Study on Support Vector Machine Combined with Infrared Spectroscopy for Timber Species Identification

Yi-Dan Sun, Jun-Yi He, Mei-Hua Wu, Jing-Jing Zheng, Yuan Gao, Xue-Shun Wang
Beijing Forestry University, Beijing 100083, China
13021236422@163.com

Abstract - Via infrared spectroscopy (IR) combined with support vector machine (SVM), the study on timber species identification was carried on. Ten kinds of precious timber were used as experimental materials; each timber picked three sets of samples. The corresponding spectrum was recorded by infrared spectrometer. The spectral data was pretreated by baseline correction and dimensionality reduction. Radial basis function (RBF)was selected as kernel function , and RBF coefficient (γ) was 0.01.As for cross-validation ,the model of timber species identification was respectively established by the adjustment of the training set and test set, the discriminant accuracy rate of three models were 70%, 80%, and 100 %. The optimal model was compared with the model of Cluster analysis and Bayes discriminant, which indicated that the SVM- infrared spectroscopy technology has better prediction results and certain research value for the development of the timber species identification.

Index Terms - Timber Identification, Support Vector Machine, Infrared Spectroscopy, Cluster Analysis, Bayes Discriminant

1. Introduction

It is of great practical significance to seek a kind of rapid and accurate timber identification method [1] for the protection and efficient utilization of the precious timber. Conventional timber identification methods mainly based on the structure characteristics [2], which depend on the practical experience, are adverse to the development of timber identification technology and application. At home and abroad, the genetic method (DNA markers) [3], chemical method (stable isotope)[4] and near infrared spectroscopy (NIR) technology [5] has been used in timber identification, which improve the accuracy and efficiency of recognition.

As a kind of advanced detection technology with the advantages of high characteristic, short analysis time and less sample [4], infrared spectroscopy qualitative analysis technology has received widespread attention in many fields. Recent years some experts have carried on the research upon the traditional Chinese medicinal identificat- ification [6] by the quantitative analysis of infrared spectrum, which has obtained certain achievements.

With the widespread utilization of intelligent computer aided spectrum analysis technology for timber species identification, intelligent algorithms such as support vector machine has been a new research hotspot in wood science field [7].Support vector machine is a kind of statistical estimation and prediction problems with small sample, nonlinear and high dimensional pattern recognition. Recent years many research achievements have sprung up, such as" Chunking" block[8]; Support vector machine algorithm[9]; The NPA closest point algorithm[10]; Least squares support vector machine algorithm, etc.[11]They have been successfully applied in text classification [12], biological information [12], speech recognition [14] and many fields .

This study took 10 kinds of precious timber as the research object, established a model of timber species identification based on support vector machine combined with infrared spectroscopy technology, and compared with the model of clustering analysis and Bayes discriminant analysis, which improves the discriminant accuracy, and provides a certain reference value for further study on the timber identification.

2. Materials and Methods

A. Theories of support vector machine

As a new learning algorithm that firstly presented and applied by Vapnik, support vector machine (SVM) initially solves data process of binary classification problems. SVM model give priority to solve the problem of the kernel function selection, which solves the problem of computation complexity increasing due to vector mapping from low dimensional space to high-dimensional space. Common kernel function are polynomial function, radial basis function, etc., their expression is as follows:

Polynomial function:

$$K(x_i, x_j) = \left(\frac{x_i \cdot x_j}{a} + b \right)^p$$

Radial basis function:

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$$

and γ presents the coefficient of radial basis.

B. Apparatus

Spectrometer: IR spectra were recorded on a Spectrum GX FT-IR system (PerkinElmer) equipped with a DTGS detector. FTIR spectra were recorded from a total of 16 scans in the 4000–400cm-1 range with a resolution of 4 cm-1.

Software: MATLAB 7.0, PASW Statistics 18, SAS 9.0, Excel 2010.

Computer system: WINDOWS 7, 32-bit.

