Volume 3, Issue 7, July 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Optimization of KDD Cup 99 Dataset for Intrusion Detection Using Hybrid Swarm Intelligence with Random Forest Classifier

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

Dr. A. Malathi
Assistant Professor
Government Arts College
Coimbatore-18, India

Abstract: Intrusion detection system plays a vital role in system security which operates data in real time that may leads to dimensionality problem. KDD cup 99 which widely used as a benchmark dataset to detect intrusion is analysed in this paper. The main drawback of the dataset is its redundancy and duplicate records which reduce accuracy and increase false alarm rate (FAR), so a data pre-processing is necessary to reduce obscurity and to clean network data. The Data mining algorithms does not overcome the difficulties of dataset problems, this paper proposed a new method of combining swarm intelligence (Simplified Swarm Optimization) with data mining algorithm (Random Forest) to pre-process the data. The proposed method easily evaluates the incomplete data and maintains accuracy, when a large proportion of the data are missing. By incorporating SSO with RF it discovers a better performance than any other existing data mining methods. The testing result shows that the proposed method provides a competitively high detection rates and achieves a near optimal solution.

Keywords: Swarm intelligence, Simplified Swarm Optimization, Random Forest, Data mining, Intrusion Detection.

I. INTRODUCTION

Network security becomes more important with the enormous growth of computer network usage both internally and externally. To defend against various attack lots of computer security techniques have been intensively studied in the last decade, among them network intrusion detection (NID) has been considered to be one of the most auspicious methods for defending complex and dynamic intrusion behaviors. The two main techniques of intrusion the researcher focused are misuse and anomaly detection. The misuse detection is also called as signature based IDS. On the other hand anomaly detection is used to detect audit data that differentiate abnormal data from normal one. Now- a-days the researcher mainly focused on using data mining algorithms to solve problem of network intrusion based security attack. It has Ability to process large amount of data and reduce data and by extracting specific data, which improves performance optimization of detection rules. The algorithms mainly based on Bayesian approaches [1, 29] to decision trees [2, 3], from rule based models [4] to functions studying [5]. The detection efficiencies therefore are becoming better and better than ever before, but still IDS having a high Detection Rate (DR), with a low False Alarm Rate (FAR) is a challenging task. In recent years many biology inspired approach such as Genetic Algorithm (GA) [6] [7], Genetic Programming (GP), Ant Colony (AC) [8], Immune Algorithm, Artificial Bee Colony and Swarm Intelligence (SI) [9] plays a vital role in intrusion to improve their efficiency and performance. SI based on inspiration from the behaviour 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

The KDD cup 99 dataset used for intrusion detection is a raw data which highly susceptible to noise, missing values and inconsistency [10]. The main drawback of dataset is its huge number of redundant record which makes evaluation result as biased, to improve quality of raw data, feature reduction and filtering is required which increase data efficiency. This paper proposed a novel simplified swarm optimization with Random forest classifier for pre-processing. On incorporating Random Forest (RF) with SSO it improves the performance accuracy of filtering phase than PSO-RF and other data mining classifiers. Therefore computational memory and processor utilization is analysed to show algorithm efficiency of proposed method.

The rest of the paper is structured as follows: section 2 present some related work based on KDD cup 99 dataset. Section 3 present an overview of proposed framework. Section 4 and 5 explain about data mining and swarm intelligence in intrusion detection. Section 6 explains in detail about proposed SSO-RF algorithm and its accuracy is compared with other data mining classifier. Section 7 concludes some result based on proposed work.

II. RELATED WORK ON KDD CUP99 DATASET

The raw network data has huge network traffic data size which leads to irrelevant and huge dimensionality problem. The KDD'99 [12] has been the most wildly used data set for the evaluation of anomaly detection methods is prepared by Stolfo et al, based on the data captured in DARPA'98 IDS evaluation program [15]. Agarwal and Joshi [13] proposed a two-stage general-to-specific framework for learning a rule-based model (PNrule) to learn classifier models on a data set

that has widely different class distributions in the training data. The proposed PN rule evaluated on KDD dataset reports high detection rate. Yeung and Chow [14] proposed a novelty detection approach using no-parametric density estimation based on Parzen-window estimators with Gaussian kernels to build an intrusion detection system using normal data. This novelty detection approach was employed to detect attack categories in the KDD dataset. In 2006, Xin Xu et al. [17] presented a framework for adaptive intrusion detection based on machine learning. Multi-class Support Vector Machines (SVMs) is applied to classifier construction in IDSs and the performance of SVMs is evaluated on the KDD99 dataset. In [18], Portnoy et al. partitioned the KDD data set into ten subsets, each containing approximately 490,000 instances or 10% of the data. However, they observed that the distribution of the attacks in the KDD data set is very uneven which made cross-validation very difficult for smurf and Neptune based attacks.

