International Journal of Soft Computing and Engineering (IJSCE)
Volume-1, Issue-1, March 2011

A Novel Approach to Intrusion Detection System using Rough Set Theory and Incremental SVM

Ghanshyam Prasad Dubey, Prof. Neetesh Gupta, Rakesh K Bhujade

***Abstract*— Intrusion Detection System (IDS) is software and/or hardware designed to detect unwanted attempts at accessing, manipulating, and/or disabling of computer systems, mainly through a network, such as the Internet. These attempts may take the form of attacks, as examples, by crackers, malware and/or disgruntled employees. An IDS cannot directly detect attacks within properly encrypted trafficOn detection of such sign triggers of IDS to report them generate the alerts. These alerts are presented to a human analyst who evaluates them and initiates an adequate response. In Practice, IDSs have been observed to trigger thousands of alerts per day, most of which are mistakenly triggered by begin events such as false positive. This makes it extremely difficult for the analyst to correctly identify alerts related to attack such as a true positive. Recently data mining methods have gained importance in addressing network security issues, including network intrusion detection. Intrusion detection systems aim to identify attacks with a high detection rate and a low false positive. We use RST (Rough Set Theory) and Incremental SVM (Support Vector Machine) to detect intrusions. First, RST is used to preprocess the data and reduce the dimensions. Next, the features were selected by RST will be sent to SVM model to learn and test respectively. The method is effective to decrease the space density of data. Using this method, it can overcome the shortages of SVM time-consuming of training and massive dataset storage. The simulation experiments with KDD Cup 1999 data demonstrate that our proposed method achieves the increasing performance for intrusion detection.**

***Index Terms*—Intrusion Detection, Support Vector Machine, Rough Set Theory, Data Mining**

I. INTRODUCTION

Intrusion detection systems have been an active area of research and development since 1987. This is particularly with the increase of attacks on computers and on networks in recent years improved and essentially automated surveillance has become a necessary addition to IT security .Intrusion detection is the process of monitoring the events occurring in a computer system or network and analyzing them for signs of intrusions. Intrusions are defined as attempts to compromise the confidentiality, integrity or availability of a computer or network or to bypass its security mechanisms. They are caused by attackers accessing a system from the Internet, by authorized users of the systems who attempt to gain additional privileges for which they are not authorized and by authorized users who misuse the privileges given to them.

There are generally two types of approach taken toward intrusion detection: Anomaly detection and Misuse detection. Misuse Detection depends on the pre-recording representation of specific patterns for intrusions, allowing any matches to them in current activity to then be reported. Patterns corresponding to known attacks are called signatures. A classifier is then trained to discriminate one category from the other, based on network traffic attributes.. They used a series of data mining techniques, such as frequent episodes and association rules, to help extract discriminative features, which include various network traffic statistics. Obtaining correctly- labeled data instances, however, is difficult and time intensive, especially for new attack types and patterns [2]? On the other hand Anomaly Detection amounts to training models for normal traffic behavior and then classifying as intrusions any network behavior that significantly deviates from the known normal patterns.

Traditional methods of network intrusion detection are based on the saved patterns of known attacks. They detect intrusion by comparing the network connection features to the attack pattern that are provided by human experts. The main drawback of traditional methods is that they cannot detect unknown intrusion. Even if a new pattern of the attacks were discovered, this new pattern would have to be manually updated into system. It is also capable of identifying new attacks to some degree of resemblance to the learned ones, the neural networks are widely considered as an efficient approach to adaptively classify patterns, but their high computation intensity and the long training cycles greatly hinder their applications, especially for the intrusion detection problem, where the amount of related data is very important. An increasing amount of research in the last few years has investigated the application of neural networks to intrusion detection. If properly designed and implemented, neural networks have the potential to address many of the problems encountered by traditional approaches.

In this study, we design an incremental SVM framework with Rough Set Theory for intrusion detection based on key feature selection. Our study includes several features.

First, a center sever summarizes the new samples and generate an updated SVM, which is finally deployed to distributed hosts. Second, the coordination

Manuscript received March 8, 2011.

Ghanshyam Prasad Dubey is M.Tech Scholar in Department of Information Technology, Technocrats Institute of Technology, Bhopal. INDIA.Phone: +909827555366; e-mail: ghanshyammtech@rediffmail.com

Prof. Neetesh Gupta is working as Head, Department of Information Technology, Technocrats Institute of Technology, Bhopal. INDIA. e-mail: gupta_neetesh81@yahoo.com.

