Aravind Kumar / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 1, Issue 3, pp.1098-1101

A Hybrid IDS for Detecting Intrusions based on Classification of Features and Complex Relations

Aravind Kumar.N, M.Tech(S.E)

Abstract:
Intrusion detection is the act of detecting unwanted traffic on a network or a device. A intrusion detection system (IDS) provides a layer of defense which monitors network traffic for predefined suspicious activity or patterns, and alert system administrators when potential hostile traffic is detected. Intrusion detection faces a number of challenges; an intrusion detection system must reliably detect malicious activities in a network and must perform efficiently to cope with the large amount of network traffic. Network based intrusion detection are the most deployed IDS. An IDS can be a piece of installed software or a physical appliance. Many IDS tools will also store a detected event in a log to be reviewed at a later date or will combine events with other data to make decisions regarding policies or damage control. Intrusion detection faces a number of challenges; an intrusion detection system must reliably detect malicious activities in a network and must perform efficiently to cope with the large amount of network traffic. In this paper, we address these two issues of Accuracy and Efficiency using Conditional Random Fields and Layered Approach.

Keywords – Intrusion Detection System (IDS), Layered Approach, Conditional Random Fields, Signature and Anomaly Based IDS.

1. INTRODUCTION

Several types of IDS technologies exist due to the variance of network configurations. Each type has advantages and disadvantage in detection, configuration, and cost.

NIDS (Network Intrusion Detection Systems) Network Intrusion Detection Systems are placed at a strategic point or points within the network to monitor traffic to and from all devices on the network. Ideally one would scan all inbound and outbound traffic. NIDS analyzes network traffic at all layers of the Open Systems Interconnection (OSI) model and makes decisions about the purpose of the traffic, analyzing for suspicious activity. Most NIDSs are easy to deploy on a network and can often view traffic from many systems at once.

HIDS (Host Intrusion Detection Systems) Host Intrusion Detection Systems are run on individual hosts or devices on the network. A HIDS monitors the inbound and outbound packets from the device only and will alert the user or administrator if suspicious activity is detected. HIDS analyze network traffic and system-specific settings such as software calls, local security policy, local log audits, and more. A HIDS must be installed on each machine and requires configuration specific to that operating system .

Signature Based

A signature based IDS will monitor packets on the network and compare them against a database of signatures or patterns of known malicious threats. This is similar to the way most antivirus software detects malware. The issue is that there will be a lag between a new threat being discovered in the wild and the signature for detecting that threat being applied to your IDS. During that lag time your IDS would be unable to detect the new threat.

Anomaly Based

An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is “normal” for that network, what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other and alert the administrator or user when traffic is detected which is anomalous or significantly different than the baseline.

Another approach for detecting intrusions is to consider both the normal and the known anomalous patterns for training a system and then performing classification on the test data. Such a system incorporates the advantages of both the signature-based and the anomaly-based systems and is known as the Hybrid System. Hybrid systems can be very efficient, subject to the classification method used, and can also be used to label unseen or new instances as they assign one of the known classes to every test instance.

II. RELATED STUDY

A. Association rule mining

These are based on building classifiers by discovering Relevant patterns of program and user behavior. Association rules are used to learn the record patterns that describe user behavior. These methods can deal with symbolic data, and the features can be defined in the form of packet and connection details. However, mining of features is limited to entry level of the packet and requires the number of records to be large and sparsely populated. Otherwise, they tend to produce a large number of rules that increase the complexity of the system.

B. Data Clustering Methods k-means and the fuzzy c-means

Clustering technique is based on calculating numeric distance between the observations, and hence, the observations must be numeric. Observations with symbolic features cannot be easily used for the clustering methods. It considers the features independently and is unable to capture the relationship between different features of a single record, which further degrades attack detection accuracy.

C. Naive Baye's classifiers

These make strict independence assumption between the attributes in an observation resulting in lower attack detection accuracy when the features are correlated, which is often the case for intrusion detection. Bayesian network can also be used for intrusion detection. However, they tend to be at tackspecific and build a decision network based on special characteristics of individual attacks. Thus, the size of a Bayesian network increases rapidly as the number of features and the type of attacks modeled by a Bayesian network increases . To detect anomalous traces of system calls in privileged processes, hidden Markov models (HMMs) have been applied in and However, modeling the system calls alone may not always provide accurate classification as in such cases various connection level features are ignored. Further, HMMs are generative systems and fail to model long-range dependencies between the observations.

D. Decision trees

This method selects the finest features for each decision node during the construction of the tree based on some welldefined criteria. One such criterion is to use the information gain ratio. Decision trees generally have very high speed of operation and high attack detection accuracy.

E. Neural Networks

According to Debar though the neural networks can work effectively with noisy data, they require large amount of data for training and it is often hard to select the best possible architecture for a neural network.

F. Support Vector Machines

Support vector machines have also been used for detecting intrusions. Support vector machines map real valued input feature vector to a higher dimensional feature space through onlinear mapping.This can also provide real-time detection capability, deal with large dimensionality of data,and can be used for binary-class as well as multiclass classification. Experimental results on the KDD '99 intrusion data set show that our proposed system based on Layered Conditional Random Fields outperforms other well-known methods such as the decision trees and the naive Baye's.

