DATA MINING FOR SECURITY APPLICATIONS

Asmita R Namjoshi
Asst. Prof., TMV - Pune
asmita03@gmail.com

ABSTRACT

The following paper is a worked on detection of intrusion of malicious attacks and data mining technique which allowed detecting at zero level of execution. We also summarized our study with existing tools which are used but not on appropriate level. This is necessary to detect malicious code in network traffic.

KEYWORDS: *Data Mining, Credit Card Fraud, Cyber Security, Security of application, Security of data.*

I. INTRODUCTION

Data mining is technique of detection of data. Data can be of any type for examples – we would like to have records of such candidates who are having habits of purchasing online but not everything except small groups of electronics. Now for doing this we need to have data of all purchase and with identification of purchase that who have done this transaction.

Now we will look at data it may have numerous record of transactions for any type may be grocery goods, combination of medicine and grocery, only electronic goods. Your mining of information will dig in various steps as;

1. Finding the right list of candidates that has done at least one purchase of electronics.
2. Finding list of candidate who have purchased complete purchased of electronic goods.
3. Did purchase of some 10,000 or greater.
4. Did purchase of some 10,000 or less.

Cyber Security terms relate your data to be secure while a customer/user does transaction online. This relates to secure computer networks and communication between end to end machines. Information like defense networks, proprietary research, intellectual property, and data based market mechanisms should be secured as well running without leak of information.

What is a threat [1]? There are two types of threats –

1. Real Time threats
2. Non real time threats

Real time threats are type where we have some time limit and in that frame we need to acted upon. Non real time threat having no time limit and we can act upon on But there is one thing to be considered that non real time threats can be converted into real time threats at any point of time. For example – Nation enemy will try to steal information about missile program. This is like a non-real time threats because we do not know the time frame. Now an intelligence agency added some information that it will be acted in May month as this time most people get leave. Now this convert into real time threats and we need to act on this before May month as we have time limits. So much work has been done here for applying data mining in security[1].

In this part of the paper we will discuss data mining for cyber security. In section 2 we will discuss data mining for cyber security applications. In particular, we will discuss threats to computers and networks and describe applications of data mining to detect such threats and attacks. Some of our current research at the University of Texas at Dallas will be discussed in section 3. The paper is summarized in section 4.

II. DATA MINING FOR CYBER SECURITY

OVERVIEW

Here we discuss information related terrorism. By information related terrorism we mean cyber- terrorism and data security violations through access of machine and by other means. Installing malicious software such as viruses for stealing data, in the next few subsections we discuss various information related terrorist attacks. In section A, An overview of cyber-terrorism and then insider threats and external attacks is discussed. B deals with Credit card and identity theft. Attacks on critical infrastructures are discussed in section C. Data mining for cyber security is discussed in section D.

A. CYBER THREATS INSIDE AND EXTERNAL

Any kind of terrorism is one of the major threats posed to our nation today. As we have mentioned earlier, To demolish a nation there is no need of physical attack by using an army but this can be done by following way also;

1. Downing economic infra.
2. Stealing info of proposed program.
3. Defense move info stealing.

So this threat is exacerbate by the vast quantities of information now available electronically and on the web. By making attacks on our computers, networks, databases and the Internet infra-structure could be lead to demolish of 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 Rupees. By changing the computer system Lakhs of hours of productivity could be lost, which lead to economic chaos in country. This is another type of terrorism called Cyber terrorism which required no troops. 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.

Now the situation is discussed is about the outer threat, this also can be done from inner side. Inner side threats sources are employees, employees of data centers etc. They also are vulnerable to soft data. A data center employee have access to all data like – An ABC company having customer care center for any card providing company, now If a customer reported the complaint of blocking due to theft of card then the first person who know this is that employee and this person has all the access of this customer information and he can lead this info to unauthorized person for further fraud process.

B. CREDIT CARD INFO FRAUD AS WELL PERSONAL IDENTITY THEFT

Credit card [2] is work as cash in advance from bank. To do transaction either you need to swipe or give details of card like no, CVV and expiry, which is actually part of Card itself. So if one needs to make fraud with credit card then he must be able to snatch all details and it can be done using information theft. Second thing is identity theft like making fake identity of yours.

It is simple to understand that a fake person is standing on behalf of yours documents and details. Now he can purchase a SIM card on your document. Open a bank account in your name. Do transaction and after making fraud money he withdraws and got invisible. This is called identity theft.

We need to explore the use of data mining both for credit card fraud detection as well as for identity theft. We need to start working actively on detecting and preventing identity thefts based on patterns from where they got loop holes.

