Data Preprocessing for Intrusion Detection System using Swarm Intelligence Techniques

S. Revathi
Ph.D. Research Scholar
Government Arts College
Coimbatore-18

A. Malathi,Ph.D
Assistant Professor
Government Arts College
Coimbatore-18

ABSTRACT

Due to access of malicious data in internet, Intrusion detection system becomes an important element in system security that controls real time data and leads to huge dimensional problem, so a data pre-processing is necessary to reduce haziness and to clean network data. To reduce false positive rate and to increase efficiency of detection, the paper proposed a new swarm intelligence technique to solve complex optimization problem. The paper work based on hybrid Simplified Swarm Optimization (SSO) algorithm to pre-process the data. SSO is a simplified Particle Swarm Optimization (PSO) that has a self-organizing ability to emerge in highly distributed control problem area, and is versatile, strong and cost effective to resolve complex computing environments. It recognize not only known attacks but also filters noisy and irrelevant data that may result on knowledge Discovery and Data Mining (KDDCup 1999) dataset and compared to a new hybrid Partial Swarm Optimization with Random Forest (PSO-RF) and with other benchmark classifiers. The testing result shows that the proposed method provides competitively high detection rates and produce a near optimal solution.

KEYWORD

Swarm intelligence, Simplified Swarm Optimization, Partial Swarm Optimization, Random Forest, Intrusion detection.

1. INTRODUCTION

The widespread use of computers and internet has enhanced the worth of life for many people, but it also exposed to increasing security threats both externally and internally. The security of a computer system is compromised when an intrusion takes place [1]. Different technologies have been developed and deployed to protect computer systems against network attacks, such as firewall, message encryption, secured network protocols, password protection, and so on. Despite Intrusion prevention techniques, it is nearly impossible to have a completely secured system. As a result, Intrusion Detection System (IDS) have become an essential component of security to detect these threats, identify and track the intruders. As IDS must have a high Detection Rate (DR), with a low False Alarm Rate (FAR) which is a challenging task [4].

In recent years many biology inspired approaches have made their appearance in a variety of research fields, and plays a vital role in intrusion to improve their efficiency and performance. Swarm intelligence is one of them [2]. Techniques and algorithms of this research field draw their inspiration from the behavior of insects, birds and fishes, and their unique ability to solve complex tasks in the form of swarms. Among swarm intelligence techniques Particle Swarm Optimization (PSO) is a popular heuristic techniques for optimization, but it suffers from premature convergence of high dimension multimodal problem which flops to achieve best fitness value [3].

The KDDcup99 dataset used for intrusion detection is a raw data which highly susceptible to noise, missing values and inconsistency. To improve quality of raw data, data pre-processing and filtering is required which increase data efficiency. As a result the paper proposed a novel simplified swarm optimization, to mine the raw data. The main objective of the paper is to screen incomplete data and to reduce irrelevant feature. SSO improves the performance efficiency, time and memory than PSO-RF and other classifiers for filtering data.

The rest of the paper is structured as follows: section 2 present some related work based on swarm intelligence for intrusion detection dataset. Section 3 present an overview of framework. Section 4 explains about technique in swarm intelligence and data mining. Section 5 explain in detail about proposed work of hybrid SSO algorithm and its efficiency is compared with PSO-RF and other classifier. Section 6 concludes some result based on proposed work.

2. RELATED WORK

Intrusion Detection Systems gross raw network data or audit records as input, process which leads to a huge network traffic data size and the 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 G. Sunil Kumar [5], based on hybrid classifiers on binary particle swarm optimization and random forests algorithm for the classification of probe attacks in a network. Dharmendra G. Bhatti [10] proposed a method to reduce false positive rate using data pre-processing method.

