Available online at www.sciencedirect.com

ScienceDirect

Future Computing and Informatics Journal 3 (2018) 436–442
http://www.journals.elsevier.com/future-computing-and-informatics-journal/

Antlion optimization and boosting classifier for spam email detection

Amany A. Naem [a],*, Neveen I. Ghali [b], Afaf A. Saleh [a]

[a] *Al-Azhar University, Faculty of Science, Cairo, Egypt*
[b] *Future University in Egypt, Faculty of Computers and Information Technology, Cairo, Egypt*

Received 17 July 2018; revised 17 October 2018; accepted 15 November 2018
Available online 23 November 2018

Abstract

Spam emails are not necessary, though they are harmful as they include viruses and spyware, so there is an emerging need for detecting spam emails. Several methods for detecting spam emails were suggested based on the methods of machine learning, which were submitted to reduce non-relevant emails and get results of high precision for spam email classification. In this work, a new predictive method is submitted based on antlion optimization (ALO) and boosting termed as ALO-Boosting for solving spam emails problem. ALO is a computational model imitates the preying technicality of antlions to ants in the life cycle. Where ALO was utilized to modify the actual place of the population in the separate seeking area, thus obtaining the optimum feature subset for the better classification submit based on boosting classifier. Boosting classifier is a classification algorithm that points to a group of algorithms which modifies soft learners into powerful learners. The proposed procedure is compared against support vector machine (SVM), k-nearest neighbours algorithm (KNN), and bootstrap aggregating (Bagging) on spam email datasets in a set of implementation measures. The experimental outcomes show the ability of the proposed method to successfully detect optimum features with the smallest value of selected features and a high precision of measures for spam email classification based on boosting classifier.

Keywords: Boosting classifier; Antlion optimization; Feature selection; Spam email; Classification

1. Introduction

Electronic mail (email) is significant for many kinds of group connection, which has become widely used by many people individuals and organizations. At the same time, email is one of the fast rising and costly problems linked with the internet today, in which this case it is called spam email. Spam emails are predominantly mercantile or have attractive links to famous websites but they lead to sites that are meddlesome [1].

As a result, spam emails cause a lessening in privacy, spreading viruses, occupying space in the email box, and destroying email servers. Therefore, the user wastes a lot of time in filtering email imports and cancelling the unwelcome email. The discovery of undesired emails categorizes the emails as spam or non-spam (ham), so this process is related to the classification problem [2,3].

Classification methods can be utilized to uncover spam but datasets oftentimes contain a great number of trifle features or reiterate features, which may diminish the classification precision. Feature selection is the answer to this trouble where it is utilized for picking out a subset of all the features. The goal is to find the optimal solution for the objective function, which target minimizing the dimensions of features [4,5].

Behjat et al. in Ref. [6] suggested feature selection process by utilizing particle swarm optimization algorithm to look for the optimum features in the seeking area. They utilized this suggested method to create optimum selection results with minimum selected features that result from high resolution of spam classification based on Multi-Layer Perceptron. Harvey and Todd in Ref. [7] showed a new genetic algorithm to

* Corresponding author.
E-mail addresses: manoo_basom@yahoo.com (A.A. Naem), neveen.ghali@fue.edu.eg (N.I. Ghali), afaf211@yahoo.com (A.A. Saleh).

Peer review under responsibility of Faculty of Computers and Information Technology, Future University in Egypt.

https://doi.org/10.1016/j.fcij.2018.11.006
2314-7288/

automate the design of the feature named Autofead. In this suggest, the genetic algorithm improved a number of filtered features from the store for sequence-handling functions. Shams and Mercer [8] utilized some famous methods to induce spam classifiers such as Random Forest, Support Vector Machine, and Naïve Bayes and assessed the classifier showings. Xia et al. in Ref. [9] submitted spammer detection as the answer for active online processing by unitizing data and information of the network. In this study, an enhanced feature selection wrapped by utilizing antlion optimization and boosting for detection of spam emails is suggested and named as ALO-Boosting. Where antlion optimization (ALO) algorithm is utilized to pick out preferential features. Boosting classifier is utilized to divide files to spam emails or ham emails to decrease features dimensionality and increase classification precision.

