Journal of The Korea Society of Computer and Information
Vol. 21 No. 5, pp. 91-98, May 2016

www.ksci.re.kr
http://dx.doi.org/10.9708/jksci.2016.21.5.091

# Learning Method for Real-time Crime Prediction Model Utilizing CCTV

Seung-Hwan Bang*, Hyun-Bo Cho**

## Abstract

We propose a method to train a model that can predict the probability of a crime being committed. CCTV data by matching criminal events are required to train the crime prediction model. However, collecting CCTV data appropriate for training is difficult. Thus, we collected actual criminal records and converted them to an appropriate format using variables by considering a crime prediction environment and the availability of real-time data collection from CCTV. In addition, we identified new specific crime types according to the characteristics of criminal events and trained and tested the prediction model by applying neural network partial least squares for each crime type. Results show a level of predictive accuracy sufficiently significant to demonstrate the applicability of CCTV to real-time crime prediction.

▸ Keyword : Real-time Crime Prediction, Crime Types, Criminal Records, Neural Network Partial Least Squares

## I. Introduction

오늘날의 우리 사회는 과거보다 풍족하지만, 범죄의 발생 건수가 증가하고 지능화되고 있어 범죄를 예방하고 신속히 대응하기 쉽지 않다. 이에 정부와 지자체에서는 범죄를 예측하여 사전에 대응하고 발생 범죄에 신속히 대응하여 피해를 최소화하려는 노력을 기울이고 있다[1]. 범죄의 예측 및 예방을 위해서 경찰의 운영을 효과적으로 지원하는 관제 시스템[2]과 발생하는 범죄 패턴을 분석하고 의사결정을 지원하기 위한 시스템[3]이 개발되었다. 한편 범죄의 발생을 예측하기 위해서 과거 범죄 발생 기록을 활용한 시계열 예측 방법론 기반의 연구가 많이 이루어지고 있다[4]. 또한 GIS에 기반한 범죄 지도를 활용하여 범죄 발생지역을 분석하고 범죄의 발생을 예측하여 예방에 활용하고 있다[5]. 그러나 이러한 과거 범죄발생 기록 기반의 예측 방법론들은 현재 상황을 실시간으로 반영하지 못함에 한계가 있다.

최근에는, 범죄의 발생을 실시간으로 관찰하고 위험상황 발생 시 범죄에 적극적으로 대응하기 위한 시스템들이 구축되고 있다. 각 도처에 위치한 U-City 통합관제센터에서는 CCTV를 활용하여 위험상황을 자동으로 감지하고, 경찰과의 연계를 통해 상황에 대응하고자 하는 노력을 기울이고 있다[6]. 그러나 위험상황에 대한 정확한 인지가 쉽지 않으며, 시스템 관리자가 지속적으로 화면을 주시하고 있어야 한다. 이로 인해, CCTV에서 수집되는 데이터 기반의 위험 상황 판단의 필요성이 증대되고 있으며 보다 범죄 상황을 명확히 판단하여 인적 자원의 낭비를 줄일 필요가 있다.

CCTV를 활용한 실시간 범죄발생 예측을 위해, CCTV에서 수집되는 정보를 기반으로 범죄에 전문 지식이 없는 사람이 범죄의 발생을 예측할 수 있는지에 대한 연구가 진행되었으며[7], CCTV의 어떤 영상 정보가 범죄발생 예측에 활용될 수 있는지에 관한 연구가 진행되었다[8]. 또한, 예측 모델을 활용하여 총기 범죄발

• First Author: Seung-Hwan Bang, Corresponding Author: Hyun-Bo Cho
*Seung-Hwan Bang (seunbhwanb@postech.ac.kr), Dept. of Industrial and Management Engineering, Pohang University of Science and Technology
**Hyun-Bo Cho (hcho@postech.ac.kr), Dept. of Industrial and Management Engineering, Pohang University of Science and Technology
• Received: 2016. 02. 19, Revised: 2016. 03. 19, Accepted: 2016. 05. 17.
• This research was supported by the IT R&D program of MSIP/IITP [10047146, Real-time Crime Prediction and Prevention System based on Entropy-Filtering Predictive Analytics of Integrated Information such as Crime-Inducing Environment, Behavior Pattern, and Psychological Information].

생 예측 가능성에 관한 연구가 진행되었다[9]. 그러나 기존의 연구는 범죄발생 예측을 사람에 의존하거나 특정 범죄에 한정되어 있다는 점에 한계가 있으며, 기존에 연구된 과거 범죄발생 기록 기반의 예측 방법론들을 활용할 수 없어 새로운 예측 방법론을 필요로 한다.

본 연구에서는 CCTV 활용하여 실시간으로 범죄발생을 예측하기 위한 예측 모델 학습 방법론을 제안하고자 한다. CCTV를 활용한 범죄발생 예측을 위해서는 학습을 위한 데이터를 필요로 한다. 그러나 CCTV에서 수집되는 데이터는 범죄의 발생과 관련된 움직임 정보만 있을 뿐, 범죄의 발생 여부가 기록되어 있지 않아 예측 모델 학습에는 적합하지 않다. 반면, 범죄발생 기록의 경우 예측에 필요한 행동, 주위 환경 등과 관련된 정보를 포함하고 있어 예측 모델 학습에 활용할 수 있다. 범죄발생 기록을 예측 모델 학습에 활용하기 위해 CCTV 및 CCTV를 활용한 시스템에서의 수집가능성 여부를 고려하여 적합한 형태로 데이터를 변환하였다. 또한, 데이터의 특징과 범죄 예측 환경을 분석하여 예측 모델 반영 요소를 도출하였으며, 범죄 유형을 재분류하고 재분류된 유형별로 세부 예측 모델을 생성하여 실시간 범죄발생 예측에 활용할 수 있도록 하였다.

