Vol. 7(25), Jul. 2017, PP. 3536-3544

Contents list available at IJMEC

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

Journal Homepage: www.aeuso.org

PISSN: 2411-6173 - EISSN: 2305-0543

A hybrid Approach for Spam Detection Based on Decision-tree Algorithm and Neural Network

Zahra Mashayekhi[1*] and Ali HarounAbadi[2]

[1] *Department of Computer, Institute For High Education ACECR Khouzestan, Khouzestan, Iran*

[2]*Department of Computer, Central Tehran Branch, Islamic Azad University, Tehran, Iran Department*

Phone Number: +98-9303989979

***Corresponding Author's E-mail:** mashayekhi.zahra86@gmail.com*

Abstract

Spams which are knows as unsolicited bulk email (UBE), are increasingly becoming a harmful section in email traffics. Besides taking the time of the users and required time to scan and eliminate a huge amount of received spams, the network bandwidth and the storage space are consumed by these types of e-mails and they slow down the e-mail servers as well as providing a medium for the distribution of offensive and harmful content. Filtering is a simple and effective solution for the fight against spams. Fortunately, numerous methods are offered for the classification of the e-mails in spam and valid categories which work automatically. In recent years, the machine learning classification algorithms for automatically filtering of the spams have attracted the attentions of researchers. In this paper, a method is presented for detecting spams by machine learning algorithms by combining decision-tree algorithms and neural network. For the output combination of the aforementioned categories, bagging algorithm as well as entropy method are used for the configuration of the categories. The proposed method is evaluated by a set of Lingspam data. The results of the evaluations show that the proposed method provides 4% improvement in terms of accuracy compared to other methods.

Keywords: *spam, machine learning, decision tree, neural network, bagging*

1. Introduction

One of the most important services that the internet provides for its users is the ability of using electronic mails as an interface for contacting others. This service has attracted the attention of many of the users due to its low costs. Therefore, some tried to take advantage of it for promoting products, getting personal information from the users etc. which has caused the spread of spams or unsolicited commercial emails. In the past, these spams caused troubles, but today, due to their huge volume, they have turned into a big problem which cause the waste of time and cost for many users. Spams also cause traffics and the waste of storage space and calculation power. Trying to reduce the exposure to spams as much as possible is counted as a point [1]. Spam is an unsolicited mail which is sent directly or indirectly by someone unknown who does not have any relation with the receiver of the mail [2]. Therefore, methods are needed to automatically filter these bothersome mails. In fact, filtering the spams is a computer program which has the ability of categorizing the electronic mails. This program most probably should be able to detect the spams. Most of these filters are a combination of methods such as using multiple black or white lists, using key words, rule based filters etc. for precise detection

Article History:
Received Date: Feb. 05, 2017
Accepted Date: Apr. 21, 2017
Available Online: Jul. 03, 2017

of spams. These methods can be single-handedly effective. But, in commercial applications, a combination of these method are used [3]. Some of these method are set manually by the user. The filters used in Yahoo e-mail are of these types of filters. In this filer, the user introduces addresses of the spammers and the e-mail service provider stores the received electronics spams separately or delete them base on the introduced addresses. But, these method have a major flaw which is the fixed rule used in these filters which should be determined by the user [2]. The other issue for these methods is that the spam writers could deceive these filters with different tricks.

The method which has become popular recently is the content-based detection of the spams which has had considerable progress over the recent years in separating the spams based on their contents. Separating the legitimate e-mails and spams base on their content could be counted as text classification, as the body of most of the mails are in text form and when a spam is received, its classification should be determined [1]. Machine learning algorithms are used extensively in filtering electronic spams and these methods have enhanced spam filtrations to a considerable level. These algorithms learn to classify the documents based on their contents. This action need a learning stage. In the learning stage, the teaching is supported by predetermined classes. On the other hand, spam filtering advances, spammers' methods as well as the content of the spams are developing and advancing, too. The spammers try to attack the filters and this leads to the reduced effectiveness of the filters. Therefore, a method is needed which can increase the effectiveness of the filters.

