© Universiti Tun Hussein Onn Malaysia Publisher's Office

IJIE

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

E-mail Spam Filtering using Genetic Algorithm based on Probabilistic Weights and Words Count

Pronaya Bhattacharya[1*], Arunendra Singh[2]

[1] Department of Computer Science and Engineering, Institute of Technology,
Nirma University Ahmedabad, Gujarat 382481, INDIA

[2]Department of Information Technology, Pranveer Singh Institute of Technology Kanpur,
Uttar Pradesh 209305, INDIA

*Corresponding Author

DOI: https://doi.org/10.30880/ijie.2020.12.01.004
Received 12 October 2018; Accepted 22 January 2019; Available online 31 January 2020

Abstract: Spam email filtering is a hot area of research, as they are growing with time. Most of the spam mails are promotional in nature. Therefore, spam mails are not harmful for the computers, but these mails are annoying for user. Spam mails can be filtered using spam filtering methods like Bayes and Naive Bayes classifications. Classification is done on the basis of content of the mail, or in particular on words and probability is calculated of finding a word from spam and ham classifier words. There are few words which can be found in both spam and ham mails, thus threshold based mechanism is desirable for correct classification. For correct classification using Bayes and Naive Bayes dataset should be huge ideally number of mails should be infinite. But in real applications a scheme is desired which is adaptive in nature and can provide good results with a few mails. In the similar direction, in this paper a genetic algorithm based spam detection method is detailed which is very simple and provide good results with limited dataset.

Keywords: Bayes, Naive Bayes, spam, ham, genetic algorithm

1. Introduction

At present, one of the most crucial elements of communications over internet is E-mail. In spite of this, few minutes of our precious time is spent every day in get rid of spam dealing with products advertisement, offering credit cards, banking mails etc. Though current spam filters based on rules are efficient to make the identification of spam mails and blocking them to inter mailbox. But spammers are always producing advanced methods/techniques to bypass filters and send spam messages to large group of people. It is now quite easy and inexpensive to communicate across the world due to the advancement of technology. Twitter, Facebook, and other social networks are very common means to connect with friends across world. Though, this has also opened a newer audience for spammers to misuse. Spam is not only limited to e-mail any longer, it is on Voice over Internet Protocol (VoIP) in the form of unsolicited marketing or advertising telephone calls, or marketing, publicizing and pornography links on social network. Spam is everywhere. As spams are different from virus, therefore they do not harm our PC, laptop etc. But they are unwanted message which crawl into our mail box. There is no universal definition of spam mails, as depend on user interest a mail can be classified as spam or ham mails. In spite of the fact, a number of researchers all over the world are busy in extensive research with the aim to fight spam; still an effective solution is unavailable. Due to the fact that spam filtering is complex problem, it is not possible to spam emails with one solution. As the spam emails structure is not constant, hence we require a solution which can be adapt as per the spam structure. Therefore, it is necessary that the mail classification method should be adaptive in nature, and adaptation should be in hand of the each mail user. In this paper we propose a genetic algorithm based mail classification method for correct identification of ham and spam mails. This paper is organized into six

*Corresponding author: pronoya.bhattacharya@nirmauni.ac.in

sections. Section 2 of the paper discusses the related work and Section 3 of the paper discusses Bayes and Naïve Bayes classifier. Genetic algorithm is detailed in Section 4 of the paper. In Section 5 experimental results are presented. Major conclusions of the paper are detailed in Section 6 of the paper.

2. Related Work

In past few studies are done on spam email filtering. Nearly all of them are based on content of the e-mail. In the similar context Bayes and Naive Bayes methods in different forms are used [1,2]. Past studies including, artificial intelligence [3], particle swarm optimization [4,5], trusted reports [6], machine learning [7] and genetic programming [8]. However, these papers mostly concentrate on how these methods can be used in spam filtering. The foundation of the work presented in this paper was laid down by Shrivastava et.al., [9,10] in their initial research where authors has described spam classification using genetic algorithm in detail. However, their work only relies on score point and thus accuracy is limited. Recently, Choudhary et.al., carry out extension of Shrivastava et.al., works [11] and proposed probabilistic weight method to improve classification of mails. In this work we further elaborate genetic based spam filtering technique using both weight of spam words and number of words in the test mail. Inclusion of word count increases the accuracy of the email classification. The data dictionary and mail corpus considered in this work is same as considered by Shrivastava el.al. [9] to make comparison fairer.

