Intelligent Model for Network Attack Identification

Richa Sharma, Priyanka Sharma
Raksha Shakti University
Gandhinagar, Gujarat, India

Abstract

Detection of the network attack proficiently needs to capture the large amount of the traffic as the dump that needs to be studied. This implies that the very huge amount of the network traffic generated is collected from the transactions that takes place through the network. The identification of the network attack is then performed on this network traffic thus obtained. The identification of the network attack is similar to that of the intrusion into the system which is obtained from the analysis of the data from the traffic of the network. The intelligent approach is thus needed to find the intrusion from the large amount of the dump data which makes the predictions similar to that of the intrusion detection system. For the same the NSL-KDD shall be used for the experimental purpose as it is incorporated with the large amount of the data from network, features, testing dataset, training dataset etc. in this paper the hybrid algorithm is developed which is based on generating less false alarm rate, and it can withstand with the threshold level of intrusion identification from the predefined datasets and this is based on optimized features obtained through the process of preprocessing of the dataset. The hybrid algorithm shall be having the enhanced time complexity, computational speed, efficient identification of the network-based attack. The proposed hybrid algorithm shall also be dealing with the issues of the false positive alarm and negative rates. In the proposed hybrid algorithm firstly the data of network traffic is refined using the vote algorithm and then proposed hybrid algorithm is the combination of algorithms like naïve bayes, random tree, and many more.

Keywords: Network, Detection, Intelligent, Attack Identification

1. Introduction

The intrusion detection is classified into two classes which is as follows: 1. Signature based and 2. Anomaly based intrusion detection system. The signature-based identification of the network intrusion is based on the signatures generate by the attack but the core drawback of the signature-based detection is that the attacks having the new signature remains undetected. While the anomaly-based detection identifies the attack based on the behaviour of the attack into the system and if the novel attack attempts to intrude into the system or network its get identified through the behaviour and patterns which is already identified. The valuable information always lures the attackers to intrude into the system. The attacker always tries to get into the server, system for get information which is the result of the vulnerability present into the system.

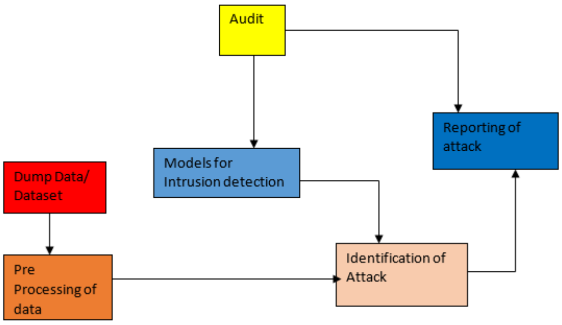


Figure 1. Intrusion Identification System

Reason for the attack is due to vital information such as online banking, confidential messages etc present into the server and the system which needs to be protected [1, 2, and 3].

The main aspect of the network-based attack detection is to have an intelligent approach that implies that the attack should be identified based on the behaviour or pattern of previous identified attacks. Thus, the intelligence approaches shall be very useful for network attacks identification. The basic steps for identifying the attack based on the network are as follows [1]:

- Identified Data: The data should be identified as it is very significant for the identification of the network attack. These data can be identified by various network command whois, nslookup, ipconfig etc.
- Types of attack: The attack which is identified from the data/dump data is classified into various kind of attack such as R2L, U2R, DoS, Probe etc.
- The various models of the intrusion detection is applied to discover the attack and their variety.
- The attack thus identified shall provide the reporting to the user or the network admin as it shall help them identify the attacker and the amount of data beach.

- At the final stage the auditing of the network is performed to find the novel cause of the attack and the modification in the intelligent model according to the behaviour of the attack.

By using the data mining techniques, the competent IDS can be developed but the deployment of such Intrusion identifying systems may be very critical as various it depends on the various parameters, competency etc [3, 4, 5]. The need of efficient system to identify the intrusion is needed as the normal system induce the large false positive. And the intelligent approach need the huge amount of the data for training the algorithm. For this purpose the NSL-KDD data set is used. Then parameters of the datasets are reduced using the correlation techniques, on the selected parameters the various machine learning algorithms are applied to identify the efficiency of each algorithm; hence using the highest efficient methods the hybrid algorithm shall be developed which shall increase the efficiency of identifying the intrusion into the network.

