IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 70-74

Role of Data Mining in Cyber Security

T.Nandhini, C.Rangarajan
Research Scholar, Pioneer College of Arts and Science
Head & Associate Professor, Pioneer College of Arts and Science,

Abstract: *Data mining is becoming a invasive knowledge in activities as varied as using historical data to forecast the success of a marketing process looking for patterns in monetary contact to discover banned activities or analyzing genome sequences From this perception it was just a material of time for the control to reach the important area of computer safety This book presents a collection of investigate efforts on the use of data mining in computer safety.*
Keywords: *Scan Detection; Virus exposure; Anomaly disclosure; Shelter*

I. Introduction

Data mining is a popular technological novelty that converts piles of data into useful knowledge that can help the data owners/users make knowledgeable choices and take stylish actions for their own advantage. In unambiguous terms, data mining looks for hidden patterns among huge sets of data that can help to understand, predict, and direct future behavior. A more technical explanation: Data Mining is the set of methodologies used in analyzing data from various size and perspectives, finding previously unknown hidden patterns, classifying and grouping the data and summarizing the standard relationships. Data mining is, at its core, pattern judgment. Data miners are experts at using specialized software to find regularities (and irregularities) in huge data sets. Here are a hardly any specific things that data mining might contribute to an disturbance detection project:

- Get rid of normal activity from terror data to agree to analysts to centre on straightforward attacks
- Identify bogus alarm generators and "bad" sensor signatures
- Find rough activity that uncovers a factual attack
- Identify long, underprovided patterns (different IP address, same activity)

To accomplish these tasks, data miners use one or more of the following techniques:

a. Data summarization with statistics, counting finding outliers
b. Visualization: presenting a graphical summing-up of the data
c. Clustering of the statistics into customary categories
d. Association rule discovery: important normal activity and enabling the discovery of anomalies [Clifton and Gengo , 2000; Barbara et al., 2001]
e. Classification: predicting the category to which a fussy record belongs [Lee and Stolfo, 1998]

Data mining has many applications in security as well as in national security (e.g., surveillance) as well as in cyber security (e.g., virus detection). The threats to national security embrace attacking buildings and destroying critical infrastructures such as power grids and telecommunication systems. Data mining techniques are being used to identify suspicious persons and groups, and to find out which folks and groups are capable of delivery out fundamental activities. Cyber security is disturbed with defensive computer and network systems from bribery due to hateful software counting Trojan horses and viruses. Data mining is also being functional to provide solutions such as disruption detection and auditing. In this paper we will focus mainly on data mining for cyber security applications. Data mining for cyber security applications For example, variance detection techniques could be used to detect unusual patterns and behaviors. Link analysis may be used to trace the viruses to the perpetrators. Classification may be used to group a variety of cyber-attacks and then use the profiles to detect an bother when it occurs. forecast may be used to settle on potential future attacks depending in a way on information learnt about terrorists from end to end email and phone conversations. Data mining is also living being applied for disturbance discovery and auditing The straight come near to securing computer systems against cyber threats is to devise mechanisms such as firewalls, certification tools, and virtual private networks that create a shielding screen. However, these mechanisms round about always have vulnerabilities. They cannot ward attacks that are continually being adapted to utilize system weaknesses, which are regularly caused by slapdash propose and routine flaws. This has created the need for interruption uncovering, security know-how that complements conventional security approaches by monitoring systems and identifying computer attacks. Traditional intrusion detection methods are based on human experts extensive Knowledge of attack

signatures which are character strings in a messages payload that designate hateful content. Signatures have several limitations. They cannot detect novel attacks, because someone must manually revise the signature database ahead of time for each new type of intrusion bare. Once someone discovers a new attack and develops its signature, deploying that signature is often delayed. These restrictions have led to an going up interest in interruption detection techniques based on data mining.

II. Data Mining For Network Security

2.1 Overview

This section discusses information associated hostility. By information related terrorism we mean cyber terrorism as well as safe haven violations through access organize and other means. Malicious software such as Trojan horses and viruses are also in turn related good hands violations, which we group into information related terrorism activities. In the next few subsections we discuss a assortment of in turn connected terrorist attacks. In section 2.2 we discussed about Anomaly Detection, in section 2.3. Profiling Network Traffic Using Clustering In Section 2.4. Scan Detection, In Section 2.5. Methodology, In Section 2.6. Cyber-terrorism, Insider Threats, and External Attacks, In Section 2.7 Credit Card Fraud and Identity Theft, in section 2.8 Attacks on Critical Infrastructures.

