Handling Intrusion Detection System using Snort Based Statistical Algorithm and Semi-supervised Approach

G.V. Nadiammai and M. Hemalatha
Department of Computer Science, Karpagam University, Coimbatore-641021, Tamil Nadu, India

Abstract: Intrusion detection system aims at analyzing the severity of network in terms of attack or normal one. Due to the advancement in computer field, there are numerous number of threat exploits attack over huge network. Attack rate increases gradually as detection rate increase. The main goal of using data mining within intrusion detection is to reduce the false alarm rate and to improve the detection rate too. Machine learning algorithms accomplishes to solve the detection problem. In this study, first we analyzed the statistical based anomaly methods such as ALAD, LEARAD and PHAD. Then a new approach is proposed for hybrid intrusion detection. Secondly, the advantage of both supervised and unsupervised has been used to develop a semi-supervised method. Our experimental method is done with the help of KDD Cup 99 dataset. The proposed hybrid IDS detects 149 attacks (nearly 83%) out of 180 attacks by training in one week attack free data. Finally, the proposed semi-supervised approach shows 98.88% accuracy and false alarm rate of 0.5533% after training on 2500 data instances.

Keywords: Application Layer Anomaly Detector (ALAD), intrusion detection, KDD cup 99 datasets, Learning Rules for Anomaly Detection (LERAD), Packet Header Anomaly Detection (PHAD), semi-supervised approach, snort

INTRODUCTION

The intrusion detection methodology was proposed by Anderson (1980). Attacks persist on computing resources and networks for many reasons such as complexity of computer and innovative hardware and software makes the system and networks easily influenced to intrusion. Bloom of internet leads to increase in network users and traffic. In recent weeks, computer hackers have attacked a Saudi Arabian oil company, a Qatari natural gas (Cyber-attack-threat, 2012) company and several American banks. Self-monitoring software such as anti-virus, spyware, pop-up blocking and anti-spam software act at the local client machine by imposing security policies. A firewall can only able to control the network access of computing resources at network level. So here we make use of intrusion detection system to detect and prevent the attacks interchangeably.

IDS can be classified as misuse and anomaly detection. Misuse approach efficiently detects the known attacks that are predefined but fails to detect the unknown attacks. One of the main advantages of using misuse approach (Ertoz et al., 2009; Ben and Kavitha, 2012; Alok and Ravindra, 2012) is producing less false alarms. In case of anomaly detection it can able to detect unknown attacks with high false alarms. Likewise there are two types of IDSs such as Host based IDSs (HIDS) and Network based IDSs (NIDS). HIDS analyses the data that comes from application programs, system logs, audit record and it checks it with the network whether any intrusion occurs to the particular host. Centralized NIDS is connected to whole ids.

Snort is the signature based anomaly detection method. It captures the incoming packets that are transmitted over the network (Roesch, 1999). It incorporates rules within it and thereby performs pre-processing by itself. It mainly reduces the burden of system administrator. New rules can be included within the rule set according to the occurrence of new attacks. Snort is used with statistical methods to improve the detection strategy in real time.

Semi-Supervised learning algorithm (Xiaojin, 2008) has a significant aspect among research community. In data mining and intrusion detection field, there is a frequent availability of unlabeled data and limited labeled data. So in these applications it is impossible to provide labels for all data and being time consuming. Usually only a small portion of data can be labeled. Based on the labeled training data, the test data can be labeled gradually. Supervised learning technique requires labeled data to learn the model and produces accurate results. Whereas unsupervised learning approaches discovers composition from unlabeled data. But semi-supervised approach utilizes the advantage of...
both supervised and unsupervised methods. One of the main advantage is that we can find meaningful structure of complicated high dimensional data.

The motivation of this study is to solve the two main issues such as, reducing the burden of security experts in case of pre-processing and lack of labeled data. Through solving these issues the performance of the IDS would be improved. Based on this dependency two algorithms have been proposed. For the first issue, hybrid IDS has been proposed based on statistical algorithms using snort. Secondly, the semi-supervised algorithm has been proposed to solve the second issue respectively.