Rakesh K Bhujade is working as Associate Professor Technocrats Institute of Technology, Bhopal , India., e-mail: rakesh.bhujade@ gmail.com.

mechanisms among the hosts and severs make an effective cooperation for sharing intrusion detection knowledge. Third, only some key features of all the 41 features of the dataset are used for making classifications. Fourth, there are five classes (normal, probe, DoS, U2R, and R2L) of patterns in the dataset. In building an Local IDS using five SVMs, each SVM can use only the important features for that class which it is responsible for making classifications. In addition, we demonstrate framework by the dataset, KDD CUP 1999 [1].

II. PREVIOUS WORK IN SVM AND ROUGH SET THEORY

Intrusion Detection System has misuse detection and anomaly detection. The known attack behaviors are constructed from misuse detection attribute database in the development stage. Misuse detection system compares user behaviors with attribute database to find intrusions. Anomaly detection system defines system exactly normal behaviors in rule. The contrast between collects system parameter and defines behaviors can find the Misbehaviors in system

A. Misuse detection

Misuse detection sets up the attack behaviors based on known attack behaviors during the Development stage. The misuse detection is similar to antivirus software. The antivirus software compares the scanned data with known virus code. If system finds un-normal attributes, the virus is existence and removes it. Hence, misuse detection collects the known attack behaviors from attribute database. If the attack behavior is similar to the one in database, the misuse detection can defend it before the intruder destroys our system.

B. Anomaly detection

Anomaly detection is different from misuse detection. The system constructs user model based on normal users have behaviors. When user has misbehaviors, the system notifies users that has an intruder. The main drawback of anomaly detection is that the detection is depended on the latest attack models, so it can't identify new attack behaviors. The intruder attack methods will be changed, so anomaly detection system collects normal behaviors and detects intruding using normal behaviors. The anomaly detection system has a party with clearly defined correct user behaviors. The problem is intruder uses normal behaviors to attack the system.

IDS system monitors the packages transmissions on the network. While malice behaviors have happen, IDS will send an alert to the network manager or use a related method to defense the attacks. Most Intrusion Detection Systems are classified as either a NIDS (Network Based Intrusion Detection System) or a HIDS (Host Based Intrusion Detection System) [6] [3]. In general, NIDS is located between host and firewall. HIDS was usually installed on a server or NIDS collects and analyzes the information at the host. NIDS could monitor the data real time on the network. If NIDS finds illegal behaviors, it will send messages to the managers. Comparatively, HIDS monitors the activities of the host. So it can determine whether an attack or not. The data of HIDS is caught by the host, so it is not easy to be influenced by some methods, just like encryption. Entropy has been used in intrusion detection for a long time. B. Balajinath et al. used entropy in learning behavior model of intrusion detection in 2001[2]. TF-IDF is often applied to IDS, too. Such as Wa-Wa Chen et al. compared SVM to ANN for intrusion detection, their methods are based on TF-IDF. The Unauthorized Access from a Remote Machine (R2L) and The Unauthorized Access to Local Super-user Privileges (U2R) both is intrusion behaviors which will be detected by HIDS.

III. THE METHODOLOGY

A. Incremental SVM

Support Vector Machines (SVMs) can be used to learn with large amounts of high dimensional data. However, computing a SVM is very costly in terms of time and memory consumption. It is a good idea to learn incrementally from previous SVM results. Compared to the number of training examples, in most cases the number of Support Vectors is very small. SVMs can compress the data of the previous batches to their Support Vectors in incremental learning. This incremental learning approach with SVMs has been investigated. Compare to nonincrementally trained SVMs, incrementally trained SVMs behave well. The data is provided in several batches. For each new batch of data a SVM is trained on the new data and the Support Vectors from the previous learning step. After each training step, a preliminary result will be produced by the algorithm. So time and memory consumption are controlled. To judge, one can comparing its results to the results of the learning algorithm trained all data simultaneously[9]. Incremental techniques have been found widespread used in SVM. Incremental SVM learning is particularly attractive in an online system, and for active learning. In an online system, the data is often collected continuously in time. Significant effort has been spent in the recent years on development of online SVM learning algorithms [10].

