Intrusion Feature Selection Using Modified Heuristic Greedy Algorithm of Itemset

Janya Onpans[1], Suwanna Rasmequan[2],
Benchaporn Jantarakongkul[3], Krisana Chinnasarn[4]
Faculty of Informatics, Burapha University
Chonburi, Thailand
[1]mai.janya@gmail.com, [2]rsuwanna@buu.ac.th,
[3]benchapornj@yahoo.com, [4]krisana@buu.ac.th

Annupan Rodtook
Department of Computer Science
Faculty of Science, Ramkhamhaeng University
Bangkok, Thailand
annupan@ru.ac.th

***Abstract*— This paper proposes the Modified Heuristic Greedy Algorithm of Itemset (MHGIS) as a feature selection method for Network Intrusion Data. The proposed method can be use as an alternative method to gain the proper attributes for the proposed domain data: Network Intrusion Data. MHGIS is modified from original Heuristic Greedy Algorithm of Itemset (HGIS) to increase efficiency for finding proper feature. In our work, we compare our result with the common method of feature selection is which the Chi-Square (Chi^2) feature selection. There are 4 main steps in our experiment: Firstly, we start with data pre-processing to discard unnecessary attributes. Secondly, MHGIS feature selection and Chi^2 feature selection have been employed on the pre-processed data, to reduce the number of attributes. Thirdly, we measure the recognition performance by using supervised learning algorithms which are C4.5, BPNN, RBF and SVM. Lastly, we evaluate the results received from MHGIS and Chi^2. From the KDDCup99 dataset, we got 13,499 randomly sampling patterns with 34 data dimensions. With the use of MHGIS and Chi^2 algorithms, we obtain 14 and 26 features respectively. The result shows that, the classification accuracies measure by C4.5 over the MHGIS selection algorithm produces better accuracies as compare to the Chi^2 feature selection and HGIS feature selection over all types of classification methods.**

Keywords—Feature Selection; Pattern Recognition; Network Intrusion Detection; Heuristc Greedy

I. INTRODUCTION

With the growing demands of computer networks made many people realize more about the significant of the security in networks. Intrusion detection is a preferable choice for most people to use in computer network security. Intrusion detection system divides into two types: misuse detection and anomaly detection. Misuse detection system is a method which is use for detecting abnormal pattern by comparing them with well-known pattern. This made it unable to detect any unknown attack that has no matched pattern found in the system. In contrast, anomaly detection system is a method that detect deviated pattern from the normal behavior. So that it can detect any new intrusion behavior. However, this method will depend pretty much on the structure design. If the structure is not well designed, then some types of attack may not be detected.

Intrusion detection system can be seen as a classification problem in which it will distinguish the attracted activities from the normal activities. Data that transmit via network is very large; as a result, it will cause the delay the indentifying of intrusion and may allow any intruder to attack the network. This problem of enormous data that may slow down the detection process needs a method to eliminate useless features of the network data to be able to increase accuracy of classification while speeding up intrusion detection process. Furthermore, a good detection system should have higher detection rate and lower false alarm. The methods that can be use to solve these problems are either feature selection or feature extraction. Feature selection is a method that will produce a subset of features whereas feature extraction will create new features. Both of these methods will discard irrelevant or redundant features, so that only essential data will be left for further processes.

It has been known that finding proper representative of data attributes is very important to the intrusion detection system development. In other word, good data representations provide higher or better degree of recognition. In our previous work [3] we have proposed HGIS to compare with feature extraction method: PCA. The result shows an improvement of 5.05% with RBF. In this work we propose MHGIS to improve the result as compare to the feature selection: Chi^2. MHGIS is a method for select feature which considers error information of each itemsets in order to give the best answer in each iteration. Itemsets used in each iteration is constructed from apriori algorithm. MHGIS method divides into two steps: (1) finding based itemset and (2) adding discard item. With this method, the discard item will be included back to reconsider. This made an extensive cover of the itemsets. In classification step, selected features from MHGIS are compared with selected features from the Chi^2 algorithm. We verify the performance of our proposed feature selection, by calibrating with the other four classic classification methods: C4.5 decision tree, back-propagation neural network (BPNN), radial basis function (RBF) network and support vector machine (SVM). Performance metrics used in this paper are accuracy rate, detection rate, false-alarm rate and CPU processing time.

