A Network Intrusion Detection System Using Simplified Swarm Optimization for Classification

C.Sudha[1] , A.Nooral shaba[2]

1 PG Scholar, Department of CSE, Al-Ameen Engineering College, Erode.
2 Assistant professor, Department of CSE, Al-Ameen Engineering College, Erode.
[1] sudha.chinnasamy@gmail.com

ABSTRACT
Intrusion detection is a mechanism of providing security to computer networks, which can effectively detect intrusion in the network. The network intrusion detection technique is important to prevent our systems and networks from malicious behaviors. Classification is the important task in data mining to classify the data. Simplified Swarm optimization (SSO) is used to classify the data with the selected features. Pruning techniques namely Induced One Rule is used for reducing the complexity of network intrusion detection system and improving its predictive accuracy by minimizing the over fitting due to noisy data.

General Terms
Data mining, Data Classification, Feature Selection.

Keywords
Network Intrusion Detection, Simplified Swarm Optimization, Induced One Rule.

I.INTRODUCTION

Data mining is the process of discovering actionable information from large sets of data. Data mining uses mathematical analysis to derive patterns and trends that exist in data. Data mining is the task of discovering interesting and hidden patterns from large amounts of data where the data can be stored in databases, data warehouses, OLAP (online analytical process) or other repository information. It is also defined as Knowledge Discovery in Databases (KDD).Data mining is used in various fields to discover the hidden patterns. Intrusion detection is one of the applications of Data mining to find various attacks in the dataset. As network attacks have increased in numbers over the past few years, Intrusion Detection Systems (IDSs) have become a necessary addition to the security infrastructure of most organizations. These systems are software or hardware schemes that automate the process of monitoring events that occur in a computer system or network and analyzing them for signs of security problems.

Network intrusion detection systems can be classified into two types which are host-based and network-based intrusion detection. Host-based detection captures and analyzes network data at the attacked system itself while the network-based detection captures and inspects online network data at the network gateway or server, before the attack reaches the end users. In addition, network intrusion detection systems can operate in two modes which are off-line detection and on-line detection. An off-line network intrusion detection system periodically analyzes or audits network information or log data to identify suspected activities or intrusions. In an on-line network intrusion detection system, the network traffic data has to be inspected as it arrives for detecting network attacks or malicious activities.NIDS are tools that collect information from a variety of system and network sources, detect, identify and respond to unauthorized or unusual activities on a target system. Basically, network intrusion detection fall into two basic categories: (1) anomaly based detection and (2) misuse-based detection. Anomaly based detection builds model of normal network behaviors (called profiles), which it uses to detect new patterns that significantly deviate from the profile. The main advantage of this anomaly detection is that it may detect novel intrusions. However, it suffers from larger false positive rate.On the other hand, misuse-based detection can detect the attacks based on well-known vulnerabilities and intrusion stored in a database. It uses the previous knowledge about known attacks and attempts to match current behavior against those attack patterns. The shortcoming of this approach is that it may not be able to alert the system administrator in case of a new attack. Ideal NIDS is one that has a high attack detection rate along with a low false positive rate. Thus, to come up with solution that can give a good accuracy while retaining low false positive rate has become a great challenge to overcome these two problems. Fig.1. shows the Overall structure of the proposed intrusion detection scheme.

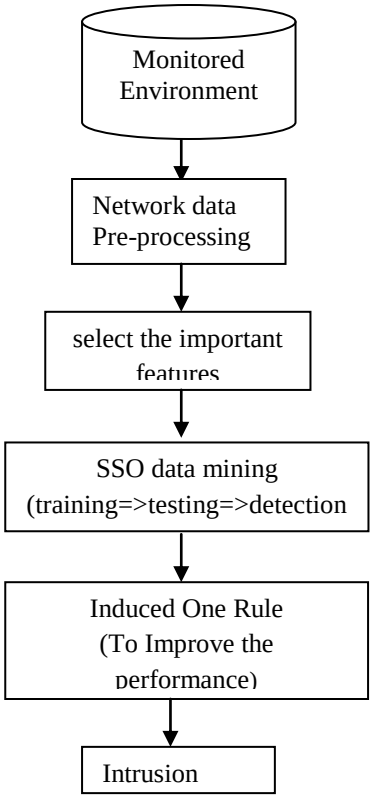

Fig.1.Overall structure of the proposed intrusion detection scheme.

