Contents list available at IJMEC

International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)

Journal Homepage: www.aeuso.org

PISSN: 2411-6173 - EISSN: 2305-0543

# Hybrid GA-AIS for Efficient Feature Extraction in E-mail Spam Detection

Zahra Razi[1*] and Seyyed Amir Asghari[2]

[1]*Department of computer, Karaj Branch, Islamic Azad University, Karaj, Iran*

[2]*Electrical and Computer Engineering Department, Kharazmi University, Tehran*

***Corresponding Author's E-mail:** Razizahra88@gmail.com*

## *Abstract*

The issue of spam has been taken into account due to the wide access to Internet. Several intelligent solutions have been presented to deal with this phenomenon. Many machine learning algorithms that are used in text classification may also be useful in identifying spam. Some of these algorithms include Support Vector Machine (SVM), genetic algorithm and immune system. In this paper, a combination of evolutionary genetic algorithm and artificial immune system is proposed that is responsible for selecting the best features of standard input of data collection. A standard method was used in the proposed system to measure the accuracy rate and error and results were tested on standard data sets of spam assassin. Finally, the results of several other algorithms, applied to the data, were compared with the outcomes obtained. Experimental results showed that the proposed method beside high speed of convergence in features extraction has acceptable accuracy about is 98%.

***Keywords:*** Spam, Feature Extraction, Genetic Algorithm, Immune System Algorithm, Support Vector Machine.

## 1. Introduction

Due to the widespread use of electronic letters an issue of spam has been arisen that causes some problems to users. Spam is called to unsolicited e-mail or UCE (Unsolicited Commercial Email) [1] which may be sent to different mailboxes of users for various purposes. Spams create many problems, some of which directly bring about economic losses. To put it more accurately, spam causes traffic and obliterates memory space and computing power. Spams make users to spend a lot of time to separate and omit them. In addition, it causes mental harassment and insecurity in users; spams also create legal problems such as pornography advertising, pyramid schemes, and scams such as phishing. Feeris Research Institute estimates that economic loss resulting from unwanted e-mails is over 50 million dollars [1].The bulk of emails exchanged are spams; 75-80% of e-mails are spams [1]. Spam detection can be divided into two general categories based on the research in the last decade. The first is called the credit-based group. These methods rely on information outside the context of personalized email messages such as: address, IP, domain and address of the sender. The second is content-based category [2]. Here, content is investigated regardless of source of message for detecting spam email. Intelligent and machine learning methods can be named for this group [3]. In the machine learning method, a set of smart algorithms are used to learn the rules through a bunch of pre-classified samples [4]. It is divided into several categories for filtering spam. Statistical filters, genetic algorithm(GA), immune system, neural networks, support vector machine, Naïve Bayes theory and combined methods like using genetic algorithm and support vector machine and composition of neural network and support vector machine in the field of email Classification. The accuracy and the calculations in each of these methods are different. Combination of the Genetic algorithm and the immune system is one

***Article History:***

**Received Date: Sep. 10, 2016**

**Accepted Date: Dec. 24, 2016**

**Available Online: Jan. 17, 2017**

of the most popular machine learning algorithms in this field [5]. The remainder of this paper is organized as follows. The related works and researches are reviewed in Section 2.Then in section 3 the proposed method are reviewed  in section 4, the results of the proposed method and Finally, the conclusions and recommendations are presented in Section 5.

## 2. Related Works

This section provides a brief description about the evolutionary algorithms. A summary of the studied algorithms are shown in Table 1.

### 2.1 The Selection Features Using the Genetic Algorithm

Genetic algorithm was invented by John Holland in 1967. Later, this method was found its place by efforts of Goldberengin 1989, and nowadays it has appropriate place among other methods in light of its capabilities. Optimization process in the genetic algorithm is based on a directed randomized process. This method has been developed based Darwin's evolution theory and fundamental ideas. In this method, a set of target parameters are generated firstly for fixed number called as population. After implementing the numerical simulator software representing the standard of deviation and/or fitness of the set of information, we assign it to that member of the mentioned population and we iterate this process for every member of the generated member.

