Crime Prediction Using Machine Learning

Student 1 **160104082**
Student 2 **160104128**
Student 3 **160104137**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Fall 2020

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2021

Crime Prediction Using Machine Learning

Submitted by

Student 1	**160104082**
Student 2	**160104128**
Student 3	**160104137**

Submitted To

Faisal Muhammad Shah, Associate Professor

Farzad Ahmed, Lecturer

Md. Tanvir Rouf Shawon, Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2021

ABSTRACT

Recognizing the patterns of criminal activity of a place is paramount in order to prevent it. Law enforcement agencies can work effectively and respond faster if they have better knowledge about crime patterns in different geological points of a city. The aim of this paper is to use machine learning techniques to classify a criminal incident. The experimentation is conducted on a dataset containing San Francisco's crime records from 2003 - 2015. For this supervised classification problem, Decision Tree, K Neighbors,Gaussian Naive Bayes, Logistic Regression,Random Forest classification models were used. But our results could be furtherrefined by pre-processing the dataset better.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Criminal activities are present in every region of the world affecting quality of life and socio-economical development. As such, it is a major concern of many governments who are using different advanced technology to tackle such issues. Crime Analysis, a sub branch of criminology, studies the behavioral pattern of criminal activities and tries to identify the indicators of such events. Machine learning agents work with data and employ different techniques to find patterns in data making it very useful for predictive analysis. Law enforcement agencies use different patrolling strategies based on the information they get to keep an area secure. A machine learning agent can learn and analyze the pattern of occurrence of a crime based on the reports of previous criminal activities and can find hotspots based on time, type or any other factor. This technique is known as classification and it allows to predict nominal class labels. Classification has been used on many different domains such as financial market, business intelligence, healthcare, weather forecasting etc. In this research, a dataset from San-Francisco Open Data[8] is used which contains the reported criminal activities in the neighborhoods of the city San Francisco for a duration of 12 years. We used different classification techniques like Decision Tree, K Neighbors,Gaussian Naive Bayes, Logistic Regression,Random Forest to find hotspots of criminal activities based on the time of day. Results of different algorithms have been compared and most the effective approach has also been documented

Chapter 2

Literature Reviews

As combating criminal activity has always been a priority for governments around the world, many researches has been done to effectively find countermeasures and indicators of crime prior to happening. Criminologists have been pursuing to identify hotspots that need major attention from law enforcement agencies.

Analyzing the usage of mobile network infrastructure and demographic information of people living in different areas of London, a group of researchers were able to predict if particular areas of London would become a criminal hotspot [1]. They have implied that anonymized data collected by mobile networks contain indicators for predicting crime levels.

Analyzing the usage of mobile network infrastructure and demographic information of people living in different areas of London, a group of researchers were able to predict if particular areas of London would become a criminal hotspot [2]. They have implied that anonymized data collected by mobile networks contain indicators for predicting crime levels.

Combining two datasets - 1990 US LEMAS and crime data 1995 FBI UCR and applying classification techniques like Decision Tree and Naive Bayesian algorithm, 83.95accuracy have been achieved when asked to predict a crime category for different states of
USA [3]. However, this paper does not disclose if there were any imbalanced classes of crime category. The same databases were also explored by Somayeh et al [4] who employed a number of machine learning algorithms, where k-Nearest Neighbor algorithm performed better than other algorithms by having an accuracy of 89.50feature to improve the feature selection.

Wang et al [5] proposed the Series Finder, a machine learning agent that tried to find patterns in crime committed by same offender or groups of offenders. Clustering has also been used to study patterns of criminal behavior and geographic criminal history.

Remond and Baveja [6] have worked on the data noise problem and studied how some

police reports or cases are idiosyncratic and do not contain good indicative matrices. Their proposed system called Case-Based Reasoning (CBR) filtered out these cases, and using this system, they were able to predict better compared to not having any filters on the data.

Chapter 3

Data Collection & Processing

The experiment is conducted on a specific dataset. The dataset is provided by SF Opendata from SFPD Crime Incident Reporting System [7]. It provides information on crime incidents that occurred in San Francisco for the period of 1/1/2003 to 5/13/2015. The dataset is a csv file containing 878049 rows. The attributes are given below.

