A NOVEL FEATURE SELECTION METHOD USING CFS WITH GREEDY-STEPWISE SEARCH ALGORITHM IN E-MAIL SPAM FILTERING

Seyed Mostafa Pourhashemi[1], Alireza Mohammad Mashalizadeh[2]

[1]*Department of Computer, Dezful Branch, Islamic Azad university, Dezful, Iran*
s.mostafa.pourhashemi@hotmail.com

[2]*Department of Mechatronic, South Tehran Branch, Islamic Azad university, Tehran, Iran*
ammashalizadeh@yahoo.com

Abstract

The purpose of this research is presenting an machine learning approach for enhancing the accuracy of automatic spam detecting and filtering and separating them from legitimate messages. In this regard, for reducing the error rate and increasing the efficiency, a new architecture on feature selection has been used. Features used in these systems, are the body of text messages. Proposed system of this research has used *Correlation-based feature selection* (CFS) with *Greedy-stepwise search* algorithm. In addition, *Multinomial Naïve Bayes* (MNB) classifier, *Discriminative Multinomial Naïve Bayes* (DMNB) classifier, *Support Vector Machine* (SVM) classifier and *Random Forest* classifier are used for classification. Finally, the output results of this classifiers methods are examined and the best design is selected and it is compared with another similar works by considering different parameters. The optimal accuracy of the proposed system is evaluated equal to 99%.

Keywords: Feature Extraction, Feature Selection, Classification, Spam Filtering, Machine Learning.

1. Introduction

In Recent years, the massive rise in Internet and low cost of E-mail have attracted a lot of attention of the most of advertisers of markets. As a result, receiving a high volume of unwanted messages which are increasing day by day, have become commonplace for users. This unwanted messages called Spam [Androutsopoulos et al., 2000]. Spams, in most cases are advertisements for advertising suspicious, plans for getting rich fast and seemingly legitimate services [Spam Abuse Corporation, Visited on 2013].
Spams are annoying for most of users, because not only beginning to diminish the reliability of e-mails, even users are affected by Spam due to the network bandwidth wasted receiving these messages and the time spent by users distinguishing between Spam and normal (legitimate) messages and damaging to the recipient system via malwares and viruses carried By spams [Androutsopoulos et al., 2000].
Nowadays, There are many ways which designed to remove spam. This methods use different techniques for analysing of E-mail and to specify that whether it is spam or legitimate mail.
Among all spam filtering approaches, Machine Learning technique has the best and high performance in spam classification. This method does not require any special rules. Instead, it needs many messages that nature of them (spam or legitimate) is identified, as training

instances for the system. An special algorithm is used for training the system for finding the rules of message classification [Tretyakov, 2004].

Ultimately, what we want to achieve is a spam filter which it can be represented as a f function which it specifies that received message m is spam or legitimate.

If we show all the received messages by M, Then we can say that we are looking for a function f defined by the equation (1).

$$f : M \rightarrow \{S,\, L\} \tag{1}$$

Figure 1 shows an overview of a spam filter that is used in most modern filters which acts based on machine learning.
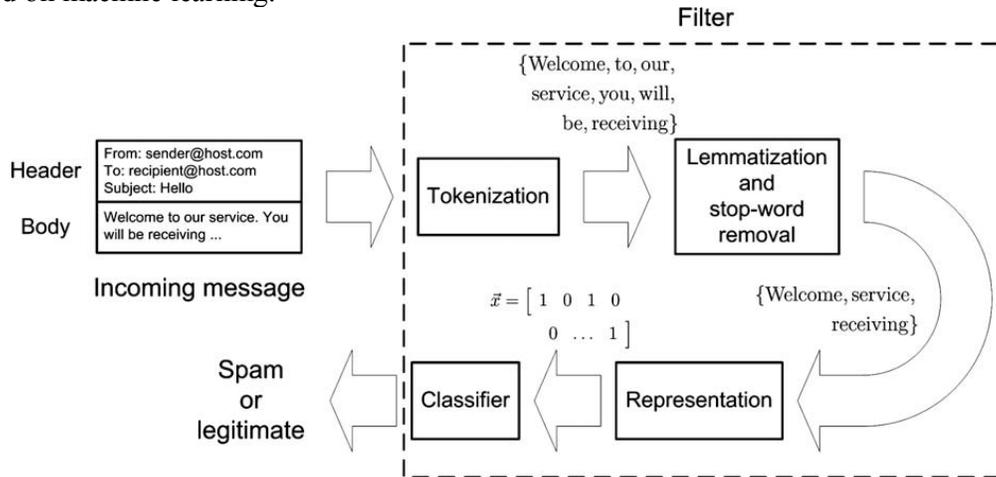

