Spam Classification Using Genetic Algorithm

Rand Ahmad Atta

randahmad_at@yahoo.com

Dr. Soukaena H. hashem **Dr.Ekhlas Khalaf Gbashi**

Soukaena.hassan@yahoo.com/ 110026@uotechnology.edu.iq/

Computer science department, University of Technology,

ABSTRACT

E-mail is the fastest way to exchange messages from one place to another across the world, the increased use of e-mail led to increase received messages in the mailbox, where the recipient receives many messages including those that cause significant and different problems such as stealing identity of recipient, losing of essential information causing losses to companies in addition to the damage to the network. These messages are so dangerous that the user is unable to avoid them especially as they take different forms such as advertisements and others. These messages are known as unwanted messages. In order to remove these spam messages and prevent them from being accessed, filtering is used. This paper aims to enhance the e-mail spam filtering by suggesting genetic algorithm classifier as a single objective evaluation algorithm problem to generate the best model to be used for classifying the e-mail messages in high accuracy. The first step in the proposal is applying normalization. The second is feature selection which is implemented to choose the best features, the third step is using genetic algorithm classifier as single objective evaluation algorithm that deal with one objective. The experimental results showed that the proposed system provides good accuracy in the first experiment (88%) and better accuracy in the second experiment (94%) and third experiment (95%).

Keywords: Spam Database, Feature Selection, Genetic Algorithm.

المستخلص

البريد الالكتروني هو أداة سريعة لتبادل الرسائل من مكان واحد إلى جميع الأماكن في العالم وان زيادة في استخدام البريد الالكتروني ادى الى زيادة استقبال عدد كبير من الرسائل في صندوق البريد ، حيث يتلقى المستلم العديد من الرسائل بما في ذلك تلك الرسائل التي تسبب مشاكل كبيرة ومختلفة مثل سرقة هوية المتلقي او فقدان المعلومات الأساسية التي تتسبب في خسائر للشركات بالإضافة إلى الأضرار التي تلحق بالشبكة ، وتعد أمر خطير للغاية حيث لا يمكن للمستخدم تجنبها كونها تأخذ مجموعة متنوعة من الأشكال مثل الإعلانات وغيرها ، وتعرف هذه الرسائل بأنها رسائل غير مرغوب فيها. من أجل ازالة هذه

الرسائل غير المرغوب فيها ومنع الوصول إليها ، يتم استخدام الترشيح . الهدف هو تحسيين من تصفية البريد الإلكتروني العشوائي. بأقتراح مقترح تصنيف الخوارزمية الجينية كخوارزمية تطورية هدف واحد لتوليد أفضل نموذج يستخدامه لتصنيف تصفية البريد الالكتروني بدقة عالية. أول خطوة في المقترح هو تطبيق Normalized. ثم ينفذ feature selection لاختيار أفضل ميزة. الخطوة الثالثة استخدام الخوارزمية الجينية كخوارزمية تطورية لهدف واحد التي تتعامل مع هدف واحد. التجارب أظهرت أن النظام المقترح يوفر دقة أفضل في التجربة 1 (88٪) ، ودقة عالية في التجربة 2 (94٪) وفي التجربة 3 (%95) .

1. Introduction

E-mail is an effective and the fastest popular communication way. Like every powerful medium, it is prone to misuse, spam is an example of misuse which led to the spread of undesirable messages to very large numbers of recipients [1], spam causes traffic clogging in the internet traffic and the main source for (spyware and viruses) [2],so all organizations should use the available tools in order to address and filter spam messages in its environment [3],the most effective method for countering spam is automatic filtering [4], the spam can be classified as: [5,6].

- Solicited e-mail: these messages sent only to recipients who have requested it and no need to reply such as commercial message, newsletter.

- **Unsolicited e-mail: one of the most spread threats on the public internet is also, known as spam, it reduces productivity for both e-mail administrator and end users.**

The proposal improves the results of email spam filtering by conducting FS and GA classifier. Also the proposal increases the accuracy. Using feature selection is very important to determine the optimal features for use it, where information gain is one of the ways of feature selection used in the proposal. The proposal suggested GA classifier to select population of solution and in the presence of both operators (crossover and mutation) to generate the best model and then evaluate the model.

2. Related Work

Razi Z. and Asghari S.A., 2017 [7], Discussed the importance of classifying messages spam and identifying them this method can be used to reduce errors for filtering systems and spam detection by proposing a system that offers a set of algorithms (genetic, artificial immune system) to extract the feature and (SVM) for classification, and compare between two methods, where tested on 1000 datasets of Spam email. The results were the accuracy of the proposed system by compared to (SVM, SVM-GA) algorithms, the accuracy of Hybrid GA-AIS is (95%), SVM-GA (91%) higher than SVM (88%), significant improvement was

observed in GA-AIS comparison to GA-AIS, and FP of Hybrid GA-AIS is (0.2%), SVM-GA (6.8%) and SVM (5.2%).

Choudhary M., et al., 2015 [8], Discussed the method of classify messages spam filter using the genetic algorithm and considered GA is good choice as the results showed. And the discovery that the accuracy of the genetic algorithm is affected by the data dictionary in the classification of messages spam filter using the GA, thus the algorithm is able to distinguish between (spam and not spam) with accuracy (81%).

Varghese L., et al., 2015 [9], Discussed how to use the genetic algorithm and the K-Means algorithm in classify messages spam filtering in order to speed up the filter process by select template messages from the training database to create good templates used in spam filtering, where the experiences gave good results, the accuracy after applying FS and GA (TP=0.94, FP=0.04, Precision=0.95, Recall=0.94, F-Measure =0.949, ROC-Area=0.99), and the accuracy after applying FS and K-Means (TP=0.91, FP=0.09, Precision=0.918, Recall=0.91, F-Measure =0.91, ROC-Area=0.955).

3. Feature selection

Feature selection, also known as a variable selection or a subset selection. This process is commonly applied in machine learning to resolve all high dimensionality problems. It chooses important features subset and excludes redundant, irrelevant and all noisy features to simplify and summarize data representation [10], some of feature selection approaches are; Filter models, Wrappers models and embedded models. The methods used in attribute evaluations are: gain ratio, information gain (IG), relief-F, and symmetrical uncertainty, chi-squared and one-R [11]. Entropy measures is the foundation in the information gain attribute ranking methods. Which distinguish the purity of examples in set of an arbitrary. It is used to measure information theory. And it is considered system's unpredictability measure [12].

The entropy of A is: $H(A) = -\sum_{y \epsilon Y} P(A) \log 2\, P(A)$ Eq. (1)

Where: P(A) = probability function for the random variable A.

There is a relationship between features "A" and "B" is:

$$H(A|B) = -\sum_{b \epsilon B} P(B) \sum_{a \epsilon A} P(A|B) \log 2\, P(A|B) \qquad \text{Eq. (2)}$$

Where: P (A|B) = probability of "A" given "B".