2.2 Anomaly Detection

Anomaly detection approaches build models of typical data and detect deviations from the normal model in observed data. Anomaly finding applied to intrusion detection and computer security has been an active area of research since it was initially proposed by Denning. Anomaly detection algorithms have the advantage that they can detect emerging threats and attacks (which do not have signatures or categorized data corresponding to them) as deviations from normal usage. Moreover, unlike misuse detection schemes (which build classification models using labelled data and then catalogue an observation as normal or attack), anomaly detection algorithms do not require an explicitly labelled training data set, which is very desirable, as labelled data is difficult to obtain in a real complex setting.

2.3 Profiling Network Traffic Using Clustering

Clustering is a roughly used data mining procedure which groups parallel items, to obtain consequential groups/clusters of data items in a data set. These clusters represent the overriding modes of behavior of the data objects strong-minded using a resemblance measure. A data forecaster can get a high level understanding of the characteristics of the data set by analyzing the clusters. Clustering provides an imaginative resolution to determine the accepted and surprising modes of behavior and to obtain a high level selfless of the network transfer.

2.4 Scan Detection

A ancestor too many attacks on networks is often a reconnaissance operation, more commonly referred to as a scan. Identifying what attackers are scanning for can alert a system commissioner or sanctuary analyst to what services or types of computers are being embattled. Knowing what services are being targeted before an attack allows an administrator to take deterrent measures to protect the resources e.g. installing patches, firewalling services from the exterior, or removing services on machines which do not need to be running them.

2.5 Methodology

Currently solution is a batch-mode carrying out that analyzes data in windows of 20 minutes. For each 20-minute close watch period, we transform the Net Flow data into a summary data set. Figure 3 depicts this process. With our meeting point on incoming scans, each new digest record corresponds to a impending scanner that is pair of external source IP and intention port (SIDP). For each SIDP, the summary record contains a set of features constructed from the unrefined Net flows on hand during the observation window. Observation window size of 20 minutes is somewhat arbitrary. It needs to be large enough to engender features that have unwavering values, but short enough so that the construction of summary records does not take too much time or memory. Above disclaimer are about the uproar detection techniques based on data mining, let us discuss the contravention various information about

- Cyber-terrorism, Insider Threats, and External Attacks
- Credit card and identity theft
- Attacks on critical infrastructures

2.6 Cyber-terrorism, Insider Threats, and External Attacks

Cyber-terrorism is one of the major fanatic threats posed to our nation today. As we have mentioned earlier, this threat is exacerbated by the vast quantities of in sequence now available electronically and on the web. Attacks on our computers, networks, databases and the Internet infra-structure could be shocking to businesses. It is imprecise 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 depository could lose millions and perhaps billions of dollars. By crippling the computer structure millions of hours of productivity could be lost, which is eventually equivalent to direct monetary loss. Even a straightforward power outage at work through some disaster could cause several hours of productivity loss and as a consequence a major monetary loss. Therefore it is essential that our information systems be secure. We discuss a choice of types of cyber-terrorist attacks. One is the propagation of malevolent mobile code that can damage or leak sensitive files or other data; another is intrusions upon computer networks. Coercion 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 stretch viruses that injure files in various computer systems. But a more threatening crisis is that of the insider threat. Insider threats are relatively well unstated in the context of non-information related attacks, but information related insider threats are often unobserved or underestimated. People inside an organization who have studied the business' practices and procedures have an massive advantage when just beginning schemes to cripple the organization's information assets. These people could be regular employees or even those working at computer centres. The problem is quite serious as a big name may be masquerading as someone else and causing all kinds of damage. In the next few sections we will examine how data mining can be leveraged to become aware of and perhaps prevent such attacks.

