Hindawi Publishing Corporation Spectroscopy: An International Journal Volume 27 (2012), Issue 4, Pages 253–264 doi:10.1155/2012/878216

Recognition of MIR Data of Semen Armeniacae Amarum and Semen Persicae Using Discrete Wavelet Transformation and BP-Artificial Neural Network

Cun-Gui Cheng,¹ Peng Yu,¹ Chang-Shun Wu,² and Jia-Ni Shou²

Correspondence should be addressed to Cun-Gui Cheng, chengcg123@sina.com

Copyright © 2012 Cun-Gui Cheng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract. Horizontal attenuation total reflection-Fourier transformation infrared spectroscopy (HATR-FT-IR) is used to measure the Mid-IR (MIR) of semen armeniacae amarum and its confusable varieties semen persicae. In order to extrude the difference between semen armeniacae amarum and semen persicae, discrete wavelet transformation (DWT) is used to decompose the MIR of semen armeniacae amarum and semen persicae. Two main scales are selected as the feature extracting space in the DWT domain. According to the distribution of semen armeniacae amarum and semen persicae's MIR, five feature regions are determined at every spectra band by selecting two scales in the DWT domain. Thus, ten feature parameters form the feature vector. The feature vector is input to the back-propagation artificial neural network (BP-ANN) to train so as to accurately classify the semen armeniacae amarum and semen persicae. 100 couples of MIR are used to train and test the proposed method, where 50 couples of data are used to train samples and other 50 couples of data are used to test samples. Experimental results show that the accurate recognition rate between semen armeniacae amarum and semen persicae is averaged 99% following the proposed method.

Keywords: HATR-FT-IR, DWT, BP-ANN, Semen armeniacae amarum, Semen persicae, Recognition

1. Introduction

Traditional Chinese medicine (TCM) is more than just the great contributions to the flourishing and prosperity of the Chinese nation. It represents an important chapter to the annuals of oriental civilization. It has fascinated more and more people in the world [1, 2].

Semen armeniacae amarum (bitter apricot kernel) is a kind of TCM, which has been known for the treatment of diseases for a long time. It is used for treating a variety of coughs and dyspnea, and treating cough caused by wind heat. It is also used for treating dryness syndrome of the intestines with constipation [3].

¹College of Chemistry and Life Sciences, Zhejiang Normal University, Jinhua 321004, China

² Chuyang Honors College, Zhejiang Normal University, Jinhua 321004, China

In Chinese medicine market, it can be difficult to be determined the authenticity of TCM. This involves the scientization of TCM quality appreciation system. While looking for the TCM quality appreciation system, more and more importance was attached to modern instrumental analysis; its application was deemed to the sign of TCM quality appreciation system coming to modernization, as it could offer more and more integrated information on quality than single component.

Fourier transform infrared spectroscopy (FT-IR) can get nearly all of material information about complex systems, so it has become (arguably) the most sophisticated analytical tool in the spectrum analysis and has become indispensable and definitive for many analyses [4–7]. TCM is a complicated system of mixtures, so it is difficult to be explained [8]. It is successful using FT-IR analysis in different family and genus. As two TCM samples are sibling species, they contain similar chemical composition. Therefore, the result is not ideal when MIR analysis was only adopted [9]. How to make use of the large amount of data of absorption spectra from complex system for fast qualitative and quantitative analysis effectively, and to make the information that buried in the MIR overlapping bands and the difference that existed in the infrared absorptions spectra be displayed visually for the identification of those spectra which are similar and complicated have been a goal of analytical chemists. The chemometrics was widely used in many fields because of the combination of the numerical computing technology and information from the equipment.

