E-mail Spam Filtering Using Support Vector Machines with Selection of Kernel Function Parameters

Hsu Wei-Chih
Department of Computer and Communication
National Kaohsiung First University of Science and Technology
Kaohsiung, Taiwan
weichih@ccms.nkfust.edu.tw

Tsan-Ying Yu
Institute of Engineering Science and Technology
National Kaohsiung First University of Science and Technology
Kaohsiung, Taiwan
Department of Electrical Engineering
Kao Yuan University
Lu Chu, Taiwan
yotnyg@gmail.com

***Abstract*—Support Vector Machines (SVM) is a powerful classification technique in data mining and has been successfully applied to many real-world applications. Parameter selection of SVM will affect classification performance much during training process. However, parameter selection of SVM is usually identified by experience or grid search (GS). GS is simple and easily implemented, but it is very time-consuming. In this study, Taguchi method is proposed for improving GS and used to optimize the SVM-based E-mail Spam Filtering model. It is easy to implement by orthogonal arrays without iteration. A real-world mail dataset is selected to demonstrate the effectiveness and feasibility of the method. The results show that the Taguchi method can find the effective model with high classification accuracy and good robustness.**

Keywords-e-mail; spam; support vector machine (SVM); parameter; grid search (GS);

I. INTRODUCTION

Spamming is the abuse of electronic messaging systems to send unsolicited bulk messages or to promote products or services, which are almost universally undesired. Spamming is economically viable because advertisers have no operating costs beyond the management of their mailing lists. The sender cannot be specified, because the sender of Spamming has only temporary E-mail address and the reply of them is not reached to the original sender. Therefore, undesired Emails to us have been increased everyday, so that, it is not easy to read an important E-mail.

Spam filtering based on the textual content of email messages can be regarded as a special case of text categorization (TC), with the categories being spam and legitimate (non-spam). Content-based filters can be divided into rule-based methods and probabilistic methods. Rule-based methods such as Ripper [1, 2], Decision Tree [3], Rough Sets [4], Boosting [5] and so on strongly dependent on the existence of key terms, therefore, specific terms can cause the failure of filtering. Methods based on probability and statistics such as K-Nearest Neighbor [4] and Support Vector Machine (SVM) [6] and so on. Besides, the prevailing machine learning method for spam message filtering is the Bayesian approach [7] used with good results.

SVM proposed by Vapnik [6] in 1995, has been widely applied in many applications such as function approximation, modeling, forecasting, optimization control...etc, and has yielded excellent performance. It is a statistical theory to deal with the dual categories of classification and can find the best hyperplane to partition a sample space. However, for the SVM-based model, its classification performance is sensitive to the parameters of the model, thus, parameters selection is very important. In order to enhance the accuracy of SVM, it is necessary to find the best SVM parameters (γ, C) combinations. Most of the previous researches focus on the GS (GS), pattern search based on principles from design of experiments (DOE) such as Staelin [8] and genetic algorithm (GA) [9, 10] to choose the parameters. GS is simple and easily implemented, but it is very time-consuming. DOE is like GS but it reduces the searching grid density and can reduce the computational time greatly. Although GA does not require setting an initial search range, it introduces some new parameters to control the GA searching process, such as the population size, generations, and mutation rate. Those above method may cause exhaustive parameter searches.

The Taguchi method [11] introduced to our approach, a DOE and robust design approach, uses many ideas from statistical experimental design for evaluating and implementing improvements in products, processes, and equipment. The fundamental principle is to improve the quality of a product by minimizing the effect of the causes of variation without eliminating the causes. The two major tools used in the Taguchi method are: 1) signal-to-noise ratio (S/N) which measures quality and 2) orthogonal arrays which are used to study many design parameters simultaneously.

We find the best combination of two parameters (C, γ) of SVM by Taguchi method. During making spam filtering model, the parameters (C, γ) of SVM are regarded as control factors. After selecting the parameters of SVM, we verify the classification results and compared with GS. Compared with the above mentioned methods, Taguchi method has few parameters to control the searching process. As far as we know, this maybe the first attempt to introduce Taguchi method to optimize the SVM for spam filtering models.

