IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 11, Issue 6, November-December 2024

Impact Factor: 7.394

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106044

Crime Hotspots Prediction Using Machine Learning

Mr.Lakshmipathi[1], Ms.T.Sushma[2], Ms.V. Lakshmi Manasa[3], Ms. S.Priyadarshini[4]

Assistant Professor, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India[1]

Student, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India[2,3,4]

ABSTRACT: Crime prediction is of great significance to the formulation of policing strategies and the implementation of crime prevention and control. Machine learning is the current mainstream prediction method. However, few studies have systematically compared different machine learning methods for crime prediction. This paper takes the historical data of public property crime from a section of a large coastal city in the southeast of China as research data to assess the predictive power between several machine learning algorithms. Results based on the historical crime data alone suggest that the LSTM model outperformed KNN, random forest, support vector machine, naive Bayes, and convolutional neural networks. In addition, the built environment data of points of interests (POIs) and urban road network density are input into LSTM model as covariates. It is found that the model with built environment covariates has better prediction effect compared with the original model that is based on historical crime data alone. Therefore, future crime prediction should take advantage of both historical crime data and covariates associated with criminological theories. Not all machine learning algorithms are equally effective in crime prediction.

KEYWORDS: *Crime Hotspot Prediction, Machine Learning,Support Vector Machine (SVM), Random Forest.*

I. INTRODUCTION

The research on crime prediction currently focuses on two major aspects: crime risk area prediction, and crime hotspot prediction. The crime risk area prediction, based on the relevant influencing factors of criminal activities, refers to the correlation between criminal activities and physical environment, which both derived from the ``routine activity theory''. Traditional crime risk estimation methods usually detect crime hotspots from the historical distribution of crime cases, and assume that the pattern will persist in the following time periods. For example, considering the proximity of crime places and the aggregation of crime elements, the terrain risk model tends to use crime-related environmental factors and crime history data, and is relatively effective for long-term, stable crime hotspot prediction. Many studies have carried out empirical research on crime prediction in different time periods, combining demographic and economic statistics data, land use data, mobile phone data and crime history data. Crime hotspot prediction aims to predict the likely location of future crime events and hotspots where the future events would concentrate. A commonly used method is kernel density estimation. A model that considers temporal or spatial autocorrelations of past events performs better than those that fail to account for the autocorrelation. Recently machine learning algorithms have gained popularity. The most popular methods include K-Nearest Neighbor(KNN), random forest algorithm, support vector machine (SVM), neural network and Bayesian model etc.. Some compared the linear methods of crime trend prediction, some compared Bayesian model and BP neural network and others compared the spatiotemporal kernel density method with the random forest method in different periods of crime prediction

II. LITERATURE SURVEY

U. **Thongsatapornwatana(2016)**In recent years the data mining is data analyzing techniques that used to analyze crime data previously stored from various sources to find patterns and trends in crimes. In additional, it can be applied to increase efficiency in solving the crimes faster and also can be applied to automatically notify the crimes. However, there are many data mining techniques. In order to increase efficiency of crime detection, it is necessary to select the data mining techniques suitably. This paper reviews the literatures on various data mining applications, especially applications that applied to solve the crimes. Survey also throws light on research gaps and challenges of crime data mining. In additional to that, this paper provides insight about the data mining for finding the patterns and trends in crime to be used appropriately and to be a help for beginners in the research of crime data mining.

J.M. Caplan, L. W. Kennedy, and J. Miller (2011) The research presented here has two key objectives. The first is to apply risk terrain modeling (RTM) to forecast the crime of shootings. The risk terrain maps that were produced from RTM use a range of contextual information relevant to the opportunity structure of shootings to estimate risks of future shootings as they are distributed throughout a geography. The second objective was to test the predictive power of the risk terrain maps over two six-month time periods, and to compare them against the predictive ability of retrospective

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106044

hot spot maps. Results suggest that risk terrains provide a statistically significant forecast of future shootings across a range of cut points and are substantially more accurate than retrospective hot spot mapping. In addition, risk terrain maps produce information that can be operationalized by police administrators easily and efficiently, such as for directing police patrols to coalesced high-risk areas.

M. Cahill and G. Mulligan *(2007)* The present research examines a structural model of violent crime in Portland, Oregon, exploring spatial patterns of both crime and its covariates. Using standard structural measures drawn from an opportunity framework, the study provides results from a global ordinary least squares model, assumed to fit for all locations within the study area. Geographically weighted regression (GWR) is then introduced as an alternative to such traditional approaches to modeling crime. The GWR procedure estimates a local model, producing a set of mappable parameter estimates and t-values of significance that vary over space. Several structural measures are found to have relationships with crime that vary significantly with location. Results indicate that a mixed model— with both spatially varying and fixed parameters—may provide the most accurate model of crime. The present study demonstrates the utility of GWR for exploring local processes that drive crime levels and examining misspecification of a global model of urban violence.