III.APPLYING CONDITIONAL RANDOM FIELDS FOR SELECTED FEATURES

CRFs are undirected graphical models used for sequence tagging[6]. The prime difference between CRF and other graphical models such as the HMM is that the HMM, being generative, models the joint distribution p(x,y) whereas the CRF are discriminative models and directly model the conditional distribution p(y|x), which is the distribution of interest for the task of classification and sequence labeling. Similar to HMM, the naive Bayes is also generative and models the joint distribution. Modeling the joint distribution has two disadvantages.

First, it is not the distribution of interest, since the observations are completely visible and the interest is in finding the correct class for the observations, which is the conditional distribution p(y|x). Second, inferring the conditional probability p(y|x) from the modeled joint distribution, using the Bayes rule, requires the marginal distribution p(x). This results in reduced accuracy. CRFs, however, predict the label sequence y given the observation sequence x. This allows them to model arbitrary relationship among different features in an observation x. CRFs also avoid the observation bias and the label bias problem, which are present in other discriminative models, such as the maximum entropy Markov models. This is because the maximum entropy Markov models have a per-state exponential model for the conditional probabilities of the next state given the current state and the observation, whereas the CRFs have a single exponential model for the joint probability of the entire sequence of labels given the observation sequence [4].

IV. LAYERED APPROACH FOR IDS

We now describe the Layer-based Intrusion Detection System (LIDS) in detail. The LIDS draws its motivation from what we call as the Airport Security

model, where a number of security checks are performed one after the other in a sequence.

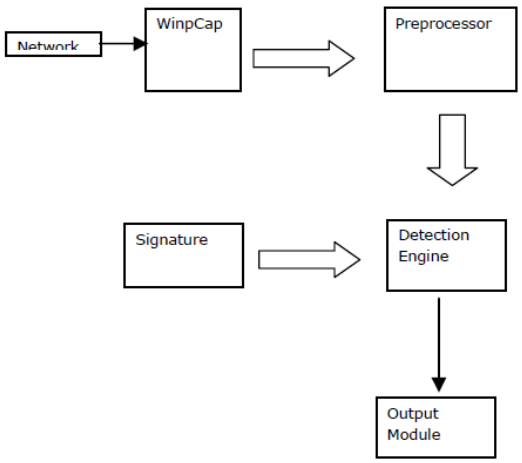

Fig. 1. Layered representation.

Similar to this model, the LIDS represents a sequential Integrated Layered Approach and is based on ensuring availability, confidentiality, and integrity of data and (or) services over a network. The goal of using a layered model is to reduce computation and the overall time required to detect anomalous events. The time required to detect an intrusive event is significant and can be reduced by eliminating the communication overhead among different layers. This can be achieved by making the layers autonomous and self-sufficient to block an attack without the need of a central decision-maker. Every layer in the LIDS framework is trained separately and then deployed sequentially. We define four layers that correspond to the four attack groups mentioned in the data set. They are Probe layer, DoS layer, R2L layer, and U2R layer.

Similar to this model, the LIDS represents a sequential Integrated Layered Approach and is based on ensuring availability, confidentiality, and integrity of data and (or) services over a network. The goal of using a layered model is to reduce computation and the overall time required to detect anomalous events. The time required to detect an intrusive event is significant and can be reduced by eliminating the communication overhead among different layers. This can be achieved by making the layers autonomous and self-sufficient to block an attack without the need of a central decision-maker. Every layer in the LIDS framework is trained separately and then deployed sequentially. We define four layers that correspond to the four attack groups mentioned in the data set. They are Probe layer, DoS layer, R2L layer, and U2R layer. Our second goal is to improve the speed of operation of the system. This results in significant performance improvement during both the training and the testing of the system. In many situations, there is a trade-off between efficiency and accuracy of the system and there can be various avenues to improve system performance.

V. HYBRID MODEL FOR IDS– INTEGRATED LAYERED APPROACH USING CONDITIONAL RANDOM FIELDS (ILACR)

A. WinPcap

As shown in the figure1, the WinPcap software provides facilities to capture raw packets, both the ones destined to the machine where it's running and the ones exchanged by other hosts (on shared media), filter the packets according to user specified rules before dispatching them to the application, transmit raw packets to the network and to gather statistical values on the network traffic.

B. Preprocessor

As mentioned in the figure1 the preprocessor defines one class called packet and this class will store all the packets that are generated by the WinPcap. It captures all the packets in the Network Interface by using Jpcap captor.

