HYBRID FEATRUE SELECTION FOR NETWORK INTRUSION

S.Sethuramalingam[1]

[1]Associate Professor and Head, Department of Computer Science
Aditanar College Tiruchendur 62815. India

Dr.E.R. Naganathan[2]

[2]Professor and Head, Department of Master of Computer Applications
Velammal Engeering College, Chennai. India

Abstract

In Computer Communications, collecting and storing characteristics about connections into a data set is needed to analyze its behaviour. Generally this data set is multidimensional and larger in size. When this data set is used for classification it may end with wrong results and it may also occupy more resources especially in terms of time. Most of the features present are redundant and inconsistent and affect the classification. In order to improve the efficiency of classification these redundancy and inconsistency features must be eliminated. In this paper, we have proposed a new algorithm based on hybrid method to identify the significance of features. The Proposed hybrid method combines Information Gain and Genetic Algorithm to select features. Clustering is carried out on selected features for classification. The experiment is conducted with NLS-KDD network intrusion data set. It classifies the data set with good accuracy.

Keywords:* *Intrusion, Hybrid Feature Selection, Information Gain, Genetic Algorithm,

I INTRODUCTION

Now a days, Information security is a critical component of any system. As a part of any secure system, detection and prevention of security violations, if any, is the need of the hours. In a computer network, there are two main intrusion detection systems - Anomaly intrusion detection system and misuse intrusion detection system. The first one is based on the profiles of normal behaviour of users or applications and checks whether the system is being used in a different manner. The second one collects attack signatures, compares behaviour with these attack signatures and signals intrusion when there is a match [1].

In a network data set, details about the connections are called connection records. Network intrusion detection classifies these records into either normal or anomaly. Classification depends on the features that adequately characterize the objects of interest. The task of identifying the features that perform well in a classification algorithm is a difficult one, and the optimal choice can be non-intuitive; features that perform are separately poor. They can often prevail when paired with other features. Many different approaches and techniques are discussed [*2*]. The filter approach [*3*] to feature selection tries to infer which features will work for the classification algorithm by drawing conclusions from the observed distributions (histograms) of the individual features. The correlation structure of the data is responsible for the success of the joint classifier, and a good classification scheme will attempt to utilize this structure.

Another technique, known as wrapper approach [*4*], uses the method of classification itself to measure the importance of a feature or a feature set. The goal in this approach is maximizing the predicted classification accuracy. This approach is computationally more expensive and tends to provide better results than the simple filter methods.

Most of the existing works are focused on the wrapper mode using different classifier methods. In this paper hybrid feature selection method is proposed. This method combines filter approach and wrapper approach. Information Gain is used as filter approach and Genetic algorithm is used as wrapper approach. Features are selected first with information gain and then with Genetic Algorithm. The clustering is carried out on the selected features to classify the data sets into normal and anomaly.

The remainder of this paper is organized as follows: Section 2 gives a review of related works in the feature selection using hybrid approach. Section 3 describes the proposed methodology. The arrived results are discussed in Section 4 and Section 5 has conclusions and future work.

II RELATED WORK

Filter approach selects features based on their characteristics, therefore it requires less computational resources whereas wrapper approach selects features based on the classification itself hence it requires more computational resources. In order to get the benefit of both, hybridization is done. Hybrid features selection is applied in micro array data, machine learning, image processing and intrusion detection. In [5] authors have selected two feature selection method Bayesian Networks and Classification and Regression Trees (CART) for network intrusion. Features selected using ensemble of the both, shows better performance than when they are applied as individual one. The SVM is used to measure the performance of the above ensemble approach. Hybrid of Information Gain and Genetic Algorithm is applied to select the features in Microarray Data is discussed in [6]. The first level of features is selected with Information Gain. The second level of features is selected using Genetic Algorithm. The K-NN algorithm is used for classification. The Symmetrical Uncertainty and Genetic Algorithm have been used to filter unwanted features [7]. In the first stage Symmetrical Uncertainty is used to filter features which have low values. In the next stage Genetic Algorithm is applied to select features. Another type of hybridization is embedding different search techniques to select the optimal feature set. In [8] local search operations are embedded in the Genetic Algorithm to fine tune the selection process.

