Journal of Theoretical and Applied Information Technology
20th January 2013. Vol. 47 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ENHANCED RULE APPROACH FOR NETWORK INTRUSION DETECTION USING EFFICIENT DATA ADAPTED DECISION TREE ALGORITHM

[1]G.V.NADIAMMAI and [2*]M.HEMALATHA

[1, 2*]Department of Computer Science, Karpagam University, Coimbatore

E-mail: [1]gvnadisri@gmail.com, [2]csresearchhema@gmail.com

ABSTRACT

Data mining has been used extensively and broadly by several network organizations. Intrusion Detection is one of the high priorities & the challenging tasks for network administrators & security experts. Intrusion detection system is employed to protect the data integrity, confidentiality and system availability from attacks. IDS use the data mining techniques to analyze the resources from the database over a network. It is also necessary to develop a robust algorithm to generate effective rules for detecting the attacks. In this paper, Classification based optimization algorithms have been used to detect attacks over KDD CUP 99 dataset. Based on this dependency, an improved Efficient Data Adapted Decision Tree algorithm is proposed to overcome the drawbacks of the existing algorithm. The experimental results clearly show that the proposed EDADT algorithm achieved higher accuracy, less alarm rate & capable of detecting new type of attack efficiently**.**

Keywords:- *Machine Learning, Data Mining, Intrusion Detection, Classification Algorithms, Efficient Data Adapted Decision Tree (EDADT) Algorithm, Optimization Technique, KDD CUP 99 Dataset.*

1. INTRODUCTION

Nowadays, the usage of network is growing rapidly and as well as it provides information rapidly. Likewise security violation like hackers, viruses, worms etc., also spread even faster across the network. Since the Intruder can compromise the confidentiality, integrity and availability of network resources. For the past few years, firewall acts as a gate wall for security measures. But it fails to detect the network traffic that has been done by the specific port or from the legitimate user port. So intrusion detection [1] is necessary to handle the hackers from exploiting the data. Intrusion detection is performed to analyze and monitor the activities done by the individual host or by the network. Our main motivation is to safeguard the system from threats. It is possible only through Intrusion Detection System. An Intrusion Detection System was first coined by Anderson (1980) [2] in a technical report. Surveillance system acts as an effective tool against threat and attacks. It monitors the behavior of the user & detects corresponding masqueraders who accessed the system dishonestly. The three important steps in intrusion detection system is,

- Monitor and examine network traffic
- If any exists, analyze the abnormal activities
- Raising alarms to handle the situation

Intrusion detection system is a passive method. It just monitors the information over network or hosts and raises alarms when any intrusion happens. But data mining based ids can identify these data when it arrives and forecast it on its own, thus by gaining the function of active approach. Data mining has been popularly recognized as an important way to analyze useful information from large volumes of data that are noisy, fuzzy & random [3]. Extracted patterns can be used to improve the business activities like sales, marketing and customer management. IDS are of two types namely Host and Network based IDS. In HIDS [4] the data come from audit record, system logs, application program etc, by comparing with network IDS to analyze network attack or an intrusion happened to particular hosts. Whereas the encrypted packets passes over the network from the system files and then decrypted in host machine. So the data are not affected and it does not require any special kind of hardware than monitoring system installed in specific host. In network based ids [5] commonly one Intrusion Detection System is enough for the whole LAN. It is of low cost & capable of analyzing many attacks like DoS, DDoS, etc., but HIDS fails to analyze those attacks.

Intrusion detection system has traditionally been classified into two classes namely anomaly detection and misuse/signature based detection.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Misuse detection compares the upcoming network traffic to the database of known attack with the help of signatures to detect intrusions. It works efficiently in analyzing known attacks that are stored in the database. But it cannot detect new attacks that are not predefined. On the other hand, the anomaly detection approach creates a profile (normal) based on the network and hosts under inspection & raises alarms or some kind of notification to make the administrator to handle the situation. However they have being able to detect new & unusual attacks. There are two types of false alarms in determining the any deviations from normal pattern false positive and false negative. The main goal is to keep these alarms as low as possible. Data mining techniques such as association, classification, clustering and neural networks have been used in intrusion detection [6, 7].