Figure. 1. An illustration of some of the main steps involved in a spam filter

A brief description of the various parts of Figure 1 is as follows:

- Preprocessing: At this phase, first all the words in the message are separated, then based on an preliminary analysis, Stop words like a-are-is-of… which do not help in classification, are separated among them and the remaining words use to determine that whether it can be a appropriate feature in classification or not, and these are sent to the next stage if these have the right conditions.
- Feature Extraction and Selection: In this section, Preprocessing phase output words, are examined based on some primary filters and the rules and conditions which designer Considers. Finally, specified number of words are selected as the main features. The selected features which are used in training the system and message classification, have important roles in the final performance of filter.
- Training the system: After selecting optimal features, we need to train the system. In this phase, from training instances, a database will be created based on optimal features, which the system is trained by it.
- Classification: In this phase, system decides whether or not it's spam, by checking the input message and based on the training that the system has been.
- Spam / Legitimate: Based on the final result of filter, Message is placed in the appropriate folder [Guzella and Cominhas, 2009].

2. Theory

In the general case, the problem of spam filtering can be displayed as eqation (2).

A novel feature selection method using CFS with Greedy-stepwise search algorithm in e-mail spam filtering

$$F(m,\theta) = \begin{cases} C_{spam} & if \quad m \quad is \quad spam \\ C_{leg} & if \quad m \quad is \quad legitimate \end{cases} \qquad (2)$$

While m is the message should be classified, θ is vector of parameters, C_{spam} and C_{leg} are labels which are assigned to message. In most of spam filters which act based on machine learning, θ is the result of the training of classifier on pre-collected data set. Specifications of the whole system is introduced by equation (3).

$$M = \{(m_1, y_1), ..., (m_n, y_n)\}, \quad y_i \in \{C_{spam}, C_{leg}\} \qquad (3)$$

While $m_1, m_2, ..., m_n$ are marked as spam or legitimate by $y_1, y_2, ..., y_n$ labels, and θ is training function [Blanzieri and Bryl, 2008].
A filter which acts based on machine learning, uses a set of labeled data for training and analysing (a set which previously collected and the judgement has been performed about them, whether they are spam or legitimate).

2.1. Related works

In reference [Zhu and Tan, 2011], by using Sliding Window and appropriate method in counting of word frequencies on spam and legitimate messages, and using variance of event frequencies for feature selection and by using SVM & Naive Bayes classifiers, the performance reached to 96.8 %.
In reference [Besavaraju and Prabhakar, 2010], by using appropriate preprocessing based on clustering, and using KNN (K-Nearest Neighbours) classifier, good results are obtained after classification.
In reference [Michelakis et al., 2004], the authors have developed a system called Filtron, in which by appropriate using of n-gram method and Information Gain (IG) and Black-White Lists and using by Flexible Bayes, good results are obtained with uni-word terms.
In reference [Beiranvand et al., 2012] by using a hybrid feature selection system based on document frequency and IG method, and using Adaboost for classification has very good results, and the performance reached to 98.3 %.

3. Methods

In this section, by considering mentioned topics in sections 1 & 2, we will describe proposed methods (included Preprocessing, Feature Extraction and Feature Selection and used classifiers), and more we will review how to create and operation of spam filter which acts based on machine learning.

3.1. Preprocessing

The first phase should be done in order to create a filtering system, is Preprocceing. In this paper, we use the body of message which includes the main text of the message, for analysing messages.
The method which we have used to display features, is *N*-Gram with values *N*=1,2,3 which uni-word and dual-word and trey-word terms should be extracted among the body of text messages, to achieve this goal.

