Effective Data Mining Approach For Crime-Terrorpattern Detection Using Clustering Algorithm Technique

Neha Gohar Khan
P.R.Patil College of Engg& Technology,
Amravati.

Prof.V.B.Bhagat(Mate)
P.R.Patil College of Engg& Technology,
Amravati.

Abstract

Crimes are a social nuisance and cost our society dearly in several ways hindering the peace within the nation. About 10% ofthe criminals commit about 50% of the crimes.The concern about national security has increased significantly since the 9/11 attacks.Any research that can help in solving crimes faster will pay for itself.A major challenge facing all law-enforcement and intelligence-gathering organizations is accurately and efficiently analyzing the growing volumes of crime data.Althoughvarious means have be adopted to help the law enforcement agencies to identify terrorist and to counter-terrorism. One of such measure is the use of computer technology and computer analysis for effective analysis of criminal activities. Data mining applied in the context of law enforcement and intelligence analysis holds the promise ofalleviating such problemsby applying the various data mining techniques.Data mining is a powerful tool that enables criminalinvestigatorswho may lack extensive training as data analysts to explore large databases quickly and efficiently. In this paper, we analyzed how data mining techniques can be adopted by law enforcement agencies in tracking the activities of criminals; also the paper examines crime data mining techniques and presents four case studies of the COPLINK project. Finally, we showed the use of clustering algorithm for a data mining approach to help detect the crime patterns and speed up the process of solving crimes.

1. Introduction

Historically solving crimes has been the prerogative of the criminal justice and law enforcement specialists. Local law enforcement agencies 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. 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 study an effective data mining approach between computer science and criminal justice that can help solve crimes faster. More specifically, we will use clustering based models to help in identification of crime patterns.

Firstly, 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 theperson that is believed to have committed the crime whomay be identified or unidentified and is not a convict until proved guilty. 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.Homicides refers to manslaughter or killing someone. Within homicides there may be categories like infanticide, eldercide, killing intimates and killing law enforcementofficers. 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 overlayedon the map of the police jurisdiction. 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 DCSniper, a serial-rapist or a serial killer. These crimes may involve single suspect or may be committed by a group of suspects.The below figure shows the plot of geo-spatial clusters of crime.

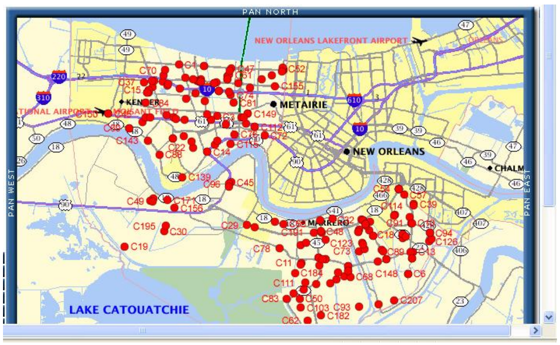

Fig 1 Geo-spatial plot of crimes, each red dot represents a crime incident.

2. Data Mining Process

Data mining is a promising tool in the fight against terrorism and crime. It plays a number of important roles in counter terrorism including locating known suspects, identifying and tracking suspicious financial and other transactions, and facilitating background checks. Often used as a means for detecting fraud, assessing risk, and product retailing, data mining involves the use of data analysis tools to discover previously unknown, valid patterns and relationships in large datasets.

The process of data mining is simply the collection of data into a single repository where data mining algorithms are applied for knowledge discovery and pattern recognition. A data warehouse provides the base for the powerful data analysis techniques that are available today such as data mining and multidimensional analysis, as well as the more traditional query and reporting. It also helps to set the stage forKnowledge Discovery Database (KDD). The most data-mining algorithms such as statistics, pattern recognition, and machine learning assume that data are in the main memory. However most of the data mining can exist with a data warehouse. Figure 1 shows the simple process of data mining, data from various sources are gathered together in a repository commonly known as data warehouse before data mining techniques are applied for pattern evaluation, recognition and analysis.

