Network Intrusion Detection using Hybrid Simplified Swarm Optimization Technique

S. Revathi[1], A. Malathi[2]
Ph.D. Research Scholar[1], Assistant Professor[2]
PG &Research department of Computer Science
Government Arts College (Autonomous)
Coimbatore-18

***Abstract*----** Network security risks grow tremendously in recent past, the attacks on computer networks have enhanced hugely and need economical network intrusion detection mechanisms. Data processing and machine-learning techniques are used for network intrusion detection throughout the past few years and have gained abundant quality. In this paper, we propose an intrusion detection mechanism based on simplified particle swarm optimization (SSO) is used to investigate the performance of various dimension reduction techniques along with a set of different classifiers including the proposed approach. SSO is used to find more appropriate set of attributes for classifying network intrusions, and also used as a classifier. In preprocessing step, we reduce the dimensions of the dataset by using various dimension reduction techniques, and then this reduced dataset is offered to the proposed hybrid SSO approach that further optimizes the dimensions of the data and finds an optimal set of features. SSO is an optimization method that has a strong global search capability and is used for dimension optimization. The analysis performed on standard KDD cup99 dataset which contain various kind of intrusion. The experimental results shows the worth of the proposed approach by using different performance metrics.

***Keywords*--** Swarm intelligence, Simplified Swarm Optimization, optimization, Data mining, Intrusion Detection.

I. INTRODUCTION

The security of a computer system or network is compromised when an intrusion takes place. As a result, Intrusion Detection System (IDS) have become an essential component of security to detect threats, and to track the intruders. As IDS must have a high attack Detection Rate (DR), with a low False Alarm Rate (FAR) at the same time, construction of IDS is a challenging task [1]. Therefore, a new generation of computational techniques and tools is

required to detect intrusion from the rapidly growing volume of data. Hence, data mining becomes the reliable solution for elucidating the patterns that underlie it. Data mining is the application of specific algorithms that has been widely used for extracting patterns or models from data [2]. Two main aspects in data mining are data classification and feature selection. However, existing traditional data classification and feature selection techniques used in data management are no longer enough to detect intrusive data. This deficiency has prompted the need for a new intelligent technique based on stochastic population-based optimization that could discover useful information from data [3].

The above causes are the main inspiration for interesting and powerful algorithms used in the search for optimal solutions of highly non-trivial problems (CHENG et al., 2009; KARABOGA, 2009; KARABOGA; BASTURK, 2007; TOKSARI, 2006; WANG et al., 2007) for finding the global minimum of nonlinear functions, This field of study is known as “swarm intelligence” and has attracted an increasingly number of researchers towards Particle Swarm Optimization (PSO) algorithm and Ant Colony Optimization (ACO) Algorithm (DORIGO et al., 1996).

To overcome certain difficulties in PSO [4] this paper proposed a novel Simplified Swarm Optimization (SSO) algorithm as a rule-based classifier and for feature selection. SSO is a simplified Particle Swarm Optimization (PSO) that has a self-organizing ability to emerge in highly distributed control problem space, and is flexible, robust and cost effective to solve complex computing environments [21]. The proposed SSO classifier has been implemented to classify Intrusion data.

The rest of the paper is structured as follows: section 2 explains important of feature selection techniques in KDD cup 99 dataset. Section 3 present an overview of proposed framework. Section 4 explain about data mining in intrusion detection. Section 5 explains in detail about proposed hybrid SSO algorithm and its accuracy is compared with other data mining classifier. Section 6 concludes some result based on proposed work.

II. FEATURE SELECTION OVERVIEW

Feature selection plays an important role in data pre-processing technique for data mining [2]. It is a process of finding a subset of features from the

original set and forming patterns in a given dataset to obtain the optimal solution. The raw network data has huge network traffic data size which leads to irrelevant and huge dimensionality problem. The KDD'99 dataset has been the most wildly used for the evaluation of anomaly detection methods prepared by Stolfo et al, based on the data captured in DARPA'98 IDS evaluation program. To reduce the number of features, removes irrelevant, redundant, or noisy data and to improve mining performance such as classification accuracy, feature selection techniques used. In the context of classification, feature selection can be structured into three fractions: filter method, wrapper method and embedded method [5]. Filter methods rely on the intrinsic properties of the training data to select some features without involving any learning algorithm. Various filter methods are used to univariate the intrinsic features in intrusion dataset.

