JOURNAL OF PURE AND APPLIED MICROBIOLOGY, May 2015. Vol. **9**(Spl. Edn. 1), p. 673-680

Medical E-mail Spam Classification using a Score Based System and Immune System Embedded with Feature Selection Process

Khongbantabam Susila Devi[1] and R. Ravi[2]

[1]Department of Information and Communication Engineering, Anna University, Chennai, India
[2]Department of Computer Science and Engineering,
Francis Xavier Engineering College, Tirunelveli, India.

(Received: 08 January 2015; accepted: 24 March 2015)

E-mail becomes a most important communication source in day today's life. Medical industry has emerged with the facility of transmitting the patients' prescription through the Email service. Spam is a very serious problem in medical email communication that has been gradually afflicting the user of the medical web portal. Programs named as spam filters are applied to assist the medical web portal users in deciding if an email is valuable for reading or not. For instance a hospital web portal has been taken for the email communication process. The incoming mails to the hospital web portal may include the spam mails which will be efficiently categorized into the spam and non-spam mails. The proposed spam filtering system is used for a variety of medical web portal users. Thus, in this paper, the Artificial Immune System with Local Feature Selection (AISLFS) is proposed for the classification of spam and ham mails and it has the special characteristic of inbuilt feature selection process. The Score based system is used to generate the rules, the incoming mails are about to answer the rules and a separate score is maintained for each feature of the email. Then the AISLFS classifier classifies the email using the significant features and the feature scores by comparing the training dataset which is trained using the Resilient Back Propagation Neural Network Algorithm.

Key words: Email service, medical web portal, Apriori algorithm, Artificial Immune System with Local Feature Selection (AISLFS), Score based system, Resilient Back Propagation Neural Network Algorithm.

The modern medical practice has emerged with the patient-doctor email communication. The patients can able to communicate with the medical experts regarding their prescription, lab test results and schedules and so on. These emails are sent to the hospital web portal email services and emails will be sent to the corresponding medical experts based on the email content. A separate admin is allotted for doing this task. The email service brings the patients, another possible place of access[1]. A quality protocol is maintained in each hospital/ medical web portal server for blocking the unwanted bulk emails. The significant process is to make sure that the mail ids of all the registered patients are recorded and the mail id of a patient and the corresponding doctor is included as a part of the patients' medical record.

The main focus of the medical email maintenance service is to make the email service easy and simple to use, allows the user to send the vast category of mails throughout the internet in a reliable, efficient and convenient manner. At the same time, the vast amount of unwanted bulk emails are spread in the internet and reached various medical web portal email server. Both the wide category of mailcauses danger to the email system

* To whom all correspondence should be addressed.
E-mail: susilaphd14@gmail.com

of the hospital. It may be spam mails or virus[2]. The incoming emails can able to render only limited information and the admin of the medical web portal email have to manage and prioritize the vast amount of emails. These emails appear in the various types, formats and the same attachments with edited content. The composition of internet connection with high bandwidth, increase in email users, low cost, reducing the cost per megabyte of storage gives rise to discharge of emails content[3]. Considering all these data, the classification of legitimate email from the spam mails is a dreadful activity.

A modernisticResilient Neural Network based Artificial Immune System with Local Feature Selection (AISLFS) approach is proposed for the medical web portal email classification.Before the classification process begins, the Resilient Back Propagation Neural Network Algorithm is used to train the pre-classified dataset from which the AIS classifier learns the rules for classification. The incoming mails are initially processed using the ideal Score Based System (SBS) where in a set of rules are generated and the email is inspected based on its characteristics. According to the inspected results, a separate score is maintained for each characteristic and finally combined to get the score of an email. This identifies and eliminates the spam mail.The mails which crossed this stage are preprocessed and given to the optimal feature selection section of the AISLFS classifier where the features from the preprocessing stage are examined and the relevant, significant features for the classification are mined out. Finally the classification section of AISLFS classifier intakes the significant features and classifies the incoming into either spam or ham mail based on the rules from the resilient network.

