International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 5 Issue XII December 2017- Available at www.ijraset.com

A Model for Intrusion Detection based on Negative Selection Algorithm and J48 Decision Tree

Sanjay Sharma[1], R.K. Gupta[2]
[1,2] *Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior (M.P.), India*

Abstract: ***Now a day's IDS becomes an important component of any network in the world of internet. An intrusion can be defined as a breach in the security system. For this reason, intrusion detection generally refers to the mechanisms that are developed to identify the breaches in security system. In recent times, data mining techniques have gained significant value in providing the valuable information which can enhance the decision making capability on identifying the intrusions. That's why, IDS with Data Mining has gone through several revisions in order to meet the current requirements for efficient detection of intrusions. A number of models have been proposed for enhancing the performance of the system. In order to enhance the performance, the paper presents a new hybrid model for intrusion detection. This hybrid model uses the combination of Negative Selection Algorithm (NSA) and J48 Classification Algorithm. The main concern of the proposed model is to increase the accuracy and decrease the false alarm rate.***
Keywords: Data Mining, Intrusion Detection System (IDS), Classification, NSA, J48 Decision Tree.

I. INTRODUCTION

Intrusion detection system has become an active area of research and development over the past few decades. This is mainly because of the rising of attacks on computer and network systems. An intrusion detection system monitors the activities of a given network and identifies that whether these activities are malicious (attack) or legitimate. An intrusion can be defined as "a set of events or a type of attack that attempt to compromise the integrity, availability or confidentiality of a computer or network resource" [1], and hence, intrusion detection generally refers the mechanisms that are developed to identify the breaches in the security of the systems. Current approaches to identify intrusions make use of data mining techniques such as naive Bayesian, decision tree, artificial neural network, support vector machine, fuzzy logic, genetic algorithm and so on. To a greater extent, IDS has justified its purpose but still there is a need for enhancement in order to meet the current requirements such as high detection accuracy, low false positives etc. Important functions of the IDS: monitor, detect and respond to illicit activities. In today's world of internet, IDS is increasingly becoming a vital component to secure the network.

Furthermore enhancements are done regularly in IDS to build up a more efficient system. So that it can detect the known and unknown attacks precisely. The purpose of this paper is to propose a new hybrid model for intrusion detection based on the combination of Negative Selection Algorithm and J48 Decision tree to improve the detection capabilities in terms of accuracy as well as false alarm rate. Also the proposed model is analyzed with some other data mining techniques in respect of performance evaluation for intrusion detection.

The rest of the paper is organized as follows: section II discusses the related works; section III presents the overview of intrusion detection systems; section IV describes the proposed methodology; section V explains the proposed model; experimental analysis is discussed in section VI and finally section VII concludes the paper.

II. RELATED WORKS

Om H and Kundu A [2] in 2012, proposed a hybrid model that combines k-Means and two classifiers: K-nearest neighbor and Naive Bayes to detect anomaly. Author used an entropy based feature selection algorithm for feature selection. Dewan Md. Farid et al [3] in 2010, given a model that uses Naive Bayesian classifier and ID3 algorithm for effective intrusion detection. Thakur M R and Sanyal S [4] in 2008, proposed a multi-dimensional approach for intrusion detection. Pathak V and Ananthanarayana V. S [5] in 2012, proposed a multi-threaded K-Means clustering approach in that they used six threads and all run in parallel. Barot V and Toshniwal D [6] in 2012, suggested a hybrid model that combines Naive Bayes and Decision Table Majority (DTM) approaches. Here author used correlation based feature selection method for the attribute selection.

III.OVERVIEW OF INTRUSION DETECTION SYSTEMS

This section contains a short overview of intrusion detection approaches, classification of IDS and categories of attacks.

