3rd International Conference on Knowledge-Based Engineering and Innovation (KBEI-2016)
Dec. 30th, 2016
(Tehran Province Payame Noor University) – Tehran, Iran

Providing an Improved Feature Extraction Method for Spam Detection Based on Genetic Algorithm in an Immune System

Zahra Razi
Department of computer, Karaj Branch, Islamic Azad University, Karaj, Iran
Razizahra88@gmail.com

Seyyed Amir Asghari
Electrical and Computer Engineering Department, Kharazmi University of Tehran, Iran
Asghari@khu.ac.ir

www.irantract.com

Abstract*—The increasing use of e-mail in the world because of its simplicity and low cost, has led many Internet users are interested in developing their work in the context of the Internet. In the meantime, many of the natural or legal persons, to sending e-mails unrelated to mass. Hence, classification and identification of spam emails is very important. Many studies on spam indicate that it costs organizations billions of dollars annually. We introduce a Genetic Algorithm (GA) assisted Artificial Immune System AIS in spam detection, and compare between two methods. Results were tested on 1000 standard datasets of Spam Assassin email. The proposed method may be used in conjunction with other filtering systems to minimize errors and Run time algorithms.

Keywords*—Spam, Classification, Genetic Algorithm, Artificial Immune System, Email.

I. Introduction

E-mails are one of the most important forms of communication; e-mails are simple, effective, and a cheap type of communication for almost all computer users. This simplicity and cheapness are prone to a lot of threats. One of the most significant of them is spam; spam e-mails are a problem that almost every e-mail user suffers from [1]. The word spam usually denotes a particular brand of luncheon meat, but in recent times, spam is used to represent a variety of junk, unwanted e-mails [2].

It is now possible to send thousands of unsolicited messages to thousands of users all over the world at approximately no cost. As a result, it is becoming common for all users worldwide to receive hundreds of spam messages daily. There are several approaches which try to stop or reduce the huge amount of spam which target individuals. These approaches include legislative measures such as worldwide anti-spam laws [3]. Other techniques are known as Origin-Based filters which are based on using network information and IP addresses in order to detect whether a message is a spam or not [4]. The most common techniques are filtering techniques, attempting to identify whether a message is spam or not based on the content and other characteristics of the message. In spite of the large number of methods and techniques available to combat spam, the volumes of spam on the internet are still rising. This work presents a new solution for spam inspired by Artificial Immune System model (AIS). With the help of Genetic Algorithm (GA) a lot of modifications on standard artificial immune system are sought in order to make it work more efficiently [5].

This work also includes a comparison study between genetic optimized spam detection using AIS and standard AIS for spam detection. In the following, the previous works and researches are reviewed in Section 1.Then in section 1 different phases in the spam detection are investigated , the details of the proposed method is examined in Section 1evaluation and the results of the proposed method and finally the conclusion are discussed through Sections 4 and 5.

II. Literature Review

SVM (Support Vector Machine) was used to classify emails [14] and its performance was compared with Ripper, Rocchio and decision trees. The results showed that increasing the BOOSTING trees and SVM presented acceptable performance in terms of accuracy and speed. However, the training time of boosting trees is very long. Email classification system based on SVM allows users to quickly recognize unwanted emails. A combination of new filtering of spam was performed by composition of NB-AIS and analysis of these two algorithms. In [7] NB used anti-spam method in the mentioned study on the basis of statistics and the gained classification accuracy rate was very high. But it's self-learning and self-matching is very weak. Artificial immune system benefits from excellent performance in recognition, learning and recording in the memory. The mechanism related to NB and the immune system have been combined together and gained a spam filtering algorithm,

Universal Scientific Organization ©KBEI-2016

then have solved the related issues and important problems of algorithm. The results demonstrated that the algorithm not only achieved high classification accuracy at first, but also took advantage of self-learning, self-adaptability (self-adjustment) and robustness as well. The artificial immune system was also used to detect spam [8]. In particular, the immune system has been tested by multiple methods of e-mail messages classification with the detector caused by the immune system.

The resulting system classified the messages with similar accuracy compared to other spam filters but used fewer detectors, so it was taken into account as a considerable solution in the processing time of spam and partly solved a lack of optimization methods in SAIS(Simple Artificial Immune System). According to [9], SVM was much more efficient than other non-parametric classifications like neural networks, K nearest neighbor(KNN), in terms of classification accuracy, computational time and parameter settings, but it performed weakly in classifying large datasets with high dimensions and a large number of features.

