An Efficient Approach towards Crimes against Women using Time Series Algorithm

Mayank Motwani
Final Year Student
Department of CSE
Galgotias College of Engineering and Technology, Greater Noida, India

Pratha Purwar
Final Year Student
Department of CSE
Galgotias College of Engineering and Technology, Greater Noida, India

Rachit Mathur
Final Year Student
Department of CSE
Galgotias College of Engineering and Technology, Greater Noida, India

Aatif Jamshed
Assistant Professor
Department of CSE
Galgotias College of Engineering and Technology, Greater Noida, India

ABSTRACT
One of the major issues in every nation these days is the rise in crime against women. Every day we come across various cases of abuse against women. Study of past crime data can help us in analysing crime patterns and important hidden relations between the crimes. So, crimes predicting model can be simulated which will study verified past crime records and predict future criminal activities. In recent past, there has been an increased interest in time series research. This has been used particularly for finding useful similar trends in multivariate time series in various applied fields such as environmental research, agriculture, sales and finance. This paper elaborates upon the use of time series algorithm in accurately predicting and extracting patterns that occur frequently within a dataset to obtain useful hidden information.

General Terms
Data Mining, Time Series Algorithm

Keywords
Crime prediction, time series, clustering, multivariate time series

1. INTRODUCTION
India is a vast country and also one of the most diverse nations in the world. It is a place where women have been considered as goddesses from the early times. However the current scenario depicts a different picture. The status of women has undergone many changes over the past few decades. There has not only been a decline in the status of women but also there has been an incidence in the crimes against women. According to National Crime Records Bureau, crime against women has significantly increased in recent years. It is high time we start taking this problem seriously and come up with solutions to curb this inhuman behaviour against women. It has become of utmost importance to enforce law & order to reduce this increasing rate of the crime against women. This is where criminology comes into picture. Criminology is scientific study of crime and criminal behaviour in order to detect crime characteristics including its disclosure and legal aspects. Data Mining is a detailed process of analyzing large amounts of data and extracting the relevant information Use of data mining techniques can produce important results from crime dataset. Crime analysis is the field of exploring; inter relating and detecting relationships between various crimes. The police are responsible for maintaining criminal records. These records contain huge amount of data set with complex relationships. Proper analysis and study of this dataset using data mining techniques can produce important results by finding undiscovered patterns. The knowledge extracted from the dataset can be a great tool and support to the police department to prevent crimes. An ideal crime analysis tool should be able to identify crime patterns quickly and in an efficient manner for future crime pattern detection and action. All the above challenges motivated this research work to focus on providing solutions that can enhance the process of crime analysis for identifying and reducing crime against women.

The present research work proposes the use of an amalgamation of data mining techniques that are linked with a common aim of developing such a crime analysis tool. The main aim of this research work consists of using time series methods that can systematically address the complex problem related to various form of crimes against women.

2. RELATED WORK
Several data mining algorithms have been compared by researchers using various real life applications. Comparison of three prominent data mining techniques (Decision Trees, Apriori and K-NN) for analyzing crimes against women [1] shows that Decision tree is better than other two techniques. The elapsed time for decision tree is the minimum. Apriori Algorithm is also one of the good techniques. The accuracy of both decision tree and Apriori is same. However the performance of K-NN Algorithm is less in comparison to Apriori and Decision tree with the given training set.

The main reasons for increasing women crimes in China are overall negligence of women's survival and education, 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 [5] have been proposed with the aim to reduce female crimes.

Three different data mining classification algorithms for prediction namely decision tree, Naïve Bayes, and K-Nearest Neighbor [8] with the help of WEKA (Waikato Environment for Knowledge Analysis),which is an open source software, have been compared for prediction of cancer. It has been concluded that Naïve Bayes is a superior algorithm compared to the two others.

Apriori Algorithm [2] is the most popular and useful algorithm of Association Rule Mining of Data Mining. Apriori algorithm is used in mining association rules from a dataset containing women crime data. The main motive is to apply Apriori on real dataset against crimes on women which

extracts hidden information that where the real culprit is hiding.

A univariate time series model [7] takes the price of a product as a parameter that systematically influences the prediction. The price influence is computed based on historical sales data to identify products with comparable history. Compared to other techniques this approach is easy to compute.

