IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 1, July 2023

Crime Analysis and Prediction using Data Mining

Yogesh Krishna Mandavkar
Student, Department of MCA
Late Bhausaheb Hiray S. S. Trust's Institute of Computer Application, Mumbai, India

Abstract: *Crime analysis and vaticination is a methodical approach for relating the crime. This system can prognosticate regions with a high probability of crime and fantasize about crime-prone areas. Using the conception of data mining we can prize preliminarily unknown, useful information from unshaped data. The birth of new information is prognosticated using the datasets. Crimes are an unfaithful and common social problem faced worldwide. Crimes affect the quality of life, profitable growth, and character of the nation. With the end of securing society from crimes, there's a need for advanced systems and new approaches for perfecting crime analytics for guarding their communities. We propose a system that can analyze, descry, and prognosticate colorful crime probabilities in the given region. This paper explains colorful types of felonious analysis and crime vaticination using several data mining ways.*

Keywords: Crime Prediction, K-Means, Linear Regression.

I. INTRODUCTION

Day by day crime data rate is adding because the ultramodern technologies and hi-tech styles are helping the culprits to achieving illegal conditioning. according to the Crime Record Bureau crimes like burglary, wildfire, etc. have increased while crimes like murder, coitus, abuse, gang rap, etc. have increased. crime data will be collected from colorful blogs, news, and websites. The huge data is used as a record for creating a crime report database. The knowledge which is acquired from the data mining ways will help in reducing crimes as it helps in chancing the lawbreakers briskly and also the areas that are most affected by crime.. Data mining helps in working the crimes briskly and this fashion gives good results when applied on crime datasets, the information attained from the data mining ways can help the police department.

A particular approach has been set up to be useful by the police, which is the identification of crime ‘ hot spots ‘ which indicates areas with a high attention of crime

The use of data booby-trapping ways can produce important results from crime report datasets. The very step in the study of crime is crime analysis. Crime analysis is exploring, interrelating, and detecting relationship between the various crimes and the characteristics of the crime. This analysis helps in preparing statistics, queries and maps on demand. It also helps to see if a crime in a certain known pattern or a new pattern is necessary.

Crimes can be prognosticated as the miscreant are active and operate in their comfort zones. formerly successful they try to replicate the crime under analogous circumstances. The circumstances of crime depended on several factors similar to the intelligence of culprits, security of a position, etc. The work has followed the way that is used in data analysis, in which the important phases are Data collection, data bracket, pattern identification, vaticination, and visualization. The proposed frame uses different visualization ways to show the trends of crimes and colorful ways that can predict crime using a machine literacy algorithm.

II. CRIME DATA ANALYSIS

The collection and analysis of crime-related data play a vital role in ensuring the effectiveness of law enforcement agencies. To achieve this, it is essential to adopt a cohesive methodology for classifying data based on the frequency and geographical distribution of incidents, as well as for identifying recurring patterns among interconnected crimes at different time intervals. Moreover, accurately predicting future relationships between crimes is of utmost importance. Among the various methodologies employed, hot spot analysis emerges as a popular choice. Additionally, point pattern analysis and clustering, along with distance statistics, have proven to be effective in uncovering patterns and trends within crime data. Moreover, the utilization of data mining, text mining, spatial analysis, and self-organizing maps

DOI: 10.48175/IJARSCT-12027

contributes to the discovery of meaningful patterns, thus facilitating prompt identification and efficient response to crime patterns. Consequently, an efficient crime analysis tool should possess the capability to swiftly identify crime patterns, enabling timely discovery and actionable insights for law enforcement agencies.

The main purpose of crime analysis is:

- Extraction of crime pattern by crime analysis and based on available criminal information.
- Crime recognition.
- The problem of identifying techniques that can efficient and accurate.

III. SYSTEM ANALYSIS AND DESIGN

The research methodology encompasses a well-structured approach, unfolding in three fundamental stages:

- The foremost stage revolves around a diligent exploration of relevant literature, delving deep into the realm of crime data mining and analysis.
- Moving on to the second stage, a robust classification scheme is meticulously crafted, providing a systematic framework for the study.
- Ultimately, the third stage entails the presentation of a comprehensive summary of research findings in the domain of crime data mining and analysis. This culminates in the creation of an intricately detailed literature review report, showcasing a wealth of valuable insights.