To determine that which features are useful for the system, the first thing to be done is preprocessing, that stop words which are not effective removed, and the words are tokenized (for example, elimination of *ing* & *ed* from end of verbs); as a result, the computational load of these features into the system, and the volume of preliminary information are reduced.

After the above preprocessing steps, we need a way to initialize the features. To do this, *Term Frequency* technique has been used. In this method, for each document, first, frequencies of each features are calculated and finally for the document, a vector is formed which included features with their frequencies [Chang and Poon, 2009]. The continuation of data mining is done by processing of these vectors.

3.2. Feature selection

Feature selection is the most important phase in data mining and machine learning. Feature selection is used to reduce the main extracting data, to be improved both in terms of computational load and achieves the highest performance.

3.3. The used feature selection method in this paper

We are dealing with a very large number of features, so for achieving the best result, we use an appropriate feature selection method, that includes the methods which are handled in "Filter" approach.

Filter model selects features based on separate specifications of features and well-being of a feature. Independent tests are implemented on information and also function evaluation selects output subset [Geng et al., 2007]. The proposed process is shown in Figure 2.

F**igure. 2.** Process of implementation and filtering in proposed method

As it can be seen in Figure 2, first primary data enters into filter 1. In this filter, Stop words, worthless words and Tokens are removed, it makes the original data size is somewhat reduced.

In filter 2, we use a filter which acts based on "Filter" method, and it has more precision than filter 1. This filter is used to decrease the features, to find the optimal subset and to increase the performance of classifier. In classification phase, four classifiers (DMNB, MNB, SVM and Random Forst) are used that the output results of this filter and the results of reviewed classifiers will be presented in section 4.

3.3.1. Filter 1

The overall messages placed at filter1 as text documents and this filter uses Bag of Words (BoW) to show words per document. Term Frequency method is used to extract the words and to recognize the usefulness of them, that the frequency of each word per document is calculated and the features which are repeated lower that a threshold, will be removed.

Then, we should separate more usefull features by special techniques. First we should calculate and analyze frequencies of each word in spam class and legitimate class separately. So we change the method of calculating the number of occurrences by defining two new parameter (according to equation 4) for each feature.

A novel feature selection method using CFS with Greedy-stepwise search algorithm in e-mail spam filtering

$$C_{s,x} = \frac{N_{s,x}}{N_s} \quad , \quad C_{h,x} = \frac{N_{h,x}}{N_h}$$

(4)

The $C_{h,x}$ and $C_{s,x}$ parameters are calculated for each features. In above-mentioned equation, N_h and N_s represent the total number of legitimate (ham) messages and spam respectively. $N_{h,x}$ is equal by total number of documents which contain x, and that message are one of legitimate messages. $N_{s,x}$ is equal by total number of documents which contain x, and that message are one of spam messages. After calculating the above values, and by considering a threshold, we can check the features.

For a feature, if $C_{h,x}$ and $C_{s,x}$ parameters are very close together, then it represents that feature is distributed in spam & legitimate messages equally, thus it can not be a good feature for separating both spam and legitimate classes. If $C_{h,x}$ and $C_{s,x}$ parameters have an appropriate difference (threshold) together, so the feature is repeated in one of classes more, and recognization of two classes can be done by the feature.

3.3.2. Filter 2

Correlation-based feature selection(CFS) is an effective feature selection method, and the set of features mostly related to some class can be selected from the gene expression data. It reduces the data in dimensionality by more than sixty percent in most cases without negatively affecting accuracy [Hall, 2000].
The CFS measure evaluates subsets of features on the basis of the following hypothesis: "Good feature subsets contain features highly correlated with the classification, yet uncorrelated to each other"
we have used CFS with Greedy-stepwise search algorithm as feature selector in filter 2. This filter has important role in identifying spams using proposed methods, due to the high performance of classifier and selecting optimal subset.