Related work

The negative selection algorithm[4] is proposed for the purpose of spam detection and is based on the immune system model which reduces the overheads. In addition the matching rule model and the detector model are implemented and they are well versed in process of matching the non-self and self-text. This process of matching beats the drawbacks of the ordinary negative selection algorithm. This proposed method improvises the false positive and true positive rates in better and efficient manner.

The spam message in the mobile through the Short Message Service (SMS) becomes a danger issue in recent days. Many models are proposed to prevent the mobile system from the spam message. A new methodology[5] based on the Artificial Immune System (AIS) is proposed that classifies the message into spam and non-spam by utilizing the training set. This training dataset of message has incorporated a big volume of SMS. This proposed methodology makes use of usual preprocessing steps. At last the proposed methodology is compared with the Bayesian filter for spam mail classification and proves its efficiency.

The AIS and the statistical learning methods are mostly used to prevent the legitimate message from the spam messages. Both the methods are came from different origin and attempts to resolve the problem and find the solution. In advance many hybrid methods came into play for the batter solution for the spam filtering. All the methodologies are reviewed here[6] and the immune based generic framework is proposed and is developed for the online process that constructs an intelligent email classification process.

In addition to the spam mail occurrence, the solutions to prevent form this issue also bring a very low false positive rate in the classification process. Hence a fresh method is proposed[7] with the weight aggregation function where the tree based classifier identifies the decision threshold of the messages and brings the optimized performance. All these improvements render the best decision and detection rate and compared with many existing methods where the proposed system proves better classification and the accuracy.

Every individual has some opinion about the day to day pursuing things. Hence mining these opinions may give us better idea to improve the products quality. The process traditionally intakes NaturalLanguage Processing approach and on the other hand the machine learning techniques are used to mine these opinions. But both the methods have the drawback that they take the features appropriately not in a confident manner. Therefore, artificial immune system[8] is proposed to mine the opinions and the proper feature selection is also performed.

Modernistic medical spam email classification

The day-to-day usage of Electronic communication between the medical experts and the patients are increasing progressively and it has many conceivable advantages. Even though this medical email communication has many potential advantages, the occurrence of spam mails cause trouble to the doctors who are using the medical web portal email system and also there may have the chance of doctors ignoring some patients' email mistakenly. Hence the accurate and faultless system for the classification of spam and ham mail is necessary for every medical/hospital web portal. Therefore, an intelligent medical spam mail classification system is suggested, which works based on a Score based system, resilient based training dataset and anAISLFS classifier. Training dataset (pre-classified mails) is trained using the resilient neural network algorithm and the weights in the neurons are regulated based on the spam and ham probabilities. In the testing process, the incoming mails separately fed into the Score Based System where the one whole mail is examined based on its own characteristics with the generated rules from the SBS system. This is also a filtering stage where the mails which is having least score will be marked as spam and removed. Then the filtering mails are processed using the traditional preprocessing steps where the unwanted information in the mail is removed and the process is also termed as feature extraction. The extracted features may be either significant or irrelevant for the classification process. The feature selection section of the AISLFS classifier is used to mine the significant features. Then these features are given to the classifier part of the AISLFS classifier which classifies the mails into spam and ham mails based on the rules learned from the resilient training network. The workflow of the proposed medical email categorization is depicted in the figure 1.

Resilient based training network

Basically, the resilient propagation is one of the well-organized, learning heuristic and cost effective schemes that do the weight adaptation by considering the local gradient adaptation and while comparing the resilient network with other normal back propagation algorithms. The resilient propagation provides more robustness, accuracy and convergence speed. In the proposed system, the pre-classified email dataset is fed into the resilient back propagation algorithm where in the learning process for the classification is performed. The machine learning algorithm for the final classification learns the classification rules from this learning process. The training process begins with taking the pre-classified email dataset into its input layer of resilient back propagation network and regulates the weights of the neuron in all the layers until the difference between the desired output and the neural output is less than the predefined threshold value. This RPROP [9] requires comparatively small memory and also has another advantage that the training process discovers the rules for the final classification when compared with other algorithms. The basic learning rule is represented as

$$L^{(m+1)} = L^m + \delta(G^m - W^m)p^m \quad ...(1)$$

p^m = training pattern, W^m = desired output, G^m = training output,δ = positive learning factor, L^m =previous weight, $L^{(m+1)}$= new weight.

