ISSN: 2508-7894 © 2021 KODISA & KAIA.
KJAI website: http://www.kjai.or.kr
doi: http://dx.doi.org/10.24225/kjai.2021.9.1.15

A Study on Crime Prediction to Reduce Crime Rate Based on Artificial Intelligence

Kyoung-Sook KIM[1], Yeong-Hoon JEONG[2]

Received: February 21, 2021. Revised: March 14, 2021. Accepted: June 05, 2021.

Abstract

This paper was conducted to prevent and respond to crimes by predicting crimes based on artificial intelligence. While the quality of life is improving with the recent development of science and technology, various problems such as poverty, unemployment, and crime occur. Among them, in the case of crime problems, the importance of crime prediction increases as they become more intelligent, advanced, and diversified. For all crimes, it is more critical to predict and prevent crimes in advance than to deal with them well after they occur. Therefore, in this paper, we predicted crime types and crime tools using the Multiclass Logistic Regression algorithm and Multiclass Neural Network algorithm of machine learning. Multiclass Logistic Regression algorithm showed higher accuracy, precision, and recall for analysis and prediction than Multiclass Neural Network algorithm. Through these analysis results, it is expected to contribute to a more pleasant and safe life by implementing a crime prediction system that predicts and prevents various crimes. Through further research, this researcher plans to create a model that predicts the probability of a criminal committing a crime again according to the type of offense and deploy it to a web service.

Keywords : Machine learning, Microsoft Azure, Crime Prediction, Multiclass Logistic Regression, Multiclass Neural Network

Major Classification Code : Artificial Intelligence, Machine learning, Crime Prediction

1. Introduction

In Korea, people's quality of life has improved as the trend of advanced industrialization, information, and specialization accelerated due to the continuous and rapid economic growth and the development of science and technology. Still, various problems such as poverty, unemployment, and crime are constantly occurring. Among these social problems, crime problems have become more intelligent, advanced, diversified, and broadened, making it difficult to predict and cope with crimes before they occur.

Most of the studies on the occurrence of crimes that have been conducted so far consist of a number of studies focusing on the past, such as the increase or decrease in the crime rate compared to the previous year, and studies such as predicting the occurrence of crime and preparing countermeasures are insufficient. In order to effectively cope with such diversified and intelligent crimes, it is important to analyze the current crime trends and prepare systematic and comprehensive countermeasures to meet the new trends of crime. (Joo, 2012) For all crimes, it is important to prevent them before they happen, rather than to deal with them well after they happen. What has already happened is irreversible, and the grief of the victims and their families caused by it cannot be compensated with anything. In order to prevent such damage, studies are

1 First Author, Student, Department of Medical IT, Eulji University, Korea, Email: lvx3xvl0794@naver.com
2 Corresponding Author, Vice President, Consulting & Project Management unit, LG Uplus Corp., Email: gohawk@lguplus.co.kr

mainly conducted today to prevent crime by analyzing psychological and social factors (Chung et al., 2012). From this point of view, there is no more effective and practical approach to crime prevention than knowing in advance when, where, and what crimes will occur and who will commit them and why. In line with this trend, in advanced countries such as the United States and the United Kingdom, police activity processes are based on various information and data analysis such as SMART Policing, Predictive Policing, and Data-Driven Approaches to Crime and Traffic Safety (DDACTS) are attracting attention. Most of the existing crime analysis methods only describe the past crime situation. Still, in the crime predictive analysis method, regression analysis, risk area analysis, data mining (classification technique, cluster technique), geographic profiling, hot spot analysis, proximity-repeated modeling, It is possible to make more accurate and effective crime predictions by rapidly and effectively processing a wide variety of data such as time and space analysis. Also, it is possible to predict criminals and victims, indicating the likelihood that individuals or groups will suffer crime. This is a method of analyzing targets in danger zones or targeted by crime type, including violence, robbery, and is not much different from the techniques used in criminal and crime prediction (Tak et al., 2015). Various research cases to effectively prevent crime and respond to crime occurrence can be broadly divided into several categories as follows. Studies focused on historical criminal event records and related data analysis, studies focused on criminal behavior psychology, and examples of crime prediction system services that can observe criminal events and view relevant statistics based on crime prediction techniques. Among these cases, research on data analysis technology and crime prediction system plays an essential role in reducing the incidence of crime thanks to the recent radical development of machine learning technology (Park et al., 2016). Therefore, in this thesis, to predict crime in advance and prepare countermeasures, the criminal records that occurred in the past were analyzed. Through the crime data in Baltimore between 2011 and 2016, various types of crimes and tools used in the crimes were analyzed and predicted. For this, the data world's crime data was used, and the crime types and tools were expected using Multiclass Logistic Regression and Multiclass Neural Network among various machine learning algorithms.

