Study on SVM Compared with the other Text Classification Methods

Zhijie Liu
1Chinese Information Processing Research Center, Beijing Information Science and Technology University
Beijing, China
Zhijie12315@163.com

Xueqiang Lv
1 Chinese Information Processing Research Center, Beijing Information Science and Technology University
2 Beijing TRS Information Technology Co., Ltd.
Beijing, China
lv.xueqiang@trs.com.cn

Kun Liu
1 Chinese Information Processing Research Center, Beijing Information Science and Technology University
2 Beijing TRS Information Technology Co., Ltd.
Beijing, China
liu.kun@trs.com.cn

Shuicai Shi
1 Chinese Information Processing Research Center, Beijing Information Science and Technology University
2 Beijing TRS Information Technology Co., Ltd.
Beijing, China
shi.shuicai@trs.com.cn

Abstract—**Based on the text information processing, we have made a study on the application of support vector machine in text categorization. Through introducing the basic principle of SVM, we described the process of text classification and further proposed a SVM-based classification model. Finally, experimental data show that F1 value of SVM classifier has reached more than 86.26%, and the classification results comparing to other classification methods have greatly improved, and it also proves that SVM is an effective machine learning method.**

Keywords- text classification; SVM; machine learning

I. INTRODUCTION

With the continuous development of the Internet, text information is increasing day by day, and the text information analysis will also become important. A key technology of the text information analysis is text classification. Through the automatic text classification system, we can classify text data, thereby to help people better find, filter and analyze text information resources, so it is very necessary to construct an effective text classification system. In recent years, Support Vector Machine (SVM) is particularly prominent in the text classification performance, and its classification accuracy rate, recall rate and the F1 value have achieved good results. As a new machine learning method, Support Vector Machine is proposed by Vapnik and others [1-2], which is on the basis of statistical learning theory. The advantages of the SVM are simple structure, complete theory, high adaptability, global optimization, short training time and good generalization performance [3], so the technology has become the current international research focus of machine learning community.

II. CORRELATION RESEARCHES

At present, text classification techniques mainly include Decision Tree method, KNN method, Naive Bayesian method and SVM method.

Decision Tree method is a process that using information gain in the information theory to find the properties of the field with the greatest amount of information in the sample database to build a decision tree node, and build a branch of a tree according to the properties of different values of the field; and then repeat building the next nodes and branches for each branch. It has the characteristics of fast and accurate classification. However, there are still some unresolved issues, for example, vacancies in data processing and the discretization of continuous attributes, etc.

KNN method is a theoretically more mature method. Suppose each class contains several sample data, and each data has a unique type of data standard. KNN is to obtain the k-nearest sample data from the data to be classified by calculating the distance from each sample data to the data to be classified, and which type of the k sample data occupies the majority, then the data to be classified belongs to that category. However, due to a large number of calculations, the efficiency of KNN method greatly reduced when training samples with more attributes to be classified.

Naive Bayesian method is a kind of learning method based on Bayesian theorem, and it can be used to predict the possibility of class membership, and it gives the probability of the text belonging to a particular class. According to forecast results, the sample is assigned to the highest probability of category when classified. In theory, the application premise of naive Bayesian classifier is that the sample attribute value is independent of the classification properties of the sample, but

978-0-7695-3987-4/10 $26.00 © 2010 IEEE
DOI 10.1109/ETCS.2010.248

its attribute independence assumption affects its classification performance.

SVM method is suitable for a large sample set of the classification, especially for text classification [4]. Its algorithm is based on structural risk minimization principle. First, the original data set is compressed to support vector set, then new knowledge is gained by learning to use the subset and also the rules decided by support vector are given. So it is a good classifier, and it has superior performance and a wide range of applications.

III. THE PRINCIPLE OF SVM

The main idea of SVM can be summarized as two points:(1)It is to study for the linearly separable case, and for the linearly inseparable case, we need a nonlinearity mapping to change the inseparable sample of low-dimensional sample space to high-dimensional feature space, then it will become linearly separable.(2)It is based on structural risk minimization theory to find the optimal separating hyperplane from the feature space. So learning machine can get global optimization, and the expected risk of entire sample space would meet a certain upper bound with a probability. We are going to discuss the principle of SVM under the following three cases.

