Review of modeling technologies used for Predicting Crime

Vivek Jaglan

CSE Deptt, Amity University, Gurugram, Haryana

***Abstract*—Crime against women these days has become problem of every nation around the globe many countries are trying to curb this problem. Preventive are taken to reduce the increasing number of cases of crime against women. A huge amount of data set is generated every year on the basis of reporting of crime. This data can prove very useful in analysing and predicting crime and help us prevent the crime to some extent. Crime analysis is an area of vital importance in police department. Study of crime data can help us analyses crime pattern, inter-related clues& important hidden relations between the crimes. That is why data mining can be great aid to analyses, visualize and predict crime using crime data set. Classification and correlation of data set makes it easy to understand similarities & dissimilarities amongst the data objects. We group data objects using clustering technique. Dataset is classified on the basis of some predefined condition. Here grouping is done according to various types of crimes against women taking place in different states and cities of India. Crime mapping will help the administration to plan strategies for prevention of crime, further using data mining technique data can be predicted and visualized in various form in order to provide better understanding of crime patterns.**

***Keywords*—Modeling Technologies Used For Predicting Crime Against Women**

I. INTRODUCTION

The concern about national security has increased significantly since the terrorist attacks on November 26, 2008 at Mumbai. Intelligence agencies such as the CBI and NCRB (National Crime Record Bureau) are actively collecting and analyzing information to investigate terrorists' activities [12]. Local law enforcement agencies like SCRB(State Crime Record Bureau) and DCRB(District Crime Record Bureau)/CCRB (City Crime Record Bureau) have also become more alert to criminal activities in their own jurisdictions. One challenge to law enforcement and intelligence agencies is the difficulty of analyzing large volumes of data involved in criminal and terrorist activities. Data mining holds the promise of making it easy, convenient, and practical to explore very large databases for organizations and users. In this paper, we review data mining techniques applied in the context of law enforcement and intelligence analysis.

The notion of crime forecasting dates back to 1931, when sociologist Clifford R. Shaw of the University of Chicago and criminologist Henry D. McKay of Chicago's Institute for Juvenile Research wrote a book exploring the persistence of juvenile crime in specific neighborhoods. Scientists have experimented with using statistical and geospatial analyses to determine crime risk levels ever since. In the 1990s, the National Institute of Justice (NIJ) and others embraced geographic information system tools for mapping crime data, and researchers began using everything from basic regression analysis to cutting-edge mathematical models to forecast when and where the next outbreak might occur. But until recently, the limits of computing power and storage prevented them from using large data sets.

In 2006, researchers at the University of California, Los Angeles (UCLA), and UC Irvine teamed up with the Los Angeles Police Department (LAPD). By then, police departments were catching up in data collection, making crime forecasting "a real possibility rather than just a theoretical novelty," says UCLA anthropologist Jeffrey Brantingham. LAPD was using hot spot maps of past crimes to determine where to send patrols—a strategy the department called "cops on the dot." Brantingham's team believed they could make the maps predictive rather than merely descriptive.

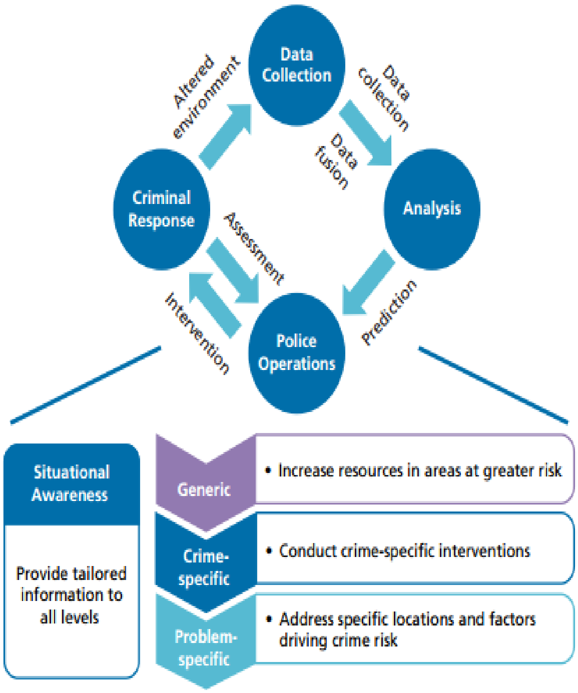

Figure 1: Process

Making "predictions" is only half of prediction-led policing; the other half is carrying out interventions, acting on the predictions that lead to reduced crime (or at least solve crimes). What we have found in this study is that predictive policing is best thought of as part of a comprehensive business process. That process is summarized in Figure S.1. We also identified some emerging practices for running this business process successfully through a series of discussions with leading predictive policing practitioners. At the core of the process shown in Figure S.1 is a four-step cycle (top of figure). The first two steps are collecting and analyzing crime, incident, and offender data to produce predictions. Data from disparate sources in the community require some form of data fusion. Efforts to combine these data are often far from easy, however. The third step is conducting police operations that intervene against the predicted crime (or help solve past crimes). The type of intervention will vary with the situation and the department charged with intervening.

