid. Finally, we could develop a crime prediction model using machine learning algorithm and visualized the crime risky areas in a map which can recalculate the model and visualize the result simultaneously as time and urban conditions change.

KEYWORDS : Crime Prediction, Machine Learning, Decision Tree, Random Forest, SVM

서 론

알파고를 계기로 인공지능(Artificial Intelligence, AI)에 대한 관심이 증가하고 있고, 핵심 기술인 머신러닝(machine learning)과 딥러닝(deep learning)에 대한 연구도 확대되고 있다. 머신러닝은 여러 분야에 매우 광범위하게 전개되고 있으며, 광고, 번역, 스팸차단, 게임, 음성 및 문자인식, 텍스트마이닝과 같은 검색엔진은 물론, 지능형 로봇, 자율주행 자동차와 같은 산업분야에서 이미 상용화되고 있다. 우리 일상생활에도 변화를 가져오고 있는데, 예를 들면 음악감상이나 라디오 청취를 위해 사용하던 스피커가 소리를 전달하는 단순한 도구에서 인공지능, 음성인식 기술 등을 활용한 AI 스피커로 발전하여 대중화되고 있다. 이뿐 아니라 글로벌 기업인 아마존은 클라우드 기반 머신러닝 기술을 이용해 야구 경기 심층 분석, 예측 등을 제공하기로 메이저리그(MLB)와 계약했고, 한국의 한 카드회사는

올해 하반기 중 머신러닝 기반 사고예측 모형을 통한 이상금융거래탐지시스템(Fraud Detection System, FDS)을 고도화할 예정이다.

도시안전 분야에서도 혁신적인 변화가 일어나고 있는데, 정보수집 및 분석기술의 향상과 더불어 미국을 중심으로 '예측적 경찰활동(predictive policing)' 시스템 개발을 서두르고 있다. 정부차원의 지원과 학계의 지속적인 연구로 범죄 예측 시스템을 개발하고 있다. 실예로 해외의 경우 민간에서 개발한 프레드폴(PredPol), RTM(Risk Terrain Modeling), HunchLab 등이 있으며, 이들은 머신러닝 기술을 이용해 예측모델을 개발하고 시스템화 하여 그 정보를 경찰과 시민에게 제공하고 있다.

이와 유사하게 우리나라에는 온라인으로 접속 가능한 행정안전부 생활안전지도 서비스가 있다. 그러나 해외사례의 경우 범죄 종류, 위치, 시간 등의 정보를 가진 빅데이터로 머신러닝 기법을 이용하여 범죄발생 예측모델을 개발하고, 블록단위의 마이크로 레벨에서 지리적으로 그 결과를 시각화하고 있다. 최근에는 ICT 기술의 발전과 더불어 빅데이터 저장 및 분석이 가능한 환경이 조성되면서 다양한 범죄예측기법들이 학계와 실무에서 시도되고 있다. 특히 인터넷과 SNS등과 같은 스마트 기기 사용이 일상화됨에 따라 실시간으로 생산되는 방대한 양의 빅데이터를 처리하여 의미 있는 정보를 찾아내어 활용하고 있다(Heo *et al.*, 2017). 하지만 한국의 경우는 개인정보 유출 등을 이유로 행정구역과 같은 매크로 레벨로 범죄 발생 건수와 같은 단순 집계자료를 제공하는 수준에 머물고 있다. 따라서 우리나라는 아직 초보 단계로 선진사례와 격차가 매우 크다고 할 것이다. 4차 산업혁명시대에 대응하고, 국민의 생명과 재산을 보호하기 위해 범죄예측 분야에 조속한 연구개발이 필요한 시점이다.

범죄발생이 우려되는 지역을 예측해서 시민들에게 제공한다면 시민들은 자신이 범죄위험에 얼마나 노출되어 있는지를 파악하여 미리 조심할 수 있는 장점이 있다. 경찰관계자들에게는 시시각각으로 변하는 위험지역 정보를 시간대별로 예측하여 제공함으로써 한정된 경찰력을 효과적으로 배치하여 범죄를 사전에 예방하고 범인 검거에 도움을 줄 수 있어 안전한 도시환경을 조성하는 데 도움이 될 것이다.

이에 본 연구는 범죄예측부터 결과의 시각화까지 자동화된 시스템을 개발하는 전 단계로 우선 빅데이터로 구축 가능한 도시 및 범죄 관련 정보를 구축하고, 머신러닝 기반 범죄발생 위험지역 예측 모형을 탐색 한 후, 시나리오 별로 범죄를 예측하여 그 결과를 지도로 시각화하는 과정을 제시하고 구현하는 것을 목적으로 한다.

이론 및 선행연구 검토

위키피디아에 따르면 머신러닝은 인공지능의 한 분야로, 컴퓨터가 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야를 말한다. 머신러닝의 학습방법은 지도학습(supervised learning), 비지도학습(unsupervised learning), 강화학습(reinforcement learning)으로 나뉜다. 지도학습은 입력된 자료 A에 대해 A라는 답(label)을 주고, A가 A임을 알 수 있도록 스스로 학습하는 방식이다. 비지도 학습은 A와 B의 두 가지 입력된 자료가 있으나 각각 A인지 B인지 모르는 상태에서 둘의 차이를 스스로 학습하여 서로 다른 집단으로 분류한다. 강화학습은 주어진 문제의 답이 명확하지는 않지만, 결과에 따라서 보상(reward)과 손실(penalty)을 주어 보상을 최대화하는 방향으로 모형을 학습하는 것이다.

본 연구에서는 주어진 범죄자료와 도시정보를 바탕으로 학습하여 범죄를 예측하는 연구이므로 연구의 목적에 따라 지도학습으로 모형개발을 진행하도록 하며, 이와 관련한 선행연구를 검토하였다. 지도학습에 속하는 머신러닝 알고리즘으로는 나이브 베이즈(Naive Bayesian)모형, 로지스틱 회귀모형, 의사결정나무(Decision Tree)모형, 랜덤포레스트(Random Forest)모형, 신경망모형, 서포트벡터머신(Support Vector Machine, SVM)모형 등이 있지만 본 연구에서는 예측 및 분류연구에 주로 사용되며 그 정확도가 다른 모형들에 비하여 높다고 선행연구에서 알려진 의

사결정나무, 랜덤포레스트, SVM 모형을 중심으로 연구대상지에 적용하여, 범죄발생 위험지역을 가장 정확하게 예측하는 모형을 탐색해 보고자 한다.

의사결정나무는 국내외에서도 급경사지의 재해 예측(Song *et al.*, 2009), 건설재해 사전 예측(Cho *et al.*, 2017) 등의 재난 · 재해 분야, 시청률 예측(Park and Kim, 2003), 대도시 주민의 우울증에 영향을 주는 요인 예측(Kim and Kim, 2013), 범죄 예측(Almanie *et al.*, 2015, Neuilly *et al.*, 2011), 질병 예측(Chaurasia *et al.*, 2013) 등 다양한 분야에 활용되고 있으며, 대부분의 연구에서 높은 정확도가 있는 것으로 보고되고 있다.

랜덤포레스트도 높은 예측력을 보이며, 무엇보다 결과를 해석하기 용이한 장점이 있다. 국내외에서 부동산 가격지수 예측(Bae and Yu, 2018), 홍수 및 산사태 취약성 분석(Lee, 2017), 화재 발생 예측(Guo *et al.*, 2016, Oliveira *et al.*, 2012), 기업신용등급 예측(Kim and Ahn, 2016, Brown *et al.*, 2012, Hajek *el al.*, 2013), 공동주택 공용관리비 추산(Jeong *et al.*, 2017)과 같은 경제와 재난분야 외에도 서로 다른 차량 환경에서의 시선 인식 시스템 개발(Oh *et al.*, 2015) 등 데이터 마이닝과 연계된 분야에서 활발하게 연구되고 있다.