The remainder of this paper is organized as follows: Brief introduction on antlion optimization, boosting classifier, feature extraction and feature selection are described in Section 2. The details of the proposed method are presented in Section 3. Section 4 shows the experimental design. Section 5 describes our experimental outcomes. Finally, we supply conclusions in Section 6.

2. Preliminaries

2.1. Antlion optimization (ALO)

Antlion Optimization (ALO) [10] is a modern meta-heuristic optimization method suggested by Mirjalili in 2015 and it is taken from the behaviour of antlions or doodlebugs in the file cycle. ALO has distinguished results in areas of local optimum congregation, exploitation, and avoidance [11–13]. Mathematical forming of the ALO algorithm can be formulated by as these steps [14,15]:

- *Random Walks*: The random walking of ants can be formulated when looking for feed in the life cycle as follows:

$$X(t) = [0, \mathrm{cumusum}(2L(s_1)-1), \mathrm{cumusum}(2L(s_2) - 1), \ldots, \mathrm{cumusum}(2L(s_m)-1)] \tag{1}$$

where m displays the topmost number of iteration, *cumusum* counts the cumulative sum, s indicates a step of random walking, and $L(s)$ denotes a stochastic function and is calculated by

$$L(s) = \begin{cases} 1, & \text{if random} > 0.5 \\ 0, & \text{if random} \le 0.5 \end{cases} \tag{2}$$

where *random* represents a random number and it falls in [0, 1].

The place of ants is created with this matrix:

$$P_{\text{Ant}} = \begin{bmatrix} E_{1,1} & \cdots & E_{1,b} \\ \vdots & \ddots & \vdots \\ E_{n,1} & \cdots & E_{n,b} \end{bmatrix} \tag{3}$$

Here P_{Ant} is the matrix for utilizing the place of every ant, $E_{i,j}$ denotes the value of the jth variable of ith ant, n displays to the number of ants, and b presents to the number of variables.

The place of antlions is created with this matrix:

$$P_{EL} = \begin{bmatrix} EL_{1,1} & \cdots & EL_{1,b} \\ \vdots & \ddots & \vdots \\ EL_{n,1} & \cdots & EL_{n,b} \end{bmatrix} \tag{4}$$

Here P_{EL} is the matrix for utilizing the place of every antlion, $EL_{i,j}$ denotes the value of the jth variable of ith antlion, n displays to the number of antlions, and b shows to the number of variable.

The random walking of ants is settled within the seeking area utilizing this equation:

$$X_i^t = K_i + \frac{\left(X_i^t - o_i\right) * \left(g_i - K_i^t\right)}{\left(g_i^t - o_i\right)} \tag{5}$$

where o_i is the smallest value of random walking of ith variable, g_i displays the topmost value of random walking of ith variable, K_i^t displays the smallest value of ith variable at tth iteration, and g_i^t displays the topmost value of ith variable in tth iteration.

- *Blockade in Pit*: Mathematical forming of ants blockade in antlion's pits is computed by these equations:

$$\begin{aligned} K_i^t &= \text{Antlion } t_i^t + k^t \\ g_i^t &= \text{Antlion } t_i^t + g^t \end{aligned} \tag{6}$$

where k^t displays the smallest value of all variables at tth iteration, g^t presents the topmost value of all variables at tth iteration, k_i^t acts the smallest value of all variables for ith ant, g_i^t is the topmost value of all variables for ith ant, and Antlion t_i^t presents the place of the chosen jth antlion at tth iteration.

- *Build blockade*: Obtain the highest probability for taking ants by utilizing roulette wheel. The fittest antlion is identified by this technique.

- *Slipping ants in the direction of antlion*: Antlions come out of the sand outside the middle of the hole, so any ant trying to escape slides down the blockade. The ant walks randomly in a hypersphere with radius, which is reduced according to these equations:

$$\begin{aligned} k^t &= \frac{k^t}{U} \\ g^t &= \frac{g^t}{U} \end{aligned} \tag{7}$$

where U is a ratio and is assigned by this equation:

$$U = 10^y * \frac{i}{j} \tag{8}$$

where i acts the actual iteration, j presents the maximum number of iterations, and y indicates to a constant and is defined based on j and i where ($y = 2$ when $i > 0.1j$, $y = 3$ when $i > 0.5j$, $y = 4$ when $i > 0.75j$, $y = 5$ when $i > 0.9j$, and $y = 6$ when $i > 0.95j$).

