ISSN:0975-9646

Automatic e-mails Classification Using genetic Algorithm

Mandeep Choudhary , V. S. Dhaka
Jaipur National University,
Jaipur, India

***Abstract*—**In the present scenario, all the email inboxes are barged with spam mails. The escape route that spam mails cater to, lies upon the fact that they are not malicious in nature and hence, generally don't get blocked with firewall or filters ; however, the darker shade is that they are unwanted emails received by any a internet user.
As per the Kaspersky, the percentage of spam in the total email traffic has risen to 70.3% in the first quarter of 2013. This paper discusses a genetic algorithm based method for spam email filtering mentioning its advantages and disadvantages. The results obtained in the paper are promising and suggest that GA can be a good option in conjunction with other e-mail filtering techniques and can provide a better solution. This paper explores the effect of the data-dictionary on the over-all efficiency of the Genetic Algorithm.

***Index Terms*—** Spam Filtering, Genetic Algorithm, SPAM and HAM.

I. INTRODUCTION

Spam can be termed as an "unsolicited email" sent against the interest and knowledge of the recipient, usually without an intention of a response other than to visit a website or sell a product. In general terms these are broadcasted messages sent to a large and varied number of people. However, it becomes significant here to differentiate between unsolicited email, which can be labeled as Spam and solicited email. Solicited emails may serve the same goal as unsolicited emails; however one may receive a solicited email that the sender has deemed to be in your interest, or related to a previous interest. Spam email, however, is usually sent without any knowledge or consideration of the recipient's interest, with the desired aim in mind.
Spams apart from being wastage of money and bandwidth are also very annoying for the users [1].
The percentage of spam amidst the total email traffic during the second quarter of this year rose to 70.7%, 4.2% higher than in the first quarter.

This increase, however, does not indicate the trend; the percentage in the first quarter was an exception to the rule, with a low of 58.3% in January, while all other months depicted spam percentage indicators lying closer to the approximated average of 70% [2] (Figure 1).

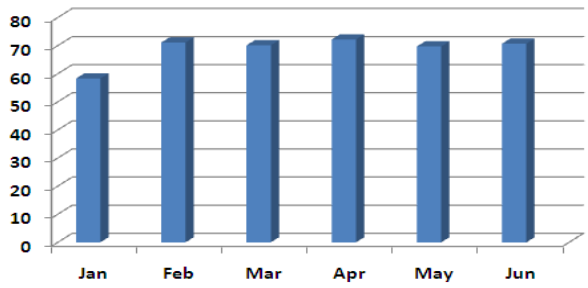

Figure 1: The percentage of spams out of total mails in first quarter 2013 [2]

These slight changes in the percentage of spam in the mail traffic point to a certain level of stabilization, after the sharp gains and falls witnessed during the recent years.

Sources of spam

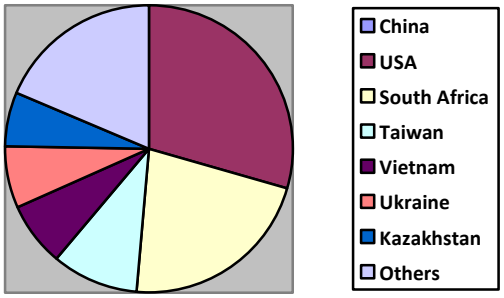

Figure 2: The country wise distribution of sources of spam, Q2 2013. Source [2]

The countries that are among the top sources of spam are the same as before, although their percentages have changed slightly:

To account for, there has been a bit of lowering in the spam originating in the countries like China, US and South Korea by 1.2%, 0.9% and 3% respectively.
Paradoxically countries like Taiwan and Vietnam witnessed a slight increase in the amount of spam (1.6% and 1.1%, respectively), making them stand at the 4th and the 5th ^{p}laces.

The situation with some former Soviet republics is also interesting. Among the three of them — Ukraine, Kazakhstan, and Belarus — the percentage of outgoing spam up surged and in the second quarter, these countries ranked 6th, 7th, and 8th places respectively among the Top 20 sources of spam, pushing Russia downward in the ratings. We also hasten to point out that not only did these three countries demonstrate an increase in the outgoing spam all at the same time, but also the dynamics of these upward movements were very similar, peaking during the month of May (Figure 3) [2].

