Advances in Engineering Research (AER), volume 124
2nd International Symposium on Advances in Electrical, Electronics and Computer Engineering
(ISAEECE 2017)

A Model for Spam Filtering Using Support Vector Machine and Artificial Immune System

Yaping Jiang[1,a], Hao Guo[2, b], Peigen Guo[3,c]

School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Henan Zhengzhou, 450002, China

[a]email:yapingjiang@163.com, [b]email：guohao3456@163.com, [c]email:guopeigen123@163.com

Keywords: Spam, Legitimate, E-mail,Similarity Coefficient

Abstract. Spam is a major problem of the Internet, because it can cause pollution and waste of resources in the network environment. Therefore,spam filtering is necessary. After spam filtering methods study, presents a support vector machine and artificial immune system combines filtration methods and filtration methods using MATLAB simulation experiments.

Introduction

The accuracy of SVM (Support Vector Machine) classifier is superior to other classification technologies while it not only provides a higher performance and accuracy, but also avoids the "dimension of disaster" [1] effectively. The Support Vector Machine uses the support vector to make decisions while it uses the training data set to create the support vector [2].

The linear classifier of the Support Vector Machine reduce the generalization error through the constraints of the decision boundary. The key concept of SVM is Statistical Learning Theory which is mainly used to determine the position of decision boundary. Key Concepts simply supporting vector machines are shown in Fig. 1.

An infinite decision boundaries of the traditional two-stage separation supporting vector machine are selected as support vectors to minimize generalization errors. The closest data points are used to determine unknown classes, which are called support vectors. Whether the two classes are linearly separable or not, the SVM method cleverly solves this problem: SVM applies the relevant mathematical theorem of the kernel function expansion so that there is no need to know the explicit expression of the nonlinear mapping. Because it is a linear learning machine established in the high-dimensional feature space, it is not only less possible to increase the computational complexity compared with the linear model, but also avoids the "dimension disaster" to a certain extent. Choosing different types of kernel functions can generate different SVM, there are the main use of four kernel functions as following:

① Linear kernel function: $k(x, y) = x * y$;

② Polynomial kernel function: $k(x, y) = [(x * y) + 1]^d$;

③ RBF-Radial Basis Function: $(x, y) = \exp(-|x - y|^2/d^2)$;

④ The kernel function layer neural network : $k(x, y) = \tanh(a(x * y) + b)$;

AIS (Artificial Immune System) is a very complex system of recognition for foreign bodies [3]. Principle of biological immune system is significant for spam filtering technology. There are many features in the spam filtering system as same as those in immune system,but these features in other intelligent immune systems are rarely or not available. In essence, the Artificial Immune System uses computer-related techniques to simulate the immune system of the organism, which has a similar fuction to that of the immune system and has the function of finding and removing “Non-self “at the same time. From the perspective of bioimmunology, mail filtering is to distinguish between "self" and "non-self". The so-called "self" is a legitimate mail while "non-self" is illegal mail. Therefore combining the related principle of Artificial Immune System with spam filtering is an important hotspot in the field of information security research now. Theri mapping is shown in table 1.

Filtering model design

In the model, five stages are designed: ①pretreatment of data set; ② feature selection; ③ feature extraction; ④ classification; ⑤ result calculation. The flow chart is shown in Fig.2.

