# A Hybrid ACO Based Feature Selection Method for Email Spam Classification

KARTHIKA RENUKA D[1], VISALAKSHI P[2]
[1]Department of Information Technology
[2]Department of Electronics and Communication Engineering
PSG College of Technology
Coimbatore - 641004, Tamilnadu, INDIA
[1]karthirenu@gmail.com, [2]visapsg@gmail.com

*Abstract:* - In recent days, Internet plays a main role. Internet mail system is a store and forward mechanism used for the purpose of exchanging documents across computer network through Internet. Spam is an unwanted mail which contains unsolicited and harmful data that are irrelevant to the specified users. In the proposed system, spam classification is implemented using Support Vector Machine (SVM), a classifier, which is a powerful non linear classifier applicable for complex classification problems. The proposed system is analyzed by calculating its accuracy. Implementation of feature selection using Ant Colony Optimization (ACO) serves to be more efficient which gives good results for the above system that has been proposed in this paper.

*Key-words:* - Ant Colony Optimization, Feature selection, Spam classification, E-mail, Spam, SVM, Spam dataset.

## 1 Introduction

Email spam or junk e-mail is one of the major problems of the today's Internet. They cause financial damage to companies and annoy individual users. It is sending unwanted email messages with commercial content to indiscriminate set of recipients.

Nowadays the percentage of people accessing the Internet has increased rapidly. Most of them are using E-mail for communication. Managing these emails becomes a significant problem for individuals and organizations. Among the traffic, most of the traffic comprises of unsolicited bulk email. Without appropriate counter-measures, the situation seems to be worse and spam e-mails may eventually undermine the usability of e-mail. Spammers collect e-mail addresses from chat rooms, websites, customer lists, newsgroups, and viruses which harvest users address books, and sell them to other spammers. Email classification presents challenges because of large and various number of features in the dataset and large number of mails. Large number of features makes most mails undistinguishable. In many emails datasets, only a small percentage of the total features may be useful in classifying mails, and using all the features may adversely affect performance. The quality of training dataset decides the performance of both the email classification algorithms and feature selection algorithms.

### 1.1 Spam classification

In Spam classification the spam email are moved to the junk email folder. Classification algorithm splits the mails into spam and ham mails. The combination of SVM and ACO for spam classification includes two phases: training phase and testing phase, where training phase involves indexing the two known data labels, which are denoted as spam and ham mails respectively. The testing phase involves the query indexing and the closest message gets retrieved from the training sets. The message which is collected by indexing based on the feature set used and the resulting query vector will be compared. If the message is closer to the spam mail contained in the spam training set, then that query message is classified as spam, otherwise it is classified as ham mails respectively.

### 1.2 Feature selection

Feature selection algorithms are divided into two categories according to the way they process and evaluate features: feature ranking methods and subset selection methods. The rank of the feature is done by the way of metric and also it eliminates all features that do not achieve an adequate score by means of feature ranking methods.

Subset selection methods search the set of possible features for the optimal subset features. A hybrid ACO- based classifier model that combines ACO and Support Vector Machine (SVM) to improve classification accuracy is proposed in this paper.

### 1.3 Objective

The main objective of the proposed system is to develop an email spam classification system in an efficient manner. This proposed system aims in classifying the input set emails into spam and ham mails.

The overall objective of this system is to execute the system in faster to provide a better classification performance with more accuracy.

### 1.4 Scope

In the proposed system the spam classification technique is applied to the spambase dataset taken from UCI repository. More efficient results can be achieved when it's applied to the real time data. The existing mail server's doesn't provide 100% accuracy in classification. The proposed system has a wide range of scope. Since the system is implemented using ACO it can process more number of mails at a time. Due to this reason the scope of the project gets widen off.

## 2 Literature Survey

### 2.1 Existing system

In an ACO based optimization method [1], the design of the pheromone update strategy, and the measurement of the quality of the solutions are critical.

