Data Preprocessing and Classification for Traffic Anomaly Intrusion Detection using NSLKDD Dataset

L. Gnanaprasanambikai and Nagarajan Munnusamy

1C.M.S. College of Science and Commerce, Bharathier University, Coimbatore, India.
gnanaambikai@gmail.com

2C.M.S. College of Science and Commerce, Bharathier University, Coimbatore, India.
mnaagarajan@gmail.com

Abstract

Network security is essential in the internet world. Intrusion Detection is one network security component. Anomaly Intrusion Detection is a type of intrusion Detection that captures the intrinsic characteristics of normal data and used it for detection process. To improve the performance of specific anomaly detector selecting the essential features of data and Generating good decision rule is important. This paper we present suitable feature extraction, feature selection and classification algorithm for a traffic anomaly intrusion detection in using NSLKDD dataset. The generated rules of classification process are initial rules of genetic algorithm.

Key Words: Traffic anomaly ids, genetic algorithm, feature extraction, feature selection, classification.
1. Introduction

With growth of Internet in the modern world, security threats to the computer systems and network has increased lot. The security threats affect the network security services. To control security threats number of technologies have been developed and deployed in organizations, for example, firewall, anti-virus software, message encryption, secured software protocols, and so on. In addition, to this Intrusion Detection is an important technology that existed for long time [1].

Intrusion Detection is a method that monitors the actions occurring in a computer system or in network and analyzing the actions for the indication of intrusions. Intrusion is any action that affects the integrity, confidentiality or availability of any network resources. To detect intrusions either in a computer system or in network, intrusion detection techniques are used [20]. Based on the monitoring activity, Intrusion Detection System (IDS) is classified into two types: Host based IDS (HIDS) and Network Based IDS (NIDS). HIDS run on individual hosts or devices in a network and monitors the incoming and outgoing packets for any undesirable actions [2]. NIDS are placed at any strategic point within the network and monitors the traffic for any undesirable actions [2]. Further, IDS are classified into two categories based on the detecting method, they are: Signature based IDS (or Misuse IDS) and Anomaly based IDS. Signature Based IDS detect intrusions based on observed data that matches pre-defined description about the intrusion action. The advantage of Signature based IDS is that accurate intrusion detection with low false positive rate and dis–advantage is lack of detecting new intrusions. Anomaly Based IDS detect intrusions by detecting anomaly from observed data which deviates from normal behavior. The advantage of Anomaly IDS is that detects new intrusions and dis–advantage generation of false positive [3].

In this paper, we present decision rules which are initial population for Genetic Algorithm process. Data Preprocessing is prior done and decision rules are created. This paper discusses the data preprocessing and decision rule generation to Genetic Algorithm process on Intrusion Detection System. Our objective is to present decision rules for Traffic anomaly Intrusion Detection, so we are experimenting with NSLKDD dataset, which is de-facto dataset anomaly intrusion detection.

Paper Organization

We have seen brief introduction of IDS. The rest of the paper is organized as follows. Section 2 Genetic Algorithm and its importance. Section 3 Related work relevant papers to this research paper. Section 4 presents Data preprocessing steps and its implementation. Section 5 presents Classification and Decision rule generation. Section 6 presents Data set used and experiment results. Section 7 presents Conclusion of the paper.
2. Soft Computing – GA

Soft Computing is a new research field which includes technologies for intelligent problem solving. The objective of Soft Computing is to exploit fault tolerance, partial truth and uncertainty problems by achieving tractability, robustness and low cost solutions [23]. Genetic Algorithm is one constituent of Soft Computing is used mimics biological evolution for finding optimal solution of a problem. In Genetic Algorithm, the individual solutions to the problem are represented as initial population. Each individual is evaluated for a fitness function. Based on fitness value, the individuals are selected as parents and new offspring are produced. The offspring to form next new generation. The above steps are repeated until a solution has found which satisfies the pre-defined termination condition. The area where GA can be applied some are medicine, machine learning, Engineering applications, networking and routing, wired and wireless communications[4]. Genetic Algorithm can be applied on number of tasks in IDS which included optimization, classification, and automatic model structure design.

3. Related Work

In this paper, we generate decision rules from the dataset. The decision rules are the initial population for the Genetic Algorithm task. A prior work to generate decision rules, Data preprocessing work is done namely feature extraction and feature selection. A number of related works to Data preprocessing, Genetic algorithm are done, these entire works main objective is to reduce false positive in intrusion detection system. In this section, the related works are categorized in two sections: Section 1 suggest papers related to Data Preprocessing and classification on a Intrusion Detection Dataset and Database. Section 2 suggests papers related to Data Preprocessing technique used in Genetic Algorithm Process in Network Intrusion Detection.