II. LITERATURE SURVEY

Abraham et al [1] proposed the development of an Intrusion Detection Program (IDP) which could detect known attack patterns. An IDP does not eliminate the use of any preventive mechanism but it works as the last defensive mechanism in securing the system. Three variants of genetic programming techniques namely Linear Genetic Programming (LGP), Multi-Expression Programming (MEP) and Gene Expression Programming (GEP) were evaluated to design IDP.

Asaka et al [2] proposed a new method for detecting intrusion based on the number of system calls during a user's network activity on a host machine. This method attempts to separate intrusions from normal activities by using discriminant analysis, a kind of multivariate analysis.

Powers and He [3] proposed a hybrid system with the aim of combining the advantages of both AIS (Artificial Immune System) and K-SOM(Kohonen Self Organising Map) approaches. Specifically, anomalous network connections are initially detected using an artificial immune system. Connections that are flagged as anomalous are then categorised using a Kohonen Self Organising Map allowing higher-level information, in the form of cluster membership to be extracted.

Two of the major challenges in designing anomaly intrusion detection are to maximize detection accuracy and to minimize false alarm rate. When addressing this issue, Zanial et al [4] proposed an ensemble of one-class classifiers where each adopts different learning paradigms. The techniques deployed in this ensemble model are; Linear Genetic Programming (LGP), Adaptive Neural Fuzzy Inference System (ANFIS) and Random Forest (RF).

Sousa et al [5] proposed particle swarm based Data Mining Algorithm for classification task. Particle Swarm Optimisers are inherently distributed algorithms where the solution for a problem emerges from the interactions between many simple individual agents called particles. Initially the size of the particle is defined and then the fitness function is evaluated and the particle position is updated using velocity and finally best solution is found based on the particle current position.

The traditional intrusion detection systems look for unusual or suspicious activity, such as patterns of network traffic that are likely indicators of unauthorized activity. However, normal operation often produces traffic that matches likely "attack signature", resulting in false alarms. One main drawback is the inability of detecting new attacks which do not have known signatures. Srinoy [6] proposed Particle Swarm Optimization (PSO) which is used to implement a feature selection, and Support Vector Machine (SVMs) with the one-versus-rest method that serves as a fitness function of PSO for classification problems.

The reason for using Data Mining Classification Methods for Intrusion Detection Systems is due to the enormous volume of existing and newly appearing network data that require processing. Srinivasulu et al [7] proposed CART, Naïve Bayesian and Artificial Neural Network Model for data mining classification methods. These are useful for gathering different knowledge for Intrusion Detection. The idea of applying data mining classification techniques to intrusion detection systems to maximize the effectiveness in identifying attacks, thereby helping the users to construct more secure information systems.

Most of existing IDs use all features in the network packet to look for known intrusive patterns. Some of these features are irrelevant or redundant. Rough Set Classification (RSC), a modern learning algorithm, is used to rank features extracted for detecting intrusions and generate intrusion detection models. Wa'el Mahmud et al [8] proposed a new hybrid model RSC-PGA (Rough Set Classification Parallel Genetic Algorithm) is presented to address the problem of identifying important features in building an intrusion detection system,

increase the convergence speed and decrease the training time of RSC.

Yeh et al [9] proposed an efficient hybrid data mining approach to separate from a population of patients who have and who do not have breast cancer. The proposed data mining approach consists of two phases. In first phase, the statistical method will be used to pre-process the data which can eliminate the insignificant features. It can reduce the computational complexity and speed up the data mining process. In the second phase, proposed a new data mining methodology which based on the fundamental concept of the Standard Particle Swarm Optimization (PSO) namely Discrete PSO (DPSO).