The elegant solution to online SVM learning is the incremental SVM which provides a framework for exact online learning. In the make of this work two extensions to the regression SVM have been independently proposed .

B. IDS Techniques

1)Statistical models

Several statistical characterizations of events and event counters, and more refined techniques, have been implemented in anomaly detection systems. These techniques include threshold measures, mean and standard deviation and multivariate models.

2) Markov process model

Markov process mode does not use system call sequences, but instead analyze the state transitions for each system call.

3) Rule-based algorithm

One of the most used rule-based algorithms in the intrusion

detection field is RIPPER [5, 6], which performs classifications by creating a list of rules from a set of labeled training examples.

4) Data mining techniques

Many recent approaches to Intrusion Detection Systems build detection models by applying data mining techniques to large data sets collected by a system.

5) Immune system approach

In the immune system approach, applications are modeled in terms of the system call sequences [2, 7, and 8].

C. Classification approaches in IDS

1) Artificial neural network

ANN is a biologically inspired form of distributed Computation. It is composed of simple processing units, or nodes, and connections between them. The connection between any two units has some weight, which is used to determine how much one unit will affect the other. A subset of the units acts as Input nodes and another subset acts as output nodes, which perform summation and threshold. The ANN has successfully been applied in different fields. The feed-forward neural network trained with the back-propagation algorithm is a common tool for intrusion detection [9, 10].

2) Support vector machine

SVM is a technique for solving a variety of learning, Classification and prediction problems [2],[7]. The basic SVM deals with two-class problems—in which the data are separated by a hyper plane defined by a number of support vectors.

Support vectors are a subset of training data used to define the boundary between the two classes. In situations where SVM cannot separate two classes, it solves this problem by mapping input data into high-dimensional feature spaces using a kernel function. In high-dimensional space it is possible to create a Hyper plane that allows linear separation. Compared with the ANN, the SVM have two advantages.

Firstly, the global optimum can be derived. Secondly, the over fitting problem can be easily controlled by the choice of a suitable margin that separates the data. Empirical testing has shown that the SVM performance is better than that for the ANN in classification and regression problems.

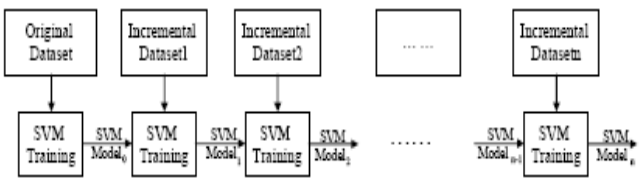

Fig. 1: The Incremental training approach

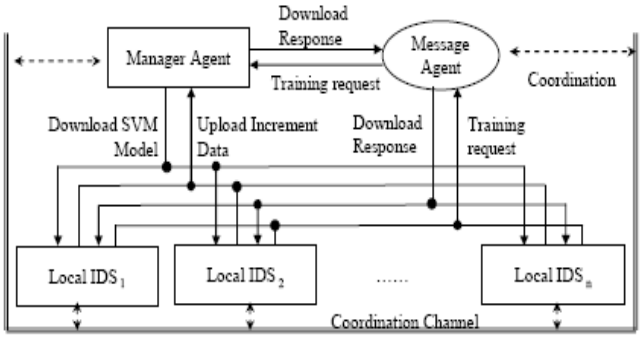

Fig.2: Cooperation of distributed classifiers in DI-IDS

In Fig.(1) the incremental process is show. The original dataset may come from the history accumulation or other network security organizations. To update the classification Engine, there are two ways. First, the classification model will be updated by the service providers. This way is delayed compared with evolution of intrusion pattern development. Second, the system can update itself in an evolutionary and real-time way to utilize the samples generated in the runtime. In engineering, the kind combination of the two approaches will produce promising effects.

D. Cooperation of distributed classifiers in DI-IDS

In Fig.(2) the coordination mechanisms among local IDSs and controller are show. The local IDS with SVM detect the intrusion by system or manual ways will produce new samples and send them to the center. The center will evaluate the samples itself by the characteristics of the samples and the self owned Information. The center informs or inquires the other local IDSs with summary of the samples. The local IDS will increase the security level after evaluating the alerts and produce response to the center. After summarizing the feedbacks and the reference information from other organization, the center will determine whether the SVM will be incrementally trained. After the regeneration of the central SVM, it will be sent to the local IDSs so that the local SVMs will be updated online without much effect to the runtime performance duce promising effects.

