Detection of Intrusions in KDD Cup Dataset using GA by Enumeration Technique

Vishal R. Chaudhary, R. S. Bichkar
Faculty, Department of Computer Engg. Modern Education Society’s College of Engg. and PG Student,
G. H. Raisoni College of Engg. Mgmt., Pune, India
Professor, Department of E&Tc. G. H. Raisoni College of Engg. & Mgmt., Pune, India

ABSTRACT: In the last decades, there has been a massive growth in network connectivity between computer system which has achieved boundless potential outcomes and opportunities. Sadly, security related issues have likewise expanded at the same rate. Computer systems become victims of such attacks. These attacks or intrusions are modified information that cause harm to the working system program or application programs, typically read or alter private data or render the system futile. Several techniques are used to prevent and detect such attacks or intrusions. This paper presents an efficient GA based methodology to produce the classification rules for Network intrusion detection system. The chromosome structure has been selected by applying enumeration technique in which the computational time required to produce the population is significantly reduced and near optimal rules are generated. These classification rules are used to find networking attacks or intrusions. The proposed system is applied on KDD Cup99 Dataset to yield more efficient and effective classification rules.

KEYWORDS: Intrusion Detection System (IDS), Genetic Algorithm (GA), Enumeration Technique, KDD Cup99 Dataset.

I. INTRODUCTION

IDS employing different techniques are available in commercial market such as IDS using data mining [15], fuzzy logic [16, 17] and hybrid technology [18, 19, 20]. These technologies have their related advantage and disadvantages.

Generally, an intrusion is an illicit action or access into system and person who tries to use intrusion to enter into the system and if becomes successful is called an intruder. Intrusion refers to any illegal action that affects integrity, confidentiality and availability of system. Detecting such malicious activities is called intrusion detection. An IDS is system software that monitors the malicious activities or policy violation and produce report [1, 2].

Network attacks can be categorized in four types [1, 2, 23]:

1. DOS: (Denial of Service) Attempts to make given service or resource to make busy. Hence legal user cannot access the system. e.g. smurf, neptune, teardrop etc.
2. R2L: unauthorised access from remote machine. e.g. guess_passwd, Imap, warezmaster, warezclient, multihop etc.
3. U2R: unauthorised access to local superusers (root) privileges. e.g. buffer_overflow, loadmodule, part, rootkit.
4. Probe: Intruder tries to gain information about system. e.g. satan, ipsweep, portsweep, nmap.

Intrusion detection systems are classified in two broad categories as Host based IDS (HIDS) and Network based IDS (NIDS) [2]. The HIDS monitors data on single or multiple host system with help of operating system, application program and files. The NIDS monitors the information across the network traffic, analyse the information that flows across such communication links.

Generally IDS is categorized by analysis approach as misuse detection and anomaly detection [2, 3]:
In misuse detection system [3], known signature or patterns are detected. However, the detection of unknown signature is the limitation of this system. In anomaly detection system, the behavior of network and change in behavior is observed. These systems are highly expensive.

In this paper, we present a GA based approach to intrusion detection (IDS). GA possesses some good characteristics. This approach has following benefits [4, 5]:

1. GA is a powerful optimization technique.
2. For one problem, many solutions are available with this system.
3. The system can be adopted new intrusions.

II. RELATED WORK

Several researchers have used GAs for IDS [6-14] for two purposes i.e. rule generation [13, 20] and feature selection [16, 18].

Pawar and Bichkar [6] have proposed an enumeration technique is used to initialize a chromosomes leading to reduction in time required for GA convergence. The paper uses DARPA Dataset and obtains remarkable detection rate.

Li [8] presents GA based approach to Network Intrusion Detection System (NIDS). The paper discusses the chromosome structure and implementation approach for rule generation.

Xia et al. [10] have used GA and genetic programming (GP) to detect NID.

Gong et al. [12] this paper have presented a software based approach to IDS using GA. The classification rules have been evolved through fitness function employing support and confidence.

Bridges and Vaughan [15] have used integrated data mining with fuzzy logic and genetic algorithm to detect network threat and anomalies. The genetic algorithm is used to identify the optimal parameter of fuzzy functions for selecting relevant network features.

Lu [21] has used genetic programming (GP) to evolve different classification rules on historical data to classify the intrusions. GP requires more data or time to train the system.

Crosbie and Spafford [22] have detected the network anomalies by different agent technique and genetic programming.

III. KDDCUP99 DATASET

KDDcup99 dataset [24] is prepared and managed by MIT Lincoln labs. This dataset includes wide variety of intrusions simulated in military network environment. Each record is information of TCP/IP connection with 41 features. The dataset contains labeled and unlabeled records.

The training dataset consist 24 attack types while additional 14 types are in test dataset.

<table>
<thead>
<tr>
<th>Category</th>
<th>Attack type</th>
<th>Training Data</th>
<th>Testing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Normal</td>
<td>97278</td>
<td>60593</td>
</tr>
<tr>
<td>DOS</td>
<td>back, neptune, pod, land, smurf, teardrop</td>
<td>391458</td>
<td>223298</td>
</tr>
<tr>
<td>U2R</td>
<td>buffer_overflow, loadmodule, perl, rootkit</td>
<td>1126</td>
<td>5993</td>
</tr>
</tbody>
</table>
Table 1. Number of records of different types of attacks in the KDDCup99 dataset

<table>
<thead>
<tr>
<th>Type</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2L</td>
<td>warezclient, warezmaster, guesspasswd, imap, ftp-write, multihop, phf, spy</td>
</tr>
<tr>
<td>Probe</td>
<td>satan, ipsweep, portsweep, nmap</td>
</tr>
<tr>
<td>Other</td>
<td>mscan, apachi2</td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>

IV. GENETIC ALGORITHM

Genetic algorithm is optimization technique based on the principle of evolutions and natural selections. The solution to the problem is encoded in chromosome like data structure and GA evolves population using operators like selection, crossover and mutation [4, 5].