#### 2.1.1 Pheromone updating

The ACOFS algorithm updates the pheromone value on each arc according to the number of solutions passing through the arc and their fitness function values [2]. For example, if an ant chooses the arc $C^i_j$, pheromone on this arc should be assigned more increment, and ants should select arc $C^i_j$ with higher probability in the next iteration. This forms a positive feedback of the pheromone system. In each iteration, the pheromone on each arc is updated according to the following formula:

$$\tau_i^j(t+1)=\rho.\tau_i^j(t)+\Delta\tau_i^j(t)+Q_i^j(t) \quad (1)$$

#### 2.1.2 The fitness function

Based on the ant's solution, which is a selected in the feature subset, the solution quality in terms of classification accuracy is evaluated by classifying the training data sets using the selected features. The calculation for accuracy is done for the number of examples that are correctly classified as spam or ham. In quality function, the number of features is also considered for the given set. The subset with fewer features could get higher quality function value. There are several ways to define such a function. One way is to define the quality function of a solution S as

$$f(S)=[\text{recall}(S)+\text{precision}(S)]/N_{feat} \quad (2)$$

where $N_{feat}$ is the number of features selected in S. Another way to define the quality function is using the redundancy rate. In this paper, the quality function of a solution S is defined as

$$f(S)=N_{corr}/(1+\lambda N_{feat}) \quad (3)$$

where $N_{corr}$ is the number of examples that are correctly classified, $\lambda$ is a constant to adjust the importance of the accuracy and the number of features selected. A solution obtaining higher accuracy and with fewer features will get a greater quality function value.

### 2.2 Spam Classification Machine Learning Methods

The classification of spam can be done in numerous ways. Initially the datasets has to be collected and should be pre-processed in order to apply the spam classification techniques. After the dataset collection, the e-mail classification phase of the process finds the actual mapping between training set and testing set. The training set data has to be processed and trained in such a way that it helps to classify the test data efficiently [3].

The machine learning approach does not require specifying any rules explicitly. Instead, a set of pre-classified training samples is needed. To gain knowledge about the classification rules, particular algorithm need to be used. In today's world, an important research issue is done over spam classification. Machine learning field is a subfield of artificial intelligence that aims to make machines able to learn like human. Statistical representation of the learned information after a strong observance is the main factor.

### 2.2.1 Naïve Bayes Classifier

Bayes' theorem is a method of getting new idea in order to update the value of the probability of occurrence of a task. The relationship between the probability that gets updated, represented as P(A | B),and the conditional probability of A which give the new value for B, and both the probabilities of A and B, and the conditional probability of B given A, P(B | A). Bayesian techniques use mathematical formulae to analyze the content of the message. In its common form bayes theorem is depicted in the equation 4.

$$P(A/B) = \frac{P(B/A).P(A)}{P(B)} \quad (4)$$

Bayesian classifier works on the events which are dependent and the probability of an event occurring in the future that can be detected from the previous occurring of the same event. Naïve bayes classifier technique has become a very popular method in the email classification techniques. The spammines are calculated using the equation 5.

$$S[T] = \frac{C_{spam(T)}}{C_{ham(T)}+C_{spam(T)}} \quad (5)$$

where $C_{spam(T)}$ and $C_{ham(T)}$ denotes the number of spam or ham mails, which contain token T respectively. The values of spamminess and hamminess calculated which gets associated with the message, determines whether probability of a mail is spam or ham mails. From the predefined spamminess and hamminess values associated with a message the probability is calculated which is then compared with a threshold value to determine it as a spam or ham.

### 2.2.2 K-Nearest Neighbour classifier

In KNN technique a document is mapped to its features and the similarity is measured to the k-nearest training documents. For each class of documents score is calculated. Based upon the similarity the specific class is classified accordingly. It is an example based classifier that means that the training documents are used for comparison. Finding the nearest neighbours can be quickened using traditional indexing methods.