A study of the real-world system developed for a large food distribution company details the system's forecasting algorithm [10] which efficiently handles several difficult requirements. The robustness of the system has been proven by its heavy and sustained use since being adopted in November 2009 by a company that serves 91 percent of the combined populations of Australia and New Zealand.

Mining sequential pattern in time series [12] data is broadly used in a variety of areas in order to make a prediction, and an appropriate model should be established before the prediction can be done, therefore, the way how to mine out time series pattern from time series database becomes extremely important. Experimental results demonstrate that this algorithm has mined out the frequent series, which meets the real-time restraints successfully.

Time series [3] is a collection of values obtained from sequential measurements over time. Time series data mining stems from the desire to reify our natural ability to visualize the shape of data. However, with the ever-growing maturity of time series data mining techniques, this statement seems to have become obsolete. Nowadays, time series analysis covers a wide range of real-life problems in various fields of research.

Some examples include economic forecasting [4], medical surveillance [6] and sales [7].

Hence on analyzing the comparisons between the various data mining algorithms mentioned above and considering the advantages of time series algorithm that has been applied in various fields such as sales, agriculture and other important areas that require working with accurate and timely information, we have decided to use time series algorithm.

3. BASIC THEORY

A time series is a sequence of observations $S_t \in R$, usually ordered in time. Time series is a group of ordered time-varying values or events. Mathematically, when a random process is sampled at a series of instants: 1 *t*, 2 *t*... n *t* (*t* is a independent variable, also, 1 *t* < 2 *t* < *n t)*, we get a set of sequential values: *t* 1 *X*, *t* 2 *X*, *t n X*, which is a discrete digital time sequence, i.e., a sample of the random process *X* (*t*). All the samples of the random process are stored in a time series database, which is broadly used in a variety of areas, such as data analysis in science experiments, fluctuating stock price prediction, etc. An appropriate model should be established before the prediction can be done, therefore, the way how to mine out time series pattern from time series database becomes extremely important.

Let an observed discrete univariate time series be $S_1 \ldots S_T$. This means that we have T numbers which are observations on some variable made at T equally distant time points, which for convenience we label 1, 2. . . T.

A fairly general model for the time series can be written as

$$S_t = g(t) + \emptyset_t \quad (1)$$

Where t = 1. . . T.

The observed series is made of two components:

Systematic part: g (t), also called signal or trend, which is a deterministic function of time.

Stochastic sequence: a residual term $\emptyset_t$, also called noise, which follows a probability law.

3.1 Simple Moving Average (SMA)

SMA is calculated by adding the number of crimes over a number of time periods and then dividing the sum by the number of time periods thus it is basically the average rate at a given time with equal weighing given to the individual crime.

$$SMA = (Sum (crime\ rate, n)) / n \quad (2)$$

Where n= time period.

3.2 Weighted Moving Average (WMA)

WMA focuses more on recent crimes than on older crimes. Each period's data is multiplied by a weight, with the weighting determined by the number of periods selected.

$$WMA = (Crime\ rate * n + Crime\ rate\ (1) * n-1 + \ldots Crime\ rate\ (n-1) * 1) / (n * (n + 1) / 2) \quad (3)$$

Where n=time period.

3.3 Welles Wilder Smoothing Average

The calculation of Wilder's Smoothing begins with an 'n' day simple moving average for the initial calculation. For the next step drops 1/14th of the previous average value and add 1/nth of the new value. These indicator smoothes price movements to help you identify and spot bullish and bearish trends. Welles Wilder is fastest among other moving averages as it's formula carries a smaller percentage of historical data in its calculation. It is basically used to identify trend direction, support and resistance level.

$$WSMA\ (i) = (SUM1-WSMA1+CLOSE\ (i))/N \quad (4)$$

Where

WSMA1 = Wilder's Smoothing for the first period,

WSMA (i) = Wilder's Smoothing of the current period (except for the first one),

CLOSE (i) = current closing,

N = smoothing period.

4. PROPOSED WORK

Time series algorithm is very useful approach that can be used for this purpose. It has been previously applied to fields such as sales [7], agriculture [4] etc. This research focuses on crimes concerning women, which extracts hidden information by comparing on the results from previous years about which areas are more prone to these kinds of crime , for this various algorithms are used and have results which satisfy this problem up to an extent but time series works faster compared to other algorithms .Euclidean distance measure is commonly used for non time series data clustering but it is not suitable for time series clustering so for this various methods of time series algorithm are used in this research which can give better results.