C. Materials and sample preparation

Sample collection: Ten kinds of precious timber for this study were provided by Research Institute of Wood Industry, Chinese Academy of Forestry, China. The detailed

information of samples is shown in Table 1. The corresponding spectrum was acquired through analyzing dynamic spectra which were collected at different temperatures ranging from 50 to 120℃ at an interval of 10℃ by IR software developed by Tsinghua University (Beijing, China).

Table 1 The details of samples

Label	Latin name	Family	Genus
1	Guibourtiaehie	Cæsalpinioideae	Guibourtia
2	Guibourtiaspp	Cæsalpinioideae	Guibourtia
3	Hymenaeaspp	Cæsalpinioideae	Hymenaea
4	Intsiaspp	Cæsalpinioideae	Intsia
5	Koompassiaspp	Cæsalpinioideae	Koompassia
6	Peltogynespp	Cæsalpinioideae	Pterogyne
7	Pterogynenitens	Cæsalpinioideae	Pterogyne
8	Sindoraspp	Cæsalpinioideae	Sindora
9	Terminaliaspp	Combretaceae	Terminalia
10	Diospyrosspp	Ebenaceae	Diospyros

Sample preparation: 2 mg of test sample was mixed with 100 mg of KBR broken crystal and the mixture was further grounded and pressed under 10 tons of pressure to produce a thin disk with 13 mm in diameter. Then the disk was put into the infrared spectrometer sample holder for testing.

D. Data preprocessing

In the process of sample preparation of potassium bromide tablet, infrared light is scattered due to the opaque tablet for no fine grinding, which raises the high frequency end of the spectrum baseline; so baseline correction is usually required for the spectral appearance, one refers to that spectrum baseline is artificially pulled back on the 0 baseline. Thus, this study obtained available spectral data for experiments by the means of baseline correction and standardization. Before baseline correction, the original spectrum was converted into absorbance spectra. Three groups of parallel experiments were made , in order to ensure the clarity of the spectrum, five kinds of timber samples of one group was selected to draw infrared spectra. The spectrum is shown in Figure 1:

Given the wide band and serious overlapping absorption peaks, IR data often do not reflect the essential characteristics of the spectral. Via the application of principal component analysis for data dimensionality reduction, multiple variables will be integrated into unrelated variables-principal components, which reflect the vast majority of information variables. Consequently, data storage and calculation is reduced mostly, and the effect of noise is removed. This study selected the wave from 800cm to 1800cm with characteristic absorption peaks, after dimensionality reduction, obtained 16 principal components, the contribution rate is 99.684%.

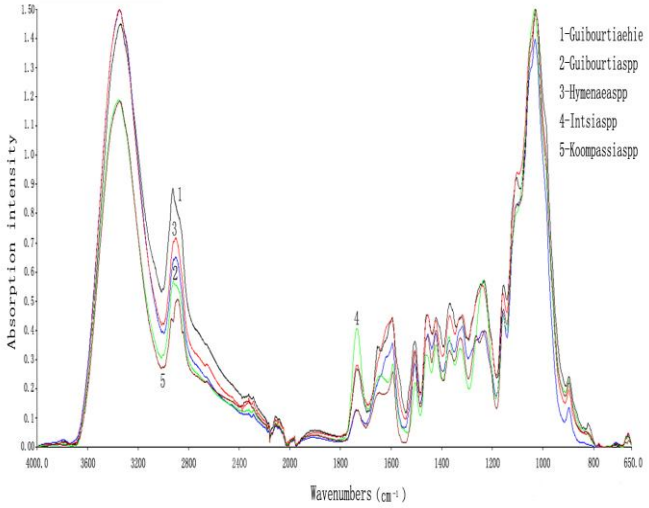

Fig. 1 The spectrum of five kinds of timber samples

3. Results and Discussion

A. The selection of kernel function type

As for the effects of support vector machine classification model, different type of kernel function makes significant difference. Three groups of timber samples were numbered sample I, sample II and sample III respectively in this study, and 10 kinds of timber are numbered from 1 to 10 in each group. Samples were randomly selected as the training set and corresponding test set. (Sample I and sample III were selected as the training set, sample II as the test set in this study), polynomial function and radial basis function was chosen as the kernel function, respectively. Other parameters were all chosen the default value (the SVM model (-s) =0, penalty parameter (-c) =1, radial basis coefficient (γ) =0.01), and then different support vector machine (SVM) models were established. The results are as follows:

The classification accuracy is 20% for the selection of polynomial kernel function (-t=1), while the rate is 70% as the kernel function was selected RBF (-t=2), the results are shown in Table 2, where the date with red box was on behalf of the correct category., the discriminant accuracy of RBF model is higher than the model of polynomial kernel function from the experimental results, due to that RBF can map the nonlinear sample data to high-dimensional feature space as to deal with the sample data with nonlinear relations. In addition that RBF values ($0 < K < 1$) is easier than polynomial function values ($0 < K$ or $K > 1$) for faster computing speed. Therefore, this experiment adopted RBF as the kernel function.

Table 2 The discriminant result of SVM model based on polynomial kernel function and RBF kernel function

Sample Tag	1	2	3	4	5	6	7	8	9	10	Discriminant accuracy (%)
Category of polynomial functions	8	8	3	8	8	8	8	8	8	8	20
Category of RBF functions	1	2	3	2	5	6	2	8	9	8	70

B. The selection of radial basis coefficient (γ)

Parameter γ controls sensitivity of SVM on the inputs change. Size of γ was determined through test, sample I and sample III was fixed as training set, samples II as test set, RBF as kernel function, other parameters were selected the default values (SVM model (-s) =0, punishment parameter (-c) =1).Via the adjustment of γ, SVM model was established respectively. Experimental results are as Table 3, which showed the change of discriminant accurate rate with the value adjustment of γ. As a result of few samples, the discriminant accuracy rate is 70% as the value of γ ranges from 0.001 to 0.05, which is the highest. However, SVM is unresponsive due to the ambassador value of γ, leading to poor discriminant results. After comprehensive consideration, this experiment selected 0.01 as the best value of γ for timber discriminant.

Table 3 Influences of γ on the result of discriminant accuracy of SVM model

The value of γ	**Discriminant accuracy (%)**
1.0×10^{-3}	70
5.0×10^{-3}	70
1.0×10^{-2}	70
5.0×10^{-2}	70
1.0×10^{-1}	60
5.0×10^{-1}	50
1.0	50
5.0	40

C. Cross validation

Through the above artificial selection on the kernel function and radial based coefficient, the model of timber identification based on SVM was established. In order to further improve the discriminant accuracy, different models were established by adjustment of the training set and corresponding test set. Then realized cross validation via comparison with the original model. The results are presented in Table 4, which showed the influence of different training set and testing set on the discriminant accuracy of the model. The discriminant accuracy rate is 100% as sample I and sample II were selected as the training sample set while sample III as the test set, which is the optimal model in the experiment.

Table 4 Influence of the training set and the test set on the results of discriminant accuracy

Training set of model	**Test set of model**	**Discriminant accuracy (%)**
Sample I, Sample II	Sample III	100
Sample I, Sample III	Sample II	70
Sample II, Sample III	Sample I	80

D. Analysis of comparison among SVM, clustering analysis and Bayes discriminant

Based on above experimental results, sample I and sample II was selected as the training set, sample III as the test set, RBF as the kernel function, 0.01as the radial basis coefficient, SVM model is established for timber identification. In order to investigate the discriminant results, SVM model was compared with clustering analysis and Bayes discriminant on discriminant error rate. The experimental results are shown in Table 5, which shows that the discriminant accuracy rate of clustering analysis is 30%, Bayes discriminant is 86.67%, and thus the discriminant effect of SVM model is significant, which has certain superiorities in small sample classification, and provides a certain referable value for the further research of timber identification.

