Term Space Partition Based Ensemble Feature Construction for Spam Detection

Guyue Mi[1,2], Yang Gao[1,2], and Ying Tan[1,2(✉)]

[1] Key Laboratory of Machine Perception (MOE),
Peking University, Beijing 100871, China
[2] Department of Machine Intelligence,
School of Electronics Engineering and Computer Science,
Peking University, Beijing 100871, China
{gymi,gaoyang0115,ytan}@pku.edu.cn

Abstract. This paper proposes an ensemble feature construction method for spam detection by using the term space partition (TSP) approach, which aims to establish a mechanism to make terms play more sufficient and rational roles by dividing the original term space and constructing discriminative features on distinct subspaces. The ensemble features are constructed by taking both global and local features of emails into account in feature perspective, where variable-length sliding window technique is adopted. Experiments conducted on five benchmark corpora suggest that the ensemble feature construction method far outperforms not only the traditional and most widely used bag-of-words model, but also the heuristic and state-of-the-art immune concentration based feature construction approaches. Compared to the original TSP approach, the ensemble method achieves better performance and robustness, providing an alternative mechanism of reliability for different application scenarios.

Keywords: Term space partition (TSP) · Ensemble term space partition (ETSP) · Feature construction · Spam detection · Text categorization

1 Introduction

Email has been an important communication tool in our daily life. However, high volumes of spam emails severely affect the normal communication, waste resources and productivity, and threat computer security and user privacy, resulting in serious economic and social problems [1]. According to Symantec Internet Security Threat Report [2], the overall spam rate of the whole email traffic all over the world in 2015 is 53 %. Meanwhile, email remains an effective medium for cybercriminals, since one out of every 220 emails contains malware. Statistics from Cyren Cyber Threat Report [3] also reveal that the average amount of spam sent per day in 2015 is up to 51.8 billion. Thus, taking measures to solve the spam problem is necessary and urgent.

Y. Tan and Y. Shi (Eds.): DMBD 2016, LNCS 9714, pp. 205–216, 2016.
DOI: 10.1007/978-3-319-40973-3_20

Machine learning based intelligent detection methods give promising performance in solving the spam problem, which can be seen as a typical binary categorization task. Machine learning techniques have been widely applied in spam classification, such as naive bayes [4–7], support vector machine [8–11], decision tree and boosting [12,13], k nearest neighbor [14–16], random forest [17,18], artificial neural network [19–22], deep learning [23,24] and so on. Besides classification technique, feature construction approach also plays an important role in spam detection, by transforming email samples into feature vectors for further utilization by machine learning methods. Feature construction approach determines the space distribution of email samples, affecting the establishment of classification model and detection performance. Effective feature construction approach could construct distinct and distinguishable features, resulting in different space distribution characteristics of different classes of email samples and complexity reduction of email classification. Research on email feature construction approaches has been focused in recent years.

In our previous work, a term space partition (TSP) based feature construction approach for spam detection [25] is proposed by taking inspiration from the distribution characteristics of terms with respect to feature selection metrics and the defined class tendency. Term ratio and term density are constructed on corresponding subspaces achieved by dividing the original term space and compose the feature vector. In this paper, we further propose an ensemble TSP (ETSP) based feature construction method for spam detection by taking both global and local features of emails into account in feature perspective. Variable-length sliding window technique is adopted for constructing local features. We conducted experiments on five benchmark corpora PU1, PU2, PU3, PUA and Enron to investigate performance of the proposed method. Accuracy and F_1 measure are selected as the main criteria in analyzing and discussing the results.

The rest of this paper is organized as follows: Sect. 2 introduces the TSP based feature construction approach. The proposed ETSP method is presented in Sect. 3. Section 4 gives the experimental results. Finally, we conclude the paper in Sect. 5.

2 Term Space Partition Based Feature Construction Approach

The TSP approach aims to establish a mechanism to make the terms play more sufficient and rational roles in spam detection by dividing the original term space into subspaces and designing corresponding feature construction strategy on each subspace, so as to improve the performance and efficiency of spam detection.

