International Journal of Future Generation Communication and Networking
Vol. 13, No. 3, (2020), pp. 27 - 35

A Survey on Crime Data Analysis and Prediction using Machine Learning

[1]Mansi.S.Bagale*,[2]Dr.(Mrs.)Sharmila .K.Wagh*

Modern Education Society's College of Engineering, Pune

Savitribai Phule Pune University, Pune Maharashtra, India

Abstract

An action taken which is against the rules and regulation of a particular country or region is called crime. A person who goes against these laws is called a criminal. In recent years it has been observed that there is rapid growth in the field of crime . Crime is considered as one of the following types for example robbery of money, murder, rape, drug trafficking, domestic violence, arms trafficking, false imprisonment, kidnapping, Robbery of business property, Card fraud etc. Individuals in India feel that the crime is expanding at untouched elevated levels. Numerous methodologies for analysis and prediction in machine learning had been carried out. But, only few determined attempts has been seen in the criminology field. Various paper uses different technologies for crime data analysis. This paper shows the review on the Crime assessment and Crime estimate using a couple of machine learning techniques. The new details are derived based on the prediction of the existing datasets. The main objective of the paper is to help law enforcement agencies or crime investigators to predict ,solve crime and identify patterns of crime at a lot quicker rate so as to diminish the crime rate.

Keywords: *Classification, patterns, Crime, Naïve Bayes, SVM, KNN, Decision Tree etc.*

1. Introduction

Crime is an infringement from the humankind perspective which is frequently punishable and chargeable by law. Criminology is a field of investigation of crime and it is interdisciplinary sciences that gathers and examine information on crime and crime execution. It aims to identify crime characteristics. Now a days the crime exercises has expanded on a large scale and it is the duty of police division to control and lessen consistently the crime exercises. Crime data analysis and prediction are the significant issues to law enforcement division as there is voluminous information of crime that exist. Prior frameworks were program or rule based however with the initiation of machine learning presently, machine can gain knowledge from the information and can act appropriately. This work will be helpful to those who carry out their research work in the crime analysis and Crime prediction using machine learning techniques. The potential crime analysis can give us the significant data about the patterns of the crime and can help in finding locations where crime activities are high or low. These days, Machine Learning based arrangements are getting well known in finding solutions and giving a lot of precise outcomes. One of the difficulties in applying Machine Learning based crime patterns analysis is, applying such models for the Indian setting where there is a huge geological information. Another difficulty is the huge size of crime datasets, and a possibly enormous accumulation of fascinating crime patterns. These Machine Learning based analysis could help the police division or law enforcement agencies to rapidly and tirelessly find significant patterns in crime events and permit police officers and experts to upgrade crime goals rate and to increase operational productivity.

ISSN: 2233-7857 IJFGCN

1.1. Types of Crime Analysis

Crime analysis relates to the group of consistently ,analytical operations that provides periodic data about crime patterns and trends correlations. Crime analysis based on its scope, analysis techniques and data is further categorized into various types.

i. Intelligent crime Analysis

The objective of Intelligent analysis is to identify network of criminals carrying out criminal activity and also to help the police in arresting those violators of law. Information in intelligent analyses is gathered by police through surveillance, participant observation, wiretapping etc. This type of information may include telephonic conversation, travel information, financial information of the offenders under investigation. The intelligent analyst works closely with police officers.

ii. **Investigative crime analysis**

It is also referred as criminal profiling. This process includes creating profiles of offenders who have committed serious crime. The main purpose of this type of analysis is to help criminal investigator recognize offender by identifying personal characteristic , social habits etc.

iii. **Tactical crime analysis**

It is study of detailed investigation and analysis of criminal incidents and activity through the examination of general characteristic such as when, how and where the incident has occurred to help in pattern development, to identify potential suspects and case clearance by linking solved cases to open cases. It also examines field data collected by patrol officers about potential criminal activity.

iv. **Strategic crime analysis**

Strategic crime analyst uses statistical methods to examine electronic databases containing huge number of records. These analysts deal with variable, date , location, time and type of incident. It has two major goals

- To help solving and identifying long term crime issues.
- To evaluate police response to crime problems

v. **Administrative crime analysis**

It is concerned with presentation of findings of crime ,research based on legal, political matter to inform citizens ,people within police administration, government etc. It is a process of choosing important findings from the past analyses and formatting correctly for target audiences. Its primary purpose is to inform audiences.