A hybrid system based on optimized colony of ants algorithm and SVM classifier was presented in [10]. In this study, the classic ant colony optimization algorithm was proposed in which the appropriate features were selected and the parameters of the SVM algorithm were optimized simultaneously. Based on [11], a hybrid classification system of genetic algorithm and SVM was proposed. In the mentioned study, genetic algorithm was offered to select suitable features and simultaneous with parameter optimization of SVM, coding of the chromosomes was also suggested.

Research results showed that SVM consumes plenty of time and memory space to classify large volumes of data sets. In this study, the experiences and the outcomes of mentioned research were implemented and it has been understood that to solve the problem of SVM classification, it is needed to select the most effective features as candidate features instead of choosing all properties in the sample space, and to use them as support vectors. So the efficiency and accuracy of SVM classification are maintained when the number of features in the data space is increased.

III. THE PROCESS OF IDENTIFYING SPAM

Figure 1 shows a system of content-based spam detection.

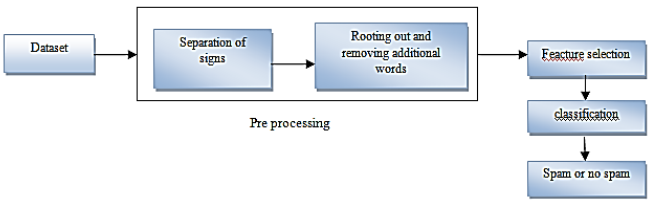

Figure 1. Identification of Spam Based on Content

A. preprocessing

Redundant words are typical words that do contain little information rate. In addition, they exist in all the texts in large numbers and have no effect on distinguishing text from others. Almost the root of word is also brought instead of derivatives of a word, thus the number of words in the document is somewhat reduced [12]. Next phase is decreasing dimensions of the vector corresponding to a letter. As a number of features or in other words the number of words in the classification are a lot, the dimensions of the problem increases to the large scale and it is a time-consuming and costly work. In order to the reduce dimensions of a vector in a letter, multiple methods of feature selection of texts can be used [13].

B. Formation of Feature vector

As the classification and filtering of spam are done in a phase of classification, it is essential to note that categorization takes place based on a series of features. The efficiency of all categories largely depends on diminishing features. Feature reduction can be executed in two ways including feature selection (on the basis of specific criteria) and feature extraction (by combining features). Existing methods use mainly one of these mechanisms to decrease their features. The proposed feature reduction method implements combination of these two methods for producing the final feature vector.

C. Selection, Extraction and Classification of Features

A subset of words is selected that contain more beneficial information. In other words inappropriate features which are extracted from the previous stage are removed from the feature vector according to the following steps: First, process of feature selection causes a subset of the initial features to be chosen, then feature extraction on the following set is applied and the final feature vector is made.

The study takes the advantage of combining of genetic algorithm and artificial immune system on extracting features. In the section based on genetic algorithm .The operation of feature extraction is executed by a cost function (optimized with genetic algorithm). It is worth noting that the selection of the best choices for the next stages is done in parallel by the algorithm of immune system in each phase. Results obtained from Implementations and analyzing the proposed method represent an increase of optimality in classifiers that utilize these features.

IV. 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

www.irantract.com

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 2, 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.

The subset of selected feature is performed based on evaluation function. It is used corresponding to the reduction of feature space and measurement to obtain the optimal solution set. When the best subset of feature is found, the output is recommended as a set of features which is used in SVM classification system or any other classifier in order to classify and predict spam emails from non-spam. As mentioned, the first generation of genetic algorithm is initialized by performing the output of response set of immune system. Then, convenient features are obtained from the next generation by applying selection, mutation, change on populations and change with the best solution using general search. Then, the next step of the algorithm of immune system is run locally and simultaneously. Immune system and genetic algorithm select a subset of features in parallel and separately; then, evaluate them based on merit criteria and select convenient features. When an optimal subset is found, it is placed in responses of immune system and in this way, its' fitness function will be updated. The number of implementing this algorithm is based on maximum number of generations.

If the number of antibodies in each division is higher than the number of generations, the number of algorithm repetition will be the number of antibodies and the number of generations will be added. If the number of generations is more than the number of genes, the number of repetition will be based on the number of generations and the next gene is created randomly. Hence, the algorithm implementation process will be continued as long as to reach its' maximum amount. After that, the algorithm implementation will be stopped and the output returns the best subset of features.

