Performances of Artificial Neural Network (ANN) and Particle Swarm Optimization (PSO) Using KDD Cup '99 Dataset in Intrusion Detection System (IDS)

Performances of Artificial Neural Network (ANN) and Particle Swarm Optimization (PSO) Using KDD Cup '99 Dataset in Intrusion Detection System (IDS)

S. Norwahidayah[1*], Noraniah[1], N. Farahah [1], Ainal Amirah[1], N. Liyana[1], N. Suhana[1].

[1]Faculty of Computer Media & Technology Management, University College TATI, 24000 Kemaman Terengganu, Malaysia. Tel./Fax +6-09-8601000/+6-09-8635863,

*Email: norwahidayah@uctati.edu.my, noraniah@uctati.edu.my.

Abstract. Nowadays, the number of attacker increasing fast due to the current technologies. Most companies or an organization use Intrusion Detection System (IDS) to protect their network system. Many researchers suggest different ways to improve the IDS such using optimization techniques. Artificial Intelligence (AI) methods also proposed in IDS to attained high accuracy of detection for example; artificial neural network, particle swarm optimization and genetic algorithm. Artificial neural network (ANN) and Particle Swarm Optimization (PSO) used in this paper to equate the method and performances in IDS environment. The ANN output value will be compared with the result where ANN supported by PSO to produce higher accurate value. KDD CUP '99 Dataset used as the benchmark of IDS and will be simulated in MATLAB Simulink 2013. 200 datasets used consists of attacks and normal activities as an input. In this paper, DoS attack which are Smurf and Neptune attacks selected for detection.

1. Introduction

Today's sphere, society use the internet as their basic requirements to interconnect, online funding, and social interacting. They did not alert that their data visible to an intruder. According to the concerns, organizations spending millions of dollars to create something that can guard the sensitive data from intruders and make sure that data transmission is safe. From the research study, organizations contribute to create an Intrusion Detection System (IDS). Recently used for real-time monitoring and detect abnormal activity. to monitor activities of host system means that made by specific hosts.

IDS have two detection methods which are signature recognition and anomaly detection. The difference between both methods is that signature recognition identifies intrusions depending on structures of known attacks while anomaly detection analyses the properties of normal behaviour. The following subcategory points out both types of recognition approaches. A probable of intruder might style its malicious activity with aspect effects which will cause odd behaviour of the system. Observation such aspect effects is tough since their location is hardly detectable. The challenges for IDS in a network environment are to track users and entities as they move across the network. The drawbacks of IDS are to identify the abnormal activities among the normal.

ANN is talented to solve the classification and regression problems. However, it has some constraint of its capability to learn from observing the dataset [4]. The goal during the learning process

in FNN is to search the best combination of linking weight and biases to achieve the smallest error. Nevertheless, most of the time FNN converge to points which are the best solution locally but not globally. The problems occur before is slow convergence and local minimal get trapped in the Back Propagation Neural Network (BPNN) [3]. The issues of artificial neural network before is its capability to learn from observing the dataset [2]. DoS attack reduces the server performances by overflowing the ICMP traffic from Feb 2000 until now it still occurred [5].

2. Artificial Neural Network

Artificial Neural Network (ANN) is an effective way to increase the performances of IDS system which are based on learning and training processes can be used for misuse detection and anomaly detection [19]. Artificial Neural Network has two types of learning process which are Supervised Training Process and Unsupervised Training Process. Supervised training process train the input and output pattern while unsupervised training process trained only by input pattern.

The main functions of a Neural Network are to learn automatically and retrain masses according to data inputs and data outputs to produce the exact result. The neural network is non-linear numerical data modelling tools. It can be used to model difficult relationships among inputs and outputs to discover design in data. When the input enters the network, the output will be generated each of the input values of the output neurons [19]. The hidden layer functioning where the weightage can be modified to get the accurate output.

3. Particle Swarm Optimization

Particle Swarm Optimization is an algorithm for population clustering using iterative method inspired by behavior of bird flocking or bird schooling. PSO is categorized in global search algorithm where it is used to commit the exploration and exploitation in order to find the position areas such as swarm movement of animals and their social interactions [13]. However, this algorithm is adopted to optimize wavelet neural network [12].

