A Rough Set based Feature Selection Algorithm for Effective Intrusion Detection in Cloud Model

S. Muthurajkumar, K. Kulothungan, M. Vijayalakshmi, N. Jaisankar and A. Kannan
Department of Information Science & Technology, Anna University, Chennai-25.
muthurajkumarrss@gmail.com, kulo_tn@annauniv.edu, vijim@annauniv.edu, jaisasi_win@yahoo.com, kannan@annauniv.edu

Abstract—There exist many problems in intrusion detection systems such as large data volume, features and data redundancy which seriously affect the efficiency of the detection algorithm. Such problems need to be addressed in developing reliable intrusion detection systems. In this paper, we propose an intrusion detection model that combines Rough Set based Feature Selection Algorithm and Fuzzy SVM for effective intrusion detection in the Cloud. The algorithm evaluates the characteristics of the attribute weights for security in the Cloud model and generates an optimal number of features in order to achieve best trade-off between detection rate and false alarm rate by using rough sets. This model solves the problem of feature redundancy and hence helps to improve the speed of evaluation. The experimental results obtained from this work show that the proposed model solves the feature selection problem and classification for intrusion detection effectively. It also achieves better detection performance for various types of attacks with reduced detection time.

Index Terms—Component, Cloud model, Feature Selection, Intrusion Detection System, Fuzzy SVM, Rough Set.

I. INTRODUCTION

Intrusion Detection System (IDS) plays an important role in cloud security. It determines the security level of the cloud through the collection and analysis of data pertaining to various threats, and logs them in record form so that it is easy to audit and evaluate the security of cloud. However, due to various problems, the false alarm rate is too high in current IDS, leading to the generation of large amount of IDS log data, with lots of redundancy. Therefore, it becomes a challenging task to optimize the data size. Feature selection is widely used to achieve the reduction of the number of attributes and also helps to reduce the number of log records. In cloud computing, there are Cloud providers, who offer a specific virtualized infrastructure [1]. An important security challenge in Cloud computing is that the user cannot control over the infrastructure being used fully since the user can only control the used services on the specific layer of the cloud. Therefore, it is necessary to provide effective security mechanisms to cloud.

Security to cloud can be provided in any one of the three ways namely key management, access control and intrusion prevention. Among them, intrusion prevention is the most important security issue. In order to provide effective intrusion prevention, efficient intrusion detection systems are necessary. Intrusion detection systems can be developed by proposing effective classification techniques. A number of classification
techniques such as, Bayesian classification, decision trees and SVM are present in the literature [8]. Among them ID3 is an important classification algorithm.

As a generalization of ID3 [8] [12], Fuzzy ID3 [9] uses the fuzzy classification entropy as the criterion to measure the uncertainty of the nodes. A similar version of fuzzy ID3 was proposed by Yuan and Shaw [10], in which the classification ambiguity is used as the criterion to measure the uncertainty of the nodes. Rough Sets proposed by Pawlak [11] are useful techniques to deal with uncertainty in knowledge. Dubois and Prade [12] first proposed fuzzy rough sets, the generalizations of rough sets to deal with both fuzziness and vagueness in data, which integrates fuzzy sets and rough sets together. It is reasonable to believe that fuzzy rough sets are better than fuzzy sets or rough sets to handle uncertainty in complex applications.

The research on fuzzy rough sets focus on approximations of fuzzy sets [13] [17], fuzzy rough attribute reductions [14] [15] [18] [19] and fuzzy rough feature selections [16] [20]. Therefore, it is essential to study how to effectively make use of multiple fuzzy attribute reductions to extract fuzzy classification rules. This is necessary because each fuzzy conditional attribute has a different significance, and therefore has different contribution to classification. More specifically, some of these attributes may have great contribution to classification, and the others may have little or even no contribution. It implies that the important fuzzy conditional attributes will play a crucial role in the inference process.

In this paper, we propose a new intrusion detection model which is the combination of new generalized Rough Set based Feature Selection Algorithm (RSFSA) and Fuzzy SVM to detect the intruders in Cloud. Firstly, Cloud model generates different clouds and evaluates the characteristics of the attribute weights. After that Rough Set based Feature Selection Algorithm generates the optimal feature subsets which achieve the best trade-off between detection rate and rate of false alarm. Experimental results show that the model can not only solve the feature selection problem of intrusion detection effectively, but also achieve balanced detection performance on different types of attacks, with better convergence at the same time.