This paper is organized as follows. Section 2 provides related work dealt with a data mining approach in intrusion detection. Section 3 includes contribution of the work. Sections 4 explain the data set used and its features in detail. Section 4 compares the classification algorithm with the optimization techniques and based on this, an algorithm is proposed. Sections 5 explain the architecture of the proposed work. Section 6 is about performance evaluation. Section 7 includes result & discussion based on the experiment. Section 8 refers to conclusion & future enhancement.

2. RELATED WORKS

M. Dorigo et al [8] proposed an meta heuristic approach to solve various optimization problems. The Meta heuristic method is robust and versatile in nature. Based on the quantity of the pheromone, ant chooses one or more path. But however it chooses the shortest path to attain the optimal solution for the given problem. In [9] author analyses the properties of both ACO and PSO and stated that this combination overcomes the limitation by improving the efficiency in the application where it has been used. Through updating the particles global best solution is found and not limited to current one.

Rafel et al [10] evaluates the best fit particles to the next candidate particle. Taking average of all candidate particles of one generation and checking whether it is better than its predecessors. This process continues until all the values are involved. Nicholas Holden et al [11] uses hybrid algorithm uses nominal attributes rather than converted into binary numbers in the preprocessing stage. It directly deals with both continuous and nominal. We compared PSO/ACO 2 with PART but PSO/ACO 2 involves a smaller set of rules than PART. The goal of this paper is to achieve more accurate and useful to the user.

Fatima Avd Jani [12] combined PSO with SVM to achieve greater classification in terms of accuracy but computation time is increased .The proposed algorithm optimizes the performance of SVM classifier. Jun Wang et al [13] used Hybrid PSO-SVM approach which is the combination of standard PSO (SPSO) and Binary PSO (BPSO) to achieve a higher detection rate than regular SVM in the same time. SPSO obtains parameters of SVM and BPSO extracts the feature as a subset of intrusion detection system. These two are useful to train data set with the SVM Classifier model. Hybrid PSO-SVM provides higher detection rate than the existing standard algorithm.

Praveen Kumar Rao et al [14] proposed a boosting ant colony algorithm to produce a reliable intrusion detection system. Boosting algorithm extracts a set of rules for classification of attacks and normal features from the network data set. In [15] the author, combined pure SVM with pure ACO and verified their functioning. From the experiments he found that CSOACN works well than existing SVM in terms of detection rate and as well as lower false negative and positive and false alarm rates.

Janakiraman et al [16] implemented the ant colony algorithm in distributed intrusion detection system to discover the intrusion in the distributed network environment. Total CPU usage time is less in ANT based DIDS than normal DIDS.

Dalila Boughaci et al [17] evaluates the effectiveness of various algorithms like FCS, EFS, C 4.5, EFRID, NB etc and also implemented a new approach FGA and conclude that FGA perform well in all classes like dos, u2r, r2l, probe and normal. Nada M.A.L. Salami et al [18] proposed a hybrid approach by combining ant colony with genetic programming to solve the optimization problem. However, ant colony optimization technique deploys genetic operations to form the best solution state. Each ant constructs a solution to a problem and fails to adapt by itself. Here the

Journal of Theoretical and Applied Information Technology
20th January 2013. Vol. 47 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

genetic programming helps to attain the optimal solution.

The main contributions of this paper is to examine the older and well known problem of intrusion over network and its solutions to present the applicability of the solutions by discovering the improved rule approach through EDADT algorithm within the context of network intrusion detection systems. EDADT algorithm reduces the input space of dataset with the regular interval of time, greater accuracy and less false alarm rate by classifying the data into attack and normal. It also shows that the EDADT algorithm is of high concern to the network intrusion detection area because of its performance characteristics in terms of feature reduction using information gain. We discuss this further in section 7.

