Implementation of Crime Patterns Prediction Using Data Mining

Sudha Rani[1] and **Shweta Roy**[2]
[1]Maharshi Dayanand University, Delhi Road, near Delhi Bypass, Rohtak 124001, Haryana India
[2]Department of Computer Science and Engineering, J.B. Knowledge Park, Menjhawali, Old Faridabad 121101 Haryana India
[2]sguddr@gmail.com

Abstract -- **Violence against women in India is social issue which has taken its root deeply due to the society norms and economic dependence. India is traditionally male-dominated country where women have to face various types of violence from the ancient times. This paper describes a summary of the methods and techniques which are implemented in crime data analysis and prediction by using data mining. Crime prediction practices on historical data and after examining data, predict the upcoming crime with respect to location, time, day, season and year.**

Keywords: Decision Tree, Naïve Bayes Algorithm, Data Mining, Crime against Women

I. INTRODUCTION

SEMANTIC meaning of 'crime against women' is direct or indirect physical or mental atrocity to women. Crimes which are 'directed specifically against women' and in which 'only women are victims' are characterized as 'Crime Against Women' [1]. Aggression, violence and crime against women that comprise about 49 percent of the population is serious issue for us.

Day by day the crime rate is increasing. Although women may be victims of any of the general crimes such as 'murder', 'robbery', 'cheating', etc, only the crimes which are directed specifically against women are characterised as 'crimes against women'[2].

Societally sanctioned rape and sexual assault is not new in India. It has been repeatedly established that Indian men assert a claim over the bodies of women because somehow, families believe that in Indian society a woman exists as an appendage to some man in her life — father, brother, son or husband. If a woman steps across an invisible line (*lakshman rekha*), where her behaviour is seen as outrageous and unacceptable, then many people still believe that she is opening herself up to sexual assault.

According to the National Crime Records Bureau of India, reported incidents of crime against women increased 6.4% during 2012, and a crime against a woman is committed every three minutes [3].

Crimes against women are broadly classified under two categories.

Crime against women classification

Crime under IPC (Indian Penal Code)	crime under SLL (special & local laws)
1. Rape	1. Dowry
2. Kidnapping & abduction	2. Sati pratha
3. Importation of girl from foreign country	

II. LITERATURE REVIEW

Gupta *et al.* [2] explained the meaning of data mining and its process, scope and various techniques. They presented security aspects and measures related with the databases for data mining. It has been suggested that a security measure should be implemented on behalf of the company policies.

Kalyani Kadam [4] tested the accuracy of classification and prediction based on different test sets. Classification is done based on the Bayes theorem which showed over 90% accuracy.

Several data mining algorithms [5] have been compared by researchers using various real life applications. In this dissertation report, three prominent data mining techniques (Decision Trees, *a priori* and K-NN) have been studied, analyzed and compared for analyzing crimes against women using MATLAB (R2015a).

Naive Bayes classifier is proposed with novel methodologies applied for the criminal prediction problem [6]. The incident-level crime data are generated synthetically by the model itself, otherwise which are hard to obtain in practice. The proposed model is practical due to the simplicity caused by the independence assumption of the Naive Bayes.

Biswas [7] shows how Naive Bayes classifier in Rapid Miner data mining presentation tool can be used to display crime data. Rapid Miner data mining tool to provide the fully automatic parameter optimization of machine learning operator provides good validation and cross validation.

Sathyadevan and Devan [8]contend that crime prediction helps people stay away from the districts at a certain time of the day, month and season along with saving living style. In addition, having this kind of knowledge would help people to improve their living and travelling place choices.

Sharma *et al*. [9] tested the accuracy of classification and prediction based on different test sets. Classification is done based on the Bayes theorem and Time Series algorithm which showed over 90% accuracy. Using this algorithm, authors trained numerous news articles and built a model.

III DATA MINING

Data mining is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use.

Those who might like to use data mining tool have many choices. Data mining technology is actually quite similar to statistics in the way it builds a predictive model from data. Often the accuracy of that prediction depends more on correct deployment of the technology and quality of data.

Now-a-days, a wide variety of data is available on different types of media. These data is in very huge quantity, so it falls under big data category and data mining is used to process and analyze such data [1].

Big data is a field that treats ways to analyze, systematically extract information from, or otherwise deal with data sets that are too large or complex to be dealt with by traditional data-processing application software.

It is an essential process in today's world because it uncovers hidden patterns for evaluation.These patterns are detected and models are created based on it. In real world, a model can be anything from mathematical model to set of rules that describes the scenario. These patterns can be used for marketing analysis, making strategies, taking decisions, to increase revenues etc. Data mining provides a number of analytical tools and algorithms for analyzing data [5].

