

Support Vector Machine Parameter Optimization for Text Categorization Problems

Mikhail S. Ageev

Boris V. Dobrov

Faculty of Mechanics and
Mathematics of Moscow State
University, Research Computing
Center of MSU, Vorobyevy
Gory, Moscow, 119899, Russia
7 (095) 9393390
ageev@mail.cir.ru

Research Computing Center
of Moscow State University
Vorobyevy Gory, Moscow,
119899, Russia
7 (095) 9393390
dobroff@mail.cir.ru

Abstract: This paper analyzes the influence of different parameters of Support Vector Machine (SVM) on text categorization performance. The research is carried out on different text collections and different subject headings (up to 1168 items). We show that parameter optimization can essentially increase text categorization performance. An estimation of range for searching optimal parameter is given. We describe an algorithm to find optimal parameters. We introduce the notion of stability of classification algorithm and analyze the stability of SVM, depending on number of documents in the example set. We suggest some practical recommendations for applying SVM to real-world text categorization problems.

1. Introduction

There are a lot of papers devoted to text categorization problem. Text categorization approaches based on machine learning show high speed of learning and high precision and recall. Several papers provide comparative analysis of different machine learning methods [Jo98] [Du98] [YL99] [Le01]. They showed Support Vector Machine (SVM) method has an advantage over other machine learning methods for text categorization. They used Reuters-21578 document collection [Reu97] that is specially developed for text categorization researches.

Text categorization for large systems of categories (more than 500 subject headings) is an important scientific problem. It is important to explore and improve machine learning methods for text categorization. Our research uses the hierarchical system of 1168 subject headings. The system of such a big number of categories requires the development of new effective methods for text categorization.

The article [Le01] describes the result of TREC-2001 batch filtering run. The author of article [Le01] used SVM with a simple parameter optimization algorithm. SVM was applied to the TREC document collection. The author achieved the best results on most

topics in three runs. We used the algorithm published in [Le01] as a basic line and improve it's method of parameter optimization.

This paper analyzes the influence of different parameters of SVM on a text categorization performance. The goal of this paper is to improve the text categorization performance of SVM. We show parameter optimization can essentially increase the text categorization performance. An estimation of range for searching optimal parameter is given. We describe an algorithm to find optimal parameters. Our algorithm estimates the bounds for searching optimal parameters. The range for searching the parameters depends on number of positive examples.

The performance of our algorithm on high-frequency topics of Reuters-21578 document collection is analogous to the performance of SVM published in [Jo98] and [Du98]. Our algorithm shows substantial performance improvement on more complex task, such as text categorization for large systems of categories (1168 subject headings).

We introduce the notion of stability of classification algorithm and analyze the stability of SVM depending on number of documents in the example set. We suggest an original algorithm to estimate classification stability.

We suggest some practical recommendations for applying SVM to real-world text categorization problems. This article is illustrated by graphs of observed effects.

This article is structured as follows:

In the section 2 we give a short description of SVM method and give a review of articles devoted to text categorization with support vector machines.

In the section 3 the main results are described. We consider a different parameters of SVM and analyze the influence of different SVM parameters on text categorization performance.

In the section 4 we analyze a stability of classification algorithm, depending on number of documents in the example set.

2. Support Vector Machines

The Support Vector Machines (SVM) method was developed by V.N.Vapnik based on structural risk minimization principle [Va95].

In their basic form, SVM learn linear threshold function $h(x) = \text{sign}((w, x) + b)$ described by a weight vector $w \in \mathbb{R}^n$ and a threshold b . The hyperplane $(w, x) + b = 0$ separates \mathbb{R}^n onto two half-spaces such that one half-space contains all positive examples and the other half-space contains all negative examples. For a given training sample, the SVM finds the hyperplane with maximum margin. Computing this hyperplane is equivalent to solving the following optimization problem [Va95]:

$$\begin{aligned}
L_D(\alpha) &= \sum_{i=1}^l \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) \rightarrow \max \\
0 &\leq \alpha_i \leq C \\
\sum_{i=1}^l \alpha_i y_i &= 0
\end{aligned} \tag{1}$$

$K(x_i, x_j)$ is called a kernel function of SVM. In the simplest case the kernel function is equals to Euclidean scalar product (x_i, x_j) . The task (1) can be solved efficiently [Jo99].