In this way, step by step stages of selecting the feature by combining the genetic algorithm and immune system is as follows:

1. The values of desired parameters of the two algorithms are initialized.
2. The subset S_i which has not been selected to determine convenient features since now and the nearest subset with contrary tag with subset S_i are selected based on the number of common features. It should be noted that this subset is selected by the immune system algorithm.
3. It initializes the test data set by accumulation of emails belonging to the subsets of S_i and S_j.
4. It generates the initial population to the size of P chromosomes made of zero and one randomly so that the length of each chromosome be equal to the length of candidate vector of subset S_i.
5. It creates the selected vector corresponding to each chromosome belonging to P population, so that the final candidate vector corresponding to each chromosome includes features of representative vector of the subset S_i which is equal to one corresponding to that feature in desired chromosome.
6. It classifies the test samples using the created candidate vector and the vector of the subset S_j.
7. After classification of all test samples, the error values of FN and FP are calculated and the fitness value for each chromosome is calculated corresponding to the final candidate vectors based on above equation.
8. The selection probability of each chromosome is calculated and r percent of output chromosomes from selection step of AIS algorithm are added to the new population of P_s with higher selection probability. After applying recombination and mutation operators, the new population of P_s is created.

www.irantract.com

9. Steps 5 to 8 are repeated per each chromosome until the difference between the highest fitness criteria in a sequence of population is lower than a small value of Ɛ.
10. If the convenient features of representative vector is per all available calculated subsets of immune system algorithm, it will be finished, otherwise, it will go to step 2.

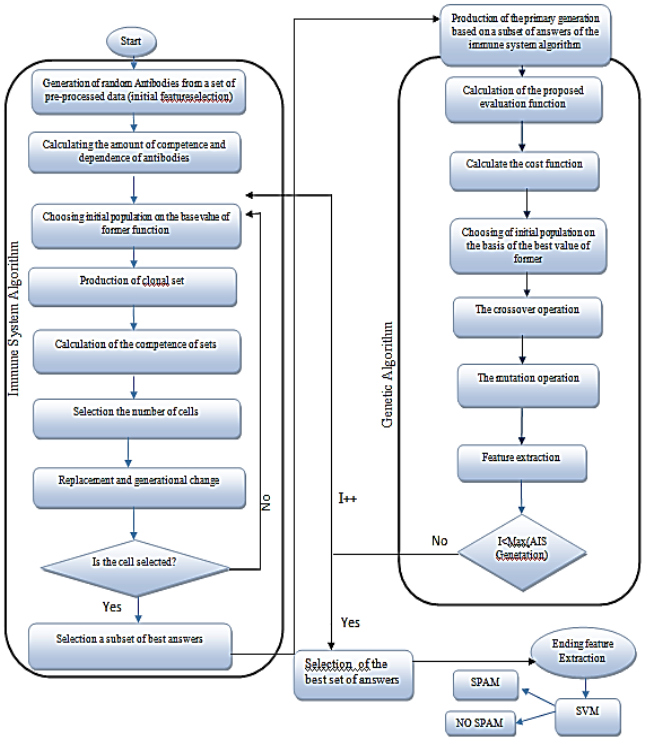

Figure 2. The Flowchart of the Proposed Algorithm

V. EXPERIMENTAL RESULTS

All of our experiments were done on a machine with 5 GHz CPU, 4 GB RAM, and the Windows 7 operating system. Classification process was implemented by the Matlab software package using data generated by a Java application. In the proposed system, criteria of accuracy and error are used for evaluating system performance which has been presented in equations of (1) and (2) [16].

Relation (1)

$$Error\ Rate = \frac{FP + FN}{TN + FP + FN + TP}$$

Relation (2)

$$Accuracy = \frac{TP + TN}{N}$$

TN: The number of records that their real group is negative and classifying algorithm has correctly diagnosed negative.
FP: The number of records that their real group is negative and classifying algorithm has mistakenly detected their class as positive group.

FN: The number of records that their real group is positive but classifying algorithm has mistakenly detected their group as negative one.

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

In order to evaluate implementation and feature selection, two algorithm sets were tested, with the relevant parameters shown in Table 1. also Each machine learning system requires a training set to train the system. In this paper, we have used Spam Assassin.

