

Escalation of Crime Trouble Diminishment via Crime Data Mining Stimulation

Apexa L. Joshi*

*(R.K. University, Asst. Prof. JVIMS-MCA, GTU, Jamnagar, India)

ABSTRACT

There had been an enormous increase in the crime in the recent past. Crime prevention is a primary concern of police as they perform their central role of protecting the lives and property of citizens. However, information and technology overload hinders the effective analysis of criminal and terrorist activities. Crime deterrence has become an upheaval task. The cops in their role to catch criminals are required to remain convincingly ahead in the eternal race between law breakers and law enforcers. Police often have at their disposal vast amounts of least utilized crime data (such as crime incident reports) which if analyzed could reveal some hidden information such as crime committing trends useful in crime prevention. Use of Information Systems techniques such as data mining and Geographic Information Systems for analyzing these data is promising in boosting the police efforts. This paper reviews the applicability of various data mining methods and Geographic Information Systems in crime analysis and visualization in mainly poor planned settings characterized by missing electronic data common phenomena in the developing countries like India. The review tends to reveal that combinations of Geographic Information Systems and data mining techniques that can work under unclear data are best suited for use in the poorly planned settings. We also use semi-supervised learning technique here for knowledge discovery from the crime records and to help increase the predictive accuracy. The results of this data mining could potentially be used to lessen and even prevent crime for the forthcoming years.

Keywords - Crime Analysis, Crime-patterns, data mining, k-means, semi-supervised learning

I. INTRODUCTION

By analyzing the historical data the rate of crimes gets increased if we are considering the top ten countries for highest crime ratio the India holds 10th position in the list which actually bad indication. With the increasing use of the computerized systems to track crimes, computer data Analysts have started helping the law enforcement officers and detectives to speed up the process of solving crimes.

Here we will take an interdisciplinary approach between computer science and criminal justice to develop a data mining paradigm that can help solve crimes faster. We will discuss some terminology that is used in criminal justice and police departments and compare and contrast them relative to data mining systems. Suspect refers to the person that is believed to have committed the crime. The suspect may be identified or unidentified. The suspect is not a convict until proved at fault. The victim is the person who is the target of the crime. Most of the time the victim is identifiable and in most cases is the person reporting the crime.

Additionally, the crime may have some witnesses. There are other words commonly used such as homicides that refer to manslaughter or killing someone. Within homicides there may be categories like infanticide, eldercide, killing intimates and killing law enforcement officers. For the purposes of our modeling, we will not need to get into the depths of criminal justice but will confine ourselves to the main kinds of crimes. Cluster (of crime) has a special meaning and refers to a geographical group of crime, i.e. a lot of crimes in a given geographical region. Such clusters can be visually represented using a geo-spatial plot of the crime overlaid on the map of the police jurisdiction.

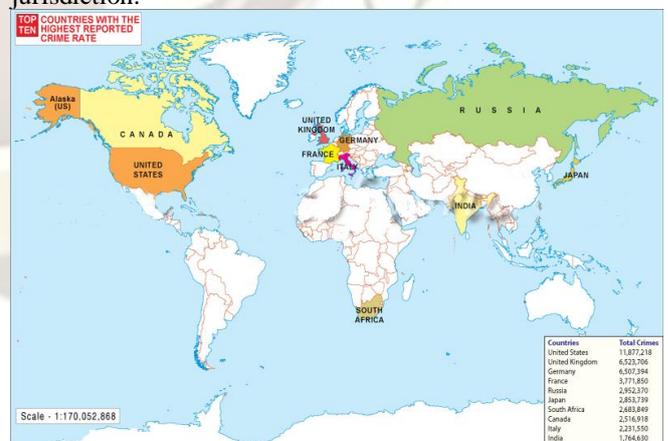


Figure 1: Top 10 countries with highest crime rate

The densely Populated group of crime is used to visually locate the 'hot-spots' of crime. However, when we talk of clustering from a data-mining standpoint, we refer to similar kinds of crime in the given geography of interest.

Such clusters are useful in identifying a crime pattern or a crime spree. Some well-known examples of crime patterns are the serial-rapist or a serial killer. These crimes may involve single suspect or may be committed by a group of suspects. The figure1 shows the top 10 countries with highest reported crime rates. So the aim of the paper is to help the 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.