A new rough set (RS) knowledge acquisition based on discrete particle swarm optimization (DPSO-RS) is proposed by Zhao et al [10] to solve feature selection strategy. Rough set is lack of the ability of anti-jamming, which is used in the information entropy, is considered as a suitable function in discrete particle swarm algorithm and make the classification rules more reliable in the case of noisy data.

Chung and Wahid [11] Proposed SSO for classification. In order to improve the performance of SSO classifier, a new weighted local search (WLS) strategy incorporated in SSO was proposed. The purpose of this new local search strategy is to discover the better solution from the neighborhood of the current solution produced by SSO.

Shehzad [12] proposed Minimum New classification and Induced One Rule pruning technique for rule induction, as well as an incremental post-pruning technique based on a misclassification tolerance.

III.DATASET DESCRIPTION

In this experiment, we are going to use a standard dataset randomly selected from 10% of KDDCup 99 which was extracted from 1998 DARPA intrusion detection evaluation program[13]. This database includes a wide variety of intrusion simulated in a military network environment that is commonly used benchmark to evaluate the intrusion detection techniques. The dataset has 41 features for each record plus one class label that contains 23 attacks which has been labeled as either normal or as an attack. All features in this dataset were labeled from A to AO. There are three main tasks involved in the experiment. In the first phase, the preprocessing task for data cleansing purposes was performed, followed by swarm intelligence classification to detect the intrusion behavior with the selected features and finally Induced One rule is used to improve the rule Quality.

IV.DATA PREPROCESSING

The original dataset consists of approximately four millions records of network attacks. Before we conduct the experiments a considerable amount of network intrusion data had to be taken into account. Apparently, the number of each attack is not consistent where there are several attacks that only consist of less than 3 records. Therefore we have selected those records that contain the attacks with majority number and ignore others which with minority number. We presume that the records that contain less attack do not have significant contribution to the classification accuracy and would not give any implication to any consequences for the detection results.Dataset with 4000 records have been randomly selected from 10% of KDDCup 99.Dataset is used to evaluate the performance of the proposed SSO classifier. In the pre-processing task,the attributes and their values were identified.Because the SSO algorithm can only deal with discrete variables, the attribute with nominal values including attribute C are needed to be mapped into discrete values. For example, in attribute C, nominal values that are tcp, udp and icmp, were mapped into discrete values 1, 2, and 3,respectively.

V.SELECTED FEATURES

Selecting the relevant features is an important task in data mining because all the features in the data set is not relevant and need to select the relevant features to reduce the complexity.Tabel1 shows the six selected features[11].These six features is selected using intelligent dynamic swarm based rough set (IDS-RS) algorithm. With these six features, the classification task has to be done.

Table 1
Detailed descriptions of six selected features
from 10% KDD Cup 99 dataset.

Label	Name	Description
c	Service	Network service on the destination, e.g.: http, telnet, etc.
E	src bytes	Numbers of data bytes from Source to destination.
F	dst bytes	Numbers of data bytes from destination to source.
AA	Rerror rate	% of connections that have “REJ” errors.
AG	dst host srv count	Sum of connections to the same destination port number.
AI	dst host diff srv-rate	% of connections from the same host with different service to the destination host during a specified time window.

VI.RULE MINING ENCODING

In the context of the classification task in data mining, knowledge discovery is represented in the form of IF-THEN prediction rules that have the advantage of being high-level symbolic knowledge representation which contribute to the ability of finding out a small number of rules with high fitness value [14]. Fig. 1 shows the form of rule mining encoding for the particles' position.

