Chinese News Text Classification Based on Machine learning algorithm

Fang Miao, Pu Zhang, Libiao Jin*, Hongda Wu
School of information engineering
Communication University of China
Beijing, 100024, China
E-mail :{ mfcuc06, zhangpu, libiao, hongda.wu} @cuc.edu.cn

***Abstract*—Text classification is the key technology for mining and organizing text information, which is the process of determining the text types automatically according to the content. Based on machine learning algorithm, text classification system includes four processes, namely text pretreatment, text representation, classifier training and classification. In this paper, a Chinese news text classification system model is designed. And in the classifier training part, we separately chose and compared K-nearest Neighbor, Naive Bayes, and Support Vector Machine as our classification algorithm. Then, we tested and analyzed these classifiers with each other and finally got a conclusion. The experimental conclusion shows that the Chinese news text classification system can get satisfied results based on the machine learning algorithm.**

Keywords-Chinese text classification; Chinese word segment; Feature extraction; Machine learning algorithm

I. INTRODUCTION

At present, Internet technology is rapidly developing and society is constantly moving towards informatization. With the explosive growth of information, people can read a lot of news texts from the Internet especially mobile phone every day. The news spreads widely, and has a great influence on the society, so it becomes particularly important to analyze and process news text. The classification of news text is a key technology to process news information, which can help organize information effectively and distinguish information categories according to the needs of users quickly. It is to label on the news text for realizing the ordering of news. By tagging news texts, each news item can be classified so that readers can choose what they are interested in. In addition, news websites can through each reader's reading records recommend the same or relevant type of news to attract more attention.

Text classification is a method, which is used to confirm the category of an unlabeled text based on the defining topic categories in advance. In mathematics, it is actually a mapping:

$$f : A \rightarrow B \qquad (1)$$

where A is the set of texts to be classified, B is the set of categories, and f is the classifier in the classification process. Text categorization can be divided into binary classification and multi-class classification. Binary classification is a method to determine the problems of Yes or No, Positive or Negative, and it is usually used in sentiment analysis. Multi-class classification problems can be divided into: single-label multi-class and multi-label classification [1]. The text classification in the news field is multi-classification, so a multi-classifier should be designed.

At present, the research about English text classification has been mature, and many classification system models have been proposed. In the 1950s, H.P.Luhn began the study of automatic classification on English texts. He proposed to apply the word frequency statistics to the text classification for the first time and made an innovative progress [2]. Maren published the first paper of automatic text classification algorithm in 1960, which promoted the development of the whole field [3]. Then, K.Spark et al. had done a lot of effective work to make the text classification technology more mature.

However, the study of Chinese text classification started late, so the methods are few, and the research is not deep enough. In the domestic, Hou Hanqing in 1981 deeply discussed the application of text classification work at first. At present, many mature algorithms have been applied in Chinese text categorization, such as K-nearest neighbor classification algorithm [4], Support Vector Machine algorithm [5], Bayesian classification algorithm [6], decision tree [7], and neural network classification algorithm [8], etc.

Chinese text classification system is mainly divided into classification method based on knowledge engineering and machine learning. The method based on knowledge engineering is classifying texts by manual according to some rules defined. This method takes a lot of time and is very inefficient. Nowadays, the most common used classification method is based on machine learning, using machine learning algorithms to train the classifier, and getting a satisfactory result.

In this paper, we introduce Chinese news text classification system flow design and discuss machine learning process. The rest of the paper is organized as follows. In the Section II, we present the process of system model and describe the each process of it. Section III shows some experiments and results analysis. Finally, Section IV concludes this paper.

978-1-5386-5836-9/18/$31.00 ©2018 IEEE
DOI 10.1109/IHMSC.2018.10117

II. THE SYSTEM MODEL

This paper introduces the principle of the text classification and designs system model based on machine learning algorithm. This model trains the classifier according to the existing data, and then classifies the unlabeled news texts by it. As shown in Fig.1, the model includes the following processes: text pretreatment, text representation, classifier training and classification.

