

The SVM With Uneven Margins And Chinese Document Categorisation

Yaoyong Li

Department of Computer Science
Royal Holloway, University of London
Egham, Surrey, TW20 0EX, UK
yaoyong@cs.rhul.ac.uk

John Shawe-Taylor

Department of Computer Science
Royal Holloway, University of London
Egham, Surrey, TW20 0EX, UK
john@cs.rhul.ac.uk

Abstract

We propose and study a new variant of the SVM — the SVM with uneven margins, tailored for document categorisation problems (i.e. problems where classes are highly unbalanced). Our experiments showed that the new algorithm significantly outperformed the SVM with respect to the document categorisation for small categories. Furthermore, we report the results of the SVM as well as our new algorithm on the Reuters Chinese corpus for document categorisation, which we believe is the first result on this new Chinese corpus.

1 Introduction

Document Categorisation (DC), the problem of assigning documents to predefined categories, is an active research area in information retrieval and machine learning. For one category, the DC problem is actually a binary classification problem by classifying a document to the category or not. Many machine learning algorithm have been applied to the DC problem, using a training set of categorised documents to obtain a classifier for one category and then judging the relevance of a document to the category by feeding the document into the classifier.

The support vector machine (SVM) is a well known learning algorithm for linear classifier and has achieved state of the art results for many classification problem, including document categorisation (see Joachims (1998), Yang and Liu (1999)). However, it has been noticed the the performance of the SVM for small category (i.e. category with a small number of relevant documents in collection) was quite poor. Small category in DC problem corresponds to the classification problem with uneven datasets, where the numbers of positive and negative examples are very different. Several kinds of adjustments have been made to the SVM to deal with the uneven datasets. The algorithm we present in this paper, the SVM with uneven margin, is the latest one.

People who are concerned with Chinese information retrieval might notice that, whereas extensive studies have been done to English document categorisation, relatively few results have been reported on Chinese document categorisation (He et al. (2003)). This is mainly because of a lack of Chinese document collection designed particularly for document categorisation. Fortunately, a multilingual corpus RCV2 created recently by Reuters contains a Chinese document collection. The Chinese corpus was manually categorised in the similar way and with the similar high quality with the Reuters English corpora such as Reuters-21578 and the RCV1 corpus. So the Reuters Chinese collection is comparable to the two commonly used English document collections with respect to DC problem. Regarding big difference between English and Chinese, it is interesting to see whether the good methods such as the SVM for English DC problem are able to achieve similar performances for Chinese as for English.

In Section 2 we overview the previous works about adapting the SVM towards uneven datasets. Section 3 describes our new algorithm the SVM with uneven margins to tackle the unbalanced classification problem and explain some interesting relationship between the SVM with uneven margins and the SVM. Section 4 presents the experimental results on the benchmark dataset, the Reuters-21578 corpus, showing our new algorithm outperforms the SVM and justifying the introduction of a margin parameter into the SVM. Section 4 also reports the results of the SVM and the SVM with uneven margins for Chinese DC problem using the Chinese document collection of the RCV2.

2 Previous Works

In the following, we assume that we are given a training set $\mathbf{Z} = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$, where \mathbf{x}_i is the n -dimensional feature vector and y_i ($= +1$ or -1) its label. Our aim is to learn the parameters $\mathbf{w} \in \mathbf{R}^n$ and $b \in \mathbf{R}$ of a linear classifier in feature space, i.e.

$$h_{\mathbf{w},b}(\mathbf{x}) := \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in \mathbf{R}^n .

We start overviewing the previous works with the SVM. The SVM belongs to a family of the so-called margin algorithms in machine learning (see Cristianini and Shawe-Taylor (2000)). The margin of a classifier with respect to a set of training instances refers to the minimal distance of the training points to the decision boundary in feature space. For a binary classification problem, the SVM tries to separate positive patterns from negative patterns using a maximal marginal hyperplane in feature space. The so-called 1-norm soft margin classifier, a commonly used SVM model, can be obtained by solving the optimisation problem (OP1)

$$\text{minimise}_{\mathbf{w}, b, \xi} \quad \langle \mathbf{w}, \mathbf{w} \rangle + C \sum_{i=1}^l \xi_i \quad (1)$$

$$\text{subject to} \quad \langle \mathbf{w}, \mathbf{x}_i \rangle + \xi_i + b \geq 1 \quad \text{if } y_i = +1 \quad (2)$$

$$\langle \mathbf{w}, \mathbf{x}_i \rangle - \xi_i + b \leq -1 \quad \text{if } y_i = -1 \quad (3)$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \dots, m \quad (4)$$

where the parameter C called the cost factor measures the cost of mistakenly classified examples in training set.

