Ensemble Learning for Text Classification

1st Jyun-Hao Lai
3rd Chang-Biau Yang*
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan

2nd Kuo-Tsung Tseng
Department of Shipping and Transportation Management
National Kaohsiung Marine University
Kaohsiung, Taiwan
tsengkt@nkmu.edu.tw

***Abstract*—In this paper, we apply ensemble learning with SVM classifiers for text classification problem. Our experimental dataset was downloaded from the Yahoo news web site. The dataset consists of about the 50,000 Chinese news in 9 classes. We constitute these news documents into five data sources: (1) full text, (2) title, (3) first paragraph, (4) full text and title, and (5) title and first paragraph. We then use three feature generation methods (a) TF-IDF, (b) χ^2 and (c) IG to produce the feature vector from each document and adopt the SVM method as our basic classifier, thus 15 SVM classifiers are trained. Next, we choose three of them to constitute an ensemble classifier by the BKS method, so totally $\binom{15}{3} = 455$ ensemble classifiers are constructed. The experimental results show that the ensemble classifier formed by (a) TF-IDF with "(4) full text and title", (b) χ^2 with "(2) title" and (c) IG with "(2) title" has a good prediction accuracy 79.04%.**

***Index Terms*—Chinese text classification, support vector machine (SVM), behavior knowledge space (BKS), ensemble learning**

I. Introduction

Given some predefined classes, the aim of the text classification problem is to classify the documents into these classes [16]. There are many applications of the text classification problem such as the advertising identification [3], the spam detection [7] and the news classification [4]. Nowadays, the number of documents grows rapidly. This is that not only it is easy for all people to publish their documents on the internet, but also almost all companies use electronic document systems for effectiveness. Therefore, how to classify such a huge number of documents automatically and effectively becomes an important issue.

Many researchers have proposed various methods to solve the text classification problem [9], [10], [19]. Yang and Pedersen compared the performances of different feature generation methods in 1997 [18]. Next year, Joachims used the *support vector machine* (SVM) with many relevant features to assign documents, i.e. news in Reuters, to classes [9].

However, there are few researches focusing on the Chinese documents classification problem. Hence, in order to fill this gap, our goal is to solve the text classification problem with Chinese documents. In fact, it is a pity that most text classification researches apply the single classifier only, thus their classification accuracies may be low in some particular types of documents. Therefore, we use the machine learning technique to analyze Chinese documents and ensemble multiple classifiers with the *behavior knowledge space* (BKS) method to integrate the predicted results for improving our prediction accuracy.

The organization of this paper is as follows. Section II provides some background information about the text classification problem and ensemble learning. In Section III, we present our method for solving the problem. Experimental results with Chinese documents are given in Section IV. Finally, we give our conclusions in Section V.

II. Preliminaries

A. The Chinese Text Classification Problem

Let $D = \{d_1, d_2, \cdots, d_n\}$ be a set of documents and $C = \{c_1, c_2, \cdots, c_m\}$ be a set of classes. The *text classification problem* is to assign the document d_i to the class c_j, where $1 \leq i \leq n$ and $1 \leq j \leq m$. For example, according to the content of a news article, the article can be assigned into the class like sports, finance, or politics.

The *Chinese text classification problem* is nothing different from the original problem but with documents written in Chinese. It is more challenging than English documents since there are no blank spaces in Chinese sentences to extract meaningful terms (words) to represent the document. In Chinese text classification problem, we should first perform the Chinese segmentation to separate Chinese sentences into a list of terms and eliminate those terms without meaning or frequently used. Afterward, we can apply the *vector space model* to represent our Chinese documents.

The vector space model was proposed by Salton *et al.* in 1975 [15] and is the most common way to solve the text classification problem. Suppose that r terms are selected to represent one document. Then, a document can be expressed as an r-dimensional vector composed of term weights. For a document d_i, we can use $\langle w_{i,1}, w_{i,2}, \cdots, w_{i,r} \rangle$ as a feature vector to represent d_i.

Table I shows an example of the vector space model representation, where there are three documents and four feature terms (t_1, t_2, t_3 and t_4). In the table, each weight in the feature vector corresponds to the number of term occurrences in the document.