Feature selection has been used as a measure to decide the importance and necessity of features. Hence, a broad range of sub-optimal feature selection methods have been developed over the last decade to overcome its limitation. Previous research on feature selection which employed a heuristic approach, such as hill-climbing methods, has proven more efficient when dealing with little noise and a small number of features. On the other hand, due to the reason that heuristic methods fail to find an optimal reduct, many researchers have shifted to metaheuristic approaches such as Genetic Algorithm (GA), Simulated Annealing (SA), Tabu Search, Ant Colony Optimization (ACO) [13] and Particle Swarm Optimization (PSO). Moreover, several hybrid methods that used SI algorithms and rough sets for feature selection have been reported. Methods which combine ACO and rough set to find reducts with promising results were proposed by Ke et al. and Chen et al [20]. Bello et al. [6] proposed a two-stage heuristic search performed by PSO and ACO in order to find the optimal feature subsets. In [7], the authors used chaotic BPSO with logistic map to determine the inertia weight to solve the feature selection problem. As a result, their approach has successfully reduced the computation time as well as improved the quality of reducts and classification accuracies. Moreover, several hybrid methods that used SI reducts and classification accuracies.

III. OVERVIEW OF PROPOSED FRAMEWORK

Simplified swarm optimization technique detect intrusion and reduce false positive rate. The proposed framework is shown in figure1. The information is obtained from KDD cup99 dataset [8], the records in the database contains 41 features in which the data may be redundant, noisy or irrelevant in nature. The proposed pre-processing approach filters data effectively and the result compares with existing approach.

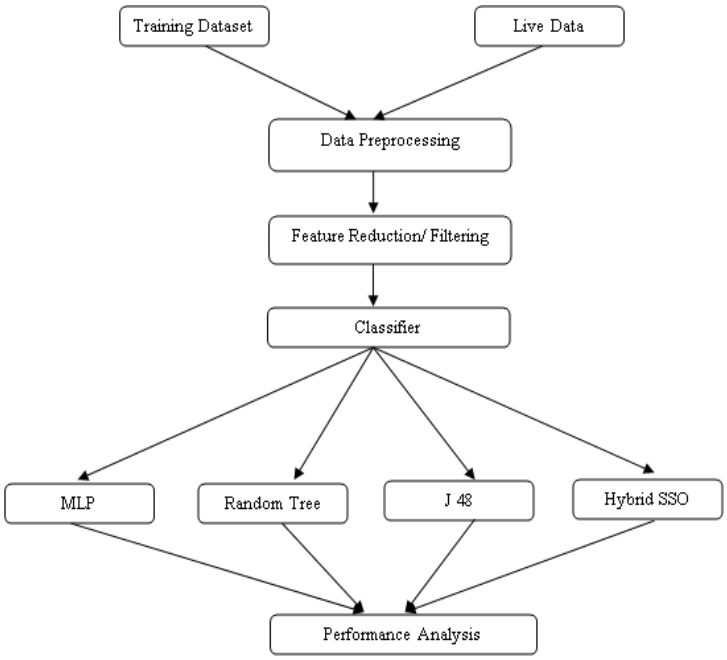

Fig. 1 Framework for proposed work

A. KDD cup 99 Dataset Description

The KDD99 dataset was used in Knowledge Discovery and Data Mining Tools Competition for building network intrusion detector, it distinguish data between intrusions and normal network connections [9]. In 1998, DARPA intrusion detection evaluation program, a simulated environment was set up to acquire raw TCP/IP dumps data for a local-area network (LAN) by the MIT Lincoln Lab to compare the performance of various intrusion detection methods. It was operated like a real environment, but being annoying with multiple intrusion attacks and received much attention in the research community of adaptive intrusion detection. TheKDD99 dataset contest uses a version of DARPA98 dataset .In KDD99 dataset, each example represents attribute values of a class in the network data flow, and each class is labelled either normal or attack. The classes in KDD99 dataset categorized into five types (normal probe, DOS, U2R, and R2L) [10].