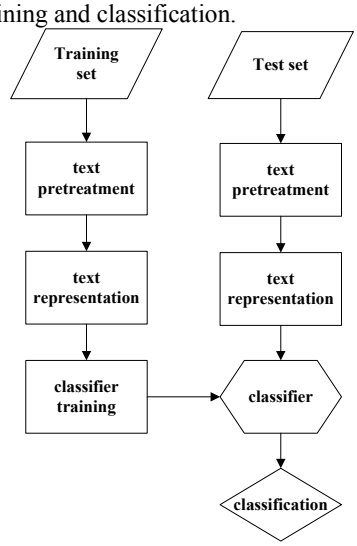

Figure 1. Text classification model

A. Text pretreatment

Text data mainly comes from the Internet and it can be crawled down by the web spiders or other web page grab tools. So far, there have been many universities or natural language processing facilities that have their own Chinese corpus and the most popular Chinese corpus include: Fudan University news text classification corpus, Sogou laboratory news corpus, Chinese Academy of Sciences Chinese and English news corpus, etc.

After obtaining the data, the data should be pretreated, and the first step is to remove the stopwords and various punctuation marks for rough dimension reduction. In the face of the huge amount of data, removing the words which have no effect on the classification result will make the text content clean, reduce the data space, and make the classification more efficient. Chinese text and English text are different. In English text, words and words are separated by spaces and easy to be extracted. In Chinese, words are formed in combination, and there haven't any inherent space, so it must to do the process of word segmentation. Chinese word segmentation is one of the foundations of information processing, and the traditional method of word segmentation is mechanical Chinese word segmentation, mainly including maximum matching method, reverse maximum matching method and full segmentation [9]. This way of word segmentation is based on a dictionary. It matches the words in the dictionary according to certain rules. If a word is found in the dictionary, then cut out this word, and then take the remaining sentences to match until the whole sentence is divided.

With the maturity of Chinese word segmentation technology, many open source segmentation tools are developed, such as SnowNLP, Jieba segmentation, THULAC, NLPIR and ICTCLAS. The dictionary of each word segmentation tools differs, so the effect of the word segmentation is different.

B. Text representation

Text representation is to translate Chinese text into binary language which can be recognized by the computer. Currently, the commonly used text representation models include Boolean model, Vector Space model (VSM) [10] and Statistical language model. The most commonly used is the Vector Space model, which is proposed by G.Salton. It is a vector space formed by a set of orthogonal vectors, and each text feature is transformed into a dimension in this space. That is to say, using a vector $(t1: w1, t2: w2 \ldots tn: wn)$ to represent a text, where *ti* stands for text feature and *wi* stands for the weight of feature. Vector Space model has good generalization ability. At present, different feature extraction methods and quantization means can be adopted to represent text. In this paper, TF-IDF algorithm is used for text vectorization.

TF-IDF technique mainly eliminates the most common words and extracts only most relevant feature words from the corpus [11]. The main idea of the TF-IDF algorithm is that if a word has the higher frequency in one text and appears in a smaller range in the certain corpus, then the word has a stronger ability to distinguish the category of texts, and it should be given a higher weight in the VSM. The calculation formula is as follows:

$$\mathrm{Wt} = \mathrm{TF}*\mathrm{IDF} = \frac{m}{N} * \log\left(\frac{N}{n} + 0.01\right) \tag{2}$$

TF is the frequency of the feature word *t* in the text *D*, reflecting the internal feature of a document. The *m* represents the number of this feature word appearing in the text, and *M* represents the total number of feature words in the text. IDF is the inverse frequency of document, and *N* is the total number of all documents in corpora, *n* is the number of documents containing the feature word *t*. IDF reflects the distribution of the feature word in the whole corpora, and reflects the feature between the documents.

This algorithm is simple, effective and easy to understand, but it is just according to the frequency of the word in a text to value its weight, did not consider the position of the words and sentence context. FastText is a fast text classifier developed by Facebook, which uses N-gram model to generate word vector according to the context of text. It not only considers the frequency of word in the document, but also considers the location and relationship between words. So, this tool gets better result in classification.

C. Training model and Classification

Text classification mainly uses classifier to label the text of unknown category, so the most important part of classification is the selection of classification algorithm. At present, the most commonly used algorithm is machine learning, which lets the computer to train the classifier according to the training set. And for a new unclassified document, the computer will make a judgment based on the previous experience, learning the rule of classification, and then give a correct category.

