J. Basic. Appl. Sci. Res., 3(1s)633-636, 2013
© 2013, TextRoad Publication
ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

The Feature Selection and Dimensionality Reduction Methods for Email Classification

Nayer TalebiBeyrami[1] and MohammadReza FeiziDerakhshi[2]

[1]Department of Computer, Shabestar Branch, Islamic Azad University, Shabestar, Iran
[2]Department of Computer, university of Tabriz, Tabriz, Iran

Received: June 10 2013
Accepted: July 16 2013

ABSTRACT

According to the increasing growth of using electronic mails, the spam emails has became a important problem. Today, there are many methods to detect and filter unwanted emails. The filtering techniques apply machine learning algorithms to classify. In email classification, usually the email term is used as a feature, but most of the methods face with the high volume problem of features' dimensions. There are a lot of methods to select feature and reduce feature dimensionality. In this paper, we will introduce some of the feature selection methods to classify emails.
KEYWORDS: Email classification, Dimensionality Reduction, Feature Selection, Text Categorization, Spam filter.

1. INTRODUCTION

Emails is one of the fastest and cheapest communication ways that today it has became the part of communication means of millions of people. Against these advantages, some unwanted emails have been created which are called spam. Spam emails lead to time consuming and waste of network resources and bandwidth. Many methods have been purposed to automatically classification of emails, among which the machine learning algorithms have had the best efficiency. The machine learning algorithms have a lot of applications in text categorization [1].

The following are the phases in designing a email filter [2]:

- Developing Email corpus.
- Splitting the content into words and special characters (Tokenization).
- Stemming.
- Feature Selection
- Classification

After feature selection, machine learning algorithms like Nave Bayesian, SVM, Decision Tree, Neural Networks and etc. can be used to email classification. Usually email words are used as feature for classification, but the high number of feature is a great obstacle for most learning methods. The appropriate feature selection solves this problem mostly. There are so many methods for feature selection, but a few of them are used for high dimensionality of the feature space in texts classification problems, where this represents this fact that most of these methods are not suitable for feature selection with high dimensionality. In feature selection methods, the aim is to reduce the features' space without losing the precision of classification [6]. Therefore, we will study the feature selection methods and related works. In this review, in section two, we define the spam emails, in section three, feature selection methods and in section four we will present the conclusion.

2. DEFINITION OF SPAM

The Spam Track at the Text Retrieval Conference (TREC) defines mail spam as [3]:

"Unsolicited, unwanted email that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the recipient."

Spammers use different types of techniques to send spam emails and pass filtering. A spam email may include text, image, virus, Trojan and etc, for example it is possible that the image is used in spam, in this method, usually text and letters in an image with GIF or JPEG formats are stored and pass through filtering by this method. The modern technology attempts to read the hidden texts within images. In some cases, spammers try to send empty emails or it may be some malware or viruses are embedded in them [4].

3. FEATURE SELECTION

As previously mentioned, to increase the efficiency of classifiers, the features of the terms must be selected. The feature selection methods are classified into two groups of filtering methods and wrapper methods. The filtering methods weighs features and then select the features based on sum or maximum weights. In wrapper methods the features are selected based on their effect on improving the classification [6].

The filtering methods are classified into two types of supervised methods and unsupervised methods, where the information gain (IG) method, mutual information (MI) method and X^2 Static (CHI) method which will be introduced in the following are included in supervised feature selection methods and document frequency (DF) method and Term

Corresponding Author: Nayer Talebibeyrami, Department of Computer, Shabestar Branch, Islamic Azad University, Shabestar, Iran, Email address: Talebi.beyrami@yahoo.com

frequency (TF) method are included in unsupervised feature selection methods. There is a threshold in whole methods where the features are selected based on this threshold [5].

In filtering methods, in general we could use two positive and negative correlations among features and class in computing. If feature f exists in document x and document belongs to class c (positive correlation), if feature f does not exist in document x and the document belongs to class c (negative correlation), if feature f exists in document x and document does not belong to class c (negative correlation) and if feature f does not exist in document x and document does not belong to class c (positive correlation) [6].

