Hindawi

Research Article

Chinese Language and Literature Online Resource Classification Algorithm Based on Improved SVM

Xiaowen Li (ID)

Cangzhou Medical College, Cangzhou, Hebei 061001, China

Correspondence should be addressed to Xiaowen Li; b20161002101@stu.ccsu.edu.cn

With the rapid development of network technology and the rapid growth of all kinds of online information resources, a large number of Chinese language and literature resources have emerged on the Internet. Online Chinese language and literature resources are becoming more and more important sources for people to obtain important information. However, existing search engines tend to provide a lot of irrelevant content when searching for information. Therefore, how to quickly and effectively obtain useful resource information and classify Chinese language and literature resources from a large number of information resources is the focus of this paper. This article mainly aims at the current development situation, by consulting a large amount of data, understanding the research status of improved SVM algorithm at home and abroad, showing the idea and advantages of SVM through the algorithm and experimental process, and further improving the SVM algorithm. The improved SVM proposed in this paper greatly improves the efficiency of classification and facilitates the rapid search of information with high matching degree among various online resources of Chinese language and literature.

1. Introduction

Since the 1990s, computer networks and database technologies have gradually matured, providing an effective platform for information sharing in various fields. Abundant information, on the one hand, brings unprecedented convenience to people's lives; on the other hand, it also raises some questions that people must face: how to effectively extract the required information from a large variety of information? How to make the extraction process more accurate and efficient? The data-based machine learning method is one of the results of people's efforts to solve such problems. The basic idea is to conduct research and analysis on known data, discover laws from it, and use it to predict unknown data. Text categorization is an important application of machine learning theory and an effective method to obtain information. From a mathematical point of view, it is a mapping process, that is, the text of an unknown category is mapped to a known category. After classification, the originally chaotic Chinese language and literature are classified into certain categories, so that people can follow

rules when looking up. When the number of Chinese language and literature resources is large, the significance of the classification of Chinese language and literature is obvious. The text classification method based on SVM overcomes the problem of data redundancy. It is based on a limited number of samples and can obtain the best classification effect under the information contained in the existing training text. At the same time, SVM is derived from VC dimension theory and structure risk minimization in statistical learning theory, referred to as SRM principle [1], which effectively solves the overlearning problem in other machine learning algorithms, that is, SVM is based on a limited number of training samples. Minimizing the training error can ensure that the test error is minimized.

2. Related Work

The research on automatic classification of foreign texts began in the late 1950s, when H. R. Luhn first proposed automatic text classification based on the idea of word frequency statistics. In 1960, Maron published the first paper

on automatic text classification. Subsequently, K. Spark, G. Salton, K. S. Jones, and others also conducted research in this field. Subsequently, a variety of new models and methods related to automatic text classification emerged in an endless stream. Compared with English text classification, the biggest difference in Chinese text classification lies in the preprocessing stage [2]. Chinese language and literature texts need word segmentation, while English texts are divided by spaces. In fact, after preprocessing the text, no matter what kind of text it is, it will be converted into the identity matrix of the sample vector, and the subsequent classification process is similar. The research of text classification technology in our country started relatively late. From the simple method of looking up the dictionary to the current text classification method based on statistics, it has gradually matured. Our country's research on automatic text classification began in the early 1980s and has generally experienced three development stages, i.e., the feasibility study-auxiliary classification system-automatic classification system [3]. In 1981, Mr. Hou Hanqing first explored the automatic classification of Chinese texts. Then, more and more domestic scholars participated in this field [4], and the main content of classification gradually changed from English to Chinese. After years of research, the gap between Chinese text automatic classification technology and the world's technical level is also constantly narrowing. More classic text classification algorithms: decision tree method, neural network method, genetic algorithm, Bayesian, and K-nearest neighbor algorithm [5], have been widely used in web text classification. There are many domestic research studies on text representation and text classification. For example, the Chinese text classification technology based on N-gram [6] has got rid of the dependence on dictionary and word segmentation and has realized the field independence and time of text classification. In recent years, domestic scholars have cited text classification techniques in different fields and have produced many new methods: group-based classification methods, fuzzy-rough set-based methods, multi-classifier fusion methods, and text classification models based on RBF networks. Ma and Tian [8] select high-frequency words that appear in the same window unit and use a prior algorithm to mine the largest frequent word co-occurrence set from these high-frequency words, thereby extending VSM to represent documents. Ma and Tian [8] compared the effect of support vector machine algorithm and traditional KNN algorithm in text classification through experiments, proposed an improved method for text pre-processing according to the shortcomings of support vector machine algorithm, and proved the effectiveness of the new method. Aiming at the vector space model in text representation, Liu Haifeng et al. [9] used the category space model to represent the text as a matrix and effectively used the category information of the text to implement a text classification algorithm based on category information [9]. Li Xiangdong et al. [10] defined the training text set, test set, and classification (column) effect evaluation required for automatic classification based on the analysis of the characteristics of the journal's permanent topic columns and used Jensen–Shannon divergence to calculate the text gap. It also improves the basic principle of the KNN algorithm in determining the k value based on the column dynamics and has achieved good results.