- *Catching ant and re-structure hole*: When the ant arrives at the down of the hole and is captured this is the end step of hunting. According to the last place, the antlions change its place by this equation:

$$\text{Antliont}_i^t = \text{Ant}_i^t \quad \text{if } f\left(\text{Ant}_i^t\right) > f\left(\text{Antliont}_i^t\right) \tag{9}$$

where Antlion_i^t acts to the place of the detected jth antlion at tth iteration, Ant_i^t acts to the place of the detected ith ant at tth iteration, and t acts the actual iteration.

- *Elitism*: Elite is of great value in the evolution method where it maintains the best solution. This can be modelled as this equation:

$$\text{Ant}_i^t = \frac{B_A^t + B_E^t}{2} \tag{10}$$

where B_A^t represents the random walking about the antlion detected via the roulette wheel at tth iteration and B_E^t displays the random walking about the elite antlion at tth iteration. Steps for ALO algorithm are appeared in Algorithm 1.

2.2. Boosting classifier

Boosting classifier is meta-algorithm prepared to develop the precision of machine learning methods and utilize in the problem of classification [16,17]. Boosting classifier is a global and efficacious way to produce a careful prediction rule by joining soft classifiers to generate a powerful classifier by weight [18]. The steps of boosting classifier are described in Algorithm 2.

Algorithm 1. Antlion optimization (ALO) algorithm

```
1. Randomly, the first population of ants and antlions is distributed
2. The fitness of ants and antlion is counted
3. Get the optimum antlions and suppose it is the elite
4. while the final criterion is not done
5. for each ant
6. A roulette wheel is used to pick an antlion
7. Change $k$ and $g$ by utilizing equation (8)
8. generate a random walking and apply it by utilizing equations (2) and (6)
9. The place of ant is updated using equation (11)
10. end for
11. The fitness of all ants is counted
12. Replace an antlion with the candidate that is calculated utilizing equation (10)
13. If an antlion is the best, the elite is changed with this antlion
14. end while
15. Return elite
```

Algorithm 2. Boosting classifier algorithm

```
1. In the training phase, building N learners by creating additional data
2. Create N training datasets using random templates with substitution over weighted data
3. Repeated in each new training dataset until the training package is expected fully or a topmost number of templates are grew
```

2.3. Feature extraction

The process of feature extraction can have an active function in enhancement procedure of classification. Extracted features are a group of objects, expirations and turn pictures to the word, which clearly distinguish if the email is ham or not. But before starting on feature extraction process, it is necessary to do the important process named by pre-processing email. Pre-processing email is applied to all emails and is implemented by decreasing high dimensionality such as HTML tags, URL, and email address [3,19].

After pre-processing process, email becomes very easy to be used in feature extraction. The step of feature extraction is chosen which words are spam and which words are non spam. The spam words were selected by choosing all words that were repeated at least 100 times in spam email messages, and they were put in the vocabulary list. Then, make an index of the words that are mentioned in the email and also in the vocabulary list, which is called by the list of word indices.

Now, each email is converted to a vector. Let jth is word in the vocabulary list. So, the feature y_j is equal 1, if the jth word exists in the email. And, the feature y_j is equal 0, if the jth word is not existing in the email. This process is called by vector space model by binary weights [19,20].

2.4. Feature selection

Feature selection is utilized to create a new structure from a group of essential features. The aim of feature selection process is to decrease the dimensional of the search area and choose features of high weight. Wrapper approach and filter approach are types of feature selection. Firstly, filter approach is not dependent on any machine learning technique and is frequently low costly and more global than the wrapper approach. Secondly, wrapper approach estimates the subset of features by utilizing a machine learning technique and always obtains better outcomes than filter approach for some troubles. Feature selection wrapper is approved to gain the better classification showing in spam email problem [3,21].

