An Efficient Feature Reduction Comparison of Machine Learning Algorithms for Intrusion Detection System

Upendra
Assistant Professor, CSE Department, NIT Raipur, C.G., India

Abstract: Organisation has come to recognize that applied science in network security has become very important in protecting its information. Intrusion detection present an important line of defense against all variety of attacks that can compromise the security and proper functioning of information system initiative. In this paper we compared the performance of intrusion detection. The evaluation of the Intrusion Detection System (IDS) execution analysis for any given security system configuration improvement is necessary to achieve real time capability. We analyse two learning algorithms (NB and C4.5) for the task of detecting intrusions and compare their relative performances.

Keywords: intrusion detection, machine learning, C4.5, NB, KDD 99

1. INTRODUCTION

Intrusion detection algorithm should consider the composite properties of attack behaviors to improve the detection speed and detection accuracy. Analyze the large volume of network dataset and the better performances of detection accuracy, intrusion detection become an important research field for machine learning. In this work we have presented C4.5 decision tree algorithm for intrusion detection based on machine learning. The Intrusion Detection System (IDS) is Process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents. IDS was first introduced in 1980 by James. P. Anderson [1] and then improved by D. Denning [2] in 1987. They are two basic approaches for Intrusion Detection techniques, i.e. Anomaly Detection and Misuse Detection (signature-based ID) [3]. Anomaly Detection is basically based on assumption that attacker behavior is different from normal user's behavior [4]. In this paper, we present the application of machine learning to intrusion detection. We analyse two learning algorithms (C4.5 and NB) for the task of detecting intrusions and compare their relative performances. There is only available data set is KDD data set for the purpose of experiment for intrusion detection. KDD data set [5] contain 42 attributes. The classes in KDD99 [6] dataset can be categorized into five main classes such as one normal class and four main intrusion classes. Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases [18]. The field of intrusion detection has received increasing attention in recent years.

2. RELATED WORK

In 2007, Panda and Patra [7] determined a method using naive Bayes to detect signatures of specific attacks. They used KDD99 dataset for experiment, in the early 1980’s, Stanford Research Institute (SRI) developed an Intrusion Detection Expert System (IDES) that monitors user behavior and detects suspicious events. Meng Jianliang [8] used the K Mean algorithm to cluster and analyze the data. He used the unsupervised learning technique for the intrusion detection. Mohammadreza Ektefa et al., [8] in 2010, compared C4.5 with SVM and the results revealed that C4.5 algorithms better than SVM in detecting network intrusions and false alarm rate. Zubair A.Baig et al. (2011) proposed An AODE-based Intrusion Detection System for Computer Networks. They suggested that the Naive Bayes (NB) does not accurately detect network intrusions [9]. In 2010, Hai Nguyen et al. [10] applied C4.5 and BayesNet for intrusion detection on KDD CUP’99 Dataset. Jiong Zhang and Mohammad Zulkernine [11] done the intrusion detection using the random forest algorithms in anomaly based NIDS. Cuixio Zhang, Guobing Zhang, Shanshan Sun [12] used the missed approach for the intrusion detection. Various paradigms namely Support Vector Machine [13], Neural Networks[14], K-means based clustering[15] have been applied to intrusion detection because it has the advantage of discovering useful knowledge that describes a user’s or program’s behavior. They are two basic approaches for Intrusion Detection techniques, i.e. Anomaly Detection and Misuse Detection. (signature-based ID) Anomaly Detection is basically based on assumption that attacker behavior is different from normal user's behavior [16]. Shadmehr et al [17] showed that the performance of Bayes algorithm is better. Recently research on machine learning for intrusion detection has standard much attention in the computational intelligence community. In intrusion detection algorithm, immense strengths of audit data must be analyzed in order to conception new detection rules for increasing number of novel attacks in high speed network. Intrusion detection algorithm should consider the composite properties of attack behaviors to improve the detection speed and detection accuracy.
3. METHODOLOGY

3.1 Naïve Bayes (NB)

A Naive Bayes classifier [19],[20] is a simple probabilistic classifier based on applying Bayes’ theorem (from Bayesian statistics) with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be “independent feature model”.

In 2004, analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of naive Bayes classifiers. Still, a comprehensive comparison with other classification methods in 2006 showed that Bayes classification is outperformed by more current approaches, such as boosted trees or random forests[25].

3.2 C4.5

Decision tree technology is a common, intuitionist and fast classification method [22]. Decision tree J48 developed by Johan Ross Quinlan [23]. C4.5 is an extension of Quinlan’s earlier the Interactive Dichotomizer3 (ID3) Algorithm. J48 builds decision trees from a set of labelled training data using the concept of information entropy. The Decision tree is a classifier expressed as a recursive partition of the instance space, consists of nodes that form a rooted tree, meaning it is a directed tree with a node called a root that has no incoming edges referred to as an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). Decision trees [24] are one of the most commonly classification methods used in supervised learning approaches.

<table>
<thead>
<tr>
<th>S No.</th>
<th>Feature name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>service</td>
</tr>
<tr>
<td>2</td>
<td>src_bytes</td>
</tr>
<tr>
<td>3</td>
<td>dst_bytes</td>
</tr>
<tr>
<td>4</td>
<td>logged_in</td>
</tr>
<tr>
<td>5</td>
<td>count</td>
</tr>
<tr>
<td>6</td>
<td>dst_host_diff_srv_rate</td>
</tr>
<tr>
<td>7</td>
<td>dst_host_srv_diff_host_rate</td>
</tr>
<tr>
<td>8</td>
<td>Class</td>
</tr>
</tbody>
</table>

3.3 Evaluation

We constructed a confusion matrix (contingency table) to evaluate the classifier’s performance.