A. Almehmadi, Z. Joudaki, and R. Jalali (2017) Social networks 1 produce enormous quantity of data. Twitter, a microblogging network, consists of over 230 million active users posting over 500 million tweets every day. We propose to analyze public data from Twitter to predict crime rates. Crime rates have increased in the past recent years. Although crime stoppers are utilizing various technics to reduce crime rates, none of the previous approaches targeted utilizing the language usage (offensive vs. non-offensive) in Tweets as a source of information to predict crime rates. In this paper, we hypothesize that analyzing the language usage in tweets is a valid measure to predict crime rates in cities. Tweets were collected for a period of 3 months in the Houston and New York City by locking the collection by geographic longitude and latitude. Further, tweets regarding crime events in the two cities were collected for verification of the validity of the prediction algorithm. We utilized Support Vector Machine (SVM) classifier to create a model of prediction of crime rates based on tweets. Finally, we report the validity of prediction algorithm in predicting crime rates in cities.

Existing System

- Liu et al. Compared the random forest and spatiotemporal KDE method, found that the random forest algorithm is more efficient than the traditional spatiotemporal KDE method in the smaller time scale and grid space unit.
- Gabriel et al. used the Gated Localized Diffusion Network for crime prediction at the street segment level.
- Compared with the traditional Network-time KDE method, the diffusion network approach significantly increased the prediction accuracy. The ability of machine learning algorithm in processing non-linear relational data has been confirmed in many fields, including crime prediction. It has a faster training speed, can handle very high-dimensional data, and can also extract the characteristics of the data.

Disadvantages

- Existing system has no research that at the same time (i) evaluates its efficiency against a traditional hotspot policing approach implemented by the police and (ii) provides a clear breakdown of the processing steps involved to implement such a predictive system.
- Small police departments, which often have more worrying demands for violence, may not be able to provide more efficient tools. If they want to build a prediction system, it can cost even more than buying one and they can take much time to build.

Proposed System

- Crime hotspot prediction aims to predict the likely location of future crime events and hotspots where the future events would concentrate. In this paper, random forest algorithm is used for crime prediction.
- The randomness of random forest is reflected in two aspects: one is to randomly select the training sample set by using bagging algorithm; the other is to randomly select the split attribute set. Assuming that the training sample has M attributes in total, we specify an attribute number F<_M, in each internal node, randomly select F attributes from M attributes as the split attribute set, and take the best split mode of the f attributes Split the nodes. The multi decision tree is made up of random forest, and the final classification result is determined by the vote of tree classifier.
- The objective would be to train a model for prediction. The training would be done using the training data set which will be validated using the test dataset. Building the model will be done using better algorithm depending

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106044

upon the accuracy. The Random Forest will be used for crime prediction. Visualization of dataset is done to analyze the crimes which may have occurred in the country.

Advantages

- The purpose of this work is to improve our previously proposed prediction framework through alternative crime mapping and feature engineering approaches, and provide an open-source implementation that police analysts can use to deploy more effective predictive policing.
- This work helps the law enforcement agencies to predict and detect crimes in India with improved accuracy and thus reduces the crime rate.

System Architecture

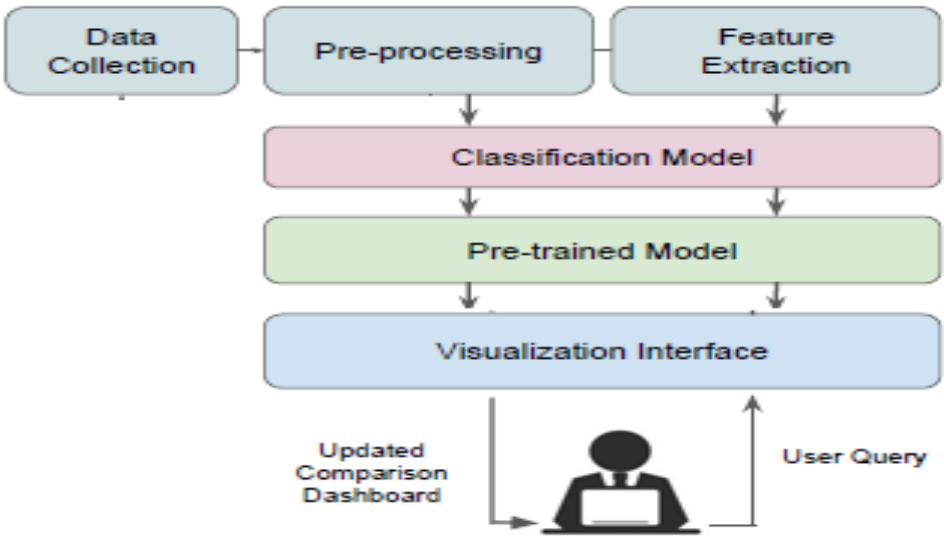

Fig 1: System Architecture

III. METHODOLOGY

Modules Name:

- **Data Collection**
- **Dataset**
- **Data Preparation**
- **Model Selection**
- **Analyze and Prediction**
- **Accuracy on test set**
- **Saving the Trained Model**

1. Data Collection

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.There are several techniques to collect the data, like web scraping, manual interventions and etc.Comparison of Machine Learning Algorithms for Predicting Crime Hotspots taken from kaggle and some other source.