Given the entropy as a criterion of not purity in a training set S, this measure is known as IG. Is measure invert additional information about "A" provided by "B" it is given by [12].

$$IG = H(A) - H(A|B) = H(B) - H(B|A) \qquad \text{Eq. } (3)$$

4. Evolutionary Algorithms

Evolutionary algorithms build adaptive systems using set of rules that employ evolutionary principals. They are probabilistic search methods and optimization heuristics which have the ability to find a solution to the search and optimization tasks by simulating the natural biological evolution, at each generation, the population evolves towards regions in the search space that are better and better by simulating the processes of "Darwinian" evolution (selection, recombination, and mutation), the best known methods in the field of evolutionary algorithms are genetic algorithms [9,14]. Genetic Algorithms (GAs) process are robust and randomly search [7, 15]. Genetic algorithm contains many steps as follow [8, 16]:

i. Initialization: give to the genetic algorithm a speed in terms both (the evolutionary and best start point process) initial population is generated randomly.

ii. Selection: is used for selection the parent and despite this the genetic algorithm employs a directionless search.

   a. Fitness based selection: is chance to direct selection for each chromosome with to its fitness. Examples of the original method of for parent selection is roulette wheel selection or fitness-based selection.

   b. Rank-based selection: the selection here based on probabilistic and relative rank instead of absolute fitness.

   c. Tournament based selection: the best one of these parents is returns by choose parents in random.

iii. Crossover: is one of important operation in GA, is process of plural of bit strings by replace of segments between pairs of chromosomes. There are types of crossover.

   a. 1-Point Crossover: is select the bit location is randomly which need to change.

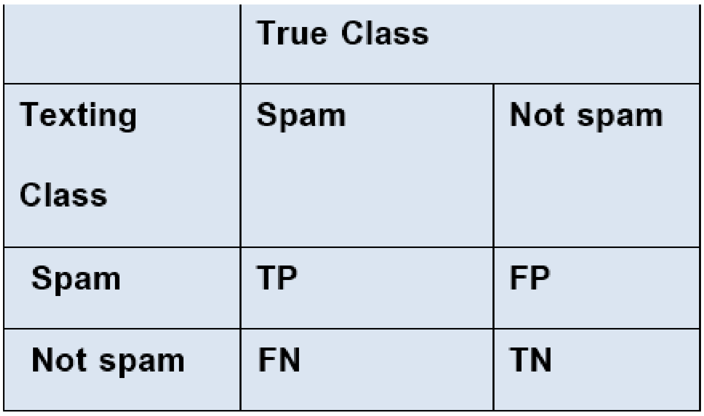
   b. 2-point Crossover: is select 2- position and the bit between 2- positions are change only.

iv. Mutation: is select any bit position randomly and changing it and has ensuring all chromosomes can keep the best gene in the new chromosomes.

5. The Evaluation Measures of Classification

The evaluation measures are defined from a matrix, which has

Only two classes – spam and not spam [17].

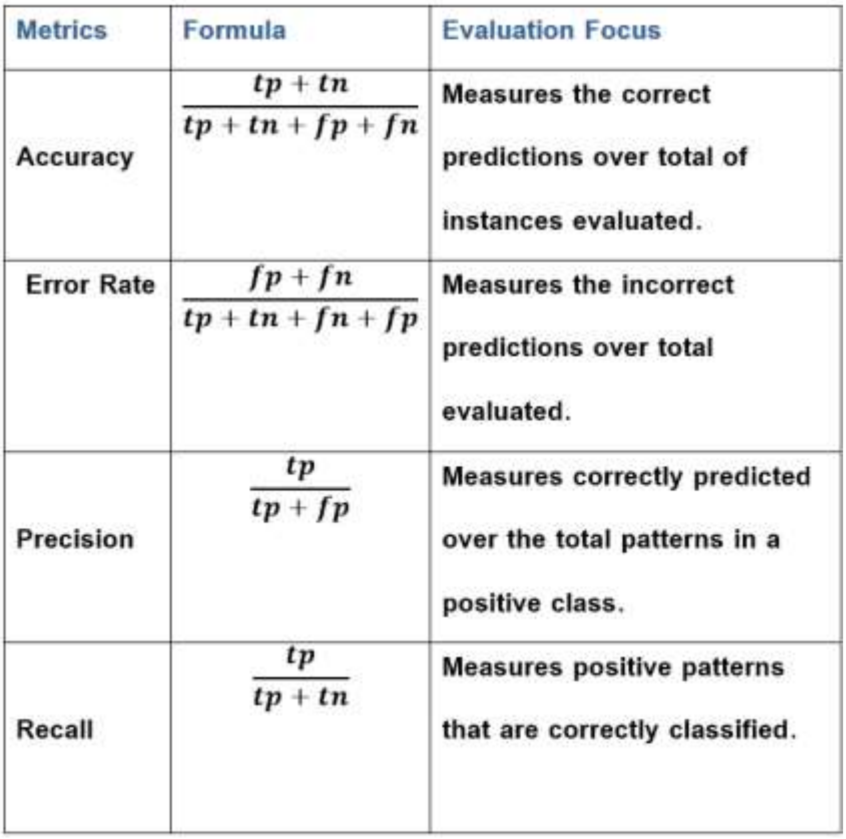


	True Class	
Texting Class	Spam	Not spam
Spam	TP	FP
Not spam	FN	TN

1. True positives (TP): e-mail in testing is spam where in true class is spam.
2. True negatives (TN): e-mail in testing is not spam where in true class is not spam.
3. False positives (FP): e-mail that is not spam in true class but testing is spam.
4. False negatives (FN): e-mail that is spam in true class but testing is not spam.

As the table (1) there are four evaluation measures that are used to evaluate the results of classifiers [18].

Table (1) Measurements Evaluation of Performance



Metrics	Formula	Evaluation Focus
Accuracy	$\frac{tp+tn}{tp+tn+fp+fn}$	Measures the correct predictions over total of instances evaluated.
Error Rate	$\frac{fp+fn}{tp+tn+fn+fp}$	Measures the incorrect predictions over total evaluated.
Precision	$\frac{tp}{tp+fp}$	Measures correctly predicted over the total patterns in a positive class.
Recall	$\frac{tp}{tp+tn}$	Measures positive patterns that are correctly classified.

6. Description of the Proposed System

The main framework of the proposed system is shown in figure below. There are three main components of system: normalization: apply normalization on spam data base to uniform the variants frequencies of words over the datasets, feature selection: to choose the best features that improve performance that contribute to raising the rate of accuracy of the model, and applied GA classifier. The following sections will explain each phase in details.

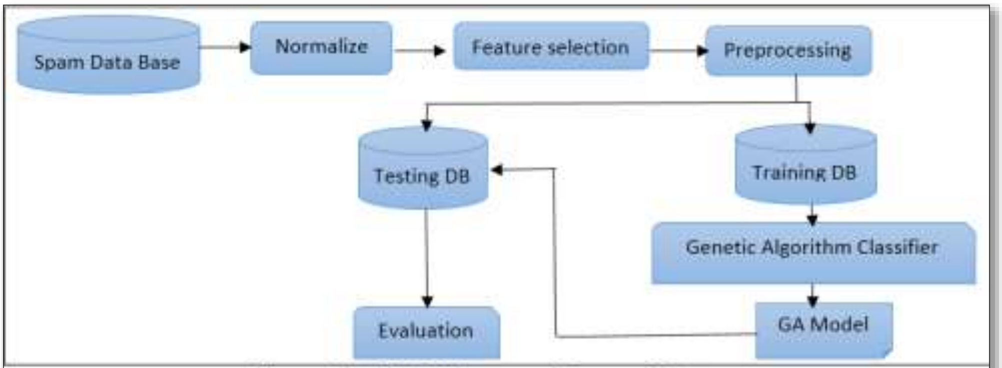


Figure (1): Block Diagram of Proposed System.