The objective of this paper is to present a hybrid approach based on decision-tree algorithm and neural network to increase the spam detection accuracy. Using the decision-tree algorithm is such that the leaves of the tree are used to categorize the legitimate emails from the spams and the neural network is employed for the email content section. The bagging technique is used for the combination of the output result. The bagging generates the random training sets by different samplings with replacement which is known as bootstrapping. It is evident that replacing different parts of data, provides the possibility of having different responses for teaching the classifiers. Also, using the entropy method is used for the configuration of the classifiers. Using the proposed model and framework as well as employing traditional classification techniques on the contents of the e-mails, by involving the non-text and domain knowledge features, a better classification could be simply implemented.

The paper is organized as follows. In section 2, the related works are presented. In section 3, the proposed method for detecting the spams is described. The evaluation of the proposed method is performed in section 4. Finally, in section 5, the paper is concluded the future works are expressed.

2. Related works

Numerous researches are carried out in the field of spam detecting using machine learning techniques. Some of these works which are carried out recently are investigated as follows.
Chang et al. [4] used three machine learning methods for the classification of the e-mails including one Bayes classifier and two k-nearest neighbor classifiers with TF-IDF and compared the results. In this research, a combination of words was used as the feature instead of using the words in the e-mail. Then, the authors studied the effects of the word sizes, sampling and neighbor sizes and proposed a number of method to enhance the accuracy of the classification. The best results obtained was reported to be 99%.

Ying et al. [5] used a hybrid classifier to filter the spams which is a combination of three methods of support vector machine, decision-tree and artificial back propagation neural network and they used the majority voting method for the hybrid results from the three methods to find out whether the e-mails are legitimate or spams. In this research, 14 features of the spams are used for the classification of the spams including empty e-mail address and empty e-mail domain. The obtained accuracy by this system was reported to be 91.78% which shows high accuracy compared to running each algorithm separately. Ciltik et al. [6] presented a spam filtering method with high accuracy and low temporal complexity. They considered Turkish e-mails for their research. They used PC-KIMMO system and a

morphological analyzer for the extraction of the form of the word roots as an input and producing word analyses as the output. This method is based on n-gram and heuristic method. They developed two models: a general class model and email specific model. The first model, classifies the e-mail as spam or legitimate e-mail by using Bayes' rule. The second model determines a correct class of a message by comparing it with similar previous messages and looking for accordance. The third model is a combination of the aforementioned models. The order of the words is used for organizing the words based on the approved order for the n-gram model. This spam filtering method is based on text content and raw content classification of the results of the e-mails.

Singh et al. [7] tried to enhance the spam classification by a hybrid algorithm. In this research, genetic algorithm was combined with simulated annealing. They used genetic algorithm for the sake of artificial neural network learning. The mostly attended to investigate different parameters, mechanisms and architectures used for the performance optimization of the networks and obtaining a practical balance between global genetic algorithm and local search technique. The effectiveness of the spam filtering of the artificial neural network was investigated in the test for the accuracy and recall parameters and shows improvements by these two parameters in detecting spams. Moreover, this hybrid technique, guides the search toward optimization and guarantees better convergence and better performance with less numbers of training steps.

Bhuleskar et al. [8] combined the advantages of different filtering techniques and introduced a hybrid filter. The hybrid method is implemented using different models considering the resources available to the server. After examining the available literature in this field and a comparative study on the advantages and disadvantages of each filter, an effective hybrid spam filter was proposed which could be easily implemented. The four used filtering techniques included: a) content-based filtering, b) rule-based filtering, c) forging (e-mail and server), d) whitelist filtering. This model uses a parallel structure. The parallel filters are organized and independent which are transferred to an array. This array is then sent to the processor to generate the spam average. This model performs six consecutive operations on each input e-mail to confirm its validity. The effectiveness of this proposed method was found to be 98.73%. In this hybrid filtering method, the effectiveness is higher than the method in which each filter is used separately.

In [9], a method that uses SVM and TF-IDF algorithms is presented which uses TF-IDF for the extraction of the features and SVM is used for detecting whether an e-mail is a spam. The results of the test show the good performance of this system in terms of the accuracy in spam detection.
In [10], a 3-step method is used for spam filtering. In this system which does not rely on any specific classifier, the authors have introduced a method in which a decision is made at the end of the three steps using Bayes' theorem based on penalizing a wrong decision and in each step, the classifier presents the probability of whether an e-mail is legitimate or spam. This system provides a good performance in detecting spams.