II. CRIME REPORTING SYSTEM

The data for crime often presents an interesting dilemma. While some data is kept confidential, some becomes public information. Data about the prisoners can often be viewed in the county sites. However, data about crimes related to narcotics or juvenile cases is usually more restricted. Similarly, the information about the sex offenders is made public to warn others in the area, but the identity of the victim is often prevented. Thus as a data miner, the analyst has to deal with all these public versus private data issues so that data mining modeling process does not infringe on these legal boundaries.

Most office and police departments use electronic systems for crime reporting that have replaced the traditional paper-based crime reports. Here with this paper we pre-assumed that the entire departments are working on electronic systems. These crime reports have the following kinds of information categories namely - type of crime, date/time, location etc. Then there is information about the suspect (identified or unidentified), victim and the witness. Additionally, there is the narrative or description of the crime and Modus Operandi (MO) that is usually in the text form. The police officers or detectives use free text to record most of their observations that cannot be included in checkbox kind of pre-determined questions. While the first two categories of information are usually stored in the computer databases as numeric, character or date fields of table, the last one is often stored as free text.

The challenge in data mining crime data often comes from the free text field. While free text fields can give the newspaper columnist, a great story line, converting them into data mining attributes is not always an easy job. How to arrive at the significant attributes for the data mining models.

III. OVERVIEW OF CRIME

Crime is an offence against a person (for example murder, and robbery), or his/her property (for example, theft and property damage) or the State regulation (for example traffic violations) (Usul N. 2004) [1]. In non-legal terms crime is a set of acts that violate socially accepted rules of human ethical or moral behavior (Usul N. 2004) [1]; for example acting against a ritual in some society.

Crime occurs in a variety of forms which police informally categorizes as being either major or volume. Major crime consists of the high profile crimes such as murder, armed robbery and non-date rape. These crimes can either be one-offs or serial. In the case of serial crimes it is relatively easy to link crimes together due to clear similarities in terms of modus operandi or descriptions of offenders. This linking is possible due to the comparatively low volume of such crimes. Major crimes usually have a team of detectives allocated to conduct the investigation. In contrast volume crimes such as burglary and shoplifting are far more prevalent. They are usually serial in nature as offenders go on to commit many such crimes. Property crimes, such as domestic burglary offences, committed by different individuals are highly similar and it is rare to have a description of the offenders (R.William and Adderley Musgrove, 2001) [2]. Table 1 shows the classification of crime (Thomas M. Green, (2002) [3].

Table 1. Crime types at different levels. Source: (Thomas M. Green, 2002) [3].

Crime Type	Description
Traffic Violations	Driving under the influence of alcohol, fatal/personal injury/property damage traffic accident, road rage
Sex crime	Sexual offences
Fraud	Forgery and counterfeiting, frauds, embezzlement, identity deception
Arson	Arson on buildings
Gang / drug offences	Narcotic drug offences
Violent crime	Criminal Homicide, armed robbery, aggravated assault, other assaults
Cyber crime	Internet frauds, illegal trading, network intrusion /hacking, virus spreading, hate crimes, cyber piracy, cyber pornography, cyber-terrorism, theft of Confidential information.

Also a single detective officer may have a large number of different volume crimes to investigate at any point in time. With a view to satisfying this demand, police forces around the world employ specialist crime analysts, people who have specialist training in a variety of disciplines including investigation techniques, criminal psychology and information technology. It is their

task to assist investigating officers by analyzing crime trends and patterns, identifying links between crimes and producing packages which target an individual or group of offenders linking them to a series of crimes (R. William and Adderley Musgrove, 2001) [2].

Most, if not all, of current systems both manual and computerized revolve around the investigation of crimes already committed. They are, therefore, reactive. In the developed countries, a majority of crime prevention forces use different types of relational database management systems (RDBMS) for recording and subsequent analysis of crime. Standard or interactive queries are written to produce patterns of crime, offending and various statistics (R. William and Adderley Musgrove, 2001) [2]. but it's a common phenomena in the developing countries to find mainly manual criminal record books used alongside the pin-up maps for crime incidence location.