SVM 모형은 고지혈증 유병 예측(Lee and Shin, 2018), 부동산 가격지수 예측(Bae and Yu, 2018), 기업 부도 예측(Choi and Lim, 2013), 기업신용등급 예측(Kim and Ahn, 2016, Ahn, 2014), 재무정보를 이용한 주가 예측(Heo and Yang, 2015) 등의 연구에서 활용되고 있으며, 이 모형도 비교적 높은 정확도를 보이는 것으로 보고되고 있다.

관련 선행연구 중 모형의 정밀도와 예측력에 대한 사례와 방법을 조금 더 구체적으로 살펴보면, Song *et al.*(2009)는 의사결정나무모형을 이용하여 급경사지 재해 정밀예측모형을 개발하였으며, 기존 모형과 비교한 결과, 결정질암지역에서 기존 예측모형보다 정확도가 높다고 하였다. Chaurasia *et al.*(2013)은 심장질환환자의 조기 예측을 위한 최상의 분류기를 찾기 위하여 의사결정나무, 규칙기반분류기로부터 파생된 CART(Classification and Regression Tree), ID3(Iterative Dichotomized 3) and decision table(DT) 모형을 이용하여 예측모형을 개발하고 비교 · 분석하였다. 그 결과 세 가지 모형 중 CART에 의한 예측 분류모형이 기존의 분류방법을 크게 향상시킬 수 있다고 하였다.

Cho *et al.*(2017)는 건설재해사전예측 모형 개발을 위하여 모형의 해석이 쉽고 변수의 상호작용효과 해석이 용이한 의사결정나무 기법을 사용하였으며, 의사결정나무 모형의 현장 활용성 검토를 위하여 판별분석 모형과 비교 · 분석하였다. 분석 결과 의사결정나무 모형의 예측 정확도가 80%, 판별분석 모형의 예측 정확도가 63%로 나타나 의사결정나무 모형의 예측 정확도가 더 높다고 하였다.

Kim and Kim(2013)은 대도시 주민의 우울감에 영향을 주는 요인을 예측하기 위하여 로지스틱 회귀모형과 의사결정나무모형을 활용하여 예측모형을 제시하였다. 그 결과 민감도와 분류정확도의 경우에는 의사결정나무모형보다 로지스틱 회귀모형이 높지만, 의사결정나무모형은 분석의 정확도보다는 분석과정 중 특정 경로의 설명이 필요한 경우에 로지스틱 회귀모형보다 유용하게 활용할 수 있는 것으로 해석하였다.

Yoo(2015)는 의사결정나무모형을 기저로 하며 무작위성을 최대로 부여함으로써 예측오차를 줄이는 방법인 랜덤포레스트와 의사결정나무모형을 비교 · 분석하였다. 그 결과 랜덤포레스트가 가지치기 이후의 의사결정나무와 비슷한 예측력과 높은 안정성을 보이는 것으로 나타났다. 즉, 랜덤포레스트는 설명변수가 많고 설명변수간 상호작용이 복잡한 자료에도 적용 가능하며, 예측력과 안정성이 높음을 실증하였다.

Yoo *et al.*(2018)는 대형마트와 전통시장을 대상으로 소비자의 소매유형 선택 및 교차쇼핑 정도를 예측하는 모형 개발을 시도하였다. 이를 위하여 로지스틱 회귀모형, 의사결정나무모형, 랜덤포레스트, 부스팅, 신경망모형을 비교 · 분석

하였으며, 예측력을 비교한 결과 랜덤포레스트와 부스팅 방식이 다른 모형들에 비해 높다고 하였다.

Guo *et al.*(2016)은 인위적인 산불 발생을 예측하기 위하여 로지스틱 회귀모형과 랜덤포레스트 모형을 비교·분석하였다. 그 결과 연구대상지의 인위적 산불 발생 예측 시 로지스틱 회귀모형보다 랜덤포레스트 모형의 예측 정확도가 높다고 하였다.

Jeong *et al.*(2017)은 데이터마이닝 기법인 랜덤포레스트 모형을 활용하여 공동주택 공용관리비(일반관리비, 청소비, 경비비, 소독비, 승강기 유지비) 추산모형을 개발하였으며, 구축한 모형의 유효성을 확인하기 위하여 사례검증을 실시하고, 다중회귀분석 추산모형과 비교한 결과 경비비를 제외한 모든 공용관리비에서 랜덤포레스트 방식이 다중회귀분석보다 우수하다고 하였다.

Oh *et al.*(2015)는 랜덤포레스트 모형을 이용하여 스마트 폰 카메라를 통해 입력된 얼굴의 구성요소에서 특징 벡터를 추출하고, 미리 만들어진 랜덤포레스트 학습모형과 비교하여 시선 인식 실험을 수행하였다. 그 결과 84.77%의 높은 시선 인식 성능을 확인하였다.

Kim and Ahn(2016)는 1,295개의 국내 상장 기업을 대상으로 기업신용등급 평가를 위해 랜덤포레스트 모형을 사용하였다. 이를 다중판별분석, 인공신경망, 다분류 SVM 모형을 비교·평가하여 랜덤포레스트 모형이 가장 우수하고, 다음으로 다분류 SVM의 정확도가 높은 것으로 분석하였다.

Heo and Yang(2015)는 재무정보를 기반으로 주식의 상승과 하락 등의 변화를 예측하기 위하여 SVM모형을 사용하였으며, 전문가 점수 및 인공신경망, 의사결정나무, 적응형부스팅 모형의 예측 정확도와 비교·분석하여 SVM이 가장 우수하다는 결과를 보고하였다.

Lee and Shin(2018)는 고지혈증을 예측하는 분류모형 개발을 위해 SVM과 meta-learning 알고리즘을 이용하여 성과를 비교하였다. 여기서는 분석을 위하여 투입되는 변수를 다르게 하여 그 결과를 비교하였는데, 투입되는 변수에 따라 SVM과 meta-learning의 정확도에서 차이가 있어 투입되는 변수에 따라 최적화 모형이 다를 수 있다고 하였다.

이상과 같이 머신러닝 알고리즘의 종류로서 의사결정나무, 랜덤포레스트, SVM은 다양한 분야에서 활용되고 있으며, 대개 정확도가 비 머신러닝 모형들보다 높은 것으로 나타났다. 그러나 투입되는 데이터의 종류나 분야 등에 따라 최적의 모형이 모두 다른 점을 주목할 필요가 있다. 즉 주어진 문제를 풀기 위해 주어진 상황과 데이터에 따라 최적 모형을 일일이 탐색할 수밖에 없다는 결론에 도달하게 된다. 따라서 본 연구에서 개발되는 범죄발생 예측모형도 전국 어디서나 적용 가능하다기 보다 연구 대상지에 특화된 모형이 될 것이다.