6.1 Description the Spam base and Converting from Text File into Excel

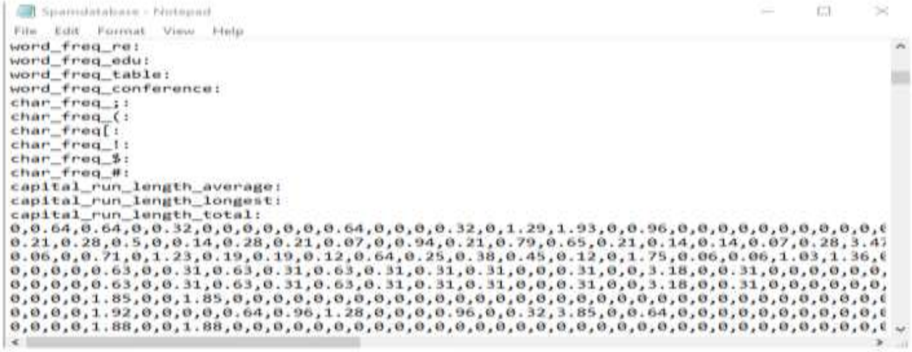
The dataset of the system is spam e–mail database that classifies e–mails as spam or non–spam. Consisting of 4601 instances (1813 Spam,2788 Not spam) and (57) features in addition to e–mail class type most features represented as certain "word" or "character" that are appear frequently in an e–mail. The definitions of the features are:

1) The (48) features are continuous real numbers. The range is [0–100] kind "word_freq_WORD" = words percentage that appear in e–mail and correspond to "WORD", i.e. (100) multiply (times number "WORD" appearance in the e–mail) divided by total characters in e–mail. A "word" in this case is any string of alphanumeric characters bounded by non–alphanumeric characters or end–of–string.

2) The next (6) features are continuous real numbers. The range is [0−100] kind "char_freq_CHAR"= characters percentage that appear in e−mail and correspond to "CHAR",i.e. (100) multiply (number of "CHAR" appearances) divided by total characters in e−mail.
3) The (55th) feature is continuous real kind "capital_run_length_average" = "average length of capital letters uninterrupted sequences".
4) The (56th) feature is continuous integer kind "capital_run_length_longest" = "length of capital letters uninterrupted sequence".
5) The (57th) feature is continuous integer kind "capital_run_length_total" = the sum of length of capital letters uninterrupted sequences = "total number of e−mail capital letters".
6) The last column is not mainly a feature. It is titular class type, which indicate whether the e−mail is considered a spam "1" or not "0".

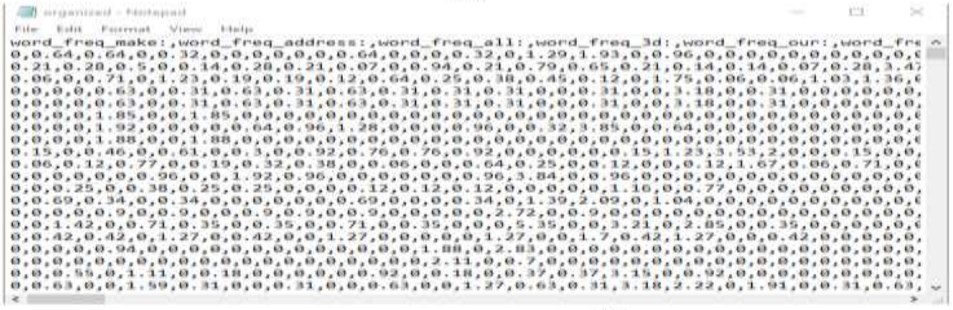
Figure (2.a) shows s

pam base text file data that consists of words and symbols and real number values that separated by a comma. Where the text file data are arranged and organized manually as shown in figure (2.b). Each of words and symbols are represented as features, the real number values are represented as features values that will be converted into excel database by using Microsoft office where each row in excel database contains of 58 features values as shown in figure (2.c).



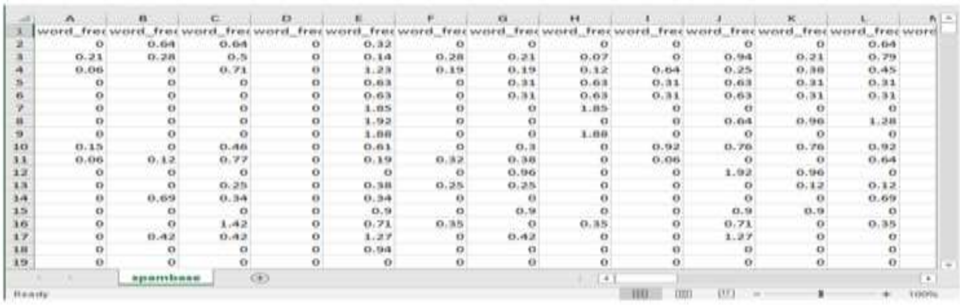
```
word_freq_re:
word_freq_edu:
word_freq_table:
word_freq_conference:
char_freq_;:
char_freq_(:
char_freq[:
char_freq_!:
char_freq_$:
char_freq_#:
capital_run_length_average:
capital_run_length_longest:
capital_run_length_total:
0,0.64,0.64,0,0.32,0,0,0,0,0,0,0.64,0,0,0,0.32,0,1.29,1.93,0,0.96,0,0,0,0,0,0,0,0,0,0,
0.21,0.28,0.5,0,0.14,0.28,0.21,0.07,0,0.94,0.21,0.79,0.65,0.21,0.14,0.14,0.07,0.28,3.4
0.06,0,0.71,0,1.23,0.19,0.19,0.12,0.64,0.25,0.38,0.45,0.12,0,1.75,0.06,0.06,1.03,1.36,
0,0,0,0,0.63,0,0.31,0.63,0.31,0.63,0.31,0.31,0.31,0,0,0.31,0,0,3.18,0,0.31,0,0,0,0,0,0,
0,0,0,0,0.63,0,0.31,0.63,0.31,0.63,0.31,0.31,0.31,0,0,0.31,0,0,3.18,0,0.31,0,0,0,0,0,0,
0,0,0,0,1.85,0,0,1.85,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,1.92,0,0,0,0,0.64,0.96,1.28,0,0,0,0.96,0,0.32,3.85,0,0.64,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,1.88,0,0,1.88,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
```

