Random Subspace Method in Text Categorization

Mehrdad J. Gangeh, Mohamed S. Kamel
Department of Electrical and Computer Engineering
University of Waterloo
Waterloo, Canada
{mgangeh, mkamel}@pami.uewtareloo.ca

Robert P.W. Duin
Pattern Recognition Laboratory
Delft University of Technology
Delft, the Netherlands
r.duin@ieee.org

***Abstract*—In text categorization (TC), which is a supervised technique, a feature vector of terms or phrases is usually used to represent the documents. Due to the huge number of terms in even a moderate-size text corpus, high dimensional feature space is an intrinsic problem in TC. Random subspace method (RSM), a technique that divides the feature space to smaller ones each submitted to a (base) classifier (BC) in an ensemble, can be an effective approach to reduce the dimensionality of the feature space. Inspired by a similar research on functional magnetic resonance imaging (fMRI) of brain, here we address the estimation of ensemble parameters, i.e., the ensemble size (L) and the dimensionality of feature subsets (M) by defining three criteria: usability, coverage, and diversity of the ensemble. We will show that relatively medium M and small L yield an ensemble that improves the performance of a single support vector machine, which is considered as the state-of-the-art in TC.**

Keywords-random subspace method; text categorization; support vector machine; ensemble of classifiers

I. Introduction

Text categorization (TC) is one of the fast paced applications of machine learning and data mining [1]. There are many applications in natural language processing and information retrieval employing TC techniques [1]. For example, news stories are typically organized by subject categories (topics) or geographical codes; academic papers are often classified by technical domains and sub-domains; patient reports in health-care organizations are indexed from multiple aspects, using taxonomies of disease categories, types of surgical procedures, and so on.

Two main issues in TC are the high dimensional feature space and high feature-to-instance ratio, both related to using term frequency as features that generates a very high dimensional feature space. Two main solutions for these problems in the literature are: first, feature selection to reduce the dimensionality of feature space [2]. However, even using feature selection, the dimensionality of feature space remains high unless a very aggressive feature selection technique is used that degrades the performance of the classification systems [2]. Second, using classifiers that behave well in high dimensional feature space. This is one of the reasons that among many classifiers reported in the literature for TC, support vector machines (SVM) are considered as the-state-of-the-art with higher performance than others [3].

The third technique, which is the focus of this paper and not very much investigated in the literature for text categorization, is random subspace method (RSM), which is a combining technique [4]. RSM is used for efficient handling of high dimensional feature space by using randomized classifier ensembles [5]. In a high dimensional feature space, a classifier usually suffers from the 'curse of dimensionality' [6], i.e., many data samples are needed to adequately train the classifier. By dividing the feature space randomly to some subsets and submitting each one to a (base) classifier (BC), we reduce the dimensionality of feature space of each classifier while use the same number of samples for training. This can effectively resolve the problem of suffering from the 'curse of dimensionality'. RSM is mainly investigated on weak classifiers such as linear discriminant classifiers (LDC) to improve their performance [7]. However, it is not extensively investigated on strong classifiers such as SVM (that perform rather well in high dimensional feature space) nor in the area of text categorization. Recently, the RSM is investigated on SVM for the classification of brain images obtained from functional magnetic resonance imaging (fMRI) [8]. The same as TC, in brain images of fMRI, the dimensionality of feature space is very high and the feature-to-instance ratio is extremely large.

Two main parameters of RSM are the number of BCs in the ensemble and the dimensionality of each random subspace submitted to the BCs. Inspired by the work in [8], we address estimation of these two parameters in the context of text categorization. We show that RSM improves the performance of the classification system comparing to a single SVM.

II. Theoretical Background

In this section, first the feature selection techniques used in this research are explained. Then the RSM and the probabilistic framework for choosing the number of BCs and feature subset size are presented.