II. ROLE OF THEORY IN PREDICTIVE MAPPING

As described in detail below, various means of forecasting crime events and locations exist, and not all of them can be considered "modeling." Some methods are strictly atheoretical, relying on past events to predict future ones. Other methods, however, are developed by modeling the behavior of likely offenders, making it important to review the theory underlying these efforts because theory can play an important role in guiding the selection of independent variables, or leading indicators. Perhaps the most germane theories for forecasting purposes are the rational choice perspective and routine activities theory. Both assume that crime is purposive and that individuals are selfdetermining: when people commit crime, they are seeking to benefit themselves, and certain calculations are involved in determining whether the criminal act will yield positive results (Clarke, 1997). Thus, offenders are influenced by situational and environmental features that provide desirable — or undesirable — offending opportunities. These theories are based upon the belief that criminals engage in rational (if bounded) decision-making (Becker, 1968; Cornish and Clarke, 1986), and that characteristics of the environment offer cues to the offender that promising opportunities for crime exist (Brantingham and Brantingham,1978, 1981; Newman, 1972; Cohen and Felson, 1979; Harries, 1980; Wilson and Kelling, 1982). The practical implications of these theories are that even motivated criminals may nonetheless be deterred from committing crime if they perceive a potential target to be too risky, to involve too much effort, to yield too meager a profit, or induce too much guilt or shame to make the venture worthwhile (Clarke, 1997; Clarke and Homel, 1997). From a predictive modeling perspective, then, these theories have the potential to guide the selection of independent variables with a focus on those that characterize desirable targets — and in turn, desirable locations — of crime. Further, theory-based modeling enables us to identify which factors influence crime target selection, and thus inform crime prevention efforts. The models described below include an assessment of whether they are supported by theory, and the extent to which they inform prevention efforts.

III. DATA MINING

In this paper we review the Crime Data Mining in two directions

1. Crime Types and security concerns
2. Crime Data Mining Approaches and technique

Crime types and security concerns Crime is defined as “an act or the commission of an act that is forbidden, or the omission of a duty that is commanded by a public law and that makes the offender liable to punishment by that law” (Webster Dictionary). An act of crime encompasses a wide range of activities, ranging from simple violation of civic duties (e.g., illegal parking) to internationally organized crimes (e.g., the 9/11 attacks). The following are the different types of crimes

- Property crime
- Violent Crime
- Crime against Women and Child
- Traffic Violations. Cyber Crime and
- Others Crime data mining approaches and techniques

Data mining is defined as the discovery of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data [1]. Data mining in the framework of crime and intelligence analysis for national security is still a young field. The following describes our applications of different techniques in crime data mining. Preprocessing has been used to keep the data set ready for the process. Entity extraction has been used to automatically identify person, address, vehicle, and personal properties from police narrative reports [2]. Clustering techniques has been used to cluster the city crime data mining depends on the crimes. Classification has been used to detect criminal data from the city crime data base. Social network analysis has been used to analyze criminals’ roles and associations among entities in a criminal network [9].

IV. PREDICTION OF CRIME TRENDS

The next task is the prediction of future crime trends. This meant we tracked crime rate changes from one year to the next and used data mining to project those changes into the future. The basic method here is to cluster the cities having the same crime trend, and then using ”next year” cluster information to classify records [11]. This is combined with the state poverty data to create a classifier that will predict future crime trends. Few "delta" attributes were applied to city crime clustering: Murder for gain, Dacoity, Prep.&Assembly For Dacoity, Robbery, Burglary, Theft, Murder, Attempt to commit murder, C.H.Not Amounting to murder, Hurt/Grievous Hurt, Riots, Rape, Dowry Death, Molestation, Sexual Harassment, Kidnapping & Abduction of others, Criminal Breach of Trust, Arson, Cheating, Counterfeiting, and Others IPC crimes. These attributes were clustered using ‘Weka 3.5.8’s, Simple EM (expectation maximization)’ with parameters of "EM -I 100 -N 4 -M 1.oE-6 -S 100" [4]. EM is a deviation of K-Means clustering. Four clusters were chosen because it produced a good distribution with a relatively easy to interpret set of clusters [5]. Usually, the high level interpretation of clusters from an unsupervised algorithm is not easily defined. However, in this case, the four clusters produced had the following attributes: Note: The clusters are ordered from best to worst.