II. RELATED WORKS

There are many works on cloud security that are present in the literature [1][2][6][10]. Liu-Hong Zhou et al [1] proposed an approach called EFSA-CP to intrusion detection based on Cloud model and improved multi-objective Particle Swarm Optimization. Sebastian Roschke et al [2] focused on proposing deployment architecture of Intrusion Detection Systems in the Cloud. They discussed and listed several existing threats for a Cloud infrastructure and are motivated to use Intrusion Detection Systems (IDS) and its management in the Cloud.

Saman Taghavi Zargar et al [3] proposed a Intrusion Detection system which is distributed in nature is capable of preventing intrusions effectively. Moreover, it has been deployed cloud in order to provide security to the cloud resources. It works by collaborating with peers to avoid distributed Denial of Service attacks. Yiming Liu et al [4] proposed an intrusion detection for cloud environment which uses a statistical waveform techniques for classification. Hisham and Fabrizio [5] proposed a cloud based intrusion detection system for cloud security which can overcome the deficiencies of current IDSs.

Yimei Jin [7] proposed a new intrusion detection system architecture, which can be used by cloud intrusion detection. Their new architecture uses the inaccurate hashing to obtain a fast packet inspection and applies a new stateful rule DFA to track session inspection process. S. Ganapathy et al [8] have developed an intelligent multi level classification algorithms that focuses on intrusion detection in Mobile Ad-hoc Networks. which developed by combining decision tree with an Enhanced Multiclass SVM algorithm. Jaisankar et al [9] proposed a new IDS in which the used fuzzy rough sets with C4.5 for improving the detection accuracy. Comparing with all these algorithms and methods, the IDS proposed in this paper is different since it uses rough sets for feature selection and classification using new fuzzy classifier.

III. SYSTEM ARCHITECTURE

The Architecture of the intrusion detection system proposed in this paper is shown in fig. 1. This intrusion detection system consists of four modules namely User Interface Module, Feature Selection Module, classification Module and Prevention module.

The user interface module collects the networks data from the data set. The feature selection module selects the necessary features based on genetic algorithms. The Classification module is used to classify the data by using the Fuzzy SVM [21]. The validation module decides whether the decision made by the classification module on the first set of records is valid when it is applied on the other records.
IV. ROUGH SET BASED FEATURE SELECTION ALGORITHM

In this work, we propose a rough set based feature selection algorithm using the definitions provided in [11] for Information System (IS). According to this, an information system IS is defined as IS = (U, W, A, V, F) where U and W are finite universal sets in which A and v are attributes and values. F is a function from U to power set of W.

We apply the rough set theory relation of operators in describability relation of A called IND (A). Moreover, generalized rough sets are defined in this work based on the IS described above. We apply reduction and detection algorithm in rough sets for effective decision making under uncertainty.

The steps of the proposed rough set based feature selection algorithm are as follows:

For each attack type I do
Begin
   Step 1: Read the attributes A of U and W
   Step 2: Compute IND (B)
   Step 3: Compute IND (B) * (Dk) for every Dk ∈ U/D.
   Step 4: Compute the matrix Ci for decision parameters.
   Step 5: Form the key attribute using CORE.
   Step 6: Remove those elements Ci = NULL or K with non empty overlap with the core.
   Step 7: Define f0 (IS).
   Step 8: Compute g0 (IS) using f0 (IS)
   Step 8: Find the Set of reduced attributes using g > 0.6
   Step 9: Output the set of reduced attributes for the attack type i.
End

V. RESULTS AND DISCUSSION

A. Dataset

Experiments have been carried out on data from cloud dataset, which is widely used to analyze IDS log data. Each sample in database contains 42 attributes, first 41 conditions attributes, where exists 9 discrete types, and 32 continuous type. In order to verify the validity of EFSA-CP algorithm, we randomly collect 10% of the dataset which contains 492 000 records. We choose a dataset comprises of 5000.

B. Experiment and analysis

We use RSFSA algorithm to search for feature subsets of 6 kinds of dataset, the results shown in Table 1. In the left column of table, ALL includes Normal class, ATTACK represents type of attack (including DOS, PROBE, R2L, U2R), the other as a separate type of attack. Right column is the subsets selected for each type of attack. For example, for U2R attack selected feature subset is 1, 2, 23, 31, where figures indicate the ID of the 41 features (initially 0). As we see from the table, properties of service and srv_count are the common properties of selected subsets, besides the number of feature subset is 5 or so, and the maximum number of selected features is for ALL.

