Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

A Kind of Automatic Classification for Chinese Books Based on Ensemble Learning

Mingjie Zhou [1, a], Enke Li [2, b, *] and Yadan Tian [1, c]

[1]School of Economics and management, Xidian University, Xi'an, China

[2]Center of Journal Publications, Xidian University, Xi'an, China

[a]642067870@qq.com, [b]ekli@xidian.edu.cn, [c]179075482@qq.com

---

Abstract: In order to classify Chinese books accurately, this paper proposes an ensemble learning method for Chinese books automatic classification. This paper tries to use the ensemble learning and compare the efficiency of different base learners, such as Decision Tree, Support Vector Machines, Naïve Bayes and Back Propagation Nerve Net, then builds an effective Chinese books classification system. It was found that the classification accuracy of Back Propagation Nerve Net reaches 72% under the framework of ensemble learning algorithm Bagging. We only analyzed five categories of Chinese bibliographies, and the granularity of classification was coarse. By comparing with the traditional methods, the effectiveness of using ensemble learning in Chinese books automatic classification is proved, which improves the accuracy and reliability of automatic classification.
Keywords: Book Automatic Classification, Machine Learning, Ensemble Learning, Bagging algorithm.

---

1. INTRODUCTION

With the rapid development of information technology and automation technology, the digitalization of library resources and the automation of Library workflow have become the primary task of intelligent library construction. In the daily work of library, cataloguing is the basis of digitalization of library resources, and is also a complicated work. It needs under the existing knowledge classification system, such as "Chinese Library Classification" , to determine the category number of books, so as to realize effective management of massive books. With the increase of marginal subjects and cross disciplines, cataloguing is becoming more and more difficult. In the past, the Chinese bibliographic classification number was usually given by the author or the book cataloguer was manually determined. There are some problems in this method. First, the determination of the book classification number has a higher requirement for professionalism, while the author of the book has subjectivity, and its professionalism in the bibliographic classification needs to be considered. Second, the cataloguer can get the classification number by reading the contents of the books. It not only takes time and effort, but also works very efficiently, resulting in a waste of human resources, and there are

differences in the judgement of different librarians. The consistency of classification results is difficult to guarantee. In recent years, many scholars began to use knowledge engineering or machine learning methods to study Chinese Bibliographic automatic classification. However, there is little application of Ensemble Learning to this problem. Therefore, under the framework of ensemble learning, the algorithm of classifying the highest accuracy is determined by comparing the classification effect of different algorithms, so as to build an optimal classifier, so that the work of book classification is completely automated, and the efficiency of cataloguing work is improved, which provides an effective way for the construction of intelligent library.

2. RELATED WORK

At present, the research methods of Chinese Bibliographic automatic classification at home and abroad can be roughly divided into two categories: bibliographic classification method based on knowledge engineering, and bibliographic classification method based on machine learning.

The method of bibliographic classification based on knowledge engineering is to set up a classification knowledge base and classify by imitating the expert's classification idea. In 1980s, some scholars tried to introduce the expert system into bibliographic classification, and then continuously devoted to the design and implementation of bibliographic classification expert system. Zhang Hui put forward the method of using classification production rules to build knowledge base of books automatic classification, which reduced the difficulty of system construction to a certain extent. As an early artificial intelligence method, the expert system can basically realize the automatic classification of books, but there are still many problems in its construction, and the classification rules are determined by artificial experience. The level of the system depends on the organization of the knowledge base to a certain extent, and the accuracy and stability of the classification are difficult to guarantee. The classification of books in different fields requires experts from different fields, and the rules of classification are not easy to expand. The huge library classification knowledge base also increased the workload of manually determining classification rules.

