E-mail Classification using Fuzzy Fusion of Average and Probabilistic Methods

Mandeep Singh
Research Scholar, Department of Computer Science and Engineering,
Jaipur University, Jaipur, Rajasthan, India.
Orcid Id: 0000-0002-1484-4795

Prashant Sahai Saxena
Professor, Department of Computer Science and Engineering,
Jaipur University, Jaipur, Rajasthan, India.

Abstract

In spite of the fact, a number of researchers everywhere throughout the globe are busy in broad research with the aim to overcome the problem of spam; although an effective solution is yet to be discovered. Due to the fact that spam filtering is perplexing issue, it is unrealistic to spam messages with one arrangement. As the spam emails structure is not constant, hence we require a solution which can be adapt as per the spam structure. In this work, spam filtering methods based on genetic algorithm is highlighted, two methods based on probabilistic and average based is detailed and the results are compared in terms of recognition rate.

Keywords: Spam filtering, recognition rate, probabilistic and average methods.

INTRODUCTION

At present, the spam has turned out to be as one of the greatest irritating issues of the internet. A number of spams are received by each email users day by day and till date we do not have any appropriate way to shield a proper email from turning into a spam target. Even with the defences, there is a possibility unauthentic persons may get the address with the help of a link on a particular url, participation in a contest, subscription to a newsletter, a worm or virus on the pc of a companion [1]. There exists numerous conceivable manners in which an address could act as a source of a number of mails, and as soon as this takes place, there is no real way to prevent those publicists from spreading a surge of "great deals". It is likewise extremely irritating for home clients to stop spam. Organizations make compromises with the spam in numerous ways. Workers squander their quality time in scanning the useful content from genuine mails. Even with this process, we can miss a crucial mail, which enhance along with the spams. The load on the servers goes on increasing with the piling of spam mails. Due to this increased load on servers, the speed of the system will go down and hence to a diminished adequacy of the organization's work process. In this work, a spam filtering method based on genetic algorithm is highlighted, here two methods based on probabilistic and average based are detailed and the results are compared in terms of recognition rate [2].

1. RECENT NOTEWORTHY CONTRIBUTIONS

In the field of spam e-mail filtering a very limited number of papers are available after 2007. This is because spam e-mail filtering is a complex problem and to obtain 100% recognition rate is impossible. Moreover, only one technique cannot work well on all type of e-mails. Still in this section we have discussed some of the notable methods which are emerged over a decade.

SOcial network Aided Personalized and effective spam filter (SOAP) (2011):

Recently social relationship among emails for detecting spam was proposed. SOAP does not rely on keywords but it uses the social relationship among emails to detect the spam [3]. SOAP uses Bayesian filter: social closeness and social interest based spam filtering. However, this technique is not effective in the case of un-correlated mails.

Supervised machine learning approaches for spam e-mail filtering (2012):

In this paper supervised machine learning techniques such as Bayes algorithms, neural network, lazy algorithms, tree algorithms, support vector machines and others for classifying a spam e-mail are discussed [4]. All these techniques rely on learning mechanism, and finally on testing of learned database for spam mail identification.

Binary PSO with mutation operator for feature selection using decision tree (2014):

In this paper, author of the paper proposed a novel spam detection method that focused on reducing the false positive error of wrongly identifying non spam mail as spam. In this method fourfold strategy is considered for various classifications. Finally, the binary PSO with mutation operator is used. This method reduces false positive error without compromising with the sensitivity and accuracy values.

Research Gap and Possible Solution:

E-mail spam filtering is a complex problem, and big giant companies like; Gmail, Yahoo etc. are searching for methods which can control spam. As this problem is industry oriented therefore, most of the e-mail providers do not disclose their spam fighting methods. It is also noticeable that simple one method cannot provide robust solution. Moreover spam mails structure is changing continuously thus it is also very difficult to fight spam in ever changing environment. In this work we focused on an approach which is adaptive in nature and changes done in e-mail structure can be directly incorporated in the proposed solution. Still it is also noticeable that in addition to other technique this technique can improve results significantly.