Note that the SVM treats positive and negative training examples equally, which may result in poor performance when applying the SVM to some very unbalanced classification problems. A few approaches have been proposed to adapt the SVM to classification problem with uneven dataset. One was presented in Morik et al. (1999), where the cost factor for positive examples was distinguished from the cost factor for negative example to adjust the cost of false positive vs. false negative. The corresponding SVM classifier corresponds to the optimisation problem (OP2)

$$\text{minimise}_{\mathbf{w}, b, \xi} \quad \langle \mathbf{w}, \mathbf{w} \rangle + C_+ \sum_{i:y_i=+1} \xi_i + C_- \sum_{i:y_i=-1} \xi_i \quad (5)$$

$$\text{subject to} \quad \langle \mathbf{w}, \mathbf{x}_i \rangle + \xi_i + b \geq 1 \quad \text{if } y_i = +1 \quad (6)$$

$$\langle \mathbf{w}, \mathbf{x}_i \rangle - \xi_i + b \leq -1 \quad \text{if } y_i = -1 \quad (7)$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \dots, m \quad (8)$$

This approach was implemented by the SVM package *SVM^{light}* (Joachims (1999)) in which an optional parameter j ($= C_+/C_-$) was provided to control different weightings of training errors

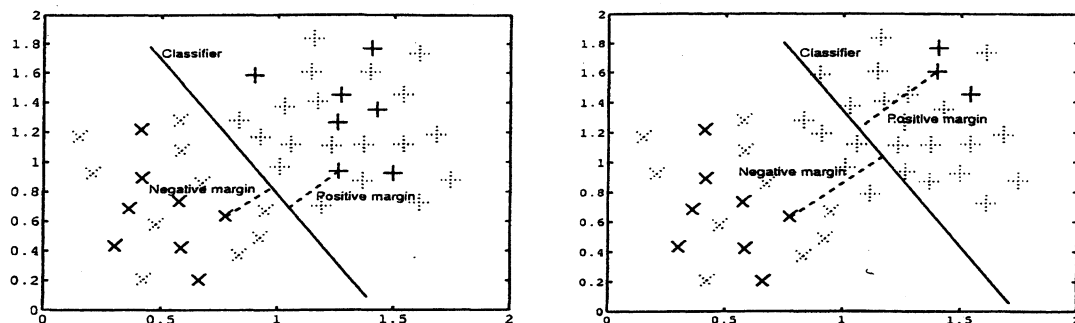


Figure 1: A toy 2-dimensional classification problem and the SVM classifiers. The two graphs illustrate two different kinds of training set. The training set in the left graph is representative of whole dataset but the training set in the right graph is unrepresentative. In both graphs every '+' represents a positive example and every 'x' is for a negative example, and the symbols '+' or 'x' with the solid line represent training examples.

on positive examples to errors on negative examples. Therefore, we denote this method as *the SVM with j -trick*.

Another approach to improve the SVM for very unbalanced problems was based on the observation that, although the SVM outperformed competing methods in document categorisation, it chose a poor threshold when the numbers of positive examples and negative examples were very different (see Lewis et al. (2003) and Zhang and Oles (2001)). Hence it was suggested that, after training the SVM, the bias term b of the classifier should be replaced by a better one, which was obtained by some thresholding strategy (Yang (2001)). The experiments by Lewis et al. (2003) and Yang (2001) showed that one thresholding strategy, the *Scut*, was consistently superior to other strategies on the two Reuters corpora — the Reuters-21578 and the RCV1. So we will only present the results of the *Scut* when comparing the thresholding strategy approach with the SVM with uneven margins.

In our experiments we used the *Scut* procedure described in Lewis et al. (2003). It has two steps. The first step is to choose the optimal threshold by the 5-fold cross-validation on training set. In detail, the training set is split into 5 subsets. Five runs are done, each using four of those subsets for training and the other subset (the *validation* set) for finding the threshold that gives the best performance on the validation set. The second step is to train the SVM on the entire training set, and then to replace the bias term b with a threshold equal to the mean of the thresholds found on the 5 cross-validation folds. Actually the *Scut* can be seen as a heuristic method for estimating the SVM with uneven margins as discussed in next section.

3 The SVM With Uneven Margins

The SVM classifier corresponds to a hyperplane in feature space with maximal margin. Recall that the margin is the distance of training set to the hyperplane. The margin can be regarded as a measure of the error-tolerance ability of classifier, since a classifier is more likely to classify a test instance correctly if it has larger margin. Generally, if the training set is representative of the whole dataset, a classifier with larger margin with respect to training set would have better generalisation performance. However, if the training set is unrepresentative, we should take great care of the margin in the margin learning algorithms, because the maximal margin classifier learned from an unrepresentative training set may have poor generalisation performance as illustrated in Figure 1.