2. Literature Survey

In this proposed work various clustering algorithm like k-mean clustering ,agglomerative clustering are used for analyzing the area of the crime in order to decrease city crime rates. Various visualization techniques are used to create graphical pictures which aid in the grasping of complicated, often huge representation of data[1]. The goal of the proposed work is to study dataset which contain various crime records and predict the kind of crime based on various condition which may take place in future. The crime dataset is obtained from official portal of Chicago police. K-NN and various other algorithms are used to test the crime prediction and one with greater accuracy is used for training .The complete agenda of this paper is to explain the use of machine learning algorithm to law enforcement agencies to predict and resolve crimes at a speedy rate reducing the city crime rate[2]. The NCRB (National Crime Records Bureau) website collects , maintains and publish the crime data. In this paper k-mean clustering algorithm is

used on criminal dataset. WEKA a Software is used to construct cluster zones. It builds model with high, low, medium crime zone. zones of state. This information can be helpful to police to increase or decrease level of preventive actions[3]. In this proposed work data is collected from government sources in csv format. This data is pre-processed in R. The software used in this paper for mining various crime patterns are WEKA and R tool. The output is represented in graphical form such as charts indicating high or low crime region[4]. Crime investigation [5]need to be quick and beneficial. Since a lot of data is collected during crime investigation, data mining techniques such as Naïve Bayes, JRip, J48 on sample criminal data. There performance are compared and the one that performs best is used against test crime and criminal database to recognize possible suspects of crime. [6] In this proposed work data is collected from Chicago Portal from 2010 to 2012 i.e. 2 years related to different crime that has been committed in different region of Chicago city. Two approaches has been used. First one is clustering using k-mean, to identify different places of crime for which WEKA tool is used. Second one is Spatial mining to locate hot spot of crime. Hot spots detection helps in detecting and investigating crimes more quickly. [7]In this proposed work the dataset is collected from Chicago Police Department. It contains data from 2001 to 2017 in .CSV format. The file contains both categorical and numerical value. The algorithms that are used on the sample crime dataset are Random Forest, Decision Tree and various ensemble methods such as AdaBoost ,bagging and Extra trees. These algorithm performance are compared and the one which provides the highest accuracy in terms of performance is used to train the model to find which category of crime will occur probably at a particular place at a specific time in Chicago city.[8]In this proposed work k -mean clustering data mining technique has been used on crime dataset collected from New South Wales region of Australia. The areas with high crime rates and the most common type of crimes are identified. David et al.[9] has presented a detail survey on crime data analysis and prediction with the help of different supervised and unsupervised data mining techniques for the purpose of criminal identification.[10] In this paper, model based on clustering algorithm such as k-mean is used to identify crime pattern. Main aim is to helps in boosting the process of crime solving. Various weights are allotted to different attributes based on the crime types being clustered. The resultant clusters that are generated contains the feasible crime patterns and are map with the help of geospatial plot. In [11] this proposed model data is collected from Libyan police Department. K-mean algorithm for clustering purpose and Apriori algorithm for data association is used to recognize possible suspects of crime. Chen et al.[12] has come up with a framework where the data is gathered from Tucson Police department in USA, consisting of 1.3 Million suspects and criminal records from year 1970.This framework was proposed to establish a relationship between data mining techniques and crime types. In this proposed work [13] k-mean clustering algorithm is used to analyse and predict, visualize patterns and trends in different areas of Chicago. In this proposed model [14] data is collected from United State City in Northeast. The dataset contains related events and aggregated counts of crime. Classification methods are used for crime forecasting. Number of techniques are used to find the correct forecasting approach to achieve the best result in terms of accuracy. Navjot Kaur [15] has provided an overview of different data mining techniques and its comparison is provided. An open source software known as WEKA tool is used to produce efficient result. Louise Underson [16] has built a model that is based on the concept of cognitive psychology and data mining techniques to construct a predictive model of crime. Given the mean air temperature, location and time it use to predict the type of item stolen with the help of different techniques such as association rules to build independent set of crime and classification and regression tree. In this proposed work[17] prediction of crime in states of India is analyzed using machine learning

