IOP Conference Series: Earth and Environmental Science

PAPER • **OPEN ACCESS**

Chinese Text Classification System on Regulatory Information Based on SVM

To cite this article: Mingyue Zhang *et al* 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **252** 022133

View the article online for updates and enhancements.

This content was downloaded from IP address 120.234.63.196 on 29/09/2019 at 09:34

Chinese Text Classification System on Regulatory Information Based on SVM

Mingyue Zhang, Xinbo Ai* and Yanzhu Hu

School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China

*Corresponding author e-mail: axb@bupt.edu.cn

Abstract. The rapid development of computer technology has brought a large amount of text information. This paper aims to classify the text in view of the increasing number of Chinese texts in the process of supervision and the increasing demand of processing Chinese texts, and to improve the efficiency of information query and management.In this paper, a Chinese text brake classification system based on support vector machine is designed and implemented. The text is represented mathematically by vector space model, and the classifier is trained to classify the text based on the principle of support vector machine.In this paper, the classification performance of the text system is tested and evaluated by using the file of regulatory information. The experimental results show that the classification accuracy of the text system is relatively high and has certain practicability.The system in this paper can improve the efficiency of information inquiry and management and make it possible to manage massive text information.

1. Introduction

In the current society, there are different types of enterprises and different types of hidden dangers. All kinds of potential accidents and safety risks are intertwined and superimposed. Besides, the supervision system and legal system are imperfect [1]. Supervision department and supervisors also differ in many ways. There is a large amount of information can be obtained in the supervision of production safety, among which, a lot of relevant information could be found, which can be used to analyze the differential factors of regulatory power in the process of safety supervision[2]. It is of great significance to analyze the current situation of safety supervision in the city, explore the related factors of the effectiveness of safety supervision, strengthen the key points of supervision and rationally allocate the safety supervision force [3].

The key step of Chinese text classification is text classification. Text classification is a basic subject in the field of Chinese information processing and the key to intelligent Chinese information processing [4].There are three main methods of text classification: K-Nearest Neighbors algorithm, Naïve Bayes algorithm and Support Vector Machine [5].

In order to achieve the classification effect of Chinese text information in the respect of the ability to supervise enterprises, we first need to match the word segmentation for Chinese with continuous state, and mean the independent word terms, and then use the traditional vsm-based machine learning algorithm to train classifiers. In order to obtain the maximum classification accuracy, the bidirectional matching save and same disambiguation segmentation algorithm with higher word segmentation accuracy is used as the Chinese word segmentation processing of the system in this paper, and then the

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

SVM with higher comprehensive performance in the current machine learning algorithm is used as the classifier learning algorithm.

Therefore, text classification system mainly includes two parts: 1.The text classifier is generated by training text set. In this process, the text classification system facilitates the generation of structured data by necessary preprocessing of the given text. At the same time, feature selection should be carried out to represent the text in the form of feature vector, and then SVM algorithm should be executed to generate classifier.2. Use the classifier generated by SVM to classify unknown text. In this process, the system also needs to preprocess the text into structured data and map it to feature vector space, so that the classifier can be used to classify the classified text.

2. Experimental Design

2.1. System Architecture Design

The system in this paper is composed of two main modules: text preprocessing and SVM classification. At the same time, it can be divided into four sub-modules: text segmentation, feature representation, SVM training and SVM testing. The overall framework of the system is shown in Figure 1.

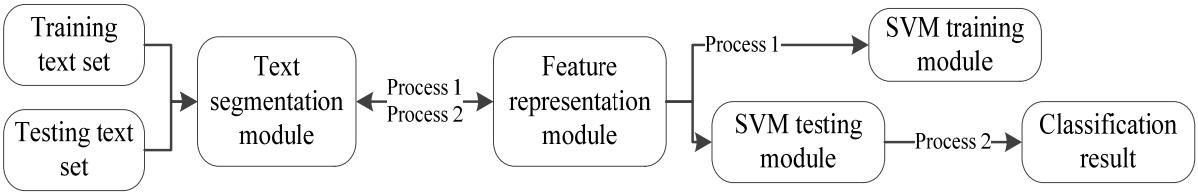


Figure 1. Overall framework of SVM Chinese text classification system

2.2. Text Segmentation Module

Chinese word segmentation needs to solve problems such as new word recognition and ambiguity recognition. The most common Chinese word segmentation algorithm are the Maximum Matching Method (MM) and the Reveres Maximum Matching Method (RMM).In order to get better effect of word segmentation, this paper adopts bidirectional matching, common storage and disambiguation algorithm.

