Advances in Intelligent Systems Research, volume 147

International Conference on Network, Communication, Computer Engineering (NCCE 2018)

Comparative Research on Automatic Classification Algorithms Based on Chinese Medical Literature

Kai An, Yunqiu Zhang [a)], Xiaoyang Wang, Zhe Jiang, Chenglong Wang and Xiang Zhu

School of Public Health, Jilin University, Changchun 130000, China.

[a)] yunqiu@jlu.edu.cn

Abstract. With the development of electronic periodicals, it is unavoidable that there are some classification management problems. But currently the classification management of papers basically majors in manual classification. Based on Chinese medical literature, this essay compares and analyzes these automatic classification algorithms: support vector machine (SVM), BP neural network, and random forest. It is found that SVM is more suitable for automatic classification of Chinese medical literature.

Key words: Chinese Medical Literature, Automatic Classification, SVM, BP Neural Network, Random Forest.

INTRODUCTION

From the point of view of the process of text classification, research on the text classification has been started since the 1960s of the last century. But from the 1960s until the end of the 1980s, the text classification system was always the classification system based on Knowledge Engineering which is constructed by experts [1]. Its typical application was the Construe system developed by Carnegie Group for Reuters [2]. It mainly guided classification through some classification rules written by professionals. At that time, it was very effective and accurate on some of the Reuters corpus. The average accuracy and recall rate can reach approximately 90%. However, building such an effective classification system requires many experienced experts, and it takes a lot of manpower and time. Also, it is difficult to update the system and apply it to different fields of application [3]. In the early 1990s, the classification technology based on machine learning began to replace the method based on knowledge engineering and gradually became the mainstream technology of text classification. The idea of machine learning was applied to the field of automatic classification of text. Classifiers are built automatically by summarizing the features of the text set. The learning and classification process comes from the machine's autonomous learning of the training texts. As a result, there is no need for the support of domain experts and no manual intervention is required, so that the classification efficiency can be improved and a large amount of manpower and material resources can be saved [4, 5].

From the point of view of research on automatic text classification of machine learning in foreign countries, over the years, researchers have proposed a variety of classification models and classification algorithms for machine learning techniques. Since Vapink et al. of the AT& Bell Laboratory proposed Support Vector Machines(SVM) machine learning algorithms in the 1990s, with the maturity of natural language processing techniques and methods, various classification algorithms have been widely applied to the research of automatic classification and have achieved Certain results [6]. Based on the Support Vector Machines, M. Corney classified the email by the author's gender [7]. In addition, such as the Rocchio classification algorithm based on vector space model and its series of improved algorithms, K-nearest neighbor algorithm, Decision Tree, Naive Bayes, Neural Network and so on, these methods have been widely used in the classification of English and European texts and have achieved good results.

However, it is still not so mature in the automatic classification of Chinese texts. Chinese has many different features from English, making it more difficult to classify Chinese texts. For example, the written form of Chinese is written continuously, and there is no natural boundary between words and words [8]. Before classifying text, it is first necessary to segment the text [9]. In addition, syntactic analysis and semantic analysis account for different proportions in the research of different languages. In English, syntactic analysis is more important than semantic analysis. Semantic analysis of Chinese plays a decisive role in Chinese research, and its proportion is much larger than syntactic analysis. This makes it more difficult to grasp the content of the text by grammar-based means such as syntactic analysis in Chinese text classification. In China, automatic text classification technology cannot copy foreign research results. Therefore, it is necessary to research and develop a practical Chinese text automatic classification system [10].