2. Literature Review

2.1. Machine Learning tool

Azure is Microsoft's cloud computing platform that has been in service since 2010. The Azure platform provides more than 600 services, and you can create a model through Azure Machine Learning Studio, easily build a web service, and apply it to various devices. In addition, unlike existing cloud platforms and machine learning libraries and tools, it provides an easy-to-access GUI environment in consideration of user convenience. At Microsoft Ignite, It announced Azure Machine Learning designer's general availability, the drag-and-drop workflow capability in Azure Machine Learning studio, which simplifies and accelerates building, testing, and deploying machine learning models for the entire data science team, from beginners to professionals (azureml.net). Azure Machine Learning Studio natively supports data input, output, and visualization, and representative machine learning algorithms that data scientists love are prepared. Unlike existing machine learning tools and libraries, it has the advantage of being able to easily create a model by dragging and dropping blocks. In addition, scripts written in R and Python languages can be inserted and utilized in block form, and the results can be checked through visualization. Thanks to this easy structure, anyone who knows how to use it can easily create and deploy predictive models (Kang et al., 2018).

2.2. Machine Learning

Machine learning is a field of artificial intelligence, the study of computer algorithms that automatically improve through experience. In other words, it is the field of developing algorithms and technologies that enable computers to learn. Also, Machine learning can also use algorithms to analyze data, learn through analysis, and make decisions or predictions based on learning. Testing of machine learning models is done through validation errors on new data, not theoretical tests to validate null hypotheses. Machine learning is an iterative approach that learns from data, making it easy to automate. Then, iterate through the data and discover strong patterns. Interest in machine learning can be seen in advances in technologies such as Bayesian analysis and data mining. Increasing the volume and variety of available data, more powerful analytics, lower storage costs, and reduced analytics costs are driving continued interest in machine learning. (sas, 2021) Machine learning is largely classified into unsupervised learning and supervised learning, and learning includes classification algorithms, regression algorithms, and deep learning (Kim & Song, 2018).

2.3. Machine Learning Algorithms

2.3.1. Multiclass Logistic Regression

Regression analysis can be used for statistical

predictions such as data changing over time, certain effects, hypothetical experiments, and causal modeling. Logistic regression is a well-known method of statistics used to predict the probability of an outcome and is heavily used in classification tasks (Wikipedia, 2020). Logistic regression analysis, like linear regression analysis, is a statistical technique that analyzes the causal relationship between independent and dependent variables, and is especially used when the dependent variable has only binary categories (Baek, 2013). In other words, logistic regression analysis is a method of regression analysis that can be used to estimate the relationship between dependent variables having bivariate values such as nominal scales and independent variables. Logistic regression analysis is mainly used to model a model that predicts and analyzes to which group individual observations can be classified when objects are divided into two or more groups (Park, 2014).

2.3.2. Multiclass Neural Network

Neural network theory has been studied to simulate the process of storing and learning information by biological brain neural networks. Such research has become a cornerstone of research in many different fields, including social science, psychology, cognitive science, and network system theory. Neural network theory is a computational model in which connection weights are assigned in an appropriate form according to the connection structure to be trained. The lengths between the learned weights and input vectors are calculated. The output is classified into relative values. The learning ability of neural network theory can also serve as a significant computational model in solving philosophical problems such as classification and recognition, optimization problems, and prediction. The most commonly used multilayer neural network is a feed-forward neural network composed of several layers, that is, layers, and the learning algorithm uses a backpropagation algorithm. A multilayer neural network consists of an input layer, a hidden layer, and an output layer (Kim, 2018). The neural network model has the advantage that it is easy to deal with problems from various angles and leads to good results even in complex domain data. However, there is a disadvantage in that all input and output values must be defined in the range between 0 and 1, and the process of deriving the result is not clear (Kim, 2000).

3. Related Research

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 H S Yoon and many other "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).

In this paper, we intend to analyze and predict crime data using Multiclass Logistic Regression and Multiclass Neural Network among machine learning algorithms. Among various crime data, tools and crime types used in crime were analyzed. Through this, a crime prediction system is developed to help establish crime prevention and safety measures.

4. Data Set

4.1. Dataset description

The data set used in this paper was taken from the City of Baltimore Open Data by referring to the dataworld site.

This data set records crime data in the city of Baltimore between 2011 and 2016.

In this paper, we used a victim-based crime data set to predict crime types and the tools used in the crime. This dataset consists of 12 columns and about 286000 rows.

As can be seen in Table 1, 12 variables were used to predict the crime type and the tools used in the crime.