(a)



```
word_freq_make:,word_freq_address:,word_freq_all:,word_freq_3d:,word_freq_our:,word_fre
0,0.64,0.64,0,0.32,0,0,0,0,0,0,0.64,0,0,0,0.32,0,1.29,1.93,0,0.96,0,0,0,0,0,0,0,0,0,0,
0.21,0.28,0.5,0,0.14,0.28,0.21,0.07,0,0.94,0.21,0.79,0.65,0.21,0.14,0.14,0.07,0.28,3.4
0.06,0,0.71,0,1.23,0.19,0.19,0.12,0.64,0.25,0.38,0.45,0.12,0,1.75,0.06,0.06,1.03,1.36,
0,0,0,0,0.63,0,0.31,0.63,0.31,0.63,0.31,0.31,0.31,0,0,0.31,0,0,3.18,0,0.31,0,0,0,0,0,0,
0,0,0,0,0.63,0,0.31,0.63,0.31,0.63,0.31,0.31,0.31,0,0,0.31,0,0,3.18,0,0.31,0,0,0,0,0,0,
0,0,0,0,1.85,0,0,1.85,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,1.92,0,0,0,0,0.64,0.96,1.28,0,0,0,0.96,0,0.32,3.85,0,0.64,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,1.88,0,0,1.88,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0.15,0,0.46,0,0.61,0,0.3,0,0.92,0.76,0.76,0.92,0,0,0,0,0,0.15,1.23,3.53,2,0,0,0.15,0,0,
0.06,0.12,0.77,0,0.19,0.32,0.38,0,0.06,0,0,0.64,0.25,0,0.12,0,0,0.12,1.67,0.06,0.71,0,
0,0,0,0,0,0,0.96,0,0,1.92,0.96,0,0,0,0,0,0,0.96,3.84,0,0.96,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0.25,0,0.38,0.25,0.25,0,0,0,0.12,0.12,0.12,0,0,0,0,0,1.16,0,0.77,0,0,0,0,0,0,0,0,0,
0,0.69,0.34,0,0.34,0,0,0,0,0,0,0.69,0,0,0,0.34,0,1.39,2.09,0,1.04,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0.9,0,0.9,0,0,0.9,0.9,0,0.9,0,0,0,0,0,2.72,0,0.9,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,1.42,0,0.71,0.35,0,0.35,0,0.71,0,0.35,0,0,0,5.35,0,0,3.21,0,2.85,0,0.35,0,0,0,0,0,
0,0.42,0.42,0,1.27,0,0.42,0,0,1.27,0,0,0,0,0,1.27,0,0,1.7,0.42,1.27,0,0,0.42,0,0,0,0,0,
0,0,0,0,0.94,0,0,0,0,0,0,0,0,0,0,0,0,0,1.88,0,2.83,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2.11,0,0.7,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0.55,0,1.11,0,0.18,0,0,0,0,0,0.92,0,0.18,0,0.37,0.37,3.15,0,0.92,0,0,0,0,0,0,0,0,0,
0,0.63,0,0,1.59,0.31,0,0,0.31,0,0,0.63,0,0,1.27,0.63,0.31,3.18,2.22,0,1.91,0,0.31,0.63,
```

(b)



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	word_frec	word_frec	word_frec	word_frec	word_frec	word_frec	word_frec	word_frec	word_frec	word_frec	word_frec	word_frec	word
2	0	0.64	0.64	0	0.32	0	0	0	0	0	0	0.64	
3	0.21	0.28	0.5	0	0.14	0.28	0.21	0.07	0	0.94	0.21	0.79	
4	0.06	0	0.71	0	1.23	0.19	0.19	0.12	0.64	0.25	0.38	0.45	
5	0	0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31	0.31	
6	0	0	0	0	0.63	0	0.31	0.63	0.31	0.63	0.31	0.31	
7	0	0	0	0	1.85	0	0	1.85	0	0	0	0	
8	0	0	0	0	1.92	0	0	0	0	0.64	0.96	1.28	
9	0	0	0	0	1.88	0	0	1.88	0	0	0	0	
10	0.15	0	0.46	0	0.61	0	0.3	0	0.92	0.76	0.76	0.92	
11	0.06	0.12	0.77	0	0.19	0.32	0.38	0	0.06	0	0	0.64	
12	0	0	0	0	0	0	0.96	0	0	1.92	0.96	0	
13	0	0	0.25	0	0.38	0.25	0.25	0	0	0	0.12	0.12	
14	0	0.69	0.34	0	0.34	0	0	0	0	0	0	0.69	
15	0	0	0	0	0.9	0	0.9	0	0	0.9	0.9	0	
16	0	0	1.42	0	0.71	0.35	0	0.35	0	0.71	0	0.35	
17	0	0.42	0.42	0	1.27	0	0.42	0	0	1.27	0	0	
18	0	0	0	0	0.94	0	0	0	0	0	0	0	
19	0	0	0	0	0	0	0	0	0	0	0	0	

(c)

Figure (2) a) Spambase Text File, b) Arrange and Organize Spam base Text File and c) Spambase Excel database after Conversion.

6.2 Normalization

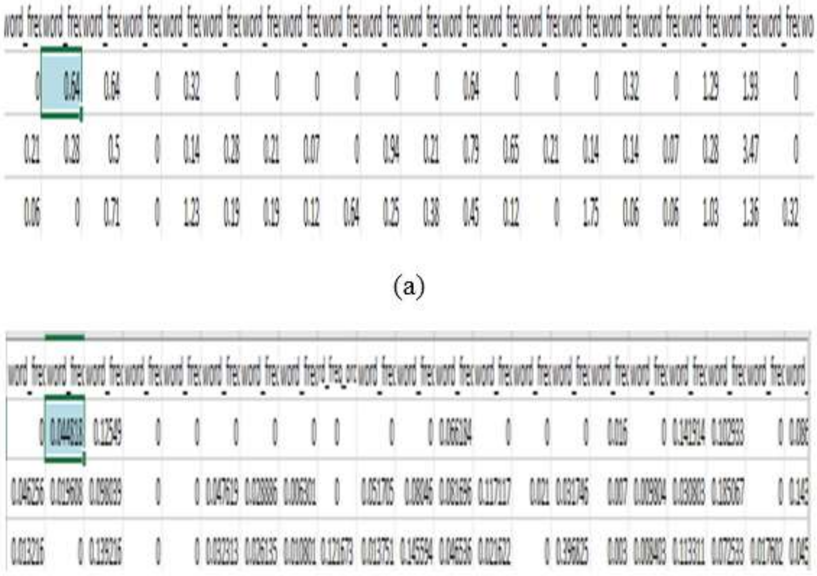
After converting the spam database text file into excel database, each row in excel database contains 57 features values and one class email type (spam and not spam), the values of spam database have different rang as in figure (3.a), so the normalization applied on a set of values to uniform range from [0] to [1] by using equation (4). As figure (3.b).

$$new_{value} = old^{value} - min^{value} \Big/ max^{value} - min^{value} \qquad Eq.(4)$$

For example:

$max^{value} = 14.28,\ min^{value} = 0,\ old^{value} = 0.64$

$$\mathrm{new}_{\mathrm{value}} = \frac{\mathrm{old}^{\mathrm{value}} - \mathrm{min}^{\mathrm{value}}}{\mathrm{max}^{\mathrm{value}} - \mathrm{min}^{\mathrm{value}}} \qquad new_{value} = \frac{0.64-0}{14.28-0} = 0.0448179$$