There exist generalizations of basic SVM method. In the case when there are no separating hyperplane the goal of SVM is to maximize the margin and to minimize the number of errors. The solution is a trade-off between the largest margin and the lowest number of errors. The SVM can also find an optimal nonlinear decision function in a set of nonlinear separating surfaces. This can be done by using nonlinear kernel functions (see [Va95] [Bu98]).

Before you apply the SVM to real task you must define a mapping of objects you want to classify to the space \mathbb{R}^n . This mapping is called a *feature vector representation* of subject area. The mapping depends on a subject area. SVM can be applied to many different tasks by choosing an appropriate feature vector representation. SVM have been used for pattern recognition, speech recognition, and text categorization.

The papers [Jo98] [Du98] [YL99] provide comparative analysis of SVM and other machine learning methods. The comparative analysis is made on text categorization task. The methods were tested on Reuters-21578 document collection [Reu97]. An advantage of SVM method over other machine learning methods on text categorization task was experimentally proved.

3. Different parameter optimization

There are many different parameters of SVM, which have an influence to the text categorization performance. These parameters are set by user. The strategy of parameters tuning depends on subject area. Our task is to improve SVM performance on text categorization task with a large number of different subject topics. In this chapter we describe an algorithm for parameter optimization of SVM.

Our research uses a sub-collection of RF legal acts of 2001 year from University Information System RUSSIA (Russian inter-University Social Sciences Information and Analytical consortium, <http://www.cir.ru/eng>). Further we shall reference this collection as FRF-10372. This collection consists of 10372 documents. All the documents were categorized by specialists using the hierarchical system of 1168 subject headings, adopted by the presidential decree [PD00]. Further we shall reference these subject headings as PRESRF-1168.

SVM_light [Jo99] is a free implementation of SVM. We adopt SVM_light v. 3.50 for our experiments.

We map documents to feature vectors as follows:

- 1) each word is converted to normal form (stemming).
- 2) each distinct word stem corresponds to a feature with its TF*IDF score as the value [CCH92].
- 3) the word is included into feature vector representation if it has document frequency bigger than five.

We used F-measure to evaluate the classification performance. Let p be a precision, and let r be a recall. Then F-measure is

$$F = \frac{1 + \beta^2}{\frac{\beta^2}{p} + \frac{1}{r}}$$

Here β is a parameter. β determines the relative importance of the precision and the recall. In our experiments we used $\beta = 1$ and $\beta = \sqrt{1/3}$ (i.e. recall is three times more important than precision).

We used 70% of documents to learn SVM classifier and 30% of documents to estimate the performance.

3.1. Verification of method on Reuters-21578 dataset

In order to verify the validity of our program, we have reproduced the results of other researchers on Reuters-21578 dataset. The results are agreed with published by other researchers. Table 1 shows our results (first column), and the results published in [Jo98] and [Du98].

		Joachims [Jo98a]	Dumais et.al. [Du98]
earn	97,79	98,20	98,00
acq	95,69	92,60	93,60
money-fx	72,83	66,90	74,50
grain	89,00	91,30	94,60
crude	82,82	86,00	88,90
trade	77,45	69,20	75,90
interest	75,57	69,80	77,70
ship	74,55	82,00	85,60
wheat	89,59	83,10	91,80
corn	86,31	86,00	90,30

Table 1: SVM performance on Reuters-21578 dataset (first 10 topics).
Our results is in the first column.

The performance of our algorithm on high-frequency topics of Reuters-21578 document collection is analogous to the performance of SVM published by other researchers. Our goal is to improve text categorization performance of SVM on more complex tasks.

Text categorization on large system of subject headings, such as our PRESRF-1168, is more complex task than the text categorization on Reuters subject headings.

We have observed for some categories there are very simple rules to be assigned to a text. But in most cases a border between different categories was difficult to determine, therefore level of personal subjectivity and inconsistency in such a large category system became very high. It is very difficult to provide sufficient example sets for each category.

Only 47 categories had more than 100 documents (1% of collection) in their example sets, only 200 categories had more than 20 documents in the example sets. Therefore SVM without parameter optimization show very little performance on PRESRF-1168.

To find optimal parameters, we did multiple SVM_light runs for each topic with different parameters. Then we chose the best parameter.