- Denial of Service Attack (DoS): is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies legitimate users access to a machine.
- User to Root Attack (U2R): is a class of exploit in which the attacker starts out with access to a normal user account on

the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.

- Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a 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.
- Probing Attack: is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls.

A complete KDD'99 dataset contains five millions connection records where 4,898,431 are labelled connections that divided into 22 different attack classes that are tabulated in Table 1.

Table I
Detail of Attacks of Labelled Records

There are 41 input attributes in KDD99 dataset for each network connection that have either discrete or continuous values and divided into three groups namely basic features, content feature and statistical features.

Category of Attack	**Attack Name**
Normal	Normal
DoS	Neptune,Smurf,Pod,Teardrop,Land,back
Probe	Portsweep, IPsweep, Nmap, satan
U2R	Bufferoverflow,LoadModule,Perl,Rootkit
R2L	Guesspassword,Ftpwrite,Imap,Phf, Multihop,Warezmaster,Warezclient

IV. DATA MINING TECHNOLOGY IN INTRUSION DETECTION

The goal of predictive data mining is to produce a model expressed as an executable code which can be used to perform data mining tasks such as classification, prediction or other similar tasks. Classification plays a vital role in intrusion to detect accuracy which increases detection rate and reduce false alarm rate. The two main type of classification are

- Unsupervised classification and
- Supervised classification

A. *Unsupervised classification*:

A method of unsupervised classification uses only the matrix of dissimilarity. No information of the class of an object is provided to the method. The objective is then to build a set of groups where each group contains a set of sufficiently similar objects. In data mining, unsupervised classification is also known as clustering [12]. Clustering divides the data into groups of similar objects and each group consists of objects that are similar between themselves and dissimilar to objects of other groups. There are many clustering algorithms that exist such as hierarchical clustering, K-Means and Fuzzy C-Means.

B. *Supervised classification*:

On the other hand, in a method of supervised classification, the objects are labelled and each value of the label represents a class. The objective of this method is to build hyper planes separating the objects according to their class. Supervised classification is performed on a set of training examples. Each training example (xi, yi) is composed of a feature vector, xi and a corresponding class label, yi [11]. The feature vector contains measurable characteristics of the object under consideration. Intrusion dataset based on class label in training dataset. Therefore supervised classification algorithm such as Support Vector Machines, kNearest Neighbor, Decision Tree (i.e.: PART and J48) and Naive Bayes classifiers are used to classify normal and abnormal data and also to detect accuracy and to reduce false alarm rate in intrusion detection.

V. PROPOSED HYBRID SIMPLIFIED SWARM OPTIMIZATION ALGORITHM FOR INTRUSION DETECTION

The paper proposed a new swarm intelligence approach based on simplified swarm optimization used to preprocess and to classify intrusion. It filters data and reduce irrelevant and dimensionality problem for both discrete and continuous variables in dataset [14]. This approach is significantly different from other research work which had combine only data mining and PSO. The proposed method produce high efficiency and produce near optimal solution for pre-processing phase.

The traditional pre-processing algorithms are not adaptive to the situations when kdd99 dataset is large. This may result in the false recommendations. In this paper, a new proposed hybrid swarm intelligence technique is used. SSO is a simplified version of PSO and can be used to find the global minimum of nonlinear functions. This approach is used to solve classification

problem and reduce dimensionality of dataset [15]. The introduction of SSO algorithm is as follows

Initially, the number of swarm population size, the number of maximum generation, and three parameters are determined. In every generation, the particle's position value in each dimension will be kept or be updated by its pbest value or by the gbest value or be replaced by new random value according to the procedure depicted in equation (1).

$$x_{id}^{t} = \begin{cases} x_{id}^{t-1} & if \quad rand() \in [0.\, c_w) \\ p_{id}^{t-1} & if \quad rand() \in [c_w,\, c_p) \\ g_{id}^{t-1} & if \quad rand() \in [c_p,\, c_g) \\ x & if \quad rand() \in [c_g,\, 1) \end{cases} \quad (1)$$

Where i = 1; 2; m, where m is the swarm population. Xi = (xi1; xi2; xiD), where xiD is the position value of the i-the particle with respect to the D-the dimension of the feature space. Cw, Cp and Cg are three predetermined positive constants with Cw < Cp < Cg. Pi = (pi1; pi2; : : : ; piD) denotes the best solution achieved so far by itself (pbest), and the best solution achieved so far by the whole swarm (gbest) is represented by Gi = (gi1; gi2; : : : ; giD). The x represents the new value for the particle in every dimension which are randomly generated from random function rand (), where the random number is between 0 and 1.