E-mails Filtering Procedure

The procedure of E-Mail filtering uses various methods to deal with spam mails. In other words, both e-mail addresses and the content is filtered to classify emails more correctly. In any case, both the methodologies need effectiveness and versatility for the basic fact that for latest and developing spam, they should be physically re-altered to adjust to the new alterations [6-8]. As we know that spammers and the processes of diversification are going on increasing day by day, the conventional filter based method is not able to deal with the present spam mails. The steps fixed for spam mails are produced by making use of the genetic algorithm due to the fact that they are advantaged with the reason that any optimization issue which we can portrait with chromosome encoding can be promptly and effortlessly solved.

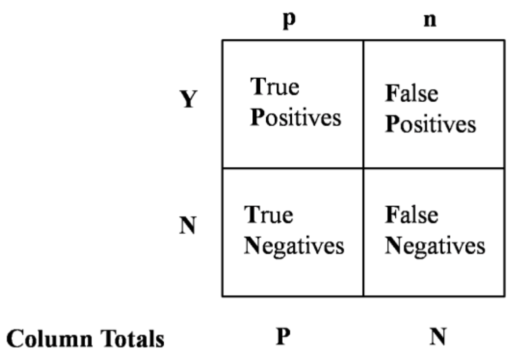
In HAM and SPAM classifications confusion matrix is used (Figure 1). In this case four possible cases are possible:

True positives (TP): These are cases in which we predicted and get HAM mail.

True negatives (TN): We predicted no, and does not receive HAM mail.

False positives (FP): We predicted yes, but they don't actually have the HAM mail. (Also known as a "Type I error.")

False negatives (FN): We predicted no, but they actually do have the HAM mail. (Also known as a "Type II error.")



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

Figure 1: Confusion matrix

$$fp\ rate = \frac{FP}{N},\ tp\ rate = \frac{TP}{P},\ fn\ rate = \frac{FN}{N},\ tn\ rate = \frac{TN}{P},$$

In SPAM and HAM classification, it is NOT possible to declare HAM mails as SPAM as this will lead to the blockage of legitimate mails. Thus *fn-rate* should be zero.

Let the total number of mails are *N*, then we have

N=TP+TN+FP+FN, as *FN* should be zero, then we have

$$N=TP+TN+FP \tag{1}$$

As the arriving mails are independent and identically distributed, therefore pdfs of HAM and SPAM can be considered to be Gaussian, and HAM and SPAM classification depends on the threshold. *TP* and *TN* are true cases, thus only adjustment is possible with *FP*. Therefore by adjusting *fp-rate*, recognition rate can be altered.

Email Classification Process

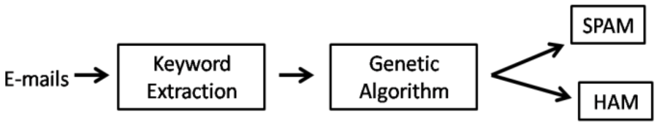


Figure 2: Schematic layout of GA based spam classifier

In the case of genetic algorithm, there exists no significance to the extent regarding the header. So far as that is concerned, the message is contemplated. We extract words from the mail content and these words are used for further processing. Amid the extraction articles and numerical numbers are disposed of [9-11].

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

$$P(M) = P(w_1, w_2, ...w_n) \qquad (2)$$

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) and then various filtering methods are applied. A general layout of Genetic Algorithm based SPAM classifier is shown in Figure 2.

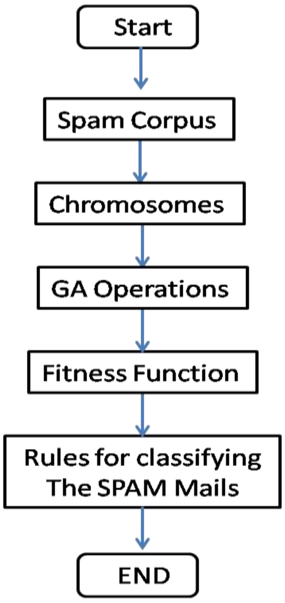


Figure 3: Genetic Algorithm steps for spam classification

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.

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 10n.

When, chromosomes are developed for the approaching mail, the procedure of hereditary calculation begins and crossover happens (Fig. 3). The crossover took into account bits of gene specifically class as it were. In every era of chromosomes just 12% of the whole is crossed. Next takes after mutation, to recuperate a portion of the lost genes. For the situation given above, just 3 % of genes are HAM sends mutated.

