ISSN 2395-1621

Survey on Crime Analysis and Prediction using Machine Learning

Anjali Kumar, Gauri Kurpaskar, Shradha Naik, Ruksaar Anchanal, Melancy Marcarenhas

anjalikumar2567@gmail.com
gaurikurpaskar3351@gmail.com
shradhanaik6199@gmail.com
ruksaaranchanal111@gmail.com
melancy26@gmail.com

Information Technology, Shree Rayeshwar Institute of Engineering and Information Technology, Shiroda, India.

ABSTRACT

The crime incidents occurred and reported in the country has increased drastically .The safety of the civilians is the major issue in today's world. The most crime investigating agencies till today uses the paper based system which is time consuming and also require man power to analyze existing crime datasets. So there is requirement for efficient crime analysis system. Machine learning is an application of Artificial Intelligence which focuses on the development of a program that can access data and used to learn by themselves. The information which is extracted is predicated using existing dataset. This paper highlights the use of machine learning techniques and clustering for effective investigation of crimes. Further the existing crime information will be used to predict the approximate growth of the crime rate in near future.

Keywords—Crime Analysis, Machine Learning, K-Means, Clustering, Dataset, Centroid.

ARTICLE INFO

Article History

Received: 8th March 2020

Received in revised form :

8th March 2020

Accepted: 10th March 2020

Published online :

11th March 2020

I. INTRODUCTION

Research in the area of crime analysis has been used for lessening of crime and public safety. In the past, with use of massive data, public data availability, there has been an increase in the creation of data analytics and visualization tools for the policing of safety systems.

In this survey, we see different aspects of past and present research crime analytics. Few researchers have studied the causation of one or more factors in terms of the crime rate, other researchers have also studied the effects such as sudden increase in the crime rate, while some research has examined the underlying patterns or co-relation associated with the crime rate. Prediction of future crime occurrence is a non- trivial task. Some researchers are trying to predict crimes from news feeds, complaint filed, online resources and also research work has focused on the crime occurred in a particular group categories such as race and gender.

II. EFFECT

The television, newspapers and social medias have its effect on crime and its rate. Earlier researchers had studied crime cultivation through television viewing and newspapers. The amount of television viewing has a remarkable impact on one's fear of victimization.

Other research suggests that the crime risk may be smaller than one would think , while others did calculable studies on threatening trends in the news and concludes news makes the situation worse. There are many television and short films being made on the crimes being execute on the women in the society. These films make us realize some cruel realities and need us to avoid the prevailing violence. This has lead to people forming many NGO groups to help out those tortured women.

In summary, media might just overemphasize the perception of actual crime events. Actual crime predicted will be more accurate by using huge data. There are four factors - increased incarceration, more police, the decline of crack and legalized abortion. Other factors, such as a strong economy, which are generally seen as a major factor.

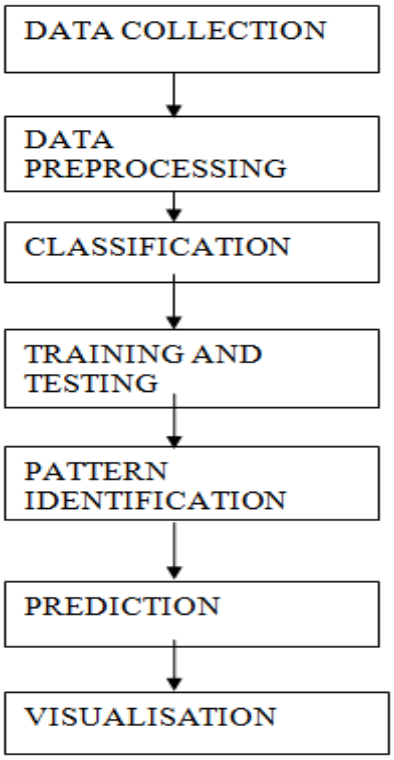
The knowledge that is gained from Machine Learning algorithms is a very useful which can help and support the crime investigators. We can use classification and clustering based models to help in identification of crime

underlying patterns and criminals. Machine Learning has a wide range of applications in the criminology which has made it an important field for the researchers. Machine Learning systems have played as a major role in assisting humans in this forensic and criminology domain. This makes it one of the most reliable decision-making environments for research.