A. Two Types of Linearly Separable Case

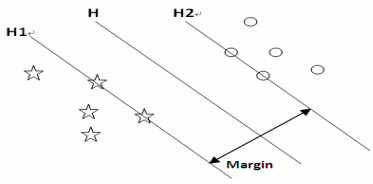

Figure 1. The case of linear classification

As shown in Fig.1, the five-pointed stars and the dots represent two different types of samples, where H is classified line. H1 and H2 are straight-lines, which pass through different types of samples distancing the classified line recently, and parallel to classified line H. The distance between line H1 and line H2 is called class interval (margin). The so-called optimal separating line refers to the classification line which will not only be able to separate two types of samples correctly, but also to make class interval maximum [5]. Classification line equation can be expressed as: $w \cdot x+b=0$; Linear sample set can be expressed as: $(x_i, y_i), i = 1, 2 \cdots n, x \in R^d, y_i \in \{-1, 1\}$. Normalize classification line to make linear sample set satisfy the condition:

$$y_i[(w \cdot x_i) + b] - 1 \geq 0, i = 1, 2, \cdots n \tag{1}$$

Class interval can be got by calculating, and it is equal to $\frac{2}{\| w \|}$. For maximum class interval is equivalent to $\| w \|^2$ minimum. Optimal separating surface is a category surface which satisfies the conditions (1) and is able to make $\frac{1}{2}\| w \|^2$ minimum. Here we need to explain that H1 and H2 are support vector, because they support the optimal separating surface. As the objective function and constraints are convex, according to optimization theory, this problem exists unique global minimum solution. Apply Lagrange method to convert it into dual problem:

$$\begin{cases} Max: W(a) = \sum_{i=1}^{n} a_i - \frac{1}{2}\sum_{i,j=1}^{n} a_i a_j y_i y_j (x_i \cdot x_j) \\ \text{constraint condition}: \sum_{i=1}^{n} y_i a_i = 0, ai \geq 0, i = 1, 2 \cdots n \end{cases} \tag{2}$$

Where a_i is the Lagrange multiplier for each sample. If $a*$ is the optimal solution to (2), then

$$W^* = \sum_{i=1}^{n} a_i^* y_i x_i \tag{3}$$

The optimal classification function is:

$$f(x) = \operatorname{sgn}\{(w^* \cdot x) + b^*\} = \operatorname{sgn}\{\sum_{i=1}^{n} a_i^* y_i (x_i \cdot x) + b^*\} \tag{4}$$

Where $\operatorname{sgn}()$ is the sign function.

B. Two Types of Linearly Inseparable Case

Some classification problem is linearly inseparable. Therefore, the classification hyperplane capacity would be limited. In this case, we can introduce a slack variable $\varepsilon_i, i = 1, 2, \cdots n$. Constraint condition in the previous section (1) becomes

$$y_i[(w \cdot x_i) + b] \geq 1 - \varepsilon, i = 1, 2, \cdots n \tag{5}$$

Then the objective function is replaced by

$$L(w, \varepsilon) = \frac{1}{2}\| w \|^2 + C \mid \sum_{i=1}^{n} \varepsilon_i \mid, i = 1, 2, \cdots n \tag{6}$$

Therefore, a broad optimal separating surface can be got by comprehensive consideration between the minimum misclassified samples and the maximum class interval. Where C is a constant and it is greater than zero, known as the penalty coefficient, which controls the degree of punishment of right or wrong samples. Then the optimal hyper-plane function is gained as follows:

$$f(x) = \operatorname{sgn}(\sum_{i=1}^{n} a_i y_i \phi(x) \cdot \phi(x_i) + b) \tag{7}$$

C. Nonlinear Case

1) Nonlinear SVM

The real value of Support Vector Machine is used to solve nonlinear problems [6]. The method is through a nonlinear mapping φ to map the sample space to a high-dimensional or even infinite dimensional feature space, so that linear SVM method in the feature space can be applied to solve the non-linear classification problems in the sample space. The nonlinear mapping from the sample space to the feature space is shown as Fig.2.

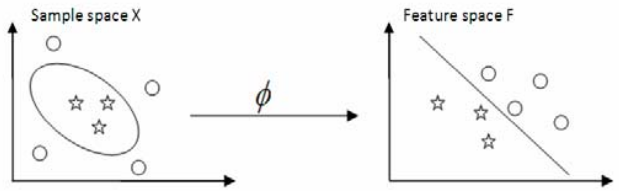

Figure 2. The non-linear mapping from sample space to feature space

2) Kernel Function

$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ is called kernel function. According to the relevant functional theory [7], if the kernel function satisfies the Mercer condition, it corresponds to the inner product in some transformation space, $\phi(x) \cdot \phi(x_i) = K(x, x_i)$. Therefore, the use of appropriate kernel function can be an alternative to non-linear mapping in high-dimensional space, to achieve the linear classification. For kernel function $K(x_i, x)$, there are three common types of SVM. They are given as follows:

a) Polynomial kernel:

$$k(x, x_i) = (\gamma x \cdot x_i + r)^d \tag{8}$$

b) RBF kernel:

$$k(x, x_i) = \exp(-\frac{1}{2\delta^2} \| x - x_i \|^2) \tag{9}$$

c) Sigmoid kernel:

$$k(x, x_i) = \tanh[\gamma(x \cdot x_i) + r] \tag{10}$$

IV. TEXT CLASSIFICATION ALGORITHM BASED SVM

A. The General Process of Text Classification

The general process of text classification [8] mainly consists of four modules: text preprocessing, feature extraction, training classifier and training model, shown as Fig.3.