In the study of Chinese text classification algorithms, the research on various algorithms and the construction of classification systems have gradually matured. The domestic research on the automatic classification of Chinese texts has generally gone through three aspects of feasibility analysis, auxiliary classification system and automatic classification system, and has achieved a lot of results in theoretical research and practical application. The Chinese text classification model is basically a vector space model. Xiujuan Liang uses a vector space model to represent texts based on SVM, uses a combination of mutual information and word frequency to extract features from texts, and uses feature vectors to represent them. Two parallel classifiers were trained, and 500 Cross-disciplinary and marginal disciplines were used as test texts for verification [11]. Lin Zhai and Yajun Liu analyzed the characteristics of SVM and conducted experiments on the classification corpus provided by Fudan University with 400 to 2700 feature items respectively and achieved satisfactory results [12]. In the comparison of the classification performance of Bayesian algorithm and SVM, after the word segmentation and elimination of stop words, the information gain was used for feature selection, and then the feature weights were calculated. From these two experiments, the recall rate, the accuracy ratio and the f-value were compared, and it is found that the Naive Bayes algorithm is easily affected by the distribution of document categories, its recall rate and precision rate vary greatly from category to category. The reason for the analysis is the calculation of Prior Probabilities in Naive Bayes algorithm depended on the number of category documents. The calculation of the prior probability is dependent on the number of category documents. If the number of documents in this category is large, there is a large prior probability. The experimental results show that the accuracy of classification of text by SVM is higher than that of naive Bayes. Some domestic scholars have optimized the SVM. For example, Tinghui Zhang proposed an improved model, introduced the idea of high-frequency words, used the Apriori algorithm to find co-occurrence sets of the largest frequent words, and used the most frequent words and keywords as a text feature. According to the characteristics of Chinese texts, another group of scholars proposed several new automatic classification models: radial neural network model, fuzzy rough sugar set model, latent semantic classification model, binary tree based multi-class support vector machine classification algorithm, and based on this, these has been improved [13]. SVM discriminates the final category of text according to the best aspects, has a high classification accuracy, and is Suitable for occasions where the text classification effect is high. It has obvious advantages in classifying Chinese texts [14]. Another good classification algorithm is the Decision Tree. The biggest advantage of the Decision Tree is its clear structure and easy understanding. Decision classification process is very clear. From the root node to the tree leaf node, the entire classification process is intuitive and facilitates tracking. This point not only has some benefits for debugging code and scalability, but more importantly, this method is easy to understand, and it is of great value for carrying out collaborative research and for specific applications [15]. The classification method has excellent data analysis efficiency, noise robustness, and ability to learn antisense expression. And the Decision Tree also has a unique advantage: Easy to extract intuitive and easy to understand classification rules make the decision tree more suitable for text classification [16]. In addition, the domestic research on feature selection and weight calculation has been relatively stable. Among them, the commonly used feature selection methods mainly include word frequency method, information gain, mutual information, and chi-square fitting test method. The commonly used features of the weight calculation methods are Boolean method, word frequency method, method and weighted method.

In the research on the classification of various literatures based on the Chinese Library Classification, most of the previous studies were based on the collective classification of related literatures of various disciplines. Due to the increasing number of new medical disciplines, there are many infiltration and cross-cutting relationships between disciplines. The classification has a larger dimension and involves more extensive aspects, which makes the classification more difficult [17]. The effect of medical classification with other disciplines is not satisfactory. Among the few classifications of medical literature, the literature we can find is related to the selection of support vector machine algorithms for the nine categories under medical journal R7 using mutual information, chi-square statistics, cross entropy, and evidence weights as feature selection methods for automatic classification [18].Until

2010, there were literatures for automatic classification of medical literature, which are about the research on Automatic Classification Using Machine Learning Method Under Chinese Classification System. Therefore, for the relatively vacant research in the medical field, our research becomes increasingly valuable.

RESEARCH ON AUTOMATIC CLASSIFICATION ALGORITHM

The Research Contents Mainly Include

To Investigate the Main Methods of Automatic Literature Classification, Which Mainly Based on Machine Learning Methods, Including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naive Bayes (NB), Decision Trees (Dts), Neural Networks (Nns) And Other New Methods. The Research Range Covers the Principle, The Scope of Application, The Relative Merits, Etc.

To Investigate the Status at Home and Abroad, focusing on the Research on Automatic Literature Classification Based on Chinese Library Classification (CLC).

To Combine the Features of The Database Platform to Investigate the Methods of Automatic Literature Classification Used by Major Database Platforms at Home and Abroad.

