A new Approach for Intrusion Detection in Computer Networks Using Data Mining Technique

Aida Parbaleh¹, Dr. Heirsh Soltanpanah ²*

¹Department of Computer Engineering, Islamic Azad University, Sanandaj Branch, Sanandaj, Iran.
²Department of Industrial Engineering, Islamic Azad University, Sanandaj Branch, Sanandaj, Iran.
* Corresponding author: Dr. Heirsh Soltanpanah

ABSTRACT

"Intrusion" signifies certain acts that are meant to hazard the integration, privacy and availability of a source. An intrusion detection system (IDS) can limit the access of user to the computer by implementation of some specific rules. These rules are based on the knowledge of the expert. In this research a model is suggested that is based on data analysis with the aim of increasing accuracy. This model consists of two layers. At the first one, the clustering algorithm is used with the aim of simplifying input space and decreasing calculations. At the second layer, classification algorithm is used to classify available samples in the cluster. The results show that this method has a better performance on “KDD CUP” data set in comparison with previous methods.

Key words: Intrusion detection, intrusion detection system, classification, clustering, training data set, test data set

Introduction

"Intrusion" is a set of operations that try to hazard the integration, privacy and availability of a source. An “intrusion detection system” can limit the access of user to the computer by the implementation of some specific rules. These rules are based on the knowledge of the expert. These experts are people who make attack scenarios. The system identifies all violations and accomplishes necessary steps for stopping attack on the database. In this regard, the problem of intrusion detection in computer security is widely investigated. Intrusion detection systems are divided into two categories based on the host and network. A system based on network can verify network traffic for specific parts of the network and detect any suspicious activity with analyzing them. The most significant variable problems concerned with IDS are [1]:

Wideness and variability of attacks: Computer systems are faced with a wide range of attacks. Variability of attacks causes long time detection. On the other hand, attacks do not have a specified pattern and everyone has a different behavior. So some attacks with unknown pattern are being expected.

Likeness of behaviors of some attacks to regular activities behavior in the network:

Some of attacks have a pattern similar to the typical pattern of activities in the network. The detection of these attacks is more difficult than others.

Methodology of intrusion detection is divided into 3 categories [1], [3] and [4]:

1.1. signature-based detection
A signature is a pattern that corresponds to a known attack. In this method, they obtain the patterns of occurred processes and compare each with the existing attack pattern. If it matches, it means that intrusion has occurred, if not the system is in a safe mode. The main advantages of this method are their simplicity and efficiency in detecting known attacks and the main disadvantage is that it cannot detect new attacks.

1.2. anomaly-based detections
An unconventional behavior is a behavior that has some deviations from a conventional behavior. Conventional behaviors (normal behaviors) form the profiles. Patterns of these behaviors can be obtained from observing legal activities in a computer system. The profiles can be “static” or “dynamic”. Anomaly-based detection systems compare ongoing processes behaviors with profiles and will announce intrusion if they observe deviation. Some sources call these methods “behavior-based detection.” The main advantage of these methods is their ability to detect...
a new attack. The other advantage is their independence of the operating system. Their major shortcoming is their production of false alarms and their low performance in case where there is no good example.

1.3. Statement of protocol analysis
These methods are based on knowing and tracing the state of protocols. They act like anomaly-based detection methods with one difference and that is related to the profiles. In this method, the profiles are based on standard rules that legislate in the network. In this paper there is a new method which is based on anomaly detection for intrusion detection in system. This model has two layers. In the first layer, the clustering algorithm is used with purpose of simplification of input space and decreasing calculations. In the second one, classification algorithm is used to classify available samples in the cluster.

2. Research Background
“Li” uses a model of “signaling game” for intrusion detection in wireless networks. In this method, it is assumed that there is a central distributor network which any node of that has IDS but just IDS embedded in central cluster has the task of detection. The above method uses statistical method’s techniques for this operation. Tian et al. utilized genetic algorithm [4]. First of all, labeled data set (normal or abnormal samples of network traffic data that can be collected by surveillance tools) is being analyzed and then some primary rules are being legislated. Then these primary rules can be used for the initial population. In the following, it is trying to extract some appropriate legislations for intrusion detection by use of genetic operations. Gosh et al. used forward multilayered perception network and also Elman returning network with the aim of classification of system calls[4]. They reported their results of the DARPA 1999 data set. Their results showed that “Elman network” has a better performance in comparison with multilayered perception network. This method is for intrusion detection and normal behavior modeling. With respect to the above case, Ghosh et al. used a clustering method based on grid for modeling the normal behaviors [4]. In this method, normal behaviors of users can be observed online. Afterward, based on them, normal patterns are being extracted and they divide into several clusters. A sample will be known as an “abnormal “ one if it doesn’t match with any cluster. A recent approach has been illustrated for intrusion detection in systems that use clustering and classifying models together. As an example, Park et al. had used a combination of clustering and classifying models for intrusion detection [5]. At first, with the usage of Fary clustering, divides data samples into several clusters and then tries to classify the samples by using artificial neural network as a classifier. Wang et al. also used a combined model like above one [7]. This model uses two algorithms; decision tree classification algorithm and Bayesian clustering networks. The purpose of this method is reducing the complexity of classification. With clustering at first, it is tried to make some categories of data samples and then each category will be classified independently.

Figure 1-2 shows an abstract of expressed methods.

3. Proposed model
Combined models which incorporate clustering and classifying cause data reduction and so classifier algorithm can classify the samples better. As an example, SVM classifier algorithm uses relationship optimization. This equation is “LaGrange equation”. Suggested methods for solving LaGrange equation are sensitive to the number of samples. If there are many samples, they can’t get the answer. Due to the above points, in this paper, a combined model will be presented which divides data into several classes and then uses classification algorithm for each category. The proposed model is as follows:

Step 1: First of all, the best properties will be selected by feature selection methods.
Step 2: Data is being clustered by using the clustering algorithms. Top-down hierarchical algorithm was used in this paper.
Step 3: Finally, combination of several classifiers is used for classification of samples in any cluster. Categories used in this paper are as follow:
- Support vector machine with Gaussian kernel
- Support vector machine with polynomials
- Support vector machine with quadratic kernel
- Decision tree
- Multilayered perception network

Figure 1-3 shows the above model. In the following, we will explain all the steps.
3.1. Feature selection
Any sample in data set is exhibited by features that express the data set. Feature selection means to select such properties that can determine a sample better and can perform properly in separating that one than another sample.
Different ways are suggested for feature selection and we use principle components analysis in this research.
Principle component analysis: this method that is considered as a dimension reduction technique tries to combine features with the propose of reducing the m. Indeed, this method reduces dimension by dimension transfer. PCA technique is the best way to reduce the data size linearly. It means that by omitting insignificant coefficients obtained from this transformation, missing information is less than other methods.

3.2. Clustering
Clustering is one of unsupervised branches of learning. This process is automatic and during that, samples are being divided into some categories that their members are similar to each other. These categories are called “cluster”. So cluster is a set of objects in which they are similar to each other and they are different from objects in another cluster. We can consider various criteria for similarity. For example, the distance criterion can be used and objects that are near to each other can be put in one cluster. Clustering algorithms includes two categories; hierarchical algorithm and dividing algorithm. Hierarchical algorithms construct clusters gradually (grow like crystals) but dividing algorithms carry out clustering directly [8]. Hierarchical algorithm: is one of the clustering algorithms which with regards to some criteria, analyzes data hierarchically and then, tries to create various clusters by dividing and conquer methods. These algorithms divide into two groups:
1- Aggregation algorithms (down to up): in these algorithms, at first, it is assumed that number of clusters is equal to the number of samples. In each step two or more clusters combine with each other and make a new cluster.
2- Dividing algorithms (up to down): in these algorithms, first of all, it is assumed that data set has one cluster and then in each step it will be tried to divide one cluster into two. This process carries out repeatedly to satisfy a certain criteria. In the present paper, a kind of hierarchical algorithm is suggested and it was used for clustering data samples. The performance of the algorithm is as follows:

Step 1: In the first step, we assume the data set forms a cluster and we call it “C1”. Then we will try to decompose the data.
Decomposition is as follows:
- We use below equation to solve the cluster center:
  \[ M_{i} = \sum_{j=1}^{n} f_{ij} \]
  Fi,j is determiner of ith feature magnitude in j th sample.
- For each sample in cluster we measure its Euclidean distance from center.
- Separate the sample which has the maximum distance from center and put it in a new cluster (Cnew)
- For each sample in cluster C1, measure its distance from center of the new cluster (Dnew). Then if this distance is smaller than distance from its own center, assign that sample to the new cluster. It means that
  \[ \text{If} D_{\text{new}} < D_{1}, \text{Then assign} S_{i} \text{to} C_{\text{new}} \]
- We check whether breaking cluster causes increasing efficiency or not (evaluation criteria). At this stage, by using a cluster we can classify the sets C1 and Cnew and measure the accuracy. If the accuracy of the set Cnew is more than C1, then breaking the two clusters is efficient and we continue this process. Otherwise, we don’t divide upper cluster into tow. Classifier used in this research is support vector machine that will be explained later.

Step 2: First step is learning phase of the above model, in the second stage, we should evaluate this model. In this stage one sample of total test (unseen) is given to this model. The distance of this sample from every cluster is being measured. Then, sample will be attributed to a cluster which has minimum distance. And so this sample is classified by using classification which was trained on that cluster and its normality or offensiveness will be determined...
3.3. Classification

Classification algorithms try to classify samples of a set based on their features. Many classification algorithms had been developed. At the present model, SVM classification algorithm is used for classifying and a standard for breaking cluster. After that the targeted cluster breaks into two, we measure the accuracy of the SVM classifier on the two separated clusters and the main one. If average size of the obtained accuracy of two separated clusters is more than the main cluster then breaking causes efficiency and we follow the process. Otherwise we stop fraction process.

The next usage of SVM classifier is after creation of clusters. After obtaining clusters, samples in Every cluster should be classified. If a cluster’s samples have an equal label, it doesn’t need classification and any sample that belongs to this cluster has that label. But if samples in the cluster have different labels (normal tag or attack tag), a combination of support vector machine classification, decision tree and MLP artificial neural network is used with the aim of assortment and obtained boundary for that cluster will be resulted.

3.4. Combination of classifications

In this research, two models of combination for combining classifiers were been used; Combiner model and selection model. We will illustrate both of them later.

3.4.1. Combining:

Combining the output of the classifiers means that the final decision is being influenced by output of every classifiers. Some methods are suggested for this operation that in this research, majority of votes is used and below equation shows it:

\[ H(t) = \max_{i=1}^C \sum_{t=1}^T d_{i,t} \]

3.4.2. Selection:
Selection means that for a specific data sample just one of the classifiers those participating in Forum section will decide. Indeed, Forum section output is the output of one of classifiers. The key question is that which of classifiers should be selected for a sample. Several methods were expanded that we will express them in the following.

There are various methods for selection. In this paper the nearest neighbor mechanism that is more common than other methods is used.

The nearest neighbor: in this method, for deciding on a sample, parameter K of the nearest neighbor of the targeted sample (in the training data set) is being obtained and then accuracy of any classifiers will be measured on K of the sample. Finally, any classifier that has higher accuracy is being selected for commenting. Neighbors are obtained on the base of the distance criterion. It means that K of any neighbor that has near distance will be measured.

4. Experiments

4.1. Data set

In this section KDD CUP data set at 1999 was used for targeted purpose. Of course only some parts of KDD data set were used. This is because the comparison with other methods.

Wang et al. after presentation of their method used 18285 samples of KDD data set for comparing their method with others [7]. These samples were selected randomly but they have some features that there are in table 3-4, as it is visible; this subset includes all samples of attack like Probe, R2L and also U2R. But because of large number of Dos samples, just 10000 samples were selected. Also 3000 normal samples had been selected randomly.

<table>
<thead>
<tr>
<th>Connection Type</th>
<th>Training Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>3000</td>
</tr>
<tr>
<td>Dos</td>
<td>10,000</td>
</tr>
<tr>
<td>Probe</td>
<td>4107</td>
</tr>
<tr>
<td>R2L</td>
<td>1126</td>
</tr>
<tr>
<td>U2R</td>
<td>52</td>
</tr>
</tbody>
</table>

Table (4-1) tested data set

4.2. Evaluation criteria

In this paper several evaluation criteria had been used. They are as follow:

Precision standard: equation 1 shows how this criterion can be measured.

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{1}
\]

Recall standard: this criterion is correct classification rate of positive samples and equation 2 shows it:

\[
\text{Recall} = \text{TPrate} = \frac{TP}{TP + FN} \tag{2}
\]

F-measure standard: this standard is used for inclusion the accuracy and TP rate of a single criterion. Equation 3 shows it:

\[
F = \frac{(1 + \beta^2).(\text{Precision}.\text{TPrate})}{\beta^2\text{.Precision} + \text{TPrate}} \tag{3}
\]

4.3. Results

With the aim of evaluating the purposed method, the above model had been tested on the data set determined in chart 4-1 and based on expressed criteria. Chart 4-2 till 4-6 show obtained results from purposed model with other methods.

Table(4-2) The results of experiment on Normal data sample.

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>Naïve Bayes</th>
<th>BPNN</th>
<th>FC-ANN</th>
<th>Our Model-Selection</th>
<th>Our Model-Combiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>91.22</td>
<td>89.22</td>
<td>89.75</td>
<td>91.32</td>
<td>100.0</td>
</tr>
<tr>
<td>Recall</td>
<td>99.41</td>
<td>97.70</td>
<td>98.20</td>
<td>99.08</td>
<td>96.81</td>
</tr>
<tr>
<td>F-value</td>
<td>95.14</td>
<td>93.27</td>
<td>93.79</td>
<td>95.04</td>
<td>98.37</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>Naïve Bayes</th>
<th>BPNN</th>
<th>FC-ANN</th>
<th>Our Model-Selection</th>
<th>Our Model-Combiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>99.84</td>
<td>99.69</td>
<td>99.79</td>
<td>99.91</td>
<td>97.73</td>
</tr>
<tr>
<td>Recall</td>
<td>97.24</td>
<td>96.65</td>
<td>97.20</td>
<td>96.70</td>
<td>100.0</td>
</tr>
<tr>
<td>F-value</td>
<td>98.52</td>
<td>98.15</td>
<td>98.48</td>
<td>98.28</td>
<td>98.37</td>
</tr>
</tbody>
</table>

327
Table (4-3) The results of experiment on Dos data samples

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>Naïve Bayes</th>
<th>BPNN</th>
<th>FC-ANN</th>
<th>Our Model-Selection</th>
<th>Our Model-Combiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>50.00</td>
<td>52.61</td>
<td>60.94</td>
<td>48.12</td>
<td>99.29</td>
</tr>
<tr>
<td>Recall</td>
<td>78.13</td>
<td>88.13</td>
<td>88.75</td>
<td>80.00</td>
<td>99.85</td>
</tr>
<tr>
<td>F-value</td>
<td>60.98</td>
<td>65.89</td>
<td>72.26</td>
<td>60.09</td>
<td>99.57</td>
</tr>
</tbody>
</table>

Table (4-4) The results of experiment on Probe data samples

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>Naïve Bayes</th>
<th>BPNN</th>
<th>FC-ANN</th>
<th>Our Model-Selection</th>
<th>Our Model-Combiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>33.33</td>
<td>46.15</td>
<td>57.14</td>
<td>93.18</td>
<td>100.0</td>
</tr>
<tr>
<td>Recall</td>
<td>1.43</td>
<td>8.57</td>
<td>5.71</td>
<td>58.57</td>
<td>87.07</td>
</tr>
<tr>
<td>F-value</td>
<td>2.74</td>
<td>14.58</td>
<td>10.39</td>
<td>71.93</td>
<td>92.80</td>
</tr>
</tbody>
</table>

Table (4-5) The results of experiment on R2l data samples

<table>
<thead>
<tr>
<th>Decision Tree</th>
<th>Naïve Bayes</th>
<th>BPNN</th>
<th>FC-ANN</th>
<th>Our Model-Selection</th>
<th>Our Model-Combiner</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>50.00</td>
<td>25.00</td>
<td>50.00</td>
<td>83.33</td>
<td>90.0</td>
</tr>
<tr>
<td>Recall</td>
<td>15.38</td>
<td>7.69</td>
<td>23.08</td>
<td>76.92</td>
<td>51.15</td>
</tr>
<tr>
<td>F-value</td>
<td>23.53</td>
<td>11.76</td>
<td>31.58</td>
<td>80.00</td>
<td>65.37</td>
</tr>
</tbody>
</table>

Table (4-6) The results of experiment on U2R data samples

5. Conclusion

In this paper we stated intrusion detection in computer networks and expressed methods for that. It was stated that most of existing methods are based on machine learning.

In this research, a combined model based on clustering and classification was presented. This model includes two layers. In the first layer clustering algorithm had been used with purpose of simplifying the input space and also decreasing calculations. In the second layer classification algorithm had been used for classifying existing samples in a cluster.

For evaluating this method, a part of KDD CUP data set was used.

References


