IMPLEMENTATION OF DECISION TREE ALGORITHM AFTER CLUSTERING THROUGH WEKA

Nitin Mohan Sharma¹ Kunwar Pal²

¹(Department of computer science & engineering Lovely Professional University Jalandhar Punjab India
²(Department of computer science & engineering Lovely professional University Jalandhar Punjab India

ABSTRACT

Intrusion detection systems have been used widely in most of the organizations to secure their data from intruders. To provide more security an Intrusion Detection system have tighter rule that means it alerts more time to the system. In this way some time it alerts the alarm unnecessarily. These alarms are false positive. We want to reduce the false alarms of the intrusion detection system. To increase the efficiency and accuracy of the system we have to filter false alarms. To reduce false alarms some machine learning algorithms have been used till date but the efficiency rate is not up to mark. In adaptive false alarms filter a combination of machine learning algorithms is used to increase the classification accuracy of the system. In this system best algorithm is used on the basis of decision value. First we have to do clustering of the data then on the result we have to apply classification algorithm. K means algorithm is used for clustering purpose. While decision tree algorithm used for the classification purpose.

Keywords: Clustering, Classification, False Alarms, Machine Learning Algorithms, Decision Tree.

1. INTRODUCTION

Network Security is the primary concern of any network. As we know this is the age of internet, almost out of 10 person 5 use internet. They share important information, do online shopping, and transaction in the bank, these all type of transaction need the security. Our network is not secure, intruders are there. These intruders want to hack the network and want to steal the information [1]. Some times as in the case of phishing attack an intruder makes a illusion in web address. For example in case of state bank of India correct web
address is statebankofindia.co.in, an intruder may change it as statebankindia.co.in. Sometimes we cannot distinguish between these types of attack. Intruder takes the benefit of this situation. So we need to identify such type of attacks and resolve them. Suppose a customer done his/her transaction with fake web address. This fake web address may promote customer to do the transaction with fake web address. Intruder may know the password of the customer. Once the intruder knows the password of the customer intruder can log in from the customer account. This is only one type of attack exist in this the environment. there may be another type of attack which we want to resolve. A secure network is desirable by every person [2]. As we know there are various type of attack on the internet these are primarily two types viz active and passive. Active attack means to change the content of the message while in case of passive attack we cannot change the content. We can say passive attacks are difficult to detect while active attack easy to detect. The network security has become an urgent problem to be addressed [3]. Many security tools available in the market such as anti viruses, firewall etc. anti viruses and firewall provide the internal security, when we need security from outside the organization. Than we need intrusion detection system. The intrusion detection system is the first line of defense against network security. It collects the information of network flow and warns for possible attacks.

2. EARLY WORK

Machine learning algorithms have been already used to reduce false alarms in intrusion detection system. These work shows that use of machine learning algorithms is good to reduce false alarms. Lee and Staflo explain how to extract features of audits and how to apply machine learning algorithm to process these extracted features [4]. After this work use of machine learning algorithm becomes popular for classification purpose. Pietraszek made a system for adaptive learner for alert classification to reduce false alarms into two categories to help analyst. Law and kwok [7] used k nearest neighbor to classify alarms into two categories. They first analyze normal profile of the system than find out deviation from the normal behavior.

2.1 Approaches Of Various Machine learning Algorithms

Till now many machine learning algorithms are used to classify alarms into two categories. Different algorithms behave differently in different environments. Thus we cannot determine which algorithm is best. A unified platform is required for comparing different algorithms. WEKA [9] is a tool which has capability to compare different algorithms on the same platform. First of all WEKA requires a labeled dataset, this dataset acquired through WIRESHARK [10].

2.2 Selection of features

When WIRESHARK [10] detects sign of attack it through alarms in specified format. How it determines features it is described with the help of following example:-

03-01-13 13:00 response activity invalid URL [classification attempt to unauthorized access to URL] Priority 3197.7.248.153-> 172.16.113.204

This example shows TCP packet alert for invalid URL it belongs to the classification of attempted information leak. Where first part that is 03-01-13 13:00 shows time of attack, type of attack invalid URL its priority is 3 and last two part shows destination and source IP address.
3. **DATASETS CONSTRUCTION AND METHODOLOGY**

This dataset contain different data of different weeks [5]. To find out false alarms we need to resend packets of three weeks through WIRESHARK. All triggered alarms are considered as false alarms. WIRESHARK uses color coding to show sign of attack. Suppose for TCP protocol blue color is defined for normal activity. If few protocols shows color other than blue than it will be considered as a sign of attack.

**Table 1 Number of false alarms**

<table>
<thead>
<tr>
<th>Week days</th>
<th>Week 1</th>
<th>Week 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>2749</td>
<td>1243</td>
</tr>
<tr>
<td>Tuesday</td>
<td>2090</td>
<td>1015</td>
</tr>
<tr>
<td>Wednesday</td>
<td>1018</td>
<td>1178</td>
</tr>
<tr>
<td>Thursday</td>
<td>1218</td>
<td>1311</td>
</tr>
<tr>
<td>Friday</td>
<td>1299</td>
<td>1656</td>
</tr>
<tr>
<td>Total number</td>
<td>8374</td>
<td>6403</td>
</tr>
</tbody>
</table>

On the basis of above table we could see how many false alarms are occurred on the network. We make following two data sets for the analyses of false alarms.