Figure 1 shows a toy 2-dimensional binary classification problem together with two kinds of training set and the corresponding SVM classifiers. The training examples in the left graph of

Figure 1 are representative of the whole dataset, and hence the maximal margin classifier learned from the training set has good generalisation capability. In contrast, the right graph illustrates a situation where the training set is not typical because the number of positive training examples is very small and the training examples aggregate in a small region of feature space. In this case the SVM classifier with maximal margin is not good at prediction since it classifies some positive examples mistakenly. However, if the classification boundary was properly moved away from the positive training examples, the classifier would have better generalisation performance. Hence, if a margin based classifier has to be learned from a very unbalanced training set which has only a few positive examples, it may be beneficial to require the learning algorithm to set the margin with respect to the positive examples (the positive margin) be some larger than the margin with respect with negative examples (the negative margin).

A document categorisation problem with only a few relevant documents for training may have similar situation with that in the right graph of Figure 1. Therefore, in order to achieve better generalisation performance, it is wise to distinguish the positive margin from the negative margin when training the SVM. We introduce a margin parameter τ into the optimisation problem (1) of the SVM to control the ratio of positive margin over negative margin in the SVM, and obtain the following optimisation problem (**OP3**)

$$\text{minimise}_{\mathbf{w}, b, \xi} \quad \langle \mathbf{w}, \mathbf{w} \rangle + C_\tau \sum_{i=1}^l \xi_i \quad (9)$$

$$\text{subject to} \quad \langle \mathbf{w}, \mathbf{x}_i \rangle + \xi_i + b \geq 1 \quad \text{if } y_i = +1 \quad (10)$$

$$\langle \mathbf{w}, \mathbf{x}_i \rangle - \xi_i + b \leq -\tau \quad \text{if } y_i = -1 \quad (11)$$

$$\xi_i \geq 0 \quad \text{for } i = 1, \dots, m \quad (12)$$

The solution of the optimisation problem (9) corresponds to a new classifier — *the SVM with uneven margins*. And τ is the ratio of negative margin to positive margin of the classifier. Our experiments on document categorisation for small category demonstrated that the SVM with uneven margins had much better generalisation performance than the SVM. Before presenting the experimental results, we would like first to describe an interesting relationship between the SVM with uneven margins and the corresponding SVM, which is stated in the following theorem.

Theorem 1 Suppose the training set $\mathbf{Z} = ((\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m))$ be non-trivial. Let $(\mathbf{w}_1^*, b_1^*, \xi_1^*)$ be the solution of the optimisation problem OP1. Then, for any margin parameter $\tau > -1$, the solution $(\mathbf{w}_2^*, b_2^*, \xi_2^*)$ of the optimisation problem OP3 with the parameter $C_\tau = \frac{1+\tau}{2}C$ can be obtained from $(\mathbf{w}_1^*, b_1^*, \xi_1^*)$ by the following transformation

$$\mathbf{w}_2^* = \frac{1+\tau}{2} \mathbf{w}_1^* \quad (13)$$

$$b_2^* = \frac{1+\tau}{2} b_1^* + \frac{1-\tau}{2} \quad (14)$$

$$\xi_2^* = \frac{1+\tau}{2} \xi_1^* \quad (15)$$

The proof can be found in Appendix A. Here are a few comments. First, the theorem shows that the classifier of the SVM with uneven margins can be obtained from the corresponding SVM classifier. It means that, in order to obtain a classifier of the SVM with uneven margins for a classification problem, we can first compute the corresponding SVM classifier by for example running some SVM package and then transform it using the transformation (13) – (15). This is quite helpful to computation.

Secondly, the transformation (13) – (15) shows that the classifier of the SVM with uneven margins is nothing but a shift of the bias term b of the corresponding SVM (up to a scale factor which has no effect on both the classifiers). As we described in Section 2, the Scut is also to find out an optimal bias term b of the SVM for a classification problem. Note that the cost parameter C of the SVM corresponding to the SVM with uneven margins is dependent upon the margin parameter τ , but the cost parameter C in the Scut is fixed when optimising the bias term b . Hence, the Scut can be seen as a heuristic method to estimate the SVM with uneven margins. On the one hand, the close relationship between the Scut and the SVM with uneven margins can give an explanation of the effectiveness of the Scut method for the document categorisation problem for small category, following the arguments given above about the advantage of introducing a margin parameter into the SVM. On the other hand, the margin parameter τ , the ratio of the negative margin to the positive margin, is more sensible than the bias term. Actually we can take advantage of the meaning of margin parameter in some classification problems. For instance, given a classification problem with only a few positive training examples, we know that a small value of marginal parameter should be used in the SVM with uneven margin, but we have no clear idea about what is a suitable value of threshold for the SVM.

Thirdly, note the theorem is for the so-called 1-norm soft margin SVM classifier, a commonly used SVM model. We know there are other models of the SVM such as the maximal margin classifier and the 2-norm soft margin SVM (Cristianini and Shawe-Taylor (2000)). We can also introduce margin parameter into those SVM models and can have similar transformations with that in Theorem 1.

