On Feature Extraction for Spam E-Mail Detection

Conference Paper *in* Lecture Notes in Computer Science · September 2006

DOI: 10.1007/11848035_84 · Source: DBLP

4 authors, including:

Serkan Gunal
Anadolu University
31 PUBLICATIONS **447** CITATIONS
SEE PROFILE

Omer N. Gerek
Anadolu University
168 PUBLICATIONS **1,437** CITATIONS
SEE PROFILE

On Feature Extraction for Spam E-Mail Detection

Serkan Günal[1], Semih Ergin[1], M. Bilginer Gülmezoğlu[1], and Ö. Nezih Gerek[2]

[1] Eskişehir Osmangazi University,
The Department of Electrical and Electronics Engineering, Eskişehir, Türkiye
{SGunal, SErgin, BGulmez}@OGU.edu.tr
[2] Anadolu University,
The Department of Electrical and Electronics Engineering, Eskişehir, Türkiye
ONGerek@Anadolu.edu.tr

Abstract. Electronic mail is an important communication method for most computer users. Spam e-mails however consume bandwidth resource, fill-up server storage and are also a waste of time to tackle. The general way to label an e-mail as spam or non-spam is to set up a finite set of discriminative features and use a classifier for the detection. In most cases, the selection of such features is empirically verified. In this paper, two different methods are proposed to select the most discriminative features among a set of reasonably arbitrary features for spam e-mail detection. The selection methods are developed using the Common Vector Approach (CVA) which is actually a subspace-based pattern classifier. Experimental results indicate that the proposed feature selection methods give considerable reduction on the number of features without affecting recognition rates.

1 Introduction

Electronic mails (e-mails) provide great convenience for communication in our daily work and life. However, in recent years, spam e-mail became a big trouble over the Internet. Spam is an e-mail message that is unwanted – basically it is the electronic version of junk mail that is delivered by the postal service. There are many serious problems associated with increase of spam. When the usage of network resources is considered, spam e-mails consume most of the bandwidth resource. Moreover spam e-mails can also quickly fill-up server storage space, especially at large sites with thousands of users who get duplicate copies of the same spam e-mail. Lastly, the boost in the number of spam e-mails can make the use of e-mail for communication tedious and time consuming. As a result, spam is a major concern of governments, Internet Service Providers (ISPs), and end users of internet. Accordingly, automated methods for filtering spam e-mail from legitimate e-mail are becoming necessary. Most of the commercially available filters currently depend on simple techniques such as white-lists of trusted senders, black-lists of known spammers, and hand-crafted rules that block messages containing specific words or phrases [1].

The problems with the manual construction of rule sets to detect spam point out the need for adaptive methods. Currently, the most effective solution seems to be the

B. Gunsel et al. (Eds.): MRCS 2006, LNCS 4105, pp. 635 – 642, 2006.

Bayesian filter which constitutes the main core of many spam filtering software [2]. Applying several layers of filtering consequently improves the overall result of spam reduction [3]. Several types of anti-spam filters were proposed in the literature [1,2,4,5]. These filters try to detect spam e-mails in the network according to the type of e-mail flows and their respective characteristics, using various classifiers. Interesting approaches such as identifying spam at the router level and controlling it via rate limiting exist [6]. A novel anti-spam system which utilizes visual clues, in addition to text information in the e-mail body, to determine whether a message is spam was also developed [7]. Several machine learning algorithms and other recent approaches for this purpose were reviewed in [8] and [9].

Different types and number of features were used in the methods mentioned above. The recognition of spam e-mails with minimum number of features is important in view of computational complexity and time. To achieve this goal, two different feature selection methods are proposed to select the most discriminative features while eliminating irrelevant ones among arbitrarily constructed e-mail feature sets. The feature selection idea depends on a new classification method known as the Common Vector Approach (CVA) which is a successful subspace classifier previously used on different pattern recognition applications [10-12]. The CVA method not only classifies the input e-mail as spam or non-spam, but also provides a measure of how relevant each element on the feature vector is. The recognition performances of the proposed methods are tested using CVA on the widely used SpamAssassin e-mail database [18].

This paper is organized as follows: In the next section, CVA classifier is briefly reviewed and the feature selection methods are presented in Section III. The experimental study and the conclusions are given in Sections IV and V respectively.