LITERATURE REVIEW

Anomaly Detection can be done from attack free data. Network anomaly detectors usually models low level attributes. Association rule learning approach (Barbara et al., 2001) used widely among data mining techniques to build IDS. Caswell and Beale (2004) 10 uses a misuse model for IDS approach. Burroughs et al. (2002) and Cuppens and Miege (2002), tries to solve the intrusion by detect and prevent attacks in future from distributed IDS. According to statistical based approach, the network traffic is analyzed through quasi stationary approach. But this approach is not applicable in real time and also leads to huge false alarm rate. Since the behavior of internet varies over time, this helps the attack to easily spread attack over the network. Cai et al. (2007) and Floyd and Paxson (2001) proposed that the intrusion detection must be taken place on connection features like transport and application layers respectively.

Kai et al. (2007) uses ADS to analyze anomalous traffic from the internet. It is correlated with snort to detect unusual attacks and thereby increasing the detection rate. ADS provides better performance of 33% while used to snort. The call sequence method has been modeled using n gram and neural networks approach (Forrest et al., 1996; Anup and Schwartzbard, 1999). Zhenwei et al. (2007) present a tunning process which will automatically perform detection and corresponding reports are generated by system operator if any false judgment is made. Mahoney and Chan (2003) and Kai et al. (2007) used real time and DARPA dataset and showed that the simulated data detects better when compared with both datasets. PHAD by Matthew and Philip (2001) 20 uses a non stationary method based on time when compared to average frequency and detects of attacks efficiently. A filter is used to find the hostile events.

The author Denis (2009) indulges LERAD algorithm for better accuracy in offline method through generating minimum rules and decrease in detection rate. Matthew (2003) uses four statistical based anomaly detection algorithms to overcome the detection problem mainly in data link layer, application layer and header files in order to extract better rules from a poor set of rules. Mahoney and Chan (2003) presents an automatic adaptable traffic model in case of analyzing the network traffic to improve the detection rate in real time. Aydin et al. (2009) proposed hybrid IDS by combining packet header anomaly detection and network traffic anomaly detection with snort to improve the performance of ids. The proposed IDS detects 72% of attack in one week data respectively.

Machine learning techniques towards reducing false positive is a common concept. There are many semi-supervised learning methods in practice. Scudder (1965) proposed a Bayes approach in 1965 and analyses the probability of error. He shows that the approach works well for unknown pattern. Pavan et al. (2009) proposed a boosting algorithm for semi-supervised approach. The semi-Boost algorithm shows better performance for base classifier using unlabeled samples. Yi et al. (2010) uses networks connection feature instance (NCF instance) to determine the cluster of alerts. Here he used RSVM algorithm to reduce the false alarms. The experimental results show that both detection rate and false alarm rate has been improved by using unlabeled data by filtering 65% false alarms and less than 0.1% true attacks in the filtered alarms. Accuracy is improved by 89%.

Ching-Hao et al. (2009) proposed a co-training framework to leverage unlabeled data to improve intrusion detection. However the co-training framework provides lower error rate than single view method and thereby incorporating an active learning method to enhance the performance. Gao (2010) proposed a cluster algorithm based on semi-supervised approach. The result shows that the applicability of this method is much better than the others. Qiang and Vasileios (2005) introduced a new clustering algorithm fuzzy correctness based clustering on the concept of fuzzy correctness to calculate the similarity among data instances where seed points are dynamically assigned to each cluster dependency upon the fuzzy correctness value. The parameters such as number of seed points and neighbors does not affect the performance of detecting rate and False alarm rate is below 4%.

