Detection of Spam Email by Combining Harmony Search Algorithm and Decision Tree

Mehdi Zekriyapanah Gashti
Department of Computer Engineering
Payame Noor University
Iran
gashti@pnu.ac.ir

***Abstract*—Spam emails is probable the main problem faced by most e-mail users. There are many features in spam email detection and some of these features have little effect on detection and cause skew detection and classification of spam email. Thus, Feature Selection (FS) is one of the key topics in spam email detection systems. With choosing the important and effective features in classification, its performance can be optimized. Selector features has the task of finding a subset of features to improve the accuracy of its predictions. In this paper, a hybrid of Harmony Search Algorithm (HSA) and decision tree is used for selecting the best features and classification. The obtained results on Spam-base dataset show that the rate of recognition accuracy in the proposed model is 95.25% which is high in comparison with models such as SVM, NB, J48 and MLP. Also, the accuracy of the proposed model on the datasets of Ling-spam and PU1 is high in comparison with models such as NB, SVM and LR.**

Keywords- Spam Email; Harmony Search Algorithm; Decision Tree

I. INTRODUCTION

Spam spotting is a significant task for all web actions and especially for email clients [1]. Spam emails consume a considerable amount of traffic volume and also may carry viruses [2]. Since the security of computer systems is based on the three principles of prevention, diagnosis, and response, if all the security risks were identified and prevented, there obviously would be no need for reaction. Hence, identification is a vital method for providing email users' security and it can be the front defensive line of for any computer system. Spam senders are always using more complicated tools and methods for getting through the spam filters. Hence, emails' security against attacks on email servers and the trial for accessing valid usernames or email addresses through accessing mail servers is a critical requirement [3]. One way for keeping unauthorized access to users' email accounts is a two-step verification of identity. Two-step verification is a security mechanism that uses a second keyword or phrase in addition to the password [4].

Spotting spam email is mainly based on characteristics and features written in the subject field of emails. Most spams use similar subjects. Therefore, this is a unique feature for identifying. The present work utilizes the Spam-base dataset [5] including the two classes spam and non-spam for the purpose of spam identification based on a combination of HSA [6] and ID3 [7]. In addition, the Ling-spam [8] and PU1 [8] datasets were used for assessment and comparison. HSA was used for selecting characteristics, increasing accuracy in the final solution, and for determining the features that take the fitness function to its optimal situation. ID3 was used for class recognition and final categorization.

II. RELATED WORKS

Various techniques have been introduced for identifying spams, including statistical techniques, expert systems, Bayesian networks, neural networks, fuzzy logic, and collective intelligence algorithms. A Radial Basis Function (RBF) model is set forward alongside with Support Vector Machines (SVM) technique for identifying spam emails in [9]. RBF, an artificial neural network model, is used for training and testing data; and SVM, a classification technique, is used for mapping the features. Assessment is done on Double Bounce Email, a linear dataset. In addition, preprocessing and determining the frequency of words according to (1) is carried out by use of TF-IDF. Results show an identification accuracy of approximately %84.

$$TF-IDF(t,d,D)=\frac{1}{\sum_{t} f_d(t)}\times TF\times IDF=\frac{1}{\sum_{t} f_d(t)} f_d(t)\times\log\left(\frac{|D|}{1+|\{d\in D: t\in d\}|}\right) \quad (1)$$

In [10], a Bayesian Additive Regression Trees (BART) model was carried out on Ling-spam, PU1, and Spam-base. BART model is a binary DT with linear regression correlations at every end node that are able to predict numerical values. The most important and main criterion for assessment of the DT is the error rate created in the tree. For calculating the total error rate of the tree the weighted total of the error rate of the leaves is calculated. In order to prevent low quality laws being created, some branches are pruned. Although this pruning causes a higher error rate, it will stop inefficient laws being created. Results show that in the models CBART, Random Forests (RF), BART, and Classification and Regression Trees (CART) the accuracy on Ling-spam and PU1 is 100%. Also, on Spam-base, the highest amount in RF model is 98.61%.

CART model reduces the error rate by 2.2% in comparison with BART.