Next, we build intrusion detection model respectively for the data without feature selecting and selected feature subsets. The results are shown in Table 2 from which we can see that, the model of RSFSA id better on detection accuracy rate. Table 3 shows detection rates of different algorithm. We can notice that RSFSA algorithm have much higher detection rate generally on 4 kinds of attacks.
TABLE I. ATTRIBUTE SUBSETS SELECTED FOR DIFFERENT KINDS

<table>
<thead>
<tr>
<th>Type Attribute</th>
<th>Subsets selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>DoS</td>
<td>[protocol_type, service, srv_count, same_srv_rate, dst_host_count]</td>
</tr>
<tr>
<td>Probe</td>
<td>[service, root_shell, srv_count, same_srv_rate, dst_host_count]</td>
</tr>
<tr>
<td>R2L</td>
<td>[service, num_file_creations, srv_count, same_srv_rate]</td>
</tr>
<tr>
<td>U2R</td>
<td>[protocol_type, service, srv_count, dst_host_count]</td>
</tr>
</tbody>
</table>

TABLE II DETECTION ACCURACY COMPARISONS BEFORE AND AFTER SELECTING ATTRIBUTES

<table>
<thead>
<tr>
<th>Types of Attacks</th>
<th>Detection accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before Selecting</td>
</tr>
<tr>
<td>Normal</td>
<td>87.8</td>
</tr>
<tr>
<td>Probe</td>
<td>88.0</td>
</tr>
<tr>
<td>DoS</td>
<td>84.9</td>
</tr>
<tr>
<td>U2R</td>
<td>85.4</td>
</tr>
<tr>
<td>R2L</td>
<td>87.5</td>
</tr>
</tbody>
</table>

TABLE III COMPARISON OF DETECTION RATES CORRESPONDING TO DIFFERENT ALGORITHMS

<table>
<thead>
<tr>
<th>Type of Attacks</th>
<th>Wenke Lee’s (%)</th>
<th>BP Network (%)</th>
<th>EFSACP + FSVM (%)</th>
<th>RSFSA + FSVM (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probe</td>
<td>90.0</td>
<td>97.47</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>DoS</td>
<td>79.9</td>
<td>92.71</td>
<td>94.85</td>
<td>99.2</td>
</tr>
<tr>
<td>U2R</td>
<td>75.0</td>
<td>48.0</td>
<td>96.0</td>
<td>97.7</td>
</tr>
<tr>
<td>R2L</td>
<td>60.0</td>
<td>95.73</td>
<td>92.86</td>
<td>96.2</td>
</tr>
</tbody>
</table>

Table 2 shows that the detection accuracy between before and after feature selection. From Table 2, it can be observed that the detection accuracy is gradually increased after feature selection. Table 3 shows the detection rates for the proposed intrusion detection model. From Table 3, it can be observe that the detection rate is better when it is compared with the existing methods after combining the Fuzzy SVM Classifier.

Fig. 2 shows the detection accuracy of before feature selection and after feature selection. From fig. 2, after feature selection detection accuracy is better than the before selecting the optimal number of features.
Fig. 3 shows the performance analysis of proposed feature selection algorithm and existing feature selection algorithm. From fig. 3, the performance of RSFSA is better when it is compared with the performance of EFACP.

From the results of experiments (2) and (3), for those containing only attack types and those also including Normal type, RSFSA gets much bigger feature subset, smaller fitness value and faster convergence speed from the former dataset than that of the latter. Therefore, the algorithm processes much better on the data containing only attack types. In conclusion, under the premise of effective classification, RSFSA algorithm can extract a smaller number of feature subset, which contains those important features reflecting the security of the system. Therefore, it improves the speed of detection and safety analysis.

VI. CONCLUSION AND FUTURE ENHANCEMENTS

This paper presents a new intrusion detection model which is the combination of rough set based feature selecting algorithm and Fuzzy SVM for cloud. The proposed model provides high detection rate and reduces the false alarm rate. The major advantages of the proposed model are effective feature selection to reduce the number of decision attributes and size of log data and better classification to enhance the detection accuracy. Future works in this direction can be the use of intelligent agents for effective decision making in classification.

REFERENCES


