Role of Data Mining in Cyber Security

K. Madhu Shre[1] , K.Sophia[2]
Students of ECE Department,
Pits, Thanjavur.

Abstract: **Data mining is becoming a pervasive technology in activities as diverse as using historical data to predict the success of a marketing campaign looking for patterns in financial transactions to discover illegal activities or analyzing genome sequences From this perspective it was just a matter of time for the discipline to reach the important area of computer security This book presents a collection of research efforts on the use of data mining in computer security.**

Keywords: Scan Detection; Virus Detection; Anomaly Detection; Security

I. INTRODUCTION

Data mining is a popular technological innovation that converts piles of data into useful knowledge that can help the data owners/users make informed choices and take smart actions for their own benefit. In specific terms, data mining looks for hidden patterns amongst enormous sets of data that can help to understand, predict, and guide future behavior. A more technical explanation: Data Mining is the set of methodologies used in analyzing data from various dimensions and perspectives, finding previously unknown hidden patterns, classifying and grouping the data and summarizing the identified relationships. Data mining is, at its core, pattern finding. Data miners are experts at using specialized software to find regularities (and irregularities) in large data sets. Here are a few specific things that data mining might contribute to an intrusion detection project:

- Remove normal activity from alarm data to allow analysts to focus on real attacks
- Identify false alarm generators and "bad" sensor signatures
- Find anomalous activity that uncovers a real attack
- Identify long, ongoing patterns (different IP address, same activity)
- To accomplish these tasks, data miners use one or more of the following techniques:
- Data summarization with statistics, including finding outliers
- Visualization: presenting a graphical summary of the data
- Clustering of the data into natural categories [Manganaris et al., 2000]
- Association rule discovery: defining normal activity and enabling the discovery of anomalies [Clifton and Gengo, 2000; Barbara et al., 2001]
- Classification: predicting the category to which a particular record belongs [Lee and Stolfo, 1998]

Data mining has many applications in security including in national security (e.g., surveillance) as well as in cyber security (e.g., virus detection). The threats to national security include attacking buildings and destroying critical infrastructures such as power grids and telecommunication systems. Data mining techniques are being used to identify suspicious individuals and groups, and to discover which individuals and groups are capable of carrying out terrorist activities. Cyber security is concerned with protecting computer and network systems from corruption due to malicious software including Trojan horses and viruses. Data mining is also being applied to provide solutions such as intrusion 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, anomaly 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 various cyber-attacks and then use the profiles to detect an attack when it occurs. Prediction may be used to determine potential future attacks depending in a way on information learnt about terrorists through email and phone conversations. Data mining is also being applied for intrusion detection and auditing The conventional approach to securing computer systems against cyber threats is to design mechanisms such as firewalls, authentication tools, and virtual private networks that create a protective shield. However, these mechanisms almost always have vulnerabilities. They cannot ward attacks that are continually being adapted to exploit system weaknesses, which are often caused by careless design and implementation flaws. This has created the need for intrusion detection, security technology 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 indicate malicious content. Signatures have several limitations. They cannot detect novel attacks, because someone must manually revise the signature database beforehand for each new type of intrusion discovered. Once someone discovers a new attack and develops its signature, deploying that signature is often delayed. These Limitations have led to an increasing interest in intrusion detection techniques based on data mining.

II. DATA MINING FOR NETWORK SECURITY

2.1 Overview

This section discusses information related terrorism. By information related terrorism we mean cyber terrorism as well as security violations through access control and other means. Malicious software such as Trojan horses and viruses are also information related security violations,

which we group into information related terrorism activities. In the next few subsections we discuss various information related 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 normal data and detect deviations from the normal model in observed data. Anomaly detection applied to intrusion detection and computer security has been an active area of research since it was originally proposed by Denning. Anomaly detection algorithms have the advantage that they can detect emerging threats and attacks (which do not have signatures or labeled data corresponding to them) as deviations from normal usage. Moreover, unlike misuse detection schemes (which build classification models using labeled data and then classify an observation as normal or attack), anomaly detection algorithms do not require an explicitly labeled training data set, which is very desirable, as labeled data is difficult to obtain in a real network setting.

2.3 Profiling Network Traffic Using Clustering

Clustering is a widely used data mining technique which groups similar items, to obtain meaningful groups/clusters of data items in a data set. These clusters represent the dominant modes of behavior of the data objects determined using a similarity measure. A data analyst can get a high level understanding of the characteristics of the data set by analyzing the clusters. Clustering provides an effective solution to discover the expected and unexpected modes of behavior and to obtain a high level understanding of the network traffic.

2.4 Scan Detection

A precursor 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 administrator or security analyst to what services or types of computers are being targeted. Knowing what services are being targeted before an attack allows an administrator to take preventative measures to protect the resources e.g. installing patches, firewalling services from the outside, or removing services on machines which do not need to be running them.

2.5 Methodology

Currently solution is a batch-mode implementation that analyzes data in windows of 20 minutes. For each 20-minute observation period, we transform the Net Flow data into a summary data set. Figure 3 depicts this process. With our focus on incoming scans, each new summary record corresponds to a potential scanner that is pair of external source IP and destination port (SIDP). For each SIDP, the summary record contains a set of features constructed from the raw Net flows available during the observation window. Observation window size of 20 minutes is somewhat arbitrary. It needs to be large enough to generate features that have reliable values, but short enough so that the construction of summary records does not take too much time or memory. Above specifications are about the intrusion detection techniques based on data mining, let us discuss the intrusion 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 terrorist threats posed to our nation today. As we have mentioned earlier, this threat is exacerbated by the vast quantities of information now available electronically and on the web. 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. We discuss various types of cyber-terrorist attacks. One is the propagation of malicious mobile code that can damage or leak sensitive files or other data; another is intrusions upon computer networks. 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 be leveraged to detect and perhaps Prevent such attacks.