[*] Corresponding author: cbyang@cse.nsysu.edu.tw

TABLE I
AN EXAMPLE OF THE FEATURE VECTORS FOR REPRESENTING THE DOCUMENTS.

Terms Documents	t_1	t_2	t_3	t_4	Feature vector
d_1	1	0	1	2	$\langle 1, 0, 1, 2 \rangle$
d_2	2	0	3	1	$\langle 2, 0, 3, 1 \rangle$
d_3	0	4	2	2	$\langle 0, 4, 2, 2 \rangle$

B. Feature Selection Methods

When applying the vector space model to solve the text classification problem, the fundamental issue is that how we choose those r terms to represent our documents. Thus, we need some feature selection methods to select more significant terms in the documents through the probability or statistic since some terms in the documents are not so meaningful. The details of our feature selection methods are described as follows.

1) Term Frequency-Inverse Document Frequency (TF-IDF): A common way of feature generation methods in the text classification problem is to use *term frequency-inverse document frequency* (TF-IDF) which was proposed by Salton *et al.* [15] in 1975. The TF-IDF calculates the frequency of each term and the number of its appearances in the text.

The idea of TF-IDF is that if the frequency of a term is high in a class and it is rare in other classes, then this term is significant and should be used as a feature term for the class. The term weight of TF-IDF can be calculated by the following equations, where n is the number of documents, $tf(d_i, t_j)$ is the number of occurrences of the term t_j in a document d_i, and $df(t_j)$ is the number of documents that contain term t_j.

$$idf(t_j) = \ln(\frac{n}{df(t_j)}) + 1. \tag{1}$$

$$tf\text{-}idf(d_i, t_j) = tf(d_i, t_j) \times idf(t_j). \tag{2}$$

2) Chi-square Statistic (χ^2): The χ^2 statistics is used to measure the relationship between term t_i and class c_j, where $1 \leq j \leq m$. If the term t_i is not related to class c_j, then the χ^2 statistic value is zero. The χ^2 statistics between term t_i and class c_j is given as follows [18]:

$$\chi^2(t_i, c_j) = \frac{n \times (o_{ij} r_{ij} - p_{ij} q_{ij})^2}{(o_{ij} + q_{ij}) \times (p_{ij} + r_{ij}) \times (o_{ij} + p_{ij}) \times (q_{ij} + r_{ij})}, \tag{3}$$

where $n = o_{ij} + p_{ij} + q_{ij} + r_{ij}$ is the total number of documents, and the meanings of other variables are shown in Table II.

TABLE II
THE MEANINGS OF VARIABLES FOR CALCULATING $\chi^2(t_i, c_j)$.

# of documents that	$\in c_j$	$\notin c_j$	total
$\supset t_i$	o_{ij}	p_{ij}	$o_{ij} + p_{ij}$
$\not\supset t_i$	q_{ij}	r_{ij}	$q_{ij} + r_{ij}$
total	$o_{ij} + q_{ij}$	$p_{ij} + r_{ij}$	n

3) Information Gain (IG): The idea of *information gain* (IG) is to measure how much information is carried by a term. It is possible that the dimension of the feature vector may be reduced if the selected terms are with high information gain. The information gain of a term t_i is calculated as follows [18].

$$\begin{aligned} IG(t_i) = & -\sum_{j=1}^{m} P(c_j) \log P(c_j) + \\ & P(t_i) \sum_{j=1}^{m} P(c_j|t_i) \log P(c_j|t_i) + \\ & P(\bar{t_i}) \sum_{j=1}^{m} P(c_j|\bar{t_i}) \log P(c_j|\bar{t_i}), \end{aligned} \tag{4}$$

where m is the number of classes, $P(t_i)$ is the probability of t_i appearing in a document, calculated as follows:

$$P(t_i) = \frac{1}{n} \sum_{k=1}^{n} \begin{cases} 1, & \text{if } t_i \text{ appears in } d_k \\ 0, & \text{otherwise.} \end{cases}$$

$\bar{t_i}$ denotes the case where the term t_i does not appear in a document, whose probability is equal to $1 - P(t_i)$. Information gain considers the influence of a term when we perform the classification.