Comparison of Algorithms

Data Preprocessing

The data of journal articles of Wanfang Medical were selected as experimental samples. The samples were preprocessed, including word segmentation and data cleaning, and further divided into training samples and test samples. The ratio was 9:1.

Text Representation

Vector Space Model (VSM) was used to represent the samples to transform into a processible form.

Feature Selection

Keywords, titles and abstracts was used as features.

Construction of Classifiers

Classifiers were built separately according to automatic literature classification algorithm.

Test Evaluation

The evaluation indicators were used to compare the results of the classification algorithm, which including the precision and the time-consuming.

The Key Issues to Be Solved

The selection of automatic classification methods and the construction of automatic classifiers.

Features and Innovations

There are a huge number of medical literatures referred to complicated classification number. However, there were few researches on automatic classification of medical literatures. This research compares several automatic literat

ure classification algorithms to select the Perfect algorithm for automatic classification of medical literatures, which provides reference for automatic classification in practice, and improves the efficiency of automatic literature classif ication greatly.

RESEARCH PROGRAM IMPLEMENTATION

Text Preprocessing

Extract the Valid Attribute Column

The three fields of title, keyword, and abstract in the data can all reveal the category of the literature. Some researchers showed that the keyword played a more significant role. Therefore, this study used keywords as the object of feature selection and extracted three attribute columns that the number, keywords, and classification codes required for the research.

Select Valid Data

The data used in this study was the medical literature data provided by wan fang, which contained the following five attributes: number, title, abstract, keyword, and classification codes divided by Chinese Library Classification. As the basis for classification, the large variety of leaf nodes in the taxonomy would increase the complexity of classification calculation, so the partition of the same level should be taken into account when categorizing. Synthesizing each kind of situation, the literature data of R5 category was selected to compare the classification in this article. Pick in R5's collection, using sorting function of Excel to delete data containing more than one classification codes or more than six keywords to form a data format with each number corresponding to up to 5 keywords and a category number. At this point, the data can be scrambled using the RAND () function of Excel to draw data at random for comparison of the classification under different sample sizes. Finally, the data was composed of three columns of “number-keyword-classification codes” by copy-paste and deleted null data by positioning nulls of Excel.

Word Segmentation

Columns containing multiple keywords were separated into columns containing only one keyword by the columnar function of Excel.

Deal with Dirty Data

The punctuation marks were deleted by replacement function of Excel and processed the classification codes, which converted the classification codes to "1, 2, 3, 4, 5, 6, 7, 8, 9, 0" to refer to ten sub-categories of R5. Finally, according to data reorganization function of SPSS, the repetitive keywords of the same that article were deleted. The data was shown in Figure 1:

ID	keyword	class
1	titration	8
1	CCB	8
1	diabetic nephropathy	8
1	valsartan	8
1	ACEI	8
2	ARB	1

FIGURE 1. Text data

Representation of Text

Since the keywords were textual information which are linearly inseparable. BP Neural Network, Random Forest, and Support Vector Machine algorithms can only handle data in the form of matrices. Therefore, the key issue was how to turn the text data into identifiable feature matrices which was Vector Space Model.

This study used Excel to convert text data into feature matrices. Based on its pivot table, the matrix combined with positioning null function or programming "number-keyword". That was, the leftmost column was a complete and non-repetitive number (this column can be deleted), the second column on the left was the classification of the corresponding numbers added, and the top row was all keywords that were not repeated. The value in the matrix was a lot of "0" s and a small amount of "1" s, and finally converted into a CSV file, which can be easily read in MATLAB. The final matrix was shown in Figure 2:

class	OCT	125OH2D3	13C-Urea Breath	16S ribosome	Type 1 Diabetes	24h ECG
1	0	0	0	1	0	0
7	0	0	0	0	0	0
4	0	0	0	0	0	0
4	0	0	0	0	0	0
4	0	0	0	0	0	0
7	0	0	0	0	0	0
7	0	0	0	0	0	0
1	0	0	0	0	0	0
8	0	0	0	0	0	0
7	0	0	0	0	0	0
4	0	0	0	0	0	0
4	0	0	0	0	0	0
4	0	0	0	0	0	0
4	0	0	0	0	0	0
7	0	0	0	0	0	0
8	0	0	0	0	0	0
9	0	0	0	0	0	0
4	0	0	0	0	0	0

