Hybrid Intrusion Detection SystemBased on Bayesian Network

Khin Khattar Myint, Nang Saing Moon Kham
University of computer Studies, Yangon
khinnkhattarrmyint@gmail.com

Abstract

Now day's security is the primary concerned in the field of computer science.With quickly growing unauthorized activities in network Intrusion Detection as a part of defense is extremely necessary because traditional firewall techniques cannot provide complete protection against intrusion.The primary goal of an Intrusion Detection System (IDS) is to identify intruders and differentiate anomalous network activity from normal one. Intrusion detection has become a significant component of network security administration due to the enormous number of attacks persistently threaten our computer networks and systems.This paper illustrates the benefit of hybrid intrusion detection system that can detect both known and unknown attacks. The system includes two phases: (1) If the attack is known attack then signature intrusion detection handles and performs appropriate action. (2)If the attack is unknown attack then anomaly intrusion detection use frequent pattern matching process and generate the signature that can handle the next attack. Our proposed system may be more accurate and better performance than traditional intrusion detection system.

Keywords: Intrusion Detection System(IDS),Bayesian Network, Naïve Bayes, Frequent pattern mining**.**

1. Introduction

Nowadays, with the network development, secured information communication has becoming more vulnerable because of threats from unknown sources and therefore the requirement for secured information assumes greater importance. Attacks on network infrastructure presently are main threats against network and information security. With rapid growth of unauthorized activities in network, Intrusion Detection (ID) as a component of defense is very necessary because traditional firewall techniques cannot provide complete protection against intrusion.
IDSs are split into two primary types of groups:

- host-based (HIDS)
- network-based (NIDS)

A HIDS resides on the particular host and search for signs of attacks with that host. An NIDS resides on the separate system that watch network traffic, searching for signs of attacks that traverse in the area of the network. The present trend in ID would be to mix both host based and network based information to build up hybrid systems.The techniques for the intrusion detection can be divided into two categories: Anomaly Intrusion Detection and Misuse Intrusion Detection. They were classified based on approaches like Statistics, Data mining, Neural Network Based and Self Organizing Maps Based approaches etc.[2]

Our proposed system has many advantages:It will be adaptive in nature and adapts the changes in user and system behavior. It will run constantly with minimal human supervision. It will create signatures of new attacks. Design of our hybrid IDS makes it fault tolerant, so that it will be able to recover from crashes. It will be able to get its prior state and resume its operation without any adverse effect. It will be able to monitor itself and detect attacks on it. It will consume less memory to operate. It will be accurate and thereby there will be less number of false positives and false negatives.

The remainder of this paper is organized as follows. The next section presents the related work. Section 3 describesbackground theory of our systemand Section 4 describes about architecture of our proposed system. Section 5explains experimental result. Finally, we conclude the paper in Section 6.

2. Related Work

Kayvan proposed the hybrid intrusion detection system using SVM with GA. He proved that enhancing the SVM with GA can reduce false alarmsand mean square error (MSE) in detecting intrusion[2].Kumar proposed a security model by

combining network based and host based with efficient data mining approach to detect any type of intrusion which coming from public network or occurring in computer system[8].Aditi and Hitesh proposed a hybrid learning approach that combines K-Mean clustering, Naive Bayes (statistical) also known as KMNB with Decision Table Majority (rule based) approaches [1].