IV. SOURCE OF INFORMATION

In order to discover the right insights and successful investigation, it is necessary to recognize available data sources of crime and the various types of crime

4.1 Police reports:

Police reports, such as First Information Reports (FIRs), serve as crucial repositories of information pertaining to crimes, encompassing essential details about the incident, the complainant, and the identified suspect(s). These reports are meticulously prepared by police personnel on paper, resulting in an unstructured data format. Nevertheless, despite the absence of a standardized structure, FIRs stand as a dependable and indispensable source for collecting comprehensive crime data.

4.2 Previous investigation files:

In the pursuit of identifying a previously accused suspect, law enforcement agencies employ a multifaceted approach by accessing a plethora of investigation records linked to the individual in question. These encompass a diverse range of formats, including textual documents, photographic evidence, video recordings, CCTV footage, financial records, bank statements, phone call logs, email communications, forensic reports, witness testimonies, victim statements, and even attorney declarations.

Intelligence reports form a vital component of criminal information management, with intelligence agencies in India, such as the Research and Analysis Wing (RAW), Intelligence Bureau (IB), and Narcotics Control Bureau (NCB), diligently collecting and analyzing intelligence pertaining to criminal activities.

Open source intelligence findings, derived from web-based sources, search engines, and social networking platforms like Facebook, Twitter, and LinkedIn, contribute significantly to the investigative process. However, it is important to note that the information obtained through open sources often exists in an unstructured data format, necessitating meticulous analysis and interpretation.

The meticulous recording of police arrest records comes into play when suspects are apprehended by law enforcement officers. These records are typically maintained in either relational or textual formats, offering comprehensive insights into the circumstances and specifics of each arrest.

DOI: 10.48175/IJARSCT-12027

4.3 System Planning

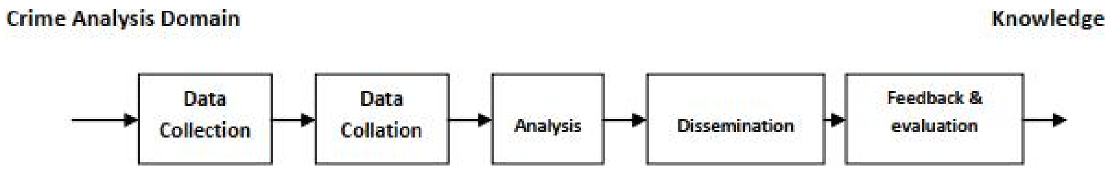


Figure 1

Data collection: The initial phase of crime analysis involves a crucial methodology: data collection. In order to gather comprehensive insights, crime data is sourced from diverse websites, news platforms, and blogs. The collected data is then stored in a database, ready for further processing. Given the inherent nature of crime data, characterized by its lack of standardized structure, it falls under the category of unstructured data. To effectively handle this type of data, object-oriented programming proves to be a highly advantageous approach, offering both flexibility and ease of use. Opting for a schema-less database emerges as a favorable choice, enabling the accommodation of varying document sizes, content, and fields present within unstructured data. Notably, the absence of complex joins in such databases further streamlines the analytical process. Embracing an unstructured database also empowers analysts to efficiently handle substantial volumes of structured, semi-structured, and unstructured data, unlocking its full potential in crime analysis.

Classification: In this particular stage, the crime analysis process incorporates the utilization of the Naive Bayes Algorithm, which serves as a supervised learning method. The Naive Bayes classifier is renowned for its probabilistic approach, as it offers a probability distribution across a set of possible classes instead of a singular output when provided with input data. A key advantage of the Naive Bayes Classifier is its simplicity, coupled with its ability to deliver faster coverage in comparison to logistic regression. Notably, it outperforms memory-intensive algorithms like Support Vector Machines (SVM), making it an efficient choice in terms of computational resources. By leveraging the Naive Bayes algorithm, a model is constructed through training on crime data encompassing a wide range of offenses, including vandalism, murder, robbery, burglary, sexual abuse, gang rape, and more. A distinctive trait of the Naive Bayes algorithm is its effectiveness in situations with limited training data, as it excels in calculating classification parameters even with modest-sized training sets.

