
Evaluation of kernel function modification in text classification using SVMs

Yangzhe Xiao

Department of Computer Science, Rutgers University. NJ 08854 USA

YXIAO@PAUL.RUTGERS.EDU

Abstract

In this study, we propose a method to modify the RBF kernel function by incorporating category related information into it in text classification. We use two approaches to measure the relative power of a term in discriminating a category from others. Such information is incorporated in mapping from input space into feature space by modifying the RBF kernel function. This allows the use of both labeled and unlabeled documents in calculating TFIDF. We apply the kernel modification methods on different sizes of training sets from Reuter-12578. We also compare our methods with a Chi-square feature selection method. The new kernel modification methods give at least as good results as the Chi-square feature selection.

1. Introduction

The research of automated categorization (or classification) of texts into predefined categories has focused on machine-learning techniques, such as Decision Tree, Naïve Bayes probabilistic classifiers, Linear Least Square Fitting, the Rocchio Method, Neural Networks, k Nearest Neighbors and Support Vector Machines (SVMs). These techniques generally infer class identity of unclassified documents by learning characteristics of categories from preclassified documents. Machine-learning techniques have proven effective and feasible (Sebastiani, 2002).

Support vector machines (SVMs) were introduced for text classification by Joachims (1998). Their generalization performance either matches or is significantly better than that of competing methods. Although the popularity of SVMs exploded in the late 90's, the subject can be said to have started in the late 70's. It is based on the concept of VC dimension and structural risk minimization. The basic idea is to find a decision surface from all the surfaces in multi-dimensional space that separates positive training examples from negative ones by the widest possible

margin (Burges, 1998). Kernel functions are used in SVMs to improve accuracy. It is possible to transform samples from the original input space, which are not linearly separable, into a feature space defined by the kernel function where they become linearly separable. The most common kernel functions are polynomial, radial basis function (RBF) and hyperbolic tangent kernels. Other kernel functions can be used as long as they satisfy Mercer's condition. If a kernel does not satisfy Mercer's condition, the quadratic optimization problem solved in SVMs will have no solution.

SVMs have advantages in text categorization. They tend to be fairly robust to overfitting and can scale up to considerable dimensionalities (Joachims, 1998). However, one of the limitations of SVMs lies in the choice of the kernel. Good choices of kernels can make linearly inseparable data become separable or increase the margin in the feature space. Therefore, good kernels can help in decreasing the generalization error. Some work has been done in improving the performance of SVMs by constructing appropriate kernel functions. One example is utilizing the prior knowledge of locality in images to create kernels (Scholkopf, Simard, Smola, et al, 1998). In natural images correlations over short distances are much more reliable as features than long-range correlations, so there is an advantage for using the constructed kernel corresponding to a dot product in a space that is spanned mainly by local correlations between pixels. Text classification with common kernels is based on distances between documents in vector space. The semantic relations between terms are not taken into account. Documents that talk about related topics using different terms are possibly mapped to very distant regions in feature space. Siolas and D'Alche-buc (2000) encoded a semantic network into kernel. So documents are implicitly mapped into a "semantic space". The semantic distances among terms are derived from a hierarchical semantic database of English words, WordNet. Another approach to get semantic information is Latent Semantic Indexing (LSI), which extracts the semantic relations between terms (Christianini, Shawe-Taylor & Lodhi 2001). Their results all showed improved performance over the standard kernels. This study is motivated by embedding

prior knowledge into the kernel function. We incorporate category discriminating information into the RBF kernel and evaluate if the modification of the kernel can improve the accuracy of SVMs in text classification.

2. Modification of the kernel function

In text categorization, the first question that needs to be answered is how to represent documents. Usually documents are represented as vectors. The dimensionality of a vector is the same as the number of candidate terms selected according to some criteria. One popular method is using TFIDF value for each term with respect to each document, which is defined as

$$tfidf(t_k, d_j) = tf(t_k, d_j) * \log(|D| / df(t_k)).$$

where $tf(t_k, d_j)$ denotes the number of times t_k occurs in document d_j , and $|D|$ is the cardinality of the set of documents D . $df(t_k)$ is document frequency of term t_k . i.e. the number of documents that have term t_k . In order for the values to fall in the $[0, 1]$ interval and to reduce the influence of different length of documents, the values of TFIDF are normalized as

$$w_{kj} = \frac{tfidf(t_k, d_j)}{\sqrt{\sum_{s=1}^{|T|} (tfidf(t_k, d_j))^2}},$$

where $|T|$ is the dimensionality of the document vectors (Sebastiani, 2002).