For a dataset, whether a classification algorithm is appropriate for this data requires data metrics to evaluate the results, such as precision (P), recall (R), and F-measure (F).

Precision represents how many of the positive samples classified are actually positive samples. It is based on our prediction results. There are two possibilities for positive prediction. One is to predict the positive class as a positive class, and the other is to predict the negative class as a positive class. On the other hand, recall indicates how many positive examples in the sample are correctly predicted. There are also two possibilities. One is to predict the positive class as a positive class, and the other is to predict the positive class as a negative class. The precision and recall are defined as follows:

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{3}$$

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{4}$$

where *TP*, *FP* and *FN* are the number of true positives, false positives and false negatives, respectively. In this paper, precision can be denoted the proportion of the document number that the classifier correctly judges them to be the category and the document number in this category. And recall is the proportion of the document number that the classifier correctly judges them to be the category and the document number in the sample. In general, the precision and recall rate are contradictory, so the results can be evaluated by the F-measure, which is the weighted average of precision and recall rate of the model, and its calculation formula is as follow:

$$\mathrm{F} = \frac{2 * P * R}{P + R} \tag{5}$$

where *P* is precision and *R* is recall rate.

In this paper, K-nearest neighbor algorithm, Naive Bayesian algorithm and Support Vector Machine algorithm (SVM) are used to train classifier according to a specific data set and then evaluate three algorithms through P, R and F.

The idea of the K-nearest neighbor algorithm is: for the new input instance, we calculate the Euclidean distance between the test point and all the reference points in order to find nearest neighbors, and then rank the obtained distances in ascending order and take the reference points corresponding to the smallest Euclidean distances.

The Naive Bayesian algorithm is a very simple classification algorithm based on Bayes theory. Bayes theory is by calculating the frequency of occurrence of something in the past to estimate the future probability of its occurrence. Its results indicate that the probability distribution for the random variable and can also be interpreted as the possibility of different level of trust. The Bayesian formula is:

$$\mathrm{P}(A \mid B) = \frac{P(AB)}{P(B)} = \frac{P(B \mid A) * P(A)}{P(B)} \tag{6}$$

That is the probability of event *A* occurring on the premise that event *B* has already occurred. In general, we usually use another form:

$$\mathrm{P}(B \mid A) = \frac{P(AB)}{P(A)} = \frac{P(A \mid B) * P(B)}{P(A)} \tag{7}$$

In the Naive Bayesian classification algorithm, if the unclassified feature set is represented by *A*, the classification result set is represented by *B*, then:

$$P(category \mid feature) = \frac{P(feature \mid category) * P(category)}{P(feature)} \tag{8}$$

Thus, we can find out the category that makes $P(category \mid feature)$ the biggest, this category is the classification result of the text.

SVM is a potential classification technology proposed by Vapnik et al, which is a supervised learning model with associated learning algorithms that analyze data used for classification and regression analysis [12]. Its main idea is: for a multi-dimensional sample set, each sample is indicated as a point in space. And the system then randomly generates a hyper plane which moves continuously and classifies the samples until the points belonging to the same category distribute on the same side of the hyper plane exactly. There are many hyper planes that satisfy this condition, and what we need is to find such a plane to maximize the blank area between sides of it in order to achieve the optimal classification of these samples. For a new data to be classified, we map it to the same space and predict the category based on its location.

III. EXPERIMENTS AND RESULTS ANALYSIS

This system used on the Pycharm platform and design in python language. Chinese word segmentation tool is Jieba segmentation. We adopt Fudan University news corpus, and the data in corpus is divided into training set and test set. Each set has nine categories, and the corresponding news content is included in each category. This system uses the K-neighbor classifier and SVM classifier and Naive Bayesian classifier in package named Sklearn to classify. The result of classification is as follows:

TABLE I. THE RESULT OF CLASSIFICATION

Category	Number of the Test set	Classifier in KNN	Classifier in Naive Bayes	Classifier in SVM
Agriculture	1022	1076	981	1062
Economy	1601	1635	1682	1595
History	468	356	442	418
Environment	1218	1193	1149	1199
Politics	1026	1067	1076	1038
Sports	1254	1209	1230	1238
Art	742	804	803	782
Computer	1358	1386	1354	1365
Space	642	605	614	634