또한 해외의 다양한 시스템 사례연구를 토대로 한국의 실정에 맞는 연구의 방법 및 데이터를 구축하고자 하였다. 먼저 해외 범죄 관련 정보를 제공하는 시스템 및 웹사이트는 범죄 발생 현황을 제공하는 것과 범죄 발생 위험지역을 예측하여 제공하는 것으로 나뉜다. 범죄 발생 현황을 제공하는 서비스는 대표적으로 뉴욕시의 NYC crime map, 북미의 Crimereports, 영국의 Metropolitan Police 등이 있으며, 각 범죄 종류에 따른 통계 데이터 및 범죄 발생 위치, 발생시간 등의 정보를 제공하는 등 다양한 범죄 발생 정보를 제공하고 있다. 예측시스템은 프레드폴(PredPol), RTM(Risk Terrain Modeling), PCA(Predictive Crime Analytics platform), HunchLab 등이 있으며 이와 같은 대표적인 범죄 예측 시스템은 범죄 발생 정보를 시민에게 제공하는 NYC crime map, Crimereports 등과는 달리 범죄 발생 위험지역을 예측하고 시각화하여 제공하기 때문에 일반 시민들에게 공개되지 않고 경찰관, 공공기관 등에서 시스템을 구매하고 데이터를 구축한 관계자를 포함한 전문가에게만 예측 정보를 제공하고 있다. 범죄 발생 가능성이 높은 지역을 예측하고 순찰을 강화하는 등의 사전 대응으로 범죄율을 줄일 수 있는 Predictive Policing System은 지속적으로 제작 및 업데이트 되고 있다. 대부분의 예측

시스템은 민간에서 제작하여 경찰서, 시 등 공공이 이용하고 있으며 실제로 범죄율을 줄이는 등의 효과를 보여주고 있다. 각 예측 시스템은 범죄데이터만을 사용하거나, 범죄데이터가 아닌 범죄 위험 요소를 사용하는 등 사용 데이터와 분석방법, 표현방법 등에 있어 차이점을 가지고 있다. 선행 연구 및 사례연구 검토 결과 한국에서는 통합적 방범정보를 제공하고 있는 사례는 극히 드문 것으로 파악되며, 범죄발생 위험지역 예측 시스템은 시민의 입장에서는 범죄에 대한 두려움 감소효과를 누릴 수 있고 공공은 범죄발생으로 인한 사회적 비용발생을 절감할 수 있다는 점에서 시도해볼 충분한 가치가 있는 연구로 판단된다.

연구방법 및 데이터 구축

1. 연구방법

범죄발생 위험지역 예측의 정확도뿐만 아니라 범죄발생의 지역적 패턴을 분석하기 위해서 개발되는 모형은 해석하기 쉬워야 한다. 본 연구에서는 지도학습 모형 중 활용도와 정확도가 높은 것으로 알려진 의사결정나무, 랜덤포레스트, SVM을 적용하여 비교 · 분석하였다. 분석도구로는 MATLAB 패키지를 활용하였다.

의사결정나무 방식은 대량의 데이터 집합에서 유용한 정보를 추출하는 데이터마이닝 기법 중 하나로서 의사결정규칙을 나무구조로 구조화하여 대상이 되는 집단을 분류(classification) 또는 예측(prediction)하는 방법이다. 분석과정이 나무구조에 의하여 도표화되기 때문에 신경망, 회귀분석 등에 비해 분석과정의 이해 및 해석이 용이하다는 장점이 있다(Choi and Seo, 1999, Song *et al.*, 2009).

의사결정나무는 종속변수를 가장 잘 설명하는 독립변수 쪽으로 가지가 뻗어 나가도록 한다. 의사결정나무에서 값의 결정은 마디(node)로 표현되며, 각 노드를 찾는 방법에 따라 모형이 결정된다. 여기서 맨 위의 마디(노드 1)를 뿌리마디라고 하며 이는 모든 데이터를 포함한다. 노드 2, 3과 같이 하나의 마디에서 분기되는 마디를 자식마디라고 하며, 노드 1, 2, 3, 6과 같이 하위에 자식마디를 두는 마디를 부모마디, 노드 4, 5, 7, 8, 9와 같이 더 이상 분기되지 않는 마디를 최종마디라고 한다(그림 1 참조).

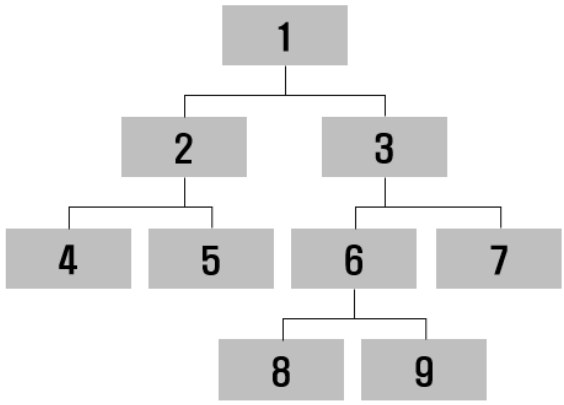


FIGURE 1. Decision tree

랜덤포레스트는 최근에 주목받는 데이터 마이닝 기법으로, 데이터를 학습하여 다수의 의사결정나무 모형들을 만든 후 그것을 결합하여 더 정확한 예측을 하는 것을 목적으로 한다(Yoo, 2015, Song and Song, 2018). 랜덤포레스트 모형은 훈련자료에서 붓스트랩(bootstrap) 표본을 다수 생성하여 입력변수들 중 일부만 랜덤으로 뽑아 의사결정나무를 생성하고, 이를 선형결합하여 최종 학습모형을 만들며, 분류 정확도가 우수하고 매우 안정적인 모형을 제공한다는 장점이 있다(Song and Song, 2018).

랜덤포레스트는 입력된 데이터 중에서 데이터를 임의의 수만큼 추출하여 의사결정나무 형태의 예측기를 만든 후, 생성된 모형들을 앙상블(ensemble) 기법을 통해 결합하여 최종모형을 만들기 때문에 예측기들은 초기 데이터셋에서 랜덤하게 선택된 데이터로부터 독립적으로 선택된다. 변수 선택의 임의성이 있기 때문에 예측기 변수는 N개의 총 변수 중 $\log_2(N+1)$개만큼 선택하게 되며, 랜덤포레스트 알고리즘의 정확도를 나타내는 평균 제곱오차는 식 1과 같다

(Lee, 2017).

$$\epsilon = (\nu_{observed} - \nu_{response})^2 \quad (1)$$

ϵ : 알고리즘의 평균 제곱 오차

$\nu_{observed}$: 알고리즘에 사용된 학습데이터 변수 값

$\nu_{response}$: 예측결과에서 나타난 변수 값

또한 각 트리에서 예측되는 값 반응치의 평균은 다음 식 2와 같다.

$$S = \frac{1}{K}\sum K^{th}\nu_{response} \quad (2)$$

여기서 S는 랜덤포레스트 알고리즘의 예측값이며, K는 랜덤포레스트의 각 트리에 적용된다. 이후 기존 데이터와 변수에 따라 미리 결정한 의사결정나무의 수만큼 각 노드에서 정보를 최대로 획득할 수 있도록 학습한다(Lee, 2017).

SVM은 Cortes and Vapnik(1995)이 제시한 지도학습 머신러닝 방법으로 분류 또는 회귀 모두에 이용 가능하며, 두 집단 경계에 존재하는 데이터 사이의 거리 차(margin)를 최대가 되도록 하는 최적의 초평면(hyperplane)을 찾는 것을 목표로 하는 방법이다. SVM 모형에서 두 집단을 분류하기 위한 분류식은 식 3과 같다.

$$f(x) = w \cdot x + w_0 \quad (3)$$

w : 추정모수

x : 입력값

$\cdot$: 벡터기호$(w_1x_1 + w_2x_2 + ... + w_nx_n)$

w_0 : 편의(bias)

$f(x)$: 분류함수

SVM은 확률 추정을 하지 않고 직접 분류 결과에 대한 예측만 수행하므로, 빅데이터에서 분류 효율 자체만을 보면 확률추정 방법보다 예측력이 전반적으로 높다는 장점이 있다(Lee and Shin, 2018, Song and Song, 2018).

2. 데이터 구축

범죄는 그 지역의 사회경제적, 인구학적, 물리적, 환경적 특성들에 영향을 받기 때문에 무작위로 발생하는 것이 아니라 상황에 따라 집중과 반복하여 발생하는 패턴이 있다(Newburn *et al*, 2004). 본 연구에서는 실험적으로 우리나라에서 사례지역의 환경적 특성이 반영된 빅데이터를 구축하고, 머신러닝 기반으로 범죄발생 위험지역 예측 모형을 탐색하고자 한다.