Now-a-days biological inspired approaches have been extensively instigated in network intrusion pattern detection. This field of study is known as “swarm intelligence” and has attracted an increasingly number of researchers since the proposal of Particle Swarm Optimization (PSO) [27] algorithm and also of

the Ant Colony Optimization (ACO) Algorithm (DORIGO et al., 1996). Michailidis et al. were among the first who merged the two aforementioned soft computing techniques to create an improved system for intrusion detection (Michailidis et al., 2008).In [6][7], 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.[8]. 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 [9] proposed a new Network intrusion detection using hybrid binary PSO and random forests algorithm which reduce the dimension of the data and finds optimal solution, similarly Tie-Jun Zhou et al. (2009) proposed an ID Based on Particle Swarm Optimization (PSO) and Support Vector Machine (SVM). The use of PSO-SVM in Intrusion Detection established a classification model; at the same time verified the validity of the model.

Category Of Attacks	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

3. OVERVIEW OF FRAMEWORK

Simplified swarm optimization technique detect intrusion and reduce false positive rate. The framework is shown in figure1. The information is obtained from KDD cup99 dataset [11], the records in the database contains 41 features in which the data may be incomplete, noisy or duplicate in nature. The proposed pre-processing approach filters data effectively and the result compares with existing approach.

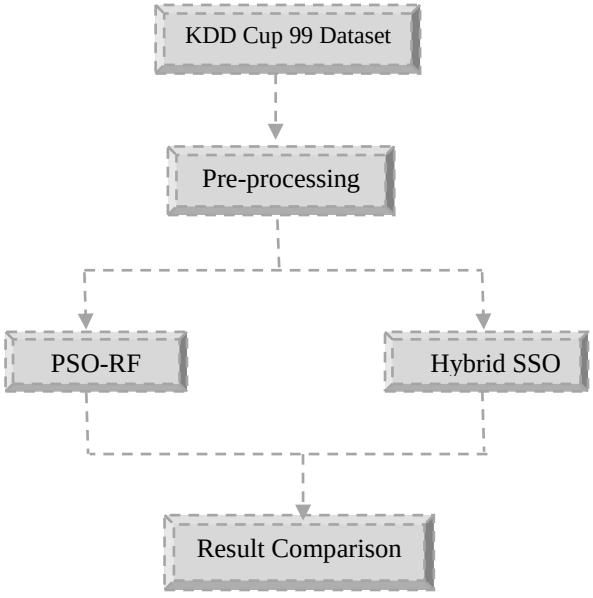

Figure1 Proposed Framework

3.1 Dataset Description

In this paper KDD cup99 dataset is used which is derived from DARPA98 TCP dump data prepared and managed by MIT Lincoln Laboratory. A complete KDD'99 dataset contains five millions connection records where 4,898,431 are labelled connections with attacks used for experimentation [12]. The attacks in the dataset fall into four categories: DoS (Denial of Service), R2L (unauthorized access from a remote machine), U2R (unauthorized access to root privileges), and probing are divided into 22 different attack classes that are tabulated in Table 1 [5].

Table 1. Detail of Attacks of Labelled Records

Where,

- DoS attacks: use of resources or services is denied to authorize users.
- Probe attacks: information about the system is exposed to unauthorized entities.
- User to Remote attacks: access to account types of administrator is gained by unauthorized entities.
- Remote to Local attacks: access to hosts is gained by unauthorized entities.

3.2 Analysis of Pre-processing Step

Data Pre-processing (DP) Phase, in order to reduce data as much as possible without any information loss, and required specialized planning, training and testing .The issues derived from the system analysis are [14]:

- To provide an optimal and efficient computing data for IDS.
- To filter false rates and improve detection rates.
- To discover attack patterns and display appropriate data types for administrators to make policies.

4. SWARM INTELLIGENCE IN INTRUSION DETECTION

A swarm can be considered as a group of cooperating agents to achieve some purposeful behavior and task [13]. It is introduced by Beni and Wang (1989) has received widespread attention in research, mainly as Ant Colony Optimization (ACO), Particle swarm Optimization (PSO) and Bee Colony Optimization (BCO) based on several collective behavior like birds flocking, ant colonies, social insects and swarm theory.