IV. DATA MINING AND CRIME PATTERN

Data mining deals with the discovery of unexpected patterns and new rules that are "hidden" in large databases. It serves as an automated tool that uses multiple advanced computational techniques, including artificial intelligence (the use of computers to perform logical functions), to fully explore and characterize large data sets involving one or more data sources, identifying significant, recognizable patterns, trends, and relationships not easily detected through traditional analytical techniques alone. This information then may help with various purposes, such as the prediction of future events or behaviors. (Reza et al, 2001)[3].

The development of new intelligent tools for automated data mining and knowledge discovery has led to the design and construction of successful systems that show early promise in their ability to scale up to the handling of voluminous data sets.

Theories of crime and delinquency tend to be discipline-specific and are dominated by psychological, sociological, and economic approaches (Reza et al, 2001)[3].

We will look at how to convert crime information into a data-mining problem (Hsinchun Chen, 2004) [4], such that it can help the detectives in solving crimes faster. We have seen that in crime terminology a cluster is a group of crimes in a geographical region or a hot spot of crime. Whereas, in data mining terminology a cluster is group of similar data points – a possible crime pattern. Thus appropriate clusters or a subset of the cluster will have a one-to-one correspondence to crime patterns.

Thus clustering algorithms in data mining are equivalent to the task of identifying groups of records that are similar between themselves but different from the rest of the data.

In our case some of these clusters will be useful for identifying a crime spree committed by one or same group of suspects. Given this information, the next challenge is to find the variables providing the best clustering. These clusters will then be presented to the detectives to drill down using their domain expertise. The automated detection of crime patterns, allows the detectives to focus on crime sprees first and solving one of these crimes results in solving the whole "spree" or in some cases if the groups of incidents are suspected to be one spree, the complete evidence can be built from the different bits of information from each of the crime incidents. For instance, one crime site reveals that suspect t has black hair, the next incident/witness reveals that suspect is middle aged and third one reveals there is tattoo on left arm, all together it will give a much more complete picture than any one of those alone. Without a suspected crime pattern, the detective is less likely to build the complete picture from bits of information from different crime incidents. Today most of it is manually done with the help of multiple spreadsheet reports that the detectives usually get from the computer data analysts and their own crime logs.

We choose to use clustering technique over any supervised technique such as classification, since crimes vary in nature widely and crime database often contains several unsolved crimes. Therefore, classification technique that will rely on the existing and known solved crimes, will not give good predictive quality for future crimes. Also nature of crimes change over time, such as Internet based cyber crimes or crimes using cell-phones were uncommon not too long ago. Thus, in order to be able to detect newer and unknown patterns in future, clustering techniques work better.

V. DATA VISUALIZATION

Visual methods are powerful tools in data exploration because they utilize the power of the human eye/brain to detect structures. A number of data mining tools for visualization exist, a histogram and Kernel plots being the most basic used for displaying single variables. Scatter plots for the display two variables at a time and reveal correlation, if any, between them. And for more than two variables, scatter plot matrices are often used (Smyth Padhraic, 2001) [5]. GIS also provides a powerful visualization tool through display of maps that allow the exploration of spatial patterns in an interactive fashion (Pfeiffer, 1996) [6].

VI. CLUSTERING TECHNIQUE

We will look at some of our contributions to this area of study. We will show a simple clustering example here. Let us take an oversimplified case of crime record. A crime data analyst or detective will use a report based on this data sorted in different orders,

usually the first sort will be on the most important characteristic based on the detective's experience.

We look at table 2 with a simple example of crime list. The type of crime is robbery and it will be the most important attribute. The rows 1 and 3 show a simple crime pattern where the suspect description matches and victim profile is also similar.

Table 2 Simple Crime Example

Crime Type	Suspect Race	Suspect Sex	Suspect Age gr	Victim age gr	Weapon
Robbery	B	M	Middle	Elderly	Knife
Robbery	W	M	Young	Middle	Bat
Robbery	B	M	?	Elderly	Knife
Robbery	B	F	Middle	Young	Piston

The aim here is that we can use data mining to detect much more complex patterns since in real life there are many attributes or factors for crime and often there is partial information available about the crime. In a general case it will not be easy for a computer data analyst or detective to identify these patterns by simple querying. Thus clustering technique using data mining comes in handy to deal with enormous amounts of data and dealing with noisy or missing data about the crime incidents.

