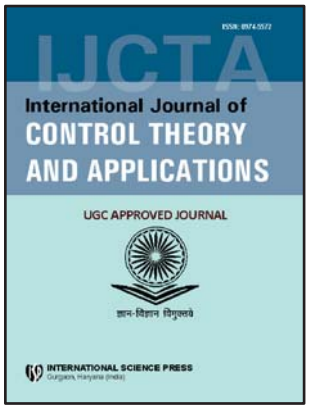

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 30 • 2017

Hybrid CGA Based Naïve Bayes Classifier For E-mail Spam Classification

P.U. Anitha[a] C.V. Guru Rao[b] and D.Suresh Babu[c]

[a]*PhD Scholar, Dept of Computer Science and Engineering, PhD Scholar, JNTU, Hyderabad*

E-mail: anitha_podishetty@yahoo.co.in

[b]*Professor, Dept. of Computer Science and Engineering, S.R. Engineering College, Warangal, Telangana, India*

E-mail: guru_cv_rao@hotmail.com

[c]*H.O.D., Department of Computer Science and Engineering, Kakatiya Government College, Hanamkonda,*

E-mail: sureshd123@gmai

Abstract: In this paper, an efficient spam classification technique is proposed using Naïve Bayes classifier and CGA algorithm. The proposed email spam classification system consists of two phases, such as training phase and testing phase. At first, input email data are given to the feature selection to select the suitable feature for spam classification. Here, Cuckoo search and Genetic algorithm is effectively hybridized to select the suitable features form higher dimensional space using correlation-based fitness function. Once the best feature space is identified through hybrid algorithm, the spam classification is done using the Naïve Bayes classifiers. The experimental validation of the proposed technique is done through evaluation metrics namely, sensitivity, specificity, accuracy. We can also see that our proposed email classification system have outperformed the existing technique in terms of accuracy.

Keywords: *Cuckoo search, Genetic algorithm, lazy classifier and neural network classifier, Email spam.*

1. INTRODUCTION

In recent years, e-mails have become a common and important medium of communication for most Internet users. This fact ensures the efficacy of the advertising messages sent through Internet e-mail [1], [2]. Malicious usage of the electronic data distribution and all other forms of unsolicited communications, also designated as spam, has reached scales never seen before. Every day e-mail users receive lots of messages containing unsolicited, unwanted, legal and illegal offers for commercial products, drugs, fake investments, etc. Spam traffic has increased exponentially in the last few years. During September 2010, the percentage of spam deliveries accounted for about 92% of all Internet e-mail traffic [3]. The average email messages sent daily have reached 3.4 billion in 2012 [4], [5]. According to the recent research from one of the biggest internet service companies, 84.4% of total mail were spam mails [6]. Spam ties up more network resources, reduces the operating efficiency of networks, and consumes a considerable amount of time, money, and energy of receivers; sometimes spam

contains malicious content such as fraud and sexually explicit images, which have a harmful effect on society [7], [8]. The explosive growth leads to several problems, such as unsolicited commercial email, heavy network traffic, and computer worms which are frequently spread via emails. Therefore, a big challenge is to manage the huge number of emails efficiently. To solve this problem, an Email Filtering approach is significantly required [5].

Common email filters should filter incoming email automatically. The filtering process results in a set of categories or classifications, such as spam and ham. The filtering process can be decided and executed based on the email origin or header [9] (*i.e*. source) or based on the email content [5], [10]. Generally speaking, there are two kinds of spam filtering techniques: (*i*) collaborative systems and (*ii*) content-based approaches. The former are based on sharing identifying information about spam messages within a filtering community [2]. The collaborative system monitors the source of the e-mail, which is stored in the domain name and address of the sender device. Such filtering preserves two types email; white-list and black-list. Usually, the new email source is compared with a database to know how it is classified (*i.e*. spam history). In such technique, however, spammers regularly change the email source, address and IP. Therefore, content-based filtering is the second way to review the email content depending on a proposed analysis technique. The content-based filtering are based on a deep analysis of the message content in order to identify its class (usually using machine learning techniques). This work is focused in Naïve Bayes filters, a well-known content-based technique able to accurately combine the probability of finding terms in spam and legitimate messages.

