Implementation of Crime Patterns Prediction Using Data Mining

Sapna Sharma, Monika Gupta, Dr. Parul Gupta
Student, Assistant Professor, Assistant Professor
Department of Computer Engineering,
YMCA University of Science and Technology, Faridabad, India

***Abstract*:** Crime against women is an age old phenomenon. A total of 3,09,547 cases of crime against women were reported according among the country among the year a mixture of a combine of 2013 as compared to 2,44,271 in 2012, therefore showing a rise of 26.7% throughout the year 2013. With the increase in rate of crimes against women, there is imperative ought to analyses the data and develop tools & techniques that can help the concerned authorities to suitable measures to mitigate increasing crime against women. A number of algorithms have already been designed in the data mining field. The objective of this paper is to study and analyze the performance of distinguished data processing techniques viz. Naive Bayes & Time Series Algorithms for Predict crimes against women. The performance is measured in terms of your time taken, properly and incorrectly classified instances and accuracy. From the experimental results, it absolutely was found that Naïve Bayes &Time Series Algorithms square measure higher than alternative algorithms.

Keywords: **Decision Tree, Naïve Bayes & Time Series Algorithm, Data Mining, Crime against Women.**

I.INTRODUCTION

I.1 Crimes Against Women

Crime against women has become a prominent topic of discussion in India especially after Nirbhaya's incident on 16th of December, 2012 in Delhi. The issue comes forth time to time in the form of gang-rape, sexual harassment, acid attack, dowry death, domestic violence, human trafficking and forced prostitution, marital rapes, honour killings, stalking etc. It is deeply rooted in our Indian society despite increasing literacy rate. The major reasons of it are male dominated social and political structures, inefficient legal justice system and weak rules of law. Apart from this, social negligence of women's survival, her development and economic rights, and women's own ignorance and disregard of their own rights are also among the major reasons.

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.

In 2012, there were a total of 244,270 reported incidents of crime against women, while in 2011, there were 228,650 reported incidents. Of the women living in India, 7.5% live in West Bengal where 12.7% of the total reported crime against women occurs. Andhra Pradesh is home to 7.3% of India's female population and accounts for 11.5% of the total reported crimes against women [9].

With the increase in reporting of crimes against women, there is need for accurate and timely information to react to women crime such as identifying the age group of those who are mostly involved in crime, relation of the accused with victim. Analysis can be made regarding which age groups of girls are the main target of criminals. Apart from this, there is need to recognize public areas especially the dark areas which have high probability of crime rate so that suitable steps can be taken to prevent the same. By analyzing previous similar crime cases, we can identify the criminal or his attributes such as age group, relation etc. in new crime cases. Thus there is urgent need to analyze the data and develop tools & techniques that can help the concerned authorities to suitable measures to mitigate increasing crime against women.

I.2 Data Mining

In modern world a huge amount of data is available which can be used effectively to produce vital information. The information achieved can be used in many real life applications especially in the field of Medical science, Education, Business, Agriculture and so

on. As huge amount of data is being collected and stored in the databases, traditional statistical techniques and database tools are no longer adequate for handling this huge amount of data [10].

I.2.1 Introduction

Data mining refers to extracting or “mining" knowledge from large amounts of data. Data mining which is also called as “Knowledge discovery from data or KDD” is the process of discovering interesting patterns and relations from voluminous amount of data. It is an essential process in today’s world because it uncovers hidden patterns for evaluation. These patterns can then 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. It provides various functionalities to data like multidimensional views of data, pre-processing of data, classifying data into classes according to their features, clustering the data etc.

I.2.2 Why use data mining?

Two main reasons to use data mining:

Too much data and too little information.

Need to extract useful information from the data.

Dealing huge volumes of data with no special tools make human analysts’ work very difficult. Data mining is used especially in science and business areas where there is need to analyze voluminous amount of data to discover patterns which they could not otherwise find. Besides these, data mining can be used in any field like banking, finance, retail, engineering, medical, web etc. [8].

I.2.3 Data Mining Process

Data mining consists of five major elements as explained in figure 1.