## II. Crime Prediction Method

본 연구에서는 실시간 범죄발생 예측을 위해 데이터 수집, 데이터 변환 및 CCTV 영상 데이터를 입력으로 받을 수 있는 범죄발생 예측 모델 학습 방법을 제안하고자 한다.

### 1. 활용 데이터

실시간 예측을 위해서는 예측 모델 학습을 위한 데이터가 필요하다. 실시간 예측은 CCTV 데이터를 활용하여 이루어지지만, 영상 데이터와 범죄기록을 매칭하여 수집하기 쉽지 않다. 그러나 보호관찰소에서 보관하고 있는 범죄기록은 가해자와 피해자의 인상 정보에서부터 가해자의 사전행동, 범죄 행동 및 사후 판결에 관한 기록이 포함되어 있다. 이중 사전행동 및 범죄 행동 기록은 가해자의 움직임에 관한 데이터로 CCTV에서도 획득 가능할 수 있다. 따라서 범죄기록을 활용하여 CCTV 데이터에 준하는 데이터를 수집하고 예측 모델을 학습 및 적용한다면 실시간으로 범죄발생 예측이 가능하다.

### 2. 범죄발생 예측 활용인자 선정

범죄발생 기록의 경우, 일반적으로 평문으로 작성되어있으며 모델을 학습하기에는 적합하지 않다. 따라서 적합한 변수를 정의하여 범죄기록을 예측 모델에 적합한 형태로 변환해야 한다. 정의된 변수는 범죄기록의 모든 정보를 데이터의 손실 없이 변환할 수 있어야 하며 범죄기록의 작성자마다 기록하는 정보 내용의 차이가 있어 많은 변수가 필요하다.

Table 1. The Possibility of Real-time Data Acquisition of Crime Prediction Factors

| | | Criminal Environmental Factors | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Absence of capable guardian | Offender | Victim | Nodes | Paths | Edges | Spatio-temporal | Legal |
| Crime Prediction Factors | Race | | | | ○ | ○ | ○ | | |
| | Gender | | ● | ● | | | | | |
| | Age | | ● | ● | | | | | |
| | Population | | | | ○ | ○ | ○ | ○ | |
| | Average income | | | | ○ | ○ | ○ | ○ | |
| | Academic ability | | | | ○ | ○ | ○ | ○ | |
| | Occupation | | | | ○ | ○ | ○ | ○ | |
| | Behavior Patterns | | ● | ● | | | | | |
| | Crime rate | | | | ○ | ○ | ○ | ○ | ○ |
| | Time | | | | | | | ● | |
| | Space | | | | ● | ● | ● | ● | |
| | Criminal tools | | ● | | | | | | |
| | Type of crime | | ● | ● | | | | | |

그러나 변수의 수를 증가시킴에 따라 상관성이 높은 변수

가 사용되어 예측 모델의 정확도에 영향을 미치게 된다. 따라서 데이터의 손실 없이 범죄기록을 예측 모델 학습에 적합한 형태로 변환하기 위해서는 최대한의 변수를 사용하여 범죄기록을 적합한 형태로 변환하고 변수간의 상관성분석을 통해 변수의 수를 줄여야 한다.

범죄기록을 예측 모델 학습에 적합한 형태로 변환하기 위한 변수는 CCTV에서 실제로 수집가능한지의 여부와 범죄 환경학(Environmental Criminology)에서의 범죄 환경요소를 고려하여 정의해야 한다. 범죄 환경학과 관련된 범죄발생 예측 인자[10]의 실시간 수집가능성을 전문가 기반으로 분석한 결과는 Table 1과 같다.

범죄발생 예측 인자가 범죄 환경학에서의 범죄 환경요소와 관련이 있으면서 CCTV에서 실시간 수집이 가능할 경우에는 "●" 기호를 활용하였으며, CCTV에서 수집이 불가능하나 실시간으로 수집될 수 있는 경우에는 "○"기호를 활용하여 표현하였다. 성별, 나이 등 사람과 관련된 정보는 CCTV에서 실시간으로 수집가능하다. 그러나 인구, 수입, 학력 등의 정보는 CCTV에서 수집이 불가능하며, 공공 정보[11]를 활용하여 데이터 수집이 가능하다.

CCTV에서 수집 가능한 범죄발생 예측 인자를 고려하여 범죄기록을 예측 모델 학습에 적합한 형태로 변환하기 위한 변수를 정의할 수 있다. 변수는 일반특성, 범행예비 및 범행방법으로 구분되며 Table 2와 같다.