An Enhanced Genetic Algorithm (EGA) model was used in [11], which is an enhanced version of Genetic Algorithm (GA) through combination with Simulated annealing (SA), is proposed for spotting spam emails. Assessment is carried out on Spam Corpus with 54 characteristics. In EGA, 15 chromosomes are used with 54 characteristics and 1000 generations. Results suggest an accuracy of 99.73% and 99.86% for GA and EGA respectively. Therefore SA has been very effective in enhancing GA, and has improved its operators, and increased its accuracy. A Particle Swarm Optimization (PSO) model, a population based algorithm, has been proposed for identification of spam emails in [12]. GPU technology is used for running PSO model. GPU processes tasks in parallel and processes many tasks better and faster than CPU. Assessment is made on TREC 2015 with 48360 spam emails and 36450 non-spam emails. The probability of identifying spam emails is calculated with (2). Results suggest an accuracy of 99% in spotting non-spam emails and an accuracy of 66% to 99% in spotting spam emails.

$$\Pr(S \mid W) = \frac{\Pr(W \mid S) \times \Pr(S)}{\Pr(W \mid S) \times \Pr(S) + \Pr(W \mid H) \times \Pr(H)} \qquad (2)$$

A Bayesian classification was done on three datasets with 1000, 1500, and 2100 emails in [13]. Bayesian classification identifies and predicts data according to probabilities. Bayesian model includes the steps of preprocessing, training, testing, and classification. Results show that the accuracy in Dataset1, Dataset2, and Dataset3 is 93.98, 94.85, and 96.46 respectively. In addition, the processing time for Dataset1 is less than the other two datasets. Identification of spam emails was performed on RFC2822 with 9189 emails according to Vertex Dependency [14]. In this model, the neighboring vertices relationships are used for prediction. The data similarity distance between the vertices is calculated with (3). Results point to a maximum accuracy of 93.78%. In addition, the accuracy in identification of spam email is 80.72% and for non-spam emails, it is 98.01%.

$$sim = \frac{A.B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \qquad (3)$$

A combined model PSO+K-Means in [15] was performed on Spam-base with 57 characteristic for spam email spotting. In PSO+K-Means model, PSO is used for characteristic selection and K-means is used for data clustering. In K-Means algorithm, *k* members are selected out of *n* members randomly as cluster centers. Afterwards, the remaining *n-k* members are allocated to the closest clusters. Once all members added to clusters, cluster centers are calculated again and members are allocated to clusters anew according to the new cluster centers. This process is carried on until the cluster centers are fixed. Distance factor according to Equation (4) is used for clustering. Results suggest that the maximum accuracy of the model is 94.62%.

$$dis(x, y) = \sum_{i=1}^{n} |x_i - y_i|^2 \qquad (4)$$

A combined model Multi-Layer Perceptron Artificial Neural (PSO-MLPNN) was proposed for spam email identification in [16]. PSO algorithm is used for characteristic selection and MLPNN model is used for training, data testing. In PSO-MLPNN model the perceptron neural function is used with sigmoid activation function for the hidden layer; %80 of the data is used for training and %20 for testing. The number of the hidden layers in MLP model is considered to be between 3 and 15; and the repetition of PSO algorithm for characteristic selection is 200. Assessment was done with 481 spam emails and 2171 non-spam emails on Ling-Spam and 6000 email samples on Spam-Assassin. Assessment done on Spam-Assassin and Ling-Spam suggests that the accuracy in PSO-MLPNN model is respectively 99.98 and 99.79. Comparison shows that PSO-MLPNN model is more accurate in identification than SVM and BPNN. In Table I, a comparison of the proposed models for spam emails identification is presented.

