Crime Analysis and Forecasting Using Machine Learning

Hilal ARSLAN[1*], Aslınur DOLUCA HOROZ[2]

[1]Software Engineering, Ankara Yıldırım Beyazıt University, Turkey
ORCID No: https://orcid.org/0000-0002-6449-6952
[2] Computer Engineering, Ankara Yıldırım Beyazıt University, Turkey
ORCID No: https://orcid.org/0009-0008-9808-9168

Keywords

Machine Learning, Crime Forecasting, Crime Analysis, Comparison of Machine Learning Techniques

Abstract

Crime is one of the most common and alarming attitudes all over the world. The number of crimes is increasing day by day, which affects the life of the people negatively. Thus, analyzing and preventing crime is a crucial task. With the advent of new technologies, machine learning methods have achieved admirable performance in all fields of crime prediction. Accurate prediction of crime that may arise shortly can help police units prevent crime before it happens. The ability to forecast any crime based on location may aid in obtaining useful information regarding strategic perspective. Therefore, the analysis and prediction of the crime are significant in identifying and diminishing future crimes. In this study, we apply various machine learning algorithms to predict where crime will take place to prevent future crimes as well as diminish crime rates in society. For this purpose, we perform decision trees, k-nearest neighbor, support vector machine, neural networks, logistic regression, and ensemble learning methods. The dataset used in this study includes 49030 samples with 12 attributes including the borough of arrest, the date of the criminal's arrest, offence description, sex, age as well as race information and coordinates. Historical data on different crimes that took place in 2019 in New York State is used. When the results are evaluated in terms of time and accuracy, decision tree methods achieved higher performance in 2 seconds with an accuracy of about 99.9%. To sum up, awareness regarding risky locations aids police units in predicting future crimes in a definite location.

Research Article
Submission Date : 08.06.2023
Accepted Date : 12.06.2023

1. INTRODUCTION

Crime is a human behaviour that expresses the deliberate violation (caste) of the legal values that need to be protected in terms of the continuation of the social order or carelessness (negligence) against the rules to protect these values. Criminology or crime science is a scientific discipline that explains crime, examines the causes of criminal behaviour, and deals with the prevention of crime and the fight against crime. Criminology examines criminality, the consequences of the crime, and the effects of crime on the victim and society.

[*]Resp author; e-mail:hilalarslan@aybu.edu.tr

The existence of crime in a society is enough to make people in that community uneasy. For this reason, police units have a great responsibility to prevent crime since it may increase that intense crime is being committed in cases the lack of sufficient police force. It is crucial to predict crimes before they are committed to take precautions accordingly. At this point, machine learning can be effectively used in crime analysis and prediction, which provides parallelism with criminology. In this paper, we perform and compare various types of machine learning techniques to predict crimes. Our methods include decision stump, hoeffding tree, J48, logistic model trees, random tree, random forest, and REP tree. Furthermore, we discuss the results of Naive Bayes classifier as well as function-based methods which include support vector machine, simple logistic regression, and artificial neural networks. Finally, we discuss the results of lazy learning methods which include KStar, locally weighted learning, k-nearest neighbor, and ensemble methods. The other sections are organized as follows: Section 2 includes related studies including crime analysis. In Section 3, we introduce the dataset. Section 4 overviews machine learning methods. Section 5 summarizes the results and Section 6 concludes the study.

2. LITERATURE REVIEW

Jain et al. (Jain, Sharma, Bhaita, and Arora, 2017) used k-means clustering algorithms to detect crime-prone areas. Their method was considered to have promising value in the current complex crime scenario and can be used as a tool in crime detection and prevention by the police and law enforcement. Chun et al. (Chun, Avinash Paturu, Yuan, Pathak, Atluri, & Adam, 2019) aimed to find out whether a person would execute a crime in the near future and the level of seriousness of this crime. They used a deep learning method and their method reached an accuracy of 99.7%. Zhang et al. (Zhang, Liu, Xiao, & Ji, 2020) performed and compared the random forest algorithm, the k-nearest neighbor algorithm, and support vector machine algorithms for crime prediction. They also performed the long short-term memory method (LSTM) which is a deep learning method to determine the most effective one. Their study achieved the best results with an accuracy of 59.9% when the LSTM method was used. Llaha et al. (Llaha, 2020) also used machine learning methods to analyze crime and take prevention. As a result of their study, it was observed that the decision tree method, which is one of the methods applied in classifying crime data, reached the most efficient result with an accuracy of 76%. Safat et al. (Safat, Asghar, & Gillani, 2021) used different machine learning algorithms to further analyze accurate crime prediction. In their study, different algorithms were applied and their efficiency was compared. As a result of their study, the XGBoost achieved maximum accuracy (94%). Tamir et al. (Tamir, Watson, Willett, Hasan, & Yuan, 2021) applied k-nearest neighbors, AdaBoost, random forest, and neural network methods to predict possible crimes and their locations. Their experimental studies showed that the neural network method with an accuracy of 90.77% achieved better results than other machine learning methods concerning different performance criteria.