#### 2.1.1 Designing of Chromosome

Firstly, we select a set of features. In other words, we extract the feature. For this purpose, we extract all unique words available in data warehouse. Then, we calculate the number of repetition of each word for each E-mail and we consider it as the vector feature for each of the E-mails. In the next stage, we should select a set of appropriate features among all extracted features. As genetic algorithm is one of the most efficient and appropriate optimization algorithms and it is used for selection of feature in many cases, we gave feature set obtained from the previous stage per every Email to genetic algorithmso that the best appropriate feature subset to be selected. In the genetic algorithm, each of feature vector was considered as a chromosome [9].

#### 2.1.2 Crossover

To operate the crossover, the arithmetic operator was considered. In this operator, the children are generated from mean weight of two parents.

#### 2.1.3 Mutation

To practice the mutation, there has been used the uniform mutation. In this mutation, the selected gene is replaced by a random uniform amount specified lower and upper limits by the user.

#### 2.1.4 Fitness Function

The value of each chromosome is evaluated by the fitness function. The best chromosome would be the chromosome with the minimum error.

### 2.2 The Algorithm of Immune System

This algorithm was introduced by Dasgupta in 1999. The immune system algorithm is based on non-gender selection principle and it is population-based algorithm. The immune system algorithm is inspired by human immune system. It is a distributed compatible system, which have capacities of immunity diagnosing, reinforcement education, extraction of property, immune memory, diversity, and strength. The main search power in the immune system is based on the mutation operator [8].

Therefore, it isconsidered as decision-making factor in this technique. Steps of this algorithm are defined as follows:

Initialization of the antibodies is performed in the first stage. Antigens represent the value of the objective function to be optimized.

Generation in which proportion or proximity of any antibody is determined. According to this proportion, antibodies are reproduced that the best reproduction took place at the best state.

Super mutation is the final step of this algorithm. Clones are linked to a super mutation process in which the clones are mutated in an inverse proportion for affiliation (dependency). The best antibody clones are muted lowly and the clones with worse antibodies have the greatest mutation. Then, the clones are evaluated in line with their main antibodies, which the best antibodies are selected for the next iteration.

Table 1. A Summary Related Works of the Studied Algorithms

| Author | The method used | Result | Advantages | Disadvantages |
|---|---|---|---|---|
| [5] j.shrivastava and etal(2014) | using the adaptive genetic algorithm | Using genetic algorithms, the feature extraction leads to increase the accuracy. | The spaces of solutions are checked in several routes.<br>The coding of parameters has been performed.<br>This method works with domestic law | If proportion extraction is not selected properly.<br>Optimal solution for problem will not provide.<br>It has premature convergence problem. |
| [7] s.jiang and et al(2012) | Method improve K-NN | By combining clustering at the Training Stage in the K-NN method, better efficiency ban be obtained | Increase the performance | Dependence on the training set.<br>There is no weight difference between samples |
| [6] H.Drucker [2012] | Svm for spam categorization | There has been used the support vector machine to classify E-mails with regard to contents. It can increase the speed of identification of the spam | Hasvery high accuracy for dataset of emails data | has low accuracy for dataset with high dimensional |
| [16]R.zitar [2011] | Application of genetic optimized artificial immune system and neural networksin spam detection | highpercentage of false positive and false negative value appeared | computationally efficient | Lack of optimization methods is observed in the immune system |

## 3. The Process of Evolutionary Algorithm

Evolutionary algorithms are algorithms for searching, in which searching is done from multiple points in the response space. There are many problems that conventional solutions are not remedial for them because there is no analysis for them or their analytical solution is very difficult, and/or complexity of variables and bulk of parameters of solutions and not necessarily the answer of the problem, so evaluation of all solutions is not possible. Evolutionary algorithms are methods based on random search, modeled from modeling of natural biological evolution. They work on possible answers which possess a superior feature and enjoy more generation survival, hence, providing a closer approximation of the optimal response. Evolutionary algorithms use preliminary mechanisms and operations for problem solving and arrive at an appropriate solution during a series of iterations. These algorithms of ten start from a population of random solutions, and try to improve the solution set during each iteration stage. At the beginning, a number of people are randomly guessed, and then the fitness function is calculated for each of these individuals, and the first generation is created. If none of the criteria to end the optimization is established, a new generation will begin to be created. People are selected in terms of their competency to children. These people are considered as parents and children crossover. All children face genetic with a certain amount of probability, mutation. Then, the children's competence is determined by the fitness function, and they are replaced by parents in the community, and a new generation is created. This cycle is repeated until one of the optimization end criteria is achieved. In this study, clonal and genetic evolutionary algorithms were used [8].