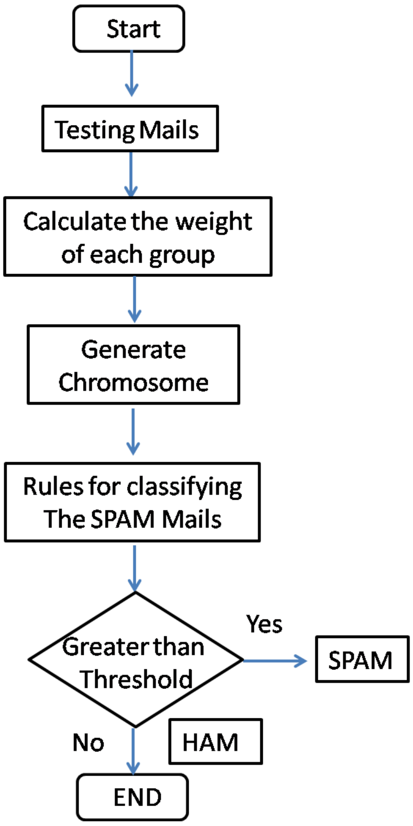


Figure 4: Procedure for spam classification using genetic algorithm

Be that as it may, as the wellness work is in itself issue subordinate and can't be settled at first in SPAM email filtering, the fundamental thought is to discover SPAM and HAM mails at first from among the mails touching base in the inbox of mail.

To define the fitness function, first we conducted experiments on 510 e-mails out of which 315 were spam mails and rest of them were ham mails. It has been found that base score point (sp) figured was 3. Subsequently, the fitness function was characterized as

$$F = \begin{cases} 1 & \text{sp} \geq 3 \\ 0 & \text{sp} < 3 \end{cases} \qquad (3)$$

RESULTS

As said earlier, in genetic algorithm, above all a database is made grouping spam and ham messages and as per the decision gets separated into a few classes. To fortify, the measure of words in the information lexicon increment with the expanded size of database. We have already discussed that the determination of classifications depends on the arrangement of the messages. Indeed, even with the smaller measure of classifications characterized; Electronic mails can at present be recognized as spam mails. Incomprehensibly the disgrace of false positive/negative additionally goes up. In our examination we considered database of 2448 messages among them 1346 were SPAM mails and the left over 1102 were Ham mails [12]. Particular to the case, the data-dictionary includes 421words, which thusly are classified into seven classes. The data dictionary is same as introduced in [13, 14].

The methodology of calculating weights for a word concerning to a specific group is explained as here under:

For instance suppose an email comprises four words, 'Beauty', 'Fitness', 'Holiday' and 'Hotel'. Among these four 'Beauty' and 'Fitness' falls under C_4 and 'Holiday' and 'Hotel' belong to categories C_5 (13, 14).

For example we consider a mail with 823 words, from these words 211 words which are in data dictionary are 'Beauty', 'Fitness', 'Holiday' and 'Hotel', with frequencies 54, 31, 79, 47 respectively are found, and rest are miscellaneous words which are not from data-dictionary.

The removed words that are extracted from the emails are initially checked if they have a place with any of the spam database classification. Weight of the words which are found in data-dictionary is calculated as shown in Table 1.

Table 1: Weights calculations under average method

Category	Word	Frequency	Proportion of a word	Weight of word	Weight of group
C_4	Beauty	54	0.1283	0.0656	0.0516
C_4	Fitness	31	0.073	0.0376	
C_5	Holiday	79	0.1876	0.096	0.0765
C_5	Hotel	47	0.1116	0.057	

The detailed procedure for calculating weights is given below:

In above the word "Beauty' occurs 54 times, hence proportion of word 'Beauty' is 54/421=0.1283.

The weight of a particular word (w_w) is obtained as under

$$w_w = \frac{f_w / t_{wd}}{\sum p_w} \times \frac{s_{wm}}{t_{wm}}, \qquad (4)$$

where

f_w **:** Spam word frequency

t_{wd} **:** Total Data dictionary words

s_{wm} **:** Spam words count in considered e-mail

t_{wm} **:** Total counts of words in e-mail

$\sum p_w$: Total probability of spam words in considered mail

The w_w for the word 'Beauty' is

$$w_w = \frac{54/421}{54/421 + 31/421 + 79/421 + 47/421} \times \frac{211}{823} \qquad (5)$$
$$= 0.0656$$

We estimate the weight of the category with the help of the category average; like the weight of category C_1 is (0.0656+ 0.036)/2=0.0516.

