Clonal Selection Approach for Network Intrusion Detection

Felix T.S. Chan, Anuj Prakash, R.K. Tibrewal, and M.K. Tiwari

Abstract—Network attacks corrupt or destroy the information and services and it affect the integrity and confidentiality of network system. It can be classified as Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L) and Probing attack. In network security, a lot of researchers attract towards intrusion attacks and normal network traffic classification problem as it is a very challenging and critical problem. This paper presents an Artificial Immune System based approach for anomaly based network intrusion detection system. Artificial Immune System (AIS) algorithm is Meta-heuristic method which is used for clustering and pattern recognition. In addition, this article explains that the Clonal Selection Classification Algorithm (CSCA) can be applied to anomaly based network intrusion detection system and it can attain better solution only with very less number of antibodies. This model is compared with other approaches like Naïve Bayes, Random Tree and Support Vector Machine (SVM) that have been used previously to solve the same problem.

Keywords—Artificial Immune System, Clonal Selection Algorithm and Anomaly Detection formatting, Network Intrusion Detection System,

I. INTRODUCTION

NORMAL. Network attacks can be characterized as a set of malevolent activities to disrupt, refute, distort or destroy the important information as well as services in networking of computers. Network attacks contain a wider range which is from annoying email to intrusion attacks on sensitive data. It also includes the damaging of computer information systems and attack on critical network infrastructure. A network attack can be classified in four categories. These categories are: Denial of Service Attack (DoS), User to Root Attack (U2R), Remote to Local Attack (R2L) and Probing attack [1].

These attacks are also called as hacking but initially it was used for finding the solutions of the technical problems in networks but in later days it is referred as the malicious activity by which the hacker can steal the confidential data or destroy the corporate network and they can also prevent the user to access the system.

The hacking for the criminal purpose is also called by various names like cyber espionage cracking, and cybercrime. To mitigate these criminal practices, intrusion detection techniques are used.

The intrusion detection can be categorized in the following category: misuse detection, specification based detection and anomaly detection. Misuse detection and specification based detection are fully effective to reveal known attacks. If the computer coded sequence is not available or not predefined by the security experts, these approaches cannot work effectively. There are also some other difficulties in applying these approaches as to define the signature of all the attacks and wrongly defined signature which can be resulted as false alarm. Therefore the effectiveness of these problems is not very much in the case of new detection. To overcome these limitations, anomaly based intrusion detection technique has been introduced [2]. The main aim of anomaly detection is to recognize such cases that are having some different nature within data. For intrusion detection, anomaly detection is a significant approach as it can detect network intrusion, financial fraud, and other events which are happened very rarely but having more importance. The typical anomaly detection system has been shown in fig 1. The main benefit to use the anomaly detection is its inherent capability to find novel attacks which cannot be identified by misuse detection. The researchers have applied various techniques like case based reasoning, logistic regression, neural network, fuzzy logic, decision trees, discriminant analysis, and genetic programming etc. for fault prediction. Nowadays, the researchers are drowning in the ocean of meta-heuristics; therefore there is a need to find such a meta-heuristic which can easily detect network intrusions within less time. To keep that in mind, CSCA has been introduced for network intrusion detection.

In the present article, anomaly based network intrusion detection has been considered as the focus of study. To detect network intrusion is very complex problem due to its wide range, variability, abuse of functionality, vulnerability; therefore, anomaly detection has been employed for network

Felix T.S. Chan is with the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, HK (852-27666605; mfchan@inet.polyu.edu.hk ).

Anuj Prakash is with the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hung Hom, HK (anujpra@gmail.com).

R.K. Tibrewal is with the Department of Industrial Engineering and Management, Indian Institute of Technology Kharagpur, Kharagpur, India (rupak.kumar04@gmail.com).

M.K. Tiwari is with the Department of Industrial Engineering and Management, Indian Institute of Technology Kharagpur, Kharagpur, India (mkt09@hotmail.com).
intrusion detection. Anomaly detection has inherent capability to find the novel attacks and these attacks are not detected by other detection techniques. 