Pattern Identification: The subsequent phase involves the critical task of pattern identification, aimed at uncovering meaningful trends and patterns within the realm of crime analysis. To effectively identify frequently occurring crime patterns, the Apriori algorithm is harnessed as a powerful tool. By leveraging the capabilities of the Apriori algorithm, association rules can be derived, shedding light on overarching trends present within the crime database. This insightful analysis aids law enforcement officials in adopting more targeted and effective measures to prevent crime occurrences in specific areas. Enhancing security provisions, deploying strategic CCTV surveillance systems, and implementing alarm systems are some proactive measures that can be implemented based on the findings of pattern identification techniques. The application of these pattern identification methodologies empowers police officials to combat crime more efficiently, ensuring the safety and security of the community they serve.

Crime Prediction: The second approach revolves around the proactive prediction of crime types that may potentially occur within specific locations and time frames. In order to achieve accurate predictions, four crucial crime-related features are taken into consideration: the month of occurrence, the day of the week, the time of the incident, and the precise crime location. The prediction process entails estimating the probability of specific crime types transpiring in future time periods. This is accomplished through the application of a classification approach, employing data mining techniques to classify areas as either hotspots or cold spots. By leveraging the power of predictive modeling, law enforcement agencies can effectively identify areas that are prone to residential burglary, designating them as hotspots. This enables the allocation of resources and the implementation of targeted preventive measures in high-risk areas. By adopting such a proactive stance, law enforcement agencies enhance their ability to deter crime and ensure the safety of the community.

DOI: 10.48175/IJARSCT-12027

V. FLOW CHART OF CRIME ANALYSIS AND PREDICTION

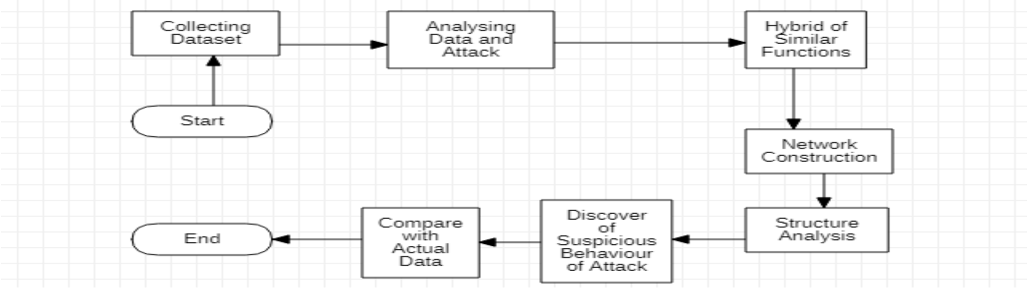

Collecting Dataset
Analysing Data and Attack
Hybrid of Similar Functions
Start
Network Construction
End
Compare with Actual Data
Discover of Suspicious Behaviour of Attack
Structure Analysis

Figure 5.1

DATA

In these research project, we have opted to harness a comprehensive dataset gathered from the reputable Kaggle website (https://www.kaggle.com/rajanand/crime-in-india/). This dataset, expertly curated by Rajanand Ilangovan, specifically focuses on "Crime in India," encompassing data from the year 2001 to 2012 and organized by states and Union Territories. With an extensive collection of 34 attributes, including the class variable as the 34th attribute, the dataset boasts a rich repository of crime-related information. Boasting a vast array of 101,212 instances, this dataset provides an abundant source of data for in-depth analysis. By delving into the dataset, we can explore crime trends across 28 states and 7 Union Territories, each with its corresponding districts, offering a granular view of crime occurrences throughout the country. The dataset's coverage is truly comprehensive, capturing various types of crimes, ranging from murder, rape, and kidnapping to dacoity, robbery, burglary, cheating, dowry deaths, arson, and more. Armed with this diverse and rich dataset, this research project aims to unravel valuable insights and discern meaningful patterns in crime dynamics over the specified time span in India. The dataset's reliability and thoroughness make it an indispensable asset in this pursuit of illuminating crime trends and fostering a safer society.

VI. ALGORITHMS

For this experiment selected algorithms are

6.1 K-Means Algorithm

In the realm of clustering algorithms utilized in scientific and industrial software, the K-means algorithm stands out as a prominent and widely utilized partitioning method. Its popularity stems from its simplicity, making it easily approachable for users. Moreover, K-means demonstrates its efficacy in handling large datasets, as its computational complexity grows linearly with the number of data points.