E. The Data Set

KDD CUP 1999 data [1] was the data set used for the Third International Knowledge Discovery and Data Mining Tools Competition. The training data set contains 494,021 connection records, and the test data set contains 311,029 records. In our experiments, we sample the data only from the training data set and use in both the training and testing stages. A connection is a sequence of TCP packets containing values of 41 features and labeled as either normal or an attack, with exactly one specific attack type. A complete listing of Features and details are in KDD CUP 1999 data [1]. There are 22 attack types in the training data[8]. In this paper, the attacks in the training data are grouped into broad classes; each neuron is then labeled as representing one of these classes. Specifically, the four broad classes of attack type defined by MIT Lincoln Labs [4] are used, as stated below:

*1) Denial-of-Service (Do*S): These are attacks designed to make some service accessible through the network unavailable to legitimate users.
2) Probe: A Probe is a reconnaissance attack designed to uncover information about the network, which can be exploited by another attack.
3) Remote-to-Local (R2L): This is where an attacker with no privileges to access a private network attempts to gain access to that network from outside, e.g. over the internet.
4) User-to-Root (U2R): The attacker has a legitimate user account on the target network. However, the attack is designed to escalate his privileges so that he can perform unauthorized actions on the network.'

IV. FEATURE SELECTION BY ROUGH SET

Using the RST reduces the attributes for SVM operation. Rough Set Theory [3][4][5] is One of data-mining methods which reduce the features from large numbers of data. Using RSTneeds to build the decision table or the information table. The decision table describes the Features of processes. Formally, an information system IS (or an approximation space) can be shown as follows:

$$IS = (U, A)$$

Where U is the Universe (a dataset of process, U={*x1,x2,x3,x4,x5,x6,…,xm*}) and A presents the attributes of a process, for instances, (A={*a1,a2,a3,a4,a5*}).

The definition of an information function is fa: U⟶ Va, Va is the set of values of the attributes. For example, the values of U and A are listed as follows and they are mapping to Vi.

U= {*x1, x2, x3, x4, x5, x6, …, xm*}
A= {*a1, a2, a3, a4, a5*}
V1= {1, 2, 3, 4}
V2= {1, 2, 3, 4, 5}
V3= {1, 2, 3, 4, 5}
V4= {1, 2, 3}

For every set of attributes B $\subseteq$ A, if b (xi) = b (xj) (every b $\subseteq$ B), there is an indiscernible relation Ind(B). Continuous, to define the basic concepts, namely the Upper approximations and Lower approximations of a set let X represent the elements of subset of the universe U (X $\subseteq$ U). The lower approximations of X in B (B $\leqq$ A) represents as *BX* such as follows.

$$\overline{BX} = \{X_i \in U \mid [X_i]_{Ind(B)} \subset X\}$$

The lower approximations of set X of process xi, which contained X of elementary set in the space B. The upper approximation of set X is BX. BX represents the union of the elementary which is a non-empty intersection with X.

$$BX = \{X_i \in U \mid [X_i]_{Ind(B)} \cap X \neq 0\}$$

For any object xi of lower approximation of X (xi $\in \overline{BX}$), it is certainly belongs to X. For object of xi of upper approximations of X (xi $\in BX$), it is called a boundary of X in U. The difference of upper and lower approximations is:

$$BNP = (BX = \overline{BX})$$

If the upper and lower approximations are identical $(BX = \overline{BX})$, the set X is definable;otherwise, set X is indefinable in U. There are four types of the set of indefinable in U. Ø represents an empty set.

If $BX \neq Ø$ and BX$\neq$ U, the set of X represents roughly definable in U;If $BX \neq Ø$ and BX=U, the set of X represents externally indefinable in U;If $BX = Ø$ and BX $\neq$ U, the set of X represents internally indefinable in U;If $BX = Ø$ and BX $\neq$ U, the set of X represents totally indefinable in U.

Using all attributes to do intrusion detection is ineffective. In this paper, RST is used to combine the similar attributes and to reduce the number of attributes. So it can enhance the processing speed and to promote the detection rate for intrusion detection.