3. Bayes and Naïve Bayes Classifiers

In our proposed work, both spam words weights and total numbers of words in an email are used for e-mail classification. For better understanding the effect of spam word and total words count on spam classification Bayes and Naïve Bayes classifiers are discussed in this section. Email filtering process is heavily dependent on the content of the mail, or more specifically, number of words and their combinations used. Let us denote number of words in a particular mail (M) as w_1, w_2, ..., w_n . Then the probability of receiving mail is equivalent of receiving words

$$P(M) = P(w_1, w_2, \ldots w_n) \tag{1}$$

But to apply Baye's theorem, all possible word and their combination are needed, therefore required a very large training set. To simplify this, the words can be considered as independent to each other i.e., w_i is independent of w_j (Naive Bayes) [1,2].

In such a case

$$P(M) = \prod_{i=1}^{n} P(w_i) \tag{2}$$

Let us define spam as M_S and ham as M_H. Then we need to compute,

$$P(M_S / M) = \frac{P(M / M_S)P(M_S)}{P(M)} = \frac{P(M_S)\prod_{i=1}^{n} P(w_i / M_S)}{P(M)} \tag{3}$$

Which represents is the probability that a given email is a spam. Similarly, probability that a given email is a ham is

$$P(M_H / M) = \frac{P(M / M_H)P(M_H)}{P(M)} = \frac{P(M_H)\prod_{i=1}^{n} P(w_i / M_H)}{P(M)} \tag{4}$$

Dividing these two equations and taking log we get

$$\log \frac{P(M_S / M)}{P(M_H / M)} = \log \frac{P(M_S)}{P(M_H)} + \sum_{i=1}^{n} \log \frac{P(w_i / M_S)}{P(w_i / M_H)}. \tag{5}$$

If $\log \dfrac{P(M_S / M)}{P(M_H / M)} > 0$, then the given mail is spam otherwise it is ham.

If we further elaborate equation 3, above formulation we get,

$$P(M_S / M) = \frac{P(M / M_S)P(M_S)}{P(M)} = \frac{P(M / M_S)P(M_S)}{P(M_S)P(M / M_S) + P(M_H)P(M / M_H)} \tag{6}$$

Using independence we get,

$$P(M_S / M) = \frac{P(M_S)\prod_{i=1}^{n} P(w_i / M_S)}{P(M_S)\prod_{i=1}^{n} P(w_i / M_S) + P(M_H)\prod_{i=1}^{n} P(w_i / M_H)} \tag{7}$$

However only the value of $P(M_S / w_i)$ is known,

$$P(w_i / M_S) = \frac{P(M_S / w_i)P(w_i)}{P(M_S)} \tag{8}$$

Therefore,

$$P(M_S / M) = \frac{P(M_S)\prod_{i=1}^{n}\left(P(M_S / w_i)P(w_i) / P(M_S)\right)}{P(M_S)\prod_{i=1}^{n}\left(P(M_S / w_i)P(w_i) / P(M_S)\right) + P(M_H)\prod_{i=1}^{n}\left(P(M_H / w_i)P(w_i) / P(M_H)\right)} \tag{9}$$

For a given mail probability of being spam is

$$P(M_S / M) = \frac{P(M_S)^{1-n}\prod_{i=1}^{n} P(M_S / w_i)}{P(M_S)^{1-n}\prod_{i=1}^{n} P(M_S / w_i) + P(M_H)^{1-n}\prod_{i=1}^{n} P(M_H / w_i)}. \tag{10}$$

If probability that a given mail (M) is spam (S) is greater than some pre-defined threshold than it is a spam mails. And probability that given mail is ham is

$$P(M_H / M) = 1 - P(M_S / M). \tag{11}$$

However, the above formulation is based on probability of finding words in different mails, so if particular word is missing then one of product term will be zero, and eventually product is zero and method fails. To improve the method probability of missing words should be left.