The motivation for proceeding with this survey work is to aid criminal analysis and prediction. The paper is ordered in such a manner to provide an accurate and deep understanding about the crime analysis procedure and then produce different types of crime analysis operations and those which can be combine together for producing an end user product which can be applied to the crime analysis in any crime investigation offices. This work will be a valuable reference to those who will lead their research work in the crime analysis and Crime prediction using Machine Learning techniques.

III. CRIME ANALYSIS PROCEDURE

Fig.1. Stages of Implementation



A. Data collection

Data Collection is the process of gathering information of crime in an organized manner. The attributes is common to all fields of research and various crime investigating offices which include the day, month, year, type of the crime etc.

B. Data pre-processing

The following techniques are used in preprocessing the data:

- Data Integration: Data with different representations are put together and the conflict within the data are resolve.
- Data Transformation: Data is normalized and aggregated.
- Data Reduction: Removes irrelevant attributes.
- Data discretization: Part of data reduction but with a particular importance.

C. Classification

Classiffication is a step-in Machine learning that allocates the data or items in the collection. We can correctly predict the destination class for each data in the data set. In this case, a classification model is used to group type of crimes based on various parameters. K-means clustering is one of the method of cluster analysis which focuses to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean.

D. Training and Testing

In Training data, the known and labeled data are used in machine learning algorithm. They find relationships and develop insights that helps in making decision. For training the model, the data is divide into training sets and testing sets. Training datasets is a sample of data used to fit the model where as testing data, on the other hand, includes only input data, not the corresponding expected output. The testing data is used to estimate how well your algorithm was trained, and to determine model properties.

E. Pattern Identification

Pattern Identification is the process of identifying patterns by using machine learning algorithm. Pattern identification can be defined as the classification of database on knowledge already gained or on information extracted from patterns. One of the important features of the pattern recognition is its application.

F. Prediction

Crime events are possible to occur in the vicinity of past crime events. The historical data of a given area where certain type of crime events happened which is analyzed to predict the crime event likely to happen in the near future.

IV. LITERATURE SURVEY

[1] Param Joshi investigates machine-learning-based crime prediction. In this work, Vancouver crime data for the last 15 years is analyzed using two different data-processing approaches. Machine-Learning predictive models, K-nearest neighbour and boosted decision tree, are implemented and a crime prediction accuracy between 39 to 44 percent is obtained when predicting crime in

Vancouver . In this work, KNN and decision-tree algorithms were used to train our model. KNN is one of the simplest classification algorithms. It assigns a sample z to class A if the most values nearer to z are from class A, otherwise, it assigns the sample to class B. In KNN, the following formula is used to calculate the probability of the test sample belonging to category train our model. KNN is one of the simplest classification algorithms. It assigns a sample z to class A if the most values nearer to z are from class A, otherwise, it assigns the sample to class B. In KNN, the following formula is used to calculate the probability of the test sample belonging to category. They applied boosted decision tree algorithm in both approaches and compared the results. For both approaches, they used the Adaptive Boosting (AdaBoost) ensemble method and learner-type decision tree. AdaBoost is a meta-algorithm that combines several weak learners to improve a weak classifier. The maximum number of splits was 20. Accuracy and training time for approach 1 was 41.9 percent 903.63 seconds, respectively, while approach 2 was 43.2 percent accurate with 459.26 sec training time. The results from both methods (KNN and boosted decision tree) are shown in Fig. 8 for both approaches.

[2]Dr Sarvanaguru RA.K mainly revolves around predicting the type of crime which may happen if we know the location of where it has occurred .The aim of this project is to make crime prediction using the features present in the dataset. The dataset is extracted from the official sites. With the help of machine learning algorithm, using python as core we can predict the type of crime which will occur in a particular area. 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 upon the accuracy. The K-Nearest Neighbor (KNN) classification and other algorithm will be used for crime prediction. Visualization of dataset is done to analyze the crimes which may have occurred in the country. This work helps the law enforcement agencies to predict and detect crimes in Chicago with improved accuracy and thus reduces the crime rate.