- Extraction and transformation of data onto the data warehouse system.
- Run data on multidimensional database system in a managed way
- providing data access to business analysts and other professionals
- Data analyzing
- Presentation of data in useful and required formats such as tables and graphs.

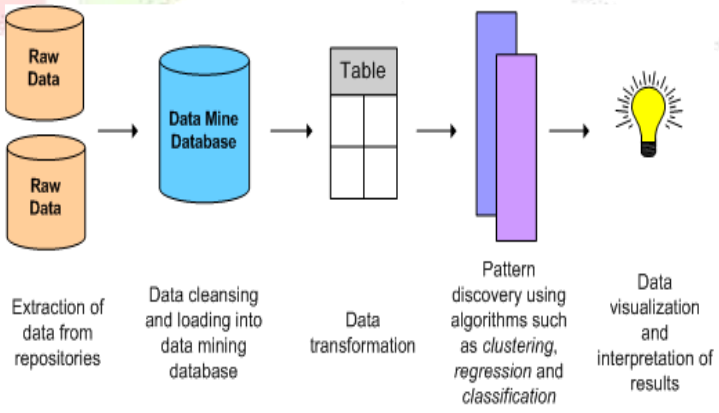

Fig 1.Data Mining Process

Since data mining is highly application-driven, it is not possible to enumerate all applications where data mining plays a critical role. Some of the notable applications of data mining are science and engineering mining, business mining, spatial data mining, visual data mining, sensor data mining, pattern mining, medical data mining, web mining etc.

II.APPROACH

In this report, three techniques (decision tree, Naïve Bayes and Time Series Algorithms) have been studied, analyzed and implemented. These algorithms are explained as follows:-

II.1.1 Decision Tree

Decision tree is a powerful classification technique. The decision trees, take the case described by its features as input, and outputs a decision. Each leaf node corresponds to a class label. A new example is classified by following a path from the root node to a leaf node, where at each node a test is performed on some feature of that example. The leaf node reached is considered the class label for that example. Figure .2 shows an example of decision tree representing root node, leaf nodes and internal nodes.

It is a flowchart-like structure in which each internal node is a test on an attribute, each branch is an outcome of test and each leaf node represents class (decision taken after computing all the attributes). A path from root to leaf represents classification rules [1]. In figure 3.shows the decision tree for analyzing crimes against women. Figure 4. Represents an activity diagram which gives estimated results of crimes against women.

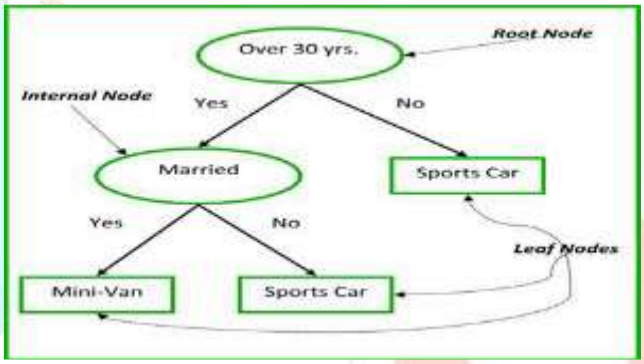

Fig 2.Decision Tree

Some of the major advantages of decision tree algorithm are:

- Simple to understand and interpret. People are able to understand decision tree models easily due to its flowchart like structure.
- **Requires little data preparation.** Other techniques often require data normalization, dummy variables etc.
- **Able to handle both numerical and categorical data.** Other techniques are usually specialized in analyzing only one type of variable.
- **Possible to validate a model using statistical tests.**
- **Robust.** Performs well even if its assumptions limited.
- **Performs well with large datasets.** Large amounts of data can be analyzed using standard computing resources in reasonable time [1].

