Improved Intrusion Detection System using cascading of C4.5 Decision Tree and Support Vector Machine

Snehal S. Somwanshi1, Daivshala R. Deshmukh2
1Post Graduate Student, 2Assistant Professor, Maharashtra Institute of Technology, Aurangabad, Maharashtra, India.

Abstract—The Computer network is a heart of the communication. lots of things individual can do on the internet such as email, credit card transaction, net banking, online shopping. Due to this increasing dependency of internet we have to significantly improve the security of internet. Hence, there is an instant need to recognize and detect the attacks. Intrusion Detection is defined as a process of diagnosing the attack and the sign of malicious activity in a computer network by evaluating the system constantly. The software that does such task can be defined as Intrusion Detection Systems (IDS). System established with the single algorithms like classification, neural networks, clustering etc. gives good detection rate and less false alarm rate. Current studies show that the flowing of multiple algorithm yields much better performance than the system developed with the single algorithm. Intrusion detection systems that uses single algorithm, the accuracy and detection rate were not up to mark. Increase in the false alarm rate was also come across. Flowing of two algorithm is implemented to resolve this problem. This paper suggests flowing of two algorithms for developing the intrusion detection system. C4.5 decision tree and Support Vector Machine (SVM) are pooled to maximize the accuracy, which is the benefit of C4.5 and decrease the wrong alarm rate which is the benefit of SVM. Outcomes illustrate the upturn in the accuracy and detection rate and less false alarm rate.

Keywords—Intrusion Detection System, Data Mining, Decision Tree, Support Vector Machine, Hybrid Algorithm, kddcup99.

I. INTRODUCTION

Heady et al 1990 defined intrusion as “any set of activities that tries to compromise the reliability, confidentially or accessibility of a resource” [3]. Damaging the information or accessing it as an unauthorized user, or maliciously using the information comes under the category of intrusion. Six types of attacks are defined by Mahoney [3] and they are viruses, worms, server attacks, client attacks, network attacks and root attacks.

It is challenging to identify such kinds of attacks in spite of of strong security strategies, anti-virus software, firewalls or other appliance because every system has some fault and bugs. Thats why the IDS are deliberate and can detect the new attacks IDS observer all traffic flow in a network, and classify the doubtful, malicious activities.

II. RELATED WORKS

An In 1980, the notion of intrusion detection initiated with Anderson’s [5] He offered a intimidation classification model that develops a security monitoring observation system based on detecting irregularities in user behaviour. There are various methods that have been used in the development of Intrusion Detection System which are categorized as single and hybrid techniques.

In this paper he concentrated on a collective model for misuse and anomaly intrusion detection. Collective behaviour rule set are developed by using the clustering analysis algorithm to detect the new unknown attacks and association rule mining algorithm was useful to detect the known attack rapidly.

M. Panda et al (2011) In this paper they have suggested a hybrid intellectual intrusion detection system by merging the two classification algorithm for creating the decision more accurate and rapid. First the classification or clustering was applied in the entire dataset and the output is applied to another classification algorithm. They applied 10-fold cross validation method, and the result realized is in the form of normal or intrusion.

Yogita B .Bhavsar et al (2013) In this paper author suggested Intrusion Detection System using data mining method SVM(Support Vector Machine) and in their suggested system SVM is used for classification and verification concerning the effectiveness of the suggested system is done by conducting some experiments using NSL-KDD Cup’99 dataset which is developed version of standard KDD Cup’99 data set. The SVM is one of the most conspicuous classification algorithms in the data mining area, but its drawback is its large training time. For this drawback they suggested system that supported several studies using NSL-KDD Cup’99 data set. The experimental results show that it can reduce widespread time required to Build SVM model by accomplishing appropriate data set pre-processing. Also when they do appropriate selection of SVM kernel function such as Gaussian Radial Basis Function because of this it create results that shows attack detection rate of SVM is enhanced and False Positive Rate (FPR) is decrease.