Key objectives of the study include the following:

(i) To design and develop an efficient methodology to quickly and effectively obtain useful resource information such as Chinese language and literature from information resources and classify them.

(ii) To improve performance of the SVM algorithm and demonstrate the idea of the improved version of the SVM along with its advantages.

3. Method

3.1. Support Vector Machine (SVM). The basic idea of support vector machine is to establish an optimal hyperplane as the decision surface [11], so that the classification interval between positive and negative samples is maximized. Because different kernel functions will lead to different SVM algorithms, choosing the right kernel function is a crucial step. There are different forms of kernel functions such as polynomial kernel, Gaussian radial basis kernel, sigmoid kernel, and so on. The advantage of support vector machine lies in its ability to learn in sparse, high-dimensional space. It can use fewer training samples to minimize the range of empirical errors and the complexity of the classifier, avoid local minima, and effectively solve the problem, with good generalization and good classification accuracy. The important point is that contrary to the neural network, its objective function is a unimodal function, so it can reach the global optimal solution to a large extent.

3.2. Improved Multi-Class SVM Algorithm. In the process of training SVM dataset, singular points with larger Lagrangian multipliers can be converted into support vectors. So many improved support vector machines are present such as RSVM [13] (robust support vector machine) is based on calculating the distance between various training data centers and data points to establish an adaptive decision surface, but the disadvantage is that it is difficult to determine additional parameters; SVND [14] (fuzzy support vector machines) can effectively extract and detect singularities from normal data, but this method is usually only applied to a class of classification problems; and FSVM [15] (fuzzy support vector machines), in which the fuzzy membership of the training data reduces the influence of outliers, but how to choose the membership function is still a big problem. Therefore, this paper proposes an improved multi-class SVM algorithm (WSVM) to solve the singular value sensitivity problem of classic SVM. The basic idea is to assign different weights to different training data, so that SVM can construct a decision-making surface according to the relative importance of each kind of data. The main idea of the SVM model is to transform the original optimization problem into a quadratic programming problem to construct the optimal decision surface. In the SVM, different kernel options are available, and we have used the Gaussian kernel for our study.

The training dataset is as in formula (1), where $(x_i, c_i), x_i \in R^N$.

$$T = (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \in R^N \times \{+1, -1\}. \quad (1)$$

The original optimization problem is as in formulas (2)–(4):

$$\min(w, \xi) = \frac{1}{2}w^2 + C\sum_{i=i}^{N} \xi_i, \quad (2)$$

$$y_i((w.x_i) + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, N, \quad (3)$$

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, N, \quad (4)$$

where $\xi = (\xi_1, \ldots, \xi_N)^T$ and $C > 0$ is a penalty parameter, not only to minimize $\|w^2\|$, but also to extremely minimize $\sum_{i=i}^{N} \xi_i$.

