A Rough Set Based Feature Selection on KDD CUP 99 Data Set

Vinod Rampure¹ and Akhilesh Tiwari²

Department of CSE & IT, Madhav Institute of Technology and Science, Gwalior (M.P), India
¹rampurevinod@yahoo.in, ²atiwari.mits@gmail.com

Abstract

In the present era as internet is growing with exponential pace, computer security has become a critical issue. In recent times data mining and machine learning have been researched extensively for intrusion detection with the aim of improving the accuracy of detection classifier. KDD CUP’ 99 Data set is the most widely used dataset in research domain. Selecting important feature on the basis of rough set based feature selection approach have lead to a simplification of the problem, faster and more accurate detection rates. In this paper, we presented an efficient approach for detecting relevant features from the KDD CUP’99 Data set.

Keywords: - intrusion detection, KDD CUP 99 intrusion detection Data set, feature relevance, information gain

1. Introduction

Internet and other area network are growing at a fast rate in current years, not just terms of shape and size, but also term of different changing the services offered. But some time is cyber-attacks by hackers and crackers misusing the internet protocol, important data and services. Several Protective techniques have been developed and implement to protect the computer system against the cyber-attack such as antivirus, firewall, encryption technique and other various protective measures. Even with all the techniques could not guarantee the full protection of the system. Hence, the need for a more active mechanism likes Intrusion Detection system (IDS) as next track of defense [13]. So the progressive use of intrusion detection system for handling the anomalies on web has caused multiple efforts arranged by the analysts. The intrusions have been found domination the internet which may be assumed as a threat to the security of authorized users. In order to meet the advantage of changing technological world, IDS has been implemented through various amendments where it is able to detecting intrusion exactly.

Therefore, intrusion detection is becoming increasingly important technique that deployed to monitor and find out the abnormal condition in the network system and identifies network intrusion such as anomalous network behavior, unauthorized network access, or malicious attack to computer system.

Intrusion detection can be categorized into two main approaches used misuse detection and second, anomaly detection. In Misuse detection, attacks can be represented in the form of pattern or a signature in order to detect or prevent same attack in future. In anomaly detection category, deviation of normal usage behavior pattern is identified in order to correctly detect the intrusion [10].

Pattern reorganization problem can be handled by intrusion detection system and it can also be classified as learning system. Selecting relevant feature is an important problem in learning systems. Bello proposed that selecting important attribute is useful for dimensionality reduction of training data sets. Speed of data manipulation and classification rate can be improved by reducing the influence of noise. Performance factor, such as, accuracy of classification is maximized in order to achieve exactly and
find a feature subset by using the concept of feature selection [9]. Feature selection is not an important issue in research domain. Selecting important features by using rough set theory makes the problem simple, faster and more accurate for detection rates. This paper explores feature selection KDD cup 99 data set by using concept of rough set theory.

This paper organized as follow: Section 2 present basic concept of rough set theory, Section 3 also present KDD CUP 99 Dataset, Section 4 explain proposed approach, Section 5 consist experiments result, finally conclusion and future work is mentioned in Section 6.

2. Basic Concept of Rough Set Theory

A rough set methodology is based on the premise that lowering the degree of precision in the data makes the data pattern more visible [1], whereas the central premise of the rough set philosophy is that the knowledge consists in the ability of classification. In other words, the rough set approach can be considered as a formal framework for discovering facts from imperfect data [3]. The results of the rough set approach are presented in the form of classification or decision rules.

2.1. Information System

Formally, an information system IS (or an approximation space), can be seen as a system [2].

\[
\text{IS} = (U, A)
\]

Where U is the universe (a finite set of objects, \( U = (x_1, x_2, \ldots, x_n) \)) and A is the set of attributes (features, variables). Each attribute \( a \in A \) (attribute a belonging to the considered set of attribute A) defines an information function \( f_a: U \rightarrow V_a \), where \( V_a \) is the set of values of a, called the domain of attribute a.

