한국빅데이터학회
제3권 제2호, 2018, pp. 101-112

기계학습을 활용한 데이터 기반 경찰신고건수 예측*

The Data-based Prediction of Police Calls Using Machine Learning

최재훈

경찰청 경찰대학 치안대학원 경감, 공학석사

요 약

본 연구는 기계학습의 하나인 신경망 분석과 음이항 회귀분석을 활용하여 경찰신고건수를 예측하고자 2016년 6월부터 2017년 5월까지 충남지방경찰청에 접수된 112신고 데이터를 이용하여 예측모델을 개발하였다. 모델을 개발하기 위해 경찰신고건수에 영향을 줄 수 있는 시간, 휴일, 휴일 전날, 계절, 기온, 강수량, 풍속, 관할면적, 인구, 외국인 수, 단독주택비율, 기타주택비율 변수 등을 활용하였다. 변수의 종류에 따라 몇몇은 경찰신고건수와 양의 상관관계 또는 음의 상관관계가 확인되었다. 사용된 두 개의 방법론을 비교한바, 신경망분석의 예측 결과는 예측 값과 실제 값의 상관계수 0.7702, RMSE 2.557이고, 음이항 회귀분석은 상관계수 0.7158, RMSE 2.831으로 나타났다. 신경망분석은 해석가능성은 낮지만, 음이항 회귀분석에 비해 예측력이 뛰어나다는 것이 확인되었다. 향후 경찰관서에서 본 연구의 예측모델을 기초로 하여 최적의 경찰력 배치를 할 수 있을 것으로 기대된다.

■ 중심어 : 경찰신고, 112신고, 신고예측, 기계학습, 신경망분석, 음이항 회귀분석

Abstract

The purpose of the study is to predict the number of police calls using neural network which is one of the machine learning and negative binomial regression, by using the data of 112 police calls received from Chungnam Provincial Police Agency from June 2016 to May 2017. The variables which may affect the police calls have been selected for developing the prediction model : *time, holiday, the day before holiday, season, temperature, precipitation, wind speed, jurisdictional area, population, the number of foreigners, single house rate and other house rate*. Some variables show positive correlation, and others negative one. The comparison of the methods can be summarized as follows. Neural network has correlation coefficient of 0.7702 between predicted and actual values with RMSE 2.557. Negative binomial regression on the other hand shows correlation coefficient of 0.7158 with RMSE 2.831. Neural network has low interpretability, but an excellent predictability compared with the negative binomial regression. Based on the prediction model, the police agency can do the optimal manpower allocation for given values in the selected variables.

■ Keyword : Police Calls, Prediction, Machine Learning, Neural Network, Negative Binomial Regression

2018년 12월 07일 접수; 2018년 12월 07일 수정본 접수; 2018년 12월 31일 게재 확정.
* 이 논문은 2018년 충북대학교 대학원 빅데이터협동과정 석사학위논문을 수정·보완한 논문입니다.

Ⅰ. 연구의 필요성 및 목적

우리나라는 세계 어느 국가와 견주어도 뒤지지 않는 뛰어난 치안환경을 갖추고 있다. 국민의 생명과 재산을 지키기 위해 지금 이 순간에도 12만 명의 경찰관들이 근무하고 있고 경찰은 시민들의 요청에 신속하게 대응하기 위해 '112 시스템'을 운영하고 있다. 112시스템은 누구든지 언제나 '112'로 전화만 하면 신고자와 가장 인접한 순찰차가 현장에 출동할 수 있도록 신고를 지령한다.

경찰은 경험적으로 평일보다는 휴일에, 주간보다는 야간에 더 많은 112신고가 접수된다는 것을 알고 있지만 인력운용의 문제로 112신고에 대응하는 경찰관을 주야, 휴일 구분 없이 거의 동일한 수준으로 배치하고 있다. 이와 같은 획일적인 경찰력 배치는 시간, 장소에 따라 변하는 치안수요를 충분히 만족시키기 어렵지만, 이에 대한 연구가 국내에서 제한적이고 국민의 생명 및 안전과 관련된 경찰력 배치를 탄력적으로 조절하는데 있어 과학적 근거가 부족하기 때문에 경찰청은 관행적으로 유지한 근무체계를 고수하고 있다.

그러나 국민의 요청에 신속히 응답할 의무가 있는 경찰은 이에 대해 좀 더 과학적으로 접근하여 가장 효율적이고 적합한 경찰력 배치를 통해 치안유지에 공백이 없도록 운용할 책임이 있다. 따라서 본 연구는 가장 대표적인 시민들의 도움요청 형태인 112신고가 어느 시간, 어느 지역적 특성에 따라 많이 발생하는지 확인하고 과거의 데이터로 미래의 경찰력 수요를 예측할 수 있는 모델을 구축하여 과학적 경찰활동에 일부 도움이 되고자 한다.

〈표 1〉 기계학습의 종류(이호현 외, 2016)

학습형태	방법론 종류	사용예시
지도학습 (Supervised Learning)	· 회귀분석 · 분류 · 퍼셉트론	· 보험, 신용카드 등 사기범죄 예측 · 음성인식
자율학습 (Unsupervised Learning)	· Self Organizing Map · 차원 감소 · 클러스터링	· 문자열 그룹핑 · 상품 추천 · 이상치 감지

Ⅱ. 이론적 배경

2.1. 기계학습

기계학습은 기계인 컴퓨터가 데이터의 특징을 발견하고 이를 검증 및 최적화하여 발전하는 과정이다. 이는 인간의 학습방법과 유사하지만 최근 정보의 양이 기하급수적으로 늘어남에 따라 생기는 인간의 한계를 컴퓨터가 보완할 수 있어 각광받고 있다(이호현 외, 2016). 기계학습은 학습하기 위한 데이터의 종류, 방법에 따라 입력 값을 이용하여 최대한 유사한 출력 값을 찾으려고 하는 지도학습, 입력된 데이터만을 이용하여 내재된 의미를 찾는 자율학습으로 크게 나눌 수 있는데, 구체적인 학습방법은 <표 1>와 같다.