0	0.64	0.64	0	0.32	0	0	0	0	0	0	0.64	0	0	0	0.32	0	1.29	1.93	0
0.21	0.28	0.5	0	0.14	0.28	0.21	0.07	0	0.94	0.21	0.79	0.65	0.21	0.14	0.14	0.07	0.28	3.47	0
0.06	0	0.71	0	1.23	0.19	0.19	0.12	0.64	0.25	0.38	0.45	0.12	0	1.75	0.06	0.06	1.03	1.36	0.32

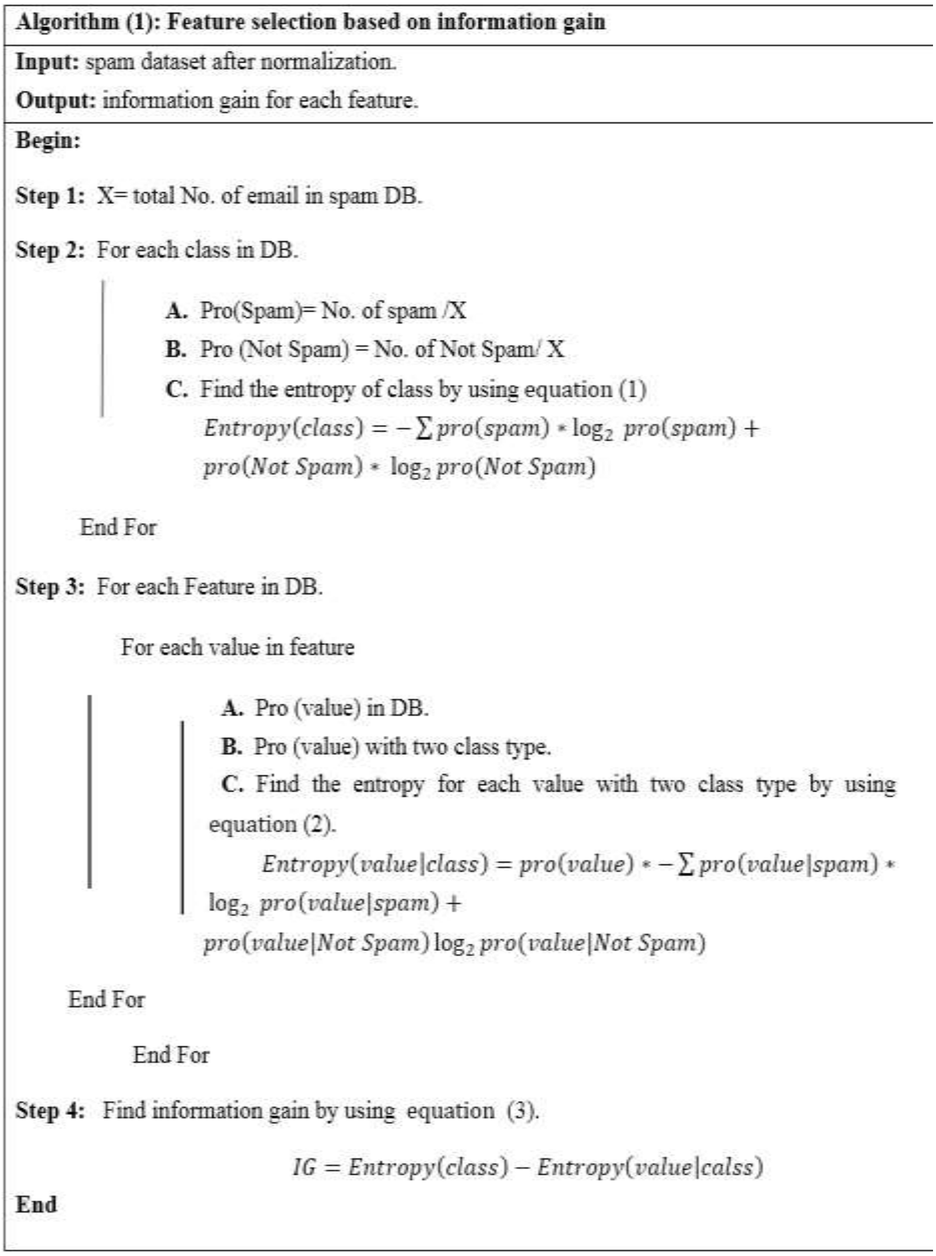
(a)

(b)

Figure (3) a) Spambase Dataset Before Normalization. b) Spambase Dataset after Normalization.

6.3 Feature Selection

Feature selection (FS) used to choose important features from spam database and remove redundant features and have a weak effect on performance. More explanation in algorithm (1).



```
Algorithm (1): Feature selection based on information gain
Input: spam dataset after normalization.
Output: information gain for each feature.
Begin:
Step 1:  X= total No. of email in spam DB.
Step 2:  For each class in DB.
            A. Pro(Spam)= No. of spam /X
            B. Pro (Not Spam) = No. of Not Spam/ X
            C. Find the entropy of class by using equation (1)
               $Entropy(class) = -\sum pro(spam) * \log_2 pro(spam) + pro(Not\ Spam) * \log_2 pro(Not\ Spam)$
      End For
Step 3:  For each Feature in DB.
            For each value in feature
               A. Pro (value) in DB.
               B. Pro (value) with two class type.
               C. Find the entropy for each value with two class type by using equation (2).
                  $Entropy(value|class) = pro(value) * -\sum pro(value|spam) * \log_2 pro(value|spam) + pro(value|Not\ Spam) \log_2 pro(value|Not\ Spam)$
      End For
            End For
Step 4:  Find information gain by using equation (3).
                  $IG = Entropy(class) - Entropy(value|calss)$
End
```

For example: Entropy (class)= $-\sum P(class)\log_2 P(class)$

$$=\left[-\frac{1813}{4601}\log_2\frac{1813}{4601}-\frac{2788}{4601}\log_2\frac{2788}{4601}\right]=0.967360237180767$$

Calculate the entropy for each value with two class type, for instance, element (0) that exists in the feature (“word_freq_table”) by using equation (2). Entropy (“word_freq_table:”|Element=0) =$[0.98(-0.60\log_2(0.60)0.39\log_2(0.39))]$ =0.937566.

Finally find information gain of feature = Entropy (class) − Entropy (feature). See table (2) displays information gain of all features.

Table (2) Information Gain of All Features.