Semi-supervised clustering algorithm based on K-Means algorithm, is proposed by Wei et al. (2007). First, cluster undergoes dynamic merging and splitting process. Secondly, a small portion of labeled samples is used in the merging and splitting stage. Finally, the algorithm models the symbolic attribute values. The experimental result shows that the algorithm provides 94.42% with FP rate of 1.52% respectively. Lane (2006) combines the strength of both misuse and anomaly detection thereby developing a Partially Observable Markov Decision Process (POMDP) to determine the cost function. The difference between training and testing data is also eliminated with more flexible intrusion detection model and that can be constantly updated as new data found.
Four kinds of semi-supervised SVM methods namely SSDC-SVM, SSR-SVM, SSDK-SVM, SSOC-SVM have been used with the intention to improve the accuracy. It is being proven that SSDC-SVM and SSDK-SVM shows better accuracy and SSD-SVM has speed metrics too. Hierarchical RSS-DSS algorithm was proposed by Jimin et al. (2010) in which the feature of dynamic filtering of large dataset based on the training pattern. Such a scheme provides the method of training genetic programming on a dataset of half a million pattern within 15 min. The cost function determines the fitness function by providing effective solutions. Yuh-Jye and Olvi (2001) compares conventional SVM with (RSVM) reduced support vector machine. RSVM algorithm randomly selects the subset of data to obtain the non linear separated surface. He found that the in this reduced data set memory usage and computational time is very less for RSVM than conventional SVM when compared to the entire dataset.

**METHODOLOGY**

**Framework of proposed Hybrid IDS:** In Fig. 1, snort is installed in the computer within the network. Once it is installed it automatically captures the network packets that are passed over the network. In this, we include (KDD Cup 99, 2009) dataset together within the snort. Since the set of rules is predefined inside the snort. It performs the preprocessing steps as per rules. Snort gives the alert message according to the information stored in the database as tcpdump files. If any attack is found then the packet is dropped otherwise it can be taken as attack free data. Here we apply the anomaly based approach such as ALAD, PHAD, NETAD or LERAD to automate the IDS by capturing the attacks synchronizing with the network. If any suspicious traffic/attack is found, it analyses the exact cause of it and creates the signature and finally included within the rule set in a snort.

**Algorithm of proposed hybrid IDS:**

*Step 1:* Input data are taken from network packet  
*Step 2:* Implementing the data into Snort  
*Step 3:* Snort performs preprocessing and analyses whether the data is attacked or a normal one  
*Step 4:* Applying ALAD algorithm to detect the attack in the application layer  
*Step 5:* Applying LERAD algorithms to perform further improvement in rule structure  
*Step 6:* Finally we perform the detection process and drop the attack packet and the new rule is generated through intrusive packets

**Framework of proposed semi-supervised approach:** According to the Fig. 2, generate input from KDDCup99 to compare the performance of various existing algorithms used in the intrusion detection systems. First, the dataset is divided into training and testing data. Training data includes both the labeled data and unlabeled data. Using the labeled data the unlabeled data can be labeled. So this kind of approach is said to be semi-supervised approach. The labeled training data is applied to the SVM classifier and the model is generated. Then, change the SVM parameters by applying the RBF kernel function and generate the model for each tuning process. Apply the training
unlabeled data to SVM model as test data and results are generated for all models. Drop the records which does not satisfy the voting results. Include the changed label as predicted labels. Randomly generate 1000 data points to find the vector distance between each support vector and the data points. This process enables the most confidential data. Provide these new data instances with the trained labeled data. At last, include the unlabeled test data to the classifier and check the accuracy rate and its corresponding false alarm rate respectively.

Algorithm of proposed semi-supervised approach:

Step 1: Select the labeled data instances $D_{train}$ which has class labels, then train the SVM using this dataset and generate training model.

Step 2: Alter the SVM parameters and the selection of the kernel and its parameter $C$ determines the accuracy of SVM. Here a Gaussian kernel parameter $\gamma$ is used for the study. The best combination of $C$ and $\gamma$ is often selected by a grid search with exponentially growing sequences of $C$ and $\gamma$, where, $c \in \{2^{-5}, 2^{-3}, \ldots, 2^{13}, 2^{15}\}$ and $\gamma \in \{2^{-15}, 2^{-13}, \ldots, 2^{1}, 2^{3}\}$. Training SVM by using the $D_{train}$ generate training model.