Table 2. Categories of Crime Prediction Variables

| Category | Variable | Value |
|---|---|---|
| General | Gender | Male/Female |
| | Birth Year | ( )year |
| | Crime Occurrence Place | (Place Name) |
| | Crime Occurrence Time | Dawn/Morning/ Afternoon/Night |
| | Crime Occurrence Date | Year/Month/Day |
| | ... | ... |
| Before Crime Info. | Conversation Before Crime | Yes / No |
| | Meal or Drinking Before Crime | Yes / No |
| | Play or Amusement Before Crime | Yes / No |
| | Approach or Trail Toward Victim | Yes / No |
| | ... | ... |
| Crime Method Info. | Violence Using Hands | Yes / No |
| | Violence Using Feet | Yes / No |
| | Violence Using Deadly Weapon | Yes / No |
| | Threat | Yes / No |
| | ... | ... |

## 3. 실시간 범죄발생 예측 모델 학습

실시간 범죄발생 예측은 범죄기록을 사전에 정의된 변수를 활용하여 수집하고 이를 활용하여 범죄 유형 분류 및 유형별 적합한 세부 범죄발생 예측 모델을 활용하여 범죄발생가능성을 예측하는 순서로 진행되며 개념도는 Fig. 1과 같다.

### 3.1 범죄 예측 환경

범죄 예측에 적합한 모델을 선정하기 위해서는 범죄 예측 환경을 분석하고 예측 모델 반영 요소를 도출하여야 한다. 범죄의 발생은 주위환경의 영향을 받으며, 인구구조학적인 특징이나 발생 시간대, 장소 등에 따라 다른 범죄 패턴을 보인다[12][13]. 이로 인해 세분화된 범죄발생 예측 모델을 필요로 한다. 또한, 범죄의 발생과 직, 간접적으로 관련된 요인들은 서로 비선형 관계이기 때문에 비선형 예측 모델을 활용해야 한다. 범죄기록은 개인정보이며 보안상 수집이 어려워 예측 모델 학습에 활용될 수 있는 데이터의 양이 작지만 범죄기록을 정보의 손실 없이 활용하기 위해서는 많은 변수를 필요로 하므로 다변량 변수에 적합한 예측 모델이 필요하다. 이외에도 범죄기록은 평문으로 작성되어 있어 범주형 데이터에 적합한 모델이어야 하며, 하나의 범죄 사건이 발생하기까지의 가해자 행동 간에 인과관계 및 전후 상관관계가 존재하기 때문에 다중공선성을 줄일 수 있는 예측 모델을 활용해야 한다.

### 3.2 범죄 유형 재분류

범죄 유형은 일반적으로 폭행, 절도, 강도 등 범죄를 규정하기 위한 범죄의 대표적인 특징에 따라 분류된다. 또한, 각각의 범죄 유형은 범죄와 직, 간접적으로 연관된 가해자의 사전행동, 범죄 현장에서의 행동[14] 등에 의해 세부적으로 분류될 수 있다. 그러나 다른 유형의 범죄들이어도 유사한 사전행동이나 행동패턴을 보이는 범죄가 존재하며, 이러한 범죄기록들을 묶어 범죄 유형을 재분류하고 유사한 특징의 데이터를 묶어 예측 모델을 학습한다면 예측 모델의 정확도 향상에 이바지할 수 있다.

본 연구에서는, 범죄 유형을 재분류하기 위해 의사결정나무를 활용하였다. 의사결정나무는 뉴럴 네트워크나 로지스틱 회귀분석 등의 방법론들보다 정확도가 낮다고 평가되나 결과를 쉽게 이해하고 설명할 수 있으며, 의사결정을 하는데 직접 사용할 수 있다. 또한, 범죄 유형은 어떤 행동을 했는지 혹은 하지 않았는지 와 같이 명확한 기준에 의해 구분되어야 하나, 뉴럴 네트워크를 사용하여 분류할 경우 명확한 분류 기준을 찾기 힘들기 때문에 범죄 유형 재분류에는 적합하지 않다.

의사결정나무의 대표적인 알고리즘은 ID3, C4.5, C5.0, CHAID 및 CART가 있다. ID3 알고리즘[15]은 C4.5, C5.0, CHAID, CART의 기반이 되는 알고리즘으로 범주형 속성에서만 사용 가능하며, 상위 노드에서 사용된 속성은 다시 사용될 수 없다. 평가지수로는 엔트로피를 활용하며 다지분리를 지원한다. C4.5 알고리즘[16]은 ID3의 단점을 보완하고 범주형 속성뿐만 아니라 연속형 속성에서 사용할 수 있으며, 무의미한 속성을 제외하고 나무의 깊이를 결정할 수 있다. 평가지수로 정보이득을 활용하며 다지분리와 이진분리를 지원한다.