3. Background Theory

3.1. Signature-based intrusion detection system

The misuse (signature-based) detection is normally used for detecting known attacks. It requires that all known threats will be defined first, and the information regarding these threats to be submitted to the IDS. Thus, the IDS is able to then compare all incoming or outgoing activity against all known threats in its knowledge base and raise an alarm if any activity matches information in the knowledge base. The information stored in this knowledge base is usually known as signatures [3]. The process for actually comparing a signature with an attack include simple string matching – which involves looking for unique key words in network traffic to identify attacks – to more complex approaches such as rule -based matching which defines the behavior of an attack as a signature. Various string-matching (or pattern-matching) algorithms are used to inspect the content of packets and identify the attacks signature in IDS. There are mainly two kinds of algorithms, viz. i) Single-keyword pattern matching algorithms viz., Brute force algorithm, Knuth-Morris-Pratt algorithm, and Boyer-Moore algorithm; and (ii) Multiple-keyword pattern matching algorithms viz., Aho-Corasick, Wu-Manber Algorithm, Horspool Algorithm, Quick search algorithm, Piranha, and E2xb. Following are the advantages of misuse detection technique: (1) Signatures are very easy to develop and understand, if we know what network behavior we are trying to identify. For instance, we might use a signature that looks for particular strings within exploit particular buffer overflow vulnerability. The events generated by signature-based IDS (SIDS) can communicate the cause of the alert. (2) It has a relatively low rate of false alarms, which means that the SIDS has a relatively high precision. This high precision is caused by the fact that a SIDS is explicitly programmed to detect certain known kinds of attacks .One big challenge of SIDS is that every signature requires an entry in the database, and so a complete database might contain hundreds or even thousands of entries. Each packet is to be compared with all the entries in the database. This can be very resource consuming and doing so will slow down the throughput and making the IDS vulnerable to DoS attacks. Some of the IDS evasion tools use this vulnerability and flood the IDS systems with too many packets to the point that the IDS cannotkeep up with the traffic, thus making the IDS time out and drop packets and as a result, possibly miss attacks [3]. On modern systems, string-matching can be done more efficiently, so the amount of power needed to perform this matching is minimal for a rule set. For example, if the system that is to be protected only communicates via DNS, ICMP and SMTP, all other signatures can be ignored.Following are the disadvantages of signature-based intrusiondetectionsystem (SIDS): (i) The detection rate of attacks is relatively low. (ii) There is a lower recall for new types of intrusions. (iii) An attacker will try to modify a basic attack in such a way that it will not match the known signatures of that attack. The attacker may insert malformed packets that the IDS will see, to intentionally cause a pattern mismatch; the protocol handler stack will then discard the packets because of the malformation. Each of these variations could be detected by IDS, but more different signatures require additional work for the IDS, which reduceperformance. (i) It cannot detect a new attack for which a signature is not yet installed in the database. Ideally, signatures should match every instance of an attack, match subtle variations of the attack, but not match traffic that is not part of an attack. However, this goal is difficult to accomplish in current IDSs. (ii) The efficiency of the SIDS is greatly decreased, as it has to create a new signature for every variation. As the signatures keep on increasing, the system performance deteriorates. Due to this, many SIDS are deployed on systems with multi processors and multi Gigabit network cards. IDS developers develop the new signatures before the attacker does, so as to prevent the novel attacks on the system. The difference of speed of creation of the new signatures between the developers and attackers determine the efficiency of the system[3].

3.2. Anomaly-based intrusion detection system

The anomaly (heuristic-based) detection is based on defining the network behavior. Instead of looking

for matches, anomaly intrusion detection looks for behavior that is suspicious [3]. Anomaly-based IDS attempt to characterize normal operation, and try to detect any deviation from normal behavior. The network behavior is in accordance with the predefined behavior, then it is accepted or else it triggers the event in the anomaly detection. The accepted network behavior is prepared or learned by the specifications of the network administrators. It builds a model of acceptable behavior and flag (i.e. label) exceptions to that model; for the future, the administrator can mark a flagged behavior as acceptable so that the anomaly-based IDS will now treat that previously unclassified behavior as acceptable.An anomaly-based detection technique compares normal behavior against the current pattern of behavior in a system. In order to achieve this task, the main challenge in anomaly detection technique is in learning what is considered ―normal behavior. The work by Axelsson describes the two main approaches which are used to achieve this goal: self-learning or programmed anomaly detection. In the self-learning approach, the anomaly detection system will begin to automatically monitor events, such as live network traffic, on the environment it has been implemented on and attempt to build information on what is considered normal behavior [3]. This is otherwise known as online learning. In the programmed approach, the anomaly-based IDS must manually learn what is considered normal behavior by having a user or some form of function ―teaching the system through input of information. This is otherwise known as offline learning, and may involve feeding the system a network traffic data set which contains normal network traffic. The major advantage of anomaly detection over misuse technique is that a novel attack for which a signature does not exist can be detected if it falls out of the normal traffic patterns. This is observed when the systems detect new automated worms. If the new system is infected with a worm, it usually starts scanning for other vulnerable systems at an accelerated rate filling the network with malicious traffic, thus causing the event of a TCP connection or bandwidth abnormality rule. Following are the disadvantages of anomaly intrusion detection: (i) there is a higher rate of false alarms, which means a lower precision. (ii) It also needs periodic online retraining of the behavior profile. (iii) It tends to be computationally expensive because several metrics are often maintained that need to be updated against every system activity and, due to insufficient data, they may gradually be trained incorrectly to recognize an intrusive behavior as normal due to insufficient data [3].