Wavelet transformation is a more effective signal-processing method than Fourier transform, the transformed results (wavelet factor) of discrete wavelet transformation (DWT) contain more valuable information, which is a relatively effective analysis method in chemometrics. The wavelet transformation is being used in chemistry and its related domains in recent years. Ehrentreich [10] pointed out that the wavelet transformation had been established with the Fourier transform as a data-processing method in analytical chemistry. Most of existing methods in chemistry are based on discrete wavelet transformation. For example, L. M. Shao et al. [11] introduced the wavelet transformation and its applications in respect of photoacoustic spectroscopy, EXAFS spectrum, NMR analysis, and Raman spectrum. In recent years, some researchers also combine the wavelet transformation with other some intelligent technique to analyze the signal of chemistry. For example, R. Tabaraki et al. [12] developed a wavelet neural network (WNN) model in quantitative structure property relationship (QSPR) for predicting solubility of 25 anthraquinone dyes in supercritical carbon dioxide over a wide range of pressures (70–770 bar) and temperatures (291–423 K).

Artificial neural network (ANN) can learn and train the information samples so that it will possess similar memories of human brain, identification capabilities, and implementation of various information-processing functions. It has good self-learning, adaptive, associative memory, parallel processing, and nonlinear conversion capabilities, which avoid complicated mathematical derivation. Even if the samples are deficient and parameters drift circumstances, the output can guarantee to be stable, thus it facilitates the theoretical analysis. Recently, DWT and back-propagation artificial neural network (BP-ANN) have been successfully applied to MIR spectroscopy analysis [13], but few studies have been reported in the MIR-DWT-ANN application to recognition TCM. Therefore, HATR-FT-IR spectroscopy combined with DWT and ANN discrimination method was proposed for the rapid and simple classification of semen armeniacae amarum and semen persicae in this study.

2. DWT and BP-ANN

2.1. DWT

In numerical analysis and functional analysis, DWT is a wavelet transformation that the wavelets are discretely sampled. As with other wavelet transformations, a key advantage it has over Fourier transformations is temporal resolution: it captures both frequency and location information. Based on this advantage, DWT has a huge number of applications in science, engineering, mathematics and computer science. Most notably, it is used for signal coding to represent a discrete signal in a more redundant form, often as a preconditioning for data compression. DWT is originated from the discretization of continuous wavelet transformation (CWT) and the common discretization is dyadic. The CWT of a function or signal, for example, can be defined as

$$W_f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \Psi^* \left(\frac{t-b}{a}\right) dt, \tag{2.1}$$

where $\Psi^*(t)$ denotes the mother wavelet function. The parameters a named as scale parameter and b named as translation parameter are, respectively, used to control the dilation and position of the mother function.

After the dyadic discretization, the function of DWT is accordingly expressed as

$$W_{\text{dwt}}(j,k) = \frac{1}{\sqrt{2^j}} \int_{-\infty}^{\infty} f(t) \Psi^* \left(\frac{t - 2^j k}{2^j}\right) dt, \tag{2.2}$$

where a and b are replaced by 2^j and 2^jk . An efficient way to implement this scheme using filters was developed in 1989 by Mallat. The original signal f(t) passes through two complementary filters and emerges as low-frequency and high-frequency signals. The decomposition process can be iterated, with successive approximations being decomposed in turn, so that a signal can be broken down into many lower-resolution components [14].

2.2. BP Algorithm

Artificial neural network has many models. It can be divided into feed-forward and feed-back based on the network structure. One of the main applications of feed-forward network is identification and classification. There is no strict distinction between the input and output layers of feedback network and we can extract the important characteristics and energy minimization of data after study [15].

When the feed-forward neural-network nodes are all used as the Sigmoid function, one hidden layer is sufficient to arbitrary classification. Figure 1 shows calculation process of the feed-forward artificial neural network. Figure 1(a) shows the first stage which includes choosing the network model, learning rules, studying input and output data (output data aka the target output data), and learning and training the network to get the neural network's node weight and node threshold. The network weights and threshold are determined on a process of constantly adjusting the network weights and threshold by

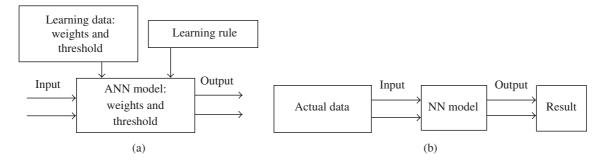


Figure 1: The artificial neural network calculation flowchart.

comparing the error between the output data of the artificial neural network and the target output data until the errors are within the allowable range.