3. KDD CUP 99 DATASET

The 1998 DARPA Intrusion Detection Evaluation Program was prepared and managed by MIT Lincoln Labs. The objective was to survey and evaluate research in intrusion detection. A standard set of data includes a wide variety of intrusions simulated in a military network environment [19]. The DARPA 1998 dataset includes training data with seven weeks of network traffic and two weeks of testing data providing two million connection records. A connection is a sequence of TCP packets starting and ending at some well defined times, between source IP address to a target IP address with some well defined protocol. Each connection is categorized as normal, or as an attack, with one specific attack type. The training dataset is classified into five subsets namely Denial of service attack, Remote to Local attack, User to Root attack, Probe attacks and normal data. Each record [20] is categorized as normal or attack, with exactly one particular attack type. They are classified as follows,

- ***Denial of Service Attack*** - Here the attacker makes the traffic busy and access the normal user system and performs all sorts of vulnerability.
- ***Probe Attack –*** the attacker gains the knowledge of the network and performs damage in future.
- ***Remote to Local Attack*** - the attacker uses the remote machine and causes some attacks to the local host machine.
- ***User to Root Attack*** - using the local machine through sniffing password the attacker exploits damages to the remote machine.

Table 1: Name Of The Attacks Classified Under 4 Groups

Denial of Service	Back, land, neptune, pod, smurf, teardrop
Probes	Satan, ipsweep, nmap, port sweep
Remote to Local	ftp_write, , imap, guess_passwd , phf, spy, warezclient, multihop, warezmaster
User to Root	Buffer_overflow, load module, Perl, root kit

4. ARCHITECTURE OF PROPOSED WORK

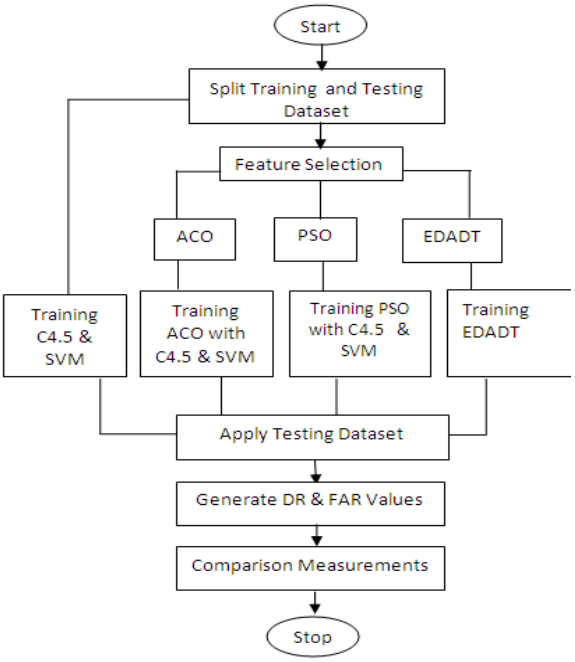

Figure 1: Work Flow of the Proposed EDADT Algorithm

According to the above figure1generate input from KDDCup99 dataset to compare the performance of various existing algorithm used in the intrusion detection systems. Apply information gain to increase the interpretability and improving accuracy of intrusion detection system only through the extraction of relevant attributes from the dataset. Then, select the training and testing dataset. Choose one of the classification methods such as C4.5, SVM, C4.5+ACO, SVM+ACO, C4.5+PSO, SVM+PSO, and Improved EDADT to classify the dataset as normal or malicious and train the model respectively. Apply it on the test dataset to evaluate detection rate (DR) and false alarm rate (FAR) values. Hence, compare the values obtained in previous step and finally represent the results by measure of performance of the model.