C. THREATS ON VITAL INFRASTRUCTURE-

A nation creates its infrastructure for mobile and telecommunication, banking system, air traffic management system, rail traffic management system, gas pipeline management system.

This all system runs on software which directs the machinery to work on scheduled way. Like in traffic management everything is going on sensor based and which signal is to be red and green, decision taken by software based on sensor.

Now if I temper the data of sensor which is sent to micro controller then this will lead to mismanagement of traffic. Similarly this is also applied to gas pipeline pressure management, anyone can understand if we temper with it then unusual pressure will be mobilized and lead to disaster. Attacks on critical infrastructures could cripple a nation and its economy.

Attacks on critical infrastructures could occur during any type of attack whether they are non-information related, information related.

Here we are not considered the attacks of nature call like flood, earth quack etc. Our goal is to examine data mining and related data management technologies to detect and prevent such infrastructure attacks which are manmade.

D. DATA MINING FOR DATA SECURITY

This is matter of research that how to prevent data from unauthorized access? By taking example one can say that a person working in data center which is responsible of input details of new credit card in to system as well password generation. There may be several people are working in different shift. Now how to identify that an action is going on is actual or attack or Trojan execution?

To find this we need to understand the pattern. Suppose a user logs on system usually in between 10 to 2 pm. But if that user logged on in evening or in night then this needs an investigation why it is happening. May be that person is there due to shift change or might be possibility that Trojan has accessed his account. Large data center where data is kept for future, warehouse, should keep an eye on this behavior. Many researchers are investigating the use of data mining for intrusion detection. While we need some form of real-time data mining, that is, the results have to be generated in real- time, we also need to build models in real-time. For example, credit card fraud detection is a form of real- time processing. However, here models are usually built ahead of time. Building models in real-time remains a challenge. Data mining can also be used for analyzing web logs as well as analyzing the audit trails. Based on the results of the data mining tool, one can then determine whether any unauthorized intrusions have occurred and/or whether any unauthorized queries have been posed.

One may need to monitor the access patterns of all the individuals of a corporation even if they are system administrators to see whether they are

carrying out cyber-terrorism activities [3], [4].

While data mining can be used to detect and prevent cyber-attacks, data mining also worsen some security problems such as inference and privacy. With data mining techniques one could infer sensitive associations from the legitimate responses. For more details on privacy we refer to [5] & [6].

III. OUR RESEARCH AND DEVELOPMENT

A. DATA MINING FOR INTRUSION DETECTION

Network intrusion detection has a versatile scope to check like there is numerous type of protocol like

1. SMTP, UDP, TCP-IP, POP3 etc.

We need to identify the pattern of genuine [9] access and intrusion. We are developing a number of tools that use data mining for cyber security applications at the TMV, including tools like;

a) Intrusion detection,
b) Malicious code detection, and
c) Botnet detection [11].

If we try to define intrusion then we can say only that an action which try to get access of particular service like SMTP [8], FTP or TCP/IP.

Intrusion can be done through to gain action by any GUI which is design to use software but due to error in coding, intruder get facilitated.

This can be done by erroneous data handling in side code from GUI to database software. These terms get name SQL injection in cyber world.

Network-based attacks make it difficult for genuine users to access network services by purposely occupying or damage network resources and services. This can be done by;

a) sending large amounts of network traffic,
b) exploiting well-known faults in networking services,
c) Overloading network hosts [10], etc.

Network-based attack detection uses network traffic data (i.e., TCP dump) to look at traffic addressed to the machines being monitored. Intrusion can also be done by move your traffic to other then your system router. We have used multiple models such as **support vector machines (SVM)**.

However we have improved SVM a great deal by combining it with a novel algorithm that we have developed. We will describe this novel algorithm as well as our approach to combining it with SVM. In addition we will also discuss our experimental results. For more details of our research we refer to [7].

B. SVM APPROACH

In this section we discussed about SVM approach by sating how it use with clustering of data. SVM is based on the idea of a hyper-plane classifier, or linearly separability.

Suppose we have N training data points { (x1,y1),(x2,y2),(x3,y3), ...,(xN,yN)}, where xi∈Rdand yi∈{+1,−1}. Consider a hyper-plane defined by (w,b), where w is a weight vector and b is a bias.Details of SVM can be found on [12].