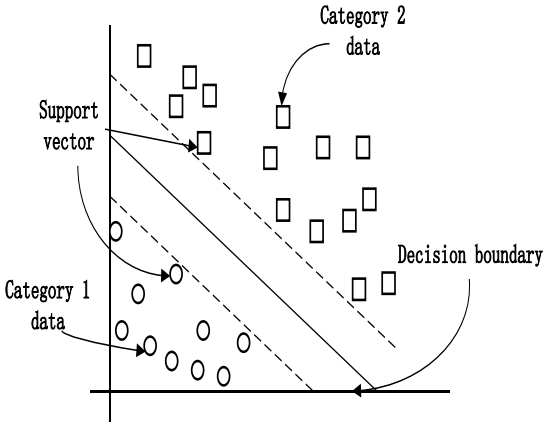

Fig.1 SVM classification overview

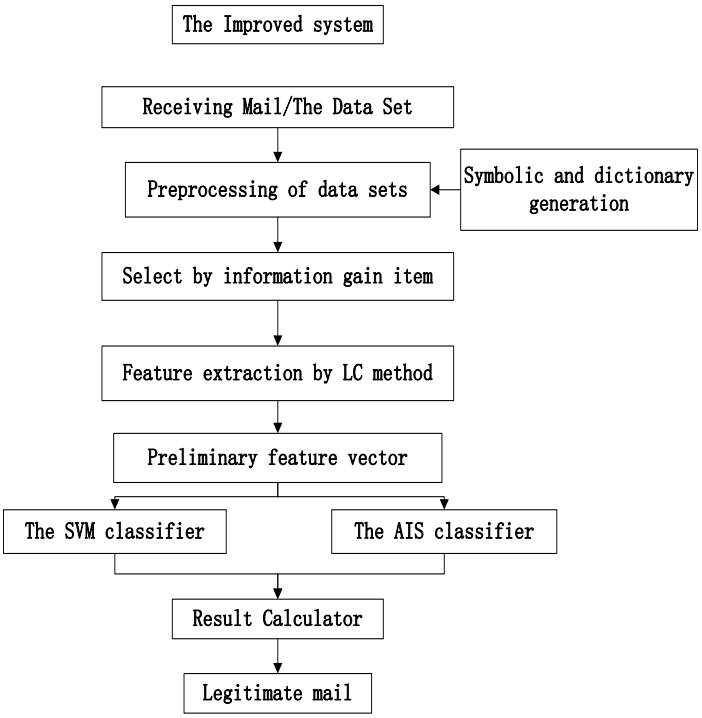

Fig. 2 improved system

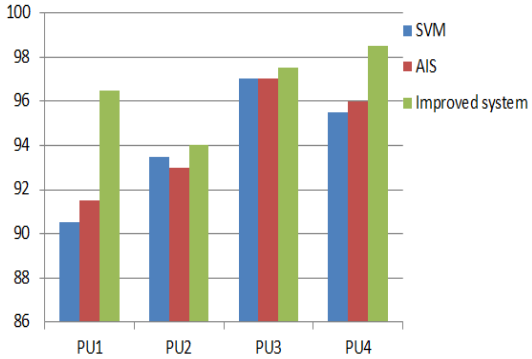

Fig. 3 Comparison of the accuracy of spam filtering methods

Table 1 entity mapping relationship between the biological immune system and the spam filtering system

Biological immune system	Spam filtering system
Antigen	Mail to be tested
Self	Legitimate mail
Nonself	spam
Antibody	Detector
Antigen recognition	Mail classification
Cell affinity	Mail similarity
Memory antibody	Known Spam Detector
Antigen presentation	Mail exception feature extraction
Vaccine	Known exception recovery strategy
B cells T cells Antibody	Antibodies expressed by eigenvectors
Antibody and antigen binding	Pattern matching algorithm
Cytokines	Mail communication system

Pretreatment of data sets. With the date preprocessed this stage, the real-time incoming e-mail preprocesses the data set through the spam filter. By using the string compiler to create a dictionary, some insignificant words are discarded at this stage. Then the data processed by the system is passed to the next stage.

Feature selection. The passing selection strategy in this stage makes a feature extraction for the previous stage and passes it to the feature extraction stage. In order to get a better choice, it is important to choose the right spam filtering strategy.

Feature extraction. At this stage, the spam filtering system analyzes the selected feature extraction conditions and makes a feature extraction through the selected keywords from the local corpus so as to determine whether the e-mail for spam sorting

Classification. At this stage, two classifiers, Support Vector Machine and Artificial Immune System, finish spam filtering through the way of parallel work to achieve higher accuracy and shorter responsive time. However, a system that forms a filter's serial combination may require a higher parallel time ratio.