Content-based spam filters analyze the words extracted from the available messages (corpus). Although each term could be a putative feature that should be analyzed for spam filtering, in practice this is not possible because of the huge amount of words extracted from the whole corpus. The usage of large feature vectors with machine learning techniques is not advisable because it can cause the loss of efficiency and accuracy in existing filters [11]. Therefore, several feature selection techniques need to be applied as a pre-processing stage previous to the construction of any spam filtering system. These techniques have been designed to perform dimensionality reduction and their goal is to discard attributes that do not provide essential information for the classification task [2]. Bayesian algorithms are the most popular method because of their convenience of design, decision features, and low storage requirements [12]. However, they also present some problems; for instance, they cannot differentiate the importance of feature words and may misidentify normal e-mail as spam. To solve these problems and improve their filtering capabilities, a new Bayesian spam filtering algorithm that carries minimum risk and is based on the weighting of feature words is proposed in this paper [8]. The remaining of the document is arranged as follows. Various researches performed in relation to our suggested work are presented in Section II. The design approach and the suggested technique are described in Section III and Section IV. The result and discussion of our suggested method are demonstrated in Section V and finally, section VI describes the conclusion of the proposed work.

2. RELATEDWORK

Jieming Yang et al. [13] proposed their binomial hypothesis based feature selection methodology that used for spam detection through the filtering process. Their proposed feature selection method takes the binomial nature of the spam mail with the aid of that it classifies the mails as spam/ham. They made the experiments with various email corpus datasets and they have proved that their proposed feature selection method performed well than the conventional feature selection such as Poisson distribution, information gain and Gini index.

Jieming Yang et al. [14] proposed a feature selection algorithm, they used the inter-category and intra-category as a feature selection method which is also known as comprehensive measure feature selection (CMFS) was introduced and compared it with other common feature selection methods such as Mutual information, Document frequency and information gain etc. They identified the document frequency only considers the intra-

category information which is not concentrated on inter-category information. The inter-category information based feature selection algorithms such as ambiguity measure, DIA association factors failed to concentrate on the intra - category information. With the motivation of this observation, they proposed CMFS method which considers both the intra and inter-category information for the feature selection process. Their experimental results showed that their proposed feature selection improves the performance of the classifier when compared with the existing feature selection algorithm.

Yuanchun Zhu et al. [15] proposed feature selection method based on local concentration, which was applied to spam filtering. This method is inspired from the biological immune system; A local concentration based feature selection was constructed by moving a window that slides over the message to compute the spam and ham concentration. A collection of such concentration is used to represent each document in the dataset. They have used naïve Bayes and support vector machine classification algorithms on the benchmark dataset such as Enron spam, PUI. A local concentration (LC) -based feature extraction method for spam filtering process. Their proposed LC approach was processed with the fixed-length sliding window and variable-length sliding window strategies for transforming every portion of a mail to a corresponding LC feature. In their proposed LC model, two types of detector sets are at first generated by using term selection methods and a well-defined tendency threshold. Then a sliding window is adopted to divide the message into individual areas. After segmentation of the message, the concentration of detectors is calculated and taken as the feature for each local area. Finally, all the features of local areas are combined as a feature vector of the message. Their experimental results depict that their proposed LC feature selection method effectively reduces the dimensions of feature and helps to improve the performance of classifier when compared to existing feature selection methods.

Shrawan Kumar Trivedi & Shubhamoy Dey [16] in their work performed a study on the effect of feature selection methods on machine learning classifiers for detecting email spams. The investigation was conducted on two feature selection methods, namely Genetic Search and Greedy Stepwise Search on Bayesian, Naïve Bayes, Support Vector Machine and Genetic Algorithm classifiers. The tests were conducted on Enron and Spam Assasin datasets. In the experiments Greedy Stepwise Search feature selection, along with Support Vector Machine classifiers resulted in highest classification accuracy of 97.8%.

Yuanning Liuet al [17] a feature selection method proposed using hybrid method, called HBM, which combines not only document frequency information, but also term frequency information in the feature selection process, is proposed. First, an optimal existing DFFS method (called ODFFS) is chosen. Then, terms are determined, whether to be selected by comparing their ODFFS values with a threshold k, which is obtained by a proposed wrapper-based parameter optimization method, called feature subset evaluating parameter optimization (FSEPO). To evaluate the HBM method, we use two classifiers: SVM and NB on four benchmark corpora (PU1, Ling Spam, Spam Assassin and Trec2007). Several DFFSs (information gain, Chi square, improved Gini index and multi-class odds ratio) and TFFSs (normal term frequency based discriminative power measure and comprehensively measure feature selection) are compared with HBM. Experiment results show that HBM is indeed effective and trustworthy.

3. PROBLEM DEFINITION

Recently, various researchers present several algorithms for email spam classification based on classification methods. But, the challenge is not only in finding the spam e-mails and also, how the dimensionality and scalability is taken into consideration for spam classification because, in reality, the processing is with a large and high dimensional data. So, (*i*) curse of dimensionality by handling these criteria, an email spam classification technique is urgently needed for improving the classification accuracy. By solving the above challenge in this research, the feature selection is the main aspect of our research. The feature selection method can solve the curse of dimensionality by identifying the suitable features.