TABLE I. COMPARISON OF PROPOSED MODELS FOR SPAM DETECTION

Models	Data set	Preprocessing	Classifier	FS	Correlation	Time Complexity
RBF [9]	●Double Bounce	X	√	√	High	Low
SVM [9]						
BART [10]	●Ling-Spam ●PU1 ●Spam-base	√	X	√	Medium	Medium
EGA [11]	●Spam Corpus	√	X	√	Medium	Low
PSO [12]	●TREC 2015		X	√	Medium	Low
Bayesian [13]	●Dataset1=1000 ●Dataset2=1500 ●Dataset3=2000	X	√	X	High	Medium
Vertex Dependency [14]	●RFC2822	X	X	X	Medium	High
PSO+K-Means [15]	●Spam-base	√	√	√	High	Medium
PSO-MLPNN [16]	●Ling-Spam ●Spam-Assassin	√	√	√	High	Low

III. PROPOSED MODEL

In the proposed model, first the data in Spam-base dataset is preprocessed. At the preprocessing level, data are controlled; because the data in the dataset might not be controlled enough and inapplicable, repeated, or erroneous values may result in invalid output. Presence of inapplicable data in most results in dysfunctions in conclusion obtained from the data. In the next step, primary vectors form in HSA. Each vector is comprised of 57 characteristic. In the vectors, a number of characteristics are selected based on HSA memory randomly and transferred to ID3. In ID3 tree classification rules according to characteristics are carried out. Characteristics that are influential in identification accuracy are saved in HM memory and used for later FS steps. Assessment function in HSA does a complete and exhaustive search in the space of the subset of characteristics until it finds the best combination of characteristics. The most important step in the ID3 tree is setting the rules. According to the root of the tree, rules are set and characteristics are compared. The most important criterion for the root node is the data rate of the tree. The characteristic that has the highest data rate is chosen as the root node and the ID3 tree is expanded based on it. Afterwards, training and testing of data is carried out. Testing is done in order to assess data and its validity. In Figure 1, the flowchart of the proposed model is shown.

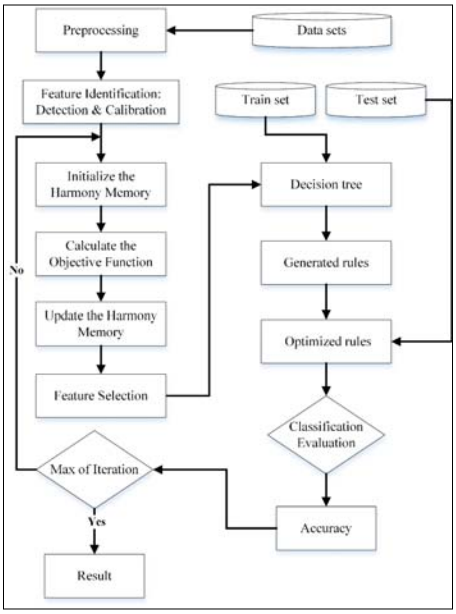
Preprocessing
Data sets
Feature Identification: Detection & Calibration
Train set
Test set
Initialize the Harmony Memory
Decision tree
Calculate the Objective Function
Generated rules
No
Update the Harmony Memory
Optimized rules
Feature Selection
Classification Evaluation
Max of Iteration
Yes
Accuracy
Result

Fig. 1. Flowchart of the Proposed Model

In HSA [6] Harmony Memory (HM) is used to maintain the best of the previous solutions. HM works as a matrix as done in Equation (5) with a solution in every line. Therefore, the number of the columns in this matrix actually shows the dimensions of the solution. The last column in the matrix is set for saving the value of the fitness function for every line. The amount of answer vectors in HM is shown using Harmony Memory Size (HMS). Output value of the fitness function for each vector is shown using *f(x)*.

$$HM = \left[\begin{array}{cccc|c} x_1^1 & x_2^1 & \dots & x_n^1 & f(x^1) \\ x_1^2 & x_2^2 & \dots & x_n^2 & f(x^2) \\ \dots & & & & \dots \\ x_1^{HMS} & x_2^{HMS} & \dots & x_n^{HMS} & f(x^{HMS}) \end{array}\right] \quad (5)$$

In HSA, first all HM lines are made and the value of the fitness function for each line is calculated and saved in the last column of the HM matrix. Then, based on the number of necessary repetition or the time repetition is finished, the entire HM is scanned and for each line the suitable value for each entry is set according to HSA parameters. Afterwards, if the value for the fitness function is better than the worst solution present in HM, replaces it. Eventually, the solution that produces the optimal value in the fitness function is chosen as the best available solution. In HSA, any value can select values randomly. Randomization is in fact utilized for increasing the variety of solutions.