Model -A description of the original historical databases from which it was built that can be successfully applied to the new data in order to make predictions about missing values or to make statements about expected values.

Pattern: An event or combination of event in database that occurs more than often expected.

Data analysis is defined as a process of cleaning, transforming, and modeling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

Here we are doing crime analysis to recognize the crime patterns.

There are five steps in doing Crime Analysis:

- Data Collection
- Classification
- Pattern Identification
- Prediction
- Visualization.

Data Collection: Data is collected from different web sites like news sites, blogs, social media, RSS feeds etc. The collected data is stored into database for further processing.

Classification: An algorithm called Naïve Bayes is used, which is a supervised learning method as well as a statistical 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 providing a single output. The algorithm classifies a news article into a crime type to which it fits the best.

Pattern Identification: Third phase is the pattern identification phase, where trends and patterns in crime are identified.

Prediction: Decision tree concept is used. A decision tree is similar to a graph in which internal node represents test on an attribute, and each branch represents outcome of a test. The main advantage of using decision tree is that it is simple to understand and interpret.

Visualization: The crime prone areas can be graphically represented using a heat map which indicates level of activity, usually darker colors to indicate low activity and brighter colors to indicate high activity [2].

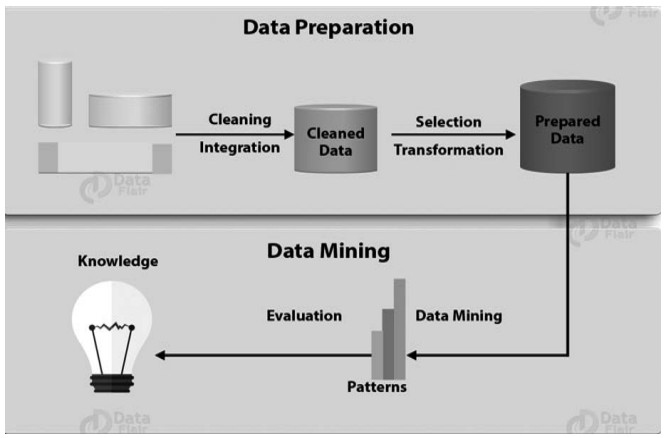


Figure 1. Crime analysis.

IV. DATA MINING TECHNIQUES

Decision Tree: A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.

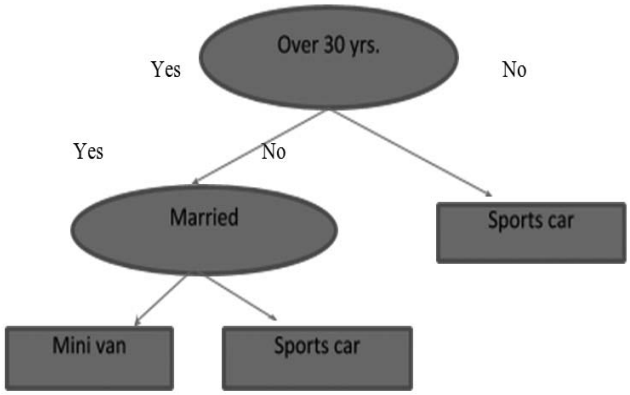
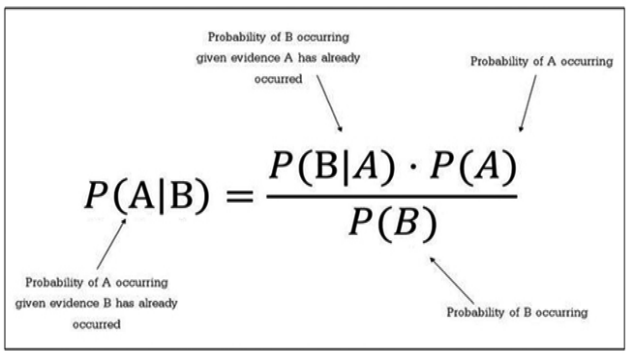


Figure 2. Decision tree.

Naive Bayes Algorithm: Naive Bayes classification is based on Bayes theorem with naive (strong) class conditional independence. Class conditional independence means the effect of an attribute value on a given class is independent of the values of other attributes [5]. The equation for Bayes theorem is given as follows [1]:



The Naive Bayes classifier is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable, hence the characterization as Naive yet the algorithm tends to perform well and learn rapidly in various supervised classification problems [5].