### 3.1 The Proposed Method

Optimized algorithms of immune and genetic systems have been proposed as hybrid and parallel algorithms for feature selection process. Genetic algorithm has been utilized to find the optimal set of features weights that improve the accuracy of classification [13]. The immune system algorithm has been taken into account because of the similar structure to genetic algorithm and in fact these algorithms are complementary. According to this technique, the algorithm is started with a group of randomly producing initial population and uses proportion value to assess the population. In both the genetic and immune systems, search methods are dependent combination of deterministic and probabilistic rules. These two algorithms are efficient and adaptive. They also own powerful search processes, produce desirable solutions and are executed in parallel, explicit and unconditional way.
The main difference between immune and genetic systems is that immune system algorithm does not have the operators of genetic algorithm such as crossover and mutation. Antibodies and antigens can update themselves with eligibility rules of an external agent. Compared to genetic algorithm, it owns very important Memory. It is more intelligent and can easily be implemented [14]. Instead, the parameters used in the genetic algorithm are fewer but it benefits from the higher convergence speed. However, if the parameters are properly set, the results can easily be optimized. The decision on the parameters of the immune system with trade-off exploration is heavily dependent on the objective function. Successful feature selection is acquired by using the memory values of immune system algorithm for the basic parameters [15].

According to Figure 1, in the initial phase in combination of immune system and genetic algorithm, the first set of outcomes of immune system (by the implementation of the first stage of the algorithm)are used as the first generation in genetic algorithm. Values of outcome Set of the immune system algorithm are searched and answers are found locally while a series of the results of genetic algorithm are quested and discovered globally in a wide variety of domains. These two algorithms choose effective features based on evaluation criterion of the adaptability in parallel with the exchange of their set of outcomes. In particular, an important advantage of immune system algorithm is that it performs only local search to achieve the optimal solution. So, the local optimality search can cause the coverage of global search with its detailed perspective that leads to optimization of ultimate solution. On the other hand, genetic algorithm covers populations and collection of all the features from the outset with a global view into performance accounts. So by combining these two methods, the algorithm can cover each other's weaknesses.