This paper has been organized as follows: In section 2 the discussion on the various aspects discovered till date shall be discussed. In section 3 the various selection features or parameters form datasets shall be explained like confusion matrix, classification techniques. In section 4 the section of datasets, application of the algorithm their efficiency and features selection shall be explained. In section 5 the hybrid algorithm is proposed. The section 6 shall be followed by conclusion and future work.

2. Related Work

Sekar et al. [7] developed a new NIDS approach based on concise specifications that can classify normal and abnormal sequences of network packet. The researchers performed the research on the known attack vectors. Studies [8] revealed that modern IDS find it difficult to handle high speed network traffic. Researchers [9] have also revealed how attackers can take advantage of this weakness to hide their exploits. They do this by using extraneous information to overload an IDS while they execute an attack. Behaviour based attack detection system were developed using various machine learning algorithms. A clear discussion of the survey for every technique as well as their pros and cons was given in [10, 11]. The survey studied in the paper discovers that the neural network can be considered as the part of machine learning algorithms identifying the network based attacks. It is a group of activities that convert a set of inputs to the expected outputs by using a set of nodes, simple processing units, and connections between them. IDS was developed using multi-layer perceptron based on supervised learning techniques [12] and self-organizing maps based on the unsupervised learning technique [13]. Using a neural network is an efficient approach that can be used to improve IDS performance based on the anomaly detection and misuse detection models [14]. To assess the performance of their developed IDS, several researchers utilized different existing datasets [15]. Saurabh Mukherjee et al. [16] introduced the methods of reducing the parameters and considering only those parameters that can efficiently identify the attacks. The anomalies in the IDS are then detected using the naïve Bayes classifier. A large amount of work is currently being performed in the field of intrusion detection. The research paper also focuses upon the various means to increase the efficiency of network and attack identification with reduced parameters. Chun Guo et al. [17] came up with a hybrid learning method called the distance sum-based SVM (DSSVM) to model an effective IDS. In DSSVM, feature dimensions of the cluster centers in the data set and the sum of the distances based on the correlation between each data sample are obtained. The SVM is then utilized as a classifier. Ravale et al. [2] explained the use of various data mining techniques to have a hybrid algorithm. The number of attributes related to every data point is reduced using the K-means clustering algorithm. Additionally, the support vector machine's (SVM) radial basis function (RBF) kernel is utilized for classification. Gaikward et al. [3] explained the intelligent way to identify the attack based on machine learning method and also find the techniques to reduce the features using the genetic algorithm.

3. Important Parameters Confusion Matrix

The confusion matrix is very important parameter which need to consider, which making the efficient hybrid algorithm.

Table 1. Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	A	b
	Positive	C	d

Where a= correct prediction for negatively occurrence. b= incorrect prediction for positive occurrences. c= incorrect prediction for negative occurrence and d= correct prediction for positive occurrence.

The confusion matrix is the information associated to actual classification and predicted classification formed by the classification system. The Table 1 gives the clear idea about the confusion matrix of how to maintain the integrity of the specifications. From the confusion matrix various parameter which are of utmost importance can be calculated such as accuracy, true positive, false positive, true negative rates, and false negative rates. Let's look into such parameters:

i. Accuracy is defined as the total number of correct predictions which can be calculated as follows:

$$AC = (a+d) / (a+b+c+d)$$

ii. True positive id defined as correctly defined positive cases which can be calculated as follows:

$$TP = d / (c+d)$$

iii. False positive is defined as correctly defined negative cases identified as positive which can be calculated as follows:

$$FP = b / (a+b)$$

iv. True negative rate is defined as negative cases identified as correctly which can be calculated as follows:

$$TN = a / (a+b)$$

v. False negative rate is defined as positive cases identified as incorrectly which can be calculated as follows:

$$FN = c / (c+d)$$

3.1. Techniques

While using the data mining technique in the way of making the system which shall identify the network-based attack the classification of the algorithm plays a very important part. To make prognosis some of the techniques can be applied on the data in numerous ways [6]. Some of the algorithms are nearest neighbor, decision tree, neural networks etc. As these techniques are efficient when they are taken into account once at the time hence the idea of hybridization the algorithm comes into picture.