Based on machine learning bibliographic classification method, this method first constructs the vector space of subject headings of each category, then calculates the subject word vector by machine learning algorithm, and finally obtains the similarity between topic similarity or topic and category, so as to realize automatic classification of bibliographies. Yang Min and Gu Jun used Support Vector Machines (SVM) to study the influence of different weights of Chinese Bibliographic titles and abstracts. Liu Gaojun and Chen Donghe put forward the automatic classification of bibliography by combining Term Frequency Inverse Document Frequency (TF-IDF) and Naïve Bayes (NB) algorithm. All of these studies have improved the Chinese bibliographic classification algorithm to some extent, but few Chinese Bibliographic classifications have been studied under the integrated learning framework. Therefore, based on the existing research, this paper attempts to study this problem from the perspective of ensemble learning.

Ensemble learning algorithm is an efficient classification algorithm in machine learning. In the field of Chinese text classification, Boosting and Bagging are the main classifiers, and many scholars have made improvements on this basis. Zhang Xiang proposed a credibility calculation for the result of K nearest neighbor classifier, and got an improved Bagging algorithm based on voting weight. Xu Kai

and other scholars build a base learner based on the Back Propagation Nerve Net (BPNN) algorithm, and form a Chinese text classification system. The above research proves that the classification task under ensemble learning can achieve higher accuracy and stability. Based on the excellent performance of ensemble learning, this paper tries to use this algorithm to learn the bibliographic text corpus. By comparing the classification effect of different basic learning devices after integration, a high accuracy Chinese Bibliographic automatic classification system is constructed.

3. CHINESE BIBLIOGRAPHIC AUTOMATIC CLASSIFICATION SYSTEM BASED ON ENSEMBLE LEARNING

3.1 System flow

The main process of the system is divided into two stages. The first stage is the training phase of the classifier, the second is the bibliographic classification stage. The specific process is shown in Figure 1,and the main tasks of each stage will be introduced in the following sections.

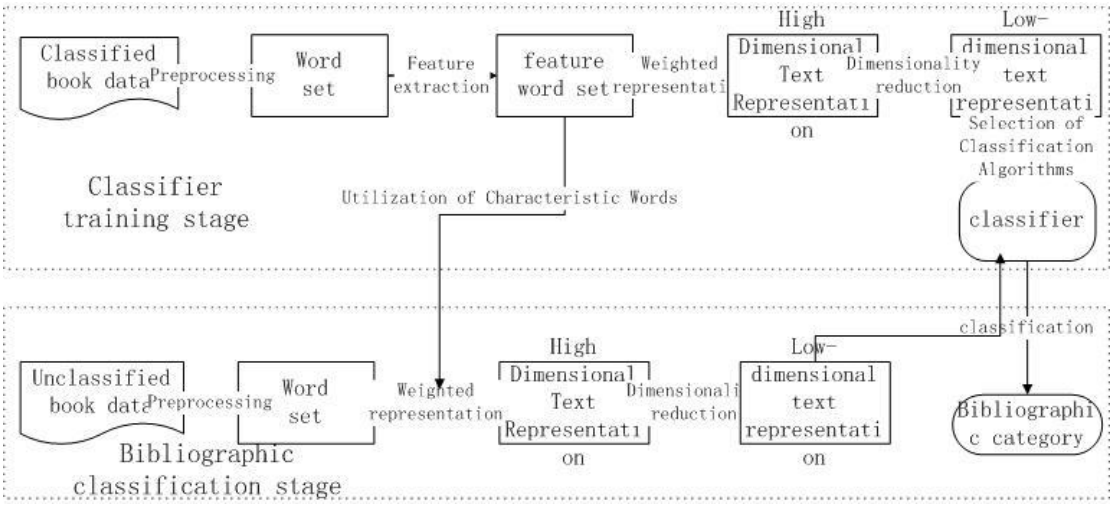

Fig. 1 Chinese Bibliographic automatic classification system flow chart

3.2 Data preprocessing

The main work of data preprocessing includes four steps:

(a)Manually delete invalid entries and duplicates in original bibliographic data to reduce the impact of data noise.

(b)Use regular expressions to eliminate punctuation and English characters in titles and abstracts of bibliographic data.

