AN ALTERNATIVE APPROACH FOR HYBRID INTRUSION DETECTION SYSTEM

[1]KHIN KHATTAR MYINT, [2]NANG SAING MOON KHAM

[1, 2]University of Computer Studies, Yangon.
E-mail: [1]khinnkhattarrmyint@gmail.com, [2]moonkhamucsy@gmail.com

Abstract- Nowadays, intrusion detection system is the most important thing for network security. There are two techniques for intrusion detection: misuse detection and anomaly detection. Most of the intrusion detection system uses all the features available in the network packet. There are many issues in intrusion detection system such as false positive rate, detection rate, memory overhead and time overhead. In this paper, a new hybrid intrusion detection system that integrates a misuse detection model and an anomaly detection model is proposed. First, a rough set based feature reduction is used to select the most significant features and then a misuse detection model is built based on the C4.5 decision tree algorithm not only to detect the known attacks but also to decompose the normal training data into smaller subsets using the model. Next, multiple one-class support vector machines models are created for each decomposed subsets. The proposed hybrid model is evaluated using the NSL-KDD data set.

Keywords- Intrusion Detection System, C4.5 algorithm, Rough set theory, One-class support vector machine.

I. INTRODUCTION

With the rapid development and application of computing and communication technologies, more and more people solve problems or handle things with the internet, using such things as email, internet banking, video conference, save personal information and so on. Therefore, internet security becomes one of the key problems in the world. Traditionally, firewall is a widely used security measure but the firewall alone does not provide enough security [1].Nowadays, much attention has been paid to intrusion detection system (IDS) which is closely linked to the safe use of network services. Intrusion Detection Systems (IDSs) are designed to defend computer systems from various cyber attacks and computer viruses. IDSs build effective classification models or patterns to distinguish normal behaviors from abnormal behaviors that are represented by network data [2]. Generally, intrusion detection algorithms are categorized into two methods: misuse detection and anomaly detection. Misuse detection algorithms detect attacks based on the known attack signatures. They are effective in detecting known attacks with low errors. However, they cannot detect newly created attacks that do not have similar properties to the known attacks. In contrast, anomaly detection algorithms analyze normal traffic and profile normal traffic patterns. The anomaly detection method is based on the hypothesis that the attacker behavior differs to that of a normal user. They classify traffic as an attack if the characteristics of the traffic are far from those of normal traffic patterns. Anomaly detection algorithms can be useful for new attack patterns, but they are not as effective as misuse detection models in the detection rate for known attacks and false positive rates, which is a ratio of misclassified normal traffic [3].The motivation for using the hybrid approach is to improve the accuracy of the intrusion detection system when compared to using individual approaches. The hybrid approach combines the best results from the different individual systems resulting in more accuracy [2].
Databases of real world contain a huge volume of information. Sometimes (or most of the time) part of the data is not interesting for the knowledge extraction. So, data must to be preprocessed in order to select their more relevant attributes for improving the prediction performance of the predictors, and the interpretability and generalization of the data [4].
In this paper, we present a new hybrid system for network intrusion detection using Rough Set Theory, C4.5 algorithm, and one-class support vector machine which has good detection rate and keeps false positive and false negative rates at an acceptable level for different types of network attacks while having the benefit of lower the memory and the time overhead for building the detection model. Rough Set Theory is used to reduce the dimension of original high dimensional data and remove noise effectively. Then we use new dataset which handled by rough set theory to make misuse intrusion detection model by C4.5 algorithm and the resulted data set is used to model the anomaly intrusion detection model by one-class support vector machine. The experimental results show that the proposed method has acceptable detection rates (DR), false positive rate (FP), false negative rate (FN) and accuracy. The rest of this paper is organized as follow: Section 2 describes some related work for this research based on feature reduction and back propagation neural network. Section 3 describes the NSL-KDD dataset. Section 4 describes the detail analysis of the proposed algorithm. Section 5 describes experiments and results followed by a conclusion in Section 6.

