# Data Mining – A necessity for Crime Detection

Prof.Neha Mishra
Master In Computer Application
Tulsiramji Gaikwad Patil College of Engineering and Technology,TGPCET
Nagpur, India
*neha.s0511@gmail.com*

Prof.Pooja Shelke
Masters In Computer Application
Tulsiramji Gaikwad Patil College of engineering and Technology,TGPCET
Nagpur,India
*poojashelkeifsc@gmail.com*

***Abstract*—**Data mining is the process of finding out interesting knowledge discovery from large amount of heterogeneous data. It involves various methods such as machine learning, summarization, artificial intelligence, neural networks etc. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing model, inference considerations, complexity considerations, post processing of discovered structures, visualization and online updating. As the internet is widely used in all areas, the proximity of attackers being able to attack has also increased. In this paper, we present recent research on types of threats that can occur on internet aiming at hampering the critical information residing, description of data mining and crime detection and their inter relationship, classification techniques and approach of data mining in various areas of crime detection.

***Keywords*—***Intrusion, threats, summarization, phishing, middleware architectures.*

*****

## I. INTRODUCTION

Data Mining can be defined as extracting or mining knowledge from large amount of data. It is considered as one of the steps in knowledge discovery process where valuable information is extracted from large database. It can also be defined as a process of extracting valid, previously unknown, non-trivial and useful information from large database. Data mining techniques are broadly classified into five categories: Dependency analysis, class identification, concept description, deviation detection and data visualization summarized below.

a. Dependency Analysis: It can be defined as a process of determining associations between entities and finding relationships among them like market basket analysis.
b. Class Identification: It can be defined as grouping various entities into classes. It includes two types of identification tasks –mathematical taxonomy and concept clustering.
c. Concept Description: In concept description groups are made on the basis of the domain knowledge and databases, without any compulsory descriptions. It includes tasks like data summarization and data comparison.
d. Deviation Detection: It includes tasks like finding changes in data and anomaly detection that is finding actions that are different from the benign actions.
e. Data Visualization: It includes finding and analyzing different patterns which are complex in nature.

In this paper, we discuss data mining approach to provide cyber security by crime detections. . In particular, we will discuss threats to computers and networks and describe applications of data mining to detect such threats and attacks.

## II. DATA MINING FOR CYBER SECURITY

### *A. Overview*

This section discuss about various threats that can be found as security violations through access controls and other means.
In next sub sections we discuss about types of attacks and there occurrences. With an increasing reliance on mobile devices it is important to be aware of new and emerging software threats that target them specifically. Mobile viruses, for example, can infect one cellular phone and then spread to other devices via the mobile phone network.

### *B. Software Threats*

Software threats are malicious pieces of computer code and applications that can damage your computer, as well as steal your personal or financial information. For this reason, these dangerous programs are often called malware (short for "malicious software.")

#### a.Mobile Viruses

Mobile devices can be infected by viruses that spread themselves via the mobile phone network. These have been a limited threat to date due to the fact that mobile phones use many different operating systems, but as a small number of systems (such as *Android* and*iOS*) become dominant, these viruses will be able to spread more widely. In all other respects these are identical to other computer viruses.

#### b.Bluejacking

Bluejacking uses a feature originally intended to exchange contact information to send anonymous, unwanted messages to other users with *Bluetooth*-enabled mobile phones or

laptops. In some cases this is used to send obscene or threatening messages or images, and it could be used to spread malware as well.

**c.Bluesnarfing**

Bluesnarfing is the actual theft of data from *Bluetooth* enabled devices (especially phones). Like bluejacking it depends on a connection to a *Bluetooth* phone being available. A *Bluetooth* user running the right software from a laptop can discover a nearby phone and steal the contact list, phonebook and images etc. Furthermore, your phone's serial number can be downloaded and used to close the phone. Again, the only current defense is to turn your Bluetooth off by setting it to "undiscoverable"