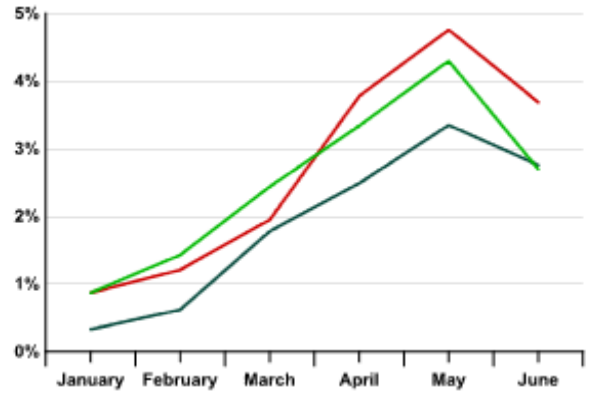

Figure 3: The changes in the percentages of spam originating in Belarus, Ukraine, and Kazakhstan during the first six months of 2013. (Blue-Belarus, Red-Ukraine and Green-Kazakhstan)

This could be a vital indicator of the emergence of new botnets in these countries or the infection of web hosting services from which spam emerges. The remarkable point here is that when we look at the sources of spam by region rather than by country, the geography is altogether different (Figure 4). In Europe, a lot of spam emanates from South Korea (47.9%) whereas the percentage of spam sent from Korea to other regions is quite low. China targets mainly Asia-Pacific (64%) and the US (21.2%), while Europe and Russia observe little to almost no spam coming from China. Most US-based spam ends up in the US (51.6%), and in Russia, spam arrives from Taiwan (12.2%), Vietnam (9.4%), and Ukraine (9%) [2].

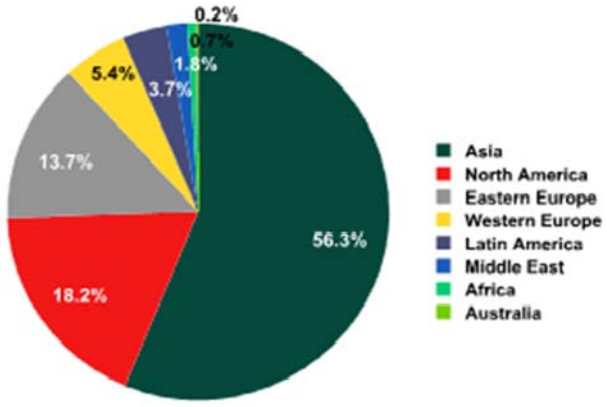

Figure 4: Sources of spam by region Source [2]

As far as the top sources of spam by region are concerned, the ratings did not experience any major changes during the first quarter, although the percentages of specific regions did change a bit. Asia's percentage rose by 4.5%, and remains the number one regional source of spam. Eastern Europe's percentage increased by 2.6% owing to greater activity in Ukraine and Belarus.% of spam originating in Western Europe was 3.7% lower akin to that originating in South America (-2.4%), which reached a new low record. It may be easily recollected that just two years ago, South America ranked second in terms of the amount of spam originating in that region. Other regions with noticeable changes are the Middle East (-0.2%), Africa (-0.6%), Australia and Oceania (-0.04%) [2].

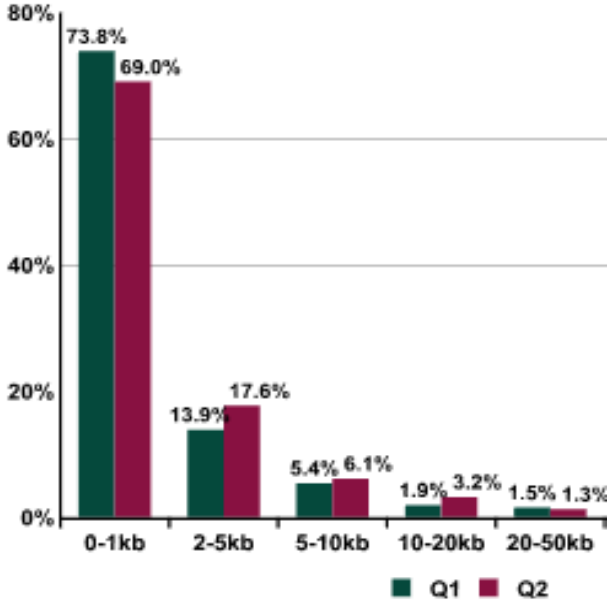

Figure 4: The size of spam emails Source [2]