지도학습의 회귀분석은 전통적인 통계적 방법론으로 사회과학에서는 가장 일반적으로 사용되는 분석방법으로 기계학습에서도 회귀분석으로 추정된 모델을 통해 예측 값을 추정할 수 있다. 그리고 퍼셉트론을 이용한 신경망분석은 주어진 독립변수와 종속변수로 상호간의 관계를 추정하는 것으로 회귀분석과 유사하지만, 제한적인 선형관계를 가정하는 회귀분석과 달리 비선형관계도 가정할 수 있다는 점에서 모델의 정확성, 예측능력을 향상시키기 위해 사용되고 있다(민대기, 2007; 이찬재 외, 2017).

많은 선행연구에서 RMSE(평균제곱오차의 제곱근, Root-Mean-Square deviation), MAPE(절대백분률오차의 평균, Mean Absolute Percentage

Error)를 이용하여 회귀분석과 신경망분석의 성능을 비교하였는데, 일반적으로 신경망분석이 회귀분석보다 모델 추정력이 더 우수한 것으로 나타났다(최형준·김주학, 2006; 김명종, 2012; 정복희, 2013; 유상록 외, 2014). 다만 신경망분석은 역전파(Backpropagation)방식으로 모델 추정값의 오차가 최소한이 되도록 끊임없이 계산을 반복하여 모델을 추정하기 때문에 과적합의 문제가 발생하는 단점이 있다. 과적합 현상은 모델을 학습시키기 위해 제공한 데이터에 지나치게 적합하게 모델링함으로써 오히려 예측하고자 하는 미래의 데이터에 대한 추정력이 떨어지는 현상을 말한다. 이를 해결하고자 전체 데이터로 모델링하지 않고 일정한 비율로 훈련데이터, 검증데이터로 구분하여 훈련데이터로 구축된 모델이 검증데이터를 추정하는데 문제가 없는지 확인하는 절차를 거치거나 Cross Validation 기법을 활용하여 과적합을 방지한다(민대기, 2007, 봉태호·김병일, 2017). 관련하여 신경망 분석은 적합정도를 모델의 정교성과 관련 있는 은닉노드수로 조절할 수 있는데 모델 구축에 가장 적합한 은닉노드수를 구하기 위해서는 각기 다른 노드수의 모델의 예측력을 비교하여 검증데이터에 대한 예측력이 떨어지지 않은 한도 내에서 가장 적합도가 높은 모델을 선택한다(이원희, 2006).

회귀분석에 대해서는 일반적으로 다중회귀분석이 많이 사용되지만, 종속변수의 분포가 정상성을 충족하지 못하는 경우에는 포아송 분포 또는 음이항 분포를 이용한 회귀분석을 사용해야 한다. 특히 종속변수의 분산이 평균보다 큰 과대산포현상이 나타나는 경우에는 음이항 회귀분석으로 모델을 추정해야 회귀계수의 불편추정량을 얻을 수 있다(신동준, 2011; 정재풍, 2013; 서영수, 2014; 김형준·최열, 2016).

2.2. 범죄의 원인

주류 범죄학은 주로 범죄발생의 원인에 대한 논의를 통해 발전되어 왔다. 우선 인간 개인의 관점에서 인간의 신체적 특성 또는 심리학적 요인으로 인해 범죄를 저지른다는 생물학적 범죄학, 심리학적 범죄학 관점과 인간이 구성하는 사회의 구조적 요인으로 범죄를 억제하는 사회기능이 저해되어 범죄가 발생한다는 사회학적 범죄학으로 크게 구분할 수 있다. 그리고 인간의 행동에 영향을 미칠 수 있는 시간, 날씨 등 환경적 요인이 범죄의 원인이 될 수 있는 관점이 있다.

사회구조적 원인이 범죄의 원인이 될 수 있다는 사회해체이론은 시카고학파의 연구에서 시작된 것으로 빈곤, 잦은 거주지의 이동, 이질적인 민족적 구성, 높은 인구밀도, 주거지역과 상업지역의 혼재, 지역의 노후화 등으로 인해 범죄가 많이 발생한다고 주장하였다(Stark 1987; Smith et al., 2000; Kubrin and Weitzer, 2003). 그리고 환경적 요인인 날씨와 시간이 범죄에 영향을 많이 미친다고 한다. 이에 대해 기온이 높아질수록 인간의 공격성이 강해져 범죄가 많이 발생한다는 연구(Anderson, 1989)가 있는 반면, 온도와 폭력성은 일정한 비례관계 있지만, 일정수준 이상에 도달되면 오히려 부정적인 환경에서 벗어나고자 하는 욕구가 강해져 폭력성이 줄어든다는 연구(Bell, 1992)가 있어 그 관계가 선형관계인지 비선형관계인지 명확하지 않다. 그 외에 풍속, 강수량, 일사량 등 여러 날씨요인이 인간의 일상생활과 관계되어 범죄발생에 영향을 미칠 수 있다는 것을 많은 선행연구에서 확인하였다(이윤호·김연수, 2010; Rotton and Cohn, 2000; Tompson and Bower, 2015). 그리고 자연적인 구분이자 사회제도적 구분인 시간의 개념과 이를 확장한 요일, 계절 등에 따라서도 범죄발생정도가 영향을 받는데, 일반적으로 주간보다는 야간에 범죄가 많이 발생하고 평일보다는

주말에, 다른 계절보다는 여름에 범죄가 더 많이 발생하는 것으로 나타났다(Cohn, 1993; Tompson and Bower, 2015).