In order to derive the dual problem of (2) and (3), the Lagrange function is introduced:

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}w^2 + C\sum_{i=i}^{N} \xi_i - \sum_{i=i}^{N} \alpha_i(y_i((wgx_i) + b) - 1 + \xi_i) - \sum_{i=i}^{N} \beta_{i\xi_i}, \quad (5)$$

where $\alpha = (\alpha_1, \alpha_2, \ldots \alpha_n)^T$ and $\beta = (\beta_1, \beta_2, \ldots \beta_n)^T$ are both multiplier vectors of the Lagrange function. Optimization problem:

$$\max -\frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} y_i y_j (x_i g x_j)\alpha_i \alpha_j + \sum_{j=1}^{n} a_j, \quad (6)$$

$$s.t \sum_{j=1}^{n} y_i \alpha_i = 0, \quad (7)$$

$$C - \alpha_i - \beta_i = 0, \quad i = 1, 2, \ldots n \quad (8)$$

$$\alpha_i \geq 0, \quad i = 1, 2, \ldots n, \quad (9)$$

$$\alpha_i \geq 0, \quad i = 1, 2, \ldots n. \quad (10)$$

Use constraint (8) to eliminate β, which becomes a problem with only variable α. This problem is reduced to a minimization problem, that is, a convex quadratic programming problem:

$$\max \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j (x_i g x_j)\alpha_i \alpha_j - \sum_{j=1}^{n} a_j, \quad (11)$$

$$s.t \sum_{j=1}^{n} y_i \alpha_i = 0, \quad (12)$$

$$0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots n. \quad (13)$$

The weighted SVM algorithm treats these data differently according to different weights of each kind of data. Weighted SVM assigns higher weight to data with important information and similarly assigns lower weight to data with less important information. Weighted training dataset:

$$T = (x_1, y_1, v_1), (x_2, y_2, v_2), \ldots, (x_N, y_N, v_N) \in R^N \times \{+1, -1\}, \quad (14)$$

where $\varepsilon < v_i < 1$ is the weight of (x_i, c_i) $(i = 1, 2, \ldots N)$, ε is a sufficiently small positive number, and $x_i \in R^N$. Like SVM, weighted SVM mainly achieves the correct rate of classification by maximizing the classification interval and minimizing the classification error rate. The difference with SVM is that the improved SVM uses a weighting function to reduce the impact of some nonimportant data on the classification results and enhance the impact of important data on the classification. In the case of data weighting, the optimal decision-making surface is constructed, and the optimization problem is transformed into formulas (15)–(17):

$$\min(w, \xi, v) = \frac{1}{2}w^2 + C\sum_{i=i}^{N} v_i \xi_i, \quad (15)$$

$$y_i((w.x_i) + b) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, N, \quad (16)$$

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, N. \quad (17)$$

It can be seen from formula (16) that the introduction of v_i greatly reduces $\sum_{i=i}^{N} \xi_i$ and reduces the influence of the slack variable ξ_i in the optimization problem. Therefore, the visible (x_i, c_i) data are not important for classification. The above weighted optimization problem can be transformed into a convex quadratic programming problem, such as formulas (18)–(20):

$$\max \sum_{i=1}^{N} a_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \ldots x_j), \quad (18)$$

$$s.t \sum_{i=1}^{N} \alpha_i y_i = 0, \quad (19)$$

$$0 \le \alpha_i \le v_i C, \quad i = 1, 2, \ldots N. \tag{20}$$

$K(.) = (x_i, x_j)$ is the kernel function and a is the inner product. This paper adopts the Gaussian radial basis function (RBF) with better learning ability and wider convergence range. It can be seen that when $v_i = 1$, the improved SVM is transformed into the original problem of the support vector machine, and the compromise of x_i in the system can be determined for different v_i. The smaller v_i is, the smaller x_i is for the construction of the maximum edge hyperplane, and vice versa.

SVM is a binary classifier and does not support multi-class classification. However, it has support for binary classification and can separate data points into two classes. Hence, for multi-class classification problems, the same principle is applied after breaking down the multi-class classification problem into multiple binary classification problems.