Journal of Theoretical and Applied Information Technology
20th January 2013. Vol. 47 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. PROPOSED EFFICIENT DATA ADAPTED DECISION TREE ALGORITHM (EDADT)

Since many data mining algorithms are available this is an innovative and efficient work towards IDS to detect the ongoing attacks over large network. This algorithm reduces the space occupied by the dataset. So it would be useful for the network administrator/manager to avoid the delay between the arrival and detection time of the attacks respectively. It also produces less false alarm rate and computation time in real time. The steps involved in this algorithm are,

1. Apply PSO technique and extract the efficient features for the given training dataset
2. Return the best solution obtained
3. Apply the best Features to the ACO as Input
4. Initialize the pheromone to obtain the optimal solution
5. Identify the local and global best values
6. Update the pheromone trail and calculate the average value
7. Return the best solution obtained
8. For each attribute a, select all unique values of a
9. Find the unique values belong to the same class label
10. If n unique values belong to the same class label
11. Split them into m intervals, and m must be less than n else
12. If the unique values belong to different c class label
13. Calculate the probability of the value belongs to each class
14. Change the class label of values with the class label with highest probability
15. Split the unique values as c interval then repeat step 10 for all values in the dataset.
16. Find the normalized information gain for each attribute
17. Create a decision node that splits on a best attribute with the highest normalized information gain
18. Recurse on the sub lists obtained by splitting on best attributes, and add those nodes as children of node
19. Repeat the process until the dataset converges
20. At last, train the EDADT model.

5.1 PSEUDOCODE OF EDADT ALGORITHM

To find local best
for each particle i = 1… n do:
for each particle dimension j=1… m do:
x_{ij}= particle position
t= number of iterations
r= random number, where r_1 and r_2 are
random number 1 and 2
b_{best}= fitness_lbest, g_{best}= fitness_gbest
x_f = fitness_x
x_b_{best} = s
x_g_{best} = p
x_{ij}= ll+ (ul-ll) * r
v_{ij} = 0
g_{best}= value
b_{best} = value
end
end
for i=1… t do:
x_f= evaluate_fitness(x)
if (x_i< $b_{best(i)}$)
b_{best} (i) = x_f(i)
s[i, j] = x(i,j)
end if
end
find global best
[min_fitness, min_fitness_index]= min (fitness_x[i])
if(min_fitness < g_{best})
g_{best}= min_fitness
for j=1…n do:
p[j] = x (min_fitness_index, j)
end
end if
particle velocity & position update
v_{ij}= w*v_{ij}+c_1*r_1+ c_1*r_1*(s - x_{ij}) + c2*r_2*(p[j] - x_{ij})
x_{ij}=x_{ij}+ v_{ij}
end
end
To find pheromone search
a= each ant completed a solution
t_l= tabu list, nm= next state to move
τ= base attractiveness, g= global best, l=local best
p_i= particle's best position (i=1, …, n)
for i < IterationMax do:
for each ant do:
choose probabilistically nm into