In order to cover the whole mail, the convergence speed of the training network should be increased. Thus the ideal weight L^i is used in the network where it contemplates the whole mail content instead of a particular portion of the email. Hence the basic learning rule is modified by taking the ideal weight and is formulated as

$$L^{(m+1)} - L^i = L^m - L^i + \delta(G^m - W^m)p^m \quad ...(2)$$

The rules for classification can be identified form the Eq. (2) only if the training pattern is correct. But the training pattern is not always correct. So let us assume that the RPROP has the misclassified training pattern p^m and the learning rule will be modified as

$$\|L^{(m+1)} - L^i\|^2 = \|L^{(m)} - L^i\|^2 + \delta^2\|p^m\|^2 + 2\delta(G^m - W^m)(L^m - L^i)p^m \quad ...(3)$$

Here, both the neural output and the desired output will not be the same, because of the misclassified pattern. So the learning continues to learn the pattern and the next iteration of learning is performed through the reinforcement learning as

$$(G^m - W^m)(L^m)^T p^m = -|(L^m)^T p^m| \leq 0 \quad ...(4)$$

$$(G^m - W^m)(L^i)^T p^m = |(L^i)^T p^m| \geq 0 \quad ...(5)$$

Adding the above reinforcement (5) and the anti-reinforcement learning (4) into the main

learning process equation which is changed as

$$\|L^{(m+1)} - L\|^2 = \|L^{(m)} - L\|^2 + \delta^2\|p^m\|^2 - 2\delta\left(|(L)^T p^m| + |(L^m)^T p^m|\right) \quad ...(6)$$

The final optimal rule will be found when the positive learning factor minimizes the Mean Square Error (MSE). But it has the problem that the positive learning cannot be found in prior. Hence the variable ∂ replaces the unknown terms in the Eq. (6) as

$$L^{(m+1)} = L^m + \frac{(G^m - W^m)(\vartheta + |L^{mT} p^m|)}{\|p^m\|^2} p^m \quad ...(7)$$

The hidden layer with unlimited neurons bring the over fitting problem. So the previous experimental results prove that the 32 neurons in a hidden layer bring better classification result in terms of accuracy and error tolerance.

Score Based System (SBS)

The Score Based System (SBS) works based on the rules and the score obtained from the rules execution. The rules are generated by inspecting the body content of the mail, keywords, and mail header details and so on. On executing each of the generated rules on the incoming mail, separate score is maintained for each rule and finally all the score of a single mail is summed up to find the spam mail initially. This is a preliminary process of identifying and eliminating the spam mail[10]. This process makes the upcoming steps for classification easier and efficient. The rules are generated by taking the features of the mail in order to identify the spam mail. Thegenerated rules are listed in the table.1.

Table 1. Generated rules For the Spam Mail Identification

S .No	Rules	Scores
1	Is the mail come from meaningful user name?	The score is “1”, if yes, otherwise “0”
2	Is the mail come from valuable domain name?	The score is “1”, if yes, otherwise “0”
3	Is the header information of the mail is original?	The score is “1”, if yes, otherwise “0”
4	Is the “from address” in the user block list?	The score is “0”, if yes, otherwise “1”
5	Is the “from and to address” be the same?	The score is “0”, if yes, otherwise “1”
6	Is the “from address” in the auto white list?	The score is “0”, if yes, otherwise “1”
7	Is the subject present in the mail?	The score is “1”, if yes, otherwise “0”
8	Is the mail a forwarded one?	The score is “0”, if yes, otherwise “1”
9	Can we reply to that particular mail?	The score is “1”, if yes, otherwise “0”
10	Are the words in the mails contains any foreign language other than English?	The score is “0”, if yes, otherwise “1”
11	Is the body of the mail present?	The score is “1”, if yes, otherwise “0”
12	Is any sensual content in the mail?	The score is “0”, if yes, otherwise “1”
13	Is this a broad cast message? (containing many to addresses)	The score is “0”, if yes, otherwise “1”
14	Is any repeated double quotes appear in the body?	The score is “0”, if yes, otherwise “1”
15	Is many blank line occurs in the body?	The score is “0”, if yes, otherwise “1”