[3] Shiju Sathyadevan tested the accuracy of classification and prediction based on different test sets. Classification is done based on the Bayes theorem which showed more than 90 percent accuracy. Using this algorithm they trained numerous news articles and build a model. For testing they are inputting some test data into the model which shows better results. Our system takes factors/attributes of a place and Apriori algorithm gives the frequent patterns of that place. The pattern is used for building a model for decision tree .Corresponding to each place we build a model by training on these frequent patterns. Crime patterns cannot be static since patterns change over time. By training means they are teaching the system based on some particular inputs. So the system automatically learns the changing patterns in crime by examining the crime patterns. Also the crime factors change over time. By sifting through the crime data they have to identify new factors that lead to crime. Since we are considering only some limited factors full accuracy cannot be achieved. For getting better results in prediction they have to predict not only the crime prone regions but also the proper time find more crime attributes of places instead of fixing certain attributes.

[4] Jyoti Agarwal focuses on crime analysis by implementing clustering algorithm on crime dataset using rapid miner tool and here we do crime analysis by considering crime homicide and plotting it with respect to year and got into conclusion that homicide is decreasing from 1990 to 2011 .From the clustered results it is easy to identify crime trend over years and can be used to design precaution methods for future. The main objectives of crime analysis include Extraction of crime patterns by analysis of available crime and criminal data. Prediction of crime based on spatial distribution of existing data and anticipation of crime rate using different Machine Learning techniques. So In this paper crime analysis is done by performing k-means clustering on crime dataset using rapid miner tool.

[5] Lenin Mookiah provides a survey of past and current research on crime, more research is needed for investigators and policy makers. Some future work that they will be investigating is the extraction of patterns based upon multiple, heterogeneous data attributes, such as crime news stories, user profiles, and social media. Analyzing such a diverse set of “big data” can not only aide in the prediction of crime, but perhaps provide law enforcement individuals with tools that can secure areas that are subject to future criminal activities, like senior-living homes and community centers. In 442 addition, we will investigate methods for extracting the reasoning for events for which users are interested. For example, if a user is interesting in swimming, they might be especially interested in safety at a local beach. Finally, we plan on implementing novel visualization techniques that will allow users to the study the evolution of patterns.

[6] Setu Kumar Chaturvedi capable to enhance the accuracy, performance, speed of predicting the crime using the techniques of Machine Learning. These techniques are effectively identify common pattern by comparing current and past crime data and predict the future value. It’s the basic step towards the Crime Prediction have been taken with demonstration and manipulation.

[7] H. Benjamin Fredrick David have Studied several methods in identification of crime and criminals which includes Text/ NLP based methods, crime patterns and crime evidence based methods, spatial and geo location based methods, communication based methods and finally Prisoner based methods. The data mining techniques studied from this survey can be applied for identifying the criminals in the society and also for providing a better future to live in. The Naïve Bayes and other algorithm will be used for crime prediction.

[8] Mugdha Sharma, has employs decision tree-based classification approach to detect e-mails in relation to criminal activities. All the e-mails were classified as suspicious, maybe- suspicious or non- suspicious. From this experiment, it is found that an Advanced Decision Tree classifier and feature selection method can provide better classification result for suspicious e-mail detection. The improved algorithm through introducing attribute-importance emphasizes on the attributes with fewer values but higher importance which solves the classification drawback of choosing attributions with more values. The experimental results show that Advanced ID3 algorithm compared with traditional ID3 Algorithm has better classification accuracy and can get more reasonable, more effective and more rational classification rules. The new tool "Z-Crime" will be helpful for identifying the suspicious e-mail and will also support the investigators to get the information in time to take necessary actions to reduce criminal activities.