Step 3: Test the same dataset $D_{train}$ by SVM with all generated trained models. Change the calls label of the $D_{train}$ by SVM predicted classes from all trained models based on majority voting. Now the dataset with changed class label $D_{train}$ is predicted.

Step 4: Train the SVM using $D_{train}$ with optimal selected parameters

Step 5: Test the dataset $D_{train}$ and predict the class label. combine $D_{train}$ and $D_{train}$ get actual trained dataset $D_{tr}$.

Step 6: Calculate the average distance $t$ of training data for every support vectors.

Step 7: Euclidean distance is calculated and the mean value is found.

Step 8: Randomly select some 500 data instance and calculate the vector distance.

Step 9: Provide a class label for extra generated using the trained SVM. Now the train dataset is combined of $D_{tr}$ and $D_{extra}$. This is called $D_{extra}$.

Step 10: Train SVM uses $D_{extra}$ then test the test dataset $D_{tr}$.

RESULTS AND DISCUSSION

Dataset description: Lippmann and Haines (2000) and Xiaojin (2008) analyzed the KDD cup dataset and provided the corresponding results. The 1998 DARPA Intrusion Detection Evaluation Program was prepared and managed by MIT Lincoln Labs. The objective was to survey and evaluate research in intrusion detection. Each record is categorized as normal or attack, with exactly one particular attack type. Table 1 specifies the four classes of attacks. The dataset which has been taken for the study are classified as follows:

- **Denial of service attack**: Here the attacker makes the traffic busy and access the normal user system and performs all sorts of vulnerability.
- **Probe attack**: The attacker gains the knowledge of the network and performs damage in future.
- **Remote to local attack**: The attacker uses the remote machine and causes some attacks to the local host machine.
- **User to root attack**: Using the local machine through sniffing password the attacker exploits damages to the remote machine.

Performance of proposed hybrid IDS: Hybrid IDS is developed to overcome the human interaction towards pre-processing. Most of the evaluation of intrusion detection is based on proprietary data and results are
not reproducible. To solve this problem, KDD Cup 99 (2009) has been used. Public data availability is one of the major issues during evaluation of intrusion detection system. Mixed dataset (real time + simulated) has been used for this study. Out of 500 data instances, 320 instances involved in the training phase and remaining 180 instances are taken for testing phase. The Fig. 3 represents the incoming data that are captured by snort in real time. Each data includes the source IP address, destination IP address, state of the packet and so on. Data can be analyzed with the help of the snort rules that are predefined within it. Figure 4 specifies the anomaly score for each packet.

Analysis is done based on the scenarios given below:

- Based on Snort
- Based on Snort + PHAD
- Based on Snort + PHAD + ALAD
- Based on Snort + ALAD + LERAD

**Performance of snort:** Snort is tested on real time traffic and simulated dataset using KDD Cup 99 (2009), Mahoney (2003) and Snort Users Manual, 2.6.1 (2006) (one week data including attack) of attacks detected on a daily basis from Fig. 5. The files have been downloaded from LAN network. Attack detected in daily order is shown in the

Fig. 4. Snort has detected 77 attacks out of 180 attacks without adding any anomaly based approach which is nearly 43% respectively.

**Performance of snort + PHAD:** Attacks detected by Snort, LERAD and NETAD on their own and results in the hybrid intrusion detection system (Snort + PHAD + NETAD) are shown in Fig. 6. It is understood that after adding PHAD with Snort it detects better than before. The number of attacks detected by Snort increases from 77 to 105 in Snort + PHAD version of IDS which is nearly 58% respectively.