978-0-7695-3873-0/09 $29.00 © 2009 IEEE

IEEE computer society

II. Support Vector Machine

The textual and non-textual features representing an email, obtained through the method mentioned previously, are as the input to the spam email filtering algorithm. In the approach, the filtering algorithm is represented by SVM.

SVM is a powerful supervised learning paradigm based on the structured risk minimization principle from statistical learning theory, which is currently placed among of the best-performing classifiers and have a unique ability to handle extremely large feature spaces (such as text), precisely the area where most of the traditional techniques fail due to the "curse of the dimensionality". SVM has been reported remarkable performance on text categorization task. In our evaluation, we used the Library for SVM [12] to build SVM models. In the following, we give a brief introduction to the theory and implementation of SVM classification algorithm.

Consider the problem of separating the set of training set vectors belonging to two separate classes in some feature space. Given one set of training example vectors:

$$(x_1,y_1),...(x_l,y_l), x_i \in R_n, y_i \in \{-1,+1\} \qquad (1)$$

we try to separate the vectors with a hyperplane

$$(w\cdot x)+b=1 \qquad (2)$$

so that

$$y_i[(w\cdot x)+b]\ge 1,(i=1,2,...,l) \qquad (3)$$

The hyperplane with the largest margin is known as the optimal separating hyperplane. It separates all vectors without error and the distance between the closest vectors to the hyperplane is maximal. The distance is given by

$$d(w,b)=\frac{2}{\|w\|} \qquad (4)$$

Hence the hyperplane that separates the data optimally is the one that minimizes the following equation:

$$Minimize\frac{1}{2}\|w\|^2 \qquad (5)$$

subject to the constraints of (4).

To solve above problem, introduce lagrange multipliers α_i, i = 1,2 …,l and define

$$w(\alpha_i)=\sum_{i=1}^{l}\alpha_i y_i x_i \qquad (6)$$

With Wolfe theory the problem can be transformed to its dual problem:

$$\max.\ W(\alpha)=\sum_i \alpha_i-\frac{1}{2}w(\alpha)\cdot w(\alpha), s.t.\ \ \alpha_i\ge 0 \qquad (7)$$

$$\sum_i \alpha_i y_i = 0 \qquad (8)$$

With the optimal separating hyperplane found, the decision function can be written as:

$$f(x)=(w_0\cdot x)+b_0 \qquad (9)$$

Then the test data can be labeled with

$$label(x)=\mathrm{sgn}(f(x))=\mathrm{sgn}((w_0\cdot x)+b_0) \qquad (10)$$

Training vectors that satisfy $y_i[(w_0\cdot x)+b_0]=1$ are termed support vectors, which are always corresponding to nonzero α_i. The region between the hyperplane through the support vectors on each side is called the margin band.

In the case of linearly non-separable training data, by introducing slack variables the primal problem can be rewritten as:

$$Min\left(\frac{1}{2}\|w\|^2+C\sum_i \xi_i\right) \qquad (11)$$

subject to $y_i[(w\cdot x)+b]\ge 1-\zeta_i,\zeta_i\ge 0$.

Similarly, we can get the corresponding dual problem

$$\begin{aligned} &\max W(\alpha)=\sum_i \alpha_i-\frac{1}{2}w(\alpha)\cdot w(\alpha),\\ &s.t.\quad C\ge\alpha_i\ge 0,\ \sum_i \alpha_i y_i=0 \end{aligned} \qquad (12)$$

Problems described as in Equation(12) and Equation(13) are typical quadratic optimization questions, and have been approached using a variety of computational techniques. Recent advances in optimization methods have made support vector learning in large-scale training data possible.

