ISSN: 2508-7894 © 2020 KODISA & KAIA.
KJAI website: http://www.kjai.or.kr
doi: http://dx.doi.org/10.24225/kjai.2022.10.1.21

Classification Model and Crime Occurrence City Forecasting Based on Random Forest Algorithm*

Sea-Am KANG[1], Jeong-Hyun CHOI[2], Min-soo KANG[3]

Received: April 07, 2022. Revised: April 15, 2022. Accepted: June 12, 2022.

Abstract

Korea has relatively less crime than other countries. However, the crime rate is steadily increasing. Many people think the crime rate is decreasing, but the crime arrest rate has increased. The goal is to check the relationship between CCTV and the crime rate as a way to lower the crime rate, and to identify the correlation between areas without CCTV and areas without CCTV. If you see a crime that can happen at any time, I think you should use a random forest algorithm. We also plan to use machine learning random forest algorithms to reduce the risk of overfitting, reduce the required training time, and verify high-level accuracy. The goal is to identify the relationship between CCTV and crime occurrence by creating a crime prevention algorithm using machine learning random forest techniques. Assuming that no crime occurs without CCTV, it compares the crime rate between the areas where the most crimes occur and the areas where there are no crimes, and predicts areas where there are many crimes. The impact of CCTV on crime prevention and arrest can be interpreted as a comprehensive effect in part, and the purpose is to identify areas and frequency of frequent crimes by comparing the time and time without CCTV.
.

keywords : Random Forest Algorithm, Crime Rates, CCTV, Machine Learning Model

Major classifications : Machine Learning, Supervised Learning, Random Forest Algorithm

1. Introduction

With AlphaGo, interest in artificial intelligence (AI) is increasing, and research on core technologies such as machine learning and deep learning is also expanding. Machine running is very widespread in many areas, and search engines such as advertising, translation, spam blocking, games, voice and door recognition, and text mining are available. It is already being commercialized in industries such as intelligent robots and self-driving cars. We're making a difference in our daily lives. According to Wikipedia, machine learning is a field of artificial intelligence that develops algorithms and technology that enables computers to learn. Learning methods of machine learning are divided into supervised learning (supervised learning), unsupervised learning (unsupervised learning), and reinforcement learning (reinforcement learning). Guidance is a way to give a label (label) for the input data A, and to learn by itself so that you know that A is A. Non-guidance learning is classified into different groups by learning the difference between the two input data of A and

* This work was supported by the research grant of the KODISA Scholarship Foundation in 2022.

1 First Author, student, Department of Medical IT, Eulji University, Email: whtodka1@naver.com
2 Second Author. Director, LG Uplus corporation, Korea, Email: nevergiveup@lguplus.co.kr
3 Corresponding Author, Professor, Department of Big data medical convergence, Eulji University, Email: mskang@eulji.ac.kr

B, but each a or b in a state that does not know whether or not. Reinforcement learning is to learn the model in the direction of maximizing the reward (reward) and loss (penalty) depending on the result, although the answer to the given question is not clear. In this study, because the study to predict the crime by learning based on the given crime data and city information to proceed with the model development by guidance learning according to the purpose of the study and reviewed the prior research in this regard. Machine learning algorithms belonging to map learning include naive Bayesian (Naive Bayesian) model, lodge stick regression model, decision tree (Decision Tree) parent type, random forest model, neural network model, support vector machine (Vector Support Machine, SVM) model, etc., but in this study it is mainly used in predictive and classification studies, and its accuracy is high compared to other older types, known in previous studies.

Random forest also shows a high predictive power, has the advantage of easy to interpret the results than anything else. Domestic and international real estate price index forecast (Bae and Yu, 2018), flood and landslide vulnerability analysis (Lee, 2017), fire forecast (Guo et al., In addition to the 2016, Oliveira et al ., 2012), corporate credit rating prediction (Kim and Ann, 2016, Brown et al., 2012, Hajek el al., 2013), public housing public administration expenses (Jeong et al., 2017), and other economic and disaster-related environments, in addition to the economic and disaster minutes, are actively researching and researching the field of industry and energy in different vehicle environments (Oh et al, 2015).