V. INTRUSION DETECTION SYSTEM WITH INCREMENTAL SVM AND ROUGH SET

So far we have gained insight in the various models of the unsupervised learning algorithms.Furthermore in relation to the Intrusion Detection Systems we have pointed out the advantages and disadvantages of the algorithms in question. Based on these considerations we can now point out an algorithm for evaluation and implementation.This choice is also inspired by the works and tests in the articles of Incremental SVM Algorithm are chosen to be used as the learning algorithm for the desired IDS. The choice is made according to the following statements :

A. Simple and easy-to-understand algorithm that works:

We have analyzed the good and bad sides of this algorithm and came to realize that it is capable of dealing with large problems that require reckoning and comparing without any complexity. As for the IDS that we want to construct, we need an algorithm that can manage to transform high dimensional data sets into a 2-dimensional data set. The simplicity of the Neural has makes it easy to implement and manage.

B. Topological clustering:

The Incremental SVM has the ability to construct a topological result. This feature will be useful during the training and test phase of the IDS in order to observe the validity of the result from the algorithm and follow up on the clustering process to check whether same patterns (e.i. features in this case) fall into the same cluster.

C. Unsupervised algorithm that works with nonlinear data

As the traffic from a network connection can be a huge amount and is most likely representing nonlinear data, we will need an algorithm, which is operational regardless of the amount and the linearity of the data sets.

The Incremental SVM has the ability to handle such data set. Another noticeable character of this algorithm is that it is unsupervised, which makes it capable of detecting intrusions without being introduced to it. So, do these statements exclude the choice of another algorithm? The answer is no, because the other algorithms with certain conditions can also work as the learning algorithm for intrusion detection.

Our choice is based on the specifications and requirements

for our IDS and therefore we decide to use the self-organizing map due to its properties listed in the statements from above. We cannot tell for sure that the self-organizing map is the best choice. The only way of finding out its quality is to compare it with the other unsupervised models is covered. A reliable way is to evaluate and implement so many models as possible, and then compare them on the efficiency. But since we are going to implement only one algorithm

VI. CONCLUSION

In recent years, the intrusion detection is increasing complex for the evolutionary attacks. In the paper, a method of detecting intrusion using incremental SVM with roughest theory based on key feature selection is proposed. The coordination mechanisms among the IDSs with SVM are kindly designed. By the KDD dataset, we propose a simulation approach to verify our researches. First, besides the incremental SVM with rough set theory will be incorporated to improve the performance and decrease the sample set. Second, the framework should be built into real IDSs to test the performance and affect the whole intrusion detection flow understanding and detection of new attack categories. Sophisticated self-labeling techniques, taking into consideration of additional network security domain knowledge, can be developed to improve the performance of clustering-based intrusion detection.

REFERENCES

[1] KDD Cup 1999 Data, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html,August 2003.
[2] M.A. Aydın, A.H. Zaim, K.G. Ceylan, A hybrid intrusion detection system design for computer network security. Computers & Electrical Engineering, 2009. 35(3): 517-526.
[3] Mukkanmala S,Sung," A Feature ranking and detection for intrusion" Proceedings of the International Conference on Information and Knowledge Engineering-IKE 2002，2002:503-509.
[4] MIT Lincoln Labs, 1999 DARPA intrusion detection evaluation,available at: http://www.mit.edu/IST/ideval/.
[5] W.T. Yue, Y.U. Ryu, The management of intrusion detection: Configuration, inspection, and investment. European Journal of Operational Research, 2009. 195(1): 186-204.
[6] B. Morin, M. Ludovic, H. Debar, M. Ducass, A logic-based model to support alert corre lation in intrusion detection. Information Fusion,2009. 10(4): 285-299
[7] P.García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges". Computers & Security, 2009. 28(1-2): 18-28.
[8] J. Yang, X. Liu, T. Li, G. Liang, S. Liu, Distributed agents model for intrusion detection based on AIS. Knowledge-Based Systems, 2009. 22(2): 115-119.
[9] R. Beghdad, Critical study of neural networks in detecting intrusions. Computers & Security, 2008. 27(5-6): 168-175.
[10] S.-J. Horng, P. Fan, Y.-P. Chou, Y.-C. Chang, Y. Pan, A feasible intrusion detector for recognizing IIS attacks based on neural networks.Computers & Security, 2008. 27(3-4): 84-100.