A Powerful Tool for Intrusion Detection & Clustering Techniques and Methodology

[1] Sonal R.Chakole, [2] Vijaya Balpande, [3] Vyenktesh Giripunge

[1, 2, 3] Computer Department, Nagpur University, Professor in Priyadarshini J. L. College of Engg.
Maharashtra , India

Abstract - As computer attacks are becoming more and more difficult to identify the need for better and more efficient intrusion detection systems increases. The main problem with current intrusion detection systems is high rate of false alarms. Distributed Denial of Service (DDoS) attacks is large-scale cooperative attack launched from a large number of compromised hosts called Zombies is a major threat to internet services. This paper presents various significant areas where data mining techniques seems to be a strong candidate for detecting and preventing DDOS attack. Purpose for this work is to examine how to integrate multiple intrusion detection sensors in the order to minimize the number of incorrect-alarms.

***Keywords* – DDOS, Intrusion, Data Mining, Zombies, IP Traceback.**

1. Introduction

The main problem with current intrusion detection systems is high rate of false alarms triggered off by attackers. Effective protecting the network against malicious attacks remains problem in both research and the computer network managing professionals. Improved monitoring of malicious attacks will require integration of multiple monitoring systems. In this paper, potential benefits of distributed multi sensor systems for intrusion detection are analyzed. The first phase deals with how to integrate data from multiple sensors, and the second phase is used to identify most important data provided by multiple sensors. A series of analytical and mathematical models are used to acquire potential benefits of multiple sensors for reducing false alarms. The purpose of this paper is to discuss implementation of prototype [1] multi sensor based intrusion detection system. Today, the number of attacks against large computer systems or networks is growing at a rapid pace.

One of the major threats to cyber security is Distributed Denial-of-Service (DDoS) [2, 3] attack. In which the victim network element(s) are bombarded with high volume of fictitious attacking packets originated from a large number of Zombies. The aim of the attack is to overload the victim and render it incapable of performing normal transactions. To protect network servers, network routers [4] and client hosts from becoming the handlers, Zombies and victims of distributed denial-of-service (DDoS) attacks data mining approach can be adopted as a sure shot weapon to these attacks. Recent rapid development in data mining [3] has made available wide variety of algorithms drawn from the fields of statistics, pattern recognition, machine learning, and database. These algorithms [4] made it possible to achieve the ultimate aim of writing this paper. The central theme of this paper is to explore areas where data mining techniques extensively gathers the audited data to compute patterns which predict the actual behavior that can be used for detecting or tracing various DDoS attacks.

2. Literature Survey

A series of research steps have shown that, out of the various categories of DDoS attacks such as flooding, software exploit, and protocol based etc. Distributed Denial of service attack is the most prominent. In fact, DDoS attack uses series of Zombies to initiate a flood attack against an unsafe single site. DDoS attack is initiated in 2-phases [Mirkovic and Reiher 2004] [Dietrich et al. 2000] i.e. Recruiting phase and Action phase. In Recruiting phase attacker initiates the attack from the master computer and tries to find some slave (Zombies) [4, 5] computers to be involved in the attack. A small piece of software is installed on the Zombies to run the attacker commands. The Action phase continued through a command issued from the attacker resides on the master computer toward the Zombies computers to run their pieces of software. The mission of the pieces of software is to send dummy traffic designated toward the victim. The result is a massive flood of packets that crashes the host or swamp down the entire network Very few networks or hosts can effectively cope with such a scale of attacks today. Most of the handler and Zombie are completely unaware of the fact that they were being used for launching of a DDoS attack. Intrusion detection [5] techniques can be classified as misuse detection [3, 5] and anomaly detection. Misuse detection systems, e.g., IDIOT [Kumar and Spafford, 1995] and STAT [Ilgun et.al., 1995], use patterns of well-known attacks or weak spots of the system to match and identify known intrusions. Anomaly detection systems, e.g., IDES [Lunt et. al., 1992] flag observed activities that deviate significantly from the established normal usage profiles as anomalies [5, 6], i.e., possible intrusions. Today the main reason of using Data Mining for intrusion detection systems is the enormous volume of existing and newly appearing network data that requires processing. Literature also provides evidence where data mining techniques are used for intrusion detection. Adaptive learning algorithms are used to improve usability that facilitates model creation and incremental update. Unsupervised anomaly detection algorithms [5] are used to reduce the reliance on labeled data. Author [Lee et. al., 2002] gives an architecture consisting of sensors, detectors, a data warehouse, and model generation components. Presented architecture facilitates the sharing and storage of audit data and the distribution of new or updated models which improves the efficiency and scalability of the IDS (Intrusion Detection System). Another similar example of Intrusion detector has been countered by [Brahmi et.al. 2010] which explains that it is a novel distributed multi-agent IDS architecture, called MAD-IDS [5].