TFIDF does not use any category information, which allows us to calculate TFIDF based on documents from both the training and testing set. By using testing documents, we hope they can give relatively more information in calculating TFIDF especially in the case when only a small amount of training samples are available. Small training-set size poses a challenge for any machine-learning technique, since the samples may represent only a portion of the underlying distribution of the data. Such incompleteness can result in poor generalization of the classifiers learned from biased samples. Using unlabeled data may alleviate this problem.

We know preclassified data contains more valuable information about the characteristics of categories than does unlabeled data. For text classification, if a term appears frequently in one category but seldom in other categories, this term is very informative in discriminating documents of that category from others. Two measurements for the importance of a term in distinguishing category c and non- c are proposed. The first one is the Chi-square test on every term. Yang and Pedersen (1997) compared five measurements in term selection, and found that the Chi-square test and information gain gave the best performance. The Chi-

square test is used instead of both, since these two measurements are highly correlated. Using the two-way contingency table of a term t and a category c , where A is the $df(t)$ in c , B is $df(t)$ in non- c , C is the number of documents of c that do not have t , D is the number of documents of non- c that do not have t , N is the total number of documents,

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)}.$$

This χ^2 statistics has a natural value of zero if t and c are independent. The higher $\chi^2(t, c)$ value is, the more important t is in discriminating c and non- c . χ^2 measurement gives equivalent weights for the presence and absence of a term.

The second one measures the distribution of t among c and non- c documents, defined as

$$E(t, c) = \frac{A}{A + C} (1 - (-p \log_2 p - q \log_2 q)),$$

where

$$p = \frac{\frac{A}{A + C}}{\frac{A}{A + C} + \frac{B}{B + D}},$$

$$q = 1 - p.$$

The parenthesized part in $E(t, c)$ is the measurement of entropy with a range from 0 to 1, measuring unbalanced distribution of a term. It has value 1 if a term only appears in c or non- c documents. This doesn't mean the term is category-relevant. For instance, a term that appears only in one document of c will have value 1 in the parenthesized part, but such term should not be valued too much. A term must appear in most documents of that category to be considered very informative, which is the role that the first part of the formula plays. Therefore, $E(t, c)$ emphasizes both the unbalanced distributions across categories and high intra-category document frequency of a term.

The two measurements mentioned above reveal the relative discriminating power of a term. We can incorporate these metrics during the mapping of input space to feature space. The RBF kernel function

$$K(x, y) = \exp(-\|x - y\|^2 / 2\sigma^2),$$

can be modified by introducing a matrix M , resulting in

$$K(x,y)=\exp(-\|Mx-My\|^2/2\sigma^2).$$

M represents semantic relations among terms in Christianini, Shawe-Taylor & Lodhi’s paper (2001). Here, we use the same RBF kernel modification but different M , which is derived from the measurements of discriminating power of terms. When using SVMs to learn a classifier between category c and non- c , $\chi^2(t,c)$ or $E(t,c)$ are used to form the matrix. Let M be a square diagonal matrix that has the same dimensionality as that of document vectors. The diagonal position of M corresponding to t has value of $(I+\chi^2(t,c))$ or $(I+E(t,c))$ depending on which measurement is used. By incorporating such matrix into the kernel, the modified kernel weighs differently for each term in mapping from the original input space to the feature space. For each classifier, we need to calculate $\chi^2(t,c)$ or $E(t,c)$ for each candidate term t in order to construct M . This adds computational complexity to the learning process.

3. Result

The Reuters-21578 data set was collected from the Reuters newswire in 1987. The “ModApte” split was used in this study, which had a corpus of 9630 training documents and 3299 test documents. Of the 135 topic categories only the most frequent 10 (acq, corn, crude, earn, grain, interest, money-fx, ship, trade, wheat) were used. All the documents are preprocessed by removing the common non-content bearing stopwords from every document.

Four methods are used for comparison. **T**: The feature set has all terms from the preprocessed training and testing documents. The aforementioned normalized TFIDF is used to represent documents. **CW**: It uses the same document vectors from method **T**. Kernel modification with term relevance values calculated with $\chi^2(t,c)$ is adopted in SVMs. **EW**: It is similar to **CW** except $E(t,c)$ measurement is used instead of $\chi^2(t,c)$. **TS**: The feature set is drawn from the training-set documents. The feature terms are selected by ranking the $\chi^2(t,c)$ value of each term from the training set. If a term has different $\chi^2(t,c)$ values for different categories, the maximum among all categories is used as the ranking value of the term. The top 3000 high-ranking terms are used as feature terms. Yang and Pedersen (1997) showed that 3000 terms selected by the Chi-square test from these documents were enough for a good result, although they used different classification algorithms. Normalized TFIDF is applied on the selected 3000 feature terms to represent documents. All the training and testing were carried out with the software package SVM-light and the standard RBF kernel was used in **T** and **TS**. For each method, ten binary-classifiers were learned corresponding to 10 categories.