As a mail contains a large number of words, it is most likely that few words are common in both spam and ham mails. Therefore, a threshold based mechanism is also desired to rule in favour of spam and ham mails. Moreover, it is also likely that, few words are more likely to be in spam mails, so weighted probability would provide better results. The total number of words in particular mails is also important, as it will affect the probability of getting words in particular mail. To enhance the accuracy, most common words which are likely to be found in both spam and ham mails can be discarded in classification.

The Baye's method heavily relies on word and not on their frequency, but finding some words does not mean that the mail under investigation is spam. For better understanding we consider below example

For excremental point of view, we have considered 2462 ham and 510 spam mails. Here, in table 1, five words 'He', 'I', 'Love', 'Free', 'Offer' are considered and their occurrence in Spam and ham mails along-with defined probability. First, we have considered only word 'Free' for evaluation. We compute probability using equation 5 we get,

Table 1 -Test mails words and probabilistic values

Word	Spam	Ham	$P(S/w_i)$
He	3	473	0.0297
I	11	1376	0.0372
Love	310	347	0.8118
Free	337	471	0.7755
Offer	107	301	0.6318

$$P(M_S)=\frac{510}{510+2462}=0.1716 \text{ and } P(M_H)=\frac{2462}{510+2462}=0.8284$$
$$P(w_i/M_S)=\frac{337}{510}=0.6608 \text{ and } P(w_i/M_H)=\frac{471}{2462}=0.1913$$
$$\log\frac{P(M_S/M)}{P(M_H/M)}=\log\frac{0.1716}{0.8284}+\log\frac{0.6608}{0.1913}$$
$$\log\frac{P(M_S/M)}{P(M_H/M)}=\log 0.2071+\log 3.45=-0.1454$$

Thus mail will be considered as ham mails.
We again compute probability using equation 11, and considering all five words.

$$P(M_S)^{1-n}\prod_{i=1}^{n}P(M_S/w_i)=0.5068$$
$$P(M_H)^{1-n}\prod_{i=1}^{n}P(M_H/w_i)=0.0309$$

We get,

$$P(M_S/M)=\frac{P(M_S)^{1-n}\prod_{i=1}^{n}P(M_S/w_i)}{P(M_S)^{1-n}\prod_{i=1}^{n}P(M_S/w_i)+P(M_H)^{1-n}\prod_{i=1}^{n}P(M_H/w_i)}=0.9426 \quad (12)$$

This mail is declared as spam. Similarly, when only first two words are considered we get,

$$P(M_H/M)=\frac{P(M_S)^{1-n}\prod_{i=1}^{n}P(M_H/w_i)}{P(M_S)^{1-n}\prod_{i=1}^{n}P(M_S/w_i)+P(M_H)^{1-n}\prod_{i=1}^{n}P(M_H/w_i)}=0.9943 \quad (13)$$

Now this mail is declared as ham. This is clear that when more than one word is considered formula 10, can be used for classification. This formula keeps the value of probability around zero for 0 and nearly 1 for spam as in above example. Thus, when all five words are considered then Naive Bayesian method, declares this mail as spam. But when only first two words are considered it decides in favour of ham. However, overall accuracy is limited as probability theory demands that the number of samples ($n\to\infty$) should be very large. This method only considers the words not their frequency in arriving any conclusion. This method is highly susceptible to words and error prone when words which are common in ham and spam are found in mail under consideration.

To solve above problems followings can be done

1. First of all identify SPAM words and develop a spam data dictionary.
2. Divide these words in some fixed groups each group containing particular types of word.
3. Compute group probability instead of individual word probability using above methodology.
4. Use heuristic to decide in favour of ham/spam mails.

This paper proposes a genetic algorithm based methods where number of spam words, frequency and total number of words in a mail are used as parameters in classification of ham and spam mails.