i. Crime is steady or dropping. The Sexual Harassment rate is the primary crime in flux. There are lower incidences of: Murder for gain, Dacoity, Preparation for Dacoity, rape, Dowry Death and Culpable Homicide.
ii. Crime is rising or in flux. Riots, cheating, Counterfeit, and Cruelty by husband and relatives are the primary crime

rates changing. There are lower incidences of: murder and kidnapping and abduction of others.

iii. Crime is generally increasing. Thefts are the primary crime on the rise with some increase in arson. There are lower incidences of the property crimes: burglary and theft.

iv. Few crimes are in flux. Murder, rape, and arson are in flux. There is less change in the property crimes: burglary, and theft. To demonstrate at least some characteristics of the clusters.

V. RELATED WORK

There has been countless of work done related to crimes. Large datasets have been reviewed, and information such as location and the type of crimes have been extracted to help people follow law enforcements. Existing methods have used these databases to identify crime hotspots based on locations. There are several maps applications that show the exact crime location along with the crime type for any given city (see Figure 1). Even though crime locations have been identified, there is no information available that includes the crime occurrence date and time along with techniques that can accurately predict what crimes will occur in the future. Figure 1. Map of downtown Denver, Colorado with crime locations[6] On the other hand, the previous related work and their existing methods mainly identify crime hotspots based on the location of high crime density without considering either the crime type or the crime occurrence date and time. For example, related research work containing a dataset for the city of Philadelphia with crime information from year 1991 - 1999. It was focusing on the existence of multi-scale complex relationships between both space and time [1]. Another research titled “The utility of hotspot mapping for predicting spatial patterns of crime” looks at the different crime types to see if they differ in their prediction abilities [7]. Other existing works explore relationships between the criminal activity and the socio-economics variables such as education, ethnicity, income level, and unemployment [1]. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.4, July 2015 3 Despite all of the existing work, none of them consider the three elements (location, time, crime type) together. In addition, there is very little research that can accurately predict where crimes will happen in the future [7]. In our study, we provide a data-mining model for crime prediction based on crime types and using spatial and temporal criminal hotspots.

D.E. Brown constructed a software framework called ReCAP(Regional Crime Analysis Program) for mining data in order to catch professional criminals using data mining and data fusion techniques. In 2009, Li Ding et al.[11] propose an integrated system called PerSearch that takes a given description of a crime, including its location, type, and the physical description of suspects(personal characteristics or vehicles) as input. To detect suspects, the system will process these inputs through four integrated components: geographic profiling ,social network analysis, crime profile, and physical matching. Essentially, geographic profiling determines where" the suspects are, while other components determine the suspects. De Bruin et. al. (2006) introduced a framework for crime trends using a new distance measure for comparing all individuals based on their profiles and then clustering them accordingly .This method also provided a visual clustering of criminal career sand identification of classes of criminals .From the literature study, it could be concluded that crime details increasing to very large quantities running into zota bytes(1024bytes). This in turn is increasing the need for advanced and efficient techniques for analysis. Data mining as an analysis and knowledge discovery tool has immense potential for crime data analysis. As is the case with any other new technology, the requirement of such tool changes, which is further augmented by the new and advanced technologies used by criminals. All these facts confirm that the field is not yet mature and needs further investigations

VI. REVIEW OF METHODS

Hot Spots

The most common method of "forecasting" crime in police departments is simply to assume that the hot spots of yesterday are the hot spots of tomorrow. Crime analysts prepare maps of crimes that have already occurred and those maps are used to deploy officers and to identify areas in need of intervention. While surprisingly scant research exists to test this assumption, the few studies we have identified suggest that the effectiveness of this approach depends upon the time period employed. Spelman (1995) found that examining past crimes over a one-month period is not a particularly powerful predictor — hardly better than chance, yet one year of data predicts with 90% accuracy.8 This suggests that hot spots may flare up and diminish over relatively short time periods, but that these flare-ups nonetheless occur in the same places over time, creating longer-term trends. Thus, law enforcement agencies that examine last week's crime statistics to deploy patrols may find it more useful to identify hot spots based on an entire year's worth of data.