A. Intrusion Detection Approaches

Intrusion detection systems can be divided in to two types according to the detection approaches:

1) *Misuse Detection:* Misuse detection is also referred as Signature-based Detection. In this approach, the known attack patterns are predefined. These predefined patterns act as signatures for the intrusions to be detected by the IDS. When a match found to the signature, an alarm gets generated. The major advantage of misuse detection is its higher accuracy to known attacks and fewer false alarms. The shortcoming of this approach is that it can detect only the known attacks.
2) *Anomaly Detection:* In anomaly detection, profiles for expected/normal behavior are defined in advance which is used to detect the anomaly within the system or network. A significant deviation from such defined expected/normal behavior beyond a certain level called threshold level is reported as anomaly. The advantage of this approach is the ability to detect novel or unknown attacks. The shortcoming of this approach is its high false alarm rate.

B. Classification of IDS

Intrusion Detection Systems can be classified into two main categories: *Host-based Intrusion Detection system (HIDS) and Network-based Intrusion Detection System (NIDS).*

1) *Host-based Intrusion Detection System (HIDS):* Host-based Intrusion Detection System mainly deals with the intrusion detection that takes place in a single host/computer system. For detection, the data is collected from a single host system. HIDS continuously monitors system logs, system based network traffic, integrity of the system, system calls and application actions [7]. Whenever, any unauthorized activity is detected, an alarm gets generated.
2) *Network-based Intrusion Detection System (NIDS):* Network-based Intrusion Detection System works on a large scale as it is installed in a network to detect intrusions. It monitors and analyzes the network traffic to protect the system from intrusions. NIDS analyses the stream of packets that flows across the network and based upon the observations it classifies the network traffic into malicious or normal.

C. Categories of Attacks

Following are the four major categories of attacks detected by IDS:

1) *Denial of Service (DoS):* In DoS attack, attacker prevents the legitimate users to access the services of a host or network resources by making resources too busy or full. Example: apache, neptune, back, mail bomb etc.
2) *Remote to Local (R2L):* Remote to Local attack is an attack, where a remote user tries to gain access to a local machine as a local user by sending packets to a local machine over the internet. An intruder explores the vulnerabilities of the system and exploits the privileges of a local user. Example: phf, guest, xlock etc.
3) *User to Root (U2R):* In User to Root attack, the attacker (non-privilege user) tries to abuse the vulnerabilities in the system in order to gain the super user privileges. Here, a local user tries to gain access to a system as a root user. Example: xterm, perl, Fd-format etc.
4) *Probing:* Probing is an attack in which attacker scans a machine or a networking device in order to gain knowledge about its vulnerabilities or weaknesses which may be exploited further to compromise the system or networking resources. Example: port sweep, nmap, mscan etc.

IV. METHODOLOGY

This section contains the description of two algorithms: Negative Selection Algorithm (NSA) and J48 Decision Tree classification algorithm. The Negative Selection Algorithm is inspired by the self/non-self discrimination behavior in the immune system. The NSA was designed for change detection, intrusion detection, similar pattern recognition and two class classification problems. While the J48 Decision Tree is an effective classification algorithm which generates a decision tree and on the basis of which rules are generated. It uses the highest information gain feature to classify the instances.

A. Negative Selection Algorithm (NSA)

NSA is one of the most popular Artificial Immune System models that have attracted much attention from researchers. Forrest et al. proposed Negative Selection Algorithm, which is based on the concept of self/non-self discrimination behavior in the immune system. It is motivated from the fact of negative selection of T cells in the thymus [8] and worked upon the immune system philosophy to recognize unknown antigens or non-self cells without reacting to self cells. It produces a set of self-patterns (strings) that define the normal network patterns (strings). This set can easily identify non-self patterns and marked them as non-self or anomalous. If any random pattern matches with any self-pattern then it is removed so that it cannot becomes a detector. A detector is a pattern or a set of patterns that only recognizes the complement of self-patterns. Those patterns that do not get matched with self patterns become detectors and signify non-self. Afterward, these detectors are used to identify non-self or anomalies. These detectors examine incoming patterns and if any new pattern matches with the detector, then this represent the detection of an anomaly. The NSA was motivated by the negative selection process occurring within the Natural Immune

System (NIS) [9]. The major component of NSA is to generate a set of detectors. Algorithm given below presents a formal description for this mechanism (*Detector Generation*):