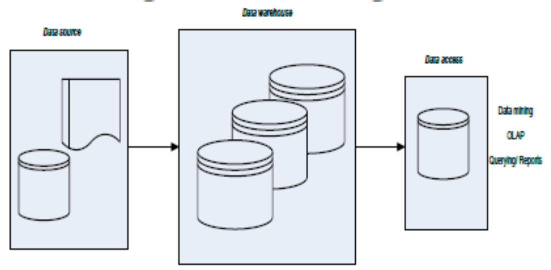

Figure1. The concept of data mining and data warehousing

In view of the above, we have to x-ray the definition of data mining and Knowledge discovery database according to Fayyad et al. (1996), Fayaad et al. defined data mining as a process in the knowledge discovery database (KDD) which is a nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Their views are diagrammatized in Figure 2 and show data mining as a continuous process; from a large dataset, valid data are selected, processed and transformed into a more useful dataset before data mining techniques are applied for valid patterns. Dissecting further the definition and concept of data mining according to Fayyad et al. (1996), the following are evident:

Datasets: Data are set of facts (database) and pattern describes a subset of the dataset.

Model: Designates extracting and fitting a model to the data

Process: The fact that KDD and data mining comprise many processes.

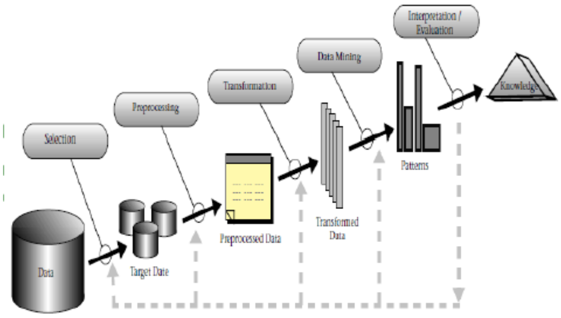

Figure 2. The process of data mining in the Knowledge Discovery Data- Base

3. Crime Data Mining

It is useful to review crime data mining in two dimensions: (1) crime types and security concerns and (2)crime data mining approaches and techniques.

3.1. Crime Types 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". 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 and 26/11 attacks).

3.2. Crime Data Mining Approaches and Techniques

Data mining is defined as the identification of interesting structure in data, where structure designates patterns, statistical or predictive models of the data, and relationships among parts of the data. Data mining in the context of crime and intelligence analysis for national security is still a young field. The following describes the applications of different techniques in crime data mining. *Entity extraction* has been used to automatically identify person, address, vehicle, narcotic drug, and personal properties from police narrative reports *Clustering techniques* such as "concept space" have been used to automatically associate different objects (such as persons, organizations, vehicles) in crime records.*Deviationdetection* has been applied in fraud detection, network intrusion detection, and other crime analyses that involve tracing abnormal activities. *Classification* has been used to detect email spamming and find authors who send out unsolicited emails. *String comparator* has been used to detect deceptive informationincriminal records.*Social network analysis* has been used to analyze criminals' rolesand associations among entities in a criminal network.

3.3. Case Studies of Crime Data Mining

Based on the crime characteristics and analysis techniques discussed above, we present four case studies of crime data mining used in the Coplink project.

3.3.1. Named-Entity Extraction

Most criminal justice databases capture only structured data that fits in predefined fields. The first data mining task involved extracting named entities from police narrative reports, which are difficult to analyze using automated investigators in crime analyses.It proposed a neural network-basedentity extractor, which applies named-entity extraction techniques to automatically identify useful entitiesfrom police narrative reports. The system has three majorcomponents: (1) *Noun phrasing*: It is a modified version of the Arizona Noun Phraserand extracts noun phrases as named entities from documents based on syntactical analysis; (2)*Finite state machine and lexical lookup*: A finite state machine compares each word in the extractedphrase, as well as the words immediately before and after the phrase, with the items in a handcraftedlexicons. Each comparison will generate a binary value (either 0 or 1) to indicate a match or mismatch;

(3) *Neural network*: The feedforward/backpropagation neural network predicts the mostlikelyentity type for each phrase.