4. Proposed System

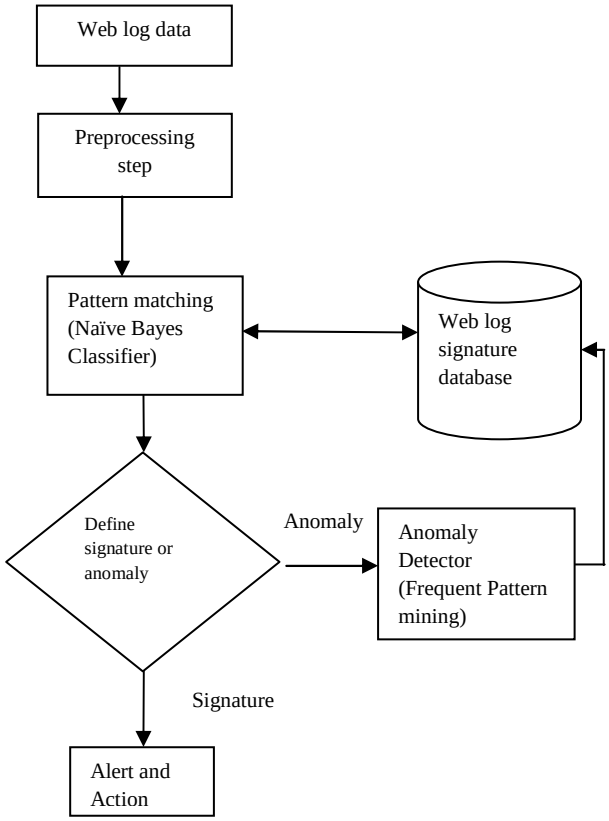

Figure1. Proposed system of hybrid IDS

In our System, web log data are entered as input to the system. Such data are preprocessed by following steps: (1)reduce and clean the noise and unimportant data, (2) identify individual user by using their IP address,(3) define the session of particular user. In our proposed hybrid-intrusion detection, it can detect both known and unknown attacks. The system used naïve bayes classifier to analyze the attack by using web log features. If the attack is known, the system displays the alert and performs the appropriate action. If the attack is unknown, the system extract rules by processing underlying feature log. And the system stores these rules to the web- log training database. So, the system can detect when the similar log are found next time.

4.1. Data Preprocessing

It is important to understand that the quality data is a key issue when we are going to mining from it. Nearly 80% of mining efforts often spend to improve the quality of data[4]. The data which is obtained from the logs may be incomplete, noisy and inconsistent. The attributes that we can look for in quality data includes accuracy, completeness,consistency, timeliness, believability, interpretability and accessibility. There is a need to preprocess data to make it have the above mentioned attributes and to make it easier to mine for knowledge. There are three steps in preprocessing of log data: data cleaning, user identification, session identification.

4.1.1. Data cleaning

The process of data cleaning is removal of outliers or irrelevant data. The Web Log file is in text format then it is required to convert the file in database format and then clean the file. First, all the fields which are not required are removed and finally we will have the fields like date, time, client IP, URL access, Referrer and Browser used/ Access log files consist of large amounts of HTTP server information. Analyzing, this information is very slow and inefficient without an initial cleaning task. Every time a web browser downloads a HTML document on the internet, the images are also downloaded and stored in the log file. This is because though a user does not explicitly request graphics that are on a web page, they are automatically downloaded due to HTML tags. The process of data cleaning is to remove irrelevant data. All log entries with file name suffixes such as gif, JPEG, jpeg, GIF, jpg, JPG can be eliminated since they are irrelevant[4].