We can classify a new object x with-

$$f(x)=sign(w\cdot x+b)=signNi\alpha iyi(xi\cdot x)+b \quad -(1)$$

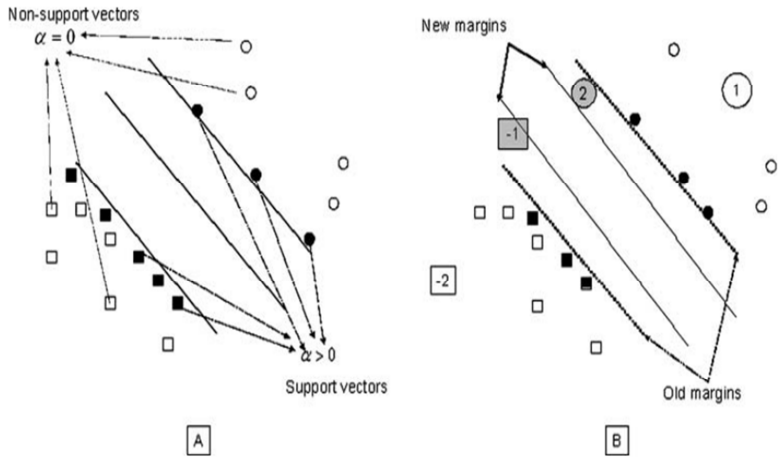

Figure 1 [A]: Value of *αi* for support vector. [B]: The Effect of adding new data points on the margins

The Lagrangian multiplier values αi shows the vital role of each data point. When the maximal margin hyper-plane is found, only points that reside near to the hyper-plane will have αi >0 and these all dots are called vectors.

All other points will have αi =0 (see Fig.1A. This means that only those points that lie closest to the hyper-plane give the representation of the hypothesis/classifier.

IV. SUMMARY AND DIRECTIONS

In this paper, we have applied decrease techniques using clustering and analysis to approximate support vectors in order to gear up non learned training process of SVM.

We have also proposed a method which is Clustering of Trees that is based upon SVM itself. This paper represents secure data transfer in between machines by clustering to support vectors.

Data mining start with support cyber, information security. We also identify anomaly in data packet in transmission. Our future endeavour will be focused on monitoring intrusion opponent data packet in defence.

REFERENCES

[1] Thuraisingham, B., "Web Data Mining Technologies and Their Applications in Business Intelligence and Counter- terrorism", CRC Press, FL, 26-june-2003,Page 333, ISBN 9780849314605 - CAT# AU1460- (Print Book)

[2] Chan, P, et al, "Distributed Data Mining in Credit Card Fraud Detection", IEEE Intelligent Systems, 14 (6), 1999.

[3] Lazarevic, A., et al., "Data Mining for Computer Security Applications", Tutorial Proc. IEEE Data Mining Conference, 2003.

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

[5] Thuraisingham B., "Database and Applications Security", CRC Press, May-26, 2005, Page 500, ISBN- 9780849322242 - CAT# AU2224, Print Book.

[6] Thuraisingham B., "Data Miming, Privacy, Civil Liberties and National Security", SIGKDD Explorations, 2002.

[7] Khan, L., Awad, M. and Thuraisingham, B. "A New Intrusion Detection System using Support Vector Machines and Hierarchica Clustering", The VLDB Journal: ACM/ Springer-Verlag, 16(1), page 507-521, 2007.

[8] Masud, M. M., Khan, L. and Thuraisingham, B. "Feature based Techniques for Auto-detection of Novel Email Worms", In Proc. 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2007), Nanjing, China, May 2007, page 205-216.

[9] Abedin, M., Nessa, S., Khan, L., Thuraisingham, B., "Detection and Resolution of Anomalies in Firewall Policy Rules", In Proc. 20th IFIP WG 11.3 Working Conference

on Data and Applications Security (DBSec 2006), Springer- Verlag, July 2006, Sophia Antipolis, France, page 15-29.

[10] Masud, M. M., Khan, L, Thuraisingham, B., Wang, X., Liu, P., and Zhu, S., "A Data Mining Technique to Detect Remote Exploits", In Proc. IFIP WG 11.9 International Conference on Digital Forensics, Japan, Jan 27-30, 2008.

[11] Masud, M. M., Gao, J., Khan, L., Han, J., Thuraisingham, B., "Peer to Peer Botnet Detection for Cyber-Security: A Data Mining Approach". In Proc. Cyber Security and Information Intelligence Research Workshop (CSIIRW 08), Oak Ridge National Laboratory, Oak Ridge, TN, May 12-14, 2008.

[12] Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer Berlin Heidelberg New York ,01-12-2000, ISBN 978-1-4757-3264-1 (E-book).