The update strategy for particles' position value in SSO is presented below.

Step 1: Initialize the swarm size (m), the maximum generation (maxGen), the maximum fitness value (maxFit), Cw, Cp and Cg.

Step 2: In every iteration, a random number R that is in the range of 0 and 1 will be randomly generated for each dimension.

Step 3: Perform the comparison strategy where:

if ($0 \le R < Cw$), then {xid = xid};
Else if ($Cw \le R < Cp$), then {xid = pid};
Else if ($Cp \le R < Cg$), then {xid = gid};
Else if ($Cg \le R \le 1$), then {xid = new (xid)};

Step 4: This process will be repeated until the termination condition is satisfied.

VI. EXPERIMENTAL RESULT

The analyses done on KDD cup 99 training dataset and the performance of the proposed method is compared with other existing algorithm. In this paper only 10% of KDD cup dataset showed in table II is employed for the purpose of training. It consists of 41 feature attributes out of which 3 are symbolic and 38 are numeric. Thus each connection is given by 41 features set. The comparison based on some popular machine learning techniques such as J48 [16], Multi- layer Perceptron [17] and Random tree [18] from Weka [19] collection to learn the overall behavior of the KDD'99 data set [8].

Table II
Attacks on 10% KDD cup dataset

Attack Name	Number of Attacks
DOS	391458
Probe	4107
U2R	52
R2L	1126
Normal	97277
Total	494020

The performance of pre-processed data is shown in figure 2 which filters noisy and incomplete data form raw data (KDD cup99 dataset) and it shows that the proposed system reduce feature selection attribute, which reduce false positive rate and improves efficiency for intrusion detection system. The proposed system can easily filters large scale dataset and removes unwanted parameters which decreases overlapping behavior of normal and intrusive data. The parameters based on SSO obtains best achievement from all data mining techniques both as preprocessor and as a classifier.

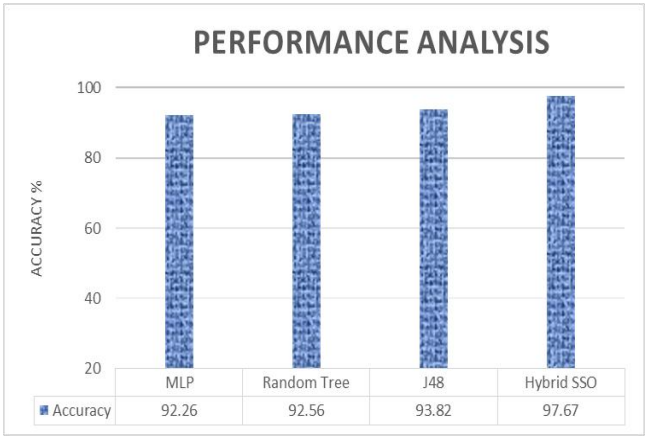

Fig. 2 Performance of Proposed Algorithm

VI. CONCLUSION

This paper analyzed the performance of the proposed algorithm for intrusion detection using KDD cup99 dataset. The attributes in the dataset are reduced used various data mining techniques. The reduced attributes are than implemented in data mining classifiers, the result is compared with our proposed new techniques. Simplified swarm

optimization is a new optimization techniques which used as preprocessor and as classifier to detect intrusion. It easily handles dimensionality problems and extract most relevant features which improves accuracy and reduce false alarm rate. Therefore swarm intelligence techniques performs quite better than other data mining classification methods. The experimental result shows that Hybrid SSO algorithm is faster in convergence and more efficient in solution. By filtering normal data we can easily detect intrusion by using various data mining and other computational intelligence technique which is a future work to be proposed to improve detection efficiency.