This algorithm is zero-order which it can be applied to a diversity issues via discontinuous or non-convex and multimodal problems [11]. The algorithm aims to find the perfect point of the particles in the problem capacity including the velocity of the particle at the outset. The velocity is measured based on its current velocity, the distance from the excellent position and populations [11]. According to [11], there are two characteristics to be considered as a comparison basis, that is:

1. Elapsed time, duration or period of algorithm takes to complete the task
2. Classification accuracy, ratio of the test data categorized by algorithm

In addition, the PSO also consider the population size, maximum and minimum inertia weight, maximum number of iterations [13] and finally release the global optimum solution and result of calculation [12]. In PSO, the input called “particles” that keep track of its coordinated in problem. The group of random particle called “solutions”. Each of iteration used to update “best” values which are pbest and gbest. When particles in the best population as topological neighbors called “lbest”. Particle need to calculate/ update velocity and positions to find the pbest and gbest.

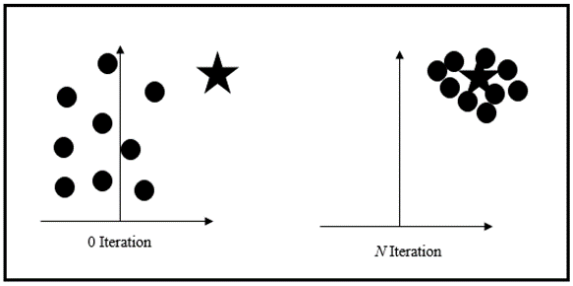

Figure 1. PSO Diagram

4. KDD Cup '99 Dataset

The 1999 version of MIT Lincoln Laboratory – DARPA (Defence Advanced Research Projects Agency) intrusion detection evaluation data was used in this research [7]. The model version of the dataset involved more than 450,000 connection records. A subsection of the data that have the wanted attack types and a reasonable number of normal events were selected manually. The dataset widely used for the evaluation of anomaly recognition methods. The attacks can be categorized as follows:

1) Denial of Service Attack (DoS): DoS attack is a harmful attack that an attacker sends many ping to make the network busy and slow. So that, the attacker can reject the authentic user access to the servers [7].
2) User to Root Attack (U2R): The attacker has a local access to the victim's host and get the source access to the system [8]
3) Remote to Local Attack (R2L): The attacker did not have a local access to the victim machine but tries to gain the local access as a user to the machine [8].
4) Probe Attack: Attacker try to gain information about the victim machine and purpose to avoid the security controls [7].

Generally, most researchers applied KDD CUP 10% Dataset as benchmark of IDS which include the total of dataset is 494,020. The distribution dataset for Normal is 97280, Probe is 4107, DoS is 391458, U2R is 52, R2L is 1124 [9]. Table 1 shows the distribution of intrusion types in datasets.

Table 1. Dataset Intrusion Categories [9].

Dataset	Normal	Probe	DoS	U2R	R2L	Total
10% Dataset	97280	4107	391458	52	1124	494020

The type's attack of Probe includes satan, ipsweep, nmap and portsweep attack. Besides, back, land, Neptune, pod, smurf and teardrop attack is categorized as DoS. R2L involved guess password, ftp write, imap, phf, multihop, warezmaster, warezclient and spy attack. Last but not least, the buffer overflow, loadmodule, perl and rootkit categorized as U2R. Table 2 show the original number of dataset model due to the attacks discussed above.

Table 2. Attack Categories and Dataset Sample in 10% KDD CUP '99.