3.4. Performance evaluation of classifier

For evaluating the performance of a classifier, there are two categories of indicators, Information Retrieval and Decision Theory. But another problem that should be noted in evaluating a classifier, is the costs for messages are being incorrectly classified. Accordingly, accuracy parameter can not be suitable for evaluating classifier solely.
In the field of decision theory, if we consider spam class as Positive class, then *TP* and *TN* parameters based on equations (5) and (6) can be defined.

$$\eta_{tp} = \frac{n_{S,S}}{n_S}$$

(5)

$$\eta_{tn} = \frac{n_{L,L}}{n_L}$$

(6)

While n_S is the total number of spams in data set, and n_L is the total number of legitimates in data set. $n_{S,S}$ is total number of spams which are correctly diagnosed, and $n_{L,L}$ is total number of legitimates which are correctly diagnosed.
In the field of information retrieval, classification be tested based on Precision & Recall parameters. Precision parameter represents the total number of positive class instances that are correctly classified to the total number of instances which have been diagnosed as

positive. Recall parameter represents the total number of positive class instances that are correctly classified to the total number of instances. Precision & Recall parameters are shown in equations (7) and (8) for spam class.

$$P_S = \frac{n_{S,S}}{n_{S,S} + n_{L,S}} \tag{7}$$

$$r_S = \frac{n_{S,S}}{n_{S,S} + n_{S,L}} \tag{8}$$

By combining Precision & Recall parameters, another parameter is defined, called F_β which β is determined for exactitude. The value of β has been equal to 1 for most of the previous works. How to calculate the F_β parameter is shown in equation (9).

$$F_\beta = (1 + \beta^2) \frac{r_s p_s}{\beta^2 p_s + r_s} \tag{9}$$

In the proposed method, the value of β has been selected equal to 1.

4. Results and Discussions

In this section, how to implement the tools has been described, and then the output of the proposed method is presented, and finally, the results are compared with some of similar previous works.
To implement different parts of the designed system, we have used MATLAB (version 7.14) for feature extracting and above-mentioned preprocessings and we have used updated version of Weka (version 3.7.9) for used filters and classifications.

4.1. Used data set

Each machine learning system requires a training set to train the system. In this paper, we have used LingSpam [LingSpam Public Corpus, Visited on 2013], as standard data set, including 2893 text messages which 2412 messages (about 83.37 %) are legitimate and 481 messages (about 16.63 %) are spam. In this data set, all of HTML tags and headers except Subject have been removed. We have used the third version of this data set. In test phase, we have used this data set (LingSpam data set) again in 10-folds cross validation mode. So we have used the LingSpam data set on both of training and testing phases.

4.2. Separation the words and features

Features are the most important part of each machine learning problem. In this paper, features are terms within text messages which should be extracted from body of text messages. To extract desired words, space character has been used as separator. In Table 1, the number of extracted features for words of length 1, 2, 3 are shown.

A novel feature selection method using CFS with Greedy-stepwise search algorithm in e-mail spam filtering

TABLE 1. EXTRACTED FEATURES

Length of terms	The number of extracted features
Uni-word	62089
Dual-word	125396
Trey-word	170341

For accurate study and better test of the proposed method, we have lengths of terms in this research between uni-word and trey-word.

4.3. Feature selection based on filter

Based on the got features in Table 1, it is necessary to eliminate redundant features. To do this, first the features have been studied by filter 1 described in the previous section. Results of the filter, are reduced set of features, which are reported in Table 2. Then the output of the filter will be given to filter 2.

TABLE 2. THE OUTPUT RESULTS OF FILTER 1

Length of terms	The number of features after applying the Filter 1
Uni-word	1540
Dual-word	1942
Trey-word	2209

Feature reduction is done in filter 1, as mentioned in the previous section also, by considering a threshold and how to repeat the features in spam and legitimate messages. Because, with increasing length of the term, frequency of features will be changed in data set also, in this phase we have used different thresholds for different lengths of terms.

In this research, we have used *CFS* with *Greedy-stepwise search algorithm* as filter 2. The output of the filter is given to four classifiers *Multinomial Naïve Bayes (MNB)* [Kibriya et al., 2004] [Schneider, 2003], *Discriminative Multinomial Naïve Bayes (DMNB)* [Hall, 2008] [Su et al., 2008], *Support Vector Machine (SVM)* [Alpaydin, 2010] [Keerthi et al., 2001] with normalized poly kernel and *Random Forest* [Biau et al., 2008] [Breiman, 2001] with 100 random trees. It should be noted that all of classifications are done in 10-fold cross validation. In filter 2 and by applying filter model, we find the final feature set. In this phase. Table 3 represents the number of final features.