3.3.2.Deceptive Identity Detection:An Algorithmic Approach

Criminals often provide police officers with deceptive identities to mislead police investigations, for example, using aliases, fabricated birth dates or addresses, etc. The large amount of data also prevents officers from examining inexact matches manually. The second data mining task involves automatically detecting deceptive criminal identities from the police departments databases, which contains information such as name, gender, address, ID number, and physical description.It was found that criminals usually made minor changes to their real identity information.Based on the taxonomy,an algorithmic approach was developed to detect deceptive criminal identities automatically.This approach utilized four identity fields: name, address, date-of-birth, and socialsecurity- number and compared each corresponding field for a pair of criminal identity records. The method employees string comparators to compare values in the corresponding fields of each record pair. Comparators measure the similarity between two strings. An overall disagreement value between the two records was computed by calculating the Euclidean Distance ofdisagreement measures over all attribute fields.A Euclidean vector norm is the square root of the sum of squared similarity measures and is also normalized between 0 and 1.The algorithm could accurately detect 94% of criminal identity deceptions.

3.3.3 Criminal Network Analysis

Criminals often develop networks in which they form groups or teams to carry out various illegal organized crimes such as narcotics trafficking, terrorism, gang-related crimes, and frauds.The fourth data mining task consists of identifying subgroups and key members in such networks and then studying interaction patterns to develop effective strategies for disrupting the networks.Social Network Analysis (SNA) has been recognized as an appropriate methodology to uncover previously unknown structural patterns from criminal networks. Four steps are involved in this task: (1)*Network extraction*: It utilizes crime incident reports as sources for criminal relationship information because criminals who committed crimes together usually were related. The concept space approach is used to identify and uncover criminal relationships; (2)*Subgroupdetection*:Itemployes hierarchical clustering to detect subgroups in a criminal network based on relational strength; (3) *Interaction pattern discovery*: It employes an SNA approach called blockmodeling to reveal patterns of between-group interaction. Given a partitioned network, blockmodel analysis determines the presence or absence of an interaction between a pair of subgroups by comparing the density of the links between these two subgroups to apredefined threshold value; (4) *Central member identification*: It employes several measures, such as degree, betweenness, and closeness to identify central members in a given subgroup. These three measures can suggest the centrality of a network member.

Figure.1 shows a narcotics network consisting of 60 criminals. It is difficult to detect subgroups, interaction patterns, and the overall structure from this original network manually. Using clustering and blockmodeling methods, however, a chain structure became apparent (Figure 1b). This system isvery useful for crime investigation and could greatly increase crime analysts' productivity.

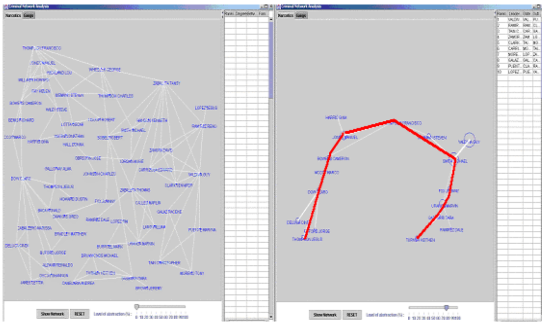

Figure 1(a). A 60-member narcotics network. (b) The chain structure (thicker links) found.

3.3.4 Authorship Analysis in Cybercrime

The large amount of cyber space activities and their anonymous nature make cybercrime investigation extremely difficult. Conventional ways to deal with this problem rely on a manual effort, which is largely limited by the sheer amount of messages and constantly changing author IDs. The next data mining task proposes authorship analysis framework to automatically trace identities of cyber criminals through messages they post on the Internet. Under this framework, three types of message features, including style markers, structural features, and content-specific features, are extracted and inductive learning algorithms are used to build feature-based models to identify authorship of illegal messages. The experimental results indicated a promising future of using our framework to address the identity-tracing problem.

4.1 Clustering

Clustering has been with us over time. People naturally cluster together based on some certain qualities, attributes and characteristics. People from the same country, religion, tribe, race etc.cluster together. According to Weiss et al. (2010), the main reason for data clustering is that it allows us to build simpler more understandable models of the world, which can be acted upon more easily. The goal of clustering is to group similar object into one cluster and dissimilar object into another cluster based on characteristics of data, also known as segregation. Clustering can easily be used to automatically segregate individual into a defined group.