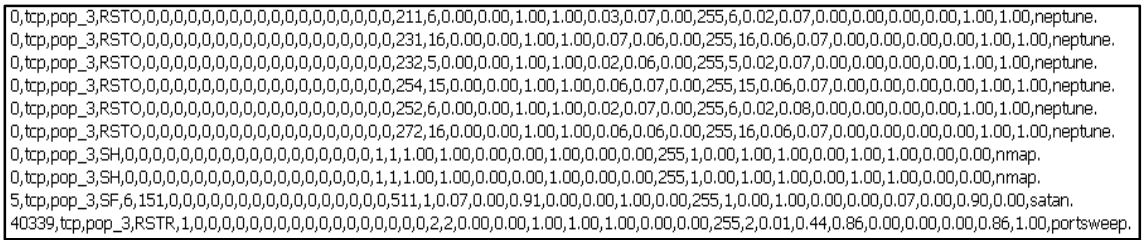

Fig. 1. Example of KDD Cup 99 Dataset

The paper is organized as follows: Related works are discussed in section II. Theoretical Background is discussed in section III. Proposed Method is presented in section IV. Experimental results are discussed in section V. Finally, the conclusion is summarized in the last section.

II. RELATED WORKS

In recent years, there is still a number of research works that try to improve the accuracy rate. The main stream of these works is focused on reducing the insignificant features as much as possible to be able to increase performance of rapid detection. So in this section, we will discuss related research works that are attempted to find proper features.

Murat Karabatake et al. [1] proposed a new feature selection base on association rules and neural network for the diagnosis of erythemato-squarmous diseases. The dimensions of input features space are reduced from 34 to 24 by using association rules. The correct classification rate of proposed system is 98.61%.

Hari Om et al. [2] proposed a hybrid system for reduction the false alarm rate of intrusion detection system. Feature selection use entropy and combines k-Means and two classifiers: K-nearest neighbor and Naïve Bayes for anomaly detection. Combination of tree classifiers, the detection rate reaches 98.18% and false alarm rate 0.83%.

Janya Onpans et al. [3] proposed the feature selection and extraction methods of network intrusion data which are the heuristic greedy algorithm of itemset (HGIS) and principal component analysis (PCA), respectively. After HGIS feature selection and PCA feature extraction steps that it obtain 19 and 13 features, respectively, and then use three standard supervised learning algorithms which are BPNN, RBF and SVM for evaluating with KDDCup99 dataset. Experimental results shown that HGIS algorithm produces better features than the PCA.

III. THEORETICAL BACKGROUND

In this section, we describe the essential background that use in the proposed method.

A. KDD Cup 99 Dataset

The Knowledge Discovery in Database (KDD) Cup data is a common benchmark for evaluation of intrusion detection system [4]. This dataset is prepared by MIT Lincoln Labs to simulate the attack to U.S. Air Force local area network. There are about 4,900,000 records. Each connection is labeled as either normal or attack that consists 41 features. Features are in form of continuous, discrete, and symbolic are shows in Fig. 1 and it divides into three categories:

- Basic Features: are fundamental features can be derived from packet data communication network such as protocol type.
- Traffic Features: are features that compute with respect to a window interval such as time in connect.
- Content Features: are features to be able to look for suspicious behavior in the data portion such as the number of failed login attempt.

The last feature is labeled as either normal or attack. Attack type that can be classified in four main categories [5]:

- DoS (Denial of Service): is class of attack that attacker makes memory resource too busy to accept legitimate request.
- U2R (User to Root): is class of attack that attacker starts with access to a normal user account and exploit some vulnerability to gain root access to the system.
- R2L (Remote to Local): is class of attack that attacker sends packets to a machine over a network, then exploit some vulnerability to gain local access to a machine.
- Probing: is class of attack that attacker attempts to gather information about a network and find the known vulnerabilities for the purpose of circumventing its security controls.

B. Heuristic Greedy Algorithm

Heuristic Greedy Algorithm is a simple and straightforward method for finding solutions such as finding the shortest path in graph, finding solution in dynamic programming problem, and so on [6]. It considers error information from current learning step in order to give the best answer at the time. It can be used for finding optimal solutions from very large database or big dataset because it does not consider all of data attributes. It is an iterative method which starts from finding the best solution in the first iteration. Then, the first solution is used for finding the best solution for the second iteration. It will be repeated until the optimal solutions are converged. Solutions from the first iteration, second iteration, third iteration, and so on are called 1-itemsets, 2-itemsets, 3-itemsets, and n-itemsets, respectively.