The Bio Inspires techniques in table2 are mainly used to solve optimization problem and are applicable in very diverse provided that the problem can be quantified into some form of fitness measure. The main strength is that they are parallel in nature [17]

Table 2. Bio-inspired approaches

Author	Year	Techniques proposed
Goldberg D. E.	1989	Genetic algorithm
John R.Koza	1994	Genetic Programming
Dorigo et al.	1999	Ant Colony Optimization
Kennedy and Eberhart	1995	Particle Swarm Optimization

GAs and PSO are commonly associated with the optimization of continuous numerical functions, and ACO with combinatorial optimization [16]. Some of the benefits of adopting such techniques are flexibility in retraining, online/continuous learning and the potential for parallelism in the algorithms, which can be exploited both in the training and detection process.

4.1 Partial Swarm Optimization

Particle swarm optimization could be a heuristic international optimization technique advises originally by Doctor Kennedy and Eberhart in 1995 [28]. It developed from swarm intelligence and is predicated on the analysis of bird and fish flock movement behavior. It shares several similarities with biological process computation techniques like Genetic Algorithms (GA). The system is initialized with a population of random solutions and searches for optima by change generations. The employment of PSO based mostly data processing in classification rule discovery was first proposed by Sousa et al [15].

Partial swarm optimization based on position of particles in D-dimensional space. The particles changed based on three conditions based on its inertia, optimist position and Swarm's optimist position [29]:

The position of every particle within the swarm is affected both by the foremost position throughout its movement and by its close. Once the complete particle swarm is close the particle, the optimist position of the surrounding is up to the one amongst the complete; this rule is named the complete PSO. In PSO rule initial population of random particles, and it moves around a multidimensional search area with velocity update supported particle's best position (pbest) or local best position (lbest), whereas the latter is named the particle's global best position (gbest) [18]. In PSO, a swarm consists of N particles traveling in an exceedingly D-dimensional searching area x_{id}, the individual best position p_{id}, and the velocity v_{id}. Particles square measure initialized with random vectors for initial positions and initial velocities. At every step, velocities and positions of the complete swarm square measure updated consequently with the principles represented in following equation (1). (Shi Y, Eberhart R C, 1998):

$$V_{id} = w.v_{id} + c1.rand1.(p_{id} - x_{id}) + c2.rand2.(p_{gd} - x_{id}) \quad (1)$$

Each particle moves to a new potential position as follows in eq (2):

$$x_{id} = x_{id} + v_{id}; d = 1,2,3....., D \quad (2)$$

Where c1 is called cognitive parameter while c2 is called social parameter, and in most cases (c1 + c2) is equal to 4. rand1 () and rand2 () are random numbers generated in the range between 0 and 1. w is an inertia weight that control the ability of global and local search.

PSO mainly based on mathematical foundation and application research to prove its convergence and robustness. It can be combined with the other intelligent optimization methods to design several compound optimization methods; PSO can be also led into scattering system, compound optimist system, non-coordinate system to develop PSO's application ranges. The main advantage of PSO is as follows [29]:

- PSO is predicated on the intelligence and applied to each research project and engineering use.
- It had no overlapping and mutation calculation. The search supported speed of the particle.
- Calculation is extremely easy and simple. It occupies the larger optimization ability
- PSO adopts the real number code, and it's set directly by the solution. The quantity of the dimension is capable to the constant of the answer.

But the drawback of PSO which leads to development of Simplified swarm optimization are

- It suffers from partial optimism which reduce its speed and regulates the direction.
- The method cannot work out the problems of scattering and optimization (Chen Yonggang, Yang Fengjie, Sun Jigui, 2006, (In Chinese)).
- Not suitable for energy field.