Repeat Victimization

The above research indicates that temporally aggregated hot spots may serve as accurate predictors of crime, but that relying on shorter previous time periods for predictive purposes is less effective. The exception to these research findings relates to "hot dots" (Pease and Laycock, 1996) rather than hot spots: that is, the repeat victim rather than the high-crime area. The concept of repeat victimization is now well established in the criminological literature (for an early review, see Farrell, 1995): those individuals or places that have been victimized once are likely to be victimized again, and the time course to subsequent victimization is a few short months (Anderson et al., 1995; Farrell and Pease, 1993; Polvi et al., 1990). This research suggests that past victimizations of individual addresses, places, and businesses can be very accurate predictors of future victimizations, even when relying on the previous month's victimization. The crime prevention benefits of focusing on repeat victims to prevent subsequent crimes is well established (for a summary of the preventive impact of repeat victimization strategies on crime, see Pease, 1998), and raises the question: Can repeat victimizations of individuals and places be used to predict not just hot dots, but hot spots? Very few researchers to date have examined the extent to which hot spots are composed of repeat victimizations, except for those who have focused on residential burglary. Both Bennett (1996) and Townsley et al. (2000) found that one-third of all burglaries reported in the hot spot areas under study were repeat burglaries. While Morgan (2001) found a lower degree of repeat victimization concentration within high-crime areas, the areas under study combined multiple census districts and thus were larger than the average hot spot. In his research, Morgan also found what he terms "near repeats": residences close to repeat victims were likely to be victimized. Finally, a recent publication by Farrell and Sousa (2001) concludes that while repeat victimizations and hot spots do coincide, some hot spots experience more repeat victimization than others and may vary based upon crime type.

Univariate Methods

There are a variety of univariate methods available to predict crime. These methods use previous values of one variable to predict its future value. They are attractive because of their straightforwardness: univariate methods require a minimum of data collection since they involve only one variable. Additionally, they are atheoretical and thus do not demand any thought about which variables should be included in the analysis. These methods range from simple random walk and naive lag 12 to more sophisticated models that incorporate both seasonally and time trends. Among police practitioners the most frequently used crime prediction methods are so-called "naive" univariate ones (Gorr et al., 2002; Gorr and Olligschlaeger, 2001). The two naive univariate methods used by police are the random walk13 or and the naive lag 12.14 The random walk method is a good predictor of series in which there are frequent pattern changes (e.g., to predict stock market behavior) because it reflects those changes immediately. However, it is a poor predictor when the series to be forecasted has seasonally or time trends (Gorr et al., 2002; Gorr and Olligschlaeger, 2001). While these basic univariate methods are by far the most straightforward methods of predicting crime, they are also unfortunately by far the least accurate (Gorr et al., 2002; Gorr and Olligschlaeger, 2001). More sophisticated univariate methods are available that more accurately predict crime levels by including seasonally in the model,15 accounting for time trends using exponential smoothing16 and pooling data17 (Gorr et al., 2002; Gorr and Olligschlaeger, 2001). However, the addition of these steps also makes the methods more complicated for the user. While all the more sophisticated univariate methods offer improvements over the simpler ones, the exponential smoothing methods have two main advantages as tools for crime forecasting. First, they offer the ability to account for changes in crime over time rather than relying on the current period to forecast the next period.

Leading Indicators

"Leading indicator" multivariate methods focus on using current and past values of independent variables to predict the future value of the dependent crime variable (Gorr and Olligschlaeger, 2001). The "leading indicators" term in the title of the model refers to specific characteristics of areas or neighboring areas (e.g., shots fired, calls for service, disorderly conduct offenses, etc.) for which their rise or fall in current and previous months can be used to predict future values of the dependent crime variable. There are three issues that must be addressed when specifying a "leading indicator" model related to crime (Gorr and Olligschlaeger, 2001). First, leading indicator methods require the identification of leading indicator variables before the model can be used. Identification of the appropriate leading indicator variables requires a thorough review of the literature, and grounds this method in theory. Developing theory-based leading indicators is a time consuming task that is critical to the success of the model.18 Second, because crime forecasts are typically done for short time periods and across smaller areas, often there are not enough events to develop robust model parameters. Thus, it is important to pick an areal unit that is large enough to provide adequate numbers of observations. In general, the greater the volume of crime the more reliable the forecasts will be, and the smaller the volume of crime the more variable the data will be and the more unreliable the forecasts. Gorr and Olligschlaeger (2001) determined that a grid with 4,000-foot cells was the smallest grid cell that would still provide reliable forecasts.