![Diagram of Anomaly Detection System]

Fig. 1 A Typical Anomaly Detection System

In this paper, CSCA, a novel approach, which is another form of AIS for negative selection, has been employed for intrusion detection. CSCA is inspired by biological immune system. The biological immune system has the capability of distinguishing between the normal component of organism and foreign material. These foreign materials can be called as antigens and these are cause of harm to human body. The capability of detection has been imitated in AIS for fault detection and this is the basic principle for clonal selection classification algorithm. The complexity of network intrusion detection has been demonstrated by a numerical example and CSCA has been applied for showing the effectiveness of the algorithm. To show the supremacy of the proposed algorithm, the results obtained are also compared with other techniques like naïve bayes, Random tree and SVM. The numerical results show the efficacy of the proposed algorithm.

The rest of the paper is organized as follows: Section II summarizes the literature in the field of anomaly based network intrusion detection. Section III delineates the complexity of the problem and the performance measure for the classifier. Section IV traces the detailed Clonal Selection Classification Algorithm. In Section V, the experimental results and comparison with other approaches is given. Finally, the paper has been concluded in section VI.

II. LITERATURE REVIEW

In literature, network intrusion detection problem has gained the attention of researchers and practitioners in late of 20th century. In 1980, Anderson’s seminal report has given an insight towards intrusion detection approaches [2]. E. Eskin [3] has proposed a mixture model for anomaly detection over noisy data using learned probability distributions. The decision support system is consistent for information security planning as well as cyber security risk. N. Zhang, W. Yu, X. Fu, S.K. Das [4] have proposed two novel algorithms for reputation establishment in anomaly detection -one for systems solely consisting of smart insiders and the other for systems in which both smart insider and attackers are present. Zeng et al. [5] studied a feedback negative selection algorithm for anomaly detection. The proposed approach uses the feedback technique to adapt the varieties of self/non-self space and build the appropriate profile of the system based on some of self elements.


III. PROBLEM DESCRIPTION

In computer networking, network intrusion detection problem has gained a lot of focus as it affects the integrity, confidentiality and availability of sensitive data, information systems as well as network infrastructure. For intrusion detection, the anomaly detection is the most capable detection approach due to its capability to find novel attacks and simultaneously it also deals with misuse detection. In this novel approach, training data and testing data forms the major input for a classifier. In this dataset, an appropriate number of records for training and testing are included quality of dataset, which is another important factor, also affects the performance measures. In other words, there should be very less or no redundancy for obtaining the better performance of the classifier. In the present study, to evaluate the performance of a classifier, four important performance measures have been taken into account. These measures are precision, sensitivity, specificity, and accuracy. These performance measures are evaluated by using the confusion matrix. The confusion matrix is the key factor in anomaly based network intrusion detection problem and it shows the prediction on instances. The confusion matrix has been shown in Table I.

From Table I, it is clear that the positive prediction is used for anomaly data where as negative prediction is used for normal data. The data is labeled in four categories: True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN). For instance, if anomaly data is predicted as anomaly, it
will be labeled as TP and if normal data is predicted as anomaly, it will be labeled as FP. After labeling of data, the above mentioned performance measures to evaluate the classifier are calculated. The definition of these performance measures with mathematical expressions has been given below.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>CONFUSION MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Confusion Matrix</td>
</tr>
<tr>
<td></td>
<td>Anomaly</td>
</tr>
<tr>
<td>Known</td>
<td></td>
</tr>
<tr>
<td>Label</td>
<td>Anomaly</td>
</tr>
<tr>
<td></td>
<td>Normal</td>
</tr>
</tbody>
</table>

1) Precision: It can be defined as the percentage of positive predictions that are correct and it can be expressed in mathematical form as:

\[
\text{Precision} = \frac{TP}{TP + FP}
\]

2) Recall/Sensitivity: The percentage of positive labeled instances that are predicted as positive and are called sensitivity. The mathematical formulation is as follows:

\[
\text{Recall} = \frac{TP}{TP + FN}
\]

3) Specificity: The specificity shows the percentage of negative labeled instances that are predicted as negative and is expressed as:

\[
\text{Specificity} = \frac{TN}{TN + FP}
\]

4) Accuracy: It is the main performance measure for evaluation of the classifier. It can be delineated as the percentage of correct prediction among the whole data set. It can be calculated as:

\[
\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}
\]

The main objective of the present study is to develop a classifier, which is worked on the basis of anomaly detection approach, to classify the data in correct manner. The classification should also improve the aforementioned performance measures. Therefore, this classification can easily detect the network intrusion. To solve such a detection problem, a Clonal Selection Classification Algorithm (CSCA) has been employed. Next section describes the proposed algorithm in detailed manner.