Attack	Samples	Categories
normal	97,277	NORMAL
back	2,203	DOS
land	21	DOS
neptune	107,201	DOS
pod	264	DOS
smurf	280,790	DOS
teardrop	979	DOS
satan	1,589	PROBE
ipsweep	1,247	PROBE
nmap	231	PROBE
portsweep	1,040	PROBE
Guess_passwd	53	R2L
ftp_write	8	R2L
imap	12	R2L
phf	4	R2L

multihop	7	R2L
warezmaster	20	R2L
warezclient	1,020	R2L
spy	2	R2L
Buffer_overflow	30	U2R
loadmodule	9	U2R
perl	3	U2R
rootkit	10	U2R

This paper focused on Neptune and Smurf types of attack. Those attack characterized as DoS Attack. Dos Attack makes the computer system to be hectic by sending many ping. Neptune attack working to SYN flood Denial of Service on one or more ports while the function of Smurf attack is Denial of Service ICMP echo reply flood. Neptune attack make the memory resources busy by sending TCP packet requesting to initiate a TCP session.
After many TCP packet sent, the computer system finally runs out of memory resources. Smurf attack is the popular type of attack sending ICMP echo request packet to middle device. ICMP packets have source address name as victim's IP address and middle device address as destination address. DoS attack reduces the performance of server by overflowing the ICMP traffic as discussed above in the Neptune and Smurf attack. So, detection of DoS attack is very significant to protect the system.

5. Methodology

The first part is collect and process dataset to be used for training ANN. The second part will focus on applying the PSO on dataset. The dataset will then also be used to classify the normal and abnormal activities. The output value of ANN will be supported by PSO to maximize the classification rate. The final parts involved the analysis and comparison of the optimization of the combination of ANN and PSO. The methodology is divided into three parts: Collecting and processing dataset and Process flow.

5.1 Collecting and Processing Dataset

In this paper, only 4 critical dataset features selected, and 200 datasets used to train ANN. Besides, one column added for categorize each record as a result which are 0 for normal and 1 for attack in the standard format step. The reason of adding this additional row is to continue and understand the error reorganization in ANN. The given attributes in the dataset are converted into double data type to make it compatible with the ANN toolbox of MATLAB because it is only support the data in integer form [10]. The "protocol_type" feature converted with values like tcp is 2 and udp is 3. Table 3 shows parameters that used in this study is number of agents, maximum number of iteration, number of training samples, inertia weight, minimum weight, maximum weight and objective function.

Table 3. Parameter Setting

Parameter	**Values**
Mass	30
Maximum of Iteration	50
Number of Training Samples	200
Inertia Weight	2
Minimum weight	0.5
Maximum weight	0.9
Objective Function	Maximization

5.2 Process Flow

First, divide the dataset into two data segments for training and testing. Then the methods develop the standard dataset format for the restructuring of ANN. After the training of the ANN completed, the ANN categorize the KDD CUP '99 testing dataset and take the accuracy output of the system detection, then plot and observe it as a result of the system. When the recognition of the ANN is finished, the ANN reorganization will attempt by PSO. After that, the results will be compared with the ANN result and ANNPSO. The purpose and methods have been illustrated in Figure 2 to support ANN in IDS optimization, the Particle Swarm Optimization (PSO) will be compared in this paper.

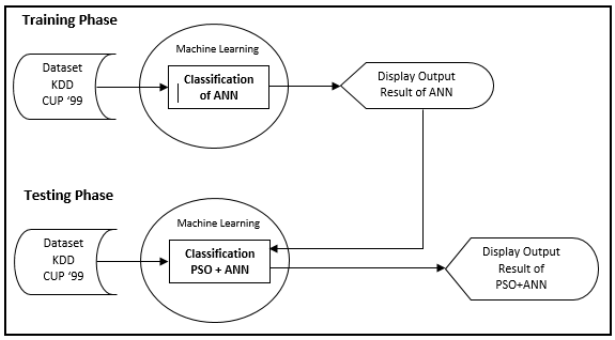

Figure 2. Process Flow

6. Result

6.1 Relevant Features Selection

In this project, only 4 critical features selected to classify the attack (Smurf attack and Neptune attack) and normal activities in intrusion detection system. The experimentation will compare the classification rate result with another best features selection.

