Research on Classification of Chinese Text Data Based on SVM

Yuan Lin, Hongzhi Yu, Fucheng Wan, Tao Xu*

*Northwest Minzu University, National languages Information Technology research institute, Lanzhou city, Gansu province, China

Abstract. Data Mining has important application value in today’s industry and academia. Text classification is a very important technology in data mining. At present, there are many mature algorithms for text classification. KNN, NB, AB, SVM, decision tree and other classification methods all show good classification performance. Support Vector Machine’ (SVM) classification method is a good classifier in machine learning research. This paper will study the classification effect based on the SVM method in the Chinese text data, and use the support vector machine method in the chinese text to achieve the classify chinese text, and to able to combination of academia and practical application.

1 SVM introduction*

With regard to text categorization, it originated from a technology in the middle of the last century. In 1957, H.P.Luhn first proposed the classification of word frequency statistics [1]. In 1970, Salton et al proposed a Vector Space Model(VSM).Vector Space Machine will be abstract text features map into a vector form, the use of similarity between the vector to measure the similarity between the texts. The abstract problem of mathematical, simplistic, to facilitate the study. In the domestic text classification research started relatively late, the general is the use of foreign mature classification technology combined with the Chinese text characteristics will be applied to the Chinese text classification. For the Chinese text, how to get a reasonable application and improve the text classification algorithm in this application, increase the classification accuracy of the algorithm and reduce the classification time of the algorithm has become a Chinese text classification problem to be solved [2].

The classification of text data is divided into two steps to complete. The first step is to create a classifier that describes the predefined data classes. This step is the training phase of the text classification, and the classifier is constructed by the analysis of the classification algorithm and the learning of the training set. The data set consists of the label and the data content of the class of the training sample.

The second step is to build a good classifier, to test the test samples , and then through the classification effect to measure the advantages and disadvantages of the classifier. At present, there are many kinds of methods of classification, but the research shows that support vector machine (SVM) classification performance, especially the generalization ability is better than the traditional classification method[3].

Support vector machine (SVM) is a supervised learning model in machine learning, usually used...
for pattern recognition, classification and regression analysis. Based on the principle of structural minimization, SVM is solved as a linear constrained convex quadratic programming problem (QP), and the solution has a unique optimal hyperplane\cite{4}. This paper will study and apply the principle of support vector machine, and apply it to the text corpus data collected from the Internet, and verify the superiority of its classification by the effect of the final classification. To study and learning support vector machine has important significance.

2. Support Vector Machine principle

Support Vector Machine (SVM) is a classification method proposed by Corinne Cortes and Vapnik in 1995 to solve the classification of binary classification. It is a supervised learning algorithm based on statistical learning theory and structural risk minimization principle \cite{5}, which is widely used in statistical classification and regression analysis. The basic idea of support vector machine (SVM) is to find the optimal high dimension classification hyper plane.

2.1 Support Vector Machine thinking

The support vector machine was originally designed to solve the problem of two classifications. The basic idea is illustrated by the two-dimensional case in Figure 1. The solid and hollow points in the figure represent two samples, \( w^*x - b = 0 \) is classification, \( w^*x - b = -1 \) and \( w^*x - b = 1 \) are the straight lines of the nearest class and are parallel to the classification line, and the distance between them is called the sorting interval. In the figure, \( 2 / ||w|| \) is the largest sorting interval of the samples, so the problem of support vector machine can be transformed into the problem of seeking \( w \) and \( b \), so that \( 2 / ||w|| \) is the largest.

![Figure 1: Binary SVM thinking](image)

The support vector machine algorithm is applied to multidimensional, and then the problem is transformed into the problem of seeking the maximum classification interval to find the optimal hyper plane. Figure 2 is extended to multi-dimensional situation.