In addition, the selection of a subset will reduce the dimensionality of the data samples and eliminate the redundancy and ambiguity introduced by some attributes. Most of the authors have used feature selection to get accurate results. Information Gain and Genetic Algorithm are used for selecting features. Support Vector Machines and Tree based classifiers are used for classification. However in this context clustering is a very limited one. The proposed method has two steps to select features. In the first step Information Gain is used to select features which are greater than the given threshold. Next, Genetic Algorithm is applied to select features from the set of features obtained from the previous step. Clustering is applied on the selected set of features. The dataset used in this work is NSL-KDD data set

III. METHODOLOGY

The main goal of this work is to use information gain to identify the significance of features and then select a set of significant features using the proposed Genetic algorithm. Using the selected set of features Clustering is performed on the data set. In the case of clustering, the selected set of features has yielded good results than that of all the features put together.

3.1 The Data Set

Since 1999, KDD'99 [*9*] has been the most wildly used data set for the evaluation of anomaly detection methods. This data set is prepared by Stolfo et al. [10] and is built based on the data captured in DARPA'98 IDS evaluation program[11]. DARAPA'98 is about 4 gigabytes of compressed raw TCP dump data of seven weeks of network traffic, which can be processed into about 5 million connection records, each with about 100 bytes. The two weeks of test data have around 2 million connection records. KDD training dataset consists of approximately 4,900,000 single connection vectors each of which contains 41 features and is labeled either normal or an attack, with exactly one specific attack type.

The first important deficiency in the KDD data set is the huge number of redundant records. Analyzing KDD train and test sets [*12*], they found that about 78% and 75% of the records are duplicated in the training and the testing data set respectively. This large amount of redundant records in the training data set will cause learning algorithms to be biased towards the more frequent records, and thus prevent it from learning unfrequented records which are usually more harmful to networks such as U2R attacks. The existence of these repeated records in the test set, on the other hand, will cause the evaluation results to be biased by the methods which have better detection rates on the frequent records.

In addition, to analyze the difficulty level of the records in KDD data set, the authors employed 21 learned machines (7 learners, each trained 3 times with different train sets) to label the records of the entire KDD train and test sets, which provides us with 21 predicted labels for each record. Surprisingly, about 98% of the records in the train set and 86% of the records in the test set were correctly classified with all the 21 learners. The reason for getting these statistics on both KDD train and test set is that in many papers, random parts of the KDD train set are used as test sets. As a result, they achieve about 98% classification rate applying very simple machine learning methods. Even applying the KDD test set will result in having a minimum classification rate of 86%, which makes the comparison of IDSs quite difficult since they all vary in the range of 86% to 100%.

The new version of KDD data set NSL-KDD is publicly available for researchers through the website [*13*]. Although, the data set still suffers from some of the problems discussed [*12*] and may not be a perfect representative of existing real networks, because of the lack of public data sets for network-based IDSs, the authors believe that it still can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods. In this work NSL_KDD dataset is used.

The data set NSL-KDD is used to test the performance of the proposed method. In the data set KDD20train.ariff consists of 25,192 records. The number of records which belongs to TCP protocol is 20526 records and that of UDP and ICMP are 3011 and 1655 respectively.

The data set has 42 features. The first 41 features describe the characteristics of connection record. The last feature is labeled either normal or anomaly. Four features of the 41 are described using discrete values. The remaining features are described using continuous values. The continuous values are separated using equal interval method.

3.2 Data Standardization

We first standardize the data set [*14*]. A collection of numeric data is standardized by subtracting a measure of central location such as mean and divided by some measure of spread such as standard deviation. This yields data with similar shaped histogram with values centered on zero.

$$x'_{ik} = \frac{x_{ik} - \bar{x}}{s_k} \qquad (i=1,2,\dots n,k=1,2..m) \qquad (1)$$

In "(1)" where $\bar{x}$ is the mean value and s_k is the standard deviation of the k^{th} dimension. m is the number of rows and n is the number of columns

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_{ik} \qquad (2)$$

$$s_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - x_k)^2} \qquad (3)$$

Standardization"(3)" transforms the mean"(2)" of the set of feature values to zero, and the standard deviation to one, but may not be in the interval [0 1]. After the following change, it "(4)" is mapped into the interval [0 1].

$$x''_{ik} = \frac{x'_{ik} - \min_{1 \le i \le n}\{x'_{ik}\}}{\max_{1 \le i \le n}\{x'_{ik}\} - \min_{1 \le i \le n}\{x'_{ik}\}} \qquad (4)$$

