A Review of Intrusion Detection Technique Based On Classification and Feature Optimization
Jaykaran Singh¹ Sunil Chavda² Mukesh Tiwari³
¹H.O.D., ²M. Tech Student, ³Director
1,2,3Electronics and Communication Depatment
1,2,3Sri Satya Sai Institute of Science & Technology, Sehore, M.P., India

Abstract— Classification and optimization play an important role in intrusion detection system. The process of classification depends on data mining and neural network approach. For the improvement of classification result, feature optimization is used. The process of feature optimization such as genetic algorithm, particle of swarm and ant colony optimization improved the detection of intrusion detection system. In this paper review of intrusion detection technique based on classification and feature optimization process, the process of classification consist of binary classifier, rule based classifier. The binary classification technique plays a major role in intrusion detection system. Feature optimization is a process of feature reduction, which gives the reduced feature set for intrusion detection. The reduced feature improves the classification ratio of intrusion detection system and decreases the false and negative detection of intrusion.

Keywords: IDS, classification feature optimization and KDDCUP99

I. INTRODUCTION

Due to the rapid increase in the illicit network activities, intrusion detection system (IDS) as a component of defense-in-depth is very necessary because traditional firewall techniques cannot provide complete protection against intrusion. Currently, the greatest threat against the security of networks and information systems, are attack on the network infrastructure. Intrusion Detection (ID) is an active and important to explore area network security [1]. There are several methods for intrusion detection expert systems such as statistical analysis and state transition implementation approaches, etc., and these several approaches on the immune system have been proposed in recent years [9, 10]. The goal of intrusion detection is to detect unauthorized access or misuse of computer systems by insiders of the system and external access and secure system integrity, privacy, usability and availability. Reduction unit means an important role in intrusion detection system [11, 12]. The various authors and explorations contribute data in the reduction of the function of penetration. The feature space with limited features that really contributes to the classification which reduce the cost of pre-processing and minimizes the impact of "peak" phenomenon in the classification [8]. The reduction process decreases the large number of attributes and improved the detection of intrusion detection system. In the process of feature reduction various algorithms are used that applied principle of component analysis and neural network. The reduction process used PCA method which is static reduction technique, reduces only fixed number of attributes. The fixed number of feature reduction process does not justify the value of feature it directly reduces the feature. On the consideration of computational time feature reduction is also an important aspect, the reduced feature increase the processing of detection ratio. Many methods have been proposed in last decades on the designs of IDSs based on feature reduction technique. For example silakari and salienda [4] proposed a generic framework for intrusion detection based on feature reduction and ensemble based classifier. On the other hand genetic algorithm is directly applied for classification in the work of Li [5]. Jain and Upendra [7] applied information gain based feature reduction for intrusion detection. They used KDDCUP’99 dataset for comparing four machine learning algorithms and they found that J48 classifier outperforms over BayesNet, OneR and NB classifiers. Muda et al. [7] also used KDDCUP’99 dataset for evaluating their K-Means and Naive Bayes based learning approach to carry out intrusion detection. Support Vector Machine (SVM) based IDS with Principal Component Analysis (PCA) dimension reduction is presented for intrusion detection in [8, 9]. Z. Xue-qin et al. [10] proposed SVM IDS with Fisher score for feature selection. Zhang and M. Zulkernine [11] applied random forests for network intrusion detection. in this paper used ID3 algorithm for feature selection. ID3 is attribute based classification technique in decision tree. Common problems with anomaly-based systems are that, they often require extensive training data for artificial learning algorithms, frequent updates. Network-based IDSs collect audit data from the network traffic. Network-based IDSs offer several advantages. First, network-based IDSs can take advantage of the standard structure of network protocols, such as TCP/IP. This is a good way to avoid the confusion resulting from heterogeneity in a distributed system. Second, network-based IDSs usually run on a separate (dedicated) computer; thus, they do not consume the resources of the computers that are being protected and most importantly it fits most the real time applications. Any of these IDs scheme is not a silver bullet. So some researchers moved to hybrid intrusion detection schemes. Section-I gives the introduction of the intrusion detection. Section-II gives the related of intrusion detection system. Section-III gives the problem formulation in intrusion detection and have been reviewed in section-III. In section IV discuss KDDCUP99 data set and empirical evaluation parameter. Finally, in section-V, conclusion and future scope are presented.