A. Feature Selection

In TC, the unique terms that occur in documents constitute the feature space. There can be hundreds of thousands terms even in a moderate-size text dataset. Hence, the dimensionality of feature space is very large. This degrades the performance of the classification system due to the 'curse of dimensionality', reduces the generalization capability of the classifier, and increases the computational cost. Feature selection can reduce these problems partially. Five feature selection techniques are investigated and compared in [2], among which document frequency (DF) and information gain (IG) are adopted in this research. These two techniques are briefly explained here.

DF is one of the simplest feature selections techniques used in TC with lowest computational cost. DF is the number of documents in which a term occurs. It is usually computed for each unique term in the dataset and those terms whose DF is less than some predetermined threshold are removed. The main assumption is that rare terms are not informative in category prediction and will not affect the global performance.

IG measures the entropy required for category prediction by knowing the presence or absence of a term in a document. If $\{\omega_1, \dots, \omega_c\}$ is the set of classes, the IG of each term t is defined as follows:

$$\begin{aligned} IG(t) = &-\sum_{i=1}^{c} \mathrm{P}(\omega_i)\, log\mathrm{P}(\omega_i) \\ &+ \mathrm{P}(t)\sum_{i=1}^{c} \mathrm{P}(\omega_i|t)\, log\mathrm{P}(\omega_i|t) \\ &+ \mathrm{P}(\bar{t})\sum_{i=1}^{c} \mathrm{P}(\omega_i|\bar{t})\, log\mathrm{P}(\omega_i|\bar{t}) \end{aligned} \qquad (1)$$

All features are ranked according to their IG values. Thus, the problem of choosing an IG threshold for feature reduction is converted to the selection of number of features to be removed. The computational complexity of IG is higher than DF, with the benefit of having control on the number of retained features.

B. Random Subspace Method (RSM)

Even after feature selection, the dimensionality of feature space can be easily few thousands. Although SVM with linear kernel is proved to perform well in such a high dimensional feature space [3], we expect that RSM can improve its performance by dividing this high dimensional feature space to smaller ones.

In RSM, after dividing the original feature space to L feature subsets of dimensionality M, each subset is submitted to a base classifier (BC) in the ensemble. The final decision on the class of the document is obtained by combining the decisions of these BCs using a combining rule such as majority vote. Since the features subsets submitted to the BCs are different, parallel ensemble is a natural choice [4] (Fig. 1).

There are two main parameters in a RSM to be determined, i.e., L (the number of BCs) and M (the dimensionality of feature subsets). Recently, the selection of these two parameters is addressed in a similar problem, i.e., the classification of brain images of fMRI [8]. The same as TC, fMRI classification suffers from high dimensional feature space as well as large feature-to-instance ratio.

Here, we briefly explain the approach in [8] for L and M selection and extend it to the TC problem.

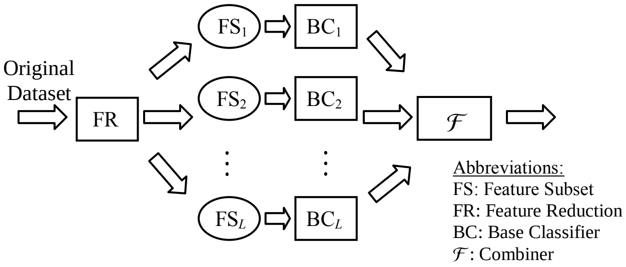


Figure 1. The structure of RSM.

Suppose $\mathbf{x} \in \mathbb{R}^n$ is the set of original feature space, among which $\mathbf{q} \in \mathbb{R}^Q$ features are important and the rest are noise. While all features in TC may contain information, some may be less important. The above assumption is made for probabilistic formulation [8].

For a good performance of the ensemble, The BCs should be as accurate and diverse as possible [4]. Hence, if all features given to a BC are noise, it may not perform better than the *prior*. On the other hand, if all BCs perform on the same important features, there is no diversity in the ensemble. Based on these observations, three criteria are defined in [8] based on which, the performance of an ensemble can be predicted, i.e., *usability*, *coverage*, and *diversity*.