Algorithm 1 Pseudocode

1. Given training data set D which consists of documents belonging to different class, say class A & B.
2. Calculate the prior probability of class A=number of objects of class A / total no of objects. Calculate the prior of class B=number of objects of class B / total no of objects.
3. Find ni, the total no of word frequency of each class. na= the total no of word frequency of class A. nb=the total no of word frequency of class B.
4. Find conditional probabilty of keyword occurance given a class.
   P(word1 / class A) = wordcount / ni(A)
   P(word1 / class B) = wordcount / ni(B)
   P(word 2/ class A) = wordcount / ni(A)
   P(word2 / class B) = wordcount / ni(B)
   P(wordn / class B)=wordcount / ni(B)
5. Avoid zero frequency problems by applying uniform distribution.
6. Classify a new document C based on the probability p(C /W).
   a) Find P(A /W)=P(A) *P(word1/ class A) * P(word2/ class A)……*P(wordn / class A).
   b) Find P(B /W)=P(B) *P(word1/ class B) * P(word2/ class B)……*P(wordn / class B).
7. Assign document to class that has higher probability.

V. OBJECTIVE AND METHODOLOGY

Objective: To study and compare some of the promising data mining algorithms for analyzing and predict the crimes against women.

Methodology: The tentative process followed during the course of Research Project: (as shown in Figure 1).

Step 1: Understanding the algorithms.
Step 2: Implementing a first draft of the algorithm step by step.
Step 3: Testing with the input files.
Step 4: Cleaning the code.
Step 5: Optimizing the code.
Step 6: Comparison of the performance with other algorithms.

VI. FUTURE SCOPE

With the increase in reporting of crimes against women, there is urgent need to develop such tools and techniques that will help the concerned authorities to get the attributes of the accused person. At the same time, it will also help in taking suitable measures to mitigate increasing crime rate against women. In future, work can be done on the following points [8]:

- To study and compare other data mining classification algorithms.
- To extend the algorithm for large data set.
- To reduce its complexity.
- To improve performance of the algorithms.

- To test the applicability of the algorithms in other real life applications.

VII. CONCLUSION

Crime is a serious problem which should be controlled by societies as well as the whole world. Huge number of peoples, society regions and world is affected with crime. Crime prediction and finding relevant information from large amount of crime data is very important but challenging. If the advanced prediction about the problem can be made then crime may be stopped.

Crime forecasting can be improved by the use of efficient data collection and data mining strategies.

In this paper we 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% accuracy. Using this algorithm, we train various news articles and create a model. For testing we are inserting some test data into the model which shows better results. The pattern is used for creating a model for decision tree.

REFERENCES

[1] Reference no 2/RN/ref/2013, "Crime against women".

[2] A. Gupta, V. Bibhu and Md. R. Hussain, "Security measures in data mining", *International Journal of Information Engineering and Electronic Business*, vol.3, pp.34-39, July 2012.

[3] S. Ram and A. Doegar, "A comparative study of data mining techniques for predicting disease using statlog heart disease database", *International Journal of Advanced Research in Computer Science and Software Engineering* , vol. 5, no.6, pp. 1202- 1210, June 2015.

[4] Kalyani Kadam, "Survey paper on crime prediction using ensemble approach", *International Journal of Pure and Applied Mathematics*, vol.118, no. 8, pp.133-139, 2018, url: http://www.ijpam.eu

[5] Decision tree, http://en.wikipedia.org/wiki/Decision_tree_learning.

[6] Mehmet Sait Vural and Mustafa Go¨k" Criminal prediction using Naive Bayes theory", https://www.researchgate.net/publication/292676932 Neural Comput & Applic DOI 10.1007/s00521-016-2205-z.

[7] S. S. Biswas, "Performance analysis of Naive Bayes algorithm on crime data using rapid miner", vol.8, no.5, *International Journal of Advanced Research in Computer Science,* May-June 2017.

[8] S. Sathyadeva and M.S. Devan, "Crime analysis and prediction using data mining", https://www.researchgate.net/publications/280722606.

[9] Sapna Sharma, Monika Gupta and Parul Gupta, "Implementation of crime patterns prediction using data mining" *, IJCRT* , vol. 6, no.2, April 2018, www.ijcrt.org.

Sudha Rani obtained BTech degree from JB Knowledge Park College, Maharshi Dayanand University, Rohtak, Haryana. Pursuing MTech at Delhi Institute of Technology, Management and Research, Faridabad. Area of interest is Data Mining.

Shweta Roy is working as Assistant professor in JBKP, Fridabad. She is pursuing PhD in computer science and engineering. She is an academician having teaching experience of over 12 years.

She taught at Ajay Kumar Garg Engineering College where, she was an Assistant professor in the Department of Information Technology. Obtained M.Tech degree from Gautam Budh Technical University in the area of Middleware Web Services. She has abiding passion for teaching and has taught a number of courses namely Computer Networks, Compiler Design, Software Engineering, Automata Theory, Java Programming and C programming Concepts.

Lambert Academic Publication House, Germany published the book, "*Neural Network Based Solution for Choice of Best Web services*" in 2012 based upon her M.Tech thesis.