C. The Ensemble Learning with the Behavior Knowledge Space Method

Ensemble learning integrates multiple learners to accomplish the learning tasks [6]. For example, three text classifiers can be integrated as a better prediction classifier. In ensemble learning, we use the training set to build some single classifiers, where each classifier may use a different learning algorithm with different features, such as *decision tree*, *k-nearest neighbor* and *support vector machine* (SVM). Finally, these predictions from different classifiers are combined to produce the final prediction result.

The *behavior knowledge space* (BKS) is a method to form an ensemble classifier with multiple classifiers [8], [14]. Assume there are m possible classes in a dataset D and δ classifiers have been built. Then the BKS table consists of m^δ entries. Each entry combines the prediction results from δ classifiers, where the document counts of with the real classes are calculated. In the testing stage, we search for the BKS table and select the prediction with the most appearance of the real class as the answer.

III. OUR METHOD

A. The Dataset

There is no public dataset for Chinese text classification problem, so we collect the news articles from Yahoo news web site during August 2016 to January 2017 [2]. The period of our collection is about once a week (on Thursday), in case that news articles in close time may be duplicate. Our dataset contains totally 52,064 news articles in 9 classes. The detail information of our dataset is shown in Table III.

TABLE III

THE DETAILED INFORMATION OF OUR DATASET. CELLS WITH TWO NUMBERS SHOW THE AVERAGE LENGTHS AND THE STANDARD DEVIATIONS OF LENGTHS, RESPECTIVELY. (1)∼(5) ARE DATA SOURCES DESCRIBED IN STEP 1 OF SECTION III-B.

Class	Art & education	Entertainment	Finance	Politics	Health	Society	Sport	Technology	Travel
(1)	528.7∥ 404.1	416.1∥ 359.9	656.2∥485.7	588.1∥465.8	770.0∥446.0	443.9∥290.3	478.6∥323.4	815.3∥730.5	490.5∥492.8
(2)	015.7∥ 004.5	018.9∥ 005.8	017.3∥005.3	017.6∥004.9	017.0∥004.0	015.8∥003.6	018.9∥004.6	022.6∥009.0	015.9∥003.8
(3)	093.3∥ 048.3	082.4∥ 048.4	090.7∥046.8	086.1∥064.5	092.9∥054.5	089.6∥042.9	089.1∥040.9	141.5∥242.6	090.0∥047.4
(4)	544.4∥ 405.0	435.0∥ 359.4	673.6∥487.1	605.7∥467.2	787.3∥467.6	459.8∥291.1	497.5∥325.0	837.9∥733.4	506.4∥293.1
(5)	109.0∥ 048.1	101.3∥ 048.4	108.0∥047.2	103.7∥065.0	109.9∥055.0	105.0∥043.0	107.9∥041.9	164.1∥245.3	105.9∥047.3
Count	2578	2687	6434	8394	2120	8817	9204	6344	5486

B. The Proposed Method

We give the detail of our proposed method as follows.

Step 1: Constitute the five sources from each document in the training dataset: (1) full text, (2) title, (3) first paragraph, (4) full text and title, and (5) title and first paragraph, and then perform Chinese segmentation to get terms and eliminate meaningless / frequently used terms. Our Chinese segmentation uses the Jieba segmentation module which is built by python [1].

Step 2: Utilize the following three feature generation methods to produce the feature vectors of each source: (a) TF-IDF, (b) χ^2, (c) IG. After this step is done, there are 15 types of feature vectors, since three feature generation methods are used for five sources.

Step 3: Build 15 SVM classifiers for all feature vectors. We adopt the linear kernel of SVM and invoke LIBSVM in our experiments [5].

Step 4: Establish a BKS table to construct an ensemble classifier for every three of all 15 SVM classifiers. Totally 455, $\binom{15}{3} = 455$, ensemble classifiers are constructed.

Step 5: Calculate the performances of the 455 ensemble classifiers to find out the superior SVM classifiers.

Figure 1 shows the architecture of our method for solving the Chinese text classification problem.

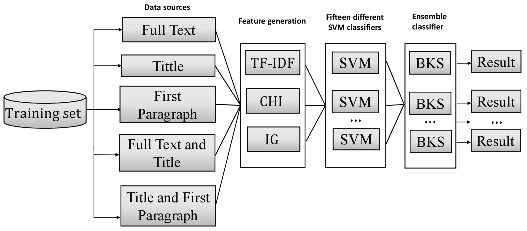

Fig. 1. The architecture of our Chinese text classification.

