A ROBUST CRIME FORECASTING AND ANALYSIS USING MACHINE LEARNING AND CV

SK. KHADER BASHA[1], K. AJAY KUMAR[2], B. YUVA KISHORE BABU[3], B. HARINADH[4], A. UPENDRA[5].

[1] Assistant Professor, CSE,Chalapathi Institute of Technology,Guntur, India
[2]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[3]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[4]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India
[5]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

ABSTRACT: Data mining and Machine learning have become a vital part of crime detection and prevention. The purpose of this project is to evaluate data mining methods and their performances that can be used for analyzing the collected data about the past crimes. Identified the most appropriate data mining methods to analyze the collected data from sources specialized in crime prevention by comparing them theoretically and practically. Some attributes of this dataset are gender, age, employment status, crime place. Methods are applied on these data to determine their effectiveness in analyzing and preventing crime. Evaluations on the data showed that the method with a higher performance is "Decision Tree".

This was achieved by some performance measures, such as the number of instances correctly classified, accuracy or precision and recall that has brought better results compared to other methods. I come to the conclusion that the data mining methods contribute to the predictions on the possibility of occurrence of the crime and as a result in its prevention.

Keywords: Crime Prediction; Machine Learning; Decision tree; J48; Artificial Intelligence; Classification Algorithms.

1. INTRODUCTION

The increase in crime data recording coupled with data analytics resulted in the growth of research approaches aimed at extracting knowledge from crime records to better understand criminal behavior and ultimately prevent future crimes.

Crime is a complex social phenomenon that has grown due to major changes in society. Law enforcement agencies need to learn the factors that lead to an increase in crime tendency. To curb this, there is always a need for strategies and policies to prevent crime. As a result of technology development, science and information, data mining and artificial intelligence tools are increasingly prevalent in the law enforcement community.

Law enforcement agencies face a large volume of data that needs to be processed and turned into useful information, and data mining can improve crime analysis by helping to predict and prevent it. By processing criminal data, law enforcement agencies can use models thatmay be important in the crime prevention process.

The use of data mining accelerates data analysis, and analysts can examine existing data to identify patterns and trends of crime. This project is structured as follows: It describes the relationship that exists between data mining, machine learning and criminology. The methodology and description of the dataset are described in it. Next, it represent a theoretical description of the methods and algorithms that will be applied practically to our data presents the results of the application of algorithms and an explanation for the algorithm with the best results.

2. LITERATURE SURVEY

Using machine learning algorithms to analyze crime data by McClendon, Lawrence, and Natarajan Meghan than.

Data mining and machine learning have become a vital part of crime detection and prevention. In this research, we use WEKA, an open source data mining software, to conduct a comparative study between the violent crime patterns from the Communities and Crime Un normalized Dataset provided by the University of California-Irvine repository and actual crime statistical data for the state of Mississippi that has been provided by neighborhoodscout.com.

We implemented the Linear Regression, Additive Regression, and Decision Stump algorithms using the same finite set of features, on the Communities and Crime Dataset. Overall, the linear regression algorithm performed the best among the three selected algorithms. The scope of this project is to prove how effective and accurate the machine learning algorithms used in data mining analysis can be at predicting violent crime patterns.

Learning to detect patterns of crime by Wang, Tong, et al.

We introduce a novel, robust data-driven regularization strategy called Adaptive Regularized Boosting (AR-Boost), motivated by a desire to reduce overfitting. We replace AdaBoost's hard margin with a regularized soft margin that trades-off between a larger margins, at the expense of misclassification errors. Minimizing this regularized exponential loss results in a boosting algorithm that relaxes the weak learning assumption further: it can use classifiers with error greater than 12. This enables a natural extension to multiclass boosting, and further reduces over fitting in both the binary and multiclass cases. We derive bounds for training and generalization errors, and relate them to Adobos.

Finally, we show empirical results on benchmark data that establish the robustness of our approach and improved performance overall. 1 Introduction Boosting is a popular method for improving the accuracy of a classifier. In particular, AdaBoost is considered the most popular form of boosting and it has been shown to improve the performance of base learners both theoretically and empirically. The key idea behind AdaBoost is that it constructs a strong classifier using a set of weak classifiers.

3. EXISTING SYSTEM:

KNN, RF, SVM and Bayes models are existing methods although studies have been done in the medical field with an advanced data exploration using machine learning algorithms, orthopedic disease prediction is still a relatively new area and must be explored further for the accurate prevention and cure. It mines the double layers of hidden states of vehicle historical trajectories, and then selects the parameters of Hidden Markov Model (HMM) by the historical data. In addition, it uses a Viterbi algorithm to find the double layers hidden states sequences corresponding to the just driven trajectory. Finally, it proposes a new algorithm for vehicle trajectory prediction based on the hidden Markov model of double layers hidden states, and predicts the nearest neighbor unit of location information of the next k stages.