2.2. Indiscernibility Relation

For every set of attributes \( A \), an indiscernibility relation \( \text{Ind}(B) \) is defined in the following way: two objects, \( x_i \) and \( x_j \), are indiscernible by the set of attributes \( B \) in A, if \( b(x_i) = b(x_j) \) for every \( b \subset B \). The equivalence class of \( \text{Ind}(B) \) is called elementary set in \( B \) because it represents the smallest discernible groups of objects [8]. For any element \( x_i \) of \( U \), the equivalence class of \( x_i \) in relation \( \text{Ind}(B) \) is represented as \([x_i]_{\text{Ind}(B)}\). The construction of elementary sets is the first in classification with rough set.

2.3. Lower and Upper Approximations

The rough sets approach to data analysis hinges on two basic concepts, namely the lower and the upper approximations of a set referring to:

- The elements that doubtlessly belong to the set, and
- The elements the possibly belong to the set.

Let \( X \) denotes the subset of elements of the universe \( U \) (\( X \subseteq U \)). The lower approximation of \( X \) in \( B \) (\( B \subseteq A \)), denoted as \( BX \), is defined as the union of all these elementary sets which are contained in \( X \) [4].

More formally:

\[
BX = \{ x_i \in U \mid [x_i]_{\text{Ind}(B)} \subseteq X \}
\]

The above statement is to be read as: the lower approximation of the set \( X \) is a set of object, which belong to the elementary sets contained in \( X \) (in the space \( B \)).
The upper approximation of the set X, denoted as BX, is the union of these elementary sets, which have a non-empty intersection with X:

$$BX = \{x_i \in U \mid [x_i] \operatorname{Ind}(B) \cap X \neq 0\}$$

For any object $$x_i$$ of the lower approximation of X (i.e., $$x_i \in \overline{BX}$$), it is certain that it belongs to X. For any object $$x_i$$ of the upper approximation of X (i.e., $$x_i \in BX$$), we can only say that $$x_i$$ may belong to X. The difference:

$$\overline{BX} = BX - BX$$

is called a boundary of X in U.

2.4. Accuracy of Approximation

An accuracy measure of the set X in BA is defined as [5]:

$$\mu_B(X) = \frac{\operatorname{Card}(BX)}{\operatorname{Card}(BX)}$$

The cardinality of a set is the number of objects contained in the lower (upper) approximation of the set X. As one can notice, $$0 \leq \mu_B(X) \leq 1$$. If X is definable in U the $$\mu_B(X) = 1$$, if X is undefinable in U then $$\mu_B(X) < 1$$.

2.5. Core and Reduct of Attributes

In rough set theory, information table is used for describe of object in the universe, it consists of two dimensions, each row is an object, and each column is an attribute. Rough set theory classifies attribute in two types according to their roles of information table: core attribute and redundant attribute. Here the minimum condition attributes set can be received, which is called reduction [6]. One information table might have a several different reduction simultaneously. The intersection of the reduction is the core of information table and the core attribute are the important attribute that influences attribute classification [7].

A subset B of a set of attribute C is the reduction of C with respect to R if and only if

$$\operatorname{POS}_B(R) = \operatorname{POS}_C(R),$$

and

$$\operatorname{POS}_{B \setminus \{a\}}(R) \neq \operatorname{POS}_C(R),$$

for any $$a \in B$$.

And the core defined by the equation given below

$$\operatorname{CORE}_C(R) = \{c \in C \mid \forall c \in C, \operatorname{POS}_C(R)\}$$

3. KDD CUP 99 Data Set

KDD Cup’99 dataset used for benchmarking intrusion detection problem is used in our experiment. These are generated by processing the tcpdump segment of DARPA 1998 evaluation data set. This data set consists of 41 feature and separate feature (42nd feature) that labels the connection as ‘normal’ or a type of attack [11]. The data set contains a total of 23 attack, these are grouped into 4 major categories:

3.1. Denial-of-Service (DoS)

In Denial-of-service attack, the attacker has the goal of limiting or denying service provided to the user, computer or network. Attacker tries to prevent genuine users from using a service. It is usually done by making the resources either too busy or too full and overflow.
3.2. Probing or Surveillance

Probing or Surveillance attacks have the main aim of gaining knowledge of the existence or configuration of a computer system or the network. The attacker then tries to harm or retrieve information about resources of the victim network [12].