2 Common Vector Approach

CVA is a subspace-based classifier utilizing the covariances of the input data, similar to the PCA or KLT. However, unlike CVA, they treat the input data according to a single projection (transformation) by selecting covariance directions that better represents the feature space. CVA, on the other hand, calculates covariances of different classes separately, applies a subspace projection, and chooses better discriminating directions during the projection, instead of the "overall" better representing directions. In CVA, it is suggested that a unique common vector can be calculated for each class if the number of feature vectors (m) in the training set is greater than the dimension (n) of each feature vector (m > n) [13]. Let the column vectors[1] $\mathbf{a}_1^c$, $\mathbf{a}_2^c$,…, $\mathbf{a}_m^c \in R^n$ be the feature vectors for a certain class *C* in the training set. The covariance matrix $\mathbf{\Phi}^c$ of class *C* can be written as

$$\mathbf{\Phi}^c = \sum_{i=1}^{m} (\mathbf{a}_i^c - \mathbf{a}_{ave}^c)(\mathbf{a}_i^c - \mathbf{a}_{ave}^c)^T \quad . \tag{1}$$

[1] For the sake of clarity in the notation, vectors will be referred to in boldface in the following discussion.

where $\mathbf{a}_{ave}^{c}$ is the average vector of the feature vectors in class *C*. Eigenvalue-eigenvector decomposition is then applied to the covariance matrix of training data of class *C* and eigenvalues (λ_j^c) of $\mathbf{\Phi}^c$ are sorted in descending order. The n-dimensional feature space spanned by all eigenvectors can be divided into (k-1) dimensional difference subspace $\mathbf{B}^c$ and (n-k+1) dimensional orthogonal indifference subspace $(\mathbf{B}^{\perp})^c$. The difference subspace $\mathbf{B}^c$ is spanned by the eigenvectors ($\mathbf{u}_j^c$, j=1, 2,…, k-1), corresponding to the largest eigenvalues; and indifference subspace $(\mathbf{B}^{\perp})^c$ is spanned by the eigenvectors ($\mathbf{u}_j^c$, j=k, k+1,…,n), corresponding to the smallest eigenvalues [13]. The purpose for the decomposition of whole feature space into two subspaces is to eliminate the part of the whole space that has large variations from the mean. In contrast to the idea of KLT, the space whose directions contain smaller variations have more common characteristics of the class, therefore they are better suited for classification.

The parameter *k* can be chosen in a way that the sum of the smallest eigenvalues is less than a fixed percentage L of the sum of the entire set [14]. Thus, we let *k* fulfill

$$\frac{\sum_{j=k}^{n}\lambda_j^c}{\sum_{j=1}^{n}\lambda_j^c} < \mathrm{L}. \tag{2}$$

If L = 5%, good performance is experimentally obtained while retaining a small proportion of the variance present in the original space [15].

The orthogonal projection of the average vector $\mathbf{a}_{ave}^{c}$ of class *C* onto the indifference subspace $(\mathbf{B}^{\perp})^c$ gives the common vector of that class, that is,

$$\mathbf{a}_{com}^{c} = \sum_{j=k}^{n}[(\mathbf{a}_{ave}^{c})^{\mathrm{T}}\mathbf{u}_j^c]\mathbf{u}_j^c \cdot \tag{3}$$

where $\mathbf{u}_j^c$'s are the eigenvectors of the indifference subspace $(\mathbf{B}^{\perp})^c$. The projection of the feature vectors of any class onto the indifference subspace will be closer to the common vector of that class. The smaller valued positions within this vector correspond to smaller amount of contribution to the overall discriminant; therefore the effects of coefficients corresponding to those vector positions are less than the effects of the rest. This immediately states the fact that features corresponding to smaller valued common vector element positions can be regarded as less important. In the remaining sections, it is illustrated that the magnitude analysis over common vectors is one of the two methods for selecting useful and useless element positions inside the feature vector.

During classification, the following criterion is used:

$$K^{*} = \underset{1\le c\le N}{argmin} \left\| \sum_{j=k}^{n} \left\{ \left[\left(\mathbf{a_x} - \mathbf{a_{ave}^{c}} \right)^{T} \mathbf{u}_j^c \right] \mathbf{u}_j^c \right\} \right\|^2 \cdot \tag{4}$$

where $\mathbf{a}_x$ is an unknown test vector and N indicates the number of classes. If the distance is minimum for any class C, the observation vector $\mathbf{a}_x$ is assigned to class C.