techniques such as KNN and Naive Bayes and it is found that Naïve Bayes gives higher performance accuracy and lesser execution time than KNN. Kianmehr et al.[18] has used two approaches first by selecting certain portion of data randomly and second by applying k-mean clustering algorithm. It was found that the performance in terms of accuracy of one class SVM for predicting the hotspot crime location gives much reliable result. Sivaranjani et al.[19] has proposed a model where the crime data is gathered from NCRB official site in India, for the state of TamilNadu. Bhanjeet Kaur et al.[20]In this paper a model is created using different kinds of data mining methods to analyse and predict violence against women. Borowik et al. [21] has shown the usefulness of machine learning algorithms in prediction crimes such as determining the area of hot spot, creating criminal profiles, and to detect trends in crime data. Garcia et al.[22] has proposed a model where a statistical analysis was made to compare among various categories of crime in Mexico city. A predictive model is built using Google Trend based on crime data of previous weekly analysis. It also explains the importance of social media platform such as Tweets from twitter can be helpful to predict in certain areas using important strategies to handle crime. Kadhim Swadi al Janabi [23] proposed a model had made used of machine learning algorithm such as k-mean to cluster crime related data and decision tree to classify crime data for analysis purpose. Bogahawatte and Adikari [24] proposed an approach and came up with a system (ICICS) Intelligent Crime Investigation System which could recognize criminals based on proof collected from the crime location, based on clustering and classification techniques. [25] In this paper a model is built by clustering criminals on the basis criminal careers.

3. Existing System

In existing system various techniques are proposed to analysis crime data. Several approaches for analysis of crime and prediction has been performed with help of WEKA tool, RapidMiner , R etc. But the existing work does not have the facility of feature engineering or exploratory data analysis. Specific crimes are analyzed one by one to get better insights. Exploratory Data Analysis is favorable as it allows to view data from various angles so that the future results is correctly depicted, and applicable to the desired business contexts. EDA also helps to discover insights which were not noticed or worth investigating for the business but it can be very much informative about a specific business.

4. System Architecture

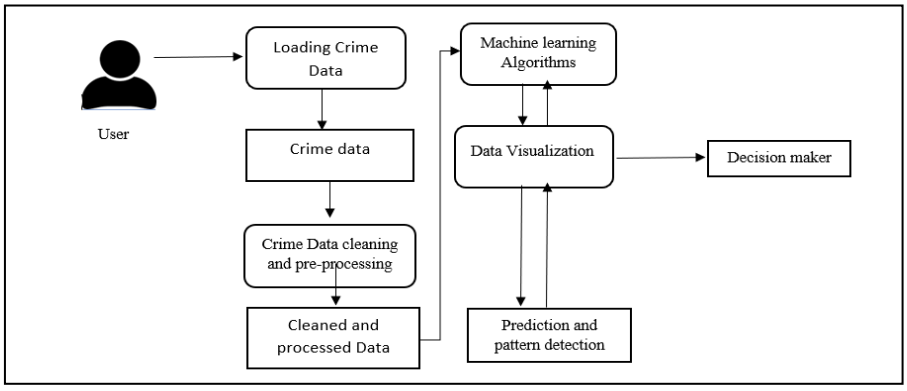

Figure 1. System Architecture

The architecture of the proposed system consist of the following stages:-

Loading Crime Data:- Firstly the user gathers the crime dataset from the portal of National Crime Records Bureau(NCRB) of India. This dataset contains entire information about different aspects of crimes that took place in India from 2001 to 2012. There are various factors that can be analyzed from this dataset. The file format for the data is .CSV. It is loaded using Pandas library in python.

Data Pre-Processing: After loading the crime data ,the next step in this model is data pre-processing. It is a way of converting data from the raw form to a much more usable form, i.e., making data more meaningful by handling the missing values, data cleaning and transformation of raw data into easy to interpret format.

Application of machine learning algorithm: Once the pre-processing is completed ,cleaned and processed data is obtained. On this classification and clustering algorithms are applied based on requirements. The classification algorithm e.g. Naive Bayes works on supervised learning concept in which random sampling need to be carried out i.e. diving the data into test and train sample for e.g. 80% train samples and 20% test samples to train the classifier to recognize the new unidentified crime record. Whereas clustering algorithm e.g. k-mean is based on unsupervised learning algorithm which splits the crime records depending upon the number of group to be generated.