The proposed approach involves mining the data sets from the past records that have been registered in the courts. The data from data.gov.in will be used for this purpose. Data.gov.in is a platform for supporting Open Data initiative of Government

of India. The portal is intended to be used by the Government of India, Ministries/ Departments, their organizations to publish datasets, documents, services, tools and applications collected by them for public use. This data is of violence against women which can fit into several broad categories. These include violence carried out by "individuals" as well as "states".After this, we compare the performance of various algorithms that have already been used for obtaining timely information so that suitable steps can be taken to reduce the crimes against women .Three promising data mining algorithms viz. Decision trees [1], Apriori [2] and K- Nearest Neighbours [1] have already been used for analysing crimes against women. Still even after using these algorithms, it is difficult to make accurate predictions or find the location of the criminal, which can be done by prediction rates. Hence we need more accurate and timely information.Now we will apply time series on real dataset against crimes on women which extracts hidden information that what age group is responsible for this and to find where the real culprit is hiding.

4.1 Dataset Preparation

Dataset contain accurate information about several crime domains for preferably up to 10 years. Example of domains such as - Murder, Rape, Drug Trafficking, Sexual Harassment, Burglary, Theft, Pickpocket etc. are the most common data types along with areas more prone with their respective rates A domain can be expressed as - $\Sigma\ (i, y, s)\ n\ y$=1 where, i= Name of the crime, y= Considered time period, s= Location.

4.2 Calculating Probability of Crime Occurrence

Using the following mathematical steps we determine the probability of occurrence of each crime –

a) Since some crimes have more importance than others, we use a precedence factor to label the extremity of a specific crime. E.g. - Murder will have larger precedence factor than Pickpocket or Snatching. It is expressed as p (i, y) = Precedence Factor of crime i in location s at a period of time y.

b) Because the crime that has been occurring at a larger scale in recent times has a higher probability of happening again than the crime that used to occur at a larger scale in previous times. So, we take into calculation a Time Impact Factor. Expressed as f (i, y, s) = Time Impact Factor of crime i in location s at a period of time y.

Then, we take into consideration the number of occurrence of a specific crime in a time period y because the crime occurring more has a bigger probability of occurring again than the crime with less number of occurrences.

4.3 Methods Used

A moving average is a technique to get an overall idea of the trends in a dataset; it is an average of any subset of numbers. Moving average is extremely useful for forecasting long term trends. It can be calculated for any period of time. Time series forecasting is used to forecast the next value(s) in the series. Here we will use the Simple Moving Average, Weighted Moving Average and Welles Wilder Smoothing for predicting future values and thus finding the trends in the occurrence of crimes so that proper steps can be taken to extract hidden information.