Each parameter in a chromosome is called as gene. Genes are selected according to our problem definition [5]. These are encoded on bits, character or numbers. The set of generated chromosome is called a population. The fitness function is used to calculated “goodness” of each chromosome [4]. The algorithm for GA is given below:

Crossover rate = 0.5  
Mutation rate = 0.015

Initialize population.  
Evaluate each individual in the population.  
for specified number of generations do  
for size of the population do  
Select two individuals (with uniform probability) as parent1 and parent2  
Apply crossover to produce a new individual (child).  
Apply mutation to child.  
Calculate the fitness of each child with (1)  
end for  
end for  
Extract the best (highly fit) individuals as final solution.

V. GA BASED IDS

We have chosen six features of KDDCup99 dataset after doing analysis as shown below:

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>protocol</td>
<td>service</td>
<td>Flag</td>
<td>Src_byte</td>
<td>Dest_byte</td>
</tr>
</tbody>
</table>

The proposed system has two phases: training and testing.  
In training phase, set of classification rules are evolved by applying GA on training KDDcup99 dataset.  
In testing phase, the rules evolved from first phase are applied on KDDcup99 test dataset to determine the accuracy of the generated rules.
Feature name | Format | Symbol
---|---|---
Duration | Int | Number
Protocol type | String | Number
Service | String | Number
Flag | String | Number
Src_byte | Int | Number
Dest_byte | Int | Number

Table 2. Features selected for our system

**Fitness function:**
It is a function which merits the value of individual relative to rest of population; it is an objective function that used to determine the merit-rule.

In our system, we have used following simple fitness function as:

\[ \text{fitness} = \frac{TP + FN}{\sum \text{Records}} \] \quad \ldots (1)

**VI. EXPERIMENTAL SETUP AND RESULTS**

The major aim of our system is to focus on detection rate and to generate best of rules set for all different categories of attack types. For this we use KDDCup99 dataset. KDDcup99 offers Training dataset as “kddcup.data_10_percent” and Testing dataset “corrected”. The GA is executed for specified number of generation on training dataset to generate classification rule. These rules are then used to detect the intrusion in test data.

<table>
<thead>
<tr>
<th>KDDcup test Dataset</th>
<th>DOS</th>
<th>R2L</th>
<th>U2R</th>
<th>Probe</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>223298</td>
<td>5993</td>
<td>39</td>
<td>2377</td>
<td>311029</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attack detected</th>
<th>DOS</th>
<th>R2L</th>
<th>U2R</th>
<th>Probe</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>222611</td>
<td>4076</td>
<td>13</td>
<td>1665</td>
<td>231769</td>
<td></td>
</tr>
</tbody>
</table>

| Percentage (%)     | 99.69%| 68.01%| 33.33%| 67.52%| 73.42%|

Table 3. The results of Proposed Enumerated system.

In our system, four classes of attacks have been detected. The overall detection rate of 73.42% and the detection rate of DOS type of attack category is remarkable. Some of the following best rule has been generated by our system are given below.

1. If duration = 0 and protocol = icmp and flag = SF and src_byte = 1032 or 520 and dest_byte = 0 then attack_type = smurf.
2. If duration = 0 and protocol = tcp and flag = SF and src_byte = 54540 and dest_byte = 8314 then attack_type = back.
3. If duration = 0 and protocol = tcp and flag = REJ and S0 and src_byte = 0 and dest_byte = 0 then attack_type = neptune.
In this paper, we present IDS with GA by enumeration technique to detect the different types of attacks. The performance of our system is measure with the help of KDDcup99 dataset. The proposed technique provides good detection rate. If we use parallelism in population generation and in detection phase, the computational time of the system will be reduced and also impact on calculation of fitness of individual. Hence with this system, we can improve the performance of proposed methodology.

**Table 4. Comparison of results obtained by proposed Enumeration based system with results of other researchers.**

<table>
<thead>
<tr>
<th>Title</th>
<th>Author and year</th>
<th>Fitness Function</th>
<th>Detection rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOS Probe</td>
<td>U2R R2L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDS using GA</td>
<td>Mohammad Hoque at el.</td>
<td>NA</td>
<td>71.1</td>
</tr>
<tr>
<td>Troubleshooting Technique for IDS</td>
<td>Shaiq Akbar at el.</td>
<td>[ \text{fitness} = \frac{\text{TP} + \text{TN}}{2 \times \text{Records}} ]</td>
<td>93.7</td>
</tr>
<tr>
<td>Our System</td>
<td>-</td>
<td>[ \text{fitness} = \frac{\text{TP} + \text{TN}}{2 \times \text{Records}} ]</td>
<td>99.9</td>
</tr>
</tbody>
</table>

**VII. CONCLUSION**

In this paper, we present IDS with GA by enumeration technique to detect the different types of attacks. The performance of our system is measure with the help of KDDcup99 dataset. The proposed technique provides good detection rate. If we use parallelism in population generation and in detection phase, the computational time of the system will be reduced and also impact on calculation of fitness of individual. Hence with this system, we can improve the performance of proposed methodology.

**REFERENCES**