Point Process Model

A new method being employed by Brown (2001) and his colleagues is based on the theory of point patterns and multivariate density estimation, and can best be described as a point process model (Brown, 2001). The modeling is akin to neural networks in that there is training involved, and past data are used to predict future events. In essence, this approach glues multivariate models together and uses notions from kriging and density estimation (Brown, 2001). Brown et al. (2000) developed this predictive model based upon the preferences of offenders, or what they term "event initiators": past behavior illustrating the preferences of offenders is used to model both when and where future crimes will occur: "...we do not regard the past crime intensity at a site as a direct factor to influence how soon criminals are going to strike again. However, this past behavior does tell us about the preferences of site selectors and we directly model those preferences..." (Brown et al., 2000:4). The output of the model is a probability surface indicating likely areas of future crimes.

Artificial Neural Networks

One of the earliest efforts to do predictive crime mapping was that of Olligschlaeger (1997), who employed a "feed-forward network with backpropogation" to predict areas where future drug markets will emerge. Best known to laypeople as artificial intelligence, the type of neural network model employed by Olligschlaeger is capable of learning extremely complex space-time patterns (Olligschlaeger, 1997). According to Olligschlaeger, "The goal is to map the input units to a desired output similar to the way in which the dependent variable is a function of the independent variables in a regression analysis. The difference is that regression analysis uses linear direct mapping whereas multi-layer feed-forward networks use non-linear direct mapping" (Olligschlaeger, 1997:325). In essence, the network is trained by feeding it past data and adjusting the weights assigned to the input units; when the network is processed, the error signal is fed back, or "backpropogated" through the network to adjust the weights until, ultimately, the error signals are minimized (Olligschlaeger, 1997). GIS was employed in conjunction with this model in order to process spatial and temporal data, including data aggregation and determination of spatial and temporal lags. This was accomplished by overlaying a grid and summing the data points that fall into each cell, as well as employing contiguity measures.

Polygon Grid/Raster GIS Methods

As stated earlier, GIS can be used throughout a research project to incorporate spatial relationships in the crime forecast. In Groff and La Vigne (2001), we used a combination of polygon grid cells and raster-based GIS to generate an opportunity surface for residential burglary. We identified a set of variables based on the theories of routine activities, rational choice and environmental criminology, and used GIS to operationalize those variables. We were very interested in modeling the effects of the values of surrounding properties on a particular property. For example, we wanted to model the effect of having a substandard housing unit or vacant unit nearby. In order to incorporate the effect of surrounding grid cell's values, each layer was recalculated using a focal neighborhood function within the GIS.20 Map algebra was used to combine the new grid cell values in each layer to produce an overall risk index surface for residential burglary. Reported burglaries were then plotted on top of this new opportunity surface to determine how well the model predicted. The percentage of cells that were accurately predicted was used to empirically compare the actual burglaries with the opportunity surface. Two categories of burglaries were examined: any burglary event and repeat burglaries.

VII. PREDICTIVE ANALYSIS FOR THE EFFECTIVENESS OF CURRENT LEGAL PROTECTION TO WOMEN VICTIMS

In this section we carry out a predictive analysis of data related to the number of cases registered under IT Act (section 67). Training dataset used for future predictions contains data from year 2002 to 2009. Then we have plotted the actual number of cases registered after 2009 and the predicted number of cases in Figure 1. From Figure1 it is clear that there is much difference between the predicted and actual number of cases registered. Hence it is proved that IT Act 2008 amendments are ineffective to curb cybercrime. By observing the Figure 3 one can say that there is a vast difference between the predicted values and the actual values. Similar data related to the number of cases registered under IT Act in Delhi was collected and predicted. against offenders committing cyber crime against women.