Datetime	A timestamp when the given crime occurred.
Category	Type of crimes. This is the target label for the data. There are 39 types ofcrime listed in the data
Crime Description	A detailed description of a specific type of crime
Day	Day of week
pDistrict	Name of police department district. There are total 10 Police Districts in thedata
Resolution	How the crime was solved. 17 types of resolution
Address	The approximate street address for the incident
X	It signifies the latitude of the location of the crime.
Y	It signifies the longitude of the location of the crime.

Table 3.1: Attributes of the crime dataset

From the list of attributes in 3.1, the features and label can be determined. The target label that needs to be predicted is the Category of a crime incident. The attributes: Crime Description and Resolution are also related to the target label. Hence, all other attributes apart from these three attributes are used as features.

There are 39 types of crime in the San Francisco Crime Dataset. These types are considered as classes and having 39 classes makes it a multi class problem.3.2

From datetime stamp, four four main features are extracted. - year, month, date, hour.Most crimes occur during afternoon-evening. From midnight to morning, reports of criminal activities are low.There is an upsurge of criminal activities at 6 PM and 8 PM. Criminal ac-

Category	Frequency
LARCENY/THEFT	174900
OTHER OFFENSES	126182
NON-CRIMINAL	92304
ASSAULT	76876
DRUG/NARCOTIC	53971
VEHICLE THEFT	53781
VANDALISM	44725
WARRANTS	42214
BURGLARY	36755
SUSPICIOUS OCC	31414
MISSING PERSON	25989
ROBBERY	23000
FRAUD	16679
FORGERY/COUNTERFEITING	10609
SECONDARY CODES	9985

Table 3.2: Frequency of top 15 crimes

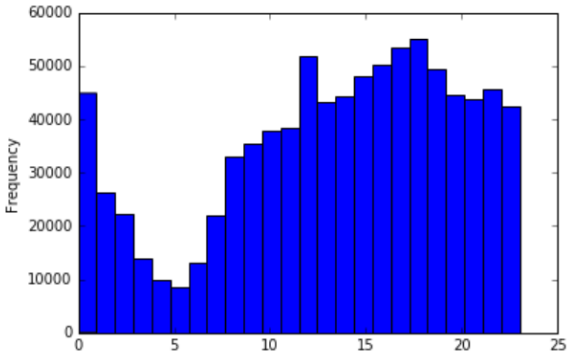

Figure 3.1: : Criminal activities occurring in different hour of day (in 24 hours format)

tivities are drastically reported around 9 AM and it continues to show a gradual increase throughout the day peaking at 6 PM, after which it starts to decrease.

'Latitude' (X), 'Longitude' (Y) have more than 30,000 unique entries of the total 878049 entries, while 'Address' has total 23228 unique Entries. One problem with the location features is, there are 26,533 entries for a specific address. Which is the location of San Francisco Police Officer's Association. This default address gives data a low variance problem.

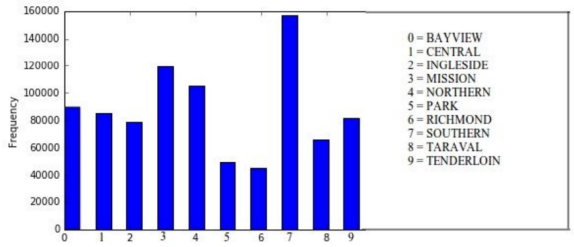

Figure 3.2: Crimes occurring in different police district

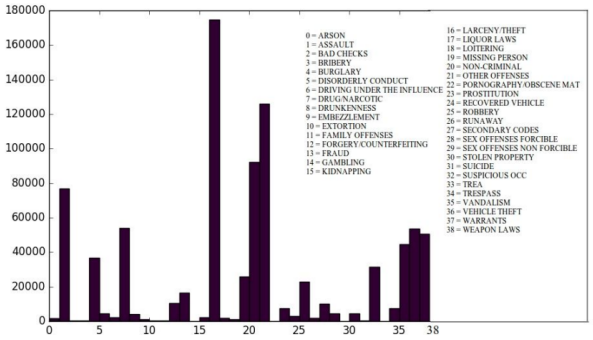

Figure 3.3: : Frequency of crime categories

Among the ten police districts, criminal activities in the southern district is higher than any other district.

Python library Scikit-learn (sklearn) is used for preprocessing the dataset. Some attributes in the csv files contains string values and others are numeric values. In order to use this dataset in machine learning models, the text features need to be converted into a numeric value.. Python library numpy is used to contain both features and label of the dataset after converting them into numeric values.

