International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 12 | Dec 2022 www.irjet.net p-ISSN: 2395-0072

Review of Algorithms for Crime Analysis & Prediction

Siddik Aiyub Patel[1], Jayanth R[2], Pawan T[3], Dr. E. Praynlin[4]

[123]*Students, Computer Science and Engineering, T. John Institute of Technology, Bengaluru, Karnataka, India*
[4]*Asst. Professor, Computer Science and Engineering, T. John Institute of Technology, Bengaluru, Karnataka, India*

Abstract - *Crime prediction is a rigorous approach to spotting patterns and trends in crime. This paper discusses various technologies that can be utilized to create a system for crime prediction. By constructing a Crime Prediction System, it accelerates the investigation of crimes and lowers the crime rate. We use a variety of methodologies, each of which is based on previously reported and recorded data containing time and place. The Crime Prediction System collects recorded data and analyses it using a variety of methodologies before forecasting the patterns and trends of crime using any of the strategies explored below.*

Key Words: Crime prediction, Clustering, Crime rate, Data Mining, Machine Learning, Deep Learning

1. INTRODUCTION

When crimes are committed regularly in a society, they in certain ways have an impact on organizations and institutions. As a result, it's important to research the connections and contributing elements to various crimes in order to effectively forecast and prevent them. The approach to predictive policing that law enforcement agencies are taking recently is becoming more pragmatic and data-driven. The foundational work of crime analysts, however, continues to be challenging and more often labouring. The objective of this paper is to present a thorough examination of theory and research about the prevention of crime in society and to put into practice various data analysis algorithms that address the relationships between crime and the patterns that crime follows. We have discussed various data mining algorithms along with deep learning technique to widen the scope of review

2. DATA MINING METHODS

The process of retrieving information from a database or any other piece of data using understanding to produce a number of outputs is known as data mining. It is nothing more than a process of gathering information and identifying patterns in the raw data in order to come to or make an effective judgement. Some fundamental steps involved are: Understanding your work target is the first step. Create a quick plan and decide on data mining objectives. In second step, the emphasis is placed on data collecting, data visualization and data examination. After reviewing the data, at third step, the data scientist starts by examining and data cleaning, keeping only the necessary data that may be used for further processes. This is known as "data preparation". Following that, modelling strategies must be chosen, at which point several tools and techniques that fall within the machine learning algorithm are introduced. Data is assessed once data modelling is applied, and then it can be used for related purpose. Machine learning uses a variety of algorithms and methods that are applied to data to gain knowledge and experience. The two fundamental types of machine learning algorithms are supervised learning and unsupervised learning. In supervised learning, an input is mapped to an output after the data is taught with correct and incorrect responses. Unsupervised learning, on the other hand, requires independent learning because there is no prior knowledge.

Data mining for crime prediction includes a number of techniques. Here, we have covered some of the commonly used techniques.

2.1 CLASSIFICATION

A classification approach represents and discerns data classes or ideas. Input data are classified into groups. Data set is divided into two types: Training data set and Testing data set. Former is used to train the model and the latter is used to test model and check its accuracy. Various approaches to this process are:

1) **Decision Tree**: A decision tree is a type of tree structure that resembles a flowchart, where each internal node represents a test on an attribute, each branch a test result, and each leaf node (terminal node) a class label or result. Data with few dozen to many thousands of dimensions can be handled by decision trees. To classify an instance, one tests the attribute given by the root node of the tree before continuing down the branch of the tree that corresponds to the attribute's value. The subtree residing at the new node is then subjected to the same procedure once more.

Yehya analysed and predicted San Francisco crime data using characteristics including longitude (X), latitude (Y), address, day of the week, date (YYYY-mm-dd: hh: MM: ss), district, resolution, and category. To classify the accuracy and prevent overfitting, the study used a variety of approaches, including principal component analysis. Additionally, he applied four alternative classifiers—K-NN, XGB Decision Tree, Bayesian, and Random Forest—to the task, with the Random Forest classifier yielding the best