Table 4. Relevant Features for Normal, Smurf Attack and Neptune Attack [14]

Class Label	Relevant Features
Normal	3,6,12,23,25,26,29,30,33,34,35,36,37,38,39
Smurf	2,3,5,6,12,25,29,30,32,36,37,39
Neptune	3,4,5,23,26.29,30,31,32,34,36,37,38,39

The 20 numbers (2, 3, 4, 5, 6, 12, 23, 25, 26, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39) of features to categorize the Neptune attack, Smurf attack and normal in IDS. The 20 of features with 200 of data samples selected and implemented in the simulation process in MATLAB. The difference analysis and result using ANN and PSO will be discussed.

6.2 Feed Forward Neural Network Result

Figure 3 shows the attack classification rate of the ANN methods achieve 93% with 50 times of iteration.

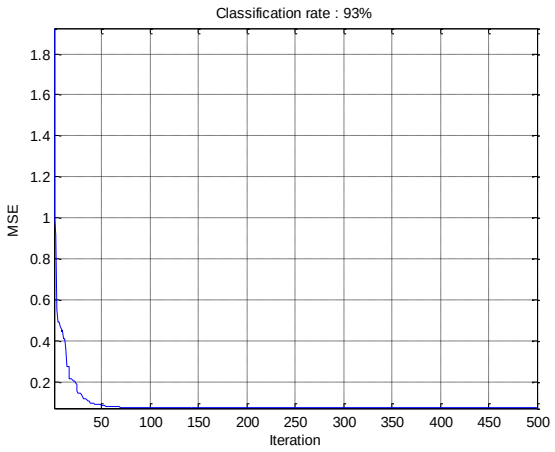

Figure 3. Classification Rate result of Artificial Neural Network

6.3 Particle Swarm Optimization Result

Figure 4 shows the classification rate of the PSO methods. The result shows 98% classification rate with 50 the number of iterations. PSO proof that its capability in attack detection more accurate than ANN.

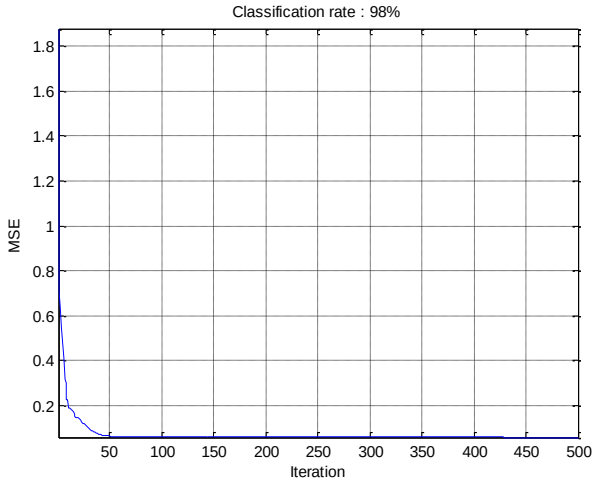

Figure 4. Classification Rate of PSO.

Figure 5 present the comparison classification rate result of ANN and PSO. According to the result, the PSO produce highest classification rate of the detection which are 98%. ANN produce 93% of the detection classification rate. The result produced by those method very high because the training sample used is small and the dataset features used is 20 that can detect the normal behaviour and focused attack in this paper. Maybe when big training sample used, the classification rate result will be decreased.