	Name of Feature	**IG**	**Name of Feature**	**IG**
1.	‘‘word_freq_over:’’	0.130485684	‘‘word_freq_857:’’	0.032794464
2.	‘‘word_freq_remove:’’	0.236046662	‘‘word_freq_data:’’	0.047894696
3.	‘‘word_freq_internet:’’	0.154923065	‘‘word_freq_415:’’	0.033700328
4.	‘‘word_freq_order:’’	0.121640238	‘‘word_freq_85:’’	0.070184526
5.	‘‘word_freq_mail:’’	0.145161942	‘‘word_freq_technology:’’	0.063495537
6.	‘‘word_freq_receive:’’	0.14121729	‘‘word_freq_1999:’’	0.101904671
7.	‘‘word_freq_will:’’	0.14121729	‘‘word_freq_parts:’’	0.012321632
8.	‘‘word_freq_people:’’	0.11276125	‘‘word_freq_pm:’’	0.048123516
9.	‘‘word_freq_report:’’	0.05673959	‘‘word_freq_direct:’’	0.046674912
10.	‘‘word_freq_addresses:’’	0.082875957	‘‘word_freq_cs:’’	0.023865182
11.	‘‘word_freq_free:’’	0.256617007	‘‘word_freq_meeting:’’	0.052205405
12.	‘‘word_freq_business:’’	0.170880595	‘‘word_freq_original:’’	0.056028803
13.	‘‘word_freq_email:’’	0.138361475	‘‘word_freq_project:’’	0.040435862
14.	‘‘word_freq_you:’’	0.345689801	‘‘word_freq_re:’’	0.122277549
15.	‘‘word_freq_credit’’:	0.102502249	‘‘word_freq_edu:’’	0.076905442
16.	‘‘word_freq_your:’’	0.391868708	‘‘word_freq_table:’’	0.00882309
17.	‘‘word_freq_font:’’	0.030244508	‘‘word_freq_conference:’’	0.027252445
18.	‘‘word_freq_000:’’’	0.169182745	‘‘char_freq_;:’’	0.091304147
19.	‘‘word_freq_money:’’	0.200598343	‘‘char_freq_(:’’	0.24228176
20.	‘‘word_freq_hp:’’	0.181635035	‘‘char_freq_[:’’	0.485103907
21.	‘‘word_freq_hpl:’’	0.133061087	‘‘char_freq__!:’’	0.485103907
22.	‘‘word_freq_george:’’	0.13708408	‘‘char_freq_$:’’	0.34147914
23.	‘‘word_freq_650:’’	0.06688124	‘‘char_freq_#:’’	0.132630197
24.	‘‘word_freq_lab:’’	0.061186553	‘‘capital_run_length_average:’’	0.718165201
25.	‘‘word_freq_labs:’’	0.070787304	‘‘capital_run_length_longest:’’	0.443300101
26.	‘‘word_freq_telnet:’’	0.048342388	‘‘capital_run_length_total:’’	0.42238831

After finding IG for each feature, it is arranged from high IG to lower in an excel database to be used in next step. The feature with high IG is ("capital_run_length_average:"= 0.718165200982574) and the lower IG is ("word_freq_table:" = 0.00882308986412284). The spam base dataset consists of (4601 records) divided into two database:

The first dataset is training dataset: it's consists of (3000) records. The second dataset is testing dataset, used to evaluate performance of system (classification): it's consists of (1601) records. That represent email and (57) columns that represent feature in addition to class type: spam and not spam.

6.4 Genetic Algorithm Classifier as Single Objective Evaluation Algorithm

The proposed algorithm to classify email spam filtering is applied on training database to get optimal model, and evaluate performance of model using the testing dataset. The GA classifier will be applying: First with all of feature (57 feature). Second with best 40 features. Third with best 20 features. Figure (4) explains genetic algorithm classifier.

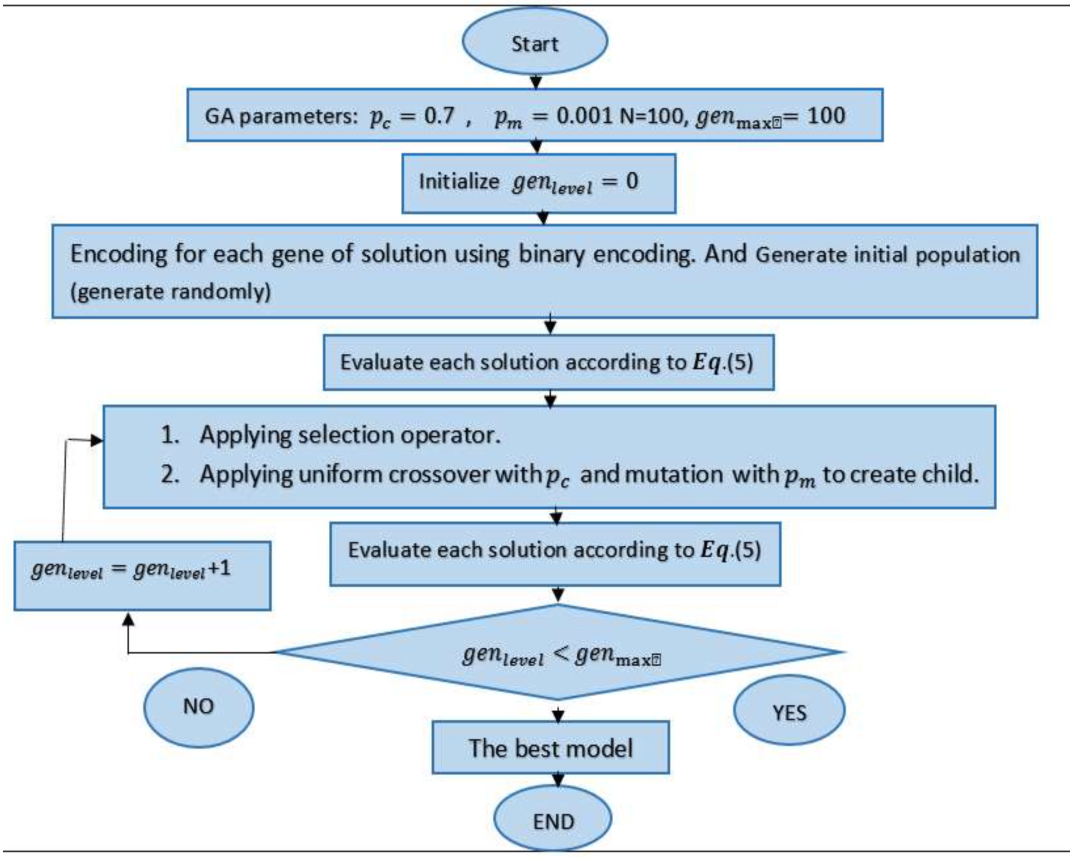

Start
GA parameters: $p_c = 0.7$, $p_m = 0.001$ N=100, $gen_{max} = 100$
Initialize $gen_{level} = 0$
Encoding for each gene of solution using binary encoding. And Generate initial population (generate randomly)
Evaluate each solution according to $Eq.(5)$
1. Applying selection operator.
2. Applying uniform crossover with p_c and mutation with p_m to create child.
Evaluate each solution according to $Eq.(5)$
$gen_{level} = gen_{level}+1$
$gen_{level} < gen_{max}$
NO
YES
The best model
END

Figure (4) Flowchart of Genetic Algorithm Classifier.

Encoding each gene of solution using binary encoding by representing values of gene that are greater than or equal to $0.1 = 1$, otherwise 0. And represented by length vector of size N, let us consider a population P of N. Can be formulated as follows P= $\{P_1, P_2, \dots, P_N\}$, where N is population size. The population starts with an initial random population p_0 and continue until it reaches the largest number of iterations specified. At each iteration in GA will be applying of three main operators: selection, crossover and mutation. After apply the operator, the fitness function used in the proposed to evaluate the quality of genetic algorithm solutions, considering the email spam filter problem, the single objective evaluation algorithm (SOEA) model is proposed to satisfy email spam filter problem challenges, it represented by maximizing “accuracy” as in the following equation:

$$\text{Maximizing SOGA (P)} = \text{Accuracy (acc)} = \frac{tp+tn}{tp+tn+fp+fn} \qquad \text{Eq. (5)}$$

6.4.1 The Fitness Function using in Classifier the Email Spam Filtering and Evaluation of the Solutions

The gene in chromosome was (1, 0), if number of gene type (1) is greater than(X), then the email is spam. Otherwise not spam and we found the minimum (X) calculated was 3 for the evolution of the fitness function an experiment was carried out on spam database, where X is number. Then comparison between new class and old class to calculate each of the (true positive, true negative, false positive, false negative) and evaluation of population using equation (5).