Section 1: [5] showed PCA as feature extraction method and KNN classifier on detecting category based attacks with reduced calculation time and accuracy. [6] proposed a combined Network IDS framework of PCA and Naïve Bayes classifiers for Intrusion Detection Dataset. [7] showed Genetic algorithm for feature selection and decision tree for classification for Signature based NIDS with experimental analysis and results. [8] proposed decision tree and CBA based classification model for misuse and anomaly detection with experimental results and analysis on KDD99 dataset, which is not effective dataset for anomaly intrusion detection. [9] surveyed decision tree algorithms for classification with various application, and suggested tools for implementing the algorithms. [10] gave descriptive review on network features and data preprocessing techniques used in anomaly intrusion detection and suggested PCA a good technique in data dimensionality compared to feature selection. [11] proposed a comparative performance on Decision tree and Rule based classifiers on multiple relational databases and their applicability on the
database with empirical results and observations. [12] proposed PCA with SVM for selecting feature subset.


4. Data Preprocessing

Data Preprocessing is one of the critical steps in data mining process which does the preparation and transformation of the original dataset. The various steps are included in Data preprocessing, they are Data cleaning, Feature reduction, Feature construction [10]. Feature Reduction includes Feature extraction and Feature selection. Feature extraction, selection and construction all are independent methods in data preprocessing. They can be combined depending on the problem analyzed like feature extraction followed by feature selection, feature construction followed by feature selection [21]. In this paper we follow the combination of feature extraction followed by feature selection.

4.1. Feature Extraction and Selection

Feature Extraction transforms data from high dimensionality to low dimensionality. Feature extraction is a process that determines what evidence that can be taken from audit data is most useful for analysis [24]. In this paper, Principle Component Analysis method is used for feature extraction. PCA is a linear method in dimensionality reduction for data analysis and compression. It is based on transforming a relatively large number of uncorrelated features by finding a orthogonal linear combinations of the original features with the largest variance [6].

Steps in PCA Algorithm

Step 1: Get the input data
Step 2: Find the mean
Step 3: Subtract the mean
Step 4: Calculate the Covariance matrix
Step 5: Calculate the Eigen vectors and Eigen values of the covariance matrix
Step 6: Sort the Eigen values in decreasing order and forming feature vector.
Step 7: Derive the new dataset with reduced features.
The input to the PCA program is the NSLKDD dataset. From the covariance matrix we will find the Eigen value and Eigen vector by using the equation

$$|A - \lambda I|\mathbf{x} = 0$$

(1)

Sort the Eigen vector corresponding to the Eigen value. The Eigen vector with the highest Eigen value represents the first principle component of the data. The K Eigen vectors with the highest Eigen values are selected for feature reduction by the equation

$$\frac{\bar{a}}{\sum_{i=1}^{k} l_j} = \text{threshold}$$

(2)

The scree plot plots the Eigen values on the graph. The K eigen values are selected by detecting elbow on the scree plot graph.

Fig 1: Scree Plot of PCA

In the above scree plot the elbow decides the selection of eigen values. There is one elbow in the screeplot, where k=4. Therefore, 41 features of NSLKDD dataset is reduced to 4 features [25].

4.2. Feature Set for Anomaly NIDS

In broad anomaly detection, separate feature sets are built for each of the anomaly detector. For, Traffic based anomaly NIDS, a feature set of Multiple connection derivative features is included, which specifies the count of connections to a particular destination IP address and port [10]. In this paper a feature set of 4 features namely protocol, src_bytes, Dst_host_count, Dst_host_same_srv_rate are constructed [10]. The feature set constructed for anomaly detection is used to determine DDOS attack [16].
5. Classification

At this stage, we generate decision rules for the records of the dataset with the reduced feature set with two classes namely normal and anomaly. [8] The decision rules of the normal record set are considered for anomaly intrusion detection. There are two ways of generating decision rules for a dataset: Direct and Indirect. Direct way is to generate decision rules directly from data. Indirect way is to generate decision rules from other classification models. We have compared Ripper Algorithm, PART Algorithm and C4.5 Algorithm decision rule generating algorithm and analyzed c4.5 Algorithm is suitable for our problem solving.

5.1. Decision Rule Generators

5.1.1. Ripper

RIPPER means Repeated Incremental Pruning to Produce Error Reduction. Ripper is a direct way rule generator for two class problem, rule is generated for one class and the other class is kept as Default class[17]. Rules are generated for class having minimum records in a Dataset, and the other class is kept as Default class[17]. Ripper Algorithm is implemented as JRip in Weka3.6.