3.1. Document Frequency (DF)

Document frequency is the number of document in with a term occurs. We computed the document frequency for each unique term in the training corpus [7].

3.2. TF-IDF

One weighting method to measure the words weight is TF-IDF method. In this method the weight of t_i in document j is calculated due to the ratio of word frequency in email (TF_{kj}) and inversion of document frequency, the equation (1) represents the criterion of this method [8].

$$a_{ij} = \frac{TF_{ij} * Log \frac{|D|}{DF_i}}{\sqrt{\sum_k \left(TF_{kj} * Log \frac{|D|}{DF_k} \right)^2}} \qquad (1)$$

Ren (2010) [8] has used TF-IDF weighting method for selection of features in emails classification and since in email classification the subject of email is important as a field, so higher weight is assigned to the selected features from the email subject and then the features are selected based on words and their weight and the given threshold and the selected features from content and email subject are combined. At last, SVM algorithm has been used for classification. Fig. 1 represents the structure of used method. To study the results, the dataset of TREC05p-1, TREC06p and TREC07p has been used where the precision values of 98.8307%, 99.6414% and 99/6327% have been obtained, respectively [8].

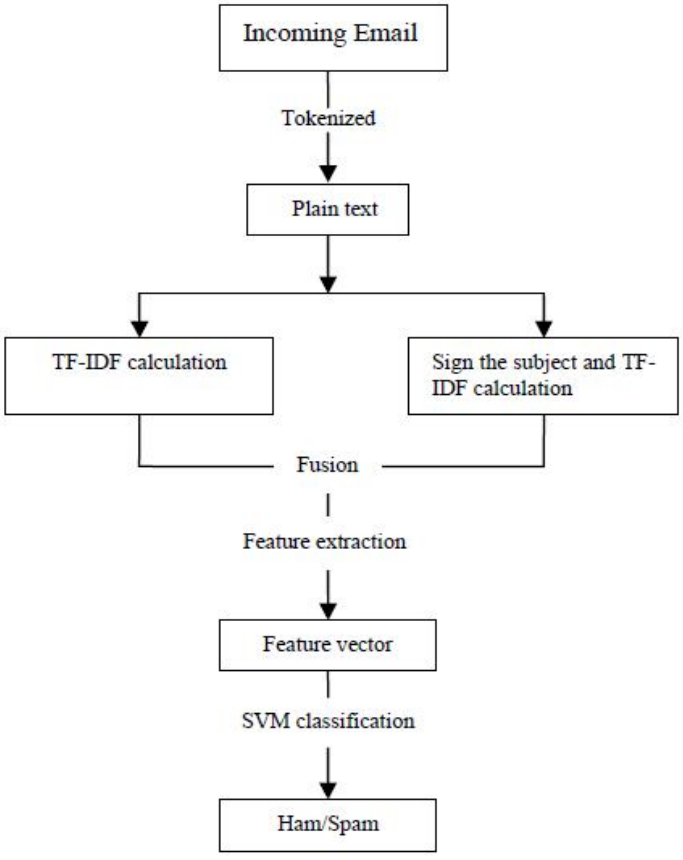

Fig 1. The structure of the combination method of context features and email subject for email classification [8].

3.3. DIA association factor

The DIA association factor criterion for word t of class c is as equation (2) [1].

$$DIA(t_k, c_i) = P(c_i \mid t_k) \quad (2)$$

To measure the score of a word for feature selection, the sum or maximum value of DIA can be combined.

3.4. Information Gain (IG)

The information gain has been used frequently in text features selection. In this method, two positive $P(t,c)$ and negative $P(\bar{t},c)$ correlation have been applied. If c and t show class and the feature, respectively, the criterion of this method is as equation (3) [6].

$$IG(t,c) = P(t,c) * Log \frac{P(t,c)}{P(t) * P(c)} + P(\bar{t},c) * Log \frac{P(\bar{t},c)}{p(\bar{t}) * P(C)} \quad (3)$$

where in this equation, $P(t,c)$ is a possibility value in which feature t has appeared in class c and $P(\bar{t},c)$ is the possibility value in which feature t has not appeared in class c and $p(\bar{t}), P(C), P(t)$ represent the occurrence possibility of class c and occurrence and not occurrence possibility of feature t, respectively [6].