Table 2. Analysis of proposed methodologies

Techniques	TN	FN	TP	FP	Specificity	Sensitivity ± SD	Accuracy ± SD
Artificial Immune system Local Feature Selection (AISLFS)	17	3±1	34±1	0	100	91.87±2.69	94.44±1.85
AISLFS + Score Based System	17	2±1	35±1	0	100	94.6±2.7	96.29±1.85

Here we take only 15 rules to identify the spam mail. When implementing this process in large medical web portal email system, the number of rules will be large. The identified spam mails are marked as spam and the details of the mail are updated in the spam mail database.

Traditional Preprocessing

The incoming mails for the classification need to be in a normalized form which makes the classification process easier. In order to make the incoming mails to be normalized, initially the preprocessing steps are to be carried out. Here,

the incoming mails are tokenized first which is similar to the morphological process and naturally viewed as independent process[11]. This process is simply taking out the tokens (words) from the running sentences. Next the stop word removal is performed where the irrelevant portion in the token are removed.In additional it removes the useless functional words, articles, prepositions, and formal grammatical words. Then the stemming process processes the words and finds all its root word. Using this stemming process, the redundancy reduces considerably[12]. The preprocessing stage removes the unwanted features from the mail and also the features (words) are extracted for the classification process.

Artificial Immune System with Local Feature Selection (AISLFS) Classifier

The classification process has the major process of classifier construction by using the training dataset through the features used to identify the mails as spam and ham mails. The paper utilizes the Artificial Immune System with Local Feature Selection (AISLFS) algorithm[13] for the purpose of classification. This algorithm has the automatic features selection process embedded in it; hence the separate feature selection is not added here. This embedded feature selection is termed as Local feature selection process which reduces the features by selecting the feature subset that explores the corresponding mail in a best manner.

Feature Selection

This feature selection method is originated by the connection of antibody to the antigen that appears between amino acid residues and a paratope. Consider the antigen denoted as AG consists of one whole mail features extracted from the previous process. Therefore the AG represents a single mail consisting of features. This AG belongs to either spam class or ham (non-spam) class. The initial population called antibody termed as AB is generated by facsimile a copy of all the AGs (pre-classified email dataset) from the training part. In addition, the class labels also copied along with the AGs. The incoming mail contained in a AG is represented as X= $[x_1,x_2,....,x_n]$ and the AB has the trained dataset which will be represented as Y=$[y_1,y_2,...,y_n]$.

Considering each AB containing the trained email and it corresponding feature set *k*, the optimal feature set (OF_k) is determined by using the tournament searching mechanism. Each AB creates "*s*" clones which is a parameter that regulates the local global search space of the algorithm and depends on the *n* solution space. The optimal feature set going to determine in this process should ensures that the maximization of function Eq. (11). Many clones are created based on the ABs and the corresponding features (φ)are selected. Let the n-element binary vector of the k^{th} clone is represented as

$$b_{k,i} = \begin{cases} 1, if\ i \in \varphi_k \\ 0, otherwise \end{cases} \quad i - 1,2, \dots, n \qquad \text{...(8)}$$