4.2 Clustering Techniques Used

Clustering is another data mining technique that can be used to detect crime and terrorism. Clustering techniques and algorithm are based on real-life model that individual with certain qualities must cluster together. Clustering algorithms will then be to identify given clusters and their areas of operation, anytime a crime is reported, law enforcement agencies can look for related clusters, and then examined them for clues. Let's study a simple clustering example by taking an oversimplified case of crime record. Acrime data analyst or detective will use a report based onthis data sorted in different orders, usually the first sortwill be on the most important characteristic based on thedetective's experience.

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

Table 1 Simple Crime Example

We look at table 1 with a simple example of crime list.The type of crime is robbery and it will be the mostimportant attribute. The rows 1 and 3 show a simple crimepattern where the suspect description matches and victimprofile is also similar. The aim here is that we can use datamining to detect much more complex patterns since in reallife there are many attributes or factors for crime and often there is partial information available about the crime. In ageneral case it will not be easy for a computer data analystor detective to identify these patterns by simple querying.Thusclustering technique using data mining comes inhandy to deal with enormous amounts of data and dealingwith noisy or missing data about the crime incidents .The technique used here is k-means clustering as it is oneof the most widely used data mining clustering technique.Next, the most important part is to prepare the data for the analysis. The operational data isconverted into denormalised data using the extraction andtransformation. Then, some checks are run to look at thequality of data such as missing data, outliers and multipleabbreviations for same word such as blank, unknown, orunk all meant the same for missing age of the person. Ifthese are not coded as one value, clustering will createthese as multiple groups for same logical value. The next task is to identify the significant attributes for theclustering. This process involves talking to domainexperts such as the crime detectives, the crime dataanalysts and iteratively running the attribute importancealgorithm to arrive at the set of attributes for the clustering of the given crime types. This is referred as the semi-supervisedor expert-based paradigm of problem solving.

Based on the nature of crime the different attributesbecome important such as the age group of victim isimportant for homicide, for burglary the same may not beas important since the burglar may not care about the ageof the owner of the house.

To take care of the different attributes for different crimes types, the concept of weighing the attributes was introducedThis allows placing different weights ondifferent attributes dynamically based on the crime typesbeing clustered. This also allows to weigh thecategorical attributes unlike just the numerical attributesthat can be easily scaled for weighting them. Using theintegral weights, the categorical attributes can bereplicated as redundant columns to increase the effectiveweight of that variable or feature. Based onthe weighted clustering attributes,the dataset is clustered for crime patterns and then the resultis presented to thedetective or the domain expert along with the statistics ofthe important attributes.The detective then looks at the clusters, smallest clusters firstand then gives the expert recommendations. This iterativeprocess helps to determine the significant attributes andthe weights for different crime types. Based on thisinformation from the domain expert, namely the detective, future crime patterns can be detected. First the future orunsolved crimes can be clustered based on the significantattributes and the result is given to detectivesforinspection. Since, this clustering exercise, groupshundreds of crimes into some small groups or relatedcrimes, it makes the job of the detective much easier tolocate the crime patterns.

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

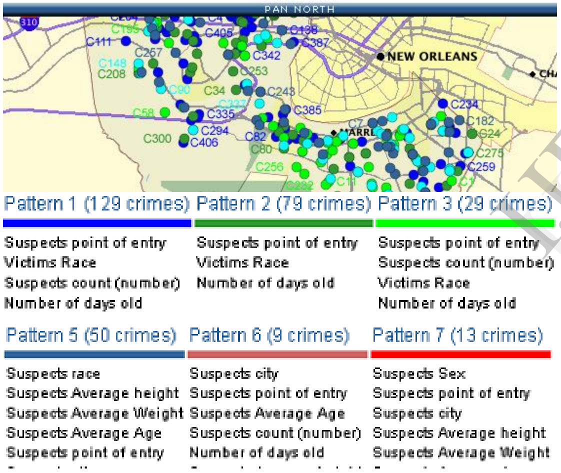

Figure 2 Plot of crime clusters with legend forsignificant attributes for that crime pattern.