5.1.2. PART

PART means Projective Adaptive Resonance Theory. Part Algorithm is a Combination of c4.5 and Ripper Algorithm. Part Algorithm is indirect way rule generator, which employs partial decision tree to generate decision rules[17]. Partial Decision tree is a regular tree that contain unexplored branches. In order to build this tree, subtree replacement is used as pruning strategy in building the tree. The Algorithm follows Divide-and-conquer strategy.

5.1.3. C4.5

C4.5 Decision Tree is a improvement to ID3 Decision Tree Algorithm. Decision tree is a indirect way of Decision rule generator. In C4.5 Decision tree algorithm, a decision tree is constructed using subtree replacement and subtree raising pruning strategies [22].

C4.5 uses Gain ratio criterion for choosing the best attribute for each decision node during construction of decision tree which gives larger information gain of an attribute[2] [22].

The process to generate Decision rules from Decision tree is straightforward. The algorithm generate decision rule for each leaf node of the decision tree. Conversion Algorithm of Decision tree to Decision rules generate the decision rules [22]. C4.5 is implemented as J48 in Weka 3.6.

Algorithm to Generate Decision Rules from a Decision Tree

Input: Decision Tree T
**Output:** Decision Rules R

**Algorithm**

R=0

For each path from root to a leaf in T do

a=True

for each non-leaf node do

a=a^(label of node combined with label of incident outgoing arc)

c=label of leaf node

R=R U r=<a, c>

5.2. Reason for choosing C4.5 Decision Tree Classifier

We used results from WEKA 3.6 for our comparison. The comparison is based on generation of maximum number of decision rules for normal class with minimum time. The table below gives the number of rules generated and time taken for it.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>No. Of Rules for Normal class.</th>
<th>Time taken</th>
<th>Accuracy of Correctly Classified Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIPPER Algorithm</td>
<td>Normal Class becomes Default class</td>
<td>2.13 seconds</td>
<td>99.0632%</td>
</tr>
<tr>
<td>PART Algorithm</td>
<td>20 rules</td>
<td>0.52 seconds</td>
<td>99.0076%</td>
</tr>
<tr>
<td>C4.5 Algorithm</td>
<td>31 rules</td>
<td>0.36 seconds</td>
<td>99.0433%</td>
</tr>
</tbody>
</table>

From the above table, we could find accuracy of correctly classified instances are same, for the above mentioned algorithms, the algorithms differ in number of rule generation. In case of RIPPER Algorithm, Decision rules are generated for anomaly class, making normal class as Default class which exploit the principle of anomaly intrusion detection [20][8]. In case of PART and C4.5 more number of normal class decision rules are generated in C4.5 as it follows Global optimization technique[27]. In our problem, we generate decision rules for initial population of Genetic Algorithm. Initial Population size of Genetic Algorithm should be large enough to explore the space of models more thoroughly and it helps in obtaining better optima [26]. Algorithm which follows Global optimization technique generates more decision rules. To our problem solving, we choose C4.5 as suitable Algorithm for classification.

6. Experiment Results

6.1. Dataset

NSLKDD dataset is de-facto dataset for anomaly intrusion detection. [19]
NSLKDD dataset is an advanced version of KDDCUP99 dataset with no redundancy, no duplication of records, and with less complexity level. NSLKDD dataset consists of 20% training dataset, full training dataset of 125973 instances, and testing dataset of 22544 instances. In this paper, we create decision rules from 20% training dataset.

### 6.2. Generating Decision rules

With PCA results and feature selection section 4, we select four features from 41 features in the dataset and form a reduced feature dataset. This reduced feature dataset serves as a dataset for generating decision rules. For our experiment, we use WEKA3.6 for classification. C4.5 decision tree is implemented as J48 tree classifier in WEKA3.6. In this paper, we use only 20% of the training dataset as input to J48 classifier and generated a decision tree below.

![Decision Tree in WEKA3.6](image)

By using the Conversion Algorithm in section 5.1.3 on the decision tree above, we could generate 31 decision rules with normal class [8] for anomaly intrusion detection. Decision rules of simple if then statement are used. The features are connected using a function in the if-then rule.

For example, one decision rule could be

If protocol<=1 and src_bytes>5 and src_bytes<=28 and dst_host_count<=254 and dst_host_same_srv_rate <= 0.67 then normal

### 7. Conclusion

In this paper, we generated decision rules for anomaly intrusion detection. To generate decision rules, a prior work of data preprocessing work is implemented.
with PCA as dataset dimensionality reduction. The features set used in decision rule serves for determining DDOS attack in turn identifying traffic volumes in the network. These Decision rules are generated for Traffic Anomaly Intrusion Detection and they serve as initial rules for Genetic Algorithm Process.

8. Future work

In this paper we generated decision rules from 20% of training dataset, as next step in future work to find the best fitness of the decision rules for Genetic Algorithm process using full training dataset and to generate fittest rules set to apply on testing dataset.

References