3. The proposed method

In this paper, a new method is proposed for detecting spams. Figure 1 depicts all the steps in this method. In the proposed method, the features of each e-mail is first extracted by the feature extractor. Then, each e-mail is digitalized and displayed by a feature vector. The feature extraction analyzes the information inside the headers and other e-mail contents. This step has an important contribution to increase the effectiveness of the system. The email content features are the very words that are used inside the e-mail text. Some of the used words are repetitive and frequent in spams. Detecting these words in the incoming e-mails considerably helps to detect the spams. Recognizing these words is achieved by experience and familiarity with spams.

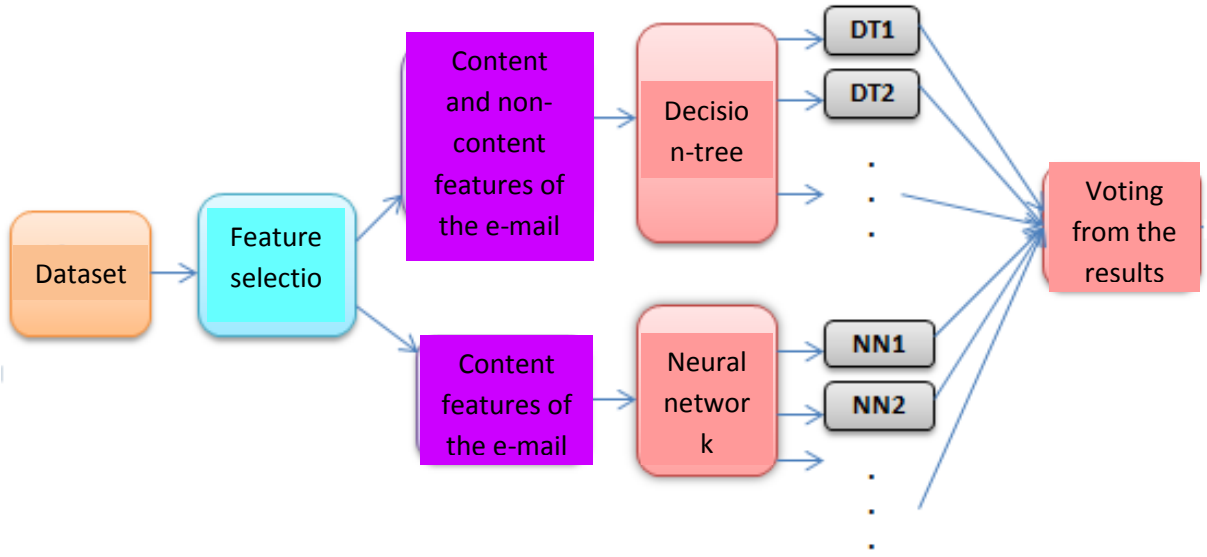

Figure 1: The proposed method framework

In the proposed method, the incoming documents which are already processed (the blacklist) and contain a series of pre-determined words are used. Then the decision-tree algorithms C4.5 and feedforward neural network are applied to a dataset of feature vectors. An answer based on the decision-tree and another one based on the neural network are produced. Then these two answers are combined based on bagging algorithm and it is finally decided whether an e-mail is spam. The details of the algorithms and techniques employed in the proposed method are described as follows.

3.1. Detecting spams by using feedforward neural network

The artificial neural networks are categorized as dynamic systems which transfer the knowledge or the hidden rule beyond the data to the network structure by processing the empirical data. The feedforward neural network is of great interest due to its special neural structure and the possibility of using error back-propagation learning algorithm. In neural networks with layer structure, the back-propagation algorithm is usually used as the learning algorithm due to its favorable effectiveness.
In the proposed neural network, investigates the word and contents of the e-mail. Words such as win$, free, many, online etc. are the words which are present in most of the spams. In the proposed method, these words are first detected in the e-mail content by neural network and determines the probability of whether an e-mail is a spam. But, the larger the range of these words, the higher the possibility of detection.