Ⅲ. 연구대상 및 방법

본 연구는 112신고에 영향을 미치는 요인과 이를 예측하기 위한 모델링을 위해 2016. 6. 1. 부터 2017. 5. 31.까지 충남지방경찰청에 접수된 112신고 데이터 507,701건을 사용하였다. 충남지방경찰청에는 16개의 경찰서가 있고 경찰서 산하에는 총 121개의 지역경찰관서가 있다. 1건의 112신고가 1건의 데이터로 기록되는데 신고와 관련된 접수일시, 신고의 분류, 관할 경찰서, 관할 지역경찰관서, 대응코드 등의 내용이 포함된다.

112신고는 신고의 내용에 따라 5가지의 대응코드로 나뉘는데(경찰청, 2017 : 20), 코드번호는 0부터 4까지 있고 숫자가 작을수록 긴급한 신고를 의미하는데, Code 3과 Code 4는 경찰관이 즉시 출동해야 할 신고가 아니라 단순 민원 또는 문의성 신고이므로 본 연구에서는 정확한 분석을 위해 제외하였다.

분석의 시간적 단위는 Tompson and Bowers (2015)가 1일을 4개의 시간대로 구분하여 더미변수로 표현한 것처럼 1일을 6개의 시간대(02:00～05:59, 06:00～09:59, 10:00～13:59, 14:00～17:59, 18:00～21:59, 22:00～25:59)로 구분하여 시간대별로 112신고 접수건수의 차이를 비교하고자 하였다. 시간을 더미변수로 표현한 이유는 신경망분석은 비선형관계도 가정할 수 있지만, 회귀분석은 단순 선형관계만 가정하므로 시간을 단순히 순서척도로 측정하게 되면 0에서 23까지 증가하는 개념으로 모델을 구축하기 때문에 실제로 순환되는 시간의 개념을 제대로 고려할 수 없어 더미화 하였다. 다만, 모든 시간을 더미화하는 하는 경우에는 더미변수가 지나치게 많아지고 1시간단위로 접수되는 112신고건수가 충분하지 않아 4시간 단위로 묶어서 분석하였다. 그리고 경찰관의 세부지역별 근무교대시간이 2시간 단위인 것을 고려하여 경찰관 근무시간 패턴에 맞춰 4시간을 하나의 시간대로 설정하여 총 6개의 시간대, 5개의 더미변수로 표현하였다.

지역적 단위는 행정구역 단위로 제공되는 사회, 날씨 등에 관한 변수와 수준을 맞추기 위해 행정구역(읍·면·동)을 기준으로 하되, 행정구역과 지역경찰관서(지구대·파출소) 관할이 일치하지 않는 경우 수개의 행정구역을 묶어 일치시키려고 노력하였으나 지역의 단위가 혼재되어 일치되지 않는 지역은 제외하였다. 그리고 시골지역으로 1년에 2,000건 미만, 4시간 기준 0.91건 미만으로 접수되는 지역은 같은 충남지역의 지역경찰관서라 하더라도 도시지역과 이질성이 굉장히 크고 대부분의 데이터가 0이므로 오히려 분석에 오류를 야기할 수 있어 제외하였다.

독립변수로 사용한 기온, 강수량, 풍속은 기상청 자료를 이용했고 인구, 면적, 외국인수, 주거유형 등의 사회적 변수는 국가통계포털과 정보공개청구를 통해 수집하였다. 주택비율은 관할 구역의 주거형태를 구분하기 위해 해당 구역의 전체 주택을 단독주택, 아파트, 기타주택(상가주택, 다가구주택, 다세대주택, 오피스텔)로 구분한 것을 비율로 구분하였는데, 3가지 주거형태 중 아파트를 기준으로 해당 지역의 단독주택과 기타주택의 상대적인 비율로 측정하였다. 그리고 종속변수가 112신고건수 이므로 인구적 요소를 통제하기 위해 인구, 관할 면적 등을 통제변수로 사용하였다.

구체적인 분석방법은 R 3.3.1.버전을 사용하였고 음이항 회귀분석은 MASS 패키지의 glm.nb 함수를 사용하였다. 신경망분석은 nnet 패키지의 nnet 함수를 사용하였고 nnet함수의 옵션으로는 초기가중치 0.1, 학습오차 0.0000001,

최대 학습회수 5,000회로 설정하여 신경망분석을 시행하였다. 그리고 오차와 활성화 함수는 가장 일반적으로 사용되는 제곱오차와 시그모이드 함수를 사용하였다. 은닉노드 개수에 대한 명확한 기준은 없지만, 입력층이 13개 노드로 구성되었기 때문에 10개, 15개, 20개, 30개로 나누어 각각 모델링을 하였다(박혜영, 2016 : 22-23). 그리고 최적의 모델을 찾기 위해 전체 데이터를 7:3으로 나누어 모델링 후 검정하였는데 분석방법에 따른 성능의 차이가 유의미한지를 확인하기 위해 모델별로 30회씩 반복하였다.

Ⅳ. 연구가설

본 연구는 112신고에 영향을 미치는 요인이 무엇인지 분석하고 이를 통해 모델을 추정하여 향후 미래에 발생한 112신고 건수를 예측하고자 한다. 먼저 영향요인에 대해서는 선행연구들을 고려하여 인간의 행동과 굉장히 밀접한 관계가 있는 시간, 날씨, 지역의 사회구조적 요인이 영향을 미칠 것이라고 예상된다.

