Abnormal Network Traffic Detection Based on Semi-Supervised Machine Learning

Min-jie ZHU\textsuperscript{1,}\textsuperscript{*} and Nai-wang GUO\textsuperscript{2,}\textsuperscript{*}

\textsuperscript{1}Department of Science, Information and Communication, SGSMEPC, Shanghai, China
\textsuperscript{2}Center of Grid Technology, SGSEPRI, Shanghai, China
\textsuperscript{*}Corresponding author

Keywords: Intrusion detection, Machine learning, Classification, Decision tree.

Abstract. Network intrusions will cause abnormal network traffic flow. The abnormal network traffic detection can be used to identify the network intrusions. The traditional intrusion detection system is based on pattern recognition which only can be used for well-known network attack behavior. Machine learning can be used to abstract the characters of a class of objects. In this paper we use machine learning classifiers to distinguish abnormal network traffic from the normal traffic background. The experiments show that this is efficient to detect new intrusion. In order to increase the accuracy for new attack detection, our scheme select the with high confidence samples from testing set to expand the training set which is a semi supervised strategy.

Introduction

Network attacks are illegal access on computer network that exploit the vulnerabilities in target systems. Most network attacks are completed remotely which will bring extra network traffic. These network attacks are called intrusion. Intrusion detection system (IDS) is used to detect such kind of network attack. Traditional IDS is based on pattern recognition technique which is applicable on known attack pattern only. However, more and more new attack technique is emerging in recent years. A new attack pattern cannot be predicted previously before it appears. New attack is called zero-day attack which means it is new for network defense system. The traditional IDS is also invalid for zero-day attack. In this paper we focus on new technique to detect unknown zero-day attack based on machine learning. The main concept is network attack will bring abnormal network traffic even it utilizes new vulnerability. The abnormal traffic can be detected by its special characters in a time window.

The main issue is how to characterize the network traffic and how to define the character of abnormal traffic. It is known that the behavior and procedure of attackers are similar even they use new vulnerability. Machine learning can be used to extract unknown knowledge from mass of data. If we use labeled attack traffic data to train a machine learning model, it can be used to detect a new attack with similar character. Specifically, we use classification model to distinguish abnormal traffic from normal traffic. It is suitable to protect the high-value network such as energy network.

Classification of network traffic is a fundamental problem in network security and management. The rapid growth of the Internet not only gives rise to new and different kinds of applications, but also leads to dynamic changes in network traffic over time. Accurate traffic classification has been already studied in a number of network security problems such as, intrusion detection, detection and prevention of various network attacks (like Denial of Service attacks), anomaly detection, lawful inspection, cybercrime forensic analyses, etc. Traffic classification also plays an important role in network management, for example in traffic prediction and Quality of Service (QoS) provisioning.
Related Work

Abnormal Traffic Detection

Traffic classification has been studied for several years. Traffic classification were mainly used to improve service quality in the first years. T. Karagiannis [1] studied how to evaluate the behavior of computer based on transport layer data. J. Erman studied several algorithms to cluster network traffic such as K-Means, DBSCAN and AutoClass. The result showed that DBSCAN performs best in this scenario. H. Kim [3] studied several classification algorithms to classify network traffic. The result showed that Support Vector Machine (SVM) performs best in this scenario. L. Grimaudo [4] studied semi-supervised machine learning method to classify network traffic. Semi-supervised machine learning can be used to improve the accuracy for new application protocol. T. Bakhshi and B. Ghita studied to detect abnormal network traffic with application information. Alina Vladuț [6] proposed a framework to classify network traffic both for supervised machine learning and unsupervised machine learning. A. L. Buczak [7] gave a survey for abnormal network traffic detection.

Training Dataset

We refer the KDD CUP 99 [8] dataset as our benchmark dataset. T. Mahbod [9] abstracted the dataset with more detailed information which is named as NSL-KDD. We use this modified dataset in the following experiment.

In the following work, the authors also used this dataset. W. Lee [10] compared several classification algorithms including random forests, SVM, J48, CARI and Naive Bayes. They showed that random forests can achieve the accuracy rate more than 96%. Later, Revathi [11] applied SOM neural network on both KDD CUP 99 and NSL-KDD datasets. R. Calix [12] tested SVM algorithm on NSL-KDD dataset. The accuracy rate is lower than 80%. M. Panda [13] compared several two-category classification. The ensemble classifier based on random forests can reach the accuracy rate as 99%.

