Comparative analysis of crime predictions using machine learning algorithms with geospatial features

Cite as: AIP Conference Proceedings **2393**, 020004 (2022); https://doi.org/10.1063/5.0074141
Published Online: 19 May 2022

M. Raja Suguna, R. Beaulah Jeyavathana and K. V. Kanimozhi

Comparative Analysis of Crime Predictions Using Machine Learning Algorithms with Geospatial Features

M. Raja Suguna[1,a)], Dr. R. Beaulah Jeyavathana[1], K.V. Kanimozhi[2]

[1]*Department of Computer Science and Engineering, Saveetha School of Engineering, Chennai, India*

[2]*Saveetha Institution of Medical and Technical Sciences, Chennai, India*

[a)] *Corresponding author: suguna.raj89@gmail.com.*

Abstract. Crime constitutes an action which is punishable by law. Crime Analysis involves the predictions of occurrence of future crime, the time and place of crime and to have insights into the trends of crime. Models are created by machine learning algorithms using the spatial, temporal and the demographic features extracted from the crime dataset. Reverse Geocoding technique is used to extract spatial features and also visualize the locations of the crime from the crime dataset using ArcGIS API of python along with WebMap and WebScene component provided by the API. Crime Analysis assists the Police, Investigation departments for the prediction of future crime and also take required actions which involves the deployment of crops in the predicted place at the time of the crime. The hotspots are identified; the hotspot is the area co-ordinates where more frequent crimes occur. After identifying hotspots, more focus is given on those crime prone areas for preventing and controlling the crime.

Keywords: Crime Category, hotspots, Geospatial feature, Crime Analysis

INTRODUCTION

Large amount of crime data is available. Advances in Machine learning and deep learning methods accelerates the data analysis of crime data. Future Crime Predictions, getting Insights into the pattern and trends of crime in a given geographical region assists the investigation Department to take preventive measures at a greater extend. These Prediction techniques doesn't predict the victim and the offender of the crime, instead helps in find out the time and the locations of future crime and the identification of the crime prone areas.Unites States Open Crime dataset for the Montgomery County is downloaded from data.World website. The dataset has recorded crime from 2016 to 2020 and it is updated daily. The dataset has the crime data under the several categories such as Homicide, Assault, Arson, Burglary, Drugs, Extort, Fraud, Robbery, Larceny, Kidnap, Weapon Offence having more than 1,67000 rows. The dataset recorded the crime under following columns Incident Id, Offense code, Number of victims, Crime Category, zip code, Block Address, latitude and longitude of the location, date and time of the crime, City, Sector, police district number and some more.For the above dataset Hierarchical Density-Based Spatial Clustering of Applications with Noise algorithm is applied to find out the Hotspots. Then spatial Features are extracted from the demographic information from the dataset. Both the hotspot information and temporal, spatial feature extracted are used to explore the future location and the time of crime. The main goal of the project is to find the primary crime clustering based on geographic, relation between clusters and different category of crime, analyses the crime patterns so that the police align their patrol dynamically for the various level of criminal activities. The rest of the paper is organized as follows section 2 discuss about the related work carried out the Crime Prediction and analysis in the literature, Section 3 gives the detail work of the proposed system Section 4 discuss about the Conclusion and the Feature Work.

RELATED WORK

The most of the work in the literature focuses on finding the Hotspots for the crime data. Hotspots are the areas with varying Geographic size having high crime rate. Brower et al [1] analyzed the crime data for Wisconsin-Madison University and Madison city for the year 2003.They used GIS (Geographical Information System) to