![Figure 2: Multidimensional SVM thinking](image)
2.2 SVM math’s model

2.2.1 Linearly separable situation. SVM proposed for the classification of the problem is a linear and non-linear classification method. For the problem of linear separable two classifications, it is assumed that the given data set \( D = \{(X_i, y_i), (X_2, y_2), \cdots, (X_n, y_n)\} \), \( X_i \) is the dimension of the training data element, and \( X \in R^d \), \( d \) is the dimension of the feature space of the sample. \( y \in \{-1, 1\} \) is the two categories of the data set \( D \) Labeling, respectively, corresponding to the two categories, the optimal classification surface should meet the formula (1)

\[
y_i \left( (w \cdot x_i) + b \right) - 1 \geq 0, i = 1, 2, \cdots, n \quad (1)
\]

In the formula (1), \( w \) is the weight vector, \( b \) is a scalar, commonly referred to as bias (bias), by definition, \( \|w\| \) is super-plane above any point to the separation of the super-plane distance. So the maximum edge is \( 2 / \|w\| \). This is necessary to meet \( 2 / \|w\| \) the largest, that is, the minimum value of the request \( \|w\| \). \( \|w\| \) is the Euclidean norm, and equal \( \sqrt{w \cdot w} \).

Constructing Lagrangian Function and KKT Condition for Optimal Classification Function:

\[
f(x) = \text{sgn} \left\{ (w^* \cdot x) + b^* \right\} = \text{sgn} \left\{ \sum_{i=1}^{n} \alpha_i y_i (x_i \cdot x) + b^* \right\} \quad (2)
\]

In the above equation (2), \( b^* \) is the threshold of classification.

2.2.2 None-linear separable situation. Above 2.2.1 describes a linear separable case, but in the case of linear indivisibility, in such a case, there is no feasible solution that SVM can not find if it can separate these classes. In this case, you can expand the linear SVM, for the inseparable data to establish a non-linear SVM. The following two steps can be achieved by expanding the linear SVM to the nonlinear SVM. The first step is to use the nonlinear mapping to transform the original data into a higher dimension space; the second step is to search for linear separation hyper plane in the new high dimensional space.

How can we use a smaller amount of computation to choose a non-linear mapping to a higher dimension space? This is a question that we have to consider. Through the previous study, the use of appropriate kernel function can be a good solution to this problem\(^6\). The following kernel functions are usually used:

H times a number of kernel functions: \( K(X, X_i) = (X \cdot X_i + 1)^{\frac{1}{2}} \) \quad (3)

Radial basis function: \( K(X, X_i) = e^{-\|X - X_i\|^2 / 2\sigma^2} \) \quad (4)

Sigmoid kernel function: \( K(X, X_i) = \tanh(kX \cdot X_i - \delta) \) \quad (5)

3. The design of chinese text classification SVM

This paper mainly introduces the selection of some parameters involved in the design of Chinese text classification and the application of some technical models. For the entire classification of the system, mainly contains the training phase and the test phase of two stages.

The training phase is the key stage of constructing the whole classifier. Since the Chinese text is different from the particularity of the English text, it is necessary to segment the whole training sample data and then remove some commonly used stop words. Because too many characteristic words will make the dimension of the text too high, it is easy to cause the dimension of the disaster. So in the sub-word text data based on the choice of feature words, commonly used feature selection methods include word frequency, mutual information, information Gain, statistics and other methods\(^7\), and then through the text representation model to convert its text corpus to represent the true meaning of the document form, this paper will use the vector space model to express the text, and then through the SVM model of text corpus training, Get training model.

In the testing phase, like the training phase, it is necessary to carry out word segmentation and filter out the non-characteristic vocabulary of those forecasting stages according to the characteristic vocabulary obtained in the training phase, so that the characteristic words of the prediction phase are derived from the feature lexicon obtained in the test phase. Then, the document that needs to be tested...
is represented by the model. Finally, the test text is classified and verified by the built classifier. According to the analysis result, it is finally possible to re-evaluate the time and accuracy Chinese text classification system.

Figure 3 is the general process of classification.