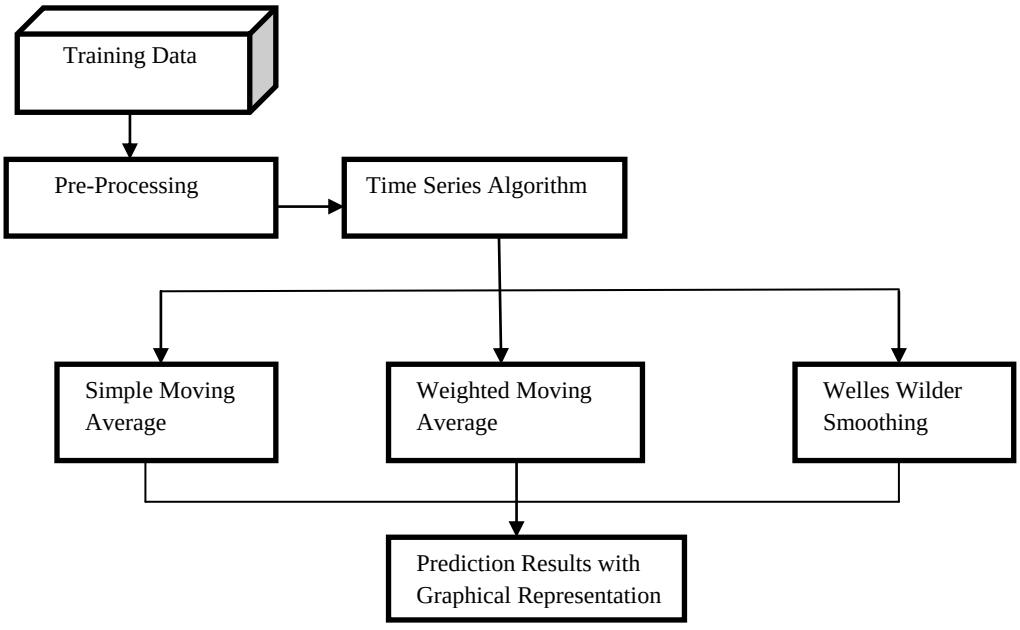

Fig 1: Overview of the Proposed System

5. EXPERIMENT RESULT

We have used three methods of time series algorithm viz. Simple Moving Average, Weighted Moving Average and Welles Wilder where prior two have been used in various fields but not used in this domain so using both the methods in addition to Welles Wilder we have got the best result of crime prediction as per future references so that we can stop or prevent crime in a particular area by prioritizing the area which we should focus more. In this domain we can't take risk by just depending on a particular method so we need comparison among the methods available in an efficient

manner so as to increase its accuracy in comparison to previous experiments.

We have analysed our result by comparing up on Naïve Bayes and Time Series i.e. the forecasting done is using both the algorithms and which is best is taken into consideration and the results coming out have been mathematically processed to get the prediction rates for future . The table 1 shown here has results based on Naïve Bayes and Time series algorithm.

Table 1. Results based on Naïve Bayes and Time Series algorithm

Instance	Value	Forecast	Error	Instance	Value	Forecast	Error
0	4816	4816	0	0	4816	4816	0
1	4849	4816	-33	1	4849	4816	-33
2	4634	4849	215	2	4634	0	-4634
3	5147	4634	-513	3	5147	4675.35	-471.65
4	4730	5147	417	4	4730	5055.15	325.15
5		4730		5		4745.4225	

Upcoming Of State:ANDHRA PRADESHof crime:ASSAULT ON WOMEN WITH INTENT TO OUTRAGE HER MODESTY4745.4225

6. RESULT ANALYSIS

In this paper, datasets from data.gov.in portal have been used as training and testing data. The crime data can be represented as a multivariate time series in terms of year, location (state) of occurrence and type of crime. The dataset of assault on woman with intent to outrage her modesty for the state Andhra Pradesh is listed in table 2.

In a similar manner, crime data for various types of crimes occurring in different locations can be processed to make accurate predictions for future references.

The data for the last 10 years has been used for predicting future crime patterns in this model.

Year	Total Crime	Id
2003	4577	1
2004	4259	2
2005	4003	3
2006	4431	4
2007	4893	5
2008	4922	6
2009	5441	7
2010	4522	8
2011	4554	9
2012	4834	10

Table 2. Test Case Data - Assault on Woman with Intent to Outrage her Modesty for the State Andhra Pradesh

This data can be represented in a graphical form as shown in figure 2. The years have been taken on the x axis and the total crime is shown on the y axis. After applying time series forecasting methods on the datasets, we obtain accurate and timely crime rate predictions of various types of crimes in different states.

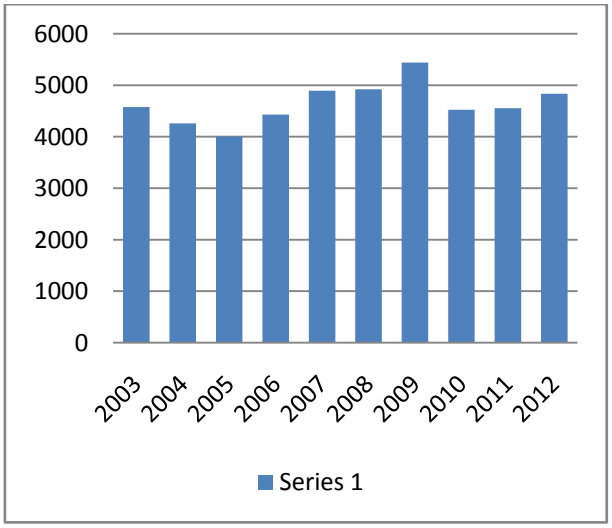

Fig 2: Graphical Representation of Test Case Data

A comparison of the results of Naïve Bayes and Time Series predictions for the assault on women with intent to outrage her modesty for the state Andhra Pradesh, given in table 1, shows that the prediction made by time series methods is more accurate.