All the training vectors corresponding to nonzero α_i are called support vectors, which form the boundaries of the classes. The maximal margin classifier can be generalized to nonlinearly separable data via transforming input vectors into a higher dimensional feature space by a map function φ, followed by a linear separation there. The expensive computation of inner products can be reduced significantly by using a suitable kernel function $K(x_i,x_j)=(\varphi(x_i),\varphi(x_j))$. We implemented the SVM classifier using the LIBSVM library [12] and adopted radial basis function (RBF) defined as the kernel $K(x_i,x_j)=\exp\left(-\gamma\|x_i-x_j\|^2\right)$. In this study, the RBF is used as the basic kernel function of SVM. There are two parameters associated with the RBF kernels: C and γ.

III. Preprocessing for Spam Mail

Vector space model is a text representing approach, which is widely used and has good performance in TC. In its simple form, spam filtering can be recast as text categorization task where the classes to be predicted are spam and legitimate. Therefore, Email can be regarded as a vector space, which is composed of a group of orthogonal key words.

For each email, its textual portion was represented by a concatenation of the subject line and the body of the message. Due to the prevalence of html and binary attachments in modern email a degree of pre-processing is required on messages to allow effective feature extraction. Therefore, we adopt the following data pre-processing steps:

1) To avoid treating forms of the same word as different attributes, a lemmatizer was applied to the corpora to convert each word to its base form (e.g., "got" becomes "get").

2) The stopping process is adopted to remove the high frequent words with low content discriminating power in a email document such as "to", "a","and","it", etc. Removing

these words will save spaces for storing document contents and reduce time taken during the subsequent processes.

3) Words were defined as contiguous strings of characters delimited by white space. All characters were converted to lowercase, but if a word consisted of all capital letters, it was effectively treated as two words: all lowercase all uppercase.

IV. Implementation

After preprocessing as previous section, we obtain word frequencies and convert into vectors. We introduce Taguchi method to our approach. In content-based spam filtering performance analysis, a commonly used evaluation criteria measuring the efficiency of the classification is accuracy (Acc). It is regarded as response variable, defined as:

$$Acc = \frac{A+D}{N} \tag{13}$$

where N is the number of all messages; A is as spam and the actual system to determine the number of spam; and D that the actual system for normal mail and e-mail to determine the number of normal.

In order to reduce the number of experiments and the cost of design, we have to choose proper orthogonal array by numbers of control factors and levels. As the search scope is suggested by Lin [12] and expand to different combinations of parameters C and γ with 32 levels: $log_2(C)$ = (-15, -14, -13, -12, ..., 15, 16) and $log_2(\gamma)$ = (-15, -14, -13, -12, ..., 15, 16) to find the best combination and obtain high accuracy.

In this work, Taguchi approach with an orthogonal table $L_{64}(2^{63})$ of 63 two-level factor is employed. To avoid interaction among columns in new orthogonal table, first, second and fourth columns and their interaction columns: third and seventh columns in the table $L_{64}(2^{63})$ are merged and converted into the first column of $L_{64}(32^1 \times 32^1 \times 2^{53})$ as depicted in Table I. For the same reason, next, we merge 8th, 16th, and 32nd columns and their interaction columns: 24th and 54th columns into the second column of $L_{64}(32^1 \times 32^1 \times 2^{53})$. We merge two groups of five columns in the interaction and create a new orthogonal table : $L_{64}(32^1 \times 32^1 \times 2^{53})$ such that it can cover 2 32-level factors and 53 two-level factors. Then, parameters $\log_2(C)$ and $\log_2(\gamma)$ with 32 levels are allocated to the first and the second column of $L_{64}(32^1 \times 32^1 \times 2^{53})$, respectively. The response variable is the accuracy.