4.1.2. User Identification

This step identify individual user by using their IP address. If new IP address, there is new user. If IP address is same but browser version or operating system is different then it represents different user. User identification an important issue is how exactly the users have to be distinguished. It depends mainly on the task for the mining process is executed. In certain cases the users are identified only with their IP addresses.User's identification is, to identify who access Web site and which pages are accessed. If users have login of their information, it is easy to identify them. In fact, there are lots of user do not register their information. There are great numbers of users access Web sites through, agent, several users use the same computer, firewall's existence, one user use different browsers, and so forth. All of problems make this task greatly complicated and very difficult, to identify every unique user accurately. We may use cookies to track users' behaviors. But considering personage privacy, many users do not use cookies, so it is necessary to find other methods to solve this problem. For users who use the same computer or use the same agent, it uses heuristic method to solvethe problem, which is to test if a page is requested that is not directly reachable by a hyperlink from any of the pages visited by the user, the heuristic assumes that there is another user with the same computer or with the same IP address. Some navigation patterns are not accurate because they only consider a few aspects that influence the process of users' identification. The success of the web site cannot be measured only by hits and page views. Unfortunately, web site designers and web log analyzers do not usually cooperate. This causes problems such as identification unique user's session, construction discrete user's sessions and collection essential web pages for analysis. The result of this is that many web log mining tools have been developed and widely exploited to solve these problems.

4.1.3. Session Identification

To group the activities of a single user from the web log files is called a session. As long as user is connected to the website, it is called the session of that particular user. Most of the time, 30 minutes time-out was taken as a default session time-out. A session is a set of page references from one source site during one logical period. Historically a session would beidentified by a user logging into a computer, performing work and then logging off. The login and logoff represent the logical start and end of the session [4].The final stepconverts the data into the format needed by the mining algorithms. If the sessions and the sequences are identified, this step can be accomplished more easily.

4.2Pattern Matching

4.2.1. Naïve Bayesian Learning

Naive Bayes is also an attractive approach in the text classification task because it is simple enough to be practically implemented even with a great number of features. This simplicity enables us to integrate the text classification and filtering modules with the existing information retrieval systems easily. Bayesian text classification uses a parametric mixture model to model the generation of documents. The model has the following form:

$$\mathrm{p(d)} = \sum_{j=1}^{|C|} \mathrm{p(c_j)p(d|c_j)} \quad (1)$$

where c_j are the mixture components (that correspond to the possible classes) and $p(c_j)$ are prior probabilities. Using Bayes' rule, the model can be inverted to get the posterior probability that d was generated by the mixture component $c_{j:}$

$$\mathrm{p(c_j}|d) = \frac{p(\mathrm{c_j})p(d|\mathrm{c_j})}{p(d)} \quad (2)$$

The prior probabilities p (c_j) is estimated from a training corpus by counting the number of training data in each class c_j. The distribution of data in each class, p ($d|c_j$), cannot be estimated directly. Rather, it is assumed that data are composed from smaller units. To make the estimation of parameters tractable, we make the Naive Bayes assumption. For reducing the problem the assumption is made that all features (w_t) of a vector are pair wise different, thus they are statistical independent. This assumption is revealed by the following formula

$$p(d|\mathrm{c_j}) = \prod_{t=1}^{|V|} p(\mathrm{w_t}|\mathrm{c_j}) \quad (3)$$

The naïve character is due to the fact that usually this assumption is not verified in practice.

4.2.2. Information Gain

In the high-dimensional vector spaces only these features should be used that are descriptive more general spoken for a category. One solution for this task is to compute the Information gain (IG) for each unique feature. After ranking thefeatures so that the highest is at the first position we can remove features that are below a predefined threshold. IG measures the number of bits of information obtained for category prediction. It is frequently employed as a feature goodness criterion in the field of machine learning. The information gain offeature t is defined as:

$$IG(t) = -\sum_{i=1}^{m} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{m} p(c_i|\mathrm{t})logp(c_i|\mathrm{t}) + p(\bar{t})\sum_{i=1}^{m} p(c_i|\bar{t})logp(c_i|\bar{t}) \quad (4)$$

is the set of categories in the target space and P (c_i) is the probability of the category c_i. P (t) is the probability that t occurs in the collection, $p(c_i|\mathrm{t})$ is the probability that a category is c_i, given the feature t appears, and$p(c_i|\bar{t})$, is the probability that a category is c_i, given the feature t does not appear.