3.3. User-to-Root (U2R)

User-to-root attack is attempts by an unauthorized user to gain administrative privileges. The attacker starts outs with access to a normal user account on the system (perhaps gained by sniffing password, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.

3.4. Remote-to-Local (R2L)

Remote-to-local attack is the kind of intrusion attack where the remote intruder consistently sends packets to a local machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.

In training data set, 23 attack that appears which is organized into 5 major class labels those are given Table 1 below such as normal, R2L, U2R, Probe and DoS.

Table 1. Class Labels and the Number of Samples that Appears in “10%” KDD Dataset

<table>
<thead>
<tr>
<th>Attack</th>
<th>Original Number of Samples</th>
<th>Class level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back</td>
<td>2,203</td>
<td>DOS</td>
</tr>
<tr>
<td>land</td>
<td>21</td>
<td>DOS</td>
</tr>
<tr>
<td>Neptune</td>
<td>107,201</td>
<td>DOS</td>
</tr>
<tr>
<td>pod</td>
<td>264</td>
<td>DOS</td>
</tr>
<tr>
<td>smurf</td>
<td>280,790</td>
<td>DOS</td>
</tr>
<tr>
<td>teardrop</td>
<td>979</td>
<td>DOS</td>
</tr>
<tr>
<td>satan</td>
<td>1,589</td>
<td>PROBE</td>
</tr>
<tr>
<td>ipsweep</td>
<td>1,247</td>
<td>PROBE</td>
</tr>
<tr>
<td>nmap</td>
<td>231</td>
<td>PROBE</td>
</tr>
<tr>
<td>portswEEP</td>
<td>1,040</td>
<td>PROBE</td>
</tr>
<tr>
<td>normal</td>
<td>97,277</td>
<td>NORMAL</td>
</tr>
<tr>
<td>Guess_passwd</td>
<td>53</td>
<td>R2L</td>
</tr>
<tr>
<td>ftp_write</td>
<td>8</td>
<td>R2L</td>
</tr>
<tr>
<td>imap</td>
<td>12</td>
<td>R2L</td>
</tr>
<tr>
<td>phf</td>
<td>4</td>
<td>R2L</td>
</tr>
<tr>
<td>multihop</td>
<td>7</td>
<td>R2L</td>
</tr>
<tr>
<td>warzemaster</td>
<td>20</td>
<td>R2L</td>
</tr>
<tr>
<td>warzclient</td>
<td>1,020</td>
<td>R2L</td>
</tr>
<tr>
<td>spy</td>
<td>2</td>
<td>R2L</td>
</tr>
<tr>
<td>Buffer_overflow</td>
<td>30</td>
<td>U2R</td>
</tr>
<tr>
<td>-----------------</td>
<td>----</td>
<td>-----</td>
</tr>
<tr>
<td>Loadmodule</td>
<td>9</td>
<td>U2R</td>
</tr>
<tr>
<td>perl</td>
<td>3</td>
<td>U2R</td>
</tr>
<tr>
<td>rootkit</td>
<td>10</td>
<td>U2R</td>
</tr>
</tbody>
</table>

By using rough set theory based proposed algorithm we can select important features in these class level which is given Table 1 above.

4. Proposed Approach

We proposed a rough set based approach for feature selection on KDD Cup’99 Data set. The proposed algorithms are described as follows:

Input: The data set values.
Output: Return the selected feature from each class level.

