International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.4 (2015), pp.283-290
http://dx.doi.org/10.14257/ijmue.2015.10.4.27

The SVM based Uyghur Text Classification and its Performance Analysis

Palidan Tuerxun[1,2], Fang Dingyi[1] and Askar Hamdulla[2]

[1] *School of information and technology, Northwestern University, Xi'an, China*
[2]*School of Software, Xinjiang University, Urumqi, Xinjiang, China*
askarhamdulla@gmail.com

Abstract

This paper mainly explores the use of Support Vector Machines (SVMs) for Uyghur text classification, presents the process of text categorization: Text preprocessing, feature dimensionality reduction, representation method and classification of text features etc., discusses the SVMs classification algorithm in the application of Uyghur text classification. Focus on the construction of text categorization model and its procedures. Experiment results show that training by using the selected training data with the guarantee of the performance of the classifier, has higher efficiency than other nearest neighbor classifier (KNN), Naive Bayes (NB) classifier with increased accuracy.

Keywords: *Uyghur Language; Text Classification; SVM; Stem Extraction; NB classifier*

1. Introduction

Text Classification is a supervised machine learning method, the research and implementation of text categorization is a branch task of natural language understanding and machine learning. Text categorization not only can be used for a single language, but also can be used for cross-language, it helps users to distinguish between a large number of information, to screen and select in accordance with the user's interest. Classification principle is valid classification mechanisms obtained by learning and summarizing the positive and negative examples of known things, thereby differentiate the unknown things.

With the rapid development of information construction in Xinjiang, digital materials in minority languages such as Uyghur, Kazakh begin to expand very rapidly. Many applications need to integrate and effectively use the mass text messages via computer by automatic classification methods. The Uyghur text classification research started relatively late, paper [1] In the Uyghur, Kazakh, Kirgiz search engine results automatic classification techniques recall of 70% (based on the results of a small amount of expected KNN). Paper [2] in the Uyghur text classification based on machine learning research results the recall rate is about 70% (based on the minimal amount of expected results). Text classification as an important information collation and collection methods have great significance to help users to accurately locate the required information and to deal with the growing clutter of information.

2. Text Categorization

Text categorization is one of the hot research topics in the field of information processing. The goal of text categorization is the classification of documents into a fixed number of predefined categories. Each document can be in multiple, exactly one or no category at all. Using machine learning, the objective is to learn classifiers from examples which perform the category assignments automatically. During the process of

ISSN: 1975-0080 IJMUE

classification, first, the training text represented as a vector in a form element (typically words), then according to some methods of feature extraction, express features by using weights, this allows the training of element - weight text vector can get the vector model, When categorizing, the classification text will be represented as text element – weight text vector, and then make a comparison with the model which is obtained after training. General text classification process is shown in Figure 1.

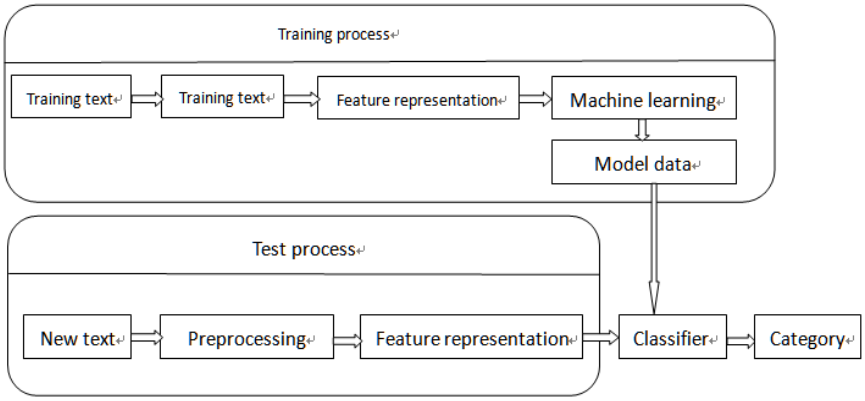

Figure 1. The Procedures of Text Classification

3. Main Title

3.1 The Features of Uyghur Language

Uyghur belongs to the Turkish language family of the Altaic language system. Its characteristics are as follows:

(1) The writing direction of the Uyghur language is from right to left, and from top to down.
(2) Uyghur uses Arabic and Persian letters, consists of 32 letters and is an alphabetical language.
(3) Uyghur is totally different from Chinese and English. It is a highly agglutinative language. word is the smallest unit in this language and which can be used independently, words are formed by affixes attaching to the stem (or root)

3.2 Pre-processing of Uyghur Texts

Text preprocessing is the first step and one of the most important parts in the Uyghur text classification; it mainly includes the text de-noising (recognition and removal of the none-Uyghur characters, stop words filtering), stemming etc.