TABLE1. THE TWO ALGORITHM SETS USED FOR EVALUATION

parameter	explanations	amount
Genetic Algorithm		
Q	Initial Population	100
S	number of repeat	100
M	number of children	8
NT	Mutation rate	٠/2
N	Recovery rate	٠/9
Immune system Algorithm		
Ncells	number of cells	10
MR	Mutation rate	٠/1
CR	Clonal rate	10
NS	number of samples	14
Sthrsh	Stimulation rate	٠/9

VI. THE RESULTS OF THE PROPOSED METHOD

The results showed the accuracy of the proposed method compared to algorithms of SVM, SVM-GA. the accuracy of mentioned method is also higher than SVM-GA and SVM when there is an increase in number of features. On the other hand, comparing the error rate, it is concluded that the noted rate in the proposed method is lower than SVM-GA.The error rate of SVM-GA is also less than SVM .therefore, error rate of the proposed method is lower than mentioned algorithms due to a combination of genetic and immune systems. In figure 3 accuracy rates of SVM, SVM-GA and proposed method are shown. According to the results, accuracy rate of the proposed algorithm is higher than of SVM-GA and SVM. In Figure 4 FP error rates of the proposed method in comparison with SVM-GA and SVM algorithms are demonstrated.

www.irantract.com

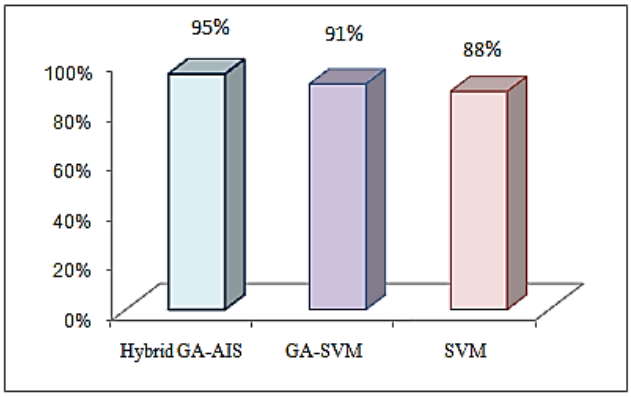

Figure 3. Accuracy rate of SVM-GA and SVM and proposed algorithms

As can be seen in figure 3, accuracy of combination of the algorithms GA-AIS is higher than the algorithms GA–SVM and SVM. Given that the algorithm SVM perform the operation of selecting features and classification by itself, it is expected that its' accuracy reduces dramatically by increasing the number of samples. It is expected that GV-SVM algorithm has reduction in accuracy by increasing the number of samples due to weakness in reviewing features in local searches.

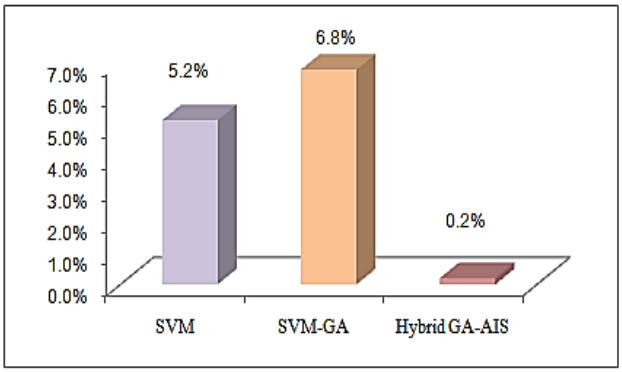

Figure 4. FP Error Rate of the Proposed Method in Comparison with SVM-GA and SVM Algorithms

In figure 4, the error rate of FP shows the Hybrid GA-AIS with SVM-GA and SVM algorithms. FP error rate in figure 4 indicates that a number of samples, which their real category was email, were incorrectly classified in spam category by classifier algorithm; in this case, SVM-GA showed a better perfon

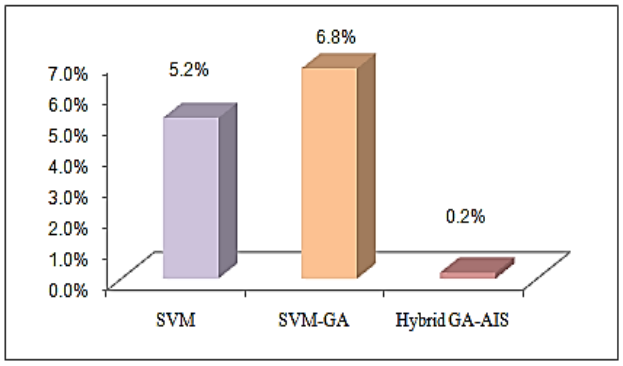

Figure 5. FN Error Rate of the Proposed Method with SVM-GA and SVM Algorithms

Figure 5 indicates the comparison of FN error rate in Hybrid GA-AIS with SVM-GA and SVM algorithms. Figure 5 illustrates that a number of samples, which their real category was spam, were incorrectly classified in email category by classifier algorithm; in this case, the proposed algorithm showed a better performance.