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
46. for each ant tj+ nm
47. repeat until a
48. end
49. for each a do:
50. update τ = a
51. end
52. if(pi<g)
53. save g=pi
54. best solution g
55. end
56. end
57. To find best attribute using information gain
58. att =attribute
59. n= set of unique values
60. m= regular _intervals
61. c= class label, where c1, c2 and cn are same,
62. different class label & child node
63. dn= decision node
64. ba = best attribute
65. if(n ε c1) then
66. split m
67. else if(n ε c2)
68. highest probability ε c
69. c= c (highest probability)
70. split c
71. update split m
72. m<n
73. end
74. until termination condition is met
75. end
76. dn = att+ highest normalized information gain
77. recurse ba
78. Cn=ba
79. repeat
80. until all ba found
81. end
```

6. PERFORMANCE EVALUATION

6.1. COMPARISON CRITERIA

The performance of classifiers involves Accuracy, Sensitivity, Specificity, computational time FAR and Receiver Operating Characteristics Curve (ROC). The accuracy, sensitivity and specificity were estimated by True Positive measure, False Positive measure, False Negative measure and True Negative measure [21] which as follows,

- Accuracy is the total number of detected attacks among all the other attack data

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN} \quad (1)$$

- Sensitivity is the possibility that the algorithms can easily predict positive instances.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2)$$

- Specificity is the possibility that the algorithms can easily predict negative instances.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

6.2 CONFUSION MATRIX

Confusion matrix is a visualized tool typically used in classification method and it is said to be a matching matrix in clustering method. A confusion matrix summarizes the number of instances predicted correctly or incorrectly by a classification model. In other words, a confusion matrix [22] is a static table for visualizing the performance of the algorithms.

Table 2: Standard Metrics For Intrusion Evaluation

Confusion Matrix (Standard Metrics)		Predicted Connection Label	
		Normal	Intrusion
Actual Connection Label	Normal	True Negative (TN)	False Alarm (FP)
	Intrusion	False Negative (FN)	Correctly Detected(TP)

- False Positive (FP)*:* Refers to the number of detected attacks but it is in fact normal or false alarm.
- False Negative (FN)*:* Relates to the number of normal instances but it is in fact an attack.
- True Positive (TP)*:* Relates to the number of detected attacks but it is in fact an attack.
- True Negative (TN)*:* Relates to the number of normal instances but it is in fact normal.

6.3 FALSE ALARM RATE

False Alarm Rate specifies the total of normal data that are mistakenly taken as attack.

$$FAR = \frac{FP}{FP + TN} \times 100$$

7. RESULTS AND DISCUSSION

Applying data mining technology to intrusion detection systems, it can mine the features of new and unknown attacks well, which is a maximal help to the dynamic defense of intrusion detection

Journal of Theoretical and Applied Information Technology
20th January 2013. Vol. 47 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

system. This work is performed using Machine learning tool with 5000 records of KDD Cup 99 dataset to analyze the effectiveness between our proposed method and the traditional algorithms. The performance of the various algorithms measured in terms of accuracy, Sensitivity, Specificity and false alarm rate. Table 1 represents the accuracy, sensitivity and specificity values for C4.5, SVM, C4.5+ACO, SVM+ACO, C4.5+PSO, SVM+PSO, and SVM and Improved EDADT algorithms. Here the ROC curve is a graphical plot of sensitivity, specificity for the attributes.