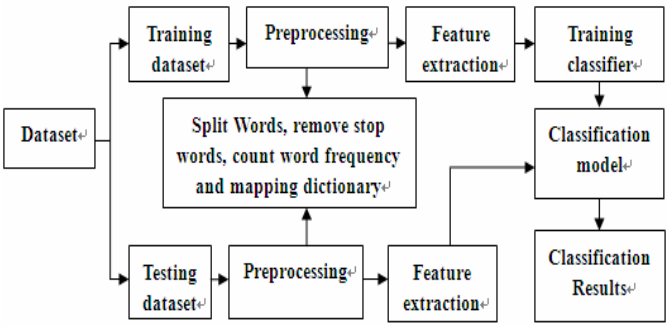

Figure 3. The general process of text classification

B. Feature Extraction

Feature extraction is one of the most important issues in text classification, and it can reduce the text vector space dimension, simplify the calculation, prevent over-fitting and so on. The implementation process of feature extraction is as follows:

- Text processing: split words→remove stop words→count word frequency.
- Dictionary generating: first, split words of all the texts and remove stop words; then use TF-IDF to count weight of each word, and take the higher value of 7834 words to constitute a dictionary.
- Feature extraction: through mapping processed text and generated dictionary mentioned above, to extract the features of the text to obtain feature vectors.

C. The Tools to Be Used

This experiment is based on the libsvm-2.89 kit, using VC++ to realize the text classification. Its processes are as follows:

- Prepare the data sets. Convert the extracted feature vector into the format that the tool kit required.
- Scale the data sets. In this experiment, the scaled range of data sets is [-1, 1]. Scaling of data sets aimed at: on the one hand, to avoid a number of feature values range too large while others are too small; on the other hand, to avoid numerical calculation difficulties caused by calculating the inner product kernel function while training
- Train the scaled data sets. The kernel function and parameters need to be set at this stage, and we used RBF kernel function in this experiment, then we selected the best parameters C and g after cross-validation, and their values were 700 and 0.5.
- Through the obtained model, we conduct testing and forecasting. Finally, we arrive at the results of the classification.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Data

Experimental data include four class documentations, and they are environment, sport, politic, and art. The training data are different from the testing data.

B. Experimental Results

Performance evaluation of text classification mainly includes accuracy rate, recall rate and F1 value. The following three sets of experimental data are the different classification results with the same training data sets and testing data sets under different classification methods.

1) The result of KNN classification is shown in Table 1:

TABLE I. THE RESULT OF KNN CLASSIFICATION

Category name	Training corpus/ Testing corpus	Accuracy rate	Recall rate	F1 value
environment	1800/200	97.44%	76.00%	85.39%
sport	1800/200	73.49%	91.50%	81.51%
politic	1800/200	78.48%	93.00%	85.13%
art	1200/200	94.30%	74.50%	83.24%

2) The result of Naive Bayesian classification is shown in Table II:

TABLE II. THE RESULT OF NAIVE BAYESIAN CLASSIFICATION

Category name	Training corpus/ Testing corpus	Accuracy rate	Recall rate	F1 value
environment	1800/200	95.92%	70.50%	81.27%
sport	1800/200	76.86%	93.00%	84.16%
politic	1800/200	91.76%	78.00%	84.32%
art	1200/200	79.67%	96.00%	87.07%

3) The result of SVM classification is shown in Table Ⅲ:

TABLE III. THE RESULT OF SVM CLASSIFICATION

Category name	Training corpus/ Testing corpus	Accuracy rate	Recall rate	F1 value
environment	1800/200	86.03%	86.50%	86.26%
sport	1800/200	86.07%	86.50%	86.28%
politic	1800/200	97.31%	90.50%	93.78%
art	1200/200	93.48%	86.00%	89.58%

Accuracy rate and recall rate reflect two different aspects of classification quality, while a comprehensive evaluation index of the two aspects is the F1 value. As shown in Fig.4, the figure reflects classification results of the various classifiers under the composite index F1 value.