Since the feature selection metrics could give terms reasonable and effective goodness evaluation, the TSP approach first performs a vertical partition of the original term space to obtain the dominant term subspace and general term subspace according to the distribution characteristics of terms with respect to feature selection metrics. Dominant terms are given high and discriminative scores by feature selection metrics and considered to lead the categorization

results. Though large amount of general terms congregate in a narrow range of the term space with similar low scores and each of them is less informative, they could contribute to spam detection integrally. The vertical partition could be performed by defining a threshold θ_{dg} with respect to the corresponding feature selection metrics employed, as shown in Eq. 1.

$$\theta_{dg} = \frac{1}{r}(\tau_{max} - \tau_{min}) + \tau_{min} \tag{1}$$

where τ_{max} and τ_{min} depict the highest and lowest evaluation of terms in the training set respectively, and variable r controls the restriction level of dominant terms. Term t_i with $\tau(t_i) \geq \theta_{dg}$ is considered as dominant term, and general term otherwise.

To construct discriminative features, a transverse partition is then performed to further divide each of the above subspaces into spam term subspace and ham term subspace according to the defined term class tendency. Term class tendency refers the tendency of a term occurring in emails of a certain class, defined as Eq. 2.

$$tendency(t_i) = P(t_i|c_h) - P(t_i|c_s) \tag{2}$$

where $P(t_i|c_h)$ is the probability of t_i's occurrence, given the email is ham, and $P(t_i|c_s)$ is the probability of t_i's occurrence, given the email is spam. Spam terms are terms that occur more frequently in spam than in ham with negative tendency, and ham terms occur more frequently in ham than in spam with positive tendency. When performing the transverse partition to separate spam terms and ham terms, terms with $tendency(t_i) = 0$ are considered useless and discarded.

In this case, the original term space is decomposed into four independent and non-overlapping subspaces, namely spam-dominant, ham-dominant, spam-general and ham-general subspaces. To construct discriminative and effective feature vectors of emails, term ratio and term density are defined on dominant terms and general terms respectively to make the terms play sufficient and rational roles in spam detection. Term ratio indicates the percentage of dominant terms that occur in the current email, emphasizing the absolute ratio of dominant terms. In this way, the contributions to spam detection from dominant terms are strengthened and not influenced by other terms. While term density represents the percentage of terms in the current email that are general terms, focusing on the relative proportion of terms in the current email that are general terms. The effect on spam detection from general terms is weakened and so is the affect from possible noisy terms. Equations 3 to 6 describe the definitions of spam term ratio, ham term ratio, spam term density and ham term density respectively.

$$TR_s = \frac{n_{sd}}{N_{sd}} \tag{3}$$

where n_{sd} is the number of distinct terms in the current email which are also contained in spam-dominant term space TS_{sd}, and N_{sd} is the total number of distinct terms in TS_{sd}.

$$TR_h = \frac{n_{hd}}{N_{hd}} \tag{4}$$

where n_{hd} is the number of distinct terms in the current email which are also contained in ham-dominant term space TS_{hd}, and N_{hd} is the total number of distinct terms in TS_{hd}.

$$TD_s = \frac{n_{sg}}{N_e} \tag{5}$$

where n_{sg} is the number of distinct terms in the current email which are also contained in spam-general term space TS_{sg}, and N_e is the total number of distinct terms in the current email.

$$TD_h = \frac{n_{hg}}{N_e} \tag{6}$$

where n_{hg} is the number of distinct terms in the current email which are also contained in ham-general term space TS_{hg}. The feature vector is achieved by combining the defined features, i.e. $\boldsymbol{v} =< TR_s, TR_h, TD_s, TD_h >$.

3 Ensemble Feature Construction Using Term Space Partition Approach

3.1 Global and Local Features

For spam detection, feature construction approaches decide the spatial distribution of email samples. Effective feature construction approaches could construct distinguishable features of emails to make the spatial distribution of spam emails apparently different from that of legitimate emails. In the TSP approach, each email sample is transformed into an individual 4-dimensional feature vector, by calculating distribution characteristics of terms in the email on the four independent and non-overlapping subspaces of terms respectively. In other words, this feature vector reflects the term distribution characteristics of the whole email in the four different subspaces of terms, which could be called global features, for each dimension of the feature vector is related to and calculated from the whole email. Global features describe the overall characteristics of each email sample. In most cases, the global features constructed by the TSP approach could successfully characterize the differences between spam emails and legitimate emails. While it should also be noted that, for some specific email samples with particularly different term distribution characteristics in some local areas of the emails, the global features constructed by the TSP approach would make the distinctive features diluted and could not well reflect the differences.

In order to solve this problem, we adopt the sliding window technique to define local areas on the whole email and further extract local features of the email by constructing TSP features on each local area. The local features and global features are combined together to form the ensemble feature vector.