At this point, after the completion of the process of the normalization, we convert the weights within the range of 0.000 to 1.000. Therefore making use of the hex representation

The chromosomes gene binaries values as per the weights can be calculated as

Weights	Binaries
0.000	0000000000
0.001	0000000001
0.002	0000000010
....	....
....	....
0.999	1111100111
1.000	1111101000

As examined over, we encode each mail into chromosomes comprising of 70 bits that are henceforth partitioned into 7 meet categories. Every category of 10 bits speaks to the hex number of the likelihood of the word lying in a specific gathering.

When, chromosomes are built for every one of the mails, the procedure of genetic algorithm begins and crossover happens. As talked about above there are different methods by which cross-over can be performed. Crossover is taken into account bits of gene in a specific class as it were. The general proficiency of the genetic algorithm construct E-mail identification relies on to the extensive number of parameters such as: email informational collection, number of words in the information dictionary, chromosome estimate, size of each group in the data dictionary and like that. The Genetic algorithm based parameters such as traverse, change, populace era technique, choice based measure and wellness work additionally influences the execution as examined in [13, 14].

In the second variation the weight of the group is previously taken by average the weight of the words lying in a group. The weight of particular category is the probabilistic value of the categories. The weight of category C_4 is

$$\frac{0.0656+0.0376}{0.0656+0.0376+0.096+0.057}=0.4028 \tag{6}$$

Table 2, details the weight calculation under probabilistic method. The algorithm was checked on large corpus and it has been found experimentally that nearly 84% mails are classified correctly as ham/spam by average method. The score point ranges from 0 to 189; while with probabilistic method obtain recognition rate is 85.11% with score point ranges from minimum value of 0 to a maximum value of 197. Therefore, probabilistic method is much superior in comparison to the average method.

Table 2: Weights calculations under probabilistic method

Category	Word	Frequency	Proportion of a word	Weight of word	Weight of group
C_4	Beauty	54	0.1283	0.0656	0.4028
C_4	Fitness	31	0.073	0.0376	
C_5	Holiday	79	0.1876	0.096	0.5972
C_5	Hotel	47	0.1116	0.057	

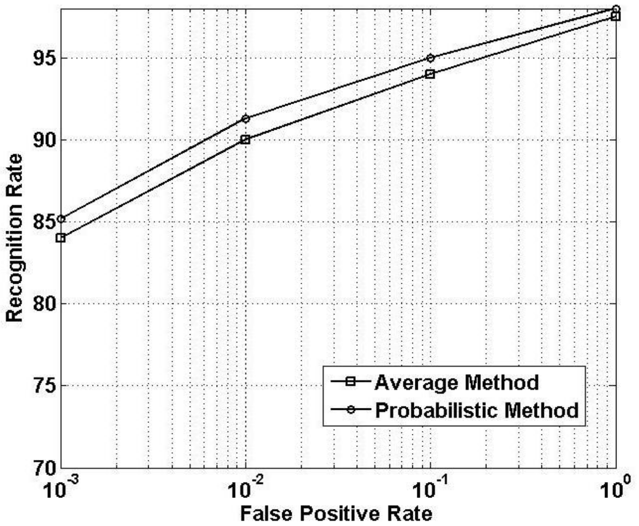


Figure 5: Spam mail recognition rate vs. FPR

In figure 5, recognition rate Vs. False Acceptance Rate (FPR) is plotted. Here at the FPR of 10^{-3} , the obtained recognition rate is nearly 85.2% which rises to 98% at the FPR of 1. Thus using the probabilistic method, recognition rates improves. However, the improvement is marginal.

2. Fuzzy Fusion

A fuzzy membership for a set of data X, can be defined as $X \rightarrow [0,1]$,i.e., to map data on a function which ranges from ‘0’ to ‘1’.