4. Genetic Algorithm

The characteristic vector in a Bayesian filter may contain the frequencies of few words normally selected by human experts. This construction indeed is now and then decisive in the output of the filter. In reference [11] a procedure to build automatically the Bayesian filter is suggested. This technique lays its foundation on the genetic programming. In the similar direction genetic algorithm is proposed. In this algorithm solution is evolved with each subsequent iteration. For this initial population is selected and re-production is performed and to improve the solution crossover and mutation is performed to generate new offspring with better fitness value, this process is iteratively repeated till solution of desired

accuracy is achieved. The GA process steps are shown in figure 1. Fitness values make sure that the offspring's traits are above some definite threshold. Fitness function or fitness value is problem dependent and threshold is also selected in such a way that we converse to the solution of given problem in least number of iterations. In our experiment fitness value depends on both score points and number of words in an email.

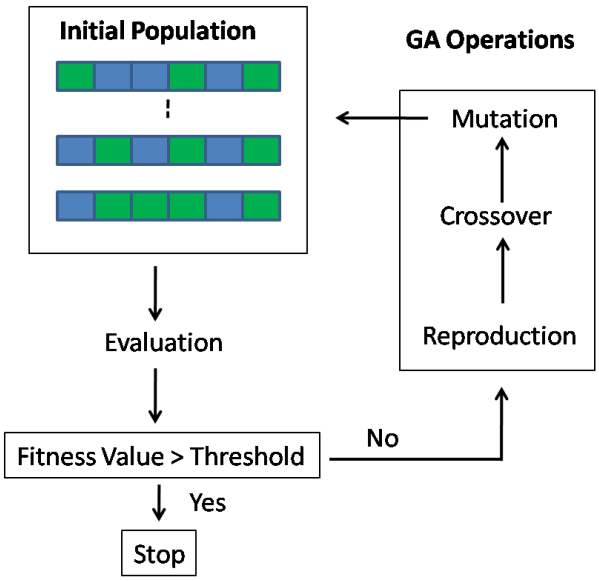

Fig. 1 - Genetic algorithm steps

In this process, first total words in e-mail are counted, and thereafter keywords extraction is done. In the subsequent step weight of each spam word is evaluated, and thereafter genetic algorithm is applied to obtain score point and on the basis this score point decision is made regarding spam or ham mail. The detailed procedure is discussed in next section.