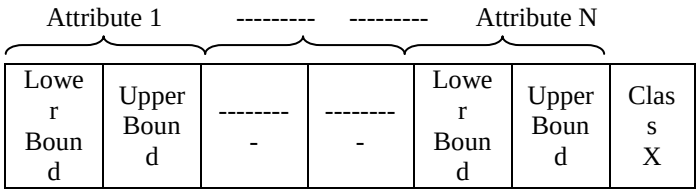

Fig. 1. Rule mining encoding.

The position of each particle contains N dimensions (attributes) except that the last cell is the predictive class which namely ClassX.The threshold for each attribute comes from the lowest data value to the highest data value of the given dataset. The former is called LowerBound and the latter is called UpperBound. The general form of the IF-THEN rules generated by PSO and SSO could be performed In all dimensions as following:

IF (LowerBound ≤ xij ≤ UpperBound) is true, THEN prediction is ClassX.

The LowerBound and UpperBound values were obtained by using Eq. (1) and Eq. (2), respectively.

$$\text{LowerBound}=x_{ij}- \text{rand}()*\text{range}(\text{Xmax} - \text{Xmin}) \quad (1)$$
$$\text{UpperBound}=\text{xij} +\text{rand}()*\text{range}(\text{Xmax} - \text{Xmin}) \quad (2)$$

where $X_i = (x_{i1}, x_{i2},. . ., x_{ij})$ denotes the i^{th} seed value of the j^{th} corresponding attribute in each dimension, rand() is a random number in the range of 0 and 1, and range((Xmax – Xmin)) is the range value of the data source in each attribute. The update strategy of Lower-Bound and UpperBound value in SSO is performed according to the comparison strategy as below:

If ($0 \le R < Cw$) then {LowerBound (x_{id}) = x_{id};
UpperBound (x_{id}) = x_{id}};
Else if ($Cw \le R < Cp$), then
{LowerBound (x_{id}) = LowerBound (p_{id});
UpperBound (x_{id}) = UpperBound (p_{id})};
Else if ($Cp \le R < Cg$), then
{LowerBound(x_{id}) = LowerBound(g_{id});
UpperBound (x_{id}) = UpperBound (g_{id})};
Else if ($Cg \le R \le 1$), then
{LowerBound =x_{id}–rand() *range (Xmax −Xmin); UpperBound = x_{id} + rand() * range(Xmax −Xmin)};

VII .INDUCED ONE RULE

In data mining, the main goal of rule pruning is to eliminate the irrelevant attributes that might have been unnecessarily included in the rule. In the rule discovery process, once we found the highest quality rule for the main class in the training set, the rule will go through a pruning process.

The principal idea of pruning process is to iteratively remove each attribute one at a time from the rule, and at the same time keep improving the quality of the rule. In the initial iteration, the process will start with the full rule. Then, each element contained by the corresponding attribute will be removed one by one, and the quality of the resulting rule is evaluated according to the rule-evaluation function.

Normal rule pruning is not as much efficient because once we remove the rule from the rule set it is necessary to check the test data every time to find the quality of the rule. The Induce one rule procedure is followed to remove the overlapping rules or overfitting rules in the dataset. It defines the two rules set namely parentruleset and childruleset. Based on this rule set the best rule is chosen to improve the accuracy.

VIII .PERFORMANCE METRICS

The following are the performance metrics used up for the evaluation of the model,
1. Accuracy - It refers to the total number of records that are correctly classified by the classifier.
2. True negative (TN): The percentage of valid records that are correctly classified.
3. True positive (TP): The percentage of attack records that are correctly classified.
4. False positive (FP): The percentage of records that wer incorrectly classified as attacks whereas in fact they are valid activities.
5. False negative (FN): The percentage of records that were incorrectly classified as valid activities whereas in fact they are attacks.
6. Precision – It refers to what percentage of positive prediction were Correct.
7. Recall – It refers to what percentage of positive cases were caught.

IX .PARAMETER SETTINGS

The Parameters going to use are Number of particle,m= 30, Maximum generation,maxGen= 20,Three pre determined constant Cw=0.1, Cp=0.4, Cg=0.9.