Beiranvand , Osareh and Shadgar (2012) [9] have used compound method of DF and IG for feature selection. The compound method is done serial and has presented good results. Fig. 2 shows the process of this compound method [9]:

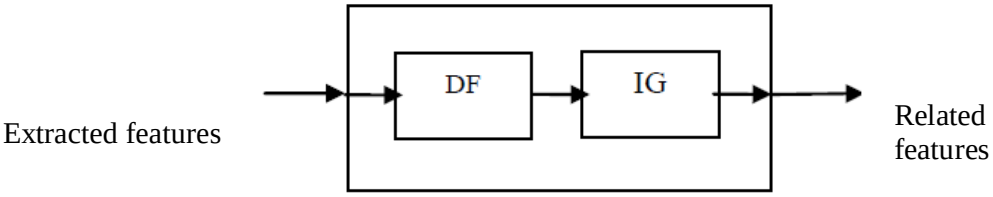

Fig 2. process compound of DF and IG for feature selection [9].

The presented combinational method over LingSpam dataset including 2412 non-spam emails and 481 spam emails has been studied and it has represented the precision of 98.2% in 1.64 seconds with Adaboost classifier [9].

3.5. Mutual Information (MI)

The mutual information method sometimes is called pointwise mutual information. Mostly, MI is applied in statistical modeling where its criterion is as equation (4). MI method uses only positive correlation $P(t,c)$ [1].

$$MI(t,c) = Log \frac{P(t,c)}{P(t) * P(C)} \quad (4)$$

If t and c are independent, the value will be MI=0. Sometimes, IG is called MI, which causes confusion. It is probably because IG is the weighted average of $MI(t,c)$ and $MI(\bar{t},c)$, where the weights are the joint probabilities $P(t,c)$ and $P(\bar{t},c)$, respectively. For each word t in class c, the value of MI is calculated and finally, its average or maximum is considered.

Chuan , Xianliang , Mengshu and Xu (2005) [10] has used MI method for emails features selection and at last the maximum value has been given as a mutual information of that word based on equation (5):

$$MI_{\max}(t) = \max_{i=1}^{m} MI(t, c_i) \quad (5)$$

To study the results of MI over Spamassasin dataset, Naïve Bayesian, ANN-BP and ANN-LVQ algorithms have been applied that presented 97.63% , 98.42% and 98.97% precision.

3.6. X^2 statistic (CHI)

CHI method applies information of every four correlation in weighing features. The criterion of this method is as equation (6), where in this equation $|M|$ is the total number of emails [1].

$$CHI(t,c) = \frac{|M| * \left[P(t,c) * P(\bar{t},\bar{c}) - P(t,\bar{c}) * P(\bar{t},c)\right]^2}{P(t) * P(\bar{t}) * P(c) * P(\bar{c})} \quad (6)$$

Yanh and O.Pedersen (1997) [7] introduce CHI method as equation (7):

$$X^2(t,c) = \frac{N * (AD - CB)^2}{(A+C) * (B+D) * (A+B) * (C+D)} \quad (7)$$

where *N* is the number of documents, *A* is the number of documents of class *c* containing the term *t*, *B* is the number of documents of other class (not *c*) containing *t*, *C* is the number of documents of class *c* not containing the term *t* and *D* is the number of documents of other class not containing t [7].

3.7. Relevance Score (RS)

This method ,measures the relation between presence of t in class c and absence of t in class other than c. The criterion of this method is as equation (8):

$$RS(t,c) = Log\frac{P(t,c)+d}{P(\bar{t},\bar{c})+d} \quad (8)$$

where d in this equation is a constant damping factor (A.Almeida, T. , Almeida, J. and Yamakami, A., 2011).