[9] Kaumalee Bogahawatte, presents a ttesh methodology of identifying a criminal by using existing evidences in situations where any witness or forensic clues are not present. The system uses an explicit clustering mechanism to segment crime data into subsets, or clusters based on the available evidences and Naive Bayesian classification has used to identify most possible suspect/ suspects for crime incidents. The ability of storing suspect details that were arrested for at least one crime incident but not proved as the responsible criminal by the court, and the efficiency in identifying possible criminals for a crime incident are the characteristics that makes ICIS perfect than other crime investigation tools. The system has used the communication power of multi agent systems to increase the efficiency in identifying possible suspects.

[10] Tahani Almanie, generated many graphs and found interesting statistics that showed the baseline to understand Denver and Los Angeles crimes datasets. Then, they applied Apriori algorithm to find frequent crime patterns in both cities. After that, they applied Decision Tree and Naïve Bayesian classifiers to help predicting future crimes in a specific location within a particular time. With that they achieved 51% of prediction accuracy in Denver and 54% prediction accuracy in Los Angeles. Finally, provided an analysis study by combining their findings of Denver crimes' dataset with its demographics information. They aimed to further understand models' findings and to capture the factors that might affect the safety of neighbourhoods.

Table

Sr. No	**Previous Work Done**		
	Name of the paper	**Advantages**	**Disadvantages**
1	Crime Analysis Through Machine Learning(2018)	Machine Learning predictive models KNN and boosted decision tree were used to obtain crime-prediction accuracy between 39% to 44%	model has low accuracy as a prediction model.
2	Crime Prediction and Analysis Using Machine Learning (2018)	machine Learning algorithms The Model predicts the type of crime with accuracy of 78%.	Data should be well organized and processed.
3	Crime Pattern Detection, Analysis & Prediction (2017)	The developed model will reduce crimes and will help the crime detection field in many ways that is from arresting the criminals to reducing the crimes by carrying out various necessary measures. methods can be applies on full data set which consists of 42 crime heads having 14 attributes to them, thus when analysed could provide more unexpected dependencies of attributes over each other.	The biggest disadvantage in the project was data acquisition and data staging
4	Crime Analysis and Prediction Using Data Mining(2014)	system takes factors/attributes of a place and Apriori algorithm gives the frequent patterns of that place. The pattern is used for building a model for decision tree.	It is providing proper security in less inhabited/ crime prone areas, increasing night patrolling and fixing CCTV's in sensitive areas
5	SURVEY ON CRIME ANALYSIS AND PREDICTION USING DATA MINING TECHNIQUES	Studied several methods in identification of crime and criminals which includes Text/ NLP based methods, crime patterns and crime	nil

	(2017)	evidence based methods, spatial and geo location based methods, communication based methods and finally Prisoner based methods.	
6	"Z-Crime: A Data Mining Tool for the Detection of Suspicious Criminal Activities based on the Decision Tree	Introducing attribute importance as a factor before information gain in the decision tree	Not giving a clear view of the processing and comparison of criminal Behaviour.
7	Evidence-based Analysis of Mentally 111 Individuals in the Criminal Justice System	Analysis for the identification of the mentally ill felony	Statistical classification of criminals missing. Could have taken more features

V. USING K MEAN ALGORITHM

K-means is useful and widely used for clustering partitioning algorithm for the dataset in which a number of clusters are required .The k-means clustering technique groups the data together by considering closeness of data . It is an iterative algorithm. Firstly, it initialize random k points which is means. Secondly categorize each item to its closest mean.

The following steps or procedures are :

- Specify number of clusters K by first shuffling the dataset initialize the centroid and then randomly select K data points for the centroids without replacement.
- Keep iterating until there is no change to the centroids. The assignment of data points to clusters isn't changing.
- K means algorithm complexity is O(tkn), where n is instances, c is clusters, and t is iterations and relatively efficient . It often terminates at a local optimum.

The advantages of using K-Means algorithm:

1) If variables are huge, than this algorithm is most of the times computationally faster than other clustering, if we keep k smalls.
2) K-Means produce tighter clusters than other clustering, especially if the clusters are globular.

Examine the crime dataset which is enormous and K-Mean algorithm can take huge data to produce accurate and tightly cluster output , so we have considered using K-Mean Algorithm is efficient.