The biological inspired approaches have been extensively instigated in network intrusion pattern detection. For example, a new collaborating filtering technique for pre-processing the probe type of attacks is proposed by G. Sunil Kumar, [19] based on hybrid classifiers on binary particle swarm optimization and random forests algorithm for the classification of probe attacks in a network. Wei-Chang yeh et.al [20] proposed new method by combining SSO with weighted exchange local search method for intrusion detection. And Dharmendra G. Bhatti [21] proposed a method to reduce false positive rate using data pre-processing method.

III. OVERVIEW OF PROPOSED FRAMEWORK

The proposed framework based on analysis of KDD cup 99 dataset. Preprocessing based on data mining algorithm still leads to dimensionality problem in detecting intrusion and leads to low efficiency. To overcome these difficulties this paper proposed a new combination of simplified swarm intelligence technique incorporated in random forest algorithm which filters data more effectively and increase detection rate. Figure 1 shows the proposed framework

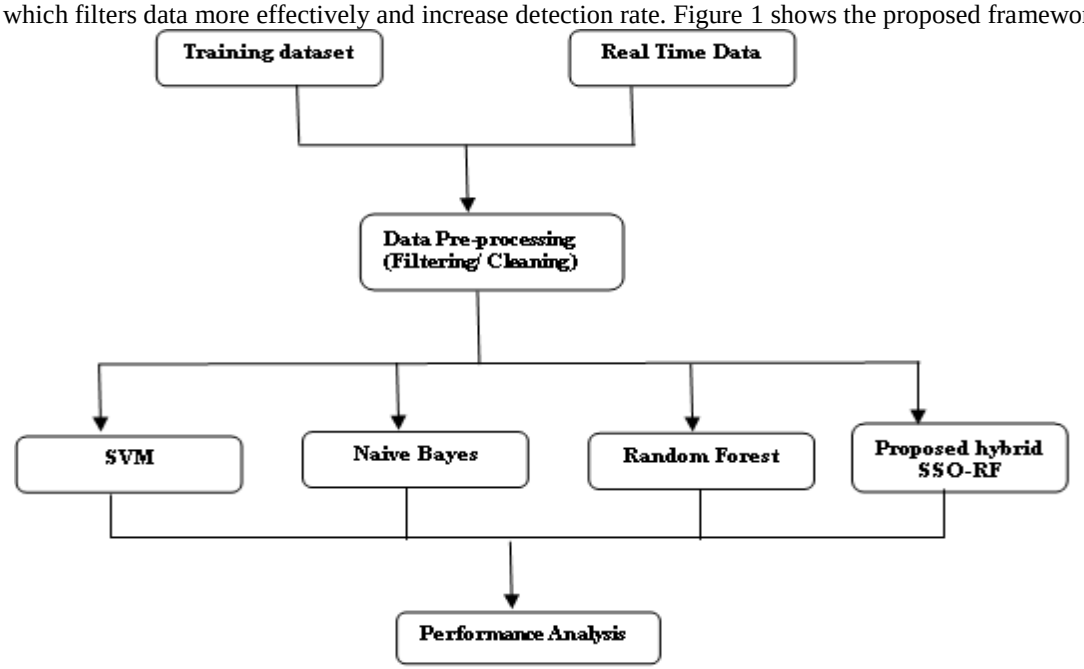

Fig. 1 Framework for Pre-Processing

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 [16]. 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).

- 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

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

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.

B. 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.

IV. DATA MINING TECHNOLOGY IN INTRUSION DETECTION

Data mining is the latest technology applied for intrusion detection [22]. The main advantage of data mining algorithm is to withdraw the need of unknown knowledge from the massive network data and from host log data. At present, data mining algorithm applied to intrusion detection mainly has four basic patterns: association, sequence, classification and clustering used to process large amount of data and to ignore hidden information [23]. Preprocessing plays a vital role in data mining techniques to reduce missing values and handles incomplete data. In this paper various data mining algorithms are used and compared with the proposed techniques. The analysis based on KDD cup 99 dataset with classification algorithm such as naive Bayes, SVM and Random Forest to filter raw data attributes. The classification algorithms are mainly used to collect audit data and to classify as normal or abnormal. It also detect individual attacks but leads to high false alarm rate.

V. SWARM INTELLIGENCE IN INTRUSION DETECTION

A swarm can be considered as a group of cooperating agents to achieve some purposeful behaviour and task [24]. It is introduced by Beni and Wang (1989) has received widespread attention in research. Now-a-days the Bio Inspired techniques listed in table2 are mainly used to solve optimization problem in intrusion and are also applicable to measure fitness function. The main strength is that they are parallel in nature [25].