X .CONCLUSION

Simplified swarm optimization strategy for intrusion data classification with the selected feature is proposed. Classification is the important task in data mining to classify the data. Simplified Swarm Optimization is used to classify the data with the selected features. Induced One Rule is used to improve the performance of rule quality. Pruning process is a technique used to iteratively remove attribute one at a time from the rule. It is not an efficient method. To overcome those problems, an approach called Induced one Rule has been proposed. This approach is expected to improve the performance of the rule quality.

ACKNOWLEDGEMENT

I take great pleasure in expressing our profound gratitude to all those who have helped for the successful completion of the manuscript.I would like to express our sincere thanks to our respected **Principal Dr.A.M.J.MD.ZUBAIR RAHMAN**, **M.E., M.S., Ph.D** for providing an opportunity to carry out this work.I am very much indebted to the **Head of the Department of Computer Science and Engineering Professor Mr.M.MOHAMED MUSTHAFA., M.Tech., (Ph.D)** for his support and guidance.I am indeed grateful to my guide **Assistant Professor Ms. A .NOORAL SABA, M.E.,** Department of Computer Science and Engineering who helped and encouraged me throughout this manuscript

REFERENCES

Abraham A., Grosan C. and Vide C.M. (2006) *'Evolutionarydesign of intrusion detection programs', International Journal of Network Security*, pp. 328–339.[1]

Asaka M. Onabura T. , Inoue T. , Okazawa S., and Goto S., (2001) ' *A new intrusion detection method based on discriminant analysis'* , IEICE Transactions on Information and System, Vol E84-D,No.5,pp. 570–577.[2]

Powers S.T., and He J., (2008) 'A *hybrid artificial immune System and self Organising Map for network intrusion detection*', Information Sciences 178, pp.3024–3042.[3]

Zainal A., Maarof M.A., and Shamsuddin S.M. , (2009) '*Ensemble classifiers for network intrusion detection system',* Journal of Information Assurance and Security 4, pp. 217–225. [4]

Sousa T., Silva A. And Neves A., (2004) ' *Particle swarm based Data mining algorithms for classification tasks* ' , Parallel Computing 6 , pp. 767–783. [5]

Srinoy S., (2007) *' Intrusion detection model based on particle swarm Optimization and support vector machine* ', Proceedings of the IEEE Symposium on Computational Intelligence in Security and Defence Applications[6] .

Srinivasulu P.,Nagaraju D., Kumar P.R. And Rao K.N.,(2009) ' *Classifying the network Intrusion attacks using data mining classification methods and their performance comparison'*, International Journal of Computer Science and Network Security 9 Vol.9,No.6,pp. 11–18. [7]

Wa'el M.M., Agiza H.N. and Radwan E (2009), '*Intrusion Detection using rough sets based parallel genetic algorithm hybrid model* ', Proceedings of the World Congress on Engineering and Computer Science , Vol 2. [8]

Yeh W.C., Chang W.W., Chung Y.Y. (2009) *'A new hybrid Approach for mining Breast cancer pattern using discrete particle Swarm optimization and Statistical method* ', Expert System with Applications pp. 8204–8211. [9]

Zhao Q., Zhao J., Meng G. and Liu L.(2010), *' A New Method Of Data Mining Based on Rough Sets and Discrete Particle Swarm Optimization'*, Proceedings of The IEEEconference on Data Mining. [10]

Chung Y.Y and Wahid N., (2012), *'A hybrid intrusion Detection system Using simplified swarm optimization* ', Applied Soft Computing 12, pp.3014-3022. [11]

Shehzad K. (2011), *'Simple Hybrid and Incremental Post-pruning Techniques for Rule Induction'* Proceedings of IEEE Transactions. [12]

KDD Cup 99 Dataset, 1999, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. [13]

Ang J.H., Tan K.C., Mamun A.A., An evolutionary *memetic Algorithm for rule extraction, Expert System with Applications* 37 (2) (2009) 1302–1315. [14]