results with a log-loss of 2.39031 which, in comparison to previously obtained result, is much more robust [1]. Based on a dataset taken from the SFPD Crime Incident Reporting System, Junbo et al. projected crime categories for the area surrounding San Francisco from 2003 to 2015. They examined the benefits and drawbacks of Naive Bayes, K-NN, and Gradient Tree Boosting classification models for that prediction problem. They found that because some features did not accurately capture the count or frequency, Naive Bayes did not function as a perfect model for that task. On the other hand, K-NN significantly enhanced the prediction outcome. Although it was a little slow, Gradient Tree Boosting was the model that fared the best in their experiment. With a score of 2.39383, the Gradient Tree Boosting model placed 93rd out of 878 teams [1]. The classification methods were the subject of an experimental investigation by R. Iqbal et al. (2013). They experimented with categorizing crimes according to the various US states. For the purpose of predicting crime, they evaluated Naive Bayes and the Decision Tree classifier. Decision Tree classifier outperformed Naive Bayes for the problems of crime categorization, which had an accuracy of 70.81% compared to 83.95% for the Decision Tree classifier [2].

2) **Nearest Neighbour**: It basically tries to find resemblance between testing and training data sets. If a train set and a test set are in close proximity, the test set will acquire the train set's class label. The only drawback of this algorithm is that it struggles when the training set has a smaller number of data records. To overcome this issue, K-NN algorithm is used. The K-NN algorithm makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like one of the existing categories, more specifically, a category which received majority of votes. When we did some research on existing systems which implements K-NN algorithm, accuracy was found to be less compared to other systems implementing different algorithms.

Assumptions vary depending on the situation because many machine learning models are constructed using datasets from various cities with various unique properties. Classification models have been used in numerous other applications, including weather forecasting, spam detection, sentiment analysis etc. Researchers from one of the studies observed six cities of Tamil Nadu state and applied crime prediction using K-NN algorithm, K-Means clustering, Agglomerative progressive clustering, and DBSCAN clustering calculation. Using features like Day, Time, Place, Year of the crime, Longitude and Latitude, the generated prediction was found to be 40% accurate [3].

3) **Random Forest**: A random forest is made up of numerous independent decision trees that work together as an ensemble. Every single tree in the random forest generates a class prediction, and the class that receives the most votes become the prediction made by our model. Many trees will be right in predicting, some may be wrong, allowing the group of trees to move in the proper direction. Random forest achieves this group of uncorrelated trees by allowing each tree to randomly select records from data set for its training.

This algorithm overall performs better just because of the way it predicts the outcome. Various studies have shown higher accuracy in crime prediction when random forest was used. The biggest drawback of random forest is that it can be too sluggish and inefficient for real-time forecasts when there are a lot of trees. Once trained, these algorithms provide predictions quite slowly, despite being quick to train.

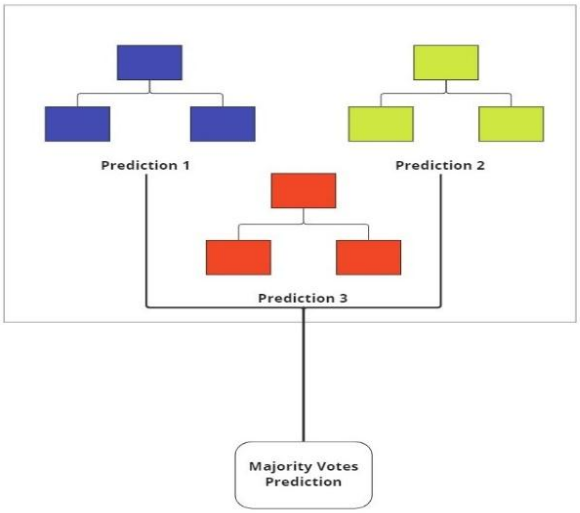

Fig -1: Random Forest Model

2.2 CLUSTERING

A machine learning approach called clustering or cluster analysis groups the unlabeled dataset. It accomplishes this by identifying comparable patterns in the unlabeled dataset, such as form, size, color, behavior, etc., then grouping the data according to the presence or absence of these patterns. Data from the same group shows more consistent features than data from other groups. A main clustering algorithm we will see is K-means clustering algorithm.

1) **K-means clustering**: One of the most straightforward and well-known unsupervised machine learning techniques is K-means clustering. Unsupervised algorithms often draw conclusions from datasets using only input vectors without taking into account predetermined or labelled results. The algorithm starts by user selecting appropriate k value. K denotes number of clusters. After that, k random centroid points are selected. Each training example will be assigned to a closest centroid. Variance is then calculated and a new centroid is calculated. This process is repeated until no assignments of data points are happening inside the cluster.