**Performance of snort + PHAD + ALAD:** When PHAD and ALAD are added to the snort it detects more attacks than before. It is clearly shown from the graph Fig. 6 that the number of attacks increases while adding PHAD and ALAD with Snort the IDS becomes powerful. The number of attacks detected by Snort + PHAD increases from 105 to 124 in Snort + PHAD + ALAD which is nearly 68% from Fig. 7. The main reason is Snort detects the attacks based on rule definition files but PHAD and ALAD detects using packet header and network protocol.
Fig. 7: Number of attacks detected by snort + PHAD + ALAD on a daily basis

Fig. 8: Number of attacks detected by snort + ALAD + LERAD on a daily basis

Table 2: Results of attacks detected by snort, PHAD, ALAD and LERAD

<table>
<thead>
<tr>
<th>Anomaly based approach</th>
<th>Detection rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Snort</td>
<td>77/180 (43%)</td>
</tr>
<tr>
<td>Snort + PHAD</td>
<td>105/180 (58%)</td>
</tr>
<tr>
<td>Snort + PHAD + ALAD</td>
<td>124/180 (68%)</td>
</tr>
<tr>
<td>Proposed hybrid IDS (snort + ALAD + LERAD)</td>
<td>149/180 (83%)</td>
</tr>
</tbody>
</table>

- **Proposed hybrid IDS (snort + ALAD + LERAD):** Attacks detected by Snort, ALAD + LERAD on their own and results in the hybrid intrusion detection system (Snort + ALAD + LERAD) are Fig. 8. After adding Snort + ALAD + LERAD, the ids give better results when compare with other methods. The number of attacks detected by Snort + PHAD + ALAD increases from 124 to 149 in Snort + ALAD + LERAD (hybrid ids) version of the IDS which is nearly 83% respectively. Table 2 shows the overall performance of snort based statistical anomaly detection algorithms.

**Performance of proposed semi-supervised approach:** It is very hard for IDS to collect and analyze the data. For this approach rule based techniques can be used. But if there is any little change in the data then the rule seems to be meaningless. To accomplish this task, we go for semi-supervised approach. In supervised approach labeled data can be taken for training phase and unlabeled data has been taken for testing phase. Usually the network data are unlabeled. It needs the security experts to label the unlabeled data which is expensive and time consuming. Because supervised approach needs the formal labeling of data to analyze whether the testing data is attacked or a normal one. But it is not realistic in real time.

So semi-supervised approach is considered as most significant one. It requires only a small quantity of labeled data with large amount of unlabeled data. This method is done on the assumption. By analyzing the distance between the data points labeling is done. These data points are considered as most confidential data. In turn these confident data are taken as training data and corresponding testing data is applied to label the unlabeled one.

Figure 8 shows the values of while tuning the parameters of SVM. The values are used between 0 to 1. This process continues till the bias value is same for many trials as shown in the following Fig. 9 and 10.
Totally 5000 dataset are taken for the study. In that 2500 are considered as training phase and remaining 2500 has been applied as testing phase dataset. Training phase includes both the labelled and unlabeled data together. Table 3 shows the accuracy of 98.88% and FAR of 0.55% respectively.

Fig. 11 shows the predicted labels after applying the testing data. Based on this result, the accuracy will be calculated. Four classes like dos, probe, r2l and u2r label has been labelled as 1, 2, 3 and 4 respectively.

In Fig. 12 and 13, the proposed semi-supervised approach is compared to the existing algorithms like RSVM, PCKCM and FCC. The proposed semi-supervised approach shows better performance in terms of accuracy and reduced false alarm rate.

**CONCLUSION AND RECOMMENDATIONS**

Mixed dataset has been used to snort based approach and KDD cup dataset is used for the semi supervised approach. To solve the overwhelming problem of supervised and unsupervised methods, the semi supervised approach has been carried out. Any number of unlabeled data can be labeled using this approach. Several tests were made in the system and overall significant results were achieved. Evaluation is
based on accuracy and false alarm rate. Proposed Hybrid IDS detects 83% of attacks in one week data and proposed semi supervised method shows better accuracy and reduced false alarm rate. Through this approach the overwhelming problem of using supervised and unsupervised method can be solved. The proposed two methods are simple and can be used in real time efficiently.

In future, a new approach has to be done regarding detection of on DOS attacks and corresponding intrusion prevention system must be designed with all necessary security measures.

REFERENCES