4. PROPOSED METHODOLOGY

The process of proposed email spam classification technique can be explained as following important steps:

1. Prepare the spam dataset (spam base and Ling-spam) for the proposed classifier.
2. Select features from the prepared spam dataset by Cuckoo Search algorithm and Genetic algorithm (CGA) for feature selection.
3. Fitness calculation is done naïve Bayes classifier.
4. Training of NB classifier is done for best feature extraction through Cuckoo Search algorithm.
5. In testing, email classification is done based on the

4.1. Preprocessing Steps

Before the feature extraction process, the email dataset is converted into further process through pre-processing. Pre-processing is done to correct the data from the different errors is taken for further processing. Here, data normalization is done before performing feature relevance analysis. Figure 2 illustrates the attribute description and frequency count for corresponding features. Using this format, discretization process is done and calculate best feature from the dataset.

Attribute number	*Type of attribute*	*Description of attribute*
A1 to A48	Word_freq_WORD	percentage of words in the e-mail that match WORD
A49 to A54	char_freq_CHAR	Percentage of characters in the e-mail that match CHAR
A55	capital_run_length-_average	average length of uninterrupted sequences of capital letters
A56	capital_run_length-_longest	Length of longest uninterrupted sequence of capital letters
A57	capital_run_length-_total	Total number of capital letters in the e-mail
A58	Class attribute	Denotes if the e-mail was spam (1) or not (0)

Figure 1: Attribute or feature description

4.1.1. Discretization

Before the feature extraction process, the training ED_{TR} and testing dataset ED_{TS} are converted the data into specific interval. General form of the training and testing email spam dataset is given by:

$$ED_{TR} = dk\ ;\ 1 \le k \le p \quad (1)$$

$$ED_{TS} = dk\ ;\ p+1 \le k \le q \quad (2)$$

		Number of features						Class (F_{58})
		F_1	F_2	F_3	…	F_{N-1}	F_{57}	
Number of mails	M_1	f_1^1	f_1^2	f_1^3	…	f_1^{56}	f_1^{57}	*Spam or Non-spam*
	M_2	f_2^1	f_2^2	f_2^3	…	f_2^{56}	f_2^{57}	*Spam or Non-spam*
	M_3	f_3^1	f_3^2	f_3^3	…	f_3^{256}	f_3^{57}	*Spam or Non-spam*
	⋮	⋮	⋮	⋮	…	⋮	⋮	*Spam or Non-spam*
	M_{k-1}	f_{k-1}^1	f_{k-1}^2	f_{k-1}^3	…	f_{k-1}^{56}	f_{k-1}^{57}	*Spam or Non-spam*
	M_k	f_k^1	f_k^2	f_k^3	…	f_k^{56}	f_k^{57}	*Spam or Non-spam*

Figure 2: Frequency count taken for each mails based on the attributes

Above training and testing dataset ED_{TR} and ED_{TS} is given to the discretization function to transfer the input data into discretized one. Discretization is a significant step in the data processing to convert the data into specific interval means that the range of values is confined into a specific interval. Here, we have used one discretization function based on the conventional way. The maximum and minimum value of every attribute is identified and the I interval is tracked by taking the ratio between the deviated value and the I value.

For example, at first, deviation is calculated as every *k* value

$$\text{Dev}(k) = \frac{\text{Max}(d_k) - \text{Min}(d_k)}{2} \tag{3}$$

After calculate deviation for each row values, values are converted to the following condition:

$$\left.\begin{array}{ll} 0, & \text{input} < 1(\text{Dev}(k)) \\ 1, & \text{input} < 2(\text{Dev}(k)) \\ 2, & \text{input} < 3(\text{Dev}(k)) \\ 4, & \text{input} < 4(\text{Dev}(k)) \end{array}\right\} \tag{4}$$

Then, every value that comes under within the range is replaced with the interval value so that the input data is transformed to the discretized data. After discretization function, the training dataset is converted to above conversion equation (4) and then each value is converted into 8421 binary conversion as discretized format ED_D.

4.2. Training Phase

The proposed email spam classification consists of two phases.

1. Training phase
2. Testing phase

1. Training Phase: Feature Extraction Using Cuckoo Search and Genetic Algorithm (CGA)

Step 1: Initialization Phase

The population (S*i*, where $i = 1, 2, \ldots n$) of host nest is commenced randomly.