(c)Segmentation is done by using the jieba word segmentation tool. The word segmentation system first uses the Trie tree structure to implement word graph scanning to generate directed acyclic graphs composed of all possible word formation patterns in bibliographic data. Then we use the dynamic programming algorithm (Dynamic Programming) to find out the maximum probability path and form the largest segmentation combination based on the word frequency of the bibliography. Finally, using the HMM model based on Chinese word forming ability, we use the Viterbi algorithm to process the unknown word , and complete the whole word segmentation process.

(d)Comprehensive use of the "stop word" of the Harbin Institute of technology, Sichuan University's machine learning intelligent laboratory suspension dictionary, Baidu stop vocabulary, and manually add some stop words suitable for Chinese bibliography, such as '第一篇', '全书', '第一章' and so on. The operation of removing the stop words from the bibliographic data completed by word segmentation forms a set of words.

3.3 Feature extraction

The existing methods of feature extraction mainly include TF-IDF method, word frequency method, mutual information, expected cross entropy, information gain method and chi square statistics.

In this paper, we use the TF-IDF method to calculate the TF-IDF value of each word, set the threshold, and retain the word whose TF-IDF is greater than the threshold to get the set of feature words. Word frequency refers to the proportion of a word in all the words in the document.

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

Among them, numerator n_{ij} represents the number of words i appears in the j document, n_{kj} represents the number of word k appearing in document j. Denominator $\sum_k n_{kj}$ represents the number of all words in the j document. Reverse file frequency refers to the total number of document sets divided by the number of documents containing the word, then logarithm.

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}+1|} \tag{2}$$

Among them, $|D|$ represents total number of documents, $|\{j:t_i \in d_j\}|$ represents the number of documents containing the word t_i. Plus 1 is to prevent the denominator from being 0. Finally, by multiplying the word frequency with the frequency of the reverse file, we can get the TF-IDF value of a word, such as formula (3). The greater the TF-IDF value of a word, the more it can show the theme of the word to its document.

$$TF - IDT = tf_{ij} * idf_i \tag{3}$$

3.4 Weighted representation

Weighted representation is mainly divided into two steps, feature word weighting and text representation.

Feature word weighting. Use formula (3) to calculate TF-IDF and assign it to each characteristic word. In order to reduce the impact of document length on weight calculation, we use normalized TF-IDF value as the weight of feature words.

Text representation uses Vector Space Model (VSM) [2]. The main idea of the model is that the N document in vector space can be expressed as a matrix, and an element in the matrix corresponds to the weight of a word item in a document. Each row is regarded as a vector representation of a document, and each column is regarded as a vector representation of a word item. As shown in Figure 2, D_n represents n documents, T_l represents he first l characteristic words, d_{nl} represents the weight of the l word in the n document.

$$\begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_n \end{array} \begin{array}{c} \begin{array}{cccc} T_1 & T_2 & \cdots & T_l \end{array} \\ \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1l} \\ d_{21} & d_{22} & \cdots & d_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nl} \end{bmatrix} \end{array}$$

Fig. 2 document set vector space model

3.5 Feature dimension reduction

Through these steps, the dimension of text collection is always very large, and there is a lot of redundancy. In order to reduce the time cost and memory cost of the subsequent computation, we need to reduce the dimensionality of the feature without reducing the accuracy. The mainstream dimensionality reduction methods include latent semantic indexing, non negative matrix factorization, principal component analysis, feature clustering and so on. These algorithms have their own advantages and disadvantages, and there is no commonly accepted best algorithm. The results are different for different tasks. Here we use Principal Component Analysis (PCA) to reduce dimensionality of high-dimensional texts. This method transforms a given set of related variables into unrelated variables by linear transformation, and sorts the variance according to variance. By selecting some variables, the cumulative variance is close to the sum of the variance of the original variable, so as to achieve the dimensionality reduction effect.

3.6 Classification algorithm selection

With the development of artificial intelligence and machine learning, there are a lot of classification algorithms in text categorization. In this paper, four algorithms, SVM, Decision Tree (DT), NB and BPNN, are used as the basic learning devices in the ensemble learning framework, and their classification accuracy is compared respectively, so as to build an optimal classifier.