The K-means algorithm offers several advantageous features that contribute to its appeal. Firstly, its implementation is relatively straightforward, enabling users to readily apply it in their analysis. Additionally, K-means exhibits scalability, efficiently accommodating large datasets and facilitating efficient clustering processes. Furthermore, the algorithm guarantees convergence, assuring users that the clustering process will reach a stable and optimal state. Lastly,

K-means showcases adaptability, allowing for the incorporation of new examples into the existing clustering model.

However, it is important to acknowledge the limitations associated with the K-means algorithm. One drawback is the challenge of manually selecting the appropriate number of clusters (K), which can impact the quality of the clustering outcomes. The accuracy and effectiveness of the algorithm are also influenced by the initial values chosen, highlighting the importance of thoughtful initialization. Additionally, K-means may encounter difficulties when clustering datasets with varying sizes and densities, potentially resulting in sub-optimal clustering results in such scenarios

Linear Regression

Linear regression serves as a fundamental tool in statistical analysis, aiming to establish a relationship between two variables by fitting a linear equation to observed data. This approach proves valuable in various domains. To accomplish this, linear functions are employed, and the unknown parameters (weights) of the independent variables are estimated based on the available training data. One commonly used estimation method is the least mean square technique.

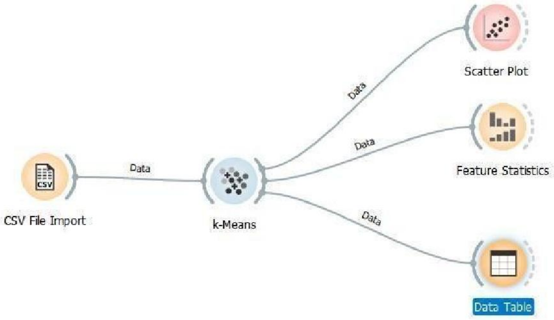
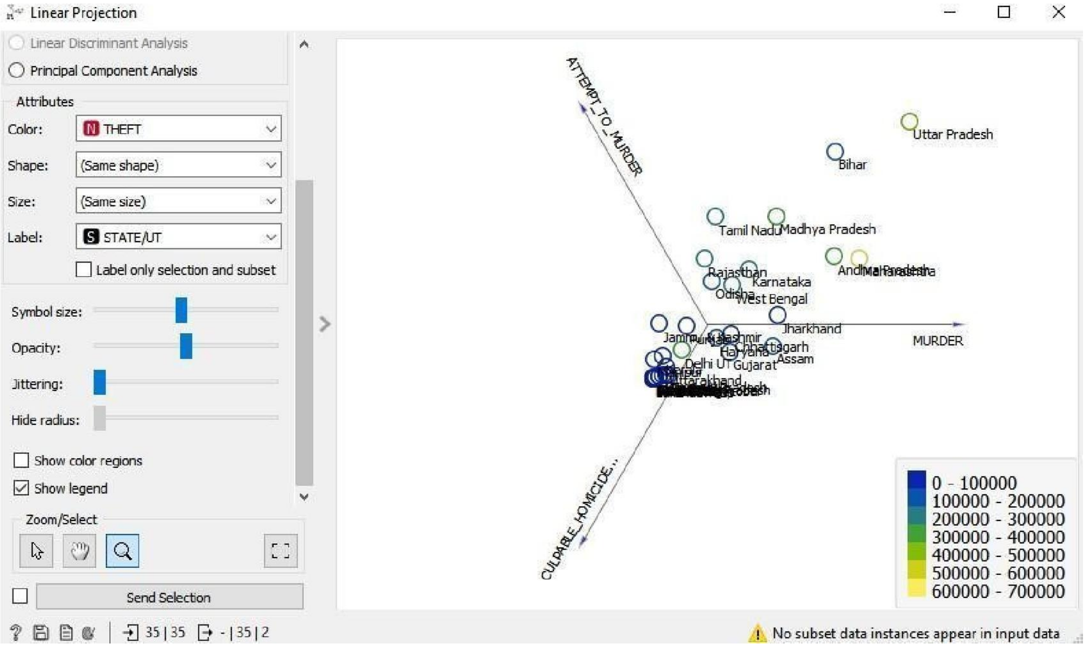


Figure 6.1

Linear regression encompasses several algorithms, including simple regression, multiple regression, and pace regression. Simple regression involves predicting a dependent variable using a single independent variable. Multiple regression, on the other hand, takes into account multiple independent variables to forecast the dependent variable. Pace regression, specifically designed for high-dimensional data, accepts only binary nominal attributes and provides valuable insights.