3.2. Different kernel functions

SVM kernel function defines scalar product in \mathbb{R}^n . SVM with nonlinear kernel function can find optimal nonlinear decision surface. For example, to find optimal second-degree decision surface in \mathbb{R}^n we can use SVM with polynomial kernel of second degree.

Here we list several examples of kernel functions:

- a) Linear: $K(x_i, x_j) = (x_i, x_j)$
- b) Polynomial: $K(x_i, x_j) = (s(x_i, x_j) + c)^d$, $d = 2, 3, 4, 5$
- c) Radial-basis: $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$, где $\gamma = 0.5, 1, 2$.

The article [Jo98] shows that using polynomial and radial-basis kernel function one can improve text categorization performance up to 1-2% on Reuters-21578 dataset.

We compared SVM performance with different kernel functions on PRESRF-1168. The experiments showed that different kernel functions give 1-5% performance increase. Only some part of topics allow to increase performance by choosing an optimal kernel function. So we choose to optimize other SVM parameters.

3.3. Feature space reduction

We map documents to feature vectors in three steps:

- 1) each word is converted to normal form (stemming);
- 2) the word is included into feature vector representation if it has document frequency bigger than five;

- 3) the third step consists of removing all the words with low document frequency. We do that to reduce the dimension of feature space.

After processing first two steps on collection FRF-10372 (introduced in ch. 3) we get 202584 distinct words. About 80% of all words have document frequency equals to 1. After removing all the words with document frequency less than or equal to 5 we get 23118 distinct words.

We compared the performance of SVM before and after the feature space reduction. After feature reduction the SVM performance was changed not more than 5% on each topic. On most topics SVM performance increased.

3.4. Relative weight of different errors

Suppose we have some machine learning classifier, and suppose we have several errors of two types:

- e_1 documents, that is truly belong to the topic, does not automatically assigned to the topic.
- e_2 documents, that is not belong to the topic, were automatically assigned to the topic.

Let *error penalty* Err be the the following value:

$$Err = \frac{j e_1 + e_2}{1 + j}$$

Here the parameter j defines the relative weight of errors e_1 and e_2 . On the other hand, the parameter j defines the dominance of recall over precision.

The optimal decision hypersurface is a trade-off between the largest margin and the lowest error penalty.

The parameter j has a strong influence on text categorization performance. Our experiments showed that for a number of low-frequency topics (20-50 documents of 10372) SVM with default parameters do not assign any documents to the topic. But SVM with $j \gg 1$ gives precision and recall 50% and higher. Further we will optimize parameter j .

The article [Le01] describes the result of TREC-2001 batch filtering run. The author of article [Le01] used SVM with a simple parameter optimization algorithm. The algorithm consists of applying SVM with different parameter j and choosing an optimal j . For each topic T_r , j is chosen over 8 different values from fixed range [0.4,8].

We used Lewis's strategy as a basic line. From result tables we see that:

- 1) for a number of topics an optimal value is equals to upper bound of search range;
- 2) the lower number of documents in category, the bigger the optimal j

The second dependence is not strict, but we can estimate this dependence numerically.

In further experiment we made brute-force search of optimal j on 350 different categories. For each category we find optimal $j \in [0.5, 100]$ over 50 values. We used sequence of j that increase in exponential way.

The computation of such a big number of variants takes several weeks of PentiumIII PC. To reduce time requirements we have analyzed an effective search range for parameter j . This search range depends on number of documents assigned to category as follows:

$$\tilde{j} \in \left[1, \max \left(1.5 \frac{\text{neg_ex}}{\text{pos_ex}}, 1 \right) \right] \quad (2)$$

Here pos_ex is the number of positive examples (i.e. documents that is assigned to category), and neg_ex is the number of negative examples (i.e. documents that is not assigned to category).

The resulting algorithm consists of finding optimal j in range (2). It is sufficient 5-10 iterations with j increased in exponential way.

Table 2 shows the result of our algorithm on several topics from PRESRF-1168 on FRF-10372 document collection. The performance metric used is F-measure with $\beta = \sqrt{1/3}$.

