Conference Paper No: SF-01

An enhanced ensemble model for crime occurrence prediction using machine learning

W. S. V. Lakshan[1*], A. T. P. Silva[2] and W. A. C. Weerakoon[1]

[1]Department of Statistics and Computer Science, University of Kelaniya, Sri Lanka
[2]Department of Computational Mathematics, University of Moratuwa, Sri Lanka
lakshanw_ps15070@stu.kln.ac.lk*

Abstract

With the rapid increase of crime, law enforcement departments are struggling to stop crimes and continuously demand automated advanced systems for crime control to provide better protection to the human being in a community. Crime predication plays a vital role in crime control. Crime analysis & prediction can reveal the complexities and hidden patterns in the crime datasets, and it can be used for early decision making. The early researchers attempted to predict the crime using a machine learning model with main factors including time, date and location but overlooked other essential factors. This paper aims to present an enhanced crime prediction algorithm based on ensemble classification technique while identifying several factors that affect the learning model's performance. The correlation of the factors versus the prediction label is analyzed using the Spearman and Pearson techniques to determine the important, influential factors. The prediction model was developed based on Ensemble techniques using the Random forest model with the Voting Classifier. Multiple decision trees had implemented the crime prediction model of this research as the base model and the Logistic Regression and K-Nearest Neighbor algorithm as the sub-models. The final classifier was developed based on using the Graphical User Interface and the REST API methods to predict the possibilities of the crime occurrences at a given specific time in each location. The proposed method can identify the likelihood of a crime in a particular location at a specific time. This helps to implement better strategic and tactical ways to minimize crimes with less risk, as the accuracy of the crime prediction algorithm was 89%.

Keywords: Correlation, Ensemble techniques, Graphical User Interface (GUI), REST API, Voting Classifier.

Introduction

For a better surveillance level of the country, the law enforcement is working as a shield for the citizen‘s protection by taking defenses against crimes and unlawful acts. Therefore, implementing a suitable crime prediction system by using machine learning tools and data mining techniques is a progressively approachable strategy towards the law enforcement. The criminology is the scientific learning of the crimes and the behaviors and intentions of the criminals. The crime prediction is the practice of forecasting the crime occurrences by analyzing huge dataset with the analyzing techniques. The motivation for implementing an enhanced algorithm to predict the future crime was the knowledge and understanding obtained from the machine learning (ML) techniques. For the crime prediction, searching the supportive information to use and assist in crime prediction problems are possible, but there is no definitive method to predict crimes with the various data (Shamsuddin et al., 2017) and the new technologies make much easier to

identify the crime hotspots where the crimes placed repeatedly over an area by focusing a specific area (Ratnayake, 2015). Data analyzing is crucial to understand the patterns of crime occurrences and several results can be obtained by using the same input factors for the different algorithms (Yuki et al., 2019). The Ensemble model had been resulted better than the other ML models and suitable model selection can be done based on the accuracy, recall and precision (Iqbal et al., 2013). Though there were many effective analyzing tools to identify the common patterns of the crime incidents, but the cost of installation is high (A & Santhosh Baboo, 2011). The Random Forest model has a better accuracy level for the pattern identification of crime types, and it can be modified with more features to predict the future crime occurrences (Alves et al., 2018). According to the facts, the main consideration of each previous research was to predict the future crime occurrences by using an algorithm to identify the patterns or distribution of crime hotspots to determine the future crime occurrences. All the previous studies revealed that the sufficient data set is essential to achieve a better performance by training the model which was implemented for the crime prediction system.

The crime occurrence prediction within concerned area at a specific time frame and identification of the new factors for a better prediction process were the main objectives of the research. The current system has been modified with the user-friendly Graphical User Interface (GUI) and Application programming Interface (API) to improve the performance of the enhanced algorithm. The crime prediction system has been implemented to predict the crime occurrence accurately based on the factors which can be give the predictions under the high measure of potential casualties (Dharmaraju, 2017). Therefore, the null hypothesis (H0) for the current algorithm building was "The crime occurrence could be predicted by the crime prediction system" and the alternative hypothesis (H1) was "The crime occurrence could not be predicted by the crime prediction system". Compared to the previous studies, this is not only a system which was built to predict the future crime occurrences, but very efficient and effective to be promoted easily among the officers of the law enforcement for encouraging the crime investigation and detection process. This study would be turning the old-cultured police stations, recording data and the activities of investigating into the new easy and effective paths which are more suitable for the present-day scenarios with the new technology.