### 2.2.3 Support Vector Machine

In SVM two sets namely training set and test sets are maintained. Vector form of representation is considered and a kernel function is used to perform the operations. The distance of separation between two classes is called as margin. The SVM text classification [4] yielded good classification results based on decision planes concepts. In this technique it finds an optimal hyperplane with maximal margin to separate two classes. SVM classifier performs best using binary features and provides satisfied test performance in terms of accuracy and speed for number of datasets[5]. However, SVM has less training time when compared to other classifiers.

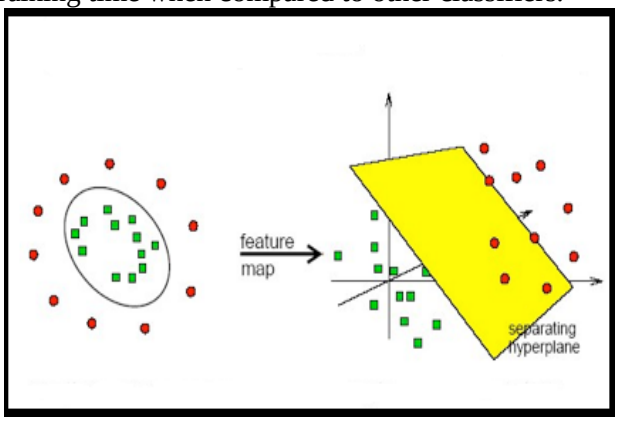


**Fig. 1: Role of non linear classification**

## 2.3 Proposed System

In the proposed system spam classification is done by implementing SVM classifier and ACO algorithm where more number of mails is handled at a time and hence speed of execution increases with better classification performance. The results are compared with the state-of-the-art methods in this challenging area[6].

### 2.3.1 Ant Colony Optimization

Feature selection selects best features from the set of extracted features from the input dataset. Filter methods, wrapper methods and embedded methods are some of the methods implied to perform feature selection [7]. Most optimization problems in engineering are nonlinear with many constraints. To find the optimal solutions to such nonlinear problems efficient optimization algorithms are required. In general, optimization algorithms can be classified into two main categories namely deterministic and stochastic.

ACO is an evolution simulation algorithm proposed by Dorigo et al [8]. Inspired by the behaviors of the real ant colony, they recognized the similarities between the ants food-hunting activities have used to find the shortest route to food source via communication and cooperation [9].

- Set the initial values of parameters.
- Starting from $v_0$, the m ants traverse on the directed graph according to the probability formula on each node. After all the m ants

reach the node $v_n$, m subsets of features are formed.
- Evaluate the fitness of the m feature subsets by classifying the training image sets.
- Select the solution with the highest fitness value found so far as Sbest.

In ACO algorithm, the artificial ants are used to travel in the graph to search for optimal paths according to the pheromone and problem-specific local heuristics information. The pheromone on each edge is evaporated at a certain rate at each iteration. Updation is done by quality of the paths containing this edge.

Artificial ants are usually associated with a list that records their previous actions, and they may apply some additional operations such as local search, crossover and mutation, to improve the quality of the results obtained. The proposed ACO-based feature selection algorithm, ACOFS, reduces the memory requirement and computation time.

In this algorithm, the artificial ants traverse on a directed graph with only $2^n$ arcs. The algorithm uses classifier performance and feature set size to guide search, and optimizes the feature set in terms of its size and classifier performance [10].

A feature selection algorithm selects a subset of important features and removes irrelevant, redundant and noisy features for simpler and more accurate data representation.

Feature selection is implemented using ACO. In feature selection, ACO can be combined with other methods to improve the quality of feature selection and classification. The proposed system presents a hybrid ACO-based classifier model that combines ACO and SVM to improve classification accuracy with a small and appropriate feature subset.