Each run of $L_{64}(32^1 \times 32^1 \times 2^{53})$ will proceed 5-fold cross validation. The S/N ratio for each run and the average S/N ratio for each level and each factor need to be evaluated. We pick the level with maximum *S/N* ratio for each factor. Therefore, we can obtain approximation results. The *S/N* ratio here can be computed using the standard greater-the-better as follows.

$$S/N = -10Log\left(\frac{1}{Acc}\right) \tag{14}$$

TABLE I. Columns In the table $L_{64}(2^{63})$ are merged into the 1st column of $L_{64}(32^1 \times 32^1 \times 2^{53})$

New Lev. Factor	1	2	3	...	31	32
Log$_2$(γ)	-15	-14	-13	...	15	16
Col. 1(Lev)	1	1	1	...	2	2
Col. 2(Lev)	1	1	1	...	2	2
Col. 3(Lev)	1	1	1	...	2	2
Col. 4(Lev)	1	1	2	...	2	2
Col. 7(Lev)	1	2	1	...	1	2

V. Experiment results and discussion

In our test, the program runs with LIBSVM toolbox provide by Lin [12] on an IBM compatible PC with AMD Athlon 64 3000+ CPU running at 1.8 GHz with 1GB RAM.

The data set used for training and testing comes from the SpamAssassin public corpus [13]. This public corpus was chosen because it is recent, contains reasonably complete header information, and contains both spam and legitimate. It includes 6047 messages with a 31% spam rate, which has been used in a considerable number of publications. This dataset is divided into three groups, denoted as A, B, C, respectively. 500 spam messages and 500 normal messages are extracted from this dataset for each group.

The results of S/N ratio of SVM for dataset A are shown in Table II. Here accuracy is desirable as larger as possible. The highest of S/N ratio -0.221 is Exp. 4. The best combination (γ, C) is $\gamma = 2^{-14}$ and $C = 2^5$. By observing the effective and variance of control factor $log_2(\gamma)$ and $log_2(C)$ for all level, as shown in Table III, it is easy to pick largest *S/N*s for each level. The difference between maximum accuracy and minimum accuracy of main effect for parameters C and γ implies the impact for accuracy. The difference of parameters γ is lager than the one of parameters C. Therefore, parameter γ is more significant than parameter C.

TABLE II. $L_{64}(32^1 \times 32^1 \times 2^{53})$ orthogonal array and experiment data

Exp.	$log_2(\gamma)$	$log_2(C)$	*Acc*	*S/N*
1	-15	-15	0.4820	-6.339
2	-15	0	0.6122	-4.262
3	-14	12	0.9709	-0.257
...	...	...	...	...
4	-14	5	0.9749	-0.221
...	...	...	...	...
63	15	-6	0.4870	-6.249
64	15	-9	0.4930	-6.143

TABLE III. EFFECTIVE AND VARIANCE OF CONTROL FACTORS FOR ALL LEVELS

Factor	Lev. 1	Lev. 2	Lev. 3	...	Lev. 21	...	Lev. 32
$Log_2(\gamma)$	-5.301	$\boxed{\mathbf{-0.239}}$	-0.283	...	-4.927	**...**	-4.804
$Log_2(C)$	-6.440	-6.431	-6.616	...	$\boxed{\mathbf{-2.497}}$	...	-4.117

Factor	Effect	Rank
$Log_2(\gamma)$	6.193	1
$Log_2(C)$	2.430	2

Compared with GS and Naïve Bayes algorithm for dataset A, the results of our confirm test are shown in Table IV. The Exp. 6 tells that the accuracy of SVM will become worse without careful selection for parameters C and γ. No parameter needs to be set up for SVM with linear kernel; the accuracy will lower than that of Naïve Bayes algorithms and our proposed method. As shown in Fig. 1, the optimal pair of (C, γ) is found at the zone with $log_2(C) = 8$ and $log_2(\gamma) = -14$, respectively. The best accuracy of the proposed method is little lower than that of GS 97.70%. GS required searching and computing 32×32 = 1024 times but our proposed method need only 64 times. Our approach is 15 times faster and our best accuracy 97.49% is very close to that of GS.