A BC is *usable* if its feature subset includes at least one important feature. The probability of having a completely usable ensemble, $P(UE)$ is

$$P(UE) = [1 - \frac{\binom{n-Q}{M}}{\binom{n}{M}}]^L \qquad (2)$$

where, n and Q are the dimensionality of the original feature space and important features, respectively [8].

Coverage is defined to take into account the percentage of important features included in the feature subsets submitted to the ensemble. The probability of complete coverage $P(CC)$, i.e., the probability of submission of all important features to the ensemble is

$$P(CC) = [1 - (1 - \frac{M}{n})^L]^Q \,. \qquad (3)$$

If all feature subsets submitted to the BCs are the same, we will not have an ensemble any more as all BCs are identical. Thus, another criteria, i.e., *diversity* has been introduced in [8] as

$$D(S_1, S_2) = |I_1 \cup I_2| - |I_1 \cap I_2| \,, \qquad (4)$$

where, S_i is the feature subset submitted to the i^{th} BC and I_i is the subset of important features in S_i. If $I_1 = I_2$ in (4), then the diversity between BCs 1 and 2 is zero. The probability of having all pairs non-identical, $P(APNI)$, in the ensemble is

$$P(APNI) = [1 - \sum_{j=1}^{\min\{Q,M\}} \frac{\binom{Q}{j}\binom{n-Q}{M-j}^2}{\binom{n}{M}}]^{\frac{L(L-1)}{2}}. \qquad (5)$$

The optimum values for the ensemble parameters, i.e., M and L, can be found by maximizing three criteria defined in (2), (3), and (5) in respect to these two parameters. However, as shown in [8], there is no single M and L for which all these criteria are maximized. Nevertheless, it is experimentally shown in [8] that relatively medium M and small L for the fMRI data yield high performance of the ensemble. We will investigate, using a grid search, the optimum values of M and L of the ensemble in TC application.

III. Experimental Setup and Results

In this section, the experimental setup and the results are presented. A text dataset is used in the experiments, which is consisting of 495 text samples from Brown corpus [9] classified into 15 categories, i.e., press: reportage, press: editorial, press: reviews, religion, skill and hobbies, popular lore, belles-letters, miscellaneous: government and industry house organ, learned, fiction:general, fiction:mystery, fiction:science, fiction:adventure, fiction:romance, and humor. The dimensionality of feature space (number of terms) is 22244 while there are only 495 documents in the dataset. This is a clear example of a text dataset with high dimensional feature space and very large feature-to-instance ratio.

To prepare the data before submission to the classifiers, two-step feature selection is performed on the dataset to reduce its dimensionality. In first step, DF is used as a coarse feature reduction technique with the threshold of 1% on document vectors. This reduces the dimensionality of Brown corpus from 22244 to 7454. In second step, IG is used for fine feature reduction. The number of features retained is selected to be the nearest thousand of 25% of the original feature space, which is 5000 for Brown corpus. It is shown in [2] that feature reduction to 25% of original space does not degrade the performance of the classification system.

As mentioned in Section 1, SVM is considered as the state-of-the-art in TC and hence is used as the type of the base classifier in the ensemble. Linear kernel is used and the trade-off parameter (C) is selected using a grid search and 5-fold cross-validation on the whole dataset as shown in Fig. 2. It can be seen from this figure that for $C > 4$, the SVM performs almost optimum. Although, this optimum value might be different for random feature subsets, since performing grid search on each BC is computationally expensive, $C = 100$ is chosen for all BCs. Majority vote is used for combining the outputs of the BCs.

The experiments are repeated 5 times and the results are averaged. In each experiment, 80% of the documents are randomly chosen as the training set and the remaining as the test set.

