An Adaptive Concentration Selection Model for Spam Detection

Yang Gao, Guyue Mi, and Ying Tan*

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

Abstract. Concentration based feature construction (CFC) approach has been proposed for spam detection. In the CFC approach, Global concentration (GC) and local concentration (LC) are used independently to convert emails to 2-dimensional or 2n-dimensional feature vectors. In this paper, we propose a novel model which selects concentration construction methods adaptively according to the match between testing samples and different kinds of concentration features. By determining which concentration construction method is proper for the current sample, the email is transformed into a corresponding concentration feature vector, which will be further employed by classification techniques in order to obtain the corresponding class. The k-nearest neighbor method is introduced in experiments to evaluate the proposed concentration selection model on the classic and standard corpora, namely PU1, PU2, PU3 and PUA. Experimental results demonstrate that the model performs better than using GC or LC separately, which provides support to the effectiveness of the proposed model and endows it with application in the real world.

Keywords: Global concentration (GC), local concentration (LC), adaptive concentration selection, spam detection.

1 Introduction

Spam has been a serious problem in the developing of internet. According to the CYREN internet threats trend report, the average daily spam level for the first quarter in 2014 was 54 billion emails per day [1]. Large numbers of spam not only consume many resources online, but also threaten security of the network, especially when they carry viruses and malicious codes. What's more, people usually take much time to handle spam, which reduces efficiency and productivity.

In the fields of spam detection, intelligent detection methods have been the most effective way to examine junk mails. On one hand, the intelligent methods have a higher degree of automation. On the other hand, these methods not only have high precision and strong robustness, but also can fit email content

* Corresponding author.

Y. Tan et al. (Eds.): ICSI 2014, Part I, LNCS 8794, pp. 223–233, 2014.

and users' interests. Now the mainstream of intelligent detection methods can be divided into two categories: machine learning and artificial immune system. Because spam email detection task is a typical classification problem, supervised learning method is general in machine learning fields, such as naive bayes (NB) [2, 3], k-nearest neighbor (KNN) [4,5], support vector machine (SVM) [6], artificial neural networks (ANN) [7, 8] and so on. And in the artificial immune system (AIS), researchers imitate the process of immune cells' recognition to antigen.

In this paper, we propose a model structure, which aims to select concentration methods adaptively. The GC approach transforms each email to a 2-dimensional GC feature vector, which may lose some important information of the email. And the LC approach extracts position-correlated information from each email by mapping it to an 2n-dimensional feature vector, which may get some redundant information. But in our model, we can adjust concentration method adaptively according to distinctive information of different emails.

The remainder of the paper is organized as follows. In Section 2, we introduce the related works. In Section 3, the proposed adaptive concentration selection model is presented in detail. Section 4 gives the detailed experimental setup and results. Finally, we conclude the paper with a detailed discussion.

2 Related Works

This section introduces term selection approaches, concentration-based methods and classifiers that have close relationship with our work.

2.1 Term Selection Approaches

Information Gain. Information gain (IG) [9] is a concept in the information theory, which gives a description of the distance between two probabilities distribution $P(x)$ and $Q(x)$. In the spam detection field, it is utilized to measure the importance of terms. The calculation formula of IG is defined as

$$I(t_i) = \sum_{C \in (C_S, C_L)} \sum_{T \in (t_i, \bar{t_i})} P(T, C) \log \frac{P(T, C)}{P(T)P(C)} \tag{1}$$

where C indicates an email's class (C_S and C_L are the spam and legitimate email classes) and T denotes the whether term t_i appears in the email or not. And all the probabilities are estimated from the whole data set.

2.2 Concentration-Based Methods

Global Concentration. Global concentration (GC) [10, 11] is an approach inspired from the human immune system, which can transform each email to a 2-dimensional feature vector. The flow chart of GC is described in Fig1. The biological immune system is a complex adaptive system, which has its unique self and non-self cells. Similar to this, the concentration approach proposed has

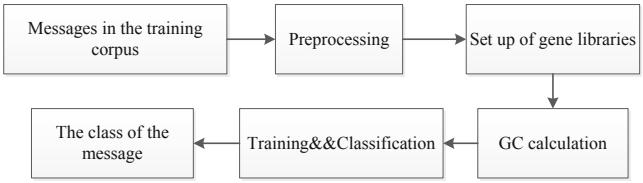

Fig. 1. Construction of GC model

two gene libraries - 'self' and 'non-self' gene libraries. The 'self' gene library is composed of words that present healthy emails. And in contrast, the 'non-self' gene library covers words that can present spam emails. So through the gene libraries, we can calculate global concentration of each email to construct its GC feature vectors.