Table II
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 [26]. 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.

VI. Proposed Hybrid Simplified Swarm Optimization With Random Forest Algorithm

The traditional pre-processing algorithms are not adaptive to the situations when kdd99 dataset is large, it may result in the false recommendations. The KDD dataset contains 41 attributes, in which the data may be incomplete, noisy or duplicate in nature. The proposed pre-processing approach filters data effectively and the result is compared with other existing data mining approach. The new proposed model namely simplified swarm optimization (SSO) is incorporated with random forest. SSO is a simplified version of Partial Swarm Optimization (PSO) and can be used to find the global

minimum of nonlinear functions [26]. This approach is used to solve classification problem and reduce dimensionality of dataset.

A. Random Forest

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 [27]:

- 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 importance 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 [28]. 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.

B. Simplified Swarm Optimization

Initially, the number of swarm population size, the number of maximum generation, and three parameters are determined [11]. 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).

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

Where i = 1; 2; : : : ;m, where m is the swarm population. $X_i = (x_{i1}; x_{i2}; xiD)$, 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 C_g 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 (pbest), and the best solution achieved so far by the whole swarm (gbest) 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 proposed SSO-RF algorithm 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: Choose m wide range of variables used to split each node. m<<M, where M is the number of input variables.

Step 5: In a growing tree at each and every node select m variables at random from M and bust them out to have the best split.

Step 6: This process will be repeated until the termination condition is satisfied.

The proposed SSO-RF method filter raw data and reduce incomplete, noisy and dimensionality problem for both discrete and continuous variables in dataset. This approach is significantly different from other research work which had combine only data mining and PSO. The proposed method produce high accuracy and produce near optimal solution for pre-processing phase.

VII. Experimental Result

To analyse the performance of the proposed method, the experimental evaluation is done on KDD cup99 training dataset and it compared with other existing data mining algorithms. In this paper only 10% of KDD cup dataset 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 cyber-attacks in the dataset is divided into three components which shown in table2. The comparison based on some popular machine learning techniques such as Support Vector Machine [32], Naïve Bayes [29, 30], and Random Forest [31] from Weka [33] collection to learn the overall behavior of the KDD'99 data set.

Table III
KDD cup99 dataset attacks in sample numbers

Dataset	Normal	Probe	DOS	U2R	R2l
10% KDD	97277	4107	391458	52	1126
Corrected KDD	60593	4166	229853	70	16347
Whole KDD	97270	41102	3883370	1126	52

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 behaviour of normal and intrusive data. The parameters are based on SSO-RF obtains best achievement from all data mining techniques.

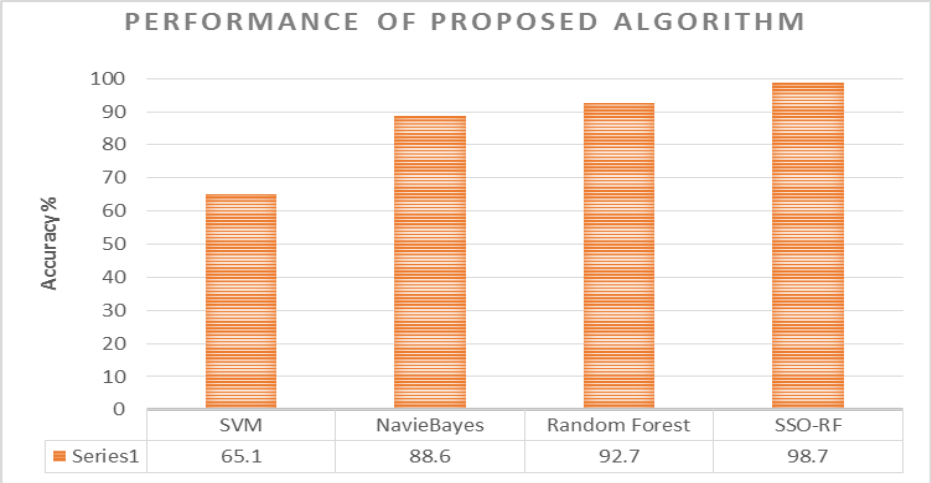

Fig. 2 Performance of Proposed Algorithm

VIII. Conclusion

The paper statistically analyzed the performance of complete KDD dataset and it proposed a new technique to overcome the difficulties. The SSO-RF method can easily reduce dimensionality and multiclass dataset problem easily than other existing data mining algorithms. The pre-processing can easily extract the most relevant feature subset form network traffic and record as normal or attack. It is clear that the proposed analysis filters normal record and reduce attribute list, thereby reducing the burden of the IDS in working with a large feature set. Therefore swarm intelligence techniques incorporated with data mining can effectively improves detection accuracy and produce optimal solution than other methods. 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-RF algorithm is faster in convergence and more efficient in solution. We still investigate new techniques for further improvement.