가설1. 112신고 접수건수는 시간, 날씨, 사회구조적 요인의 영향을 받을 것이다
- 가. 시간, 요일 및 계절에 따라 112신고 접수건수는 다르게 나타날 것이다.
- 나. 기온, 강수량, 풍속에 따라 112신고 접수건수는 다르게 나타날 것이다.
- 다. 인구밀도, 외국인에 따라 112신고 접수건수는 다르게 나타날 것이다.
- 라. 지역의 주된 주거형태(단독, 아파트, 기타)에 따라 112신고 접수건수는 다르게 나타날 것이다.
- 마. 요일과 날씨의 영향에는 일정한 교호작용이 나타날 것이다.
- 라. 계절과 날씨의 영향에는 일정한 교호작용이 나타날 것이다.
- 마. 사회구조적 요인 간에는 일정한 교호작용이 나타날 것이다.

가설2. 회귀분석을 통해 추정한 모델보다 신경망분석을 통해 추정한 모델이 예측성능에 있어 더 우수할 것이다.

그리고 요인의 영향과 그 유의정도를 추정하기 위해 회귀분석을 사용하였지만, 예측력의 측면에서는 상대적으로 신경망분석이 회귀분석에 비해 성능이 뛰어난 것을 확인하기 위해 동일한 데이터에 대해 모델링한 두 가지 모델을 성능의 측면에서 비교하고자 한다.

Ⅳ. 분석결과

4.1 기초통계량

분석에 사용된 설명변수와 종속변수의 최소값, 최대값, 평균, 표준편차를 제시하였다. 종속변수인 112신고건수는 최소 0건에서 최대 41건까지 나타나지만, 평균이 3.92로 정상성이 나타나지 않는다. 게다가 분산이 15.98로 평균인 3.92에 비해 과도하게 크므로 과대산포현상이 나타난다.

〈표 2〉 기초통계량

변수명	최소값	최대값	평균	표준편차
기온	-15.8	35.9	12.95	10.75
강수량	0	102.7	0.09	1.08
풍속	0	11.2	1.63	1.31
관할면적	1.46	222.7	64	54.6
인구	9,624	159,149	50,887	38,321
외국인 수	284	3,422	1,397	1,005
단독주택비율	3.5	67.1	26.9	18.1
기타주택비율	3.5	33	14.4	7.2
112신고건수	0	41	3.92	3.997

〈표 3〉 상관분석

	경찰신고	기온	강수량	풍속	관할 면적	인구	외국인수	단독주택비율
경찰신고	1							
기온	0.099*	1						
강수량	0.003	0.025*	1					
풍속	-0.09*	0.111*	0.027*	1				
관할면적	-0.24*	-0.011*	0.011*	0.102*	1			
인구	0.536*	0.001	-0.002	-0.079*	-0.197*	1		
외국인수	0.339*	0.007	-0.007	0.077*	-0.179*	0.617*	1	
단독주택	-0.354*	-0.013*	0.008	0.069*	0.567*	-0.527*	-0.403*	1
기타주택	-0.004	-0.011*	0.002	-0.012*	-0.171*	-0.244*	-0.202*	0.420*

유의수준: 0.05 > *

4.2 상관관계

종속변수인 경찰신고와 기온, 인구, 외국인 수가 양의 방향으로 유의미한 것으로 나타났고 풍속, 관할면적, 단독주택비율이 음의 방향으로 유의미한 것으로 나타났다. 이는 이변량 분석에서 나타난 것으로 다른 요인들은 통제되지 않은 상태의 결과라는 제한점은 있으나 각 설명변수가 종속변수에 일정한 영향을 미치는 것을 알 수 있다. 그리고 독립변수간의 상관관계가 최고 0.617로 나타나 다중공선성의 우려는 없다고 할 수 있다. 게다가 본 연구에서 사용되는 음이항 회귀분석은 최대우도추정법(MLE)에 의해 추정되므로 상대적으로 최소자승법(OLS)에 의한 추정방법보다는 다중공선성 문제로부터 자유롭다.

4.3 음이항 회귀분석

음이항 회귀모델의 분석결과에는 <표 4>와 같이 많은 시간, 공간, 구조적 변수에서 유의미한 한 것으로 나타났다. 독립변수는 변수간의 측정단위의 괴리가 크므로 표준화한 값을 사용하였다.

먼저 모든 시간대가 기준 시간인 02:00～05:59에 대비하여 유의미한 차이가 있는 것으로 나타났는데, 시간대1(06:00～09:59)은 음의 방향으로 기준시간대에 비해 20.5% 신고가 적게 접수되는 것으로 나타났고, 시간대2(10:00～13:59)는 10.8%, 시간대3(14:00～17:59)는 41.4%, 시간대4(18:00～21:59)는 75.2%, 시간대5(22:00～25:59)는 105.4% 더 많은 신고가 접수되어 야간일수록 112신고가 집중되고 새벽이 되면 줄어드는 비선형관계로 나타나는 것을 확인하였다.