We used k-means clustering technique here, as it is one of the most widely used data mining clustering technique. Next, the most important part was to prepare the data for this analysis. The real crime data was obtained from a Sheriff's office, under non-disclosure agreements from the crime reporting system. The operational data was converted into denormalised data using the extraction and transformation. Then, some checks were run to look at the quality of data such as missing data, outliers and multiple abbreviations for same word such as blank, unknown, or unk all meant the same for missing age of the person. If these are not coded as one value, clustering will create these as multiple groups for same logical value. The next task was to identify the significant attributes for the clustering. This process involved talking to domain experts such as the crime detectives, the crime data analysts and iteratively running the attribute importance algorithm to arrive at the set of attributes for the clustering the given crime types. We refer to this as the semi supervised or expert-based paradigm of problem solving. Based on the nature of crime the different attributes become important such as the age group of victim is important for homicide, for burglary the

same may not be as important since the burglar may not care about the age of the owner of the house.

To take care of the different attributes for different crimes types, we introduced the concept of weighing the attributes. This allows placing different weights on different attributes dynamically based on the crime types being clustered. This also allows us to weigh the categorical attributes unlike just the numerical attributes that can be easily scaled for weighting them. Using the integral weights, the categorical attributes can be replicated as redundant columns to increase the effective weight of that variable or feature. We have not seen the use of weights for clustering elsewhere in the literature review, as upon normalization all attributes assume equal importance in clustering algorithm. However, we have introduced this weighting technique here in light of our semi-supervised or expert based methodology. Based on our weighted clustering attributes, we cluster the dataset for crime patterns and then present the results to the detective or the domain expert along with the statistics of the important attributes.

The detective looks at the clusters, smallest clusters first and then gives the expert recommendations. This iterative process helps to determine the significant attributes and the weights for different crime types. Based on this information from the domain expert, namely the detective, future crime patterns can be detected. First the future or unsolved crimes can be clustered based on the significant attributes and the result is given to detectives for inspection. Since, this clustering exercise, groups hundreds of crimes into some small groups or related crimes, it makes the job of the detective much easier to locate the crime patterns.

The other approach is to use a small set of new crime data and score it against the existing clusters using tracers or known crime incidents injected into the new data set and then compare the new clusters relative to the tracers. This process of using tracers is analogous to use of radioactive tracers to locate something that is otherwise hard to find.



Pattern 1 (129 crimes)	Pattern 2 (79 crimes)	Pattern 3 (29 crimes)
Suspects point of entry	Suspects point of entry	Suspects point of entry
Victims Race	Victims Race	Suspects count (number)
Suspects count (number)	Number of days old	Victims Race
Number of days old		Number of days old

Pattern 5 (50 crimes)	Pattern 6 (9 crimes)	Pattern 7 (13 crimes)
Suspects race	Suspects city	Suspects Sex
Suspects Average height	Suspects point of entry	Suspects point of entry
Suspects Average Weight	Suspects Average Age	Suspects city
Suspects Average Age	Suspects count (number)	Suspects Average height
Suspects point of entry	Number of days old	Suspects Average Weight

Figure 2 Plot of crime clusters with legend for significant attributes for that crime pattern

VII. GEOGRAPHIC INFORMATION SYSTEMS

Over the past few years Geographic Information Systems (GIS) has become a standard tool for crime analysts in many police departments, regardless of their size (see for example McEwen and Taxman 1994; Rossmo, 1995). One of the inherent advantages of GIS is its ability to integrate information from a variety of sources into one user interface. In turn, this allows for spatial analyses that would



Figure 3 Geographical Information systems

either not have been possible or at a minimum far more difficult prior to the advent of GIS (Olligschlaeger, 1997)[7]. .