The majority of spam emails are still very small, weighing under 1Kb.The number of these smaller-sized emails increased from the first quarter by 4.8% and amounted to 73.8% of all spam mails. Incidentally, there was also a slight increase (+0.94%) in the percentage of emails weighing in between 50 and 100 Kb size.
This size is used primarily in emails with attachments, including malicious content [2].

The amount of malicious attachments in the second quarter was lower than the first roughly by 1% and figured out to 2.3% of all the mail traffic [2].

The most prevalent malicious programs spread by emails in the second quarter of this year were the same as that in the first quarter: Readers may recall that [Trojan-Spy.HTML.Fraud.gen], is designed to look like an html page, used as a registration form for online banking services, used by phishers to steal users' financial information.[email-worm.win32.bagle.gt] held onto its second place during the first quarter. This email worm, unlike others, replicates itself to the contacts in a user's address book and also receives remote commands to install other malware.One of the modifications of the notorious ZeuS/Zbot program – Trojan-Spy.Win32.Zbot.lbda – ranked third during the second quarter. ZeuS/Zbot is designed to steal different types of confidential information

from the computers, including credit card data among others.

Trojan-PSW.Win32.Tepfer.hjva landed in the fourth place. This malicious program is designed to steal passwords off user accounts.
Further to scroll down the list there are several more email worms and another two modifications of the ZeuS/Zbot Trojan. Another malicious program featured in the Top 10 is [Backdoor.Win32.Androm.pta].
These types of malicious programs allow malicious users to control infected computers without getting noticed. They can, for example downloads other malicious files, launch them; send a range of data from the user's computer, to name a few. Moreover, computers infected with these programs are often integrated into a botnet.
As discussed above the spam emails are increasing day by day. Till date no full-proof technique is available that can fight against spam. In past various techniques were proposed to fight spam [3], however, none of them provided a robust solution. As the type and content of the spam mails is continuously changing this makes the identification of spam mails more tedious. Therefore, in recent past, studies have been done using the adaptive techniques like artificial neural network and genetic algorithms. In this paper we also propose a simple genetic algorithm based approach for the spam identification, the results presented in this paper are our initial results and found to be promising.
The rest of the paper is organized as follows.
Section 2 of the paper accentuates genetic algorithm based e-mail spam classification. Section 3 of the paper dwells it's foundation over the various steps of the genetic algorithm. E-mail spam filtering process is discussed in section 4 of the paper whereas section 5 discusses the results and finally section 6 concludes over the major topics.

II. E MAILS SPAM CLASSIFICATION USING THE GENETIC ALGORITHM

Genetic Algorithms can identify and exploit regularities in the environment, and converge on the solutions (can also be regarded as locating the local maxima) that were globally optimal [4]. This method is highly effective and widely used to find-out optimal or near optimal solutions to a varied number of problems. Unlike other methods like gradient descent search, random search and others, these algorithms do not impose any limitations required by traditional methods. The Genetic Algorithm techniques have many advantages over the traditional non-linear solution techniques. However, both of these techniques do not always achieve an optimal solution. However, GA provides near optimal solutions easily in comparison to other methods.

A. Advantages

The GA is very different from "classical" optimization algorithms as under:
1) It does the encoding of the parameters, not the parameters itself.
2) It can solve every optimization problem which can be described with the chromosome encoding.
3) The search is more elaborative in a given amount of time.
4) Genetic algorithm is a method which is quite easy to understand and it practically does not demand the knowledge of mathematics.
5) As GA is probabilistic in nature, it may yield "different solutions on different set of simulations". To get an optimal solution Monte Carlo methods can be adopted.
6) It gives solution to the problems with multiple solutions.