3.2 Construction of Local Features

Local features are constructed on local areas of samples. In the ETSP method, the sliding window is adopted to define local areas of emails. In this case, TSP

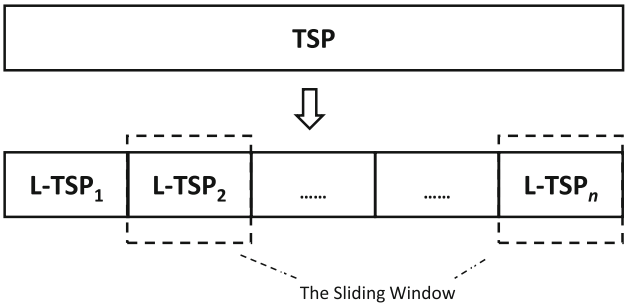

Fig. 1. Construction of local TSP feature with sliding window

features constructed on each local area could reflect independent term distribution features of each local area in the four different subspaces.

As shown in Fig. 1, independent local TSP (L-TSP) feature vector is calculated on each individual local area of the email, other than constructing global TSP feature vector on the whole email. Variable-length sliding windows are adopted to guarantee obtaining feature vectors with the same dimensionality to facilitate further use in the classification phase, for the size of email samples varies greatly. For a specific email with N_t terms, the length of corresponding sliding window utilized is defined as $\frac{N_t}{n}$, where n is constant for different email samples. To obtain independent and non-overlapping local areas, the window slides with a step of length of itself, which is $\frac{N_t}{n}$, from the beginning to the end of the email. In this case, each email is divided into n independent and non-overlapping local areas, and n individual L-TSP feature vectors are obtained. Hence, n is a core parameter during this process, determining both the granularity of local areas and dimensionality of the final feature vectors.

3.3 TSP Based Ensemble Feature Construction

Algorithm 1. Ensemble Feature Vector Construction

1: construct TSP feature vector on the given sample
2:
3: move a sliding window of $\frac{N_t}{n}$ terms over the given sample with a step of $\frac{N_t}{n}$ terms
4:
5: **for** each position i of the sliding window **do**
6: construct TSP feature vector on the current local area
7: **end for**
8:
9: combine the achieved feature vectors together to form the final feature vector

Global features and local features tend to characterize samples from different perspectives, where global features describe the overall characteristics of each sample, while local features presents the local details. Global features and local features should play different but necessary roles in depicting and classifying samples. Therefore, ensemble feature vectors of emails are constructed by calculating TSP features on both the whole email and its local areas, as presented by Algorithm 1. Finally, the global feature vector and the local feature vectors are combined together to form the feature vector of the sample, i.e. $\boldsymbol{v} =< TSP, L-TSP_1, L-TSP_2, \ldots, L-TSP_n >$.

4 Experiments

4.1 Experimental Setup

Experiments were conducted on PU1, PU2, PU3, PUA [26] and Enron-Spam [27], which are all benchmark corpora widely used for effectiveness evaluation in spam detection. Support vector machine (SVM) was employed as classifier in the experiments. WEKA toolkit [28] and LIBSVM [29] were utilized for implementation of SVM. 10-fold cross validation was utilized on PU corpora and 6-fold cross validation on Enron-Spam according to the number of parts each of the corpora has been already divided into. Accuracy and F_1 measure [30] are the main evaluation criteria, as they can reflect the overall performance of spam filtering.

4.2 Investigation of Parameters

Experiments have been conducted on PU1 to investigate the parameters of the ETSP approach by utilizing 10-fold cross validation. Besides the term selection parameter p and partition threshold parameter r in the TSP approach [25], the ETSP method has got an external parameter n, which determines the granularity of local areas those an sample is divided into and further the dimensionality of the corresponding feature vectors. Small n brings coarse-grained local areas and further low dimensionality of feature vectors, which may cause dilution of local features and could not describe the local details well. While large n may lead to incomplete and inaccurate representation of local features due to the meticulous partition of local areas, making the process of extracting local features meaningless.

Figure 2 shows the performance of ETSP under varied n, where information gain is selected as the representative feature selection metric. As is shown, the ETSP method achieves better performance with relatively smaller ns, and performs the best when $n = 2$ happens in the parameter investigation experiments, which meets our expectation well. For the specific problem of spam detection, the vast majority of email samples in the communication traffic are of relatively small lengths, no matter spam or legitimate emails, but with distinctive local characteristics of term distribution, especially spam.