References

[1]. John, G.H., Langley, P: "Estimating Continuous Distributions in Bayesian Classifiers". *In Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence* 1995.

[2]. Quinlan, J.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo 1993.

[3]. Kohavi, R.: "Scaling up the accuracy of naïve-bayes classifier: A decision-tree hybrid*". In: Proc. Of the 2nd International Conference on Knowledge Discovery and Data Mining*, pp. 202–207. AAAI Press, Menlo Park 1996.

[4]. Witten, I.H., Frank, E.: Data Mining: "Practical Machine Learning Tools and Techniques", 2nd edn. Morgan Kaufmann, San Francisco 2005.

[5]. Werbos, P.: Beyond Regression: "New Tools for Prediction and Analysis in the Behavioral Sciences". PhD Thesis, Harvard University (1974)

[6]. Al-Tabtabai H, Alex PA. "Using genetic algorithms to solve optimization problems in construction". *Eng Constr Archit Manage* 1999; 6(2):121–32.

[7]. Dharmendra G. Bhatti, P. V. Virparia, Bankim Patel, "Conceptual Framework for Soft Computing based Intrusion Detection to Reduce False Positive Rate', *International Journal of Computer Applications* (0975 – 8887), Volume 44– No13, April 2012.

[8]. Dorigo M, Di Caro G. 1999. "The ant colony optimization meta-heuristic, new ideas in Optimization", *ACM Transaction*, 11-32.

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

[10]. 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.

[11]. 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.

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

[13]. Agarwal, R., Joshi, and M.V.: PNrule: "A New Framework for Learning Classifier Models in Data Mining". Tech. Report, Dept. of Computer Science, University of Minnesota 2000.

[14]. Yeung, D.Y., Chow, C.: Prazen-"window Network Intrusion Detectors". *In: 16th International Conference on Pattern Recognition*, Quebec, Canada, pp. 11–15 August 2002.

[15]. S. J. Stolfo, W. Fan, W. Lee, A. Prodromidis, and P. K. Chan, "Cost based modeling for fraud and intrusion detection: Results from the jam project," discex, vol. 02, p. 1130, 2000.

[16]. MIT Lincoln Labs, 1998 DARPA Intrusion Detection Evaluation. Available on: http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/ index.html, February 2008.

[17]. Xu, X.: "Adaptive Intrusion Detection Based on Machine Learning: Feature Extraction, Classifier Construction and Sequential Pattern Prediction". *International Journal of Web Services Practices* 2(1-2), 49–58 2006.

[18]. L. Portnoy, E. Eskin, and S. Stolfo, "Intrusion detection with unlabeled data using clustering," *Proceedings of ACM CSS Workshop on Data Mining Applied to Security*, Philadelphia, PA, November, 2001.

[19]. G. Sunil Kumar, C.V.K Sirisha, Kanaka Durga.R, A.Devi, "Robust Preprocessing and Random Forests Technique for Network Probe Anomaly Detection", *International Journal of Soft Computing and Engineering (IJSCE)*, ISSN: 2231-2307, Volume-1, Issue-6, and January 2012.

[20]. 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 @ 201*2 issn 1349-4198. Volume 8, number 6, June 2012.

[21]. Dharmendra G. Bhatti, P. V. Virparia, Bankim Patel, "Conceptual Framework for Soft Computing based Intrusion Detection to Reduce False Positive Rate*", International Journal of Computer Applications* (0975 – 8887), Volume 44– No13, April 2012.

[22]. Ian H. Witten, Eibe Frank "Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)", Morgan Kaufmann June 2005, 525 pages Paper ISBN 0-12-088407-0.

[23]. C. Romero, S. Ventura and E. García, "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, Volume 51, Issue 1, pp. 368-384, 2008, Elsevier Science.

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

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

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

[27]. Bosh, A., Zisserman, A., Munoz, and X.: "Image classification using Random Forests and ferns". *In: IEEE ICCV* 2007.

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

[29]. G. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," *in Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 338–345, 1995.

[30]. R. Kohavi, "Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid," *in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, vol. 7, 1996.

[31]. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[32]. C. Chang and C. Lin, "LIBSVM: a library for support vector machines," 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[33]. "Waikato environment for knowledge analysis (weka) version 3.5.7." Available on: http://www.cs.waikato.ac.nz/ml/weka/, June, 2008.