2.7 Credit Card Fraud and Identity Theft

We are hearing a lot these days about tribute card fraud and identity theft. In the case of credit card racket, an attacker obtains a person's credit card and uses it to make unlawful purchases. By the time the owner of the card becomes attentive of the scam, it may be too late to reverse the harm or arrest the lawbreaker. A similar problem occurs with telephone calling cards. In fact this type of attack has happened to me personally. Perhaps while I was making phone calls using my calling card at airports someone noticed the dial tones and reproduced them to make free calls. This was my company calling card. Beneficially our telephone company detected the problem and informed my company. The problem was dealt with immediately. A more serious theft is identity theft. Here one assumes the identity of another person by acquiring key personal information such as social safekeeping number, and uses that information to carry out transactions under the other person's name. Even a single such transaction, such as selling a house and depositing the income in a deceptive depository account, can have distressing consequences for the wounded. By the time the owner finds out it will be far too late. It is very likely that the owner may have lost millions of dollars due to the identity burglary. We need to explore the use of data mining both for credit card fraud detection as well as for distinctiveness theft. There have been some efforts on detecting credit card fraud. We need to start working actively on detecting and preventing identity thefts.

2.8 Attacks on Critical Infrastructures

Attacks on critical infrastructures could cripple a realm and its financial system. Infrastructure attacks include aggressive the telecommunication lines, the thrilling, command, chatter, reservoirs and water sup-plies, food supplies and other basic entities that are critical for the manoeuvre of a nation. Attacks on critical infrastructures could occur during any type of attack whether they are non-information related, information related or bioterrorism attacks. For example, one could attack the software that runs the telecommunications engineering and close down all the wire shape. Similarly, software that runs the power and gas supplies could be attacked. Attacks could also occur through missiles and explosives. That is, the telecommunication lines could be in the flesh attacked. Attacking transportation lines such as highways and railway tracks are also attacks on infrastructures. Infrastructures could also be attacked by expected disaster such as hurricanes and earth quakes. Our main interest here is the attacks on infrastructures through malevolent attacks, both information associated and non-information related. Our goal is to examine data mining and interrelated data management technologies to detect and prevent such infrastructure attacks.

III. Data Mining Techniques

The art of data mining has been continuously emergent. There are a number of inventive and discerning techniques that have emerged that tweak data taking out concepts in a tender to give companies more all-inclusive insight into their own data with useful opportunity trends. Many techniques are engaged by the data mining experts, some of which are listed below:

3.1 Seeking Out Incomplete Data:

Data mining relies on the actual data present, hence if data is incomplete, the results would be utterly off-mark. Hence, it is of the quintessence to have the astuteness to snivel out incomplete data if possible. Techniques such as Self-Organizing-Maps (SOM's), help to map misplaced data based by visualizing the model of multi-dimensional intricate data. Multi-task learning for missing inputs, in which one existing and true to life data set along with its procedures is compared with another like-minded but lacking data set is one way to seek out such data. Multi-dimensional preceptors using intelligent algorithms to build attribution techniques can address curtailed attributes of data.

3.2 Dynamic Data Dashboards:

This is a scoreboard, on a manager or supervisor's computer, fed with real-time from data as it flows in and out of various databases within the company's upbringing. Data mining techniques are applied to give live just round the corner and monitoring of data to the stakeholders.

•Database Analysis:

Databases grip key data in a well thought-out format, so algorithms built using their own language (such as SQL macros) to find secreted patterns within embarrassed data is most useful. These algorithms are now and again ingrained into the data flows, e.g. tightly together with user-defined functions, and the findings presented in a ready-to-refer-to report with momentous analysis. A good modus operandi is to have the shot dump of data from a hefty database in a cache file at any time and then analyze it auxiliary. Similarly, data mining algorithms must be capable to pull out data from multiple, heterogeneous databases and foretell shifting trends.

• Text Analysis:

This concept is very helpful to repeatedly find patterns within the text surrounded in hordes of text files, word processed files, PDFs, and staging files. The text processing algorithms can for instance, find out repetitive extracts of data, which is moderately useful in the publishing business or universities for tracing plagiarism.

• Efficient Handling of Complex and Relational Data:

A data warehouse or bulky data provisions must be supported with interactive and query-based data mining for all sorts of data mining functions such as cataloguing, clustering, friendship, prophecy. OLAP (Online Analytical Processing) is one such useful attitude. Other concepts that make possible interactive data mining are analyzing graphs, cumulative querying, image arrangement, meta-rule guided mining, swap over randomization, and multidimensional geometric analysis.