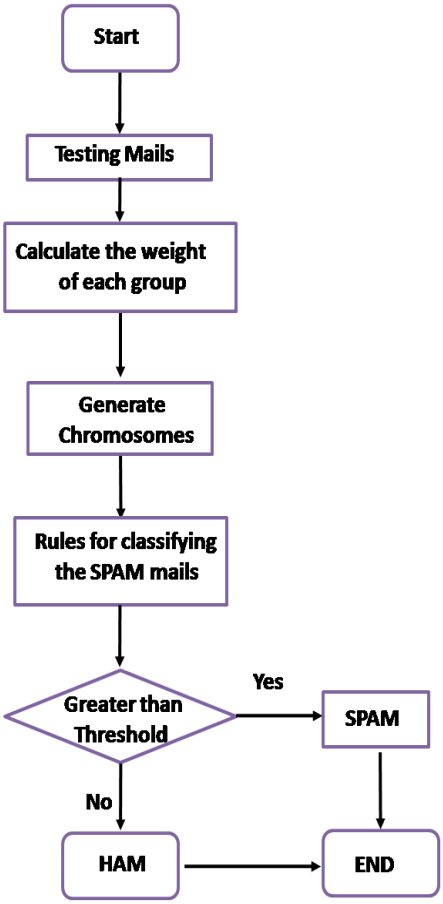

Fig. 2 - Genetic algorithm procedural steps in email classification

In the proposed method, first of all words are identified and for selected words spam data-dictionary is framed. These words are elected by considering spam database mails. The words in data dictionary are divided into several groups (G_1, G_2,…,G_n) and for each mail in database the words from each group and their frequencies are identified and weight of each group is calculated. The obtained weights are converted into binary strings of '0s' and '1s'. Thus, for a mail total numbers of '0s' and '1s' in a string are 10*n*. Let considering that there are *M* emails in database out of which *m* emails are ham and rest M-m are spam mails. This classification is available in advance from the downloaded database. Therefore, there will be *M* chromosomes representing *M* mails, next aim is to design genetic algorithm such that it correctly recognizes both spam and ham mails. For the identification of spam and ham mails concept of score is introduced which is based on string matching of test mail with already classified spam and ham emails. A single gene matching increases score point by one. However, only score point cannot be used in spam and ham classifications, as in emails the number of words varies and mail with more number of words are more likely to have larger score. Therefore, in classification both number of words in an email and score should be considered. The main steps in GA based spam detection are shown in figure 2.

Table 2 - Rules for HAM classification

Score Points	Number of words (email)
0	0-50
1-5	51-100
6-10	101-200
11-15	201-300
16-20	301-400

In spam dictionary 415 words are selected, and are divided into seven groups 'Adult', 'Financial', 'Commercial', 'Beauty', 'Travelling', 'Home based' and 'Gambling'. The number of words in each group varies and detailed in Table 3. For details of words in each group reader can refer to [10]. First of all using these words chromosome string of 70 bits is created, the weight of each group is represented by a string of 10 bits. Now using gene matching score points is evaluated and finally considering both score points and number of words classifications is performed. If criterion given in Table 2, is not satisfied than chosen mail is HAM.

Table 3 - Data dictionary words classification in each group

Group	Number of words
G_1 **(Adult)**	44
G_2 **(Financial)**	68
G_3 **(Commercial)**	86
G_4 **(Beauty)**	109
G_5 **(Travelling)**	25
G_6 **(Home based)**	□□
G_7 **(Gambling)**	20
	Total 415 Words

5. Experimental Results

Completer process is explained using below mention email example. First of all, total words in email are counted, and words which are common to data dictionary words are extracted and their frequency is counted. For a word w_i the weight W_{wi} is given by

$$W_{w_i} = \frac{f}{N} \tag{14}$$

Where, *f* is the frequency of particular word, and *N* is the total number of words in particular mail. The obtained weights for presented words in the mail are shown Table 4.

Test Mail

"Capital One Financial Corporation,.
1680 Capital One Drive,
McLean, VA, 22102-3491
United States

Final Notice of Your Transfer
Attention: Sir/madam xxxxxxxx@gmail.com

In view of your inability to receive your fund with Capital-One Bank from UK Lottery London United Kingdom, originally scheduled to be transferred to you by Capital-One Bank. This fund valued at $5,316,000.00USD [five million, three hundred and sixteen thousand dollars] has been cleared by HM treasury UK, due to incomplete bank details, your fund have not been successfully transferred.

The management, chief financial officer of Capital-One Bank wishes to inform you as the beneficiary to please re-confirm to us the following details below to enable us process and release your fund deposited here by UK Lottery London United Kingdom, we have been mandated to give you access to our data-base so you can view your on-line bank account but it will only be possible by the time you forward your details so we can cross-check with the one forwarded to us.

FULL NAME…….
CURRENT ADDRESS…….. TELEPHONE NUMBER…….
AGE……
SEX……
OCCUPATION…….
NEXT OF KIN……..
RELATIONSHIP WITH NEXT OF KIN……

If I do not remit this money urgently it would be forfeited and subsequently converted to company's fund which will only benefit only the directors of the bank.
.
Upon the receipt of these information we shall send Login Information to enable you view your account online to hence further instruction on how you can transfer the fund to any of your choice bank account. Please do give me a reply with (01helpdesk.capitalone@gmail.com) so that I can send detailed information on the modalities of my proposition to you. I completely trust you to keep this proposition completely confidential.

Yours Respectfully
Chief Financial Officer
Mr. Michael Skepper.
Capital one bank chief financial officer"

First of all we have counted total words in mail which is 295. Next number of words which are common to data dictionary are extracted and for each word frequency and weights are evaluated as shown in Table 4.