Jashan Koshal, et al(2012) This paper author signified two hybrid algorithms for emerging the intrusion detection System.C4.5 decision tree algorithm and Support Vector Machine (SVM) algorithm are cascading to maximize the accuracy and for that purpose grouping of algorithm result shows that growth in the accuracy and detection rate. And then they accomplish that the cascading of algorithm gives improving results than single algorithm in detection of Intrusion Cascading the Algorithm increases the accuracy and consistency. Thus Scientists are focused on cascading methods for developing the Intrusion Detection System.

Following figure shows percentage distribution of several data mining techniques used to detect intrusion.

### III. DATA MINING METHODS

#### A. C4.5 Decision Tree

C4.5 is a typical decision tree algorithm. It is basically an extension of ID3 algorithm. This was developed by Ross Quinlan [15]. C4.5 is also known as statistical or supervised learning classifier. In this algorithm, divide and conquer approach is used to recursively construct a decision tree of database/dataset attributes. Decision tree is built by using two methods i.e top down manner or bottom up manner. A decision tree is constructed with empty root node and then decision tree node corresponding to the algorithm is built [15][14].

To construct the decision tree following steps is followed recursively. Following are the steps [15]:

1. Information gain (IG) is computed for each and every attribute.
2. A tree like structure gets started constructing from its root node by the selection of attributes which have a maximum value of IG.
3. If the attributes are more discrete than its branch with all possible values and if they are continuous than maximum value is selected on the basis of G or cut point.
4. After splitting, consider whether or not these new nodes are leaves (their data belongs to the same type); otherwise, new nodes are the root of the sub-trees.

5. Repeating all the above steps, until all new nodes are leaves.

This partition of the training dataset is performed and a tree like structure is created. Due to selection of highest IG value it is best suited for decision making tree.

Let D is a training set which contains m distinct classes, Ci(i=1,2,...,m) Ci,D set of tuples in class Ci in D,[Ci,D] and |D| represent the number of tuples in D and Ci,D respectively and Pi is the probability that any tuple present in D belongs to class Ci, which is calculated by |Ci,D| / |D|. When an attribute A is selected, it can be used to split D into v partitions {D1,D2,....Dv} where Dj contains tuples in D that have outcome aj of A. There are some method for selecting best or splitting attributes as follows:

\[ \text{Info}(D) = - \sum_{i=1}^{m} P_i \log_2 P_i \]

\[ \text{Gain}(a,T) = \text{Entropy}(a) - \text{Info}(a,T) \]

\[ \text{Gain}(a,T) = \frac{\sum_{i=1}^{v} T_{a,v} \cdot \text{Entropy}(a_i)}{T_a} \]

\[ \text{GainRatio}(a,T) = \frac{\text{SplitInfo}(a,T)}{\text{Info}(a,T)} \]

\[ \text{SplitInfo}(a,T) = - \frac{|T_{a,v}| \log |T_{a,v}|}{|T_a| \log |T_a|} \]

B. SVM (Support vector machine)

SVM are a class of supervised learning algorithms which first introduced by Vapnik. Given a set of labeled training vectors (positive and negative input examples), SVM learn a linear decision boundary to discriminate between the two classes. The outcome is a linear classification rule that can be used to classify new test examples. SVM have demonstrated excellent generalization routine (accuracy on test sets) in practice and have strong theoretical motivation statistical learning theory. The basic idea of SVM can be illustrated in Figure below, two types of two-dimensional linearly separable case, there are many possible linear classifiers can be separated from this set of data, but only a classification of the two intervals the most that figure has.

The so-called optimal separating superficial is the classification of line not only requires appropriate parting of the sample (training error rate is 0), but also the largest class interval, these are two inconsistent goals.