FIGURE 2. Vector Matrix

In order to compare the performance of the algorithm under three different data sizes and achieve comparative analysis of the algorithm, this study formed three sparse matrices of 900*2015, 2430*3621, and 3995*5237, which can not only make data more clear and concise, but also allow the data to be basically ready for processing to enter the following modeling phase.

Training of Three Classifiers

It was found that BP Neural Network, Random Forest, and Support Vector Machine (SVM) were more representative in previous research. Three sparse matrices were trained by Mat lab, and three algorithm models were under constructed.

The Construction of BP Neural Network Classifier Model

After 900 * 2015 data were processed, the training sets and test sets generated at random. Next BP Neural Network was created, trained and simulation tested. The prediction type and the actual type of classification map was shown in Figure 3:

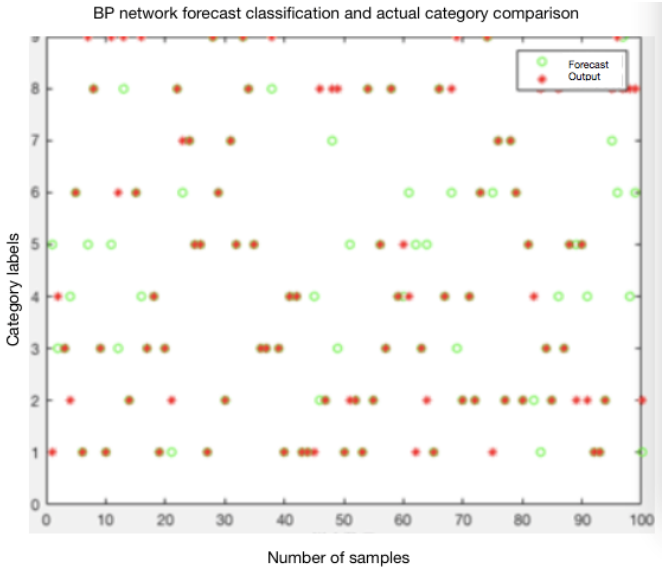

FIGURE 3. BP Network Classification Comparison

It can be seen that the test accuracy rate was 0.66, which took 13min37s. As for 2430 * 3621 data, the accuracy rate was 0.7333, which took 1h12min. And 3995 * 5237 data, the accuracy rate was 0.8027, which took 2h45min.

Construction of a Random Forest Classifier Model

After 900 * 2015 data were processed, the training sets and test sets generated at random. According to training data and testing data, Random Forest classifier was generated. After simulation testing and accuracy testing, the classification result was shown as Figure 4:

	1	2	3	4	5	6	7	8	9
1	53	0	50	23	128	29	36	8	173
2	0	0	3	20	41	309	1	5	121
3	0	95	2	82	75	7	4	14	221
4	0	0	4	38	100	11	6	17	324
5	1	0	34	122	135	13	3	31	161
6	0	0	4	35	91	8	5	111	246
7	497	0	2	0	0	1	0	0	0
8	0	0	3	7	19	2	415	2	52
9	0	0	4	57	108	11	6	18	296
10	14	1	37	103	160	7	41	9	128
11	0	0	26	6	10	2	400	1	55
12	0	0	15	45	107	11	6	18	298
13	0	0	4	45	111	11	6	18	305
14	4	0	1	111	103	63	14	21	183
15	60	0	2	50	96	34	12	12	234
16	17	482	0	0	0	0	0	0	1
17	0	0	4	44	109	11	6	18	308
18	0	0	0	1	0	496	0	1	2
19	0	0	4	45	111	11	6	18	305
20	0	0	4	45	111	11	6	18	305
21	0	365	1	17	30	1	0	3	83
22	0	0	4	45	111	11	6	18	305
23	0	0	3	34	172	9	6	17	259
24	0	0	49	15	35	2	284	5	110

FIGURE 4. RF Classification Results

It can be seen the accuracy rate was 0.6100, which took 4min36s. As for 2430 * 3621 data, the accuracy rate was 0.7015, which took 1h10min. And 3995 * 5237 data, the accuracy rate was 0.7900, which took 3h23min.