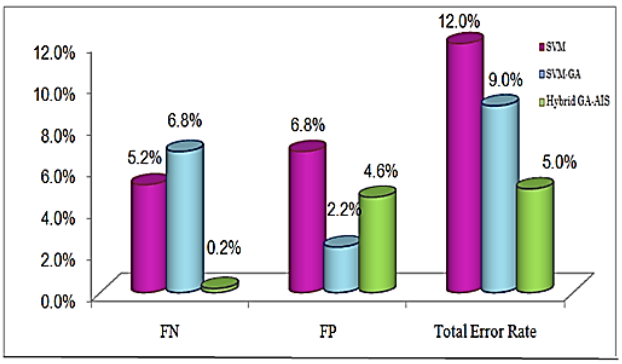

Figure 6. Total Error Rate in each of the Algorithms

It should be noted that by obtaining the results of the other tests and experiments, it is concluded that the by increasing number of features, the accuracy of the proposed method is higher rather than other algorithms. According to the conducted tests, the highest accuracy rate is in figure 3 and the lowest error rate in figure 6; identification of spams on determined datasets is related to the proposed algorithm which is shown clearly in figures. Figure7 displays the convergence of GA algorithm without using genetic algorithm and immune system in order to extract features to compare the results of problem solving period.

www.irantract.com

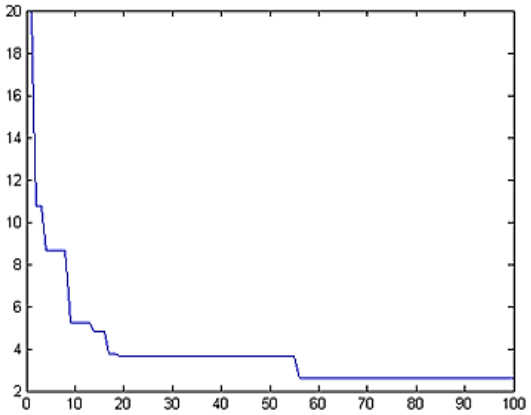

Figure 7. The Convergence of GA Algorithm in order to Features Extraction

Figure 7 indicates the convergence of GA algorithm without using the combination of two GA-AIS algorithms in order to extract the feature and compare the results of problem solving period with each other. As can be seen in figure 7, the score of the cost function of genetic algorithm chromosomes is trapped in a local optimum at the stage of about twentieth to fifty-fifth of algorithm generation; and once has got a better score of fifty-six and exit from this local optimum. Then, again it is trapped in other local optimum at stage of fifty-eight. This indicates poor performance of the algorithm in the face with local optimums along with good speed of genetic algorithm convergence.

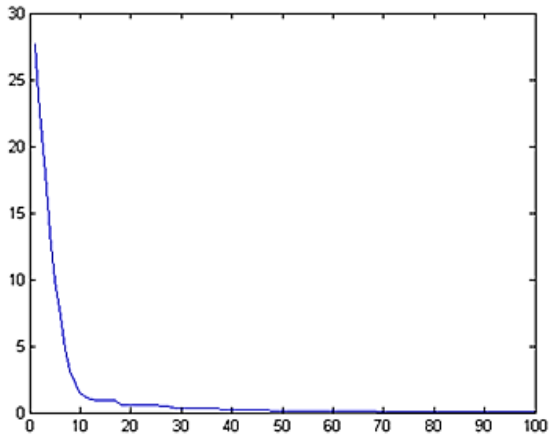

Figure 8. Convergence of Genetic Algorithm and Immune System in Order to features Extraction

Figure 8 shows the convergence of algorithm resulting from combination of genetic algorithm and immune system in order to extract features. As it is observed, the proposed algorithm has faster convergence. Cost function values of the proposed algorithm, become closer to zero considerably in 28 times of repetitions and reach to the border convergence that is another advantage of the combination of these algorithms.