2. Dataset

The dataset consists of 821 individual data. There are 27 columns in the dataset, which are described below.
STATE : State in India
DISTRICT : District in the state of India.
Year : 2001-2018
MURDER : Total number of murder rate

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106044

RAPE : Total number of rape rate
THEFT : Total number of theft rate
Total crime : Total number of total crime rate

3. Data Preparation

We will transform the data. By getting rid of missing data and removing some columns. First we will create a list of column names that we want to keep or retain.Next we drop or remove all columns except for the columns that we want to retain. Finally we drop or remove the rows that have missing values from the data set.

4. Model Selection

While creating a machine learning model, we need two dataset, one for training and other for testing. But now we have only one. So lets split this in two with a ratio of 80:20. We will also divide the dataframe into feature column and label column.

Here we imported train_test_split function of sklearn. Then use it to split the dataset. Also, *test_size = 0.2*, it makes the split with 80% as train dataset and 20% as test dataset.

The *random_state* parameter seeds random number generator that helps to split the dataset.

The function returns four datasets. Labelled them as *train_x, train_y, test_x, test_y*. If we see shape of this datasets we can see the split of dataset.

We will use Random Forest Classifier, which fits multiple decision tree to the data. Finally I train the model by passing *train_x, train_y* to the *fit* method.

Once the model is trained, we need to Test the model. For that we will pass *test_x* to the predict method.

Random Forest is one of the most powerful methods that is used in machine learning for regression problems. The random forest comes in the category of the supervised regressor algorithm. This algorithm is carried out in two different stages the first one deals with the creation of the forest of the given dataset, and the other one deals with the prediction from the regressor.

5. Analyze and Prediction

In the actual dataset, we chose only 3 features :
STATE :State in India
DISTRICT: District in the state of India.
Year : 2001-2018
Prediction :
1.Total number of murder rate
2.Total number of rape rate
3.Total number of theft rate
4.Total number of total crime rate

6. Accuracy on test set:

1. We got a accuracy of 95.1%,97.1%, 98.1%, 96.5%, on test set.

7. Saving the Trained Model:

Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 or .pkl file using a library like pickle .Make sure you have pickle installed in your environment.

Next, let's import the module and dump the model into .pkl file

Implementation

The provided Python script creates a simple Flask web application for predicting crime hotspots using a pre-trained machine learning model. It begins by importing necessary modules: This Flask application provides a platform for crime prediction and analysis using pre-trained machine learning models. It uses different models to predict crimes like murder, theft, rape, and total crimes, allowing users to upload datasets and preview them, as well as interact with individual predictive features through dedicated pages. The app includes routes for rendering various HTML templates (e.g., first, login, chart) to guide users through functionalities like uploading files, previewing data, and making predictions

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106044

Key features of this application include:

File Upload and Preview: Users can upload a CSV file, which is previewed on the "preview.html" page after being read into a pandas DataFrame.

Crime Prediction: There are routes for predicting specific crimes (murder, theft, rape) and total crime counts using machine learning models loaded via pickle.

Modular Templates: HTML templates like crime.html, murder.html, an others ensure clear navigation and separation of functionalities.

Machine Learning Integration: Models are pre-trained and saved as .pkl files (murder.pkl, theft.pkl, etc.), which are loaded and used for predictions via POST requests.

Interactive Web Application: With dynamic input handling through forms and result rendering, users can input features and get real-time crime predictions.

Data Scaling: A MinMaxScaler is loaded but not currently used in the provided code, which suggests potential normalization of features.

This application is ideal for analyzing historical crime data and making future predictions in an interactive and user-friendly manner. However, there are some corrections needed:

- Replace _name_ with __name__ to properly initialize the Flask app.
- Ensure secure handling of uploaded files and inputs to prevent potential security risks.

IV. ALGORITHM USED

Existing Algorithm

Spatio temporal Kernel Density Estimation (KDE):

It an extension of kernel density estimation that accounts for both spatial and temporal dimensions of data.

It is used to estimate the probability density function of events distributed over space and time.

This is particularly useful in fields like criminology, epidemiology, and environmental science where data is spatially and temporally explicit..