The second stage shows up as shown in Figure 1(b). The output result is generated by inputting the testing data into the network model with the chosen weights and threshold in the first stage.

We use FT-IR-discrete wavelet characteristic vector value of the TCM samples as guide data to train the network. The BP algorithm, which is mature in multilayer feed-forward networks both in theory and applications, is used.

After data preprocessing, the two congeners are mapped to the two nodes in the output layer and the wavelet FT-IR eigenvalues of ten eigenvectors are normalized into the value between 0-1.

Classification algorithm has two steps. The first step is BP network training and the second step is classification of different kinds of categories using the trained BP network (see Figure 2).

Step 1 Is Network Training. There are three layers: the input, the output, and the hidden layer. The input layer is ten normalized eigenvectors from the five regions. The output layer is used for classification with each node of its corresponding to each kind of semen which is two in this case. The hidden layer is the layer between the input layer and output layer. The number of nodes in the hidden layer has to be carefully decided. Less hidden nodes mean higher local minimum and poor fault tolerant ability. However, too much hidden nodes mean long study time and the classification result is not always the best. It is necessary to test repeatedly and choose the best number of hidden nodes. In step 2, the network model designed in step 1 is used to test the ten eigenvectors from the five normalization regions.

Sigmoid function is used as activation function. In order to make the least-square error of the corresponding input samples p minimum, we should study and amend the threshold and weights. The formula of the least square error function can be written by

$$E_p = 0.5 \times \Sigma (t_{pj} - o_{pj})^2,$$
 (2.3)

 t_{pj} is the target output value of sample p in the output layer's j node, that is, the type of the plant, and o_{pj} is the actual output value of sample p in the output layer's j node. Actual output value is calculated from the input layer to the output layer while the adjusted direction of error and weight are from the output layer to the input layer.

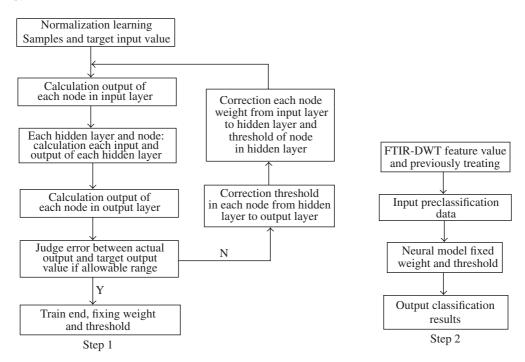


Figure 2: BP algorithm flow chart.

(1) The formula of calculation of the output value o_{pj} of the node j (the output value equal to the input value when node j is the input layer's node) is

$$o_{pj} = f(\text{net}_{pj}) = \frac{1}{(1 + e^{-\sum w_{ji}o_{pj} - \theta_j})},$$
 (2.4)

 w_{ji} is the weight value which connected nodes i and j, θ_j is the threshold of node j. Threshold values can be considered as the weight connecting an output equal to 1 to other nodes, so its adjustment process is the same as w_{ji} . The adjustment of weight w_{ji} is as following.

(2) The formula of amendment weight w_{ji} , which connects the hidden layer's node i to the output layer's node j, is as follows (when j is the output layer's node):

$$\Delta w_{ii}(n+1) = \eta \delta_{pi} o_{pi} + \alpha \Delta w_{ii}(n), \tag{2.5}$$

where η is the learning rate, α is the momentum term, δ_{pj} is the error signal of the output layer's node j. δ_{pj} is calculated as follows:

$$\delta_{pj} = (t_{pj} - o_{pj})o_{pj}(1 - o_{pj}). \tag{2.6}$$

(3) When j is not the output layer nodes, we also used the above weight amendment connected the hidden layer's node i and the hidden layer's node j. But δ_{pj} calculation becomes

$$\delta_{pj} = o_{pj}(1 - o_{pj}) \sum \delta_{pk} w_{kj}, \tag{2.7}$$

 δ_{pk} is the error message between the output with input from node j and node k, w_{kj} is the weight connected node j and k.