Where, φ_k is the selected optimal feature

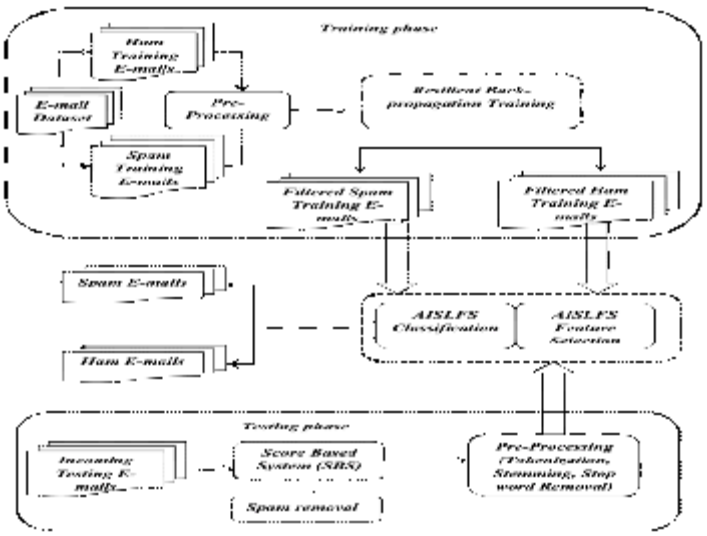

Fig. 1. Proposed Architecture

set that forms the k[th] mail.

The distance measure between the k^{th} trained dataset mail and the j^{th} testing dataset mail is calculated as

$$d(y_k, x_j, \varphi_k) = \left(\sum_{i \in \varphi_k} |y_{k,i} - x_{j,i}|^p \right)^{\frac{1}{p}} \quad ...(9)$$

The affinity measure for the mails in the trained dataset and the testing dataset is inversely proportional to the distance between them and is formulated as

$$A(y_k, x_j, \varphi_k) = \begin{cases} 0, & if\, d(y_k, x_j, \varphi_k) > r_k(\varphi_k) \quad or \quad r_k(\varphi_k) = 0 \\ 1 - \dfrac{d(y_k, x_j, \varphi_k)}{r_k(\varphi_k)}, & otherwise \end{cases} \quad ...(10)$$

Where, $A(y_k, x_j, \varphi_k) \in [0,1]$

r_k is the cross relative threshold of the AB and is regulated once the new AB is generated in such a way that an AB (training) represents any one of the classes without containing any features of the AG (testing).

Once the threshold is calculated, the number of features of AG contained in the AB is taken out. This number of features are depends on the φ_k. The fitness function of AG is evaluated as

$$F(y_k) = \left| \{ x_j : d(y_k, x_j, \varphi_k) \le r_k(\varphi_k) \} \right| \to max \quad ...(11)$$

The feature set which is having higher value is selected and replaced with the AG. If any two selected AGs are having the same highest fitness value, then any one feature set is selected regarding its size.

Classification

In the classification process, mails which are having highest fitness value features are taken from the previous features selection process is given to the immune memory. The new incoming mail AG with highest fitness value features is compared with the mail and its features in the resilient training set. Then the AB in the training set starts recognizing the testing AG. Let the stimulated AB's be F which consists of the features of the particular mail AB. Therefore the similarity between the testing AG mail and the training AB mail is measured based on the optimal features. This similarity measure is called as avidity measure and is formulated as

$$A_s'(x^*) = |F_s(x^*)| \quad ...(12)$$

Result analysis

The paper work is about to support the medical web portal email system for the purpose of