Eventually, a grid search is performed on L and M to find their optimum values. L is changed from 1 (single classifier) to $n/5$ with the steps of 100 and M is changed from 1 to n with the steps of 500, where n is the dimensionality of reduced feature space, i.e., 5000 here. This grid search is shown in Fig. 3a in whole searching space for the *accuracy* of the ensemble. It can be seen that the performance is relatively low for $M < 1000$ and hence the *error* surface on the grid search is zoomed for M larger than 1000 in Fig. 3b.

To compare the performance of the proposed approach with other classifiers used in the literature, the accuracy of two classifiers, i.e., SVM and *k*-NN on the original dataset (with all features) and on reduced feature space using two feature selection techniques described above is provided in Table I. Linear kernel is used for SVM and the trade-off parameter (C) is found using a grid search and 5-fold cross-validation with the amount indicated in the forth column of Table I. As for *k*-NN, the optimal value of k is found using leave-one-out approach.

As can be seen from Table I and also Fig. 3b, while maximum accuracy for single SVM is 52.83%, the accuracy is improved by 3.69% using the RSM. According to Fig. 3b, the error is minimum for almost the middle of grid, i.e., $M = n/2.5$ and $L = n/12.5$, which is 2000 and 400 in this experiment, respectively. The performance of *k*-NN is clearly inferior to single SVM and RSM on both original and reduced feature spaces.

TABLE I. The Performance of Single *k*-NN, Single SVM, and RSM on Brown Corpus at Different Feature Sizes.

Classifier	**Feature Selection Technique**	**No. of Features**	**Optimal Classifier Parameter (k^* or C^*)**	**Accuracy**
k-NN	None	22244	19	44.85
	DF	7454	3	20.61
	DF + IG	5000	133	17.37
SVM	None	22244	> 3	43.43
	DF	7454	11.31	48.69
	DF + IG	5000	> 11.31	52.83
RSM with SVM	DF + IG	5000	100	56.52

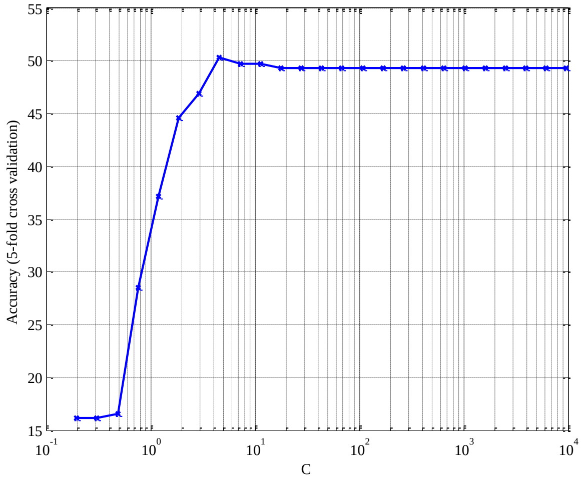
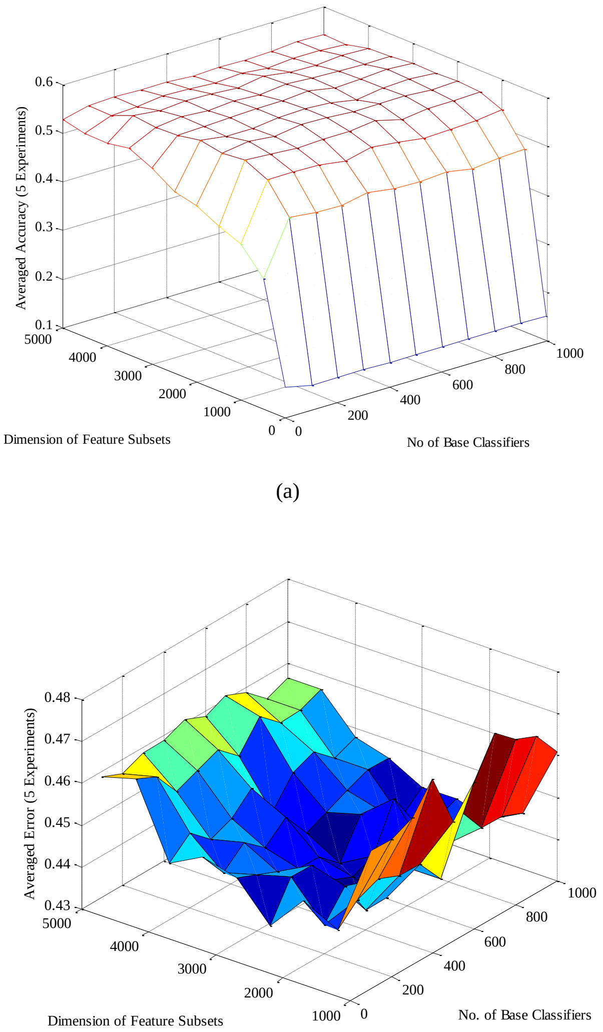