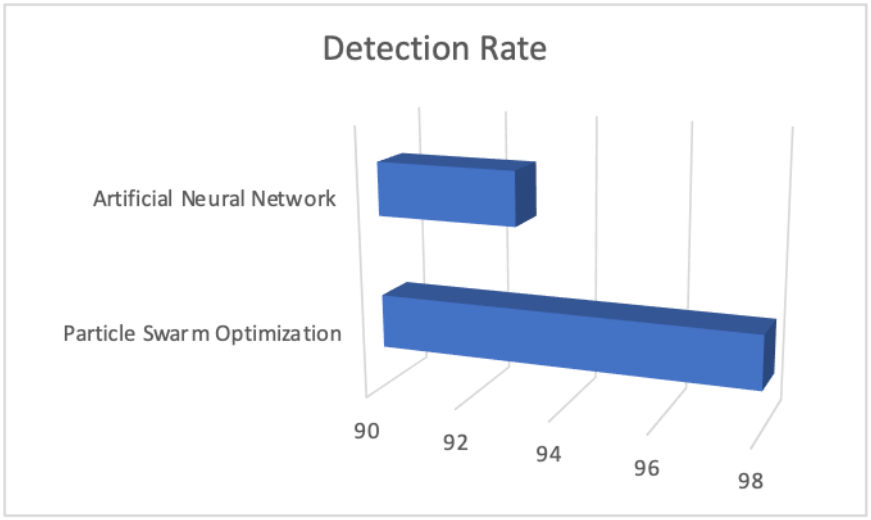

Figure 5. Result Comparison

7. Conclusions

A training optimization algorithm called hybrid of ANN and PSO are investigated. KDD CUP '99 Dataset used to represent as the packet and simulated in the MATLAB Simulink. The project stages offered as follows: collecting and processing for training ANN in IDS; applying the PSO on dataset; classify the dataset which are normal and abnormal activities.

The combination optimization method proven hybridization of ANN and PSO achieve high accuracy. For further study, other algorithm can be used to apply in IDS for attacks detection. Besides, solve the problem using other AI techniques that no need to train the input data to get the accurate output value. The result also can include the percentage of all the types of attacks detection to summarize the classification rate of the detection.

8. References

[1] Asmaa Shaker Ashoor and Sharad Gore 2011 International Journal of Scientific Engineering Research.
[2] Fatai Adesina Anifowose and Safiriyu Ibiyemi Eludiora 2012 World Applied Programming Vol (2).
[3] Seyedali Mirjalili and Ali Safa Sadiq 2011 IEEE.
[4] Amin Dastanpour, Suhaimi Ibrahim, Reza Mashinchi and Ali Selamat 2014 Smar Computing Review, Vol. 4, No. 6.
[5] Prajakta Solankar, Subhash Pingale and Ranjeet Singh Parihar 2015 (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (2).
[6] S. Sinaie 2010 Solving shortest path problem using Gravitational Search Algorithm and Neural Networks, Universiti Teknologi Malaysia (UTM), Johor Bahru, Malaysia, M.Sc. Thesis.
[7] Mehdi Moradi and Mohammad Zulkernine 2004 Natural Sciencesand Engineering Research Council of Canada (NSERC).
[8] Swati Paliwal and Ravindra Gupta 2012 International Journal of Computer Applications, Volume 60– No.19.
[9] Megha Aggarwal, Amrita 2013 International Journal of Scientific & Technology Research Volume 2, Issue 6.
[10] Indraneel Mukhopadhyay and Mohuya Chakraborty 2014 Journal of Information Security.
[11] Aburomman, Abdulla Amin, Mamun Bin, and Ibne Reaz 2016 38 Elsevier B.V.:360–72.
[12] Dongmei, Zhao, and Liu Jinxing. 2018 125 (February):764–75.
[13] Marouani, H, and Y Fouad 2019 Physica A 514. Elsevier B.V.:708–14.

[14] N. S. Chandolikar and V. D. Nandavadekar 2012 MIT International Journal of ComputerScience & Information Technology, Vol.2, No. 2.
[15] Md Nasimuzzaman Chowdhury and Ken Ferens, Mike Feren 2016 International Conference. Security and Management.
[16] Erfan A. Shams, • Ahmet Rizaner 2017, Springer.
[17] Amin Dastanpour, Suhaimi Ibrahim, Reza Mashinchi and Ali Selamat 2014 Smar Computing Review, Vol. 4, No. 6, December 2014.
[18] V.Selvi, Dr.R.Umarani 2010 Vol 5. No.4, International Journal of Computer Applications (0975 – 8887).
[19] Jose Ernesto Luna Dominguez and Anabelem Soberanes Martin 2015 International Journal of Engineering Science and Innovative Technology, Vol. 4, Issue 2.

Acknowledgments
This research is fully supported by FRGS grant, 2018/ICT02/TATI/1. The authors fully acknowledged University College TATI for the approved fund which makes this important research viable and effective.