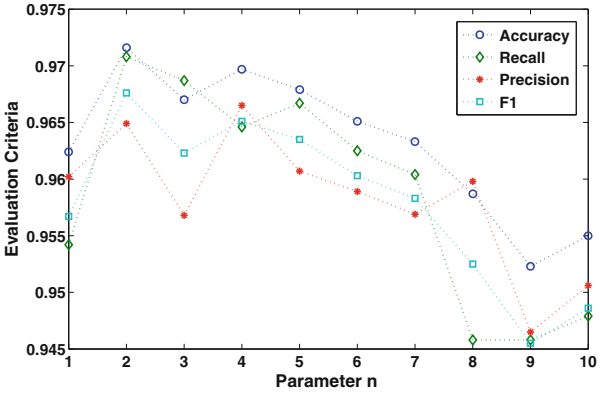

Fig. 2. Performance of ETSP under varied n

4.3 Performance with Different Feature Selection Metrics

In the TSP approach, vertical partition of the original term space is performed according to term evaluation given by feature selection metrics. Selection of appropriate feature selection metrics is also crucial to performance of the ETSP method. We selected document frequency (DF) and information gain (IG) as representatives of unsupervised and supervised feature selection metrics respectively to conduct verification experiments, which are widely used and perform well in spam detection and other text categorization issues [31].

Table 1. Performance of ETSP with different feature selection metrics

Corpus	Feature sel.	Precision (%)	Recall (%)	Accuracy (%)	F_1 (%)
PU1	DF	96.33	97.29	97.16	96.77
	IG	97.28	96.67	**97.34**	**96.95**
PU2	DF	93.95	89.29	**96.62**	**91.29**
	IG	93.87	84.29	95.63	88.23
PU3	DF	96.66	95.66	96.59	96.12
	IG	96.54	97.47	**97.29**	**96.97**
PUA	DF	96.50	96.67	**96.49**	**96.52**
	IG	96.58	94.91	95.70	95.67
Enron-Spam	DF	94.97	98.35	**97.32**	**96.57**
	IG	94.25	98.29	97.02	96.18

Performance of ETSP with respect to DF and IG on five benchmark corpora PU1, PU2, PU3, PUA and Enron-Spam is shown in Table 1. As the experimental results reveal, the ETSP method performs quite well with both DF and

IG, showing good adaptability with different kinds of feature selection metrics. Meanwhile, DF could outperform IG with ETSP as feature construction approach in more cases of the experiments, indicating that the transverse partition of the original term space is effective to make use of the information of term-class associations, as the supervised feature selection metrics provide.

4.4 Performance Comparison with Current Approaches

Experiments were conducted on PU1, PU2, PU3, PUA and Enron-Spam to verify the effectiveness of ETSP by comparing the performance with current approaches. The selected approaches are Bag-of-Words (BoW) [30], concentration based feature construction (CFC) approach [21,32], local concentration (LC) based feature construction approach [9,33] and the original TSP approach [25]. Tables 2, 3, 4, 5 and 6 shows the performance of each feature construction approach in spam detection when incorporated with SVM.

Among the selected approaches, BoW is a traditional and one of the most widely used feature construction approach in spam detection, while CFC and LC are heuristic and state-of-the-art approaches by taking inspiration from biological immune system. LC-FL and LC-VL utilize different local areas definition strategies. As we can see, ETSP far outperforms not only BoW, but also CFC and LC, in terms of both accuracy and F_1 measure. This strongly verifies the effectiveness of ETSP as a feature construction method in spam detection.

Table 2. Performance comparison of ETSP with current approaches on PU1

Approach	Precision (%)	Recall (%)	Accuracy (%)	F_1 (%)
BoW	93.96	95.63	95.32	94.79
CFC	94.97	95.00	95.60	94.99
LC-FL	95.12	96.88	96.42	95.99
LC-VL	95.48	96.04	96.24	95.72
TSP	96.90	96.67	97.16	96.74
ETSP	96.49	97.08	**97.34**	**96.95**

Table 3. Performance comparison of ETSP with current approaches on PU2

Approach	Precision (%)	Recall (%)	Accuracy (%)	F_1 (%)
BoW	88.71	79.29	93.66	83.74
CFC	95.12	76.43	94.37	84.76
LC-FL	90.86	82.86	94.79	86.67
LC-VL	92.06	86.43	95.63	88.65
TSP	94.09	83.57	95.63	88.12
ETSP	93.95	89.29	**96.62**	**91.29**