TABLE 3. THE NUMBER OF FINAL SELECTED FEATURES BY APPLYING CFS WITH GREEDY-STEPWISE SEARCH FOR FILTER

Length of terms	Number of features
Uni-word	58
Dual-word	64
Trey-word	58

According to the above table, after applying filter 2, the number of final optimal selected features in all of cases is different by the case that only filter 1 was applied.

4.4. *The output results of the proposed method*

In this system, uni-word features produced better results. The accuracy of this hybrid feature selection method for all of four studied classifiers, is shown in Table 4.

TABLE 4. THE ACCURACY OF CLASSIFIERS

Classifier	Accuracy (%)
DMNB	98.96
SVM	98.82
Random Forest	98.48
MNB	98.79

We consider the case which has most accuracy and precision on messages diagnosis, as proposed method and the output results are shown in Table 5.

TABLE 5. THE OUTPUT RESULTS OF PROPOSED METHOD

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)	FP
DMNB	98.96	99.0	99.0	99.0	7

According to the Table 5, Recall parameter for proposed method is equal to 99.0%, that represents a few number of spams which have been wrongly diagnosed as legitimates, and the Precision parameter is equal to 99.0%, that represents a few number of legitimates which have been wrongly diagnosed as spams. False Positive (FP) parameter is eual to 7, that represents only 7 messages of 2412 legitimate messages have been wrongly diagnosed as spams. By considering output results, it can be seen that proposed method is shown very good performance.

4.5.*Performance comparison of the proposed method with other references*

In Table 6, the proposed method is compared with some other methods, using different parameters. Notice that, the training and testing data set of all of following references, all similar to our dataset. It means that all of them have used LingSpam data set on training and testing phases, same to us.

TABLE 6. COMPARISON OF THE PROPOSED METHOD BASED ON DIFFERENT PARAMETERS

	Accuracy (%)	Precision (%)	Recall (%)
Proposed Method	98.96	99.0	99.0
(Zhu and Tan, 2011)	96.80	93.73	98.1
(Besavaraju and Prabhakar, 2010)	94.40	91.1	97.6
(Michelakis et al., 2004)	95.42	94.95	91.43
(Beiranvand et al., 2012)	98.30	98.3	98.3

Amount of difference between proposed method and other references is compared, and amount of performance improvement is shown in Table 7.

TABLE 7. AMOUNT OF IMPROVEMENT OF PROPOSED METHOD IN COMPARISON WITH OTHER REFERENCES

	amount of Accuracy improvement (%)	amount of Precision improvement (%)	amount of Recall improvement (%)
(Zhu and Tan, 2011)	+ 2.16	+ 5.27	+ 0.9
(Besavaraju and Prabhakar, 2010)	+ 4.56	+ 7.9	+ 1.4
(Michelakis et al., 2004)	+ 3.54	+ 4.05	+ 7.57
(Beiranvand et al., 2012)	+ 0.66	+ 0.7	+ 0.7

5. Conclusions and future works

The purpose of this paper is designing and presenting an machine learning system to increase the performance for automatic diagnosing and filtering spam messages from legitimate messages.

First, we attempted to seprate and to extract uni-word, dual-word and trey-word terms by considering the body of text messages. This terms are the features which messages can be judged by them, at next phases. For Appropriate judgment about a message, we should select the best features among all of extracted features; so, in continue, we enter the next phase called Feature Selection, which is done by two filters. In filter 1, after eliminating the stop words which are not effective and tokenizing the words, we calculated the frequencies of each features in spam and legitimate message catogories, then we deleted the features which have repeated lower than a threshold. In filter 2, we selected optimal set among reduced feature set, using filter algorithm. The performance of used classifiers, is one of parameters which helps in selecting optimal subset.

Output results of each classifiers and feature selection approaches which used in this paper, was noted in section 4, the performance of designed system was evaluated, the best design was selected and it was compared considering different parameters. Finally, what can be concluded about the designed system, it is that the proposed method in feature selection and the use of appropriate classifier can has very good performance in data mining issues.