B. Limitations

Genetic Algorithms have proven themselves as efficient 'problem solving strategy'. However, they cannot be considered as an ultimate remedy. Some limitations of GA are:
1) Certain optimization problems (termed as variant problems) cannot be solved by means of genetic algorithms.
2) Genetic Algorithm requires the Fitness function to be chosen very carefully. It should be able to evaluate correct fitness level for each set of values.
3) Genetic Algorithms adopt random parameter selection; henceforth it does not work well with the smaller population size where the rate of change is too high.
4) In Genetic Algorithm, solution is comparably better in comparison to the known solutions; it cannot make out "the optimum solution" on its own.
5) Sometimes over-fitness of the fitness function abruptly decreases the size of population leading the algorithm to converge on to the local optimum without examining the rest of the search space. Thi+s problem is also known as "Premature Convergence".

C. Steps in Genetic Algorithms

The details of how Genetic Algorithms work are explained below [5-8].

Initialization

In genetic algorithm initial population is *generated randomly*. However, some research has been done to produce a higher quality initial population more useful for a particular problem.
Such an approach is used to give the GA a comparatively better start and speed up the evolutionary process.

Reproduction

There are two kinds of reproduction: generational and steady-state.

Generational Reproduction

In generational reproduction, the complete population is replaced in each generation. In this method, two mates of the older generation are coupled together to produce two new off springs. This procedure is repeated N/2 times thereby producing N newly generated chromosomes.

Steady-state Reproduction

In this method, two chromosomes are selected randomly and a cross-over is performed producing one or more children.

In some cases mutation is also applied and after crossover and mutation the newly generated off springs are then added again to the original population; thus after some iterations older generation dies out.

Parent Selection mechanism

Generally probabilistic method is used for the parent selection with an intrinsic random nature.

However it does not question the authenticity of GA as being directionless. To be precise, the chance of each parent being selected is related to its fitness.

Fitness-based selection

The standard, original method for parent selection is Roulette Wheel selection or fitness-based selection.

Roulette Wheel picks up on every a chromosome with an equal chance of selection but with the only constraint that the criteria for selection should be directly proportional to the 'fitness' of the chromosome.

The selection rests primarily on the range of fitness values in the current population.

Rank-based selection

'Probability' is the word when it comes to 'Rank-based selection method'. Rather than basing the search on absolute fitness, the chromosomes are selected on the basis of relative rank or position in the population.

Tournament-based selection

The tournament based selection is to choose N parents randomly, finally returning the fittest among them.

Crossover Operator

The crossover is one of the most important operators in GA. From one generation to the next, the operator alters the programming of the chromosomes.

The process involves recombining bit strings through the exchange of segments between pairs of chromosomes. To account for, there are various kinds of crossovers:

One-point Crossover

One point cross-over randomly selects a bit position to change. The crossover position is decided upon by generating a random number that might fall short of the chromosome length or be equal to it.

Here, the bits before the number are kept unchanged and the bits after the crossover position are swapped between the two parents.

Two-point Cross Over

The two point cross-over, is similar to that of one-point crossover except that here two positions are selected randomly and only the bits between the two positions are swapped. This crossover preserves the first and the last parts of a chromosome and just swaps the middle part.

Uniform Crossover

In a uniform cross-over, each gene of the first parent has a definite probability (generally 0.5) of getting swapped with the corresponding gene of the second parent.

Inversion

Inversion is a type of reordering technique. Here a single chromosome is chosen and the order of the genes goes under inversion between two randomly chosen points.

As the operator is inspired by a natural biological process some additional overhead is required.

Mutation

Mutation is inspired from the concept of biological mutation. It ensures that all the possible chromosomes can maintain better genes in the newly generated ones. With crossover and even inversion, search is constrained to alleles which exist in the initial population so as to preserve initial characters.

The mutation operator overcomes this by randomly selecting any bit position in a string and changing it,

in consonance with the need. This is useful since crossover and inversion may not be able to produce new alleles if they do not appear in the initial generation and a newer type of chromosomes can be generated with older and newer characters.