The algorithm is described as follows: when scanning a string, RMM and MM algorithm are used successively to compare the segmentation results of the two sequences, then keep the same segmentation results, and disambiguate different segmentation results. The algorithm process is shown in Figure 2.

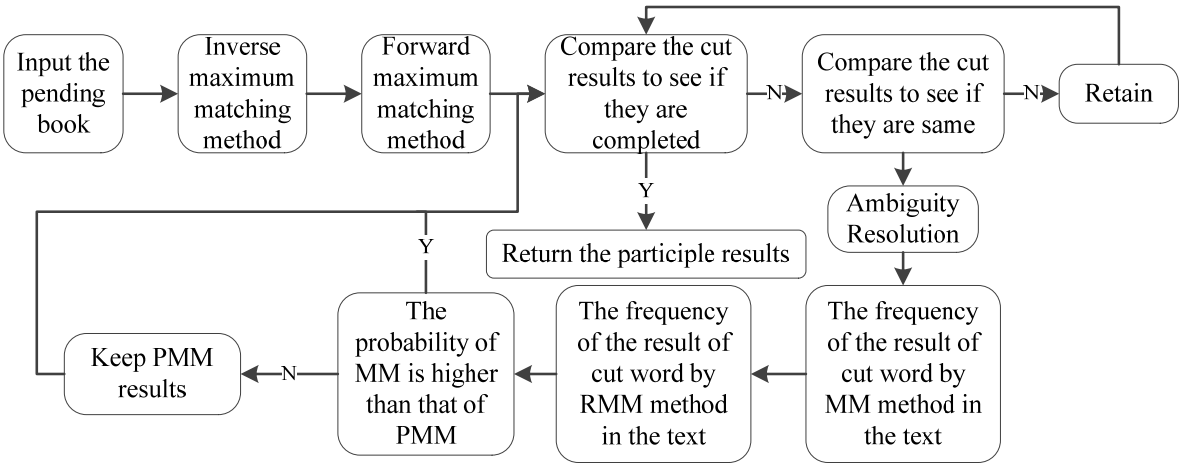


Figure 2. Flow chart of bidirectional matching and disambiguation algorithm

2.3. Feature Representation Module

The function of the feature representation module is to extract the features of the text information of the supervisory ability after the Chinese word segmentation processing, and conduct the mathematical mapping of the text based on the vector space model, so that the text classifier can be trained based on SVM principle.

The basic flow of the feature representation module are as follows.

1. Connect the text word segmentation module and read the Chinese text of regulatory information after word segmentation. The larger text capacity of the text will be stored on the hard disk, at which time reading data will need to start IO operation. The system provides different types of file read and write operations.

2. Unified processing of text content. Chinese is the main expression language in the regulatory information text, but at the same time, the possibility of traditional Chinese characters and English and other languages is high, so we need unification on heterogeneous text processing, such as , the transformation of traditional Chinese characters, the unification of English case as well as the unity of plurality etc. The transformation of Chinese traditional case in English as well as the unity of singular and plural, and so on, in order to facilitate the text representation of the system later.

3. Perform mathematical mapping for the normalized regulatory information text. Based on the principle of vector space model, the supervised information text is represented mathematically and stored.

2.4. Text Categorization

The strict definition of text classification is: the computer constructs classifier according to text features and some classification rules, and divides the text to be classified into pre-defined known categories. Figure 3 and 4 respectively describe the training and classification process of text classification.