The fuzzy systems are operated on IF-TEHN rules. Like, IF variable IS property THEN action. The general formation of rules is as under

IF *x* IS *a* AND IF *y* IS *b* THEN *z* IS *c*

$I = a \times b \rightarrow c,$

In e-mail classifications vulnerability lies when the value of score point is low. From the experiments we found that up to a score point of 5, vulnerability is high. We normalized the score point values 1 to 5 by dividing all the values by 5. Thus, five membership functions are chosen for, very low (VL), low (L), medium (M), high (H) and very high (VH) values.

Membership Function Type	Range
VL	0-0.2
L	0.2-0.4
A	0.4-0.6
H	0.6-0.8
VH	0.8-1.0

In the same range membership function for average and probabilistic approach are defined. The normalized threshold value is considered to be 0.6. In the output only two criterion need to be meet, like “mail is HAM (H)” OR “mail is SPAM (L)”

In the fuzzy rules we have considered that mail will only be considered as SPAM if any one or both criterion are satisfied.

Example:

IF Average IS VL AND IF Probabilistic IS VL THEN mail IS H

Similarly

IF Average IS VH AND IF Probabilistic IS VH THEN mail IS L

A total of (5x5) 25 rules are defined.

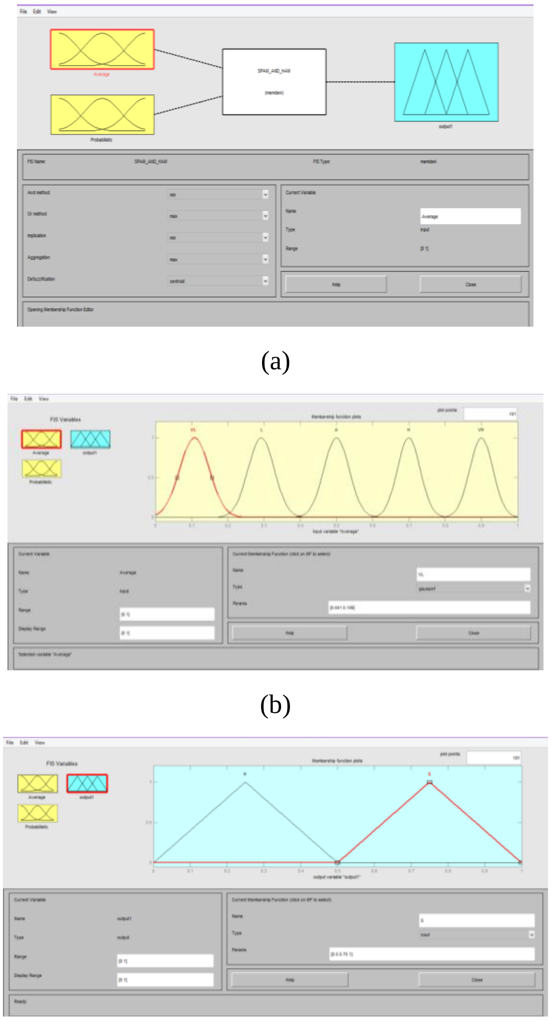
After obtaining the results for each input for defined set of rules, the individual outputs are aggregated using the max function

$\mu_A(x) = \max\{O_1, O_2 ....O_{25}\}$ which unifies the outputs. In this expression O_i is the output of 'i^{th}' rule.

