An Ensemble Method for Multi-class and Multi-label Text Categorization

Bofeng Zhang[1] **Xin Xu**[1, 2] **Jinshu Su**[1]

[1] School of Computer, National University of Defense Technology, Changsha 410073, P. R. China
[2] Institute of Automation, National University of Defense Technology, Changsha 410073, P. R. China

Abstract

A method for multi-class and multi-label automated text categorization based on twin-SVM with naïve Bayes ensemble is proposed. Twin-SVM classifiers give a solution to the multi-label problem. For multi-class situation, naïve Bayes classifier constrains the belonging scope of a testing sample within a few most likely classes and greatly reduces the number of binary classifiers needed to make the final prediction. The benefits of the ensemble method are described and preliminary results with Reuters-21578 data set are also presented.

Keywords: Text categorization, Twin-SVM, Naïve Bayes, Ensemble of classifiers, Multi-label and multi-class,

1. Introduction

Automated text categorization is defined as the task of assigning predefined class labels to text documents by learning a set of training samples to construct a classifier or mining model. Recently, lots of research applied machine learning methods in automated text categorization. The text categorization methods that have been studied include various supervised learning algorithms such as kNN, decision tree, Naïve Bayes, Rocchio, neural networks and support vector machines (SVM), etc. Among the existing text categorization methods, the applications of SVM for text categorization have obtained several state-of-art results in classification precision. However, the computation cost of multiple binary SVMs for more than two classes called multi-class problem is usually a bottleneck for mining of large-scale text documents. And also, SVM adapted to multi-label problem where a testing sample may belong to more than one class has attracted increasing attention [1] and [2].

Ensemble learning algorithms train multiple classifiers and then combine their predictions. Since the generalization ability of an ensemble classifier can be much better than a single learner, the algorithms and applications of ensemble learning have been widely studied in recent years. In many successful applications, ensemble learning classifiers usually achieve the best performance in the literature [3].

In this paper, to solve the multi-class and multi-label problem of text categorization by binary SVM, a novel text categorization method based on twin-SVM with naïve Bayes ensemble is proposed. In the proposed method, twin-SVM classifiers for multi-label prediction and naïve Bayes classifier are cascaded as an ensemble classifier, which can be viewed as new ensemble architecture and a new decomposition-based strategy for multi-class SVM such as one-vs-one and one-all, etc. To solve the multi-label problem, for each pair of classes c_1 and c_2 sharing common training samples, we proposed a twin-SVM method which respectively trains two binary classifiers SVM_1 to distinguish c_1 against c_1-c_2 and SVM_2 to distinguish c_2 against c_2-c_1. The Bayes classifier firstly selects for a testing instance the top ranked labels with the sum of whose posteriori probabilities bigger than a threshold. Then the SVM classifier makes a final decision within the selected labels. It is like that the Bayes classifier is configured to perform coarse filtering for all possible labels due to its higher efficiency. The multi-class SVM is only used to process selected part of the label set output by the naïve Bayes. The label decision strategy for SVM is based on the validation results of the naïve Bayes classifier so that possible labels with lower posteriori probabilities are refined by the SVM classifier.

This paper is organized as follows. Section 2 gives a brief introduction on the techniques involved in text categorization. Section 3 and 4 presents the principles and algorithms of multi-class and multi-label twin-SVM with Bayes ensemble. Experimental results on Reuters 21578 are given in Section 5. And some conclusions are drawn in Section 6.

2. Preliminaries

To employ automated text categorization for various applications, sampled data are collected and labeled with their corresponding classes. The sampled data are transformed to a uniform format by extracting ASCII text information from them. The transformed data are then divided into two sets for classifier training and testing based on automated text categorization. Automated text categorization usually involves three steps, namely, *document representation*, *classifier construction*, and *performance evaluation*.

