A Method of Spam Filtering Based on Weighted Support Vector Machines

CHEN Xiao-li, LIU Pei-yu, ZHU Zhen-fang, QIU Ye
Department of Information Science and Engineering, Shandong Normal University, Ji'nan , 250014, China
cxl86148@163.com; liupy@sdnu.edu.cn; zhuzhfyt@163.com; 1986zx@sina.com

Abstract

The problem of content-based spam filtering on machine learning methods actually is a binary classification. SVMs can separate the data into two categories optimally so SVMs suit to spam filtering. With used into spam filtering, the standard support vector machine involves the minimization of the error function and the accuracy of the SVM is very high, but the degree of misclassification of legitimate emails is high. In order to solve that problem, this paper proposed a method of spam filtering based on weighted support vector machines. Experimental results show that the algorithm can enhance the filtering performance effectively.

1. Introduction

With the development of network and communication technology, E-mail is becoming the most important way of communication among modern people. However, in recent years, spam has become a key problem in electronic communication. Spam has frustrated, confused, and annoyed e-mail users because it can waste time, spread virus and so on.

Today, the approach of content analysis has shown particular promise and generality for combating spam. Nevertheless, spam filtering is different from document classification; it has some characteristics of its own. The most important point is that spam filtering is an imbalance classification problem. Namely, consumers can be tolerant of receiving ten spam, but can not allow that a legitimate e-mail was discarded. Also for each legitimate e-mail, the importance is different, so the cost of their mistake classification is different as well [1].

Currently, the approach of content-based anti-spam filtering mainly has two major types, one is the rule-based, and the other is statistics-based [2].There are many rule-based methods, for example, decision tree, boosting and rough set. The main advantage of rule-based methods is that it can generate rules people can understand. The disadvantage is that the effect of methods was bad, when the regularity is not obvious, the rules are less effective. There are also many methods based on the statistics, such as the filter based on support vector machine, Bayesian Theory and neural networks. Among the filters, the one based on support vector machines is very suitable for the spam filtering, but the standard SVM is to optimize the classification accuracy. The accuracy of the SVM is very high, but the precision is low. Some legitimate emails are misclassified. In order to solve that problem, this paper proposed a method of spam filtering based on weighted support vector machines. Experimental results show that the algorithm can enhance the filtering performance effectively.

2. Standard Support Vector Machine

The Standard Support Vector Machine was proposed by Cortes and Vapnik in 1995，it has many special advantages in solving a finite data set, non-linear and high-dimensional pattern recognition, and it could be expanded to the learning problems of other machines like function fitting and so on. [3]

Based on statistical learning theory of VC dimension and structural risk minimization theory, the standard support vector machine method ,to finite set, find the best compromise between the complexity of the model (namely Accuracy) and learning ability (The capacity to identify any samples error-free), with a view to promote the ability of getting the best generalization. Viewing two sets of vector in an n-dimensional space, a SVM will construct a separating hyperplane in that space, one that maximizes the margin between the two data sets. Intuitively, a good separation distance is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin, the lower the generalization error of the classifier.

When doing the text classification, first of all, giving the training sample set T and assuming that T is a non-linear sub-training set, each sample is composed of a vector composed by the characteristics of the text and a class labels. The set as follows:

978-1-4244-3930-0/09/$25.00 ©2009 IEEE

$$T = \{(x_i, y_i) \mid i = 1,2,\ldots,l\}$$

Where x_i is the vector containing features describing element i, $x_i \in R_n$; The y_i is either +1 or −1, indicating the class which the point x_i belongs.

Formally, a Standard Support Vector Machine is defined by the following primal optimization problem [4]:

$$\begin{aligned} &\min \quad \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\xi_i \\ &s.t. \quad y_i[(\omega \cdot x_i) + b] \geq 1 - \xi_i \qquad (1) \end{aligned}$$

In the formula above, $\xi_i \geq 0$, $i = 1,2,\ldots,l$, ω is the vector of coefficients, the constant C appears as a punishment factor, and ξ_i is the parameter for handling non-separable data. It is clearly that the larger the C, the more the error is penalized. This is a quadric programming optimization problem, so the constraint along with the objective of minimizing can be solved using Lagrange multipliers. As a result, we can calculate the optimal solution: ω^*, b^* and ξ^* ,then we get the maximum-margin hyperplane, and the best classifier is obtained:

$$F(x) = \text{sgn}[(\omega^* \cdot x) + b^*]$$

In practice, the importance of each sample is different, but the standard support vector machine does not consider that. Ignoring the importance of the sample may cause an important classification being wrongly classified, which will result in the new data is wrongly classified by the discriminant function.

