www.ijiarec.com

ISSN:2348-2079

Volume-7 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

Crime analysis and prediction of crime data using mining

S.Anbukkarasi[1], S. Deepa Karthik[2], P. GopalaKrishnan[2], V.Rathika[2]

[1]Assistant Professor, Department of Information Technology, Nandha Engineering College, Erode.

[2]UG Students, Department of Information Technology, Nandha Engineering College, Erode

ABSTRACT

An approach for crime detection in India using Data mining techniques is proposed in this paper. The approach consists of the following steps - Data pre-processing, clustering, classification and visualization. Data mining techniques are often applied to Criminology as it provides good results. Criminology is a field which studies about various crime characteristics. Analyzing crime data means exploring crime data. Crime is identified using k-means clustering and the clusters are formed based on the similarity of the crime attributes. The Random Forest algorithm and neural networks are applied on the data for classification. Visualization is achieved using the Google marker clustering and the crime spots are marked on the India map. This approach will benefit the Crime department of India in analyzing crime with better prediction.

INTRODUCTION

Day by day the crime rate is growing considerably. Crime cannot be foretold since it is neither methodical nor random. Also, the up-to-date technologies and hi-tech methods help convicts in achieving their crimes. Deliberating to Crime Records Bureau crimes like burglary, arson etc. have been compact while crimes like murder, sex abuse, gang rape etc. have been augmented. Even though we cannot forecast who all may be the victims of crime but can predict the place that has probability for its rate. The predicted results cannot be assured of 100% correctness but the results show that our request assistances in reducing crime rate to a certain degree by if security in crime sensitive areas. So for structure such a powerful crime analytics tool we have to collect crime records and appraise it . It is only within the last few years that the technology made spatial data mining a practical answer for wide audiences of Law application officials which is affordable and accessible. Since the obtainability of criminal data or records is limited, we are collecting crime data from numerous sources like web sites, news sites, blogs, social media, RSS feeds etc. This large data is used as a record for creating a crime record database. So the main contest in front of us is developing a better, effectual crime pattern detection tool to identify crime patterns efficiently. The main challenges we are facing are:

Increase in crime evidence that has to be stored and analysed. Analysis of data is difficult meanwhile data is incomplete and inconsistent. Restriction in getting crime data records from Law Implementation department. Accuracy of the program depends on correctness of the exercise set. Finding the patterns and leanings in crime is a challenging issue. To categorize a pattern, crime forecasters takes a lot of time, scanning through data to find whether a specific crime fits into a acknowledged pattern. If it does not fit into an present pattern then the data must be classified as a new pattern. After detecting a pattern, it can be used to predict, forestall and prevent crime. Before this clustering algorithms have been used for crime analysis. For instance, one site it is open that

Author for correspondence:
Department of Information Technology, Nandha Engineering College, Erode

suspect has black hair and from next site/witness it is bare that suspect is youth and from third one reveals that the offender has tattoo on his left arm etc. By describing the crook details it gives a complete picture from different crime incidents. Today most of it is manually done with the help of several reports that the detectives usually get from the computer data analysts and their own crime logs. The reason for choosing this method is that we have only data about the known crimes we will get the crime pattern for a specific place. Therefore, classification technique that will rely on the prevailing and known solved crimes, will not give good predictive quality for future crimes. Also nature of crimes change over time, so in order to be able to perceive newer and unknown patterns in future, clustering techniques work better. There are steps in doing Crime Analysis:

- Data Collection
- Classification
- Prediction

RELATED WORK

In countries like England, Cambridge Police Department have done a alike one named Series Finder for finding the patterns in burglary. For achieving this they used the modus operandi of offender and they removed some crime patterns which were followed by offender. The algorithm constructs modus operandi of the offender. The M.O. is a set of ways of a criminal and is a type of behaviour used to characterize a pattern. The data comprised means of entry (front door, window, etc.), day of the week, characteristics of the property (apartment, house), and geographic proximity to other breaking. Using nine known crime series of burglaries Series Finder improved most of the crimes within these patterns and also identified nine additional crimes. The predicted result showed more than 80% accuracy. So the same concept we are smearing here i.e. find unknown patterns from known data and facts [5]. It's the first mathematically principled method to the automated learning of crime series.