3 E-Mail Feature Extraction

In recent years, many features were determined to represent spam e-mails [16,17]. In this study, we initially propose 140 features composed of both relatively discriminative features and several relatively irrelevant features. The mentioned features are

Table 1. List of features

1	Picture	**36**	'make'	**71**	'actor'	**106**	'lunatic'
2	Link	**37**	'meeting'	**72**	'amplitude'	**107**	'marble'
3	!', '?', '$', '#', '(', '[', ';'	**38**	'money'	**73**	'array'	**108**	'mentor'
4	'address'	**39**	'offer'	**74**	'balance'	**109**	'monkey'
5	'adult'	**40**	'order'	**75**	'beacon'	**110**	'nuclear'
6	'all'	**41**	'original'	**76**	'blue'	**111**	'outrage'
7	'bank'	**42**	'our'	**77**	'bubble'	**112**	'peripheral'
8	'best'	**43**	'over'	**78**	'chamber'	**113**	'potter'
9	'business'	**44**	'paper'	**79**	'clerk'	**114**	'punch'
10	'call'	**45**	'parts'	**80**	'coherent'	**115**	'render'
11	'casino'	**46**	'people'	**81**	'concert'	**116**	'revenue'
12	'click'	**47**	'pm'	**82**	'designate'	**117**	'spirit'
13	'conference'	**48**	'price'	**83**	'disposal'	**118**	'steak'
14	'credit'	**49**	'project'	**84**	'doubt'	**119**	'stunt'
15	'cs'	**50**	'promotion'	**85**	'drum'	**120**	'tape'
16	'data'	**51**	'quality'	**86**	'eagle'	**121**	'thread'
17	'dear'	**52**	're'	**87**	'egg'	**122**	'tomb'
18	'direct'	**53**	'receive'	**88**	'episode'	**123**	'tripod'
19	'edu'	**54**	'regards'	**89**	'evident'	**124**	'twinkle'
20	'email'	**55**	'remove'	**90**	'furious'	**125**	'upgrade'
21	'fast'	**56**	'report'	**91**	'flesh'	**126**	'utility'
22	'font'	**57**	'sincerely'	**92**	'fault'	**127**	'vibration'
23	'free'	**58**	'spam'	**93**	'friction'	**128**	'violence'
24	'george'	**59**	'table'	**94**	'gadget'	**129**	'vocal'
25	'hello'	**60**	'take'	**95**	'germ'	**130**	'vulture'
26	'here'	**61**	'technology'	**96**	'gun'	**131**	'wedge'
27	'hi'	**62**	'telnet'	**97**	'hammer'	**132**	'wheel'
28	'how'	**63**	'thank'	**98**	'highway'	**133**	'wolf'
29	'hp'	**64**	'think'	**99**	'humid'	**134**	'wrap'
30	'internet'	**65**	'valuable'	**100**	'indigo'	**135**	'yacht'
31	'investment'	**66**	'we'	**101**	'intiment'	**136**	'yield'
32	'lab'	**67**	'will'	**102**	'interrupt'	**137**	'zoo'
33	'let'	**68**	'x'	**103**	'kite'	**138**	'zip'
34	'low'	**69**	'you'	**104**	'lavender'	**139**	'women'
35	'mail'	**70**	'your'	**105**	'lick'	**140**	'weird'

listed in the Table 1. The occurrence count of each feature inside an e-mail is used as an element of 140 dimensional feature vector of that e-mail, and the constructed feature vector is normalized by the total text size. In the first step, the whole set of features is used. In the next stage, less important features are discarded using two different feature selection methods which are detailed in the following subsections. In order to check the efficiency of the feature selection methods, the CVA classifier is performed using the selected features over the same e-mail set.

3.1 Feature Selection Method 1

In this method, the absolute values of the common vector elements are calculated for each class. Based on the consideration of indifference subspace projection, and justified by extensive testing, it is observed that elements of the common vector, which are low-in-magnitude, correspond to relatively irrelevant features compared to the features corresponding to elements which are high–in-magnitude. In other words, the common vector elements which have large magnitudes correspond to more common, hence representative, properties of respective class. Therefore, since the elements of the common vector that have small values carry relatively small information, their use in classification is redundant.

In the feature selection process, labeling a feature vector index as useless according to the results obtained for one class is not appropriate. One useless feature for one class may be quite critical for the expression of the other class. Therefore, features can only be eliminated if they prove to be redundant for both of the classes. The experimental study in Section 4 shows practical results of eliminating irrelevant elements from a larger set of features using this method. In this study, the redundancy analysis is carried out for both spam and non-spam classes, and common redundant features are eliminated. As expected, the irrelevant tagged features are intuitively justifiable. Besides, the classification performance does not deteriorate.

3.2 Feature Selection Method 2

In the second proposed feature selection method, features are selected according to the observations that:

i) Elements with small magnitudes along directions of eigenvectors corresponding to small eigenvalues of the covariance matrix do not contribute to discrimination.

ii) Elements with larger magnitudes along directions of the eigenvectors corresponding to large eigenvalues of the covariance matrix are irrelevant to discrimination.