3. Spam Filtering Based on WSVM

Weighted Support Vector Machine (WSVM) is usually used to solve the imbalance classification problems [5], besides the sample's amount of each class, this paper argues that the importance of different classes can also lead to imbalance problems. For the issue of spam filtering, legitimate e-mails is more important than spam. In ensuring the accuracy of classification, at the same time, the SVM should try to reduce the misclassification of legitimate e-mails, so the filtering problem is also the problem of unbalanced sample classification. Weighted support vector machine introduces weight variables, σ, which reflect the importance of different classes. As we can see, it is an effective way to reduce the misclassification of legitimate e-mails by increasing the weight value of the class legitimate e-mails. In addition, taking the importance of each e-mail into account, the method introduces another variable, $s_i > 0$, which reflects the importance of each case .After those variables are added, for each E-mail, the possibility of correct classification is raised, and the degree of misclassification of important E-mails is reduced. As a result, the method improves the classification accuracy obviously.

Based on the above analysis, the equation (1) now transforms to the pattern of Spam Filtering Based on Weighted Support Vector Machines:

$$\begin{aligned} &\min \quad \frac{1}{2}\|\omega\|^2 + C\sigma\sum_{i=1}^{l}s_i\xi_i \qquad (2) \\ &s.t. \quad y_i(\omega^T\Phi(x_i) + b) \geq 1 - \xi_i \end{aligned}$$

In the equation (2), $\xi_i \geq 0$, $i = 1,2,\ldots,l$, y_i is the label for the sample i, in spam filtering, the legitimate e-mail and spam are assigned by the numerical class labels +1 and -1, respectively. $\Phi(x_i)$ is a non-linear kernel function. The kernel function may transform the data into a higher dimensional space to make it possible to perform the separation. From literature [6], we can know that Radial Basis Function $K(x,x') = \exp(-\gamma\|x - x'\|)$ is suit to spam filtering better than other kernel function, so we use Radial Basis Function in the experiments. The samples which belong to the same class have the same class weight, σ_+ or σ_- .Compared with standard support vector machine, The Weighted Support Vector Machine gives fuzzy penalties on misclassification [7], In other words, the slack variable for each sample is multiplied by the sample weight s_i and class weight σ .

Of the equation (2), The Lagrange function is constructed as follows:

$$\begin{aligned} L(\omega,b,\xi,\alpha,\beta) = &\frac{1}{2}\|\omega\|^2 + C\sigma\sum_{i=1}^{l}s_i\xi_i \qquad (3) \\ &-\sum_{i=1}^{l}\alpha_i(y_i(\omega^T\Phi(x_i) + b) - 1 + \xi_i) - \sum_{i=1}^{l}\beta_i\xi_i \end{aligned}$$

In the equation above, α_i , β_i are Lagrange multipliers, then the KKT conditions for the problem are as follows:

$$\begin{aligned} &\frac{\partial L}{\partial \omega} = \omega - \sum_{i=1}^{l}\alpha_i y_i \Phi(x_i) = 0 \\ &\frac{\partial L}{\partial b} = -\sum_{i=1}^{l}\alpha_i y_i = 0 \qquad (4) \\ &\frac{\partial L}{\partial \xi_i} = \sigma s_i C - \alpha_i - \beta_i = 0 \end{aligned}$$

The above-mentioned conditions will be incorporated into the Lagrange function, then the weighted support vector machine problem of the optimal solution for the dual problem:

$$\max_{\alpha} \quad \sum_{j=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i \cdot x_j)$$

$$s.t. \quad \sum_{i=1}^{l}\alpha_i y_i = 0 \qquad (5)$$

$$0 \le \alpha_i \le C\sigma s_i, i = 1,2,\ldots,l$$

By solving the quadratic programming equation (5), the optimal solution of Lagrange multiplier is obtained:

$$\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_l^*)^T$$

Then we consider the KKT condition (4) coupled with α^*, which leads to

$$\omega^* = \sum_{i=1}^{l}\alpha_i^* y_i \Phi(x_i)$$

$$b^* = y_i - \sum_{i=1}^{l} y_i a_i^* K(x_i \cdot x_j)$$

The classification of new patterns uses the optimum values for w^* and b^*:

$$F(x) = \operatorname{sgn}\left[\sum_{i=1}^{l} y_i \alpha_i^* K(x_i, x) + b^*\right]$$

in which $0 \le \alpha_i \le C\sigma s_i$, the upper bound of Lagrange multipliers α_i vary with the change of sample weights s_i and the class weight σ.