1) The performance of the SVM with no parameter optimization is shown in column "no optim". 2) The performance of the basic algorithm that is analogous to [Le01] is shown in column "basic". 3) The performance of our algorithm is shown in column "improved".

subject heading	doc count	no optim	basic	improved
Decrees about assignment to a position and dismissal from office	1997	88,85	88,85	88,85
Supreme organs of executive power	1630	53,48	60,29	62,57
Subjects of the Russian Federation	1211	57,83	60,72	61,82
Registration and systematization of official and legislative documents	787	55,72	59,57	64,03
Business enterprises	394	62,77	66,67	72,42
Government of the Russian Federation	366	15,60	23,00	30,35
Legislation on individual subjects of the Russian Federation	174	36,36	45,13	46,64
Formation, reorganization, and liquidation of juridical person	170	9,14	24,62	40,00
Organs of executive power of subjects of the Russian Federation	157	27,16	32,18	37,70
General norms /about taxes and duties/	143	82,67	85,99	85,99
Armament and defense technology	133	32,22	34,43	52,87
Noncommercial organizations	113	8,89	17,14	38,96
UNO and organizations that is members of UNO	111	33,33	38,59	50,42
Air transportation	87	14,16	27,12	30,25
Energetics	83	45,28	56,64	56,64
Diplomatic representations, consular institutions and others	77	65,67	75,36	75,36
Subjects of scientific and technical activity	75	10,53	25,00	35,95
Water transport	73	56,25	56,25	56,25
General norms /about education/	72	20,78	20,78	40,41
General norms /civil rights/	62	97,20	97,20	97,20
Foreign currency	62	7,40	21,06	26,08
Federal budget	22	0,00	0,00	0,00
Privatization of government and municipal...	20	0,00	68,97	88,89
Protection and usage of of historical and cultural heritage	20	12,90	34,29	34,29

Table 2. Text categorization performance of SVM with different parameter optimization strategies on several topics.

4. Stability of text categorization

In this chapter we explore the stability of text categorization performance. We will use different splits of the collection onto the training set and the test set. By *stability* we mean the situation in which the precision and recall do not change grossly depending on the split.

There are two well-known approaches to the problem of comparing different machine learning methods. The first approach is to fix the split of dataset and use this split for evaluating different methods. For example, there is a fixed split "MOD-APTE Split" of Reuters-21578 dataset [Reu97]. It is recommended to evaluate machine learning methods exactly on this split.

The second approach is to compute mean value of performance metrics for different document splits. This procedure is called *cross-validation*. The cross-validation takes a long time so it is not very popular.

For all these approaches the question about stability of categorization performance remains open.

We consider the following factors that affect classification accuracy:

- the number of positive examples for the topic
- the parameter j of SVM (see chapter 3.4)
- the split of document collection

Let T be a topic and let S be a split of document collection. Our algorithm, described in section 3.4, searches for optimal $j \in \{j_1(T), j_2(T), \dots, j_{10}(T)\}$. For each $j_i(T)$ the precision, recall, and F-measure is evaluated. We can plot a graph of this optimization process. The axes is a precision and a recall.

Figure 1 shows:

- a) Thick line. Precision-recall curve for some topic and some split.
- b) Thin lines. Contour plot of function F-measure. The bigger the precision and recall, the bigger F-measure.
- c) The point with maximum F-measure is selected by circle. The selected point corresponds to optimal j .

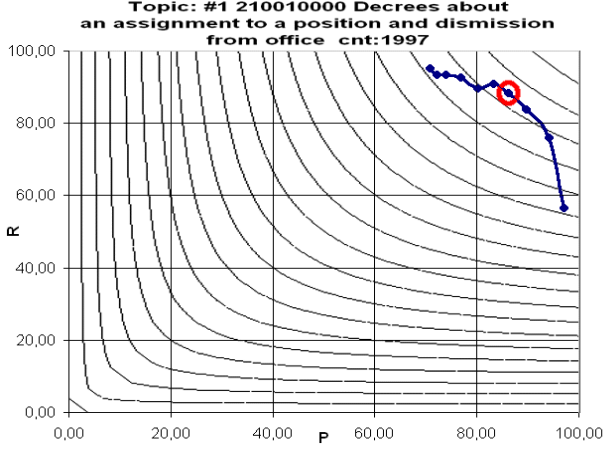


Figure 1. Precision-recall curve for some topic and some split.
The selected point corresponds to optimal j .

We selected 24 different topics with document count from 20 to 2200. For each topic we computed precision-recall curves for 10 random document splits. Each run uses 70% of documents as a training set and 30% of documents as a test set.