Data Visualization: Visualization takes a huge amount of data to represent useful data in the form of charts or graphs for quick and better understanding of information. The results can be visualized using appropriate graphs or maps showing sensitive areas of having high probability of crimes. Matplotlib library from sklearn is used for analysis of crime dataset. Visualization is basically shown with use of bar charts, boxplot, heatmap, scatterplot etc.

Pattern detection: Next phase in the methodology is pattern identification that is used to find the sequence of crimes which are similar in nature and belongs to same class. Identification of meaningful patterns can help police to develop effective crime prevention and crime reduction strategies.

Prediction: The frequent patterns obtained is used to drive models which can predict future crime.

Decision maker: The information that is obtained as result help law enforcement agencies in enhancing intervention strategies via effective preparedness.

A. Selection Of Algorithm

a) Decision tree

Decision tree classification model forms a tree like pattern from dataset. Decision tree is constructed by splitting a dataset into small pieces. At each step in the algorithm, a decision tree node is split into many branches until it reaches leaf nodes. Leaf nodes indicates the class labels or result and internal node denotes “test” on an attribute. At each and every step, decision tree chooses an attribute that best splits the data. Tree based methods allow to build predictive models with higher accuracy, invariability and ease of explanation. Unlike linear models, decision tree depict non-linear relationships very well. It is adaptable of solving various types of problem encountered i.e. classification or regression.

Entropy

Entropy is the value of unpredictability of some random variable, it denotes the impurity of an arbitrary collection of samples. If the sample is totally homogeneous, then the entropy computed is zero and if the sample is partitioned into 2 equal parts, it has entropy of one.

It is calculated using formula:-

Entropy = - p log2 p - qlog2 q (1)

Information Gain
Information gain is the effect of the change in entropy. Information gain explains the important a given feature to a feature vectors. It decides in what order the attributes need to be ordered in the nodes of a decision tree.

b) Naïve Bayes

One of the classification technique, Naïve Bayes uses probabilities already known to determine how to classify input. These probabilities are related to existing classes and what features they have. This technique is based upon Bayes' Theorem used in a many classification tasks. The algorithm does that by making an assumption of conditional independence over the training dataset. The assumption of conditional independence states that, for some random variables P and Q we say P is conditionally independent of Q , only if the probability distribution governing P is not dependent on the value of Q . It takes some input and calculate the probability of happening given that it belongs to one of the classes. These calculation is performed for each of the classes. After getting these probabilities, the one which has the largest value is taken as the prediction purpose for the class where input belongs to.

$$P(s|x) = \frac{P(x|s)\,P(s)}{P(x)} \quad (2)$$

- P(s|x) -To find probability of s, given predictor class x. It is called posterior probability of s.
- P(s) -the prior probability of target class s being true.
- P(x|s) Probability of x given target class s was true. It is called probability of likelihood of predictor class x
- P(x) is the probability of predictor class x(independent of target class s).

c) K-nearest neighbors (KNN)

K-nearest neighbors algorithm belongs to supervised machine learning algorithm that is widely used in field of pattern recognition, data mining etc. It is utilized to solve both regression as well as classification problems. K-nearest neighbors (KNN) calculation utilizes 'feature similarity' to anticipate the estimations of new item. It means that the based on measurement of similarity among data points in the training set, a new item is assigned a value. Its functioning consist of the following steps −

1. Load the training and test data
2. K value i.e. the nearest data point is chosen. It should be integer.
3. Carry out the following steps for each data point in the test data :-
   - 3.1. The distance between test data and every row of training data is determined using Euclidean or Manhattan or Hamming distance. Euclidean is the most frequently used method to calculate distance.
   - 3.2. Arrange them in ascending order, based on the distance value.
   - 3.3. The topmost K rows is chosen from the sorted array.
   - 3.4. Lastly allot a class to a data point based on most often class assigned to these rows
4. End.

d) Support Vector machine (SVM)

SVM model is a portrayal of different classes in a hyperplane in multidimensional space. The hyperplane will be produced in an iterative way by SVM with the goal that error occurrence is minimized. The point of SVM is to divide the dataset into classes in order to get to marginal hyperplane which is maximum.