Figure 2. The grid search to find optimized trade-off parameter (C) in Brown corpus.



(a)

(b)

Figure 3. The grid search on *L* (ensemble size) and *M* (dimensionality of feature subsets) for Brown corpus: (a) the averaged *accuracy* on whole grid search and (b) the averaged *error* on part of the grid that shows better performance.

IV. Discussion and Conclusion

Text categorization has the intrinsic problem with the high dimensionality of feature space and sometimes large feature-to-instance ratio.

Random subspace method (RSM) is clearly an approach that can help in these situations by dividing the original feature space to smaller ones. However, it is not yet investigated on strong classifiers like SVM in the literature in TC application. Inspired by the recent research in similar problem, i.e., fMRI classification [8], the parameters of the ensemble, *L* (ensemble size) and *M* (dimensionality of feature subsets) are formulated in a probabilistic framework by defining three criteria: *usability*, *coverage*, and *diversity*. It is shown on Brown corpus that relatively medium *M* and small *L* improves the performance of the ensemble in comparison to single SVM, which is considered as the state-of-the-art in TC.

There might be two reasons for this improvement. First, the dimensionality of feature subsets submitted to the BCs is reduced. Second, using majority vote for combining the outputs of the BCs can create an overall nonlinear decision boundary at the output of ensemble. This may help to improve the performance over linear SVM on whole feature space.

Acknowledgment

The first author is funded by the Natural Sciences and Engineering Research Council (NSERC) of Canada under Canada Graduate Scholarship (CGS D3-378361-2009).

References

[1] S. Fabrizio, "Machine learning in automated text categorization," ACM Computing Surveys, vol. 34, no. 1, 2002, pp. 1-47.

[2] Y. Yang and O. J. Pedersen, "A comparative study on feature selection in text categorization," Proceedings of the 14th International Conference on Machine Learning, 1997, pp. 412–420.

[3] C. Silva and B. Ribeiro, "RVM ensemble for text categorization," International Journal of Computational Intelligence Research, vol. 3, no. 1, 2007, pp. 31–35.

[4] L. I. Kuncheva, Combining Pattern Classifiers Methods and Algorithms, New Jersey: John Wiley & Sons, 2004.

[5] T. K. Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, 1998, pp. 832-844.

[6] A. J. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 1, 2000, pp. 4-37.

[7] M. Skurichina and R. P. W. Duin, "Bagging, boosting and the random subspace method for linear classifiers," Pattern Analysis and Applications, vol. 5, no. 2, 2002, pp. 121–135.

[8] L. I. Kuncheva, J. J. Rodriguez, C. O. Plumpton, D. E. J. Linden, and S. J. Johnston, "Random subspace ensembles for fMRI classification," IEEE Transaction on Medical Imaging, vol. 29, no. 2, 2010, pp. 531-542.

[9] W. N. Francis and H. Kucera, "Manual of information to accompany a standard corpus of present-day edited American English, for use with digital computers," 1964.