Algorithm1 Proposed Feature selection algorithm:

Step 1 Load the dataset values $N_D$.  
Step 2 Repeat step 3 for all dataset values.  
Step 3 Manipulate the values of loaded dataset.

\[ M_D = \frac{F_V - M_F}{\sigma_F} \]

Where,

$M_D =$ Manipulated Feature Values  
$F_V =$ Original feature values  
$M_F =$ Mean of row wise feature values  
$\sigma_F =$ Standard deviation of feature vectors

Step 4 Set the Manipulated data values in new variable $A_T$.  
Step 5 Round off all the variable of $A_T$.  

\[ A_{T1} = \text{Round} \ (A_T) \]

Step 6 Initialize new variable ($A_{Tnew}$) by substituting the $A_{T1}$ values with corresponding column details.

\[ A_{Tnew} = [\text{Column number} \ A_{T1}] \]

Step 7 Compare the row parameters with column parameters.  
Step 8 Get Index values if row data and column data matches.  
Step 9 Count the total feature values when reduced data is obtained under the threshold limit.  
Step 10 The final exactly selected features are obtained by removing the reduced data.
5. Experimental Analysis and Results

The Experiment is performed in MATLAB 2012a. The processor used is intel core i7 and memory required 512 MB. The input data set used is KDD CUP 99 and rough set based approach is applied for selecting the optimal feature among the given 41 feature from 10% KDD CUP 99 Data set. The training dataset consisted of 494,021 records among which 92,277 (19.69%) were normal, 391,458 (79.24%) DoS, 4,107 (0.83%) probe, 1,126 (0.23%) R2L and 52 (0.01%) U2R connections [14]. The experimental result shown in Table 2:

<table>
<thead>
<tr>
<th>Class Level</th>
<th>Total number of features</th>
<th>Name of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>DoS</td>
<td>7</td>
<td>3, 23, 29, 30, 32, 34, 35</td>
</tr>
<tr>
<td>Normal</td>
<td>24</td>
<td>1, 3, 5, 6, 10, 12, 16, 19, 23, 24, 26, 28, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 41</td>
</tr>
<tr>
<td>Probe</td>
<td>22</td>
<td>1, 3, 4, 10, 12, 23, 24, 25, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41</td>
</tr>
<tr>
<td>R2L</td>
<td>19</td>
<td>1, 6, 10, 12, 19, 22, 23, 28, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40</td>
</tr>
<tr>
<td>U2R</td>
<td>13</td>
<td>6, 11, 12, 14, 17, 24, 32, 33, 35, 36, 37, 40, 41</td>
</tr>
</tbody>
</table>

6. Conclusion and Future Work

Feature selection is a preprocessing part of an intrusion detection system. In this Paper, analysis of the various features of the KDD CUP 99 Dataset is done to find the optimal selection feature using rough set theory based approach in order to maximize the accuracy, simplify the problem and makes the processes faster for detecting the intrusions in a IDS. The basic concept of reduct and core has been applied to efficiently improve the detection rate.

We plan to extend the work in term of accuracy by focusing on fusion of classifiers after a set of optimum feature subset is obtained.

Reference


Authors

Vinod Rampure is currently pursuing the M.tech degree at the department of CS/IT from MITS Gwalior (M.P), India. He received him B.tech degree from Mahatma Gandhi Chitrakoot University, Chitrakoot Satna (M.P), India. Him current interest in data mining, rough set, classification and their application.

Dr. Akhlesh Tiwari has received Ph.D. degree in Information Technology from Rajiv Gandhi Technological University, Bhopal, M.P. (India). He is currently working as Associate Professor in the Department of CSE & IT, Madhav Institute of Technology & Science (MITS), Gwalior, India. He has guided several theses at Master and Under Graduate level. His area of current research includes Knowledge Discovery in Databases and Data Mining, Wireless Networks. He has published more than 20 research papers in the journals and conferences of international repute. He is also acting as a reviewer & member in editorial board of various international journals. He is having the memberships of various Academic/ Scientific societies including IETE, CSI, GAMS, IACSIT and IAENG.