In the proposed method, the used feedforward neural network consists of a three-layer architecture. The input vector is applied to the first layer and its effects are propagated to the outer layers from the middle layers. In this path, the network parameters are assumed to be fixed and constant. In the second path which follows the BP rule, the network parameter are changed and adjusted. This change is applied according the error-correction rule. The amount of error is equivalent to the difference between the favorable response and the actual response of the network. After the calculation of the error, it is distributed in the return path by the network layers throughout the network. The input layer consists of 43 nodes equal to the number of the features of the content of the e-mail and the outer layer consists of 2 nodes. 70% of the input dataset to the neural networks are used for learning, 15% are used for validation, and other 15% are used as the training data.

The number of the nodes in the hidden layer is a problem which is usually solved by trial and error such that the lowest classification error is achieved. As a multi-layer perceptron model could have sufficient capabilities with only one hidden layer and increasing the number of nodes of this layer, in

the proposed method, one hidden layer is considered for the neural network. 10 nodes are used in the hidden layer.

3.2. Spam detection using decision-tree

Some of the e-mail content features (the considered words in the proposed method) are examined by neural network. Afterwards, non-content features of the e-mail are examined using decision-tree. These features include: IP address and word length size. In the proposed method, C4.5 decision-tree is used. Selecting the feature that should be investigated in each node is of great importance in decision-tree. The feature should be selected which has the highest impact in the classification of the samples. As a good criterion, the information gain criterion could be considered for determining the superiority of a feature which measures the impact of a feature for the classification of the sample based on the classification of their objective function. C4.5 uses this criterion for selecting the feature in each step of the growth of the tree.

3.2.1. information gain

For an accurate definition of the information gain, the definition of another criterion called entropy, which has broad applications in information theory, is used. This criterion determines the uniformity or non-uniformity of a set of samples. For a given S set of positive and negative samples, the concept of entropy objective for the S set is defined as follows based on this logical classification:
If the feature has C different values, the entropy S for this classification with C number of cases is defined as follows:

$$Entropy(s) = \sum_{i=1}^{c} -p_i log_2 p_i \quad (1)$$

In this equation p_i is a fraction of S which has a value of i. Here, if the feature has c possible cases, the maximum entropy will be $\log_2 c$. For a given entropy as a criterion for the uniformity of the sample training set, one can obtain the effectiveness criterion of a feature in the sample training classification. As stated before, this criterion is called the information gain. The information gain is the expected decrease of the entropy value resulted from the classification based on a specific feature. To be more accurate, the information gain of the feature A on the S set, Gain(S,A), is defined as follows in terms of the available sample sets:

$$Gain(S, A) \equiv Entropy(s) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (2)$$

In this equation, Values(A) is the class of all possible values for feature A. The mean entropy in equation (2) is the entropy set for all S_y's which are multiplied by the ratio of the samples $\frac{|S_v|}{|S|}$. Therefore, Gain(S,A) is the expectation of entropy decrease by classification based on feature A. In other words, Gain(S,A) is the amount of information which is obtained regarding the value of the objective function by having the value of feature A. For each feature, the information gain is calculated and the feature with the highest value is selected as the best feature for the classification. In what follows, the information gain is obtained among other features and is set as the next nodes of the tree and this procedure continues until the features are finished.

3.3. Feature selection

One of the issues of spam filtering the high dimensions of the feature space. The feature space which is formed out of available words and terms in the documents, consists of tens of thousands or even more features. This issue is a huge obstacle for many of machine learning algorithms. Hence, a

dimension reduction step is required. The main objective of dimension reduction is to reduce the vector space without losing the effectiveness of the classification. in this regard, numerous techniques exist. The feature selection is the most common method used in for texts classification. In the proposed method, Shannon entropy method is used.

- **Weighting based on Shannon entropy method**

The main idea of this method is based on the fact that as the dispersion is higher for the values of an index, that index is of greater importance. Therefore, the weights of the indices are calculated as follows:

$$P_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}; \forall_{i,j} \quad (3)$$

$$k = \frac{l}{ln(m)} \quad (4)$$

$$E_j = -k \sum_{i=1}^{m} [p_{ij} \ln p_{ij}]; \forall_j \quad (5)$$

E_j is the entropy for the jth index and m is the number of options.

d_j expresses the value of uncertainty or deviation degree of the jth index and as the Shannon entropy method assigns the highest weight to the index with the highest deviation degree, the weight are calculated as follows:

$$w_j = \frac{d_j}{\sum_{j=1}^{n} d_j}; \forall_j \quad (6)$$

Table 1 shows the extracted features from the dataset.