SVM is a kind of classifier based on kernel technology to find out the intervals between different categories of samples in the feature space and make the intervals largest. DT is a tree structured classification and regression algorithm. The algorithm can be considered as a set of if-then rules or a conditional probability distribution in feature space and class space. Generally, a tree contains a root node, several internal nodes and several leaf nodes. The leaf nodes represent the decision results, and the internal nodes represent the attribute tests[3]. NB is a classification method based on Bayes theorem and assuming characteristic conditions are independent. For training data sets, we first learn the joint probability distribution based on the independent assumptions of feature conditions. For input, we use the Bias theorem to find the posterior probability maximum output according to the joint probability distribution. BPNN is an artificial neural network algorithm based on the error propagation model. It is an adaptive and self-organizing algorithm. Its basic idea is to use gradient descent method to minimize the mean square error of the actual output value and expected output value of the network[4].

The above classification algorithm has a good classification effect on the standard dataset. How to further improve the accuracy of classification and apply it to Chinese bibliographic classification is the focus of this paper.

Ensemble learning is an algorithm to complete learning tasks by constructing and combining multiple learning devices. It first produces a set of "individual learning devices" and combines them with some strategies. These individual learning devices, also known as basic learning devices, can be any classification algorithm, and the classification algorithm can be homogenous or heterogeneous. Unlike the standard AdaBoost algorithm, the Bagging algorithm can be applied to multiple classification and regression tasks. Because there are 22 broad categories of Chinese map classification and 22 English letters, it belongs to multi classification tasks, so it is more suitable to apply Bagging algorithm as an integrated learning framework. The above four algorithms are merged into the Bagging framework, which are referred to as BagSVM, BagDT, BagNB and BagBPNN respectively.

This paper compares the accuracy and accuracy variance of various algorithms in Chinese Bibliographic data automatic classification task to evaluate the classification effect and stability of the algorithm. The higher the accuracy, the better the classification effect; the lower the accuracy variance, the more stable the classification algorithm is.

4. CHINESE BIBLIOGRAPHIC AUTOMATIC CLASSIFICATION EXPERIMENT

4.1 Bibliographic data matrix construction

Segmenting each data in the processed data set by using the staging segmentation system. The specific process is to use the Jieba module provided by the Anaconda integration platform to segment the word, and get the data set consisting of word sets, that is, a number of words are made up of one data. Using the disabling word list containing 1717 stop words, we remove the stop words from each word set in the preceding operation, and reduce the impact of meaningless words on the theme of each data. After removing the stop word step, get the preprocessed word set.

The TF-IDF value of each word in each data is calculated by using the formula (3) in the 3.3 section. In this paper, a reasonable threshold is selected through many experiments to help extract the feature words, and set the thresholds of the range and interval 0.1 respectively, and calculate the number of feature words under the corresponding threshold. The number of feature words decreases with the increase of threshold. We know that the high number of feature words will lead to the long running time and high memory cost of the subsequent classification algorithm. The small number of feature words will lead to a large number of missing data information, resulting in poor classification accuracy. After many experiments, we choose 1428 feature words with TF-IDF threshold of 1 as the initial feature space.

Based on the feature space obtained from the above steps, we use vector space model to represent the bibliographic data. For every dimension in the feature space, that is, every word, if the word appears in one data, the TF-IDF value corresponding to the word in this data is assigned to this dimension. If the word does not appear in the data, the value is assigned to 0. Finally, the text representation of this data is obtained, that is, a vector is obtained. Repeat this step for each data, and finally get a 11418*1428 high dimensional bibliographic data matrix.

Observation shows that the matrix is a sparse matrix, that is, the elements in each row are mainly 0, and a few elements are nonzero values, which are still too high in spatial dimension, so dimension reduction is needed. In this paper, we use PCA to reduce dimension, and usually set its cumulative

variance to 80%. We can get 399 variables and form a 11418*399 low dimensional bibliographic data matrix. In this way, most of the information of variables is retained, and the redundancy of data is reduced. So far, Chinese Bibliographic data has been represented as a 399 dimensional matrix, and the following is the selection of classification algorithm.