Local Concentration. Similar to GC, the local concentration (LC) [12, 13] approach also transforms each email to a feature vector. However, the difference between GC and LC is that the LC can provide local information of a document, which can help to 'check' the email microscopically. In the process of LC, it

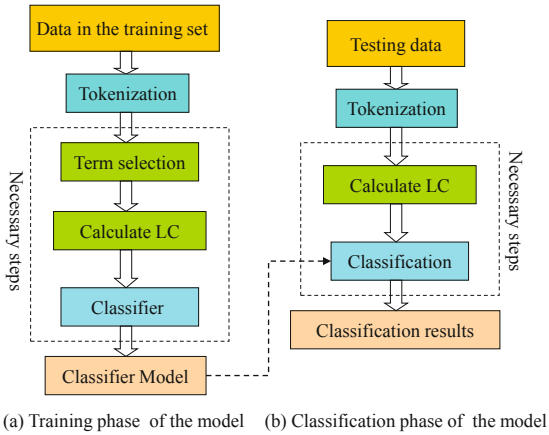

Fig. 2. Construction of LC model

mainly covers two parts: the training part and the testing part. And in both parts, tokenization is the first step to pre-process the documents. Then in the term selection step, it chooses the important terms, which can reflect the emails' tendency to spam or non-spam. After calculating the local concentration of each email, every document is represented by a 2n-dimensional feature vectors. Then the feature vectors are transported to the classifier for training or testing.

2.3 Classifier

K-Nearest Neighbor. K-nearest neighbor (KNN) [14] is a kind of basic classification and regression method, which was proposed by Cover and Hart in 1968. The central idea of KNN is that when a new testing case is fed to the classifier, we look for k cases that are nearest to the testing case, and the testing case is classified as the class that those k cases belong to. KNN can be defined as follows

$$\mathrm{y} = \arg\max_{\mathrm{C_j}} \sum_{\mathrm{X_i} \in \mathrm{N_k(x)}} \mathrm{I}\left(\mathrm{y_i} = \mathrm{c_j}\right), \mathrm{i} = 1, 2, \ldots, \mathrm{N}; \mathrm{j} = 1, 2, \ldots, \mathrm{K} \tag{2}$$

where $I(y_i = c_j)$ is a indicator function, with the value of 1 when $y_i = c_j$, and 0 otherwise, and $(y_i) \in \Upsilon = (c_1, c_2, \ldots, c_k)$. And the special situation that the k is set to 1, KNN degrades to nearest neighbor.

3 Adaptive Concentration Selection Model

3.1 Overview of Our Proposed Model

In global concentration method, we transform an email into a 2-D feature vector, which reflects the global information of the email. Similarly, we use local concentration method to reflect emails' local information. However, global concentration may be too simple to cover some 'necessary' information and local concentration may cover some 'unnecessary' information. As a result, we propose the adaptive concentration selection model to transform emails into global or local feature vectors adaptively, according to their contents.

Our method can be mainly divided into four steps. (1) Set up 'self' and 'non-self' gene library from training emails. (2) Generate global and local concentration vectors of each email, using the gene library. (3) Judge that which concentration method each email should apply. (4) Train and classify on the corpora. In this paper, we use KNN to calculate the evaluation which is the reference standard of concentration selection method.

3.2 Set Up of Gene Libraries

Intuitively, if a word appears mostly in spam emails, it belongs to the 'non-self' gene library largely. Accordingly, a word which can provide more information for spam emails than non-spam emails usually will be put into the 'non-self' gene library, and vice versa. This inspires us to calculate information gain of each word, and sort them in a decent order. Considering the amount of words is too big to build gene library, and most documents contain the same common words, we also discard 95% of the words that appear in all emails, just as the paper does [10].

Algorithm 1. Generation of gene libraries

1. Initialize gene libraries, detector DS_s and DS_l to the empty
2. Initialize tendency threshold θ to predefined value
3. Tokenization about the emails
4. **for** each word $t_k$separated **do**
5. According to the term selection method, calculate the importance of t_k and the amount of information $I(t_k)$
6. **end for**
7. Sort the terms based on the $I(t)$
8. Expand the gene library with the top $m\%$ terms
9. **for** each term t_i in the gene library **do**
10. **if** $\|P(t_i|c_l) - P(t_i|c_s)\| > \theta$, $\theta \geq 0$ **then**
11. **if** $P(t_i|c_l) - P(t_i|c_s) < 0$ **then**
12. add term t_i to the spam detector set DS_s
13. **else**
14. add term t_i to the legitimate detector set DS_l
15. **end if**
16. **else**
17. abandon this term, because it contains little information about those emails
18. **end if**
19. **end for**