Detector Generation

Input: $Data_{self}$
Output: Repertoire
Repertoire ← null
while (¬ StopCondition ( ))
Detectors ← GenerateRandomDetectors ( )
for ($Detector_n \in$ Repertoire)
if (¬ Matches ($Detector_n$, $Data_{self}$))
Repertoire ← $Detector_n$
end
end
end
return (Repertoire)

Conceptually, it is illustrated in Fig. 1 given below:

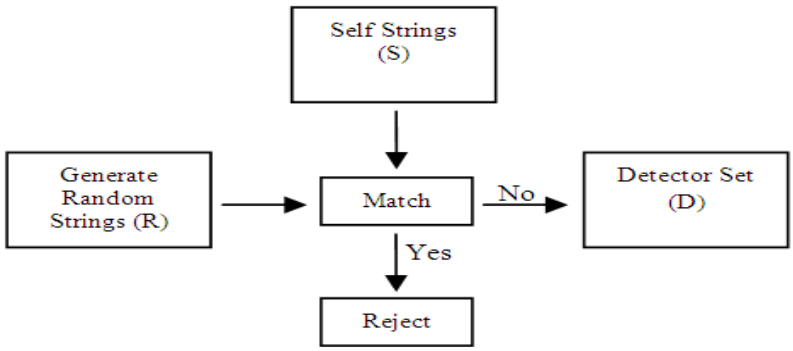

Fig. 1 Censoring phase of Negative Selection Algorithm

B. J48 Algorithm

J48 Decision Tree is a classification algorithm which is based on C4.5 classification algorithm [10] [11]. A decision tree is a predictive machine learning model that decides the value of a dependent variable (target value) of a new instance based on the values of various attributes of the available data. Different attributes are represented by the internal nodes of a decision tree, possible values that these attributes can have in the examined instances (samples) are represented by the branches between the nodes, whereas the terminal nodes represent the final value (classification) of the dependent variable or target value.

The attribute which is to be predicted is known as dependent variable because its value depends upon the values of all the other attributes. In the dataset, the other attributes which assist in predicting the value of the dependent variable (target value), are known as the independent variables.

The J48 Decision Tree classification algorithm follows the following simple steps. For the classification of a new item, it first needs to construct a decision tree based on the values of the attribute of the available training data. Hence, whenever a set of items (training dataset) encounters to it, it identifies the attribute that most clearly differentiate the various instances. This feature that gives the most information about the data instances so that it can classify the data instances the best is said to have the highest information gain. Now, if there is any value among the possible values of this feature for which there is no ambiguity, then it terminates that branch and assigns the target value that it has obtained. For other cases, it then look for another attribute that gives it the highest information gain. So it continue in this way until it either get an unambiguous decision of what combination of attribute provides it a particular target value or it run out of attributes. In the condition, when it run out of attributes or if it cannot get an unambiguous outcome from the available data, it assign this branch a target value that the most of the items posses under this branch [12].

A formal description for the J48 Decision Tree algorithm is given below:

1) Check if algorithm met any termination criteria.
2) Find the normalized information gain ratio for all attributes.

3) Choose the best attribute with the highest information gain.
4) Create a decision node on the basis of best attribute in step 3.
5) Split the dataset on the basis of newly created decision node in step 4.
6) Call the algorithm for all sub-dataset in step 5 to get a sub-tree (recursive call).
7) Add the tree obtained from step 6 to the decision node in step 4.
8) Return tree.

J48 Decision Tree

Input: an attribute valued dataset D
Tree = null
if D is "pure" or other stopping criteria satisfies then
terminate
end if
for all attribute a ∈ D do
Find normalized information gain ratio if we split on a
end for
a_{best} = Attribute with the highest information gain
Tree = Create a decision node that tests a_{best} in the root
D_i = Sub-datasets split from D based on a_{best}
for all D_i do
$Tree_i$ = J48(D_i)
Add $Tree_i$ to the concerned branch of the Tree
end for
return Tree