4.2 Algorithm parameter selection

This paper mainly explores the classification effect of various classification algorithms on Chinese Bibliographic data under the framework of ensemble learning Bagging. Four algorithms, SVM, DT, NB and BPNN, are used as basic learning devices of Bagging respectively. Different algorithms require certain parameter selection or configuration.

The most important thing in SVM is to select the kernel function. The commonly used kernel functions include linear kernel, polynomial kernel, Gauss kernel, Laplasse kernel and Sigmoid kernel, and the kernel function is obtained through the combination of the above kernel functions. For text data, linear kernel is often used, so we use linear kernel as kernel function.

The commonly used versions in DT are ID3, C4.5 and CART (Classification and Regression Tree). Their main difference is that they use information gain, information gain rate and Gini index respectively as criteria for partitioning attributes. The three decision trees have their advantages and disadvantages. In this paper, CART decision tree is used as the basic learning device in ensemble learning.

NB is mainly divided into polynomial naive Bias, Gauss plain Bias and Bernoulli plain Bias. For discrete events encountered in general document classification, polynomial distribution and Bernoulli distribution are more suitable. The attribute values in this article are all continuous values, so Gauss naive Bayes are more suitable for .

BPNN takes shallow neural network as a model, because the bibliographic data matrix has 399 dimensions, so the input layer sets 399 neurons. The number of neurons in the hidden layer has not been universally recognized by theory. The commonly used method is to set 1/2 as the number of neurons in the input layer, so it is set to 200. The output layer corresponds to the number of classes, and the number of neurons is 5. Activation functions in neural networks are mainly Sigmoid functions, tanh functions, ReLU functions. In this paper, we use the Sigmoid function as the activation function. Randomly selected 10000 of the data as training set, 1418 data as a test set, in order to avoid contingency, improve the credibility of the experimental results, calculate the average accuracy rate after 10 experiments.

4.3 Experimental results and analysis

First, we compare the average classification accuracy of four traditional classifiers in the 10 experiments with or without Bagging algorithm.

From Figure 7, we can see that without adding the Bagging framework, the classification effect of NB is the worst, only 51.78%, and the classification effect of BPNN and SVM is better, about 70%. This is because NB assumes that each characteristic word in the sample is independent, which is obviously inconsistent with the facts, so its classification effect is not good. Then we observe the accuracy rate of each algorithm when adding the Bagging framework. Among them, the accuracy rate of SVM classification has almost no enhancement effect, and even slightly decreased; the classification accuracy of NB and BPNN has slightly increased; DT's promotion effect under the Bagging

framework is the best, reaching 8.38% of the original accuracy rate, but it still does not exceed the BagBPNN algorithm.

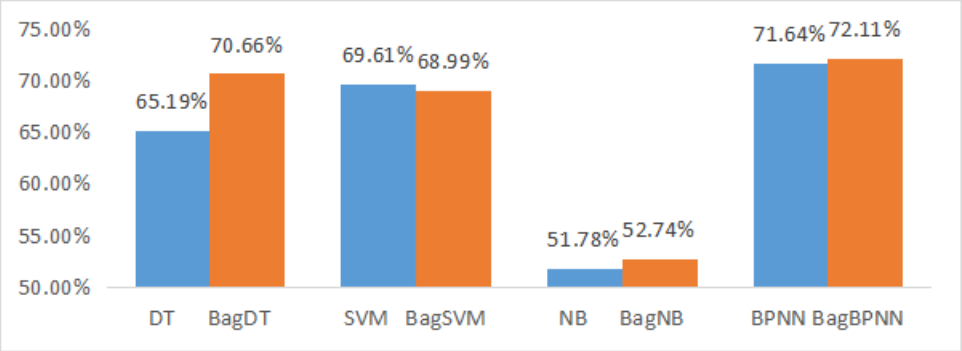

Fig. 7 classification accuracy of different basic learning devices with or without Bagging algorithm framework

Next, the learning process of four different basic learning devices is constructed from 5 to 40, and the bag sampling set with 5 intervals is trained. Each classifier is trained. The voting mechanism is used to test the test set. If there is the same number of votes, one of the results is randomly selected and the accuracy is calculated. Finally, the accuracy of these classifiers on the test set is shown in Table 1.