Table 1
Initialization members of FGSO algorithm

	A_1	A_2	A_3	...	A_m	...	A_{56}	A_{57}
Member^k_1	1	1	0	1	0	1	0	1
Member^k_2	0	0	1	1	0	1	0	1
Member^k_3	1	0	1	1	0	1	1	0
•	0	1	0	1	1	0	1	1
Member^k_n	1	0	0	1	0	1	0	0
•	0	1	0	1	0	0	1	0
Member^k_{N-1}	1	0	1	0	0	0	1	1
Member^k_N	1	0	0	0	1	1	0	1

Step 2: Fitness Evaluation Phase

Assess the fitness function based on the equation and after that choose the best one.

$$\text{Fitness} = \left[\frac{\text{Corr}[\text{I},\text{C}]}{\text{Number of } '1's}\right] + \text{classifier accuracy} \tag{5}$$

Where, $\text{Corr}[\text{I}, \text{C}] = \sum_{i=1}^{k} \text{I}_k, \text{C}_k$,

which specifies summation of correlation between selected attributes I and corresponding class.

Step 3: Updation Phase

Revise the first solution by levy flights in which cosine transform is employed. The excellence of the novel solution is assessed and a nest is chosen among randomly. If the excellence of novel solution in the chosen nest is better than the old solutions, it will be substituted by the new solution (Cuckoo). Or else, the earlier solution is set aside as the best solution. The levy flight used for cuckoo search algorithm is described in eqn.

$$\begin{aligned} S_i^* &= S_{(t)}^{(t+1)} \\ &= S_i^{(t)} + \alpha \oplus \text{Levy}\ (n) \end{aligned}$$

Step 4: Crossover operator

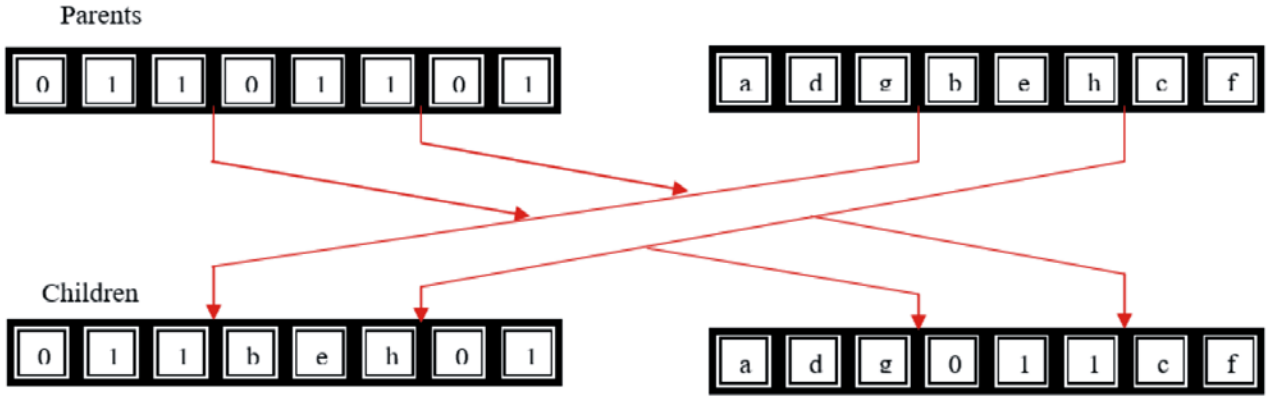

Figure 3: Single point crossover

Once a cuckoo search iteration is finished, the worst solutions are selected to improve the solution quality through cross over rate C*r* = 0.2. In our researches, the most leading genetic operator is crossover, as it generally alters the solution most. A crossover is a process of substituting some of the genes in one parent by consequent genes of the other. The weight optimization problem, the crossover operator is joining two legal parents, whose waits are ordered topologically to produce two offspring's which will as well be legal. The single point cross over operation is briefly explained in fig.3,

Step 5: Termination condition

The algorithm discontinues its execution only if maximum number of iterations is achieved and the solution which is holding the best fitness value is selected and it is specified as best feature to testing process.

4.3. Testing Phase: Email Spam Classification Using NB Classifier

In testing phase, email classification is done by NB classifier. Once the best feature space is identified through hybrid algorithm, the spam classification is done using the Naïve Bayes classifier

1. **To find NB_{err} and NB_{acc} using lazy (naïve Bayes) classifier :** Here, naïve Bayes is utilized to classify the mails and to calculate NB_{err} and NB_{acc}. Naïve Bayes is a binary classifier based on Bayes theorem of probability. By using this classifier, for each instance in the testing spam dataset with the selected significant attributes, algorithm calculates the posterior (spam) Pos(S) and posterior (non-spam) Pos (NS). The algorithm compares and determines which posterior is greater. If the posterior (spam) of the instance I_c is greater, then the instance corresponding instance related to spam mail otherwise it related to non-spam which is represented in the equation (24). The above equations helps to find Pos(S) & Pos (NS).

