Modern Applied Science
www.ccsenet.org/journal.html
Vol. 3, No. 10
October 2009

Applications of Support Vector Machine Based on Boolean Kernel to Spam Filtering

Shugang Liu & Kebin Cui
School of Computer science and technology, North China Electric Power University
Hebei 071003, China
E-mail: lsg69@sohu.com, ncepuckb@163.com

Abstract

Spam is so widely speared that has a bad effect on daily use of E-mail. Nowadays, among the primary technologies of spam filtering, support vector machine (SVM) is applied widely, because it is efficient and has high separating accuracy. The main problem of support vector machine arithmetic is how to choose the kernel function. To solve this problem people propose spam filtering arithmetic of support vector machine based on Boolean kernel. The arithmetic uses filtering methods based on attributes, such as IP address, subject words, keywords in content, enclosure information, etc. These attributes compose the feature vectors, and the vectors are classified by SVM-MDNF based on Boolean kernel. The experiment results show that this arithmetic has high separating accuracy, high recall ratio and precision ratio. The arithmetic has its value in theory and application.

Keywords: Spam, Support Vector Machine, Boolean Kernel

1. Introduction

E-mail is one of the main means for people to communicate information on Internet. As the Internet is so widely used, sending and receiving E-mail has almost become a part of considerable amount of people's daily life. However, with the convenience the Internet brings, it also brings the existence and wide spread of spams, which cause a lot of troubles to people. It is evident that people's work efficiency and their emotion will be influenced, if they have to spend time and efforts on identification E-mail every day. So to auto-distinguish spam has important meaning and applying value(Shawe-Taylor J, Cristianini N. KerneI. 2005). Spam means that publicizing E-mails, containing all kinds of publicities, such as ads, electronic publications, are not requested or accepted by receivers in advance.

To classify the technologies of spam filtering, they can be classified into two kinds: server spam filtering and client spam filtering, according to different places the filter is executed. But if we classify the technologies based on different filtering methods, there are three ways: spam filtering based on blacklist/ whitelist, spam filtering based on principles and spam filtering based content.

1) Spam Filtering Based On Blacklist/Whitelist

Any E-mails, sent by senders in the whitelist, are considered legal E-mails, while any E-mails sent by the senders in the blacklist are treated as spams. The following method is widely used in spam filtering recently. Usually it collects a blacklist and a whitelist. In these lists, the content can be E-mail addresses, the DNS of E-mail servers or IP addresses. They help receivers to check senders in real time.

2) Spam Filtering Based On Principles

This method needs people to set some principles. And the spam is the E-mail that meets one of several principles. These principles always include analysis on header, filtering on multiple send, accurate matching on keywords and other features of the E-mail.

3) Spam Filtering Based On Content

Actually, the producers who send spam vary continuously. So the blacklist/whitelist has great limitations. And spam filtering based on principles also has some disadvantages: principles are made by people, and those users who are lack of experience will affect the validity and accuracy of principles. Therefore, many experts come up with an idea that analyze the content of E-mail first, and then distinguish whether it is a spam. This method combines spam filtering with other technologies, such as text classification and information filtering. It requires the arithmetic of text classification and information filtering to be introduced into the spam filtering.

To solve this problem, a great amount of measures have been adopted, such as extension of E-mail protocols, certification of E-mail server, spam filtering and legislation. Among these measures, the spam filtering is more realistic. Nowadays, many arithmetic of text classification have been introduced into applications of spam filtering based on content, like Bayes, Decision Tree, K-Most Neighboring Arithmetic, Support Vector Machines, etc(Wang bin, Pan wenfeng. 2005). And applications of SVM are more successful in spam filtering.

2. Evaluate Standard of Spam Filtering System

The performance evaluation on spam filtering often makes use of some related indexes in text classification. The standard, which can decide whether text classification is mature or not, is the mapping accuracy and mapping speed. And the mapping speed is decided by the complexity of mapping arithmetic; the mapping accuracy is evaluated by information retrieval evaluation. The followings are the definitions about two common indexes: Recall Ratio and Precision Ratio of information retrieval in spam filtering field(C.J. van Rijsbergen. 1979).

Def 1: Recall Ratio is the ratio of the amount of spam that has been filtered to the amount of E-mails that should be filtered. The computing formula of Recall Ratio is:

$$\mathrm{Re}\,call = \frac{amount\ \ of\ \ filtered\ \ spam}{amount\ \ of\ \ E\!-\!mails\ \ that\ \ should\ \ be\ filtered} \tag{1}$$

Def 2: Precision Ratio is the ratio of the amount of spam that has been filtered to the amount of E-mails that have been filtered. The computing formula of Precision Ratio is:

$$\mathrm{Pr}\,ecision = \frac{amount\ \ of\ \ filtered\ \ spam}{amount\ \ of\ \ E\!-\!mails\ \ that\ \ have\ \ been\ filtered} \tag{2}$$