3. Experimental Section

3.1. Materials

Semen armeniacae amarum is the kernel of Prunus arminiaca L. (family Rosaceae). They were collected from Linyi (34°22′N, 117°24′E) of Shandong province, Changzhi (35°50′N, 113°01′E) of Shanxi province, Xingtai (36°50′N, 113°52′E) of Hebei province, and Luochuan (35°26′N, 109°13′E) of Shanxi province, China, in July, 2009, and they dried in sunlight, respectively. Semen persicae is the kernel of Prunus persica (L.) Batsch. (family Rosaceae). It was collected from Bozhou (32°51′N, 115°53′E) of Anhui province, Jinhua (28°32′N, 119°14′E) of Zhejiang province, and Changzhi (35°50′N, 113°01′E) of Shanxi province, China, in July, 2009, and dried in sunlight respectively. Eight kernel samples are selected randomly for one time. The samples have been grounded to fine powder in agate mortars to about 200 meshes, respectively.

3.2. Spectral Measurements

The HATR-FT-IR spectra were collected at a resolution of 2 cm⁻¹ scans using a Thermo-Electron (Madison, WI, USA) Nexus 670 FT-IR spectrometer with a room temperature deuterated triglycine sulfate (DTGS) detector, and with a single-bounce HATR (Ge) accessory, spectral range 4000–650 cm⁻¹, resolution 2 cm⁻¹, the cumulative number of scan 64 times. 8.0 mg of predisposed samples was, respectively, placed directly about 3.14 mm² on the center of the Ge crystal of the HATR accessory for measurement. To ensure good contact with the Ge crystal surface, all powder samples were pressed using a pressure tower to provide the same mechanical pressure on all samples. All obtained spectra were autobaseline corrected. No other sample preparation was required. Each species of all samples was measured three times and the averaged spectrum was used for further analysis.

3.3. Data Analysis

HATR-FT-IR of all the samples can be obtained by determination. According to the absorbance value characteristic of absorption peak, we can make the principal component analysis to the data, which are obtained by data copy in different wave bands. Then Matlab software is used to make wavelet transformation to analyze the data further. Using Morlet wavelet, which has a good detection capability of the signal singularity, as the analysis wavelet, one-dimensional DWT is done to the FT-IR spectra of samples under different scales. Then, the differences of HATR-FT-IR spectra of the samples in various

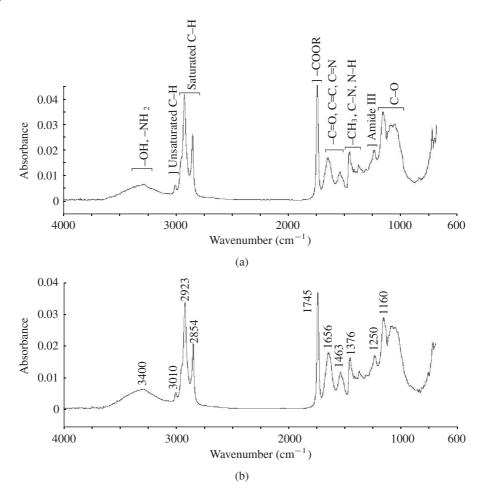


Figure 3: FT-IR spectra of (a) semen armeniacae amarum; (b) semen persicae.

scales are compared. We choose three representative scales to extract features of samples, then use BP-ANN to identify them. In the experiment, we make one-dimensional DWT to the HATR-FT-IR spectra of the samples (they are decomposed into 5 levels). We choose two scales (3 and 4) as the scales to extract the feature vector.

4. Results and Discussion

4.1. FT-IR Analysis

Figure 3 shows the typical HATR-FT-IR spectra of semen armeniacae amarum and semen persicae.