II. RELATED WORK

In this section we discuss related work in current scenario of intrusion detection technique using classification and
In this paper, authors investigate with Soft Computing paradigm of evolutionary computation techniques for synthesizing intrusion detection programs on Mobile Ad hoc Networks. We evolve the programs to detect such as Ad hoc Flooding, Route Disruption and Dropping Attacks against Mobile Ad hoc Networks. Since prevention techniques cannot be sufficient and new intrusions continuously emerges, IDS is an indispensable part of a security system. IDS detect possible violations of a security policy by monitoring system activities and response. If we detect an attack once it comes in to the network, a response can be initiated to prevent or minimize damage of the system. The two major techniques for machine learning were highlighted, with the use of Genetic Algorithm and Artificial Neural Network providing intrusion system with extra intelligence.

In this research paper, an optimized intrusion detection mechanism using soft computing techniques is proposed to overcome performance issues. The KDD-cup dataset is used that is a benchmark for evaluating the security detection mechanisms. The Principal Component Analysis (PCA) is applied to transform the input samples into a new feature space. The selecting of an appropriate number of principal components is a critical problem. So, Genetic Algorithm (GA) is used in the optimum selection of principal components instead of using traditional method. The Support Vector Machine (SVM) is used for classification purpose.

In this paper, the current work presents an evaluation of different neural networks such as Self-organizing map (SOM), Adaptive Resonance Theory (ART), Online Back propagation (OBPROP), Resilient Back propagation (RPROP) and Support Vector Machine (SVM) towards intrusion detection mechanisms using Multi-criteria Decision Making (MCDM) technique. The results indicate that in terms of performance supervised NNs are better while unsupervised NNs are better regarding training overhead and aptitude towards handling varied and coordinated intrusion.

In this paper, a hybrid approach based on GSA and k-Means (GSA-k-Means), which uses the advantages of two algorithms, is presented. The performance of GSA-k-Means is compared with the base-line k-Means and GSA-based classifier in terms of accuracy, FAR and hit rate. We also improve the performance of IDS in terms of detection accuracy and reduce FAR. Clustering is a significant and effective task in data mining used in many applications such as IDS to detect novel attacks. Clustering refers to grouping data into diverse clusters which similar data are grouped in a cluster that is dissimilar with another.

In this paper authors propose cooperative network intrusion detection Based on Fuzzy SVMs. Three types of detecting agents are generated according to TCP, UDP and ICMP protocol. This method divides the network data flow with network protocol, so it can improve the speed of each detection agent and the accuracy rate of prediction. Meanwhile, experiment results prove the method. But the accuracy rate of the UDP detection agent is low because of too lack of attack records in training set.

In this paper, a survey on intelligent techniques for feature selection and classification for intrusion detection in networks based on intelligent software agents, neural networks, genetic algorithms, neuro-genetic algorithms, fuzzy techniques, rough sets, and particle swarm intelligence has been proposed. These techniques have been useful for effectively identifying and preventing network intrusions in order to provide security to the Internet and to enhance the quality of service. In addition to the survey on existing intelligent techniques for intrusion detection systems, two new algorithms namely intelligent rule-based attribute selection algorithm for effective feature selection and intelligent rule-based enhanced multiclass support vector machine.