Fig 3: Decision Tree for analyzing crimes against women

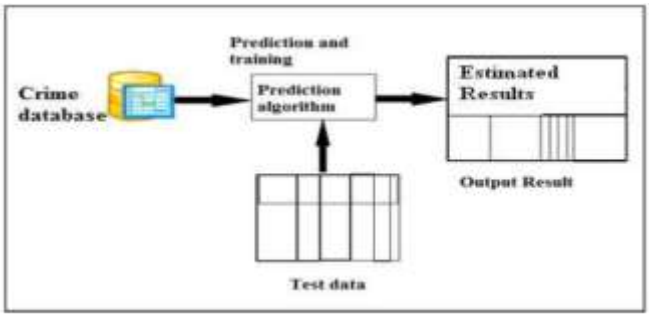

Fig 4: Activity Diagram

III.ALGORITHMS

III.1 Naïve Bayes Algorithm

For classification we are using an algorithm called Naïve Bayes and Time Series algorithms which is a supervised learning method as well as a statistical method for classification.

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 [6].

The advantage of using Naive Bayes algorithm is that it is simple, easy to use, and quicker than the other algorithms like SVM (Support Vector Machine) which takes lots of size of memory and the easy for implementation and high-performance which makes this algorithm different from other algorithms. Using Naive Bayes algorithm we create a model by teach them on particular inputs such that we can test them for unknown inputs crime data belongs to murder, robbery, burglary, sex abuse, gang rape, arson, armed robbery, highway robbery, snatching etc. For testing the accuracy of the model we apply the test data. Unlike SVM as the size of training data increases accuracy of test set also increases. Another advantage of Naïve Bayes is that it works well for small amount of training set data to calculate the classification parameters. Also it fixes the Zero-frequency problem .

Naive Bayes shows more than 90% accuracy than the other algorithms. We have shown a simple pseudo code of Naïve Bayes theorem. 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 etc.

Algorithm 1 Pseudocode

1. Given training data set D which consist 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.

III.2 Time Series Algorithm

A time series represents a collection of values obtained from sequential measurements over time. Time series data mining stems from the desired to reify our natural ability to visualize the shape of data. In fig 5.shows the time series the prediction is based on the historical data.

A time series algorithm represents a collection of values obtained from sequential measurements over time. Time series data mining stems from the desired to reify our natural ability to visualize the shape of data. In time series the prediction is based on the historical data.

Example

Calculated Percent Over Last Year

This method multiplies sales data from the previous year by a factor calculated by the system.

Required sales history: One year for calculating the forecast plus the user specified number of time periods for evaluating forecast performance .

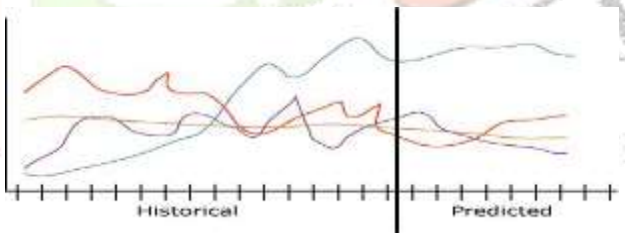

Fig 5. Time series predicted data

IV.LITERATURE REVIEW

Yu Hong et al. (2010) [4] compares four data mining techniques - logistic regression (LR), decision tree (C4.5), support vector machine (SVM) and neural networks (NN) on the basis of efficiency by applying them to two data sets of credits. The results show that the LR and SVM techniques produce the best classification accuracy, and the SVM shows the higher robustness as compared to other algorithms. On the other hand, the neural network (NN) technique performs relatively poor. its classification accuracy is unstable.

Peiying Wang (2010) [5] discusses both objective and subjective reasons for increasing women crimes in China. The prime reasons are the overall negligence of women's survival and education, her development and economic rights, and women's own ignorance and disregard of their rights. The characteristics and causes of female crimes in China are analyzed first and then appropriate strategies have been proposed with the aim to reduce female crimes.

Gupta Anish et al. (2012) [11] have explained the meaning of data mining and its process, scope and various techniques of it. The author has also discussed various security concerns of data mining and its security aspects and measures related with the databases for data mining. The security measures are very important for its applications. It has been suggested that a security measure should be implemented on behalf of the company policies.