4. Experiments and Discussion

4.1. Performance Analysis of Chinese Language and Literature Online Resource Classification Algorithm Based on Improved SVM. 1329 articles were downloaded from the Internet, which were divided into three first-level categories, of which the training document set included 1038 articles and the test set included the rest of the articles. The performance of a classifier is usually measured by some evaluation indicators. The commonly used performance evaluation indicators in the classification of Chinese language and literature include recall, precision, and *F*-measure. Different evaluation criteria have different evaluation objectives and calculation methods. Recall rate refers to the ratio of the number of related documents retrieved to the total number of related documents, and the precision rate refers to the ratio of the number of related documents retrieved to the total number of documents returned. The *F*-measure is the most commonly used method to measure the overall classification effect. Micro-average is the evaluation of the overall performance of classification, which is the arithmetic average of the performance indicators of each instance document as shown in (24):

$$\text{precision} = \frac{A}{A + C}, \tag{21}$$

$$\text{recall} = \frac{A}{A + B}, \tag{22}$$

$$F - \text{measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}, \tag{23}$$

$$r = \frac{\sum_{c_i \in C} r_c}{|C|}, \tag{24}$$

where A represents the number of documents that are correctly classified, B represents the number of documents that actually do not belong to a certain category but are classified into that category, and C represents the number of
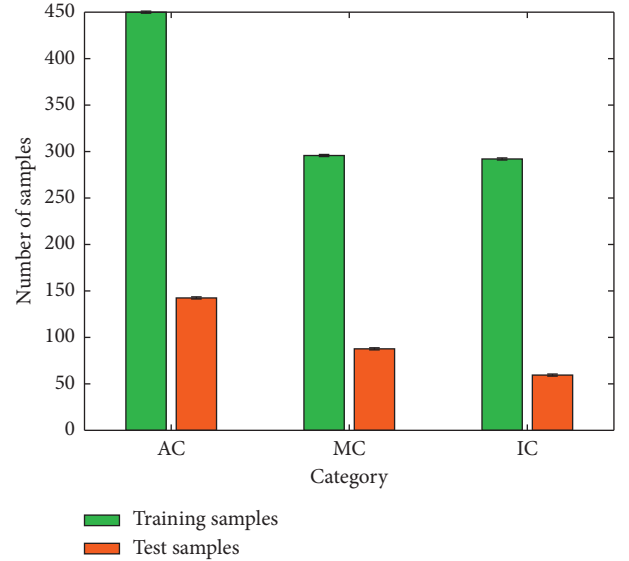

FIGURE 1: Training sample and test sample situation (AC is ancient Chinese language and literature, MC is modern Chinese language and literature, and IC is introduction to Chinese language and literature).

documents that actually belong to a certain category but are not classified. The recall rate mainly refers to the ratio of texts that meet the requirements in the set of documents and meet the requirements in all the texts that meet the requirements. The accuracy rate refers to the ratio of the documents that really meet the requirements in all the documents retrieved, and it indicates the accuracy of the classification results. Among them, the check rate and the recall rate have a constrained relationship with each other. As the accuracy rate increases, the recall rate will decrease, and vice versa. The *F*-measure combines these two indicators to evaluate the overall performance of the classifier. This experiment adopts commonly used evaluation methods, recall, precision, and *F*-measure, as shown in Figure 1.

4.2. System Design of Chinese Language and Literature Online Resource Classification Based on Improved SVM. The traditional Chinese language and literature text classification system based on SVM is divided into two modules: classification function training and test text classification. In these two modules, the processes of Chinese language and literature word segmentation, feature extraction, feature selection, and text vector representation are included. In this way, the two modules have a large number of repetitive functions and strong coupling. In the process of implementing the Chinese language and literature text classification system based on SVM, this paper has carried out a more rational module design, namely, the whole system is divided into text preprocessing module, classification function training module, and text classification module, and Chinese language and literature word segmentation, feature extraction, feature selection, and text vector representation are integrated into the text preprocessing module. It also provides a common input interface for the training text and the test

TABLE 1: Text classification experiment data.