In classification, the word can be seen as a feature unit; therefore, segmentation of the text and select proper words set which can reflect the real content of the text is the key point in text representation. It is also the hardest part of Chinese text processing. Segmentation is not a difficult task for alphabetical languages such as Uyghur; Words are separated by blanks or other punctuation. For example:

شنجاڭ جۇڭگونىڭ غەربىي قىسمىغا جايلاشقان.

In this text, five words are separated by four spaces.

For Uyghur, word segmentation is not critical; stem segmentation is the hardest task [2], extract the most important parts of words which is essential in expressing the real

meaning of words as features. Making stems as features is an effective way to reduce the number of features. for example, as follows are consists of same stem but different affix

مەكتەپتە ، مەكتەپتىن ، مەكتەپنىڭ ، مەكتەپكە ، مەكتەپنى

(At school, from school, of school, to school, the school),

The essential meaning of these words lies in its stem "مەكتەپ" (school). If we take this word as a feature, then it has 5 features, Stemming as their word as the feature item, then the median dropped to a text feature1.

On the other hand, this can effectively eliminate the negative impact of affixes to the similarity calculation of texts. If the five words above viewed by word order that is completely five different features of the word. However, according to its stem, that is completely the same feature, there is a certain correlation between the texts with same word stem.

Stop words filtering: obtaining words set after the stop words filtering. In order to make stop words filtering we prepare stop words list first which includes words that do not have much effect on the text representation and words which has same high frequency in the text. After filtering stop words, we further realize our purpose to accurately represent the content of the text. Below are some stop words in Uyghur:

ئۇ,يە نە,بىلەن,ۋە,ھەم,ياكى,بەلكى,لېكىن,بىراق,خۇىسۇسەن.....

Text preprocessing is the most important part of the Uyghur text classification; it has a direct impact on classification results. First, getting stem feature sets after filtering stop words, stemming and removing Non-Uyghur characters in Uyghur In order to achieve better classification results, select features from those stem feature sets.

3.3 Feature Selection

The Common feature selection methods are as following: document frequency (DF), information gain (IG), mutual information (MI), $\chi 2$ statistics (CHI) and so on. In this paper, we use $\chi 2$ statistics as a statistical method of feature selection; the $\chi 2$ statistic measures the lack of independence between t and c and can be compared to the $\chi 2$ distribution with one degree of freedom to judge extremeness. Using the two-way contingency table of a term and a category c, where A is the number of times t and c co-occur B is the number of time the t occurs without c, C is the number of times c occurs without t, D is the number of times neither c nor t occurs, and N is the total number of documents, the term-goodness measure is defined to be:

$$x^2(t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{1}$$

The $\chi 2$ statistic has a natural value of zero if t and c are independent. We computed for each category the $\chi 2$ statistic between each unique term in a training corpus and that category, and then combined the category-specific scores of each term into two scores:

$$x^2 \max(t) = \max_{1<i<m} \{x^2(t,c_i)\} \tag{2}$$

4. The SVM based Text Classification Algorithm

SVM is a classification method which has a good performance in pattern classification. The original SVM algorithm was invented by Vladimir N. Vapnik [3] with his teammates at Bell Laboratory, and its theory derived from the statistical learning theory proposed by V. Vapnik.

In recent years, the theory has made remarkable progress in automatic text classification. SVM is a kind of novel machine learning method based on the VC dimension theory and structural risk minimization principle of Statistical Learning Theory (SLT), is now recognized as the most effective machine learning method for text classification. The following two types of classification are given as an example to illustrate the idea of SVM. As shown in Figure 2, cross and hollow points represent two types of samples, O1 is classification line. H1 and H2 respectively the nearest and parallel samples to the classification line in each class, the distance between them is called the class interval (margin). The so-called optimal classification line (plane) is the line which can not only to separate the two classes correctly (training error rate is 0), but also with the largest classification interval. Assuming an m-dimensional classification line equation$w \cdot x + b = 0$,$w \in R^m$ and its classification interval is$2/\|w\|$, when it uses$[\![w]\!]^2$, the Classification line is the optimal, here is the need to satisfy the constraint condition:

$$y_i(w \cdot x_i + b) \gg 1, i = 1,2,\cdots,n \tag{3}$$

Wherein, n is the number of samples, $y \in \{1,-1\}$, positive example is 1, counter-examples is -1. Positive example includes features of the essence of the concept; Counterexample does not include features of the essence of the concept