4.2.3. Exampleto classify Normal and Anomaly using Naïve Bayes Classifier

We assume feature f1 as duration, f2 as protocol-type, f3 as service, f4 as flag, etc..

Features=(f_1=0, f_2=udp, f_3=ftp-data,…,f_n=0.0)

P (Normal/ Features) =

$$\frac{\mathrm{P(Features|Normal)P(Normal)}}{\mathrm{P(Features)}}$$

Bestcategory=

$$\underset{Normal}{ArgMax}\frac{\mathrm{P(Features|Normal)P(Normal)}}{\mathrm{P(Features)}}$$

Bestcategory=

$$\underset{Normal}{ArgMax}\mathrm{P(Features|Normal)P(Normal)}$$

Here, f_1, f_2, f_3… f_n are the features in the web log data.

Bestcategory=

$$\underset{Normal}{ArgMax}\mathrm{P}(f_1|\mathrm{Normal}) * \mathrm{P}(f_2|\mathrm{Normal}) * \ldots . * \mathrm{P}(f_N|\mathrm{Normal}) * \mathrm{P(Normal)}$$

In this step, the classifier classifies 2-class (normal and anomaly) by using the training web log database. And then the next step is to analyze if the class is anomaly class, the system uses the frequent pattern mining to extract rules that satisfy the next incoming new web log.

4.3Anomaly Detector using Frequent Pattern Mining

In this step, if the unknown attack is not contained in training web log database, the system cannot handle this unknown attack. So, the system use the frequent pattern mining to train this unknown attack as known attack appropriated for next time. In this work, patterns are discovered by applying the frequent pattern mining methods i.e. CBFP mining algorithmon the log data. For this reason the log data have to be converted in the preprocessing phase such that the output of the conversion can be used as the input of the algorithm. Pattern analysis means understanding the results obtained by the algorithmsand drawing conclusions. In pattern analysis, local proxy serverenables our technique to analyze information retrieved is relevant to user or not, using feedback for each accessed article.After the discovery has been achieved, the analysis of the patterns follows. The whole mining process is an illustrative task which is depicted by the feedback. Depending on the results of the analysis either the parameters of the preprocessing step can be tuned (i.e. by choosing another time interval to determine the sessions of the users) or only the parameters of the mining algorithms (In this case that means the minimum support threshold). In this work the aim of mining is to discover the generalized templates from pages frequently visited at the same time, and to discover the page for corresponding keywords. The results obtained is used to allow users to view all templates (both simple and generalized) or generalized templates only, view templates sorted by different ways, and display templates in different level of details. The discovery of the templates can provide insights to the web editors as to what topics users are mostly interested in. When incorporated with the regular search engine, those templates can improve the search speed.

CBFP mining algorithm:The algorithm uses the basic idea and techniques of Apriorialgorithm and FP treemethod. Algorithm employs level-wise and explore pattern based on downward traversal.The log obtained from preprocessing phase is given to CBFP mining algorithmfor constructing consensus tree. In the FP treetechnique, each frequent pattern obtained from the leave of tree. Each piece of user log information is meant for constructing or updating consensus tree. The path of the tree, from root to some leaf represents information about the article, used for level constraint and rule constraint. Each node of the tree represents the classification of all articles, which help in updating thearticle. A node can have path to different other node of the level next toit. A single path covers too many queries correctly since theirarticle and keywords posted by user is different, hence accuracy of accessing article based on the number of keywords and their meaning results in better accuracy. A constraint on particular node is applicableto its entire descending node, so tree also has downward