From Figure 3, we notice that the semen armeniacae amarum and semen persicae are similar of absorption peaks in the FT-IR spectra because they belong to the sibling plant kernels. They contain similar chemical composition like hydroxy of cellulose (seed coat), starch, and plant hormones β -sitosterol, and their FT-IR absorption are quite similar. Two samples generated large numbers of sharp

peaks in the FT-IR spectra region (4,000–650 cm⁻¹), which indicates the seeds have a rich chemical composition. Several absorption regions were identified, and the band assignments are labeled in Figure 3. Absorption bands located around 3400 cm⁻¹ correspond to O–H and N–H stretching vibrations that mainly occur from proteins and carbohydrates. The bands around 3010 cm⁻¹ represent unsaturated C–H stretching vibrations that are mainly caused by unsaturated compounds and unsaturated fatty acid ester. The bands around 2923 and 2854 cm⁻¹ represent C–H stretching vibrations that are mainly caused by lipid and carbohydrates. Absorption raised from C–H bending modes was located around 1,200 cm⁻¹ to 1,500 cm⁻¹, but it overlaped with other absorption bands within this region. Three absorption bands located around 1656 cm⁻¹ (mainly C=O strt.), 1463 cm⁻¹ (N–H bend), and 1,250 cm⁻¹ (C–N stret.) were largely due to amide I, II, and III modes of the proteins and lipids, respectively. Absorption bands around 1745 cm⁻¹ correspond to isolated carbonyl group (COOR), indicating ester-containing compounds commonly found in membrane lipid and cell wall pectin. Bands around 1060 cm⁻¹, 1100 cm⁻¹, and 1160 cm⁻¹ in the "fingerprint" region indicate several modes such as C–H bending vibration or C–O or C–C or P–O stretching vibration.

As semen armeniacae amarum and semen persicae are sibling species, they contain similar chemical composition. The FT-IR spectra from the different plant kernels have very close absorbance and are difficult to be distinguished by experience. So we use other methods for further classification.

4.2. Principal Component Analysis (PCA)

Although PCA itself cannot be used as a classification tool, this behavior may indicate the data trend in visualizing dimension spaces. In this paper, 6 samples of each species are randomly selected to make PCA. We have selected 20 absorption peaks in the range of 2000–650 cm⁻¹, then the absorption peaks which we select are tested by PCA, but the results are not satisfactory. The three-dimensional plot of MIR spectra of the two species based on PCA is shown in Figure 4.

Figure 4 shows that the result does not reflect clearly the real relationship of the 12 samples in the relatives, and it is not agreed with our expectations, thus this method is not satisfactory. In order to achieve our desired results, the one-dimension DWT is introduced into our study.

4.3. Feature Extraction of FT-IR in DWT Domain

We will use the DWT to detect the singularity of the curvature curve, so we should choose proper wavelet, which has similar shape to the signal to the analyzed, short compact branch set, and big vanishing moment, as wavelet basis function. Some representative wavelet basis functions include Coiflet, Symlets, Daubechies, Molet, Mexihat, and Meyer. In this paper, we choose Daubechies wavelet as analyzing wavelet.

5 scale wavelet comprssions are performed to the MIR spectral ata. Figure 5 represents the DWT coefficients at five scales. The approximation holds the low-frequency components and the detail holds high-frequency components. Even the 5th level approximation looks very similar with the original MIR spectral data, but it is smoother than the original MIR spectra after noise is removed. We choose representative two levels (scale 3 and 4) to extract their characteristics. Characteristic variable is defined as the energy of spectrum at scale 3 and 4 in the DWT.

According to Figure 5, the differences of DWT coefficients between semen armeniacae amarum and semen persicae are obvious in five regions. In order to effectively extract representative

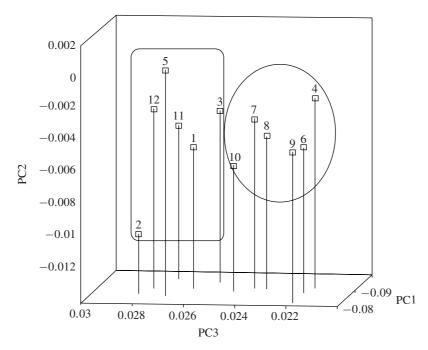


Figure 4: Three-dimensional plot of FT-IR spectra of semen armeniacae amarum (1–6) and semen persicae (7–12) based on PCA.