The followings are valuable terms in SVM :-

ISSN: 2233-7857 IJFGCN

• Support Vectors − These are data points that are closest to the hyperplane, It is hence called support vectors. These data points are utilized to describe the Separating line.

• Hyperplane − It is a plane or space which is apportioned between a lot of articles having a place with various classes.

• Margin − For a given hyperplane, the distance between hyperplane and closest data points is computed. It is measured as the perpendicular distance from the line to the support vectors. If this value is doubled then margin is obtained. Margin which is large is considered as a good margin and margin which is small considered as a bad margin.

It is done in the accompanying two stages :-The fundamental purpose of SVM is to divide the datasets into classes to watch out for a maximum marginal hyperplane (MMH) and it is done in the following two stages :-

• Initially separating lines i.e. hyperplanes are created repetitively by SVM, thus separating the classes in better way

• After that the hyperplane which isolates the classes accurately is chosen

5.Conclusion

The crime percentage in the world is expanding now a days because of many reasons, for example, increment in poverty, corruption, unemployment and so forth. If the crime has expanded important measures are taken by the police authorities to contemplate why the crime percentage has expanded and furthermore how to decrease crime percentage in that area. Numerous papers have been studied, just those papers with foundation in crime analysis and prediction are compared. Each paper has their own advantages and drawbacks. Each paper has its own individual methodology for solving crime. In our research work we are doing exploratory data analysis and feature engineering for e.g. to find out the major reason people being kidnapped in each and every state, age group wise murdered victim, juveniles family background and education etc.

Utilizing python the exactness of the proposed model is estimated and confirmed. Their output would be compared to find out which algorithm yields best result in terms of performance. The proposed model is helpful for the law enforcement agencies in taking necessary steps to reduce crime. In future we may anticipate crime hot spot that will help in deployment of police force at the most probable places of crime for some random window of time.

Acknowledgment

First of all, I would like to thank the supreme power the Almighty God who has always guided me to work on the right path of life. I am also grateful to my guide Dr. S. K. Wagh, for her constant encouragement, valuable guidance, comments, suggestions without whose my project work would have been incomplete and also thank the college authorities for providing the infrastructure and support required to complete my project work.

References

[1] AdelAli Alkhaibari , Ping-Tsai Chung, Ping-Tsai Chung, “Cluster Analysis for Reducing City Crime Rates”, in Long Island Systems, Applications and Technology Conference (LISAT), 2017.

[2] Alkesh Bharati, Dr Sarvanaguru RA.K ,”Crime Prediction and Analysis using Machine Learning” In International Research Journal of Engineering and technology (IRJET), Vol no. 05, pp. 2395-0072, 2018.

[3] Lalitha Saroja Thota, Suresh Babu Changlasetty, “Cluster based Zoning of Crime Info”, IEEE International Conference on Security and Privacy in Computing and Communications (Trust Com), 2017.

[4] Sunil Yadav, Meet Timbadia, Nikhilesh Yadav, Rohit Vishwakarma. "Crime Pattern Detection, Analysis and Prediction", In International Conference on Electronics, Communication and Aerospace Technology, 2017.

[5] S.R Deshmukh, Arun Dalvi, Tushar Bhalerao, Ajinkya Dahale, Rahul Bharati , Chaitali R. Kadam, "Crime investigation using Data mining", In International Journal of Advance Research in Computer and Communication Engineering (IJARCCE), Vol.4,Issue 3,March 2015.

[6] Ayidh alqahtani, Ajwani Garima, Ahmad Alaid, "Crime analysis in Chicago City", in 10th IEEE International Conference on Information and Communication System (ICICS), 2019.

[7] Jesia Quader Yuki, Md.Mahfil Quader Sakib, Zaisha Zamal, Khan Mohammad Habibullah, Amit Kumar Das "Predicting Crime Using Time and Location Data", ICCCM 2019 Association for Computing Machinery.

[8] Anand Joshi, A.Sai Sabitha, Tanupriya Chaudary, "Crime Analysis using k-mean Clustering", In International Conference on Computational Intelligence and Networks, IEEE, Vol. 18, No. 3, March 2017.