MAD-IDS integrate the mobile agent methodology and the data mining techniques to accommodate the special requirements in distributing IDS. It is clear that the data mining techniques and in particular the unsupervised clustering algorithm and the generic association rule mining are capable of discovering anomalous connections, as well as, generating an informative summarize. [John et.al, 2007] builds a Fuzzy Intrusion Recognition Engine (FIRE) [6, 7], which is an anomaly based intrusion detection system that uses fuzzy logic to evaluate whether malevolent activity is taking place on a network. The FIRE system applies basic data mining techniques to TCP packet data [7] for extracting metrics that are not obvious in the raw data. These metrics are then evaluated as fuzzy sets. DDoS attack detection model presented by [Zhong et.al. 2010] was based on data mining algorithm. FCM cluster algorithm [6, 7] and Apriori association algorithm used to extracts network traffic model and network packet protocol status model. Here threshold is set for detection model.

3. Major Intrusion Detection Techniques

Distributed Denial-of-Service (DDoS) attack is the one in which the victim's network elements are bombarded with high volume of fictitious attacking packets that originate from a large number of machines [Kim et.al. 2004]. A successful attack allows the attacker to gain access to the victim's machine, allowing stealing of sensitive internal data and possibly cause disruption and denial of service (DDoS) in some cases. The number of DDoS attacks grew 20 % last year - a major decrease in the rate of attacks from 2007 to 2008, when these devastating attacks increased 67 percent, according to a report.1 On the same lines, Internet Service Providers (ISPs) [7] are most worried about botnet-driven distributed denial-of-service (DDoS) attacks2.DDoS attacks launched by the group Anonymous took down the Web sites of U.K. record label Ministry of Sound and its legal firm Gallant Macmillan on 3rd Oct, 20103 contributes some latest DDoS attacks Out of the various categories of DDoS attacks such as flooding, software exploit, protocol based etc Distributed Denial of service attack [7, 8] is the most prominent.

In fact, DDoS attack uses series of Zombies to initiate a flood attack against an unsafe single site. DDoS attack is initiated in 2-phases [Mirkovic and Reiher 2004] [Dietrich et al. 2000] i.e. Recruiting phase and Action phase. In Recruiting phase [5] attacker initiates the attack from the master computer and tries to find some slave (Zombies) computers to be involved in the attack. A small piece of software is installed on the Zombies to run the attacker commands. The Action phase continued through a command issued from the attacker resides on the master computer toward the Zombies computers to run their pieces of software. The mission of the pieces of software is to send dummy traffic [5, 6, 7] designated toward the victim. The result is a massive flood of packets that crashes the host or swamp down the entire network operations Very few networks or hosts can effectively cope with such a scale of attacks today. Most of the handler and Zombie are completely unaware of the fact that they were being used for launching of a DDoS.

An authentication policy at routers, filters, firewalls with hardware security appliances, learning based mechanisms, agents based detection at host level or at immediate level etc but none of them has proved to be the best, addressing all the challenges and still there exist a gap amongst the security requirements & existing mechanisms. Therefore, a mechanism that is strong and reliable is desired. Hence the key idea is to use data mining techniques to discover consistent and useful patterns of system features that describe program and user behavior of attack. There are some significant applications which must be noted. Recently, data mining has become an important component for DDoS attack prevention. Different data mining approaches like classification, association rule, clustering, and outlier detection are the few techniques frequently used to analyze network traffic or data to gain knowledge that helps in controlling intrusion. Various applications where data mining approach can be used in prevention and detection of DDoS attacks are discuss below.

3.1 Intrusion Detection

Intrusion detection is the process of observing the events occurring in a computer system or network and analyzing them for instances which violates related security policies [7, 8] or practices. Intrusion detection techniques can be classified as misuse detection and anomaly detection.