2. Related Research

There was a variety of studies on the criminal side the related studies are largely divided into 1) one study on classifying crime types and 2) study on analyzing crime patterns through case-based reasoning techniques. This type of criminal classification and study uses the Crime and Communities data set provided by the UCI Machine Learning Repository to provide various mechanical learning models (decision tree), Support Vector Machine (supportive-machine), and Nair-neighborhood(Na-neigh-Na), beyond this Using unstructured data such as Edo text, a series of studies was conducted to determine the pattern type of crime. M. Gerber analyzes crime tracts tagged with GPS information and conducts research on crime in a specific area, (M.Alruil, 2016) conducts research to measure the details of criminal cases through news article analysis. However, existing crime-side analysis studies are performed only partially on the same data type to see the limits of accuracy. Research using a news article in an article also revealed many limitations in analyzing criminal elements due to the low efficiency of de ether processing and the low commercialization of calculated attribute values

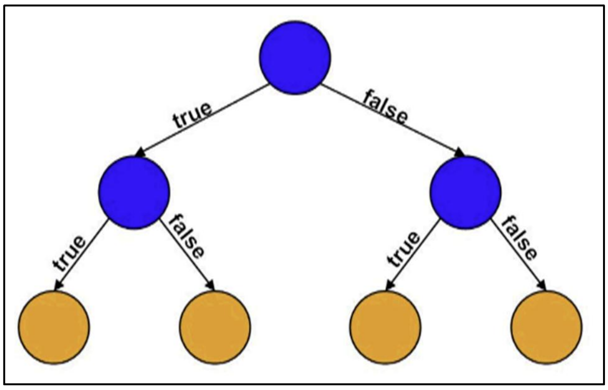


Figure 1: Decision Tree Schema

The decision tree allows branches to extend towards the independent variable that best describes the dependent variable. The determination of values in the decision tree is expressed in nodes, and the model is determined by how each node is located. where the top joint (node 1) is called the root joint, which contains all the data. MADI, such as nodes 2 and 3, is referred to as child muzzle, and words that are no longer branching, such as parenting, nodes 4, 5, 7, 8, 9 and so on, such as nodes 1, 2, 3 and 6. CCTV was selected as the target of the investigation considering that CCTV is the optimal condition for analyzing the actual impact of crime because of its high frequency of crime and narrow density compared to the area

In J M Park's “A Study on the Prediction Model of Crime Frequency Using Big Data”, a model for predicting the occurrence of crime was proposed by using actual crime data and Google trends for crime prediction. He also conducted correlation analysis by dividing into various categories to analyze the association between crimes. As a result, there was a strong connection between murder, theft and violence. Through this, it predicted the number of crimes, analyzed the association, and helped to establish a crime prevention policy (Park, 2018). In “Predicting Crime Risky Area Using Machine Learning” by S Y HEO and two others, we studied an automated system that visualizes results from crime prediction. Among the supervised learning models of machine learning algorithms, a crime prediction model was constructed and compared and analyzed using the Decision tree model, Random forest model, and Support Vector Machine model, which are known to have high accuracy and are used in various fields. As a result, a decision tree model with high predictive power due to low mean square root error was selected as the optimal model. Based on this, scenarios were created for theft and violent crimes, predicted and visualized in the form of a map (Heo et al., 2018). According to Yoon et al.(2014), "Building Crime Prevention System Utilizing Big Data (I)", it is not simply collecting and analyzing personal information, but collecting and analyzing spatial, situational, and temporal information, so that certain crimes occur in any region at any time It can be predicted statistically. In that case, violent crimes such as murder or rape and national crimes such as terrorism can be more effectively prevented in advance. In addition, by analyzing the patterns of criminals, it is possible to predict the escape route after causing a crime, or predict the habituality of criminals and the likelihood of recidivism (Yoon, et al., 2014). Sanghyuck You and Minsoo Kang are writing the following thesis. He wrote a paper on a study was conducted to find the main factors to Pima Indians Diabetes based on machine learning. It used Support Vector Machine (SVM), Decision Tree, and correlation analysis to discover three critical factors that predict Pima Indians diabetes with 70% accuracy(You & Kang, 2020).