예측 모형 탐색을 위해 범죄발생과 연관이 많은 속성들을 선별하여 머신러닝에 투입해야 한다. 그러나 범죄자료가 공개되지 않아 제한적인 우리나라의 여건상 유사하면서 구축 가능한 자료에 의존할 수밖에 없다. 이러한 한계를 감안하여 본 연구에서는 그나마 사례지역의 2년에 걸친 범죄발생 정보를 입수하였으며, 이를 기반으로 하고, 사례연구를 통해 알려진 범죄와 관련성이 높으며 머신러닝으로 분석 가능한 빅데이터 형태로 구축할 수 있는 데이터 세트를 확보하였다. 여기에는 날씨정보(기온, 강수량, 풍속, 습도, 일조, 일사, 적설, 전운량), 지역정보(평균 건폐율, 평균 용적률, 평균 높이, 총 건축물수, 평균 공시지가, 평균 주거용도면적, 평균 지상층수)를 사용하였다. 데이터의 경우 위치정보를 포함하고 있으며, 공공기관에서 지속적으로 제공하는 대규모 데이터를 활용하였다. 즉, 기존의 범죄 분석이 발생한 범죄데이터를 누적시켜 범죄 발생예측 지역을 도출했다면 분 분석에서는 기존의 분석에서 제외 되었던 범죄의 세부 속성정보와 범죄 발생과 연관성이 높은 대규모 빅데이터를 활용한 연계분석을 통해 범죄발생지역을 예측하는 차이가 있다.

범죄자료는 2008년, 2011년 4,906건의 5대 범죄(절도, 폭력, 강도, 강간, 살인)데이터를 사용하였다.

한편 날씨정보가 투입된 배경은 건폐율, 용적률 등과 같이 변화가 적은 지역정보만 사용할 경우, 범죄위험이 높다고 예측된 지역은 시간적 변화와 상관없이 1년 내내 범죄위험이 높게 예측되어 실용적이지 않았기 때문이다. 따라서 시

간적으로 변화하는 동적인 예측시스템이 요구되는데 이를 위해 시계열 변수의 도입이 필요하게 된다. 이에 사례지역에 대하여 빅데이터로 구축할 수 있는 정보로서 시시각각 변화하는 날씨, 인구, 건물, 토지 등을 우선 시범적으로 투입하여 분석해 보기로 하였다.

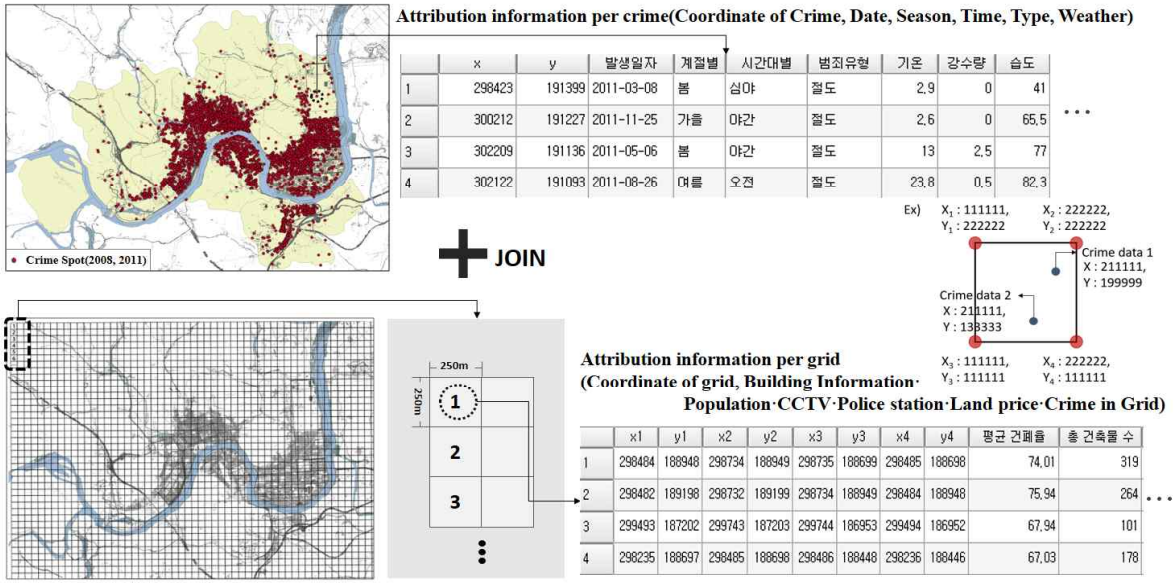
특히 날씨정보를 도입한 이유는 Ranson(2014), Horrocks and Menclova(2011), Elleng and Cohn 1990) 등 다수의 연구에서 날씨와 범죄가 관련성이 높다는 연구를 반영하였기 때문이다. 기상청에서 제공하는 자료의 갱신 주기는 1일, 인구정보는 수시(년 2회 이상), 건물정보는 월 1회, 토지정보는 연 2회(6개월 단위)로 갱신되는 정보를 활용하였다. 좋은 예측모형은 좋은 데이터를 투입하지 않으면 불가능하며, 특히 머신러닝의 경우 질 좋은 데이터의 확보가 매우 중요하다. 또한 범죄발생에 실질적으로 영향을 주는 변수의 시계열 자료를 지속적으로 업데이트하고, 주기적으로 기계학습 시켜 새로운 모형을 지속적으로 업그레이드해야 한다.

여기서 본 연구는 확보한 데이터를 기반으로 범죄발생 위험지역을 예측하고 그 장소를 지도로 표출하기 위해 사례도시를 250×250m의 격자 단위로 데이터를 구축하였으며, 국토지리정보원(https://www.ngii.go.kr)에서 제공하는 격자를 기준으로 하였다.

지역정보는 격자 내의 건축물 평균건폐율, 평균용적률, 평균높이 등의 건축물 정보와 거주인구 및 유동인구 등의 인구 정보, CCTV, 경찰서(지구대 및 파출소 포함), 공시지가 정보 등을 포함하여 데이터를 구축하였다. 유동인구는 SK텔레콤에서 휴대전화 사용인구를 기반으로 추정하여 50×50m 단위로 제공하는 데이터를 활용하였다. 구체적인 데이터 구축방법은 표 1과 같다.

TABLE 1. Data Used in Machine Learning

Predictor variables (Source)		Category (Unit)
Crime Information	Crime (National Police Agency)	Season–Spring, Summer, Fall, Winter
		Time–Morning, Afternoon, Night, Midnight
		Type–Rape, Theft, Violence, Robber, Murder
Life Information	Weather (Weather data open portal)	Temperature (° C)
		Rainfall (mm)
		Wind speed (m/s)
		Humidity (%)
		Sunshine (hr)
		Insolation (MJ/m2)
		Snowfall (cm)
		Cloud cover (coverage level)
Local Information	Building and land (National Geographic Information Institute)	Average building coverage (%)
		Average Floor Area Ratio (%)
		Average height (m)
		Number of building (building)
		Average appraised value of land (won/㎡)
		Average area of residential zone (㎡)
		Average number of ground floors (floor)
	Population (National Geographic Information Institute, SK Telecom)	Total resident population (person)
		Total floating population (person)
	CCTV (Open data portal)	Number of CCTV (unit)
	Police station (National Police Agency)	Number of police station (unit)



	x	y	발생일자	계절별	시간대별	범죄유형	기온	강수량	습도
1	298423	191399	2011-03-08	봄	심야	절도	2.9	0	41
2	300212	191227	2011-11-25	가을	야간	절도	2.6	0	65.5
3	302209	191136	2011-05-06	봄	야간	절도	13	2.5	77
4	302122	191093	2011-08-26	여름	오전	절도	23.8	0.5	82.3

	x1	y1	x2	y2	x3	y3	x4	y4	평균 건폐율	총 건축물 수
1	298484	188948	298734	188949	298735	188699	298485	188698	74.01	319
2	298482	189198	298732	189199	298734	188949	298484	188948	75.94	264
3	299493	187202	299743	187203	299744	186953	299494	186952	67.94	101
4	298235	188697	298485	188698	298486	188448	298236	188446	67.03	178

FIGURE 2. Data pre-processing

범죄 정보발생 정보와 격자별로 구축된 지역 정보 및 생활정보를 취합하여 하나의 데이터세트로 구축하기 위해 본 연구에서는 QGIS를 이용하여 격자모서리 4점의 X, Y좌표를 추출하였다. 다음으로 범죄 발생 위치 좌표를 해당하는 격자에 할당하는 방식으로 데이터 전처리 과정을 거쳤으며 구체적인 흐름은 그림 2와 같다.