Heuristic Greedy Algorithm of itemset is feature selection technique by using apriori as following [3]:

Step1: generate 1-itemsets; find support value as in (1) of each item with RBF.

$$f(\{itemset\}) = rmse(\{itemset\}) \quad (1)$$

Where $rmse$ is root mean square error.

Step 2: generate 2-candidate itemsets; find support value as equation (1) with RBF of each pair 1-itemset.

Step 3: generate 2-itemset by union of 2-candidate itemsets that have support value less than or equal to 1-itemset subsets of the 2-candidate itemsets as in (2).

$$f(\{a,b\}) \leq f(\{a\}) \; and \; f(\{a,b\}) \leq f(\{b\}) \quad (2)$$

Step 4: repeat step 2 and increase size of itemsets until cannot generate itemsets.

Step 5: Select the best itemset that has lowest root mean square error ($rmse$).

C. Itemset Creation using Apriori Algorithm

Apriori Algorithm is the most commonly used in finding association rules between data attributes. It has been known that it works iteratively. First, it finds the set of attributes of size 1-itemsets. Then, set of attributes of size 1-itemsets are used as the base for finding set of attributes of size 2-itemsets. Next, set of attributes of size 3-itemsets will be calculated based on set of attributes of size 2-itemsets. The algorithm will repeat until set of attributes of size n-itemsets could not be computed. As mention before, it can be said that Apriori algorithm is a simple technique but powerful method for creating smaller candidate subset from very large sets which were found from the previous iteration. In addition, it can be used for eliminating infrequent itemsets. Frequent itemsets are itemsets that their support values are less than or equal to support value of previous itemsets [7].

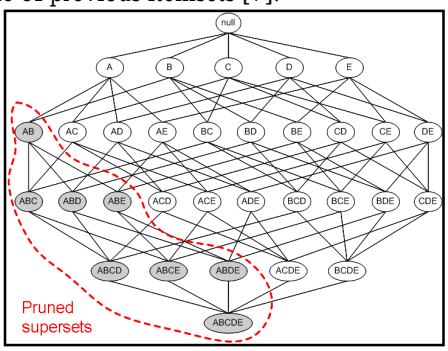

Fig. 2. Itemset lattice with eliminate Infrequent itemsets

On the other hand, support values for infrequent itemsets are higher than the previous support values. In [8] describes frequent itemsets as follow: if any itemsets are frequent itemsets then every subsets must be frequent itemsets. In other word, if a subset is infrequent itemsets then the following itemsets are infrequent itemset as well. For example in Fig.2, if {A,B} is infrequent itemset then {A,B,C}, {A,B,D}, {A,B,E}, {A,B,C,D}, {A,B,C,E}, {A,B,D,E} and {A,B,C,D,E} are infrequent itemset. Hence, we can eliminate or prune all subsets within that itemsets. This eliminating infrequent itemset step is sometime called support-based pruning algorithm. In Fig. 2 show itemset lattice with eliminate infrequent itemsets witch receive smaller candidate set. Therefore it can remove the itemsets and not consider all superset of itemsets.

D. Chi-square Feature Selection

The Chi-Square (χ^2) statistic is a common technique for find relationship between two variables. In this paper, we use for feature selection in data of high dimension. The Chi-Square feature selection algorithm evaluates the worth of a feature by computing the value of the Chi-Square statistic with respect to the class [9]. Then we remove all irrelevant and least relevant features from the dataset that considers from Chi-Square value. It is tested by Chi-Squared formula as is shown in equation (3):

$$\chi^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}} \quad (3)$$

Where: O_{ij} is the observed frequency,

E_{ij} is the expected frequency.

$$E_{ij} = \frac{(R_{T_i})(C_{T_j})}{N} \quad (4)$$

Where: R_{T_i} is number patterns in the ith interval,

C_{T_i} is number patterns in the jth class,

N is total number patterns.