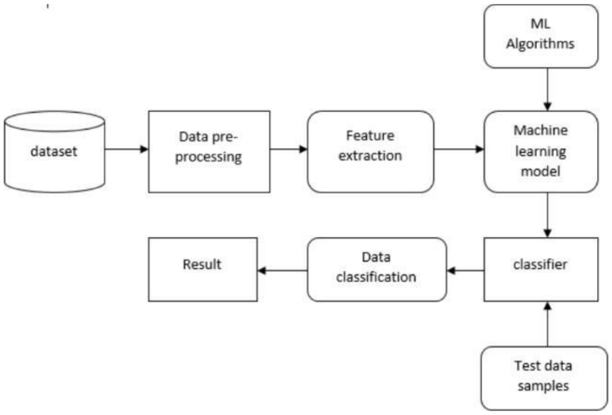
4. PROBLEM STATEMENT:

Crimes now a days are increasing day by day and with different level of Intensity and versatility. The result is a great loss to society in terms of monitory loss, Social loss and further it enhances the level of threat against the smooth livelihood in the society. To overcome this problem, the computing era can help to reduce the crime or even may be helpful in predicting the crime so that sufficient measures can be taken to minimize the loss to property and life. The crime rate prediction strategies can be applied on historical data available in the police records by examining the data at various angles like reason of crime, frequency of similar kind of crimes at specific location with other parameters to prepare the model crime prediction. It is a major challenge to understand the versatile data available with us, then model it to predict the future incidence with acceptable accuracy and further to reduce the crime rate.

5. PROPOSED SYSTEM:

The proposed system is made on the basis of the research work that is done by going through various such documentations. Nearly all of the crimes are predicting based on the location and the types of crimes that are occurring in those areas. On surveying previous works, Linear Regression, Decision Tree and Random Forest tend to give good accuracy so these models are used in this paper to predict crimes. The data set contains different types of crimes that being committed in India according to the state and year respectively. This paper takes types of crimes as input and gives the area in which crimes are committed as output. The data pre-processing involves data cleaning, features election, dropping null values, data scaling by normalizing and standardizing. After data preprocessing the data is free of null values which may alter the accuracy of the model significantly and feature selection is used to select only the required features that won't affect the accuracy of model. After data pre-processing the models chosen i.e. Logistic Regression, Decision Tree and Random Forest are trained by splitting the data into as train and test data. As the output required is a categorical value classification models are used here. Python language is used for the data prediction

6. SYSTEM ARCHITECTURE:



.

7. IMPLEMENTATION

Data Collection

The crime data collection is achieved using the third-party API provided by New York City Police Department (NYPD) at the NYC Open Data portal which is reserved for free federal data to involve civilians in the reports generated and managed by the city administration. This dataset comprises all credible transgressions delivered to the New York City Police Department (NYPD). The dataset is updated every three months. The variables stored in the dataset comprise of the following:- the name of the borough in which the incident occurred, the date and time of occurrence for the reported event, an intimation of whether the crime was interrupted prematurely, attempted but failed, or completed successfully, the level of offense, the specific location of occurrence in or around the premises, the description of the crime, the date of reporting of the event, the victim and the suspect's age group, race description, and sex description, and the latitude and longitude of the crime incident.

Data Preprocessing

Data preprocessing includes reconstructing primary data to proper data sets since machines cannot use data that they cannot interpret. Primary data is usually deficient and has incongruous formatting. The adequacy or inadequacy of data preparation is associated with the success of every project that requires analysis or prediction of data. Data Preprocessing comprises both validation and imputation of data. The purpose of validation is to evaluate whether the data is both comprehensive and precise. The purpose of the imputation of data is to rectify errors and input missing values for the preprocessing of the dataset, we split it into two separate datasets, one dataset for analysis and the other for prediction. For the prediction dataset, there were quite a few missing values. Instead of dropping entire rows, we replaced missing values with "UNKNOWN" values to avoid the loss of data. We also used the date and time to add the year, month, day of the week, the part of the day, and the hour at which the crime took place for better analysis. For the analysis dataset, we took the date, time, latitude, longitude, category, and description of the crime.

Since analysis prefers the data to be in numerical format, we created dummies for the crime categories and descriptions. We also used the date and time to add the year, month, and day of the week, part of the day, and the hour in a numerical format.

Data Set

A data set (or dataset) could be an assortment of knowledge. most ordinarily a knowledge set corresponds to the contents of one information table, or one applied mathematics information matrix, wherever each column of the table represents a specific variable, and every row corresponds to a given member of the information set in question. the information set lists values for every of the variables, like height Associate in Nursing weight of an object, for every member of the information set. every price is understood as a data point. Set might comprise data for one or additional members, appreciate the quantity of rows.