J48 Decision Tree uses the information gain as a splitting criterion and to find out the information gain entropy is used. Entropy can be defined as a measure of disorder of data. The Entropy for $\vec{p}$ can be calculated as:

$$\text{Entropy}(\vec{p}) = -\sum_{i=1}^{n} \frac{|p_i|}{|\vec{p}|} \log\left(\frac{|p_i|}{|\vec{p}|}\right)$$

Conditional Entropy can be calculated as:

$$\text{Entropy}(i|\vec{p}) = \frac{|p_i|}{|\vec{p}|} \log\left(\frac{|p_i|}{|\vec{p}|}\right)$$

At last Information Gain can be calculated as:

$$\text{Gain}(\vec{p}, i) = \text{Entropy}(\vec{p} - \text{Entropy}(i|\vec{p}))$$

V. PROPOSED MODEL

The proposed model is the combination of Negative Selection Algorithm and J48 Decision Tree Classification Algorithm which will enhance the efficiency of IDS as compare to the other existing IDS models. The model shown in Fig. 2 includes the preprocessing of KDD dataset, applying of the negative selection algorithm and then J48 classification algorithm, then evaluating the performance of this proposed model and finally visualizing the performance.

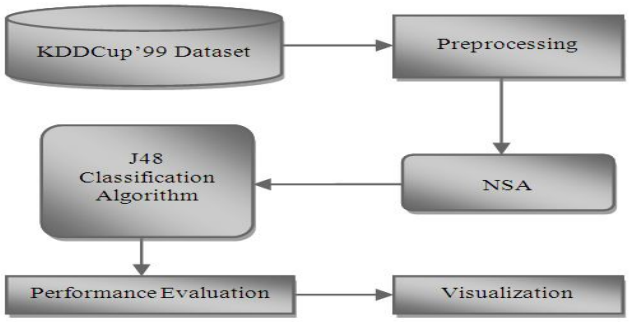

Fig. 2 Block diagram of proposed model

A. Description of the Proposed Model

1) *KDDCup'99 Dataset:* The KDDCup'99 Dataset is the most widely used dataset in the field of intrusion detection. Most of the researchers prefer this dataset to evaluate the performance of the IDS models. In 1998, DARPA intrusion detection evaluation program, a simulated environment was established by the MIT Lincoln Lab to acquire raw TCP/IP dump data for a LAN and the objective was to evaluate the performance of various intrusion detection methods [13]. The KDD99 dataset contains a standard set of data which includes a broad variety of intrusions [14].The dataset consists of 42 attributes in which 42nd attribute labels the connection as normal or a type of attack [15]. As the KDDCup'99 dataset consists of a large number of data records so we take a part of the dataset for the proposed model. There are 22 different types of attacks that are categorized into four main types tabulated in Table I.

TABLE I: DIFFERENT TYPES OF ATTACKS AND THEIR CATEGORIES

Category	Types of Attacks
Denial of Service (DoS)	neptune, land, back, pod, smurf, teardrop
Remote to Local (R2L)	ftp_write, guess_passwd, multihop, imap, phf, spy, warezmaster, warezclient
User to Root (U2R)	buffer_overflow, perl, rootkit, loadmodule
Probe	ipsweep, portsweep, nmap, satan

2) *Preprocessing:* This phase modify the selected part of the KDD dataset to make the classification simple. Here, the preprocessing of dataset is done by labeling the prediction class into four different types of attacks i.e. dos, r2l, u2r and probe.

3) *Performance Evaluation:* The performance evaluation phase evaluates the performance of the proposed IDS model by calculating the following terms:

a) *True Positive Rate (TPR):* $TPR = \frac{TP}{TP+FN}$

b) *False Positive Rate (FPR):* $FPR = \frac{FP}{TN+FP}$

Where the terms TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative) can be defined as follows [2]:

i. *True Positive (TP):* Number of malicious records that are correctly classified.
ii. *True Negative (TN):* Number of normal records that are correctly classified.
iii. *False Positive (FP):* Number of normal records that are incorrectly classified as attacks.
iv. *False Negative (FN):* Number of attack records that are incorrectly classified as normal.