Document representation can be viewed as a preprocessing process, which includes stop word elimination, stemming, feature selection and weighting. After preprocessing, a text document d is usually represented as a data vector

$$\mathbf{d} = [z_{d,1}, z_{d,2}, \ldots, z_{d,n}],$$

where $w_{d,i}$ (i=1,2,…n) are the weights of all the n document features. The feature weights are usually determined by some function of feature frequencies

$$z_{d,i} = g(t_{d,i}) \ ,$$

where $t_{d,i}$ is the occurrences of feature f_i in the document d and the selection methods of function $g(\cdot)$ include TF, TF*IDF and log(TF)*IDF, etc. The above document representation method is usually called VSM (vector space model).

In classifier construction for automated text categorization, various machine learning methods can be used to learn a classifier model based on training data. The training data are composed of preprocessed document data vectors from different classes and each data vector is labeled with the corresponding class labels.

The performance evaluation of text classifiers is conducted on a testing or validation data set which is different from the training set. The error of a classifier is one of the main criteria for performance evaluation defined as all the incorrect predictions made by the classifier dividing by the size of the according testing set.

Details of above discussions can be referred to [1].

3. Naïve Bayes and SVM method

3.1. Naïve Bayes

Naïve Bayes is a classifier based on the Bayesian theory. It is highly practical because of its assumption of terms independence, although this is often not the case. The Bayesian approach classifies a new documents d by assigning the class label in the label set C={c_1,c_2,…,c_m} with the maximum posteriori probability $P(c_k|d)$ to the given document .

By Bayes' theorem, $P(c_k|d)$ can be replaced as the following equations:

$$P(c_k \mid d) = \frac{P(d \mid c_k)P(c_k)}{\sum_{c\in C} P(d \mid c)P(c)} \ .$$

$P(c_k)$ is the probability of the class c_k, and using Laplace estimator it can be calculated by:

$$P(c_k) = \frac{1+|c_k|}{|C| + \sum_{c\in C}|c|} \ ,$$

where $|c|$ is the set size of the training documents belonging to class c.

We compute the likelihood $P(d|c)$ by the formula:

$$P(d \mid c) = \frac{h_d!}{z_{d,1}!\, z_{d,2}! \ldots z_{d,n}!} \prod\nolimits_{w_{d,i}>0, i=1,2,\ldots,n} p_{c,i}^{z_{d,i}} \ ,$$

where h_d is the length of document d and

$$p_{c,j} = \frac{1+\sum_{d\in c} z_{d,i}}{\sum_{c\in C}\sum_{d\in c} z_{d,i} + n}$$

is the Laplace estimator of $P(f_j|c)$ (j=1,2,…,n).

3.2. SVM

Based on the idea of constructing optimal separating hyper-plane to improve generalization ability, SVM are originally proposed for binary classification problems [1] and [4].

In the training of binary SVM classifiers, a hyperplane $(\mathbf{w}\cdot\mathbf{x})+b=0$ ($\mathbf{w}\in\mathbf{R}^n, b\in\mathbf{R}$) is considered to separate two classes of samples. Then the decision function can be given by $f(\mathbf{x}) = \operatorname{sgn}(\mathbf{w}\cdot\mathbf{x}+b)$.

Based on the SRM principle in statistical learning theory, the optimal separating hyperplane can be constructed by the following optimization problem:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1, i = 1,2,\ldots,N.$$

To reduce the effects of noise and outliers in real data, the following soft margin techniques are usually used, which is to solve the primal optimization problem as:

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$\text{s.t.} \quad y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1-\xi_i \quad \xi_i \geq 0, i = 1,2,\ldots,N.$$

The Lagrangian dual of soft-margin support vector learning can be formulated as:

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j (\mathbf{x}_i\cdot\mathbf{x}_j)$$

$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, i = 1,2,\ldots,N \text{ and } \sum_{i=1}^{N}\alpha_i y_i = 0.$$

Since in most real-world classification problems, nonlinear separating planes have to be constructed, a well-known 'kernel trick' is used to transform the above linear form of support vector learning algorithms to nonlinear ones. The optimization problem of SVMs for two-class soft margin classifiers with kernel is formulated as follows:

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\text{s.t.} \quad 0 \le \alpha_{\mathrm{i}} \le C, i = 1,2,\ldots,N \text{ and } \sum_{i=1}^{N}\alpha_i y_i = 0.$$

where K(·,·) is called the kernel function which, without explicitly knowing the map, compute the inner product of two mapped vectors in the space with higher (possible infinite) dimension. Then the decision function becomes:

$$f(\mathbf{x}) = \operatorname{sgn}(\sum_{i=1}^{N}\alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b) .$$

For multi-class classification cases, several decomposition-based approaches are proposed. The idea of decomposition-based methods is to divide a multi-class problem into multiple binary problems, i.e., to construct multiple standard two-class SVM classifiers and fuse their classification results. There are several strategies for the implementation of multi-class SVMs using binary algorithms, including one-vs-all, one-vs-one, and error correcting coding [4], etc. Among them, one-vs-one is the simplest approach with high effectiveness. Decomposition-based method constructs considerable number of binary SVM classifiers respectively for each pairs of two date set combined from deferent class subsets. Therefore, the multi-class classification problem is decomposed into various subtasks of training binary SVM classifiers. In the testing phase, each binary classifier votes on the according pair of label subsets and the label with the most votes from all the binary classifiers becomes the final decision.

4. SVM with naïve Bayes ensemble

The above decomposition-based strategies have some drawbacks. Firstly, any testing sample must be tested by a great deal of binary classifiers to make a final decision, i.e., *m*-1 for one-vs-all and *m*(*m*-1)/2 for one-vs-one, etc. Secondly, if there are some samples belonging to more than one class, there will be some confusion to make a correct prediction whether the class with less number of votes is also the label of those samples. The second case is the problem of so called multi-label classification.

4.1. Twin-SVM for multi-label

To solve the multi-label problem, for each pair of training sets c_1 and c_2 sharing common training samples, we proposed a twin-SVM method which respectively trains two binary classifiers SVM$_1$ to distinguish c_1 against c_1-c_2 Band SVM$_2$ to distinguish c_2 against c_2-c_1. As the fig.1 shows, the combination the two SVM called a twin-SVM may predict a sample to be classified to both classes, that is, a twin-SVM may give votes to both the two parties it wants to differentiate. Thus the combination by the decomposition-based strategies of all the twin-SVMs may classify a testing sample to more than one class. Reference [5] has given the detailed discussions.

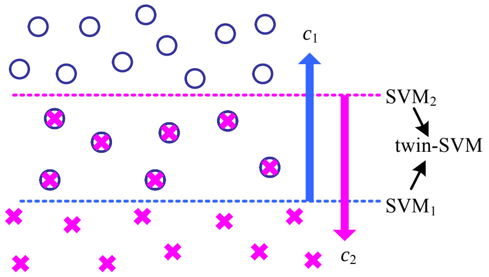

Fig. 1: Construction of twin-SVM.

4.2. Twin-SVM with naïve Bayes ensemble for multi-class

The decomposition-based strategies of binary SVMs for multi-class problem must test an unknown sample by a great deal of classifier, especially for twin-SVM the testing time spending may be doubled. However, if we firstly constrain the belonging scope of the sample within a few classes, the cost will be reduced greatly.

The main idea of the twin-SVM with Naïve Bayes ensemble is not to distinguish among all the classes but the most likely classes a testing sample may belong to.

In the training phase, we firstly train a naïve Bayes classifier for all the classes. Secondly, like one-vs-one, we train twin-SVM classifiers for every pair of all the classes.

In the testing phase, we select the top ranked classes of Naïve Bayes by the principle that the sum of their posteriori probabilities is bigger than a threshold of *θ*. The threshold results in a varying number of binary classifier for different testing sample and makes our method different from the method discussed in [3], where fixed number of binary classifiers is chosen. It is like that the Bayes classifier is configured to perform coarse filtering for all possible labels due to its higher efficiency. Then only fewer twin-SVMs involved is used to process selected part of the label set output by the naïve Bayes. The label decision strategy for twin-SVMs is based on the validation results of the naïve Bayes classifier so that possible labels with lower posteriori probabilities of the selected classes are refined by the twin-SVM classifiers. The proposed method takes advantages both of the fast speed of the naïve Bayes and the high precision of the SVM.

