FEATURE EXTRACTION USING SIM-SWADOREST OPTIMIZATION ALGORITHM FOR INTRUSION DETECTION

[1]S. REVATHI, [2]A. MALATHI

[1] Ph.D. Research Scholar, [2] Assistant Professor
PG and Research, Department of Computer Science Government Arts College Coimbatore-18
Email: revathisujendran86@gmail.com,malathi.arunachalam@yahoo.com

Abstract- Intrusion detection is a key element in system security that operates in real time network data. The Main drawback in detecting intrusion is its dataset size, which leads to a huge dimensional dataset problem. Therefore, feature extraction is necessary to reduce attribute size and improve detection rate to produce high efficiency. This paper proposed a new mechanism based on simplified swarm optimization (SSO) and Random Forest algorithm (RF) called Sim-Swadorest Optimization algorithm. SSO is used to find a more appropriate set of attributes for classifying network intrusions, and RF is used as a classifier to detect intrusion. The proposed method is investigated and assessed on live network data collected. It's clear and consistently superior to other existing methods and improves the performance accuracy of filtering phase than PSO-RF and other classifiers. The test result shows that the proposed method reduces attribute size more effectively and achieves a near optimum solution.

Index Terms—Swarm intelligence, Simplified Swarm Optimization, Particle Swarm Optimization, Random Forest, Intrusion detection.

I. INTRODUCTION

Intrusion Detection is an important part of Security Management system for computers and networks that tries to detect break-in attempts. There is no disputing the fact that the number of hacking and intrusion incidents is increasing year to year as technology rolls out, unfortunately in today's interconnected commerce world, there is no hiding place. It is a burglar alarm for computer network, to detect unauthorized access attempts. The two main intrusion-detection techniques are misuse detection and Anomaly detection. The dataset used for intrusion detection is a raw data which highly susceptible to noise, missing values and inconsistency. To improve a quality of raw data, attribute reduction is required, which increase data efficiency. Many technologies have been proposed for intrusion detection such as data mining, machine learning, and ensemble classifier, which have been widely used to solve many IDS classification problems. Recently, many biologically inspired algorithms such as Genetic Algorithm (GA), Ant Colony (AC), Artificial Bee Colony and Swarm Intelligence (SI) have been used. Among them, Particle Swarm Optimization (PSO) is a popular heuristic technique for optimization, but it suffers from premature convergence of high dimension multimodal problem, which fails to achieve better fitness optimization value.

As a result, this paper proposed a novel, Sim-Swadorest optimization algorithm for pre-processing the raw network data. The experimental analysis is evaluated using real time data collected from Linkware Technologies. The main objective of this paper is to screen redundant data and to reduce irrelevant feature. The rest of the paper is structured as follows: section 2 explains some related work based on feature selection for intrusion-detection dataset. Section 3 describes about background information of various evolutionary algorithms. Section 4 explains in detail about evolutionary analysis based on Sim-Swadorest optimization algorithm. Section 5 shows the experimental result analysis, and Section 6 draws some conclusion and future work.

II. RELATED WORK

Intrusion-Detection Systems gross raw network data or audit records as input, which leads to a huge network traffic data size and invisibility of intrusive patterns, which are normally hidden among the irrelevant and redundant features to identify it as normal or attack. Researchers have identified that pre-processing is needed for better results and used various approaches. A new collaborating filtering technique for pre-processing the probe type of attacks is proposed by Fernando used n-gram theory to identify redundant subsequence and proposed Hidden Markov Model for service selection to reduce audit data significantly.

Now-a-days biological inspired approaches have been extensively originated in network intrusion pattern detection For example, Wei-Chang yeah et.al proposed new method of combining SSO with weighted exchange local search method for intrusion detection. Banerjee et al suggested to use ACO to keep track of intruder and in addition PSO algorithms had been successfully employed to learn classification rules from Chen et al. In order to reduce the false alarm rate in IDS, Chang and Ping employed PSO algorithm with a new fitness function that proved to be suitable for IDS by extracting high-

quality rules. Arif Jamal Malik proposed a new network intrusion detection using hybrid binary PSO and a random forest algorithm, which reduce the dimension of the data and finds optimal solution, similarly Wang et al (2009) used two different senses of PSO: the Standard Particle Swarm optimization (SPSO) and Binary Particle Swarm Optimization (BPSO) for seeking optimal SVM parameters and extracting a feature subset respectively. Safaa Zaman et al. (2013) proposed feature selection for intrusion detection based on different evolutionary algorithms.