For future work we will focus on Ontology. The combination of semantic ontologies in feature selection phase, can be used to improve classifier performance. In this paper, we used body of messages for decision making; we can use another characteristics like Sender address, Recipient address and Size of message also. And also we can generalize our proposed method on another data sets used for spam filtering (like multi-language datasets), and another data sets used for another topics based on text processing (like web classification) and finally we can test them and observe the results.

References

Alpaydin, E., 2010 February, Introduction to Machine Learning, Second Edition. *The MIT Press*, pp 350-380.

Androutsopoulos, I., et al., 2000, An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA*, pp 160-167.

Beiranvand, A., et al., 2012 March, Spam Filtering By Using a Compound Method of Feature Selection. *Published in Journal of Academic and Applied Studies (JAAS)*, Vol(2), pp 25-31.

Besavaraju, M., Prabhakar, R., 2010 August, A Novel Method of Spam Mail Detection Using Text Based Clustering Approach. *Published in International Journal of Computer Applications (IJCA)*, Vol(5), pp 15-25.

Biau, G., et al. (2008). Consistency of Random Forests and Other Averaging Classifiers. Published in Journal of Machine Learning Research (JMLR), 9, 2015-2033.

Blanzieri, E., Bryl, A., 2008 March, A Survay of Learning-Based Techniques of Email Spam Filtering. Published in *Elsevier Journal: Artificial Intelligence Review*, pp 63-92.

Breiman, L., 2001 October, Random Forests. *Published in Journal of Machine Learning, MA, USA*, Vol(45), pp 5-32.

Chang, M., Poon, C.K., 2009 June, Using Phrases as Features in Email Classification. *Published in Elsevier: The Journal of Systems and Softwares*, Vol(82), pp 1036-1045.

Geng, X., et al., 2007, Feature Selection for Ranking. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA*, pp 407-414.

Guzella, T.S., Cominhas, W.M., 2009, A Review of Machine Learning Approaches to Spam Filtering. Published in Elsevier Journal: Expert System with Application, Vol(36), pp 10206-10222.

Hall, M. (2000). Correlation-Based Feature selection for discrete and numeric class machine learning. In: Langley P, et al., eds. Proc. of the 17th Intl Conf. Machine Learning. San Francisco: Morgan Kaufmann Publishers, 359-366.

Hall, M., 2008 June, Discriminative Multinominal Naive Bayes for Text Classification. Community Contribution: Pentaho Data Mining-Weka/DATAMINING-125.

Keerthi, S.S., et al. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. Published in MIT press journal, Neural Computation. 13(3), 637-649.

Kibriya, A.M., et al., 2004 December 4-6, Multinominal Naive Bayes for Text Categorization Revisited. *Proceedings of 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia*, Vol(3339), pp 488-499.

LingSpam Public Corpus, *<http://www.aueb.gr/users/ion/publications.html>*, Visited on 2013.

Michelakis, E., et al., 2004 July 30-31, A Learning-Based Anti-Spam Filter. *Proceedings on First Conference on Email and Anti-Spam (CEAS), California, USA*.

Tretyakov, K., 2004, Machine Learning Techniques in Spam Filtering. *Data Mining Problem-Oriented* Seminar, pp 62-79.

Schneider, K.M. (2003). A comparison of event models for Naive Bayes anti-spam e-mail filtering. Published in EACL '03 Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, PA, USA, 1, 307-314.

Spam Abuse Corporation, <http://spam.abuse.net/overview/whatisspam.shtml>, Visited in 2013.

A novel feature selection method using CFS with Greedy-stepwise search algorithm in e-mail spam filtering

Su, J., et al. (2008). Discriminative Parameter Learning for Bayesian Networks. Proceedings of the 25th international conference on Machine learning (ICML), New York, NY, USA, 1016-1023.

Zhu, Y., Tan, Y., 2011 June, A Local-Concentration-Based Feature Extraction Approach for Spam Filtering. *IEEE Transactions on Information Forencics and Security*, Vol(6), pp 486-497.