The statement in (i) indicates that a feature corresponding to small absolute valued elements must be ineffective in obtaining similarity between the input and the correct class common vector, and the statement in (ii) indicates that a feature which produces a large distance with the "unused" projection elements are irrelevant in the classification. The distributions of magnitudes of the values obtained from these criteria are given in Figures 1 (a) and (b) for both spam and non-spam classes. In combination of these criteria, the elements which have maximum difference between the values obtained from above two criteria carry more important information. Therefore the features corresponding to these elements are considered as the most discriminative features.

4 Experimental Study

The widely used SpamAssassin e-mail database [18] is used in our experiments. This database contains wide variety of spam emails that guarantees the proposed feature selection methods to be suitable for different types of e-mails. The number of e-mails in the training and test sets is defined as 2000 and 750 respectively for both spam and non-spam e-mail classes. Initially, CVA classification is performed based on the feature set without eliminating less discriminative features. In the next stage, features tagged as irrelevant by the first feature selection method are eliminated. The discarded features are given in Table 2. Most of the features in Table-2 are also intuitively irrelevant. This result verifies that the proposed feature selection method works in accordance with the intuition for detecting discriminative features. Then the classification procedure is repeated with the lower-dimensional new feature vectors.

Table 2. List of features discarded by feature selection method 1

11	'casino'	**83**	'disposal'	**103**	'kite'	**122**	'tomb'
45	'parts'	**84**	'doubt'	**104**	'lavender'	**123**	'tripod'
49	'project'	**85**	'drum'	**106**	'lunatic'	**124**	'twinkle'
65	'valuable'	**86**	'eagle'	**107**	'marble'	**126**	'utility'
72	'amplitude'	**88**	'episode'	**108**	'mentor'	**127**	'vibration'
73	'array'	**89**	'evident'	**110**	'nuclear'	**128**	'violence'
74	'balance'	**90**	'furious'	**111**	'outrage'	**129**	'vocal'
75	'beacon'	**91**	'flesh'	**112**	'peripheral'	**130**	'vulture'
77	'bubble'	**93**	'friction'	**113**	'potter'	**131**	'wedge'
79	'clerk'	**94**	'gadget'	**115**	'render'	**132**	'wheel'
80	'coherent'	**98**	'highway'	**117**	'spirit'	**133**	'wolf'
81	'concert'	**99**	'humid'	**118**	'steak'	**135**	'yacht'
82	'designate'	**101**	'intiment'	**119**	'stunt'	**136**	'yield'

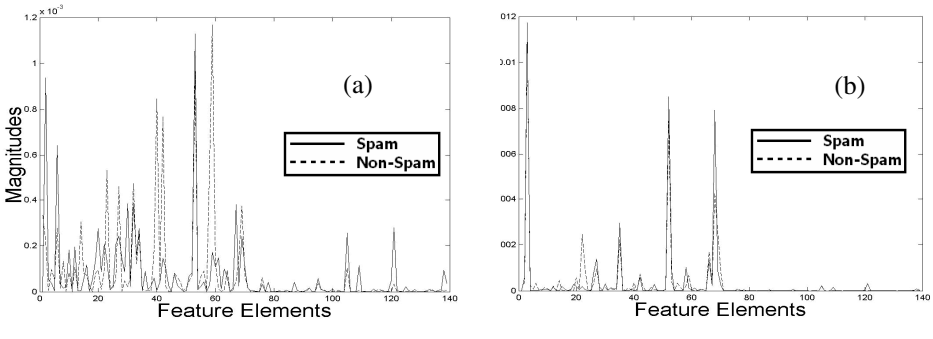

Fig. 1. Dimension element magnitudes for eigenvectors corresponding to (a) small, and (b) large eigenvalues

When the feature selection method 2 is used, the situation is almost identical with only two extra discarded features ("meeting", "sincerely") in addition to those tagged by the first method. The proposed feature selection method 2 selects the discriminative features from initial set more strongly as compared to method 1.

Recognition results obtained from the original and discriminative feature sets are given comparatively in Table 3. It is evident from the results that the proposed methods are successful at detecting irrelevant features from a large feature set.