Table 1: The extracted features from LingSpam dataset

Feature	Description	Values
Feature 1 to 4 (Checking IP)	IP classification to four classes: features 1 to 4 show the normalized IP values	[0-100]
44 features associated words	The most frequent words in e-mails	[0-100]
6 sign features	The repetition percentage of punctuations (:, ., !, ?, ;) to the total number of punctuations in each e-mail	[0-100]
One feature of uppercase letters	The longest non-stop repetition of uppercase letters	[0-…]
One feature of uppercase letters	The average number of non-stop repetition length of uppercase letters	[0-…]
One feature of uppercase letters	The percentage ratio of total uppercase letters to lowercase letters	[0-…]

Using weighted entropy method for reduction of features, the features of the dataset are reduced from 67 to 47. In this method, a weight is assigned to each feature. In the proposed method, the features with weights of more than 0.1 are considered and the ones with weights less than 0.1 are omitted. The reduced features are shown in table 2.

Table 2: The reduced features of the dataset

Feature	The feature number after LingSpam reduction
IP (1 to 4)	2,1,4
The feature regarding words (5 to 48)	6 to 11, 13 to 18, 20, 22 to 48
Punctuations feature (49 to 54)	49, 51
One uppercase feature	54
One uppercase feature	-
One uppercase feature	56

3.4. Bagging technique for the combination of classifiers

In the proposed method, bagging technique is used to combine the output of the classifiers of the decision-tree and neural networks. The bootstrap aggregating based bagging is one of simplest and yet the most successful group methods for the improvement of classification problem. This concept is used for the combination of predicted classifications from several models. For each training sample, each classifier generates an output value which determine the relevance degree of that sample to a class. Then, the output of the classifiers are combined to reach a consensus using voting method [11].

4. Evaluation results

4.1. Dataset

Lingspam dataset is used for the test [12]. This dataset includes 2412 normal messages and 481 spams. Kfold method with k=10 is used for the evaluation. In order to generate diversity in the classifiers, 70% of the training set is selected each time using bootstrap method and is applied to the classifiers. Frist, the required features of the system, being the words, are detected.

4.2. The evaluation criteria

For the evaluation of the proposed method, accuracy, precision, recall and F1 criteria are used. In what follows, the method of calculating these criteria are investigated. To calculate these criteria, the following criteria are employed:

True Positive (TP) of f++, which concerns the number of correctly predicted positive samples by the classification model (the spam which is detected correctly as a spam, meaning a test data record with original class of 1, which is classified as class 1 by the classifier.)

False Negative (FN) or f+-, which concerns the number of positive samples which are incorrectly predicted as negative by the classifier model (a spam incorrectly detected as a legitimate e-mail).

True Negative (TN) or f--, which concerns the number of number of negative samples which are correctly predicted by the classifier model (a legitimate e-mail detected as a valid e-mail).

The most important criterion for detecting the application of a classification algorithm is the accuracy criterion. This criterion calculates the total accuracy of a classifier. This criterion shows the total number of spams and correctly predicted e-mails to the total number of spams and e-mails. Equation 1-4 shows the way the accuracy criterion is calculated:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1\text{-}4)$$

The precision criterion is the ratio of the number of correctly classified messages which are from the spam classes to the total number of messages detected as spams. The way this criterion is calculated in shown in equation 2-4 as below:

$$Precision = \frac{TP}{TP + FP} \quad (2\text{-}4)$$

The recall criterion is the ratio of the total number of detected messages as spams to the total number of messages which are actually in the spam class. The way this criterion is calculated is shown in equation 3-4:

$$Recall = \frac{TP}{TP + FN} \quad (3\text{-}4)$$

The F-measure or F1 criterion is a combination of recall and precision criteria and is used in cases in which no specific importance could be regarded to precision and recall criteria. Equation 4-4 shows the way this criterion is calculated:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (7)$$

4.3. The evaluation of the proposed method

The results of implementing the proposed method on LingSpam dataset set based on accuracy, recall, F1 and accuracy criteria are shown in table 3. As observed in this table, good results are obtained by the proposed method.