III. BACKGROUND

This section explains in detail about various evolutionary algorithms such as particle swarm optimization, Simplified swarm optimization, genetic algorithm and also a Random Forest algorithm.

A. Particle Swarm Optimization

Particle swarm optimization is a heuristic optimization technique based on the behavior of birds flocking and fish schooling. In PSO the position of the particles x_n and the velocity of the particles v_n are modelled based on D dimension space. In every iteration the particles position and velocity are updated based on its pbest (personal best) and gbest (global best) values according to the following equation (1).

$$V_n^{(t+1)} = wv_n^d + c_1.rand_1(p_n^{pbest}(t) - x_n^i(t)) + c_2.rand_2(g_n^{gbest}(t) - x_n^i(t)) \quad (1)$$

$$X_n^{(t+1)} = x_n(t) + v_n^{(t+1)}$$

Where w be inertia weight to less than 1. c_1 and c_2 be cognitive and social constant that shows the particle is travel in good position. $rand_1$ and $rand_2$ be a random number generator in the range [0,1]. The personal best and global best position of the swarm be denoted as $p_n^{pbest}(t)$ and $g_n^{gbest}(t)$ in dimension D. The updated position of the particle be $X_n^{(t+1)}$ based on its velocity be $V_n^{(t+1)}$.

B. Simplified Swarm Optimization

The SSO algorithm consists of population size of swarm particles be p, the number of iterations performed be (MG), the maximum fitness be (MF) and three predetermined constant be (c_w,c_p,c_g). The particle's position is updated based on its personal best (pbest) and Global best (gbest) position for each iteration. According to the new random value for each generation the delineate equation for the procedure is expressed in equation (2) .

$$x_{nd}^t = \begin{cases} x_{nd}^{t-1} & if\ rand() \in [0.\ c_w) \\ p_{nd}^{t-1} & if\ rand() \in [c_w,\ c_p) \\ g_{nd}^{t-1} & if\ rand() \in [c_p,\ c_g) \\ x & if\ rand() \in [c_g,\ 1) \end{cases} \quad (2)$$

Where n = 1, 2,……p, where p is the number of population size. $X_n = (x_{n1}; x_{n2};\ldots. X_{nd})$, where x_{nd} is the position value of the n-the particle with respect to d-the dimension of the feature space. C_w, C_p and C_g are three pre-determined, positive constants with $C_w < C_p < C_g$. $P_n = (p; p_{n2}; : : : ; p_{nd})$ denotes the personal best solution achieved so far by itself (pbest), and the best solution achieved so far by the whole swarm (gbest) is represented by $G_n = (g_{n1}; g_{n2}; : : : ; g_{nd})$. The x represents the new value for the particle in every dimension which are randomly generated from rand (), where the random number is between.

C. Genetic Algorithm

Genetic Algorithm is an adaptive heuristic search algorithm based on natural selection. It is used to solve optimization problem based on random search represented by chromosomes called a population. Then it evaluates the fitness function based on roulette wheel selection for each chromosome and create a new population. Then two offspring is generated based on one pint crossover probability of two parents by using genetic information. Finally the mutation probability is used to change individual genes at each position in the chromosome. The process repeats until the best solution is returned in current population.

D. Random Forest

The random forests algorithm is an ensemble unpruned classification or regression algorithm that generates many decision trees. It is a combination of tree predictors such that each tree depends on values of a random vector sampled independently and with same distribution for all trees in the forests. When new records come as input, RF generate a classification and choose the class that is classified by most trees. It builds many trees which minimize the classification errors based on a bootstrap sample of training data from the original dataset using a tree classification algorithm. After the forest is created, a replacement object must be classified within its tree. Every tree provides a vote concerning the category of the object. The algorithm chooses the class with the maximum votes. By injecting randomness at each node of the grown tree, it has improved accuracy. A new sample is pushed down the tree based on the training sets in the terminal node it ends up. This procedure is iterated over all trees in the ensemble as a random forest prediction. It is applied in various research fields for modeling. Now-a-days such prediction is used to detect intrusion. Zhang and Zulkernine implemented random forest algorithm to detect hybrid Intrusion detection system. The main advantage of using RF, is it easily handles large and

imbalanced dataset more effectively and does not overfit.