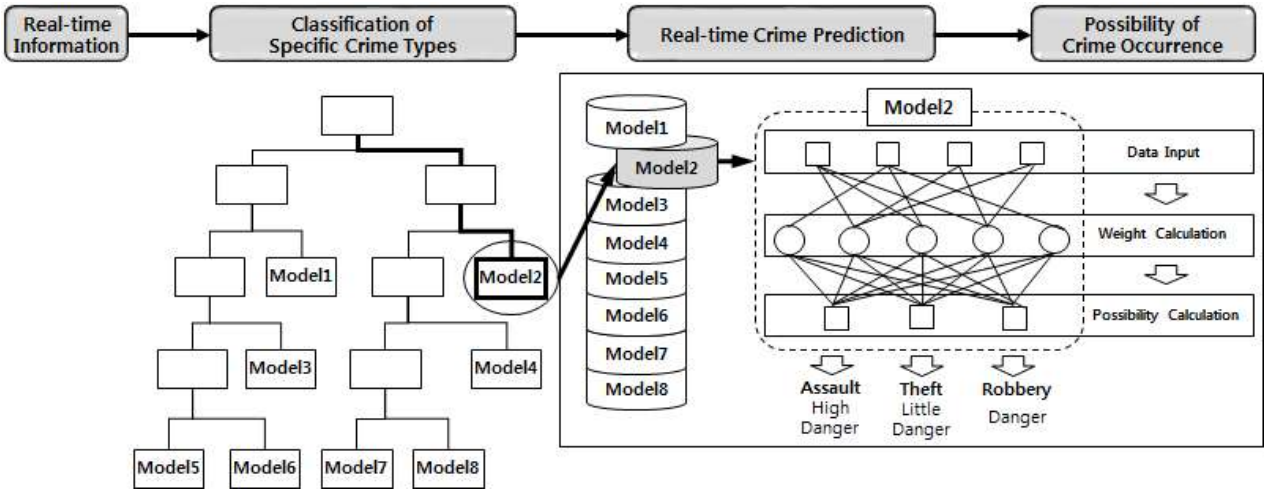


Fig. 1. Real-time Crime Prediction Process

C5.0 은 C4.5를 개선한 알고리즘으로 엔트로피 기반 이익비를 사용하여 속성을 선택하며, 명목형 속성만을 지원하나 가장 정확한 분류를 지원한다. 평가지수로 정보이득을 사용하며 다지분리와 이진분리를 지원한다. CHAID 알고리즘[17]은 범주형 속성에서만 적용가능하며 사전에 미리 정한 한곗값을 활용하여 나무의 구축 여부를 결정한다. 모형의 복잡도를 계산하기 위하여 사전 가지치기를 사용하며, 평가지수로 카이제곱 검정이나 F-검정을 사용한다. CART 알고리즘[18]은 ID3와 접근 방식은 같지만 지니 지수나 분산의 감소량을 사용하며, 사후 가지치기 방식을 사용하여 나무를 확장한다.

범죄기록의 경우, 범주형 속성과 연속형 속성의 정보들이 혼재되어 있으며 변수별로 다양한 속성값들을 포함한다. 따라서, 범주형과 연속형 속성에서 사용할 수 있어야 하며 이진분리가 아닌 다지분리를 지원해야 한다. 또한, 범죄기록이 많지 않은 경우, 의사결정나무의 깊을 너무 깊게 하여 유형을 세분화한다면 정확도는 올라갈 수 있으나, 과적합이 발생할 수 있다. 따라서, 적절한 수준에서 의사결정나무의 깊이를 결정할 수 있어야 한다. 이에 범죄기록을 활용하여 범죄 유형을 재분류하기 위해서는 C4.5 알고리즘이 가장 적합하다.

### 3.3 세부 범죄발생 예측

범죄 유형을 재분류한 이후, 각각의 유형별 적합한 범죄발생 예측 모델을 생성할 수 있다. 범죄발생 예측 모델은 범주형 및 다변량 변수에 적합해야 하며, 변수간 다중공선성을 줄일 수 있어야 한다. 위에서 언급한 내용을 반영할 수 있는 방법론으로는 부분최소자승법(Partial Least Square, PLS)이 있다. 또한 부분최소자승법은 데이터의 차원이 높은 경우, 주성분분석(Principal Component Analysis, PCA)에 비하여 예측력이 더 좋고 독립변수와 종속 변수를 동시에 잘 설명할 수 있다[19]. 부분최소자승법은 변수간 상관성이 높을 때 활용할 수 있으며[20], 독립변수 간에 다중공선성이 존재하고 잡음이 존재할 때 적합한 방법론이다[21]. 또한 부분최소자승법과 뉴럴 네트워크를 접목하여 활용한다면 독립변수와 종속변수간의 비선형 관계를 표현할 수 있다. 뉴럴 네트워크 기반 부분최소자승법의 독립변수와 종속변수는 다음과 같이 정의될 수 있다.

$$X = t_1 p_1^T + t_2 p_2^T + \ldots + t_A p_A^T + E = TP^T + E$$
$$Y = u_1 q_1^T + u_2 q_2^T + \ldots + u_A q_A^T + F = UQ^T + F$$

여기서 $\mathbf{T}(n \times A)$와 $\mathbf{U}(n \times A)$는 스코어행렬이며 $\mathbf{P}(m \times A)$와 $\mathbf{Q}(p \times A)$는 $\mathbf{X}(n \times m)$와 $\mathbf{Y}(n \times p)$의 로딩행렬이다. 뉴럴 네트워크 기반 부분최소자승법에서는 $\mathbf{t}$와 $\mathbf{u}$의 비선형 관계를 표현하기 뉴럴 네트워크를 사용하며 다음과 같이 정의된다.

$$u = N(t) + r$$

실시간으로 범죄의 발생을 예측하기 위해 뉴럴 네트워크 기반 부분최소자승법을 활용하였다. 독립변수로는 범죄의 발생과 관련된 행동인자를 활용하였으며 종속변수로는 폭행, 절도, 강도, 성폭력 및 살인의 범죄발생 가능성을 활용하였다. 범죄발생 가능성은 2개의 항목(위험/위험 없음)과 4개의 항목(매우 위험/위험/조금 위험/위험 없음)을 활용하였다. 또한 세부 범죄발생 예측 모델별 적합 범죄 유형을 도출하여 적합 범죄 유형을 예측할 수 있도록 종속변수를 설정하였다.