In selection process, the best chromosomes are selected randomly from the population where the size of the tournament (i.e.,$tournament_{size} = 2$) two chromosomes are chosen randomly. Using uniform crossover with probability p_c to create new generation (child). Applying uniform crossover on index of position of bit which has been identified. The mutation process is useful because crossover can't produce a new generation, the mutation process is done by choosing a bit position and changing it to 1 instead of 0 or 0 instead of 1, index of position of bit was selected, the mutation with probability p_m. As shown in algorithm (2). Finally, generation has the highest accuracy is identified as the best model, then we compare actual class for testing dataset with previous class for model (previous class: is classify the email to spam or N-spam by applied the proposed algorithm on training database) then evaluation. More explanation in algorithm (3). In our experiments both crossover and mutation are done. The diversity in position of bit selection (not selected random) leads to get variety in results.

```
Algorithm (2): Genetic Algorithm Classifier

Input:

•  Single Objective Problem: Maximizing SOEA (P).
•  $P_C$: Probability of Crossover. $P_m$: Probability of Mutation.
•  N:  No. Individual.
•  $gen_{max}$: Maximum No. of Generation.
Output: optimal model of GA.
```

```
Begin:
Step 1: Encoding
Each gene of solution in GA using binary encoding.
Step 2: initializing
 Generate initial population (generate randomly in GA).
Step 3: Evaluation
Evaluate each solution according to equation (3.2).
Step4: GA operators
 1. Selection with tournament size=2.
 2. Uniform crossover with pc and crossover point CP.
 3. Flip mutation with pm and mutation point MP.
Step 5: Termination Function
 If (New generation == old generation) Stop and get the result.
        Else, go to step 4.
End if
End
```

Algorithm (3): Using Genetic Algorithm Classifier as Single Objective Evaluation Algorithm

```
Input: Training and Testing dataset
Output: GA classifier.

Begin:
Step 1: Training dataset
1. Apply algorithm (2) with the all features.
2. Apply algorithm (2) with the best 40 features.
3. Apply algorithm (2) with the best 20 features.

Step 2: Testing dataset

Compare between actual classes (for testing dataset) with previous class (for model).
If  actual class=Spam && previous class= Spam
        TP++
      Else If  actual class=N-Spam && previous class= Spam
              FN++
            Else If  actual class=Spam && previous class= N-Spam
                  FP++
                Else  actual class=N-Spam && previous class= N-Spam
                      TN++
End if  End if
            End if    End if
Step 3: Evaluation
Find accuracy, Error Rate, Precision and Recall.
End
```

7. The Experimental Results

Experiments show that the difference in the index of position of bit has an effect in obtaining different results in accuracy, although each of (P_C and P_M=) is the same value has been applied in the experiments, but each time applied to the different index of position of bit. Note that CP is index of position of bit that applies crossover and MP is index of position of bit that applies mutation. And experiments are depended on the features with the highest IG. The experiments start with all (57) features, with best (40) features, then with best (20) features. In experimented, we used 3000 record as training dataset, and used 1601 record as testing dataset to evaluated the performance. In figure (5) shows classification results of highest of accuracy for three experiment.

- Experiment 1: The experiment done with 57 feature and the crossover and mutation is done but the index of position of bit determines differently each time, affecting the final results of the classification process as show in table (3). (P_C = 0.7, P_M= 0.001, N =100, K =3).

Table (3) Result of Experiment 1.

NO.	CP	MP	Accuracy
1	35	20	88%
2	31	19	83%
3	33	20	75%

- **Experiment 2: The experiment done with 40 feature and the crossover and mutation is done but the index of position of bit determines differently each time, affecting the final results of the classification process as show table (4). (P_C = 0.7, P_M= 0.001, N =100, K =3).**

Table (4) Result of Experiment 2.

NO.	CP	MP	Accuracy
1	27	9	94%
2	20	5	83%
3	26	5	82%

- **Experiment 3: The experiment done with 20 feature and the crossover and mutation is done but the index of position of bit determines differently each time, affecting the final results of the classification process as show in table (5). (P_C = 0.7, P_M= 0.001, N =100, K =3).**

Table (5) Result of Experiment 3.

NO.	CP	MP	Accuracy
1	12	3	95%
2	10	1	93%
3	16	1	76%

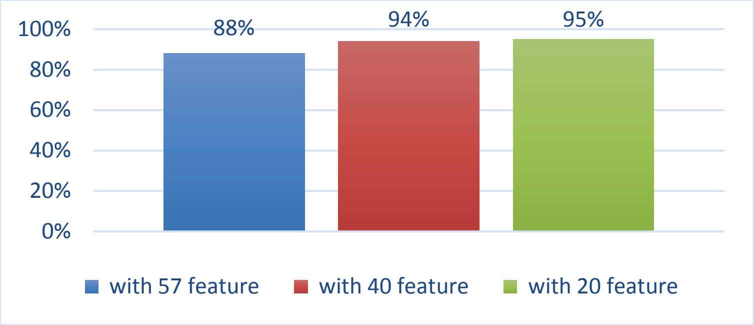
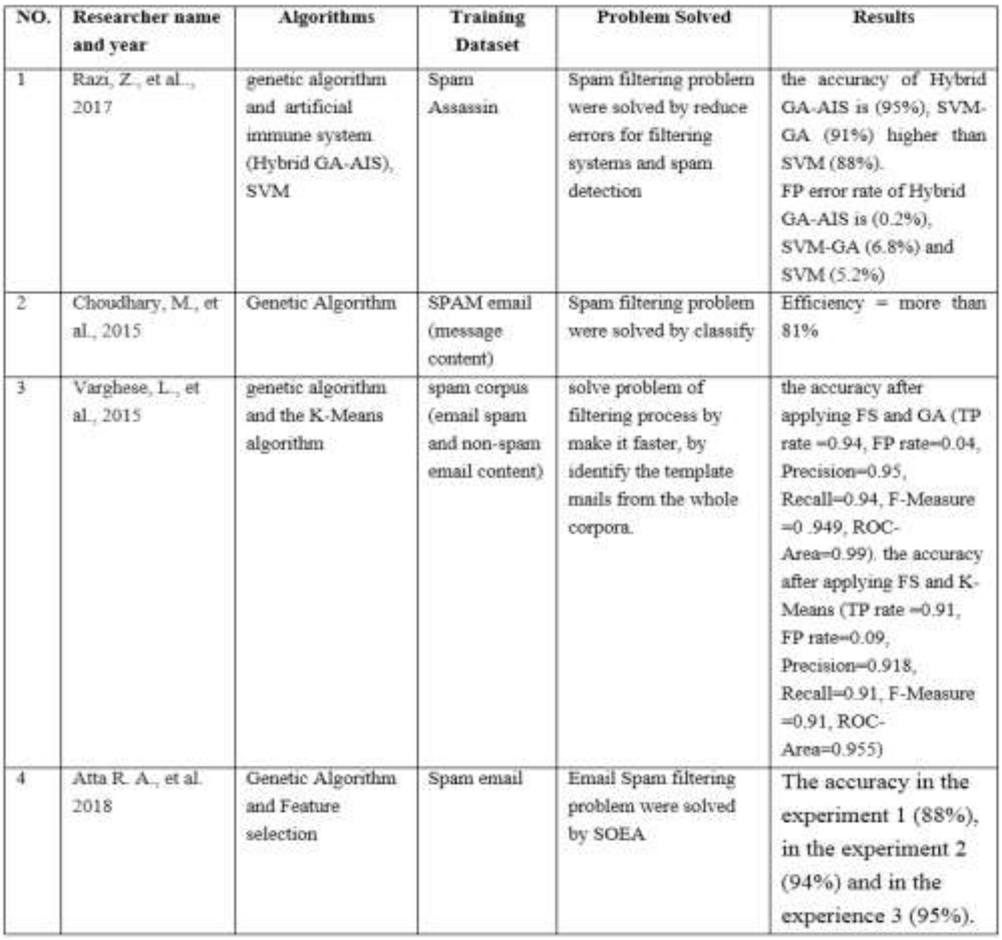