4.2 Random Forest

Random Forests were introduced by Lepetit et.al. [19] [20]. It is a generic principle classifier, that generate many decision tree classifier algorithm with different bootstrap sample that uses combination of L tree-structured base classifiers $\{h(X, \Theta_n), N=1, 2, 3\ldots L\}$, where X denotes the input data and $\{\Theta_n\}$ is a family of identical and dependent distributed random vectors. Every Decision Tree is made by randomly selecting the data from the available data. For example, a Random Forest for each Decision Tree can be built by randomly sampling a feature subset. By injecting randomness at each node of the grown tree, it has improved accuracy. The correlation between trees is reduces by randomly selecting the features which improves the prediction power and results in higher efficiency. As such the advantages of Random Forest are [21]:

- Overcoming the problem of over fitting
- In training data, they are less sensitive to outlier data
- Parameters can be set easily and therefore, eliminates the need for pruning the trees variable and accuracy is generated automatically

Random Forest not only keeps the benefits achieved by the Decision Trees but through the use of bagging on samples, its voting scheme through which decision is made and a random subsets of variables, it most of the time achieves better results than Decision Trees [22]. It can easily handle high dimensional data modelling such as missing values and can handle continuous, categorical and binary data. The bootstrapping and ensemble scheme makes Random Forest strong enough to overcome the problems of over fitting and hence there is no need to prune the trees.

5. PROPOSED SSO ALORITHM

The paper proposed a new swarm intelligence approach based on simplified swarm optimization. It filters data and reduce incomplete, noisy and dimensionality problem for both discrete and continuous variables in dataset [23]. 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. The introduction of SSO algorithm is as follows [24].

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 (3).

$$x_{id}^{t} = \begin{cases} x_{id}^{t-1} & if \quad rand() \in [0.\ c_w) \\ p_{id}^{t-1} & if \quad rand() \in [c_w,\ c_p) \\ g_{id}^{t-1} & if \quad rand() \in [c_p,\ c_g) \\ x & if \quad rand() \in [c_g,\ 1) \end{cases} \qquad (3)$$

Where i = 1; 2; m, where m is the swarm population. $X_i = (x_{i1}; x_{i2}; : : : ;x_{id})$, where x_{id} is the position value of the i-th particle with respect to the d-th dimension of the feature space. C_w, C_p and Cg are three predetermined positive constants with $C_w < C_p < C_g$. $P_i = (p_{i1}; p_{i2}; : : : ; p_{id})$ denotes the best solution achieved so far by itself (p_{best}), and the best solution achieved so far by the whole swarm (g_{best}) is represented by $G_i = (g_{i1}; g_{i2}; : : : ; g_{id})$. 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), C_w, C_p and C_g.

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

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

6. EXPERIMENTAL AND RESULT ANALYSIS

6.1 Experimentation:

This section describes the experimental results and performance evaluation of the proposed system. For experimental simulation KDD cup 99 data, which is widely used for evaluating intrusion detection system is used. The proposed system can easily filters large scale dataset. Data pre-processing removes unwanted parameters which decreases overlapping behavior of normal and intrusive data. In proposed work modern data mining [25] and swarm intelligence [26] [13] based data cleaning provides filtered data and improves detection rate. It easily process huge amount of data in real time which is a challenge faced by most intrusion detection system. The aim of our proposed work is to filter out normal data and to reduce dimensionality of the training KDD cup 99 dataset. The proposed pre-processing work reduce false positive rate and improve efficiency. It produce near optimal result than other swarm intelligent techniques. Thus, the implementation specifications and parameters involved in these methods for the above mentioned swarm intelligence algorithms have listed below. The parameters are based on SSO obtains best achievement from all possible trials.

6.2 Result Analysis

The raw data used for intrusion detection consist of duplicate and spurious data which need to be pre-processed, so that we remove inconsistent data that is not related to the purpose of intrusion detection. In this paper the proposed hybrid simplified swarm optimization technique removes such noisy and duplicate data, helps in reducing false positive rate for intrusion detection. Pre-processing mainly based on data mining techniques may lead to average efficiency in filtering raw data. In this paper newly proposed swarm intelligent technique result in better efficiency than other data mining algorithm. Figure 2 shows experimental result for between existing and proposed algorithm to filters normal data that shows better efficiency than existing method.