범죄발생 위험지역 예측 모형

앞 장에서 격자단위로 전처리한 4,906건의 실제 범죄데이터와 관련 정보를 머신러닝 훈련용(training) 데이터와 검정용(test) 데이터로 분리하였다. 이때 Ahn(2014), Kim and Ahn(2016) 등의 선행연구에서 적용하는 비율을 고려하여 랜덤하게 8:2, 7:3, 5:5로 나누어 각각 2회씩 머신러닝을 적용하고 가장 설명력이 좋은 모형을 탐색하였다. 응답변수(response variable)는 범죄 발생 위치가 속한 격자의 2년간 총 범죄 발생 건수로 설정하였으며, 예측변수(predictor variables)는 표 1에서 구축한 생활정보와 지역정보 데이터로 하였다.

이 때 머신러닝의 다양한 알고리즘 중에서 예측 정확도가 높은 알고리즘을 찾아야 되는데, 그 평가 척도는 대개 평균 제곱근 오차(Root Mean Square Error, RMSE)를 사용한다. 이는 실제값과 예측값의 오차를 나타내는 척도이며 식 4와 같다.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y_i})^2} \quad (4)$$

y_i : 실제 발생건수

$\tilde{y_i}$: 예측한 발생건수

n : 사례지역의 격자 수

구축된 데이터 셋을 사용하여 범죄발생 위험지역 예측 모형을 검증한 결과는 표 2, 3과 같다. 표 3에서 error range가 표시된 부분이 넓거나 많을수록 예측값과 실제값의 차이가 크다는 의미가 있으며, 세가지 모형 중 RMSE가 낮아 예측력이 가장 높은 모형은 의사결정나무모형인 것으로 나타났으며, SVM모형의 경우 예측력이 가장 낮은 것으로 나타났다. 즉 3개의 알고리즘 별로 학습용데이터와 검정용데이터를 8:2 비율로 나누어 훈련할 경우, 의사결정나무 방식의 2회차 실험에서 1.1241로 평균제곱근오차가 가장 낮아 예측력이 높은 것으로 나타났다.

TABLE 2. RMSE by each algorithm

Trial condition		Random Forest	Decision Tree	SVM
5:5	1st	1.9244	2.1893	10.137
	2nd	2.035	3.2756	11.426
7:3	1st	1.5485	1.6512	10.227
	2nd	1.491	2.2678	9.5387
8:2	1st	1.3713	1.615	9.224
	2nd	1.5192	1.1242	9.1156

TABLE 3. Response Plot by Primary Model

	Random Forest 8:2-1st	Decision Tree 8:2-2nd	SVM 8:2-2nd
Response plot			
RMSE	1.3713	1.1242	9.1156

다음 단계로서 학습된 알고리즘을 이용하여 실제 범죄 위험지역을 예측해야 한다. 그 방법으로는 범죄 발생 위험지역을 범죄발생 빈도로 표현하는 방법과 범죄발생 확률로 나타낼 수 있다. 그러나 본 연구는 범죄위험지역을 예측하는 성격을 띠고 있으므로 식 5와 같이 각 격자별 범죄 발생 확률로 결과를 시각화하고자 한다. 즉 사례지역 격자 전체에서 발생할 것으로 예측된 범죄 건수의 합을 분모로 하고, 각 격자의 범죄 발생 예측 건수를 분자로 하여 각 격자별 범죄 발생 확률을 산정하였다.

$$\frac{B}{A} \tag{5}$$

A : 전체 격자의 범죄 발생 예측 건수
B : 개별 격자의 범죄 발생 예측 건수

구축된 모형을 활용하여 범죄 발생 위험지역을 예측하고 이를 지도에 시각화하여 표현하면 사용자가 이해하기 쉬울 것이다. 또한 조건의 변화(예를 들면 날씨나 시간대의 변화)에 따라 즉각적으로 예측결과를 시각화하여 주는 자동화 시스템이 필요하지만, 본 연구는 아직 이 단계에 도달하지 못해, 다음과 같이 시나리오를 가정하여 각 시나리오에 따른 범죄발생 위험지역을 예측하고, 이를 시각화하는 방법으로 진행하고자 한다. 범죄유형은 그림 3과 같이 대상지역에서 2년 동안 절도 2,710건, 폭력 2,081건으로 강간 · 추행, 강도, 살인에 비하여 현저하게 많이 발생하고 있기 때문에 절도와 폭력으로 하였다.