## III. Case Study

본 장에서는 실제로 우리나라에서 발생한 범죄기록을 활용하여 범죄 유형을 분류하고 범죄 유형별 예측 모델을 학습하였다. 의사결정나무를 활용하여 범죄 유형을 재분류하였으며,

재분류된 유형별로 범죄발생 예측 모델을 학습하여 실시간 범죄발생 예측에 활용하였다. 또한 실시간 범죄 예측을 위해 CCTV 영상 데이터의 활용가능성을 확인해보고자 한다.

### 1. 활용 데이터

본 연구에서 예측 모델의 학습을 위해 국내에서 발생했던 5대 범죄(폭행, 절도, 강도, 성폭행 및 살인) 기록을 수집하여 활용하였다. 데이터는 서울 보호관찰소에서 관리되고 있는 1624건의 가석방 및 집행유예를 포함한 실제 범죄기록을 활용하였다. 데이터 수집 결과 폭행 629건, 절도 755건, 강도 77건, 성폭행 144건 및 살인 19건의 기록을 수집할 수 있었으며, 데이터의 수집은 서울 보호관찰소를 직접 방문하여 평문으로 작성된 범죄기록을 예측 모델 학습에 적합한 형태로 변환하는 방법으로 진행되었다. 데이터 수집에는 전문가 의견을 기반으로 정의된 65개의 변수를 활용하였다.

데이터 수집에 활용된 변수 모두를 실시간 범죄 예측에 활용할 수 없어 예측 모델 학습에 제외하였다. 예를 들어, 가해자의 인적 정보(예. 학력, 전과 여부 등)와 사후 판결 정보(봉사시간, 수강시간 등)는 실시간으로 수집할 수 없어 제외하였다. 또한, 기존에 정의된 변수의 데이터를 CCTV에서 획득할 수 없는 경우(예. 만남에서 이동방법은 현존의 CCTV 기술로는 획득 불가)도 제외하였다. 변수의 실시간 수집 활용가능성을 고려하여 변수를 제외하고 의미가 중복된 변수(예. 범행시각 및 범행 시간대)를 제외하여 실시간 예측에 활용할 17개의 변수를 정의하였다 [Table 3].

Table 3. Using Variables to Train the Prediction Model

| Variable | Value |
|---|---|
| Gender | Male/Female |
| Age | ( )year |
| Place | Home/Public House/ BackStreet/Parking Lot/ School/Park/Amusement Center |
| Time | Dawn/Morning/ Afternoon/Night |
| Day | Mon/Tue/Wed/Thu/ Fri/Sat/Sun |
| Drinking Before Crime | Yes / No |
| Approach or Tail Toward Victim | Yes / No |
| Using Tools | Yes / No |
| House Breaking | Yes / No |
| Violence Using Hands or Feet | Yes / No |
| Kidnapping | Yes / No |
| Hiding | Yes / No |
| Attack While Walking | Yes / No |
| Disability of Victims | Yes / No |
| Attempting to Trespass by Using Instruments | Yes / No |
| Grabbing | Yes / No |
| Physical Fight | Yes / No |

또한, 범죄기록과 관련된 데이터 중 모든 변수에 대한 정보를 포함하고 있지 않은 경우, 그 수의 많지가 않아 예측 모델 학습에는 활용하지 않았다. 변수를 정의한 이후, 데이터의 품질을 향상시키고 예측 모델의 정확도를 높이기 위해 데이터 전처리를 진행하였다. 데이터의 전처리는 기존에 연구된 방법론[22][23]을 활용하여 진행하였다.

### 2. 실험 및 결과

본 연구에서 C4.5 알고리즘 기반의 의사결정나무를 활용하여 범죄 유형을 재분류 하였으며, 재분류된 범죄 유형별로 뉴럴 네트워크 기반 부분최소자승법을 활용하여 예측 모델을 학습하였다. 네트워크 기반 부분최소자승법의 적합성을 검증하기 위해 로지스틱 회귀분석과 뉴럴 네트워크와의 비교분석을 진행하였다. 전체 데이터 중 60%를 모델 학습, 20%을 테스트, 나머지 20%를 검증에 활용하였으며, 학습, 테스트 및 검증에 활용되는 데이터를 각각의 비율에 맞춰 임의로 구분하여 10회 반복하였다. 분석 결과는 Table 4와 같다.

로지스틱 회귀분석 및 뉴럴 네트워크와 뉴럴 네트워크 기반 부분최소자승법의 예측 정확도를 비교해 보았을 때, 예측 정확도의 큰 차이는 없다. 그러나, 예측모델 학습을 위한 데이터의 양이 증가할 경우를 고려하여 범죄예측모델로 뉴럴 네트워크 기반 부분최소자승법이 가장 적합하다고 판단하였다.