# 3 Architecture

## 3.1 High Level Design

The high level design of the proposed system in a elaborated manner is shown in figure 2. Initially spam datasets are trained. In every stage input is processed and the reducer gives the final output. The tokenized words are nothing but the features extracted from the input dataset. The selected features are given as input to the SVM classifier. Feature selection is performed to select the best set of features out of the extracted features[11]. Selected best set of features is the input given to the SVM classifier. N-fold cross validation is performed to evaluate the system by means of various evaluation measures.

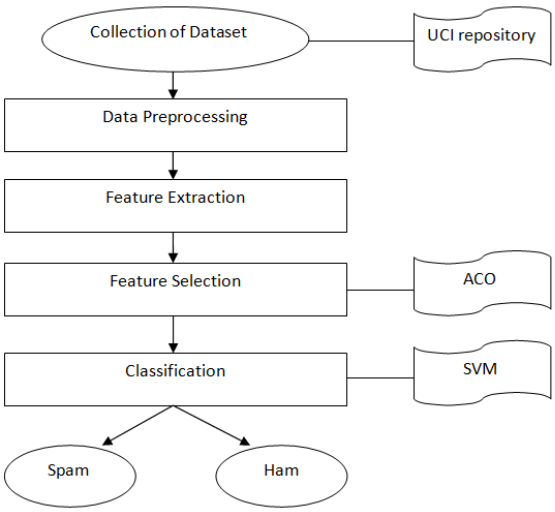


**Fig. 2: Design of the proposed system**

## 3.2 Module Description

### 3.2.1 Collection of datasets

The dataset consists of data for one or more members, with respect to the number of rows. From UCI Machine Repository, Spambase dataset is collected. The Spambase dataset contains 58 attributes, where last column denotes whether it is spam or ham mails respectively. The particular variable for each column represented in the spambase dataset gives the way to the individual database table and statistical data, and each row represents a given member of the dataset. The dataset lists height and weight of an object which determines the values for each of the variables, for each member of the dataset. The value is determined to be as datum.

### 3.2.2 Data preprocessing

In data mining process, data pre-processing is main step for all the machine learning projects. Data-gathering methods leads to loosely controlled, ends with out-of-range values, improper data combinations, incomplete values and so on. Misleading results can produce if analyzing of data is not identified keenly. Hence before running an analysis, the representation and quality of data need to carried over earlier stage itself.

The training of dataset is key for data pre-processing. The data which is in incomplete manner need to be pre-processed so that it allows the whole data set to be processed by means of supervised machine learning algorithm[12]. However most of

recent machine learning algorithms are able to extract knowledge from dataset and stores features in discrete manner. The algorithms can be integrated with a discretization algorithm that transforms them into discrete attributes [13].

### 3.2.3 Feature Selection

A feature selection algorithm selects a subset of important features and removes irrelevant, redundant and noisy features for simpler and more accurate data representation[14]. Feature selection algorithms can also be divided into two categories according to the way they process and evaluate features: feature ranking methods and subset selection methods. As a result, saving in the computational resource, storage and memory requirements could be achieved.

### 3.2.4 SVM Classifier

SVM are supervised learning models with associated learning algorithms that analyze data and it recognize the patterns which can be used for classification and regression analysis[15]. The basic SVM takes a set of input data and predicts, each given input, into two possible classes that forms the output, making it a non-probabilistic binary linear classifier[16].