Both of Recall Ratio and Precision Ratio reflect the quality of E-mail classification. They should be considered together rather than only one of them is paid attention. So F1 Test Value is often used to plan the classification result of evaluation E-mails as a whole. The computing formula of F1 Test Value is:

$$F1 = \frac{\mathrm{Pr}\,ecision\ Ratio \times \mathrm{Re}\,call\ Ratio \times 2}{\mathrm{Pr}\,ecision\ Ratio + \mathrm{Re}\,call\ Ratio} \tag{3}$$

And there are Micro Average and Macro Average to calculate Recall Ratio, Precision Ratio and F1 Test Value. Micro Average counts respectively every kind's recall ratio, precision ratio and n test value; and Macro Average unifiedly calculates all kinds' recall ratio, precision ratio and n test value. It is evident that all E-mail filtering arithmetic is aimed at reaching the performance requisition of recall ratio and precision ratio in E-mail classification in the end.

3. Support Vector Machine Based on Boolean Kernel Function

Support vector machine (SVM)(Zhang Yang, Li Zhanhuai, Tang Yan, Cui Kebin, DRC-BK. 2004) is a learning method proposed by Vapnik and the research group, which is led by him in Bell Laboratory. And this method is based on statistics.

SVM is developed from Optimal Separating Plane on linear classifying. The basic idea of it is maximum-separation (margin). The so called optimal means that separating plane is required not only to separate two kinds of text correctly, but also to find a max margin.

Actually, the maximum-margin is the control of promotion ability. Linear support vector machine separates the "yes" and "no" examples, through constructing optimal hyperplane $\langle W, X\rangle + b = 0$ in input space. Here the "<,>" represents the inner product; $W \in R^n$, $b \in R$, to make that:

$$y_i\left[\left\langle W, X_i\right\rangle + b\right] - 1 \geq 0 \qquad i = 1, 2, ..., d \tag{4}$$

It can be proved that the optimal separating plane is what leads to minimum $\frac{1}{2}\|W\|^2$ in input space. To solve this problem we need to transform it to dual form with Lagrange Optimization. The dual form can also be called constraints:

$$\sum_{i=1}^{d} y_i \alpha_i = 0 \quad i = 1, 2, ..., d \tag{5}$$

The solving is as follows:

$$Q(\alpha) = \arg\max_{\alpha} \sum_{i=1}^{d} \alpha_i - \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \alpha_i \alpha_j y_i y_j \left\langle \vec{X}_i, \vec{X}_j \right\rangle \tag{6}$$

α_i is the corresponding Lagrange multiplier of constraint (5) in primary problem. This is a problem of seeking optimization for quadratic function on the constraint of inequality and it has unique answer. It can be proved easily that only part (often a little part) of α_i answers are not equal to zero, and the corresponding examples are the support vector. Through working out the above-mentioned problem, we get the optimal separating function. That is:

$$f(X) = \mathrm{sgn}(\sum_{i=1}^{d} \alpha_i y_i \left\langle \vec{X}_i, X \right\rangle + b) \tag{7}$$

In the function: in fact, the summation only works in support vector. The b is separating threshold. It can be worked out with any support vector (satisfying formula 5th) or through the median of any pair of support vectors in two classes.

$$b = y_s - \sum_{i=1}^{d} \alpha_i y_i \left\langle \vec{X}_i, \vec{X}_s \right\rangle, \quad \alpha_s \neq 0 \tag{8}$$

Here, the sgn() is a symbol function.

With Non-linear-Mapping ϕ, vectors of input space can be transformed to vectors of higher-dimension space, which is named as feature space. The feature space has a higher dimension than the input space.

Non-linear SVM makes use of Non-linear-Mapping ϕ to transform vectors of input space to vectors of high-dimension space. Therefore, $\vec{X}_i, X$ in the above equation are respectively replaced by $\phi(\vec{X}_i), \phi(X)$. So we can get that:

$$f(X) = \mathrm{sgn}(\sum_{i=1}^{d} \alpha_i y_i \left\langle \phi(\vec{X}_i), \phi(X) \right\rangle + b) \tag{9}$$

In the function:

$$b = y_s - \sum_{i=1}^{d} \alpha_i y_i \left\langle \phi(\vec{X}_i), \phi(\vec{X}_s) \right\rangle, \quad \alpha_s \neq 0 \tag{10}$$

We name the function like $K(x, y) = (\phi(x), \phi(y))$ as kernel function. Some common kernel functions include:

1) Gaussian Radial basis functions: $K(x, x_i) = \exp(-\frac{\| x - x_i \|^2}{2\sigma^2})$

2) Polynomial: $K(x, x_i) = ((x, x_i) + 1)^d$, for d=1,..,N

3) Hyperbolic tangent: $K(x, x_i) = tamj(\beta x_i + b)$

4) Spline kernel functions: $K(x, x_i) = B_{2n+1}(x - x_i)$

Choosing different kernel functions, you can get different Non-linear support vector machine.