IV. EXPERIMENTAL RESULTS

In Section III-B, Step 1 and 2, we give the short representations of our data sources and feature selection methods. Thus, for easy explanation, if we use the method "TF-IDF" with the data source "full text and title" to train an SVM classifier, then we denote the SVM classifier as $a4$. Table IV shows the average accuracies of 15 SVM classifiers.

TABLE IV

THE AVERAGE ACCURACIES OF 15 SVM CLASSIFIERS.

Methods Sources	(a) TF-IDF	$(b)\chi^2$	(c) IG
(1) full text	$a1$, 75%	$b1$, 75%	$c1$, 73%
(2) title	$a2$, 64%	$b2$, 73%	$c2$, 71%
(3) first paragraph	$a3$, 71%	$b3$, 72%	$c3$, 71%
(4) full text & title	$a4$, 76%	$b4$, 75%	$c4$, 74%
(5) title & first paragraph	$a5$, 72%	$b5$, 73%	$c5$, 72%

Table V shows the 10 ensemble classifiers with the highest accuracies. In the table, each ensemble classifier is denoted as a 6-letter code, since each ensemble classifier is built by three SVM classifiers. For example, $a4c2c5$ means the three SVM classifiers $a4$, $c2$ and $c5$. The prediction accuracy is computed by the following equation.

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 1, & \text{if real class}_i = \text{predicted class}_i \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

where n is the number of testing documents.

TABLE V

THE TOP 10 ENSEMBLE CLASSIFIERS WITH THE HIGHEST ACCURACIES.

Ensemble classifier	Accuracy
a4c2c5	79.08%
a4c2c3	79.08%
a4c1c2	79.06%
a4b5c2	79.06%
a4b5c4	79.04%
a4b2c2	79.04%
b2b3c1	79.04%
a4c2c4	79.04%
a4b3c2	79.04%
a4b1c2	79.00%

It is interesting that in Table V, SVM classifiers $a4$ and $c2$ appear 9 and 8 times among 10 entries, respectively. This means that they may be significant classifiers for constituting the ensemble classifiers.

To explore the performance of the 15 SVM classifiers in more details, in Figure 2, we show the appearance amount of each SVM classifier constituting the ensemble classifiers of the top 10% accuracies to top 30% accuracies among the 455 combinations. Note that the accuracy of each of the 455 combinations is averaged from 5-fold cross validation for repeating 10 times.

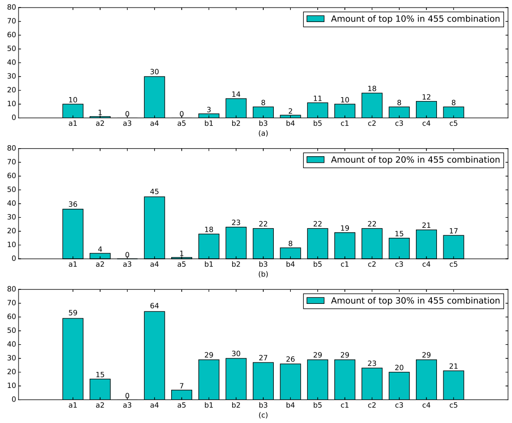

Fig. 2. The appearance amount of each SVM classifier constituting the ensemble classifiers among the 455 combinations, where the x-axis denotes SVM classifiers and y-axis denotes the appearance amount of each SVM classifier. (a) Top 10% accuracies of 455 combinations. (b) Top 20% accuracies of 455 combinations. (c) Top 30% accuracies of 455 combinations.

It is clear to see in Figure 2 that SVM classifier $a4$ is always greater than the others. A possible reason is that the title terms are often the keywords of a piece of news, so these keywords may appear many times in the full text. The "TF-IDF" method intends to increase the weights of these keywords, so it is natural to find better keywords than other methods, with the help of titles. On the other hand, "χ^2" and "IG" treat the title and the full text as an article and they only concern whether a term appears in a document or not, but they do not calculate the occurrence counts of one term.

