Journal of Academic and Applied Studies
Vol. 2(3) March 2012, pp. 25-31
Available online @ www.academians.org
ISSN1925-931X

Spam Filtering By Using a Compound Method of Feature Selection

Azadeh Beiranvand, Alireza Osareh, Bita Shadgar

Beiranvand_a@yahoo.com , Shahid Chamran University, Ahvaz, Iran

Alireza.Osareh@scu.ac.ir , Shahid Chamran University, Ahvaz, Iran

Bita.Shadgar@scu.ac.ir , Shahid Chamran University, Ahvaz, Iran

Abstract

Nowadays, the increase volume of Spams has been annoying for the internet users. In the recent years, the applying of machine learning techniques has attracted many researches' attention for automatic filtering of Spams. In this article, a system of spam filtering has been presented based on Adaboost algorithm. In the proposed method, the available terms in email have been used as the basic features in classifying email issues. That is why the feature selection has an important role in effective improvement of Spam filtering In the proposed filtering system, a compound method has been used to identify related features and remove unrelated features, and the results have been tested and compared on a standard data set of Ling-Spam. Finally, to compare the obtained results, several other algorithms have been applied on the data and their results are compared with the obtained results. The results of the experiments clear the fact that this system has an acceptable efficiency about 0,983.

Keywords: Spam Filtering, Feature Selection, Machine Learning, AdaBoost.

I. Introduction

Nowadays, A problem named Spam has come to existence with regard to the widespread using of electronic mails (E-mails) which not only wastes the time of the users, but also it brings about other problems such as influencing on bandwidth and misusing of storage space. However, it is of high significance to plan an efficient system to effectively filter the Spams. Up to now, various methods have been presented in order to fight and spam filtering (Guzella, 2009). In this article, we aim to analyze, interpret, plan and apply an effective and flexible system to fight Spams based on statistic and machine learning machine methods.

In this respect, the emphasis is on using hybrid multi-layer architectures instead of individual classifyiers with regard to innate complications of Spam filtering problem and using clear methods by spammer in order to prevent recognizing and identifying such emails.

In (Chang, 2009), the three machine learning methods i.e. one Naïve Bayes classifier and two K-Nearest Neighbour classifier have been used and the results have been compared in order to classify electronic mails. In this article, the combination of words has been used instead of the available words in the mails. Then the authors of this article, have considered the influence of word length, the size of sampling and the size of neighboring and have proposed several methods

for improving the accuracy of classifying. The best results obtained have been reported to be 99%. In (Ying, 2010), in order to filtering Spams a compound classifying made up of the three support vector Machine, Decision Tree and Back Propagation neural network in addition to Majority Voting for combination of applied results have been applied.

In this research, the Spam features such as lacking email address, lacking of email amplitude and etc. have been applied to classify Spams. The best results obtained have been reported to be 91%. In (Su, 2010), used the advantages of the two methods of the Decision Tree, Neural Network and Neural Tree have been applied to classify Spams. In this research, the letter Header has been used and 38 features such as the date of letter sending, the topic of the letter and etc.have been extracted to classify letters. The best result obtained have been reported to be 99%.

In this article, first, we have focused on extracting words as features then we have introduced a method consisting of several feature selection methods and have evaluated the influence of reducing feature space dimensions and finally we have used Adaboost algorithm to learn about the system.

In section 2, we have presented pre-processing stages and word selection. In section 3, Feature Selection method has been described. In section 4, the applied learning algorithm has been presented and the obtained results have been shown in section 5 and at last, the conclusion has been mentioned in section 6.

II. Preprocessing stages

Presenting documents is an important part in filtering process, or generally, the text classification. In Spam filtering, the texts usually are extracted from the message body although; it is also possible to use the topic or even the message header fields in this regard. One of the most famous displaying method is Bag-Of-Word which can be named as Vector-Space. In this method, there is a set of words obtained from the documents ti in which each d document is shown as a dimention N vector from these phrases, $\vec{x} = [x_1, x_2, \ldots, x_N]$, that the amount of xi depends on the number of word occurrence ti in the d document. Suppose we have a document for instance as follow:

This is a document.

The set of extracted features from this document is as:

{this, is, a, document}

In this research, we have used the Bag-Of-Words method in which, first, we have extracted all the words in the training documents then we have omitted the common words and those words, which have been occurred many times, then each document has been shown as a vector of these features. The amount of each feature for each document is equal to the number of occurrence of that feature in that document.

III. Feature selection

One of the main chracter or the problem of Spam filtering is its high dimension of space feature. The feature space that contains words or phrases in the documents has more than ten thousands features, which is a great preventive problem for many of the machine learning algorithms. For this reason, we need a reducing stage of dimensions.