Table 1: Training and testing results by BP-ANN.

	semen armeniacae amarum (%)	semen persicae (%)
Training samples (50 pairs)	100	100
Testing samples (50 pairs)	100	98

characteristics within two scales of DWT, the spectra in each scale is divided into five representative regions, respectively. Figure 6 is the division diagram of the feature regions. Ten feature regions of two scales in the DWT domain, whose feature values are the spectra energy in the ten feature regions, form the feature vector.

4.4. Identified Network and Application of the Results

After testing, we define the structure of BP network as ten nodes in the input layer, 12 nodes in the hidden layer, and two nodes in the output layer, the error is 0.05, α is 0.8, and η is 0.02.

For the training process, we use ten input layer nodes of BP network structure, followed by normalized ten feature vectors. The output layer nodes are divided into category (1) semen armeniacae amarum, and category (2) semen persicae. The trained network is used to verify the 200 different sample data. The input data is the eigenvector extracted from the wavelet transformation of the original FT-IR. The results are showed in Table 1.

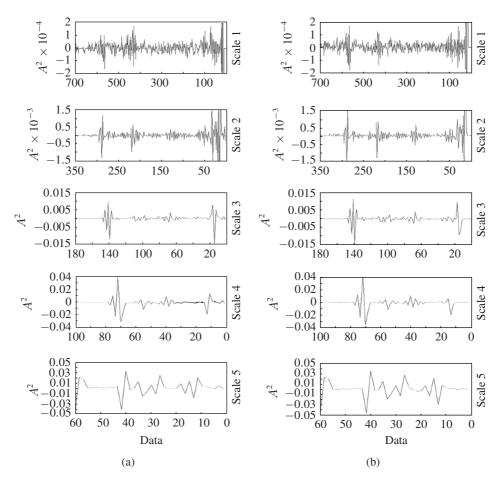


Figure 5: Coefficients of five level approximation/detail of (a) semen armeniacae amarum and (b) semen persicae in DWT domain.

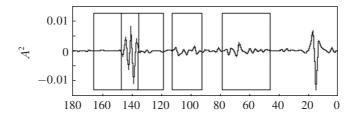


Figure 6: Division of feature region in the DWT domain.

Based on the results in Table 1, basically, the two different types of kernel (the semen armeniacae amarum and the semen persicae) are correctly identified. The eigenvectors of the different kernel that extracted from the wavelet transformation of the MIR have significant difference so that the high accuracy of the classification can be achieved.

5. Conclusion

- (1) Using the sample kernel as material to MIR analysis is an effective way to classify plants because the kernel as an organ of reproduction contains more stable characters than the vegetative organ.
- (2) The technique of HATR has the advantage of measuring directly, does not damage the samples, and has better repeatability than traditional methods such as solvent extraction and direct compression process. So to the plant kernel samples, HATR is a terrific method to determinate.
- (3) Discrete wavelet transformation as an important technique of chemometrics can remain the information unchanged, make variables reduced, data processing process simplified, and make a great help to distinguish similar samples.

The proposed method has a high recognition rate to the MIR data of the semen armeniacae amarum and its confusable derivatives semen persicae by combining BP-ANN with the DWT features of FT-IR of samples.

Acknowledgment

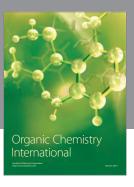
The authors would like to thank Dr. Jian-Hua Chen of Department of Biology, Zhejiang Normal University, China, for identifying the samples.