Table 1 classification accuracy of different basic learning devices under different sampling sets

Base learner accuracy / Sampling sets	BagDT	BagSVM	BagNB	BagBPNN
5bag	68.8223%	68.9986%	52.1157%	71.8900%
10bag	69.8731%	69.9577%	52.1298%	71.5444%
15bag	70.4020%	69.5346%	51.7560%	71.3752%
20bag	70.6559%	68.9915%	52.7433%	72.1086%
25bag	71.1707%	69.6827%	51.6855%	72.3343%
30bag	71.4386%	69.2102%	52.0733%	72.2285%
35bag	70.9661%	69.6474%	51.2200%	72.4894%
40bag	71.2412%	69.7391%	51.5303%	72.4753%

We can see that the classification accuracy of DT basically converges after 25 sets of samples, and reaches a peak value of 71.4386% when the 30 sets are sampled, and it is more sensitive to the number of sampling sets. The classification accuracy of SVM and NB is basically not affected by the number of sampling sets, which is stable at 69% and 52% respectively. The classification accuracy of BPNN converges around 30 sets of samples and reaches a peak value of 72.4894% when the 35 samples are collected. It is the best classifier in all classifiers and is most suitable for Chinese Bibliographic automatic classification system.

5. CONCLUSION

In this paper, the Bagging algorithm in ensemble learning is used as the integration framework of different kinds of algorithms. A Chinese Bibliographic automatic classification system is constructed to classify Chinese Bibliographic data. Experiments show that the Chinese Bibliographic automatic

classification system can classify Chinese bibliographies. The accuracy of classification can reach 72%, and the stability is high. It can improve the efficiency of Chinese bibliographic classification and save labor cost to some extent. The Bagging algorithm framework has certain effect in improving decision tree, naive Bayes and BP neural network, and improves the decision tree. The effect is the most obvious. In the four algorithms, the stability of BP neural network and decision tree algorithm is most benefit from the Bagging algorithm framework. Whether there is or without Bagging algorithm framework, BP neural network algorithm can better adapt to Chinese Bibliographic data, and the accuracy of classification is the highest.

The shortcomings and future work of this paper are: how to improve the impact of insufficient data on the accuracy of classification; how to further improve the method of feature extraction; and to classify the Chinese Bibliographic data only coarsely, we need to make a more detailed classification. These questions will be discussed in the following contents.

REFERENCES

[1] Tu S, Huang M. Mining microblog user interests based on TextRank with TF-IDF factor[J]. Journal of China Universities of Posts & Telecommunications, 2016, 23(5):40-46.
[2] Salton G, Wong A, Yang C S. A vector space model for automatic indexing[J]. Communications of the Acm, 1974, 18(11):613-620.
[3] Quinlan J R. Induction on decision tree[J]. Machine Learning, 1986, 1(1):81-106.
[4] HechtNielsen. Theory of the backpropagation neural network[J]. Neural Networks, 1988, 1(1):445-445.
[5] Schapire R E, Freund Y, Bartlett P, et al. Boosting the margin: a new explanation for the effectiveness of voting methods[J]. The Annals of Statistics, 1998, 26(5):1651-1686.
[6] Schapire R E. A Brief Introduction to Boosting[C]// Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc. 1999:1401-1406.
[7] Freund Y. A detection-theoretic generalization of on-line learning and application to boosting[J]. Journal of Computer & System Sciences, 2005, 13(5):663-671.
[8] Schapire R E. The Boosting Approach to Machine Learning: An Overview[M]// Nonlinear Estimation and Classification. Springer New York, 2003:149-171.
[9] Breiman L. Bagging predictors[M]. Kluwer Academic Publishers, 1996.