$$\text{if Pos(S) } (I_c \mid S) \;>\; \text{Pos}(I_c \mid NS) \text{ then } I_c \rightarrow \text{Spam else } I_c \rightarrow \text{Non} - \text{Spam}$$

$$\text{Pos}(I_c \mid S) = \frac{P(S_{tst})\prod_{i=1}^{m} P\left(a_i(I_c)\right|(S))}{\text{evidence}(I_c)}$$

$$\text{Pos}(I_c \mid NS) = \frac{P(NS_{tst})\prod_{i=1}^{m} P\left(a_i(I_c)\right|(NS))}{\text{evidence}(I_c)}$$

$$P(a_i(I_c) \mid (S)) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-\left(d_{bi}-\mu(a_i)\right)^2}{2\sigma^2}\right)$$

$$P(a_i(I_c) \mid (NS)) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-\left(d_{ci}-\mu(a_i)\right)^2}{2\sigma^2}\right)$$

$$\text{evidence } (I_c) = P(S)\prod_{i=1}^{m} P\left(a_i(I_c)\right|(S)) + P(NS)\prod_{i=1}^{m} P\left(a_i(I_c)\right|(NS))$$

Mean : $$\mu(a_i) = \frac{1}{m}\sum_{i=1}^{m} a_i$$

Variance : $$\alpha(a_i) = \frac{1}{m}\sum_{i=1}^{m}\left(a_i-\mu(a_i)\right)$$

The value of probability of spam $[P(S)]_{NB}$ and probability of non-spam $[P(NS)]_{NB}$ of naïve bayes classifier is calculated from the training and testing spam dataset. Finally, error rate NB_{err} is obtained from the equations (34) and (35),

$$[P(S)]_{NB} = \frac{N(S)}{N(M)}$$

$$[P(NS)]_{NB} = \frac{N(NS)}{N(M)}$$

$$NB_{err} = \frac{\text{No. of incorrectly classified mails}}{\text{Total no. of mails in the class}}$$

$$NB_{acc} = \frac{\text{Correctly classified mails}}{\text{Total no. of mails in the class}}$$

After calculated accuracy, finally, email spam classification is done through following equation (44),

$$\text{Result} = \begin{cases} \text{Spam} ; A > T_1 \\ \text{Non-spam} ; A < T_1 \end{cases}$$

Where,

$T_1 \rightarrow$ Threshold

$A \rightarrow$ Accuracy

5. RESULT AND DISCUSSION

This section presents the results obtained from the experimentation and its detailed discussion about the results. The proposed approach of email spam classification is experimented with the Spambase, CSDMC2010 SPAM corpus Datasets and the result are evaluated with the sensitivity, specificity and accuracy and computation time.

5.1. Database Description

Spambase Dataset: The spambase dataset is taken from UCI machinery [31]. The dataset contains 4601 instances of email, of which about 39% is spam. Each instance corresponds to a single email, and is represented with 57 attributes plus a class label (1 if spam, 0 if not). The data files contain one instance per line, and each line has 58 comma delimited attributes, ending with the class label. Most of the attributes indicate whether a particular word or character was frequently occurring in the e-mail; frequency is encoded as a percentage in [0,1]. A few attributes measure the length of sequences of consecutive capital letters.

5.2. Performance Measures

The evaluation of proposed technique in email dataset is carried out using the following metrics as suggested by below equations,

$$\text{Sensitivity} = \frac{\text{Number of True Positives}}{\text{Number of True Positives} + \text{Number of False Negatives}}$$

$$\text{Specificity} = \frac{\text{Number of True Negatives}}{\text{Number of True Negatives} + \text{Number of False Positives}}$$

$$\text{Accuracy} = \frac{\text{Number of True Positives} + \text{Number of True Negatives}}{\text{Number of True Positives} + \text{False Negatives} + \text{True Negatives} + \text{False Positives}}$$

In this paper, we have compared proposed technique of email spam classification technique against our previous technique and some existing techniques such as particle swarm optimization (PSO), neural network. The performance analysis has been made by plotting the graphs of evaluation metrics such as sensitivity, specificity and accuracy. By analyzing the plotted graph, the performance of the proposed email spam classification technique has significantly improved.

