IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Crime Data Prediction Based On Geographical Location

Mr.Kamlesh Singh [1], Asst.Prof. Nimesh Vaidya [2], Dr. Vijaykumar B Gadhavi[3]

[1]PG Scholar – Faculty of Engineering, Computer Engineering Department Swaminarayan University, India

[2]Assistant Professor & HOD - Faculty of Engineering, Computer Engineering Department Swaminarayan University, India

[3]Associate Professor & Dean –Faculty of Engineering(I/C), Computer Engineering Department Swaminarayan University, India

ABSTRACT

To have a better response towards criminal activity, it is very important that one should understand the patterns in crime. We analyze this pattern by taking crime datasets from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. This dataset includes different blocks of the city of Chicago. The major aim of this mission is to expect which category of crime is most probably to take place at a detailed time and places in Chicago. Finally, this paper uses a different algorithm like Random Forest, Decision Tree and different ensemble methods such as Extra Trees, Bagging and AdaBoost to evaluate the accuracy given by each algorithm.

Key words: distance-decay principle, space tendency, transition density mode

1. INTRODUCTION

Crime as the word suggests it is the violation that people does, and it is usually performed against the laws and it is punishable. Crime Analysis is a part of criminology studies where a various pattern of activities involving criminology are studied and tries to find the indicators of occurred events. Criminal activities are a regular occurrence all over the world. Governments spend a huge amount of time to use technology to tackle criminal activities. Machine learning of its works with data and through reading records helps to predict. This paper uses machine-learning techniques to analyze the previous crime datasets and predicts the hotspots for the crime based on time and location.

The motivation of this paper is to explore the special device studying strategies to investigate the crime patterns so that this could help regulation enforcement organization to behavior their operations. Our approach will help regulation enforcement agencies to have assumption ahead about the possibility of crime, which could arise at a place in a given time. This will assist them to resolve the instances faster than earlier.

2. RELATED WORKS

Criminal activities are common around the world. Therefore, researchers have completed many works on this subject matter. Researches have been analyzing the relation among criminal activities and socio-economic variables like unemployment, earnings level, level of schooling and so forth.

Researchers like Bharati et al worked-on Crime Prediction and Analysis Using Machine Learning where the author used machine learning and data mining for prediction of crimes in Chicago [1]. The datasets include information like vicinity description, type of crime, date, time, range, longitude, etc. They used the K-Nearest Neighbor (KNN) classification and plenty of different algorithms to test for crime prediction. A classifier that gives higher accuracy is used for further training. Also, Sangani et al worked on similar paper on Crime Prediction and Analysis where they used Simple K-Means clustering techniques and algorithm for predicting Crimes

3. METHODOLOGY

This paper uses a specific dataset to train the algorithm. The algorithm that is used to train the dataset are Random Forest, Decision Tree and different ensemble methods such as Extra Trees, Bagging and AdaBoost.

The following steps are followed for all the implemented algorithms:

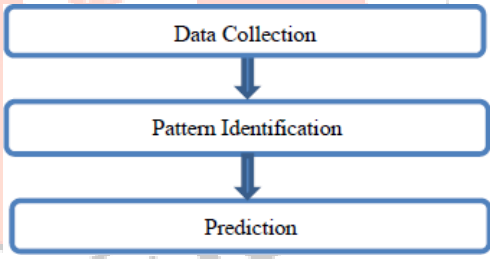

Figure 1. Proposed system.

3.1. Features

The dataset that is used is collected from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. The datasets consist of facts on crime prevalence that has taken area in Chicago over the time of 1/1/2001 to 1/1/2017. The dataset that we used is in a CSV format, which contains more than 6,000,000 records/rows.

There are different attributes of the dataset. The attributes that are used in this paper is given in the table:

Table 1. Attributes that is used from the datasets

Location Description
Description
FBI Code
Block
Location
Year
Latitude
Longitude
Month
Day
Hour
Minute
Second
Primary Type

3.2 .Data Preprocessing

For preprocessing the dataset, this paper used Tabular software and Python library Scikit-learn (sklearn).

1. The dataset consists of some attributes, which are string values, and other attributes are in numeric values. To train the model, the text features in this paper's dataset needs to be converted right into a numeric value. This conversion is dealt with by means of the usage of Python library NumPy.
2. Attributes in our dataset with string type are "Date", "Location", "Location Description" etc. Using python, this paper assigned numeric values for those capabilities.
3. Since time is considered as the main factor thus "Date" has been split into "Day", "Month", "Year", "Hour", ".