3. Algorithm for Predicting Criminal

3.1. Data Set

Table 1: Example of a crime rate data set

Criminal Classification	Seoul-jungno	Seoul-jung-gu	Seoul-yongsan
theft	1488	1858	1133
stolen goods	12	11	8
scam	1761	2206	1535
seizure	664	709	693

In this paper, the values of the dependent variable (violent crime) and the independent variable (CCTV) are included. The dependent variable uses correlation analysis to select a highly relevant property out of 16 properties. Unnecessary attributes such as crime motivation, date, time, and place were removed to identify only the frequency and region of the crime through preprocessing. In the selected attributes, violent crime is murder, murder, sex crime, theft, burn, assault, fraud, multiple cities.

3.2. Random Forest Algorithm

This study suggests a crime type classification model using random forest techniques. Random post-base method is a type of decision-making tree technique, because it is possible to solve the overfitting (Overfitting) problem of the model because it forms a model by integrating the grain of each crystal tree, because it consists of crystal trees that do not correlate with each other, there is a strong advantage for the noise of the training data. As the first step in creating a proposed crime type classification model, a classification model is created for criminal record data using random post techniques. In the second step, the urethra of the crime device lock property is calculated within the generated classification model to extract the upper N urethra property as an experiment. The third step analyzes the news articles collected to determine the rate of the analysis of the crime type of the extracted property. In the fourth step, the property value of the crime history data is adjusted by multiplying the value of each property calculated. The final step is to regenerate the new crime type classification model using the adjusted property values.

Random Forest is after creating a predictor of the shape of the decision tree by extracting the data from any number of inputs, because the generated models are combined through the ensemble (ensemble) technique to create the final model predictors are selected independently from the data randomly selected from the initial dataset. Because there is the randomness of the variable selection, the predictor variable is selected by log (N+2) of the total number of N variables, and the average square error indicating the accuracy of the random forest algorithm is equal to Expression 1.

$$€ = (v_{observed} - v_{response})^2 \quad (1)$$

€ : Average square error of the algorithm.

$v_{observed}$: The variation value of the training data used in the algorithm.

$v_{response}$: Variable values shown in the prediction results.

In addition, the average value of the reaction predicted in each tree is the same as the following expression 2.

$$S = \frac{1}{K} \sum K^{th} v_{\,response} \qquad (2)$$

Where is the predicted value of the random forest algorithm, K is applied to each tree of the random forest. As many decision trees as you have decided in advance based on existing data and variables, you have learned to obtain the most information from each node (Lee, 2017)

4. Research

<Table 1> shows the experimental environment used to evaluate the performance of the model for predicting violent crimes using the date of occurrence of the crime. And the dataset consists of 587 sets, which were tested separately with training data and 20% test data to avoid overwriting. The experiment was designed for Microsoft's zurel and began using machine learning with Microsoft Azure Machine Learning Studio. I will remove data that is not used in the experiment and participate in the experiment.

This study uses data through normalization to improve accuracy. Normalized data has applied with random forest of each violent crime numerical value and cctv numerical attribute.

Table 2 : Selected Properties

Attribute	Properties Description
Murder	The number of murders of each city.
Murder spree	The number of murders spree of each city.
Sex crime	Sexual offenses in each city.
Thief	Theft offenses in each city.
Burn	Arson crimes in each city.
Assault	Assault crimes in each city.
Scam	Fraud occurs for each city.
CCTV	The number of closed-circuit televisions of each city.