그리고 여름, 휴일 전날, 휴일, 기온에 따라 각각 9.9%, 9.7%, 7.8%, 15.9% 더 많은 신고가 접수되는 것으로 나타나 112신고건수가 계절과 날씨의 영향을 받는 동시에 요일에 따른 영향도 받는 다는 것을 알 수 있었다. 그러나 계절, 요일과 날씨간의 교호작용에 의하면, 여름에 기온이 높은 경우에는 오히려 신고가 15.2% 줄어드는 것으로 나타나 여름과 기온이 112신고와의 상대적 관계에서 양의 방향으로 나타나기는 하지만 두 요인이 중첩되는 경우에는 오히려 음의 영향을 주어 여름철 지나치게 높은 기온은 오히려 사건사고를 감소시키는 것으로 나타났다. 그리고 휴일, 휴일전날, 가을과의 풍속의 교호작용에서는 각각 음의 방향으로 나타나 의무적인 활동을 하는 평일보다는 여가활동을 주로 하는 주말에는 날씨가 사람들의 외부활동에 영향을 미치고 이에 따라 112신고에 영향을 미치는 것으로 나타났다. 그리고 가을철의 풍속이 높은 것은

〈표 4〉 음이항 회귀분석 결과

	Coef	OR	Sig		Coef	OR	Sig
(Intercept)	0.7999	2.225	***	관할면적x인구	-0.0315	0.969	*
시간대1 (06:00～09:59)	-0.2299	0.795	***	면적x외국인수	0.3586	1.431	***
시간대2 (10:00～13:59)	0.1022	1.108	***	단독주택x인구밀도	-0.2897	0.748	***
시간대3 (14:00～17:59)	0.3464	1.414	***	기타주택x인구밀도	0.1260	1.134	***
시간대4 (18:00～21:59)	0.5609	1.752	***	휴일전날x기온	-0.0012	0.999	
시간대5 (22:00～25:59)	0.7197	2.054	***	휴일전날x강수량	-0.0003	1.000	
여름	0.0942	1.099	***	휴일전날x풍속	-0.0306	0.970	**
가을	0.0162	1.016		휴일x기온	-0.0125	0.988	
겨울	-0.0022	0.998		휴일x강수량	-0.0133	0.987	
휴일 전날	0.0928	1.097	***	휴일x풍속	-0.0425	0.958	***
휴일	0.0747	1.078	***	여름x기온	-0.1646	0.848	***
기온	0.1479	1.159	***	가을x기온	-0.0106	0.989	
강수량	-0.0107	0.989		겨울x기온	-0.0141	0.986	
풍속	-0.0147	0.985		여름x강수량	0.0196	1.020	
관할면적	-0.0054	0.995		가을x강수량	0.0218	1.022	
인구	0.4269	1.532	***	겨울x강수량	0.0064	1.006	
외국인수	-0.0046	0.995		여름x풍속	-0.0066	0.993	
단독주택비율	-0.3370	0.714	***	가을x풍속	-0.0327	0.968	**
기타주택비율	0.2871	1.333	***	겨울x풍속	-0.0013	0.999	
인구밀도	-0.1989	0.820	***	유의수준: 0.05 > * > 0.01 > ** > 0.001 > ***			

태풍 또는 악천후를 의미하므로 날씨가 좋지 않으면 사람들이 외부활동을 자제한다는 것을 알 수 있다.

구조적 요인으로 외국인의 수는 사회해체이론의 내용과는 달리 유의미하지 않은 것으로 나타났는데, 면적과 외국인의 수의 교호작용에서는 굉장히 강한 양의 상관관계를 나타나 상대적으로 관할면적이 넓은 시골지역의 외국인의 수가 112신고에 미치는 영향이 크다는 것을 알 수 있었다. 그리고 주거의 형태로 단독주택비율은 음의 상관관계를 가지고 기타주택비율은 양의 비율을 상관관계를 가지는데, 주거지역이 안정적인 단독주택지역보다는 다세대주택이 많은 기타주택지역에서는 상업시설과 혼재되어 있는 경우가 많으므로 112신고가 더 많이 접수되는 것을 추정할 수 있다. 그리고 인구밀도와 관련하여 인구밀도가 높은 곳이 오히려 신고가 적게 접수되었는데, 인구밀도와 주거형태와 교호작용과 관련하여서는 단독주택지역에서 더 적게 나타나고 기타주택에서는 더 크게 나타나 인구밀도가 높은 곳에서 주거유형이 더 큰 영향을 미치는 것으로 나타났다.

그렇다면 음이항 회귀분석을 통해 도출된 연구모형이 실제로 경찰신고를 예측하는데 얼마

나 효용이 있는지에 대해 확인을 할 필요하다. 본 연구에서는 모델을 처음 만들 때부터 과적합 문제 및 예측력 검정을 고려하여 전체 데이터를 7:3으로 임의로 나누어 70%의 데이터로 모델을 만들고(민대기, 2007) 모델링 후 남은 30%의 데이터를 모델에 대입하여 모델을 평가하였는데 그 결과는 <표 5>과 같다.

〈표 5〉 음이항 회귀분석 예측성능

데이터		상관계수	RMSE
훈련데이터 평균		0.7160	2.773
검정 데이터	평균	0.7158	2.831
	최고성능	0.7265	2.737
	최저성능	0.7031	2.846
	표준편차	0.0051	0.026

유상록 외(2014)는 회귀분석과 신경망분석의 예측력을 비교하기 위해 RMSE와 MAPE을 이용했지만, 본 연구에서는 RMSE만 활용하고 MAPE은 이용하지 않았다. 왜냐하면, MAPE의 경우에는 종속변수의 실제 값 중 0이 있으면, 계산이 불가능하기 때문이다. 대신 종속변수인 112신고 건수의 실제 값과 모델을 통한 예측 값 간의 상관계수를 활용하였다. 그리고 모델의 안정성을 확인하기 위해 모델링을 30회에 걸쳐 반복하였다. 그 결과 검정데이터 상관계수 기준으로 최저 0.7031에서 최고 0.7265로 안정적인 예측 결과가 도출되는 것을 확인하였다.