2.7 Credit Card Fraud and Identity Theft

We are hearing a lot these days about credit card fraud and identity theft. In the case of credit card fraud, an attacker obtains a person's credit card and uses it to make unauthorized purchases. By the time the owner of the card becomes aware of the fraud, it may be too late to reverse the damage or apprehend the culprit. 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. Fortunately 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 security 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 fraudulent bank account, can have devastating consequences for the victim. 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 theft. We need to explore the use of data mining both for credit card fraud detection as well as for identity 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 nation and its economy. Infrastructure attacks include attacking the telecommunication lines, the electric, power, gas, reservoirs and water sup-plies, food supplies and other basic entities that are critical for the operation of a nation. Attacks on critical infrastructures could occur during any type of attack whether they are non-information related, information related or bio-terrorism attacks. For example, one could attack the software that runs the telecommunications industry and close down all the telecommunication lines. Similarly, software that runs the power and gas supplies could be attacked. Attacks could also occur through bombs and explosives. That is, the telecommunication lines could be physically attacked. Attacking transportation lines such as highways and railway tracks are also attacks on infrastructures. Infrastructures could also be attacked by natural disaster such as hurricanes and earth quakes. Our main interest here is the attacks on infrastructures through malicious attacks, both information related and non-information related. Our goal is to examine data mining and related data management technologies to detect and prevent such infrastructure attacks

III. DATA MINING TECHNIQUES

The art of data mining has been constantly evolving. There are a number of innovative and intuitive techniques that have emerged that fine-tune data mining concepts in a bid to give companies more comprehensive insight into their own data with useful future trends. Many techniques are employed 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 completely off-mark. Hence, it is imperative to have the intelligence to sniff out incomplete data if possible. Techniques such as Self-Organizing-Maps (SOM's), help to map missing data based by visualizing the model of multi-dimensional complex data. Multi-task learning for missing inputs, in which one existing and valid data set along with its procedures is compared with another compatible but incomplete data set is one way to seek out such data. Multi-dimensional preceptors using intelligent algorithms to build imputation techniques can address incomplete 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 environment. Data mining techniques are applied to give live insight and monitoring of data to the stakeholders.

Database Analysis:

Databases hold key data in a structured format, so algorithms built using their own language (such as SQL macros) to find hidden patterns within organized data is most useful. These algorithms are sometimes inbuilt into the data flows, e.g. tightly coupled with user-defined functions, and the findings presented in a ready-to-refer-to report with meaningful analysis. A good technique is to have the snapshot dump of data from a large database in a cache file at any time and then analyze it further. Similarly, data mining algorithms must be able to pull out data from multiple, heterogeneous databases and predict changing trends.

Text Analysis:

This concept is very helpful to automatically find patterns within the text embedded in hordes of text files, word-processed files, PDFs, and presentation files. The text-processing algorithms can for instance, find out repeated extracts of data, which is quite useful in the publishing business or universities for tracing plagiarism.

Efficient Handling of Complex and Relational Data:

A data warehouse or large data stors must be supported with interactive and query-based data mining for all sorts of data mining functions such as classification, clustering, association, prediction. OLAP (Online Analytical Processing) is one such useful methodology. Other concepts that facilitate interactive data mining are analyzing graphs, aggregate querying, image classification, meta-rule guided mining, swap randomization, and multidimensional statistical analysis.

Relevance and Scalability of Chosen Data Mining Algorithms:

While selecting or choosing data mining algorithms, it is imperative that enterprises keep in mind the business relevance of the predictions and the scalability to reduce costs in future. Multiple algorithms should be able to be executed in parallel for time efficiency, independently and without interfering with the transnational business applications, especially time-critical ones. There should be support to include SVMs on larger scale.

Popular Tools for Data Mining:

There are many ready-made tools available for data mining in the market today. Some of these have common functionalities packaged within, with provisions to add-on functionality by supporting building of business-specific analysis and intelligence.

LISTED BELOW IS SOME OF THE POPULAR MULTI-PURPOSE DATA MINING TOOLS THAT ARE LEADING THE TRENDS:

Rapid Miner (erstwhile YALE):

This is very popular since it is a ready-made, open source, no-coding required software, which gives advanced analytic s. Written in Java, it incorporates multifaceted data mining functions such as data preprocessing, visualization, predictive 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 tool, which is free to use. It includes visualization and predictive analysis and modeling 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 programming language/platform. Hence, it is used to make statistical and analytical software for data mining. It supports graphical analysis, both linear and nonlinear modeling, classification, clustering and time-based data analysis.

Python based Orange and NTLK:

Python is very popular due to ease of use and its powerful features. Orange is an open source tool that is written in Python with useful data analytic s, text analysis, and machine-learning features embedded in a visual programming interface. NTLK, also composed in Python, is a powerful language processing data mining tool, which consists of data mining, machine learning, and data scraping features that can easily be built up for customized needs.

Knime:

Primarily used for data preprocessing – i.e. data extraction, transformation and loading, Knime is a powerful tool with GUI that shows the network of data nodes. Popular amongst financial data analysts, it has modular data pipe lining, leveraging machine learning, and data mining concepts liberally for building business intelligence reports. Data mining tools and techniques are now more important than ever for all businesses, big or small, if they would like to leverage their existing data stores to make business decisions that will give them a competitive edge. Such actions based on data evidence and advanced analytics have better chances of increasing sales and facilitating growth. Adopting well-established techniques and tools and availing the help of data mining experts shall assist companies to utilize 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] Thuraisingham, B., “Web Data Mining Technologies and Their Applications in Business Intelligence and Counter-terrorism”, CRC Press, FL, 2003.

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

[8] 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.