Huang Shin-Chen et al. (2013) [6] conduct a comparative analysis on the accuracy of data mining classification techniques namely, support vector machine, decision tree, neural network and logistic regression for credit check in banking and reduce the credit risk. The support vector machine model has higher accuracy rates and therefore outperforms other classification methods in the context of credit risk in banking.

Shah Chintan et al. (2013) [7] have used three different data mining classification algorithms for prediction of breast cancer namely decision tree, Naïve bayes, and K-Nearest Neighbor with the help of WEKA (Waikato Environment for Knowledge Analysis),which is a open source software. Different parameters have been compared for prediction of cancer. But, for superior prediction, accuracy and lowest computing time have been focused. It has been concluded that Naïve Bayes is a superior algorithm compared to the two others because it takes the lowest computing time and at the same time provides highest accuracy.

Uppal Veepu et al. (2013) [10] describes the method to solve the problems faced in library because of the huge growth of library data and to improve the quality of managerial decisions. In this paper, various data mining techniques have been used that are helpful in predicting the allocation of books in library , need of the department , analysis of book circulation by time series and pattern identification of inventory loss. The main motive is that book occurrences in frequent sequences, layout of books should be arranged such that readers can easily find the books.

Bansal Divya et al. (2013) [3] has elaborated the use of association rule mining for extracting patterns within a dataset. The implementation of Apriori algorithm on a dataset containing crimes against women has been shown. For this, WEKA tool has been used for extracting results. A comparison analysis between Apriori and Predictive Apriori Algorithm has been done. The Apriori Algorithm is more efficient than Predictive Apriori Algorithm.

Ngaruiya Njeri et al. (2015) [2] uses two promising data mining tools (R Environment and WEKA) to derive patterns in Prostate Cancer. A tool has been built for identification of the Gleason score. It helps in deciding the treatment technique for Prostate Cancer. The WEKA and R Algorithm used gives almost the same results but the R Algorithm is an easier tool to learn and its representation of data is much efficient and easy to read. The patterns achieved will assist the GOK (Government of Kenya) for correctly placing the cancer diagnosis and treatment equipment which were launched by the National Government of Kenya in early 2015.

Ram Shrawan et al. (2015) [12] have done a comparative study and evaluation of decision tree and Artificial Neural Network with the help of Statlog Heart Diseases Database collected from UCI machine learning repository. These algorithms have been compared on the basis of classification accuracy and performance matrices.

Objective and Methodology

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

2 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.

V.IMPLEMENTATION AND RESULTS

The implementation of the three promising data mining techniques namely, Decision Tree, Naïve Bayes and Time Series Algorithm has been shown using Visual Studio 2010. The implementation has been carried out using Windows Operating system on a PC with Intel® Core CPU running at 2.00 GHz, with 4 GB of RAM. Below Figures shows the working of Naïve Bayes and Time Series Algorithm in Command Window of Visual Studio 2010. The predicted results cannot be assured of 100% accuracy but the results shows that our application helps in reducing crime rate to a certain extent by providing security in crime sensitive areas. So for building such a powerful crime analytics tool we have to collect crime records and evaluate it.

Fig 6.Upcoming of state DELHI of Crime KIDNAPPING and ABDUCTION362.8625

Fig 7.Upcoming of State CHANDIGARH of Crime DOWERYDEATHS45375

Fig 8.Upcoming of state HARYANA of Crime ASSAULTON WOMEN WITH INTENT TO TOUCH OUTRAGE HERMODESTY 537.7300

VI.PROBLEM FORMULATION

Crime against women is an alarming public issue not only in India but in the worldwide too. There has been a massive increase the crime rate against women. There is need for accurate and timely information to react to women crime such as identifying the age group of those who are mostly involved in crime, relation of the accused with victim, whether accused is stranger or known to the victim etc. can be of immense help. By analyzing previous similar crime cases, we can identify the criminal or his attributes such as age group, relation etc. in new crime cases.

Analysis can be made regarding which age group of girls are the main target of criminals. Apart from this, there is need to recognize public areas specially the dark areas which have high probability of crime rate so that suitable steps can be taken to prevent the same. Such information can be helpful for the Government, society and police to suggest measures to be taken towards creating a peaceful society. It will also help in the appalling situation of women in society. Thus, a comparative study of Data classification algorithms has been implemented for predict the crime against women.