If x and y in the kernel functions above are Boolean, then we can suppose that $U \in \{0,1\}^n, V \in \{0,1\}^n$, $\sigma > 0, p \in N$, for I represents unit vector. So:

$$K_{MDNF}(U, V) = -1 + \prod_{i=1}^{n} (\sigma U_i V_i + 1) \tag{11}$$

We call K_{MDNF} as Monotone Disjunctive Normal Form (MDNF) kernel function. MDNF kernel function is the kernel function we use in this paper as SVM arithmetic.

4. SVM Spam Filtering Based on Boolean Kernel Function and the Experiment Results

4.1 The Strategy of SVM Spam Filtering Based On Boolean Kernel

This experiment adopts Enron-spam E-mail dataset. And the dataset includes two parts: “pre-processed” is the set of E-mails that have been pretreated, and the part “raw” are pretreated based on needs to get “preprocessed”. Our experiment cramps out some “preprocessed” as training set, and some as testing set. We select 2000 E-mails. Among these E-mails, 1100 are spam and 900 are normal E-mails.

The specific procedures of the strategy of SVM spam filtering based on Boolean kernel are as followings:

1) Firstly, we process the dataset with standard. Wipe off the noise words (such as spelling mistakes, etc), and

filter words whose text frequency are between 2 and 8000; set different weighing to the subject and text content of every E-mail, and the subject is set higher weighing to concern the words appearing in the E-mail subject. Taking subject, text content and many other features of the E-mail into consideration, we will get the feature vector of every E-mail.

2) Make binaryzation towards the features in the feature vector. That is to give every feature the value "0" or "1". Since we use Boolean kernel MDNF here, there is a need to transform the feature vector to Boolean feature vector.

3) Filter spam with SVM based on MDNF Boolean kernel. In order to verify whether the arithmetic is valid or not, we use k cross for our experiment. K cross is to separate E-mails into k parts. We make use of the k-1 parts for training, and the remaining for testing. The procedure loops k times, so every part has been tested. Finally, the average of tests' values is used as the result of test for evaluation. Here we make k equal 10.

4.2 Experiment Result and Analysis

In this experiment, we compare the separating accuracy of the spam filtering arithmetic based on Boolean kernel SVM with that of some arithmetic-Naïve Bayes, linear SVM and Non-linear SVM based on radial basis functions. The result is shown is the table 1:

From the comparison result of separating accuracy, it is evident that the highest is SVM based on MDNF Boolean kernel. Second top is the Non-linear SVM based on radial basis functions. The lowest is Naïve Bayes.

During the evaluation of the efficiency of E-mail separating arithmetic, it cannot evaluate the arithmetic completely only to compare the separating accuracy. So we evaluate the arithmetic further using precision ratio, recall ratio and F_1 given in the Section 2.

In table 2, it compares the recall ratio, precision ratio and F_i. And from these targets, we can evaluate the validity of the arithmetic in a more comprehensive way. From the experiment result, we can find that SVM based on MDNF Boolean kernel has the best spam filtering effect, comparing with the other three.

5. Conclusion

After the analysis of all the characteristics of spam, we propose the SVM based on MDNF Boolean kernel spam filtering arithmetic when we make the feature vector using E-mail subject, text content, etc. The experiment shows that this arithmetic has higher separating accuracy, and has better spam filtering effect in recall ratio and precision ratio, comparing with Naïve Bayes, Linear SVM and SVM based on radial basis functions. And in the experiments thereafter, we will apply SVM with more Boolean kernels to spam filtering, and look forward a better effect.

References

C.J. van Rijsbergen. (1979). *Information Retrieval* (2nd edition), Butterworths, London, 1979.

http://www.cs.cmu.edu/~enron/

Shawe-Taylor J, Cristianini N. KerneI. (2005). Methods for Pattern Ana1ysis. Bei jing: China Machine Press, 2005:60-74.

Wang, bin, Pan, wenfeng. (2005). Content-based spam filtering technology. *Journal of Chinese Information Processing*. Bei jing: 2005, 19(5):1-10.

Zhang Yang, Li Zhanhuai, Tang Yan, Cui Kebin, DRC-BK. (2004). Mining Classification Rules with Help of SVM. In the Proceedings of the 8th Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD'04), Lecture Notes in Artificial Intelligence, Volume 3056, Springer-Verlag Press, 2004.

Table 1. Comparison of separating accuracy

Classify algorithm	Classify accuracy
NB	92.5%
Liner SVM	93.8%
RBF kernel SVM	94.7%
MDNF-SVM	97.8%

Table 2. Comparison of recall, precision and F$_i$

Classify algorithm	recall	precision	F_1
NB	90.4%	88.7%	89.5%
Liner SVM	91.2%	90.5%	90.8%
RBF kernel SVM	92.2%	92.5%	92.3%
MDNF-SVM	94.2%	95.5%	94.8%