Category	Number of training texts	Number of test texts
Data structure	300	100
Computer network	300	100
Artificial intelligence	300	100

text, and the preprocessing system outputs the feature vector of the training text and the feature vector of the test text according to different situations. In this way, the text vector representation function does not need to be involved in the training module and the classification module, which is beneficial to the development and maintenance of the system and makes the system have better performance. At the same time, this system is a Chinese language and literature text classification system based on the nonlinear separable SVM algorithm. It uses the polynomial kernel function $\ker(x, x_i) = (xx_i + 1)^q$ and uses the feasible direction method to solve the algorithm.

4.3. Implementation of the System of Chinese Language and Literature Online Resource Classification Based on Improved SVM.
System development is implemented in the Windows environment using Java language, which specifically includes the following: Development platform: Microsoft Windows, Eclipse 3.2+MyEclipse5.0, jsdk1.5, and development language: Java.

In the traditional Chinese language and literature text classification system based on the SVM algorithm, it mainly includes a classification function training module and a text classification module, which correspond to the training of the classification function and the classification of the test text, including the conversion of the text into a feature vector. In the classification function training module, the training text is converted into a feature vector with category information, and in the text classification module, the test text is converted into a feature vector without category information. In fact, these two processes are very similar, and the only difference is whether the obtained feature vector contains known category information. In view of this, this paper adjusts and improves the structure of the traditional SVM-based Chinese language and literature text classification system. The process of expressing training text and test text into feature vectors is integrated into the text preprocessing module. In this way, after the text vector representation process is completed in the text preprocessing module, they can be substituted into the classification function training module and the text classification module for calculation, which is beneficial to the development and maintenance of the system. After modular design, this system is mainly composed of text preprocessing module, classification function training module, and text classification module.

4.4. Experimental Results and Algorithm Performance Analysis.
The improvement of this system is mainly reflected in the structure of the system, and its classification efficiency and classification accuracy should be equivalent to the traditional SVM-based Chinese language and literature
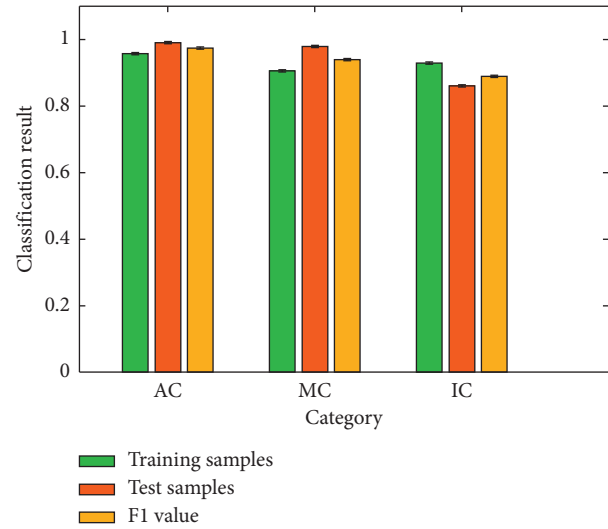

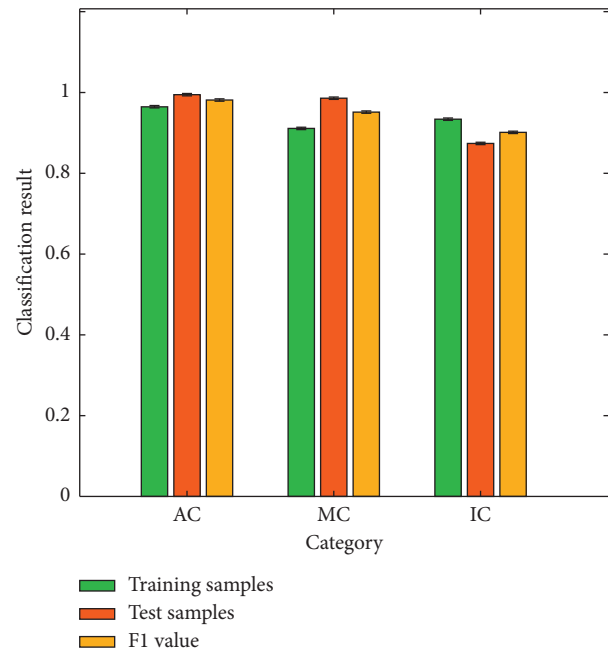FIGURE 2: Multi-class SVM algorithm classification results.