Methodology/materials and methods

The model implementation and the algorithm development for the crime prediction system has been done based on the Python Language as it is an object-oriented scripting language which stimulates code reuse with the modules for data analysis. The Google Colaboratory Notebooks has been used for executing the code on Google`s cloud servers and the Spyder application has been used for training the model with the GUI. The Django and Postman frameworks had been used for building and testing the REST API respectively. The crime data for the model implementation found by "Kaggle" open database, website of "Statistics Canada" and the Census data (2011 and 2016) from the "Census Profile of the Statistics of the Canada" website to get the data related to the Vancouver city as required to the factors. While preprocessing the data, data integration has been performed by combining the data collected from each source, then data cleaning has been performed to remove the missing values and data transformation has been done by labelling such as that number of 'crime types' as 1-11, number of 'neighborhoods' as 1-26 and 'class' of crime occurrence as "if the crime occurred then 1 if not as 0" then split the data into two phases as training phase and testing phase. This large dataset has been summarized into smaller one that contains the information and extract the suitable

features using the Principal Component Analysis. The Pearson Correlation and Spearman`s Ranking are the correlation tests had been performed to test the mutual relationship precisely between each input feature and the 'class' variable. Then the algorithm has been implemented based on the Random Forest classifier with the two sub models called Logistic Regression algorithm and K– Nearest Neighbors (KNN) algorithm which have been chosen based on the accuracy level and the Support Vector Machine model did not result better accuracy. Finally, Max Voting has been chosen as the suitable Ensemble classifier for a better accurate future crime prediction.

The Random Forest, Neural Network and KNN models had been used to predict the 'Crime type' with the other input factors (without the previous 'class' variable). The levels of accuracy were lower than 50% and the correlation coefficients between the 'Crime type' variable and other factors were very small. Through the R&Ds, checked for the possibilities to predict the level of the severity of sending back the offender to the society and found that the current factors were not suitable to use as the inputs. There are new various factors which can affect the level of severity and crime type as shown in the table 1 below, but the data are not available as required.

Table 1. *Factors related to both Crime Type Prediction and Severity Level Prediction*

Feature Category	**Factors for Crime Type Prediction**	**Factors for Severity Level Prediction**
Historical features	Crime type Event Trend	First time or a repeat offence
	Crime Rate of Surrounding Regions	Main offender or an Accessory
	Seasonal Pattern of each crime type	Contribution of the offender
Demographical features	Gender	Gender of the offender
	Age group	
	Income level	Age of the offender
	Number of Residence	
Dynamic features	Diversity of visitors in a location	Committed under Personal stress or Duress
	Visitor Ratio	Committed without intention (or self-defensing purpose)
	Count of unique visitors	Attempted with Destructive/ Vindictive intention
Geographical features	Venue Category distribution	
	Density	
	Reginal Diversity	

Therefore, the prediction system for Crime Occurrence was developed with the GUI by using the "Tkinter package (Tk interface)" and Tk GUI toolkit makes it easier to develop suitable standard Python Interface for the algorithm which is as shown in the figure 1. Using the Django framework implemented the REST API locally. The ".joblib" extension

used to serialize the model and the "python manage.py runserver" command to make server up in the Anaconda prompt and produced a local IP address. Users will be able to access the web GUI of the REST API by using IP address. The "Basic Authentication" facility has been created for the security checking process while accessing the REST API. The postman framework used to test the efficiency. Finally, the Local Endpoint would be ''http://127.0.0.1.8000/api/predict/'' and the default port was 8000 used by the local host.

Figure 1. *Graphical User Interface.*

Results and Discussion

According to the correlation coefficients as shown in table 2 and table 3, identified the new four (04) factors which can be affected for the crime occurrences.

Table 2. *Correlation checking for the Current Features Used in Previous Studies*

According to Class	Type	Date	Month	Year	Time	Area
(r)	-0.014	0.005	-0.014	-0.022	-0.086	-0.026
(rs)	0.081	0.066	-0.004	0.029	-0.004	0.032

The above factors were taken as the input variables for the previous research works. Therefore, those factors will be considered for the current model building by concerning the graph of the variable importance of the factors towards the model as shown in the figure 2.

Table 3. *Correlation checking for the New Features selected for the Current study.*

According to Class	Near to Main Road or Not	Near to Commercial Centre or not	Income status of the House Owner	Population of Area	Knowledge of House Owner
(r)	-0.003	-0.122	0.017	-0.128	0.066
(rs)	-0.086	0.068	-0.001	-0.002	-0.047

The "Income status of the house owner" variable had been removed due to weak relationship (as the values are closer to the 0). Then, checked for the variable importance of the selected variables as follows.