Journal of Academic and Applied Studies
Vol. 2(3) March 2012, pp. 25-31
Available online @ www.academians.org
ISSN1925-931X

The main aim of reducing dimension is to reduce the vector space without losing the efficiency of the classification. There are many techniques in this regard. In the text categorization, the commonest used method is feature selection. Based on (Blum, 1997), the feature selecting methods are divided into three sets: Embedded, Wrapper, and Filter. In the first approach, the feature selecting process has been embedded in the result base algorithm. The wrapper approach selects a subset of features by means of evaluating function in learning algorithm and these features are used in a similar algorithm. In the third method, that is, filter, a subset of features is selected by means of evaluating function, which is independent from the learning method. The filter method is the most popular and the quickest method calculatedly.

In this article, a compound method has been presented for this stage, which is a combination of the two methods of filtering Documents Frequency(DF) and Information Gain(IG). This compound method is done serial and after extracting of the features, first, the DF method is run and those sets of features, which have a few numbers of occurrences, are deleted and the remaining features are given to the next filter. In the next filter, an efficient filter is used which has shown useful results in many researches and has kept the related features and has deleted the remaining features which have no influence in classification on the bases of informative utility. The figure (1) illustrates the running process of this stage.

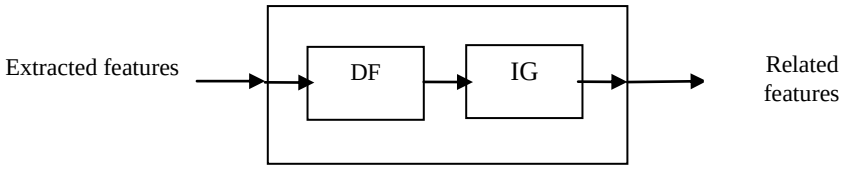

Figure. 1. process of feature selection.

These two used methods in this stage, will be explained in this section

A. Documents Frequency

In this method, the numbers of documents, which have the same feature, are calculated for each feature and those features for which this amount is less than the threshold extent, which was defined in advance, are deleted from the feature space.

B. Information Gain

The information gain is a information theoretical method is data, which mostly is used in machine learning. This method is based on the amount of data that a feature by itself can classify for a system to measure. The amount of greatness calculated for a feature shows the effective amount of that feature. In the issues of binary classification like Spam filtering, the information gain of a feature ti is calculated as below (Yang, 1997):

$$IG(t_i) = \sum_{t\in\{t_i,\bar{t}_i\}} P(t,s)\log\left(\frac{P(t,s)}{P(t)P(s)}\right) + \sum_{t\in\{t_i,\bar{t}_i\}} P(t,h)\log\left(\frac{P(t,h)}{P(t)P(h)}\right) \quad (1)$$

That s and h show the set of Spam and the set of legitimate mails respectively. $\bar{t}_i$ shows that the feature ti has not occurred. P(t,s) shows the probable occurrence of the feature t and

belongs to the set of Spam. P(t,h) shows the probable occurrence of feature t that can occur and belongs to the set of legitimate mails. P(s) and P(h) show the probable of being Spam and legitimate respectively and P(t) shows the probable occurrence of feature t.

IV. Classification

Many researches have shown that the combining of classifiers can improve the total accuracy of the system. Ensemble methods are groups of machine learning methods that cooperate and increase the total results of the classifier system. A general survey of the various methods of ensemble classifier construction together with definition about them will be presented in (Dietterich, 2000).The boosting and bagging methods are classified in set (Neumayer, 2006). The bagging method combines the results of instructed classifiers on the sampling subsets. The boosting method gives a weigh to each sample in the training sets. The sample weigh are updated based on effective classifier after each course of train. At first, the weigh are equal but after each course, the samples weigh, which have been classified wrongly, increase so that the classifier will be forced to focus on hard samples in the training sets (Tulyakov, 2008). A certain sort of algorithm is Adaboost algorithm. This algorithm produces the set of classifiers and their weighs. The changing of samples weigh is based on previous classifier function. The aim is to reduce the final classifier errors (Freud, 1996). Three criteria such as precision, recall and F1 are used in the proposed system for evaluating the performance of the classifier. Precision (P_s) is the number proportion of messages that have been classified correctly in Spam sets to the total number of messages recognized as Spam. Recall (r_s) is the total number proportion of recognized messages as Spam to the total number of messages, which really belong to Spam sets.

$$r_s = \frac{n_{s,s}}{n_{s,s} + n_{s,h}}, \quad p_s = \frac{n_{s,s}}{n_{s,s} + n_{h,s}} \tag{2}$$

$$F1 = \frac{2 r_s p_s}{p_s + r_s} \tag{3}$$

That $n_{s,s}$ is the number of Spams that have been recognized correctly. $n_{s,h}$ is the number of Spams that have been recognized as legitimate and $n_{h,s}$ is the number of legitimate mails which have been classified as Spams. The F1 messure is a combination of recall and precision criteria. So, each classifier that has a high F1 indicates that it has a good performance.