References

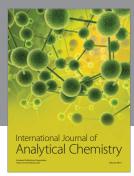
- [1] C. G. Cheng, Y. M. Tian, and C. J. Zhang, "Research of recognition method between semen cuscutae and its sibling plant Japanese dodder seed based on FTIR-CWT and ANN classification method," *Acta Chimica Sinica*, vol. 66, no. 7, pp. 793–798, 2008.
- [2] T. J. Zhang, "Evaluation of Chinese materia medica quality," *Chinese Traditional and Herbal Drugs*, vol. 42, no. 1, pp. 1–9, 2011.
- [3] The Pharmacopoeia Committee of the People's Republic of China, in *Chinese Pharmacopoeia*, vol. 1, pp. 187–188, China Medical Science Press, Beijing, China, 2010.
- [4] C. Cheng, J. Liu, C. Zhang, M. Cai, H. Wang, and W. Xiong, "An overview of infrared spectroscopy based on continuous wavelet transform combined with machine learning algorithms: application to chinese medicines, plant classification, and cancer diagnosis," *Applied Spectroscopy Reviews*, vol. 45, no. 2, pp. 148–164, 2010.
- [5] S. J. Xu, Y. P. Xie, and C. X. Xu, "Routine classification of food pathogens Staphylococcus and Salmonella by ATR-FT-IR spectroscopy," *Spectroscopy*, vol. 26, no. 1, pp. 53–58, 2011.
- [6] L.-H. Wang and W.-C. Sung, "Rapid evaluation and quantitative analysis of eugenol derivatives in essential oils and cosmetic formulations on human skin using attenuated total reflectance-infrared spectroscopy," *Spectroscopy*, vol. 26, no. 1, pp. 43–52, 2011.
- [7] S. Dharmaraj, L.-H. Gam, S. F. Sulaiman, S. M. Mansor, and Z. Ismail, "The application of pattern recognition techniques in metabolite fingerprinting of six different Phyllanthus spp.," *Spectroscopy*, vol. 26, no. 1, pp. 69–78, 2011.
- [8] C. Cheng, J. Liu, W. Cao, R. Zheng, H. Wang, and C. Zhang, "Classification of two species of Bidens based on discrete stationary wavelet transform extraction of FTIR spectra combined with probability neural network," *Vibrational Spectroscopy*, vol. 54, no. 1, pp. 50–55, 2010.

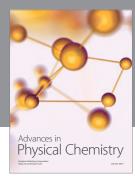
- [9] C. Cheng, J. Liu, H. Wang, and W. Xiong, "Infrared spectroscopic studies of chinese medicines," *Applied Spectroscopy Reviews*, vol. 45, no. 3, pp. 165–178, 2010.
- [10] F. Ehrentreich, "Wavelet transform applications in analytical chemistry," *Analytical and Bioanalytical Chemistry*, vol. 372, no. 1, pp. 115–121, 2002.
- [11] L. M. Shao, X. Q. Lin, and X. Ĝ. Shao, "A wavelet transform and its application to spectroscopic analysis," *Applied Spectroscopy Reviews*, vol. 37, no. 4, pp. 429–450, 2002.
- [12] R. Tabaraki, T. Khayamian, and A. A. Ensafi, "Wavelet neural network modeling in QSPR for prediction of solubility of 25 anthraquinone dyes at different temperatures and pressures in supercritical carbon dioxide," *Journal of Molecular Graphics and Modelling*, vol. 25, no. 1, pp. 46–54, 2006.
- [13] T. Hu, W. Y. Jin, and C. G. Cheng, "Classification of five kinds of moss plants with the use of Fourier transform infrared spectroscopy and chemometrics," *Spectroscopy*, vol. 25, no. 6, pp. 271–285, 2011.
- [14] S. Prabhakar, A. R. Mohanty, and A. S. Sekhar, "Application of discrete wavelet transform for detection of ball bearing race faults," *Tribology International*, vol. 35, no. 12, pp. 793–800, 2002.
- [15] S. B. G. Ha, J. W. Ma, Z. J. Zhou, and Q. Q. Li, "Artificial neural network method used for weather and AVHRR thermal data classification," *Journal of the Graduate School of the Chinese Academy of Sciences*, vol. 20, no. 3, pp. 328–333, 2003.

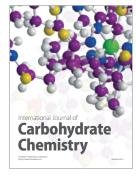
















Submit your manuscripts at http://www.hindawi.com