![Text classification process](image)

Figure 3: Text classification process

4. Experiment platform
Cpu: Intel(R) Core(TM) i3-6300 CPU @3.80GHz
Run memory: 4.00GB
Operating System: Microsoft Windows7
Software Environment: Jet Brains PyCharm Community Edition 2016.3.
Development language: Python 2.7.13

5. Experimental analysis and results
5.1 Experimental analysis
The experiment was based on the open source system tmsvm and python language to complete. In the system of Taiwan University Professor Lin Zhiren developed Libsvm system, and the experiment was classified. Using the tmsvm system to provide the python interface, call this interface on the text automatically SVM model training. Including the selection of Libsvm, Liblinear package, word segmentation, dictionary generation, feature selection, SVM model training process. And then use the generated model to predict the unknown text and return the predicted label and the membership score. The effect of the model is evaluated by calculating the F value, recall rate, accuracy rate, Macro and micro index of the predicted results.

5.2 Experimental results
(1) Data set1
Data set 1 is a test of two categories of samples, contains a total of two categories, a total of 3,000 training samples, including training samples for the 2000, the test sample of 1000, where the category 1 on behalf of economic data, category -1 on behalf of sports data. The data analysis results are shown in Table 1.

| Table 1 Two categories of problem validation |
|-------------------------------------------|
| Cross validation accuracy                  |
| Label 1 | 0.945 |
| Label 2 | 0.950 |
| Macro   | 0.948 |
| Micro   | 0.948 |
Note: Macro is Macro average, Micro is Micro average
(2) Data set 2
Data set 2 is a multi-sample classification of the verification, contains a total of four categories, a total of 6000 data samples, including training samples for the 4000, the test sample of 2000, of which Category 1 is economic data, category 2 is health Class data, category 3 is the sports class data, category 4 is the tourism class data. The data analysis results are shown in Table 2

| Label 1 | Label 2 | Label 3 | Label 4 |
|---------|---------|---------|---------|
| Cross validation accuracy | 0.912 | 0.93 | 0.988 | 0.94 |
| Macro | 0.945 |
| Micro | 0.945 |

5.3 Experimental Conclusion
It can be seen from Table 1 and Table 2 that SVM has achieved high accuracy for two classification problems and multi-classification problems. This proves that the SVM model has good classification performance in the text classification. It is very important to apply the SVM model to the classification of Chinese text, which is of great significance to the theoretical research and practice of Chinese text classification.

5.4 Shortcomings and improvements
Although this study achieved good classification performance, but in terms of the selection of data sets, there are certain problems. In the paper, the SVM is function selection is the method of RBF. RBF method in stability is obvious, but in training and testing data sets used in the time longer. So, in subsequent studies, I will use different kernel function to compare the performance of the classifier. And then select a can in stability, robustness and time performance superior kernel function to supplement and improve the papers.

Acknowledgements
This research supported by the Fundamental Research Funds for the Central Universities (NO. Yxm2015186).

References
[1] Luhn, H. P. (1958). Luhn, h.p.: the automatic creation of literature abstracts. ibm journal of research and development 2(2), 157-165. Ibm Journal of Research & Development, 2(2), 159-165.
[2] Han J, Kamber M. Data Mining: Concepts and Techniques [J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2004, 5(4):1 - 18.
[3] Liu H. Dealing with the End Issue of EMD Based on Polynomial Fitting Algorithm[J]. Computer Engineering & Applications, 2004, 40(16):84-86.
[4] Cai D M, Gokhale M, Theiler J. Comparison of feature selection and classification algorithms in identifying malicious executables [J]. Computational Statistics & Data Analysis, 2007, 51(6):3156-3172.
[5] Cortes C, Vapnik V. Support-Vector Networks.[J]. Machine Learning, 1995, 20(3):273-297.
[6] Han J, Kamber M. Data Mining: Concepts and Techniques [J]. Data Mining Concepts Models Methods & Algorithms Second Edition, 2004, 5(4):1 - 18.
[7] Yang Y, Pedersen J O. A Comparative Study on Feature Selection in Text Categorization[C]// Fourteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 1998:412-420.