IV. CLONAL SELECTION CLASSIFICATION ALGORITHM

The proposed algorithm is inspired by biological immune system to detect the unusual data. To demonstrate the inherent capabilities of the proposed algorithm, the basic theory of immune system with clonal selection is given below.

A. Clonal Selection Theory

An immune system identifies the harmless and harmful antigens or disease causing bacteria, viruses, incompatible blood cells foreign materials etc and simultaneously protects the body from these harmful foreign materials. The immune system is divided into two main parts, innate immunity and acquired immunity. The innate immunity is the basic elements of immune system such as skin; whereas, acquired immunity is acquired through antigens and provide the immune response accordingly the severity of antigens [12]. The clonal selection theory is inspired by acquired immunity and it diversifies the antibodies to shield the organs from the attack of disease causing harmful material [13], whereas antibody is a produced by white blood cells or B lymphocyte cells and each antibody is unique for a specific type. According to clonal selection theory, the antibodies recognize the antigens due to specificity and chemically bind, replicate and produce more antibodies. To provide the variations in replicated cells, the mutation will occur. The selection of antibodies for proliferation or replication is based on Darwin’s rule of survival of fittest. The immune system also has the innate ability to respond to the foreign materials and regulate the system itself to protect the organism.

In this paper, the capability of differentiating between self and non-self antigens has been employed for anomaly based network intrusion detection.
B. Steps of Algorithm

In this section, the working mechanism of Clonal Selection Classification Algorithm (CSCA) for anomaly based network intrusion detection problem has been described in detailed manner (Figure 2). The full step-wise procedure of CSCA for network intrusion detection problem is given below.

**STEP 1:** Convert the training and testing data into binary form.

**STEP 2:** Initialize the antibody pool of size N having equal bits to as in antigens.

**STEP 3:** Calculate the affinity of each antigen with respect to all antibodies.

**STEP 4:** Select the best antibody for each antigen i.e. the antibody having highest affinity.

**STEP 5:** Mutate the selected antibodies and classify them into two pool i.e. normal pool and anomaly pool.

**STEP 6:** Clone all the antibodies of two pools and put them in memory pool.

**STEP 7:** Select an antigen from testing data and calculate the affinity of that antigen with respect to all the antibodies in both normal and anomaly pool.

**STEP 8:** Select the two best antibodies for each antigen from the normal and anomaly pool.

**STEP 9:** Repeat step 2 to step 8 and classify the problem data set after the successful completion of training of classifier.

V. RESULTS AND DISCUSSIONS

In the present network intrusion detection problem, a clonal selection classification algorithm has been successfully implemented to classify the data. The proposed algorithm has been applied on KDD CUP-99 dataset to show the efficacy and compared with other recognized approaches to show the supremacy over other algorithm. The KDD CUP-99 dataset and its features with the performance measures obtained by applying CSCA are as follows:

**A. KDD CUP-99 Dataset**

The anomaly detection of novel attacks allured a lot of researchers and many researches were conducted in the last decade. Various researches have been conducted to mitigate the different weaknesses of intrusion detection systems, which is signature based. To evaluate the accuracy of different classifiers, KDDCUP’99 is the dataset which is used by most of the researchers. The detailed description of KDDCUP’99 data set is identical as [14]. In this dataset, there two types of data: training data and testing data. The training data have 24 data of attack whereas testing data contain 14 types of data. On applying CSCA on KDD-CUP 99 dataset the encouraging results are obtained for various performance measures which are depicted in Table II. The various features of the presenting dataset (KDDCUP’99) are divided in three different groups [15]. These groups are:

**TABLE II**

<table>
<thead>
<tr>
<th>MEASURES</th>
<th>PERCENTAGE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>100</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>90.21</td>
</tr>
<tr>
<td>Specificity</td>
<td>100</td>
</tr>
<tr>
<td>Accuracy</td>
<td>94.6</td>
</tr>
</tbody>
</table>