Figure (8) Classification Results of Highest of Accuracy.

8. Comparison with Related Works

The table (6) lists of the related work, which are collected through the research survey. The table (6) displays the used algorithms and accuracy results.

Table (6) Comparison with Related Works.



NO.	Researcher name and year	Algorithms	Training Dataset	Problem Solved	Results
1	Razi, Z., et al., 2017	genetic algorithm and artificial immune system (Hybrid GA-AIS), SVM	Spam Assassin	Spam filtering problem were solved by reduce errors for filtering systems and spam detection	the accuracy of Hybrid GA-AIS is (95%), SVM-GA (91%) higher than SVM (88%). FP error rate of Hybrid GA-AIS is (0.2%), SVM-GA (6.8%) and SVM (5.2%)
2	Choudhary, M., et al., 2015	Genetic Algorithm	SPAM email (message content)	Spam filtering problem were solved by classify	Efficiency = more than 81%
3	Varghese, L., et al., 2015	genetic algorithm and the K-Means algorithm	spam corpus (email spam and non-spam email content)	solve problem of filtering process by make it faster, by identify the template mails from the whole corpora.	the accuracy after applying FS and GA (TP rate =0.94, FP rate=0.04, Precision=0.95, Recall=0.94, F-Measure =0 .949, ROC-Area=0.99). the accuracy after applying FS and K-Means (TP rate =0.91, FP rate=0.09, Precision=0.918, Recall=0.91, F-Measure =0.91, ROC-Area=0.955)
4	Atta R. A., et al. 2018	Genetic Algorithm and Feature selection	Spam email	Email Spam filtering problem were solved by SOEA	The accuracy in the experiment 1 (88%), in the experiment 2 (94%) and in the experience 3 (95%).

9. Conclusions

In this work, genetic algorithm classifier as (SOEA) is proposed for enhanced email spam filtering.

1. Where a database has been prepared by applying normalization to get uniform for the values between [0, 1].
2. The suggested system used IG to choose important features from spam database.
3. Then implementing genetic algorithm classifier where the system will pass in two stage: in the first stage the model was trained with all 57 features and with different indexes of positions of bits by applying (crossover and mutation), then training will be done with best 40 features and with different indexes of positions of bits with applying (crossover and mutation) then training will be done with best 20 features and with different indexes of positions of bits that applies (crossover and mutation). In the second stage evaluation the performance will be done for each part (57, 40, and 20). In order to get result of high of accuracy from each part as shown in figure (8).
4. This work gives very good and excellent accuracy results of classification as shown in tables (3), (4) and (5). The contrast results in accuracy in each experiment are due to the feature

selection, since some features prevented results from reaching higher efficiency, and also (crossover and mutation) on position of bit in the chromosome showed a significant impact on the accuracy of the proposed system. Subsequently the algorithm of system succeeded in classifying email spam filtering to (an annoying email and an unobtrusive email) with better accuracy in the experiment 1 (88%), very high accuracy in the experiment 2 (94%) and in the experience 3 (95%).

REFERENCES

1. Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C. D., & Stamatopoulos, P. (2000), “Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach”, arXiv preprint cs/0009009.
2. Banday, M. T., & Jan, T. R. (2009), “Effectiveness and limitations of statistical spam filters”. arXiv preprint arXiv:0910.2540.
3. Christina, V., Karpagavalli, S., & Suganya, G. (2010), “A study on email spam filtering techniques. International Journal of Computer Applications”. 12(1), 0975-8887.
4. Awad W.A. and ELseuofi S.M., “MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION”, Vol 3, No 1, Feb 2011.
5. 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.

6. Blanzieri, E., & Bryl, A. (2008), "A survey of learning-based techniques of email spam filtering", Artificial Intelligence Review, 29(1), 63-92.
7. Razi, Z., & Asghari, S. A. (2016), "Providing an Improved Feature Extraction Method for Spam Detection Based on Genetic Algorithm in an Immune System".
8. Choudhary, M., & Dhaka, V. S. (2015), "Automatic e-mails Classification Using Genetic Algorithm".
9. Varghese, L., Supriya, M. H., & Jacob, K. P. (2015), "Finding Template Mails from Spam Corpus Using Genetic Algorithm and K-Means Algorithm", Training, 50, 50.
10. De Silva, A. M., & Leong, P. H. (2015), "Grammar Based Feature Generation", In Grammar-Based Feature Generation for Time-Series Prediction (pp. 35-50). Springer, Singapore.
11. Novakovic, J. (2009, November), "Using information gain attribute evaluation to classify sonar targets", In 17th Telecommunications forum TELFOR (pp. 1351-1354).
12. Cover, T. M., & Thomas, J. A. (2012), "Elements of information theory", John Wiley & Sons.
13. Kumar, V., & Minz, S. (2014), "Feature selection", SmartCR, 4(3), 211-229.
14. Holland, J. H., & Goldberg, D. (1989), "Genetic algorithms in search, optimization and machine learning", Massachusetts: Addison-Wesley.

15. Li, W. (2004), “Using genetic algorithm for network intrusion detection”, Proceedings of the United States Department of Energy Cyber Security Group, 1, 1-8.
16. Ferragina, P., & Grossi, R. (1999), “Improved dynamic text indexing”, Journal of Algorithms, 31(2), 291-319.
17. Costa, E., Lorena, A., Carvalho, A. C. P. L. F., & Freitas, A. (2007), “A review of performance evaluation measures for hierarchical classifiers”, In Evaluation Methods for machine Learning II: papers from the AAAI-2007 Workshop (pp. 1-6).
18. Hossin, M., & Sulaiman, M. N. (2015), “A review on evaluation metrics for data classification evaluations”, International Journal of Data Mining & Knowledge Management Process, 5(2), 1.