FIGURE 3: Weighted multi-class SVM algorithm classification.

text classification system. In order to accurately understand its performance, this paper selects 3 sets of data and 1200 Chinese language and literature texts related to various subjects for testing and compares the test results with the test results of the traditional Chinese language and literature text classification system based on SVM algorithm. The specific distribution of text data is shown in Table 1.

During the experiment, when one type of training text is regarded as a positive type, any other type is regarded as a negative type. The specific experimental steps are as follows:

(1) The training text is input into the preprocessing subsystem, and after Chinese word segmentation, feature extraction, feature selection, and text vector representation, the training text feature vector with

category mark recognition is obtained, and the training vocabulary is established. Then, save the obtained feature vector and training vocabulary as a text file and put it into the specified file.

(2) Read the training text feature vector, input it into the classification function training subsystem to solve the quadratic programming problem, determine the classification function parameters w and b, and construct a specific classification function.

(3) The test text is input into the preprocessing subsystem, and after Chinese word segmentation, feature extraction, feature selection, and text vector representation, the feature vector of the test text without category identification is obtained. Then, save the obtained text vector as a text file and put it into the specified folder.

(4) Read the test text feature vector, input it into the classification function, and calculate the function value. Then, output the text category according to the function value.

(5) Statistic experiment results: plot 3 sets of data into tables for comparative analysis.

The final result is shown in Figures 2 and 3.

From the experimental results, it can be found that the recall and precision of the Chinese text classification system based on SVM before and after the improvement are equivalent, and the error between them is related to the experimental environment and other factors. But after adjusting and improving the structure of the system, the internal cohesion of each module is stronger, and the coupling between the modules is weaker, which is conducive to the development and maintenance of the system.

5. Conclusions

With the development and perfection of computer network and database technology, abundant resources make text classification technology gradually become the focus of people's attention. Because of its solid theoretical foundation and excellent performance, support vector machines have become a research hotspot in text classification and other data mining fields. This article focuses on the key technologies and text classification algorithms of Chinese language and literature text classification. Compared with the traditional SVM-based Chinese language and literature text classification system, this system is unique because of the following reasons. (1) The process of Chinese language and literature word segmentation, feature selection, feature extraction, and text vector representation is integrated into the preprocessing module. A common input interface is provided for the training style and the test text, and the preprocessing system outputs the feature vector of the training text and the feature vector of the test text according to different situations. In this way, the text vector representation function does not need to be involved in the training module and the classification module, so that the coupling between the modules is weakened and the independence is

enhanced, which is beneficial to the development and maintenance of the system. (2) We use the feasibility direction method to solve the secondary programming problems involved in the training process. Since the training of the classification function is the key to this system, the following briefly analyzes the time complexity of the classification function obtained by the system from the training library. In this system, suppose the number of training texts is n, the average number of feature words in each text is k, and the text is represented as a t-dimensional vector. In the process of training the SVM classification function, the running time from text segmentation to text vector representation is at least $5n * k + (nk)^2$, and the running time is consumed in the process of establishing the training vocabulary and text vector representation. Then, the feasibility direction method is used to solve the quadratic programming problem involved in the SVM algorithm to complete the training process of the classification function. Its running time is $C * [(3t + 2)n + t]$, where C is a positive integer much smaller than n, so the total time complexity of training the classification function in this system is $O(5n * k + (nk)^2 + C * [(3t + 2)n + t])$, that is, $O((nk)^2)$. In addition, in the actual application process, the training time of the SVM algorithm is also related to the quality of the training library, the dimension of the text vector, and the configuration of the computer [15-23].

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony, "Structural risk minimization over data dependent hierarchies," *EEE Transactions*, vol. 44, 1998.

[2] X. Y. Zhang, "Research on automatic classification of Chinese text based on rough set theory," *Master's thesis*, vol. 6, no. 5, pp. 10-11, 2005.