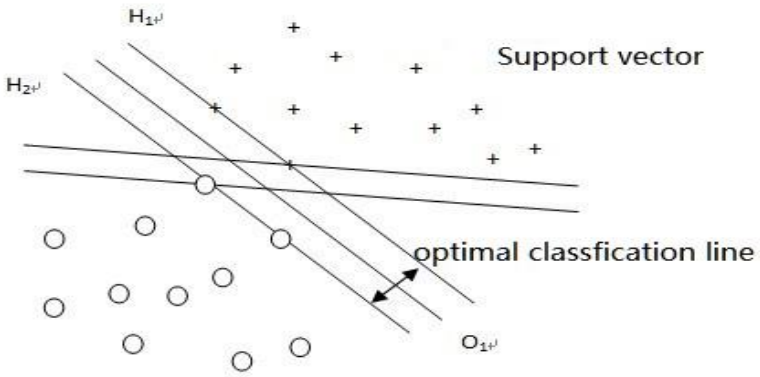

Figure 2. Data Points around the Classification Line Sets

At this point, the sample points on H1 and H2 are called support vectors. Using Lagrange optimization method we can transform the above optimal classification face problem into the dual problem as follows:

$$maxQ(\alpha) = \sum_{i=1}^{n} \alpha - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j) \tag{4}$$

Among $0 \ll \alpha_i \ll \gamma$; $i = 1,2,\cdots,n$; $\sum \alpha_i y_i = 0$, $\alpha = (\alpha_1, \alpha_2, \cdots \alpha_n)^T$. This is a quadratic optimization problem with inequality constraints, and there exists a unique optimal solution$(\hat{a}, \hat{b})$, which corresponds to the optimal decision function as below:

$$f(x) = sgn\{\sum_{i=1}^{n} \hat{a}_i k(x_i, x) + \hat{b}\} \tag{5}$$

SVM method is different from conventional statistical and neural network method; it does control the complexity of the model by reducing the number of features. SVM provides a functional complexity method which has nothing do with the dimension of the problem it introduces a high-dimensional feature space, transform the non-linear decision boundaries in input space into high linear decision

boundaries in feature space, the use of a linear function of the dual-core, solved the numerical optimization quadratic programming problems. The commonly used kernel functions are mainly three types: polynomial kernel, radial basis kernel function and S forms the kernel function. Depending on the classification problem, you can choose a different kernel functions.

5. Experimental Results and Analysis

5.1 Experimental Corpus

Since Uyghur text categorization started relatively late, therefore, has not publicly released any text corpus, therefore, we collect a lot of the Uyghur texts from Internet and then established large-scale text corpus through manual classification. The corpus includes real estate, health, education, military, tourist attractions, star (art class), people, phone (common sense, market, trading), art, politics, computers, transportation, economy, history, national language and customs, car (common sense, market, trading), social and legal system, sports, recruitment, religion, etc. each category contains 300 texts, a total of 6000 texts. For convenience, these corpuses are divided into four categories; Table 1 shows Experimental corpus species.

Table 1. Experimental Corpus

Corpus name	Number of categories	Total number of documents	Number of training documents	Number of test documents
D5	5	1500	220 in each categories with total 1100	80 in each categories with total 400
D10	10	3000	220 in each categories with total 2200	80 in each categories with total 800
D15	15	4500	220 in each categories with total 3300	80 in each categories with total 1200
D20	20	6000	220 in each categories with total 4400	80 in each categories with total 1600

5.2 Experimental Environment

The experiments were run on a PC which has a configuration as below:
CPU: Intel dual-core E7300 2.66GH
Hard drive: 250GB
RAM: 2GB
Operating System: Windows 7
Experiment Software: Microsoft Visual C # 2010 Express

5.3 Results of Classification based on the SVM

SVM classifier divides the texts into 5,10,15,20 categories As shown in Figure 3. Experimental results show that the above figure is the latest result of Uyghur text classification achieved by the NB classifier having 5 categories, 10 categories, 15 categories, 20 categories, etc. The results are 95.65%, 93%, 77%, 76%. SVM classifier is more complex than KNN, NB since there are many parameters, classification results depends on the parameters variety.

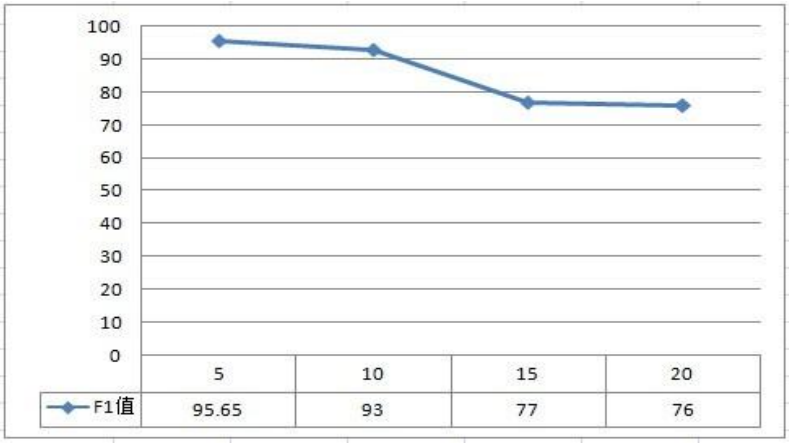

Figure 3. SVM Classifiers Results

5.4 Comparison Results of SVM, KNN and NB

Performance comparison of KNN, NB, SVM classifiers (according to the accuracy), As shown in Figure 4. The results show that the above figure is the latest result of Uyghur text classification by the KNN, NB, SVM classifier, have 5, 10, 15, 20 categories. Experimental results show that the SVM is the best classifier.