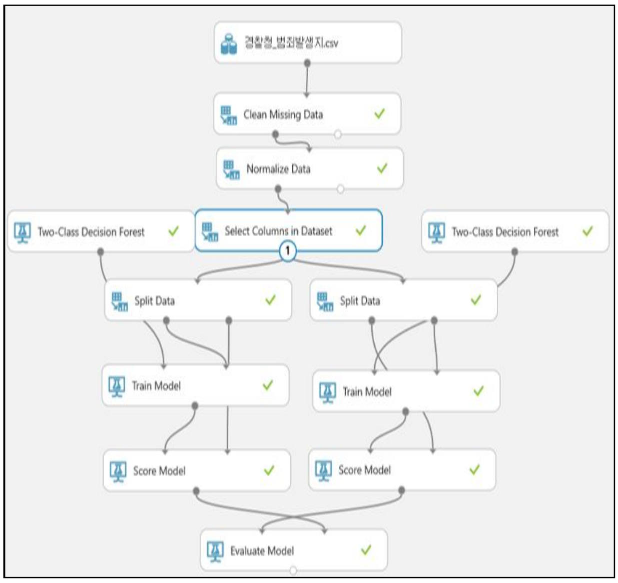


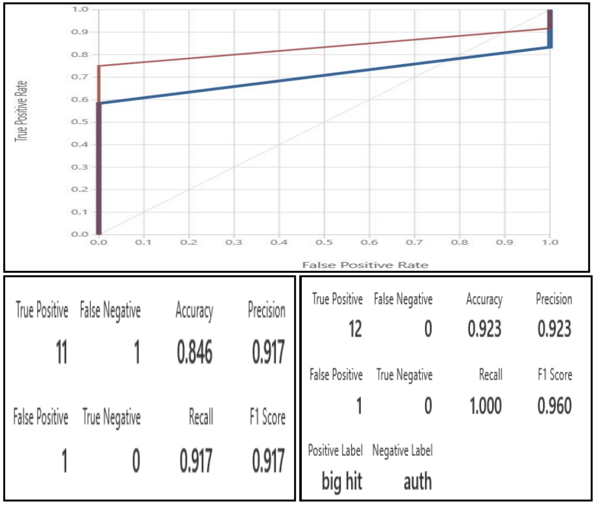
Figure 2: Experiment of Microsoft Azure Machine Learning Studio.

4.2. Experimental Process

Using data from 17 cities, this experiment uses TWO-Class Decision Forest with Normalize data to optimize data from violent crime data and data from each attribute, and extends the data set to 1185 datasets. To improve the accuracy of the study, each training eliminated missing data values. This experiment was conducted through ensemble learning and improved performance and accuracy. Data has been partitioned to prevent overfitting. In order to confirm the relationship between violent crime and CCTV, the left train model was trained as a violent crime column and CCTV was installed in the right train model. The relationship between violent crime and CCTV will be confirmed through the data trained.

4.3. The Result of an Experiment

The experimental results are the same as in <Figure 3>, and the evaluation model is visualized and expressed as a numerical value. The characteristics of the prediction accuracy were 84.6%. And the second train was 46.2 percent accurate.



True Positive	False Negative	Accuracy	Precision
11	1	0.846	0.917
False Positive	True Negative	Recall	F1 Score
1	0	0.917	0.917

True Positive	False Negative	Accuracy	Precision
12	0	0.923	0.923
False Positive	True Negative	Recall	F1 Score
1	0	1.000	0.960
Positive Label	Negative Label		
big hit	auth		

Figure 3: Result of Two-Class Multi Decision Forest

The first prediction is the probability of a crime in the absence of CCTV. And the second prediction is the probability of a crime in the presence of CCTV. In addition, Jeju Island is a place where violent crimes occur compared to the population. The accuracy and reproducibility of the study is 91.7%. The second train model has an accuracy of 92.3%, accuracy of 92.3%, and recovery rate of 100%. The F1 score value is 91.7% higher than the number of attributes. Like the figures, high level of accuracy could be confirmed, and different parts could be confirmed. According to the experimental results, the crime rate in areas without CCTV was 84.6%, 38.4% higher than in areas without CCTV. It has been confirmed that there is a clear correlation between CCTV and crime rate in terms of high accuracy and precision. Jeju Island has been identified as the area where violent crimes occur the most compared to the population