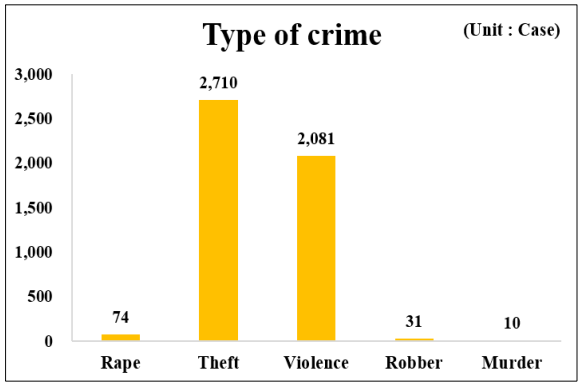


FIGURE 3. Status by Crime Type

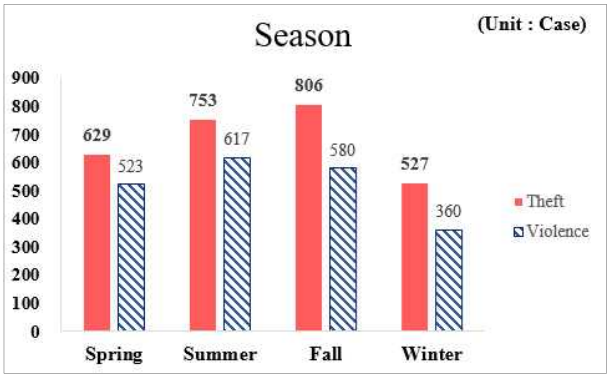


FIGURE 4. Crime Status by Season

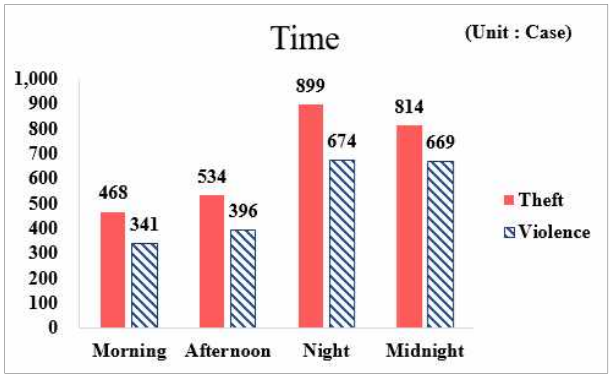


FIGURE 5. Crime Status by Time

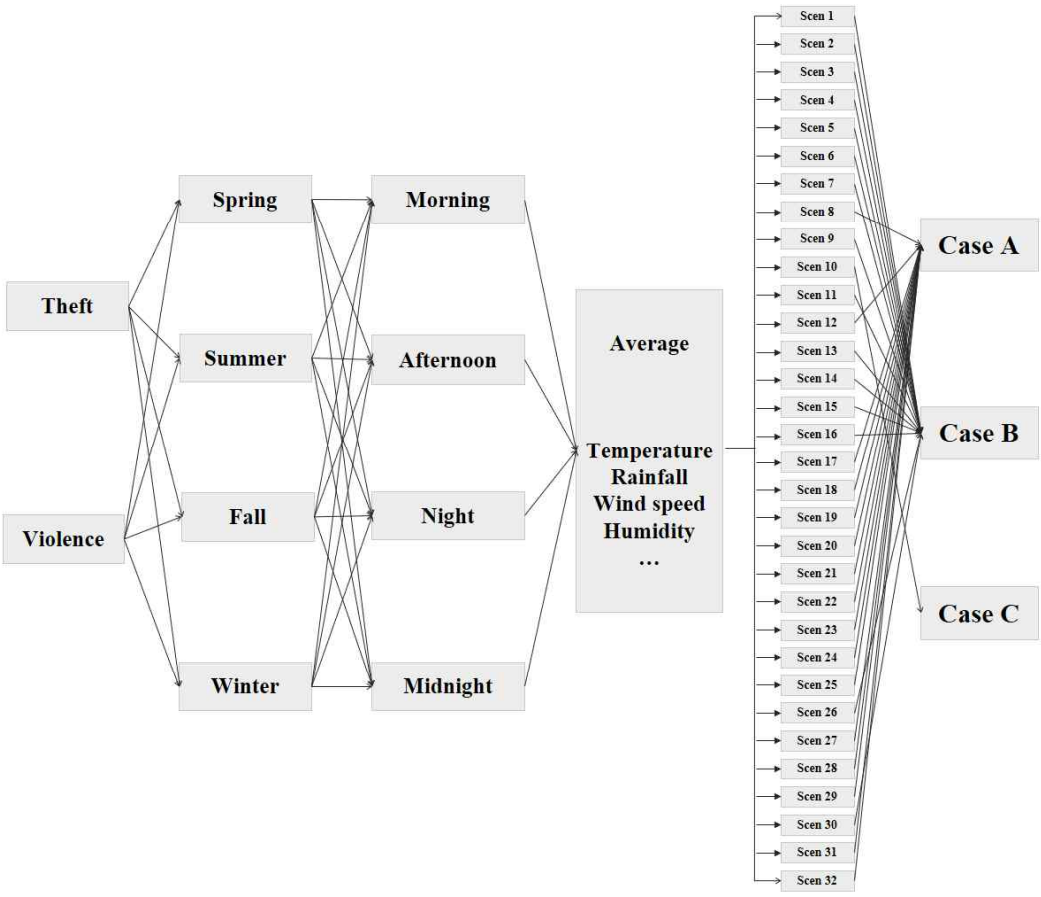


FIGURE 6. Scenario Setting Process

날씨의 경우는 시간에 따라 유동적으로 변화하는 변수이다. 동일한 계절이어도 오후인지, 심야인지에 따라 날씨 변수는 큰 차이를 보이며, 범위도 다양하게 분포한다(그림 4, 5 참조). 따라서 계절, 시간대에 따른 날씨별 평균값을 사용하여 시나리오를 작성하였다.

지역의 특성을 감안하여 시나리오 설정은 그림 6과 같다. 대상지인 J시에서 절도와 폭력범죄를 대상으로 계절별, 시간대별, 날씨정보를 바탕으로 J시의 상황에 맞는 32개 시나리오를 설

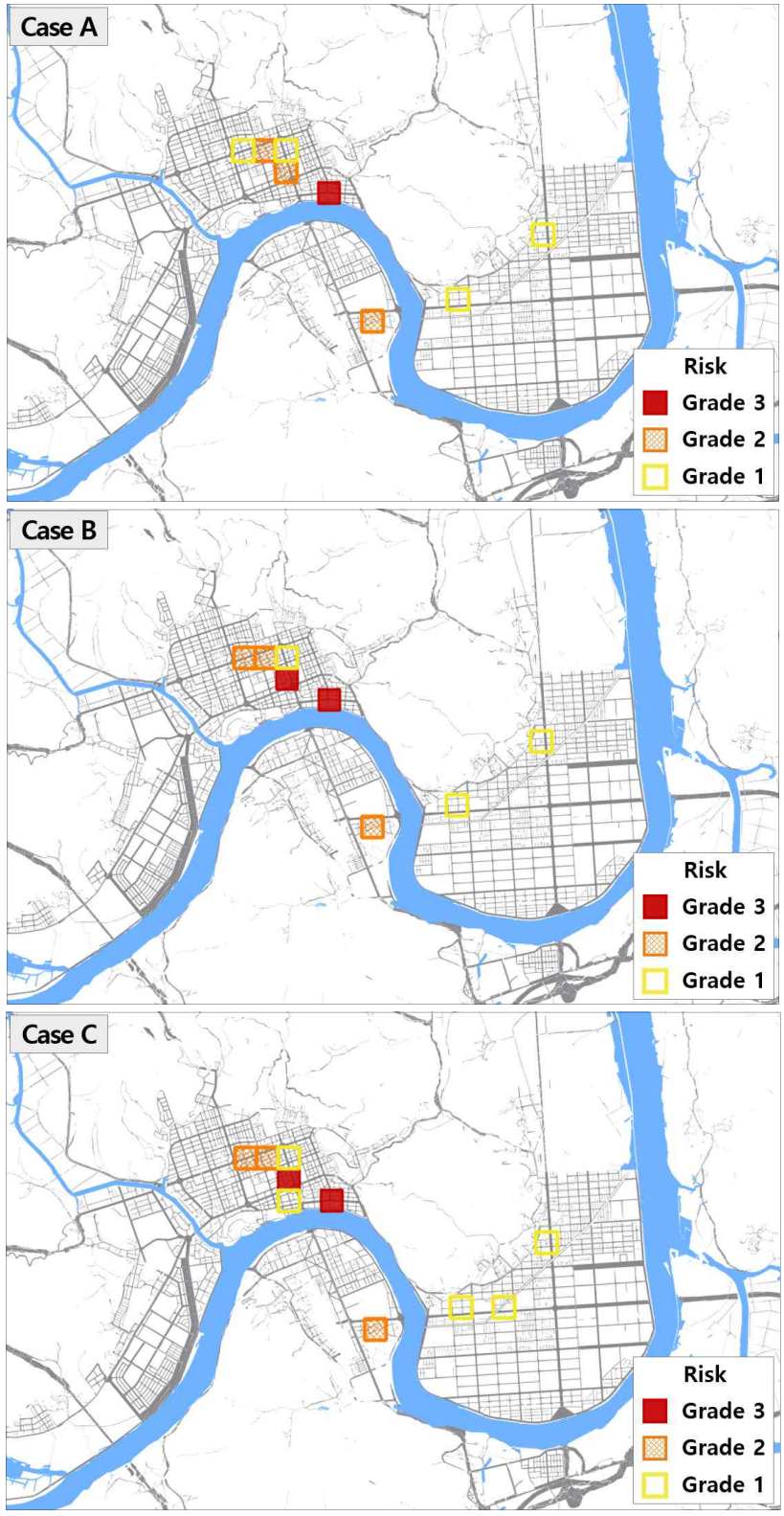


FIGURE 7. Visualization of Predicted Crime Risky Area by Case

정하였다.