METHODOLOGY

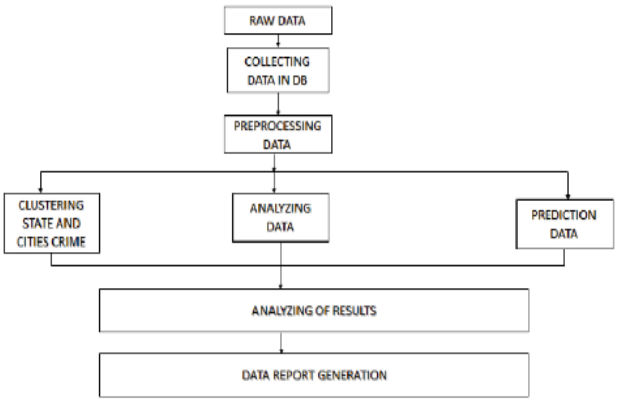
Data Collection In data collection step we are collecting data from dissimilar web sites like news sites, blogs, social media, RSS feeds etc. The collected data is stored into database for more process. Since the collected data is unstructured data, we use XAMPP. Crime data is an unstructured data since the no of field, content, and size of the document can differ from one document to extra the better option is to have a schema less database. Also, the nonattendance of joins reduces the complexity. Other aids of using an unstructured database is that Large volumes of structured, semi-structured, and unstructured data. Object-oriented programming that is easy to use and elastic. The advantage of MySQL database over SQL database is that it allows supplement of data without a predefined schema. Unlike SQL database it not needs to know what we are storing in advance, specify its size etc.

Classification For classification we are using an algorithm called Naïve Bayes which is a supervised learning method as well as a numerical method for classification. Naive Bayes classifier is a probabilistic classifier which when given an input gives a probability distribution of set of all classes rather than provided that a single output. The algorithm classifies a news article into a crime type to which it fits the best. Compared to other algorithms like SVM (Support Vector Machine) which takes lot of memory the easiness for implementation and high performance makes it different from other algorithms. Also, in case of SVM as size of training set increases the speed of implementation decreases. Using Naive Bayes algorithm we create a model by training crime data related to vandalism, murder, robbery, burglary, sex abuse, gang rape, arson, armed robbery, highway robbery, snatching etc. By training means we have to teach them on particular inputs such that we can test them for unknown inputs. For testing the correctness of the model we apply test data. Unlike SVM as the size of training data increases accuracy of test set also increases. Another benefit of Naïve Bayes is that it works well for small amount of training to compute the classification parameters. Also it fixes the Zero-

frequency problem i.e. while estimating probability sometimes while checking a probability P(A) * P(B/D) * P(C/D) * P(E/D) where P(C/D)=0. So the assessed probability result always gives zero which leads to uncertainty in results. To avoid this condition, we add +1 to the count of every zero value classes to achieve uniform distribution. Test results shows that Naive Bayes shows more than 90% accuracy since it categorises each word as tokens and removing frequent words like "the", "and", "of" etc which improves accuracy. A word is automatically terminated if it occurred fewer times or less than 3 times. We are also integrating the concept of Named Entity Recognition (NER) in the crime trainings. NER also known as Entity Extraction finds and classify elements in text into predefined categories such as the person names, organizations, locations, date, time .So by using this concept in crime article we can get more details related to crime like victim and offender names, location of crime, date, time.

Proposed approach

The first step in the model is Data pre-processing (DP) which comprises filling the missing values, data cleaning and transformation of data. The k-means algorithm is used to replace the missing values and it is replaced with the mean/method value of the corresponding attributes instance. The k-means algorithm categorizes the crime instances into clusters with alike attributes by performing the required number of repetitions. The k-means clustering is then trailed by the classification and this process is divided into two steps firstly, house a model and then using the model for classification. Classification helps in finding a set of models which can be used for future prediction of unknown class labels. By using the training dataset, the predictive accuracy of the model is slow. The correctness depends on how many instances are correctly classified and if the accuracy is good the model can be used for upcoming prediction. In this approach two organization algorithms have been used to check which gives better results for the chosen dataset and its kind. The algorithms used are Random Forest and Neural Networks (Multi-layer Perceptron). The Random Forest algorithm is a administered learning technique which creates a forest with the number of trees. It can be used for regression, organization and other tasks. The Random forest works by creating multiple decision trees during training. Using Random forest, the variables can be ranked on the basis of their priority. Neural Networks which is a nonlinear model helps in exhibiting real world complex relationships.