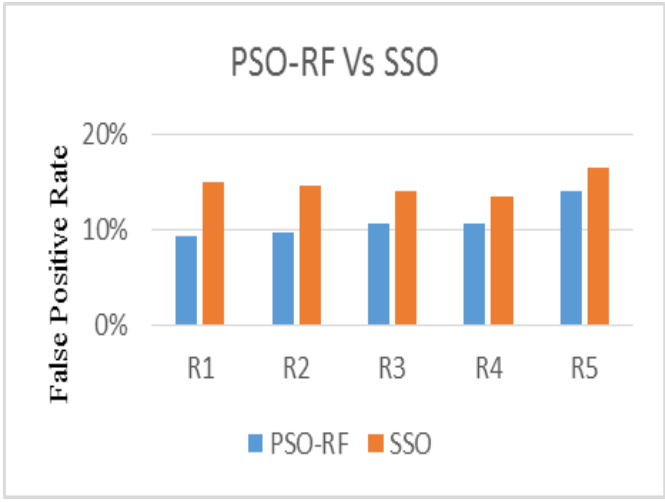

Figure 2 Comparision between PSO-RF VS SSO

The performance of preprocessed data is shown in figure 3 which filters noisy and incomplete data form raw data (KDD cup99 dataset) and it shows that the proposed system reduce feature selection attribute that reduce false positive rate and improves efficiency for intrusion detection system

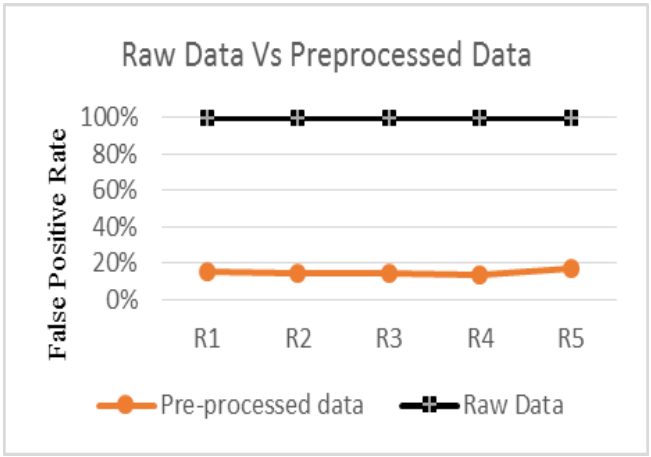

Figure 3 Raw data VS Pre-processed data

7. CONCLUSION

SI techniques have established themselves as a solid option for any contemporary IDS. PSO based IDS have been extensively studied in combination with other machine learning techniques constantly providing solid DR rates. To overcome the difficulties of PSO the paper proposed a new simplified version of the PSO algorithm known as simplified swarm optimization. The algorithm proposed here is very simple to implement and the performance is compared with hybrid PSO-RF. KDD cup 99 dataset is used to test the algorithm in search for global optimal solution. By generating an optimal solution in pre-processing module makes intrusion detection more accurate and reduce false positive rate. The experimental result shows that 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.

8. REFERENCE

[1] R. Heady, G. Luger, A. Maccabe, and M. Servilla. The architecture of a network level intrusion detection system. Technical report, Computer Science Department, University of New Mexico, (August 1990).

[2] Khaled Sellami, Rachid Chelouah, Lynda Sellami, Mohamed Ahmed-Nacer, Intrusion Detection Based on Swarm Intelligence using mobile agent, ICSI 2011: International conference on swarm intelligence, Cergy, France, June 14-15,(2011).

[3] P.Amudha, H.Abdul Rauf Ph.D, A Study on Swarm Intelligence Techniques in Intrusion Detection, IJCA Special Issue on "Computational Intelligence & Information Security" CIIS (2012).

[4] 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).

[5] 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).

