A rapid identification of four medicinal chrysanthemum varieties with near infrared spectroscopy

Bangxing Han[1,3], Hui Yan[2], Cunwu Chen[1,3], Houjun Yao[1,3], Jun Dai[1,3], Naifu Chen[1,3]

[1]College of Biological and Pharmaceutical Engineering, West Anhui University, Anhui Province, Lu'an, [2]School of Biological and Environmental Engineering, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, [3]Engineering Technology Research Center of Plant Cell Engineering, Anhui Province, Lu'an, People's Republic of China

Submitted: 01-02-2013 Revised: 14-03-2013 Published: 24-07-2014

ABSTRACT

Background: For genuine medicinal material in Chinese herbs; the efficient, rapid, and precise identification is the focus and difficulty in the filed studying Chinese herbal medicines. *Chrysanthemum morifolium* as herbs has a long planting history in China, culturing high quality ones and different varieties. Different chrysanthemum varieties differ in quality, chemical composition, functions, and application. Therefore, chrysanthemum varieties in the market demands precise identification to provide reference for reasonable and correct application as genuine medicinal material. **Materials and Methods:** A total of 244 batches of chrysanthemum samples were randomly divided into calibration set (160 batches) and prediction set (84 batches). The near infrared diffuses reflectance spectra of chrysanthemum varieties were preprocessed by first order derivative (D1) and autoscaling and was built model with partial least squares (PLS). **Results:** In this study of four chrysanthemum varieties identification, the accuracy rates in calibration sets of Boju, Chuju, Hangju, and Gongju are respectively 100, 100, 98.65, and 96.67%; while the accuracy rates in prediction sets are 100% except for 99.1% of Hangju. **Conclusion:** The research results demonstrate that the qualitative analysis can be conducted by machine learning combined with near infrared spectroscopy (NIR), which provides a new method for rapid and noninvasive identification of chrysanthemum varieties.

Key words: *Chrysanthemum morifolium*, near infrared spectroscopy, rapid detection

Access this article online

Website: www.phcog.com

DOI: 10.4103/0973-1296.137378

Quick Response Code:

INTRODUCTION

The Chinese herbal medicines are the material basis of traditional Chinese medicine (TCM) in disease prevention and treatment, whose quality directly influences the clinical results. During the long-term medical practice, "genuine medicinal material" has become the synonym of high quality herbal medicines and the comprehensive criteria of quality evaluation.[1] Generally, for geoherbs in Chinese herbs, the efficient, rapid, and precise identification is the focus and difficulty in the filed studying Chinese herbal medicines.[2]

Chrysanthemum species for medicine is derived from the capitulum of chrysanthemum; for expelling wind and heat, calming the hyperactive liver, and improving acuity of vision; as food and herbs. Chrysanthemum species as herbs has a long planting history throughout China, culturing high quality ones and different varieties, such as Chuju, Boju, Hangju, and Gongju.[3] They are cultured in Bozhou City, Chuzhou City, Huangshan City, and Tongxiang City, respectively. These different varieties differ in quality, chemical composition, functions, and application.[4] Therefore, chrysanthemum varieties in the market demand precise identification to provide reference for reasonable and correct application as geoherbs.

Currently, the identification of chrysanthemum genuine medicinal material mainly depends on the observation of properties,[4] the chemical compositions,[5] and molecular biology.[6] However, these methods have inevitable shortages, such as being difficult in promotion, complicate analysis, time lasting, and high expense.[7] Accordingly, it is essential to develop an efficient, rapid, and comprehensive method to detect the information of genuine chrysanthemum with low cost.

Near infrared spectroscopy (NIR) is between visible range and mid-infrared spectral region and its spectral

Address for correspondence:
Dr. Bangxing Han, College of Biological and Pharmaceutical Engineering, West Anhui University, Anhui Province, 237 012 Lu'an, People's Republic of China. E-mail: hanbx1978@sina.com

range is 4,000-12,500 cm^{-1}, which is primarily the frequency multiplication and combination frequency absorption of hydrogen-containing radicals like C-H, N-H, and O-H. By scanning samples using NIR, the information of the samples can be obtained, including the chemical compositions, physical and chemical properties, and even biological properties.[8] Together with the identification techniques of computer, stoichiometry, and pattern recognition technology; NIR can rapidly, efficiently, and correctly analyze samples with easy sample processing without reagent or pollution and multi-component detection, so it can be extensively applied in many fields, including TCM, with good results.[9-13] In recent years, due to the development of computer technology and chemometrics softwares, especially the in-depth research and wide application of stoichiometry, NIR has become one of the most eye catching spectroscopic technologies.