Table 4 -Word weight in considered mail

Group	Word	Frequency	Weight
G_1	Sex	1	0.00338
G_2	bank	8	0.02711
G_2	benefit	1	0.00338
G_2	fund	6	0.02033
G_3	online	1	0.00338
G_3	only	3	0.01016
G_3	release	1	0.00338
G_6	receive	1	0.00338

Table 5- Group weight of each group

Group	Weight
G_1 **(Adult)**	0.00338
G_2 **(Financial)**	0.05082
G_3 **(Commercial)**	0.01692
G_4 **(Beauty)**	0
G_5 **(Travelling)**	0
G_6 **(Home based)**	0.00338
G_7 **(Gambling)**	0

The combined weight of each group is shown in Table 5. As in this mail no words are present therefore weight for group G_4, G_5 and G_7 is zero. The weight of the group is represented by 10 bits, where the precision is of the order of 10^{-3}. The chromosome structure using above analogy is

G_1	G_2	G_3	G_4	G_5	G_6	G_7
000000011	0000110010	0000001001	×	×	000000011	×

Fig. 3- Chromosome structure for considered email

where '×' represents '0000000000'.

The experimental results are obtained through computer simulation, for this code is written in Matlab. This code is run on simple machine with i5 processor with 8 GB RAM, the run time is only 3.04678 seconds. For the considered mail score point is 24, thus using Table 2, considered mail is SPAM.

In ham and spam classifications confusion matrix is used (Figure 4). In this case four possible cases are possible:
True positives (TP): HAM mails are correctly identified.
True negatives (TN): SPAM mails are correctly identified.
False positives (FP): HAM mails are in-correctly identified.
False negatives (FN): SPAM mails are in-correctly identified.

	p	n
Y	True Positives	False Positives
N	True Negatives	False Negatives
Column Totals	P	N

Fig. 4- Confusion matrix

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN} \quad (15)$$

$$\text{F-measure} = \frac{2}{\frac{1}{\text{Precession}} + \frac{1}{\text{Recall}}}, \quad (16)$$

F-measure is a test of accuracy, in binary classification. It depends on both precession and recall. Precession is a measure which accuracy and recall is a measure of broadness of correctness. In our case precision is a measure of how accurately spam is identified, and recall measures how many spam mails are correctly identified. Accuracy considers only true cases in classification, while F-measure considers both true and false values in classification. F-measure is direct measure of classifier correctness. F-measure is a broader measure and take into account the variation in data and classification.

Our method is tested on 1100 mails, and obtained results are shown in Table 6. Out of 1100, correctly identified HAM mails are 1009, 59 spam mails are identified as spam mails. 27 ham mails are identified as spams. 5 spam mails are identified as HAM. Precession is 0.9733 and recall is 0.9477.

Table 6- Test Results

Parameters	Value
TP	1009
TN	59
FP	27
FN	5
Precession	0.9733
Recall	00.9477
F-score	0.96

The accuracy is defined as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = 0.971$$

F-measure = 0.96.