VI. SCOPE

Crime analysis should provide currently useful information to aid in controlling the crime and prevent objectives by recognizing and analysing methods of operation of crimes.

Crime analysis and prediction can be useful to the Department's of long-range planning efforts by providing estimates of future crime trends. Collecting, managing and analysing large volumes of accurate data will help the department.

By combining prior knowledge and maintaining crime analysis resource which will enhance the accuracy, performance, speed of predicting the crime.

VII. ACKNOWLEDEGMENT

An Endeavour over a period of time can be successful only with the constant support and guidance from our well-wishers, hence we take this opportunity to express our gratitude to all those who encouraged us in getting a head start on our project work .We are immensely grateful to our Principal Dr. Anurag Jain, for providing the needed facilities creating a comfortable environment required for the project.

We also thank Mrs. Manjusha Sanke, Head of the Department of Information Technology who assured us every support from the institute.

Our sincere gratitude to our Internal guide, Miss. Melancy Mascarenhas Assistant Professor, IT Department, for providing us with excellent guidance, direction and constant support in organization, planning and scheduling the project phases.

We would like to thank Mrs. Prajakta Tanksali, project coordinator, IT Department, for her support in organization, planning and scheduling the project phase.

Besides we are also grateful to our friends and well-wishers who provided us with lots of support.

Lastly, we are grateful to whole of the Information Technology staff for providing us with the lab facilities and co-operation for our project.

VIII. CONCLUSION

The research has been evident that the basic details of a criminal activities in an area contains indicator that can solved by machine learning agents. It is used to classify criminal activities in a particular location .Using the massive input(crime data-set) was taken and based on that input, with machine learning algorithms dataset was analysed to identify underline patterns .The accuracy depends upon the data collection and the algorithms that is used for better and accurate prediction. The model can be used for geographical areas .This could also help to analyse crime occurring in different locations, so that

crime investigator can work on that case for future betterment**.**

IX. REFERENCES

[1] Param Joshi, Parminder Singh Kalsi, and Pooya Taheri Suhong Kim.-"Crime Analysis Through Machine Learning" International Research Journal of Engineering and Technology (IRJET) 2018.

[2] Alkesh Bharati, Dr Sarvanaguru RA.K.-"Crime Prediction and Analysis Using Machine Learning " in International Research Journal of Engineering and Technology (IRJET) ,Volume: 05 Issue: 09 | Sep 2018.

[3] Sathyadevan, Shiju, Devan M.S, and Surya Gangadharan S.. "Crime analysis and prediction using Machine Learning", First International Conference on Networks & Soft Computing (ICNSC2014).

[4] Jyoti Agarwal, Renuka Nagpal, Rajni Sehgal.-"Crime Analysis Using K-MEANS Clustering" International Journal of Computer Applications (0975 – 8887) Volume 83 – No4, December 2013

[5] Lenin Mookiah, William Eberle and Ambareen Siraj.- "Survey of Crime Analysis and Prediction" Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference

[6] Nikhil Dubey ,Setu Kumar Chaturvedi.-"A Survey Paper on Crime Prediction Technique Using Machine Learning" Int. Journal of Engineering Research and Applications www.ijera.com ISSN : 2248-9622, Vol. 4, Issue 3(Version 1), March 2014, pp.396-400

[7] H. Benjamin Fredrick David and A. Suruliandi.- "Survey On Crime Analysis And Prediction Using Data Mining Techniques" Ictact Journal On Soft Computing, April 2017, Volume: 07, Issue: 03

[8] Mugdha Sharma, "Z-Crime: A Data Mining Tool for the Detection of Suspicious Criminal Activities based on the Decision Tree", International Conference on Data Mining and Intelligent Computing, pp. 1-6, 2014.

[9]Kaumalee Bogahawatte and Shalinda Adikari, "Intelligent Criminal Identification System", Proceedings of IEEE International Conference on Computer Science and Education, pp. 633-638, 2013.

[10] Tahani Almanie, Rsha Mirza and Elizabeth Lor.- " Crime Prediction Based On Crime Types And Using Spatial And Temporal Criminal Hotspots", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.4, July 2015.