**d.Cyber-terrorism, Insider Threats, and External Attacks**

Cyber-terrorism is one of the major threats and it is of major concern. This threat is exacerbated by the large amount of information which is easily available on Internet. Attacks on our computers, networks, databases and the Internet infra-structure could be devastating to businesses. It is estimated that cyber-terrorism could cause billions of dollars to businesses. A classic example is that of a banking information system. If terrorists attack such a system and deplete accounts of funds, then the bank could lose millions and perhaps billions of dollars. By crippling the computer system millions of hours of productivity could be lost, which is ultimately equivalent to direct monetary loss. Even a simple power outage at work through some accident could cause several hours of productivity loss and as a result a major financial loss. Therefore it is critical that our information systems be secure. Threats can occur from outside or from the inside of an organization. Outside attacks are attacks on computers from someone outside the organization. We hear of hackers breaking into computer systems and causing havoc within an organization. Some hackers spread viruses that damage files in various computer systems. But a more sinister problem is that of the insider threat. Insider threats are relatively well understood in the context of non-information related attacks, but information related insider threats are often overlooked or underestimated. People inside an organization who have studied the business' practices and procedures have an enormous advantage when developing schemes to cripple the organization's information assets. These people could be regular employees or even those working at computer centers. The problem is quite serious as someone may be masquerading as someone else and causing all kinds of damage. In the next few sections we will examine how data mining can serve as a productive tool to control all such attacks.

## III. DATA MINING TECHNIQUES

Data mining provides various techniques which are very helpful in finding out abnormal activities in day to day operations. Entity extraction has been used to automatically identify person, address, vehicle or personal properties from criminal record book. Clustering techniques have been used to automatically associate different objects (such as persons, organizations, vehicles) in crime record. Deviation detection has been applied in fraud detection, network intrusion detection, and other crime analysis that involve tracing abnormal activities. Classification techniques are widely used to detect email spamming and an authentication of sender. String comparator has been used to detect deceptive information in criminal records.

## IV. CRIME DETECTION

Government and Intelligence agencies continuously investigate and monitor large amount of data so that they can keep track on all illegal activities that are carried out. Local law enforcement agencies have also become more alert to criminal activities in their own jurisdictions. It is possible to control these crimes if we can identify and control them at local level. When the local criminals are identified accurately and restricted from their crimes, then it is possible to reduce the crime rate.

Illegal activities such as crime are generally carried out in a network. Therefore, data mining can be helpful as it can identify the sub-groups in a network and the key persons responsible for handling such group and then studying interaction patterns to develop effective strategies for disrupting the networks. Data is used with a concept to extract criminal relations from the incident summaries and create a likely network of suspects. Co-occurrence weight measures the relational strength between two criminals by computing how frequently they were identified in the same incident.

## V. DATA MINING IN CRIME DETECTION

Data mining offers various techniques and algorithms to analyze and scrutinize the data. However, depending on the situation, the technique to be used solely depends upon the circumstance. Also one or more data mining techniques could be used if one is inadequate. Data mining applications also uses a variety of parameters to examine the initiation point causing such attack and the individuals who might be responsible for such attack. We have stated that crime investigation remains the prerogative of the law enforcement Agencies concern, but computer and computer analysis can be useful in detection.

## VI. CLASSIFICATION IN DATA MINING

Classification in a broad sense is a data mining technique that produces the characteristics to which a population is divided

based on the characteristic. The idea is to define the criteria use for the segmentation of a population. Once this is done, individuals and events can then fall into one or more groups naturally. Classification divides the population (dataset) based on some predefined condition. When classification is used, existing data set can easily be understood and it will undoubtedly help to predict how new individual or events will behave based on the classification criteria. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long-distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier.

## VII. DISCOVERING DISCRIMINATION

Discrimination discovery is about finding out discriminatory decisions hidden in a dataset of historical decision records. The basic problem in the analysis of discrimination, given a dataset of historical decision records, is to quantify the degree of discrimination suffered by a given group (e.g. an ethnic group) in a given context with respect to the classification decision (e.g. intruder yes or no). Discrimination techniques should be used in training data are based towards a certain group of users (e.g. young people). The learned model will show discriminatory behavior towards that group and most requests from young people will be incorrectly classified as Intruders. Additionally, anti-discrimination techniques could also be useful in the context of data sharing between IDS (intrusion detection systems. Assume that various IDS share their IDS reports that contain intruder information in order to improve their respective intruder detection models.