The research studies the four most famous chrysanthemum varieties for herbs, analyzes the data of NIR of chrysanthemum samples, takes classification accuracy rate as evaluation parameter, and establishes the disaggregated model of discriminant partial least squares classification algorithm.

MATERIALS AND METHODS

Apparatus

WQF-400N FT-NIR analyzer (Beijing Rayleigh Analytical Instrument Corporation) was used to collect near infrared spectrum, and the range is 10,000-3,500 cm^{-1}. Lead sulfide (PbS) probe was selected to diffuse reflect loading attachments.

Sample collection and preparation

Chuju comes from Liji village, Shiji town, Chuzhou City of Anhui province; Gongju is from Jinzhu village, Beian town, Xi County, Huangshan City of Anhui province; Boju is collected from Qiaodong Medicinal Botanical Garden of Qiaodong town, Bozhou City of Anhui province; Hangju is from Minlian village, Shimen town, Tongxiang City of Zhenjiang province. They were identified as *Chrysanthemum morifolium* Ramat by Prof. Dequn Wang of Anhui College of Traditional Chinese Medicine.

Boju: 30 samples, 20 were randomly selected as calibration set and 10 as prediction set. Chuju: 62 samples, 40 were randomly selected as calibration set and 22 as prediction set. Gongju: 62 samples, 40 were randomly selected as calibration set and 22 as prediction set. Hangju: 90 samples, 60 were randomly selected as calibration set and 30 as prediction set. All samples were naturally dried and shattered into power (40 mesh).

Collect NIR data

Ambient temperature of 20°C, relative humidity of 45%, scanned area of 10,000-3,500 cm^{-1}, 32 times of scanning, resolution factor of 4 cm^{-1}, and light source was 10 W/6 V halogen tungsten lamp with air in the background. To avoid the errors caused by uneven samples, samples cell would rotate at 120 degrees each time the samples were tested and the spectral data took the average value of three-time sampling.

Spectral data preprocessing

The methods taken in preprocessing are standard normal variable transformation, multiplicative scatter correction, first derivative (D1), and second derivative (D2). By comparing the four preprocessing methods on grain sizes, processing environment, and the machine's noise; the best preprocessing method would be obtained.

Partial least squares discriminant analysis modeling

Partial least squares discriminant analysis (PLS-DA) is a regression method based on characteristic variable. As a stable discriminant statistical analysis method, it fits for the use in the situation with many variable numbers and multicollinearity, few samples of observation, and interference noise.

PLS-DA is a partial least squares algorithm based on discriminant analysis and takes Y variable as categorical variable replacing concentration variable. Generally, 0, 1, and 2 represent the category of samples. Similar to quantitative calibration, PLS-DA simultaneously decomposes spectroscopy array and category array, focusing on the function of category information in spectral decomposition, so as to the spectral information most related to sample category; that is, to maximumly abstract the differences between spectrums of varied categories. Therefore, PLS-DA usually can achieve better categorical and discriminant results than principal component analysis (PCA) does. Spectra preprocessing and PLS-DA were carried out by PLS-toolbo × 5.0 (American Eigenvector).

RESULTS

Spectra preprocessing

The original spectra collected by equipment [Figure 1], contains the information related to sample composition and the noise signals produced by different factors. The noise signals would interfere the spectrum information, which is, sometimes, even serious, and thus influences the establishment of calibration model and the prediction of unknown samples compositions and properties. Hence, spectra preprocessing mainly aims at the filtration of spectral noise, screening of the data, optimizing spectral range and eliminating the influences of other factors on data, so as to

lay the foundation of further establishment of calibration model and the precise prediction of unknown samples.

The original spectra were preprocessed by D1, D2, standard normal variate (SNV), and multiplicative scatter correction (MSC), and calibration set was model established by PLS-DA, while prediction set was used for testing the preciseness of model. The results demonstrate that D1 + autoscale is the best, achieving 100% prediction accuracy in calibration set (leave-one-out cross-validation) and prediction set. The spectra preprocessed by D1 + autoscale are seen in Figure 2. Comparing Figures 1 and 2, the preprocessed spectra have many additional peaks in all bands, highlighting the spectral information.

PLS-DA modeling

In order to determine class attribution, the array should be able to describe the samples of some specific category. Generally, a critical value is set to determine the attribution.

Figures 3 and 4 are the PCA distribution of the four chrysanthemum varieties as herbs; first two and first three latent variable (LV) scores, respectively. From the figures, factors distribute disorderly, which are not sufficient to distinguish the four varieties. More LV values are needed.