3.3 Information Gain for feature selection

The computation of the Information Gain for only one attribute according to the classes is given below: let S be a set of training set samples with their corresponding labels. Suppose there are m classes and the training set contains s_i samples of class I and s is the total number of samples in the training set expected information needed to classify a given sample is calculated by "(5)":

$$I(s_1,s_2,\dots\dots s_m) = -\sum_{i=1}^{m}\frac{s_i}{s}\log 2\frac{s_i}{s} \qquad (5)$$

Feature F with values ($f_1,f_2,f_3,\dots\dots,f_v$) can divide the training data set into v subsets $\{S_1,S_2,\dots,S_v\}$ where S_j is the subset which has the value f_j for the feature F. Furthermore let S_j contain s_{ij} samples of class i. Entropy of the feature F is given in "(6)"

$$E(F) = -\sum_{i=1}^{m}\frac{s_{ij}+s_{2j}+s_{3j}+\dots\dots+s_{mj}}{s}I(s_{1j},s_{2j},\dots\dots\dots s_{mj}) \qquad (6)$$

Information gain for F can be calculated using “(7)”

$$\text{Gain}(F)= I(s_1,s_2,\ldots\ldots s_m)-E(F) \qquad (7)$$

The value of the gain as given above computes the information gain of a feature F with regard to all the classes. If we want to measure the gain of the feature for a given class k, we shall consider the problem as a binary classification one. We consider two classes: the class normal (s_k) and the remaining will constitute another class anomaly ($s_{k'}$). So the expected Information Gain needed to classify a given sample will be:

$$I(s_k,s_{k'})= -\frac{s_k}{s}\log 2\left(\frac{s_k}{s}\right)\frac{s_k}{s}-\frac{s_{k'}}{s}\log 2\frac{s_{k'}}{s} \qquad (8)$$

where k’ denotes the complemented class of the class k. The entropy of a feature F according to the class k is

$$E(F)=-\sum_{i=1}^{m}\frac{s_{kj}+s_{k'j}}{s}I(s_{kj},s_{k'j}) \qquad (9)$$

Information Gain for F can be calculated using “(10)”

$$\text{Gain}(F)=I(s_k,s_{k'})-E(F) \qquad (10)$$

This gain measure gives the significance of the features. The following algorithm selects features which are greater than threshold value from the data set.

```
Algorithm Feature Selection Using Information Gain
    //sf1 is used to store selected set of features. Initially it is empty
    // th conatins threshold value.
    // f(i) contains i^th feature of the data set
sf1={};
for i= 1 to number of features in the data set
    inf=compute Information Gain for the feature
    gain(i)=inf
end for
th= threshold value
for i= 1 to number of features
    if gain(i)>th then
        sf1=sf1+f{i}
    end if
end for
end
```

3.4 Genetic Algorithm for Feature Selection

All feature selection methods needs to use an evaluation function together with a search strategy to obtain the optimal feature set. It is unfeasible to search all subsets to find out a optimal subset since it requires a large amount of computational effort. A wide range of heuristic search strategies have been used including forward selection, backward elimination, hill-climbing, branch and bound algorithms, and the stochastic algorithms like simulated annealing and genetic algorithms [15].

The First step of the Genetic Algorithm is designing a chromosome. Each chromosome is a group of genes. In this problem, each gene represents a feature and a chromosome represents a set of features. To indicate whether a particular feature is present or not in the chromosome, one and zero are used. One in a gene position indicates that a particular feature is present and zero indicates that it is absent. The next question how many features and what are the features that are to be present in a chromosome is guided by the Information Gain. Initial population is created using randomization of the values present in the chromosome. Having generated the population, the individuals are evaluated using fitness function. The K-means clustering is used as a fitness function. The false positive and false negative values are measured for each chromosome. The chromosome which has a minimum value of false positive and false negative is considered as Elite one.

In the next step, crossover and mutation are to be applied on the chromosomes which have the highest fitness value. Crossover operation is performed on these two chromosomes. Location of crossover point may be decided randomly using the one point crossover, two point crossovers, or homologue crossover. In this problem one point crossover is applied. Mutation genetic operator is responsible for maintaining diversity in the population. For mutation select one position randomly from the chromosome and the value is changed in that position by inversing. The mutation and cross over operations are shown in the following figure 1. In the figure mutation bit is shown in italics.