IV. Proposed Method

In this section, we will elaborate on our proposed method for intrusion detection which consists of four phases as illustrated in Fig. 3. Four phases of our proposed method used in this paper are: data preprocessing, feature selection, classification, and system evaluation. Firstly, insignificant attributes will be removed and then applied with data sampling technique. Secondly, feature selection based on Chi^2, HGIS and MHGIS feature selection are to get the optimal attributes. Thirdly, data classification based on C4.5, BPNN, RBF and SVM are used to classify the network data. In the last phase, accuracy rate, detection rate, false alarm rate and CPU processing time are calculated to evaluate the performance of intrusion detection.

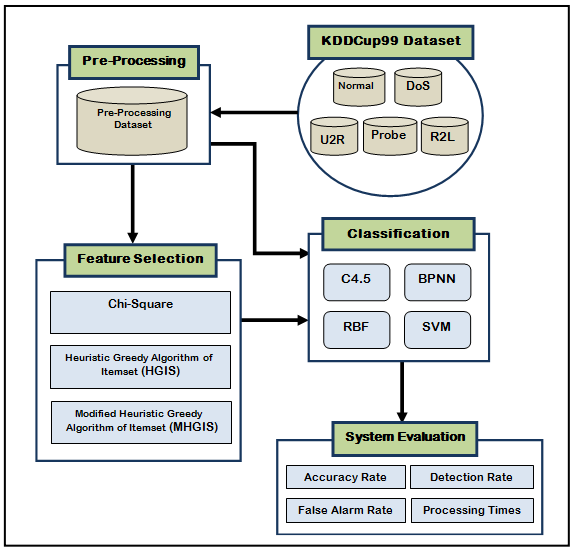

Fig. 3. Proposed model of intrution detection

A. Data Preprocessing Phase

The KDD Cup 99 dataset is standard dataset used for evaluating intrusion detection algorithms. It is about 5 million records. It can be said that it is a very large dataset. It will consume a lot of CPU time if we use all of them in recognition procedure. Then, many researchers recommend to choose only 10 percents and then sampling about 13,499 patterns for learning and testing the performance of the recognition system [11][12][13]. KDD Cup 99 consists of 5 groups. After the last sampling step, numbers of instants in each group are shown in TABLE II. In this paper, we discard some basic features and zero value features because these features have no significant effect to the learning performance. Hence, the remaining features for next step are only 34 features.

TABLE I. AMOUNT DATA OF EACH GROUP

Class Name	Amount
Normal	4,107
DoS	4,107
U2R	4,107
R2L	1,126
Probe	52

B. Feature Selection Phase

We propose two algorithms for extracting and selecting features, detailed as following:

- Feature Selection using Chi-square:

Step 1: Compute chi-square value every pair of attribute and class as in equation (3) then sort value in descending order.

Step 2: remove irrelevant features from dataset that there chi-square value equals zero.

Step 3: remove least relevant feature from the dataset that satisfy the condition

$$\frac{\chi_i^2 \times \log(N^2)}{\sum \chi_i^2 \times N} \times 100 < \delta \qquad (5)$$

Where we set $\delta = 0.1$ to satify our criterion, χ_i^2 is chi-square value for feature in consideration and N is the total number of attributes.

- Feature Selection using Modified Heuristic Greedy Algorithm of Itemset:

In feature selection step, we used MHGIS for select proper feature. Difference between HGIS and Modified HGIS criterion for itemsets generation and itemsets addition step that added to modified HGIS are shows in TABLE II. Improvement of criterion for itemset generation more elaborate in modified HGIS is added with delta. Final step of modified HGIS is item addition that dropped off by finding consequents item that it can reduce $rmse$.

TABLE II. DIFFERENCE BETWEEN HGIS AND MODIFIED HGIS

	HGIS	Modified HGIS
Criterion for itemset generation	$f(\{a,b\}) \le f(\{a\})$ *and* $f(\{a,b\}) \le f(\{b\})$	$f(\{a,b\}) \le f(\{a\}) + \alpha$ *and* $f(\{a,b\}) \le f(\{b\}) + \alpha$
Itemset dddition	-	Finding consequent item that it can reduce rmse

As described in previous section, 34 features have been selected. Then, feature selection using modified heuristic greedy algorithm by apriori algorithm consists two steps will be described as following:

1. Finding based itemset step

Step1: generate 1-itemsets; find support value as in (1) of each item with RBF.

Step 2: generate 2-candidate itemsets; find support value as equation (1) with RBF of each pair 1-itemset.