V. The experiments and the results obtained

In this section, the evaluation of the obtained results is presented. In subsection 5-1 the details of the applied dataset is presented. In the next subsection, the results of performing algorithm have been mentioned and compared for each of the feature selection method

A. Dataset and the extracted features

The data used in this article, are the dataset available in LingSpam (LingSpam, 2000). This set consists of 2412 legitimate messages and 481 Spams. The needed features of the system that are in fact the same words are recognized after applying the pre-processing actions. The

number of extracted words from this dataset is 23094 words, which are regarded as the features of the system.

B. feature selection and result evaluation

After recognizing the features and gaining the vector space of the regarded data set, the feature selection stage is done and the related features are recognized in relation to the rest of feature, which have more ability of classifying. In this regard, the features are given to the first filter i.e. DF and the number of 658 features is remained after omitting those features that rarely appear in the dataset. Then this number of features is given to the next filter i.e. IG and after performing this method repeatedly, the number of errors suitable for the features are gained by try and error . With a look at table I which shows the results of performing this filter with different number of final features and by applying a simple learned alogorithm, we come to this conclusion that the number of 50 features will be suitable.

TABLE I
THE COMPARISON OF THE NUMBER OF THE FEATURES

FP rate	Time(second)	Number of feature
0.01	0.33	25
0.006	**0.45**	**50**
0.006	0.95	100
0.006	1.79	200
0.007	2.62	300
0.007	3.48	400
0.007	4.48	500

With regard to the results in table 1, the number of 50 features is suitable. The more the features are, the more will be the time of the model construction with this number of features and the higher will be the number of emails that have been Spams wrongly.

In section three, from among the ten thousands features extracted, 50 features are chosen then we regard them together with the samples as instructional data of learned alogarithm.

In the next stage, we calculate the regarded criteria with different repetitions by AdaBoost algorithm. The table II shows the value of criteria of recall, precession and F1 for the repeated numbers of algorithm.

TABLE II
RECALL ،PRCESION AND F1

Time(s)	Recall	Precesion	F1	repeated numbers of algorithm
0.08	0.833	0.694	0.758	1
0.45	0.972	0.972	0.971	10
1.64	0.982	0.982	0.982	40

2.53	**0.983**	**0.983**	**0.983**	**60**
4.15	0.983	0.983	0.983	100
8.67	0.985	0.984	0.984	200

With regard to the shown results and the amount of time for performing each repetition number of algorithm, the repetition number 60 is suitable.
The table III shows the results of the advantage of Adaboost algorithm in relation to other comparable methods among several performed methods.

TABLE III
ADABOOST COMPARED WITH SEVERAL OTHER METHOD

F1	algorithm
0.983	**AdaBoost**
0.968	SVM
0.974	Decision Tree
0.962	MultiLayerPerceptron
0.971	K-Nearest Neighbour

VI. Conclusions

In this article, a system has been proposed to spam filtering based on Adaboost combined algorithm and a combined method based on several filters has been used to select the effective features. With regard to the obtained results, we can come to this conclusion that the use of a combined method is an effective solution to filtering and selecting the best features.
Together with it and by comparing the obtained results from running of several Adaboost learning algorithms, which is a combined method to instruct the data, the results will be better.

References

Blum A L, Langley P (1997), Selection of relevant features and examples in machine learning, Artificial Intelligence, Vol(97), pp. 245-271.
Chang M, Poon C K (2009), Using phrase as features in email classification, The Journal of Systems and Software, Vol(82), pp. 1036-1045.
Dietterich T G (2000), Ensemble methods in machine learning, Lecture Notes in Computer Science, pp. 1–15.
Freund Y, Schapire R (1996), Experiments with a New Boosting Algorithm, Proc. of 13th Int. Conf. on Machine Learning, Bari, Italy, pp. 148‑156.
Guzella T S, Caminhas T M (2009), "A review of machine learning approaches to Spam filtering", Expert Systems with Application, Vol(36), pp. 10206-10222.
LingSpam public corpus, (2000), <http://www.aueb.gr/users/ion/publications.html>, [visited on july 2009].
Neumayer R (2006), Clustering based ensemble classification for Spam filtering, Proceedings of the 6th Workshop on Data Analysis. Elfa Academic Press, pp. 11-22.
Tulyakov S, et al (2008), Review of Classifier Combination Methods, Studies in Computational Intelligence (SCI), Vol(90), pp. 361‑386.
Yang y, Pedersenl J O(1997), A comparative study on feature selection in text categorization, Proceeding of the Fourteenth International Conference on Machine learning, pp. 412-420.

Ying K C, et al (2010), An ensemble approach applied to classify spam e-mails, Expert Systems with Application, Vol(37), pp. 2197-2201.
Su M C, et al (2010), A neural tree and its application to spam e-mail detection, Expert Systems with Applications, Vol(37), pp. 7976-7985.