Table 4. Results of Prediction Model Test

| | | | | | | |
|---|---|---|---|---|---|---|
| Logistic R Misclassification rate(%) | #1 | #2 | #3 | #4 | #5 | |
| | 49.0 | 45.8 | 50.0 | 54.5 | 57.7 | |
| | #6 | #7 | #8 | #9 | #10 | Avg. |
| | 50.0 | 53.9 | 52.9 | 56.4 | 53.8 | 52.4 |
| Neural Networks Misclassification rate(%) | #1 | #2 | #3 | #4 | #5 | |
| | 45.2 | 44.5 | 46.8 | 49.4 | 50.0 | |
| | #6 | #7 | #8 | #9 | #10 | Avg. |
| | 50.0 | 46.8 | 52.3 | 54.5 | 48.8 | 48.8 |
| NN PLS Misclassification rate(%) | #1 | #2 | #3 | #4 | #5 | |
| | 43.9 | 44.5 | 44.9 | 44.2 | 52.6 | |
| | #6 | #7 | #8 | #9 | #10 | Avg. |
| | 50.0 | 44.8 | 51.6 | 50.6 | 48.7 | 47.6 |

범죄 유형을 재분류하고 범죄예측을 위한 모델을 학습하기 위해 1624건의 사건 기록 중 35건의 기록을 제외하고 1589건의 기록만을 활용하였다. 예측 모델의 검증을 위해서 CCTV에서 획득되는 데이터를 활용해야 하나, 본 연구에서는 수집된 범죄기록 중 CCTV에서 획득 가능한 변수에 해당하는 데이터만을 활용하여 검증하였다.

의사결정나무의 정확도는 의사결정나무 리프 노드들의 분류 순수도를 활용하여 평가하였다. 전체 데이터 중 60%를 모델 학습, 20%을 테스트, 나머지 20%를 검증에 활용하였으며, 학습, 테스트 및 검증에 활용되는 데이터를 각각의 비율에 맞춰 임의

로 구분하여 10회 반복하였다. 의사결정나무의 정확도를 평가한 결과 정확도는 약 71% 수준이며 반복 횟수별 결과는 Table 5와 같다.

Table 5. Results of Specific Crime Type Classification

| | #1 | #2 | #3 | #4 | #5 | |
|---|---|---|---|---|---|---|
| Classification Purity(%) | 73 | 69 | 69 | 73 | 73 | |
| | #6 | #7 | #8 | #9 | #10 | Avg. |
| | 72 | 72 | 69 | 74 | 70 | 71 |

범죄 유형을 재분류한 결과, 8개의 세부 유형이 생성되었으며 이때 활용된 분류 기준으로는 폭행 특성, 피해자 성별, 고함욕설여부, 고함욕설여부, 음주여부 및 폭행도구가 활용되었다. 분류 결과는 Table 6과 같다. 기존 5개의 범죄 유형은 8개의 범죄 유형으로 재분류되었다. 유형 1의 경우 절도 범죄가 대부분이나 폭행 12건 강도 8건 및 성폭행 2건이 같은 유형으로 분류되었다. 이처럼 유형 1에서 유형 4는 한 가지 유형의 범죄가 대부분을 차지했지만, 유형 5에서 유형 8은 다양한 유형의 범죄가 골고루 분류되었다.

Table 6. The Number of Criminal Records of Each Crime Types

| | Assault | Theft | Robbery | Sexual Offence | Murder | Sum |
|---|---|---|---|---|---|---|
| **Model 1** | 12 | 536 | 8 | 2 | 0 | 558 |
| **Model 2** | 366 | 21 | 41 | 2 | 7 | 437 |
| **Model 3** | 5 | 117 | 1 | 22 | 0 | 145 |
| **Model 4** | 193 | 7 | 6 | 19 | 4 | 229 |
| **Model 5** | 0 | 8 | 0 | 46 | 0 | 54 |
| **Model 6** | 4 | 40 | 0 | 10 | 1 | 55 |
| **Model 7** | 12 | 3 | 11 | 4 | 5 | 35 |
| **Model 8** | 28 | 1 | 7 | 38 | 2 | 76 |

재분류된 각각의 범죄 유형별로 세부 범죄발생 예측 모델을 생성할 수 있으며 각 예측 모델별 예측에 적합한 범죄 유형을 선정하였다. 적합한 범죄 유형은 하나의 유형에 포함된 범죄기록의 수를 기준으로 판단하였으며, 본 연구에서는 하나의 유형에 포함된 기록이 10건 이하일 경우 예측에 적합한 범죄 유형에서 제외하였다. 유형 1의 경우 폭행, 절도, 강도 및 성폭행에 관한 4가지 유형의 범죄가 포함되어 있으나 강도 및 성폭행의 경우 범죄기록 건수가 적어 적합 범죄 유형에서 제외하였다. 8가지 유형별 예측 모델의 적합 범죄 유형은 Table 7과 같다.

Table 7. Suitable Crime Types of Each Detailed Crime Prediction Model

| Model(Type) | Suitable Crime Types |
|---|---|
| Model 1(Type 1) | Assault, Robbery |
| Model 2(Type 2) | Assault, Theft, Robbery |
| Model 3(Type 3) | Theft, Sexual Offence |
| Model 4(Type 4) | Assault, Sexual Offence |
| Model 5(Type 5) | Sexual Offence |
| Model 6(Type 6) | Theft, Sexual Offence |
| Model 7(Type 7) | Assault, Robbery |
| Model 8(Type 8) | Assault, Sexual Offence |

각 유형별 범죄기록을 활용하여 세부 범죄발생 예측 모델을 학습하였으며, 사전에 정의한 값과 예측된 값을 비교하여 평가하였다. 종속변수로는 범죄 발생가능성을 활용하였으며, 발생가능성은 앞에서 정의한 2가지 및 4가지 항목을 사용하였다. 전체 데이터 중 60%를 모델 학습, 20%를 테스트, 나머지 20%를 검증에 활용하였으며, 학습, 테스트 및 검증에 활용되는 데이터를 각각의 비율에 맞춰 임의로 구분하여 10회 반복하였다. 예측 모델의 성능 평가 결과는 Table 8과 같다.