The proposed system uses Real time network data collected from live network server (NSC) based on transmission of data packets for every 2 Sec and stored them on disk. These unclassified data are further categorized into various attacks that come under 4 major attack types. Such classified training and testing dataset are collected from Linkware Technologies, which contain 41 attributes. The real-time data consists of more than 1000 attacks that fall under DOS, probe, U2R and R2l types. The proposed system handles these live dataset more effectively. The figure 1 depicts the block diagram of the proposed system for preprocessing.

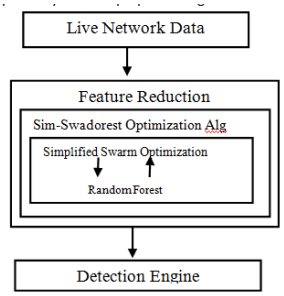


Fig 1. Proposed Framework

IV. EVOLUTIONARY ANALYSIS OF SIM-SWADOREST OPTIMIZATION METHOD

There are several pre-processing algorithms that are not adaptive when dataset size is large which may result in the false endorsement. In this paper, a new proposed hybrid technique is used; Simplified Swarm Optimization is mainly used to find a global optimum solution for non -linear function. This approach is used to solve the classification problem and reduce the dimensionality of the dataset. The random forest classifier is used to split the dataset and to identify the most important attribute to detect intrusion. On preprocessing the proposed method choose the most suitable attribute for identifying whether the record is normal or attack. This section explains in detail about an overview of the implementation is described as follows.

A. Representation of Individuals

The set of all 41 features is represented by a vector of 41 bits, each bit value represent as either 1 (selected attribute) or 0 (non-selected attribute) from the current set of all attributes.

B. Initialization of the Particles

The value of each element in the vector is represented by a random binary variable between 0and 1. For experimentation, we have set 50 as the maximum number of particles and 100 as maximum iteration. Each particle consists of 41 attributes that are initialized with random binary values based on the presence or absence of the features.

1	2	...	...	...	...	..41
1	0	0		1		1

C. Fitness Evaluation

The fitness value of each attribute is determined using accuracy results from the random forest classification algorithm. The feature with bit value 1 is selected for calculating accuracy. If suppose the bit values be 111001 means that feature1, feature2, feature3 and feature6 are selected. Based on 1's and 0's the global best particle will know about the best attribute selected by the random forest classifier. In the next generation the global best particle may be changed based on its position of 1's which may reduce the attribute as compared to the input set.

D. Evolution Method

The evolutionary algorithm for proposed Sim-Swadorest optimization algorithm are as follows.

Step1: Let the swarm population size be p, maximum generation be (MG), maximum fitness function be (MF) and the three predetermined constant be c_w,c_p,c_g be initialized.

Step2: Generate random number R=0 to 1 for d dimension data.

Step3: perform the comparison strategy where

if ($0 \le R < C_w$), then $\{x_{nd} = x_{nd}\}$;
Else if ($C_w \le R < C_p$), then $\{x_{nd} = p_{nd}\}$;
Else if ($C_p \le R < C_g$), then $\{x_{nd} = g_{nd}\}$;
Else if ($C_g \le R \le 1$), then $\{x_{nd} = \text{new } (x_{nd})\}$;

Step4: RF (n, d) =1/exp (x (n, d))
Update (x_{nd});

Step5: Update pbest

Step6: Update gbest

Step7: Process will repeat until terminated condition is satisfied or maximum IDR is reached.

V. EXPERIMENTAL RESULT ANALYSIS

For evaluating the performance of the proposed algorithm, the real time dataset has been used. We compare the different evolutionary algorithms (PSO, GA, SSO) with random forest classifier in this section. We initially describe the content of real time server data and the experimental setting; followed by some experimental result and discussions.

A. Dataset Description

To make this system more efficient, real time server data collected from Linkware Technologies and a dedicated server of Cluster Research Desk are used to train and test the system. To measure the efficiency of intrusion detection system four categories of attacks such as DOS (Denial of Service), Probe, U2R (User to Root) and R2L (Remote to User). The actual

dataset from a dedicated server handles multiple mixed mode operation like Web hosting, Mail Portals, Cloud Storage, Social Media websites and Data logging. This environment enabled us to stumble across and to deal with all the above listed attacks in a random manner. Though the data structure is similar to KDD cup that contain 41 attribute, the format of the live data set is Real-time Server Data (RSD) that holds contents in binary format for ease of access. The dataset contains a total of 50,000 samples of which 25000 are normal samples (50%) ad 25000 are attacks sample (50%)