Recent Trends in Science and Engineering
AIP Conf. Proc. 2393, 020004-1–020004-4; https://doi.org/10.1063/5.0074141
Published by AIP Publishing. 978-0-7354-4198-9/$30.00

extract spatial And Temporal Features from the Crime Data to find hotspots. different Category of crime associated with different geographical region called hotspots. They find out the serious crime occurs at the peak hour bar closing time at 2:00 to 3:00 AM and less serious crime at 12:00 to midnight. Students from the university related to the alcohol crime. P. Thongtae et al [2] gave a detailed comprehensive analysis of data mining algorithms such as Clustering and classification in predicting the crime and to find the escape offender. Malachi et al [3] proposed new modified K-Nearest Neighboor Algorithm for handling missing values in the dataset and used k-means clustering algorithm and the Density Based Spatial Clustering Application with Noise (DBScan) algorithm to identify crime hotspots and the crime patterns for the crime dataset. Keivan Kianmehr et al [4] integrated k-means clustering algorithm and one-class SVM (Support Vector Algorithm) and two-Class SVM Algorithm for prediction of Crime Hotspots. In first step they form the initial level hotspots with randomly selected crime rates datapoints and k-means clustering algorithm. Then, output of the first step is used to train the One-Class SVM and two Class SVM Algorithm to form different shaped Hotspots for the given crime dataset. Their approach performed well compared to Statistical Regression Approach and Neural Network Method.Murray et al [5] had a detailed study over the use of non-hierarchical clustering algorithm for the hotspot identification. They reviewed the shortcomings of k-means clustering algorithm for the crime hotspot generation. Proposed improvised Sum of Squares Clustering Model (SSCM) by including Spatial lag called Spatial Lag Clustering Model - Median (SLCM-M) for the better generation oh hotspots with different geographical region.Natalya et al [6] used space-tome 3D cube for visualizing and crime cluster movement by combining Space Time Kernel Density Estimation and Space time Scan Statistics and a promising approach for 3D-GeoVisualization. Wang et al [7] used novel pre-processing techniques to take crime rates over only the specific longitude and latitude coordinates (specific geographic region) and implemented state of art Station-Temporal Deep Neural networks called ResNet one with convolution layer and other without the convolution layer. The heat map showing the crime cluster distribution in the particular region is given s input to the Deep Neural Networks, the predicted map without convolution layer shows the greater accuracy in predicting crime cluster distribution over the particular time. Wolff M et al [8] developed a 3D-GeoVisualization App with most of the features of 3D-GIS framework, which extensively used for extensive Exploratory Crime Data Analysis for the different Crime Mapping by geo-profiling experts.Yu, Chung-Hsien et al [9] used ensemble method of datamining algorithms such as j48, Neural Network, SVM and 1NN for predicting the crime hotspots. INN algorithm shows the improved accuracy and f-score for crime hotspot predictions.

PROPOSED WORK

The proposed method consists of the following steps, Data Pre-processing, Engineering Spatio-Temporal Features, Finding Hotspots, and Classification Model for Feature crime prediction. Fig 1 shows the general architecture of the proposed system.

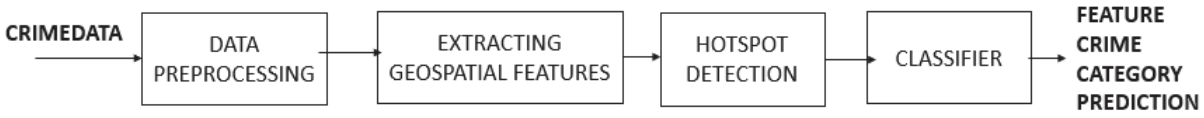

FIGURE 1. System Architecture

The Montgomery County open crime dataset is downloaded from data. world website. The data contains more than 16 lakes rows having crime data collected under the columns Incident ID, offender Id, date of dispatch, number of victims, Crime Place, location, Block Address, City, zip code, address number, street type, latitude, longitude, Police District number under 30 columns. The dataset is filtered by considering the following crime category of Arson, Assault, Theft, Burglary, Drugs, Forgery, Fraud, Homicide and Threat for the year 2019. The data rows with missing rows are removed from the dataset. All the thirteen columns are not taken for exploratory Data AnalysisEngineering Geo-Spatial Features:Using Geocoding Technique, Spatial feature is extracted for the crime data. ArcGIS API of python combines demographic information such as block Address, Street name, Place of the