Closure property. Level-wise search strategy used for finding frequent articles and corresponding keywords. FP treelike TRIE structure (called consensus tree) for shared representation obtaining generalized templates and level-wise search strategy for predicting the future user's interestingness of an article.The consensusgeneralize one-keyword queries pairing, since it would inevitably produce an overly generalized template The level-wise search of article along with their keywords starts from the root of the tree. Each node in level-1 of the treerepresents the different type of search engine (we can have another level to represents the time-zone), and level-2 shows the number of keywords used in a query. Each leaf contains pointer to all articles belonged to it. Each leaf is organized and contains pointer to the bucket where the article resides. Each leaf may split and combine based on extendible hashing technique. Extendiblehashing technique is used for storing article, each article being connected to a link list. The link list contains keyword ofuserquery, corresponding to user click through on that article. Depth of consensus treeis same for all nodes. The growth of theconsensus treeis constrained using level-wise and rule constraint. CBFP mining algorithm is as a reasonable time complexity and space complexity.

5. Experimental Result

We examine the accuracy of our system by using Precision, Recall and F-score. Precision and recall are more suitable in such system because they measure how precise and how complete the classification is on the positive class. The following table describes the confusion matrix of a naïve bayes classifier. The classifier uses the useful data after preprocessing step. So the system's accuracy is higher than other non preprocessing system.

	Classified positive	Classified negative
Actual positive	TP	FN
Actual negative	FP	TN

Where
TP: the number of correct classifications of the positive examples(true positive)
FN: the number of incorrect classifications of the positive examples (false negative)
FP: the number of correct classifications of the negative examples (false positive)
TN: the number of correct classifications of the negative examples(true negative)

$$p = \frac{TP}{TP + FP} \quad (5)$$

$$r = \frac{TP}{TP + FN} \quad (6)$$

$$F = \frac{2pr}{p + r} \quad (7)$$

Where, p is precision and r is recall for these equation(5) and (6). We use these equations to evaluate the performance of our proposed system.

Now, I can't show detail experimental result of the system. Because this is my ongoing research.

6. Conclusion

Traditional Network IDS suffer from different problems that limit their detection effectiveness and efficiency.In this paper, we have been discussed signature-based and anomaly-based approaches for network intrusion detection. We suggested that a combination of both approaches may overcome the limitations in current Network IDS and leads to high performance including the intrusion detection accuracy by reducing the false positives.And it also deals with the problem of detecting both known and unknown attacks from large amount of Web log data collected by web servers. The contribution of the paper is to introduce the process of web log mining, and to show how to reduce irrelevant raw data to get more accurate classification. The result may have high Accuracy to detect intrusion while using hybrid model rather than primary algorithms, also result may show good percentage of alarms in terms of: False positive, True positive, False Negative and True Negative when researcher use hybrid model.

REFERENCES

[1]Aditi purohit, Hitesh Gupta "Hybrid Intrusion Detection System Model using Clustering, Classification and Decision Table" 2013, IOSR Journal of Computer Engineering.

[2]Kayvan Atefi, Saadiah Yahya, Ahmad Yusri Dak, and Arash Atefi "A Hybrid Intrusion Detection System Based On Different Machine Learning Algorithms" 2013, International Conference on Computing and Informatics.

[3]Kanubhai K. Patel, Bharat V. Buddhadev " An Architecture Of Hybrid Intrusion Detection System" 2013, International Journal of Information & Network Security.

[4]Manisha Valera*, Kirit Rathod(Guide) "A Novel Approach of Mining Frequent Sequential Pattern fromCustomized Web Log Preprocessing" 2013, International Journal of Engineering Research and Applications (IJERA).

[5] M.Moorthy,S.Sathiyabama "A Hybrid Data Mining based Intrusion Detection System for Wireless Local Area Networks" 2010,International Journal of Computer Applications.

[6]Nareshkumar D. Harale, B B Meshram "Hybrid Design Approach for Efficient Network Intrusion Detection using Data Mining and Network Performance Exploration" 2013, International Journal of Innovative Research in Engineering & Science.

[7]Ramachandra. V. Pujeri, G.M. Karthik "Constraint based frequent pattern mining for generalized query templates from web log" 2010,International Journal of Engineering, Science and Technology.

[8]Sandeep Kumar Singh, Nishant Chaurasia, Pragya Sharma "Concept & Proposed Architecture of Hybrid Intrusion Detection System using Data Mining " 2013, International Journal of Engineering and Advanced Technology.