Table 8. Test Results of Each Detailed Crime Prediction Model (Unit: Accuracy)

| | Assault | | Theft | | Robbery | | Sexual Offence | | Murder | |
|---|---|---|---|---|---|---|---|---|---|---|
| **# of Class** | **2c** | **4c** | **2c** | **4c** | **2c** | **4c** | **2c** | **4c** | **2c** | **4c** |
| **Model 1** | 98 | 98 | 96 | 48 | 98 | 98 | 100 | 100 | - | - |
| **Model 2** | 86 | 59 | 96 | 96 | 90 | 89 | 99 | 100 | 98 | 99 |
| **Model 3** | 97 | 98 | 81 | 44 | 98 | 99 | 80 | 85 | - | - |
| **Model 4** | 82 | 68 | 98 | 97 | 97 | 97 | 92 | 92 | 98 | 98 |
| **Model 5** | - | - | - | 88 | - | - | 94 | 55 | - | - |
| **Model 6** | 90 | 94 | 69 | 51 | - | - | 83 | 97 | 98 | 99 |
| **Model 7** | 71 | 74 | 86 | 87 | 69 | 66 | 89 | 90 | 80 | 81 |
| **Model 8** | 74 | 75 | 99 | - | 87 | 93 | 63 | 52 | 97 | 97 |
| **Avg.** | 84 | 64 | 91 | 50 | 88 | 88 | 82 | 69 | 93 | 94 |

회색 음영인 부분은 유의미한 예측 정확도이며 음영이 없는 부분은 데이터의 수가 적어 예측 모델이 과적합(Over fitting)된 경우이다. 또한 값이 없는 부분은 해당 범죄 유형의 기록이 없어 예측 모델을 학습하지 못한 경우를 의미한다.

총 8개의 세부 범죄 예측 모델의 정확도를 가중 평균한 결과, 범죄 발생가능성을 2개의 항목으로 구분하였을 경우에는 87.54%, 4개의 항목으로 구분하였을 경우에는 59.44%의 정확도를 보였다. 단순히 2개의 항목으로 구분하여 범죄의 발생과 발생하지 않음과 같이 그 기준이 명확한 경우에는 예측 정확도가 높으나, 4개의 항목으로 구분할 경우에는 그 기준이 명확하지 않아 예측 정확도가 높지 않다. 특히, 폭행이나 절도의 경우에는 그 예측 정확도가 다른 범죄 유형에 비하여 더 낮은데, 이는 범죄의 특성상 특정한 패턴이 일정하지 않고 우발범죄가 많기 때문이다.

## IV. Conclusions

범죄를 예방하고 발생한 범죄에 대해 신속한 대응 필요성이 증가하고 있다. 이로 인해, CCTV에서 수집되는 영상 데이터를 입력으로 받아 범죄가 일어나기 직전 혹은 일어났을 때 범죄를 예측 및 판단이 필요하다. 이로 인해 CCTV에서 수집되는 데이터를 활용할 수 있는 실시간 범죄발생 예측 방법에 관한 연구를 진행하였다.

실시간 범죄발생 예측 모델을 학습하기 위해 CCTV 데이터가 아닌 서울 보호관찰소에서 관리 중에 있는 범죄기록을 활용하였으며, 예측 모델에 적합한 형태로 변환하여 사용하였다. 실시간 범죄발생 예측 모델은 범죄 유형을 분류하여 각 유형별 세부 범죄발생 예측 모델을 학습하는 방법으로 진행되었다. 범죄 유형을 재분류하기 위해 실제 범죄가 일어나기까지 가해자의 행동을 활용하였으며, 실제 범죄의 행동 특징이 유사하면 다른 범죄 유형이라도 같은 범죄 유형으로 분류되었다. 재분류된 유형별로 세부 범죄발생 예측 모델을 학습하였으며, 예측 모델은 데이터의 특징과 범죄 예측 환경에 적합한 뉴럴 네트워크 기반 부분최소자승법을 활용하였다. 다른 비선형 예측 방법론과 비교하여도 예측 정확도가 높음을 보였으며, 지속적인 활용 가능성 측면에서의 적합성을 보였다. 또한, 범죄기록에 작성된 범죄 유형이 아닌 실제로 가해자의 행동 특징에 따른 유형별로 예측 모델을 생성하여 정확도를 높일 수 있었다.

본 연구는 CCTV를 활용하여 범죄의 발생을 예측할 수 있는지를 알아보았다는 점에서 의의를 갖는다. 실제 CCTV 데이터를 활용하여 검증하지는 않았지만 예측 모델에서 활용되는 변수를 CCTV에서 수집가능한지 판단하여 정의함으로 추후 CCTV를 활용한 관제시스템에서 바로 활용할 수 있도록 하였다. 본 예측모델을 활용하여 범죄발생 가능성을 예측한다면, 기존의 통합관제시스템에서의 위험상황 판단 정확도를 더욱 높일 수 있으며, 이로 인해 효율적인 인력 운영이 가능해질 것이다. 또한, 실시간으로 수집되는 데이터뿐만 아니라 과거 범죄 통계 데이터 및 인구구조학적, 지리적 등의 데이터를 함께 고려한다면 더욱 정확한 예측이 가능할 것으로 예상된다.