각 시나리오에 따라 J시의 격자별 범죄 발생 위험 확률을 산정한 뒤 도출된 결과를 바탕으로 위험지역을 예측하였다. 조건마다 결과치가 약간 차이를 보이기는 하지만 범죄 발생 확률이 가장 높은 경우는 0.014~0.05 정도인 것으로 나타났다. 우리나라의 지방도시의 경우 범죄가 그다지 많지 않아 발생확률도 낮게 예측되었다. 하지만 본 연구에서는 상대적인 개념에서 위험 확률이 높은 순으로 Grade 1, 2, 3으로 분류하였다.

Grade 3은 범죄 발생 확률이 가장 높은 0.015와 다음으로 높은 0.01의 차이를 고려하여 0.01로 설정하고, Grade 2와 Grade 1은 Grade 3의 기준인 0.01에서 0.002씩 적어지는 약 0.008과 약 0.006을 기준으로 설정하였다. 따라서 각 격자별 산출되는 예측결과가 기준치보다 높을 경우에는 해당하는 단계의 위험지역으로 표시되도록 하였다.

그 결과를 지도상에 분포상황을 보면 그림 6, 7과 같으며, 32개의 시나리오 중 J시는 위험지역이 다르게 나타나는 패턴이 3가지 유형으로 구분되었다. 도출된 3가지 유형을 각 Case A, B, C로 할 때, Case A는 시나리오 8, 12, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 28, 29, 31, 32의 결과이며, Case B는 시나리오 1, 2, 3, 4, 5, 6, 7, 9, 11, 13, 14, 15, 16, 26, 30의 결과이고, Case C는 시나리오 10의 결과이다.

각 사례별로 그 속성을 구체적으로 살펴보면, Case A는 계절과 시간의 구분 없이 폭력범죄의 발생 위험이 높은 지역이며, Grade 3 수준의 범죄 발생 위험 지역이 1 곳으로 다른 결과에 비하여 적게 나타났다. 실제로 Case A에서 Grade 3 수준으로 범죄 발생 위험이 예측된 지역은 터미널, 음식점, 유흥주점 등이 위치하여 있어 유동인구가 많은 지역으로 통상적인 인식과 유사한 결과이다. 또한 Grade 2 수준의 범죄 발생 위험 지역도 Grade 3과 마찬가지로 음식점과 유흥주점이 밀집한 곳이다. 이들 지역의 실제 범죄도 위에서 Grade 3으로 예측된 지역에서 현저하게 많이 발생하고 있으며, 이를 통해 본 결과가 실제 지역의 특성을 잘 반영하고 있는 것으로 판단된다.

Case B는 계절과 시간의 구분 없이 주로 절도범죄의 발생 위험이 높은 지역으로서, Case A와 비교하였을 때, Grade 3 수준의 범죄 발생 위험지역이 1 곳 더 많은 2 곳으로 나타났으며, Case A에서는 Grade 2 수준의 범죄 발생 위험지역이었던 곳이 Case B에서는 Grade 3 수준의 범죄 발생 위험지역으로 변화한 것을 확인할 수 있다. 또한 Case A에서는 Grade 1 수준이었던 범죄 발생 위험지역이 Case B에서는 Grade 2로 변화하여, J시에서는 절도와 폭력범죄에 있어서 범죄 발생 위험 지역은 유사하나, 그 위험 정도에 있어 차이를 보이는 것을 확인할 수 있다.

Case C는 가을 오후에 발생하는 절도 범죄(날씨 요소는 실제 기존 가을 오후에 발생한 절도 범죄의 평균값)의 위험 지도로서 다른 날씨, 계절, 시간대, 유형 별 위험 지역 예측결과와 비교하였을 때, Grade 1 지역이 추가된 것을 확인할 수 있다. 이는 같은 종류의 범죄이거나 같은 계절 · 시간대이라고 해도 특정 계절, 시간대, 범죄 유형, 날씨 조건 등에 있어 위험 지역이 달라짐을 의미한다.

결 론

현재 경찰에서 가지고 있는 범죄자료는 텍스트로 되어 있는 범죄일지 수준에 지나지 않아 체계적인 데이터 구축과 4차 산업혁명시대에 대응하는 첨단 기법을 개발하고, 이를 이용한 공간분석을 통해 실효성 있는 범죄 예측과 정보제공이 필요하다. 최근 AI기술이 혁신적으로 발전함에 따라 이러한 기술들을 통해 파생되는 실시간 정보들을 방범전략에 적극적으로 활용해야 하며, 특히 빅데이터 축적이 가능한 환경이 조성되었기 때문에 방범분야에서도 활용 방안을 모색할 필요가 있다.

이에 본 연구에서는 지역 내 범죄 발생 데이터와 범죄발생과는 직접적인 연관성이 있지만 그동안 기술부족으로 고려되지 않았던 범죄 연

관 정보를 활용하여 빅데이터를 활용한 머신러닝 기법으로 범죄발생 위험지역 예측모형을 구축하고 검증해 보았다.

선행연구 및 사례에서 범죄발생에 영향을 미치는 요인 중 빅데이터로 구축 가능한 범죄정보, 날씨정보, 지역정보를 머신러닝에 활용할 수 있도록 데이터 처리를 하였으며, 이를 랜덤포레스트, 의사결정나무, SVM 방식으로 예측력을 각각 비교 · 분석하였다. 그 결과 3개의 알고리즘 별로 학습용데이터와 검정용데이터를 나누어 분석한 18개의 모형 중 RMSE가 가장 낮아 정밀도가 높은 의사결정나무모형(학습용:검증용 데이터 비율을 8:2로 하고, 2차 실험한 경우)이 최적모형인 것으로 분석되었다.

이를 바탕으로 가장 빈번하게 발생한 절도와 폭력범죄를 대상으로 시나리오를 작성하여 범죄발생 위험지역을 예측한 결과, J시는 시나리오별로 위험지역이 다르게 나타나는 3가지 패턴으로 도출되었다, 위험도 예측 결과는 발생 확률을 3개의 등급으로 시각화하여 시민들과 경찰관계자에게 도움이 되는 정보로 제공할 수 있게 하였다.

이상과 같은 과정을 보다 실용화하기 위해서는 상황의 변화에 따라 실시간으로 예측 가능한 예측 시스템(predictive policing system)을 개발해야 하지만 본 연구는 그 전단계의 성과이며, 향후 연구로 자동화 시스템 개발을 추진할 예정이다. 한편 예측 모델의 정밀도를 향상시키기 위해서는 머신러닝의 성격상 질 좋은 데이터를 대량으로 투입할 수 있어야 한다. 이 또한 지속적으로 향상시켜야 할 부분이다. 정확한 예측을 위해서는 본 연구에서 사용된 데이터 외에 생활기상지수, 건축물 1층 용도, 도시시설물 위치 등과 같이 범죄발생에 영향을 미치는 정보가 필요하고, 빅데이터 시대에 맞추어 지역정보를 시/군/구, 읍/면/동, 집계구/블록별 등 조금 더 세분화된 형태의 빅데이터로 구축한다면 좀 더 좋은 결과를 얻을 수 있을 것이다. 또한 실시간 뉴스, 지역 이벤트, 도시계획 및 개발 현황, SNS 및 온라인 커뮤니티 정보 제공 등의 실시간 데이터를 추가하여 종합적으로 분석할 수 있는 응용 모형 개발도 필요하며, 이들 과제는 향후 본격적인 시스템 구현과 함께 추가로 연구를 수행하고자 한다. KAGIS

REFERENCES

Almanie, T., R. Mirza and E. Lor. 2015. Crime prediction based on crime types and using spatial and temporal criminal hotspots. International Journal of Data Mining & Knowledge Management Process 5(4):1-19.

Ahn, H.C. 2014. Optimization of multiclass support vector machine using genetic algorithm : application to the prediction of corporate credit rating. Information Systems Review 16(3):161-177 (안현철. 2014. 유전자 알고리즘을 이용한 다분류 SVM의 최적화 : 기업신용등급 예측에의 응용. 한국경영정보학회지 16(3): 161-177).