Algorithm 2. Construction of feature vectors based on global concentration

1. **for** each term t_j in the email **do**
2. calculate the matching $M(t_j, DS_s)$ between term t_j with spam detector set;
3. calculate the matching $M(t_j, DS_l)$ between term t_j with legitimate detector set
4. **end for**
5. According to 3, calculate the concentration of spam detector set SC;
6. According to 4, calculate the concentration of legitimate detector set LC;
7. Combine the above concentration values to construct the global concentration feature vectors $< SC, LC >$

3.3 Construction of Feature Vectors Based on the Immune Concentration

After we have got the gene library, we can construct the feature vectors. According to the generation of detector set, it is obvious that the DS_s can match spam emails and the DS_l can match the legitimate emails with large probability. As a result, the match between two detector sets and emails can reflect the class information of emails, and the two detector sets have complementary advantages with each other, which provides a guarantee for the effectiveness of detection.

$$SC_i = \frac{\sum_{j=1}^{\omega_n} M(t_j, DS_s)}{N_t} \tag{3}$$

where N_t is the number of distinct terms in the window, and $\mathrm{M}(t_j, DS_s)$ is the matching function which is used to measure the matching degree of term t_j and detector DS_s.

$$LC_i = \frac{\sum_{j=1}^{\omega_n} M(t_j, DS_l)}{N_t} \tag{4}$$

where $\mathrm{M}(t_j, DS_l)$ is the matching function which is used to measure the matching degree of term t_j and detector DS_l.

Algorithm 3. Construction of feature vectors based on local concentration

1. According to the length of each email and preset number of windows to calculate the value of ω_n
2. Move the ω_n-term sliding window to separate the email, with each moving length being ω_n
3. **for** each moving window **do**
4. **for** each term in the moving window **do**
5. calculate the matching $M(t_j, DS_s)$ between term t_j with spam detector set;
6. calculate the matching $M(t_j, DS_l)$ between term t_j with legitimate detector set;
7. **end for**
8. According to 5, calculate the concentration of spam detector set SC_i;
9. According to 6, calculate the concentration of legitimate detection set LC_i
10. **end for**
11. Combine local concentration values in each sliding window to construct the local concentration feature vector $< (SC_1, LC_1), (SC_2, LC_2), \ldots, (SCn, LCn) >$

$$SC_i = \frac{\sum_{j=1}^{\omega_n} M(t_j, DS_s)}{N_t} = \frac{\sum_{j=1}^{\omega_n} \sum_{d_k \in DS_s} M(t_j, d_k)}{N_t} \tag{5}$$

$$LC_i = \frac{\sum_{j=1}^{\omega_n} M(t_j, DS_l)}{N_t} = \frac{\sum_{j=1}^{\omega_n} \sum_{d_k \in DS_l} M(t_j, d_k)}{N_t} \tag{6}$$

3.4 Implementation of Our Model

Global concentration reflects entire features of emails and the local concentration reflects local characteristics. However, the GC lacks some detailed information and the LC separates the emails quite meticulously. As a result, we propose our model to combine their advantages and make up for their disadvantages. The key point of our model is the evaluation which is used to determine concentration methods. In our paper, we use KNN to calculate the evaluation. As we all know, the main idea of KNN is to count the numbers of neighbors belonging to different kinds of classes. However, if the numbers of different classes are close, it is hard to judge which class the undetermined point belongs to. So in our model, we take use of this characteristic of KNN and adapt it to determine concentration methods.

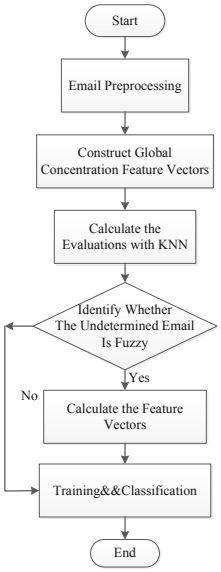

Fig. 3. Implementation of our model

Firstly, after preprocessing, we convert all data to GC feature vectors and use KNN classifier to evaluate them. During the evaluation, if the number of a particular class, which belongs to the neighbors of a undetermined point, is larger than a certain proportion, we can classify the point to this class. But if the number is less than the proportion, we consider this point as a fuzzy one. Secondly, for those fuzzy points, we convert them to LC feature vectors which can reflect their details and evaluate them with KNN classifier again. Thirdly, we manipulate all classification results and assess them with precision, recalls, accuracy and F_1 measure.