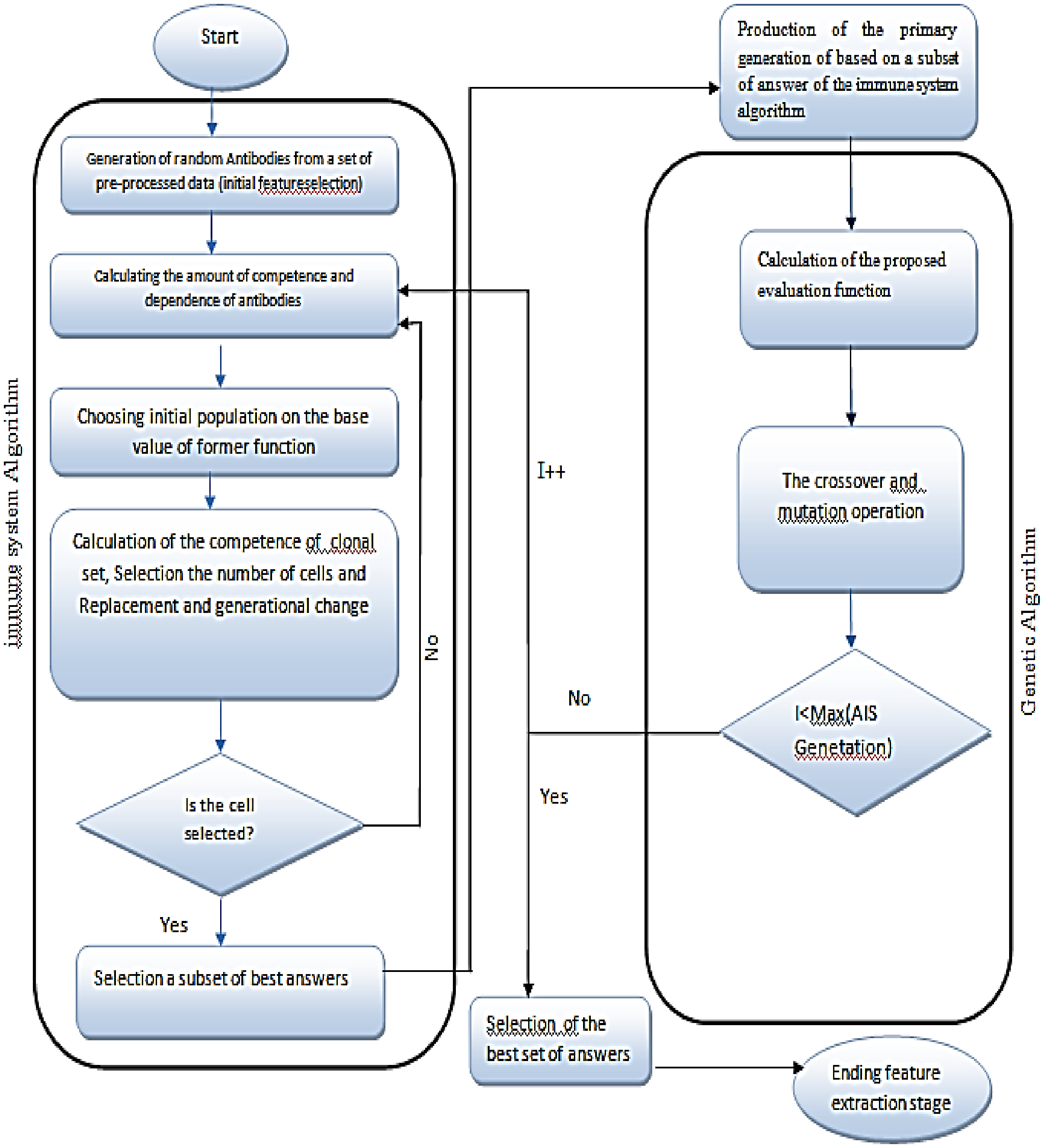
Start
Generation of random Antibodies from a set of pre-processed data (initial featureselection)
Calculating the amount of competence and dependence of antibodies
Choosing initial population on the base value of former function
Calculation of the competence of clonal set, Selection the number of cells and Replacement and generational change
Is the cell selected?
Yes
No
Selection a subset of best answers
immune system Algorithm
Production of the primary generation of based on a subset of answer of the immune system algorithm
Calculation of the proposed evaluation function
The crossover and mutation operation
I<Max(AIS Genetation)
Genetic Algorithm
I++
No
Yes
Selection of the best set of answers
Ending feature extraction stage

Figure 1. The Flowchart of the Proposed Algorithm

## 4. Experimental Results

In the proposed system, criteria of accuracy and error are used for evaluating system performance which has been presented in equations of (1) [10].

$$Accuracy = \frac{\text{TP+TN}}{\text{N}} \qquad \text{Relation} \qquad (1)$$

**TP:** The number of records that their actual group is positive and classifying algorithm has diagnosed positive.

**TN:** The number of records that their real group is negative and classifying algorithm has correctly diagnosed negative.

The results of experiments showed the accuracy of the proposed method compared with SVM-GA, PSO-SVM, KNN, Naïve Bayes, and NN SVM algorithms and were assessed and compared on dataset of 1000 Spam Assassin email samples (Figure 2).The results revealed that the accuracy of the proposed method was more when features increased.

In this study, in order to extract features from sample emails, the combination of GA-AIS algorithms with SVM algorithm was used for classification. The quantitative results of the proposed algorithm, compared with a number of conventional algorithms in this field, including neural network and Bayesian algorithm, indicate the acceptable accuracy and speed of this algorithm.