Sl. No.	YEAR	Actual number of cases Registered	Predicted cases to be registered
1	2002	11	-
2	2003	8	-
3	2004	34	-
4	2005	37	-
5	2006	37	-
6	2007	56	-
7	2008	63	-
8	2009	95	-
9	2010	221	117.77
10	2011	414	142.2
11	2012	587	169.9
12	2013	1116	208.8
13	2014	-	244.5
14	2015	-	280.97
15	2016	-	325
16	2017	-	374.48
17	2018	-	422.63
18	2019	-	473.61

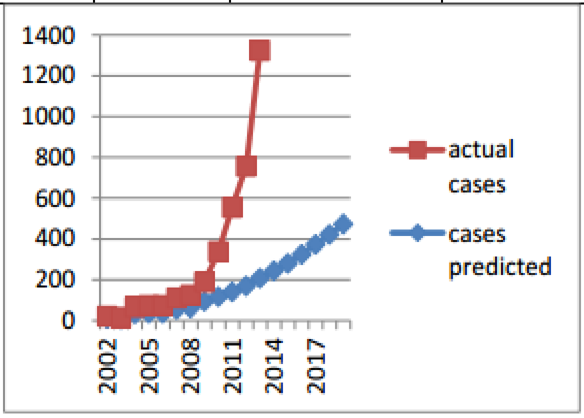

Figure 2: Comparison of actual cases registered vs. cases predicted

Table 4: predicted number of cases to be registered under IT Act in Delhi

We can see the predictions in Figure 4 which shows a dip in 2013 but actually there is a rise in number of cases registered under IT Act in Delhi. The dataset contains the number of cases registered under section 67 of IT Act in all over India from 2002 to 2013. The source of data is Crime in India(2002-2013), National Crime Records Bureau. We have used weka tool to predict the number of cases which will be registered till 2019.

VIII. CONCULSION AND FUTURE SCOPE

As mentioned at the beginning of this paper, our purpose was to review the current methods employed for predictive crime mapping, from basic approaches currently used by crime analysts, to sophisticated models developed by researchers. Our method was to assess each forecasting approach on the basis of accuracy, data requirements, hardware and software requirements, and ease of use. What we learned is that in many respects, this review is premature. The more sophisticated approaches described in this chapter are still very much in the development stages and can best be considered "alpha versions" that have yet to be tested by the end users.

There are several conclusions to be drawn from this review of crime forecasting methods. First, the more complicated methods are not always better predictors. More research is needed that evaluates the relative performance of methods. Second, many questions surrounding the choices made in sophisticated models must be empirically answered before the models will accurately and consistently perform (e.g., size of grid cell size and spatial lag). Third, additional research is needed to identify the input variables in the multivariate models. Choice of variables is critical to the success of the model and must be informed by theory. Finally, the connection between the output of models and how they translate into practice is extremely important. In fact, perhaps the most important measure of a crime forecasting technique may be whether it aids in crime control and prevention.

IX. References

[1] K.S.Arthisree, M.E, A.Jaganraj, Arulmigu Meenakshi, "Identify Crime Detection Using Data Mining Techniques" ,International Journal of Advanced Research in Computer Science and Software Engineering, Amman College Of Engineering, Thiruvannamalai district, Near Kanchipuram, India,volume 3,Issue 8,August 2013.

[2] B.Chandra and M. P.Gupta, "Crime Data Mining for Indian Police Information System Manish Gupta" ,Indian Institute of Technology Delhi, Hauz Khas, New Delhi - 110 016,India.

[3] D.E. Brown, "The regional crime analysis program (RECAP): A frame work for mining data to catch criminals" , In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, pp. 2848-2853, 1998.

[4] Ram Singla, Deepak Dembla, Yogesh Chaba and Kevika Singla CITM, "An Optimal KD Model for Crime Pattern Detection Based on Semantic Link Analysis",A Data Mining Tool C,Faridabad.

[5] Malathi. ,Dr. S. Santhosh Baboo Reader, "An Enhanced Algorithm to Predict a Future Crime using Data Mining" ,International Journal of Computer Applications (09758887),Volume 21 No.1, May 2011.

[6] TongWang1, Cynthia Rudin1, Daniel Wagner, and Rich Sevieri, "Learning to detect patterns of crime", JECET, 1:124131, 2012.

[7] Revatthy Krishnamurthy,J. Satheesh Kumar, "Survey Of Data Mining Techniques On Crime Data Analysis" , International Journal of Data Mining Techniques and Applications Vol 01, Issue 02, December 2012.

[8] Shyam Varan Nath Florida, "Crime Pattern Detection Using Data Mining", Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops) (WI-IATW06).

[9] Priyanka Gera, Rajan Vohra, "City Crime Profiling Using Cluster Analysis" ,International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014.

[10] Malathi. A and Dr. S. Santhosh Baboo, "Enhanced Algorithms to Identify Change in Crime Patterns", International Journal of Combinatorial Optimization Problems and Informatics, Vol. 2, No.3, Sep-Dec 2011, pp. 32- 38, ISSN: 200