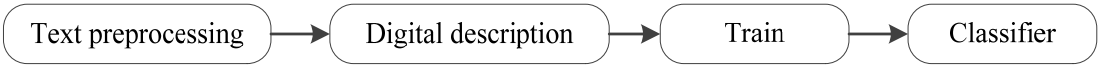


Figure 3. Text classification training process

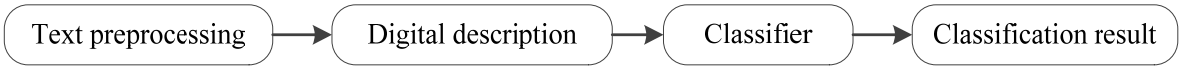


Figure 4. Text classification sorting process

The third step is to find out the pattern of hidden dangers in different types of enterprises. According to the classification in a single enterprise, the number of hidden dangers is converted into the proportion corresponding. Synthesize the hidden dangers of each enterprise under the single enterprise classification, and finally use the Apriori algorithm to obtain the association rules between them.

After the above three steps, an enterprise can be classified. After the classification, other potential risks can be predicted through the detected hidden dangers to provide data support for the hidden danger investigation work.

2.4.1. Linear Separable Problems. SVM was initially applied to the biclassification problem, that is, there is known training set (x_i, y_i) , $x \in R^d$, $y \in \{-1,+1\}$, where x_i is the training sample, and y_i is the category label of sample x_i. We illustrate the basic idea of SVM with the two-dimensional model as an example. As shown in Figure 5, the solid sphere and the fork symbol respectively represent two types of samples, and there are multiple groups of different classification hyperplanes (in the case of two-dimensional hyperplanes are straight lines), which can separate the two types of samples with zero error. We define the distance between the classification hyperplane and the nearest sample in each class as the geometric interval of the linear classifier. According to the statistical learning theory, the classification

hyperplane with maximum geometric spacing has the best classification generalization ability, that is, the optimal classification hyperplane.

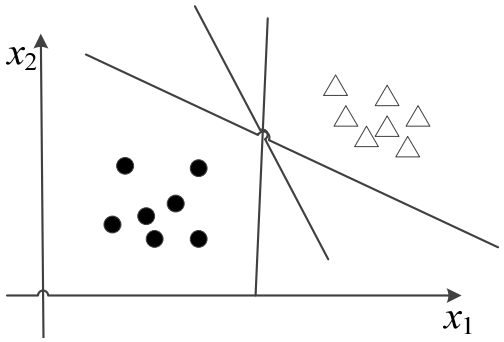


Figure 5. Classification hyperplane

It is noted that the classification hyperplane can be obtained by multiplying by any non-zero coefficient, so it can be normalized to make it satisfy equation 1.

$$yi(< wi, xi > +b) \geq 1, i = 1, 2, \cdots, \quad l \tag{1}$$

Where <*a*, *b*> represents the inner product operation between a and b, and l represents the training set capacity. As shown in Figure 5, the geometric interval after normalization is equal to 0.5ω. Maximizing the geometric spacing is equivalent to minimizing ω, or minimizing 0.5ω.

At this point, the SVM classification problem is transformed into the optimal problem of the variables a and b of equation 2.

$$\text{g(x)}= \min \frac{1}{2}\omega^2 \tag{2}$$

The optimal solution ω* and b* can be obtained by solving equation 3, and then the optimal classification hyperplane $< \omega^*, x > +b^* = 0$ and the decision function $f(x) = \text{sgn}(< \omega^*, x > +b^*)$ can be obtained.

Transform equation 3 into Wolfe dual problem and define Lagrange function.

$$L(\omega, \alpha, b) = \frac{1}{2}\omega^2 - \sum_{i=1}^{l} \alpha_i y_i (< wi, xi > +b) + \sum_{i=1}^{l} \alpha_i, \alpha_i \geq 0, i = 1, 2, \cdots, \quad l \tag{3}$$

where α_i is Lagrange multiplier, partial differential is taken to L and set it equal to 0， then the original problem can be transformed into its dual problem (equation 4).

$$\text{g(x)}= \min \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j < x_i, x_j > - \sum_{j=1}^{l} \alpha_j \tag{4}$$

The optimal solution should satisfy the Kuhn-Tucker condition, namely the complementary relaxation factor of equation 5:

$$y_i(< \omega, x_i > +b - 1) = 0 \quad \text{i=1, 2, … ,l} \tag{5}$$

Finally, the optimal solution ω^* and b^* can be expressed as.