We used the four methods on different sizes of training samples and the results are listed as in Table 1. The effectiveness measurement used here is F value defined as

$$F = \frac{2\pi\rho}{\pi + \rho},$$

where precision π and recall ρ are averages over the 10 categories (Sebastiani, 2002). F depends equally on π and ρ , and has the range from 0 to 1.

Table 1. F values of four methods.

SampleSize	T	CW	EW	TS
20	0.185	0.211	0.223	0.197
100	0.574	0.602	0.613	0.593
250	0.690	0.726	0.728	0.717
500	0.779	0.790	0.791	0.788
1000	0.810	0.819	0.820	0.814
1880	0.837	0.855	0.851	0.855
5100	0.830	0.861	0.867	0.880

In Table 1, the 5100 training samples contain all the training documents from all ten categories, which have different numbers of training documents per category. For other cases, a training set consists of an equal number of documents randomly drawn from each category. The best F value for each sample size is listed in bold font. Although **EW** has more best-values than others, the 4 methods (**T**, **CW**, **EW**, **TS**) have very similar results on different training sample sizes. The difference among the four methods is small. As the size of training set increases, the results of all 4 methods get better

4. Discussion

To the best of our knowledge, the idea of incorporating category-discriminating capability of each term into the kernel mapping in SVMs are first proposed in this paper. We used all terms from the training and testing sets as features in the three methods-- **T**, **CW**, and **EW**. We intended to alleviate the small training-set problem by using terms from the testing set. However, the results were not very encouraging, since the results of these three methods were not much better than that of **TS** (only using 3000 terms from the training set). Therefore, we did not expand our experiment.

Based on the results in Table 1, there is no significant difference among all the 4 methods. If this is the fact, it is a surprise to see **T** and **TS** have the same results. Since **T**

uses all terms from the testing set beside the training set to form a feature set, it is very likely that some features (terms) of the testing set do not exist in the training documents. These features do not provide any useful information in the training process. Even more, they may add noise to the classification problem especially when using the algorithms that have a constraint on dimensionality. SVMs are not quite sensitive to dimensionality and overfitting, which may help to explain the observed similar performance of **T** and **TS**.

Our modification to the kernel incorporates the category-discriminating information of terms measured by the $\chi^2(t,c)$ or $E(t,c)$. These two measurements use only labeled documents from the training set, since they need category information about each document. The modification method gives different weights to the terms according to the two category-based measurements. The terms that are not in the training-set documents get less emphasized in our kernel modification method. We expected that such kernel modification could counteract the possible noise introduced by using terms from the test set and could bring better results. Unfortunately, the results of the kernel modification methods, **CW** and **EW**, were not much better than those of **T** and **TS**. It seemed that the RBF kernel modification methods didn't have significant influence to the classification results. Apparently, the RBF kernel modification with the two measurements did not perform well enough to meet our expectation.

None of the methods performed well at extreme small sample size (20 training samples, 2 from each category). It is very easy to understand the reason. The $\chi^2(t,c)$ and $E(t,c)$ are both statistical measurements. The accuracy of SVMs algorithm also relies on the accurate representation of underlying category distribution by training samples. At very small sample-size it is very likely to have a partial, incomplete, or even distorted picture of the underlying problems. Hence, it is not surprised to see poor generalization performance. Incorporating category information from human experts instead of using statistic measurements like $\chi^2(t,c)$ and $E(t,c)$ may help to improve performance in problems with small training sample size.

References

- Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and Knowledge Discovery* 2, pp 121-167.
- Christianini, N. Shawe-Taylor, J. & Lodhi, H. (2001) Latent semantic kernels. *Proc. of the 18th International Conference on Machine Learning*, pp 66-73. Morgan Kaufman.

Joachims, T. (1998). Text Categorization with support vector machines: Learning with many relevant features. *Proceedings of ECML-98*. pp 137-142.

Joachims, T. (1999). Transductive inference for text classification using Support Vector Machines. *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pp 200-209.

Sebastiani, F. (2002). Machine Learning in automated text categorization. *ACM computing Surveys*. 34. pp1-47.

Scholkopf, B. Simard, P. Smola, A. & Vapnik, V. (1998) Prior knowledge in support vector kernels. *Advances in Neural Inf. Proc. Systems*, volume 10. pp 640-646.

Siolas, G. & D'Alche-buc, F. (2000). Support vector machines based on a semantic kernel for text categorization. *In Proceedings of the International Joint Conference on Neural Networks, IJCNN, Como., IEEE. Vol.5*, pp 205-209.

Yang, Y. & Pedersen. J. (1997). A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97)*, pp 412-420.