A confusion matrix using the above depicted terms (TP, TN, FP and FN) can be represented as shown in Table II.

TABLE II: CONFUSION MATRIX

Actual	Predicted Normal	Predicted Attack
Normal	True Negative (TN)	False Positive (FP)
Attack (intrusion)	False Negative (FN)	True Positive (TP)

Here, the row represents the actual class of an instance where as the column represents the predicted class of the instances. Because of its tabular form, it is easy to understand the performance of the model.

4) *Visualization:* This phase visualizes the results of the proposed IDS model by means of graph, text, table etc. By observing the visualized results, one can easily understand the performance of the model.

VI. EXPERIMENTAL RESULT AND ANALYSIS

This segment contains the experimentation results obtained from the proposed IDS model and also it is compared with another classifier called Bayes Net. A subset of the KDDCup'99 dataset is used for the experiment purpose and it contains 292300 instances. The Java programming language is used to develop the system. Proposed intrusion detection approach is

implemented to detect four different classes of attacks and a normal class from the dataset i.e. DoS, R2L, U2R, Probe and normal. From experiment, it has been observed that the proposed model gives more reasonable results than the Bayes Net based approach. It performs well in terms of accuracy also. So, it can be said that the proposed approach (NSA+J48 Decision Tree) is more effective than the Bayes Net based approach.

Comparison of TPR and FPR between the two approaches is shown in Table III.

TABLE III: CLASS WISE COMPARISON OF TPR AND FPR

CLASS	NSA + J48		Bayes Net	
	TPR	FPR	TPR	FPR
DoS	1.000	0.000	0.993	0
R2L	0.989	0.000	0.984	0
U2R	0.667	0.000	0.769	0.006
Probe	0.987	0.000	0.986	0
Normal	0.999	0.001	0.991	0.009

Comparison of correctly classified instances and incorrectly classified instances between the NSA+J48 based approach and Bayes Net based approach is shown in Table IV.

TABLE IV: COMPARISON OF CORRECTLY & INCORRECTLY CLASSIFIED INSTANCES

Approach	Correctly Classified	Incorrectly Classified
NSA + J48	292188	112
Bayes Net	287878	4422

Fig. 3 shows the comparison of accuracy between the NSA+J48 based IDS model and the Bayes Net based IDS model where NSA+J48 exceeds Bayes Net with higher value of accuracy.

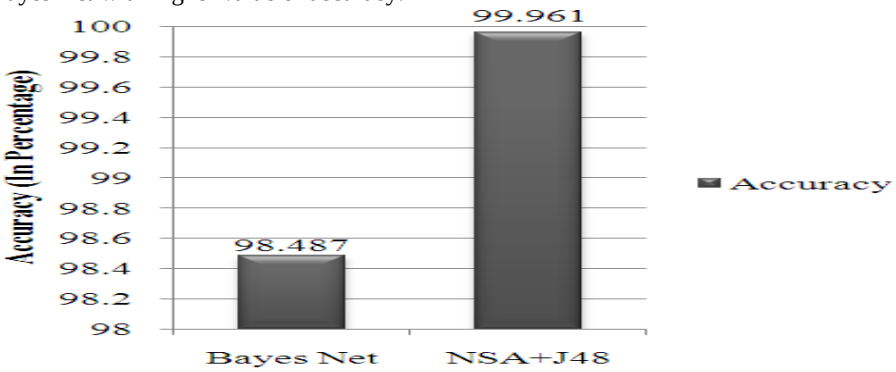

Fig. 3 Accuracy of Bayes Net and NSA+J48

When the proposed (combination of NSA & J48) IDS model is compared with the other prevalent data mining approaches used for intrusion detection then the maximum accuracy result was achieved by the proposed model. The comparison is shown in Fig. 4.