In PLS-DA analysis, spectral data are alternated to get LV score. Low LV score reflects the information hidden in original spectra, so as to achieve dimensionality reduction. The cumulative contribution rate of LV of the tested spectra is shown as Figure 5. The first five LV contributes a lot, while 6-10 LV contributes less. Ten LV are adopted to establish, when the optimal accuracy is obtained. From Figure 6, the error rate of model prediction decreases along with the increase of LV number. When 10 LV are used, the average category error is the lowest. It means when the principal component selection is 10, correction and

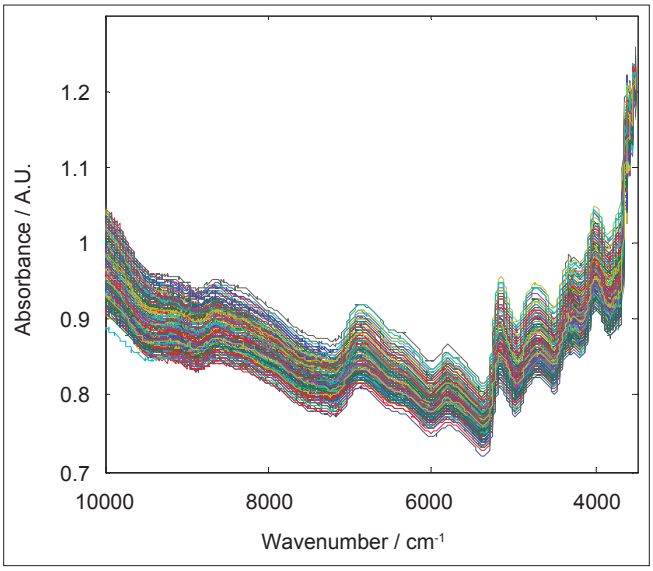

Figure 1: Original spectra

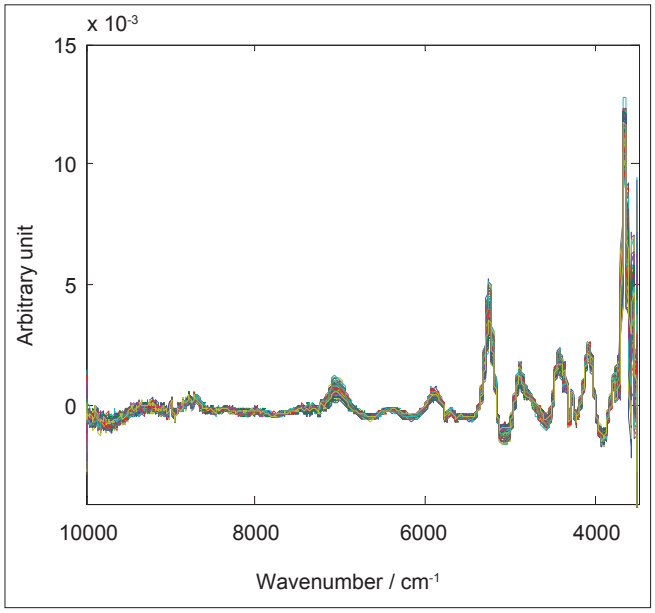

Figure 2: The spectra preprocessed by D1 + autoscale

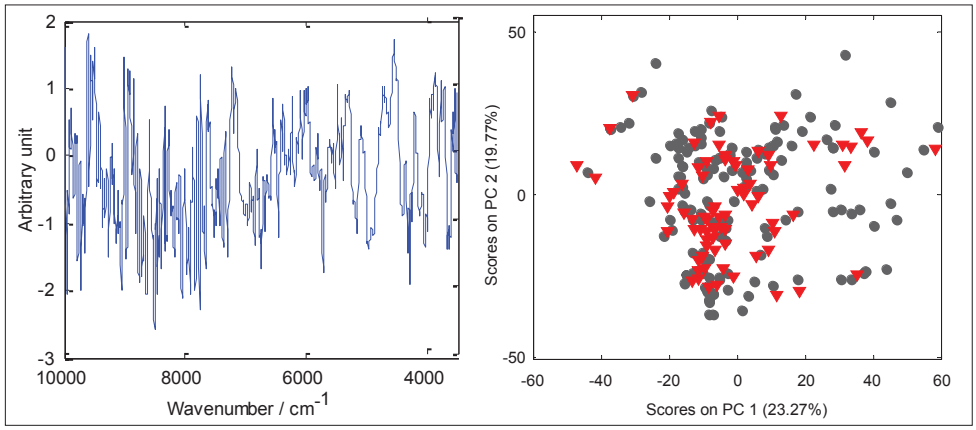

Figure 3: Principal component analysis (PCA) distribution 1

prediction sets reach the best accuracy and classification accuracy is the best. If 10 LV are shown in advance, the information related to the place of production is reflected.

