Survey on Data Mining Techniques in Intrusion Detection

Amanpreet Chauhan, Gaurav Mishra, Gulshan Kumar

Abstract-Intrusion Detection (ID) is the main research area in field of network security. It involves the monitoring of the events occurring in a computer system and its network. Data mining is one of the technologies applied to ID to invent a new pattern from the massive network data as well as to reduce the strain of the manual compilations of the intrusion and normal behavior patterns. Keeping in mind, data mining techniques are practiced significantly intrusion detection and prevention. This article reviews the current state of art Data mining techniques with ID in brief and highlights its advantages and disadvantages.

Keywords: - Network Intrusion, Decision Trees, Naïve Bayes, Fuzzy Logic, Support Vector Machines, Data Clustering, Data Mining.

I. INTRODUCTION

The internet has become a part of daily life and an essential tool today. It aids people in many areas, such as business, entertainment, and education etc. Most traditional communications media including telephone, music, film, and television are being reshaped or redefined by the Internet [1]. Newspaper, book and other print publishing have to adapt to Web sites and blogging. The Internet has enabled or accelerated new forms of human interactions through instant messaging, Internet forums, and social networking. Online shopping has boomed both for major retail outlets and small artisans and traders. Business-to-business and financial services on the Internet affect supply chains across entire industries. In particular, internet has been used as an important component of business models. For the business operation, both business and customers apply the internet application such as website and e-mail on business activities. Therefore, information security of using internet as the media needs to be carefully concerned [2].

II. INTRUSION AND INTRUSION DETECTION

Intrusion, in simple words, is an illegal act of entering, seizing, or taking possession of another's property (in this case the property being the computer system). It means a code that disables the proper flowing of traffic on the network or steals the information from the traffic [3].

- *Amanpreet Chauhan, Student B.tech(C.S.E.) Malout institute of management & Information Technology, Malout (Pb.) INDIA.*
  PH: - +91-8699609173, ciseapc@gmail.com
- *Gaurav Mishra, Student B.tech(C.S.E.) Malout institute of management & Information Technology, Malout (Pb.) INDIA.*
  PH: - +91-9958374852, mishra2k7@gmail.com
- *Gulshan Kumar, Assistant Professor C.S.E. deptt. Malout institute of management & Information Technology, Malout (Pb.) INDIA.*
- *PH: - 8146550540, gulshanahuja@gmail.com*

The other most common names for intrusion are virus, Trojan etc. The various divisions of intrusions can be listed as follows:-

DoS Attack - A denial-of-service attack (DoS attack) or distributed denial-of-service attack (DDoS attack) is an attempt to make a computer resource unavailable to its intended users. Although the means to carry out, motives for, and targets of a DoS attack may vary, it generally consists of the concerted efforts of a person or people to prevent an Internet site or service from functioning efficiently or at all, temporarily or indefinitely. Perpetrators of DoS attacks typically target sites or services hosted on high-profile web servers such as banks, credit card payment gateways, and even root name servers.

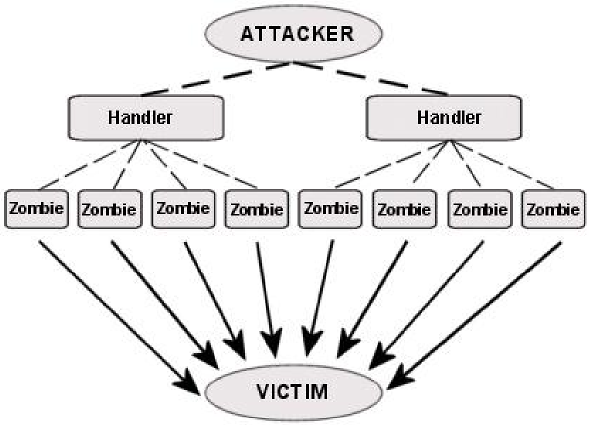

Figure 1 DDoS attack scenario

Remote to User (R2L) – This kind of attack describes the unauthorized access from a remote machine into the super user (root) account of the target system. It is a class of attack where an attacker sends packets to a machine over a network, then exploits the machine's vulnerability to illegally gain local access as a user. There are different types of R2U attacks; where the most common attack in this category is the art of social engineering.

User to Root (U2R) – User to root attack defines the unauthorized access to local super user (root). These exploits are classes of attacks which an attacker starts out with access to a normal user account on the host system and is able to exploit vulnerability to gain root access to the system. Most common exploits in this class of attacks are regular buffer overflows, which are caused by regular programming mistakes and environment assumptions.