Step 3: generate 2-itemset by union of 2-candidate itemsets that have support value less than or equal to 1-itemset subsets of the 2-candidate itemsets as in (6).

$$f(\{a,b\}) \le f(\{a\}) + \alpha \; and \; f(\{a,b\}) \le f(\{b\}) + \alpha \qquad (6)$$

where we set $\alpha = 0.2$ to satify our criterion.

Step 4: repeat step 2 and increase size of itemsets until cannot generate itemsets.

2. Itemset addition step

The best itemset that derived from the finding based itemset step are used as the base for finding consequent set that it can reduce $rmse$.

Step 1: get 1-itemsets with $rmse$ is less than the based itemset add to based itemset and find support value as in (1).

Step 2: select lowest $rmse$ itemset is the base for finding consequent itemset.

Step 3: add item that its $rmse$ is less than the $rmse$ in previous step.

Step 4: repeat step 2 until get the best itemset that it is the lowest $rmse$.

C. Classification Phase

In classification phase, we compare the performance metrics with four classification methods named C4.5 [13], BPNN [14], RBF [15] and SVM [16]. Learning Parameters of each method are defined as follows:

- BPNN
  Number of hidden Layers = (attributes + classes) / 2
  Learning Rate = 0.3
  Momentum = 0.2
- SVM
  The polynomial kernel
- RBF
  Gaussian function
- C4.5
  Confidence threshold for pruning = 0.25

Minimum number of instances per leaf = 2
Number of folds for reduced error pruning = 3

In classification experiment, data is divide into 10 folds cross validation for training and testing.

D. Performance Evaluation

The standard metrics and most to use that have been developed for intrusion detection system evaluation are accuracy rate, detection rate and false alarm rate.

Accuracy rate is computed as the ratio between amounts of correctly classified and total number of all data can be found from:

$$\text{Accuracy } Rate = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (7)$$

Detection rate is computed as the ratio between the number of correctly detected attacks and the total number of attacks can be represented by:

$$Detection\ Rate = \frac{TP}{TP+FN} \quad (8)$$

False alarm rate is computed as the ratio between the numbers of normal connections that is incorrectly misclassified as attack and total number normal connection by following equation:

$$False\ Alarm\ Rate = \frac{FP}{FP+TN} \quad (9)$$

Definitions of variables are show in TABLE III.

TABLE III. CONFUSION MATRIX

Predicted / Actual	Normal	Attack
Normal	True Negative (TN)	False Positive (FP)
Attack	False Negative (FN)	True Positive (TP)

V. EXPERIMENTAL RESULTS

After proposed feature selection, we use four standard supervised learning algorithms which are C4.5, BPNN, RBF and SVM for evaluating the significance of the selecting features. From the KDDCup99 with 34 data dimensions based on Chi^2 and MHGIS algorithms, we obtain 26 and 14 features, respectively. Furthermore, we compared with original HGIS feature selection that it gets 13 features.

TABLE IV. ACCURACY RATE

Learning Method	Accuracy Rate (%)			
	All (34)	Chi^2(26)	HGIS (13) [4]	MHGIS (14)
C4.5	99.43	99.44	98.76	99.52
BPNN	98.69	96.54	97.34	97.43
RBF	93.69	94.52	95.53	95.59
SVM	97.02	95.04	94.32	95.24

In TABLE IV shows accuracy rate of experimental. It can be seen that most accuracy rate using MHGIS feature selection batter than other feature selection. And MHGIS with C4.5 decision trees classification is the best with accuracy rate 99.52%. MHGIS with RBF has accuracy rate more than original data 1.9%.

TABLE V. DETECTION RATE

Learning Method	Detection Rate (%)			
	All (34)	Chi^2(26)	HGIS (13) [4]	MHGIS (14)
C4.5	99.57	99.53	99.30	99.63
BPNN	99.33	98.17	98.41	99.05
RBF	93.05	95.50	96.40	95.37
SVM	98.75	97.85	96.87	97.72

Detection rate of experiment displays in TABLE V. that MHGIS is better than HGIS and resemble Chi^2. MHGIS feature selection and C4.5 classification has the best result detection rate 99.63%.