ID3 tree [7] is a method for presenting a set of rules that result in a group or a value. In ID3 tree, a statistical value is used called data rate for clarifying how much a character is effective in the final identification. In ID3 tree, at first, the amount of disorder for every characteristic is calculated using entropy; and using its value for each characteristic, the data rate is calculated. Entropy displays the randomness as a mathematical figure. If the set S includes positive and negative samples of a set, the entropy of S in relation to Boolean categorization is defined as (6).

$$E(S) = -P\oplus \log_2 P\oplus - P\Theta \log_2 P\Theta \quad (6)$$

In (6), $P_{\oplus}$ is the ratio of positive samples to the all the samples, and $P_{\ominus}$ is the ratio of negative samples to the all the samples. The decision of which characteristic is to be in the root of the tree depends on the data rate of each characteristic. Equation (7) is used for calculation of the data rate [7].

$$IG(S,A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \times E(S_v) \quad (7)$$

In (7), the Values (A) parameter denotes sum of the A values and S_V is a subset of S for which A gives value V. The criteria used for the assessment of the proposed model are Precision, Recall, F-Measure, and Accuracy; accuracy is the most important item among those criteria [17, 18].

$$precision = \frac{TP}{TP+FP} \quad (8)$$

$$recall = \frac{TP}{TP+FN} \quad (9)$$

$$F - Measure = \frac{2 * precision * recall}{(precision + recall)} \quad (10)$$

$$accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (11)$$

$$AUC = \left(\left(\frac{TP}{TP+FN}\right) + \left(\frac{TN}{TN+FP}\right)\right) / 2 \quad (12)$$

TN represents the number of records that factually pertain to the negative set and were identified correctly as negative by the algorithm. TP represents the number of records that factually pertain to the positive set and were identified correctly as positive by the algorithm. FP represents the number of records that factually pertain to the negative set and were identified wrongly as positive by the algorithm. FN represents the number of records that factually pertain to the positive set and were identified wrongly as negative by the algorithm.

IV. Evaluation And Results

Assessment and results of the proposed model are obtained in MATLAB 2015 on Spam-base. Spam-base dataset includes 4601 samples with 57 characteristics. In addition, in order to illustrate its efficiency and accuracy, the proposed model was run on Ling-Spam and PU1 datasets and was compared with other models. Ling-Spam datasets includes 481 spam and 2412 non-spam emails. PU1 dataset includes 481 spam and 618 non-spam emails. The maximum repetition in HSA is 150 and to the end of maintaining variety and optimal solutions, the technique of unfit vectors omission was put to use. In Table II, assessment of the proposed model on Spam-base dataset with different number of characteristics is shown. As could be seen in Table II, the number of characteristics is very influential in identification accuracy; also, the type of the chosen characteristic is influential in the accuracy of the results. In Figure 2, the comparison of the number of characteristics in identification accuracy in the dataset Spam-base is shown in a graph. In Table III, the comparison of the proposed model with other models is presented. As can be seen in Table III the accuracy of the proposed model is 95.25%. The proposed model has a higher accuracy than models such as Naive Bayes (NB), SVM, J48, and most other models; but it also has a lower accuracy than Random Forest and Random Tree models. In Table IV the comparison of the proposed model with other models on Ling-Spam dataset is presented. The accuracy of the proposed model in Ling-Spam dataset is 99.80%, which is higher than models NB, SVM, NNET, and LR.