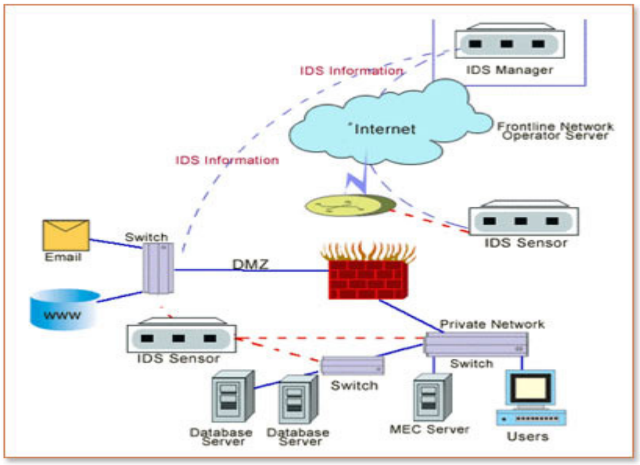

Figure 1. Intrusion Detection System

3.1.1 Misuse Detection

Catch the intrusions in terms of the characteristics of known attacks or system vulnerabilities. It is based on known attack actions. This makes feature extract from known intrusions. It also Integrate the Human knowledge. The rules are pre-defined. Only disadvantage is it cannot detect novel or unknown attacks.

3.1.2 Anomaly Detection

Detect any action that significantly deviates from the normal behavior based on audit data collected over a period of normal operation. When a noise (intrusion) data in the training data, it will make a mis-classification. The features are usually decided by domain experts. It may be not completely.

3.2 IP Trace Back

DDoS is rapidly growing problem. IP Trace back is the ability to trace IP packets from source to destination. This is a significant step towards identifying and thus stopping attackers. The IP Trace back is an important mechanism in defending against DDoS attacks. Lot of techniques and methodologies are used to trace the DDoS attacks An approach suggested by Sager [6] and Stone [7] is called 'Logging' that is to log packets at key routers and then use data mining techniques to determine the path that the packets traversed. This scheme has the functional property that it can trace an attack long after the attack has completed. However, it also has obvious drawbacks, including potentially enormous resource requirements (possibly addressed by sampling) and a large scale inter provider database integration problem. The data mining techniques are providing very efficient way for discovering useful knowledge from the available information. There is also a system which uses packet marking mechanisms along with the Intrusion Prevention Systems for efficient IP traceback.Thus,this approach proposes data mining techniques to be applicable to the data collected from the packet marking scheme for detecting attack and therefore, resultant database of knowledge can be further used by network Intrusion prevention systems for decision making.

4. Intrusion Detection Techniques

There are number of intrusion detection techniques, which are used in data mining applications and are as follows:

- Pattern Matching
- Measure Based method
- Data Mining method
- Machine Learning Method

4.1 Pattern Matching

- KMP-Multiple patterns matching Algorithm
- Using keyword tree to search
- Building failure link to guarantee linear time searching
- Shift-And(Or) pattern matching Algorithm

A classical approximate pattern matching algorithm Karp-Rabin fingerprint method using Modular arithmetic and Remainder theorem to match pattern such as regular expression pattern matching.

4.2 Measure Based Method

Statistical methods & information-theoretic measures define a set of measures to measure different aspects of a subject of behavior. Define Pattern generate an overall measure to reflect the abnormality of the behavior.

For example:

Statistic $T2= M12+M22 +\ldots+Mn2$
Weighted intrusion score $= \Sigma Mi*Wi$
Entropy: $H (X|Y) = \Sigma \Sigma P (X|Y) (-\log (P (X|Y)))$

Define the threshold for the overall measure Association Pattern Discover Goal is to derive multi-feature (attribute) correlations from a set of records. An expression of an association pattern example. Association pattern detecting statistics approaches the model of constructing temporal statistical features from discovered pattern. Using measure-based method to detect intrusion pattern matching nobody discuss this idea. Advanced techniques such as machine learning method, time-based inductive machine like Bayes Network and also the use of probability and a direct graph to predict the next event instance based learning are prominent in this context. A distance is used to measure the similarity between featuresd, vectored neural network.

4.3 Classification

This is supervised learning. The class will be predetermined in training phase. Define the character of classes in training phase. A common approach in pattern recognition system

4.4 Clustering

This is unsupervised learning. There are not predetermined classes in data. Given a set of measurement, the aim is that establishes the class or group in the data. It will output the character of each class or group. In the detection phase, this method will get more time cost (o (n2)). This method is suggested only use in pattern discover phase.