Finally the aggregated output is de-fuzzified using centroid area calculation as

$$X_{COA} = \frac{\int \mu_A(x)xdx}{\int \mu_A(x)dx} \quad (7)$$

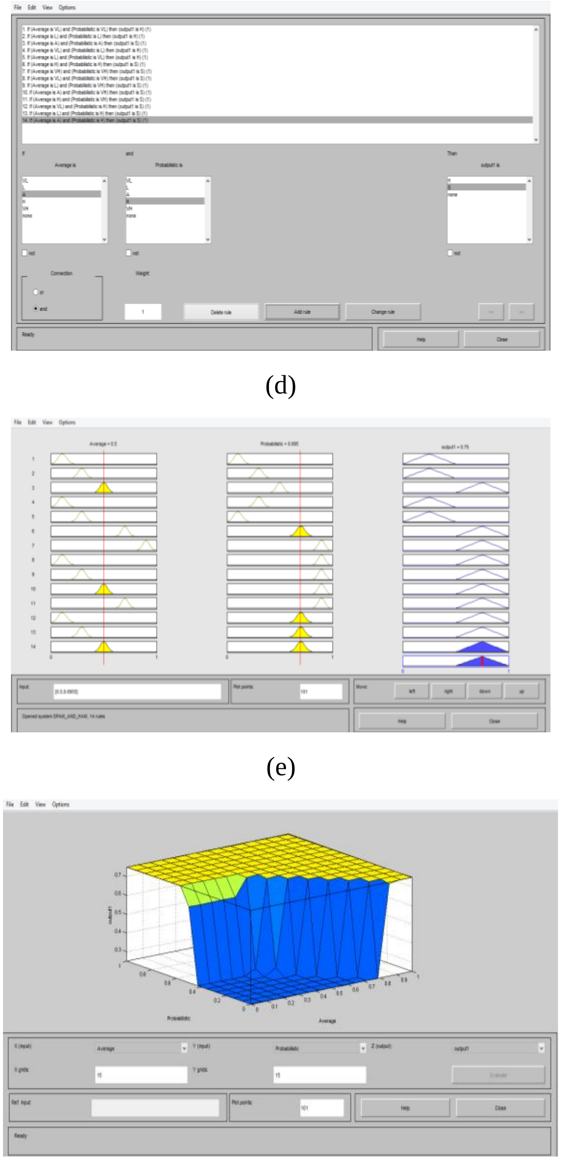
In Figure 6, Fuzzy implementation of fusion is shown. In Figure (a) basic GUI of fuzzy system is shown. The basic system is Mamdani with two inputs one for average and other one for probabilistic measures. For each method (average/probabilistic) five membership functions (Figure b) and two fuzzy output membership functions (Figure c) are shown. In Figure (b) Gaussian membership function is chosen for both the methods is shown. The fuzzy rules set is shown in Figure (d), here in the Figure only 14 rules are considered for example, however a total of 25 rules are constructed. In Matlab the vertical red bars can be changed and corresponding changes in the output can be observed. The output is shown in Figure (e). The surface plot is shown in Figure (f).



(a)

(b)

(c)


1. If (Average is VL) and (Probabilistic is VL) then (output1 is H) (1)
2. If (Average is L) and (Probabilistic is L) then (output1 is H) (1)
3. If (Average is A) and (Probabilistic is A) then (output1 is S) (1)
4. If (Average is VL) and (Probabilistic is L) then (output1 is H) (1)
5. If (Average is L) and (Probabilistic is VL) then (output1 is H) (1)
6. If (Average is H) and (Probabilistic is H) then (output1 is S) (1)
7. If (Average is VH) and (Probabilistic is VH) then (output1 is S) (1)
8. If (Average is VL) and (Probabilistic is VH) then (output1 is S) (1)
9. If (Average is L) and (Probabilistic is VH) then (output1 is S) (1)
10. If (Average is A) and (Probabilistic is VH) then (output1 is S) (1)
11. If (Average is H) and (Probabilistic is VH) then (output1 is S) (1)
12. If (Average is VL) and (Probabilistic is H) then (output1 is S) (1)
13. If (Average is L) and (Probabilistic is H) then (output1 is S) (1)
14. If (Average is A) and (Probabilistic is H) then (output1 is S) (1)

(d)

(e)

(f)

Figure 6: Fuzzy Fusion Implementation in Matlab

Using the fuzzy fusion method only 231 mails out of 1100 mails were tested, this process is applied mails whose score point was in the range of 1 to 5. Rest of the mails were correctly identified by average and probabilistic methods. Figure 7, shows the recognition rate for Average, Probabilistic, and for the fusion of these two processes. It is clear from the Figure as in alone, average and probabilistic methods produces the recognition rate of around 85%. However the fuzzy fusion the recognition rates enhances to nearly 90%, to be prices 89.7%. This happens because mails which are the boundary of HAM and SPAM mails are not correctly characterized by either of the methods. And some of the mails are correctly characterized by one method and not by other. However, using the fuzzy method, performance at the HAM and SPAM boundary can be enhanced.