III. E MAILS FILTERING PROCESS

The process of E-Mail filtering works on two aspects of filtering; one filters e-mail addresses while the other, the e-mail content. However, both the approaches lack intelligence and adaptability for the simple reason that for newer and emerging spam, they must be manually re-amended to adapt to the new modifications. With spammers and means of diversification sprouting up, the traditional filter based approach finds it difficult to adapt to the newly generated spam mails. The rules set for spam mails are developed using the genetic algorithm as they are privileged with the fact that any optimization problem which can be described with chromosome encoding can be readily and easily solved with them and furthermore these algorithms are apt in solving multiple solution problems.

A. Rules for classifying the emails:

The weight of words of gene in the test mail is compared with those in the spam mail prototypes and the matching gene is found. If the matched gene is greater than some number let say 'x' then mail is considered as spam.

Fitness Function:

$$F = \begin{cases} 1 & \text{SPAM mail} \\ 0 & \text{Ham mail} \end{cases}$$

However, as the fitness function is in itself problem dependent and cannot be fixed initially in SPAM email filtering, the basic idea is to find SPAM and HAM mails initially from among the mails arriving in the mail box.

For the evolution of the fitness function an experiment was carried out on 500 mails which consisted of pool of 300 SPAM and 200 HAM mails and the minimum score point calculated was 3. Hence, the fitness function was defined as

$$F = \begin{cases} 1 & \text{Score point } \geq 3 \\ 0 & \text{Score point } < 3 \end{cases}$$

Procedure:

Figure 5: Schematic layout of GA based spam classifier

There stands no relevance as far as the header is concerned, in genetic algorithm. For that matter, the message is taken into consideration. From the body of the mail, words are extracted. During the extraction articles and numerical numbers are discarded.

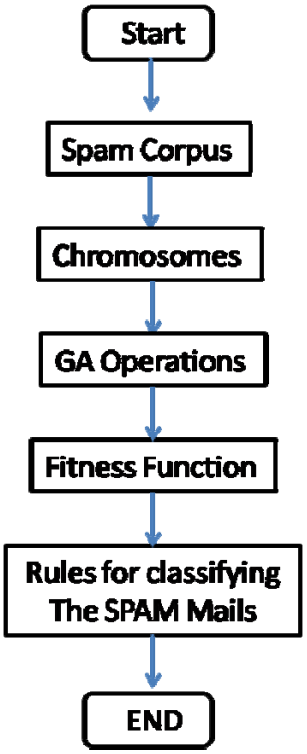

Figure 6: Flow diagram for the GA based spam classifier

To start with, database is created, classifying spam and ham emails, and as per choice database gets divided into several categories. The point of emphasis here is that as the size of the database increases, the number of words in the data dictionary increase with the increasing size of the database. Email's classification decides upon the selection of the categories. However, if lesser number of categories is defined, still email can be identified as spam mail, whereby the chances of false positive/negative increases. Once, chromosomes are constructed for the incoming mails the process of genetic algorithm starts and crossover takes place. As discussed above there are various ways by which cross-over can be performed The crossover is only allowed for bits of gene in particular category only. In our algorithm, both multi-point and single point crossover is done and positions of bits are selected randomly. In each generation of chromosomes only 12% of the total is crossed. Next follows mutation, to recover some of the lost genes. In the case given above, only 3 % of genes are HAM mails mutated.