3. DATA AND PREPROCESSING

The historical data on 62 different crimes that took place in 2019 in New York State was acquired from NYC open data (NYPD, 2020). The dataset includes 49,030 instances and 12 attributes. The features and their explanations are given in Table 1. We choose an equal number of instances for each class to obtain a balanced dataset. Furthermore, we removed the entries with missing values.

Table 1. Dataset description

Feature Name	Description
ARREST KEY	ID value that differs for each arrest
ARREST DATE	The date of the criminal's arrest
KY CD	Crime classification code for each crime type
OFNS DESC	Description of the crime classification code
LAWCATCD	Grade of offense: F(Felony), M(Misdemeanor), V(Violation), I(Infraction)
ARREST BORO	District of the arrest. B(Bronx), S(Staten Island), K(Brooklyn), M(Manhattan), Q(Queens)
ARREST PRECINCT	Police station where the arrest took place
AGE GROUP	Perpetrator's age within a category 1(-18), 2(18-24), 3(25-44), 4(45-64), 5(65+)
PERP SEX	Perpetrator's sex description
PERP RACE	Perpetrator's race description
XCOORDCD	X-coordinate the place where the crime took place
YCOORDCD	Y-coordinate the place where the crime took place

4. METHODOLOGY

This section summarizes the machine learning methods we implement for crime prediction. The methods are implemented using Waikato Environment for Information Analysis (WEKA) (Witten, Frank, Trigg, Hall, & Cunningham, 1999) which is a comprehensive application written entirely in Java and incorporates many machine learning and data mining methods developed by the University of Waikato.

Decision tree techniques are widely used for classification. In these methods, the data is classified as root nodes, internal nodes, and leaf nodes as if it were a tree (Song, & Ying 2015). In this study, the decision tree methods we implement are Decision Stump, Hoeffding Tree, J48 (C4.5), LMT, Random Forest, Random Tree, and REP Tree. On the other hand, Bayesian classifiers assign a particular instance defined by the feature vector to the most probable class (Rish, 2001). In this study, we performed Bayes Net, Naive Bayes, and Naive Bayes Updateable methods. We also performed function-based methods including Logistic Regression, Support Vector Machine, and Artificial Neural Networks. Furthermore, we apply lazy learning methods. The K-nearest neighbor (KNN) classifier (Dasarathy, 1991) is known as the basis of many lazy learning algorithms. The KNN method stores the entire training set and defers all efforts to inductive generalization until classification time (Wettschereck, Aha, & Mohri, 1997). In this study, the lazy learning methods we implement are KNN, K-Star, and LWL (Locally Weighted Learning) methods. Finally, we performed ensemble methods which are machine learning methods that create a set of classifiers and then make predictions by classifying new data points by taking the (weighted) vote of their predictions (Dietterich, 2000). In this study, the ensemble methods we implement are AdaBoost, Bagging, LogitBoost, Multi Scheme, Random Committee, Random SubSpace, Stacking, and Vote.

5. RESULTS

Machine learning techniques performed in this study are compared and evaluated using accuracy, precision, recall, and F-measure metrics. We also applied 10-fold cross-validation to avoid overfitting. First of all, we evaluate the results of the tree-based methods. The accuracy, precision, recall, and F-measure values of the decision tree methods are listed in Table 2. The J48 method achieves remarkable results when compared to other decision-based methods based on accuracy, precision, recall, and F-measure values. On the other hand, the decision stump method has lower performance. Second, the accuracy, precision, recall, and F-measure values of the Naive Bayes classifier methods are listed in Table 3. The Bayes Net method has better results than Naïve Bayes and achieves 99.97% accuracy, and 1.0 precision, recall, and F-measure values.

Table 2. Results of decision tree methods

Method	Accuracy (%)	Precision	Recall	F-Measure
J48	99.99	1.000	1.000	1.000
LMT	99.98	1.000	1.000	1.000
Random Forest	99.98	1.000	1.000	1.000
REP Tree	99.98	1.000	1.000	1.000
Hoeffding Tree	98.99	0.990	0.990	0.990
Random Tree	97.37	0.974	0.974	0.974
Decision Stump	40.78	0.344	0.408	0.155

Table 3. Results of naïve bayes classifiers

Method	Accuracy (%)	Precision	Recall	F-Measure
Bayes Net	99.97	1.000	1.000	1.000
Naive Bayes	99.34	0.994	0.993	0.993

Table 4. Results of lazy learning methods

Method	Accuracy (%)	Precision	Recall	F-Measure
KStar	98.35	0.984	0.983	0.983
LWL	84.41	0.882	0.844	0.841
KNN	79.08	0.794	0.791	0.792

Table 5. Results of function-based methods

Method	Accuracy (%)	Precision	Recall	F-Measure
SVM	99.98	1.000	1.000	1.000
Simple Logistic	99.97	1.000	1.000	1.000
Logistic Regression	99.65	0.997	0.997	0.997
ANN	20.00	0.200	0.200	0.186