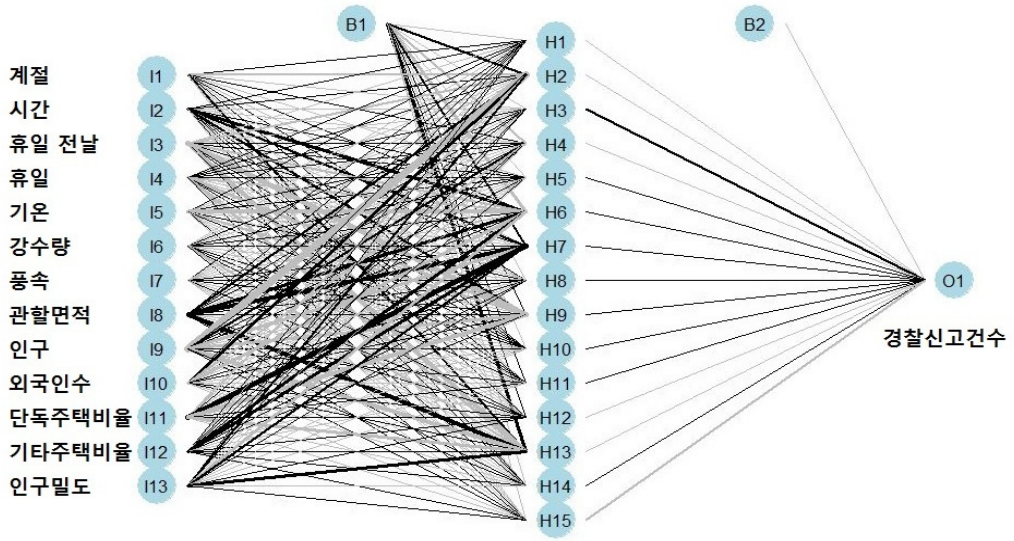
4.4 신경망 분석

신경망 분석은 많은 연구 등을 통해 회귀분석보다 예측성능이 뛰어난 것으로 알려져 있다. 그러나 회귀분석은 유의미한 변수를 찾아내고 변수별 영향의 방향 및 크기를 쉽게 알 수 있다는 점에서 해석가능성은 매우 높지만, 신경망 분석의 경우 인간의 인지능력으로는 매개변수인 은닉층의 역할을 해석하기가 거의 불가능하다(최형준·김주학, 2006; 김명종, 2012; 정복희, 2013). 그리고 신경망 분석은 과적합의 위험이 있으니 이를 해결하고자 여러 개의 모델 중에 가장 적합한 모델을 선택해야 하는데 본 연구에서 구성한 모델의 결과는 <표 6>과 같다.

은닉노드가 많아질수록 모델의 훈련데이터에 대한 적합도는 높아져 실제 값과 예측 값의 상

〈표 6〉 신경망 분석 예측성능

은닉노드	데이터		상관계수	RMSE	은닉노드	데이터		상관계수	RMSE
10	훈련데이터		0.7659	2.558	20	훈련데이터		0.7818	2.481
	검정 데이터	평균	0.7583	2.609		검정 데이터	평균	0.7706	2.545
		최고성능	0.7750	2.529			최고성능	0.7809	2.496
		최저성능	0.7400	2.683			최저성능	0.7507	2.610
		표준편차	0.0074	0.035			표준편차	0.0055	0.032
15	훈련데이터		0.7760	2.528	30	훈련데이터		0.7873	2.453
	검정 데이터	평균	0.7702	2.557		검정 데이터	평균	0.7521	2.742
		최고성능	0.7800	2.510			최고성능	0.7783	2.505
		최저성능	0.7612	2.644			최저성능	0.2899	7.982
		표준편차	0.0047	0.034			표준편차	0.0891	0.995



〈그림 3〉 신경망 분석 모델(은닉노드: 15)

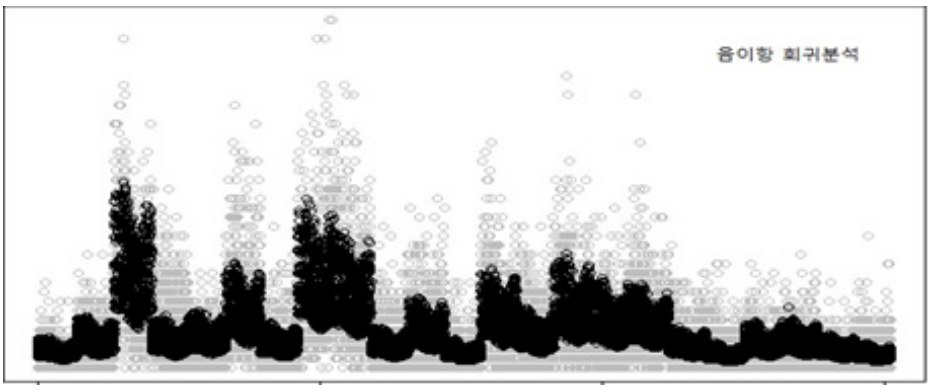
〈표 7〉 음이항 회귀분석과 신경망 분석의 예측성능 비교

		음이항 회귀분석		신경망분석(H: 15)	
데이터		상관계수	RMSE	상관계수	RMSE
훈련데이터 평균		0.716	2.773	0.7760	2.528
검정 데이터	평균	0.7158	2.831	0.7702	2.557
	최고성능	0.7265	2.737	0.7800	2.510
	최저성능	0.7031	2.846	0.7612	2.644
	표준편차	0.0051	0.026	0.0047	0.034