TABLE II. ASSESSMENT OF THE PROPOSED MODEL ON SPAM-BASE WITH DIFFERENT NUMBER OF CHARACTERISTICS

FS	Precision	Recall	F-Measure	Accuracy
15	94.66	97.30	95.96	98.05
17	94.54	95.65	95.09	97.65
20	94.21	92.15	93.17	98.00
23	93.84	95.80	94.81	95.08
28	93.99	94.77	94.38	94.15
30	95.31	93.10	92.70	96.72
33	91.48	94.48	92.96	98.32
35	92.03	93.41	92.71	96.61
38	94.69	95.34	95.01	94.90
40	93.11	90.57	91.82	93.11
45	92.89	93.68	93.28	96.46
50	93.34	94.35	93.84	95.12
53	95.78	92.22	93.97	93.15
57	90.15	91.23	90.69	95.25

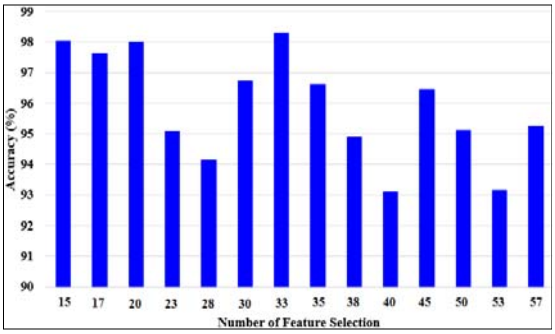

Fig. 2. Comparison Graph of the Influence of the Number of Characteristics on Detection Accuracy on Spam-base

TABLE III. COMPARISON OF THE PROPOSED MODEL WITH OTHER MODELS ON SPAM-BASE

Refs	Models	Accuracy (%)	Percentage of Comparisons
[19]	NSA-PSO	91.22	4.03
	PSO	81.32	13.93
	NSA	68.86	26.39
[20]	NB	79.3	15.93
[21]	SVM	90	5.25
[22]	DFS-SVM	71	24.25
[23]	ANN	86	9.25
[24]	Type-2 Fuzzy Set	86.9	8.35
[15]	PSO-Fuzzy	92.55	2.7
		94.62	0.63
[25]	NSA-PSO	83.20	12.05
[26]	Bayes Net	88.56	6.69
	Logic Boost	89.7	5.55
	Random Tree	91.54	3.71
	JRIP	92.32	2.93
	J48	92.34	2.91
	MLP	93.28	1.97
	KSTAR	93.56	1.69
[27]	NB	90.19	5.06
	Logistic	94.45	0.8
	KSTAR	93.84	1.41
	Filtered Classifier	92.76	2.49
	PART	93.91	1.34
	J48	92.97	2.28
[28]	NB	78.93	16.32
	Bayes Net	92.71	2.54
	SVM	86.54	8.71
	FT	95.54	-0.29
	J48	95.65	-0.4
	Random Forest	99.54	-4.29
	Random Tree	99.71	-4.46
	Simple Cart	93.93	1.32
-	**Proposed Model**	**95.25**	-

In Table V, the comparison of the proposed model with other models on PU1 dataset is presented. The accuracy of the proposed model in PU1 dataset is 97.12%, which is higher than models NB, and NNET. In Table Vi, the comparison of the proposed model with other models is shown on the dataset Spam-base. In the proposed model the accuracy in the dataset Spam-base is 93.25%, which is higher than NB models.

TABLE IV. COMPARISON OF THE PROPOSED MODEL WITH OTHER MODELS ON LING-SPAM

Ref	Models	Precision	Recall	F-Measure	AUC
[29]	CBART	100.00	100.00	100.00	100.00
	RF	100.00	99.57	99.78	100.00
	BART	100.00	100.00	100.00	100.00
	SVM	98.39	93.00	94.35	99.63
	NNET	96.97	46.22	62.18	99.68
	CART	100.00	100.00	100.00	100.00
	LR	99.28	97.86	98.54	99.77
	NB	100.00	33.48	49.71	66.55
Proposed Model		**98.77**	**99.35**	**99.06**	**99.80**

TABLE V. COMPARISON OF THE PROPOSED MODEL WITH OTHER MODELS ON PU1

Ref	Models	Precision	Recall	F-Measure	AUC
[29]	CBART	100.00	100.00	100.00	100.00
	RF	99.36	100.00	99.68	100.00
	BART	100.00	100.00	100.00	100.00
	SVM	97.94	94.78	95.92	99.25
	NNET	94.71	55.13	69.51	95.09
	CART	100.00	100.00	100.00	100.00
	LR	98.83	96.82	97.78	98.16
	NB	99.02	37.86	54.58	68.97
Proposed Model		**96.90**	**98.88**	**97.88**	**97.12**