VIII. CRIME ANALYSIS

Crime being inherently spatial phenomena (Ratcliffe, 2004)[8], taking place within a given location and at a specific time, it is logical to analyse crime in terms of spatial crime analysis. The advent of geographic information systems (GIS) has given behavioural scientists the ability to map and track a wide range of social phenomena such as crime incidences. Much more information about what is happening “on the ground” is now available. This enhanced analytic capacity has presented police departments with the opportunity to better serve and protect the people and places in their care. Furthermore, as GIS assists one in seeing and understanding behaviour patterns, it also provides stakeholders with opportunities to join together in partnerships for the common good (Holzman et al, 2003)[8]. A data pattern is an expression in some language describing a subset of the data or a model applicable to that subset (Fayyad

et al, 1996)[6] There are at least three types of police crime information (primary, secondary, and tertiary), intelligence (prospective, retrospective, and applied), and operational strategies (preventive, prospective, and reactive), each of which interacts in a complex fashion with technology (Manning, 1992) [9]. Crime data analysis aids police in transforming this information from one type to another.

IX. CRIME SPATIAL DATA ANALYSIS

Spatial data analysis utilizes statistical analysis methods that address specific issues relating to spatial data, including spatial dependence (autocorrelation) and spatial heterogeneity. These issues run counter to traditional statistical assumptions of heterogeneity and independence of sample data. If these issues are ignored, then analysis results might not be valid. In combining the powerful tools of GIS to integrate and manipulate spatial data with rigorous statistical methods, spatial data analysis shows great promise for criminology, criminal justice, and law enforcement research and practice. (NIJ, 2006)[10].

X. POINT PATTERN

Spatial point patterns (SPP) are based on coordinates of events such as locations of crime incidences and may also include the time of occurrence. All or a sample of point pattern may be plotted on the map. The aim of SPP analysis is to detect whether the point pattern is distributed at random, clustered or regular. SPP is typically interpreted as analysis of clustering. A dot map is commonly used to represent SPP. The tool effectively used for analysis of clustering effects is the K function. This method assesses clustering of crime incidences in detection of hot spots (Kingham et al, 1995) where time and space relationship analysis is required, the methods used are Knox’s method, Mantel’s Method and K-nearest neighbour method. All the three methods require the production of distance matrices of the spatial as well as temporal relationship between crime incidences. Knox’s method requires critical distance in time as well as space defining closeness has to be set but the determination of these critical distances requires subjective decision. Mantel approach does not however require use of critical distances but uses both time and space matrices. It is however insensitive to non-linear associations. The K-nearest neighbor is based on the approximate randomization of the Mantel product statistic (Pfeiffer, 1996) [6].

XI. CONCLUSION AND FUTURE WORK

Exploratory data analysis makes few assumptions about data and it is robust to extreme data values. It is possible to use simple analytical models. The methods that are robust to missing data are useful in the data mining of crime data where data is not so precisely collected. The distances between crime locations are normally not easily available to

the police in the areas that are not well planned. The poorly planned areas are best represented by dividing them into area clusters and the analysis is done based on the clusters. The methods that support clustering are therefore best suited for the crime analysis of the poorly planned settings. The manual pin maps are best replaced by the use of GIS.

We looked at the use of data mining for identifying crime patterns crime pattern using the clustering techniques. Our contribution here was to formulate crime pattern detection as machine learning task and to thereby use data mining to support police detectives in solving crimes. We identified the significant attributes; using expert based semi-supervised learning method and developed the scheme for weighting the significant attributes. Our modeling technique was able to identify the crime patterns from a large number of crimes making the job for crime detectives easier.

Some of the limitations of our study include that crime pattern analysis can only help the detective, not replace them. Also data mining is sensitive to quality of input data that may be inaccurate, have missing information, be data entry error prone etc. Also mapping real data to data mining attributes is not always an easy task and often requires skilled data miner and crime data analyst with good domain knowledge. They need to work closely with a detective in the initial phases.



Figure 4: First step toward justice by using intelligent mining tools

As a future extension of this study we will create models for predicting the crime hot-spots [11] that will help in the deployment of police at most likely places of crime for any given window of time, to allow most effective utilization of police resources. We also plan to look into developing social link networks to link criminals, suspects, gangs and study their interrelationships. Additionally the ability to search suspect description in regional, FBI databases [12], to traffic violation databases from different states etc. to aid the crime pattern detection or more specifically counter terrorism measures will also add value to this crime detection paradigm.