Author, in this paper, propose a new approach, called FC-ANN, based on ANN and fuzzy clustering, to solve the problem and help IDS achieve higher detection rate, less false positive rate and stronger stability. The general procedure of FC-ANN is as follows: firstly fuzzy clustering technique is used to generate different training subsets. Subsequently, based on different training subsets, different ANN models are trained to formulate different base models. Finally, a meta-learner, fuzzy aggregation module, is employed to aggregate these results. Moreover, other data mining techniques, such as support vector machine, evolutionary computing, outlier detection, may be introduced into IDS. Comparison of various data mining techniques will provide clues for constructing more effective hybrid ANN for detection intrusions.
I. In this paper, author proposes a soft Computing technique such as Self organizing map for detecting the intrusion in network intrusion detection. Problems with k-mean clustering are hard cluster to class assignment, class dominance, and null class problems. The network traffic datasets provided by the NSL-KDD Data set in intrusion detection system which demonstrates the feasibility and promise of unsupervised learning methods for network intrusion detection.

III. PROBLEM FORMULATION

The soft computing and data mining approach of network intrusion detection system suffered from detection rate and false alarm generation. The process of mining not conformed how many classifier are ensemble for the process of classification of data. The nature of intrusion data is mixed data type but 90% mining technique perform only numerical data for analysis. The conversion of data into one form to another form takes more time and suffered from grouping. The soft computing approach such as neural network and heuristic function compromised with selection of neural network model and appropriate feature selection algorithm for process of classification. Some problem in concern of intrusion detection apply both these technique found in review process [8, 9, 12].

1) The pre-processing of KDDCUP99 takes more time.
2) The rate of false alarm generation is high.
3) Some data mining classifiers create ambiguous situation for selection of base classifier.
4) Entropy based intrusion detection system suffered by high false rate.
5) The detection of dynamic feature evaluation as confusion matrix.

IV. KDDCUP99 DATA SET AND EMPIRICAL EVALUATION PARAMETER

To check performance of the soft computing and data mining algorithm for intrusion detection and classification, we can evaluate it practically using KDD’99 intrusion detection datasets [1]. In KDD99 dataset these four attack classes (DoS, U2R, R2L, and probe) are divided into 22 different attack classes that are tabulated in Table I. The 1999 KDD datasets are divided into two parts: the training dataset and the testing dataset. The testing dataset contains not only known attacks from the training data but also unknown attacks. Since 1999, KDD’99 has been the most wildly used data set for the evaluation of anomaly detection methods. This data set is prepared by Stolfo et al. [11] and is built based on the data captured in DARPA’98 IDS evaluation program [12]. DARPA’98 is about 4 gigabytes of compressed raw (binary) tcpdump data of 7 weeks of network traffic, which can be processed into about 5 million connection records, each with about 100 bytes. For each TCP/IP connection, 41 various quantitative (continuous data type) and qualitative (discrete data type) features were extracted among the 41 features, 34 features (numeric) and 7 features (symbolic).

<table>
<thead>
<tr>
<th>4 Main Attack Classes</th>
<th>22 Attack Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denial of Service (DoS)</td>
<td>buck, land, Neptune, pod, smurf, teardrop</td>
</tr>
<tr>
<td>Remote to User (R2L)</td>
<td>ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster</td>
</tr>
<tr>
<td>User to Root (U2R)</td>
<td>buffer_overflow, perl, loadmodule, rootkit</td>
</tr>
<tr>
<td>Probing(Information Gathering)</td>
<td>Ipsweep, nmap, portsweep, satan</td>
</tr>
</tbody>
</table>

Table 1: Different types of attacks in kdd99 dataset

To analysis the different results using some standard parameter such as Precision- Precision measures the proportion of predicted positives/negatives which are actually positive/negative. Recall -It is the proportion of actual positives/negatives which are predicted positive/negative. Accuracy-It is the proportion of the total number of predictions that were correct or it is the percentage of correctly classified instances. False-negative rate (FN) is the percentage that attacks are misclassified from total number of attack records. False-positive (FP) is the percentage that normal data records are classified as attacks from total number of normal data records. Below we have shown how to calculate these parameters by the suitable formulae and the graph for that particular data set.

\[
\text{Precision} = \frac{TP}{TP+FP}
\]

\[
\text{Recall} = \frac{TP}{TP+FN}
\]

\[
\text{Accuracy} = \frac{TP+TN}{FP+TN+FN+TP}
\]

\[
\text{FPR} = \frac{FP}{FP+TN}, \quad \text{FNR} = \frac{FN}{FP+FN}
\]

The assessment matrices are computed for testing dataset in the testing phase and the obtained result for all attacks and normal data are given in table 2, which is the overall classification performance of the proposed system on KDD cup 99 dataset. By analyzing the results, the overall performance of the proposed system is improved significantly and it achieves more accuracy for all types of attacks.