• Relevance and Scalability of Chosen Data Mining Algorithms:

While selecting or choosing data mining algorithms, it is of the essence that enterprises keep in mind the business relevance of the predictions and the scalability to reduce expenditure in future. Multiple algorithms should be able to be executed in parallel for time efficiency, independently and without interfering with the conglomerate business applications, more than ever time-critical ones. There should be sustain to include SVMs on well-built scale.

• Popular Tools for Data Mining:

There are many ready-made tools to be had for data mining in the bazaar today. Some of these have common functionalities packaged within, with provisions to add-on functionality by taking sides building of business-specific analysis and cleverness.

Listed Below Is Some Of The Popular Multipurpose Data Mining Tools That Are Leading The Trends:

• Rapid Miner(erstwhile YALE):

This is very admired since it is a ready-made, open source, no coding required software, which gives difficult analytics. Written in Java, it incorporates all-round data mining functions such as data pre-processing, mental picture, projecting analysis, and can be easily integrated with WEKA and R-tool to directly give models from scripts written in the former two.

• WEKA:

This is a JAVA based customization utensil, which is at no cost to use. It includes spectre and prophetic chemical analysis and modelling techniques, clustering, association, regression and classification.

- **R-Programming Tool:**

This is written in C and FORTRAN, and allows the data miners to write scripts just like a brainwashing language/raised area. Hence, it is used to make geometric and undercover software for data mining. It supports graphical analysis, both linear and nonlinear modelling, cataloguing, clustering and time-based data analysis.

•Python based Orange and NTLK:

Python is very all the rage due to ease of use and its prevailing features. Orange is an open source tool that is written in Python with useful data analytics, text analysis, and machine learning features surrounded in a visual programming crossing point. NTLK, also composed in Python, is a authoritative language processing data mining tool, which consists of data mining, machine learning, and data scraping features that can easily be built up for made to order needs.

•Knime:

Primarily used for data pre-processing – i.e. data withdrawal, restoration and loading, Knime is a overriding tool with GUI that shows the set of connections of data nodes. Popular in the midst of pecuniary data analysts, it has modular statistics pipe lining, leveraging machine learning, and data mining concepts abundantly for building big business cleverness reports. Data mining tools and techniques are now more urgent than ever for all businesses, big or small, if they would like to weight their existing data provisions to make business decisions that will give them a ready for action edge. Such actions based on data authentication and advanced analytics have better chances of greater than ever sales besides facilitating growth. Adopting well time-honoured techniques and tools and availing the help of data mining experts shall give a hand companies to make use of relevant and powerful data mining concepts to their fullest potential.

Reference:

[1]. Data Mining for Security Applications : Bhavani Thuraisingham, Latifur Khan, Mohammad M. Masud, Kevin W. Hamlen

[2]. Rakesh Agrawal, Tomasz Imieliski, and Arun Swami. Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data.

[3]. Daniel Barbara and Sushil Jajodia, editors. Applications of Data Mining in Computer Security. Kluwer Academic Publishers

[4]. Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and J Sander. Lof: identifying density-based local outliers. In Proceedings of the 2000 ACM SIG-MOD international conference on Management of data, pages

[5]. Varun Chandola and Vipin Kumar. Summarization {compressing data into an informative representation. In Fifth IEEE International Conference on Data Mining, pages.

[6]. [1] Thuraisingham, B., "Web Data Mining Technologies and Their Applications in Business Intelligence and Counterterrorism", CRC Press, FL, 2003.

[7]. Chan, P, et al, "Distributed Data Mining in Credit Card Fraud Detection", IEEE Intelligent Systems.

[8]. [3] Lazarevic, A., et al., "Data Mining for Computer Security Applications", Tutorial Proc. IEEE Data Mining Conference, 2011.

[9]. Thuraisingham, B., "Managing Threats to Web Databases and Cyber Systems, Issues, Solutions and Challenges", Kluwer, MA 2004 (Editors: V. Kumar et al).

[10]. Thuraisingham B., "Database and Applications Security", CRC Press, 2005.

[11]. Thuraisingham B., "Data Miming, Privacy, Civil Liberties and National Security", SIGKDD Explorations, 2012.