Finally, we know that the SVM without the bias term is more useful than the SVM with bias term for some problems like the adaptive document filtering (Cancedda et al. (2003)). We can introduce a margin parameter into the SVM without the bias term as well, However, obviously we are unable to apply the Scut method to the SVM without the bias term at all.

4 Experimental Results

We conducted experiments on document categorisation to test the SVM with uneven margins on two document collections — the well known Reuters-21578 collection, and the Chinese collection of RCV2. The Reuters-21578 collection, containing English news stories from Reuters news agency, is a benchmark in recent studies of document categorisation. It is fair to implement some experiments on this collection at first to compare the SVM with uneven margins with other related algorithms. The Chinese collection of RCV2, also from Reuters news agency, was created recently and it is quite suitable for studying Chinese document categorisation. It is interesting to see whether we can obtain similar results of the SVM and its variants for document categorisation in Chinese as in English. For this purpose, we tried to use the same experimental settings on the Chinese collection as on the Reuters-21578 as much as possible.

There were many categories annotated in the Reuters-21578 as well as in the Chinese collection of RCV2. We conducted the experiments on some of the categories (as described below). For each category we learned a binary classifier and evaluated its performance on test set by computing a value of F_1 . The F_1 , a common performance measure in information retrieval, is computed by $2pr/(p+r)$, where p is Precision and r is Recall. Then we can obtain a macro-averaged value of F_1 by averaging the F_1 over a group of categories. In order to gain a better insight into the behaviour of the algorithms with respect to category size, we report three different averages, the average over all categories (ALL), the 10 largest (TOP10), and the categories with less than 10 relevant training documents (the 30 smallest (LAST30) for Reuter-21578 and the 10 smallest (LAST10) for Chinese collection of the RCV2).

One main objective of the experiments was to justify the introduction of the margin parameter

into the SVM. To this end, we report the results of comparison of the SVM with uneven margins with other three related algorithms described in Section 2, as well as the results of other two particular experiments. We also present the experimental results on Chinese document categorisation. In the following we first describe the experimental settings used in all the experiments. Then we present the particular details and results of the experiments in the subsections.

The first step of the experiments was to determine an appropriate value of model parameter in learning algorithms. The model parameter refer to the marginal parameter τ in the SVM with uneven margins, the threshold b in the Scut, and the parameter j in the SVM with j -trick. In order to make a fair comparison among the algorithms, we adopted the same process of model selection for both the SVM with uneven margin and the SVM with j -trick as for the Scut. In particular we used the 5-fold cross-validation on training set to choose the optimal values of the model parameters for each of categories considered. The optimal value of margin parameter τ was picked up from the 15 values: $\{-0.1, 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 2\}$, and the best j from another 15 values: $\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 2, 3, 4, 5, 8\}$. In contrast, the optimal b was chose from every possible values in the Scut. Then a classifier was learned from the whole training set using the optimal parameter and finally it was evaluated on test set.

One problem about model selection using the cross-validation is that the value of model parameter obtained might not be trustworthy for a category which has only very small number of positive training examples, because probably there is even none positive example in a validation set. For the SVM with uneven margins, since a large positive margin is helpful, we simply set the marginal parameter τ be as small as 0.1 for the categories with less than 5 positive examples in training set. However, for both the Scut and the SVM with j -trick, we have no idea about what value of the threshold b or the parameter j is good for a very small category. So we had to use the values of the two parameters obtained by cross-validation for all categories.

In all the experiments described here, the package *SVM^{light}* (version 3.50)¹ was used to train an SVM classifier. All the parameters were left as default values, except the optional parameter j in the experiments for the SVM with j -trick where an optimal value of j should be found.

4.1 Comparing the four algorithms on the Reuters-21578 Collection

We used the commonly used “ModeApte” split of the Reuters-21578 collection. This split led to 9603 training documents, 3299 test documents, and 90 categories with relevant training documents ranging from 1 to 2877. The documents were preprocessed in the usual way, consisting of downcasing the text, replacing numbers and punctuations with whitespace, removing the words occurring in a stop list, applying Porter stemmer to the words, and finally breaking the text into terms at whitespace. This resulted in 20494 unique terms. Documents were then represented according to the normalised $tf * idf$ scheme: each document d in the collection was represented by a 20494-dimensional vector \mathbf{x}_d , the i th element of which was computed by $\log(1 + tf_{d,i}) * \log(m/df_i)$, where $tf_{d,i}$ was the number of occurrence of the term i in the document d , and df_i was the number of documents in the collection containing the i th term, and m was the total number of documents in the collection. Finally all vectors were normalised to have unit length.

We applied the SVM with uneven margins and other three algorithms to the feature vectors. The results are presented in Table 1. First of all, the SVM with uneven margins gave better results than the SVM and the SVM with j -trick, in particular for the small categories. This means that introducing a marginal parameter into the SVM is indeed quite useful to document categorisation for small categories. Secondly, note that the results for the SVM with uneven margins is similar with those for the Scut. This is not surprising because the Scut is a good

¹ Available from http://www.joachims.org/svm_light

Table 1: Compare the four algorithms on Reuters-21578 dataset. In the table “SVMUM” refer to the SVM with uneven margins and “ j -trick” for the SVM with j -trick.