C. Signature Database

It is a specially prepared pattern database. Every incident is analyzed to get a regular expression describing the type of attack attempt. There is lot of signatures in the database for such analysis. We have used signatures of the project Snort because the database is still being developed by the Snort Project Team, so updates are often released. Snort uses a simple , lightweight rules description language that is flexible and quite powerful. There are a number of simple guidelines to remember when developing Snort rules. The first is that Snort rules must be completely contained on a single line, the Snort rule parser doesn't know how to handle rules on multiple lines. Snort rules are divided into two logical sections, the rule header and the rule options. The rule header contains the rule's action, protocol, source and destination IP addresses, net masks, source and destination ports information. The rule option section contains alert messages and information on which parts of the packet should be inspected to determine if the rule action should be taken.

D. Detection Engine

It takes packets from preprocessor and compares them with special signatures from the database . Result of the compassion is sent to the output module, where a report is prepared. The detection engine compares the packets in the preprocessor and in the signature database. The comparison takes place at different layers. To compare the content in this paper we are using layered approach algorithm, considering different attributes at each layer. Finally it would detect whether the packet has any attack or not.We now describe the Layer-based Intrusion

Detection System (LIDS) in detail. The LIDS draws its motivation from what we call as the Airport Security model, where a number of security checks are performed one after the other in a sequence. Similar to this model, the LIDS represents a sequential Layered Approach and is based on ensuring availability, confidentiality, and integrity of data and (or) services over a network. The goal of using a layered model is to lessen the computations and the overall time required t o detect anomalous events. The time required to detect an intrusive event is significant and can be reduced by eliminating the communication overhead among different layers. This can be achieved by making the layers autonomous and self-sufficient to block an attack without the need of a central decision-maker.

VI SUGGESTIONS FOR FUTURE WORK

1.Many of the attacks are successful because the attackers enjoy anonymity and they can launch attacks from spoofed sources, making it very hard to trace back the true source of the attack. However, if there is a reliable method to trace back the packets to their actual source, many of the attacks can be prevented.The problem is to identify the true source of attack without affecting the performance of the overall system.

2.Security policy plays an important role in a network and describes the acceptable and non acceptable usage of the resources. There are two major issues in defining the security policy; first, the security policy must be complete and second, the policy must be clear and unambiguous. Hence, the problem is to clearly define the acceptable and the unacceptable usage of every resource.

3. Many systems are based upon authenticating a user. However, authentication mechanisms such as the use of login and password are weak and can be compromised. Multi factor authentication and use of biometric methods have been introduced but they can also be bypassed. The problem is how to link the supplied credentials with the actual human user? Methods based on user profiling can be used which learn the normal user profile and then can be used to detect significant deviations from the learnt profile.

VII. CONCLUSION

This paper mainly deals with the Accuracy and Efficiency problems for devising a robust and efficient intrusion detection systems. Our experimental results how that CRFs are very effective in improving the attack detection rate and decreasing the FAR. Having a low FAR is very important for any intrusion detection system. We compared our approach with some well-known methods and found that most of the present methods for intrusion detection fail to reliably detect R2L and U2R attacks, while our integrated system can effectively and efficiently detect such attacks giving an improvement of 34.5 percent for the R2L and 34.8 percent for the U2R attacks. Our system can help in identifying an attack once it is detected at a particular layer, which expedites the intrusion response mechanism, thus minimizing the impact of an attack. We showed that our system is robust to noise and performs better than any other compared system even when the training data is noisy. The areas for future research include the use of our method for extracting features that can aid in the development of signatures for signature-based systems. The signature-based systems can be deployed at the periphery of a network to filter out attacks that are frequent and previously known, leaving the detection of new unknown attacks for anomaly and hybrid systems. Sequence analysis methods such as the CRFs when applied to relational data give us the opportunity to employ the Integrated Layered Approach, as shown in this paper. This can further be extended to implement pipelining of layers in multicore processors, which is likely to result in very high performance.

REFERENCES

1.P. Dokas, L. Ertoz, A. Lazarevic, J. Srivastava, and P.-N. Tan, “Data Mining for Network Intrusion Detection,” Proc. NSF Workshop Next Generation Data Mining (NGDM ’02), pp. 21-30, 2002.

2 Y. Gu, A. McCallum, and D. Towsley, .“Detecting Anomalies inNetwork Traffic Using Maximum Entropy Estimation,” Proc.Internet Measurement Conf. (IMC ’05), pp. 345-350, USENIX Assoc.,2005.

3 K.K. Gupta, B. Nath, and R. Kotagiri, “Conditional Random Fieldsfor Intrusion Detection,” Proc. 21st Int’l Conf. Advanced Information Networking and Applications Workshops (AINAW ’07), pp. 203-208,2007.

4 J. Lafferty, A. McCallum, and F. Pereira, “Conditional Random
Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” Proc. 18th Int’l Conf. Machine Learning (ICML ’01), pp. 282-289, 2001.

5. E. Tombini, H. Debar, L. Me, and M. Ducasse, “A Serial Combination of Anomaly and Misuse IDSes Applied to HTTP Traffic,” Proc. 20th Ann. Computer Security Applications Conf. (ACSAC ’04), pp. 428-437, 2004.

6. Kapil Kumar Gupta, Baikunth Nath, Ramamohanarao Kotagiri “Layered Approach Using Conditional Random Fields for Intrusion Detection”