3. Proposed method

This study suggests a novel computational method, ALO-Boosting for spam email detection. ALO-Boosting consists of two main stages. The first stage, ALO with feature selection is utilized to look for the optimum features. The second stage, efficacious and influential boosting classifier is conducted based on the optimum feature subset gain in the

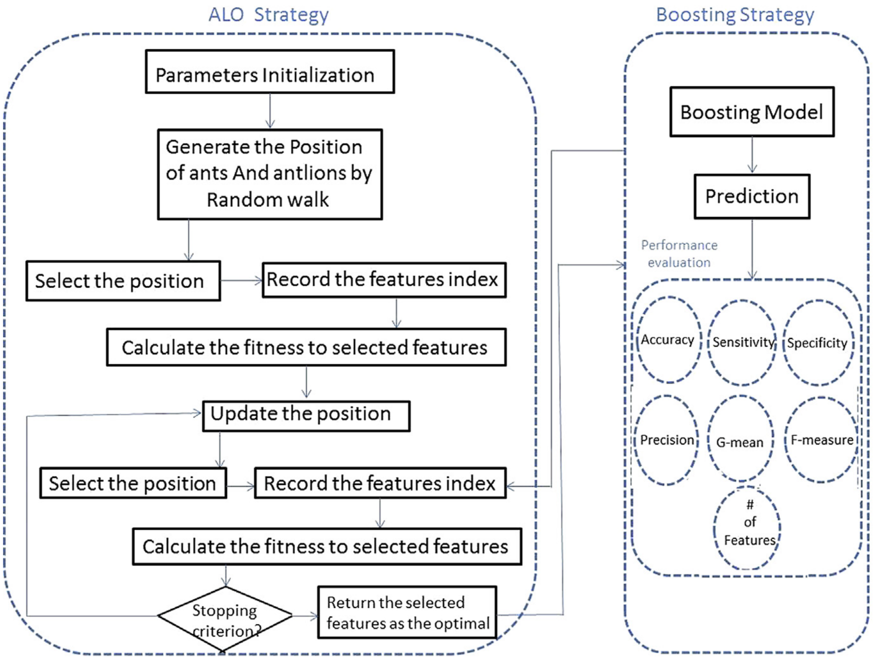
ALO Strategy
Parameters Initialization
Generate the Position of ants And antlions by Random walk
Select the position
Record the features index
Calculate the fitness to selected features
Update the position
Select the position
Record the features index
Calculate the fitness to selected features
Stopping criterion?
Return the selected features as the optimal
Boosting Strategy
Boosting Model
Prediction
Performance evaluation
Accuracy
Sensitivity
Specificity
Precision
G-mean
F-measure
# of Features

Fig. 1. ALO-boosting method.

past stage. A detailed of ALO-Boosting is presented in Fig. 1. The ALO is fundamentally utilized to be fit search in the feature area for the optimum features. The optimum features are the smallest value of features and achieve the topmost of precision of classification. The objective function used in ALO to estimate the selected features is displayed as this equation:

$$\text{Objective Function} = \gamma D + \delta\frac{N-L}{N} \tag{11}$$

where γ and δ are two parameters matching to the classification precision weight and the quality of the number of selected features, γ falls in interval [0,1] and $\delta = 1 - \gamma$, N denotes to the number of all features, D indicates to the classification precision, and L indicates to the selected feature length.

Vector model for feature selection is explained in Fig. 2. The vector consists of a series of binary values of 0 and 1 indicates all features in the dataset. If the value is 1, the feature is selected. If the value is 0, the feature is not selected.

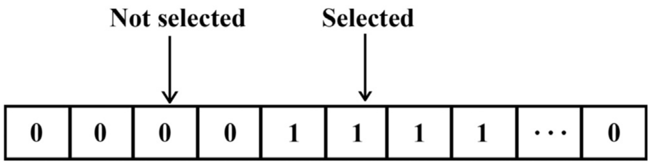

Fig. 2. Vector model for feature selection.

4. Experimental design

4.1. Data description

CSDMC2010 dataset is one of the datasets for the data mining opposition associated with ICONIP 2010. The total

Table 1
Confusion table.

		Factual Class	
		Spam	Ham
Guessed class	Spam	(RP)	(FP)
	Ham	(FN)	(RN)

Table 2
Definitions of measures.