II. RELATED WORKS

There has been much research on hybrid intrusion detection system by researchers using different

techniques and approaches to overcome the limitations of both misuse detection and anomaly detection.

Patel [8] proposed a novel and efficient hybrid intrusion detection system (IDS). An architecture and implementation details of their hybrid IDS are presented. The proposed design of hybrid IDS aims to be more accurate and it does not require more processing resources, thus offering both speed and accuracy to detect the intrusions.

Depren [5] proposed an intelligent hybrid intrusion detection system that consists of an anomaly detection model, a misuse detection model, and a decision support system. They modeled the anomaly detection model with a self-organization map (SOM) and the misuse detection model with a decision tree. Each model is trained independently, and then the decision support system simply combines the classification results of both models.

Goel [6] also proposed a novel hybrid model for misuse and anomaly detection. C4.5 based binary decision trees are used for misuse and Classification Based Association based classifier is used for anomaly detection. Model's performance is evaluated on DARPA KDD CUP99 benchmark data set.

Lakhina and Joseph [7] proposed a hybrid algorithm using principal component analysis and neural network algorithm. The PCA transform used to reduce the feature and trained neural network is used to identify the any kinds of new attacks. Test and comparison are done on NSL-KDD dataset.

Chen [1] proposed a new hybrid system for network intrusion detection system using principal component analysis and C4.5 to get a good detection rate and keep false positive and false negative rate at an acceptable level for different types of network attacks.

III. NSL-KDD DATA SET

KDD Cup 99 is the mostly widely used data set for the evaluation of the intrusion detection system. But it has the huge number of redundant records. In KDD train and test set, 78% and 75% of the records are duplicated. This large amount of redundant records in the train set will cause learning algorithms to be biased towards the most frequent records. The existence of these repeated records in the test set will cause the evaluation results to be biased by the methods which have better detection rates on the frequent records. NSL-KDD is the enhanced version of the KDD CUP 99 intrusion detection data set. It is generated by removing redundant instances in both the training and testing data of the KDD 99 dataset. NSL-KDD consists of the same features as KDD 99. The KDD99 data set consists of 41 features and one class attribute. The data set contains 24 attack types. All these attacks fall into four main categories.

1. **Denial of Service (DOS):** In this type of attacks an attacker makes some computing or memory resources too busy or too full to handle legitimate requests, or denies legitimate users access to a machine. Examples are Apache2, Back, Land, Mailbomb, SYN Flood, Ping of death, Process table, Smurf, Teardrop.
2. **Remote to User (R2L):** In this type of attacks an attacker who does not have an account on a remote machine sends packets to that machine over a network and exploits some vulnerability to gain local access as a user of that machine. Examples are Dictionary, Ftp_write, Guest, Imap, Named, Phf, Sendmail, Xlock.
3. **User to Root (U2R):** In this type of attacks an attacker starts out with access to a normal user account on the system and is able to exploit vulnerability to gain root access to the system. Examples are Eject, Loadmodule, Ps, Xterm, Perl, Fdformat.
4. **Probing:** In this type of attacks an attacker scans a network of computers to gather information or find known vulnerabilities. An attacker with a map of machines and services that available on a network can use this information to look for exploits. Examples are Ipsweep, Mscan, Saint, Satan, Nmap.

Table.I. Data Distribution in NSL-KDD data set

NSL-KDD dataset	Normal	Attack	Total
Training	67343	58630	125973
Testing	9711	12833	22544

IV. PROPOSED HYBRID INTRUSION DETECTION SYSTEM

In this section, Rough set theory, decision tree and one-class support vector machine algorithms that are required in order to build the misuse detection model and anomaly detection model are briefly introduced. Then, the integration of these models is explained.

4.1. Rough Set Theory

Rough set theory (RST) is a useful mathematical tool to deal with imprecise and insufficient knowledge, find hidden patterns in data, and reduce dataset size [9]. Also, it is used for evaluation of significance of data and easy interpretation of results. RST contributes immensely to the concept of reducts. A reduct is a minimal subset of features with the same capability of objects classification as a whole set of features.