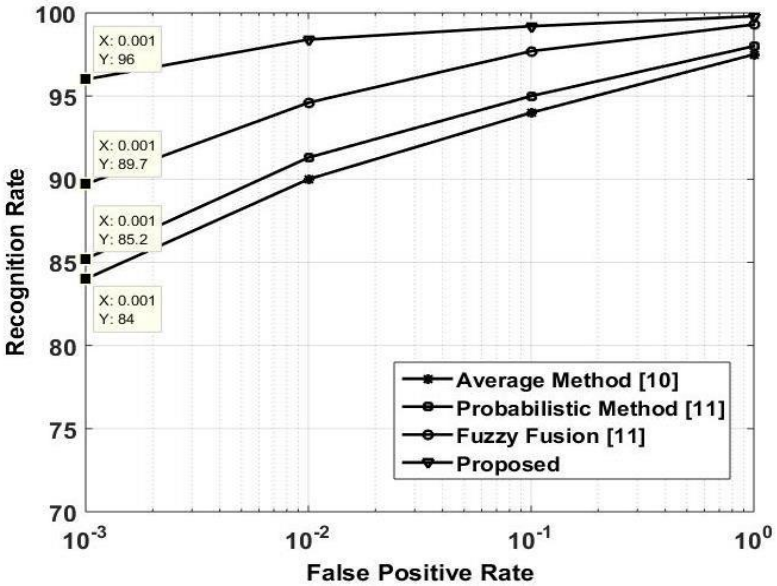

Fig. 5- Recognition rate vs. false positive rate (comparative results)

In figure 5, recognition rate vs. false positive rates are plotted under various methods. Here, as false positive rates are increased, than the recognition rate also improves. In this figure we have compared our results with recently published results. Average method which was proposed by Srivastava et.al., [9] the recognition rate is 84%, with probabilistic method recognition rate is 85.2% [10], and with fuzzy fusion of both average and probabilistic method the recognition rate is 89.7%, [11] while with proposed method the recognition rate is 96%. Therefore, we conclude that in addition to spam words weights the total words count in an email is also very important while deciding in favour of ham or spam mails.

6. Conclusions

This paper set out to investigate the role that genetic algorithm can play in Spam e-mail filtering. As till date spam filtering is an open problem and it is very hard to solve this problem with 100% satisfaction. We are able to get a FScore of 0.96 with GA based method. The major findings of the work are:

1. In SPAM email filtering the fitness function is important and should be selected very carefully.
2. SPAM database is also very important in classifying mail as SPAM and HAM mails.
3. The threshold value of the fitness function cannot be set in advance, it varies with data and type of problem.
4. The word of the data dictionary should be chosen very carefully, as on the basis of these words SPAM and HAM mails will be classified.

Acknowledgement

I would like to thank the reviewers for their valuable comments in improving the quality of this work.

References

[1] Schneider, K. M. (2003). A comparison of event models for Naive Bayes anti-spam e-mail filtering. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, 1, 307314.
[2] Metsis, V., Androutsopoulos, I., & Paliouras, G. (2006). Spam filtering with naive bayes-which naive bayes?. In CEAS, 17, 28-69.
[3] Fdez-Riverola, F., Iglesias, E. L., Díaz, F., Méndez, J. R., & Corchado, J. M. (2007). Applying lazy learning algorithms to tackle concept drift in spam filtering. Expert Systems with Applications, 33, 36-48.
[4] Zhang, Y., Wang, S., Phillips, P., & Ji, G. (2014). Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowledge-Based Systems, 64, 22-31.
[5] Zheleva, E., Kolcz, A., & Getoor, L. (2008). Trusting spam reporters: A reporter-based reputation system for email filtering. ACM Transactions on Information Systems (TOIS), 27, 3.
[6] Lai, C. C., & Tsai, M. C. (2004). An empirical performance comparison of machine learning methods for spam email categorization. In Fourth International Conference on Hybrid Intelligent Systems (HIS'04), 44-48.
[7] Ahluwalia, M., Bull, L., & Banzhaf, W. (1999). A Genetic Programming-based Classifier System. In GECCO, 1118.
[8] Katirai, H. (1999). Filtering Junk e-mail: A performance comparison between genetic programming and naïve bayes. Unpublished manuscript: citeseer. nj. nec. com/katirai99filtering. html, 10.
[9] Shrivastava, J. N., & Maringanti, H. B. (2014). E-mail spam filtering using adaptive genetic algorithm. International Journal of Intelligent Systems and Applications, 6, 54.
[10] Shrivastava, J. N., & Bindu, M. H. (2013). E-mail classification using genetic algorithm with heuristic fitness function. International Journal of Computer Trends and Technology (IJCTT), 4(8), 2956-2961.
[11] Singh, M., & Saxena, P. S. (2017). E-mail Classification using Fuzzy Fusion of Average and Probabilistic Methods. International Journal of Applied Engineering Research, 12, 7816-7822.