Our Contribution

We verify the feasibility of using machine learning and statistical features to detect abnormal traffic through experimental methods. Through experiments, we found that the combination of multiple classifiers can significantly improve the accuracy of classification. The ensemble classifier can achieve satisfied accuracy based on weak classifiers.

The algorithms we use are decision tree and KNN algorithm. The training process of decision tree algorithm consumes less computing resources. The training process of KNN algorithm is to add new samples to the existing sample set. Therefore, these two algorithms are efficient algorithms to save computing resources, but the ensemble classifier can achieve good accuracy.

By simulating the network attack environment, we collect new network attack traffic. We use the ensemble classifier trained on the NSL-KDD dataset on this new traffic. By append the new trading data with high confidence to the training set and then retain the classifier to detect new attack adaptively, we get semi-supervised classifier. The experiment shows that this semi-supervised classifier can improve the accuracy to detect zero-day Attack.

System Model

We first introduce the training dataset and data preprocessing procedure.

Dataset and Data Preprocessing

There are many application scenarios in the field of network data mining such as QoS management, situation analysis. In the field of abnormal traffic detection or intrusion detection we studied, related studies have been carried out since the end of last century. In 1998 and 1999, DARPA obtains a batch of tagged data sets by attacking and auditing in a simulated military network. These data were used not only as standard data in the 1999 KDD competition, but also as a standard for evaluating
algorithms in subsequent studies. We selected the KDD data set as the test data in the experiment, and used the characteristics defined by it as the basis for our classification.

1) KDD Feature description

In order to investigate and evaluate intrusion detection system, Lincoln laboratory in MIT established 1998 DARPA intrusion detection and evaluation project.

This project provides data to be used as a benchmark. These data contain a wide range of attack types, which are generated in a network by simulating various attacks. The 1999 KDD intrusion detection competition uses a derived version of the DARPA1998 dataset.

From the original network traffic to the connection record definition defined by KDD, the processing process is as follows:

- Traffic is processed into connection records, where basic features, such as protocol type, connection duration, and so on, can be obtained.
- Another type of connection feature is obtained by checking the payload contents of datagrams, which are called content features.
- The last feature is derived from multiple join records. This is because a lot of attacks involve more than one connection, and most of these connections are continuous. There are two statistical methods, one is based on the two second time window, and the other is based on the previous one hundred connections.

It is important to note that because of the widespread use of various encryption methods in the IP load, the intrusion detection based on context detection is now not considered an effective method. So these features from payload contents have not been extracted in our experiments.

These features were defined by W. Lee[^9].

2) Data Classification

There are two options on the classification of the attack sample:

- Two classifications, all the data are classified into normal or abnormal.
- Multi classification, all the data are classified into four types. The four types are Probing attack, Denial of Service (DoS) attack, User-to-Root (U2R) attack and Remote-to-Login (R2L) attack.

The two classification and multiple classifications described in the following experimental procedures refer to these two schemes.

3) The Characters of KDD and NSL-KDD Dataset

The KDD dataset is very large, and there are nearly five million records in the original dataset. Using all of the original data for training becomes a way to get stuck in a situation where the computing resources are not abundant. Thus, the usual approach is to sample a small fraction from the random set as a training set or a test set. The usual practice is to select 10% randomly, that is the training is carried out on a sample set of nearly five hundred thousand records. The problem with this is that classifiers are not able to make full use of dataset. There is a risk that the classifier that is not selected due to the relevant sample fails to obtain the rules represented by this part of the sample.

On the other hand, the KDD data set is very unevenly distributed on the label. And there are a lot of duplicate redundant records in the samples of some tags. In NSL-KDD[^8], the authors showed that there are more than 90% redundant data in the original attack data and there are more than 15% redundant data in the normal data. In order to make classifier training more, they eliminate the redundant data from the dataset. Then the samples are randomly extracted according to the inverse probability of distribution. That is NSL-KDD dataset. We use NSL-KDD dataset as our training data.

4) Data Preprocessing

We chose sklearn[^1] based numpy as the algorithm library for machine learning. Because of the most basic data structure of ndarry, numpy requires that its elements must be homogeneous. On the other hand, there are some columns of data in the NSL-KDD dataset is discrete string type, we used the method of coding each column to convert them into integer types

5) Features Extraction of Simulation Attack Traffic

We will test the generalization ability of classifiers trained based on NSL-KDD dataset. In order to prove that the machine learning classification algorithm can detect zero-day attack, we generate simulating attack traffic with the tools provided by Kali Linux on a simulated network. The tools include ident-user-enum, massscan, metasploit, nbtscan-unixwiz, nmap and zenmap. We capture the traffic on the simulated network. We use kdd99_feature_extractor\(^1\) to extract the features of attack traffic.