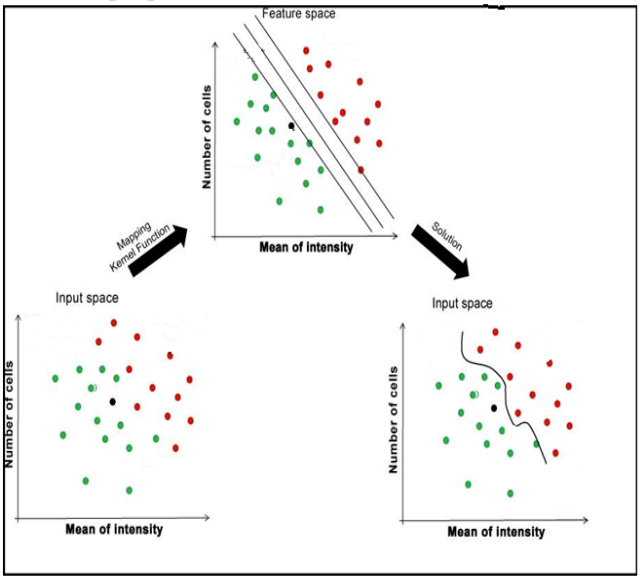


**Fig. 3: Hyperplane in Support Vector Machine**

# 4 Detailed Description

## 4.1 Dataset Processing

The dataset consists of data for one or more members, with respect to the number of rows.

- Dataset characteristics : multivariate
- Attribute characteristics : integer, real
- Associated tasks : classification

Spambase dataset is the dataset taken for training and testing the SVM classifiers, where classification is done using SVM classifier, classify as spam or ham.

## 4.2 ACO for Feature Selection

A feature selection algorithm selects a subset of important features and removes irrelevant, redundant and noisy features for simpler and more accurate data representation. As a result, saving in the computational resource, storage and memory requirements could be achieved. Correctly identifying the relevant features in a text is the vital importance to the task of text classification. Additionally, other methods for reducing features, such as pruning and clustering, can improve performance of text classification [17].

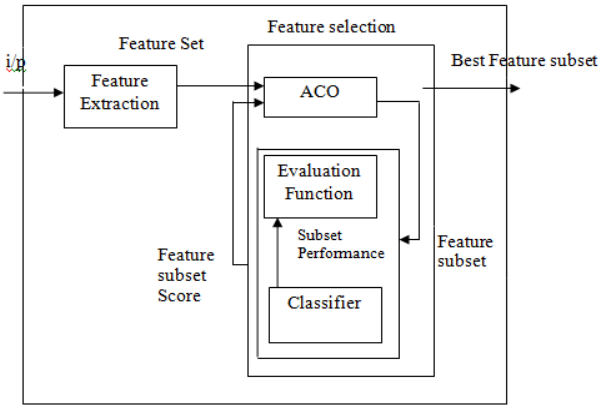


**Fig. 4: Subset Generation in subset generation**

Feature selection algorithms can also be divided into two categories according to the way they process and evaluate features: feature ranking methods and subset selection methods. Ranking of features by a metric way and eliminate all features by means of feature ranking methods which do not achieve an adequate score. Subset selection methods search the set of possible features for the optimal subset. To find the optimal feature subset, the entire search space contains all the possible subsets of features. Feature selection is implemented using Ant Colony Optimization. In feature selection, ACO is also combined with other methods to improve the

quality of feature selection and classification. The proposed system presented a hybrid ACO-based classifier model that combines ACO and SVM to improve classification accuracy with a small and appropriate feature subset. The pheromones are used to determine the transition probability. The classification accuracy and the weight vector of the feature provided by the SVM classifier are both considered to update the pheromone.

### 4.3 Implementation of SVM classifier

Most powerful classifier SVM is implemented over the Spambase dataset. Feature extraction and selection selects the best features from the Spambase datasets. The features are given as input to the SVM classifier. SVM classifier is implemented through N-fold cross validation method in which each mail contained in the dataset has an opportunity to act as both training set and testing set. The entire dataset is divided into N partitions which run through N iterations where in each iteration N-1 partitions are treated as training set and the remaining dataset is treated as testing set. Evaluation measures such as precision, recall and fitness are used to evaluate the system.

## 5 Performance Analysis

On implementing email spam classification system using SVM classifier, an analysis was made over the performance of both ACO-SVM and SVM classification. Performance analysis was made based on the evaluation measures such as accuracy, precision and recall. On analyzing the performance over the above measures, ACO-SVM showed better results when compared to SVM classification on implementing them over Spambase dataset.