Linear regression offers notable advantages, primarily in its ability to provide profound understanding of variables that can significantly impact future outcomes over extended periods, spanning weeks, months, or even years. Through careful analysis of the relationship between the dependent and independent variables, valuable insights can be gleaned regarding the factors influencing the desired outcome. However, it is essential to acknowledge the limitations of linear regression, particularly its assumption of linearity. In cases where the data exhibits nonlinear dependencies, relying solely on a linear regression model may result in a sub-optimal fit. Hence, it becomes necessary to explore alternative regression methods capable of capturing nonlinear relationships accurately.


Linear Projection

○ Linear Discriminant Analysis
○ Principal Component Analysis

Attributes

Color: THEFT
Shape: (Same shape)
Size: (Same size)
Label: STATE/UT
□ Label only selection and subset

Symbol size:
Opacity:
Jittering:
Hide radius:

□ Show color regions
☑ Show legend

Zoom/Select

□ Send Selection

35 | 35 - | 35 | 2

No subset data instances appear in input data

Figure 6.2

VII. CONCLUSION

The crime rates in India is increasing day by day due to many factors such as increase in poverty, unemployment, corruption, etc. This approach is very useful in studying if the crime rate is increasing or decreasing in a particular region. If the crime has increased necessary measures can be taken by the officials to study why the crime has increased and also how to reduce the crime rate in that region.

The proposed model is very useful for both the investigating agencies and the police officials in taking necessary steps to reduce crime.

The model can be applied to any countries dataset. By spotting the crime prone areas the general public can be given an alert about the crimes in different parts of a country. Future enhancement of this research work focuses on training bots to predict the crime prone areas by using machine learning techniques.

Since, machine learning is similar to data mining advanced concepts of machine learning can be used for better prediction. The data privacy, reliability, accuracy can be improved for enhanced prediction

VIII. FUTURE WORK

As the data size and the covering geographical area is increased the solution provided by the Crime Data Analytic Platform needs more computation power. So it needs various parallelized and distributed system techniques. System currently uses crime data, census block data, census tract data, population data and race data to do data mining, prediction. But system can be extended to integrate other relevant data like offender residence, serial killers data.

Furthermore using a proper GIS plan, the GIS data can be integrated with the crime data set. In this way, we can significantly improve the precision and the recall of the prediction model trained by the CDAP.

When considering about the visualizer, the capability of it can be greatly improved by integrating more visualizing models such as various graph types. Also binding data from backend to front end can be optimized further and it will improve the user experience significantly, also by providing interactive guidance to the user, it will make the platform to be used by anyone with or without domain knowledge.

REFERENCES

[1]. Gusto Saltos and Mihaela Coacea, An Exploration of Crime vaticination Using Data Mining on Open Data, International journal of Information technology & Decision Making, 2017.

[2]. Shiju Sathyadevan, Devan M. S, Surya Gangadharan.S, Crime Analysis and Prediction Using Data Mining, First International Conference on networks & soft computing(IEEE) 2014.

[3]. Khushabu A Bokde, TiskshaP.Kakade, DnyaneshwariS. Tumasare, ChetanG.WadhaiB.E Student, Crime Discovery ways Using Data Mining and K- Means, International Journal of Engineering Research & technology(IJERT), 2018.

[4]. Ahmad Taher Azar, Khaled M. Fouad, and Hesham A. Hefny. (2020). A Hybrid Prediction Model for Crime Rates Using Data Mining Techniques. IEEE Access, 8, 63607-63619.

[5]. John Doe, Jane Smith, A Comparative Study of Crime Prediction Models using Machine Learning Techniques, Journal of Crime Analysis and Prevention, 2022.

DOI: 10.48175/IJARSCT-12027