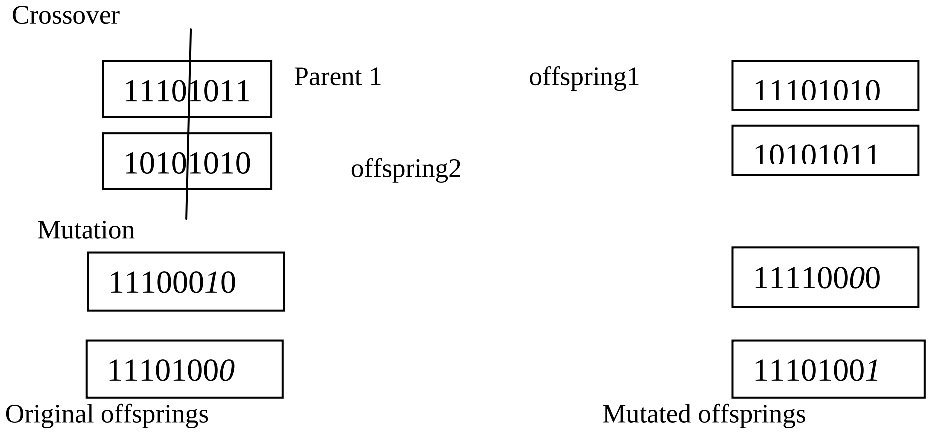

Figure 1 Mutation and Crossover operations

The main goal of feature selection is to use minimum number of features to obtain the same or better results. Fitness function is one of the most important parts in genetic search. This function has to evaluate the effectiveness of each individual in a population, so it has an individual as an input and it returns a numerical evaluation that must represent the goodness of the feature subset. The crossover and mutation functions are the main operators that randomly impact the fitness value. Fitness is equal to the value of False Positive and False Negative obtained from clustering. The following algorithm selects feature from the set of features which are obtained by Information Gain.

```
        Algorithm: feature selection using Genetic Algorithm
    // sf1 is set of features used as input to this algorithm
        //sf2 contains selected features at the end of the algorithm
    // psize    population size
//a is an array is used to store the populations
    gen =0   // number of generations
    Design chromosome
    While gencount<=gen
    Randomly generate psize population and store it into the array a.
    For i=1 to psize
    fitnessfunction(a(i))
    Store the fitness value    //value of false positive(fp) and false negative(fn)
    End for
    Select two chromosome say ch1 and ch2 which has minimum fp and fn values
    crossover(ch1,ch2)            //cr1,cr2 are  outputs  of  crossover
    randomly mutate one bit in cr1 //mutation operation
    randomly mutate one bit in cr2 //mutation operation
    gencount=gencount+1
    end while
    end
```

3.5 K-means Clustering

Clustering refers to identifying the number of subclasses of c clusters in a data universe X comprising of n data samples, and partitioning X into c clusters. There are two kinds of c-partitions of data, hard and soft. For numerical data one assumes that the members of each cluster bear more mathematical similarity to each other than to the members of other clusters.

One important issue to consider is how to measure the similarity between pairs of observations. One of the simplest similarity measures is distance between pairs of features in the record. In the clustering, Euclidean Distance measure is used to measure the similarity. d_{ik} is a Euclidean distance measure between the kth data sample x_k and ith cluster center v_i is given by “(11)”

$$d_{ik} = d(x_k - v_i) = \| x_k - v_i \| = \left[\sum_{j=1}^{n} (x_{kj} - v_{ij})^2 \right]^{1/2} \qquad (11)$$

since each data sample requires n dimensions to describe its location in the dataset, each cluster center also requires n dimensions to describe its location in the data set. Therefore

$$Vi = \{v_{i1}, v_{i2}, \ldots\ldots. v_{in}\}$$

where the jth coordinate is calculated by “(12)”

$$v_{ij} = \frac{\sum_{k=1}^{n} Xik \; . \; xkj}{\sum_{k=1}^{n} Xik} \qquad (12)$$

The step by step procedure is given below:

1. Fix the number of clusters(c) and initialize the partition (U) matrix
2. Initialize the K cluster centroids. This can be done by arbitrarily dividing all objects into K clusters, computing their centroids and verify that all centroids are different from each other. Alternatively, the centroids can be initialized to K arbitrarily chosen different objects.
3. Iterate over all objects and compute the distances to the centroids of all clusters. Assign each object to the cluster with the nearest centroid.
4. Recalculate the centroids of both modified clusters
5. Repeat step 3 until the centroids do not change any more

A confusion matrix as shown in the Table 1 is typically used to evaluate the performance of the algorithm.