The Construction of Support Vector Machine Classifier Model

After 900 * 2015 data were processed, the training sets and test sets generated at random. According to training data and testing data, the data normalized. Then looking for the best c / g parameters and creating/training the SVM model. After simulation testing and accuracy testing, the classification result was shown as Figure 5:

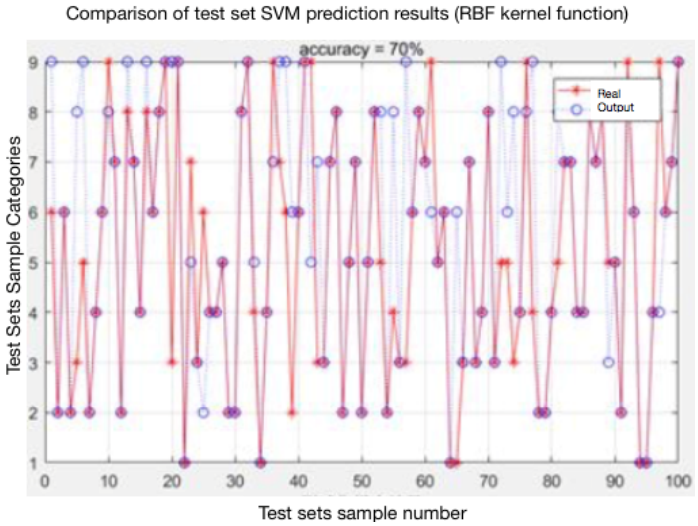

FIGURE 5. SVM Prediction Results

It can be seen the accuracy rate was 0.7000, which took 2min10s. As for 2430 * 3621 data, the accuracy rate was 0.7520, which took 1h30min. And 3995 * 5237 data, the accuracy rate was 0.8200, which took 3h30min.

COMPARISON AND ANALYSIS OF THREE CLASSIFIER ALGORITHMS

This study compares and analyzes the algorithms of the three classifiers in terms of accuracy and time-consuming, and comprehensively judges their performance:

Accuracy

As shown in Figure 6, as the amount of data increases, the accuracy of the three algorithms gradually increases. The overall precision of the SVM is at the top, followed by the BP neural network, and the random forest is the last.

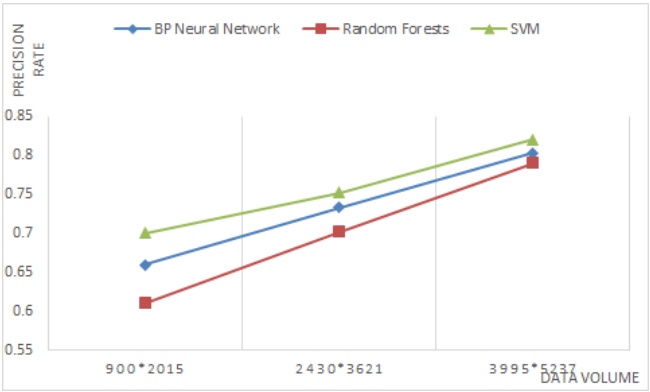

FIGURE 6. Comparison of accuracy

Time Consuming

As shown in Figure 7, as the amount of data increases, the time consumption of the three classification algorithms gradually increases. The BP neural network consumes a relatively short time for a large amount of data. The support vector machine is more conducive to smaller data, and the random forest is centered.