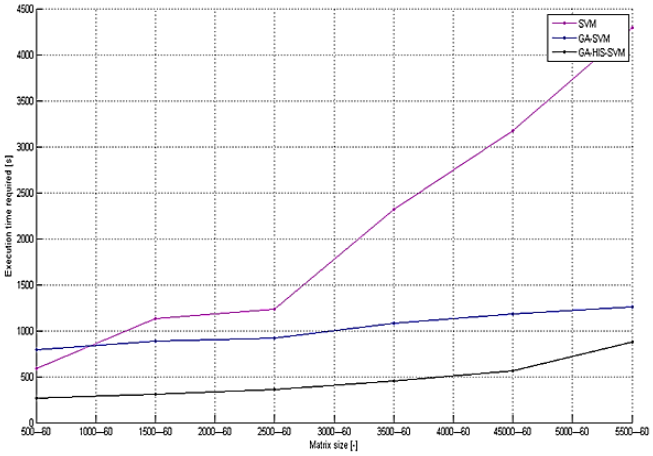

Figure 9. Run time algorithms (feature dimensions is 5500)

Figure 9 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 SVM 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.

VII. DISCUSSION AND FUTURE WORKS

In this paper, 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 Spam Assassin. The proposed method has the lower FN error rate than SVM-GA and also lower FP error rate than SVM. With the combination of GA-AIS, spam detection accuracy rate is %95 in which a significant improvement is observed compared to the SVM-GA. In future research, more investigations should be performed to study the choice of fitness function, enhance of the operators, increase effectiveness of methods and reduce more features. Research into the possibility of using the algorithm of intelligent particles optimization PSO to select and separate spams, is one of the areas of future studies.

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, 6(1) (2013) 43-56.

[2] S. Guzella, M. Caminhas, {A review of machine learning approaches to spam filtering}, Expert systems with applications, 3(6) (2009)10206-10222.

[3] H. Alkahtani, P. Gardner-Stephen, R. Goodwin, {A Taxonomy of Email SPAM Filter}, (2011) 356-363.

www.irantract.com

[4] W. A. Awad, S.M. Elseuofi, {Machine Learning Methods for E-mail Classification}, International Journal of computer Applications, 6(1) (2011) 39-45.
[5] J. N. Shrivastava, H. B. Maringanti, {E-mail Spam Filtering Using Adaptive Genetic Algorithm}, Intelligent Systems and Applications, 2(2014) 54-60
[6] H. Drucker, V. N. Vapnik, {Support Vector Machines for Spam Categorization}, IEEE Transactions on Neural networks, (2012) 1048-1054.
[7] M. Zavvar, M. Rezaei, {Email spam detection using combination of particale swarm optimization and artificial neural network and support vector machine}, modern education and computer science 7, (2016) 68-78.
[8] p. saurabh, B. Verma, {an efficient proactive artificial immune system based anomaly detection and prevention system}, expert systems with applications, 6(2016)311- 320.
[9] L. Wei, Y. Yang, R.M. Nishikawa, M.N. Wernick, A. Edwards,{Relevance vector machine for automatic detection of clustered micro calcifications}, Medical Imaging, IEEE Transactions, 2(4) (2005) 1278-1285.
[10] C.L. Huang, {ACO-based hybrid classification system with feature subset selection and model parameters optimization}, 7(3) (2009) 438-448.
[11] C. Huang, L. Wang, {A GA-based feature selection and parameters optimization for support vector machines}, Expert Systems with applications, 3(1) (2006) 231-240.
[12] B. Agarwal, N. Mittal, {Text Classification using machine learning methods}, proceedings of the second international conference on soft computing, (236) (2014)701-709.
[12] F. Temitayo, Q. Stephe, A. Abimbola, {Hybrid GA-SVM for Efficient Feature Selection in E-mail Classification}, Computer Engineering and Intelligent Systems, 3(3) (2012) 17-29.
[13] J. Perez, J. Basterrechea, {Comparison of Different Heuristic Optimization Methods for Near-Field Antenna Measurements}, IEEE Transaction on Antennas and Propagation, 5(5) (2007) 549-555.
[14] E. Chandra, K. Nandhini, {Learning and Optimizing the Features with Genetic Algorithms}, International Journal of Computer Applications, 9(6) 32(10) 1-50.
[15] I. Idris, A. Selamat, {Improved email spam detection model with negative selection algorithm and particle swarm optimization}, Applied soft computing 22 (2014) 11-27.

www.irantract.com