Proposed Algorithm

Random Forest algorithm:

- Random Forest works by constructing a multitude of decision trees during training. Each tree is trained on a subset of the training data, sampled with replacement (bootstrap sample), and a subset of features randomly selected at each node. This randomness ensures that each tree in the forest learns different patterns from the data.
- During prediction, the output of each decision tree is aggregated to make the final prediction. For classification tasks, the most common class among the predictions of individual trees is selected as the final prediction. For regression tasks, the average of the predictions from all trees is taken as the final prediction

V. EXPERIMENTAL RESULTS

This project is implements like web application using Python and the Server process is maintained using the SOCKET & SERVERSOCKET and the Design part is played by Cascading Style Sheet.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 7.394| A Bi-Monthly, Double-Blind Peer Reviewed & Referred Journal |

|| Volume 11, Issue 6, November-December 2024 ||

DOI:10.15680/IJARETY.2024.1106044

VI. CONCLUSION

With the help of machine learning technology, it has become easy to find out relation and patterns among various data's. The work in this project mainly revolves around predicting the type of crime which may happen if we know the location of where it has occurred. Using the concept of machine learning we have built a model using training data set that have undergone data cleaning and data transformation. The model predicts the type of crime with Good Accuracy. Data visualization helps in analysis of data set. The graphs include bar, pie, line and scatter graphs each having its own characteristics. We generated many graphs and found interesting statistics that helped in understanding Indian crimes datasets that can help in capturing the factors that can help in keeping society safe.

VII. FUTURE ENHANCEMENT

For the future research, there are still some aspects to be improved. The first is the temporal resolution of the prediction. Felson et al. revealed that the crime level changes with time. Some studies have shown that it is useful to check the variation of risks during the day. We chose two weeks as the prediction window. It does not capture the impact of crime changes within a week, let alone the change within a day. The sparsity of data makes the prediction of crime event difficult if the prediction window is narrowed down to day of a week or hour within a day. There is no viable solution to this challenging problem at this time. The second is the spatial resolution of the grid. In this paper, the grid size is 150m * 150m. Future research will assess the impact of changing grid sizes on prediction accuracy. Third, the robustness and generality of the findings of this paper needs to be tested in other study areas. Nonetheless, the findings of this research have proven to be useful in a recent hotspot crime prevention experiment by the local police department at the study size.

REFERENCES

[1] U. Thongsatapornwatana, ``A survey of data mining techniques for analyzing crime patterns," in Proc. 2nd Asian Conf. Defence Technol. (ACDT), Jan. 2016, pp. 123128.
[2] J. M. Caplan, L. W. Kennedy, and J. Miller, ``Risk terrain modeling: Brokering criminological theory and GIS methods for crime forecasting," Justice Quart., vol. 28, no. 2, pp. 360381, Apr. 2011.
[3] M. Cahill and G. Mulligan, ``Using geographically weighted regression to explore local crime patterns," Social Sci. Comput. Rev., vol. 25, no. 2, pp. 174193, May 2007.
[4] A. Almehmadi, Z. Joudaki, and R. Jalali, ``Language usage on Twitter predicts crime rates," in Proc. 10th Int. Conf. Secur. Inf. Netw. (SIN), 2017, pp. 307310.
[5] H. Berestycki and J.-P. Nadal, ``Self-organised critical hot spots of criminal activity," Eur. J. Appl. Math., vol. 21, nos. 45, pp. 371399, Oct. 2010.
[6] K. C. Baumgartner, S. Ferrari, and C. G. Salfati, ``Bayesian network modeling of offender behavior for criminal proling," in Proc. 44th IEEE Conf. Decis. Control, Eur. Control Conf. (CDC-ECC), Dec. 2005, pp. 27022709.
[7] W. Gorr and R. Harries, ``Introduction to crime forecasting," Int. J. Fore- casting, vol. 19, no. 4, pp. 551555, Oct. 2003.
[8] W. H. Li, L.Wen, and Y. B. Chen, ``Application of improved GA-BP neural network model in property crime prediction," Geomatics Inf. Sci. Wuhan Univ., vol. 42, no. 8, pp. 11101116, 2017.
[9] R. Haining, ``Mapping and analysing crime data: Lessons from research and practice," Int. J. Geogr. Inf. Sci., vol. 16, no. 5, pp. 203507, 2002.
[10] S. Chainey, L. Tompson, and S. Uhlig, ``The utility of hotspot mapping for predicting spatial patterns of crime," Secur. J., vol. 21, nos. 12, pp. 428, Feb. 2008.
[11] S. Chainey and J. Ratcliffe, ``GIS and crime mapping," Soc. Sci. Comput. Rev., vol. 25, no. 2, pp. 279282, 2005.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 7.394

www.ijarety.in editor.ijarety@gmail.com