[9] Benjamin Fredrick David, A.Suruliandi, "A Survey on Crime Analysis and Prediction using Data Mining" Techniques", ICTACT Journal on Soft Computing, April 2017 vol. 07, ISSUE.03.

[10] Shyam Varan Nath, "Crime Pattern Detection using Data Mining", Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, pp. 1-4, 2006.

[11] DR: Zakaria Suliman Zubi, Ayman Altaher Mahmmud, "Crime Data Analysis using Data mining to improve Crime Prevention", International Journal Of Computers, 2014, vol. 08.

[12] Hsinchun Chen, Wingyan Chung, Jennifer Jie Xu, Gang Wang Yi Qin and M.Chau. "Crime Data Mining: A general Framework and some examples", IEEE Computer Society, vol. 037, no.4, pp. 50-56 April 2004.

[13] Md Abu Saleh, Ihtiram Raza Khan, "Crime Data Analysis using K-Means clustering", International Journal for Research in Applied Science & Engineering Technology (IJRASET), April 2019, vol. 07, ISSUE.04.

[14] Chung-Hsien Yu, Max W. Ward, Melissa Morabito, and Wei Ding," Crime Forecasting Using Data Mining Techniques", IEEE International Conference on Data Mining Workshops, pp.779-786,2011.

[15] Navjot Kaur," Data Mining Techniques used in Crime Analysis:- A Review" , International Research Journal of Engineering and Technology (IRJET), vol. 03, Issue: 08 , Aug 2016.

[16] Louise F.G underson," Using Data Mining and Judgment Analysis to Construct a Predictive Model of Crime", IEEE SMC, 2002.

[17] Mrinalini Jangra and Shaveta Kalsi, "Naïve Bayes Approach for the crime Prediction in Data Mining", International Journal of Computer Applications, vol.178 - No.14, May 2019.

[18] Keivan Kianmehr, Reda Alhajj, "Crime Hot-Spots Prediction Using Support Vector Machine", IEEE , 2006.

[19] S.Sivaranjani, Dr.S.Sivakumari, Aasha.M," Crime Prediction and Forecasting in TamilNadu using Clustering Approaches", International Conference on Emerging Technological Trends [ICETT],IEEE,2016.

[20] Bhanjeet Kaur, Laxmi Ahuja, Vinay Kumar, "Crime Against Women: Analysis and Prediction Using Data Mining techniques", International Conference on Machine Learning, Big Data and Parallel Computing,IEEE,2019.

[21] Grzegorz Borowik, Zbigniew M. Wawrzyniak, and Paweł Cichosz, "Time series analysis for crime forecasting", European Union,2018.

[22] C.A. Pina Garcia, Leticia Ramirez , "Exploring Crime patterns in Mexico City", Journal of Big Data, Springer Open,2019.

[23] Kadhim Swadi al-Janabi, "A proposed Framework for Analyzing Crime Dataset using Decision Tree and Simple k-mean Mining Algorithm", Journal of Kufa for Mathematics and Computer, vol.1, No.3, May 2011.

[24] Kaumalee Bogahawatte and Shalinda Adikari, "Intelligent Criminal Identification System", Proceedings of 8th IEEE International Conference on Computer Science and Education, pp. 633-638, 2013.

[25] Jeroen S. De Bruin, Tim K. Cocx, Walter A. Kosters, Jeroen F. J. Laros and Joost N. Kok, "Data Mining Approaches to Criminal Career Analysis", Proceedings of 6th IEEE International Conference on Data Mining, pp. 1-7, 2006.

[26] Sathyadevan, S. and Gangadharan, S., 2014, August. Crime analysis and prediction using data mining. In 2014 First International Conference on Networks & Soft Computing(ICNSC2014) (pp. 406-412). IEEE.

[27] https://scikitlearn.org/stable/modules/preprocessing.html.

[28] https://pandas.pydata.org/pandas-docs/stable/user_guide/ missing_data.html.

[29] National Crime records bureau, Ministry of home affairs, India, Software for data analysis, Crime Info(Crime in India) , http://ncrb.nic.in/index.htm.

[30] J. Han, M. Kamber and J. Pei and M. Kamber, "Data Mining, Concepts and Technologies", 3rd Edition, The Morgan Kaufmann, 2011.