   The dataset that we used is imbalance as 20000 (Theft) have the highest frequency for the specific crime occurrences and the lowest frequency is six for another crime occurrences (Stalking). That is why; we want to lessen the training, which has a lower frequency to end up with the balanced dataset.

   Only the top 18 crime classes are utilized to try to reduce the range of instructions out of 30 categories. All other crime lessens are categorized through "Other offenses".

	Primary Type	Amt
9	GAMBLING	232
15	LIQUOR LAW VIOLATION	206
12	INTERFERENCE WITH PUBLIC OFFICER	198
0	ARSON	158
10	HOMICIDE	124
14	KIDNAPPING	89
13	INTIMIDATION	64
27	STALKING	41
19	OBSCENITY	6
18	NON-CRIMINAL	1
24	RITUALISM	1
11	HUMAN TRAFFICKING	1
4	CONCEALED CARRY LICENSE VIOLATION	1

Figure shows 13 classes of "Other crimes".

In Fig above shows specific crimes which are considered as “Other crimes. This is because the number of occurrences involving those crimes are lower, compared to other classes of crimes available in the dataset.

4. RESULTS

This paper uses dataset which contains both the mixture of categorical and numeric values. Thus, the paper mainly focuses on those algorithms which can work on the combination of both categorical and numeric values. Also, keeping in mind that, the algorithm performs well for our classification problem. Therefore, several algorithms are chosen to serve the purpose such as Decision Tree, Random forest and several ensemble methods such as Bagging, AdaBoost and ExtraTree Classifier.

The main motive of this paper is to use algorithms on these datasets to classify the type of crime occurring based on time and location. The chosen algorithms are applied where it provides a simple and fast way of learning a function. This is where the algorithm maps data x to outputs y, where x is a mixture of categorical and numeric variables and y is the categorical value for classification. The applied algorithm gives better performance for any classification problem.

The result after reducing the classes is shown in the below table for all algorithms.

Table 2. Comparison of accuracy

Algorithm	Accuracy
Random Forest	95.99%
Decision Tree	99.88%
AdaBoost	74.78%
Bagging	99.92%
Extra Tree	97.10%

The table 2 shows that Bagging gives the highest accuracy and AdaBoost gives the lowest accuracy among the fixed algorithm that is used.

AdaBoost creates a strong classifier from several weak classifiers. AdaBoost generally works best for binary classification. The dataset used in this paper consists of different categories of crimes. Therefore, it is a multiclass classification problem. Thus, AdaBoost provides lowest accuracy among all the algorithm used.

5. CONCLUSION

This paper uses five different types of algorithms to predict the type of crime that might occur based on time and location. The algorithm involving trees showed that the predicted results is very much closer to the actual results. Thus, the dataset used, provides the maximum correct result with higher accuracy when implemented with different tree classifiers. The stated results in this paper show that Bagging method works best and AdaBoost works least well for predicting crimes using time and location. The results in this paper provides similar results when implemented with tree-based algorithms. Therefore, this paper expects to get more variation in the results when implemented with other classifying algorithms in the future.

6. REFERENCES

[1] Alkesh Bharati and Dr Sarvanaguru RA.K. 2018. Crime Prediction and Analysis Using Machine Learning.

[2] Ankit Sangani, Vijaya Pinjarkar and Chirag Sampat. 2019. Crime Prediction and Analysis, *2nd International Conference on Advances in Science & Technology*.

[3] Ayisheshim Almaw and Kalyani Kadam. 2018. Survey Paper on Crime Prediction using Ensemble Approach, International *Journal of Pure and Applied Mathematics*, 118 (8), 133-139

[4] Christian Tabedzki, Amruthesh Thirumalaiswamy and Paul van Vliet. 2018. Yo Home to Bel-Air: Predicting Crime on The Streets of Philadelphia

[5] Clifton Phua, Damminda Alahakoon and Vincent Lee. 2004. Minority Report in Fraud Detection: Classification of Skewed Data, *Sigkdd Explorations*, 6(1),51-5

[6] Tayal, D.K., Jain, A., Arora, S., Agarwal, S., Gupta, T. and Tyagi, N., 2015. Crime detection and criminal identification in India using data mining techniques. *AI & society*, *30*(1), pp.117-127. https://doi.org/10.1007/s00146-014-0539-6

[7] Hyeon-Woo Kang and Hang-Bong Kang. 2018. Prediction of crime occurrence from multimodal data using deep learning.
DOI= https://doi.org/10.1371/journal.pone.0176244

[8] Irina Matijosaitiene, Peng Zhao, Sylvain Jaume and Joseph W. Gilkey Jr. 2018. Prediction of Hourly Effect of Land Use