4.5 Ideas for Improving Intrusion Detection

1) Idea 1: Association Pattern Detecting

Using the pattern matching algorithm to match the pattern in sequent data for detecting intrusion. No necessary to construct the measure. But its time cost is depending on the number of association patterns. It possible constructs a pattern tree to improve the pattern matching time cost to linear time.

2) Idea 2: Discover Pattern from Rules

The exits rules are the knowledge from experts' knowledge or other system. The different methods will measure different aspects of intrusions. Combine these rules may find other new patterns of unknown attack [8]. For example: Snort has a set of rule which come from different people. The rules may have different aspects of intrusions. Data mining or machine learning method can be used to discover the pattern from these rules.

1) Classical Techniques: Statistics, Neighborhoods and Clustering
2) 1.1. The Classics

These two sections have been broken up based on when the data mining technique was developed and when it became technically mature enough to be used for business, especially for aiding in the optimization of customer relationship management systems. Thus this section contains descriptions of techniques that have classically been used for decades the next section represents techniques that have only been widely used since the early 1980s.This section should help the user to understand the rough differences in the techniques and at least enough information to be dangerous and well armed enough to not be baffled by the vendors of different data mining tools. The main techniques that we will discuss here are the ones that are used 99.9% of the time on existing business problems. There are certainly many other ones as well as proprietary techniques from particular vendors - but in general the industry is converging to those techniques that work consistently and are understandable and explainable.

3) 1.2. Statistics

By strict definition "statistics" or statistical techniques are not data mining. They were being used long before the term data mining was coined to apply to business applications. However, statistical techniques are driven by the data and are used to discover patterns and build predictive models. And from the users perspective you will be faced with a conscious choice when solving a "data mining" problem as to whether you wish to attack it with statistical methods or other data mining techniques. For this reason it is important to have some idea of how statistical techniques work and how they can be applied.

4) What is different between statistics and data mining?

I flew the Boston to Newark shuttle recently and sat next to a professor from one the Boston area Universities. He was going to discuss the drosophila (fruit flies) genetic makeup to a pharmaceutical company in New Jersey. He had compiled the world's largest database on the genetic makeup of the fruit fly and had made it available to other researchers on the internet through Java applications accessing a larger relational database. He explained to me that they not only now were storing the information on the flies but also were doing "data mining" adding as an aside "which seems to be very important these days whatever that is". I mentioned that I had written a book on the subject and he was interested in knowing what the difference was between "data mining" and statistics. There was no easy answer.

The techniques used in data mining, when successful, are successful for precisely the same reasons that statistical techniques are successful (e.g. clean data, a well defined target to predict and good validation to avoid overfitting). And for the most part the techniques are used in the same places for the same types of problems (prediction, classification discovery). In fact some of the techniques that are classical defined as "data mining" such as CART and CHAID arose from statisticians.

So what is the difference? Why aren't we as excited about "statistics" as we are about data mining? There are several reasons. The first is that the classical data mining techniques such as CART, neural networks and nearest neighbor techniques tend to be more robust to both messier real world data and also more robust to being used by less expert users. But that is not the only reason. The other reason is that the time is right. Because of the use of computers for closed loop business data storage and generation there now exists large quantities of data that is available to users. IF there were no data - there would be no interest in mining it. Likewise the fact that computer hardware has dramatically upped the ante by several orders of magnitude in storing and processing the data makes some of the most powerful data mining techniques feasible today. The bottom line though, from an academic standpoint at least, is that there is little practical difference between a statistical technique and a classical data mining technique. Hence we have included a description of some of the most useful in this section.

5) What is statistics?

Statistics is a branch of mathematics concerning the collection and the description of data. Usually statistics is considered to be one of those scary topics in college right up there with chemistry and physics. However, statistics is probably a much friendlier branch of mathematics because it really can be used every day. Statistics was in fact born from very humble beginnings of real world problems from business, biology, and gambling! Knowing statistics in your everyday life will help the average business person make better decisions by allowing them to figure out risk and uncertainty when all the facts either aren't known or can't be collected. Even with all the data stored in the largest of data warehouses business decisions still just become more informed guesses. The more and better the data and the better the understanding of statistics the better the decision that can be made.

Statistics has been around for a long time easily a century and arguably many centuries when the ideas of probability began to gel. It could even be argued that the data collected by the ancient Egyptians, Babylonians, and Greeks were all statistics long before the field was officially recognized. Today data mining has been defined independently of statistics though "mining data" for patterns and predictions is really what statistics is all about. Some of the techniques that are classified under data mining such as CHAID and CART really grew out of the statistical profession more than anywhere else, and the basic ideas of probability, independence and causality and overfitting are the foundation on which both data mining and statistics are built.