B. Result Analysis

The experimental setup is split into two main steps. First, we implement the proposed evolution algorithms to reduce the feature set for the IDS. Secondly, we validate the result using random forest classifier. The evolutionary algorithm is repeated 10 times over both training and testing dataset. The experimental analysis has been evaluated for all 50 particles and the iteration are repeated for 100 generations or till maximum IDR reached. The proposed work filters attributes and reduce the dimensions of the live data set. These attributes are further used in the detection engine to analyze intrusive behavior. It produces near optimal result than other swarms intelligent techniques. The table 1 shows the 11 reduced features of the proposed method and compared with other existing methods.

Table 3: Reduced features using the Sim - Swadorest Method

Proposed Sim-Swadorest
1. Protocol Type
2. Service Name
3. Flag status-
4. Size of Source in Bytes
5. Size of Destination in Bytes
6. Emergency Flag
7.Connection Status
8. Number of Outbound commands
9. Accessed Files Count
10. Host Login Status
11. Guest Login Status

These Proposed works are further analyzed with other evolutionary and machine learning techniques based on its accuracy and false alarm rate. The detection rate (accuracy) is calculated using the number of correctly classified instances with total number of instances and for False Alarm Rate using the number of mis-classified instances with total number of instances. The records filtered based on number of normal connections. By reducing, the attribute may lead to more detection rate and reduce alarm rate in anomaly detection. The combination of SSO with RF reduce attribute more effectively in a live dataset which makes the detection process more secure. The result of SSO-RF is compared with other existing method are presented in table2 based on its performance accuracy, number of features selected, false positive rate.

Table 2: Overall performance between three methods

Methods	No.of. features	DR	FPR
GA-RF	19	75.42	3.09
PSO-RF	17	86.92	1.34
Proposed SSO-RF	11	89.62	0.98

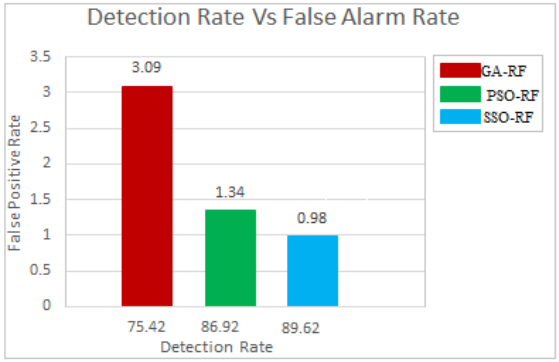


Fig 2. Average Result

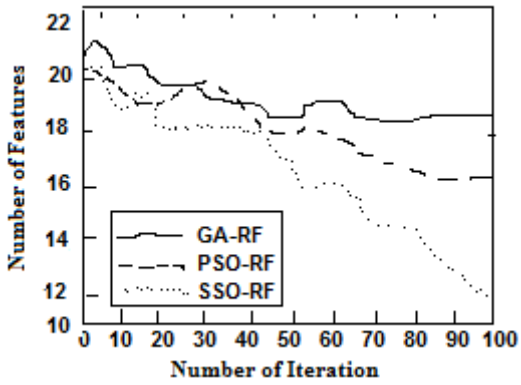


Fig 3. Feature convergence with respect to number of iteration

Figure 2 illustrates the average result of accuracy with a false alarm rate of three algorithms. The figure shows that the proposed methods have better performance than the other two methods and improve detection rate significantly. Figure 3 shows the number of feature convergence based on three methods with respect to the number of iterations. The above figure shows that the proposed methods SSO-RF are the best algorithm in reducing the number of features than PSO-RF and GA-RF.

The features that are consistently detected by a certain algorithm based on three runs out of five are as follows. For GA-RF the dominant features were 1, 2, 3, 4, 5, 7, 8, 11, 12, 13, 14, 16, 20, 23, 25, 27, 37, 38 and 40. We observed the features that are detected by PSO-RF are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 16, 17, 18, 19, 20 and 21. Finally, for SSO-RF, the dominant features are 1, 2, 3, 4, 5, 8, 9, 20 and 21. Consequently, SSO-RF is considered to be the best evolution algorithm for feature selection compared to PSO-RF and GA-RF. It gives the best values based

on a number of features, accuracy and false positive rate.