REFERENCE

1. Deris tiawan, Abdul Hanan Abdullah, Mohd. Yazid dris, "Characterizing Network Intrusion Prevention System", *International Journal of Computer Applications* (0975 – 8887), Volume 14– No.1, (January 2011).
2. J. Han and M. Kamber, *Data Mining: Concepts and Techniques. San Fransisco: Morgan Kaufmann*, 2001.
3. C. Grosan, A. Abraham, and M. Chris, "Swarm Intelligence in Data Mining," *Studies in Computational Intelligence*, vol. 34, pp. 1-20, Springer-Verlag: Berlin Heidelberg, 2006.
4. R. C. Eberhart and Y. Shi, "Particle swarm optimization: developments, applications and resources," *in Proceedings of the 2001 Congress on Evolutionary Computation*, Seoul, South Korea, May 27-30, vol. 1, pp. 81-86, 2001.
5. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157 - 1182, Mar.2003.
6. R. Bello, Y. Gomez, A. Nowe, and M. M. García, "Two step particle swarm optimization to solve the feature selection problem," *in Proceedings of The 7th International Conference on Intelligent Systems Design and Applications*, Rio de Janeiro, Brazil, Oct. 22-24, pp. 691-696, 2007.
7. C. S. Yang, L. Y. Chuang, J. C. Li, and C. H. Yang, "Chaotic maps in binary particle swarm optimization for feature selection," *in Proceedings of the 2008 IEEE Conference on Soft Computing on Industrial Applications*, Muroran, Japan, June 25-27, pp. 107-112, 2008.
8. KDDCUP 99 dataset, available at: http://kdd.ics.uci.edu/dataset/kddcup99/kddcup99.html.
9. MIT Lincoln Labs, 1998 DARPA Intrusion Detection Evaluation. Available on: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/ index.html, February 2008.
10. G. Sunil Kumar, C.V.K Sirisha, Kanaka Durga.R, A.Devi, "Robust Pre-processing and Random Forests Technique for Network Probe Anomaly Detection", *International Journal of Soft Computing and Engineering (IJSCE)*, ISSN: 2231-2307, Volume-1, Issue-6, (January 2012).
11. Jiawei Han and Micheline Kamber "Data mining concepts and techniques" Morgan Kaufmann publishers an imprint of *Elsevier* .ISBN 978-1-55860-901-3. Indian reprint ISBN 978-81-312-0535-8.
12. I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Toolsand Techniques*, Second ed. San Fransisco: Morgan Kaufmann, 2005.
13. L. Ke, Z. Fenga, and Z. Rena, "An efficient ant colony optimization approach to attribute reduction in rough set theory " *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1351-1357, July 2008.
14. W.C. Yeh, W.W. Chang, Y.Y. Chung, "A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method", *Expert System with Applications* 36 (May (4)) (2009) 8204–8211.
15. Yuk Ying Chung, Noorhaniza Wahid: "A hybrid network intrusion detection system using simplified swarm optimization (SSO)". *Appl. Soft Computing*. 12(9): 3014-3022 (2012).
16. J. Quinlan, C4.5: "Programs for Machine Learning". Morgan Kaufmann, 1993.
17. D. Ruck, S. Rogers, M. Kabrisky, M. Oxley, and B. Suter, "The multilayer perceptron as an approximation to a Bayes optimaldiscriminant function," *IEEE Transactions on Neural Networks*, vol. 1, no. 4 pp. 296–298, 1990.
18. D. Aldous, "The continuum random tree. I," The Annals of Probability, pp. 1–28, 1991.
19. "Waikato environment for knowledge analysis (weka) version 3.5.7." Available on: http://www.cs.waikato.ac.nz/ml/weka/, June, 2008.
20. Y. Chen, D. Miaoa, and R. Wanga, "A rough set approach to feature selection based on ant colony optimization," *Pattern Recognition Letters*, vol. 31, no. 3, pp. 226-233, Feb. 2010.
21. Noorhaniza Wahid "A Novel Approach To Data Mining Using Simplified Swarm Optimization ", *Thesis submitted to university of Sydney*, January 2011