Research on Document Content Classification on Mathematical Regression Model

Hua Long[1, a], Baoan Li[1, b*]

[1]Computer School, Beijing Information Science and Technology University, Beijing, China

[a]longh2015@163.com, [b]liba2010@139.com

*Corresponding author

Keywords: Document Classification; SVM (Support Vector Machine); CHI (Chi-square Statistic); Mathematical Regression Model

Abstract. To improve the document classification problem, this study proposes a classification algorithm based on mathematical regression model, making Chinese document classification get rid of the dependence on traditional dictionary method. The method of extracting high frequency keywords, establishes the appropriate matrix model, making a high-dimensional document change into a low-dimensional document, and then use mathematical regression model to give a comprehensive feature weighting function by corpus training. It explored an approach to avoid the traditional method of the problem of curse of dimensionality.

Introduction

With the rapid increase of document information, document classification becomes the key technology of a large number of document data. In the field of document classification, news classification and search engine in corpus processing and information retrieval and so on are widely used.

The current document classification technology is basically based on words or word series information, because words or word strings and the concepts of the class of documents are closely related. This makes document classification need the help of dictionary words and using special extraction technology. Word extraction in Western processing is called taking stem, while in Chinese language processing as cutting word. On Chinese document classification, the cutting word is a very complex task. The Chinese word segmentation systems are generally more complex and large, cutting the word slowly, and the accuracy is not high. There are many common classification methods with Support Vector Machine (SVM), such as K Nearest Neighbor (KNN) and Bayesian (Bayes theorem) etc [1]. The traditional method requires the support of document classification dictionary, and some need a lot of tedious matrix operations, making the time of classification system in the classification take more, and the effect of classification is general [2]. If the documents can be classified without Chinese dictionary support, document classification system has a more extensive applicability and lasting vitality [3].

Based on the above consideration, this study puts forward a new method for the classification of the document content: Math linear regression model. When classifying documents, the model does not need the support of the dictionary, and make the high-dimensional document dimension reduction into a low-dimensional document, and the classification effect and the running speed has a very big enhancement. So the document classification system is undoubtedly a more broadly applicable.

Document Classification of SVM

In many years ago, people whose research based on the statistical learning theory of linear classifier proposes a best design practices. Speaking from the principle can be divided into linear, then extended to the case of linear inseparable. Even it is extended to the use of non-linear function, so the classification is known as SVM (Support Vector Machine).

SVM classification method implementation can follow the speech filter to select only certain parts of speech as classification features. Such as, nouns and verbs are chosen as the classification feature words. The accuracy of text classification feature words with the number of classification continues to increase, and it is generally to 2000 words.

CHI measures the dependencies between the term and the class. If the term and class are independent of each other, then the value is close to zero.

Methods of feature selection of SVM have CHI (Chi-square Statistic) and IG (Information Gain), such as CHI of feature selection method is described below.

Table 1. Statistical Variable Definition Table of CHI

	In the class	Not in the class	sum
Contain the word	a	b	a+b
Not contain the word	c	d	c+d
sum	a+c	b+d	a+b+c+d=n

Among them:

a: The number of co-occurrence of the term and class.

b: Does not belong to class, there are number of term appears.

c: Without term, the number belongs to the class.

d: Without term, nor the number of class.

n: The total number of documents.

Therefore, a word term of CHI statistical formula was showed as Eq. 1.

$$\mathrm{chi}_{\mathrm{statistics}(\mathrm{term},\mathrm{class})} = \frac{n\times(ad-bc)^2}{(a+c)\times(b+d)\times(a+b)\times(c+d)} \qquad (1)$$

Document Classification on Mathematical Regression Model

Mathematical regression model is a good assessment of the weight of the text feature [4]. It is starting from a set of sample data, to determine the mathematical relationship in the variables, and the credibility of those relationships is done by various statistical tests to find out the impact of a particular variable from the many variables affecting significant and no significant variables [5]. Using of those, the method accords to the value of one or a few variables to predict or control the other particular variable values. Mathematical regression model can make use of the mathematical relation between the input and the output, to establish a corresponding matrix model, and finally get a relationship model of input and output. In this study, the text characteristic of artificial summary is an independent input variable, and the keyword weight is an output variable, to get a comprehensive feature weighting function, to get a set of optimal scaling factors through the weighted function to training corpus. The concept represented by a matrix of mathematical regression model was showed as Eq. 2.

$$\begin{pmatrix} W(S_1) \\ \vdots \\ W(S_i) \\ \vdots \\ W(S_m) \end{pmatrix} = \begin{pmatrix} W_F(s_1) & W_L(S_1) & W_C(S_1) & W_I(S_1) & W_T(S_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ W_F(S_i) & W_L(S_i) & W_C(S_i) & W_I(S_i) & W_T(S_i) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ W_F(S_m) & W_L(S_m) & W_C(S_m) & W_I(S_m) & W_T(S_m) \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \\ \varepsilon \end{pmatrix} \qquad (2)$$

The W(si) is the output vector, representing the comprehensive weight of key si, (WF(si),WL(si),WC(si),WI(si),WT(si)) is the input vector, representing feature vector of keyword si; $\alpha, \beta, \gamma, \delta, \varepsilon$ are scaling factors. In conclusion, the comprehensive keyword weighting function as Eq. 3.

$$W(S_i) = \alpha W_F(S_i) + \beta W_L(S_i) + \gamma W_C(S_i) + \delta W_I(S_i) + \varepsilon W_T(S_i) \qquad (3)$$

The W (si) represents the comprehensive feature weights of the keywords si.