## VIII. LINK ANALYSIS

Link Analysis (LA) is another data mining technique that is useful in detecting valid and useful patterns. The theoretical framework of Link Analysis (LA) is based on the fact that events are linked to one another and are hence mutually exclusive. Link Analysis framework is that if A is linked to B and B is linked to C and C to D then A could be linked to D. link analysis can be employed by enforcement investigators and intelligence analysts to connect network of relationships and contacts that are hidden in data. In Link analysis, one needs to reduce the graphs so that the analysis is manageable and not explosive. Link analysis can then be used to analyze the activities of individuals by forming a link of their activities. These links might be in form of telephone conversation, places visited, bank transactions etc.

FINANCIAL CRIME DETECTION

Financial crime here refers to money laundering, violative trading, and insider trading. The Financial Crimes operates with an expert system with Bayesian inference engine to output suspicion scores and with link analysis to visually examine selected subjects or accounts. Supervised techniques such as case-based reasoning, nearest neighbor retrieval, and decision trees were seldom used due to propositional approaches, lack of clearly labeled positive examples, and scalability issues. Unsupervised techniques were avoided due to difficulties in deriving appropriate attributes. It has enabled effectiveness in manual investigations and gained insights in policy decisions for money laundering.

Use of large amounts of unstructured text and web data such as free-text documents, web pages, emails, and SMS messages, is common in adversarial domains but still unexplored in fraud detection literature. Link Discovery on Correlation Analysis uses a correlation measure with fuzzy logic to determine similarity of patterns between thousands of paired textual items which have no explicit links. It comprises link hypothesis, link generation, and link identification based on financial transaction timeline analysis to generate community models for the prosecution of money laundering criminals. Relevant sources of data can decrease detection time, expert systems and clustering for finding and predicting early symptoms of insider trading in option markets before any news release.

## IX. CRIME REPORTING SYSTEMS

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 or sheriff‟s 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. 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. We will look at how to arrive at the significant attributes for the data mining models.

## X. CONCLUSION

Data mining applied in the context of law enforcement and intelligence analysis holds the promise of alleviating crime related problem. Using a wide range of techniques it is possible to discover useful information to assist in crime matching, not only of single crimes, but also of series of crimes. In this paper we use a clustering/classify based model to anticipate crime trends. The data mining techniques are used

to analyze the crime data from database. The results of this data mining could potentially be used to lessen and even prevent crime for the forth coming years. We believe that crime data mining has a promising future for increasing the effectiveness and efficiency of criminal and intelligence analysis.

## XI. REFERENCES

[1] Levine, N., 2002. CrimeStat: A Spatial Statistic Program for the Analysis of Crime Incident Locations (v 2.0).Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC.May 2002.
[2] Ewart, B.W., Oatley, G.C., & Burn K., 2004.Matching Crimes Using Burglars‟ Modus Operandi: A Test of Three Models.
[3] [3] Ratcliffe J.H., 2002.Aoristic signatures and spatiotemporal analysis of high volume crime patterns. J.Quantitative Criminology.
[4] Chen, H., Chung, W., Xu, J.J., Wang,G., Qin, Y., & Chau, M., 2004. Crime data mining: a general framework and some examples. IEEE Computer April 2004, Vol 37.
[5] Polvi, N., Looman, T., Humphries, C., & Pease, K., 1991.The Time Course of Repeat Burglary Victimization. British Journal of Criminology
[6] Brown, D.E.The regional crime analysis program: A frame work for mining data to catch criminals," in Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics.
[7] Abraham, T. and de Vel, O. Investigative profiling with computer forensic log data and association rules, Proceedings of the IEEE International Conference on Data Mining
[8] Townsley, M. (2006) Spatial and temporal patterns of burglary: Hot spots and repeat victimization in an Australian police divison. Ph.D. thesis, Griffith University, Brisbane, Australia
[9] Mannheim, H. Comparative Criminology (Volume 1). London: Routledge and Kegan Paul.