The performance of the proposed system compared with the state of art methods like Naïve Bayes Classifier [18], KNN [19] and SVM is tabulated in the Table 1. The comparative analysis show in figure 5 revels that the proposed ACO-SVM based method provides better performance.

**Table 1: Performance analysis of the proposed method**

| PERFORMANCE MEAURES | KNN | NB | SVM | ACO-SVM |
|---|---|---|---|---|
| ACCURACY | 75.25 | 76.24 | 79.5 | 81.25 |
| PRECISION | 70.65 | 70.59 | 79.02 | 87.02 |
| RECALL | 72.05 | 74.96 | 68.67 | 75.1 |

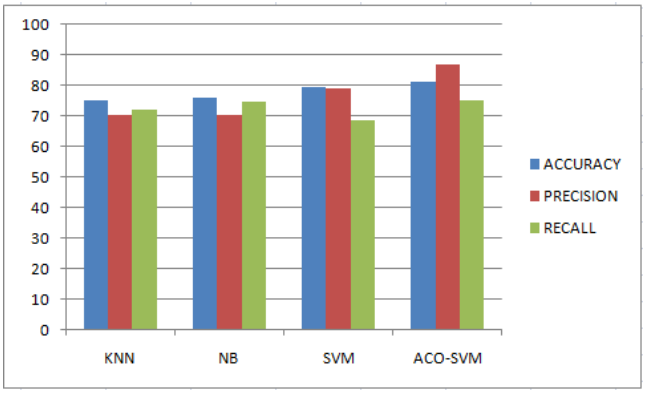


**Fig.5 Comparative Analysis of the proposed method with state of art methods**

The time complexity of proposed algorithm is O(Imn), where I is the number of iterations, m the number of ants, and n the number of original features. In the worst case, each ant selects all the features. As the heuristic is evaluated after each feature is added to the candidate subset, this will result in n evaluations per ant. After the first iteration in this algorithm, mn evaluations will have been performed. After I iterations, the heuristic will be evaluated Imn times.

## 6 Conclusion

Email spam classification done in an efficient manner. This classification done with more than one mail at a time and hence the speed of execution increased. Various factors such as accuracy, precision and recall when considered for both ACO-SVM and SVM classification, ACO-SVM classification shows better results with accuracy of the system as 81% and SVM classification gives an accuracy of 77%. ACO optimization parallelizes the activities which enable the system to classify the test set emails into spam and ham more accurately with better speed in execution time[20]. In the current work a specific dataset is taken for the analysis with SVM classifier. As further enhancements the spam classification technique can be deployed in the mail server and has to be trained to classify the incoming mails efficiently. This enhancement increases the scope further which benefits a large number of users in real time. As a further enhancement the feature selection can be done with other optimization techniques like Firefly [21], Genetic Algorithm, Group Search Optimizer ect. Many other classifiers such as KNN, roughest classifiers can be implemented and the results can be analyzed.

*References:*


[1] A.A. Mousa, Waiel F. Abd El-Wahed, R.M. Rizk-Allah, A hybrid ant colony optimization approach based local search scheme for multi objective design optimizations, *Electric Power Systems Research*, Vol.81, No. 4, 2011, pp. 1014 - 1023.

[2] Bolun Chen, Ling Chen and Yixin Chen, Efficient ant colony optimization for image feature selection, *Signal Processing*, Vol. 93, No. 6, 2013, pp. 1566 -1576

[3] Chi-Yao Tseng, Pin-Chieh Sung, and Ming-Syan Chen, A Collaborative Spam Detection System with a Novel E-Mail Abstraction Scheme, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23, No.5, 2011, pp. 669 - 682.

[4] CL. Huang and CJ. Wang, A GA-based feature selection and parameters optimization for support vector machines, *Expert Systems with Applications*, Vol. 31, No. 2, 2006, pp. 231 - 240.