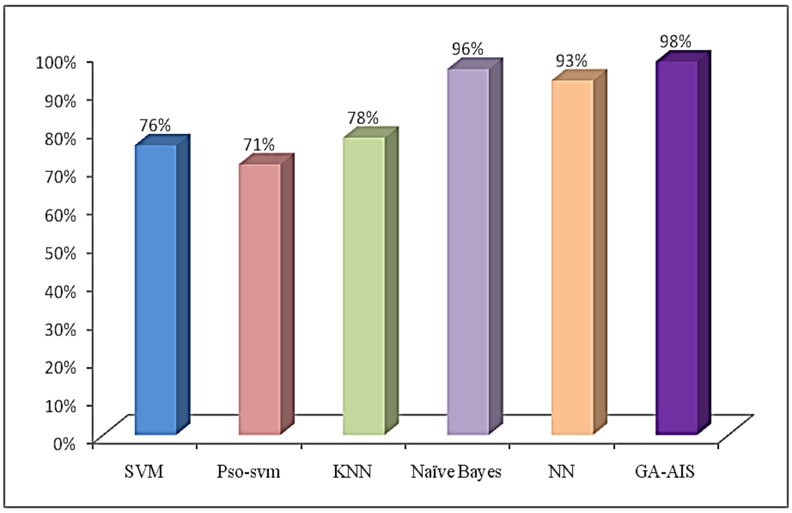

Figure2- Comparing the Proposed Models with Other Algorithms

The figures 3,4 and 5 demonstrate the convergence of factors in GA-AIS, PSO and GA algorithms in100 iterations. The results indicate the high speed of convergence of the proposed algorithm for extracting features.