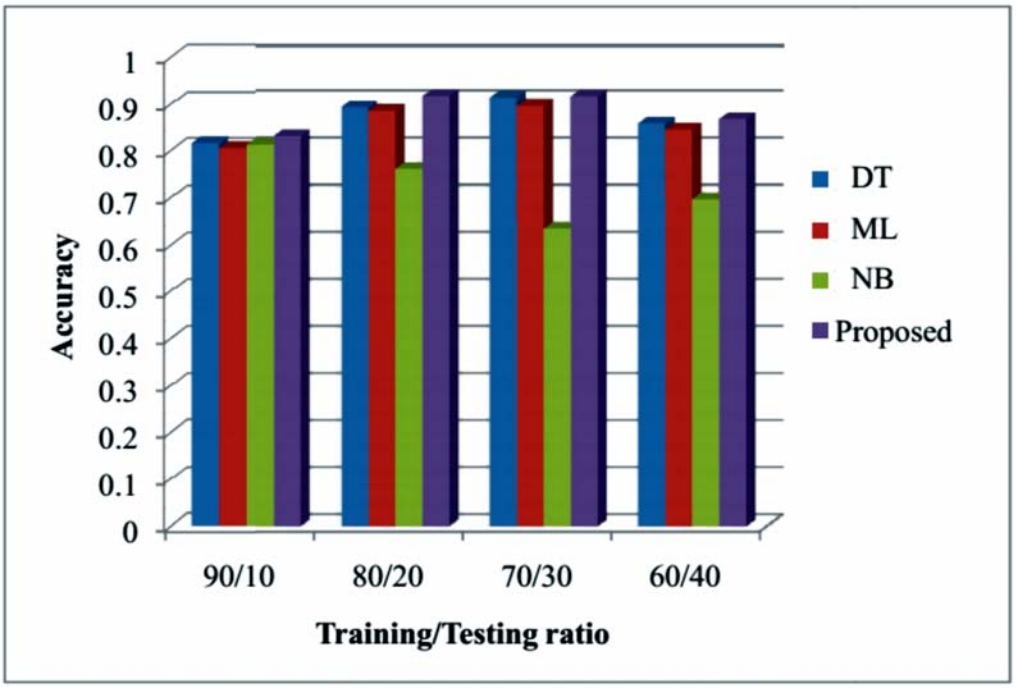

Figure 5: Classification performance of spam base dataset

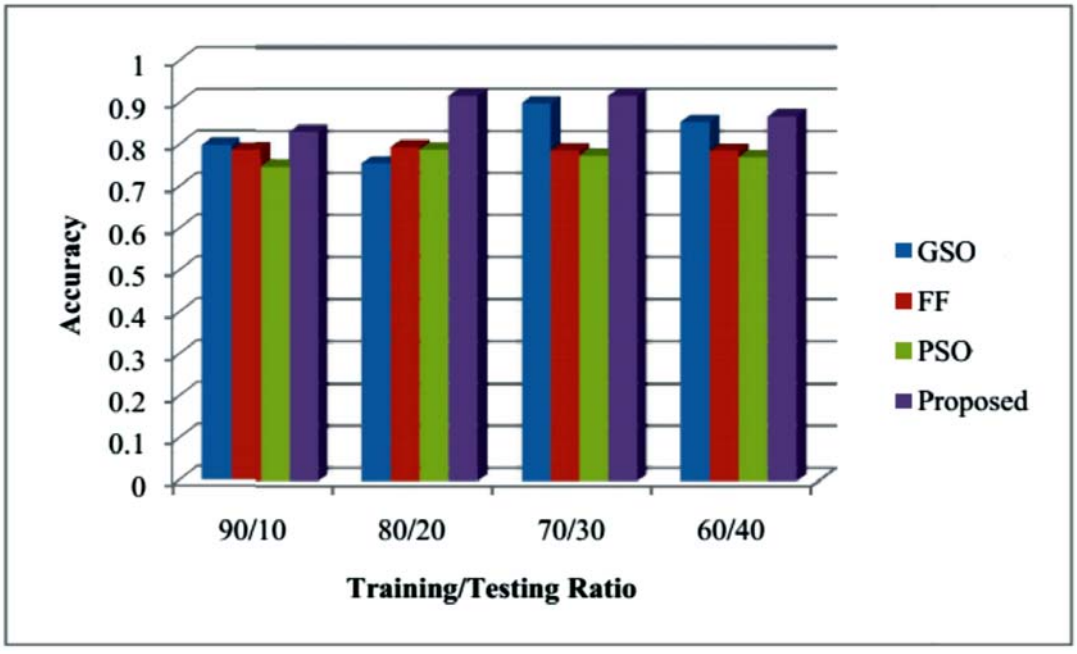

Figure 6: Feature extraction performance of spam base dataset

Table 2
Performance analysis on Classification Accuracy

Techniques	*Training Samples*			
	90:10	*80:20*	*70:30*	*60:40*
DT	0.8	0.87	0.9	0.84
ML	0.79	0.86	0.87	0.82
NB	0.8	0.75	0.62	0.68
Proposed Method	0.82	0.9	0.9	0.85

Table 3
Performance analysis on classification Accuracy

Techniques	*Training Samples*			
	90:10	*80:20*	*70:30*	*60:40*
GSO	0.78	0.73	0.88	0.84
FF	0.76	0.76	0.78	0.77
PSO	0.7	0.75	0.76	0.75
Proposed Method	0.83	0.9	0.908	0.86

Analyzing spam base dataset and figure 5, the proposed approach is achieved the accuracy of about 90.88%, where our previous method has achieved only 79.24%, PSO based technique has achieved only 78.59% and neural network has achieved only 57.19% in training testing ratio 80-20. Fig. 7 and 8 indicates the sensitivity and specificity analysis of proposed method with Neural network and PSO. Table I and II represents the comparative analysis of existing algorithms with proposed work.