4.3. Results of Clustering Crime Pattern Analysis

The above system is used along with the geo spatial plot. The crime analyst may choose a time range and one or more types of crime from certain geography and display the result graphically. From this set, the user mayselect either the entire set or a region of interest. The resulting set of data becomes the input source for the data mining processing. These records are clustered based on the predetermined attributes and the weights and have the possible crime patterns which are plotted on the geo-spatial plot. The results are shown in the figure below. The different clusters or the crime patterns are color-coded. For each group, the legend provides the total number of crimes incidents included in the group along with the significant attributes that characterize the group. This information is useful for the detectives to look at when inspecting the predicted crime clusters. The above results were validated for the detected crime patterns by looking the court dispositions on these crime incidents as to whether the charges on the suspects were accepted or rejected. So to recap the starting point is the crime incident data (some of these crimes already had the court dispositions/ rulings available in the system), which weremeasured in terms of the significant attributes or features or crime variables such as the demographics of the crime, the suspect, the victim etc. No information related to the court ruling was used in the clustering process. In this case, we looked at the crime patterns, as shown in same colors above and looked at the court dispositions to verify that some of the data mining clusters or patterns were indeed crime spree by the same culprit(s).

5. Conclusions

In this paper, we looked at the use of data mining for identifying crime patterns and presented four Coplink case studies. We also presented an overview of clustering techniques for a data mining approach which was able to identify the crime patterns from a large number of crimes making the job for crime detectives easier. Fromthe encouraging results, we conclude that crime data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis. Many future directions can be explored in this still young field. For example, more visual and intuitive criminal and intelligence investigation techniques can be developed for crime pattern and network visualization.

However, Crime pattern analysis and detection can only help the law enforcement agencies and are not intended to replace them. Also, data mining techniques are not going to crop up and say that the bad guy is this or that rather it will help the detectives and law enforcement agencies in crime fighting.

Data mining in not all to counter-terrorism as there are various drawbacks which included the issue of skilled manpower, inadequate investment in telecommunication and IT infrastructure, inadequate data mining policies and above all legal issues that characterize unwanted tracking of innocent citizens.

Data mining techniques and clustering algorithms can be applied in most cases to solve the issue of crime by identifying the activities of these criminals and tracking them down. A well implemented data mining algorithm will definitively go a long way in this quest to minimize crimes.

6. References

[1] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer JieXu, Gang Wang, RongZheng, HomaAtabakhsh, “Crime Data Mining: An Overview and Case Studies”, AI Lab, University of Arizona, proceedings National Conference on Digital Government Research,2003, available at:http://ai.bpa.arizona.edu/

[2] Hsinchun Chen, Wingyan Chung, Yi Qin, Michael Chau, Jennifer JieXu, Gang Wang, RongZheng, HomaAtabakhsh, “Crime Data Mining: A General Framework and Some Examples”, IEEE Computer Society April 2004.

[3]Hauck, R.V., Atabakhsh, H., Ongvasith, P., Gupta, H., & Chen, H. (2002).Using Coplink to analyze criminal-justice data. IEEE Computer, 35(3), 30-37.

[4] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.

[5] Alex Berson, Stephen Smith and Kurt Thearling (2010). “An Overview of Data Mining Techniques” retrieved from the web 14-12-2010

[6]DeRosa Mary (2004). “Data mining and Data Analysis for counter terrorism”, CSIS report-2004

[7]Carlile of Berriew Q.C. (2007). “Data mining: The new weapon in the war on terrorism” retrived from the Internet on 28-02-2011

[8]S. V. Nath, “Crime Pattern Detection Using Data Mining Florida Atlantic University / Oracle Corporation,”
vol. 1, no. 954, pp. 1–4, 2006

[9]AravindanMahendiran, Michael Shuffett, SathappanMuthiah, Rimy Malla, GaoqiangZhang,”Forecasting Crime Incidents using Cluster
Analysis and Bayesian Belief Networks”

[10] J. Jie and M. Chau, “Crime Data Mining : A General Framework,” IEEE Computer Society, no. 4, 2004.