Bae, S.W. and J.S. Yu. 2018. Predicting the real estate price index using machine learning methods and time series analysis model. Housing Studies 26:107-133 (배성완, 유정석. 2018. 머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측. 주택연구 26:107-133).

Brown, I. and C. Mues. 2012. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications 39(3):3446-3453.

Chaurasia, V. and S. Pal. 2013. Early prediction of heart diseases using data mining techniques. Caribbean Journal of Science and Technology 1:208-217.

Cho, Y.R., Kim, Y.C. and Y.S. Shin. 2017. Prediction model of construction safety accidents using decision tree technique. Journal of the Korea Institute of Building

Construction 17(3):295-303 (조예림, 김연철, 신윤석. 2017. 의사결정나무기법을 이용한 건설재해 사전 예측모델 개발. 한국건축시공학회지 17(3):295-303).

Choi, H.N. and D.H. Lim. Bankruptcy prediction using ensemble SVM model. Journal of the Korean Data & Information Science Society 24(6):1113-1125 (최하나, 임동훈. 2013. 앙상블 SVM 모형을 이용한 기업 부도 예측. 한국데이터정보과학회지. 24(6):1113-1125).

Choi, J.H. and D.S. Seo. 1999. Decision trees and its applications. Statistical Analysis Studies 4(1):61-83 (최종후, 서두성. 1999. 데이터마이닝 의사결정나무의 응용. 통계분석연구 4(1):61-83.)

Cortes, C. and V. Vapnik. 1995. Support-vector networks. Machine learning 20(3):273-297.

Elleng G. and COHN. 1990. Weather and crime. The British Journal of Criminology 30(1):51-64.

Guo, F., L. Zhang., S. Jin., M. Tigabu., Z. Su and W. Wang. 2016. Modeling anthropogenic fire occurrence in the boreal forest of China using logistic regression and random forests. Forests 7(11):250.

Hajek, P. and K. Michalak. 2013. Feature selection in corporate credit rating prediction. Knowledge-Based Systems 51:72-84.

Heo, J.Y. and J.Y. Yang. 2015. SVM based stock price forecasting using financial statements. KIISE Transactions on Computing Practices (KTCP) 21(2):167-172 (허준영, 양진용. 2015. SVM 기반의 재무 정보를 이용한 주가 예측. 정보과학회 컴퓨팅의 실제 논문지 21(3):167-172).

Heo, S.Y., J.Y. Kim and T.H. Moon. 2017. Crime incident prediction model based on Bayesian probability. Journal of the Korean Association of Geographic Information Studies 20(4):89-101 (허선영, 김주영, 문태헌. 2017. 베이지안 확률 기반 범죄위험지역 예측 모델 개발. 한국지리정보학회지 20(4):89-101).

Horrocks, J. and A.K. Menclova. 2011. The effects of weather on crime, New Zealand Economic Papers 45(3):231-254.

Jeong, J.H., J.H. Kim., J.H. Choo., S.H. Lee. and C.T. Hyun. 2017. Common maintenance cost estimation model using random forest for multi-family housing. Journal of the Architectural Institute of Korea 33(3):19-27 (정진호, 김종협, 추재호, 이승훈, 현창택. 2017. 랜덤 포레스트 기반 공동주택 공용관리비 추산모델. 대한건축학회 논문집-계획계 33(3):19-27).

Kim, S.J. and H.C. Ahn. 2016. Application of random forests to corporate credit rating prediction. Industrial Innovation Studies 32(1):187-211 (김성진, 안현철. 2016. 기업신용등급 예측을 위한 랜덤 포레스트의 응용. 산업혁신연구 32(1):187-211).

Kim, S.J. and B.Y. Kim. 2013. Comparative analysis of predictors of depression for residents in a metropolitan city using logistic regression and decision making tree. Journal of The Korea Institute of Building Construction 13(12):829-839 (김수진, 김보영, 2013, 로지스틱 회귀분석과 의사결정나무 분석을 이용한 일 대도시 주민의 우울 예측요인 비교 연구. 한국콘텐츠학회 논문지 13(12):829-839).

Lee, S.K. and T.S. Shin. 2018. Development and application of prediction model of

hyperlipidemia using SVM and meta-learning algorithm. Journal of Intelligence and Information Systems 24(2):111-124 (이슬기, 신택수. 2018. SVM과 meta-learning algorithm을 이용한 고지혈증 유병 예측모형 개발과 활용. 지능정보연구 24(2):111-124).

Lee, S. M. 2017. Spatial analysis of flood and landslide susceptibility in Seoul using random forest and boosted tree models. Master. Thesis, Univ. of Seoul, Seoul, Korea. 78pp (이선민, 2017, 랜덤 포레스트와 부스티드 트리 모델을 적용한 서울의 홍수와 산사태 취약성 분석, 서울시립대학교 석사학위논문. 78쪽).

Neuilly, M.A., K.M. Zgoba., G.E. Tita and S.S. Lee. 2011. Predicting recidivism in homicide offenders using classification tree analysis. Homicide studies 15(2): 154-176.

Newburn, T. and R. Sparks.(eds.). 2004. Criminal Justice and Political Cultures: National and international dimensions of crime control. Willan Publishing. UK.

Oh, B.H., K.W. Chung. and K.S. Hong. 2015. Gaze recognition system using random forests in vehicular environment based on smart-phone. The Journal of The Institute of Internet. Broadcasting and Communication 15(1):191-197 (오병훈, 정광우, 홍광석. 2015. 스마트폰 기반 차량 환경에서의 랜덤 포레스트를 이용한 시선 인식 시스템. 한국인터넷방송통신학회 논문지 15(1):191-197).

Oliveira, S., F. Oehler., J. San-Miguel-Ayanz., A. Camia and J.M. Pereira. 2012. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. Forest Ecology and Management 275:117-129.

Park, W.K. and S.Y. Kim. 2003. A Study on TV program rating prediction : Emphasizing the comparison of prediction capability between regression model and data mining model. Advertising Research 58:61-79 (박원기, 김수영. 2003. 시청률 예측에 관한 연구 : 회귀모형과 데이터마이닝 모형의 예측력 비교를 중심으로. 광고연구 58:61-79).

Ranson, M. 2014. Crime, weather, and climate change. Journal of Environmental Economics and Management 67(3):274-302.

Song, J.Y. and T.M. Song. 2018. Crime prediction using Big Data. Hwangsogeoleum academi. Seoul. 414pp (송주영, 송태민. 2018. 빅데이터를 활용한 범죄예측. 황소걸음아카데미. 서울. 414쪽).

Song, Y.S., Y.C. Cho., Y.S. Seo and S.R. Ahn. 2009. Development and its application of computer program for slope hazards prediction using Decision Tree Model. Journal of The Korean Society of Civil Engineers 29(2):59-69. (송영석, 조용찬, 서용석, 안상로. 2009. 의사결정나무모형을 이용한 급경사지재해 예측프로그램 개발 및 적용. 대한토목학회논문집 29(2):59-69).

Wikipedia. https://www.wikipedia.org/ (위키피디아).

Yoo, B.K., K.Y. Choi and D.K. Kim. 2018. An study on shopper' s retail format choice via Machine Learning Method : Based on national chain market and traditional market. The Journal of Business Education 32(1):155-174. (유병국, 최규영, 김대관. 2018. 기계학습기법을 활용한 소비자의 소매유형 선택 연구 : 대형마트와 전통시장을 중심으로. 상업교육연구

32(1):155-174).

Yoo, J.E. 2015. Random forests, an alternative data mining technique to decision tree. Journal of Educational Evaluation 28(2):427-448 (유진은. 2015. 랜덤 포레스트 : 의사결정나무의 대안으로서 데이터 마이닝 기법. 교육평가연구 28(2): 427-448).