CONCLUSION AND FUTURE WORK

In this paper, we instigated and compared three different evolutionary algorithms based on random forest classifier for feature reduction problem in IDS. This paper proposed a new concept Sim-Swadorest optimization algorithm by combining SSO with RF classifier to reduce features of live data set based on anomaly detection. The experimental analysis has been compared with other existing methods such as GA-RF and PSO-RF, The result showed that proposed SSO-RF method is clearly and consistently superior compared to other problems, both in respect to classification accuracy and false positive rate.
It is observed that the proposed hybrid intrusion detection mechanism achieved average DR and average FPR than other evolutionary methods. The main advantage of Simplified Swarm Optimization than other technique is no mutation and crossover are required for processing individual particles which may reduce time effectiveness. The future work will be focused on using fuzzy rule generation to generate various rules based on its membership function and improve detection rate in real time scenario in detecting intrusion.

REFERENCE

[1] Arif Jamal Malik, Waseem Shahzad, Farrukh Aslam Khan, "Binary PSO and Random Forests Algorithms for PROBE attacks Detection in a network". In Proceedings of IEEE Congress on Evolutionary Computation, 662-668, (2011).

[2] Breiman.L ,"Random forests," Mach. Learn., vol. 45, pp. 5–32, (2001).

[3] Changseok bae, Wei-Chang yeh, Noorhaniza wahid,yuk ying chung and yao liu, "A new simplified swarm optimization (sso) using exchange local search scheme". ICIC International issn 1349-4198. Volume (8- No.6), (2012).

[4] Chen G, Chen Q, Guo W, "A PSO-based approach to rule learning in network intrusion detection", in: Advances in Soft Computing, (vol. 40), Springer, Berlin- Heidelberg, pp. 666–673, (2007).

[5] Darpa Intrusion Detection datasets [Online]. Available:www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/index.html.

[6] Dorigo M, Di Caro G, "The ant colony optimization meta-heuristic, new ideas in Optimization", 11-32, (1999).

[7] Denning D.E, "An intrusion detection model," IEEE Trans. Software. Engineering.,vol. SE-13, no. 2, pp. 222–232, (1987).

[8] Fernando Godinez and Dieter Hutter and Rahul Monroy, "Service Discrimination and Audit File Reduction for Effective Intrusion Detection", Proceedings of the 5th international conference on Information Security Applications, Springer-Verlag Berlin, Pages 99-113, (2005).

[9] https://www.linkware.in/

[10] http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol1/hmw/article1.html

[11] http://www.clusterresearchdesk.org

[12] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, October (2007).

[13] Kennedy J, Eberhart R, "Particle swarm optimization". Proceedings of the IEEE international conference on neural networks (Perth, Australia), 1942–1948. Piscataway, NJ: IEEE Service Center, (1995).

[14] Lu W and Traore I, "Detecting new forms of network intrusion using genetic programming," Computational. Intelligence. vol. 20, no. 3, pp. 474–494, (2004).

[15] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani,"A Detailed Analysis of the KDD CUP 99 Data Set", In the Proc. Of the IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA), pp. 1-6, (2009).

[16] "Nsl-kdd data set for network-based intrusion detection systems." Available on: http://nsl.cs.unb.ca/KDD/NSLKDD.html, March (2009).

[17] Safaa Zaman, Mohammed El-Abed, Fakhri Karray, "Features Selection Approaches for Intrusion Detection Systems based on Evolution Algorithms", ICUIMC(IMCOM)'13, Kota Kinabalu, Malaysia, ACM 978-1-4503-1958-4…..(2013).

[18] X. Li, Particle Swarm Optimization, a tutorial prepared for GECCO'07, 2007.

[19] Yao Liu, Yuk Ying Chung, Wei-Chang Yeh, "Simplified Swarm Optimization with Sorted Local Search for golf data classification". IEEE Congress on Evolutionary Computation, 1-8,(2012).

[20] Yeh W.C, Chang W.W, Chung Y.Y, A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method, Expert System with Applications 36 (May (4)) 8204–8211, (2009).

[21] Yuk Ying Chung, Noorhaniza Wahid, "A hybrid network intrusion detection system using simplified swarm optimization (SSO)". Appl. Soft Computing. 12(9): , 3014-3022, (2012).

[22] Zhang.J and Zulkernine.M, "Network intrusion detection using random forests," in Proc. 3rd Annu. Conf. Privacy, Secur. Trust (PST), St.Andrews, NB, Canada, pp. 53–61, (2005).

★★★