TABLE VI. FALSE ALARM RATE

Learning Method	False Alarm Rate (%)			
	All (34)	Chi^2(26)	HGIS (13) [4]	MHGIS (14)
C4.5	0.96	0.84	1.56	0.84
BPNN	1.50	2.87	3.55	2.11
RBF	16.49	10.52	11.49	10.29
SVM	2.77	4.91	7.47	5.31

In TABLE VI, Selected data with MHGIS 9 features and Chi^2 26 features have minimum false alarm rate 0.84%. The best performance of processing times in all pattern recognitions is HGIS with 13 features that show in TABLE VII.

TABLE VII. PROCESSING TIMES

Learning Method	Processing Times(s)			
	All (34)	Chi^2(26)	HGIS (13) [4]	MHGIS (14)
C4.5	38.63	31.39	16.61	18.43
BPNN	166.68	114.87	65.28	74.90
RBF	47.28	38.98	12.74	15.62
SVM	41.48	32.96	13.18	17.05

VI. CONCLUSION

In this paper, the proposed method of feature selection using modified Heuristic Greedy Algorithm of Itemset (MHGIS) implementing with apriori algorithm to improve the detection rate, accuracy rate and false alarm rate, as compare to feature selection using Chi^2 and HGIS. The experimental results indicate that the feature selection based on MHGIS yields a better all of assessment but processing times are a

little lower than HGIS. The most results have good result with decision tree C4.5 because KDDCup99 dataset has high spreading. But for the detection rate and the false alarm rate are improved only in some cases. So in the future works, we intend to improve an algorithm to have a better performance on both detection and false alarm rate. In our future work, we are interested in find out the algorithm that can improve the computing time during the selection process too.

ACKNOWLEDGMENT

This work is funded by the National Research Council of Thailand (NRCT), fiscal year 2012.

REFERENCES

[1] Murat Karabatak, M. Cevdet Ince, "A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases", Expert Systems with Applications, Volume 36, pp. 12500–12505, 2009.

[2] Hari Om , Aritra Kundu, "A Hybrid System for Reducing the False Alarm Rate of Anomaly Intrusion Detection System", 1st Int'l Conf. on Recent Advances in Information Technology, 2012.

[3] Janya Onpans, Annupan Rodtook , Suwanna Rasmequan, Benchaporn Jantarakongkul and Krisana Chinnasarn, "Intrusion Feature Selection using Heuristic Greedy Algorithm of Item Set", Knowledge and Smart Technology (KST) 5th, pp.22-29, 2013.

[4] KDD'99 datasets, The UCI KDD Archive, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, Irvine, CA, USA, 1999.

[5] Mahbod Tavllaee, Ebrahim Bagheri, Wei Lu Ali A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data set", Proceeding of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA), 2009.

[6] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms (3e), p.360, 2001.

[7] P. N. Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining, Pearson International Edition, ISBN: 0-321-42-52-7, 2006.

[8] K. P. Soman, S. Diwakar, and V. Ajay, "Insight into Data Mining Theory and Practice", Prentice-Hall of India, ISBN: 81-203-2897-3, 2006.

[9] H. Liu, R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," IEEE 7th International Conference on Tools with Artificial Intelligence, pp. 338-391, 1995.

[10] Ranjit Abraham , Jay B. Simha, S. Sitharama Iyengar, "Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical datamining", International Journal of Computational Intelligence Research, ISSN 0974-1259 Vol.5, No.2, pp. 116–129, 2009.

[11] Amir-Massoud Bidgoli, Mehdi Naseri Parsa, "A Hybrid Feature Selection by Resampling, Chi squared and Consistency Evaluation Techniques", World Academy of Science, Engineering and Technology 68, 2012.

[12] M.Revathi, T.Ramesh, "Network Intrusion Detecion System Using Reduced Dimensionality", Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976-5166, vol. 2, no.1, 2011.

[13] J. Ross Quinlan, "C4.5: programs for machine learning", Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, ISBN: 1-55860-238-0, 1993.

[14] Robert Hecht Nielsen, Theory of the back propagation neural network in Proceedings 1989 IEEE IJCNN, pp. 1593–1605, IEEE Press, New York, 1989.

[15] S.Chen, C. F. N. Cowan, P. M. Grant, "Orthogonal Least Squares Learning Algorithm for Radial Basis Function Networks" IEEE transactions on neural networks, vol. 2, no.2, 1991.

[16] M. Hearst, ed., "Support Vector Machines," IEEE Intelligent Systems Magazine, Trends and Controversies, Marti Hearst, ed., vol 13, no 4, 1998.