Probing – Probing is a class of attack where an attacker scans a network to gather information or find vulnerabilities. An attacker with a map of machines and services that are available on a network can use the information to look for exploits. There are different types of probes: some of them abuse the computer's legitimate features; some of them use social engineering techniques. This class of attack is the most common and requires very little technical expertise.

In information security, intrusion detection is the act of detecting actions that attempt to compromise the confidentiality, integrity or availability of a resource. The recent advances of computer communication infrastructure realizes the era of computer-based data processing. The need for intrusion detection arises from the fact that computer systems are used in each and every aspect of this age of mainstream science and technology [4]. There broadly exist the following intrusion detection methodologies:-

1. **Anomaly Detection -** it refers to detecting patterns in a given data set that do not conform to an established normal behavior. The patterns thus detected are called anomalies and often translate to critical and actionable information in several application domains. Anomalies are also referred to as outliers, surprise, aberrant, deviation, peculiarity, etc. It actually refers to storing features of user's usual behaviors into database, then comparing user's current behavior with those in the database. If there occurs a divergence huge enough, it is said that the data tested is abnormal. The advantage of anomaly detection lies in its complete irrelevance of the system, its strong versatility and the possibility to detect the attack that was never detected before. Anomaly-based intrusion detection is about discrimination of malicious and legitimate patterns of activities (system or user driven) in variables characterizing system normality [5]. But due to the fact that normal contour conducted cannot give a complete description of all user's behaviors in the system, moreover each user's behavior changes constantly, its main drawback is the high rate of false alarm (a failure of an IDS to detect an actual attack).
2. **Misuse Detection -** In misuse detection approach, we define abnormal system behavior at first, and then define any other behaviour, as normal behavior. It assumes that abnormal behavior and activity has a simple to define model. Its advantage is simplicity of adding known attacks to the model. Its disadvantage is its inability to recognize unknown attacks. Misuse Detection refers to confirming attack incidents by matching features through the attacking feature library. It advances in the high speed of detection and low percentage of false alarm. However, it fails in discovering the non-pre-designated attacks in the feature library, so it cannot detect the numerous new attacks.

III. DATA MINING

Data mining is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. A primary reason for using data mining is to assist in the analysis of collections of observations of behavior. An unavoidable fact of data mining is that the (sub-) set(s) of data being analyzed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships and behaviors that exist across other parts of the domain [6].

Data mining technology is advanced for:

- It can process large amount of data
- It can discover the hidden and ignored information

Data mining commonly involves four classes of tasks:-

1. **Clustering** – it is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
2. **Classification** – it is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, Naive Bayesian classification, neural networks and support vector machines.
3. **Regression -** Attempts to find a function which models the data with the least error.
4. **Association rule learning -** Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.

IV CLASSIFICATION TECHNIQUES

The various classification techniques in this study are as follows:-

1. Decision trees: A decision tree performs the classification of a given data sample through various levels of decisions to help reach a final decision [7]. Such a sequence of decisions is represented in a tree structure. The classification of a sample proceeds from the root node to a suitable end leaf node, where each end leaf node represents a classification category. A decision tree with a range of discrete (symbolic) class labels is called a classification tree, whereas a decision tree with a range of continuous (numeric) values is called a regression tree. CART (Classification and Regressing Tree) is a well-known program used in the designing of decision trees (Breiman, Friedman, Olshen &

Stone, 1984). Decision trees make use of the ID3 algorithm and the J48 algorithm which is an enhanced version of C4.5.

1.1 ID3 Algorithm **-** it is considered to be a very useful Inductive Logic Programming method developed. ID3 is an attribute based machine-learning algorithm that constructs a decision tree which is said to be based on a given training set data. The formulation of ID3 is: (i) initially the attribute is determined which has the highest information gain on the training set. (ii) His attribute is used as the root of the tree, an ranches are created for each of the values that it accepts. (iii) This process is repeated for each of the branches with the subset of the training set.

1.2 J48 Algorithm – the J48 algorithm is based on the algorithm designed with features which easily address the loopholes that were earlier present in ID3. This algorithm was primarily designed as the enhanced version of C4.5 as the principal disadvantage of C4.5 was the amount of CPU time it took and the system memory it required.

Now if we consider a set A of case objects, J48 initially grows a tree and uses divide-and-conquer algorithm as follows: (i) if all the cases in A belong to the same class or if the set is a small one, the tree is leaf labeled with the most frequent occurring class in A. (ii) or, a test is selected based on a single attribute with two or more outcomes. This test is made the root of the tree with each branch as one outcome of the test. Further the same procedure is applied recursively for each subset.