Confusion Matrix (standard metrics)		Predicted connection label	
		Normal	Intrusion (Anomaly)
Actual connection Label	normal	True Negative (TN)	False Alarm (FP)
	Intrusion (anomaly)	False Negative (FN)	Correctly detected (TP)

Table 1 Standard metrics for evaluation of intrusions

From Table 1, recall and precision may be defined as follows

Precision=TP/(TP+FP)

Recall=TP/(TP+FN)

IV RESULTS and DISCUSSION

Using NSL-KDD dataset [13] the experiment is conducted. The data set has 20526 TCP connection records, 3011 UDP connection records and 1655 ICMP connection records. In TCP, 10681 records are belonging from normal and 9685 records are belonging from anomaly connections. In UDP number of normal connection and anomaly connection records are 2507 and 504 respectively. In ICMP number of normal and anomaly connection records are 261 and 1394 respectively. In these data sets information gain for each feature is computed as discussed in section 3.3. The following table 2 shows features which are having

Information Gain greater than zero. Those features which are having Information Gain value as zero does not play any significance role in the classification. Therefore they are eliminated at the first level.

Protocol Type	Features greater than Information Gain zero
ICMP	5,24,29,30,34,36,32,33,31,37,23,35,38,40,8,25,27
UDP	5,6,34,36,33,40,24,8,23,38,31,35,29,30,25,27,32,37
TCP	5,30,29,34,6,33,23,12, 39,25,38,26,35,37,32, 36,31,41,27,24,40,28,10,13,16,19,22,17,15,14,18,11,9,7

Table 2: protocol type and features have Information Gain greater than zero

This set of features is the input to the algorithm for feature selection using genetic algorithm discussed in section 3.4. The parameter values taken as population size is 100, number of generations is 7, the crossover takes place at the middle position and mutation is done at one point randomly. The results for the classification on the training dataset are given below. Table 3 shows that the classification has good precision and recall when the selected features are used. In the case of ICMP using Information gain, number of features selected for classification is eight. Accuracy and Recall are also better. When features are selected with Information Gain and Genetic Algorithm, five features are selected and the results of accuracy and recall are better than Information Gain which alone is used for selecting features. The same observation is present in UDP and TCP number of features selected is less when the proposed hybrid approach is applied. Table 3 shows the features and their false positive, true negative, precision and recall [*16, 17, 9*]. They are metrics used to measure the performance of classification.

protocol type	Method of Feature selection	features Selected	Actual		predicted				Precision	Recall
			Normal	anomaly	TN	FN	FP	TP		
ICMP		All	261	1394	213	48	235	1159	0.83142	0.83142
ICMP	Information Gain	5,24,29,30,34,36,32,33,31,37,23 35,38,40,8,25,27	261	1394	213	48	235	1159	0.83142	0.96023
ICMP	Information Gain and Genetic Algorithm	24,29,30,34,36	261	1394	211	50	94	1300	0.93256	0.96296
UDP		All	2507	504	2235	272	176	328	0.65079	0.54666
UDP	Information Gain	5,6,34,36,33,40,24,8,23,38,31,35,29,30,25,27,32,37	2507	504	2235	272	176	328	0.65078	0.54666
UDP	Information Gain and Genetic Algorithm	6,34,36,24,8,23,38,29,30,25	2507	504	2391	116	163	388	0.70417	0.85462
TCP		All	10681	9845	10096	585	434	9411	0.95591	0.94147
TCP	Information Gain	5,30,29,34,6,33,23,12,39,25,38,26,35,37,32,36,31,41,27,24,40,28,10,13,16,19,22,17,15,14,18,11,9,7	10681	9845	10096	585	434	9411	0.95591	0.94147
TCP	Information Gain and Genetic Algorithm	5,29,34,6,33,12,39,25,38,26,35,32,36,31,41,27	10681	9845	10096	584	434	9411	0.95591	0.94147

Table 3 showing the precision and recall value for ICMP, UDP and TCP with all features and selected features using different methods of feature selection

V. CONCLUSIONS and FUTURE WORK

In this work, a new Hybrid approach for selecting the best discriminate features using information gain and genetic algorithm is presented. From the Results of Table 3 the classification with selected features shows better results. The selection method using Information Gain and Genetic Algorithm shows better results in terms of number of features selected and accuracy than applying methods individually.