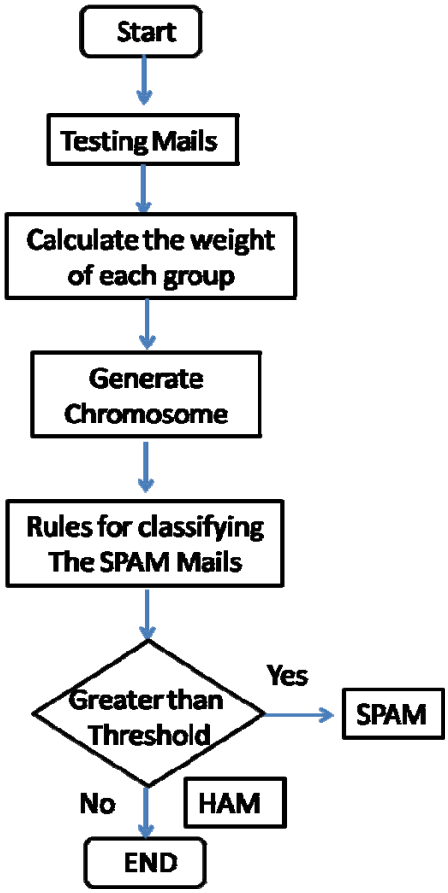

Figure 7: Flow diagram for the Genetic approach for the spam classification

As specified in the rules for classifying mails, in our particular example also, the weight of the words of gene in test mail and those in the spam mail prototypes are compared to find the matching gene. If number of matched gene, is greater than or equal to three, then spam mail prototype shall receive one score point.

On the other hand, if the score point is greater than some threshold then the mail is considered to be a spam. However, the threshold point can be manually adjusted to get the appropriate results as we 'parked' it by performing an experiment on 500 emails.

IV. RESULTS

As mentioned previously, in genetic algorithm, first of all a database is created classifying spam and ham emails and in accordance with the choice gets divided into several categories. To reinforce, the number of words in the data dictionary increase with the increased database size. As discussed previously, the selection of categories is based on the classification of the emails.

Even with the lesser number of categories defined,

Electronic mails can still be identified as spam mails.

Paradoxically the stigma of false positive/negative also increases. In our experiment we considered database of 2448 emails out of which 1346 were SPAM mails and rest 1102 were Ham mails. Specific to the case, the data-dictionary comprises of 421words, which in turn are divided into seven categories.

The data dictionary is presented in appendix A.
The procedure of calculating weights for a word belonging to a particular group is detailed below:
As an example let an email consist of four words, 'adult', 'porn', 'free' and 'offer'. Out of these four 'adult' and 'porn' belong to categories C_1 and 'Free' and 'offer' belong to categories C_3 (see Appendix –A in [10]).

Let us consider an email with 797 words, out of which 685 words are 'adult', 'porn', 'free' and 'offer', with frequencies of occurrence to be 107, 31, 466, 81 respectively.

These words are taken so large in number, so as to make sure that the mail in consideration is a spam mail, as the spam database is very small containing only 421 words.
The extracted words from the emails are first marked whether they belong to any of the spam database category. In case the words in the email match with those in the spam data dictionary, then the probability of getting a word from the spam database is obtained by dividing the frequency of a spam word by total number of words in the data dictionary.

In our case "adult' occurs 107 times, hence probability of getting 'adult' word is 107/421=0.254.
The weight of the word (W_w) is calculated as under

$$W_w = \frac{F_w / T_{WD}}{\sum p_W} \times \frac{S_{WM}}{T_{WM}}$$, where

F_w **:** Frequency of spam word

T_{WD} **:** Total word in data dictionary

S_{WM} **:** Total spam word in e-mail

T_{WM} **:** Total word in e-mail

$\sum p_W$: Probability of getting a word

The p_w f or the word 'adult is

$$W_w = \frac{F_w / T_{WD}}{\sum p_W} \times \frac{S_{WM}}{T_{WM}}$$

$$W_w = \frac{107/421}{107/421 + 31/421 + 466/421 + 81/421} \times \frac{685}{797}$$

$$W_w = 0.134$$

The weight of the category is calculated by taking the average of the category; as an example the weight of category C_1 is (0.156+ 0.045)/2=0.101.

Thus the obtained weights of each of the words are tabulated in the underlying Table:

Table 1: Calculation of weights under average weight age method

Group	Word	Frequency	Probability of getting a word	Weight of word	Weight of group
C_1	adult	107	0.254	0.156	0.101
C_1	porn	31	0.074	0.045	
C_3	Free	466	1.107	0.680	0.399
C_3	offer	81	0.192	0.118	

Then after the normalization the weights are converted in the range of 0.000 to 1.000. Thus using the hex representation we have:
The weight of the gene can be encoded as
Binary 0000000000 represents weight 0.000
Binary 0000000001 represents weight 0.001
Binary 0000000010 represents weight 0.002
……
…….
Binary 1111100111 represents weight 0.999
Binary 1111101000 represents weight 1.000