Measure	Definition
A	$\frac{\mathrm{RP}+\mathrm{RN}}{\mathrm{FP}+\mathrm{FN}+\mathrm{RP}+\mathrm{RN}}$
SN	$\frac{\mathrm{RP}}{\mathrm{FN}+\mathrm{RP}}$
SP	$\frac{\mathrm{RN}}{\mathrm{FN}+\mathrm{RN}}$
P	$\frac{\mathrm{RP}}{\mathrm{FP}+\mathrm{RP}}$
GM	$\sqrt{\mathrm{SN}*\mathrm{SP}}$
FM	$\frac{(Q^2+1)*P*\mathrm{SN}}{Q^2*p+\mathrm{SN}}$
Features No.	L

Table 3
Experimental results of four classifications with optimization system ALO and without optimization system ALO on CSDMC2010 dataset.

Measure	Method							
	KNN	ALO-KNN	SVM	ALO-SVM	Bagging	ALO-Bagging	Boosting	ALO-Boosting
A	0.9534	0.9700	0.9316	0.9606	0.9515	0.9931	0.9800	**0.9980**
SN	0.9022	0.9190	0.9187	0.9255	0.9730	0.9840	0.9799	**0.9900**
SP	0.9867	0.9934	0.9691	0.9816	0.9404	0.9968	0.9735	**0.9998**
P	0.9036	0.9836	0.9300	0.9675	0.8944	0.9919	0.9697	**1.0000**
GM	0.9435	0.9555	0.9436	0.9531	0.9566	0.9904	0.9767	**0.9949**
FM	0.9029	0.9502	0.9243	0.9460	0.9320	0.9879	0.9748	**0.9950**
Features No	507	245	507	242	507	236	507	**211**

Table 4
Experimental results of four classifications with optimization system ALO and without optimization system ALO on SpamAssassin dataset.

Measure	Method							
	KNN	ALO-KNN	SVM	ALO-SVM	Bagging	ALO-Bagging	Boosting	ALO-Boosting
A	0.9334	0.9480	0.9316	0.9537	0.9532	0.9745	0.9735	**0.9891**
SN	0.9432	0.9630	0.9521	0.9790	0.9627	0.9808	0.9800	**0.9996**
SP	0.9155	0.9249	0.9395	0.9462	0.9487	0.9667	0.9699	**0.9783**
P	0.9603	0.9795	0.9711	0.9800	0.9875	0.9929	0.9891	**0.9988**
GM	0.9292	0.9438	0.9458	0.9625	0.9557	0.9737	0.9753	**0.9889**
FM	0.9517	0.9712	0.9615	0.9795	0.9749	0.9868	0.9845	**0.9991**
Feature No	507	255	507	249	507	241	507	**221**

number of emails is 4327, of which 2949 are solicited emails and 1378 are unsolicited emails [22].

SpamAssassin dataset is one of the most popular public datasets. The total number of emails is 6047, of which 4150 are solicited emails and 1897 are unsolicited emails [23].

4.2. Performance evaluation

A confusion table includes basics of a factual class and guessed class done by the classification system. In our method, all measures are calculated based on the confusion table, presented in Table 1.

Where RP is the number of right guessing that have a positive instance, FP is the number of false guessing that have a positive instance, FN is the number of false guessing that have a negative instance, and RN is the number of right guessing that have a negative instance.

In this study, the prediction methods are evaluate by utilizing the various rating measures for the classification technique, listed in Table 2. There are accuracy (*A*), sensitivity (SN), specificity (SP), precision (*P*), G-mean (GM), F-measure (FM) where *Q* takes a value from 0 to infinity and in is put equal 1, and number of selected features (Features No.) where *L* indicates to the length of selected feature subset.

5. Experimental outcomes

Comparative experiments were performed between ALO-Boosting and the other competitive methods (including KNN, SVM, Bagging, Boosting, ALO-KNN, ALO-SVM, and ALO-Bagging) and in order to judge the performance of the submitted method for CS-DMC2010 dataset and SpamAssassin dataset respectively.

Table 3 illustrates the detailed classification outcomes of the eight methods in terms of Accuracy, Sensitivity, Specificity, Precision, G-mean, F-measure, and Number of selected features on the CSDMC2010 dataset. From the analysis of Table 3, the ALO-Boosting achieves the best values of 99.80% accuracy, 99.00% sensitivity, 99.98% specificity, 100% precision, 99.49% G-mean, and 99.50% F-measure and with the least number of selected features.