	ALL	TOP10	LAST30
Macro-Average F_1			
SVM	0.401	0.850	0.000
j -trick	0.436	0.862	0.000
Scut	0.644	0.874	0.462
SVMUM	0.650	0.872	0.484

approximation of the SVM with uneven margins. However, the SVM with uneven margins gives a bit better result for small categories than the Scut, since we set the marginal parameter τ be a small value as 0.1 for very small categories in the experiments but we can not find a better way to pick up a value for the threshold b in the Scut rather than using cross-validation on training set.

4.2 More results to demonstrate the advantages of the SVM with uneven margins

Recall that, in order to justify the introduction of the margin parameter into the SVM, we speculated in Section 3 that the margin parameter would be beneficial to the classification problem with unrepresentative training set. In order to check this speculation, we present the results of two more experiments which investigated the relationship between the optimal margin parameter of the SVM and the number of positive training examples.

In the first experiment we considered one large category of Reuters-21578 dataset — the category “acq” which has 1650 relevant documents in training set. We formed several classification problems with different kinds of unbalanced training set from the categorisation problem “acq” by removing different number of positive examples from training set. The SVM with uneven margins was then applied to each of the problems. In detail, ten different sizes of training set were obtained by keeping all the negative training examples and picking up some positive training examples from training set randomly. The number of the positive examples in new training set is in $\{2, 5, 10, 20, 50, 100, 200, 500, 1000, 1500\}$. For every size of training set, we conducted ten experiments by choosing the positive examples randomly in ten times and averaged the results over them. The results of the ten classification problems, including the optimal values of margin parameter τ , the two values of F_1 respectively for the optimal τ and for $\tau = 1$, are plotted in the left graph of Figure 2. We can see that the optimal value of τ for the small number of positive training examples is much less than 1 which is used in the SVM, and, correspondingly, the F_1 for the optimal τ is quite larger than that for $\tau = 1$. And the difference between the two values of F_1 decreases as the number of positive examples increases.

The right graph of Figure 2 shows the results for several groups of the Reuters-21578 categories. Actually we divided the 90 categories of Reuters-21578 into ten groups according to their sizes, in order to keep consistent with the above experiment. The i th group contained such the categories that each of them has m_i positive training examples, where m_i satisfies $N(i) \leq m_i \leq N(i+1)$ and $N(i) \in \{0, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2877\}$. Ten averages were obtained by averaging the results over each of the ten groups and were plotted in the right graph in Figure 2. The results are the optimal margin parameter τ and two values of F_1 respectively for the optimal τ and $\tau = 1$. We see three similar curves in this graph with those in the left graph.

The result of the two experiments consistently show the relationship we speculated between the margin parameter and the number of positive training examples, that is, a small number of positive examples corresponds to a small value of margin parameter. And the performance of

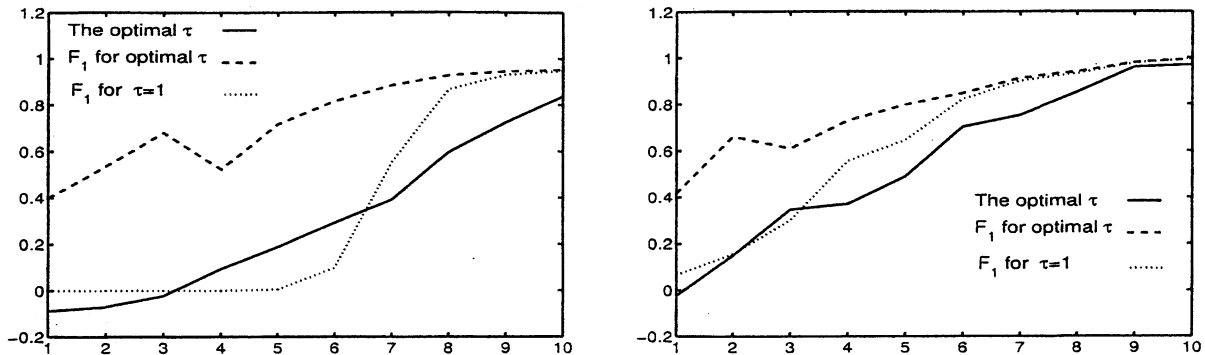


Figure 2: Two experiments demonstrate the benefit gained from the SVM with uneven margins. The left graph shows the results of classification problems with ten different numbers of positive training examples: $\{2, 5, 10, 20, 50, 100, 200, 500, 1000, 1500\}$, which were derived from the Reuters-21578 categorisation problem “acq”. And the right graph shows the averaged results over each of 10 groups of the Reuters-21578 90 categories with different sizes. For each of the two experiments, we plot the optimal value of the margin parameter τ , the two values of F_1 for the optimal τ and $\tau = 1$, respectively.

the SVM with the optimal margin parameter is significantly better than the SVM for document categorisation for small categories.