TABLE II. A SAMPLE CONFUSION MATRIX

<table>
<thead>
<tr>
<th>Actual Class Positive</th>
<th>Predicted Class Positive</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>A + D</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>A + B</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>A + C</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>A + D</td>
</tr>
</tbody>
</table>

In this confusion matrix, the value A is called a true positive and the value D is called a true negative. The value B is referred to as a false negative and C is known as false positive.

3.3 Accuracy

This is the most basic measure of the performance of a learning method. This measure determines the percentage of correctly classified instances. From the confusion matrix, we can say that:

\[
\text{Accuracy} = \frac{A + D}{A + B + C + D}
\]

\[
\text{Precision} = \frac{A}{A + C}
\]

\[
\text{Recall} = \frac{A}{A + B}
\]

The Tables III and IV Shows the performance of C4.5 and NB classification methods based on accuracy, Learning time, Error rate, Average true positive rate, Average False positive rate, Average precision and Average F-Measure respectively. The comparison is performed for 41, 11 and 7 attributes. The C4.5 and NB classifier models on the dataset are built and tested by means of 10-fold cross-validation. The Java Heap size was set to 1024 MB for WEKA 3.6.2, the simulation platform is an Intel™ Core i3-2100 processor system with 3 GB RAM under Microsoft Windows XP™ Service Pack-2 operating system, 3.10 GHz with 500 GB memory. C4.5 was evaluated on the dataset by taking into account 7 feature reductions using the Information Gain measure. The results of this test are summarized in the following table.

TABLE III. RESULT OF C4.5 WITH SELECTED 7 ATTRIBUTES

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>99.8901 %</td>
</tr>
<tr>
<td>Learning Time</td>
<td>12.67 sec</td>
</tr>
</tbody>
</table>
Similar to the C4.5 tests, NB was also evaluated by taking into account 7 features of the dataset. The results of this evaluation are summarized in the table below.

**TABLE IV. RESULT OF NB WITH SELECTED 7 ATTRIBUTES**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>93.5698 %</td>
</tr>
<tr>
<td>Learning Time</td>
<td>1.3 sec</td>
</tr>
<tr>
<td>Error Rate</td>
<td>6.4302 %</td>
</tr>
<tr>
<td>Average True Positive Rate</td>
<td>0.936</td>
</tr>
<tr>
<td>Average False Positive Rate</td>
<td>0.054</td>
</tr>
<tr>
<td>Average Precision</td>
<td>0.949</td>
</tr>
<tr>
<td>Average Recall</td>
<td>0.936</td>
</tr>
<tr>
<td>Average F-Measure</td>
<td>0.942</td>
</tr>
</tbody>
</table>

In this paper, we have done the feature reduction to 7 attribute and gave the result. Now we compare the result of the C4.5 and NB algorithms with reduced 7 attribute than only we conclude that which one algorithm is good best for the intrusion detection.

Now the figure 1. given below show the comparison of the accuracy of C4.5 and NB.

**Figure 1.** Comparison of accuracy for C4.5 and NB using all attribute, selected 11 attribute and selected 7 attribute.

From above figure 1. It is clear that information gain feature reduction method gives the better accuracy which is desirable for good Intrusion Detection System. Especially in the case of C4.5 accuracy is 99 % for all attribute, 11 attribute and 7 attribute.

Now we compare the True Positive Rate of the C4.5, and NB algorithm with selected 7 attributes.

**Figure 2.** Comparison of TPR for C4.5 and NB with selected 7 attribute.

For good IDS True Positive Rate should be high. Above figure 2. Shows that True Positive Rate of the C4.5 algorithm is higher when we reduce the feature of the data set using information gain. Especially in the case of C4.5 True Positive Rate is 1 and Figure 2. also shows that TPR of the C4.5 is higher than the NB algorithm with selected 7 attribute.

Now we compare the False Positive Rate of the C4.5 and NB algorithm with selected 7 attributes. Types of attack group such as four main categories: (Probing, DoS: denial-of-service, R2L: remote to local, and U2R: user to root) and one normal class.

**Figure 3.** Comparison of FPR for C4.5 and NB using selected 7 attribute.
For a better Intrusion Detection System, false positive rate should be low. Above figure 3. Shows that FPR of the C4.5 algorithm is lower when we reduce the 7 feature of the data set using information gain. Especially in the case of C4.5 FPR is 0. In the case of NB algorithm False Positive Rate of the greater than 0 with selected 7 attribute.

![Figure 4](image)

**Figure 4.** Comparison of time taken to build model for C4.5 and NB using all attribute, selected 11 and 7 attribute.

From above figures 1, 2, 3 and 4 it is clear that C4.5 algorithm Accuracy, TPR and FPR is better than NB algorithms. So we can say that reduction of the feature using information gain is better technique.

4. **CONCLUSIONS**

In this paper we have showed C4.5 selected 7 attribute technique for intrusion detection and performed feature set reduction and evaluated their performance. From the result, it is observed that after applying the feature selection from 41 attributes to 11 and 7 attributes. The overall performance of C4.5 has increased their performance than NB. The comparison results show that, in general, the C4.5 has the highest classification accuracy performance with the lowest error rate. C4.5 achieves better detection rates than NB and increased true positive rate. On the other hand, we also found that drastically decreased in learning time of the algorithm. We evaluated two machine learning algorithms C4.5 and NB on the dataset built in this exercise. Based on the accuracy, true positive rate, false positive rate, error rate and learning time C4.5 performed best classifier.

**REFERENCES**