Performance Test

Test document library

Currently, there are no accepted standards of Chinese document classification test libraries, so the study use its own document libraries to test document classification system. Test document library documents is from People's Daily and Dispatch of Xinhua News Agency corpus, which are divided into forty categories, and we take the number of documents which contain up to 9 categories of the test. The number of documents and their class names of these classes included in Table 2 below.

Table 2. Test Document Library

Class	Polity	Sport	Economy	Environment	Art	Education	Medicine	Traffic
Documents No.	619	366	276	162	150	138	173	116
Total	2000							

Evaluation index

The study adopts a general evaluation of document classification system to evaluate the classification performance. In experiment, we take 70% of the documents for training documents, and 30% for test documents. The study uses precision (p), recall(r) and micro-average F1 to evaluate document classification system. These are defined as Eq. 4 - Eq. 6.

$$\text{precision} = \frac{\text{N}}{\text{D}_1} \times 100\% \tag{4}$$

$$\text{recall} = \frac{\text{N}}{\text{D}_2} \times 100\% \tag{5}$$

$$\text{F}_1 = \frac{2\times\text{precisio n}\times\text{recall}}{\text{precision}+\text{recall}} \times 100\% \tag{6}$$

Among them:
N: The correct number of documents divided into some kind of documents
D_1: The total number of documents divided into the class of test set
D_2: The total number of documents belonging to the category of test set

Test results

Table 3. Comparison of Two Algorithms Classification Performance

Classification	Precision	recall	F_1
CHI	91.3	84.2	87.6
Mathematics Regression Model	95.5	88.6	91.9

According to Table 3, the precision, recall and micro-average of the linear regression model are generally higher than CHI. Because the linear regression model in the training focuses on high-dimensional documents, changing to use a low-dimensional categorize documents, and sets up a matrix model of high-frequency keywords, to get a comprehensive feature weighting function to avoid the SVM training a large number of matrix operations which is very time consuming.

Seen from the results in Fig. 1, with the two algorithms comparing the document bibliography on running time, the running time of mathematical linear regression models is lower than CHI method, and with the growth in number of the document is relatively flat, thereby text shows the proposed classification algorithm has better expansibility when dealing with large data sets. The reason is that the linear regression method effectively reduces the feature space dimension of documents, to avoid the high-dimensional document caused by the problem of *curse of dimensionality*. In addition, the mathematical regression model does not use dictionary classification and avoids a lot of calculations, increasing the operating speed of classification.

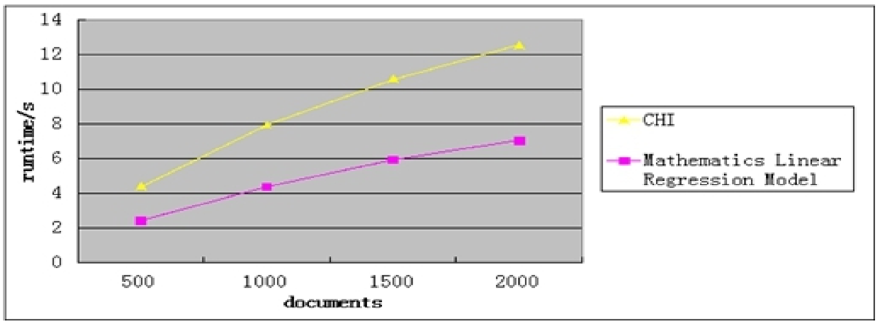

Fig. 1. Comparison of Two Algorithms

Conclusion

Mathematical regression model does not use a traditional method, and effectively solves to the problem of the *curse of dimensionality*. So the experiments show that the method can improve the accuracy and speed of document classification. In future, the study will use mathematical linear regression model of high-frequency keywords and other grammatical and semantic analysis technology [6], to explore multi-level classification technology [7], to extract keywords more efficiently, to improve the quality of document classification.

Acknowledgement

The work was supported by Project of Construction of Innovative Teams and Teacher Career Development for Universities and Colleges under Beijing Municipality (Grant No. IDHT20130519), and Project of the Specialty Construction and Comprehensive Reform under Beijing Municipality (Grant No. 71M1510818).

References

[1] Wang, Ziqiang, Web document classification algorithm based on IB and LapSVM, Journal of Central South University (Science and Technology), v 42, n SUPPL. 1, pp.731-736, September 2011 Language: Chinese (2011)

[2] Bekkerman, Ron, Gavish, Matan, High-precision phrase-based document classification on a modern scale, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.231-239 (2011)

[3] Wang, Weihong, Research on methodology of document classification based on statistical learning theory, 2005 International Conference on Services Systems and Services Management, Proceedings of ICSSSM'05, v 2, pp.1107-1109 (2005)

[4] Král, Pavel, Lenc, Ladislav, Confidence measure for czech document classification, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v 9042, pp.525-534 (2015)

[5] Sun, Jiantao, Shen, Dou; Lu, Yuchang; Shi, Chunyi, Web document classification techniques, Journal of Tsinghua University, 44(1), pp.65-68, January 2004 Language: Chinese (2004)

[6] Sung, Li-Chun, Kuo, Chin-Hwa; Chen, Meng Chang; Sun, Yeali, Progressive analysis scheme for web document classification, Proceedings - 2005 IEEE/WIC/ACM InternationalConference on Web Intelligence, WI 2005, v 2005, pp.606-609 (2005)

[7] Calvo, Rafael A., Ceccatto, H.A. Intelligent document classification, Intelligent Data Analysis, 4(5), pp.411-420 (2000)