[5] Harris Drucker, Wu Donghui and Vladimir N. Vapnik, Support Vector Machines for spam Categorization, *IEEE transactions on Neural Networks*, Vol.10, No. 5, 1999, 1048 – 1054.

[6] LY. Chuang, HW. Chang, CJ. Tu, CH. Yang, Improved binary PSO for feature selection using gene expression data, *Journal on Computational Biology and Chemistry*, Vol. 32, No.1, 2008, pp. 29-38.

[7] Marcus Randall and Andrew Lewis, A Parallel Implementation of Ant Colony Optimization, *Journal of Parallel and Distributed Computing*, Vol. 62, No. 9, 2002, pp. 1421–1432 .

[8] M.Dorigo, L.M Gambardella, Ant colony system: a cooperative learning approach to the traveling salesman problem, IEEE Transactions on Evolutionary Computation, Vol.1 No, 1, 1997, pp. 53-66.

[9] Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee and Mohammad Ehsan Basiri, Text feature selection using ant colony optimization, *Expert Systems with Applications*, Vol. 36, No. 3, 2009, pp. 6843-6853.

[10]Rahul Karthik Sivagaminathan and Sreeram Ramakrishnan, A hybrid approach for feature subset selection using neural networks and ant colony optimization, *Expert Systems with Applications*, Vol. 33, No. 6, 2007, pp. 49 - 60.

[11] Shahla Nemati, Mohammad Ehsan Basiri, Nasser Ghasem-Aghaee and MehdiHosseinzadeh Aghdam, A novel ACO–GA hybrid algorithm for feature selection in protein function prediction, *Expert Systems with Applications*,Vol. 36, No. 10, 2009, pp. 12086 - 12094.

[12]W.A. Awad and S.M. ELseuofi, Machine learning methods for Spam Email Classification, *International Journal of Computer Science and Information Technology*, Vol 3, No 1, 2011, pp. 173 -184.

[13] Androutsopoulos, I, Koutsias, J, Chandrinos, K.V, Paliouras, G and Spyropoulos, An evaluation of Naïve Bayesian anti-spam filtering, *Proceedings of the Workshop on Machine Learning in the New Information Age*, Barcelona, Spain, 2000, pp. 9-17.

[14] Chih-chin Lai and Ming-hi Tsai, An empirical performance comparison of machine learning methods for spam e-mail categorization, *Fourth International Conference on Hybrid Intelligent Systems*, 2004, pp. 44-48.

[15] Md. Rafiqul Islam, Morshed U. Chowdhury and Wanlei Zhou, An Innovative Spam Filtering Model Based on Support Vector Machine, *International Conference on Computational Intelligence for modeling, Control and Automation*, 2005, pp. 348 – 353.

[16] G. Rios and H. Zha, Exploring support vector machines and random forests for spam detection, *In Proceedings of the First International Conference on Email and Anti Spam*, 2004.

[17] E. Giacoletto and K. Aberer, Automatic expansion of manual email classifications based on text analysis, *Lecture Notes in Computer Science*, 2003, pp. 785-802.

[18] Guzella, T. S. and Caminhas, W. M, A review of machine learning approaches to Spam filtering, *Expert Systems with Applications*, Vol.36, No. 7, 2009, pp 10206 – 10222.

[19] D. C. Trudgian and Z. R. Yang, Spam classification using nearest neighbor techniques, *In Proceedings of Fifth International Conference on Intelligent Data Engineering and Automated Learning*, 2004, pp. 578-585.

[20]Karthika Renuka Dhanaraj and Visalakshi Palaniswami, Firefly and BAYES Classifier for Email Spam Classification in a Distributed Environment, *Australian Journal of Basic and Applied Sciences*, Vol.8, No.17, 2014, pp 118-130.

[21] Karthika Renuka D and Visalakshi P, Blending Firefly and Bayes Classifier for Email Spam Classification, *International Review on Computers and Software*, Vol. 8, No. 9, 2013, pp. 2168 - 2177.