4.3 Results on the Chinese Collection of RCV2

The Chinese collection of RCV2 comprises 28964 Chinese news stories, spanned from 20 August 1997 until 19 August 1998. We used the 15278 documents in the first six months for training and the others for test. There are more than one thousand categories annotated in this collection. In our experiments only the 61 categories was used, which exist in the RCV1 collection as well and each of which has at least one relevant document in both training and test sets.

We tried to use the so-called “bag of words” model, which was mostly used in English document categorisation, to represent Chinese document. A problem we faced was, unlike English and other western languages, there is no natural delimiter between two neighbouring words in Chinese sentence. Therefore, we had to segment Chinese sentence into words. Chinese word segmentation has been extensively studied but perfect performance has not been reached².

In our experiments a free Chinese segmentation package³ compiled by Zhibiao Wu was used to segment Chinese document into a set of words. This package consists of a Perl script and a frequency dictionary of 44405 Chinese words. Since the package is for simplified Chinese but the Chinese document in the RCV2 is in traditional Chinese, we had to transfer the documents into simplified Chinese at first. Note that both the transfer and the word segmentation would introduce errors into the representation of document (but this kind of errors did not occur in the representation of English document). After word segmentation and removing the words appearing less than three times in the collection, we obtained 13720 words (including some English words). Unlike in the preprocessing of English document, we did not use any tokenisation nor stop list to process Chinese document. Finally we computed an normalised $tf * idf$ representation for every document in the usual way, as described in Subsection 4.1.

We applied the SVM with uneven margins and other three algorithms to the feature vectors

²Actually, so far there has not yet been an agreement on definition of Chinese word among the researchers of Chinese language processing. Given a sequence of Chinese characters, some researchers think it is a word but others may think it is a combination of two or more words.

³Available from <http://www ldc.upenn.edu/Projects/Chinese/>

Table 2: Experimental results on the Chinese collection of RCV2. “SVMUM” refer to the SVM with uneven margins and “ j -trick” for the SVM with j -trick.

	ALL	TOP10	LAST10
Macro-Average F_1			
SVM	0.374	0.867	0.000
j -trick	0.436	0.872	0.000
Scut	0.502	0.876	0.109
SVMUM	0.506	0.874	0.140

of the Chinese documents. The experimental results were summarised up in Table 2. Firstly we compare the results of Chinese document categorisation with those of English document categorisation. We can see that the algorithms achieved similar performance for large categories, but obtained quite worse results for small categories for Chinese. Regarding that some extra errors were introduced when preprocessing the Chinese documents, we can speculate that the SVM and the variants are quite robust to the noise in feature vectors for large categories but are sensitive to the data noise for small categories. Then we make a comparison among the four learning algorithms themselves. We can see that they have the similar behaviours for Chinese categorisation with those for English. That is, the two algorithms the Scut and the SVM with uneven margins achieved better performance in small categories than other two algorithms, and the SVM with uneven margins is slightly better than the Scut.

5 Conclusion

Machine learning concerns learning general target functions from specific training examples. Learning algorithm and training set are two fundamental elements in machine learning. Good results can often be achieved for a learning problem if the training example are representative of whole dataset. However, for some learning problems only some unrepresentative examples are available for learning. In this case some specific strategy should be adopted in learning algorithm to compensate for the poor quality of training examples.

Introducing a margin parameter into the SVM, which led to a new learning algorithm — the SVM with uneven margins, is a strategy for the SVM to deal with some kind of poor quality of training examples. In particular, if we want to use the SVM to solve a very unbalanced binary classification problem where only a few positive training examples are available and are very unlikely to be representative, a larger positive margin in the SVM classifier should be beneficial to the generalisation performance of the SVM, as we argued in Section 3. The margin parameter τ in the SVM with uneven margins is the ratio of negative margin to positive margin. The optimal value of τ can result in much better results than the default value 1 in the SVM for the document categorisation for small categories. Our experiments on document categorisation have justified the introduction of the margin parameter into the SVM, as shown in Section 4.

The Scut is a effective method to adapt the SVM to the classification problem with very unbalanced training set. We can regard the Scut as a heuristic and good approximation of the SVM with uneven margin, which can explain why the Scut have quite better performance than the original SVM for the document categorisation for small categories.

In this paper we also reported the results of the SVM and the SVM with uneven margins on Chinese document categorisation. We believe the result is among the first ones on the new Chinese document collection — the Chinese collection of the RCV2. Our results showed that the SVM and the variants achieved similar performance on Chinese with those on English for document categorisation for large categories, despite big difference between English and Chinese.

However, the results of Chinese document categorisation for small categories were much worse than those for English. We think that reducing the noise in feature vectors of Chinese documents, which needs better preprocessing of Chinese document than what we did, may lead to better results.