## REFERENCES


[1] C. Shu, A. Hampapur, M. Lu, L. Brown, J. Connell, A. Senior, and Y. Tian, “IBM Smart Surveillance System(S3): A Open and Extensible Framework for Event based Surveillance”, IEEE Conference on Advanced Video and Signal Based Surveillance, pp. 318-323, Como, Italy, September 2005.

[2] Predpol, *http://www.predpol.com/*

[3] J. W. Brahan, K. P. Lam, H. Chan and W. Leung, “AICAMS: Artificial Intelligence Crime Analysis and Management System”, Knowledge-Based Systems, Vol. 11, No. 5, pp. 355-361, November, 1998.

[4] Y. S. Chung, J. M. Kim and K. R. Park, “A Study of Improved Ways of the Predicted Probability to Criminal Types”, Journal of The Korea Society of Computer and Information, Vol 17, No. 4, pp 163-172, April, 2012.

[5] Y. H. Kim and J. M. Mun, “A Study on the Development of Crime Prediction Program(CPP)”, Journal of The Korea Society of Computer and Information, Vol. 11, No. 4, pp. 221-230, December, 2006.

[6] Daejeon Metropolitan City, http://www.daejeon.go.kr/uic/index.do

[7] T. Troscianko, A. Holmes, J. Stillman, M. Mirmehdi, D. Wright and A. Wilson, “What Happens Next? The Predictability of Natural Behaviour Viewed Through CCTV Cameras, Perception, Vol. 33, No. 1, pp. 87-101, January, 2004.

[8] D. Grant and D. Williams, “The Importance of Perceiving Social Contexts When Predicting Crime and Antisocial Behaviour in CCTV Images”, Legal and Criminological Psychology, Vol. 16, No. 2, pp. 307-322, September, 2011.

[9] I. Darker, A. Gale, L. Ward and A. Blechko, “Can CCTV reliably detect gun Crime?”, IEEE Conference on Security Technology, pp 264-271. October, 2007.

[10] S. H. Bang, T. H. Kim and H. B. Cho, “A Study on the Applicability of Data Mining for Crime Prediction : Focusing on Burglary”, Journal of The Korea Society of Computer and Information, Vol.

19, No. 12, December, 2014.
[11] Seoul Statistics, http://stat.seoul.go.kr
[12] L. Cohen, and M. Felson, "Social Change and Crime Rate Trends: A Route Activity Approach," American Sociological Review, Vol. 44, No. 4, pp. 588-608, August, 1979.
[13] P. Brantingham, and P. Brantingham, "Environmental Criminology," Wavelend Press Inc, pp. 27-54, 1991.
[14] S. Y. Ko, K. O. Kim, Y. D. Jung and D. H. Choi, "A Study to Classify Serial Sex Offenders Based on Crime Scene Actions", The Korean Journal of Forensic Psychology, Vol. 1, No. 3, pp. 171-183, November, 2010.
[15] J. R. Quinlan, "Introduction of decision trees", Machine learning, Vol. 1, No. 1, pp. 81-106, March, 1986.
[16] J. R. Quinlan, "C4.5: Programs for Machine Learning", Elsevier, 2014.
[17] G. V. Kass, "An Exploratory Technique for Investigating Large Quantities of Categorical Data", Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 29, No. 2, pp. 119-127, 1980.
[18] L. Breiman, J. Friedman, C. Stone and R.A. Olshen, "Classification and Regression Trees", CRC press, 1984
[19] Y. D. Kim, C. H. Jun and H. S. Lee, "A New Classification Method Using Penalized Partial Least Squares", Journal of the Korean Data & Information Science Society, Vol. 22, No. 5, pp. 931-940, October, 2011.
[20] E. Malthouse, A. Tamhane and R. Mah. "Nonlinear Partial Least Squares". Computers and Chemical Engineering, Vol. 21, No. 8, pp. 875-890, April, 1997.
[21] S. Wold, M. Sjostrom and L. Eriksson, "PLS-regression: a Basic Tool of Chemometrics", Chemometrics and Intelligent Laboratory Systems, Vol. 58, No. 2, pp. 109-130, October, 2001.
[22] F. Famili, W. M. Shen, R. Weber and E. Simoudis, "Data Pre-processing and Intelligent Data Analysis", International Journal on Intelligent Data Analysis, Vol. 1, No. 1, pp. 1-28, March, 1997.
[23] S. Zhang, C. Zhang and Q. Yang, "Data Preparation for Data Mining", Applied Artificial Intelligence, Vol. 17, No. 5-6, pp. 375-381, November, 2003.

## Authors

Seung Hwan Bang received the B.S degree in Industrial and Management Systems Engineering from Khyung Hee University, in 2013. In addition, he has been a graduate student and a Ph.D. candidate at Pohang University of Science and Technology since 2013.
Mr. Bang joined as a graduate student of the Department of Industrial and Management Engineering at Pohang University of Science and Technology, Pohang, Korea, in 2013. He is interested in data analytics, data modeling and smart system.

Hyun Bo Cho received the B.S. and M.S. degrees in Industrial Engineering in Seoul National University, Korea, in 1986 and 1988. In addition, he received the Ph.D. degree in Industrial Engineering from Taxas A&M University, Texas, in 1993.
Dr. Cho joined the faculty of the Department of Industrial and Management Engineering at Pohang University of Science and Technology, Pohang, Korea, in 1993. He is currently a professor in the Department of Industrial and Management Engineering. He is interested in data analytics on smart system and smart manufacturing.