In future work, this approach can be tested for other type data sets and to explore the possibilities of other methods of selecting optimal feature set. Identify other methods of classification which will improve the results. The algorithm can be applied to different type of data sets which are required for dimensionality reduction and classification.

6. REFERENCES

[1] Fayyad.U.M, Piatetsky-Shapiro.G, and Smyth.P, .From data mining to knowledge discovery in databases. AI Magazine, vol. 17, no. 3, pp. 37.54, 1996

[2] Yang.J. and Honsvar.V. Feature subset selection using genetic algorithm. In IEEE Intelligent Systems, Volume13 page 44-49, 1998

[3] John.G, Kohavi.R and Pfleger.K Irrelevant features and the subset selection problem. In the 11th International Conference on Machine Learning, pages 121-129, 1994.

[4] Kohavi.R and John.G Wrappers for feature subset selection. Artificial Intelligence Journal, Volume 97, special issue on relevance, pp 273-324. Dec. 1997

[5] Srilatha Chebrolu, Ajith Abraham and Johnson P Thomas “Hybrid Feature Selection for Modeling Intrusion Detection Systems”, Department of Computer Science, Oklahoma State University, USA ajith.abraham@ieee.org, jpt@okstate.edu

[6] Cheng-Huei Yang Li-Yeh Chuang, Cheng-Hong Yang “IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data” Journal of Medical and Biological Engineering, 30(1): 23-28

[7] Bai-Ning Jiang Xiang-Qian Ding Lin-Tao Ma “A Hybrid Feature Selection Algorithm:Combination of Symmetrical Uncertainty and Genetic Algorithms” The Second International Symposium on Optimization and Systems Biology (OSB’08) Lijiang, China, October 31– November 3, 2008 Copyright © 2008 ORSC & APORC, pp. 152–157

[8] Il-Seok Oh, Jin-Seon Lee, Byung-Ro Moon, "Hybrid Genetic Algorithms for Feature Selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1424-1437, November, 2004

[9] Claise.B, Bryant.S, Sadasivan.G, Leinen.G, and T. Dietz, .IPFIX Protocol Speci_cations,. Internet Draft, work in progress, draft-ietfip_x-protocol-24, Nov. 2006

[10] Lee.W and Stolfo.S, .Data mining approaches for intrusion detection, in Proceedings of the 7th USENIX Security Symposium, San Antonio, TX, 1998

[11] Dokas P, Ertoz.L,Kumar.V,Lazarevie.A,Srivastava.J and Tan P.N., “Data mining for network intrusion detection”, Proceedings of the NSF Workshop on nxt generation Data Mining, Nov. 2002

[12] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani . A Detailed Analysis of the KDD CUP 99 Data Set Proceeding of the IEEE Symposium on Computational Intelligence in Security and Defence Applications (CISDA 09)

[13] http://nsl.cs.unb.ca/NSL-KDD/

[14] Hai Jin Jianhua Sun, Han Chen, Zongfen Han Cluster and Grid Computing Lab. Huazhong University of Science and Technolory, Wuhan 430074 China. A Fuzzy Data Mining Based Intrusion Detection Model. Proc. Of the 10th IEEE International Workshop on Feature Trends of Distributed Computing Systems (FTDCS’04)@2004 IEEE

[15] Aouatif Amine Ali Elakad Mohammed Rziza Driss Aboutajdine “ GA-SVM and Mutual Information based Frequency Feature Selection for Face Recognition” GSCM-LRIT, Faculty of Sciences, Mohammed V University, B.P. 1014 Rabat, Morocco

[16] Provost.F and Fawcett.T, Robust Classification for Imprecise Environments. Machine Learning, vol. 42/3 pp. 203-231, 2001

[17] Joshi.M, Kumar.V, Agarwal.R Evaluting Boosting Algorithm to Classify Rare Classes: Comparison and Improvements First IEEE International conference on Data Mining San Jose, CA, 2001