6) Data, counting and probability

One thing that is always true about statistics is that there is always data involved, and usually enough data so that the average person cannot keep track of all the data in their heads. This is certainly more true today than it was when the basic ideas of probability and statistics were being formulated and refined early this century. Today people have to deal with up to terabytes of data and have to make sense of it and glean the important patterns from it. Statistics can help greatly in this process by helping to answer several important questions about your data:

5. Proposed Methodology for Working

Data mining is becoming a persistent technology in activities as diverse as using historical data to predict the success of a marketing campaign, looking for patterns in network traffic to discover illegal activities or analyzing sequences [Sundari and Thangadura, 2010]. From this outlook, the approach is gaining importance in the field of DDoS attacks. Data mining is, at its core, pattern finding. Data miners are proficient at using specialized software to find regularity (and irregularities) [6] in large & complex data sets. Data mining applications are computer software programs or packages that enable the extraction and identification of patterns from stored data. A data mining application is typically a software interface which interacts with a large database containing Network traffic parameters or other important data. Database technology and artificial intelligence technology to promote the organic integration of knowledge discovery in databases (KDD) [7] technology generation. Knowledge discovery in databases, also known as data mining, is from a large database or data warehouse extraction of implicit, unknown, a potential value of the information or model approach.

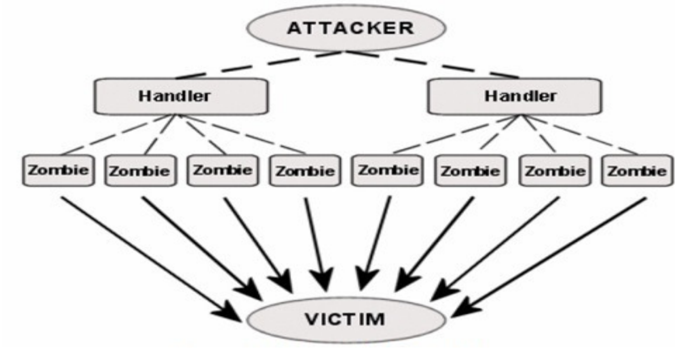

Figure 2. DDoS attack system

Data mining has the advantage of the ability to handle large amounts of data associated with data analysis capabilities. Network database [7, 8] using this technology to achieve security at home and abroad is a new attempt. Data mining technology used in intrusion detection system based on a database, you can audit the data from a lot of knowledge that can help detect and rules, so as to effectively prevent the intrusion of emerging. Classification algorithm used to construct the invasion of the main features, and by association rules from frequent episodes of the series of the extracted pattern mining properties [8, 9] in an easy people to understand the characteristics of heuristic rules to describe the attack and build the corresponding classifier, and finally classified using the trained To execute the corresponding function of misuse detection. Purpose of classification is to use the database features of the data item attributes [9], generate the classification model or classification function, the model can put the items in the database map to a given category in a classification.

6. Conclusion

DDoS attacks are quite complex methods of attacking a Computer network, ISP, individual system makes it ineffectual to legitimate network users. These attacks are an aggravation at a minimum, and if they are against a particular system, they can be brutally destroying. Loss of network resources costs money, delays work, and interrupts Communication between various legal network users. The drastic consequences of a DDoS attack make it important that strict and productive solutions and security measures must be made to prevent these types of attacks. Detecting, preventing, and mitigating DDoS attacks is important for national and individual security. This paper discussed various detection algorithms which are using data mining concepts & algorithms for DDoS detection & prevention. But with the improvement in technology new areas are emerging where data mining techniques can be utilized for handling DDoS attacks.

References

[1] Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[3] K. Elissa, “Title of paper if known,” unpublished.

[4] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[5] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[6] M. Young,Sager The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[7] Stone, Electronic Publication: Digital Object Identifiers (DOIs):

[8] D. Kornack and P. Rakic, “Cell Proliferation without Neurogenesis in Adult Primate Neocortex,” Science, vol. 294, Dec. 2001, pp. 2127-2130, doi:10.1126/science.1065467.

[9] H. Goto, Y. Hasegawa, and M. Tanaka, “Efficient Scheduling Focusing on the Duality of MPL Representatives,” Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS 07), IEEE Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670.