VII.FUTURE SCOPE

Given databases of huge size and quality, data mining technology provides new opportunities in the research field.

1 Automation in prediction of behavior and trends

Data mining automates the process of finding information in large databases. Traditionally methods of data mining required extensive analysis by humans' hands and with data mining it has become direct to answer the predictions. An example of the same is targeted marketing where it uses data of past promotional mailing system to identify the customers who will probably maximize the return on investment in future mailings. Other examples include insurance analysis for prediction and decision making, income tax department of government for fraud discovery.

2 Automated discovery of previously unknown patterns

Data mining tools sweep through databases and identify hidden information in them. An example of pattern discovery is to identify items that are often purchased together from sales data. Other examples include detecting fraudulent credit card transactions. Data mining techniques can also be implemented on new systems as well as on existing platforms. Data mining tools can analyze massive databases in minutes. Larger databases, in turn, yield improved predictions.

VIII.CONCLUSION

In this paper we have 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 more than 90% accuracy. Using this algorithm we trained numerous news articles and build a model. For testing we are inputting some test data into the model which shows better results. 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 we 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 we 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 we have to find more crime attributes of places instead of fixing certain attributes. Till now we trained our system using certain attributes but we are planning to include more factors to improve accuracy.

Our software predicts crime prone regions in India on a particular day. It will be more accurate if we consider a particular state/region. Also another problem is that we are not predicting the time in which the crime is happening. Since time is an important factor in crime we have to predict not only the crime prone regions but also the proper time.

REFERENCES

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

[2] Njeri Ngaruiya and Christopher Moturi (2015), **"Use of Data Mining to Check the Prevalence of Prostate Cancer: Case of Nairobi County"**, IST-Africa 2015 Conference Proceedings, Paul Cunningham and Miriam Cunningham (Eds), IIMC International Information Management Corporation, 2015 (ISBN: 978-1-905824-51-9)

[3] Divya Bansal and Lekha Bhambhu (2013), **"Execution of Apriori Algorithm of Data Mining directed towards tumultuous crimes against women",** *International Journal of Advanced Research in Computer Science and Software Engineering 3(9),* Sept-2013, pp.54-62 (ISSN: 2277 128X).

[4] Hong Yu and Xiaolei Huang (2010), **"A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation"**, *IEEE International Conference on Management of e-Commerce and e-Government* (ISBN: 978-0-7695-4245-4/10)

[5] Wang Peiying (2011), **"Research on Current Female Crime Control and Prevention Strategies"** (ISBN: 978-1-61284-109-0/11)

[6] Shin-Chen Huang and Min-Yuh Day (2013) , **"A Comparative Study of Data Mining Techniques for Credit Scoring in Banking" ,** *IEEE IRI 2013, August 14-16, 2013, San Francisco, California, USA* (ISBN : 978-1-4799-1050-2/13)

[7] Chintan Shah and Anjali g. Jivani (2013), **"Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction",** *IEEE International Conference on Computing, Communications and Networking Technologies (ICCCNT),* 4-6 July, 2013 (IEEE-31661)

[8] http://www.unc.edu/~xluan/258/datamining.html

[9] Crimes against Women, http://ncrb.gov.in/CD-CII2013/Chapters/5-Crime%20against%20Women.pdf

[10] Veepu Uppal and Gunjan Chindwani (2013) **, "An Empirical Study of Application of Data Mining Techniques in Library System",** *International Journal of Computer Applications (0975 – 8887) Volume 74– No.11, July 2013.*

[11] Anish Gupta, Vimal Bibhu, Md. Rashid Hussain (2012) , **"Security measures in Data Mining"** , *International Journal of Information Engineering and Electronic Business*, Volume 3, Pages 34-39, July 2012

[12] Shrawan Ram and Amit Doegar (2015) , **"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* , Volume 5, Issue 6, June 2015 *5(6), June- 2015, pp. 1202- 1210.*