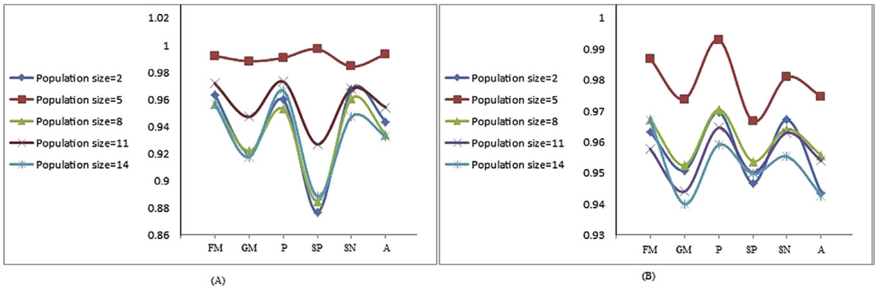

Fig. 3. Experimental results of ALO-Boosting measures with different population size to CSDMC2010 dataset and SpamAssassin dataset respectively.

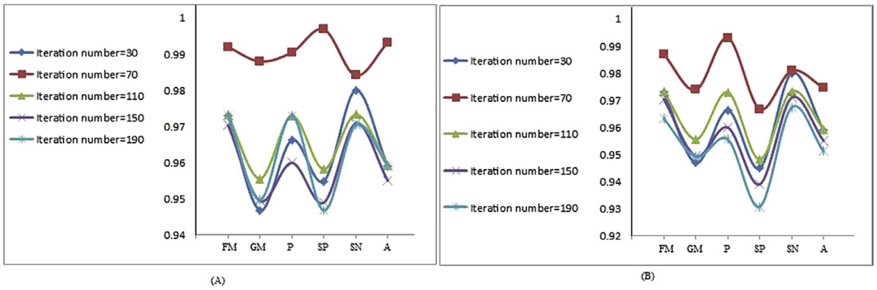

Fig. 4. Experimental results of ALO-Boosting measures with different iteration number to CSDMC2010 dataset and SpamAssassin dataset respectively.

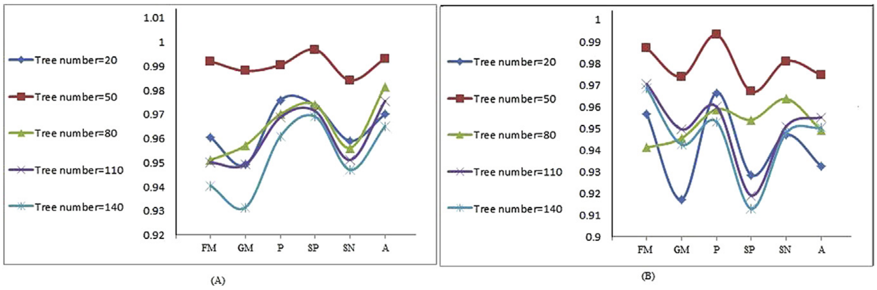

Fig. 5. Experimental results of ALO-Boosting measures with different tree number to CSDMC2010 dataset and SpamAssassin dataset respectively.

Table 4 illustrates the detailed classification outcomes of the eight methods in terms of Accuracy, Sensitivity, Specificity, Precision, G-mean, F-measure, and Number of selected features on the SpamAssassin dataset. From the analysis of Table 4, the ALO-Boosting achieves the best values of 98.91% accuracy, 99.96% sensitivity, 97.83% Specificity, 99.88% precision, 98.89% G-mean, and 99.91% F-measure and with the least number of selected features.

The population size, the iteration number, and tree number are three key factors in proposed method; thus their suitable values were investigated on our datasets. Firstly, to find the best value of the population size, different sizes of population from 2 to 14 were taken when the number of iterations and the number of trees were fixed to 70 and 50 respectively. Secondly, to find the best value of the iteration number, the size of population and the number of trees were fixed to 5 and 50 respectively and different numbers of iterations from 30 to 190 were taken. Thirdly, to find the best value of the tree number, different numbers of tree from 20 to 140 were taken when the number of iterations and the population size were fixed to 70 and 5 respectively. As shown in Figs. 3−5 ALO-Boosting achieves the best performance when the population size, the iteration number, and the tree number are equal to 5, 70, and 50 respectively.