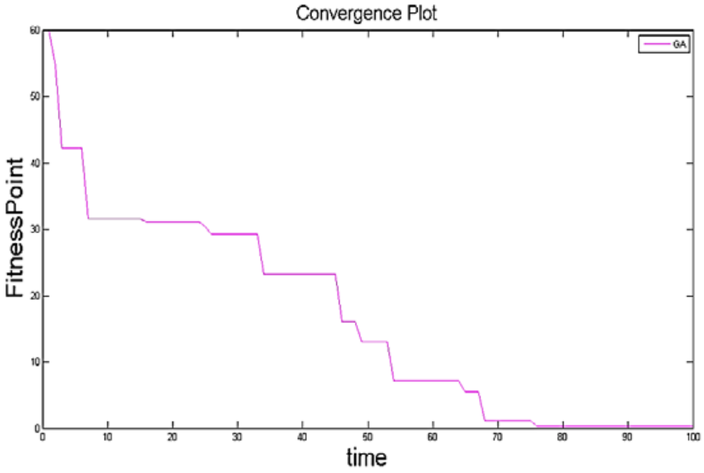

Figure 3- The Convergence of GA Algorithm in Order to Features Extraction

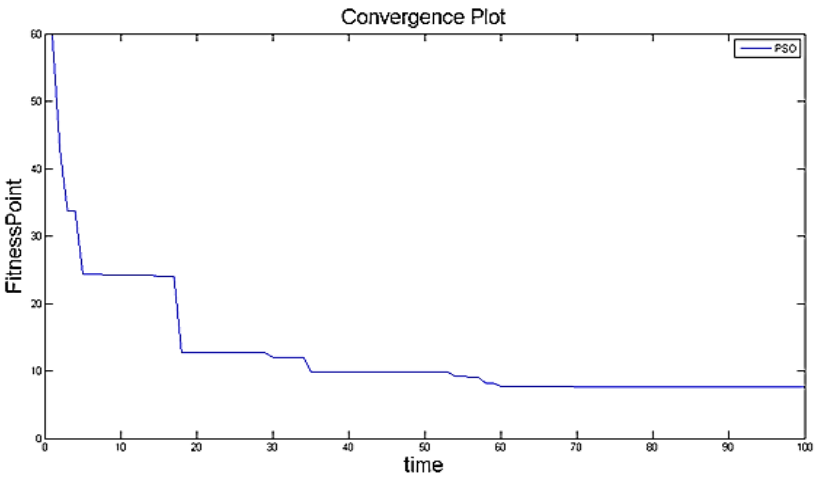

Figure 4- The Convergence of PSO Algorithm in order to Features Extraction

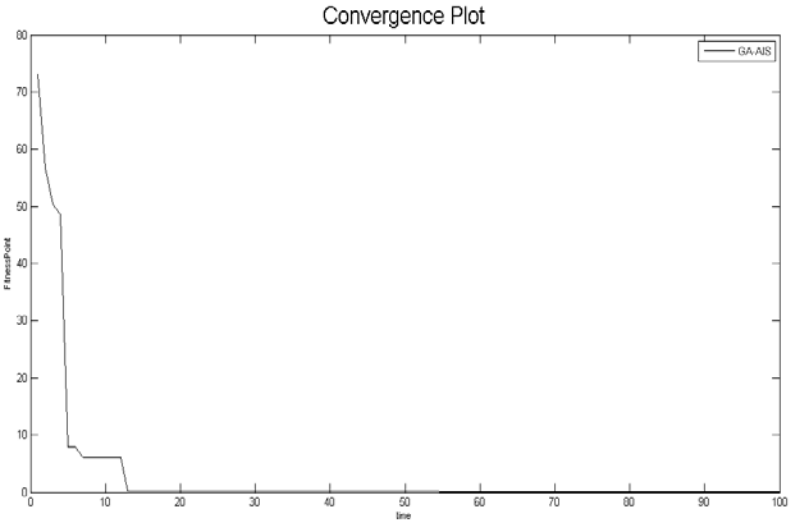

Figure 5- Convergence of Genetic Algorithm and Immune System in order to Features Extraction

Figure 6 shows the runtime of algorithms at the training and testing phase and the number of features' entries in various aspects. With increase of the number of features, computational complexity of features extraction increases, hence, algorithms' running time increases. For example, although the accuracy of Bayesian algorithm is assumed high, with increase of the number of input features, its run-time increases significantly. In contrast, assuming the high accuracy of the proposed algorithm, with increase of the number of input features, it has the lowest run-time.

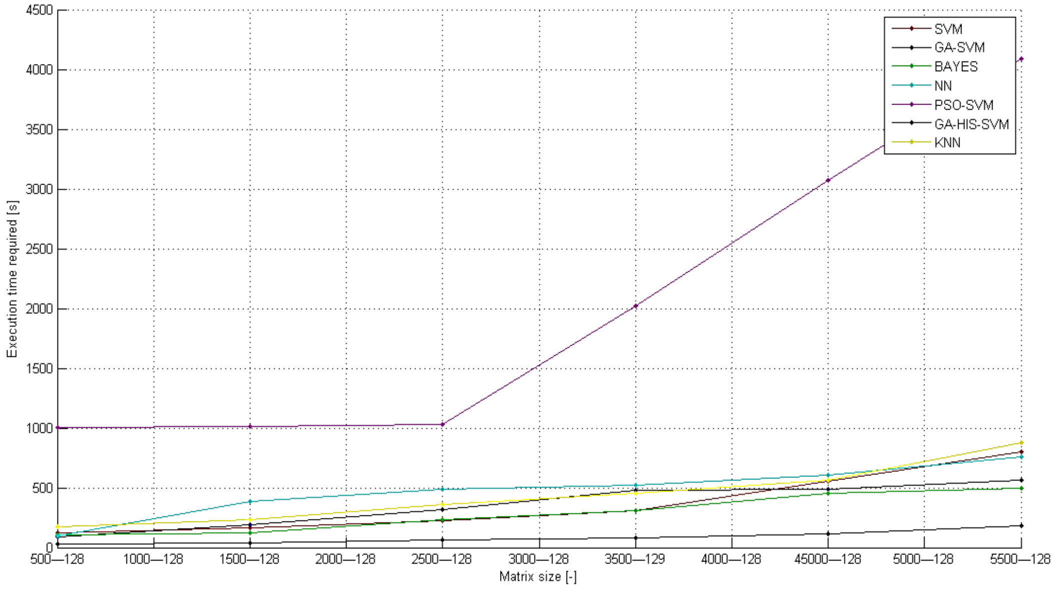

Figure6- Run time Algorithms(feature dimensions is 5500)