A.Almeida , Almeida and Yamakami (2011) [1] have studied the filtering methods for feature dimensionality reduction, to study the results different types of Bayes classifiers have been applied where experiments have been done on six well-know Enron datasets. MCC criterion has been applied to evaluate the results. Equation (9) represents MCC evaluation criterion [1]:

$$MCC = \frac{(|TP|*|TN|)-(\lambda|FP|*|FN|)}{\sqrt{(|TP|+\lambda|FP|)*(|TP|*|FN|)*(|TN|+\lambda|FP|)*(|TN|*|FN|)}} \quad (9)$$

where, in this equation, we have $\mathrm{TP} = \mathrm{True\ Positive}$, $\mathrm{TN} = \mathrm{True\ Negative}$, $\mathrm{FP} = \mathrm{False\ Positive}$ and $\mathrm{FN} = \mathrm{False\ Negative}$, λ is the independency ratio.

The MCC return a real value between +1 and -1, if it is +1, then the prediction is perfect, 0, an average random prediction and if MCC equals to -1, the prediction has been done inversely (A.Almeida, T. , Almeida, J. and Yamakami, A., 2011).

In studies, $\mathrm{DIA}_{\max}$ has represented the best result for Enron1 dataset with MCC=0.885, for Enron 2 data set, $\mathrm{OR}_{\max}/\mathrm{OR}_{\mathrm{sum}}$ method with MCC=0.952 had the best prediction, for Enron 3 and Enron 4 dataset, the best algorithm was GI. $\mathrm{OR}_{\max}/\mathrm{OR}_{\mathrm{sum}}$ method with MCC=0.970 and $\mathrm{OR}_{\mathrm{wsum}}$ with MCC=0.92 had the best efficiency for datasets of Enron 5 and Enron 6, respectively (A.Almeida, T. , Almeida, J. and Yamakami, A. , 2011).

4. Conclusion

In this article, briefly we have introduced feature selection methods for emails classification. The related words represent that the presented methods may show different results according to datasets and classification method. Results show that among feature selection techniques, the document frequency, IG and X^2 statistic methods have had the best effect on eliminating the words without losing the precision of classification. In turn, the MI methods have had the weakest result in classification.

REFERENCES

1. A.Almeida, T. , Almeida, J. , Yamakami, A, Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers, J Internet Serv appl, 2011. p. 183-200.
2. Srikanth, K. , Ramakrishna, S. , Sarma, K. V. S, An improved statistical filter for spam detection combining bayesian method and regression analysis, International Journal of Information Technology and Management, 2012. 5(1). p. 169-175.
3. V.Cormack, G, Email Spam Filtering: A Systematic Review, 2008. 1(4).
4. Saad, O. , Darwish, A. , Faraj, R, A survey of machine learning techniques for spam filtering, International Journal of Computer Science and network Security, 2012. 12(2).
5. Liu, T. , Liu, Sh. , Chen, Zh, An evaluation on feature selection for text clustering, Proceeding of the Twentieth International Conference on Machine Learning (ICML-2003), Washington, 2003.
6. Jalili, S. , Bitarafan, M, Increase the efficiency of text categorization based on the improved feature selection method,، University College of Engineering , 2006. 4(3), p. 313-328.(In Persian).
7. Yanh, Y. , O.Pedersen, J, A comparative study feature selection in text categorization. Proceedings of the Fourteenth International Conference on Machine Learning (ICML), Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 1997. p. 412-420.
8. Ren, Q, Feature-Fusion Framework for Spam Filtering Based on SVM. CEAS-Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference, Washington, US, 2010.
9. Beiranvand, A. , Osareh, A. , Shadgar, B, Spam Filtering By Using a Compound Method of Feature Selection, Journal of academic and applier studies, 2012. 2(3).
10. Chuan, Zh. , Xianliang, L. , Mengshu, H. , Xu, Zh, A LVQ-based neural network anti-spam email approach. ACM, New York, NY, USA, 2005. 39(1), p. 34-39