Several techniques exist for determining the ideal number of clusters. The Silhouette Coefficient or the Elbow method are most widely used to determine K in K-means. Any one of the preferred methods can be used by an analyst. The main distinction between elbow and silhouette method is that while silhouette accounts for characteristics like volatility, skewness, high-low differences, etc., elbow just calculates the Euclidean distance. For datasets with a lesser size or more time complexity, elbow is a better option than silhouette score due to its easiness of calculations. The optimal value of k is the "elbow" on the arm in the Elbow plot, when an SSE line plot is made, if the line chart resembles an arm. It is the point where the SSE decline begins to appear linear. When adopting the Silhouette plot, count the number of clusters where each cluster's plot is above the average in terms of thickness and does not show a lot of size variation. For the sake of ensuring that we choose the most advantageous number of clusters for K-means clustering, we might as well use both the Elbow plot and Silhouette plot. Formula used for single silhouette coefficient is:

$$\frac{b-a}{\max(a,b)}$$

Where, a = mean intra-cluster distance and b = mean nearest-cluster distance.

An average score that falls between [-1, +1] has been generated after factoring in each silhouette coefficient. The number of ideal clusters is determined by this average silhouette score. The elbow method does not perform well if the data are not highly clustered. You cannot determine the number of clusters if the graph is not skewed. Additionally, the skewed graph might not always provide the proper solution. There is a chance that the elbow approach won't produce accurate results if there exists a lot of duplicate data. The silhouette coefficient performs better when there are overlapping data because it can spot the data redundancy. In contrast, the number of clusters determined by the silhouette score is independent of the graph's degree of skewness. It depends on the silhouette score; the closer it is to +1, the greater the likelihood that it will approach the ideal value.

The elbow method's effectiveness is dependent on the dataset's characteristics. The elbow approach is effective if the relevant dataset's pattern is favorable. The silhouette score, on the other hand, is independent of the dataset's characteristics. It is a distance-based method. The mean distances between the intra-cluster objects and the closest cluster are utilized to calculate the silhouette score.

It can be claimed that the silhouette co-efficient approach is more appropriate when taking into account the increased precision of getting a distinct number of clusters. The elbow technique will, however, perform effectively in the event of a clean, noise-free dataset that doesn't have duplicate data in the training set.

3. DEEP LEARNING

Deep learning is in fact a neural network with "deep" referring to a greater number of hidden layers in the neural network. Traditional neural networks typically contain 2 to 3 layers but deep networks can have 5x amount of those present in the former. Deep Learning technique is employed in this case to create a model of the world with a variety of crimes and establish relationships between these crimes. Deep Learning employs a number of algorithms to transform the raw data into better representations.

Datasets and information are used to create a visual depiction of hotspots. The algorithm may pick up on patterns in events, and this knowledge can be used to various crimes occurring in various places and at various times. The directionality of the growth of crimes is not seen by the standard correlation analysis. Graph-based progression analysis with pairwise progression between two crimes type was conducted in order to determine the crime progression that might help in prediction of crimes that occur.

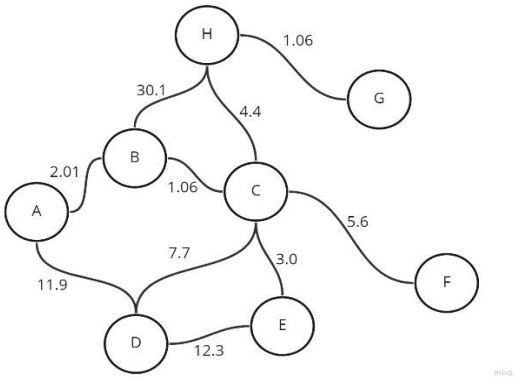

Fig – 2: Graph based model

Findings of one of the studies conducted show that spatial crime forecasting research has grown significantly. It also asserts that the prediction of crime hotspots is the most common kind of crime forecasting research. The extent of the area for which predictions are produced varies for different studies. The usefulness of crime projections depends on how far into the future they project their predictions. The intervals for hotspot forecasting using NNs range greatly, from one hour to one day to one month to one year. Another area where the prediction of criminal recidivism, or the chance that a criminal would either repeat an offence or be arrested for a new, more serious offence. Over the past two decades, NNs have emerged as one of the most widely used data mining approaches for analyzing criminal relapse. The different NN models' training approaches may change depending on the kind of issue they are modelling. While supervised learning approaches are utilized for classification and forecasting issues, and supervised learning Recurrent Neural Network (RNN) models are used for problems that are based on time, such as

time-series forecasting, Self-organizing Maps (SOM) and Cellular Neural Networks (CNN) are often employed in other domains for media classification.