Table 5 Comparison of discriminant accuracy rate among SVM, clustering analysis and Bayes discriminant

Discriminant method	Discriminant accuracy (%)
Clustering analysis	30.00
Bayes discriminant	86.67
Support vector machine	100.00

4. Conclusions

The model of timber species identification based on support vector machine (SVM) combined with infrared spectroscopy (IR) technology was used to identify the target of ten kinds of precious timber. Via artificial selection as well as cross-validation, the optimal model was set by selecting RBF as kernel function, 0.01 as radial coefficient, sample I and sample II as training set, sample III as test set. The results indicated that SVM-IR technology can be applied to the quantitative analysis of timber identification. SVM has certain advantages in classification of small samples; the discriminant accuracy is associated with its parameter selection. Study on parameter optimization of SVM and neural network algorithm of swarm intelligence for further improvement of timber discriminant accuracy, will be the main direction of future research.

Acknowledgment

The authors would like to acknowledge the Students'Scientific Research Innovation Projects(SRIP)of Beijing Forestry University(BJFU) for financing the project, as well as Research Institute of Wood Industry, Chinese Academy of Forestry(CRIWI) for providing the timber samples. Thanks also are due to the Tsinghua University chemistry laboratory, of the Infrared Spectrograph service, for allowing experiments to be performed on their premises.

References

[1] Vlandimir N Vapnik. The nature of statistical learning theory. Zhang Xuegong, eds. Beijing: Tsinghua university press, 1999. nik. The nature of statistical

[2] Jiang Xiaomei, Yin Yafan, Liu Xiaoli etc. IAWA softwood microstructure features list. Journal of wood industry, 2004, 18 (4): 37, 38. (in Chinese)

[3] Jiang Zehui. Thorough going efforts to promote scientific and technological innovation to support leading modern forestry urban forestry in China 2009: 4-8 (in Chinese)

[4] Wang Hangjun, Zhang Guangqun, Qi Heng nian, etc. Wood identification method research overview of zhejiang forestry college journal 2009 (6) 896-896-26 (in Chinese)

[5] Jiang Xiaomei YanYaFang, Liubo Timber species recognition technology present situation, development and prospect of wood industry 2010:36-39 (in Chinese)

[6] Deng Gugang, QinJieping etc. Based on the infrared spectrum data of traditional Chinese medicine medicinal identification research when Jane GuoYi characters 2010, 21 (3) (in Chinese)

[7] Ren Honge, Gao Jie, Ma Yan. New progress of wood identification technology in China. Journal of wood processing machinery, 2007 (4): 38-41. (in Chinese)

[8] V.Vapnik, The nature of statistical learning theory.Springer,1999

[9] C.Cortes, V.Vapnik. Support Vector Networks. Machine Learning, 1995, 20(3):273~297.

[10]S.S.Keerthi, S.K.Shevade, C.Bhattacharyya, et al.A Fast Iterative Nearest Point Algorithm for Support Vector Machine Classifier Design. IEEE transactions on neural networks, 2000, 11(1):124~136.

[11]J.A.K.Suykens. Least squares support vector machines classifier. Neural Processing Letters, 1999, 9(3):293~300.

[12]Yuan Ailing, JiWei, Qian Xu. Manifold regularization based support vector machine (SVM) text classification software in February 2013 (in Chinese)

[13]Wang chunli, Zhang Shijiang, Wang Ting etc. The use SVM to predict potential drug targets proteins Advances in Artificial Intelligence (Volume 3) -- Proceedings of 2011 International Conference on Management Science and Engineering (MSE) 2011

[14]Zhang Ling. Support vector machine (SVM) in The Application of speech recognition Proceedings of The 3 rd 2010 International Conference on Computational Intelligence and Industrial Application (Volume 8)