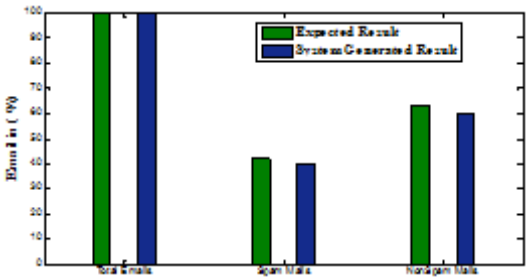

Fig. 2. Comparison of proposed and expected result

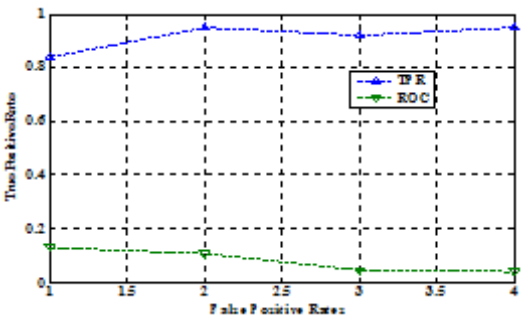

Fig. 3. Classification accuracy range

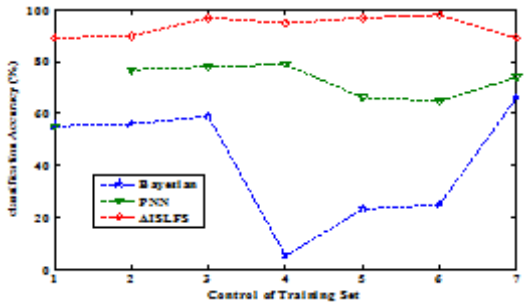

Fig. 4. Classification methods comparison

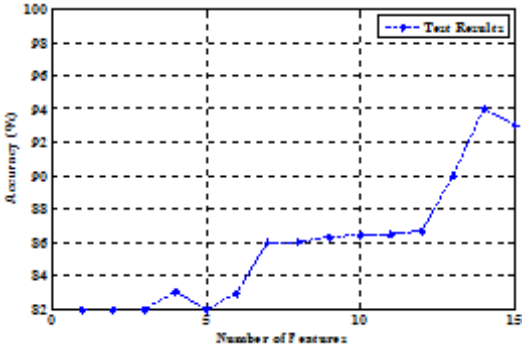

Fig. 5. Accuracy of Feature Selection

classifying the spam emails from the legitimate emails that helps the email servers of the hospital. Therefore the proposed medical email classification system is evaluated by comparing with the results of the proposed system and the expected results. It is depicted in the figure 2.

The figure 2 explains that the expected results and the results obtained from the proposed system bring more or less same result in the sense that the proposed system assures better accuracy in classification.

The figure 3 proves that the true positive rate of the classification process is comparatively higher than the ROC value that means that proposed system works better in finding the true spam and non-spam mails.

The classification method AISLFS used in the proposed system is compared with the previously used methods for classification namely Bayesian filter and the Probabilistic neural network (PNN) in terms of classification accuracy and the effect of training set. The proposed AISLFS proves better classification accuracy.

The features selected among set of different features shows better accuracy and the accuracy starts from the 82 % which shows the features selection process embedded in the classification works good than the individual process.

Theperformance evaluation of the methods employed in the proposed system is done in this section. The examination of these methods is exposed in terms of accuracy, specificity and the sensitivity. These three assessment terms are specified in the following forms.

$$\text{Accuracy} = ((TP+TN)) / ((TP+TN+FP+FN)) * 100\% \quad ...(12)$$

$$\text{Specificity} = TN / (TN+FN) * 100\% \quad ...(13)$$

$$\text{Sensitivity} = TP / (TP+FN) * 100\% \quad ...(14)$$

Where,

TP (True Positives) = correctly classified positive cases,

FP (False Positives) = incorrectly classified negative cases,

TN (True Negative) = correctly classified negative cases,

FN (False Negative) = incorrectly classified positive cases.

Accuracy is chance of the methods performed correctly. Specificity also called as true negative fraction is that the results of the tests are negative (i.e.) the person is not having the particular disease. Sensitivity is also called as true positive fraction is that the results of the tests are positive (i.e.) the person is having the chance of that particular disease.

The methodologies used in the paper are to classify the emails reaching the medical web portal email system into spam mails and the legitimate emails. The table 2 shows the proposed system classification range by considering the parameters taken for determining the sensitivity, accuracy and the specificity.