So the aim of the paper is to use the advance technology find the crime and criminals from the world and provide the crime free world.

REFERENCES

- [1] Akpınar E. and Usul N. (2004). "Geographic Information Systems Technologies in Crime Analysis and Crime Mapping".
- [2] Adderley R. William and Musgrove Peter, (2001), "Police crime recording and investigation systems: A user's view", An International Journal of Police Strategies and Management, Vol. 24 No. 1, pp. 100-114.
- [3] Reza Fadaei-Tehrani, Thomas M. Green, (2002) "Crime and society" International Journal of Social Economics Volume 29 Number 10 pp. 781-795
- [4] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer Jie Xu, Gang Wang, Rong Zheng, Homa Atabakhsh, "Crime Data Mining: A General Framework and Some Examples", IEEE Computer Society April 2004.
- [5] Hand David, Mannila Heikki, Smyth Padhraic., (2001), "Principles of Data Mining" Prentice Hall.
- [6] Pfeiffer, D. U. (1996), "Issues related to handling of spatial data". Massey University, Palmerston North, New Zealand.
- [7] Olligschlaeger Andreas M (1997), "Weighted Spatial Adaptive Filtering and Chaotic Cellular Forecasting with Applications to Street Level Drug Markets" PhD dissertation on Spatial Analysis of Crime Using GIS-Based Data., Carnegie Mellon University.
- [8] Ratcliffe Jerry H. (2004), "Geocoding crime and first estimate of the minimum acceptable hit rate". International Journal of Geographical Information Science, Vol. 18, No. 1, pp 61-72
- [9] Manning Peter K., (1992), "Information Technologies and the Police" Crime and Justice, Vol. 15, Modern Policing, pp. 349-398
- [10] National Institute of Justice "Spatial Data Analysis", (January 2006), USA <http://www.ojp.usdoj.gov/nij/maps/research.html> - Accessed February 2006
- [11] C McCue, "Using Data Mining to Predict and Prevent Violent Crimes", available at: <http://www.spss.com/dirvideo/richmond.htm?source=dmpage&zone=rtsidebar>
- [12] Whitepaper, "Oracle's Integration Hub For Justice And Public Safety", Oracle Corp. 2004, available at: http://www.oracle.com/industries/government/IntegrationHub_Justice.pdf
- [13] For the geo spatial information available at: http://www.geospatialworld.net/index.php?option=com_content&view=article&id=20403&Itemid=1391

- [14] Details related to the crime rates <http://www.mapsofworld.com/world-top-ten/countries-with-highest-reported-crime-rates.html>
- [15] Data Mining concepts and Techniques by Jiawei Han, Micheline Kamber– Elsevier.
- [16] Adderley, R. & P.B. Musgrove, P.B. 2001. Data Mining Case Study: Modeling the Behavior of Offenders Who Commit Serious Sexual Assaults, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining.
- [17] Shyam Varan Nath, Crime Pattern Detection Using Data Mining, Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, 41- 44, 2006
- [18] Chainey, S.P., S. Reid, and N. Stuart. 2002. “When Is a Hotspot a Hotspot? A Procedure for Creating Statistically Robust Hotspot Maps of Crime.” In Innovations in GIS 9. London: Taylor & Francis.
- [19] Ratcliffe, J. 1999a. “Spatial Pattern Analysis Machine Version 1.2 Users Guide.” <http://jratcliffe.net/ware/index.htm>
- [20] Ratcliffe, J., and M.J. McCullagh. 1998. “Hotbeds of Crime and the Search for Spatial Accuracy.” Paper presented to the Second Crime Mapping Research Center Conference: Mapping Out Crime, Arlington, Virginia, December 10–12, 1998.
- [21] Harries, K. 1999. Mapping Crime: Principle and Practice. Washington, DC: U.S. Department of Justice, National Institute of Justice.
- [22] Anselin, Luc, Ibnu Syabri, and Youngihn Kho. (forthcoming). “GeoDa: An Introduction to Spatial Data Analysis.