$$\left\{ \omega^* = \sum_{i=1}^{l} \alpha_i y_i x_i \quad b^* = y_i - \sum_{i=1}^{l} y_i \alpha_i < x_i, x_j > \right\} \tag{6}$$

2.4.2. Linear Indivisible Problems. If there is no classification hyperplane that can separate the training sample set with error of 0, it is called linear indivisibility problem. At this point, we hope to get the classification hyperplane that minimizes the classification error. With the introduction of slack variable ξ_i and penalty factor C, the original problem changed into a new optimization problem as shown in equation 7.

$$\text{g(x)} = \min \frac{1}{2} \omega^2 + C \sum_{i=1}^{l} \xi_i \tag{7}$$

Among them, the penalty factor C is a predefined constant, which is used to indicate the tolerance degree of the classifier to the error. Lagrange dual problem transformation is performed for the optimization problem of equation 8, and the results are as follows:

$$\text{g(x)} = \min \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle - \sum_{j=1}^{l} \alpha_j \tag{8}$$

2.4.3. Kernel Method. Due to the complexity of the real world, there are a lot of problems that cannot be expressed in a linear way. As the expression ability of linear function is limited, we can increase the expression ability of linear function by mapping original data to higher-dimensional space, but there is no corresponding theoretical guidance and systematic method for specific mapping methods. For the SVM model, as shown in equation 9, only the inner product operation of the sample exists, so we only need to get the inner product after the vector is mapped to the higher-dimensional space, without considering the specific mapping. We will be able to receive lower dimensional space vectors and calculate the mapping in higher dimensional space inner product value of the function called kernel function, that is $K(x, x')$. At this point, the original problem can be changed into.

$$\text{g(x)} = \min \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K \langle x_i, x_j \rangle - \sum_{j=1}^{l} \alpha_j \tag{9}$$

2.5. Multiple Classification Strategy

The SVM model considers the problem of binary classification, while in the Chinese text classification of regulatory information, it is basically a problem of multiple classification. In order to apply the SVM to the multi-classification problem of the regulatory nformation system in this paper, we need to adopt certain strategies to train the multi-classifier. Common methods include one to the rest, one to one and DAG strategies.

2.5.1. One to the Rest Strategy. One to the rest still adopt the idea of biclassification. Suppose there are No different text categories. First, i is selected as a positive sample, and the rest of the samples are considered as negative samples. This classifier can determine whether the input samples belong to i class. The classification process is shown in Figure 6. If a classifier determines that the input sample belongs to this class, the classification results are obtained.

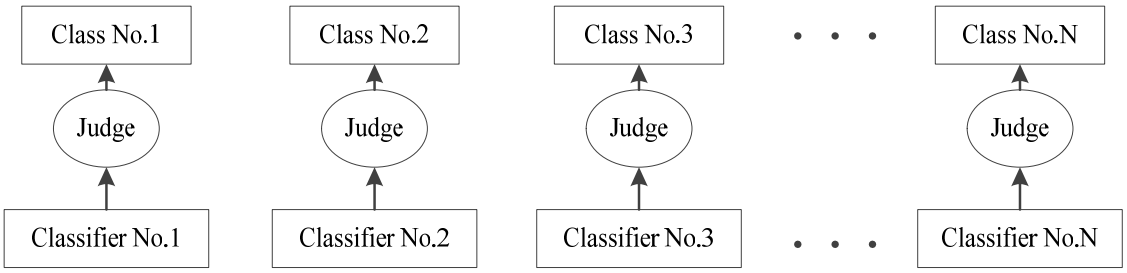


Figure 6. One to the rest Classification strategies

2.5.2. One to One Strategy. Suppose there are No different text categories, and one-to-one multi-classification is based on the following strategies: As shown in figure 7, two classifiers are trained. Each classifier is responsible for determining whether the object belongs to class i or class j (where i and j are arbitrary category labels), and the results of the classifier are counted. If the output of a classifier is i, then the number of votes in class i is increased by 1. In the end, the result of classification is considered as the largest number of votes. The one-to-one strategy classifier only involves related categories in training, while the training speed of a single classifier is relatively fast, and there is no leakage in the classification process. However, with the increase of the number of categories, the number of classifiers increased too fast, which greatly extended their training and classification speed.