[6] S. Banerjee, C. Grosan, A. Abraham, P.K. Mahanti, Intrusion detection on sensor networks using emotional ants, International Journal of Applied Science and Computations 12 (3) 152–173, (2005).

[7] S. Banerjee, C. Grosan, A. Abraham, IDEAS: intrusion detection based on emotional ants for sensors, in: Proceedings of the 5th International Conference on Intelligent Systems Design and Applications (ISDS'05), Wroclaw, Poland, pp. 344–349, (2005).

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

[9] 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).

[10] Dharmendra G. Bhatti, P. V. Virparia, Bankim Patel, Data Pre-processing for Reducing False Positive Rate in Intrusion Detection, International Journal of Computer Applications (0975 – 8887) Volume 57– No.5, November (2012).

[11] KDDCUP 99 dataset, available at: http://kdd.ics.uci.edu/dataset/kddcup99/kddcup99.html.

[12] Bhawana Pillai, Uday Pratap Singh, NIDS for Unsupervised Authentication Records of KDD Dataset in MATLAB, (IJACSA) International Journal of Advanced Computer Science and Applications, Special Issue on Wireless & Mobile Networks, Page 57 – 61, ISSN 2156-5570 (Online), (2011).

[13] S. H. Zahiri and S. A. Seyedin, Swarm intelligence based classifiers, Journal of the Franklin Institute, vol.344, no.5, pp.362-376, (2007).

[14] Salem benferhat, karima sedki, karim tabia, pre-processing rough network traffic for intrusion detection purposes, iadis international telecommunications, networks and systems (2007).

[15] T. Sousa, A. Silva, A. Neves, Particle swarm based data mining algorithms for classification tasks, Parallel Computing 6 (May/June) (2004) 767–783.

[16] K. Shafi, H.A. Abbass, Biologically inspired complex adaptive systems approaches to network intrusion detection, Information Security Technical Report 12 (4) (2007) 209–217.

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

[18] Angeline P J. (1999). Using selection to improve Particle Swarm Optimization of the 1999 Congress on Evolutionary Computation. Piscataway. NJ: IEEE Press, (1999):84-89.

[19] Lepetit, V., Fu, P.: Key point recognition using randomized trees. IEEE Trans. Pattern Anal. Mach. Intel. 28(9), 1465– 1479 (2006).

[20] Ozuysal, M., Fua, P., Lepetit, V.: Fast key point recognition in ten lines of code. In: IEEE CVPR (2007).

[21] Bosh, A., Zisserman, A., Munoz, X.: Image classification using Random Forests and ferns. In: IEEE ICCV (2007).

[22] Introduction to Decision Trees and Random Forests, Ned Horning; American Museum of Natural History's.

[23] 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.

[24] Yuk Ying Chung, Noorhaniza Wahid: A hybrid network intrusion detection system using simplified

swarm optimization (SSO). Appl. Soft Computing. 12(9): 3014-3022 (2012).

[25] Tadeusz Pietraszek and Axel Tanner, Data Mining and Machine Learning - Towards Reducing False Positives in Intrusion Detection, Information Security Tech. Report (Elsevier Advanced Technology Publications Oxford, UK), Volume 10 Issue 3, Pages 169-183, (January 2005).

[26] Zheng Zhang, Jun Li, C.N. Manikopoulos, Jay Jorgenson, Jose Ucles, HIDE: a Hierarchical Network Intrusion Detection System Using Statistical Pre-processing and Neural Network Classification, Proceedings of the 2001 IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY, 5-6 (June, 2001).

[27] Banks A, Vincent J, Anyakoha C. A review of Particle Swarm Optimization, Part I: Background and Development, Natural Computing, vol.6, 467–484. (2008).

[28] Kennedy J, Eberhart R. Particle Swarm Optimization. In Proceedings of IEEE International Conference on Neural Networks, 1942-1948. (1995).

[29] Qinghai Bai. Analysis of Particle Swarm Optimization Algorithm. In Computer and information Science, Vol 3, No1. (Feb 2010).