4 Experiments

4.1 Experimental Setup

In this paper, experiments were conducted on PU series email corpora, which contains PU1, PU2, PU3 and PUA and were collected and published by Androutsopoulos [15] in 2004. The PU series email corpora were widely utilized in spam detection related experiments. To ensure the objectivity, all the experiments are organized with 10-fold cross validation. At the stage of classification, we choose the KNN method to verify the spam and legitimate emails. Besides, we use recalls, precision, accuracy and F_1 measure to assess the results. Among them, the F_1 measure is taken as the most important evaluating indicator, for its reflection of the recalls and precision. All experiments were conducted on a PC with Intel P7450 CPU and 2G RAM.

4.2 Experiments of Parameter Selection

Proportion of term selection. In the term selection stage, we choose top $m\%$ of the terms according to their information quantity, which decides the size of the gene library. When we screen the terms, on one hand, we need to cut off those noise terms, and on the other hand, the important terms should be held back. In the practical application, this parameter can be adjusted based on the need of time and space complexity.

According to the paper written by Zhu [13], when the parameter m is set to 50%, the performance of experiments can achieve optimal. Therefore, the value of m is set to 50% in our experiments.

Tendency threshold. Tendency function is mainly used to measure the difference between the terms and the two kinds of emails and add corresponding terms to the related detector. In Zhu's paper [13], with the increasing of the tendency threshold θ, the whole performance of the algorithm degrades. As a result, the value of θ is set to 0.

Dimension of feature vectors. In the global concentration method, each email is reconstructed with the self and non-self concentration, which means the dimension is two. And in the local concentration method, this paper adopts variable length sliding window strategy, which means that if we assume N is the number of sliding window, each email is transformed into an $2N$-dimensional feature vector. In this paper, we set the parameter N to 3, according to [10]. As a result, the dimension of local concentration method is 6.

Parameter k in KNN. We have done some experiments to determine the value of parameter k. And the results are shown as follows. As mentioned above, the PU2 is a corpus containing only English emails and the PUA contains not only English emails but also other languages. So the experiments on these two corpora reflect general characteristics. Besides, we find that different experiments based on different values of parameter k perform similarly And as we all know, if the value of k is set too large, the computation complexity will increase. Consequently, we choose a moderate value, which sets the value of k to five.

4.3 Experiments of the Proposed Model

In this paper, we conducted comparison experiments of the model with selection method IG and mainly compares the performance among GC, LC and our model. These experiments are mainly conducted on corpora PU1, PU2, PU3 and PUA using 10-fold cross-validation. The average performance experiments are reported in Table 1 to Table 4.

Compared to GC and LC, the proposed adaptive concentration selection model achieves a better performance on the four corpora. Although in the experiment with PU1, the precision and recall indexes are less than GC or LC, the

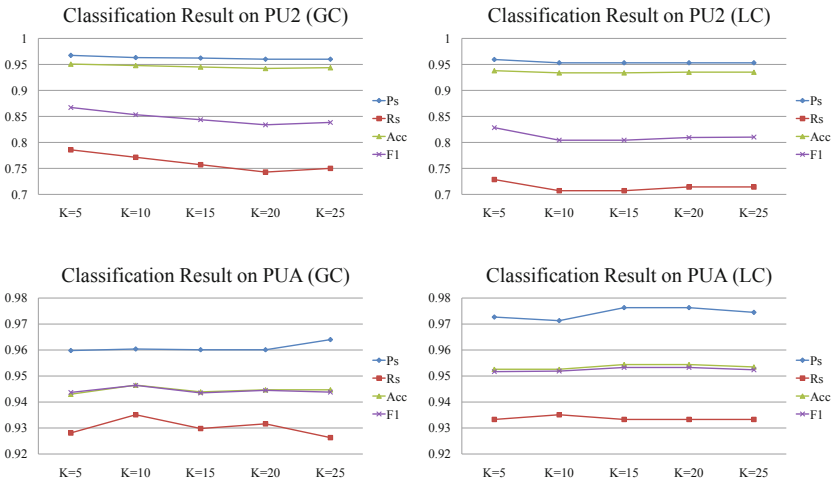

Fig. 4. Classification results on PU2 and PUA

Table 1. Performance of three feature construction methods on PU1

Corpus	Approach	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
PU1	Global Concentration	95.59	94.37	95.60	94.97
	Local Concentration	96.54	92.92	95.41	94.69
	Adaptive Concentration	96.18	94.17	**95.78**	**95.16**