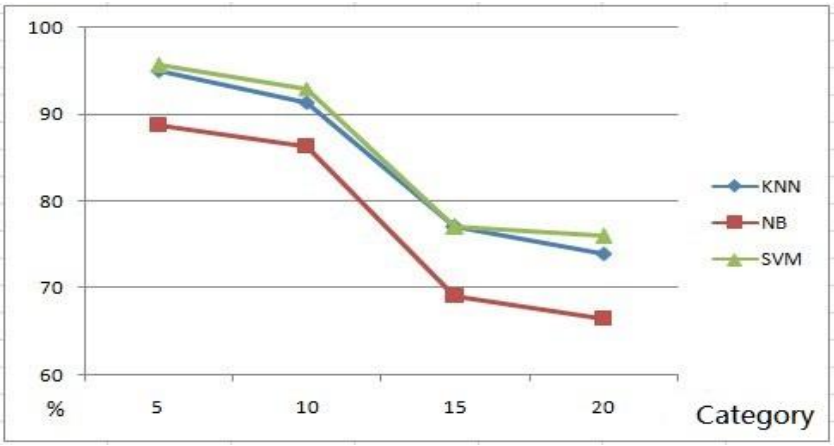

Figure 4. Performance Comparison Results of KNN, NB, SVM Classifiers (on Accuracy)

6. Conclusions

In this paper, according to the characteristics and writing rules of the Uyghur, we established (includes 20 categories, 300 texts in each category) large text corpus. Through studying and considering the features and syntax rules of the Uyghur and analyzing the impact of stem extraction to the accuracy and speed of Uyghur text classification. Due to the reduction of Vector space dimension in text categorization is a very important issue, for this, in this paper we employed stemming approach according to the lexical rules the Uyghur, this method does not affect the Uyghur text classification accuracy while achieving a good dimension reduction purposes. We employed CHI statistical feature selection method to extract features. By using large-scale text corpus, conducted Uyghur text classification experiments with SVM method and analyzed the performance of KNN, (NB), SVM, etc. on Uyghur texts.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China under the grant number of 61163033.

References

[1] Z. Wang and W. Musa, "Automatic classification technology of search engine results", Master's Thesis, Xinjiang University, Urumqi (**2010)**.
[2] A. Aysa, T. Ibrahim, H. Omar and M. Ali, "The Uyghur text classification based on machine learning research", Computer engineering and application, vol. 5, (**2012**).
[3] V. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, New York (**1995**).
[4] T. Tohti, W. Musajan and A. Hamdulla, "Uygur language, ha, key full-text search engine retrieval device key technology", Computer Engineering, vol. 21, (**2008**).
[5] X. Li and L. Fei, "Data mining and knowledge discovery", Higher Education Press, Beijing (**2003**).
[6] Bi. ng Liu, "Web data mining", Tsinghua University Press, Beijing(**2009**).
[7] P. Zhang, "Chinese text classification feature selection method based on X^2 statistics research", Doctoral Dissertation, Chongqing University, Chongqing (**2008**).
[8] P. YingBo, "Chinese text classification feature selection method in the research and implementation", Doctoral Dissertation, Northwest University, Xi'an (**2010**).
[9] X. Wang, "Chinese text classification feature selection method research", Southwest University, Xi'an (**2010**).

Authors

Palidan Tuerxun, she received her M. S. degree in 1996 from Liaoning University, China. She is currently working toward PhD. degree in Northwestern University, China. Since 1992, she has been working as a teacher at Xinjiang University, and since 2004, she was an associate professor in school of software of Xinjiang University. Her research interests are machine learning and Uyghur natural language processing.

Fang Dingyi, he currently is a Professor in the School of information and technology, Northwestern University, Xi'an, China. His research interest includes networking and information security.

Askar Hamdulla, he received B.E. in 1996, M.E. in 1999, and Ph.D. in 2003, all in Information Science and Engineering, from University of Electronic Science and Technology of China. In 2010, he was a visiting scholar at Center for Signal and Image Processing, Georgia Institute of Technology, GA, USA, tutored by Professor Biing-Hwang (Fred) Juang. Currently, he is a professor in the School of Software Engineering, Xinjiang University. He has published more than 120 technical papers on speech synthesis, natural language processing and image processing. He is a senior member of CCF and an affiliate member of IEEE.