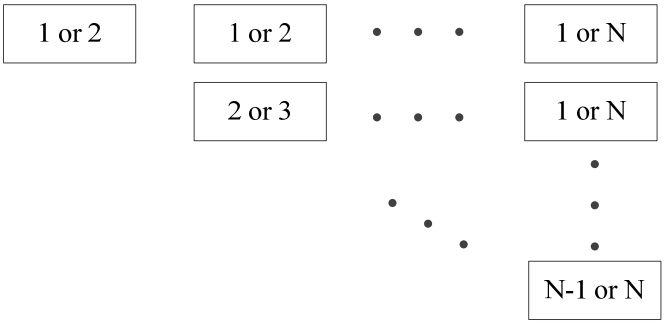


Figure 7. One-to-one policy classifier

2.5.3. DAG Strategy. The DAG multi-classification strategy adopts the same training method as the one-to-one strategy. Who Classify according to the method shown in Figure 8 $N*(N-1)/2$ classifiers constitute a directed acyclic graph, and each classifier is a node in the graph. Each sample to be classified passes through a node to get an output category. The direction of the graph is determined according to the output category until the traversal reaches every leaf node.

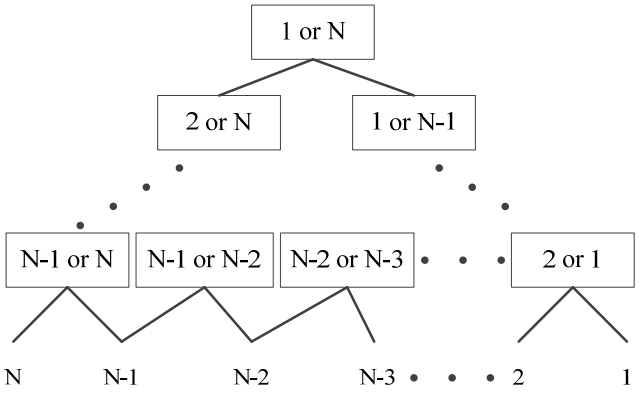


Figure 8. DAG classification strategy

2.6. Performance Evaluation Criteria

The performance evaluation criteria of the regulatory information classification usually refers to the use of some quantitative indicators in the training test process (the training sample number is l and the category number is n), which can be used to evaluate the classification performance of the regulatory information classifier. Common criteria include precision, recal and F1 values.

Table 1. Cross matrix

Actual value	Predicted value	
	Output the number of texts that belong to this class	Output the number of texts that do not belong to this class
Output the number of texts that belong to this class	TP	FN
Output the number of texts that do not belong to this class	FP	TN

As shown in table 1, the introduction of cross matrix, Precision (P) and Recall (R) reflected the classification accuracy and completeness of the text classification system. The expression are as follows:

$$P = \frac{TP}{TP + FN} \tag{10}$$

$$R = \frac{TP}{TP + FP} \tag{11}$$

The accuracy rate and recall rate focus on different aspects of the evaluation of classification quality, and the comprehensive classification rate, namely F1 value, can be obtained by taking both aspects into consideration. The formula is as follows:

$$F1 = \frac{2 \times P \times R}{P + R} \tag{12}$$

To evaluate the overall classification performance of the regulatory power text classification system, Macro and Micro concepts were introduced to calculate the overall precision and recall and F1 values. The formula are as follows:

$$MAP = \frac{1}{l}\sum_{i=1}^{l} P_l \quad MAR = \frac{1}{l}\sum_{i=1}^{l} R_l \quad MAF1 = \frac{2 \times MAP \times MAR}{MAP + MAR} \tag{13}$$

$$MIP = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i} \quad MIR = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i} \quad MAF1 = \frac{2 \times MIP \times MIR}{MIP + MIR} \tag{14}$$

3. Experiment Procedure

3.1. Results of Different Chinese Word Segmentation Algorithms

Different Chinese word segmentation algorithms were selected to compare the classification performance of SVM algorithm. Based on the principle of SVM, this paper adopts the one-to-one multi-

classification strategy and respectively uses forward matching algorithm (MM), reverse matching algorithm (RMM) and two-way sample-and-reserve matching algorithm to test the precision, recall and F1 values of Chinese text. The results are shown in Figure 9-11.