Table 1: Victim_Based_Crime_Data.csv Data Description

Variable Name	Variable Content
CrimeDate	The date the crime occurred
CrimeTime	The time the crime occurred
CrimeCode	The code that classified the crime
Location	Where the crime occurred
Description	Type of crime
Inside/Outside	Whether the crime took place indoors or outdoors
Weapon	Tools used in crime
Post	Post
District	Area where the crime occurred
Neighborhood	Neighborhood
Location 1	The latitude and longitude where the crime occurred
Total Incidents	Total number of accidents

As can be seen from Table 2, in the Victim Based Crime Data.csv data set, the tools used for crime consist of 4 and the crime types consist of about 14. For these two variables, the crime type and the crime tool were predicted using Multiclass Logistic Regression and Multiclass Neural Network algorithm.

Table 2: Description of the variable to be analyzed

Variable	Example
Weapon	FIREARM, HANDS, KNIFE, OTHER
Description	AGG. ASSAULT, ASSAULT BY THREAT, AUTO THEFT, COMMON ASSAULT, LARCENY, LARCENY FROM AUTO, RAPE, ROBBERY–CARJACKING, ROBBERY–COMMERCIAL, ARSON, ROBBERY–RESIDENCE, BURGLARY, ROBBERY–STREET, SHOOTING

5. Experimental model

5.1. Experimental model process

The process of constructing a crime prediction model through the Microsoft Azure Machine Learning Studio program is as follows.

First, for Victim Based Crime Data.csv, take the Select Columns in Dataset block from the navigation palette and select all 12 properties through the Launch Colum Selector. After that, it was preprocessed using Clean Missing Data. By setting the Minimum missing value to 0 and the Maximum missing value to 1, the missing value is cleaned up even if there is only one missing value. In addition, the Cleaning Mode was set to Remove entire row, and the row with the missing Species attribute was deleted.

Next, for the existing data set to be applied to the machine learning algorithm through Split Data, the test data was set to 30% (0.3) and the training data was set to 70% (0.7), and the random seed value was randomly set to separate the data. Also, in the Train Model for Multiclass Logistic Regression and Multiclass Neural Network algorithm, the desired predicted values, Weapon and Description, were selected through Launch Column Selector.

Among the machine learning algorithms, the Multiclass Logistic Regression algorithm was selected. This algorithm used a single parameter, and the regularization weight and memory size were set to 1 and 20, respectively. And the Multiclass Neural Network algorithm was selected. This algorithm used Single Parameter and Fully-connected case, and the Number of hidden nodes and Learning Rate were set to 100 and 0.1, respectively.

6. Experimental model evaluation

6.1. Weapon prediction

Figures 1 and 2 show that the Multiclass Logistic Regression algorithm has higher accuracy, precision, and recall for weapon prediction. In addition, when comparing the actual and predicted values, FIREARM, HANDS, and OTHER are predicted with a relatively high probability with a probability of more than 70%. However, it can be seen that the case of KNIFE as the criminal tool is very low with a probability of 13.3%.

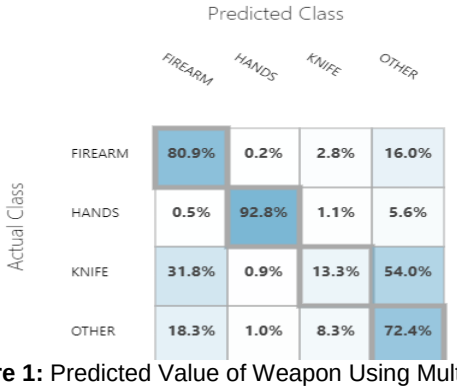

Figure 1: Predicted Value of Weapon Using Multiclass Logistic Regression

The results of the weapon evaluation model using the Multiclass Logistic Regression algorithm are as follows. Overall accuracy: 0.789046, Average accuracy: 0.894523, Micro-averaged precision: 0.789046, Macro-averaged precision: 0.64088, Micro-averaged recall: 0.789046, Macro-averaged recall: 0.64856.

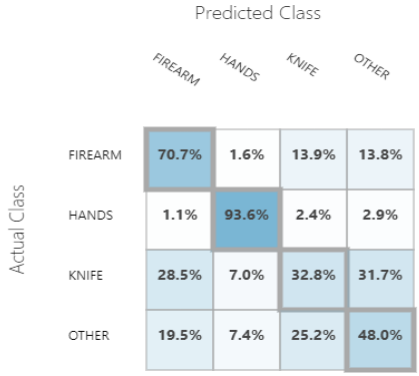

Figure 2: Predicted Value of Weapon Using Multiclass Neural Network

The results of the weapon evaluation model using the Multiclass Neural Network algorithm are as follows. Overall accuracy: 0.754892, Average accuracy: 0.877446, Micro averaged precision: 0.754892, Macro-averaged precision: 0.608943, Micro-averaged recall: 0.754892, Macro-averaged recall: 0.612495.