6. Conclusion

In this work, an ALO-Boosting method is explained in detail. This proposed method consists of two main stages: feature selection and classification. First, the ALO was suggested for searching for the optimum feature in seeking area. Second, the boosting classifier was utilized to predict based on the features obtained in the first stage. The proposed method is matched with famous classification methods including KNN, SVM, and Bagging. These methods were applied on CSDMC2010 dataset and SpamAssassin dataset to detect the spam email. The experimental outcomes present that the proposed method gets a low number of selected features and achieves a high degree of classification precision.

References

[1] Renuka K, Hamsapriya T. Email classification for spam detection using word stemming. Int J Comput Appl 2010;5(1):45−7.

[2] Bayati M, Jabbar S. Developing a spam email detector. Int J Eng Innov Technol 2015;5(2):16−21.

[3] ZhiWei M, Singh M, Zaaba Z. Email spam detection: a method of metaclassifiers stacking. In: The 6th international conference on computing and informatics, vol. 16; 2017. p. 750−7.

[4] Awad W, ELseuofi S. Machine learning methods for spam e-mail classification. Int J Comput Sci Inf Technol 2011;3(1):173−84.

[5] Prilepok M, Jezowicz T, Platos J, Snasel V. Spam detection using compression and PSO. In: Computational aspects of social networks (CASoN), 2012 fourth international conference on IEEE; 2012. p. 263−70.

[6] Behjat R, Mustapha A, Nezamabadi-pour H, Sulaiman N, Mustapha N. A PSO-based feature subset selection for application of spam/non-spam detection. In: Soft computing applications and intelligent systems. Springer; 2013. p. 183−93.

[7] Harvey D, Todd M. Automated feature design for numeric sequence classification by genetic programming". IEEE Trans Evol Comput 2015; 19:474−89.

[8] Shams R, Mercer R. Classifying spam emails using text and readability features. In: Data mining (ICDM), 2013 IEEE 13th international conference; 2013. p. 657−66.

[9] Xia H, Jiliang T, Huan L. Online social spammer detection. Association for the Advancement of Artificial Intelligence; 2014. p. 59−65.

[10] Mirjalili S. The ant lion optimizer. Adv Eng Softw 2015;83:80−98.

[11] Patil S, Patel D. The ant lion optimization algorithm for flexible process planning. J Prod Eng 2015;18(2):65−8.

[12] Christaline J, Ramesh R, Vaishali D. Bio-inspired computational algorithms for improved image steganalysis. Indian J Sci Technol 2016;9(10):1−10.

[13] Esha G, Akash S. Performance evaluation of antlion optimizer based regulator in automatic generation control of interconnected power system. J Eng 2016:1−14.

[14] Shivani M, Meenakshi M. Ant lion optimization for optimum power generation with valve point effects, computer methods in applied mechanics and engineering. Int J Res Appl Sci Eng Technol 2015;3:1−6.

[15] Petrović M, Petronijević J, Mitić M, Vuković N, Plemić A, Miljković Z, et al. The ant lion optimization algorithm for flexible process planning. J Prod Eng 2015;18(2):65−8.

[16] Vasconcelos N, Saberian J. Boosting classifier cascades. Adv Neural Inf Process Syst 2010:2047−55.

[17] Kim T, Cipolla R. Multiple classifier boosting and tree-structured classifiers. Springer-Verlag Berlin Heidelberg; 2012. p. 163−96.

[18] Machov K, Puszta M, Barčák F, Bednr P. A comparison of the bagging and the boosting methods using the decision trees classifiers. Comput Sci Inf Syst 2006;3(2):57−72.

[19] Vinod P, Divakar S, Anju S. A novel technique of email classification for spam detection. Int J Appl Inform Syst 2013;5(10):15−9.

[20] Bagus P. Answer search Indonesian language Hadith using vector space model in pdf document. Int Res J Comput Sci 2017;8(4):1−5.

[21] Qiang L, Huiling C, Hui H, Xuehua Z, ZhenNao C, Changfei T, et al. An enhanced grey wolf optimization based feature selection wrapped kernel extreme learning machine for medical diagnosis. Comput Math Methods Med 2017:1−15.

[22] http://csmining.org/index.php/spam-email-datasets-.html; 2017. Accessed September 2017.

[23] http://csmining.org/index.php/spam-assassin-datasets.html; 2017. Accessed September 2017.