Acknowledgements

The work described in this paper has been supported by the European Commission through the IST Programme under Contract IST-2000-15431 (KerMIT).

A Proof of Theorem 1

Proof. First we will show that any feasible solution of the optimisation problem OP1 can be transformed into a feasible solution of problem OP3 by using the transformation in Theorem 1. Let $(\mathbf{w}_1, b_1, \xi_1)$ be a feasible solution of the optimisation problem OP1. We will prove that $(\mathbf{w}_2, b_2, \xi_2)$ obtained by the transformation (13) – (15) in Theorem 1, namely,

$$\mathbf{w}_2 = \frac{1+\tau}{2}\mathbf{w}_1, \quad b_2 = b_1 + \frac{1-\tau}{2}(1-b_1), \quad \xi_2 = \frac{1+\tau}{2}\xi_1 \quad (16)$$

is a feasible solution of the optimisation problem OP3.

Since $(\mathbf{w}_1, b_1, \xi_1)$ satisfies the constraints inequalities in (2) and (3) of OP1, not loss of generality, we can assume that we have the following equations and strict inequalities,

$$\langle \mathbf{w}_1, \mathbf{x}_i \rangle + \xi_{1i} + b_1 = 1 \quad \text{for } i = 1, \dots, m_0 \quad (17)$$

$$\langle \mathbf{w}_1, \mathbf{x}_i \rangle + \xi_{1i} + b_1 > 1 \quad \text{for } i = m_0 + 1, \dots, m_1 \quad (18)$$

$$\langle \mathbf{w}_1, \mathbf{x}_i \rangle - \xi_{1i} + b_1 = -1 \quad \text{for } i = m_1 + 1, \dots, m_2 \quad (19)$$

and

$$\langle \mathbf{w}_1, \mathbf{x}_i \rangle - \xi_{1i} + b_1 < -1 \quad \text{for } i = m_2 + 1, \dots, m \quad (20)$$

where $1 \leq m_0 \leq m_1 \leq m_2 \leq m$. Furthermore, we know from the constraints (4) of the OP1 that

$$\xi_{1i} \geq 0 \quad \text{for } i = 1, \dots, m \quad (21)$$

By using the transformation (16) and relationships (17) – (21), we can prove that $(\mathbf{w}_2, b_2, \xi_2)$ satisfies all the constraints of the OP3. Firstly, for $i = 1, \dots, m_0$ we have

$$\begin{aligned} \langle \mathbf{w}_2, \mathbf{x}_i \rangle + \xi_{2i} + b_2 &= \frac{1+\tau}{2} \langle \mathbf{w}_1, \mathbf{x}_i \rangle + \frac{1+\tau}{2} \xi_{1i} + b_1 + \frac{1-\tau}{2}(1-b_1) \\ &= \frac{1+\tau}{2} (1-b_1) + b_1 + \frac{1-\tau}{2}(1-b_1) \\ &= 1 - b_1 + b_1 \\ &= 1 \end{aligned}$$

where the equations (16) and (17) was used.

Secondly, for $i = m_0 + 1, \dots, m_1$, we have,

$$\begin{aligned} \langle \mathbf{w}_2, \mathbf{x}_i \rangle + \xi_{2i} + b_2 &= \frac{1+\tau}{2} \langle \mathbf{w}_1, \mathbf{x}_i \rangle + \frac{1+\tau}{2} \xi_{1i} + b_1 + \frac{1-\tau}{2}(1-b_1) \\ &> \frac{1+\tau}{2} (1-b_1) + b_1 + \frac{1-\tau}{2}(1-b_1) \\ &= 1 - b_1 + b_1 \\ &= 1 \end{aligned}$$

where equations (16) were used in the first equality and the inequalities (18) and the relationship $\tau > -1$ assumed in Theorem 1 were used in the second inequality.

Thirdly, for $i = m_1 + 1, \dots, m_2$, using (16) and (19) we have,

$$\begin{aligned} \langle \mathbf{w}_2, \mathbf{x}_i \rangle - \xi_{2i} + b_2 &= \frac{1+\tau}{2} \langle \mathbf{w}_1, \mathbf{x}_i \rangle - \frac{1+\tau}{2} \xi_{1i} + b_1 + \frac{1-\tau}{2} (1-b_1) \\ &= \frac{1+\tau}{2} (-1-b_1) + b_1 + \frac{1-\tau}{2} (1-b_1) \\ &= -\tau - b_1 + b_1 \\ &= -\tau \end{aligned}$$