Table 2. Performance of three feature construction methods on PU2

Corpus	Approach	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
PU1	Global Concentration	96.74	78.57	**95.07**	**86.71**
	Local Concentration	95.95	72.86	93.80	82.83
	Adaptive Concentration	96.74	78.57	**95.07**	**86.71**

Table 3. Performance of three feature construction methods on PU3

Corpus	Approach	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
PU1	Global Concentration	96.14	93.57	95.40	94.84
	Local Concentration	96.95	92.86	95.47	94.86
	Adaptive Concentration	96.78	94.07	**95.91**	**95.41**

overall evaluation index, F_1 measure, is better than GC and LC. And we can still conclude that our model performs better on this corpus.

As a result, we can come to a conclusion that the proposed model combines the advantages of GC and LC, and it can enhance the experimental effects so as to classify emails more precisely.

Table 4. Performance of three feature construction methods on PUA

Corpus	Approach	Precision(%)	Recall(%)	Accuracy(%)	F_1(%)
PU1	Global Concentration	95.98	92.81	94.30	94.37
	Local Concentration	97.27	93.33	**95.26**	95.17
	Adaptive Concentration	97.44	94.21	94.65	**95.79**

4.4 Discussion

We have proposed our model for adaptively taking use of concentration methods' feature construction characteristics. The improvement of the model can be explained with the defects of GC and LC. Although GC approach extracts global information of emails into 2-dimensional feature vectors, it may miss some information because of its rough data processing. To the contrary, LC processes data in detail, which may be too excessive to retain some noise terms. By contrast, our proposed model first uses GC feature vectors to evaluate data, and divide all data into two parts: certain classes and fuzzy ones. For those fuzzy ones, the proposed model further takes use of the detailed information based on LC feature vectors and finally we get better performance according to the experimental results. Generally speaking, the model combines both advantages of GC and LC, and avoids large computational complexity of only LC method.

5 Conclusion

In this paper, we present a spam filtering system that combine GC and LC feature construction methods that further makes the system adaptive to different emails. In the stage of feature extraction, we use IG to estimate terms' importance and concentration methods to transform emails into reconstructed feature vectors. And in the classification, according to different characteristics of emails, the system adaptively chooses feature construction methods and the performance is promising.

In the future, we intend to convert emails into variable length future vectors according to the length of emails' messages and study its performance.

Acknowlegements. This work was supported by the National Natural Science Foundation of China under grants number 61375119, 61170057 and 60875080.

References

1. CYREN: Internet threats trend report: April 2014. Tech. rep. (2014)
2. 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, pp. 98–105. AAAI Technical Report WS-98-05, Madison (1998)
3. Ciltik, A., Gungor, T.: Time-efficient spam e-mail filtering using n-gram models. Pattern Recognition Letters 29(1), 19–33 (2008)

4. 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)
5. Sakkis, G., Androutsopoulos, I., Paliouras, G., Karkaletsis, V., Spyropoulos, C., Stamatopoulos, P.: A memory-based approach to anti-spam filtering for mailing lists. Information Retrieval 6(1), 49–73 (2003)
6. Drucker, H., Wu, D., Vapnik, V.: Support vector machines for spam categorization. IEEE Transactions on Neural Networks 10(5), 1048–1054 (1999)
7. Clark, J., Koprinska, I., Poon, J.: A neural network based approach to automated e-mail classification. In: Proceedings of the IEEE/WIC International Conference on Web Intelligence, WI 2003, pp. 702–705. IEEE (2003)
8. Wu, C.: Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. Expert Systems with Applications 36(3), 4321–4330 (2009)
9. Yang, Y.: Noise reduction in a statistical approach to text categorization. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 256–263. ACM (1995)
10. Tan, Y., Deng, C., Ruan, G.: Concentration based feature construction approach for spam detection. In: International Joint Conference on Neural Networks, IJCNN 2009, pp. 3088–3093. IEEE (2009)
11. Ruan, G., Tan, Y.: A three-layer back-propagation neural network for spam detection using artificial immune concentration. Soft Computing 14(2), 139–150 (2010)
12. 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)
13. Zhu, Y., Tan, Y.: A local-concentration-based feature extraction approach for spam filtering. IEEE Transactions on Information Forensics and Security 6(2), 486–497 (2011)
14. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
15. Androutsopoulos, I., Paliouras, G., Michelakis, E.: Learning to filter unsolicited commercial e-mail. "DEMOKRITOS". National Center for Scientific Research (2004)