IV. Dataset Description

**KDD Cup 1999 Dataset**

KDD Cup 1999 dataset is the most popular dataset that is used for evaluating the anomaly type intrusion. In 1998 DARPA conducted an evaluation program for intrusion detection in the MIT Lincoln Labs. The main objective of this program was to evaluate the intrusion [7] [8]. In this a standard dataset was provided which simulated a wide variety of intrusions on the perspective of a military network environment. This dataset was prepared by Stotfo et al[8]. DARPA 1998 dataset consist of 4 gigabytes of compressed tcp dump data which was collected over the period of 7 weeks. In this dataset, around 5 million connections and each one connection having 100 bytes were records. KDD Cup 1999 having 4,90,000 single vector connection [9]. This dataset contains 41 features which were labeled as normal or as abnormal[10-12].

In our experiment we used 4178 records. Among that 4178 2089 is used for training and 2089 is used for testing.

<table>
<thead>
<tr>
<th>TOTAL NO OF RECORDS IN ORIGINAL DATASET</th>
<th>TRAINING DATASET</th>
<th>TESTING DATASET</th>
</tr>
</thead>
<tbody>
<tr>
<td>4178</td>
<td>2089</td>
<td>2089</td>
</tr>
</tbody>
</table>

**V. Attack Categorisation**

Attacks are grouped into following four categories:

**Denial of Service Attack (DOS):** It is a malicious attempt in which the attacker makes server or network resource unavailable or too busy to handle requests. Due to the unavailability of the resources, any other kind of request made by legitimate users gets unprocessed.
User to Root Attack (U2R): It is a type of exploit in which attacker has a local normal user account access on the system. But an attacker takes the advantage of present vulnerabilities in the system like sniffing passwords, a dictionary attack or social engineering and gets the super user privilege access.

Remote to Local Attack (R2L): is a type of attack in which an attacker machine is able to send a packet remotely but does not have an account on the victim machine. So by taking advantage of any vulnerabilities on the victim machine like clock, guest, xnsnoop, phf, sendmail, the attacker gets access to the victim machine.

Probing Attack: It is a very basic and initial step of exploiting any system. The attacker scans a machine to find out the weakness or vulnerabilities in the network using saint, portsweep, mscan, nmap etc. To exploit the victim machine.

VI. EXPERIMENTAL RESULTS

The proposed system is developed by using cascading of two algorithm C4.5 decision tree and support vector machine. for that purpose experiments done on standard dataset i.e. kdd cup 99 dataset. these dataset containing 4178 records and among that records 2089 is selected for training and 2089 is selected for testing. for training purpose c4.5 and svm algorithm is applied on dataset to train the dataset about normal and abnormal traffic, after training testing is done on that dataset to check the efficiency of model.

When algorithm is applied on the dataset, dataset gives results in the form of normal and malicious records. these records are obtained from each algorithm such as c4.5, svm and hybrid algorithm which is a combination of c4.5 and svm algorithm. After getting normal and abnormal results of data, from that results confusion matrix is drawn. Confusion matrix has following values

False positive (FP): Or false alarm. Corresponds to the number of detected attacks but it is in fact normal.

False negative (FN): Corresponds to the number of detected normal instances but it is actually attack, in other words these attacks are the target of intrusion detection systems.

True positive (TP): Corresponds to the number of detected attacks and it is in fact attack.

True negative (TN): Corresponds to the number of detected normal instances and it is actually normal.

After getting confusion matrix values performance of proposed model is tested by using some performance evaluation parameter.

We checked performance of model by using following parameter

Accuracy: This refers to the ability of the classifiers to correctly measure the intrusions from the training dataset. This is defined as the ratio of correctly classified data to the total classified data.

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

Results of accuracy of c4.5, svm and hybrid model shows following table.