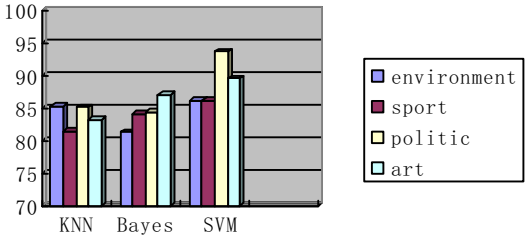

Figure 4. The comparison of Comprehensive index F1 value

C. Experimental Analysis

By comparing and analyzing Table Ⅰ, Table Ⅱ and Table Ⅲ, we can draw the following conclusions:

From the view of accuracy rate, SVM classification method all achieved 86.03%; although the accuracy rate of KNN classification method in environment and art is higher than SVM, classification results are lower than 80% in sport and politic; similarly, the accuracy rate of Naive Bayesian classification method is higher than SVM only in environment, the other three types are not better than SVM.

From the view of recall rate, SVM classification method has also reached 86%; for the KNN classification method and Naive Bayesian classification method, the recall rate has a fluctuation up and down, and the difference is obvious. That is to say, the overall effect is not better than SVM.

From the view of F1 value, Fig.4 has made an intuitive comparison of classification results for different classifiers. We can clearly see that, SVM classification method in the four types of texts is higher than the other two classification methods.

As a comprehensive evaluation index for text classification, F1 test value is better to reflect the effects of a good or bad classifier, so as a whole, SVM classification method is superior to other classification methods.

VI. CONCLUTIONS

Support Vector Machine is based on structural risk minimization principle and the limited sample of information, and it searches the best compromise between the complexity of the model (the learning accuracy of a particular training sample)and learning ability(error-free ability to identify any samples) in order to expect the best generalization ability [9]. In this paper, the experimental data well verify the value of support vector machine. However, as an emerging technology, support vector machine still has some requirement in-depth study and improvement in theory and practical application. On the one hand, the user of SVM must be given an error parameter C or g, and this parameter has a great impact on the results [9]. But the C value and the g value are highly subjective setting, and the user must guess the value of the various possible to find the best results. On the other hand, in the kernel function choice, the researchers studied the choice of kernel function via a priori knowledge, but how to choose the best kernel for the specific problem is still a difficult problem [10]. So how to promote the SVM to text classification applications better, will remain our ongoing efforts to study direction in the future.

ACKNOWLEDGMENT

The research work is supported by 863 Key Program of China (2006AA010105), National Natural Science Foundation of China (60772081, 60872133), Beijing Municipal National Natural Science Foundation (4092015), Funding Project for Academic Human Resources Development in Institutions of Higher Learning Under the Jurisdiction of Beijing Municipality (PXM2007_014224_044677, PXM2007_014224_044676), and Scientific Research Common Program of Beijing Municipal Commission of Education (KM200910772022).

REFERENCES

[1] Vapnik V N.The Nature of Statistical Learning Theory[M].New York:Springer Verlag,1995.

[2] Cortes C,Vapnik V.Support Vector Networks[J].Machine Learning,1995,20:273-297.

[3] Gao Yuan, Dazhong Liu. A Comparison Study of Chinese Text Categorization[J](in chinese).Science&Technology Information(Science teaching and research),2008,(02):7-8.

[4] T-Y Kwok. Automatic Text Categorization Using Support Vector Machine[C]. Proc. Int. Conf. on Neural Information Processing, 1998. 347-351.

[5] Jichao Chen. Technology and Application of Support Vector Machine[J](in chinese). Science & Technology Information(Science teaching and research),2007,(25):490-491.

[6] Yongyi Chen, Xiaoding Yu, Xuehao Gao, Hanzhong Feng. A New Method For Non-Linear Classify And Non-Linear Regression Ⅰ: Introduction To Support Vector Machine[J](in chinese). Quarterly Journal of Applied Meteorology, 2004,15(03):345-354.

[7] Xiaodan Wang, Jiqin Wang. Research and Application of Support Vector Machine[J](in chinese). Journal of Air Force Engineering University(Natural Science Edition), 2004,5(03):49-55.

[8] Liu Ke. Text Classification based on KNN algorithm[J](in chinese).Science & Technology Ecnony Market,2009,(06):12-13.

[9] Li Gang, Shubao Xing, Huifeng Xue. Comparison on pattern analysis performance of SVM and RVM based on RBF kernel[J](in chinese). Application Research of Computers, 2009,26(05):1782-1784.

[10] Changchun Cui, Wenlin Liu, Junzhe Zheng. Theory and application of support vector machine[J](in chinese). Journal of Shenyang Institute of Engineering(Natural Science),2007,3(02):170-172.