TABLE IV. RESULTS FOR DATASET A

Exp.	Method	Average of Acc.
1	Naïve Bayesian	84.56%
2	SVM (Linear)	73.76%
4	SVM (Taguchi Method $\gamma=2^{-14}$, $C=2^{5}$)	97.49%
5	SVM (GS $\gamma=2^{-14}$, $C=2^{8}$)	97.70%
6	SVM ($\gamma=2^{-1}$, $C=2^{-12}$)	44.39%

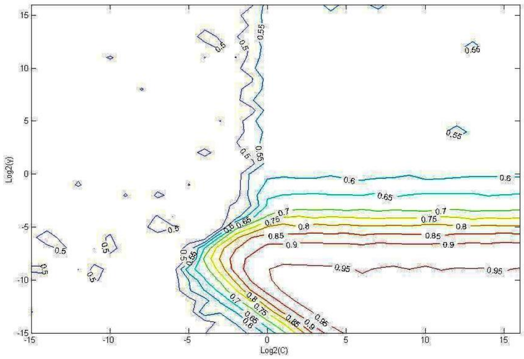

Figure. 1 Grid-search on C = (2-15, 2-14, 2-13, 2-12, ..., 215, 2-16) and γ = (2-15, 2-14, 2-13, 2-12, ..., 215, 2-16)

Table V shows the result of both our search method and GS for each dataset using RBF kernels. The experimental results show that our proposed method can select good parameters for SVM with kernel RBF and the accuracy is very close to that of GS.

TABLE V. RESULTS FOR DIFFERENT DATASET

Data	Taguchi			GS		
	$log_2(C)$	$log_2(\gamma)$	Acc.	$log_2(C)$	$log_2(\gamma)$	Acc.
A	5	-14	0.9749	8	-14	0.9770
B	10	-13	0.9759	9	-14	0.9819
C	10	-11	0.9880	13	-15	0.9900

VI. CONCLUSIONS AND FUTURE WORK

Our proposed approach base on Taguchi method does'nt like other approximation methods or heuristics may cause exhaustive parameter searches. On the other hand, our proposed approach sometimes may obtain approximation results but not optimal. Compared with much computational time to find the optimal parameter values by the grid-search, it is worth for our methods to obtain approximation results with sacrificing little accuracy.

In our method the parameter selection by orthogonal table $L_{64}(32^{1} \times 32^{1} \times 2^{53})$ will obtain high accuracy. If we would like to obtain higher accuracy, we could enlarger orthogonal table such as L_{128} to promote the accuracy.

REFERENCES

[1] W. W. Cohen, "Fast effective rule induction," 1995, pp. 115–123.

[2] W. W. Cohen, "Learning rules that classify e-mail," 1996, pp. 8-25.

[3] X. Carreras and L. Marquez, "Boosting trees for anti-spam email filtering," *Arxiv preprint cs.CL/0109015,* 2001.

[4] I. Androutsopoulos, G. Paliouras, V. Karkaletsis, G. Sakkis, C. D. Spyropoulos, and P. Stamatopoulos, "Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach," *Arxiv preprint cs/0009009,* 2000.

[5] I. Androutsopoulos, G. Paliouras, and E. Michelakis, *Learning to filter unsolicited commercial e-mail*: " DEMOKRITOS", National Center for Scientific Research, 2004.

[6] V. N. Vapnik, "The Nature of Statistical Learning Theory [M]," New York: Springer-Verlag, 1995.

[7] J. Provost, "Naive-bayes vs. rule-learning in classification of email. The University of Texas at Austin," *Artificial Intelligence Lab. Technical Report AI-TR-99-284,* 1999.

[8] C. Staelin, "Parameter selection for support vector machines," *Hewlett-Packard Company, Tech. Rep. HPL-2002-354R1,* 2003.

[9] C. L. Huang and C. J. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Systems With Applications,* vol. 31, pp. 231-240, 2006.

[10] T. Howley and M. G. Madden, "The genetic kernel support vector machine: Description and evaluation," *Artificial Intelligence Review,* vol. 24, pp. 379-395, 2005.

[11] G. Taguchi and S. Chowdhury, *Robust engineering*: McGraw-Hill Professional, 2000.

[12] C.-C. C. a. C.-J. Lin, "LIBSVM -- A Library for Support Vector Machines," 2008.

[13] M. Justin, "Spamassassin public corpus," *URL http://www. spamassassin. org/publiccorpus/. Accessed,* vol. 27, pp. 02-04.