Next, we will introduce our classification model.

**Classification Model**

First, we validate the feasibility of the decision tree and the KNN algorithm by cross validation on the NSL-KDD dataset. After obtaining the satisfied results, we verified that the noise in the sample did not appear overfitting during the training process by investigating the dataset. We construct an ensemble voting framework based on multiple classifiers. Then, the performance of the ensemble classifier is verified by the samples obtained from the simulated attack environment.

1) Preliminary experiments and results

In order to verify the effectiveness of machine learning method, we use cross validation to test the decision tree and KNN algorithm on NSL-KDD data set where we set the parameter k=10. Further, the experimental data are randomly divided into 33% test sets and 67% training sets, and the confusion matrix of the prediction results is calculated as in Table 1 and Table 2.

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>Normal</th>
<th>probe</th>
<th>dos</th>
<th>r2l</th>
<th>u2r</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>22153</td>
<td>16</td>
<td>5</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>probe</td>
<td>16</td>
<td>3863</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dos</td>
<td>2</td>
<td>3</td>
<td>15105</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>r2l</td>
<td>22</td>
<td>0</td>
<td>0</td>
<td>332</td>
<td>2</td>
</tr>
<tr>
<td>u2r</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>KNN</th>
<th>Normal</th>
<th>probe</th>
<th>dos</th>
<th>r2l</th>
<th>u2r</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>22127</td>
<td>30</td>
<td>31</td>
<td>15</td>
<td>3</td>
</tr>
<tr>
<td>probe</td>
<td>45</td>
<td>3739</td>
<td>96</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dos</td>
<td>14</td>
<td>57</td>
<td>15043</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>r2l</td>
<td>17</td>
<td>0</td>
<td>1</td>
<td>338</td>
<td>0</td>
</tr>
<tr>
<td>u2r</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 2. Confusion matrix of two classification based on decision tree and KNN algorithm.

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>Normal</th>
<th>attack</th>
<th>KNN</th>
<th>normal</th>
<th>attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>22147</td>
<td>59</td>
<td>normal</td>
<td>22126</td>
<td>80</td>
</tr>
<tr>
<td>attack</td>
<td>43</td>
<td>19323</td>
<td>attack</td>
<td>84</td>
<td>19282</td>
</tr>
</tbody>
</table>

In the preliminary experiment, we found that the performance of the depth learning algorithm is not as good as that of the simple learning algorithm. Combined with the previous results of other researchers, we only used decision tree and KNN for our machine learning algorithm in the latter experiment. The accuracy rate is in Table 3.

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>KNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two Classification</td>
<td>99.78%</td>
</tr>
<tr>
<td>Multi Classification</td>
<td>99.78%</td>
</tr>
</tbody>
</table>

Table 3. Cross validation results of different classification algorithms on NSL-KDD.

From the confusion matrix, we can see that the decision tree classifier performs well both on the two indexes of accuracy and recall rate. In order to verify that the classifier does not appear
overfitting, we first test the trained decision tree on the training set and the test set respectively. The accuracy rate is similar, so it can be considered that there is no obvious overfitting.

We examine the redundancy of the NSL-KDD dataset. If the data in the training set and the test set appear similar, it will cause the false positive rate of the accuracy. First, hash values are calculated for each sample in the dataset and stored in a table. Then select the random number of subscript, the number of the original set 10% of the test set, traverse the hash table for comparison, if found the same hash value, the original data is further compared to each component. If exactly the same, the output finds duplicate data.

The result of the examination is to find 6 pairs of duplicate data. For there are more than 12000 data in the NSL-KDD dataset, the duplicate data is negligible. We get the conclusion that decision tree algorithm is an effective classification method in the field of abnormal traffic detection.

2) Ensemble Classification

Previous experiments have obtained relatively successful results on known datasets. We will carry a further test on the new dataset, which is to detect the unknown abnormal network traffic. This paper uses the current new attack tools to simulate attack behavior in the test environment, which is a new attack traffic relative to the NSL-KDD dataset.

There are two kinds of errors in machine learning: bias and variance. Different classifiers have their own unique tendency in these two kinds of errors, so there are different performances on the same training set. Taking decision tree algorithm as an example, if the convergence condition is too harsh to define, it is easy to cause the phenomenon of overfitting, the performance is too large variance. The decision tree model takes the rules from the noise in the training set into the process of prediction, and the decision tree classifier is likely to make wrong classification.