Through the classifier, the tests were stored into the corresponding categories. As is shown in Table Ⅰ, the Chinese news text classification system gets satisfied results . And then we can figure out the precision value, recall and F-value of classifier respectively. The results are as follows:

TABLE II. PERFORMANCE EVALUATION

	Running time	P	R	F
K-nearest neighbor	19.60s	0.920	0.920	0.919
Naive Bayesian algorithm	3.42s	0.921	0.920	0.920
SVM	484.75s	0.957	0.957	0.957

From experiment data shown in Table Ⅱ, SVM algorithm's precision value, recall and F-value are highest. However, this algorithm takes more time than the other two algorithms, because it iterates more times. Therefore, the SVM algorithm is only fit for the small amount of data.

K-nearest neighbor and Naive Bayesian algorithm have similar results in P, R and F. K-nearest neighbor classification method is mainly based on the limited nearby samples, so it is more suitable for the sample set whose categories overlap more. The Naive Bayesian method algorithm takes less time, and it is easy to implement. However, the Naive Bayesian classification model needs to meet the requirement that the conditional independence assumption between features. But even so, it can still get better classification result in many areas.

Overall, the news text classification system got satisfactory results and could achieve expected experiment purposes.

IV. CONCLUSION

This paper is constructing a Chinese news text classification system model based on Machine learning algorithm. This paper expounds the K-nearest neighbor, Naive Bayesian algorithm and SVM algorithm, and describes every aspect of the system model in detail, giving the evaluation metrics. The results show that support vector classifier with the tf-idf feature attain the highest accuracy and is the most stable in small dataset. However, there are many problems with the Chinese text classifier like that the small class produces better results than big class, and that semantic information of text features is insufficient. Thus, how to make the classification more effective is the next major content of the study.

REFERENCES

[1] X. Wang *et al*., "Research and Implementation of a Multi-label Learning Algorithm for Chinese Text Classification," *2017 3rd International Conference on Big Data Computing and Communications (BIGCOM)*, Chengdu, 2017, pp. 68-76.

[2] H. P. Luhn, "A Business Intelligence System," in *IBM Journal of Research and Development*, vol. 2, no. 4, pp. 314-319, Oct. 1958.

[3] M.E. Maron and J.L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *Journal of the ACM*, vol. 7, pp. 216–244, 1960.

[4] Q. Xu and Z. Liu, "Automatic Chinese Text Classification Based on NSVMDT-KNN," *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, Shandong, 2008, pp. 410-414.

[5] S. Wei, J. Guo, Z. Yu, P. Chen and Y. Xian, "The instructional design of Chinese text classification based on SVM," *2013 25th Chinese Control and Decision Conference (CCDC)*, Guiyang, 2013, pp. 5114-5117.

[6] Z. Gong and T. Yu, "Chinese Web Text Classification System Model Based on Naive Bayes," *2010 International Conference on E-Product E-Service and E-Entertainment*, Henan, 2010, pp. 1-4.

[7] D. E. Johnson, F. J. Oles, T. Zhang and T. Goetz, "A decision-tree-based symbolic rule induction system for text categorization," in *IBM Systems Journal*, vol. 41, no. 3, pp. 428-437, 2002.

[8] H. Zhuang, C. Wang, C. Li, Q. Wang and X. Zhou, "Natural Language Processing Service Based on Stroke-Level Convolutional Networks for Chinese Text Classification," *2017 IEEE International Conference on Web Services (ICWS)*, Honolulu, HI, 2017, pp. 404-411.

[9] H. Gong, C. Zhou, "Chinese word segmentation system research", *Journal of Beijing Institute of Machinery*, vol. 3, 2004.

[10] J. Geng, Y. Lu, W. Chen and Z. Qin, "An Improved Text Categorization Algorithm Based on VSM," *2014 IEEE 17th International Conference on Computational Science and Engineering*, Chengdu, 2014, pp. 1701-1706..

[11] P. Bafna, D. Pramod and A. Vaidya, "Document clustering: TF-IDF approach," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, 2016, pp. 61-66.

[12] B. Xu, S. Chen, H. Zhang and T. Wu, "Incremental k-NN SVM method in intrusion detection," *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2017, pp. 712-717.