As discussed above, each mail is encoded into chromosomes consisting of 70 bits, which are hence divided into 7 equal groups. Each group of 10 bits represents the hex number of the probability of the word lying in a particular group.
Once, chromosomes are constructed for all the mails, the process of genetic algorithm starts and crossover takes place. As discussed above there are various ways by which cross-over can be performed. Crossover is only allowed for bits of gene in a particular category only.
Our algorithm use both multi-point and single point crossovers. Positions of bits are selected randomly. In each generation of the chromosomes only 12% are crossed. Next follows mutation so as to recover some of the lost genes or in our case it is done to recover some of the lost data. Specific to our case, only 3 % of the genes are mutated.
The over-all efficiency of the genetic algorithm based E-mail identification depends on the large number of parameters like: e-mail data set, number of words in the data dictionary, chromosome size, size of each group in the data dictionary and so on.
On the other hand the type of mail also affects the performance of GA based filtering techniques like url, image type, text type etc.
The Genetic algorithm based parameters like cross-over, mutation, population generation method, selection based criterion and fitness function.
Many re-searchers all over the world are therefore trying to investigate some good solution to such a complex problem.
This paper presents the method of the spam identification.
In table 2, the results are presented by varying the total number of words in the data dictionary.
However, the point of emphasis here is that we have kept on the total seven groups and each group contains nearly same number of words. The total numbers of tested mails are 659 and the obtained results are depicted below in table 2.

Table 2: Efficiency of the GA method by varying the number of words in data dictionary

Number of words in data-dictionary	Efficiency
100	24.11
150	36.23
200	47.56
250	56.34
300	63.97
350	72.09
421	81.7

V. CONCLUSIONS

This paper presents a Genetic Algorithm based 'e-mail spam classification algorithm' along with some basic results. The algorithm is able to successfully distinguish between spam and ham e-mails. The efficiency of the process depends on the dataset and the GA parameters and turns out to be more than 81%.

It is also shown that the number of datasets in the data dictionary have a deep impact on the over-all efficiency of the genetic algorithm based e-mail classification.

It is intended to minimize the false positive/negative results in the future. Also some advanced results shall be presented, pertaining to the characterization of the GA parameters.

Hence, GA in conjunction with other e-mail filtering techniques can provide more accurate SAPM filtering techniques.

REFERENCES

[1] Enrico Blanzieri and Anton Bryl, "A Survey of Learning Based Techniques of Email Spam Filtering," Conference on Email and Anti-Spam,2008

[2] http://usa.kaspersky.com/internet-security-enter/threats/spam-statistics-report-q2-2013.

[3] Gordon V. Cormack, Email Spam Filtering: A Systematic Review

[4] G. J. Koprowski. Spam accounts for most e-mail traffic, Tech News World (2006). http://www.technewsworld.com/story/51055.html

[5] K.S. Tang et.al., "Genetic Algorithm and Their Applications"IEEE Signal Processing magazine, pp.22-37, Nov. 1996.

[6] Goldberg, D. E. 1989a. Genetic Algorithms in Search, Optimization, and Machine Learning.Addison−Wesley

[7] Goldberg, D. E., and Deb, K. 1991. A comparitive analysis of selection schemes used in genetic algorithms.In G. Rawlins, Foundations of Genetic Algorithms. Morgan Kaufmann.

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

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

[10] UsaratSanpakdee,et.al., Adaptive Spam Mail Filtering Using Genetic Algorithm" ICACT 2006"

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

AUTHORS' PROFILES

Mandeep Choudhary

Is a Research Scholar at the Jaipur National University, India. He is M Tech in Computer Science. His current research interests include Artificial Intelligence, genetic algorithm based optimization. Spam and Ham e-mail classification.

V. S. Dhaka

Prof(Dr) V S Dhaka is M Tech & PhD in Computer Science. He is on Editorial and Subject Expert Board of Various Research Journals including IJAECT, Manthan and other international Research Journal. He has guided over 8 Ph D scholars and more than 24 Masters' projects. He has 35 research publications in his name. He has authored 3 books. His current research interests include Pattern Recognition, Artificial Intelligence, Human Computer Interface, Automation, Wireless ATM, mobile ad-hoc networks, multimedia networking and performance evaluation