7. CONCLUSION

Research is an essential and powerful tool in leading mankind towards progress. The purpose of any research is to find solutions to problems and thus advance knowledge through the application of scientific procedures. In this research, Time Series Algorithm is used to uncover and understand the underlying patterns in the court's records from their data in various sections. Cruelty and crime against women are rampant not only in India but also in most advanced countries. Hence, there is a need for accurate and timely information to assist in changing this pathetic condition of women. This will be helpful for the government, society and police to take actions against those responsible and come up with measures to curb these crimes against women. Result of this research will be used to analyse and predict crimes from the huge data set available. Results will be in the form of relation between various crimes, types of crime and location of crime i.e. state/city.

8. FUTURE WORK

With the rise in reporting of crimes against women, there is an urgent need to come up with such models and techniques that will help the concerned authorities to get the attributes of the accused person. This will help the government in directing their efforts in a definite direction. In future, we can also correlate crime on the basis of age group, location of crime & type of crime. Prediction of crime will be displayed using various diagrams pie charts, heat maps, spikes and graphs which will help in better understanding of the crime patterns. In future, work can also be done to study and compare other

data mining classification algorithms. We can also extend the algorithm for large data set.

9. ACKNOWLEDGMENTS

The authors are grateful to the referee for valuable comments and suggestions in improving the paper.

10. REFERENCES

[1] Aarti Bansal, Comparative Study of Data Mining Algorithms for Analyzing Crimes against Women. International Journal of Innovations & Advancement in Computer Science, Volume 4, Issue 9 September 2015.

[2] Divya Bansal and Lekha Bhambhu, Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013.

[3] Philippe Esling and Carlos Agon, Time series data mining, ACM Computing Surveys, 2013.

[4] Girish K. Jha and Kanchan Sinha, Agricultural Price Forecasting Using Neural Network Model: An Innovative Information Delivery System, Agricultural Economics Research Review, Vol. 26 (No.2) July-December 2013 (pp 229-239).

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

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

[7] Michael Schaidnagel, Christian Abele, Fritz Lauxy, Ilia Petrovy, Sales Prediction with Parameterized Time Series Analysis, The Fifth International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA), 2013.

[8] Chintan Shah and Anjali g. Jivani, 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) 2013.

[9] Veepu Uppal and Gunjan Chindwani, 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

[10] Neal Wagner and Zbigniew Michalewicz, Intelligent techniques for forecasting multiple time series in real-world systems, IJICC, 2011.

[11] Xiangchun Xiong and Yangon Kim , Analysis of Breast Cancer Using Data Mining & Statistical Techniques , IEEE Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Networks (SNPD/SAWN'05) (ISBN: 0-7695-2294-7/05) 2005.

[12] Chong Zhu , Xiangli Zhang , Jingguo Sun , Bin Huang, Algorithm for Mining Sequential Pattern in Time Series Data, International Conference on Communications and MobileComputing,2009.

[13] Aatif Jamshed and Pawan Singh Mehra (2012), "Modified Block Playfair Cipher using Random Shift Key Generation", International Journal of Computer Applications, Vol. 58, pp. 2012/1/1.

[14] Aatif Jamshed ,Surbhi Chandhok and Romil Anand (2017), "Analysis of Sequential Mining Algorithms", International Journal of Computer Applications, Vol. 165, pp. 12-2017/5.

[15] Aatif Jamshed ,Surbhi Chandhok and Romil Anand (2017), "An Analysis of Sentimental Data using Machine Learning Techniques", International Journal of Computer Applications, Vol. 166, pp. 3-2017.

[16] Aatif Jamshed ,Garima Verma(2013), "Mobile Devices integration with Grid by Using Efficient Scheduling for Local Resource", Journal of Advanced Computing and Communication Technologies (ISSN: 2347 - 2804)Volume No. 1 Issue No.2, December 2013.