2. Support Vector Machines: SVM first maps the input vector into a higher dimensional feature space and then obtain the optimal separating hyper-plane in the higher dimensional feature space. an SVM classifier is designed for binary classification. The generalization in this approach usually depends on the geometrical characteristics of the given training data, and not on the specifications of the input space. This procedure transforms the training data into a feature space of a huge dimension [8]. That is, to separate a set of training vectors which belong to two different classes [9].

3. Fuzzy logic: It processes the input data from the network and describes measures that are significant to the anomaly detection [8]. Fuzzy logic (or fuzzy set theory) is based on the concept of the fuzzy phenomenon to occur frequently in real world. Fuzzy set theory considers the set membership values for reasoning and the values range between 0 and 1. That is, in fuzzy logic the degree of truth of a statement can range between 0 and 1 and it is not constrained to the two truth values (i.e. true, false) [10].

A fuzzy system comprises of a group of linguistic statements based on expert knowledge. This knowledge is usually in the form of if-then rules. A case or an object can be distinguished by applying a set of fuzzy logic rules based on the attributes' linguistic value [11].

4. Naïve Bayes: It can be easily considered the upgraded version of the Bayes Theorem due to the fact that it assumes independence of attributes [12].There are many cases where we know the statistical dependencies or the causal relationships between system variables. However, it might be difficult to precisely express the probabilistic relationships among these variables. In other words, the prior knowledge about the system is simply that some variable might

Influence others. To exploit this structural relationship or casual dependencies between the random variables of a problem, one can use a probabilistic graph model called Naïve Bayesian Networks (NB). The model provides an answer to questions like ‘‘What is the probability that it is a certain type of attack, given some observed system events?” by using conditional probability formula. The structure of a NB is typically represented by a directed acyclic graph (DAG), where each node represents one of system variables and each link encodes the influence of one node upon another (Pearl, 1988). Thus, if there is a link from node A to node B, A directly influences B.

V. CLUSTERING TECHNIQUES

Clustering basically means that we have to make the group (clusters) from our data so that we can easily find our required data or we can say that it is a classification of similar objects into different groups and the partitioning of dataset into subsets or clusters so that data in each subset share some common trait. It is usually applied in the statistical data analysis which can be made use of in many fields, for instance, machine learning, data mining, pattern recognition, image analysis and bioinformatics (Yusufovna, 2008). Clustering is useful in intrusion detection as malicious activity should cluster together, separating itself from non-malicious activity. Frank (1994) divides clustering techniques into the following areas: hierarchical, statistical, exemplar, distance and conceptual clustering, each of which has their own different ways of obtaining of cluster membership and representation. Clustering is an effective way to find hidden patterns in data that humans might otherwise miss. Clustering provides some significant advantages. It is advantageous over the classification techniques as it does not require any use of labeled dataset for training.

Generally the pattern clustering activity involves the following steps:-

(i) Representation of the pattern
(ii) Definition of a pattern proximity
(iii) Clustering

VI CONCLUSION

Data mining is the modern technique for network intrusion detection. Ready-made data mining algorithms are available. .large amount of data can be handled with the data mining technology. It is still in developing state, it can become more effective as it is growing rapidly. Our main task is to get the more correct rate of intrusion detection, to reduce the rate of false negatives. As the data mining is still is developing state so more study and research is to be done.

REFERENCES

[1] Alexander D. Korzyk. A Forecasting Model For Internet Security Attacks.

[2] Simon Hansman and Ray Hunt (2004). A Taxonomy of Network and Computer Attacks

[3] Mrityunjaya Panda and Manas Ranjan Patra. A Comparative Study of Data Mining Algorithms for Intrusion Detection.

[4] Eric Knight (2000). Computer Vulnerabilities

[5] Jose F. Nieves (2009). Data Clustering for Anomaly Detection in Network Intrusion Detection.

[6] Ian H. Witten and Eibe Frank. Data Mining : Practical Machine Learning Tools and Techniques.

[7] Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin, Wei-Yang Lin (2009). Intrusion detection by Machine Learning : A Review

[8] Srinivas Mukkamala, Andrew H. Sung, Ajith Abraham (2004). Intrusion Detection Using an Ensemble of Intelligent Paradigms
.

[9] V. Vapnik (1998). Statistical Learning Theory. New York: John Wiley

[10] H. Zimmerman (2001). Fuzzy Set Theory and Its Applications. Kluwer Academic Publishers.

[11] Ajith Abraham and Ravi Jain. Soft Computing Models for Network Intrusion Detection Systems.

[12] Mrutyunjaya Panda, Manas Ranjan Patra. A Comparative Study of Data Mining Algorithms for Intrusion Detection.