In addition, because the type of attack is changing rapidly, we hope that the classification model can be constantly updated, so as to continue to constantly change the type of attack to better detection. We take the idea of semi-supervised learning on this issue. In the data classification process, if found the classifier outputs classification results of a dataset with high confidence, we can think that the judgment is accurate, then put the dataset to the training set. If the new dataset contains relatively new attack traffic, then we can achieve the goal of updating the classifier. If multiple classifiers work at the same time when the new dataset is predicted, it is possible to obtain the data with high confidence level described above.

Based on the above discussion, we propose the scheme of ensemble classification. By training multiple classifiers individually, and then considering their prediction results on unknown data, the prediction error is reduced, and the confidence is increased. The purpose of updating training set can be achieved.

3) Semi-Supervised Strategy

Our original training set is based on NSL-KDD. We train an ensemble classifier with decision tree and KNN algorithms. In order to increase the accuracy, we apply semi-supervised strategy on it. We collect a testing data set in the experiments. We use attack tools in Kali Linux to simulate attack behaviors in a testing network. The network traffic is collected to build a testing data set without label. We use the ensemble classifier to detect abnormal traffic. If a traffic data is voted as an attack by more than 2/3 independent classifiers in the ensemble classifier, we take it as an abnormal traffic with high confidence. All of the data with high confidence is merged into the training data set to refine the classifiers. The experiment shows that this semi-supervised strategy will increase the accuracy.

4) Schemes

First, in order to obtain multi classifiers, two algorithms of decision tree and KNN are selected, each classifier has three instances. Then, 30% of the NSL-KDD dataset is randomly sampled as training sets for each classifier. At the same time, in order to test the performance of two classification and multi classification, we trained two classifiers and multi classifiers. What we specify is that if there are 1/3, most classifiers assume abnormal traffic, then attack behavior is judged.
Experimental Results and Discussion

1) Preliminary Experimental Results
The results of preliminary experiments on NSL-KDD datasets show that the decision tree and KNN algorithm are feasible in the detection of abnormal traffic.
For the existing traffic, the accuracy is very close to 1. This is because the decision tree algorithm can produce nonlinear rules, so its assumptions are more flexible, more able to describe the law of data sets. Furthermore, because the NSL-KDD data set itself comes from DARPA1998 data, its attack and normal traffic are generated at the same time period, and there is a certain similarity in itself, so it is easier to produce higher scores. In fact, in [8], the author also mentioned that one of the goals of proposing NSL-KDD data sets is to make the scores of different classifiers widen.
The experiment result can be found in Figure 1.

2) Experimental Results for Ensemble Classification
The classification results of ensemble classifiers are analyzed, and it is easy to find the following points:
● For some attacks, such as ident-user-enum, Metasploit, nbtscan and so on, the decision tree algorithm has a very high accuracy. This is because the attack patterns of these tools are close to those of the traditional attack, so it is easier to determine the abnormal traffic by the rules extracted by the decision tree algorithm from the training set.
● In some attacks, such as the zenmap model of intense, the KNN algorithm which is not outstanding in most cases shows much better accuracy than the decision tree algorithm.
● Observing the accuracy of an ensemble classifier and the performance of a single classifier can find that the accuracy of integration is very close to the accuracy of the best performing members. The above two points verify the rationality of the ensemble classifier.
For voting rules, we choose 1/3 as threshold. If all members have more than 1/3 to represent the data to be measured as attack data, they are abnormal data. The reason to choose a threshold lower than 1/2 is that the data set is not balance. Attack traffic is much less than normal traffic in the network.

3) Experimental Results for Semi-Supervised ML
As in Figure 2, if we append new simulation data with high confidence to the training data set to get a semi-supervised classifier, the accuracy is improved. Especially for new attack technique, such as Masscan. This shows that semi-supervised ML will help the classifier to detect new attack traffic efficiently.
Conclusion

By using the basic and statistical attributes of network traffic, we explore a scheme to detect abnormal traffic using machine learning. In particular, we investigate the advantages of ensemble classifiers for improving classifier performance on the whole. In the early stage of the experiment, we validate the NSL-KDD data set as an anomaly detection field of standard data sets, and in this data set to verify the performance of decision tree algorithm and KNN algorithm, which verifies the feasibility of the abnormal flow of machine learning. Furthermore, we use the ensemble classifier trained on the NSL-KDD dataset to validate the extracted traffic in the simulated attack environment. It is proved that the machine learning method does have the ability to resist zero-day attack. The ensemble method can classify several weak classifier sets into a strong classifier.

References