5. Conclusions

The quality of life is improving due to rapid economic growth and the development of science and technology, but various social problems are also rapidly increasing, and crime is a serious problem, especially. Crime problems occurring in recent years are becoming more intelligent and diversified than before, and violent crimes are increasing rapidly. Crimes that have already occurred are irreversible, and even if a criminal is punished for the crime he committed, the pain of the victim and their family cannot be healed.In this paper, we predicted the characteristics of violent crime using random forests by predicting the national crime data set and the data of places with many crimes, which identified the relationship between CCTV and crime rate. The algorithm used a two-step decision forest with Microsoft Azure Studios. And this property used 17 properties. Since it was difficult to achieve good results, we removed the missing parts from the dataset and performed machine learning through ensemble learning. According to the forecast, the crime rate was more than 40% lower in places with CCTVs than in places without CCTVs, and Jeju Island had the highest crime rate in Korea, but crime surged in proportion to the CCTV population. According to the experimental results, the crime rate in areas without CCTV was 84.6%, 38.4% higher than in areas without CCTV. It has been confirmed that there is a clear correlation between CCTV and crime rate in terms of high accuracy and precision. Jeju Island has been identified as the area where violent crimes occur the most compared to the population. However, since it is a data value as of the end of 2017, the set value is limited to 2021. For more accurate data, I would like to investigate the new data at the National Statistical Office and increase the accuracy.

References

Chung, Y. S, Kim, J. M., & Park, K. R. (2012). A study of improved ways of the predicted probability to criminal types, *Journal of The Korea Society of Computer and Information, 17*(4), 12-21.

Heo, S. Y., Kim, J. Y., & Moon, T. H. (2018). Predicting Crime Risky Area Using Machine Learning, *Journal of the Korean Association of Geographic Information Studies, 21*(4), 34-43.

Joo, I. Y. (2012). A Case Study on Crime Prediction using Time Series Models, *Korean security science review, 30*, 139-169.

Kim, K. P., & Song, S. W. (2018). A Study on Prediction of Business Status Based on Machine Learning, *Korean Journal of Artificial Intelligence, 6*(2), 23-27.

Kang, M. S., Kang, H. J., You, K. B., Lim, C. H., & Choi, E. S., (2018). Getting started with machine learning with Microsoft's Edge Machine Learning Studio, *Hanti Media, 1*(1), 89-132.

Kim, M. S., & Kang, T. W. (2018). Proposal and Analysis of Various Link Architectures in Multilayer Neural Network, *Journal of KIIT, 16* (4), 11-19.

Kim, O. H. (1999). *Case analysis using neural network data analysis technique,* Domestic Master's Thesis Graduate School, Ewha Womans University, Seoul.

Park, J. H. (2014). *Comparing performances of logistic regression and decision tree for classifying infection risk with patients in chemotherapy,* Domestic Master's Thesis Graduate School, Kyunghee University, Seoul

Park, J. M. (2018). *A Study on the Prediction Model of Crime Frequency Using Big Data*, Domestic doctoral dissertation Graduate School, Kongju National University, Chungcheongnam-do.

Park, J. Y., Chae, M. S., & Jung, S. K. (2016). Classification Model of Types of Crime based on Random-Forest Algorithms and Monitoring Interface Design Factors for Real-time Crime Prediction, *KIISE Transactions on Computing Practices, 22*(9), 455-460

Tak H. S., Park, J. H., Jeong, J. S., & Yoon, J. W. (2015). Building Crime Prevention System Utilizing Big Data(II), Seoul, Korea: Korean Institute of Criminology.

Yoon, H. S. Jeon. H. W., Yang, C. S. Kim B. S., & Kim K. B. et al., (2014). *Building Crime Prevention System Utilizing Big Data(Ⅰ),* Seoul, Korea: Korean Institute of Criminology. Doi: https://doi.org/10.23000/TRKO201500001310

You, S. H., & Kang, M. S. (2020). A Study on Methods to Prevent Pima Indians Diabetes using SVM. *Korean Journal of Artificial Intelligence 8*(2), 7-10.