The table explains that theArtificial Immune system Local Feature Selection (AISLFS) with the Score Based System brings the high accuracy classification and the minimized computational cost. The accuracy rate is ranging from 96% to 98% which is not possible in the existing hybrid methodologies. This makes the additional advantage to the proposed system.

CONCLUSION

In this paper, the classification system model for the medical web portal email system is proposed for the purpose of classification incoming emails into spam mails and the legitimate mails. This system helps the hospital or the medical web portal email system for classifying the spam and legitimate mails. The proposed model initially utilized a Score Based System that generates the rules to validate the emails where the email characteristics are verified using the generated rules. Then identified spam mails are removed from the process and the other mails are assumed as legitimate mails which are preprocessed and given to the Artificial Immune System Local Feature Selection (AISLFS) where the embedded process of optimal feature set selection and the classification process is performed using the resilient neural network training set. This classification model is compared and evaluated in terms of many process and methods. The stimulated results proves that the proposed combination of methods classifies the emails arrives in the medical web portal emails system in efficient and accurate manner.

REFERENCES

1. Leong SL, Gingrich D, Lewis PR, Mauger DT, George JH., "Enhancing doctor-patient communication using email: a pilot study", *Journal of the American Board of Family Practice*, 2005; **18**(3).
2. Ismaila Idris, "E-mail Spam Classification With Artificial Neural Network and Negative Selection Algorithm", *International Journal of Computer Science & Communication Networks*, 2014; **1**(3): 227-231.
3. Foruzan Kiamarzpour1 , Rouhollah Dianat2 , Mohammad bahrani3 , Mehdi Sadeghzadeh4, "Improving the methods of email classification based on words ontology", *International Journal of Computer Science*, 2013; **10**(4): 262-266.
4. Ismaila Idris, "Model and Algorithm in Artificial Immune Systemfor Spam Detection", *International Journal of Artificial Intelligence & Applications* (IJAIA), 2012; **3**(1): 83-94.
5. Tarek M Mahmoud and Ahmed M Mahfouz, "SMS Spam Filtering Technique Based on Artificial ImmuneSystem", *International Journal of Computer Science Issues*, 2012; **9**(2), ISSN (Online): 1694-0814, pp. 589-597, March 2012.
6. Ying Tan, Guyue Mi, Yuanchun Zhu, and Chao Deng, "Artificial Immune System Based Methods for SpamFiltering", *IEEE International Symposium on Circuits and System*, 2013; 2484-2488.
7. Yang Song, Aleksander KoBcz and C. Lee Giles, "Better Naive Bayesclassification for high-precisionspam detection" *Software—Practice and Experience*, 2009; **39**: 1003-1024.
8. Norlela Samsudin, Mazidah Puteh, Abdul Razak Hamdan, Mohd Zakree Ahmad Nazri, "Immune Based Feature Selection for OpinionMining", Proceedings of the World Congress on Engineering, 2013; **3**: ISSN: 2078-0958.
9. Reyadh Shaker Naoum and Zainab Namh Al-Sultani, "Hybrid System of Learning Vector Quantization and Enhanced Resilient Back-propagation Artificial Neural Network for Intrusion Classification", IJRRAS, 2013; 14(2).
10. V. Christina, S. Karpagavalli and G. Suganya, "Email Spam Filtering using Supervised Machine Learning Techniques", *International Journal on Computer Science and Engineering*, ISSN: 0975-3397, 2010; **2**(9): 3126-3129, .
11. Ahmed H. Aliwy, "Tokenization as Preprocessing for Arabic Tagging System", *International Journal of Information and Education Technology*, 2012; **2**(4): 348-353.
12. M. Basavaraju, Dr. R. Prabhakar, " A Novel Method of Spam Mail Detection using Text Based Clustering Approach", International *Journal of Computer Applications, Foundation of Computer Science*, 2010; **5**(4): 15-25.
13. GrzegorzDudek, "An Artificial Immune System for Classification with Local Feature Selection", *IEEE Transactions on Evolutionary Computation*, 2012; **16**(6): 847-860.