Fourthly, for $i = m_2 + 1, \dots, m$, using equations (16), inequalities (20) and $\tau > -1$ we have

$$\begin{aligned} \langle \mathbf{w}_2, \mathbf{x}_i \rangle - \xi_{2i} + b_2 &= \frac{1+\tau}{2} \langle \mathbf{w}_1, \mathbf{x}_i \rangle - \frac{1+\tau}{2} \xi_{1i} + b_1 + \frac{1-\tau}{2} (1-b_1) \\ &< \frac{1+\tau}{2} (-1-b_1) + b_1 + \frac{1-\tau}{2} (1-b_1) \\ &= -\tau - b_1 + b_1 \\ &= -\tau \end{aligned}$$

Finally, as $\tau > -1$, from the transformations (16) and inequalities (21) we have

$$\xi_{2i} \geq 0 \quad \text{for } i = 1, \dots, m$$

We have shown that $(\mathbf{w}_2, b_2, \xi_2)$ satisfied all the constraints of OP3. So, $(\mathbf{w}_2, b_2, \xi_2)$ is a feasible solution of the optimisation problem OP3.

Similarly, we can prove that, if $(\mathbf{w}_2, b_2, \xi_2)$ is a feasible solution of OP3, then $(\mathbf{w}_1, b_1, \xi_1)$ obtained by

$$\mathbf{w}_1 = \frac{2}{1+\tau} \mathbf{w}_2, \quad \xi_1 = \frac{2}{1+\tau} \xi_2, \quad b_1 = \frac{2}{1+\tau} b_2 + \frac{\tau-1}{\tau+1} \quad (22)$$

is a feasible solution of OP1.

Now we are in the position to prove Theorem 1, namely, if $(\mathbf{w}_1^*, b_1^*, \xi_1^*)$ is the optimal solution of the problem OP1, then $(\mathbf{w}_2^*, b_2^*, \xi_2^*)$ obtained by the application of transformation (16) to $(\mathbf{w}_1^*, b_1^*, \xi_1^*)$, i.e.

$$\mathbf{w}_2^* = \frac{1+\tau}{2} \mathbf{w}_1^*, \quad b_2^* = b_1^* + \frac{1-\tau}{2} (1-b_1^*), \quad \xi_2^* = \frac{1+\tau}{2} \xi_1^* \quad (23)$$

is the optimal solution of the problem OP3. For any feasible solution $(\mathbf{w}_2, b_2, \xi_2)$ of the optimisation problems OP3, we know that $(\mathbf{w}_1, b_1, \xi_1)$ obtained by the transformation (22) is a feasible solution of OP1. Using the two transformations (22) and (23), we have

$$\begin{aligned} \langle \mathbf{w}_2, \mathbf{w}_2 \rangle + C_\tau \sum_{i=1}^l \xi_{2i} &= \left(\frac{1+\tau}{2} \right)^2 \langle \mathbf{w}_1, \mathbf{w}_1 \rangle + C_\tau \frac{1+\tau}{2} \sum_{i=1}^l \xi_{1i} \\ &\geq \left(\frac{1+\tau}{2} \right)^2 \left(\langle \mathbf{w}_1^*, \mathbf{w}_1^* \rangle + C \sum_{i=1}^l \xi_{1i} \right) \\ &= \langle \mathbf{w}_2^*, \mathbf{w}_2^* \rangle + C_\tau \sum_{i=1}^l \xi_{2i}^* \end{aligned}$$

where the second inequality used the fact that the $(\mathbf{w}_1^*, b_1^*, \xi_1^*)$ is the optimisation solution of OP1 and the relationship $C_\tau = \frac{1+\tau}{2} C$ assumed in Theorem 1. The above relationship shows that the $(\mathbf{w}_2^*, b_2^*, \xi_2^*)$ is indeed the solution of the optimisation problem OP3. \square

References

- Cancedda, N., N. Cesa-Bianchi, A. Conconi, C. Gentile, C. Goutte, T. Graepel, Y. Li, J.M. Renders, and J. Shawe-Taylor. 2003. Kernel methods for document filtering. In E. M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Eleventh Text Retrieval Conference (TREC 2002)*. The NIST.
- Cristianini, N. and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK.
- He, J., A.-H. Tan, and C.-L. Tan. 2003. On machine learning methods for chinese documents classification. *Applied Intelligence*, 18:311–322.
- Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, pages 137–142, Berlin. Springer.
- Joachims, T. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods—Support Vector Learning*, pages 169–184, Cambridge, MA. MIT Press.
- Lewis, D.D., Y. Yang, T. Rose, and F. Li. 2003. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, to appear.
- Morik, K., P. Brockhausen, and T. Joachims. 1999. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *Proc. 16th Int'l Conf. on Machine Learning (ICML-99)*, pages 268–277, San Francisco, CA. Morgan Kaufmann.
- Yang, Y. 2001. A study on thresholding strategies for text categorization. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01)*, pages 137–145, New York. The Association for Computing Machinery.
- Yang, Y. and X. Liu. 1999. A re-examination of text categorization methods. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49.
- Zhang, T. and F.J. Oles. 2001. Text categorization based on regularized linear classification methods. *Information Retrieval*, 4(1):5–31.