4. CONCLUSION

Discovering patterns and information that may be beneficial for future prediction in crime analysis and behavior classification is now simple and effective thanks to major developments in data science technology, particularly in machine learning. There are several methods for crime analysis and prediction, some of which are covered in this paper: Sentimental analysis, deep learning, crime casting, and data mining. The aforementioned strategies each have advantages and disadvantages. Any one of them demonstrates superiority in a certain situation.

ACKNOWLEDGEMENT

We sincerely appreciate Ms. Suma R, the department head of computer engineering, as well as our project coordinators Dr. John T. Mesia Dhas and Dr. E. Praynlin for their inspiration, ongoing support, and insightful suggestions during our research on "Review of Algorithms for Crime Analysis & Prediction", which would have seemed challenging without them.

REFERENCES

[1] Y. Abouelnaga, "San francisco crime classification," 2016, arXiv preprint arXiv:1607.03626.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] R. Iqbal, M. A. Azmi Murad, A. Mustapha, P. H. Shariat 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, 2013K. Elissa, "Title of paper if known," unpublished.

[3] M. V. Barnadas, Machine learning applied to crime prediction, Thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, Sep. 2016.

[4] Varshitha D N, Vidyashree K P, Aishwarya P, Janya T. S., K. R. Dhananjay Gupta and Sahana R "International Journal of Engineering Research & Technology (IJERT)", Department of Information Science and Engineering, Vidyavardhaka college of Engineering, Mysuru, 2017.

[5] Snehal Dhaktode, MiralDoshi, Neeraj Vernekar and Ditixa Vyas "IOSR Journal of Engineering (IOSR JEN)" Computer, Atharva College of Engineering, University of Mumbai, India, 2019.

[6] Gouri Jha, Laxmi Ahuja and Ajay Rana "8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)" AIIT, Amity University, Amity, Noida, Uttar Pradesh, 2020

[7] J. Kiran and K Kaishveen "The Second International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2018)" Department of Information Technology Guru Nanak Dev Engineering College, Ludhiana

[8] Sunil Yadav, Meet Timbadia, Ajit Yadav, Rohit Vishwakarma and Nikhilesh Yadav "International Conference on Electronics, Communication and Aerospace Technology ICECA, 2017" Department of Information Technology, University of Mumbai, Shree L.R Tiwari College of Engineering, Thane, India.

[9] Akash Kumar, Aniket Verma, Gandhali Shinde, Yash Sukhdeve, Nidhi Lal "International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)", Department of Computer Science and Engineering, IIIT Nagpur, 2020.

[10] Shiju Sathyadevan and Surya Gangadharan S. "First International Conference on Networks & Soft Computing" Amrita Vishwa Vidyapeetham Amritapuri, Kerala, India. 2014.

[11] Krishnendu S.G, Lakshmi P.P and Nitha L "Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020)" Department of Computer Science and IT, Amrita School of Arts and Science, Kochi.

[12] Priyanka Gera and Rajan Vohra "International Journal of Computer Science and Information Technologies" Vol. 5(4), 2014, 5145-5148, PDM College of Engineering ,Sector 3A, Sarai Aurangabad, Bahadurgarh, Haryana, India.

[13] Arpita Nagpal, Aman Jatain and Deepti Gaur "IEEE Conference on Information and Communication Technologies (ICT 2013)" Computer Science & IT Department, ITM University, Gurgaon.

[14] Prtibha, Akanksha Gahalot, Uprant, Suraina Dhiman and Lokesh Chouhan "Crime Prediction and Analysis" Department of Computer Science and Engineering, National Institute of Technology, Hamirpur, Himachal Pradesh, India.

[15] Chhaya Chauhan and Smriti Sehgal "International Conference on Computing, Communication and Automation (ICCCA2017)" Department of Computer Science and Engineering, Amity University Uttar Pradesh, India.