Calculation results. The result is stored in a binary array of numbers, and the elements stored in the array (0 or 1) specify the result of the classification. With 0 for spam, 1 for legal e-mail and weighted average calculation [5], it assumes that the weight is accurate. Use the following formula for weighted average calculation:

$$M = \frac{\frac{a_1 \times F_1 + a_2 \times F_2}{F_1 + F_2}}{\frac{\beta_1 \times \overline{F_1} + \beta_2 \times \overline{F_2}}{\overline{F_1} + \overline{F_2}}}$$

In the above formula, α is the efficiency of spam, β is the legal e-mail cooperation efficiency, $F_1$and F_2 are corresponding to the SVM filtering rules and AIS filtering rules, M is the meaning of spam.

Simulation Experiments and Analysis

In the experiment using a computer simulation experiment, its CPU frequency is 2.4GHz, memory is 4GB, the operating system environment is Windows7, the experimental software is MATLAB.

We conducted our experiments on four benchmark test corpus PU1, PU2, PU3, PU4. These corpus preprocesses eliminate HTML tags, attachments, and header fields. In the 1099 e-mails of PU1, 480 e-mails are spam and 619 is legal. The 721 e-mails in PU2, 149 are spam and 572 are

legal. In the 4139 e-mails of PU3, 1826 e-mails are spam, 2313 are legal. In the 1142 e-mails of PU4, 572 are spam, 570 are legal. The simulation results are shown in Table 2.

Analysis of the experimental results. The simulation experiment proposed for this model divides date into training set and test set. The average recall rate and the average precision rate are the correct rate and average precision of the model [6]. All filtering methods mainly focus on accuracy. The comparison results of the Support Vector Machine Artificial Immune System are shown in Fig.3.

Table 2 Simulation of experimental results statistics

Method / Data Set	SVN		AIS		Improved system	
	Spam	Legitimate mail	Spam	Legitimate mail	Spam	Legitimate mail
PU1(1099)	434	665	428	671	480	619
PU2(721)	149	572	156	565	149	572
PU3(4139)	1828	2311	1809	2330	1826	2313
PU4(1142)	570	572	569	573	572	570

Conclusion

This paper presents a spam filtering model based on support vector machine and artificial immune system and realizes the simulation experiment of the filter model by using MATLAB. The model can be used to solve some related problems in spam filtering , which is a practical significance.

Acknowledgements

This work was financially supported by the National Natural Science Foundation (No.61272038); Henan Science and Technology Agency-funded science and technology research projects (No.0624220084); Henan Science and Technology Department of Basic and cutting-edge technology projects (NO. 122300410255).

Refereces

[1] Qing J J,Mao R L,Bie R F,et al.An AIS-based E-mail Classification Method[C]. The 2009 International Conference on Intelligent Computing, Ulsan,Korea,2009:492-499.

[2] Secker A,Freitas A,Timmis J. AISEC:an Artificial Immune System for E-mail Classification[C].The 2003 Congress on Evolutionary Computation, California, USA,2003:131-138.

[3] Liu Fengling, Yang Guangming, Wang Xinyan, Liu Ying. Research and application of ARTIS artificial immune model in mail filtering [J]. Small Microcomputer System, 2007,28 (7): 1293-1296.

[4] Liang Gang, Liu Xiaojie, Li Tao, Jiang Yaping, Yang Jin, Gong Xun. NSC: A new type of spam filter [J]. Small microcomputer system, 2008,29 (1): 158-161.

[5] Huang Jue, Chen Bing, Liao Changwu.Improved Artificial Immune Spam Filtering Algorithm [J]. Computer Engineering and Applications, 2011,47 (30): 72-74.

[6] Li Xia, Jiang Shengyi. A fast identification method of spam [J]. Small micro-computer system, 2013,34 (3): 498-502.