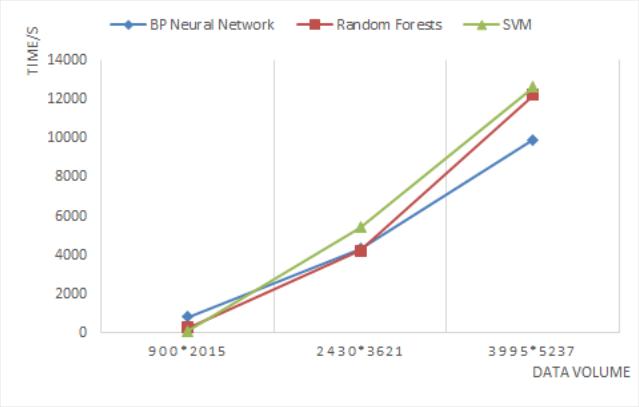

FIGURE 7. Time Consuming Comparison

CONCLUSION

According to the comparison and analysis of the three classifier algorithms, it is not difficult to see that the three automatic classification algorithms have their advantages and disadvantages: SVM precision rate is in the first place, but it is more conducive to the processing of small data; BP Neural Network precision rate is the second, which is more conducive to the processing of big data; the precision of Random Forests is poor, and the processing efficiency of data is medium. As a whole, SVM is more conducive to automatic classification of Chinese medical literatures. Next is BP Neural Network. Random Forests is the worst.

ACKNOWLEDGEMENTS

Project 2017025 Supported by Graduate Innovation of Jilin University.

REFERENCES

1. Huang Hua. Research on classification of scientific literature based on decision tree and SVM fusion learning [D]. Henan University of Technology, 2011.
2. Vidin Kumar. Text Categorization Using Weight Adjusted K-Nearest Neighbor elssifieation[C]. In Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and data mining. 2001.
3. Zhang Zenoah. Research and implementation of key technologies for automatic classification of Chinese texts. Zhejiang University of Technology [D], 2013.
4. Rung Gang. Research on Chinese Text Classification Method [D]. Shandong Normal University. 2009.
5. Ye Peng. Research on Automatic Classification of Chinese Journal Papers Based on Machine Learning [D]. Nanjing University, 2013.
6. VAPNIK V. The nature of inductive learning theory [M]. New York: Springer—Verlag, 2000: 89.
7. CORNEY M. Gender-preferential text mining of e-mail discourse [EB / OL]. Http: / / www. Assay. Org /2002 / papers.
8. Wang Hai, Ye Peng, Deng Swansong. The application of machine learning in automatic classification research of Chinese journal articles. Modern Library Information Technology [J], 2014, 30 (3):80-87.
9. Ma Janna, Tina Dagan. Research on Chinese Text Automatic Classification Based on SVM [J]. Computer and Modernization, 2006, 138 (8): 5.

10. Chinese Library Classification [EB/OL]. [2010-07-20]. Http: // www.ztflh.com/.
11. Liang Xiujuan. Research on multi-category text classification based on SVM [D]. Wuhan: Hungnam University of Economics and Law, 2008.
12. Yu Lin, Liu Cajun. Research on Chinese Text Classification Based on Support Vector Machines [J]. Computer and Digital Engineering, 2005, 33 (3): 21-23.
13. Hu Yan, Xing Haiyang, Fu Xinyang. Research and Improvement of SVM Algorithm for Linear Separable Text [J]. Computer and Digital Engineering, 2008, 36 (3): 18-20.
14. Ma Jian in, Li Jin, Tang Guava, Wang Fang, Zhao Yang. Comparative Study of KNN and SVM Algorithms in Chinese Text Classification Technology [D]. 2009:54-57.
15. Wang Qian. Application of Decision Tree in Text Classification [J]. Science & Technology Information Development and Economy, 2007, 17(17): 197-198.
16. Yang Cubing, Zhang Jun. Decision tree algorithm and its core technology [J]. Computer Technology and Development, 2007, 17(1):43-45.
17. Xu Na. A tentative study on the influence of Chinese Library Classification on medical classification [J]. Modern Information. 2006.10:26-27.
18. Gao Yuan, Wang Baoding. Design and Implementation of Library Document Classification System Based on "Chinese Library Classification"[J]. Science & Technology Information, 2009 (28): 120-121.