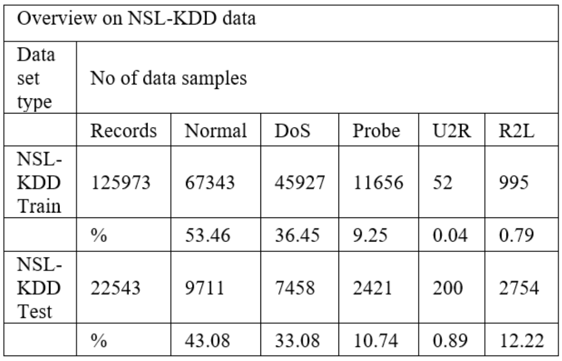
4. Basics for hybrid algorithm

To have the hybrid algorithm which shall identify the network attack efficiently for that purpose the study should be directed on the predefined entities. The following steps were considered in development of the hybrid algorithm which shall efficiently identify the network-based attack.

4.1. Selection of NSL-KDD dataset

NSL-KDD is the enhanced version of KDDcup 99 dataset [6]. This dataset is made up of the large number of data, this data compromise of the NSL-KDD Train, NSL-KDD Test, NSL-KDD 20% Train and many more.

Table 2. NSL-KDD Dataset

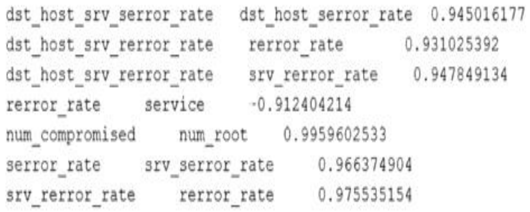


Overview on NSL-KDD data						
Data set type	No of data samples					
	Records	Normal	DoS	Probe	U2R	R2L
NSL-KDD Train	125973	67343	45927	11656	52	995
	%	53.46	36.45	9.25	0.04	0.79
NSL-KDD Test	22543	9711	7458	2421	200	2754
	%	43.08	33.08	10.74	0.89	12.22

This data also consists of large number of records of attack like Normal, Probe, U2R, R2L, DoS apart from the generalized data. Table 2 shows the overview of the mentioned data.

4.2. Selection of features

For experimental purpose only 8 features out of the 41 parameters had been elected which is based on the normalization have IG over .50 were chosen. The one of the important task to remove the unwanted features or parameters this can be done by calculating the variance of the feature.



```
dst_host_srv_serror_rate   dst_host_serror_rate  0.945016177
dst_host_srv_rerror_rate    rerror_rate       0.931025392
dst_host_srv_rerror_rate    srv_rerror_rate    0.947849134
rerror_rate     service     -0.912404214
num_compromised     num_root    0.9959602533
serror_rate     srv_serror_rate     0.966374904
srv_rerror_rate     rerror_rate     0.975535154
```

Figure 2. Variance of parameters

As these chosen parameters will have more efficient attack identification rate, low false positive, low false alarm rate. Table 3 shows the description of selected features.

Table 3. Selected Parameters

Feature	**Description**
5 (src bytes)	Number of data bytes transferred from source to destination in single connection
3 (service)	Destination network service used
6 (dst bytes)	Number of data bytes transferred from destination to source in single connection
4 (flag)	Status of the connection − Normal or Error
30 (diff srv rate)	The% of connections that were to different services, among the connections aggregated in count
29 (same srv rate)	The% of connections that were to the same service, among the connections

	aggregated in count
33 (dst host srv count)	The% of connections that were to the same service, among the connections aggregated in dst host count
34 (dst host same srv rate)	The% of connections that were to different services, among the connections aggregated in dst host count

Table 4. Comparison of values of confusion matrix

Comparison of algorithms				
Algorithms	TP	FP	Correctly classified Instance	Incorrectly Classified Instance
J48	0.997	0.003	99.74	0.26
Random Tree	0.997	0.003	99.747	0.253
Naive Bayes	0.903	0.102	90.2876	9.7124

Finding the accuracy for detection of attack classification and the test accuracy. Table 5 shows the values of the test accuracy and attack detection rates.