In order to estimate the stability of classification numerically, we use the following measure:

$$\begin{aligned}
 Disp_{T_r} &= Avg_k \left(StdDev_i \left(F(T_r, S_i, j_k) \right) \right) = \\
 &= \frac{1}{m} \sum_{k=1}^m \sqrt{\frac{1}{n} \sum_{i=1}^n F(T_r, S_i, j_k)^2 - \left[\frac{1}{n} \sum_{i=1}^n F(T_r, S_i, j_k) \right]^2} \quad (3)
 \end{aligned}$$

Here $F(T_r, S_i, j_k)$ is an estimation of F-measure made on topic T_r , split S_i , and with SVM parameter $j = j_k$. The number of document splits n is equal to 10. The number of steps m for each split is equal to 10. If $Disp_{T_r}$ is low then the stability of classification for topic T_r is high and vice versa.

The experiment has shown that stability is high when the number of positive examples is high. Conversely, when the number of positive examples is low, the stability is likely to be low.

Table 2 shows the result of this experiment.

Figure 2 illustrates the dependency between the number of positive examples and the value of $Disp$.

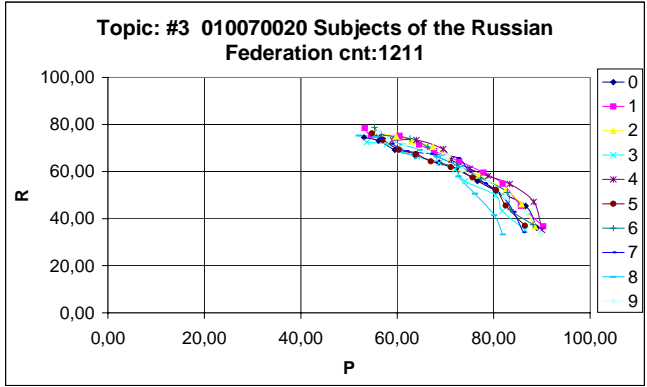


Figure 3. The precision-recall curves for different document splits. For a topic with a big number of positive examples (1211) there is the stability of classification performance.

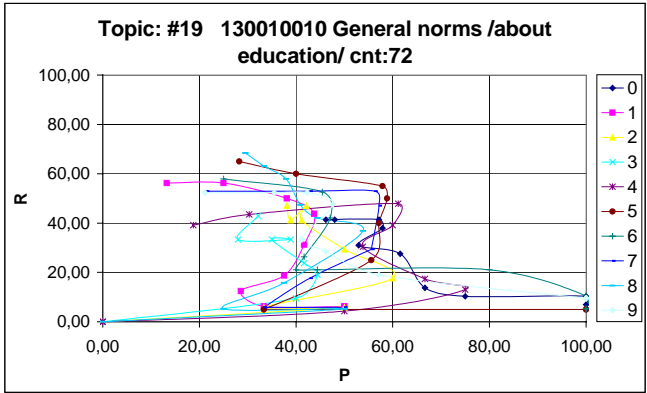


Figure 4. The precision-recall curves for different document splits. For a topic with a low number of positive examples (72) there is the instability of classification performance.

5. Conclusions

This paper analyzes the influence of different parameters of SVM on text categorization performance. The experimental results show that parameter optimization can essentially increase text categorization performance. We describe an algorithm for finding optimal parameters.

References

- [Bu98] Burges C.J.C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2(2):955-974, 1998.

- [CCH92] J. P. Callan, W. B. Croft, and S. M. Harding, "The INQUERY Retrieval System", in Proceedings of the 3rd International Conference on Database and Expert Systems, 1992.
- [Du98] Dumais S., Platt J., Heckerman D., Sahami M. Inductive learning algorithms and representations for text categorization. In Proc. Int. Conf. on Inform. And Knowledge Manage., 1998.
- [Jo98] Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998.
- [Jo99] Joachims T. Making Large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.
- [Le01] Lewis D. Applying Support Vector Machines to the TREC-2001 Batch Filtering and Routing Tasks. Proceedings of TREC-2001 conference.
- [PD00] RF President Decree #511 of March 15, 2000.
- [Reu97] Reuters-21578 text categorization test collection. Distribution 1.0. 1997. <http://www.research.att.com/~lewis>.
- [Va95] Vapnik V. The Nature of Statistical Learning Theory. Springer-Verlag, New York, 1995.
- [YL99] Yang Y., Liu X. A re-examination of text categorization methods. 22nd Annual International SIGIR 1999.