Table 3: The results of the evaluation of different criteria for the proposed method

k-fold	Recall	Precision	F1	Accuracy
1	0.86	0.86	0.86	**0.89**
2	1	1	1	**0.93**
3	1	0.94	0.97	**0.86**
4	0.94	0.94	0.94	**0.93**
5	0.74	1	0.85	**0.93**
6	0.8	1	0.89	**0.97**
7	0.86	0.86	0.86	**0.86**
8	1	0.94	0.97	**1**
9	0.93	1	0.96	**0.79**
10	0.79	0.85	0.82	**0.90**
Mean	**0.89**	**0.94**	**0.91**	0.92

4.4. The comparison of the methods

In this section, the proposed method in this paper with the proposed svm based method in [9] and the 3-step NB based method in [10] as well as using each method of neural network and decision-tree are compared separately for spam detection. The performance results of these five methods are presented in table 4 based on accuracy and F1. According to the table 3, it is clear that the proposed method has higher accuracy and F1 criterion compared to the other four methods. As seen in the table,

the combination of the two decision-tree and neural network methods lead to better results compared to using each algorithm separately for spam detection.

Table 4: Comparing the proposed method with other methods

Approach	F1	Accuracy
[9]	0.87	0.85
[10]	0.88	0.87
FeedForward Neural network	0.70	0.61
C4.5	0.85	0.88
Proposed Method	0.91	0.92

The most important compared criterion is the accuracy criterion for which the results show a 4% improvement for the proposed method compared to other methods. The evaluation shows that combining the classifiers could lead to better results.

Conclusion

In this paper, a method is presented for spam detection using the combination of decision-tree and artificial neural networks and bagging technique. In the proposed method, different classifiers are first used to detect spams and then, bagging technique is used to combine the results of these two classifiers. The results from the evaluation on LingSpam dataset shows the power and capability of this method in detecting spams. In order to improve the results of the proposed method, one can combine it with other classification methods such as genetic algorithm or other types of decision-tree and neural networks. Also, one can use other aggregating techniques such as boosting in the algorithms combination step.

References

[1] E. Ghanbari, “Spam detection with incremental learning approach”, M.Sc. thesis, (Sharif University of Technology), 2010, p. 113.

[2] X. Tang, "Hybrid Hidden Markov Model and artificial neural network for automatic speech recognition". In Pacific-Asia Conference on Circuits, Communications and Systems, 2009. PACCS'09., 2009, pp. 682-685.

[3] L. Nosrati, “Unsolicited e-mails detection using content changes”, M.Sc. thesis, (Sharif University of Technology), 2010, p.86.

[4] M. Chang and C. K. Poon, "Using phrases as features in email classification". Journal of Systems and Software, 82(6), 2009, pp. 1036-1045.

[5] K. C. Ying, S. W. Lin, Z. J. Lee, and Y. T. Lin, “An ensemble approach applied to classify spam e-mails”, Expert Systems with Applications, 37(3), 2010, pp. 2197-2201.

[6] A. Çıltık, and T. Güngör, "Time-Efficient Spam E-mail Filtering Using n-gram Models". Pattern Recognition Letters, 29(1), 2008, pp. 19–33.

[7] S. Singh, A. Chand and S. P. Lal," Improving Spam Detection Using Neural Networks Trained by Memetic Algorithm". In Fifth International Conference on IEEE Computational Intelligence, Modelling and Simulation (CIMSim), 2013, pp. 55-60.

[8] R. Bhuleskar, A. Sherlekar and A. Pandit, "Hybrid spam e-mail filtering" In First International Conference on IEEEComputational Intelligence, Communication Systems and Networks, 2009, pp. 302-307.

[9] K. D. Renuka and P. Visalakshi, “Latent Semantic Indexing Based SVM Model for Email Spam Classification”, Journal of Scientific & Industrial Research, 73(7), 2014, pp. 437-442.

[10] X. Jia, K. Zheng, W. Li, T. Liu and L. Shang, "Three-way decisions solution to filter spam email: an empirical study", International Conference on Rough Sets and Current Trends in Computing. Springer Berlin Heidelberg, 2012, pp. 287-296.

[11] M. SanieeAbadeh, “practical data mining”, niyaz danesh press, tehran, 2012, pp. 123-125.

[12] M. C. Su, H. H. Lo and F. H. Hsu, “A neural tree and its application to spam e-mail detection”, Expert Systems with Applications, 37(12), 2010, pp. 7976-7985.