Different wavenumbers affect the LV scores greatly, being vital for judging the variety and origin of chrysanthemum, and also beneficial for understanding mechanism of model discriminant. The relation of variable importance in projection (VIP) scores of origin and wave number is shown in Figure 7. As shown in the figure, information is in the interval of 1,200-1,600. The VIP scores wavenumber of Chuju (Y2) differs from other chrysanthemum varieties mostly. Especially, there is no VIP score in the interval of 1,200-1,300, where others have VIP scores, Boju (Y1) and Chuju (Y2) differs from others in high VIP score wavenumber of 200-280. Gongju (Y3) and Hangju (Y4) have similar VIP score wavenumber of spectral classification, with small differences in the interval of 1,300-1,400. Gongju (Y3) has higher VIP score, while Hangju (Y4) has lower ones.

The different VIP scores may be the basis of model's distinguishing origins, meanwhile, may be because the different scores at diverse wavenumber are caused by the different molecular group vibrations, including the varied varieties and quantity. These demonstrate that origins affect the chemical compositions of chrysanthemum to some degree.

Figure 8 (three-dimensional (3D) distribution) indicates that different varieties have been well-distinguished. As shown in Figure 9, the reciever operating characteristic (ROC) curve area shows that sensitivity has basically reached 100%. The data manifest that the accuracy rates of calibration sets of Boju, Chuju, Hangju, and Gongju are 100, 100, 98.65, and 96.67%, respectively. Except for 99.1% of Hangju, the others reached 100% accuracy rate in prediction sets.

DISCUSSION

TCM, is the valuable experience of disease prevention and treatment for several thousand years in China, as well as the gorgeous treasures of oriental civilization. The identification method, technology and quality control of herbs are important for the modernization and safety of TCM, as well as developing TCM theory.

In the study of the four best chrysanthemums' identification, the accuracy rates of calibration sets of Boju, Chuju, Hangju, and Gongju are 100, 100, 98.65, and 96.67%, respectively; while the accuracy rates of calibration sets are 100% except for 99.1% of Hangju. These data manifest that nonlinear classification model is of high classification accuracy.

The research results demonstrate that the qualitative analysis can be conducted by machine learning combined