Table 4. Performance comparison of ETSP with current approaches on PU3

Approach	Precision (%)	Recall (%)	Accuracy (%)	F_1 (%)
BoW	96.48	94.67	96.08	95.57
CFC	96.24	94.95	96.05	95.59
LC-FL	95.99	95.33	96.13	95.66
LC-VL	95.64	95.77	96.15	95.67
TSP	96.37	97.09	97.05	96.69
ETSP	96.54	97.47	**97.29**	**96.97**

Table 5. Performance comparison of ETSP with current approaches on PUA

Approach	Precision (%)	Recall (%)	Accuracy (%)	F_1 (%)
BoW	92.83	93.33	92.89	93.08
CFC	96.03	93.86	94.82	94.93
LC-FL	96.01	94.74	95.26	95.37
LC-VL	95.60	94.56	94.91	94.94
TSP	95.91	96.49	96.05	96.11
ETSP	96.50	96.67	**96.49**	**96.52**

Table 6. Performance comparison of ETSP with current approaches on Enron-Spam

Approach	Precision (%)	Recall (%)	Accuracy (%)	F_1 (%)
BoW	90.88	98.87	95.13	94.62
CFC	91.48	97.81	95.62	94.39
LC-FL	94.07	98.00	96.79	95.94
LC-VL	92.44	97.81	96.02	94.94
TSP	94.29	98.21	97.02	96.14
ETSP	94.97	98.35	**97.32**	**96.57**

Compared with the original TSP, ETSP achieves not only better but also more balanced performance on different corpora in the experiments. This demonstrates that taking both global and local features into account in feature perspective could bring both better performance and better robustness. It is worth mentioning that ETSP could perform better than TSP mainly on some specific email samples with particularly different term distribution characteristics in some local areas of the emails. Thus, the performance improvement of ETSP approach compared with TSP approach could not be dramatically. The ETSP approach could be an alternative implementation strategy of TSP with better robustness.

In the experiments, we conducted parameter investigation on a small corpus and applied the selected group of parameter values on all the benchmark

corpora with different sizes and email sample length distributions utilized for performance verification. From the experimental results, the ETSP method possess good parameter generalization ability and this further endows it with adaptivity in real world applications.

5 Conclusions

In this paper, a term space partition based ensemble feature construction method for spam detection was proposed by taking both global and local features into account in feature perspective. The experiments have shown: (1) utilization of sliding window successfully constructs local features of email samples; (2) the ETSP method cooperates well with different kinds of feature selection metrics and shows good parameter generalization ability, endowing it with flexible applicability in real world; (3) the ETSP method shows better performance and robustness by taking both global and local features into account during spam detection.

Acknowlegements. This work was supported by the Natural Science Foundation of China (NSFC) under grant no. 61375119 and the Beijing Natural Science Foundation under grant no. 4162029, and partially supported by National Key Basic Research Development Plan (973 Plan) Project of China under grant no. 2015CB352302.

References

1. Research, F.: Spam, spammers, and spam control: a white paper by ferris research. Technical report (2009)
2. Corporation, S.: Internet security threat report. Technical report (2016)
3. Cyren: Cyber threat report. Technical report (2016)
4. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A Bayesian approach to filtering junk e-mail. In: Learning for Text Categorization: Papers from the 1998 Workshop, vol. 62. AAAI Technical Report WS-98-05 98–105, Madison (1998)
5. Almeida, T., Almeida, J., Yamakami, A.: Spam filtering: how the dimensionality reduction affects the accuracy of naive bayes classifiers. J. Internet Serv. Appl. **1**(3), 183–200 (2011)
6. Zhong, Z., Li, K.: Speed up statistical spam filter by approximation. IEEE Trans. Comput. **60**(1), 120–134 (2011)
7. Trivedi, S.K., Dey, S.: Interaction between feature subset selection techniques and machine learning classifiers for detecting unsolicited emails. ACM SIGAPP Appl. Comput. Rev. **14**(1), 53–61 (2014)
8. Drucker, H., Wu, D., Vapnik, V.: Support vector machines for spam categorization. IEEE Trans. Neural Netw. **10**(5), 1048–1054 (1999)
9. Zhu, Y., Tan, Y.: A local-concentration-based feature extraction approach for spam filtering. IEEE Trans. Inf. Forensics Secur. **6**(2), 486–497 (2011)
10. Duan, L., Tsang, I.W., Xu, D.: Domain transfer multiple kernel learning. IEEE Trans. Pattern Anal. Mach. Intell. **34**(3), 465–479 (2012)
11. Li, C., Liu, M.: An ontology enhanced parallel SVM for scalable spam filter training. Neurocomputing **108**, 45–57 (2013)