Based on values obtained, the accuracy of C4.5 is 93.23%, the accuracy of SVM is 87.18%, the accuracy of C4.5+ACO is 95.06%, the accuracy of SVM+ACO is 90.82%, the accuracy of C4.5+PSO is 95.37%, the accuracy of SVM+PSO is 91.57% and the accuracy of Improved EDADT is 98.12%. Finally, an Improved EDADT took highest accuracy percentage when compared to all six classification based algorithms. Fig. 2 specifies the corresponding chart for the result obtained in table 3.

Table 3: Comparison Based On Accuracy, Sensitivity, Specificity And FAR Values

Algorithms	Sensitivity (%)	Specificity (%)	Accuracy (%)	FAR (%)
C4.5	86.57	82	93.23	1.56
SVM	83.82	64.29	87.18	3.2
C4.5+ACO	88.26	84.42	95.06	0.87
SVM+ACO	87.42	67.95	90.82	2.42
C4.5+PSO	92.51	88.39	95.37	0.72
SVM+PSO	89.06	70.80	91.57	1.94
Proposed EDADT	95.86	89.36	98.12	0.18

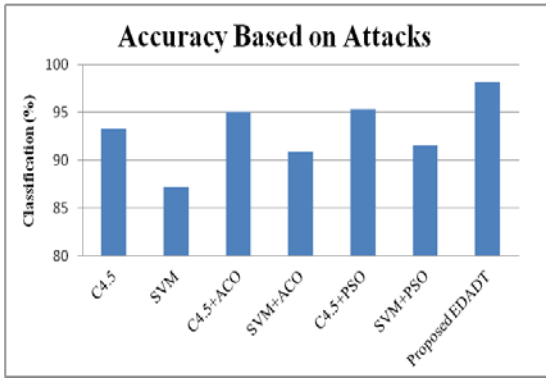

Figure 2: Chart Showing The Results Obtained By The Existing And Proposed Algorithm

Figure 3, illustrates the build time of C4.5, SVM, C4.5+ACO, SVM+ACO, C4.5+PSO, SVM+PSO and Improved EDADT algorithms. C4.5+PSO take more time to build the model. SVM + ACO take less time than the proposed algorithm but provide less accuracy percentage than the other. However the improved EDADT takes less time when compared to C4.5+PSO and SVM +PSO & provides better accuracy in terms of all existing algorithms.

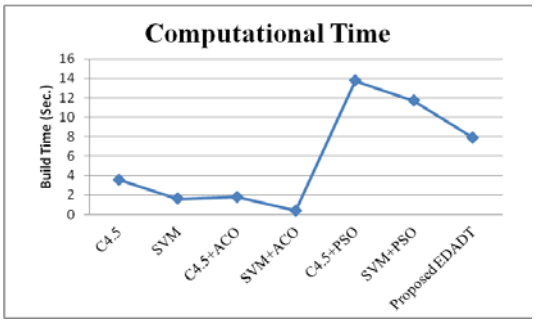

Figure 3: Computational Time Taken For The Existing And Improved EDADT Algorithm

In the below figure 4, the Receiver Operating Characteristic (ROC) curve is used to measure the performance of algorithms in terms of False Alarm Rate (FAR) & Detection Rate (DR). According to the graph, the parameter for the X axis is the False Positive Rate (Specificity) and the Y axis takes the detection Rate (Sensitivity) in fractions respectively.

While comparing the existing algorithm, the proposed Efficient Data Adapted Decision Tree algorithm shows the Sensitivity is 96.86%, Specificity is 92.36% and False Alarm Rate is 0.18 %, followed by C4.5+ PSO with 92.51 % sensitivity, 88.39% specificity and False Alarm Rate is 0.72% respectively. Thus, a proposed EDADT algorithm effectively detects the attack with less computational time and False Alarm Rate.

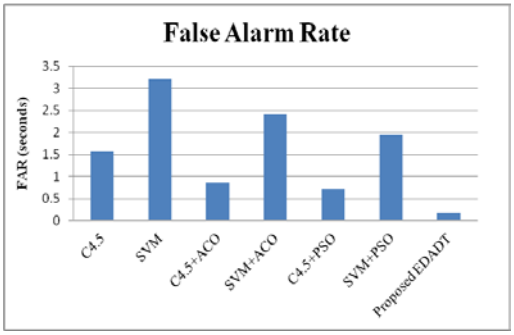

Figure 4: False Alarm Rate

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

8. CONCLUSION & FUTURE SCOPE

In this paper, a novel Efficient Data Adapted Decision Tree algorithm for NIDS has been proposed for intrusion detection system. NIDS monitors packets and status from network and application layers. The results of experiments on KDD Cup data sets demonstrate that the improved EDADT algorithm can much quickly discover classification rules which are roughly competitive in rule predicative accuracy and simplicity. The proposed EDADT algorithm is 19.4% better than C4.5, 18.8% better than SVM, 19.6% better than C4.5+ACO, 19.2% better than SVM+ACO, 19.7% better than C4.5+ PSO, 19.3% better than SVM + PSO in terms of accuracy. Thus the proposed EDADT algorithm reduces the actual size of the dataset and helps the administrator to analyze the ongoing attacks efficiently with less false alarm rate respectively.

In future, a hybrid intrusion detection system can be developed based on data mining algorithms which would be fast and robust in identifying the huge variety of new and unusual attacks.

ACKNOWLEDGMENT

We thank the Karpagam University for providing motivation, encouragement and support to complete this research work

REFERENCES:

[1] The Intrusion-Detection [online]. Available from: http://en.wikipedia.org /wiki/intrusion-detection [last cited on 2012 June 15].