**TABLE III**

<table>
<thead>
<tr>
<th>Approach</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>J48</td>
<td>93.82</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>81.66</td>
</tr>
<tr>
<td>NB Tree</td>
<td>93.51</td>
</tr>
<tr>
<td>Random Forest</td>
<td>92.79</td>
</tr>
<tr>
<td>Random Tree</td>
<td>92.53</td>
</tr>
<tr>
<td>Multi-Layer Perception</td>
<td>92.26</td>
</tr>
<tr>
<td>SVM</td>
<td>65.01</td>
</tr>
<tr>
<td>CSCA</td>
<td>94.6</td>
</tr>
</tbody>
</table>

1) **Basic features:** These are the most important features which summarized most of the attributes of a TCP/IP connection. These features are mainly responsible for the delay in detection and this delay is implicit.

2) **Traffic features:** These features are related to service connection and these are evaluated in time window. It can be classified into two categories, which are as follows:
   a) “Same host” features: This feature observed only the connection host that it is having the same destination as it was before 2 seconds.
   b) “Same service” features: In this feature, only the service provided in last seconds are observed that it is same as the current connection or not.

3) **Content features:** These features are required for detecting R2L and U2R attacks and it observes any suspicious behavior in pattern of collected data. For instance, it is required for observing about number of failed login attempts. From Table 2, it is clear that the obtained value of precision and specificity is 100%. Therefore, it can be said that the classifier is 100 % correct in classifying the normal class, as the value of FP is coming out to be zero. The accuracy is 94.6% i.e. only 5.4% data are not correctly revealed. Thus it concluded that CSCA based classifier is enough sufficient for tackling such datasets.

The CSCA algorithm has been compared with other techniques like Random Tree, Multi-Layer perception, Support Vector Machines (SVM) etc. The comparison has been made on the basis of accuracy of the algorithm for classifying the data. The accuracy is an important performance measure as it shows the correctly labeled data. The data of accuracy for other approaches has been taken from [28]. The comparison for KDD-CUP 99 dataset has been shown in Table III. From this table it is clear that in the traditional technique SVM gives only 65.01% accuracy where as J48 provides 93.82% accuracy. The proposed algorithm classifies 94.60% data correctly. For better understanding, the results are graphically depicted in Figure 3 as a bar chart.
Another study has been carried out to show the effect of initial number of antibodies on the accuracy of classifier. The results have been depicted in graphical form in Figure 4. The result shows that the best result is obtained with 20 numbers of initial antibodies. On increasing by 1 the accuracy dropped by 0.10%. From the Figure 4, there is very less variation in accuracy up to 100 numbers of initial antibodies. There is a big drop in the curve at 200 numbers of initial antibodies and later on it improved slightly. Therefore, the graph can prove the inherent strength or capabilities of proposed algorithm for searching and achieving the better solution. The strength of proposed algorithm is that it can explore the search space and achieve better solution only with very less number of antibodies.

The proposed algorithm has been coded in MATLAB programming language and the experiment has been carried out on Intel core i3 2.13 Ghz processor. In addition, from the results, this conclusion is extracted that the proposed algorithm is having the supremacy over the existing approaches as it is having the strong exploring capability of search space.

VI. CONCLUSION

In the present article, classifier problem of network intrusion detection by applying anomaly based approach has been taken into consideration. To classify the problem dataset, clonal selection classification algorithm, which is worked on the basis of negative selection through artificial immune system, has been applied. The classifier has been evaluated on the basis of various performance measures like precision, sensitivity, specificity and accuracy. The proposed algorithm has been tested on KDD CUP’99. From the results, it is clearly shown that the quality of the data set is a key factor in classifying or detection of network intrusion. Clonal Selection Classification Algorithm (CSCA) has also been compared with other approaches to show the supremacy of the proposed algorithm. The present research has enough scope for future extension. Some other objectives can be taken into account for further research. The proposed algorithm can be modified by applying fuzzy logic for controlling the hyper mutation and proliferation rate. The problem can be modified by using a quality check over the redundant data. This research can also provide an insight for carrying on the new researches in network intrusion detection area.

ACKNOWLEDGMENT

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 510410). The authors also would like to thank The Hong Kong Polytechnic University Research Committee for the financial and technical support.

REFERENCES