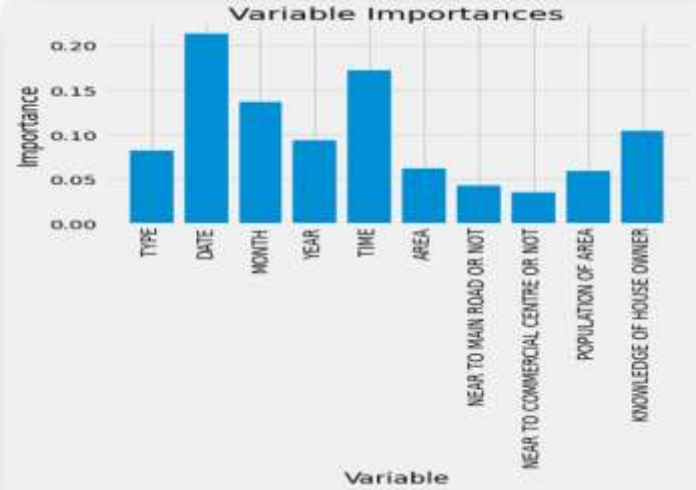

Figure 2. *Graph of the Variable Importance.*

For the base model, the Random Forest algorithm was trained with six test cases by increasing the number of decision trees such as 10, 50, 100, 120, 150 and 200, respectively. The optimum result was obtained as 90% of accuracy with 100 decision trees. Accuracy of the Logistic Regression algorithm and the K – Nearest Neighbor algorithm were obtained as 90% and 87.8% respectively. Each algorithm has been used for the Voting classifier without boosting since the accuracy of each model did not change with Adaptive Boosting technique and resulted in 89.8% of accuracy for the final enhanced Ensemble model. The precision, recall and F1 score obtained as 90.4%, 99.5% and 94.7% for the final model respectively. Then the Ensemble model tested with 06 months of real data and obtained 92% of accuracy predicting 5011 incidents out of 5432 crimes exactly. The response time of the REST API is the time elapsed between sending the JSON Request and the response from REST API was 200 – 250 ms and the outputs obtained through the REST API as follows in the figure 3.

Figure 3. *Possible Outputs from the REST API.*

Incomplete dataset was the main issue and after the data cleaning phase there were less amount of data. Therefore, the results had been less reliable and difficult to identify crime patterns. The use of sensible data is essential as it plays a major role in the crime prediction and not having such data is a limitation to identify many hidden patterns with the increased number of input factors while training the model. Multiple models were implemented to overcome the main drawbacks of the decision tree algorithm such as overfitting and biasness of the trained model. In a Deep Neural Network it needs a lot of resources, data, and computational power for the model training phase. Therefore, Ensemble model has been chosen as it is capable to obtain a higher percentage of accuracy with less data.

Conclusion

The main aim was achieved by implementing an enhanced Ensemble model to predict the occurrence of the crimes with the 89% of accuracy level. The accuracy of the trained model resulted as 92% of accuracy for real time data. Identified new four (04) factors which can be used for the model building for a crime prediction system. The user-friendly GUI and the REST API had been developed for the user interaction and to reduce the cost of hardware implementation. The law enforcement will be able to use the current crime prediction system in multiple locations at the same time concurrently. Therefore, each objective of the study has been achieved. The output of this research will be a benefit for a better performance in the crime detection and analyzing process. For further development, the current model will be enhanced to predict the type of crime and the level of severity by finding the related data. Not only the researchers can be focused on building more reliable models with the collaboration of the law enforcement collecting sensitive data records, but for focusing on other metrics to measure the performance of the algorithm except the accuracy.

Acknowledgment

We thank for the support given by the Department of Statistics & Computer Science, Faculty of Science, University of Kelaniya, Sri Lanka.

References

A, M., & Santhosh Baboo, S. (2011). An Enhanced Algorithm to Predict a Future Crime using Data Mining. *International Journal of Computer Applications*, *21*(1), 1–6. https://doi.org/10.5120/2478-3335

Alves, L. G. A., Ribeiro, H. V, & Rodrigues, F. A. (2018). Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and Its Applications*, *505*, 435–443. https://doi.org/10.1016/j.physa.2018.03.084

Dharmaraju, G. (2017). Analysis of Crime data using data mining. *International Journal of Engineering, Science and Mathematics*, *6*(8), 1059–1071.

Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., & Khanahmadliravi, N. (2013). An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology*, *6*(3), 4219–4225. https://doi.org/10.17485/ijst/2013/v6i3/31230

Ratnayake, R. (2015). Distribution Pattern of Urban Crimes in Sri Lanka; with Special Reference to Mirihana Police Division. *International Journal of Multidisciplinary Research and Development*, *2*(8), 441–453.

Shamsuddin, N. H. M., Ali, N. A., & Alwee, R. (2017). An overview on crime prediction methods. *6th ICT International Student Project Conference: Elevating Community Through ICT, ICT-ISPC 2017*, *2017-Janua*(May), 1–5. https://doi.org/10.1109/ICT-ISPC.2017.8075335

Yuki, J. Q., Mahfil Quader Sakib, M., Zamal, Z., Habibullah, K. M., & Das, A. K. (2019). Predicting crime using time and location data. *ACM International Conference Proceeding Series*, *July*, 124–128. https://doi.org/10.1145/3348445.3348483