Chapter 4

Methodology

Attributes with string data type are “Day”, “Category”, “Addresss” columns. Scikit-learn has a preprocessing package that converts string data into numeric data. This package gives an integer value to each unique item after sorting items in ascending alphabetical order. Datetime attribute is also a string data type, however this is converted into a datetime object and four different attributes are obtained from it: “Hour”,”Date”,”Month” and “Year”.
To avoid overfitting and getting more realistic accuracy, the dataset is divided into two portion: testing dataset and training dataset. Training dataset contains all features along with the target label. Testing dataset only contains the features from which a machine learning model predicts the target label. Scikit-learn’s modelselection package contains a class testtrainsplit that splits the original dataset into testing and training dataset. The default value of the test dataset size is 25 percent of the original dataset. This default value is used in the conducted experiments. After that different classifiers are used and measured the performance.

Chapter 5

Experiments and Results

Decision Tree:
Sklearn.tree module provides DecisionTreeClassifier class 5.1. Among many parameters of this class, two parameters are useful in this case: min samples split indicates the number of splits to make at each step of building a decision tree and criterion indicates the function to measure the quality of split.

F1 Score	Test data accuracy	Train data accuracy
23.465	23.465	23.665

Table 5.1: Decision Tree Classification result

LogisticRegression:
sklearn linear model LogisticRegression5.2 class is used for this model. The parameter multi class is set to ovr which provides one vs the rest scheme, as multi class model is needed. Class weight parameter makes classes balanced in case of imbalanced classes.

F1 Score	Test data accuracy	Train data accuracy
21.116	21.116	21.307

Table 5.2: Logistic Regression result

k-Nearest Neighbor:
KNearestNeighbors class in sklearn.neighbors module provides supervised nearest neighbors classification models using k nearest neighbors.5.3

F1 Score	Test data accuracy	Train data accuracy
18.662	18.662	41.453

Table 5.3: k-Nearest Neighbor result

Random Forest: sklearn.ensemble provides RandomForestClassifier5.4

F1 Score	Test data accuracy	Train data accuracy
27.605	27.605	87.529

Table 5.4: Random Forest result

Gaussian Naive Bayes:

Sklearn.naive-bayes provides GaussianNB class.5.5

F1 Score	Test data accuracy	Train data accuracy
19.824	19.824	19.951

Table 5.5: Gaussian Naive Bayes result

Comparing the results Random forest gives the best result.

Chapter 6

Future Work and Conclusion

For this research, only crime data has been used, but as many researched have showed that a particular area's socio-economic standard is also a key indicator of possible criminal activity. This machine learning agent could incorporate those data and might perform better. This model can be also used for other geographic locations. This would also help to analyze crimes occurring in different locations and build a better understanding of different crimes and its relation with particular demography. 50 Also, there are many advanced machine learning approaches that can be explored. Deep Learning Neural Networks can provide a more balanced understanding of criminal activities. As it has been seen on this research, imbalanced classes has been a major issue in dealing with the particular database. Advanced techniques to deal with imbalanced classes are also something that remains to be explored

References

[1] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: Towards crime prediction from demographics and mobile data," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, (New York, NY, USA), p. 427–434, Association for Computing Machinery, 2014.

[2] A. Bogomolov, B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland, "Once upon a crime: towards crime prediction from demographics and mobile data," in *Proceedings of the 16th international conference on multimodal interaction*, pp. 427–434, 2014.

[3] R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. S. Panahy, and N. Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," *Indian Journal of Science and Technology*, vol. 6, no. 3, pp. 4219–4225, 2013.

[4] A. Nasridinov and Y.-H. Park, "A study on performance evaluation of machine learning algorithms for crime dataset," *Adv. Sci. Technol. Lett.-(Networking Commun. 2014)*, vol. 66, pp. 90–92, 2014.

[5] X. Wang, M. S. Gerber, and D. E. Brown, "Automatic crime prediction using events extracted from twitter posts," in *International conference on social computing, behavioral-cultural modeling, and prediction*, pp. 231–238, Springer, 2012.

[6] M. Redmond and A. Baveja, "A data-driven software tool for enabling cooperative information sharing among police departments," *European Journal of Operational Research*, vol. 141, no. 3, pp. 660–678, 2002.

[7] Kaggle, *San Francisco Crime Classification*. PhD thesis, https://www.kaggle.com/c/sf-crime/overview, 2017.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Sunday 26th September, 2021 at 7:57am.