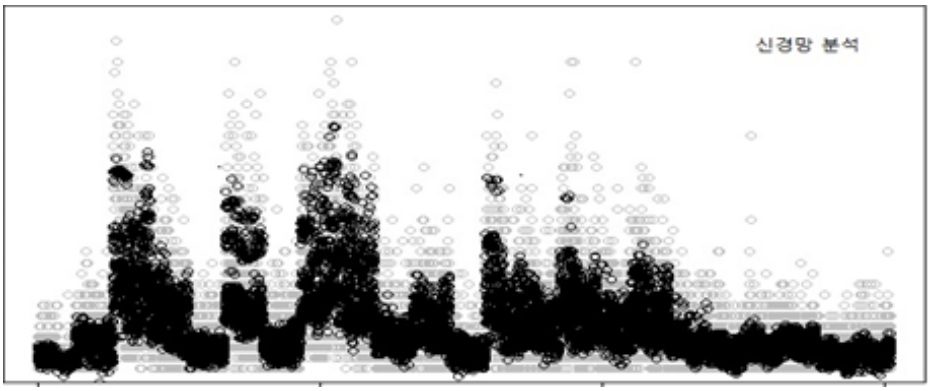
관계수는 커지고 RMSE는 작아진다. 그러나 이를 검정데이터에 적용할 경우에는 은닉노드가 20개인 경우에 15일 때보다 성능은 조금 좋아지지만 그 차이는 거의 없고 오히려 은닉노드가 많아질수록 예측결과의 편차나 과적합의 위험이 커지기 때문에 본 연구에서는 은닉노드가 15개인 모델을 선택하는 것이 더 적절한 것으로 보인다. 은닉노드 30개인 경우에는 30개의 모델 중 가장 성능이 나쁜 모델의 상관계수가 0.2899이고 RMSE가 7.982이므로 모델별로 편차가 너무 크다는 것을 알 수 있다. 그러므로 본 연구에서는 데이터에 대한 적합도를 최대한 높이면서 예측의 안정성과 과적합의 위험을 고려하여 <그림 3>의 은닉노드 15개 모델을 선정하였다.

4.5 예측성능 비교

양자의 예측성능을 비교하기 위해 수행한 훈련데이터 및 30회에 걸친 검정데이터에 대한 결과는 <표 7>과 같다. 신경망 분석은 해석능력이 떨어진다는 단점이 있지만 상관계수, RMSE 모든 측면에서 음이항 회귀분석에 비해 월등하다는 것을 알 수 있다.



〈그림 4〉 112신고 예측 그래프 – 음이항 회귀분석



〈그림 5〉 112신고 예측 그래프 – 신경망 분석

〈표 8〉 음이항 회귀분석과 신경망 분석성능의 t검정 결과

구분	방법론	등분산 여부	평균	표준편차	t-value
상관계수	음이항 회귀분석	p-value: 0.6683 등분산 가정	0.7158	0.0051	-43.042 ***
	신경망분석		0.7702	0.0047	
RMSE	음이항 회귀분석	p-value: 0.5948 등분산 가정	2.831	0.026	29.344 ***
	신경망분석		2.557	0.034	

이를 <그림 4>, <그림 5>와 같이 시각화를 해보면 그 차이를 더 명확히 알 수 있다. <그림 4>, <그림 5>는 각각 음이항회귀분석과 신경망분석의 예측성능을 나타내는 것으로 X축은 지역관서(23개)별로 시간대(2,189개)를 병렬적 표시한 것이고 Y축은 해당 시간대의 112신고 건수이다. 회색으로 표시된 실제 112신고 건수와 검은색으로 표시된 예측 값의 분포를 비교해보면, 음이항 회귀분석의 그래프보다 신경망 분석의 그래프가 품질이 좋다는 것을 알 수 있다.

그러나 음이항 회귀분석과 신경망 분석결과 간의 성능차이가 통계적으로 유의미한지 확인할 필요가 있다. 이를 위해 <표 8>와 같이 방법론별로 30회 반복하여 나온 결과를 t-검정하였는데 그 차이가 유의미한 것으로 나타났다. 그러므로 신경망 분석의 예측능력이 음이항 회귀

분석의 예측능력보다 통계적으로 우수하다는 것을 확인할 수 있었다(김명종, 2012).

Ⅴ. 한계 및 결론

국민의 생명과 재산을 지키기 위해 국민이 언제 어디서든지 신고만 하면 즉시 도움을 줄 수 있는 경찰의 112신고시스템은 매우 효율적으로 관리되어야 한다. 그러므로 경찰은 국민들이 경찰의 도움을 언제 어디에서 가장 많이 요청하는지 미리 분석하여 경찰력을 적절하게 배치하여야 할 것이다. 경찰은 경험적으로 평일보다는 주말에, 주간보다는 야간에, 주거지역보다는 상업지역에서 더 많은 112신고 건수가 접수된다는 것을 알고 있었지만 이때까지 이에 대한 심도 있는 연구가 없었기에 과학적으로 대응하지 못했던 점이 있었다.

그런데 본 연구를 통해 112신고가 시간, 날씨, 사회구조적 요인 등에 의해 어떻게 영향을 받는지 확인하였고 이를 예측하기 위해서 기계학습을 이용하여 모델링을 할 수 있다는 것을 확인하였다. 다만, 데이터의 문제와 연구방법의 한계로 몇 가지 연구상 한계점이 존재한다. 먼저 112신고는 세부적으로 다양한 범죄와 민원성 신고가 혼재되어 있고 이들은 각각 처리하는데 필요한 경찰력과 시간이 다름에도 불구하고 단순히 건수를 기준으로 치안수요를 측정한 점에 있어 한계가 있다. 그리고 실질적인 현장적용성을 높이기 위해 112신고가 거의 없는 지역경찰관서를 제외하고 분석하였는데, 이는 연구대상을 선택하는데 있어 연구자의 임의성이 과도하게 개입될 수 있는 부분이라 일정한 한계점을 가진다.