4. Evaluations and Analysis

4.1. Evaluation indicators

In this paper, in order to verify spam-filtering performance, some evaluation indicators are introduced. We assume that there are N e-mails to be tested. The definitions of several variables are as follows:

Table 1. Variable definition table

	spam	legitimate emails
Be classified as spam	a	b
Be classified as legitimate e-mail	c	d

From the above table, it is easy to get the equation N=a+b+c+d. The indicators are given as follows:

1. Recall(R): $R = \frac{a}{a+c}$,the higher the recall rate, the less the misclassification of spam.

2. Precision (P): $P = \frac{a}{a+b}$, the higher the precision rate, the less the misclassification of legitimate emails.

3. Accuracy (A): $A = \frac{a+d}{N}$,For all mail, it represents the rate of classification.

For the spam filtering models, the precision is more important than the recall rate, because consumers prefer to receive spam, rather than being tolerant of misclassifying the legitimate emails as spam.

4.2. Experiment and Analysis

In this paper, in order to verify the spam-filtering model has a good performance, we carry out several sets of experiments by choosing the widely used e-mail corpus ZH1. In the experiments, from the Chinese spam corpus ZH1, we select 200 legitimate emails and 200 spam as a training set, and use the same method to select another 400 E-mails as a test set. In order to verify the influence of the degree of misclassification of legitimate emails and the accuracy of the spam-filtering model after the adjustment of the class weight, we set up three different groups of σ_+ and σ_- in the experiment. Our algorithm refers to the algorithm developed by Chang and Lin's Libsvm, and the experimental results are as follows:

Table 2. Experimental results

σ+	σ-	R	P	A
2	1	96.50%	97.47%	97.00%
5	1	93.00%	98.41%	95.75%
10	1	89.50%	99.44%	94.50%

From the table we can see that the weighted support vector machine can control the class error rate of classification, and with the increase of the weight of legitimate e-mails, the precision increases, while the recall rate decreases slightly. This is because the improved legitimate e-mail weight results in the change of classification hyperplane margin. Experimental results show that the filtering algorithms can effectively improve the precision.

5. Conclusions

In this paper, the method of weighted support vector machine is used to spam filtering, and we obtain the new pattern of spam filtering based on weighted support vector machines. The experimental result shows that the method reduces the degree of misclassification of legitimate emails effectively while the accuracy of classification reduces a little. Therefore, this method has good application prospects in spam filtering.

Acknowledgement

the National Natural Science Foundation of China (No.60873247); the Natural Science Foundation of Shandong Province of China (No.Y2006G20); High-tech self-innovation project of Shandong Province (No. 2008ZZ28)

6. References

[1] BIAN Ji-rong, “Study of spam filter algorithm based on CS-SVM and Bagging”,*Ningxia Engineering Technology*,2008, PP.67-69.

[2]JIAN Yan-ying,LIN Min,“Spam-filtering Techniques”,*CommunicationsTechnology*,2008,PP.158-160.

[3]C.-C.Chang,C.-W.Hsu,C.-J.Lin,“The analysis of decomposition methods for support vector machines”,*IEEE Transactionsons on Neural Networks* ,2000,PP.1003-1008.

[4]Drucker H,Vapnik V. “Support Vector Machines for Spam Categorization”.*IEEE Transactionsons on Neural Networks*, 1999, PP. 1048-1056.

[5]YANG zhi-min,LIU guang-li.*The Theory of Fuzzy SVM and Its Application*,Science Press of China,Beijing,China , 2006.

[6]DONG jian-she,YUAN zhan-ping,ZHANG qiu-yu. “Application of various kernel function based on SVM in spam filtering”.*Journal of Computer Applications*, 2008, PP.425-427.

[7]Kubat,M.,Matwin,S.,“Addressing the Curse of Imbalaneed Training Sets:One-Side Selection”,*Proceedings of the 14th hitemational Conferenee on Machine Learning,*1997,PP.217-225.