**3.1 Dataset 1**

We select from table 1 8374 false alarms and by selecting 7288 true alarms. Here ratio of true and false alarms is 1:1. This dataset will be used when number of false alarm and true alarms are equal in number.

**3.2 Dataset 2**

As shown in table 1 there are 14777 false alarms and 7288 true alarms. Here ratio of true and false alarms is 2:1. This dataset would be used when numbers of false alarms are double than true alarms [7].

Performance of machine learning algorithm would be evaluated on the basis of following parameters:

(i) Classification accuracy
(ii) Precision of false alarms
(iii) Recall of false alarms

Precision of false alarm = number of false alarms classified as false alarm/ number of alarm classified as false alarm (1)

Recall of false alarm= the number of false alarms classified as false alarm/ number of false alarms (2)

A best algorithm should have precision of 1 to show it does not classify any true alarm as false alarm, also a recall of 1 to show it can classify every false alarm as false alarm [6]. In fact best algorithm would have all the three parameters that classification accuracy should also good and remaining two factors also better.
4. PROPOSED WORK

Under this section we have developed new algorithm. This algorithm is basically a mixture of two algorithms. First one is k means algorithm [7] and second is decision tree algorithm. Our approach is to combine both of them for making a hybrid algorithm.

(Some notation used in the following algorithm
X- it is set of data, Dj- training set, entropy- it is the measure of impurity in a collection of training sets, info gain- it is used to select a particular attribute to be a decision node of tree.
Entropy(S) = -P (positive) log2 P (positive)-P (Negative) log2 P (Negative) )
Step (1) For each process in the test data do.
Step (2) If x has new data call then x is cluster 1
Step (3) Else then for each process Dj in training data do
Step (4) Calculate similarity (x,Dj)
Step (5) If similarity (x,Dj) equals 1 then x is cluster 0 and exit
Step (6) Find k biggest score of similarity (x,Dj)
Step (7) Calculate similarity average for k nearest neighbor
Step (8) If similarity average>threshold then x is cluster 0
Step (9) Else x is cluster 1
Step (10) Create decision tree root node as cluster 1 and add training set S into root node as its subset
Step (11) For root node compute entropy (root node.subset) first
Step (12) If entropy (root node.subset)==0 then root node.subset consists of records all with the same value, return a leaf node with attribute value.
Step (13) if entropy (root node.subset) != 0 then compute info.gain for each attribute left. Find attributes with maximum gain(S,A) create child nodes of this root node and add to root node in the decision tree.
Step (14) for each child of the tree apply recursively until reach node entropy ==0 or reach leaf node.

5. EXPERIMENTAL METHODOLOGY

Under this section we produce architecture of false alarms filter that will use best algorithm on the basis of decision value. First all passed through WEKA [9] which is a platform to compare various machine learning algorithms. In this package we have to add new algorithm name as classification after clustering [8]. The experimental methodology shows classification accuracy better than previous system available in the market. Comparison among inbuilt algorithms and new algorithm are done on the basis of decision value. Decision value can be described with the help of following equation

Decision value =0.5*C.A. + 0.5*PFA (3)

Where C.A is the classification accuracy
P.F.A. is the precision of false alarms.
This methodology described in the following diagram.
6. RESULTS AND DISCUSSIONS

Under this part we run our sets constructed from WIRESHARK through WEKA. Algorithms are chosen than shows the output of the classification accuracy, because our decision value depends on classification accuracy and precision of false alarms. By choosing customized algorithm of classification after clustering algorithm we see classification accuracy is increased. We can see it in the following table:

<table>
<thead>
<tr>
<th>Machine learning algorithm</th>
<th>Classification accuracy</th>
<th>Precision of false alarm</th>
<th>Recall of false alarms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification after clustering</td>
<td>96.75</td>
<td>0.867</td>
<td>0.976</td>
</tr>
</tbody>
</table>

6. CONCLUSION

By using such a methodology we have reduced false alarms. It also reduces burden of security officer. Before reduction of false alarms security officer needs to analyze more alarms, now if security officer suppose to analyze less alarms in comparison to previous work than it definitely saves the time of security officer. IDS have the efficiency to detect malicious activity and report it to the user. When IDS do not capable to identify all the attacks in that case the developer of the IDS make the security rules tighter than previous one. Due to these tighter rules sometimes IDS alarms unnecessarily. Actually these are false positive alarms. So no need to alert the user while such type of signs is found by IDS, in other case if the developer looses the security rule than it may happen some true attacks are not identified by the IDS. Due to the inherent limitations of the IDSs we have to filter the false
alarms from the system. Our aim is to make our system hundred percent secure. So here we want to make a system which can easily identify attack and make the system in such a way that it can identify maximum number of attack. So we want to develop such a system which does not provide hundred percent securities but enhance the security of the network.

REFERENCES

Journal Papers:

Web Resource: