DoS ATTACK DETECTION SYSTEM BASED ON MULTIVARIATE CORRELATION ANALYSIS

Ms. Supriya S. Thakare, Ms. Parminder Kaur

Department of Computer Science and Engineering MGM’s JNEC, Aurangabad

ABSTRACT:

In today’s world the number of online applications are increasing. With the increasing number of online applications the threat to the security of these applications is also increasing. The increasing number of cyber attacks have challenged the security of these online applications. DoS is one such type of cyber-attack which aims at making the website and the resources of the server unavailable to the intended users. In this paper we propose an anomaly based detection for both known and unknown attacks. We propose an MCA based system for accurate traffic characterization by extracting the geometrical correlation between network traffic. We also propose a triangle area based technique which enhances and speedup the MCA process. For evaluating the system KDD Cup 99 dataset is used. KDD Cup 99 dataset has 41 features out of which 10 features with maximum statistical variance are selected.

Keywords: Denial-of-Service attack, network traffic characterization, multivariate correlation, triangle area.

[1] INTRODUCTION

The increasing number of cyber attacks has increased the threat to the security of online applications. One type of cyber-attack is DoS attack which affects the websites running on the server and the online resources of the server. Other kinds of cyber attacks are usually launched to cause a longtime effect and to hijack sensitive information, denial-of-service attack do not aim at breaching your security parameters. Rather it aims at making the website and the resources of the server unavailable to the intended user. In some cases DoS attack can also be used as a smokescreen for other malicious activities and to take down security applications. DoS attack is a highly noticeable event as it impacts the entire online user base. DoS attacks can last for several days, week or even months making it more and more destructive to any online organization. Different examples of DoS attacks includes: disrupting services of a specific user or system, flooding a network with traffic to prevent legitimate traffic from flowing, preventing a person from accessing a particular service and disrupting the connection between two specific machines thereby interrupting a service. Another example of DoS attack is e-mail bomb in which a large number of spam emails are sent indoor to disrupt the normal functioning of mail servers.
In DoS attack only a single internet connection is used to exploit software vulnerabilities or to exhaust server resources. In DDoS a attack is launched using multiple connections or a distributed internet network. Usually DoS attack attempts to target network infrastructure to saturate it with huge volumes of traffic.

DoS attack can be an application level attack or network layer attack. In application layer attack large number of request are send to the server which requires resource intensive handling and processing which makes the resources of server like RAM, memory unavailable to the intended user. In network layer huge amount of traffic is sent through the connection of network creating a clog and disrupting the communication in the network.

[2] RELATED WORK

J. Huff, Q. A. Zhu [2] proposed an anomaly based detection using hierarchical kohenen net. The objective of this intrusion detection system was to detect as many types of attacks as possible. In context of IDS for network traffic there are specific features in the packet header which are indicators of abnormal behavior. This IDS makes use of these packet features as input vectors for grouping the packets into normal and attack. The issue with this technique is selection of set of features which are to be used at each detection level of multilayer k-map.

Weiming Hu[3] proposed an Ada Boost based algorithm to construct an IDS with low computational complexity, high detection rate and low false alarm rate. Ada boost based algorithm is one of the most popular machine learning algorithms. In this algorithm decision stumps are used as weak classifiers. Decision rules are provided for both categorical and continuous features. These decision rules are used as weak classifiers. These weak classifiers for continuous and categorical features are used to form strong classifiers. In this way relation between these two types of features are identified without applying any forced conversion between continuous and categorical features. Overfitting of some weak classifier can be done easily.

In 2012 author Mohammd Sazzadul Hoque presented an IDS using genetic algorithms. Genetic algorithms are inspired by Darwin's theory of evolution. Solution to a problem solved by genetic algorithms uses an evolutionary process (it is evolved). Algorithm begins with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are then selected to form new solutions (offspring) are selected according to their fitness - the more suitable they are the more chances they have to reproduce. This is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied.

R. Shanmugavadivu[4] presented an anomaly based IDS. KDD cup 99 dataset is used as input. This dataset is divided into two subsets training dataset and testing dataset. The first component of the system classifies the input data into multiple classes. The number of classes is equal to the number of different attacks involved in IDS. As the KDD cup 99 dataset contains four different attacks, the training dataset is divided into five classes considering a separate class for normal data. The next step is construction of fuzzy rules. This system makes use of automatic methods for generation of rules. 1-length frequent items from attack and normal data are mined which are used to identify important attributes. Using these identified attributes and deviation methods set of definite and indefinite rules are generated. The definite rules generated are used and by fuzzifying these rules we obtain a set
of fuzzy if-then rules. These fuzzy rules obtained are then fed to fuzzy rule base for learning the system. In the testing phase, the test input data is data applied to fuzzifier. The output of this fuzzifier is passed on to inference engine. This inference engine compares the test input with rule base. Rule base contains a set of rules obtained from definite rules. The output of the definite engine is one of the value from the set {low, high}. Using a defuzzifier this value is converted into crisp values varied between 0 and 1. 0 denotes data is completely normal and 1 denotes data as completely attacked data.

Author Yu Chen[5] in 2008 gave an idea of collaborative detection of DDoS attack over multiple network domains. One type of DDos attack is flooding attack. The consequences of flooding attack are congestion on communication links, overflow in half open queues, imbalance between input output traffic on gateway. The main issue of flooding attack is, damage is already done when the consequences are observed. Therefore it is highly desirable to detect this attack the earliest possible time. The author in this paper has proposed a lightweight and low complexity architecture called distributed-change-point detection(DCD) using change aggregation tree(CAT). The DCD system architecture considers multiple autonomous system (AS). Each domain has central CAT server. The system has a hierarchical detection architecture with three layers in it.

Dheeraj pal[6] has proposed attribute subset selection with information gain and Genetic algorithm for creation of rules for detection. In this approach, GA is used to detect the attack with help of rules with less attribute, on Network Security Laboratory-Knowledge Discovery and Dat

[3] PROPOSED SYSTEM

Here we propose a more sophisticated DoS detection approach using multivariate correlation analysis (MCA). It is an anomaly based detection approach for detecting both known and unknown attacks. We also propose a triangle area map technique which will enhance and speedup the process of MCA. Features are extracted which are further used to classify the traffic as legitimate. In MCA approach geometrical correlation between network traffic features are extracted which are further used to classify the traffic as legitimate or as an attack.

FrameWork:

The whole detection process can be divided into 3 major steps.

Step 1 : Network traffic enters the internal network where protected servers are present. These servers are used to form traffic records for a well-defined time interval. Monitoring and analyzing at the destination reduces the overhead of detecting malicious activities by concentrating only on relevant inbound traffic.

Step 2: Here MCA technique is applied using the TAM generation module. TAM is used to enhance and speed up the MCA process. MCA technique is applied to the traffic coming from step 1 to extract the correlation between two distinct features within each traffic records. Any kind of network intrusion changes these correlation between features. These changes in the correlation are used to detect any intrusive activity.
All the extracted correlations which are called as triangle areas and are stored in triangle area map (TAAM). These TAM are used to replace the original basic features to represent the traffic record. This provides higher discriminative information to differentiate between legitimate and illegitimate records.

**Step 3:** This is the decision making step where we have two phases, training phase and testing phase. In the training phase legitimate traffic records are considered to generate normal profiles. These generated normal profiles are stored in a database. In the testing phase profiles for individual traffic records are generated. These tested profiles are compared with the respective stored normal profile. Any kind of dissimilarity between these two profiles indicates an intrusive behavior. A threshold based detection mechanism is used. If the dissimilarity is greater than a predetermined threshold the traffic record is flagged as an attack otherwise it is labeled as legitimate traffic record.

**Fig1: Framework for the proposed DoS attack detection system**

**KDD Dataset:**

The KDD cup 99 dataset is used to evaluate the proposed DoS attack system. KDD cup 99 dataset is a publicly available labeled benchmark dataset and it has been widely used
in the domain of intrusion detection research. The records of the KDD cup 99 dataset has in all 41 features. Out of these 41, 34 are real valued features.

**Multivariate Correlation Analysis:**

The behavior of the DoS attack traffic is different from the legitimate network traffic. This difference is seen in the statistical properties of the network traffic. These differences in the statistical properties can be used to detect any intrusive activity. MCA approach using triangle area map technique well describes these statistical properties. It can effectively extract the correlative information between the features within an observed data object that is a data record.

We consider an arbitrary data set $X=\{x_1, x_2, x_3, ..., x_n\}$, where $x_i=[f_1, f_2, ..., f_m]^T$, $(1 \leq i \leq n)$ represents the $i^{th}$ m-dimensional traffic record. We apply the concept of triangle area to extract the geometrical correlation between the $j^{th}$ and $k^{th}$ features in the vector $x_i$. The two features $f_i$ and $f_j$ can be interpreted as a 2D column vector, which can also be defined as a point on the Cartesian coordinate system. Then, on the Cartesian coordinate system, a triangle is formed using the origin. The area of the triangle $T_{j,k}^i$ is defined as:

$$T_{j,k}^i = \frac{1}{2} \left| (f_i,0) \cdot (0,0) - (0,0) \right|$$

where $(1 \leq i \leq n)$, $(1 \leq j \leq m)$, $(1 \leq k \leq m)$ and $j \neq k$. The proposed system selects a set of $s$ features from the set of $m$ features where $s < m$. To make a complete analysis all possible permutations of any two distinct features from the set of $s$ features are extracted and the corresponding triangle areas are computed for each distinct pair. A Triangle Area Map (TAM) is constructed in the form of a matrix and the triangle areas are arranged on the map with respect to their indices. For example, the $T_{j,k}^i$ is positioned on the $j^{th}$ row and $k^{th}$ column of the map $TAM^i$. The values of all the diagonal elements are set to zero. It can be seen that for every nondiagonal element we have $T_{j,k}^i = T_{k,j}^i$, $j \neq k$ both the positions represents areas of the same triangle. Therefore $TAM^i$ is a symmetric matrix having diagonal elements zero. Therefore for further calculations we consider only the lower triangle matrix. This reduces the time required for comparing the two TAMs.

The lower triangle of the $TAM^i$ is converted into a new correlation vector $TAM^i_{\text{lower}}$ lower denoted as follows:

$TAM^i_{\text{lower}} = [T_{1,2}^i, T_{2,3}^i, ..., T_{m-1,m}^i, T_{1,3}^i, T_{2,4}^i, ..., T_{m-2,m}^i, ..., T_{m-1,m-1}^i]$

We calculate $TAM_{\text{lower}}$ for every record $x' \in X$ which can be represented as $X_{TAM_{\text{lower}}} = \{TAM_{\text{lower}}^1, TAM_{\text{lower}}^2, ..., TAM_{\text{lower}}^n\}$

[4] EXPERIMENTATION
Training legitimate traffic records:

In the training phase we generate normal profiles for legitimate training traffic records. Suppose we have \( g \) legitimate training traffic records \( X_{normal} = \{x_{1,normal}, x_{2,normal}, \ldots, x_{g,normal}\} \).

The triangle area based MCA technique is applied to these legitimate traffic records and lower triangles of TAMs are generated for each legitimate record. The set of all lower triangles of TAMs are denoted by \( X_{TAM_{lower}} = \{TAM_{lower}^{normal,1}, TAM_{lower}^{normal,2}, \ldots, TAM_{lower}^{normal,g}\} \).

To measure the dissimilarity between the traffic records we use Mahalanobis Distance. MD distance measures the distance between two multivariate data objects by considering the correlation between variables. Following is the algorithm for normal profile generation:

**Required:** \( X_{TAM_{lower}}^{normal} \), it has \( g \) TAM_{lower}^{normal} one for each record

1. Calculate mean of all the TAM_{lower}^{normal} g records using: \( \bar{TAM_{lower}^{normal}} = \frac{1}{g} \sum_{i=1}^{g} TAM_{lower}^{normal,i} \)
2. Generate covariance matrix Cov for \( X_{TAM_{lower}}^{normal} \)
3. Calculate MD distance for all g legitimate records using \( MD_{normal,i} = MD(TAM_{lower}^{normal,i}, TAM_{lower}^{normal}) \)
4. Calculate the mean of all MD distances using: \( \mu = \frac{1}{g} \sum_{i=1}^{g} MD_{normal,i} \)
5. Calculate standard deviation using MD distance and mean MD distance \( \sigma = \sqrt{\frac{1}{g-1} \sum_{i=1}^{g} (MD_{normal,i} - \mu)^2} \)
6. \( Pro \leftarrow (N(\mu, \sigma^2), TAM_{lower}^{normal}, Cov) \)
7. Return \( Pro \)

The normal profile \( Pro \) consist of density estimation of the MDs between individual legitimate training traffic records \( TAM_{lower}^{normal,i} \) and the expectation \( TAM_{lower}^{normal} \) of the \( g \) legitimate training traffic records.

The MD is computed using the following equation:

\[
MD_{normal,i} = \sqrt{\frac{TAM_{lower}^{normal,i} - TAM_{lower}^{normal}}{Cov} (TAM_{lower}^{normal,i} - TAM_{lower}^{normal})}
\]

The covariance between two arbitrary elements in the lower triangle of a normal TAM is defined:

\[
\sigma(T_{i,j}, T_{k,l}) = \frac{1}{g-1} \sum_{i=1}^{g} (T_{i,j} - \mu_{i,j}) \cdot (T_{k,l} - \mu_{k,l})
\]

The distribution of the MDs is described by two parameters, namely \( \mu \) the and the \( \sigma \) of the MDs.
Threshold Selection:

The threshold can be calculated using following equation:

\[
\text{Threshold} = \mu + \sigma \times \alpha
\]

\(\alpha\) is ranged from 1 to 3 with an increment of 0.5.

Thus, if the MD between an observed traffic record \(x^{\text{observed}}\) and the respective normal profile is greater than the threshold, it will be considered as an attack.

Attack Detection:

The proposed triangle area based MCA approach is applied to the observed traffic records and the lower triangles \(\text{TAM}_{\text{lower}}^{\text{observed}}\) and \(\text{TAM}_{\text{lower}}^{\text{normal}}\) is generated. The MD distance between the \(\text{TAM}_{\text{lower}}^{\text{observed}}\) and \(\text{TAM}_{\text{lower}}^{\text{normal}}\) is computed using following equation:

\[
\text{MD}_{\text{observed}} = \sqrt{\left(\frac{\text{TAM}_{\text{lower}}^{\text{observed}} - \text{TAM}_{\text{lower}}^{\text{normal}}}{\text{cov}}\right)^2}
\]

We consider a threshold based anomaly detector. Normal profiles are generated from purely legitimate network traffic records. These normal profiles are used for comparing with the incoming traffic records. The dissimilarity between incoming traffic record and the respective normal profile is examined by the detector. If the dissimilarity is greater than the threshold, the record is labeled as an attack, otherwise it is labeled as a legitimate traffic record.

The detection algorithm is as follows:

**Required:** Observed traffic records \(x^{\text{observed}}\), normal profile \(\text{Pro}: (N(\mu, \sigma^2), \text{TAM}_{\text{lower}}^{\text{normal}}, \text{cov})\) obtained from the normal profile generation algorithm and parameter \(\alpha\) for threshold.

1. Generate \(\text{TAM}_{\text{lower}}^{\text{observed}}\) for the observed traffic record \(x^{\text{observed}}\)
2. Calculate MD distance for the observed traffic record using:
   \(\text{MD}_{\text{observed}} \leftarrow \text{MD}((\text{TAM}_{\text{lower}}^{\text{observed}}), (\text{TAM}_{\text{lower}}^{\text{normal}}))\)
3. Check whether the calculated MD falls within the threshold:
   if\((\mu - \sigma \times \alpha) \leq \text{MD}_{\text{observed}} \leq (\mu + \sigma \times \alpha)\) then
   4. return normal
   5. else
   6. return attack
   7. End if