TABLE VI. COMPARISON OF THE PROPOSED MODEL WITH OTHER MODELS ON SPAM-BASE

Ref	Models	Precision	Recall	F-Measure	AUC
[29]	CBART	95.24	92.72	93.96	96.40
	RF	95.42	92.64	93.99	98.61
	BART	94.78	88.43	91.48	97.60
	SVM	93.04	90.88	91.92	97.83
	NNET	93.17	92.89	93.01	97.77
	CART	88.38	85.82	87.05	94.73
	LR	92.32	88.98	90.60	97.19
	NB	57.52	95.54	71.81	88.74
Proposed Model		**90.15**	**91.23**	**90.69**	**95.25**

V. CONCLUSION AND FUTURE WORKS

Spam emails consume a huge bulk of email storage and compromise users' security. Therefore, measurements such as identification filters should be adopted. Content based filters that use emails' content are the main and most common type of spam filters. In most content based methods, machine learning and data mining are used. In addition, many identification filters scrutinize the content and subject of emails for existence of key words or phrases used in spam emails frequently. In conclusion, at first identification is the best method for avoiding spams. In the present paper, a model for spam identification was proposed based on the combination of HSA and ID3. Assessment was carried out on datasets Spam-base, Ling-base, and PU1. Results suggest that the proposed model has higher identification accuracy in comparison with models SVM and NB; and compared to most models increases the identification accuracy up to 15%. In spam email identification one of the main problems faced is selection of the type of the characteristic. In conclusion, in order to eliminate the problem, algorithms should be used that are capable of FS and can enhance identification accuracy.

REFERENCES

[1] S. Liu, Y. Wang, J. Zhang, C. Chen, Y. Xiang, "Addressing the class imbalance problem in twitter spam detection using ensemble learning", Computers & Security, 2016 (in press)

[2] A. Heydari, M.A. Tavakoli, N. Salim, Z. Heydari, "Detection of review spam: A survey", Expert Systems with Applications, Vol. 42, No. 7, pp. 3634-3642, 2015

[3] T. Ouyang, S. Ray, M. Allman, M. Rabinovich, "A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise", Computer Networks, Vol. 59, pp. 101-121, 2014

[4] N. Perez-Diaz, D. Ruano-Ordas, J. R. Mendez, J. F. Galvez, F. Fdez-Riverola, "Rough sets for spam filtering: Selecting appropriate decision rules for boundary e-mail classification", Applied Soft Computing, Vol. 12, No. 11, pp. 3671-3682, 2012

[5] https://archive.ics.uci.edu/ml/datasets/Spambase

[6] Z. W. Geem, J. H. Kim, G. V. Loganathan, "A New Heuristic Optimization Algorithm: Harmony Search", Simulation, Vol. 76, No. 2, pp. 60-68, 2001

[7] J. R. Quinlan, Induction of Decision Trees, Machine Learning, Vol. 1, No. 1, pp. 81-106, 1986

[8] http://www.csmining.org/index.php/spam-email-datasets-.html

[9] S. Ali, S. Ozawa, J. Nakazato, T. Ban, J. Shimamura, "An autonomous online malicious spam email detection system using extended RBF network", 2015 IEEE International Joint Conference on Neural Networks (IJCNN), pp. 1-7, 2015

[10] S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, "Bayesian Additive Regression Trees-Based Spam Detection for Enhanced Email Privacy", IEEE Third International Conference on Availability, Reliability and Security, pp. 1044-1051, 2008

[11] S. Salehi, A. Selamat, M. Bostanian, "Enhanced genetic algorithm for spam detection in email", IEEE 2nd International Conference on Software Engineering and Service Science, pp. 594-597, 2011