6.2. Description prediction

Figures 3 and 4 show that the Multiclass Logistic Regression algorithm has higher accuracy, precision, and recall for Description prediction. Also, if you compare the actual and predicted values, AGG. ASSAULT, COMMON ASSAULT, and SHOOTING are predicted with a relatively high probability with a probability of over 70%. However, the type of crime is predicted to be very low with 19.5% for HOMICIDE, 1.6% for RAPE, and 14.8% for ROBBERY – CARJACKIN.

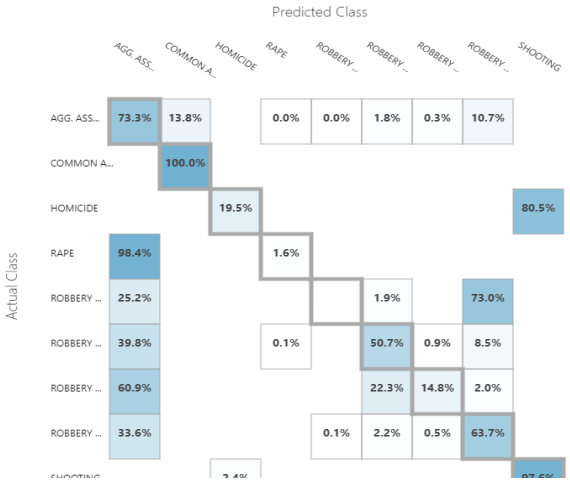

Figure 3: Predicted Value of Description Using Multiclass Logistic Regression

Figure 4: Predicted Value of Description Using Multiclass Neural Network

The result of the evaluation model of Description using Multiclass Logistic Regression algorithm is as follows. Overall accuracy: 0.818921, Average accuracy: 0.95976, Micro-averaged precision: 0.818921, Macro-averaged precision: 0.620467, Micro-averaged recall: 0.818921, and Macro-averaged recall: 468079.

The result of the evaluation model of Description using Multiclass Neural Network algorithm is as follows. Overall accuracy: 0.76762, Average accuracy: 0.94836, Micro-averaged precision: 0.76762, Macro-averaged precision: 0.552257, Micro-averaged recall: 0.76762, Macro-averaged recall: 0.518997.

7. Conclusion

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. Therefore, in this paper, research was conducted to prevent and respond to crimes by predicting crimes based on artificial intelligence. In order to predict the occurrence of crime in advance and prepare countermeasures, we analyzed the crime records that occurred in the past. Using crime data from the city of Baltimore between 2011 and 2016, we analyzed and predicted various types of crimes and tools used in the crimes. Among the various machine learning algorithms, we predicted crime tools and crime types by using the Multiclass Logic Regulation, which can increase model stability through fast learning time and unnecessary variable selection, and the Multiclass Neural Network, which can create accurate and highly complex models. In predicting the types of crime and the tools used in the crime, Multiclass Logistic Regression showed higher accuracy, precision, and recall than Multiclass Neural Network. Especially when comparing the actual and predicted values of the tools used in crime, HANDS was predicted with the highest probability of over 90%. Also, when comparing the true and predicted values of the crime type, COMMON ASSAULT was predicted with the highest probability. These prediction results can help people live comfortably and safely by implementing a crime prediction system that predicts and prevents various crimes. And it can be used to prepare preventive measures by recognizing the importance of preventing crimes rather than coping well after crimes. Through further research, this researcher plans to create a model that predicts the probability of a criminal again committing a crime according to the type of crime and deploy it to web services.

References

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

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.

Tak H. S. (2015). Building Crime Prevention System Utilizing Big Data(II), *Korean Institute of Criminology*, 1, 703.

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.

Kang, M. S., Kang, H. J., Yoo, K. B., Ihm, C. H., & Choi, E. S. (2018). *Getting started with Machine Learning using Azure Machine Learning studio*. Seoul, Korea: Hanti media.

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.

Dataworld. (2016). City of Baltimore Crime Data. Retrieved October, 2020 from https://data.world/data-society/city-of-baltimore-crime-data

Wikipedia. (2020). Retrieved October, 2020 from https://en.wikipedia.org/wiki/Multinomial_logistic_regression

Baek, J. R. (2013). Development of accident prediction model for military aircraft by using logistic regression, *Domestic Master's Thesis Graduate School, Yonsei 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.*

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

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.

Yoon, H. S et al., (2014). Building Crime Prevention System Utilizing Big Data(Ⅰ), *Korean Institute of Criminology.*

Microsoft Azure Machine Learning Studio (classic) (azureml.net).

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.