Table 3. Recognition results obtained from the original and discriminative feature sets

	Original Feature Set	New Feature Set by Method 1	New Feature Set by Method 2
Feature Vector Dimension	140	88	86
Spam Recognition Rate	90%	90%	92%
Non-Spam Recognition Rate	98%	98%	97%

5 Conclusion

The problem imposed by spam e-mails is obvious. Therefore automatic prevention or filtering of spam e-mails is essential for users and ISPs. Although the features used in e-mail classification widely vary among different approaches in the literature, classification with a small set of discriminative features is preferred in view of processing complexity. In this paper, two feature selection methods are proposed to determine the most discriminative features. The CVA classifier used in this work not only provides successful detection results, but also establishes a measure of how well each element within the feature vector performs. In our experiments, 88 and 86 discriminative features are selected from 140 features for the first and second feature selection methods respectively using CVA. It was observed that spam recognition rate obtained by these features is equal or slightly better than those obtained using 140 features for considered e-mail database. This result points out that the proposed feature selection methods are substantially deterministic about detection of discriminative features.

As a future work, different variations of the feature selection methods will be developed to detect more discriminative features in a comparative manner. The effects of feature selection methods will be tested with performance analysis over different classification algorithms such as Bayesian filtering, SVM, etc.

References

1. Qiu X., Jihong H., Ming C., “Flow-Based Anti-Spam”, Proceedings IEEE Workshop on IP Operations and Management, pp. 99 – 103, 2004.
2. Pelletier, L., Almhana, J., Choulakian, V., “Adaptive Filtering of SPAM”, Proceedings of Second Annual Conference on Communication Networks and Services Research, pp. 218 – 224, 2004.
3. Sahami, M., S. Dumais, D. Heckerman ve E. Horvitz, “A Bayesian Approach to Filtering Junk E-Mail”, Proc. of AAAI–98, Workshop on Learning for Text Categorization, Madison, WI, 1998.

4. Michelakis E., I. Androutsopoulos, G. Paliouras, G. Sakkis ve P. Stamatopoulos, "Filtron: A Learning-Based Anti-Spam Filter", Proc. of the 1st Conf. on E-mail and Anti-Spam (CEAS–2004), Mountain View, CA, 2004.
5. Drucker H. D., D. Wu ve V. Vapnik, "Support Vector Machines for Spam Categorization", IEEE Transactions on Neural Networks, pp. 1048–1054, Vol.10, No.5, 1999.
6. B., Agrawal, Kumar N., Molle, M., "Controlling Spam Emails at the Routers", IEEE International Conference on Communications, Vol. 3, pp. 1588 – 1592, 2005.
7. Ching-Tung W., Cheng K-T., Zhu Q., Wu Y-L., "Using Visual Features for Anti-Spam Filtering", Proceedings of IEEE International Conference on Image Processing (ICIP-2005), Vol. 3, pp. 509 – 512, 2005.
8. C-C., Lai, Tsai M-C., "An Empirical Performance Comparison of Machine Learning Methods for Spam E-mail Categorization", Proceedings of Fourth International Conference on Hybrid Intelligent Systems, HIS 2004, pp. 44 – 48, 2004.
9. X-L., Wang, Cloete, I., "Learning to Classify Email: A Survey", Proceedings of 2005 International Conference on Machine Learning and Cybernetics, Vol. 9, pp. 5716 – 5719, 2005.
10. Gülmezoğlu MB, Dzhafarov V, Keskin M, Barkana A (1999) A novel approach to isolated word recognition. IEEE Trans. on Speech and Audio Processing 7: 620-628
11. Gülmezoğlu MB, Dzhafarov V, Barkana A (2001) The common vector approach and its relation to the principal component analysis. IEEE Trans. on Speech and Audio Processing 9: 655-662
12. Çevikalp H, Neamtu M, Wilkes M, Barkana A (2005) Discriminative common vectors for face recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 27: 1-10
13. Gülmezoğlu MB, Dzhafarov V, Barkana A (2000) Comparison of the Common Vector Approach with the other subspace methods when there are sufficient data in the training set. In: Proc. of 8th National Conf. on Signal Processing and Applications. Belek, Turkey, June 2000, pp 13-18
14. Oja E (1983) Subspace methods of pattern recognition, John Wiley and Sons Inc., New York.
15. Swets DL, Weng, J (1996) Using discriminant eigenfeatures for image retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence 18: 831-836
16. Vaughan-Nichols, S. J., "Saving Private E-mail", IEEE Spectrum Magazine, August 2003, pp.40-44.
17. Günal, S., Ergin, S., Gerek, Ö. N., "Spam E-mail Recognition by Subspace Analysis", INISTA – International Symposium on Innovations in Intelligent Systems and Applications, pp. 307-310, 2005
18. I. Katakis, G. Tsoumakas, I. Vlahavas, "On the Utility of Incremental Feature Selection for the Classification of Textual Data Streams", 10th Panhellenic Conference on Informatics (PCI 2005), P. Bozanis and E.N. Houstis (Eds.), Springer-Verlag, LNCS 3746, pp. 338-348, Volos, Greece, 11-13 November, 2005.