[12] M. Prilepok, T. Jezowicz, J. Platos, V. Snasel, "Spam detection using compression and PSO", IEEE Fourth International Conference on Computational Aspects of Social Networks (CASoN), pp. 263-270, 2012

[13] S. B. Rathod, T. M. Pattewar, "Content based spam detection in email using Bayesian classifier", IEEE International Conference on Communications and Signal Processing (ICCSP), pp. 1257-1261, 2015

[14] M. Prilepok, M. Kudelka, "Spam Detection Based on Nearest Community Classifier", IEEE International Conference on Intelligent Networking and Collaborative Systems, pp. 354-359, 2015

[15] S. Salehi, A. Selamat, O. Krejcar, K. Kuca, "Fuzzy Granular Classifier Approach for Spam Detection", Journal of Intelligent & Fuzzy Systems, vol. 32, no. 2, pp. 1355-1363, 2017

[16] A. R. Behjat, A. Mustapha, H. Nezamabadipour, M. Nasir Sulaiman, N. Mustapha, "A PSO-Based Feature Subset Selection for Application of Spam/Non-spam Detection", in Soft Computing Applications and Intelligent Systems, Communications in Computer and Information Science, Vol. 378, Springer, Berlin, Heidelberg, 2013

[17] R. S. Michalski, I. Bratko, M. Kubat, Machine Learning and Data Mining: Methods and Applications, New York: Wiley, 1998

[18] D. Francois, Binary classification performances measure cheat sheet, 2009

[19] I. Idris, A. Selamat, "Improved email spam detection model with negative selection algorithm and particle swarm optimization", Applied Soft Computing, Vol. 22, pp. 11-27, 2014

[20] Y. Zhang, H. Y. Li, M. Niranjan, P. Rockett, "Applying cost-sensitive multiobjective genetic programming to feature extraction for spam e-mail filtering", Lecture Notes in Computer Science, Genetic Programming, Berlin/Heidelberg, Springer, Vol. 4971, pp. 325-336, 2008

[21] T. Fagbola, S. Olabiyisi, A. Adigun, "Hybrid GA-SVM for efficient feature selection in e-mail classification", Comput. Eng. Intell. Syst, Vol. 3, No. 3, pp. 17-28, 2012

[22] A. K. Uysal, S. Gunal, "A novel probabilistic feature selection method for text classification", Knowl. Based Syst., Vol. 36, pp. 226-235, 2012

[23] L. Ozgur, T. Gungor, F. Gurgen, "Spam mail detection using artificial neural network and bayesian filter:, in: Z. Yang, H. Yin, R. Everson (Eds.), Intelligent Data Engineering and Automated Learning- IDEAL 2004, Springer, Berlin/Heidelberg, 2004, pp. 505-510, 2004.

[24] R. Ariaeinejad, A. Sadeghian, "Spam Detection System: A New Approach based on Interval Type-2 Fuzzy Sets", 24th Canadian Conference on Electrical and Computer Engineering (CCECE, 2011), 2011

[25] I. Idris, A. Selamat, N.T. Nguyen, S. Omatu, O. Krejcar, K. Kuca, M. Penhaker, "A Combined Negative Selection algorithm-Particle Swarm Optimization for an Email Spam Detection System", Engineering Applications of Artificial Intelligence, Vol. 39, pp. 33-44, 2015

[26] S. Sharma, A. Arora, "Adaptive Approach for Spam Detection", International Journal of Computer Science Issues, Vol. 10, No. 4, No 1, pp. 23-26, 2013

[27] S. S. Shinde, R. Patil, "Improving Spam Mail Filtering using Classification Algorithms with Discretization Filter", International Journal of Emerging Technologies in Computational and Applied Sciences, Vol. 10, No. 1, pp. 82-87, 2014.

[28] M. Rathi, V. Pareek, "Spam Mail Detection through Data Mining-A Comparative Performance Analysis", International Journal of Modern Education and Computer Science, Vol. 12, pp. 31-39, 2013

[29] S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, "Bayesian Additive Regression Trees-Based Spam Detection for Enhanced Email Privacy", IEEE Third International Conference on Availability, Reliability and Security, pp. 1044-1051, 2008