Table 6. Results of ensemble methods

Method	Accuracy(%)	Precision	Recall	F-Measure
AdaBoost	40.78	0.344	0.408	0.155

Bagging	99.98	1.000	1.000	1.000
LogitBoost	99.98	1.000	1.000	1.000
Multi Scheme	20.39	0.204	0.204	0.134
Random Committee	99.97	1.000	1.000	1.000
Random SubSpace	99.97	1.000	1.000	1.000
Stacking	20.39	0.204	0.204	0.134
Vote	20.39	0.204	0.204	0.134

Third, the accuracy, precision, recall, and F-measure values of the lazy learning methods are listed in Table 4. The best results are obtained when the KStar is used and achieves an accuracy of 98.4%, a precision of 0.98, a recall of 0.98, and an F-measure of 0.98. On the other hand, the KNN has a lower performance and reaches an accuracy of 79.08%, a precision of 0.79, a recall of 0.79, and an F-measure of 0.79. Fourth, the accuracy, precision, recall, and F-measure values of the functions-based methods are listed in Table 5. While the ANN has the lowest performance, the SVM method achieves the best results. The SVM method achieves an accuracy of 99.98%, and full precision, recall, and F-measure values. Finally, the accuracy, precision, recall, and F-measure values of ensemble methods are listed in Table 6. Bagging and LogitBoost achieve the best results with an accuracy of 99.98% and full precision, recall, and F-measure.

6. CONCLUSION

Crime is a type of behavior that causes bad effects on people. To eliminate these bad effects, predicting a crime before it happens improves social life in a good way. With the developing technology, crime prediction can be made with machine learning methods. This study aims to predict crime by using various machine learning methods. When the accuracy and time results of machine learning methods are evaluated, decision tree methods achieve 99.99% accuracies in about 2 seconds. Experimental results demonstrate that the machine learning methods achieved admirable performance in crime prediction. We believe that these methods will also enable the police units to develop a new strategy by saving time to prevent crime.

Conflict of Interest

Authors declare that there is no conflict of interest.

Contribution of Authors

[Author 1's Hilal Arslan]: Contributed to the study design, method implementation, and revised the manuscript for important content, research supervisor, and wrote the manuscript.

[Author 2's Aslınur Doluca Horoz]: Conceived to the idea and designed the study, collected and analyzed the data, literature review, and wrote the manuscript.

REFERENCES

Chun, S. A., Avinash Paturu, V., Yuan, S., Pathak, R., Atluri, V., R. Adam, N. (2019, June). Crime prediction model using deep neural networks. In Proceedings of the 20th Annual International Conference on digital government research (pp. 512-514). doi: https://doi.org/10.1145/3325112.3328221

Dietterich, T. G. (2000, June). Ensemble methods in machine learning. In International workshop on multiple classifier systems (pp. 1-15). Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/3-540-45014-9_1

Gardner, M. W., Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment, 32(14-15), 2627-2636. doi: https://doi.org/10.1016/S1352-2310(97)00447-0

Jain, V., Sharma, Y., Bhatia, A., Arora, V. (2017). Crime prediction using K-means algorithm. GRD Journals-Global Research and Development Journal for Engineering, 2(5), 206-209. https://grdjournals.com/uploads/article/GRDJE/V02/I05/0176/GRDJEV02I050176.pdf

Llaha, O. (2020). Crime Analysis and Prediction using Machine Learning. In 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO) (pp. 496-501). IEEE. doi: https://doi.org/10.22214/ijraset.2023.50310

NYPD Arrests Data Historic 2006 - 2020, List of every arrest in NYC going back to 2006 through the end of the year 2020. [Online]. Available: https://www.kaggle.com/datasets/okettaeneye/nypdarrests-data-historic-2006-2020

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). https://www.cc.gatech.edu/home/isbell/classes/reading/papers/Rish.pdf

Safat, W., Asghar, S., Gillani, S. A. (2021). Empirical analysis for crime prediction and forecasting using machine learning and deep learning techniques. IEEE Access, 9, 70080-70094. doi: https://doi.org/10.1109/ACCESS.2021.3078117

Song, Y. Y., Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130. doi: https://doi.org/ 10.11919/j.issn.1002-0829.215044

Tamir, A., Watson, E., Willett, B., Hasan, Q., Yuan, J. S. (2021). Crime Prediction and Forecasting using Machine Learning Algorithms. International Journal of Computer Science and Information Technologies, 12(2), 26-33. https://ijcsit.com/docs/volume12/vol12issue02/ijcsit2021120201.pdf

Wettschereck, D., Aha, D. W., Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. Artificial Intelligence Review, 11(1), 273-314. https://link.springer.com/article/10.1023/A:1006593614256

Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations. https://researchcommons.waikato.ac.nz/bitstream/handle/10289/1040/uow-cs-wp-1999-11.pdf?sequence=1&isAllowed=y

Zhang, X., Liu, L., Xiao, L., Ji, J. (2020). Comparison of machine learning algorithms for predicting crime hotspots. IEEE Access, 8, 181302-181310. doi: https://doi.org.tr/ 10.1109/ACCESS.2020.3028420