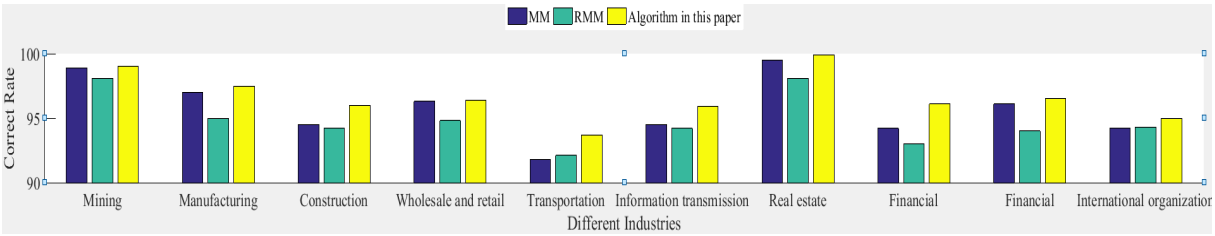


Figure 9. Precision of different segmentation algorithms

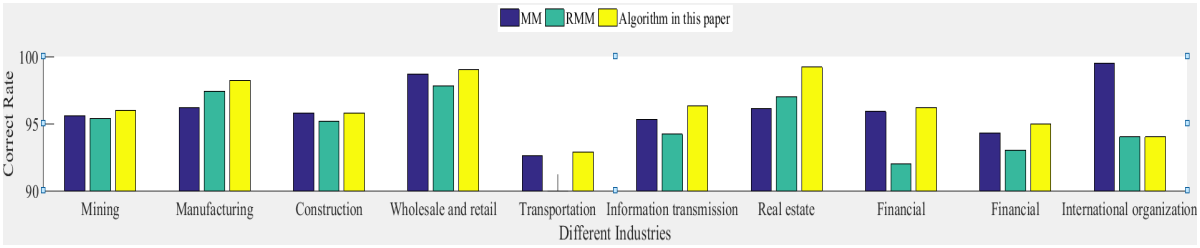


Figure 10. Recall rate of different word segmentation algorithms

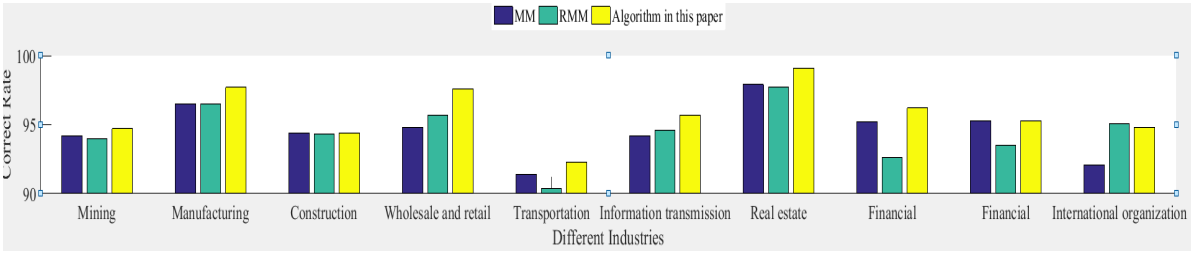


Figure 11. F1 values of different word segmentation algorithms

Experimental results show that the proposed algorithm is generally superior to MM and RMM in terms of three performance evaluation indexes, namely, the precision, recal and F1 values. It's just slightly lower on individual measures, which proves that the Chinese word segmentation algorithm adopted in this paper has better word segmentation ability.