Table 5. Test accuracy of attacks

Algorithm	**Class Name**	**Test Accuracy**
J48	Normal	99.8
	DoS	99.1
	Probe	98.9
	U2R	98.7
	R2L	97.9
SVM	Normal	98.8
	DoS	98.7
	Probe	91.4
	U2R	94.6
	R2L	92.5
Naïve Bayes	Normal	74.9
	DoS	75.2
	Probe	74.1
	U2R	72.3
	R2L	70.1

5. Proposed Algorithm

The quasi code for the proposed hybrid algorithm *to identify network attack is as follows:*

- Model ()
- Input Fn = NSL-KDD dataset having 41 features (1 to 42)
- Reduction of to 8 features
- Use IG and variance
- Development of model M
- Provide Fn to high efficient algorithm using NSL-KDD Train+20%
- Calculate A_n (Accuracy of various efficient algorithms)
- H= Ensemble representing efficient algorithm using NSL-KDD Train+20%
- Compare the accuracy of A_n and, E
- Select the best model M=H

Our proposed algorithm shall help to identify the network-based attack through the behavior learning. The pattern of network attack thus identified using this algorithm can be updated in the cloud and the future network attacks having a similar kind of the behavior can be detected. The estimated accuracy to identify the behavior of network attack shall be around 98 percentage. The T`able VI shows the comparison of the selected algorithm with the hybrid algorithm; from this comparison table the conclusion can be drawn that the hybrid algorithm thus developed is more accurate as it yields the high rates for the true positive and correctly classified instances with reference to the low rates for the false positives and incorrectly classified instances. The comparison can be evidently seen from Figure 3.

Table 6. Test accuracy of attacks

Comparison of algorithms				
Algorithms	TP	FP	Correctly classified Instance	Incorrectly Classified Instance
J48	0.997	0.003	99.74	0.26
Random Tree	0.997	0.003	99.747	0.253
Naive Bayes	0.903	0.102	90.2876	9.7124
Hybrid Algorithm	0.999	0.001	99.89	0.3

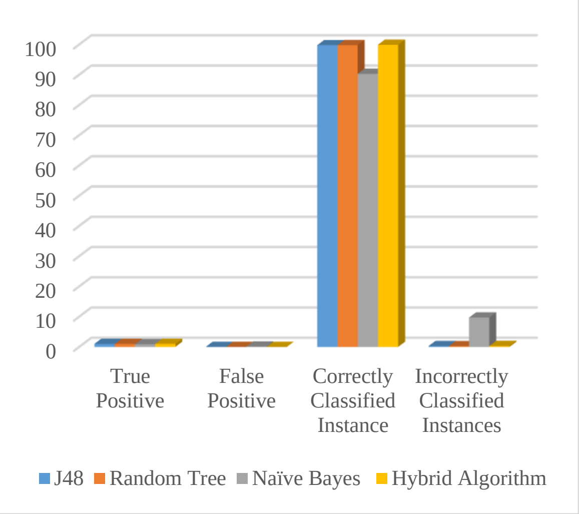


Figure 2. Test Accuracy of Attacks (Comparison)

Table 7. Test accuracy of attacks

Algorithm	**Class Name**	**Test Accuracy**
J48	Normal	99.8
	DoS	99.1
	Probe	98.9
	U2R	98.7
	R2L	97.9
SVM	Normal	98.8

[16] S. Snapp, J. Brentano, G. Dias, et al., DIDS (distributed intrusion detection system)—motivation, architecture, and an early prototype, Proceedings of the14th National Computer Security Conference, October (1991).

[17] Saurabh Mukherjee, Neelam Sharma, Intrusion detection using naive bayes classifier with feature reduction, in: proceedings in 2nd International Conference on Computer, Communication, Control and Information Technology, C3IT-2012, Procedia Technol. 4 (2012) 119–128 (Elsevier).
.