The dataset consists of the following details about the crime incidents:

Category - category of the crime incident. This is the target variable which is going to be
predicted.

Descript - description of the crime incident.

Day Of Week - the day of the week.

PD District - name of the Police Department District

Resolution - how the crime incident was resolved.

Address - the approximate street address of the crime incident.

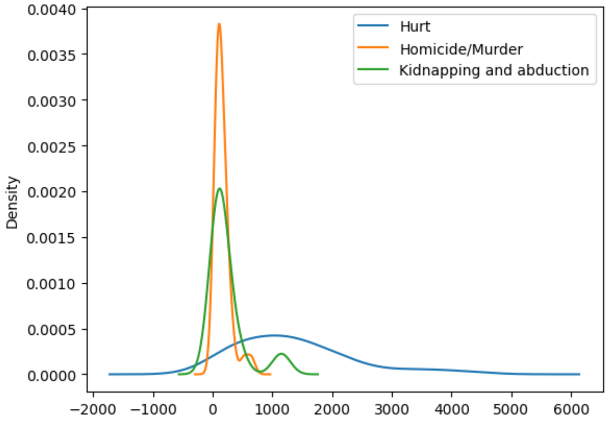
FEATURE EXTRATION

Next thing is to do Feature extraction is an attribute reduction process. Unlike feature selection, which ranks the existing attributes according to their predictive significance, feature extraction actually transforms the attributes. The transformed attributes, or features, are linear combinations of the original attributes. Finally, our models are trained using Classifier algorithm.

EVALUATION MODEL

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and over fitted models. Performance of each classification model is estimated base on its averaged. The result will be in the visualized form. Representation of classified data in the form of graphs

8. SCREEN SHORT



9. CONCLUSION

Crime prediction is one the current trends in the society. Crime prediction intends to reduce crime occurrences. It does this by predicting which type of crime may occur in future. Here, analysis of crime and prediction are performed with the help of various approaches. From the results obtained we saw that the training time of SVM is very high thus it should be avoided for this dataset. However which model will work best is totally dependent on the dataset that is being used. In this system, we get to classify and cluster to improve the accuracy of location and pattern-based crimes. This software predicts frequently occurring crimes, especially for particular state, and occurrences.

FUTURE SCOPE

As of now, the project relies on manual input from a human (a police officer) in order to enter details in the database. If we can make this a centralized system and connect it to all the police stations countrywide and make FIR reporting digital, then it would be quite easier to predict crimes in that particular location and recognize patterns in them. It would also encourage citizens to track their E-FIR online. We can also avoid corruption as the government can keep a track on the number of cases registered and their solvability rate which can help them utilize their resources better.

REFERENCES

[1] K. Zakir Hussain, M. Durairaj and G. R. J. Farzana, "Criminal behavior analysis by using data mining techniques," IEEE-International Conference on Advances in Engineering, Science and Management (ICAESM -2012), Nagapattinam, Tamil Nadu, 2012, pp. 656-658

[2] Keyvanpour, Mohammad & Javideh, Mostafa & Ebrahimi, Mohammadreza. (2011). Detecting and investigating crime by means of data mining: A general crime matching framework. Procedia CS. 3. 872-880. 10.1016/j.procs.2010.12.143.

[3] Ioannis Kavakiotis OlgaTsave Athanasios Salifoglou Nicos Maglaveras Ioannis Vlahavas Ioanna Chouvarda, Machine Learning and Data Mining Methods in Diabetes Research, Computational and Structural Biotechnology Journal Volume 15, 2017, Pages 104-116.

[4] Frank, Eibe & Hall, Mark & Holmes, Geoffrey & Kirkby, Richard & Pfahringer, Bernhard & Witten, Ian & Trigg, Len. (2010). Weka-A Machine Learning Workbench for Data Mining. 10.1007/978-0-387- 09823-4_66.

[5] Pang-Ning Tan; Michael Steinbach; Anuj Karpatne; Vipin Kuma Introduction to

Data Mining 2 nd ed, Publisher: Pearson, 2019, Print ISBN: 9780133128901, 0133128903 e-text ISBN: 9780134080284, 013408028.

[6] M. Kantardzic, Data Mining Concepts, Models, Methods, and Algorithms, 2 nd ed, John Wiley & Sons, Inc., Hoboken, New Jersey 2011, ISBN 978-0-470-89045-5, oBook ISBN: 978-1-118-02914-5, ePDF ISBN: 978-1-118-02912-1, ePub ISBN: 978-1-118-02913-8.