Input to the System

The input to the system is KDD cup 99 dataset. The records of the KDD cup 99 dataset has in all 41 features. Out of these 41, 34 are real valued features. A subset of 10 features with maximum variance is selected. Following is the list of 10 selected features:
1. src_bytes : No of data bytes from source to destination (feature 5)
2. dst_bytes : No of data bytes from destination to source (feature 6)
3. wrong_fragment : No of wrong fragments (feature 8)
4. num_failed_logins : No of failed login attempts (feature 11)
5. logged_in : 1 - if success, 0 – if fail (feature 12)
6. num_access_files : No of operations on access control files (feature 19)
7. is_guest_login : 1 if guest login else 0 (feature 22)
8. dst_host_diff_srv_rate : % of connections to different hosts (feature 35)
9. dst_host_serror_rate : % of connections that have “SYN” error (feature 38)
10. dst_host_srv_rerror_rate : % of connections that have “REJ” error (feature 41)

KDD Cup 99 dataset has 6 types of attack (Teardrop, Smurf, Back, Pod, Neptune, Land) out of these we are detecting Teardrop, Smurf, Back.

[5] RESULT AND ANALYSIS

We calculate the detection rate for the three attacks. For analysis we are comparing the result of the proposed system with the result of existing system. Table1 gives the detection rate for the three attacks when evaluated with the existing system. Table 2 gives the detection rate for the same three attacks when evaluated with the proposed system.

The three graphs show a graphical representation of the comparison between the detection rates of the existing system and the proposed system for different values of alpha..

The first graph compares the detection rate for back attack. From the first and the second graph we see that the detection rate for back and smurf attack of proposed system has reduced as compared to the existing system. Whereas From the third graph we see that the detection rate for teardrop attack has increased.

<table>
<thead>
<tr>
<th>Records</th>
<th>1σ</th>
<th>1.5σ</th>
<th>2σ</th>
<th>2.5σ</th>
<th>3σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treadrop</td>
<td>84.61</td>
<td>84.61</td>
<td>89.61</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Smurf</td>
<td>92.0</td>
<td>96.0</td>
<td>97.0</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Back</td>
<td>93.75</td>
<td>95.75</td>
<td>97.75</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

Table1: Results of proposed system

<table>
<thead>
<tr>
<th>Records</th>
<th>1σ</th>
<th>1.5σ</th>
<th>2σ</th>
<th>2.5σ</th>
<th>3σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treadrop</td>
<td>71.50</td>
<td>63.92</td>
<td>57.93</td>
<td>52.81</td>
<td>48.45</td>
</tr>
<tr>
<td>Smurf</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Back</td>
<td>99.96</td>
<td>99.82</td>
<td>99.58</td>
<td>99.44</td>
<td>99.31</td>
</tr>
</tbody>
</table>

Table1: Results of existing system
Fig 2: DR comparison for Back

Fig 3: DR comparison for Smurf

Fig 4: DR comparison for Teardrop

[6] CONCLUSION
In this paper we have considered a threshold based anomaly detector. Normal profiles are generated from purely legitimate network traffic records. These normal profiles generated are used for comparing with the incoming traffic records. The dissimilarity between the incoming traffic records and the respective normal profiles is examined by the detector. If the dissimilarity is greater than the threshold, the record is labeled as an attack. Otherwise it is labeled as a legitimate traffic record.

In the proposed system we selected a set of features instead of considering all the features of a record. Because of this the detection rate for back and smurf reduced. So we concluded that there are some features which are targeted in a particular attack. The exclusion of these features affects the detection rate of that particular attack.

In the part of the future work, we will carry out analysis to find out the relation between the features of the record and a particular attack.

REFERENCES
Author[1]: Supriya Thakare pursuing ME CSE from MGM’s JNEC affiliated to Dr BAMU university. Completed BE CSE from Dr. BAMU university.

Author[2]: Parminder Kaur Prof. working with MGM’s JNEC CSE Department since 15 years. Completed BE from Pune university, ME from Dr. BAMU and PhD. pursuing in the area of spatial data mining and remote sensing.