V. Conclusion

There are many papers proposed for solving the text classification problem with English text, but few with Chinese text. In this paper, we study the Chinese text classification problem. We collect about 50,000 news articles from Yahoo website as our experimental dataset. We constitute five different data sources from the dataset. According to our experimental results, the feature generation methods "χ^2" and "IG" perform well with the data sources of news "titles". Moreover, the feature generation method "TF-IDF" is good with long text like the data source: "Full text and title". At last, we establish ensemble classifiers to improve the prediction accuracy. The accuracy of our suggesting classifier $a4b2c2$ assigning a news article to the correct class (totally 9 classes) is 79.04%. We do not compare our performance with others since there is no public dataset for our research.

It is interesting that all our classifiers with data sources about the "first paragraph", i.e. (3) and (5), do not perform well. We think that the first paragraph of a news article should be the abstract of the news. However, since the writing styles differ among journalists, the first paragraphs in our dataset are not abstracts as we have expected. The experimental results indicate that the full text of a document is better than a part of the document. In addition, if the document has a title, keywords or something like that, they should also be involved.

The Chinese classification problem still has a long way to go since the accuracies with Chinese text are not high enough. If it can be done with high accuracies, it may be applied to analyze if the last 40 chapters of *Dream of Red Chamber* (DRC) were written by Cao Xueqin [17], or to predict the stock trading if a company's financial news are given [11]–[13]. It is an interesting and worthy problem to study.

References

[1] "Jieba chinese text segmentation." https://github.com/fxsjy/jieba.

[2] "Yahoo news." https://tw.news.yahoo.com/.

[3] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel, "A semantic approach to contextual advertising," *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, USA, pp. 559–566, 2007.

[4] C.-H. Chan, A. Sun, and E.-P. Lim, "Automated online news classification with personalization," *In Proceedings of the 4th International Conference of Asian Digital Library (ICADL2001)*, Bangalore, India, pp. 320–329, Dec. 2001.

[5] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, No. 27, pp. 1–27, 2011.

[6] T. G. Dietterich, "Ensemble methods in machine learning," *Multiple Classifier Systems*, Vol. 1857, pp. 1–15.

[7] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *IEEE Transactions on Neural networks*, Vol. 10, No. 5, pp. 1048–1054, 1999.

[8] Y. S. Huang and C. Y. Suen, "The behavior-knowledge space method for combination of multiple classifiers," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '93)*, pp. 347–352, June 1993.

[9] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Proceedings of the 10th European Conference on Machine Learning (ECML)*, Chemnitz, Germany, pp. 137–142, 1998.

[10] E. Leopold and J. Kindermann, "Text categorization with support vector machines. how to represent texts in input space?," *Machine Learning*, Vol. 46, pp. 423–444, 2002.

[11] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Waha, and D. C. L. Ngo, "Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment," *Expert Systems with Applications*, Vol. 42, No. 24, pp. 306–324, 2015.

[12] T. H. Nguyena, K. Shirai, and J. Velcinb, "Sentiment analysis on social media for stock movement prediction," *Expert Systems with Applications*, Vol. 42, No. 24, pp. 9603–9611, 2015.

[13] W. Nuij, V. Milea, F. Hogenboom, F. Frasincar, and U. Kaymak, "An automated framework for incorporating news into stock trading strategies," *IEEE transactions on knowledge and data engineering*, Vol. 26, No. 4, pp. 823–835, 2014.

[14] Š. Raudys and F. Roli, "The behavior knowledge space fusion method: Analysis of generalization error and strategies for performance improvement," *Multiple Classifier Systems* (T. Windeatt and F. Roli, eds.), Vol. 2709 of *Lecture Notes in Computer Science*, pp. 55–64, 2003.

[15] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, Vol. 18, No. 11, pp. 613–620, 1975.

[16] F. Sebastian, "Machine learning in automated text categorization," *ACM Computing Surveys*, Vol. 34, No. 1, pp. 1–47, 2002.

[17] H. C. Tu, *A Text-Mining Approach to the Authorship Attribution Problem of Dream of the Red Chamber*. National Taiwan University, 2014.

[18] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text classification," *Proceedings of the 14th international conference on machine learning(ICML)*, Nashville, Tennessee, USA, pp. 412–420, 1997.

[19] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of TF*IDF, LSI and multi-words for text classification," *Expert Systems with Applications*, Vol. 38, No. 3, pp. 2758–2765, 2011.