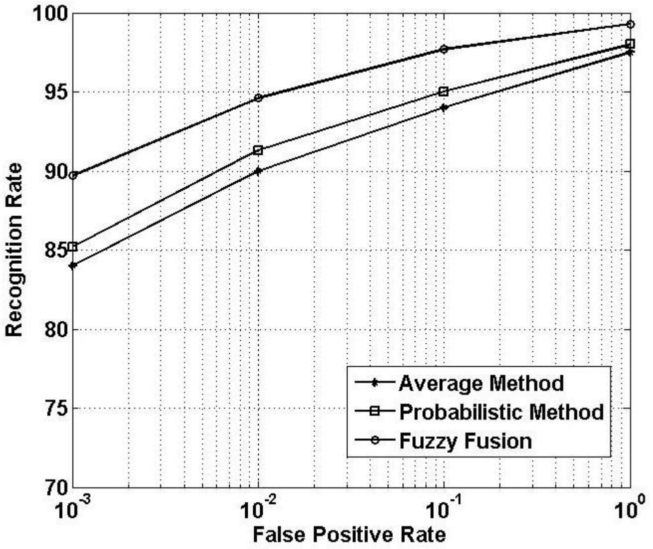


Figure 8: FPR vs. Recognition Rate

Still some of the mails cannot be correctly identified. This result is on expected side as discuss in review chapters, it is not possible to get 100% recognition rate with one method. Fusion of more than one SPAM classification method is necessary to increase recognition rate.

CONCLUSION

In this work for SPAM and HAM mails classifications two methods average and probabilistic is detailed and it has been found that the performance of probabilistic method is much superior to average method. The score point varies from 0 to 197 which in case of probabilistic method while for average method score vary from 0 to 189. The use of fuzzy system in e-mail classification is described. Finally, results of individual process along-with fuzzy fusion are presented. This has been found that using fuzzy fusion the recognition rate can be increased upto nearly 90%.

REFERENCES

[1] E. Blanzieri and A. Bryl, “A survey of learning-based techniques of email spam filtering,” Artificial Intelligence Review, Vol. 2, No. 1, pp.63-92, 2008.

[2] G. Hongyu et al., 2012, “Towards Online Spam Filtering in Social Networks” *NDSS*,12,2012..

[3] Li, Ze, and Haiying Shen. "Soap: A social network aided personalized and effective spam filter to clean your e-mail box." In INFOCOM, 201, Proceedings IEEE, pp. 1835-1843. IEEE, 2011.

[4] Gharibian, Farnaz, and Ali A. Ghorbani. “Comparative study of supervised machine learning techniques for intrusion detection,” In Communication Networks and Services Research, 2007. CNSR'07. Fifth Annual Conference on, pp. 350-358. IEEE, 2007.

[5] Zhang, Yudong, Shuihua Wang, Preetha Phillips, and Genlin Ji. "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection." Knowledge-Based Systems Vol. 64,pp. 22-31,2014.

[6] A. Almomani et al. "A survey of phishing email filtering techniques," IEEE communications surveys and tutorials, Vol. 15, No. 4, pp.2070-2090, 2013.

[7] De Wang, D. Irani, and P. Calton "A study on evolution of email spam over fifteen years." Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference Conference on. IEEE, 2013.

[8] K.S. Tang et.al., “Genetic Algorithm and Their Applications,” IEEE Signal Processing magazine, pp.22-37, 2006.

[9] J. R. Koza, “Genetic Programming: On the Programming of Computers by Means of Natural Selection,” MIT Press, 1992.

[10] J. R. Koza, “Genetic Programming II: Automatic Discovery of Reusable Programs,” MIT Press, 1994.

[11] U. Sanpakdee et.al., “Adaptive Spam Mail Filtering Using Genetic Algorithm,” ICACT 2006.

[12] Spam Assassin,http://spamassassin.org.

[13] M. Choudhary and V. S. Dhaka, “Automatic e-mails Classification using genetic Algorithm” Vol. 6, No. 6, pp**.** 5097-5103, 2015.

[14] M. Choudhary and V. S. Dhaka “E-mail Spam Filtering Using Genetic Algorithm: A Deeper Analysis,” Vol. 6, No. 5, pp. 4266-4270,2015**.**