K-MEANS IMPLEMENTATION

In this model the existing patterns and the relations are searched in the dataset by using k-means and Google Map Marker clustering. This technique helps in providing an indication of the dataset and hence, helps in searching, handling and retrieving of the required or desired information. From the dataset 10 states are particular for formulating the clusters and the selection is done on the basis of the average IPC value of each state during 2001-2012. Since the locations are clear on the map it helps in analysing the states. The information is quite helpful for inspecting agencies and the police officials. Case 1 crime detection in India during 2001 to 2012: K-means helps in grouping objects (crimes in Indian states during 2001 – 2012) into clusters and here we are denoting each object using A, B, C, … L. Clusters are formed using the two crime attributes "year" and "total IPC crime" (Fig 5). The number of clusters by default is 2. Totally 9 iterations are performed on the dataset for case 1. Case 2 crime detection in Uttar Pradesh and Delhi during 2001 to 2012: To analyse the number of crimes during 2001 to 2012 in Uttar Pradesh and Delhi clusters are generated. The attributes "year" and "total IPC crime" are used to create the clusters and it is independent of the crime type. The clusters are generated in the same way for other states and union territories. Case 3 crime detection on Murder and Rape in India during 2001 to 2012: Here clusters are created to find the number of crimes of a particular type (rape or murder) during the 12 years (2001 to 2012) in various states and union territories of India. The attributes that are considered are "year", "crime type" and the attribute "state" is not considered for this case. Similarly, the clusters are generated for other crime types in India. Case 4 crime identification detection of type rape in Uttar Pradesh during 2001 to 2012: The clusters are created for this case using the attributes "crime type" and "district". This helps in knowing which region has the highest crime rate in a state (Uttar Pradesh). In the same way clusters are generated for other states and union territories.

CONCLUSION

The crime rates in India are growing day by day due to many factors such as increase in poverty, unemployment, corruption, etc. The 10 Indian states or union territories selected are chosen on the basis of their crime rate. This approach is very useful in studying if the crime rate is increasing or decreasing in a precise region. If the crime has increased necessary measures can be taken by the spokespersons to study why the crime has increased and also how to reduce the crime rate in that region. In this research the crime rates during 2001 to 2012 are analysed and this has helped in ranking the states and union grounds on the basis of their average IPC crime rate. The accuracy of the planned model is measured and verified. A good accuracy of 99.93 % is obtained and this verifies the accuracy of the instances.

The proposed model is very useful for both the inspecting agencies and the police officials in taking necessary steps to reduce crime. The model can be applied to any country's 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 development of this research work focuses on training both to predict the crime prone areas by using machine learning techniques. Since, machine learning is like to data mining advanced concepts of machine learning can be used for improved prediction. The data privacy, reliability, accuracy can be enhanced for enhanced prediction.

REFERENCES

[1]. Md. Abdul Awal,Jakaria Rabbi, Imran Rana 'Using Data Mining Technique to Analyze crime of Bangladesh', 2017.
[2]. Aarathi Srinivas Nadathur, Gayathri Narayanan Indraja Ravichandran, Srividhya.S, Kayalvizhi.J 'Crime analysis and prediction using big data', 2018.
[3]. Sarpreet Kaur, Dr. Williamjeet Singh 'Systematic Review of Crime Data Mining', 2017.
[4]. Rasoul Kiani, Amin Keshavarzi 'Analysis and prediction of crime clustering and classification' 2016.

[5]. Malathi. A, Dr. S. Santhosh Baboo 'An Enhanced Algorithm to Predict using data mining', 2018.
[6]. Anisha agarwal, dhanashree chougule, arpita agrawal, divya chimote 'Application for analysis and prediction of crime data using data mining', 2016.
[7]. Adewale opeoluwa ogunde Gabriel opeyemi Ogunleye, 'A Decision Tree Algorithm Based System for predicting crimes', 2017.
[8]. Ila Savant,Mayur Gade,Rohan Kalap, Ajay Kamble, Kamran Khan 'Analysis of crime data using mangoDB', 2017.
[9]. Deepika K.K, Smitha Vinod 'Crime analysis in India using data mining techniques', 2018.