12. Carreras, X., Marquez, L.: Boosting trees for anti-spam email filtering. Arxiv preprint cs/0109015 (2001)
13. DeBarr, D., Wechsler, H.: Spam detection using random boost. Pattern Recogn. Lett. **33**(10), 1237–1244 (2012)
14. Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Sakkis, G., Spyropoulos, C., Stamatopoulos, P.: Learning to filter spam e-mail: a comparison of a naive Bayesian and a memory-based approach. Arxiv preprint cs/0009009 (2000)
15. Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C., Stamatopoulos, P.: A memory-based approach to anti-spam filtering for mailing lists. Inf. Retr. **6**(1), 49–73 (2003)
16. Jiang, S., Pang, G., Wu, M., Kuang, L.: An improved k-nearest-neighbor algorithm for text categorization. Expert Syst. Appl. **39**(1), 1503–1509 (2012)
17. Koprinska, I., Poon, J., Clark, J., Chan, J.: Learning to classify e-mail. Inf. Sci. **177**(10), 2167–2187 (2007)
18. Amin, R., Ryan, J., van Dorp, J.R.: Detecting targeted malicious email. IEEE Secur. Priv. **10**(3), 64–71 (2012)
19. Clark, J., Koprinska, I., Poon, J.: A neural network based approach to automated e-mail classification. In: Proceedings. IEEE/WIC International Conference on Web Intelligence, 2003, WI 2003, pp. 702–705. IEEE (2003)
20. Wu, C.: Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. Expert Syst. Appl. **36**(3), 4321–4330 (2009)
21. Ruan, G., Tan, Y.: A three layer back-propagation neural network for spam detection using artificial immune concentration. Soft Comput. Fusion Found. Methodol. Appl. **14**(2), 139–150 (2010)
22. Li, C.H., Huang, J.X.: Spam filtering using semantic similarity approach and adaptive BPNN. Neurocomputing **92**, 88–97 (2012)
23. Mi, G., Gao, Y., Tan, Y.: Apply stacked auto-encoder to spam detection. In: Tan, Y., Shi, Y., Buarque, F., Gelbukh, A., Das, S., Engelbrecht, A. (eds.) ICSI-CCI 2015. LNCS, vol. 9141, pp. 3–15. Springer, Heidelberg (2015)
24. Gao, Y., Mi, G., Tan, Y.: Variable length concentration based feature construction method for spam detection. In: 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–7. IEEE (2015)
25. Mi, G., Zhang, P., Tan, Y.: Feature construction approach for email categorization based on term space partition. In: The 2013 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE (2013)
26. Androutsopoulos, I., Paliouras, G., Michelakis, E.: Learning to filter unsolicited commercial e-mail. DEMOKRITOS, National Center for Scientific Research (2004)
27. Metsis, V., Androutsopoulos, I., Paliouras, G.: Spam filtering with naive bayes-which naive bayes. In: Third Conference on Email and Anti-spam (CEAS), vol. 17, pp. 28–69 (2006)
28. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: an update. ACM SIGKDD Explor. Newsl. **11**(1), 10–18 (2009)
29. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
30. Guzella, T., Caminhas, W.: A review of machine learning approaches to spam filtering. Expert Syst. Appl. **36**(7), 10206–10222 (2009)
31. Yang, Y., Pedersen, J.: A comparative study on feature selection in text categorization. In: Machine Learning-International Workshop Then Conference, pp. 412–420. Morgan Kaufmann Publishers, Inc. (1997)

32. Tan, Y., Deng, C., Ruan, G.: Concentration based feature construction approach for spam detection. In: International Joint Conference on Neural Networks, 2009, IJCNN 2009, pp. 3088–3093. IEEE (2009)
33. Zhu, Y., Tan, Y.: Extracting discriminative information from e-mail for spam detection inspired by immune system. In: 2010 IEEE Congress on Evolutionary Computation (CEC), pp. 1–7. IEEE (2010)