3.2. Results of Different Multi-classification Strategies

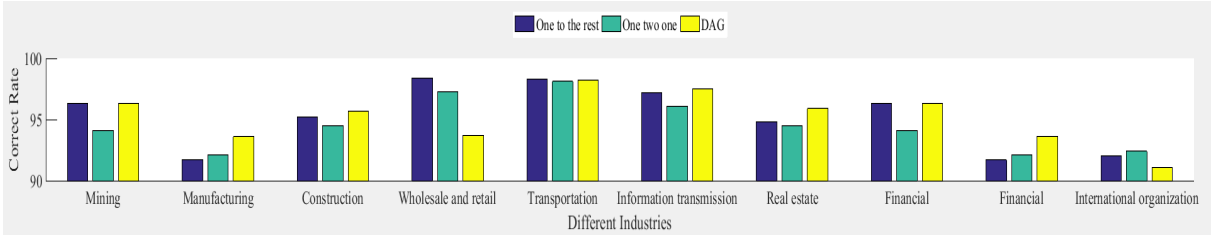


Figure 12. Precision of different multi-classification strategies

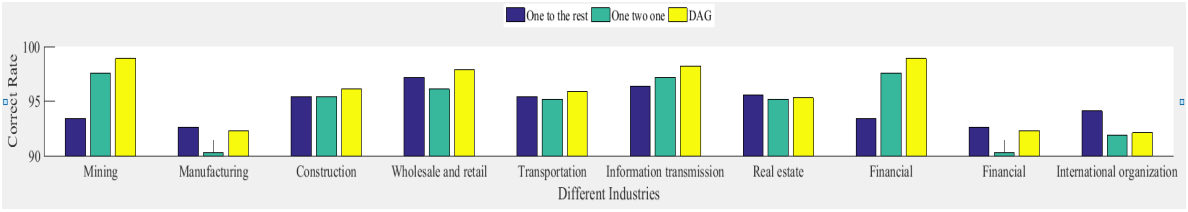


Figure 13. Recall rates of different multi-classification strategies

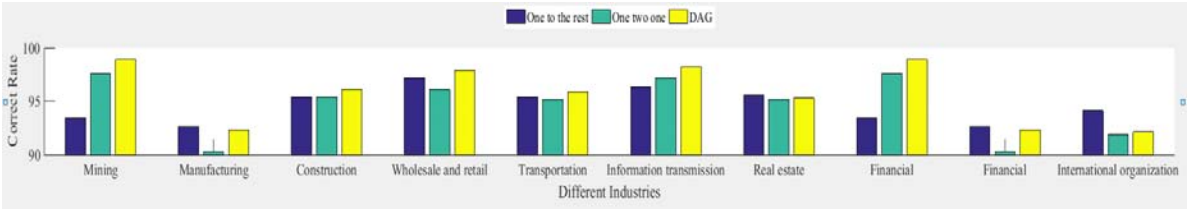


Figure 14. F1 values of different multi-classification strategies

The experimental results show that the one-to-one, one-to-one and DAG strategies achieve similar classification performance, while the one-to-one strategy performs slightly better than the latter two in F1 value, which proves that the one-to-one multi-classification strategy is superior in the overall classification performance of regulatory information. Therefore, it is the preferred multi-classification strategy for the system in this paper.