## Conclusion and Future Works

In the current study, the combination of two genetic and immune system algorithms for feature extraction and SVM for classification were presented. This approach was evaluated considering a number of other algorithms and standard datasets of SpamAssassin. The evaluation results showed that the proposed methodology for the dataset has the best performance compared to theses algorithms. Also, the proposed algorithm had a suitable convergence speed in feature extraction, compared to PSO, GA the use of similar combinations with other evolutionary algorithms such as gradual annealing of SA instead of immune system algorithm is one of the areas of future research of the present study.

## References

[1] S. Mohammed, O. Mohammed, J. Fiaidhi, S. Fong, T. H. Kim, “Classifying Unsolicited Bulk Email (UBE) using Python Machine Learning Techniques”, International Journal of Hybrid Information Technology, Vol. 06, No. 01, pp. 43-56, 2013.

[2] F. D. Garcia, J. H. Hoepman, J. V. Nieuwenhuizerr, ”spamfilter analysis”, SEC, pp. 395-410, 2004.

[3] H. Alkahtani, P. Gardner-Stephen, R. Goodwin, ”A Taxonomy of Email SPAM Filter”, pp. 356-363, 2011.

[4] W. A. Awad, S. M. Elseuofi, ”Machine Learning Methods for E-mail Classification, International Journal of computer Applications”, Vol.16, No.01, pp. 39-45, 2011.

[5] J. N. Shrivastava, H. B. Maringanti, ”E-mail Spam Filtering Using Adaptive Genetic Algorithm”, Intelligent Systems and Applications,Vol. 02, pp. 54-60, 2014.

[6] H.Drucker, V.N.Vapnik, ”Support Vector Machines for Spam Categorization. IEEE Transactions on Neural networks 10, 5. FANO R. 1961, pp. 1048–1054, 2012.

[7] U.Sanpakdee, A.Walairacht and S.Walairacht, “Adaptive Spai Mail Filtering Using Genetic Algorithm”, International Conference Advanced Communication Technology, Vol. 1, pp. 441-445, 2006.

[8] I.Idrise, A.Selamat, N.Nguyen, O,Krejcar “ A combined negative selection algorithm- particale swarm optimization for an email spam detection system” Engineering Application of Artificial intelligence 39, pp. 33-44, 2015.

[9] L.Wei, Y.Yang, R.M.Nishikawa, M.N.Wernick, A. Edwards,” Relevance vector machine for automatic detection of clustered microcalcifications”, Medical Imaging, IEEE Transactions, Vol. 24, pp.1278-1285, 2005.

[10] V.P.Deshpande, R.I.Erbacher,” An Evaluation of Naive Bayesian Anti-Spam Filtering Techniques”, Information Assurance and Security Workshop, PP.333-34, 2007.

[11] C.Huang, L.Wang, "A GA-based feature selection and parameters optimizationfor support vector machines", Expert Systems with applications, Vol. 31, pp. 231-240, 2006.

[12] F.Sebastiani, "Machine learning in automated text categorization", A CM Computing Surveys, Vol. 34, No. 1, pp.1-47, 2000.

[13] F.Temitayo,Q.Stephe, A.Abimbola," Hybrid GA-SVM for Efficient Feature Selection in E-mail Classification", Computer Engineering and Intelligent Systems, Vol 3, No.3, pp. 17-29, 2012.

[14] J.Perez, J.Basterrechea," Comparison of Different Heuristic Optimization Methods for Near-Field Antenna Measurements", IEEE Transaction on Antennas and Propagation, V[15], 2007.

[15] E.Chandra, K.Nandhini," Learning and Optimizing the Features with Genetic Algorithms" , International Journal of Computer Applications, Vol.9, No.6, pp. 1-5ol.55, pp. 549-55, 2010.