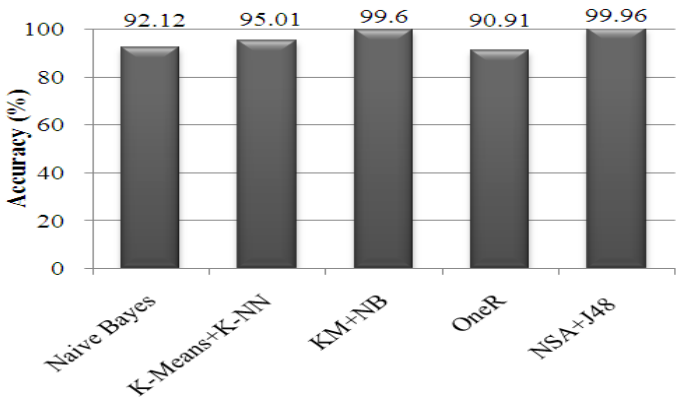

Fig. 4 Accuracy comparison with existing techniques

VII. CONCLUSION

This paper presents an efficient hybrid approach for intrusion detection by making use of Negative Selection Algorithm along with J48 Decision Tree classification algorithm. When the proposed approach is compared with another classification approach called Bayes Net then the proposed approach gives better accuracy and detection rate for different types of attacks discussed in the paper and also reduces the false alarm rate. Also in terms of accuracy, it proves itself as the most efficient approach when compared with some other existing techniques of data mining used for intrusion detection. In future research work, focus will be on improving the detection rate for U2R attack to make more efficient IDS model.

REFERENCES

[1] Kapil Wankhade, Sadia Patka, Ravindra Thool, "An Efficient Approach for Intrusion Detection Using Data Mining Methods", In Proceedings of 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), IEEE 2013, pp.1615-1618.
[2] Hari Om and Aritra Kundu, "A Hybrid System for Reducing the False Alarm Rate of Anomaly Intrusion Detection System", In proceedings of First International Conference on Recent Advances in Information Technology (RAIT), IEEE 2012.
[3] Dewan Md. Farid, Nouria Harbi and Mohammad Zahidur Rahman, "Combining naive bayes and decision tree for adaptive intrusion detection", International Journal of Network Security & Its Applications (IJNSA), Volume 2, Number 2, April 2010.
[4] Thakur M R and Sanyal S, "A Multi-Dimensional approach towards Intrusion Detection System", International Journal of Computer Applications 48(5):34-41, June 2012.
[5] Pathak V and Ananthanarayana V. S, "A novel Multi-Threaded K-Means clustering approach for intrusion detection", Software Engineering and Service Science (ICSESS), IEEE 3rd International Conference on 22-24 June 2012, pp. 757 - 760.
[6] Barot V and Toshniwal D, "A New Data Mining Based Hybrid Network Intrusion Detection Model", IEEE International Conference on 18-20 July 2012.
[7] V Jaiganesh, S Mangayarkarasi and P Sumathi, "Intrusion Detection Systems: A Survey and Analysis of Classification Techniques", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 4, April 2013, pp.1629-1635.
[8] https://en.wikipedia.org/wiki/T_cell#Negative_ selection.
[9] Delona C Johny, Haripriya P V and Anju J S, "Negative Selection Algorithm: A Survey", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 6, Issue 4, April 2017, pp.711-715.
[10] N S Chandolikar and V D Nandavadekar, "Efficient Algorithm for Intrusion Attack Classification by Analyzing KDD Cup 99", In Proceedings of 2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN), IEEE 2012, pp. 1-5.
[11] Abdulrazaq Almutairi and David Parish, "Using classification techniques for creation of predictive intrusion detection model", In proceedings of 2014 Ninth International Conference for Internet Technology and Secured Transactions (ICITST), IEEE 2014, pp. 223-228.
[12] http://www.d.umn.edu/~padhy005/Chapter5.html.
[13] "DARPA Intrusion Detection Evaluation Data Set 1998", MIT Lincoln Laboratory. https://ll.mit.edu/ideval/data/1998data.html.
[14] The KDD Archive. KDD Cup 1999 data set. http://kdd.ics.uci.edu/databases/kddcup99/kdd cup99.html.
[15] Mradul Dhakar and Akhilesh Tiwari, "A New Model for Intrusion Detection based on Reduced Error Pruning Technique", I. J. Computer Network and Information Security, 2013, pp. 51-57.