3.3. Results of Different Classification Algorithms

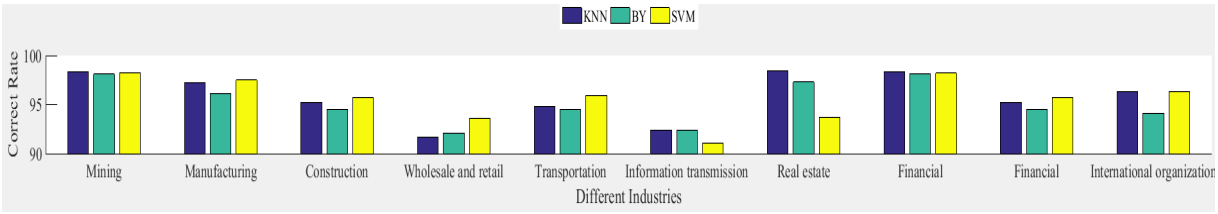


Figure15. Precision of different classification algorithms

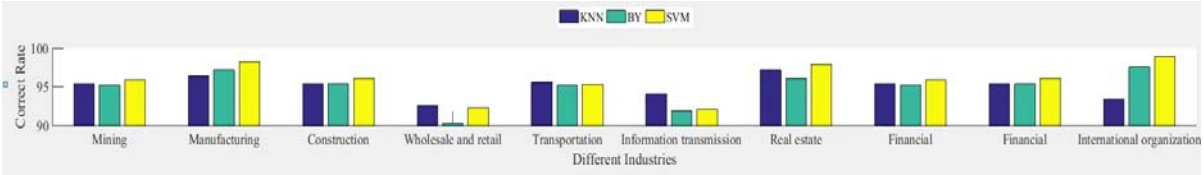


Figure 16. Recall rate of different classification algorithms

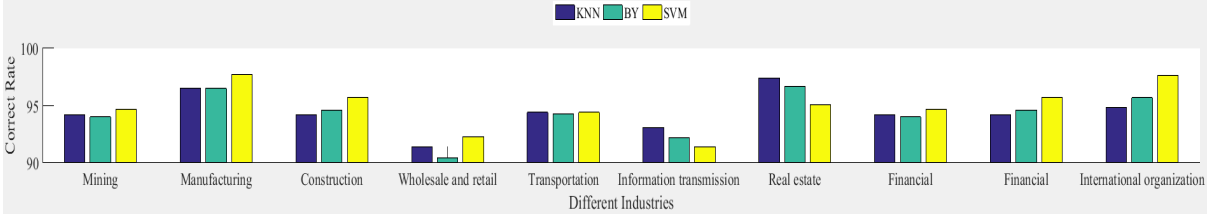


Figure 17 .F1 values of different classification algorithms

Experimental results show that the SVM algorithm is superior to KNN and BY algorithms in terms of three performance evaluation indexes, namely, accuracy, recall rate and F1 value, which proves the superior performance of SVM algorithm and the effectiveness and practicability of the Chinese text classification system of the regulatory information in this paper.

4. Conclusion

The Chinese text classification system of this paper adopts modular programming mode, which is easy to be reconstructed and extended. As the field of machine learning continues to develop at both the theoretical and application levels, the classification of text texts based on the Chinese regulatory power is also progressing. Opportunities and challenges coexist. In order to achieve higher accuracy rate of classification of regulatory power, new and efficient classification algorithm will become the future research direction.

Acknowledgments

This work was financially supported by Beijing Municipal Science and Technology Project (No. Z181100009018003).

References

[1] Feng Q, Chen H. The safety-level gap between China and the US in view of the interaction between coal production and safety management [J]. Safety science, 2013, 54: 80 - 86.
[2] Kitajima T, Shimada Y, Fuchino T, et al. Production Management Integrating Safety, Quality and Productivity [J]. Kagaku Kogaku Ronbunshu, 2014, 40 (3): 187 - 193.
[3] Hong C M, Chen C M, Chiu C Y. Automatic extraction of new words based on Google News corpora for supporting lexicon-based Chinese word segmentation systems [J]. Expert Systems with Applications, 2009, 36 (2): 3641 - 3651.
[4] Zhang L, Li Y, Meng J. Design of chinese word segmentation system based on improved chinese converse dictionary and reverse maximum matching algorithm [C]//International Conference on Web Information Systems Engineering. Springer, Berlin, Heidelberg, 2006: 171 - 181.
[5] Li Z, Sun M. Punctuation as implicit annotations for chinese word segmentation [J]. Computational Linguistics, 2009, 35 (4): 505 - 512.
[6] Zhang L, Li Y, Li D, et al. Call-Center-Oriented Method of Chinese Words Automatic Segmentation for Cotton Knowledge [J]. Sensor Letters, 2013, 11 (6-7): 1251 - 1255.
[7] Yang C C, Li K W. A heuristic method based on a statistical approach for Chinese text segmentation [J]. Journal of the American Society for Information Science and Technology, 2005, 56 (13): 1438 - 1447.
[8] Meng X, Liu S, Wang S. Word segmentation based on database semantics in NChiql [J]. Journal of Computer Science and Technology, 2000, 15 (4): 346.
[9] Zhang M, Lu Z, Zou C. A Chinese word segmentation based on language situation in processing ambiguous words [J]. Information Sciences, 2004, 162 (3-4): 275 - 285.
[10] Yang C C, Li K W. Segmenting Chinese unknown words by heuristic method[C]//International Conference on Asian Digital Libraries. Springer, Berlin, Heidelberg, 2003: 510 - 520.