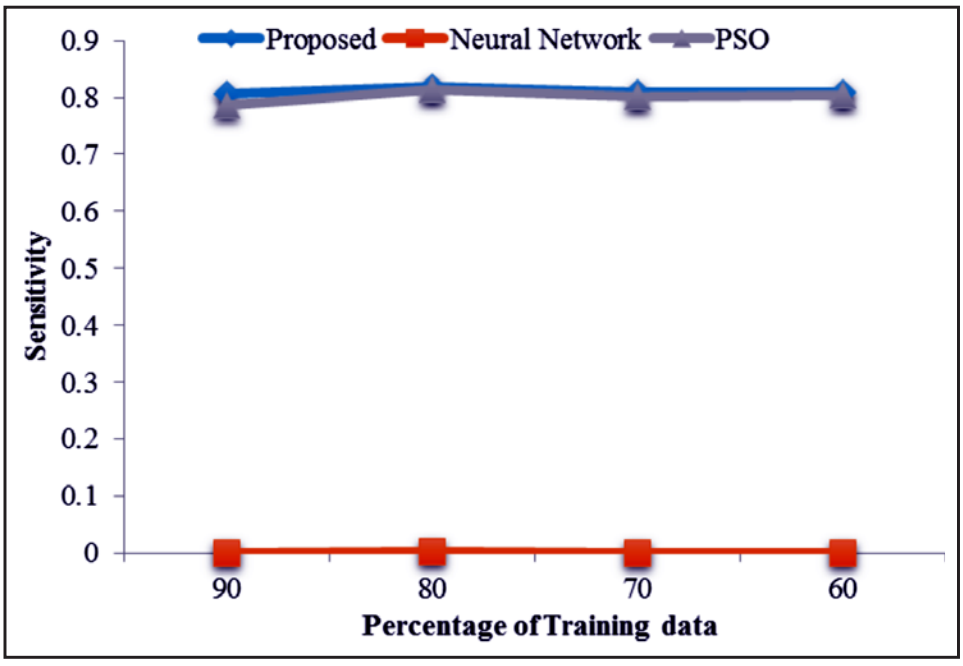

Figure 7: Comparative analysis of sensitivity on spam base dataset

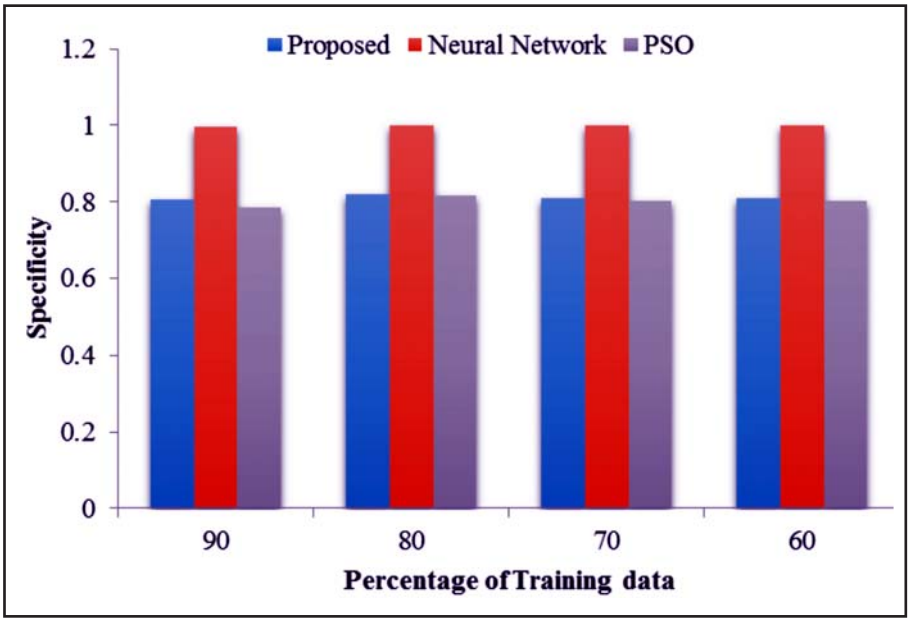

Figure 8: Comparative analysis of specificity on spam base dataset

6. CONCLUSION

E-mail has become one of the fastest and most economical forms of communication. Email is also one of the most ubiquitous and pervasive applications used on a daily basis by millions of people worldwide. However, the increase in email users has resulted in a dramatic increase in spam emails during the past few years. This paper proposes a new email classification system using CGA and NB classifier. The proposed email classification system comprises into two phases, like as training and testing phase. In the training phase, a best feature is selected through optimization algorithm. Once the best feature space is identified through hybrid algorithm, the spam classification is done using the naïve Bayes classifiers. Performance was compared with that of previous techniques and PSO and neural network. The comparison indicates that the proposed email classification system provide better classification accuracy.

In our present work, content-based approach for spam classification, which has earlier brought laurels as regards recognition of the most dangerous gatecrasher, email spam. Whereas earlier investigations on spam classification have invested time and effort focusing mainly on a text based single classifier. This will not accurately identify the spams. So in future we will use multi-classifier for spam classification.

REFERENCE

[1] P. Cunningham, N. Nowlan, S.J. Delany, M. Haahr, "A Case Based Approach to Spam Filtering than Can Track Concept Drift," Proc. 5th International Conference on Case Based Reasoning, Workshop of Long-Lived CBR Systems, pp. 115-123, 2003.

[2] J.R. Méndez, I. Cid, D. Glez-Peña, M. Rocha, F. Fdez-Riverola, "A Comparative Impact Study of Attribute Selection Techniques on Naïve Bayes Spam Filters," Lecture Notes in Computer Science, vol. 5077, pp. 213-227, 2008.

[3] I. Yevseyeva, V.B. Fernandes, and José R. Méndez, "Survey on Anti-spam Single and Multi-objective Optimization," Communications in Computer and Information Science, vol. 220, pp. 120-129, 2011.

[4] S. Radicati, Q. Hoang, "Email statistics report," The Radicati Group Inc., London, 2012