이러한 문제점에도 불구하고 국내에는 현재 112신고 건수를 분석하여 예측한 연구가 제한적인 상황에서 본 연구는 향후 4차 산업혁명시대에 발맞춰 획일적인 경찰력 배치를 지양하고 치안 수요에 맞는 인력운용을 할 수 있도록 도움을 줄 수 있는 탐색적 연구로서의 의의를 가진다고 할 것이다. 따라서 향후에는 본 연구의 한계점을 보완한 발전된 연구가 지속적으로 수행되기를 바란다.

참 고 문 헌

[1] 김명종(2012), “로지스틱 회귀분석과 인공신경망을 적용한 내부회계관리제도 평가모형의 성과비교," 국제회계연구, 46(1), 1-30.

[2] 김형준, 최열(2016), “음이항 회귀모형을 이용한 공간구문론 및 도시특성요소가 범죄발생에 미치는 영향 연구,” 대한토목학회논문집, 36(2), 333-340

[3] 경찰청(2017), 112신고처리매뉴얼, 서울, 경찰청.

[4] 민대기(2007), “SAS E-MINER를 이용한 고객 패턴 분석을 통한 신경망과 로지스틱 회귀의 비교,” Journal of the Korean Data Analysis Society, 9(4), 1861-1873.

[5] 박혜영(2016), 신경회로망을 이용한 데이터 분석, 서울, 카오스북

[6] 봉태호, 김병일(2017), “다중회귀분석 및 인공신경망을 이용한 자갈다짐말뚝 개량지반의 극한 지지력 예측,” 한국지반공학회논문집, 33(6), 27-36.

[7] 서영수(2014), “자동차보험 데이터를 이용한 포아송과 음이항 분포모형 비교연구,” 한국과학예술포럼, 18, 336-343.

[8] 신동준(2011), “범죄와 비행 연구의 가산자료 회귀분석 모형 활용에 대한 검토,” 범죄와 비행, 1(1), 121-137.

[9] 유상록, 김종수, 정중식, 정재용(2014), “인공신경망과 시계열 분석을 이용한 해상교통량 예측,” Journal of the Korean Society of Marine Environment & Safety, 20(1), 33-41.

[10] 이원희(2006), 신경망(NeuralNetworks)분석을 이용한 축구경기 순위예측모형 개발, 명지대학교 대학원 박사학위 논문.

[11] 이윤호, 김연수(2010), "날씨 및 요일특성과 범죄발생의 관계분석: 서울시 겨울철 범죄발생을 중심으로," 한국범죄심리연구, 6(1), 207-237.

[12] 이찬재, 김경도, 김용혁(2017), "뜰개 이동 예측을 위한 신경망 및 통계 기반기계학습 기법의 성능 비교," 디지털융복합연구, 8(10), 45-52.

[13] 이호현, 정승현, 최은정(2016), "기계학습 응용 및 학습 알고리즘 성능 개선방안 사례연구," 디지털융복합연구, 14(2), 245-258.

[14] 정복희(2013), 다중회귀분석과 신경망모델을 이용한 축제 만족도 평가방법, 중부대학교 대학원 박사학위 논문.

[15] 정재풍(2013), 교통사고건수에 대한 포아송 회귀와 음이항 회귀모형 적합, 고려대학교 대학원 석사학위 논문.

[16] 최형준, 김주학(2006), "인공신경망(Artificial Neural Network)을 이용한 2005년도 영국 윔블던 테니스 대회의 경기결과 예측에 관한 연구," 한국체육학회지, 45(3), 459-467.

[17] Anderson, C. A. (1989), "Temperature and Aggression: Ubiquitous Effect of Heat on Occurrence of Human Violence," Psychological Bulletin, 106(1): 74-96.

[18] Cohn, Ellen G.(1993), "The Prediction of Police Calls for Service : The Influence of Weather and Temporal Variables on Rape and Domestic Violence," Journal of Environmental Psychology, 13(1), 71-83.

[19] Kubrin, Charis E. & Weitzer, Ronald. (2003), "New Directions in Social Disorganization Theory," Journal of Research in Crime and Delinquency, 40(4): 374-402.

[20] Rotton, James and Ellen G. Cohn(2000), "Weather, Disorderly Conduct, and Assaults: From Social Contact to Social Avoidance," Environment and Behavior, 32(5), 651-673.

[21] Smith, William R., Frazee, Sharon G. & Davision, Elizabeth L. (2000), "Furthering the Integration of Routine Activity and Social Disorganization Theories : Small Units of Analysis and the Study of Street Robbery as a Diffusion Process," Criminology, 38(2): 489-524.

[22] Stark, Rodney. (1987), "Deviant Places : a Theory of the Ecology of Crime," Criminology, 25(4): 893-909.

[23] Tomson, L. A. and K. J. Bowers(2015), "Testing Time-sensitive Influences of Weather on Street Robbery," Crime Science, 4(1), 1-11.

저 자 소 개

최 재 훈(Jaehun Choi)

- 2004년 : 경찰대학 법학과(학사)
- 2018년 : 충북대학교 빅데이터협동과정(석사)
- 현재 : 경감, 경찰대학 치안대학원 범죄학과 석사과정 파견교육 중
- 관심분야 : Big data analytics, Data mining, Criminology