[2] Anderson. J.P, "Computer Security Threat Monitoring & surveillance "*Technical Report*, James P Anderson Co., Fort Washington, Pennsylvania, 1980.

[3] Jiawei Han and Kamber," Data Mining: Concepts and Techniques", 2nd Edition, Morgan Kaufman Publishers, Elsevier Inc, 2006.

[4] Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P., Srivastava, J., Kumar, et al," The MINDS-Minnesota intrusion detection system", *Next generation data mining, MIT Press*, 2009.

[5] Bace, Rebecca G, NIST, special publication on "intrusion detection systems", 2002.

[6] Ben Sujatha. B, Kavitha .V, "Survey on intrusion detection approaches", *International Journal of Advanced Research in Computer Science*, vol. 3, no. 1, pp.363-371, 2012.

[7] Alok Ranjan, Ravindra S. Hegadi, "Emerging Trends in Data Mining for Intrusion Detection", International *Journal of Advanced Research in Computer Science*, vol. 3, no. 2, pp.279-281, 2012.

[8] Marco Dorigo and Gianni Di Caro, "Ant Colony Optimization: A New Meta-Heuristic," New York: McGraw-Hill, pp. 11–32, 1999.

[9] Selvi.V, Umarani R, "Comparative Analysis of Ant Colony and Particle Swarm Optimization Techniques", *International Journal of Computer Applications*, vol.5, no.4, pp.2010.

[10] Rafael S. Parpinelli, Heitor S. Lopes, and Alex A. Freitas, "Data Mining With an Ant Colony Optimization Algorithm", *IEEE Transactions on Evolutionary Computing*, vol.6, no. 4, pp.2002.

[11] Nicholas Holden and Alex A. Freitas,"A Hybrid PSO/ACO Algorithm for Discovering Classification Rules in Data Mining", *Journal of Artificial Evolution and Applications, Hindawi Publishing Corporation*, pp.1-11, 2008.

[12] Fatima Ardjani, Kaddour Sadouni, "Optimization of SVM Multiclass by Particle Swarm (PSO-SVM)", I.J. Modern Education and Computer Science, no.2, pp. 32-38, 2010.

[13] Jun Wang, Xu Hong, Rong-Rong Ren, Tai-hang Li, "A Real-time Intrusion Detection System Based on PSO-SVM ", Proceedings of *International Workshop on Information Security and Application* Qingdao, China, November 21-22, pp.319-321, 2009.

[14] Praveen Kumar. K, kamakshi. P, "Ant colony optimization algorithm for computer intrusion detection", 2006.

[15] Qinglei Zhang, Wenying Feng, "Network intrusion detection by support vectors and ant colony", *Proceedings of the International Workshop on Information Security and Application*, Qingdao, China, November 21-22, pp.639-642, 2009.

[16] Janakiraman.S, vasudevan.V, "ACO based Distributed Intrusion Detection System", *International Journal of Digital Content Technology and its Applications*, vol.3, No.1, pp.66-72, 2009.

[17] Dalila Boughaci, Mohammed Lamine Herkat, Mohamed Amine Lazzazl," A Specific Fuzzy Genetic Algorithm for Intrusion Detection", *Second International Conference on Communication and Information Technology*, pp.6-11, 2012.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[18] Nada M.A. AL Salami "Ant Colony Optimization Algorithm", *UBICC Journal*, vol. 4, no.3, pp- 823-826, 2009.

[19] Stolfo. S.J, Fan. W, Lee. W, Prodromidis. A, and Chan. P. K, "Cost based modeling for fraud and intrusion detection: Results from the JAM project", *proceedings of DARPA Information Survivability Conference and Exposition*, vol.2, pp.130 -144, 2000.

[20] KDD Cup 99 intrusion Detection Dataset. Available from: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html[last cited on 2012 June 15].

[21] Sethuramalingam.S, "HYBRID FEATURE SELECTION FOR NETWORK INTRUSION", *International Journal of Computer Science and Engineering*, vol.3, no.5, pp.1773-1779, 2011.

[22] The Wikipedia Website [online]. Available from: http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html[last cited on 2012 June 15].