crime and the coordinates of the crime place and outputs the 108 bytes of locations that can be visualized in the WebMap and WebScene Component of the python API. All the locations given by API is categorized into 12 categories of locations such as pubs, police stations, parks, hospitals etc., in the map of Montgomery County.Hotspot is the area or coordinates of location in the city map having a greater number of crimes. Hotspot detection is the important step in Crime Prediction.Hierarchical Density Based Spatial Clustering of Application with Noise is used for cluster generations. The clusters are the hotspots identified. After this, to have a greater insight into the crime trends over a particular cluster, an exploratory crime data Analysis is needed. Spatial Lag Cluster Model -Median (SLCM-M) is used for analyzing crime trends over a particular hotspot. Spatial lags are the spatial Proximity metric, which tends to find the average number of crime surrounding a hotspot identified. Spatial lag Clustering Model is non-hierarchical clustering technique includes all the crime events in any one of the hotspots identified. A crime event will not be the members of two different hotspot cluster. It resolves the local minima problem and also enables for comparative analysis of the neighborhood Hotspots and also resolves spatial bias problem. The central part of each hotspot is identified called hotpoints. These Hotpoints and the engineered geospatial feature achieved a increased accuracy in prediction of feature crime Category.

CLASSIFICATION MODEL

Naïve Bayes and Support Vector Classifier are trained by the crime dataset with the 10-Fold Cross validation. An Ensemble Classifier of both the Support Vector Machine and Naïve Bayes Classifier is also trained. Crime Category is predicted by the shortest distance from the Hotpoints of the cluster hotspots. Table 1 shows the accuracy of the classifiers for the four category of crime such as Arson, Assault, Theft, Burglary, Drugs, Forgery, Fraud, Homicide and Threat. The Ensemble Classifier shows the increase in Accuracy of predictions.

TABLE 1. Accuracy of the Classifier

S.NO	Crime Category	Accuracy of the Classifier		
		Naïve Bayes	SVM	Ensemble Classifier
1	Assault	80.1	82.5	86.5
2	Burglary	79.3	81.22	82.5
3	Drugs	85.5	84.9	86.5
4	Fraud	79.32	88.14	88.56
5	Homicide	75.48	79.65	80.81

CONCLUSION

In this paper, the geospatial features extracted from the crime data and the hotspot detection shows the increased in Accuracy of the Prediction of Crime Category by the Naïve Bayesian, Support Vector Machine and the Ensemble of this two classifiers. The future work is extended to include all types of crime category with Neural Network Architecture for the future crime prediction.

REFERENCES

1. Brower, A. and Lisa Carroll., Journal of American College Health **55**, 267 – 275 (2007).
2. P. Thongtae and S. Srisuk, in Computer and Information Technology, IEEE 8th International Conference on, null, 2008.
3. Malathi. A and Dr. Santhosh S Baboo., International Journal of Computer Applications **21**(1),1–6(2011)
4. KeivanKianmehr&RedaAlhajj, Applied Artificial Intelligence, **22**(5), 433-458 (2008).
5. Simran Agarwal, M. Ramachandran, International Journal of Research in Ayurveda and Pharmacy**7**(2), 67-70 (2016).
6. Murray A.T., Grubs T.H., Geotechnologies and the Environment, Springer, Dordrecht,**8**(2013)

7. Nakaya, T., Yano, K.: Visualising crime clusters in a space-time cube,T. GIS **14**(3), 223-239 (2010)
8. Wang, B., Zhang, D., Zhang, D., Brantingham, P. J., and Bertozzi, A. L., "Deep Learning for Real Time Crime Forecasting2017.
9. Wolff M., Asche H., Lecture Notes in Computer Science, **6016** (2010).
10. DeepaliMor, M Ramachandran, PramodRaichurkar, Nature Environment and Pollution Technology **16**(1), 247-250 (2017).
11. Yu, Chung-Hsien et al. "Crime Forecasting Using Data Mining Techniques." IEEE 11th International Conference on Data Mining Workshops 779-786 (2011).