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>ACCURACY</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5</td>
<td>98.32</td>
</tr>
<tr>
<td>SVM</td>
<td>98.08</td>
</tr>
<tr>
<td>C4.5 + SVM</td>
<td>99.42</td>
</tr>
</tbody>
</table>

From the above figure it is clear that hybrid model gives better result in terms of accuracy than single model.
Detection Ratio: It is defined as the ratio of detecting attacks to total no of attacks. This is the best parameter to measure the performance of the model.

\[
\text{Detection Ratio} = \frac{TP}{TP + FN}
\]

Following table shows results of detection ratio of three algorithm

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>DETECTION RATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5</td>
<td>96.55</td>
</tr>
<tr>
<td>SVM</td>
<td>88.88</td>
</tr>
<tr>
<td>C4.5+SVM</td>
<td>98.89</td>
</tr>
</tbody>
</table>

False Positive Rate: This is one of the main parameters to find out the effectiveness of various models and also the major concern while network setup. A normal data is considered as abnormal or attack type data. It is defined as:

\[
FPR = \frac{FP}{TN + FP}
\]

Following table shows false positive rate of three algorithm

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>FALSE POSITIVE RATE (FPR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5</td>
<td>1.45</td>
</tr>
<tr>
<td>SVM</td>
<td>1.35</td>
</tr>
<tr>
<td>C4.5 + SVM</td>
<td>0.53</td>
</tr>
</tbody>
</table>

Fig: shows graphical representation of results of accuracy
Fig: shows graphical representation of results of detection rate
Fig: shows results of false positive rate of three algorithm
**Precision Ratio:** It is also known as positive predictive value (PPV). It measures the relevant instance that is retrieved after classification. A classifier that has high precision means that classifiers or algorithm returns more relevant results.

\[
PPV = \frac{TP}{TP + FP}
\]

Following table shows results of precision value

<table>
<thead>
<tr>
<th>ALGORITHM</th>
<th>PRECISION</th>
</tr>
</thead>
<tbody>
<tr>
<td>C4.5</td>
<td>0.38</td>
</tr>
<tr>
<td>SVM</td>
<td>0.39</td>
</tr>
<tr>
<td>C4.5 + SVM</td>
<td>0.66</td>
</tr>
</tbody>
</table>

**VII. CONCLUSION**

As overall study shows that hybrid approaches gives better results than single algorithm, therefore researchers are now focusing on two hybrid algorithm for greater results.

In this work Experimental results concludes that the hybrid approach is implemented which is combination of two algorithm namely C4.5 and SVM algorithm.
The results of the proposed approach are compared with the results of single data mining techniques such as C4.5 and SVM, then it shows that the HYBRID approach is effective during detection of attacks, Accuracy and to reduce false positive. The DETECTION RATE of the proposed algorithm is 98.89 which is better than single data mining techniques c4.5 & svm which is having detection rate 98.32,98.08 respectively. The ACCURACY of proposed algorithm is 99.42 which is also higher than c4.5 and svm which is having 98.32, 98.08 respectively. The FALSE POSITIVE RATE of Proposed System is 0.53 lower than c4.5 and svm which is having 1.42,1.35 respectively.

REFERENCES
[1] Intrusion Detection Using Data Mining Techniques, Mohammadreza Ektefa, Department of IS UPM Serdang Malaysia mrektefa@gmail.com., Sara Memar Department of CS UPM Serdang, Malaysia sr.memar@gmail.com
[2] Intrusion Detection System Using Data Mining Technique: Support Vector Machine Yogita B. Bhavsar1, Kalyani C.Waghmare2 Post Graduate Student, Assistant Professor, Pune Institute of Computer Technology, Pune, Maharashtra, India
[3] Cascading of C4.5 Decision Tree and Support Vector Machine for Rule Based Intrusion Detection System Jashan Koshal, Monark Bag Indian Institute of Information Technology Allahabad, Uttar Pradesh211012,Indiajashankoshal@gmail.com,monarkbag@gmail.com
[16] Choosing an Intrusion Detection System that Best Suits your Organization Dennis Mathew GSEC Practical v1.4b Option A

Acknowledgements
I would like to take this opportunity to express my solicitous appreciation and deep regard to my Project Guide Prof. D. R. Deshmukh mam, for her valuable guidance, appreciated feedback and constant encouragement for above work. As well thankfull to Prof.K.V.Bhosale mam, Prof.S.R.Chaudhary mam, Prof. V.Kala mam for their valuable guidance.