4.2. C4.5 decision tree

Decision tree learning is one of the most widely used techniques for classification. Its classification accuracy is competitive with other learning methods, and it is very efficient. It is a flowchart-like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. It operates in a divide and conquer manner, which partitions the training data set based on its attributes until the

stopping conditions are satisfied. A decision tree node has its corresponding data set; this specifies the attribute to best divide the dataset into its classes. Each node has several edges that specify possible values or value ranges of the selected attributes on the node. The primary issue of the decision tree algorithm is to locate the attribute that best divides the data into their corresponding classes.C4.5 builds decision trees from training data sets using the concepts of information entropy. It is based on the highest gain of each attribute.

The gain is calculated using the formula:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} f_s(A_i) \times \text{Entropy}(S_{A_i})$$

$$\text{Entropy}(S) = -\sum_{j=1}^{m} f_s(j) \times \log_2 f_s(j)$$

4.3. One-class Support Vector Machine

One-class SVMs attempt to learn a decision boundary that achieves the maximum separation between the points and the origin. Interestingly this was the initial idea from which traditional super-vised SVMs emerged. The idea was hindered by the inability to learn non-linear decision boundaries as well as the inability to account for outliers. Both of these problems were solved by the introduction of kernels and the incorporation of soft margins. A one-class SVM uses an implicit transformation function $\emptyset(.)$ defined by the kernel to project the data into a higher dimensional space. The algorithm then learns the decision boundary (a hyperplane) that separates the majority of the data from the origin. Only a small fraction of data points are allowed to lie on the other side of the decision boundary. Those data points are considered as outliers.

4.4. Proposed hybrid intrusion detection method

The proposed system is shown in the figure 1. Firstly, rough set theory is used as a feature extraction method because some of the attributes of the NSL-KDD dataset are redundant. It may consume the memory and CPU time for building the intrusion detection system. Therefore, rough set theory is used as a feature extraction method to select the most significant features. The output of the feature extraction is the reduced dataset. Then, the resulted reduced dataset is used as an input to the misuse-based model. Decision Tree based misuse detection model is built using reduced dataset containing both normal and attack connections .It can not only detect the known attack but also divide the normal data connections into subsets using the divide and conquer strategy. Finally, a one-class SVM model is trained for each normal training data set which is decomposed by the decision tree model. The data patterns of each decomposed subset are less complex than those of the whole data set. The training and testing time of the anomaly detection model is reduced.

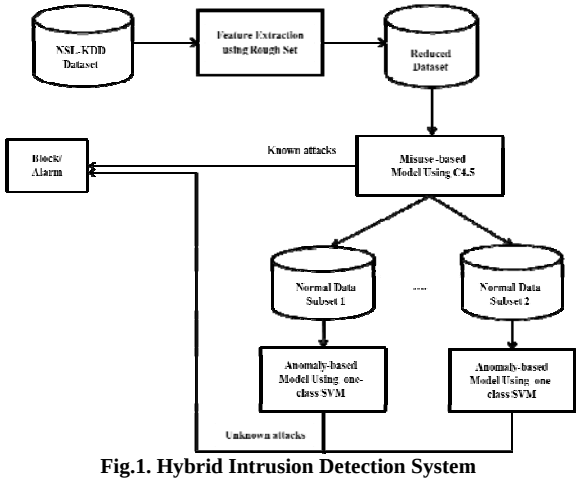


Fig.1. Hybrid Intrusion Detection System

V. EXPERIMENTAL

The experiments were run on a system with a 1.70GHZ core i5 processor and 4GB of RAM running windows 7.NSL-KDD data set is used to train and test the intrusion detection system. It has 41 features and one class attribute. Some features are redundant and it may cause the memory overhead problem. In our proposed system, the resulting reduced features using the rough set theory are only 7 features given in table II.

Table.II. Significant Features reduced by RST

No.	Attribute Name
1	Src_byte
2	Dst_byte
3	Count
4	Srv_count
5	Dst_host_count
6	Dst_host_srv_count
7	Dst_host_same_src_port_rate

The performance of the proposed model is calculated by the following measures.

- True Positive Rate(TPR) :TPR $= \frac{TP}{TP+FN}$
- False Positive Rate(FPR) :FPR $= \frac{FP}{TN+FP}$
- TP: correctly identified attack traffic.
- TN: correctly identified normal traffic.
- FP: normal traffic being misclassified as an attack.
- FN: attack instances being misclassified as normal.

Table III shows the confusion matrix.

Table.III. Confusion Matrix

Actual Class	Classified as	
	Normal	**Attack**
Normal	True Negative (TN)	False Positive (FP)
Attack	False Negative (FN)	True Positive (TP)

Weka 3.6 is used to evaluate the performance of the proposed method. Table IV shows the comparison of detection accuracy. The experiment demonstrates that the proposed hybrid intrusion detection method is better in computation time than the conventional hybrid methods.

Table.IV. Comparison of Detection Accuracy

	No. of attributes	Time taken to build the model	Detection rate
Without Feature Selection	41	45.83 seconds	99.7555 %
With Feature Selection	7	6.83 seconds	98.818 %

CONCLUSIONS

The proposed hybrid intrusion detection system combines both misuse and anomaly intrusion detection systems in a decomposition structure. Rough set theory is used to select out feature to cope the problem of dimensionality reduction as a feature extraction. Then, C4.5 is used as a misuse detection model not only to detect the known attack but also to decompose the normal training data into smaller subsets. One-class support vector machine is used to create an anomaly detection model in each decomposed region. It is expected to get a high detection rate and low false alarm rate in comparison with the existing individual and hybrid IDSs. It is also expected to reduce both memory and CPU time for building the hybrid IDS. In our research, we found that the computation time is reduced seven times than the model without using feature extraction.

REFERENCES

[1]. Zhi-Guo Chen, Sung-Ryul Kim, "A Hybrid System for Reducing Memory and Time Overhead of Intrusion Detection System," IFIP International Federation for Information Processing 2014.

[2]. Vahid Golmah, "An Efficient Hybrid Intrusion Detection System based on C5.0 and SVM," International Journal of Database Theory and Application, vol.7, No.2, 2014.

[3]. Gisung Kim, Seungmin Lee, Sehun Kim, "A Novel Hybrid Intrusion Detection Method Integrating Anomaly Detection with Misuse Detection," Expert Systems with Application 41, 2014.

[4]. Fabio Henrique Pereira, Renato Jose Sassi, "Rough Sets and Principal Components Analysis: A Comparative Study on Customer Database Attributes Selection," African Journal of Business Management, vol.6 (10), 14 March, 2012.

[5]. Ozgur Depren, Murat Topallar, Emin Anarim, M.Kemal Ciliz, "An Intelligent Intrusion Detection System for Anomaly and Misuse Detection in Computer Networks," Expert Systems with Applications 29, 2005.

[6]. Rakhika Goel, Anjali Sardana, Ramesh C. joshi, "Parallel Misuse and Anomaly Detection Model," International Journal of Network Security, vol.14, no.4, July 2012.

[7]. Shilpa lakhina, Sini Joseph, Bhupendra Verma, "Feature Reduction using Principal Component Analysis for Effective Anomaly Based Intrusion Detection on NSL-KDD," International Journal of Engineering Science and Technology, vol.2 (6), 2010.

[8]. Kanubhai K. Patel, Bharat V. Buddhadev, "An Architecture of Hybrid Intrusion Detection System," International Journal of Information & Network Security (IJINS), vol.2, no.2, April 2013.

[9]. Rowayda A.Sadek, M.Sami Soliman, Hagar S.Elsayed, "Effective Anomaly Intrusion Detection System based on Neural Network with Indicator Variable and Rough Set Reduction," International Journal of Computer Science Issues, vol.10, Issue 6, No.2, November 2013.

★★★