[5] E. M. Bahgat, S. Rady and W. Gad, "An E-mail Filtering Approach Using Classification Techniques," Advances in Intelligent Systems and Computing, vol. 407, pp. 321-331, 2016.

[6] H.J. Kim, H.N. Kim, J.J. Jung, and G.S. Jo, "Spam Mail Filtering System Using Semantic Enrichment," Lecture Notes in Computer Science, vol. 3306, pp. 619-628, 2004.

[7] P. Graham, Will filters kill spam? December 2002," 2003.

[8] H. Wang, G. Zheng, and Y. He, "The Improved Bayesian Algorithm to Spam Filtering," Lecture Notes in Electrical Engineering, vol. 355, pp. 37-44.

[9] C.C. Lai, M.C. Tsai, "An empirical performance comparison of machine learning methods for spam e-mail categorization," Proc. Fourth International Conference on Hybrid Intelligent Systems HIS'04, pp. 44-48, 2004.

[10] M.D. del Castillo, J.I.L. Serrano, "An interactive hybrid system for identifying and filtering unsolicited e-mail," Intelligent Data Engineering and Automated Learning–IDEAL," Lecture Notes in Computer Science, vol. 4224, pp. 779-788, 2006.

[11] A. Jain, and D. Zongker, "Feature Selection: Evaluation, Application, and Small Sample Performance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 2, pp. 153-158, 1997.

[12] X. Zhang, W. Dai, G.R. Xue, "Adaptive email spam filtering based on information theory," Berlin: Springer, pp. 159-70, 2007.

[13] J. Yang, Y. Liu, X. Zhu, X. Zhang, "New feature selection based on binomial hypothesis," Knowledge-Based Systems, vol. 24, no. 6, pp. 904-914, , August 2011.

[14] J. Yang, Y. Liu, X. Zhu, X. Zhang, " A New feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization," Information Processing & Management, vol. 48, no. 4, pp. 741-754, July 2012.

[15] Y. Zhu and Y. Tan, "A Local-Concentration-Based Feature Extraction Approach for Spam Filtering," IEEE Transactions on Information Forensics and Security, vol. 6, no. 2, pp. 486-497, 2011.

[16] S.K. Trivedi, S. Dey, "Effect of Feature Selection Methods on Machine Learning Classifiers for Detecting Email Spams," In Proc. Research in Adaptive and Convergent Systems, pp. 35-40, 2013.

[17] Y. Liu, Y. Wang, L. Feng, X. Zhu, "Term frequency combined hybrid feature selection method for spam filtering," Pattern Analysis and Applications, vol. 19, no. 2, pp. 369-383, May 2016.