Fig. 1: Classification process of kddcup99 dataset

Figure 1 shows that classification process of kddcup99 dataset. In this process we used 10,000 data instant. In
A Review of Intrusion Detection Technique Based On Classification and Feature Optimization

(IJSRD/Vol. V/Issue 12/2014/0069)

10000 data instant, there were 7000 instant abnormal data and 3000 instant data of normal data instant.

<table>
<thead>
<tr>
<th>Data-Set</th>
<th>algorithm</th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kddcup9/9</td>
<td>Fuzzy-GA</td>
<td>92.14</td>
<td>87.24</td>
<td>84.43</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>97.14</td>
<td>96.11</td>
<td>94.10</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>89.90</td>
<td>84.32</td>
<td>83.23</td>
</tr>
<tr>
<td></td>
<td>ENSEMBLE</td>
<td>95.23</td>
<td>92.14</td>
<td>91.21</td>
</tr>
<tr>
<td></td>
<td>SOM</td>
<td>91.34</td>
<td>86.14</td>
<td>85.11</td>
</tr>
<tr>
<td></td>
<td>RBF</td>
<td>95.12</td>
<td>93.21</td>
<td>91.13</td>
</tr>
<tr>
<td></td>
<td>HYBRID</td>
<td>92.22</td>
<td>88.21</td>
<td>87.66</td>
</tr>
<tr>
<td></td>
<td>Reinforced</td>
<td>97.13</td>
<td>94.52</td>
<td>93.67</td>
</tr>
</tbody>
</table>

Table 2: Classification performance of data mining and soft computing algorithm

<table>
<thead>
<tr>
<th></th>
<th>Accuracy (%)</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fuzzy-GA</td>
<td>92.14</td>
<td>87.24</td>
<td>84.43</td>
</tr>
<tr>
<td>SVM</td>
<td>97.14</td>
<td>96.11</td>
<td>94.10</td>
</tr>
<tr>
<td>KNN</td>
<td>89.90</td>
<td>84.32</td>
<td>83.23</td>
</tr>
<tr>
<td>ENSEMBLE</td>
<td>95.23</td>
<td>92.14</td>
<td>91.21</td>
</tr>
<tr>
<td>SOM</td>
<td>91.34</td>
<td>86.14</td>
<td>85.11</td>
</tr>
<tr>
<td>RBF</td>
<td>95.12</td>
<td>93.21</td>
<td>91.13</td>
</tr>
<tr>
<td>HYBRID</td>
<td>92.22</td>
<td>88.21</td>
<td>87.66</td>
</tr>
<tr>
<td>Reinforced</td>
<td>97.13</td>
<td>94.52</td>
<td>93.67</td>
</tr>
</tbody>
</table>

Fig. 2: comparative result analysis of soft computing and data mining technique for intrusion detection and classification of attack

Figure 2 shows that comparative result analysis of soft computing and data mining technique for intrusion detection and classification of attack. Result gives information about deviation of method SVM is better in all mining technique and also Fuzzy-GA is better in soft computing technique.

V. CONCLUSION AND FUTURE SCOPE

In this paper we review current method of data mining and soft computing of intrusion detection, in particular, this paper reviews recent papers used for classification and feature optimization. In addition, we consider a large number of data mining techniques used in the intrusion detection domain for the review including clustering, classification, and ensemble technique. Regarding the comparative results of related work, found that the improvement in process of intrusion detection systems still needs to be researched. The applied process of soft computing approach such as fuzzy and genetic are more effective than other method but still suffered from problem of alarm generation.

REFERENCES

intrusion detection: Techniques, Systems and challenges” in Elsevier Ltd. All rights reserved 2008.