5. Experiment

The Reuters-21578 corpus [6] was used to evaluate the performance of the twin-SVM with naïve Bayes ensemble method. As in several other studies, only the ten most populous classes were used and the classifiers were trained and tested with ModApte split [1].

For each binary SVM we use linear kernel with C=1 and for Naïve Bayes to ensemble we set θ=0.9, resulting average 2.8 classes to be distinguished for every samples.

We compare the testing errors of the Naïve Bayes, one-vs-one of SVM and our ensemble method in table 1. From table 1 we can see that that our twin-SVM with Naïve Bayes ensemble method outperforms the other two methods in multi-label and multi-class classification. Although for single label and multi-class classification, one-vs-one SVM is very effective, it can rarely make precise perditions for multi-label samples because there will be only one label getting the most votes from all the binary classifiers. This results the lower performance of one-vs-one than that the combination of twin-SVM classifiers where likely labels may get the same number of votes.

class	Naïve Bayes	One-vs-one SVM	Twin-SVM+ NB
acq	7.72%	4.93%	**4.07%**
corn	6.99%	4.58%	**3.93%**
crude	7.16%	4.66%	**4.15%**
earn	7.48%	4.91%	**4.34%**
grain	8.04%	5.72%	**4.62%**
interest	7.97%	5.30%	**4.89%**
money	8.95%	6.64%	**5.23%**
ship	5.30%	3.37%	**1.96%**
trade	8.87%	6.49%	**6.07%**
wheat	7.91%	5.17%	**4.50%**
full set	34.8%	27.81%	**18.39%**

Table 1: Comparison of the testing error.

The comparison of the time for classifying all the testing samples of different method is shown in fig.2. One-vs-one method need far more the time spending than our twin-SVM with naïve Bayes ensemble method. Though the ensemble method spent more than twice the time needed by naïve Bayes, it was acceptable considering its remarkable effectiveness.

6. Conclusions

In this paper, to solve the multi-class and multi-label problem of text categorization by binary SVM, we proposed a novel text categorization method based on twin-SVM with naïve Bayes ensemble. Experiments have shown that the ensemble of twin-SVM and naïve Bayes was highly more effective in multi-class and multi-label than using SVM or naïve Bayes only. The testing efficiency was also acceptable.

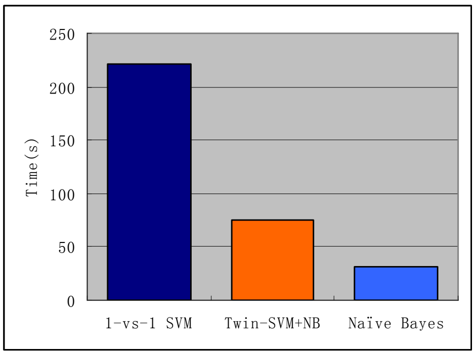

Fig. 2: Comparison of the testing time.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China under Grants 90604006 and Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20049998027.

References

[1] F. Sebastiani, Machine learning in automated text categorization, *ACM Computing Surveys*, 34:1-47, 2002.

[2] J.S. Su, B.F. Zhang, and X. Xu, Advances in machine learning based text categorization, *Journal of Software*, 17:1848-1859, 2006.

[3] Y.G. Wei and J.J. Tsay, A study of multiple classifier systems in automated text categorization, Master Thesis, College of Engineering, National Chung Cheng University, 2002.

[4] C. Hsu and C. Liu, A comparison of methods for multi-class support vector machines, *IEEE Transactions on Neural Networks*, 13:415-425, 2002.

[5] B.F. Zhang, Twin-SVM for multi-label text categorization, Technical Report, CS-615, Changsha China: Computer School, National University of Defense Technology, 2006.

[6] http://www.daviddlewis.com/resources/testcollections/reuters21578/