[3] J. B. Tan, "Design and implementation of an automatic text classification system for online education resources," *Technology Application*, vol. 4, no. 4, pp. 68-69, 2009.

[4] X. G. Zhang, "About statistical learning theory and support vector machines," *Acta Automatica Sinica*, vol. 26, no. 1, pp. 32–38, 2000.

[5] F. Huang, "Research on web document feature extraction in basic education search engine," Master's Thesis, Nanjing Normal University, Nanjing, China, 2006.

[6] X. Yao, X. D. Wang, Y. X. Zhang, and W. Quan, "Feature selection algorithm based on approximate Markov Blanket and dynamic mutual information," *Computer Science*, vol. 39, no. 8, pp. 1046–1050, 2012.

[7] T. H. Zhang, H. T. Geng, and Q. S. Cai, "An improved SVM and its application in automatic text classification," *Microelectronics & Computer*, vol. 22, no. 12, pp. 24–27, 2005.

[8] J. N. Ma and D. G. Tian, "Research on automatic classification of Chinese text based on support vector machine," *System*

Engineering and Electronic Technology, vol. 29, no. 3, pp. 475–478, 2007.

[9] H. F. LiuS. S. Liu et al., "An automatic text classification model based on category information," *Modern Library and Information Technology*, vol. 5, no. 4, pp. 72–76, 2010.

[10] X. D. LiP. Xu et al., "Research on Automatic Text Classification Method Based on KNN Algorithm—Taking Automatic Classification of Academic Journals as an Example," *Library, Information and Knowledge*, vol. 5, no. 4, pp. 71–76, 2010.

[11] J. Yuan, J. Li, and Bo Zhang, "Learning concepts from large scale imbalanced data sets using support cluster machines," in *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pp. 441–450, ACM, Santa Barbara, CA, USA, Octobe 2006.

[12] Q. Song, W. Hu, and W. Xie, "Robust support vector machine with bullet hole image classification," *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, vol. 32, no. 4, pp. 441–447, 2002.

[13] C.-Fu Lin and S.-De Wang, "Training Algorithms for Fuzzy Support Vector Machines with Noisy data," in *Proceedings of the 2003 IEEE 8th Workshop on Neural Networks for Signal Processing(IEEE Cat. No.03TH8718)*, pp. 517–526, Toulouse, France, September 2003.

[14] Li J. Cao, H. P. Lee, and W. K. Chong, "Modified support vector novelty detector using training data with outliers," *Pattern Recognition Letters*, vol. 24, no. 14, pp. 2480–2484, 2003.

[15] G. Y. Jiang, Y. Zhang, and M. Yu, "Multi-mode and multi-view video coding based on correlation analysis," *Chinese Journal of Computers*, vol. 30, no. 12, pp. 2206–2208, 2007.

[16] J. Q. Wang, "Research on the construction of featured database of university library—take Minjiang University for example," *Journal of Minjiang University*, 2013.

[17] C.-Fu Lin and S.-De Wang, "Fuzzy support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 464–471, 2002.

[18] J. C. Burges, "A tutorial on support vector machine for pattern recognition," *Knowledge Discovery and Data Mining*, vol. 2, no. 2, pp. 135–140, 1998.

[19] B. Fei and J. Liu, "Binary tree of SVM: a new fast multiclass training and classification algorithm," *IEEE Transactions on Neural Networks*, vol. 17, no. 3, pp. 696–704, 2006.

[20] Yu Lei and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Machine Learning Research*, vol. 5, no. 1, pp. 1207–1220, 2004.

[21] D. Koller and M. Sahami, "Toward optimal feature selection," in *Proceedings of International Conference on Machine Learning*, pp. 162–187, Bari, Italy, July 1996.

[22] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, NY, USA, 1995.

[23] J. Pei, K. Zhong, J. Li, and X. Wang, "ECNN: evaluating a cluster-neural network model for city innovation capability," *Neural Computing & Applications*, vol. 6, pp. 1–13, 2021.