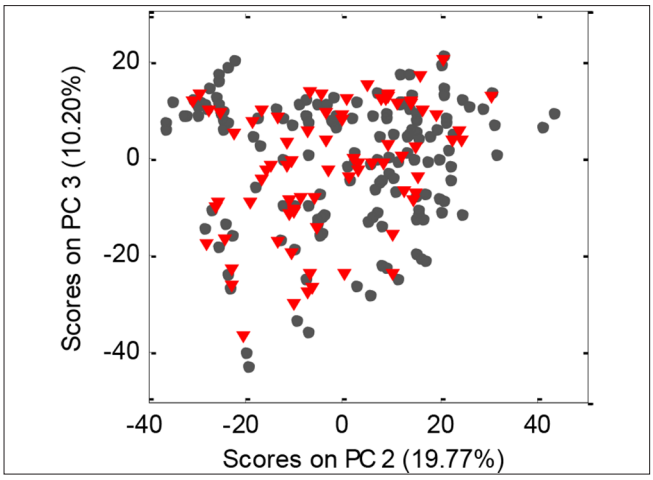

Figure 4: PCA distribution 2

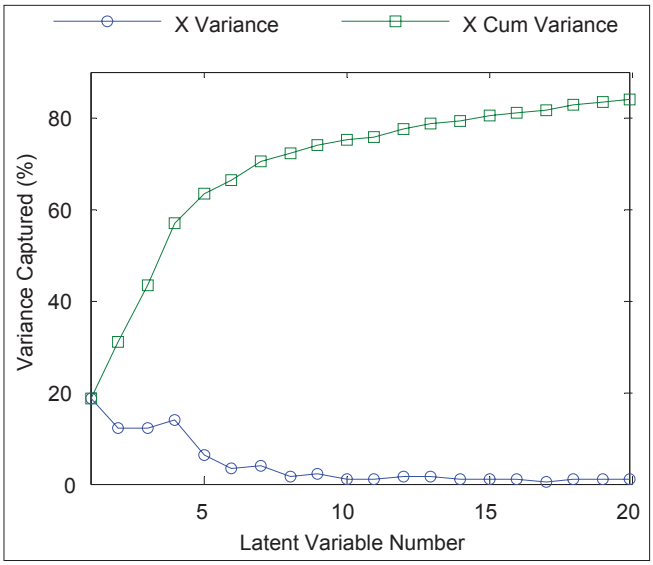

Figure 5: Relationship of latent variable number and variance

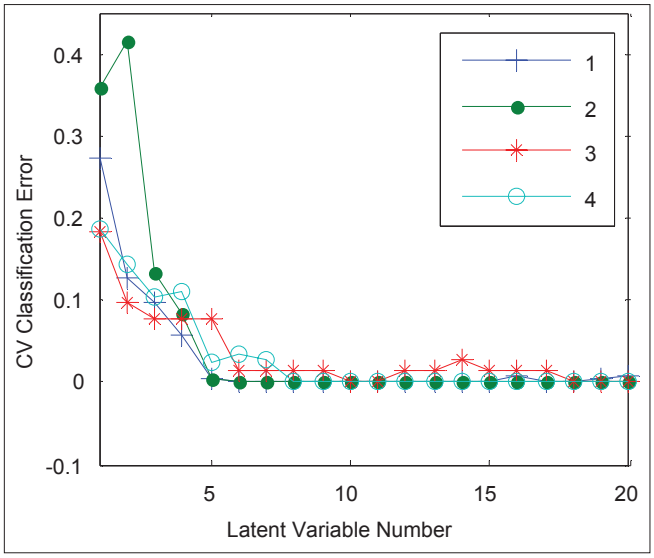

Figure 6: Classification average errors

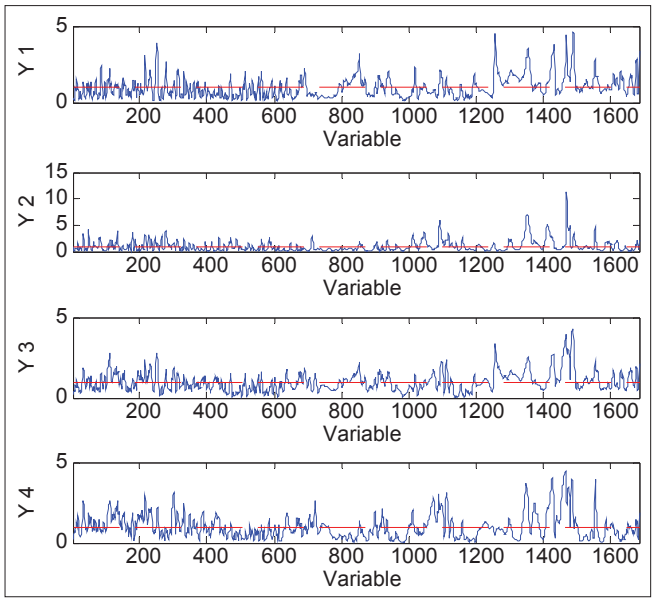

Figure 7: Deviation weighting diagram

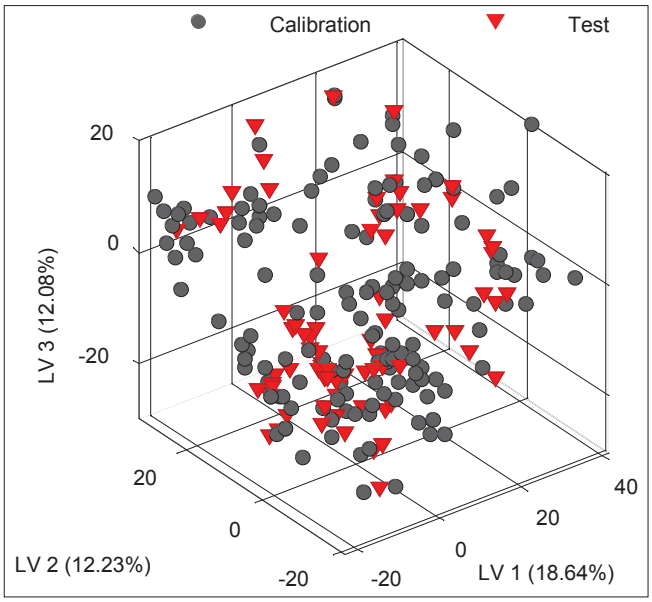

Figure 8: Three-dimensional (3D) distribution

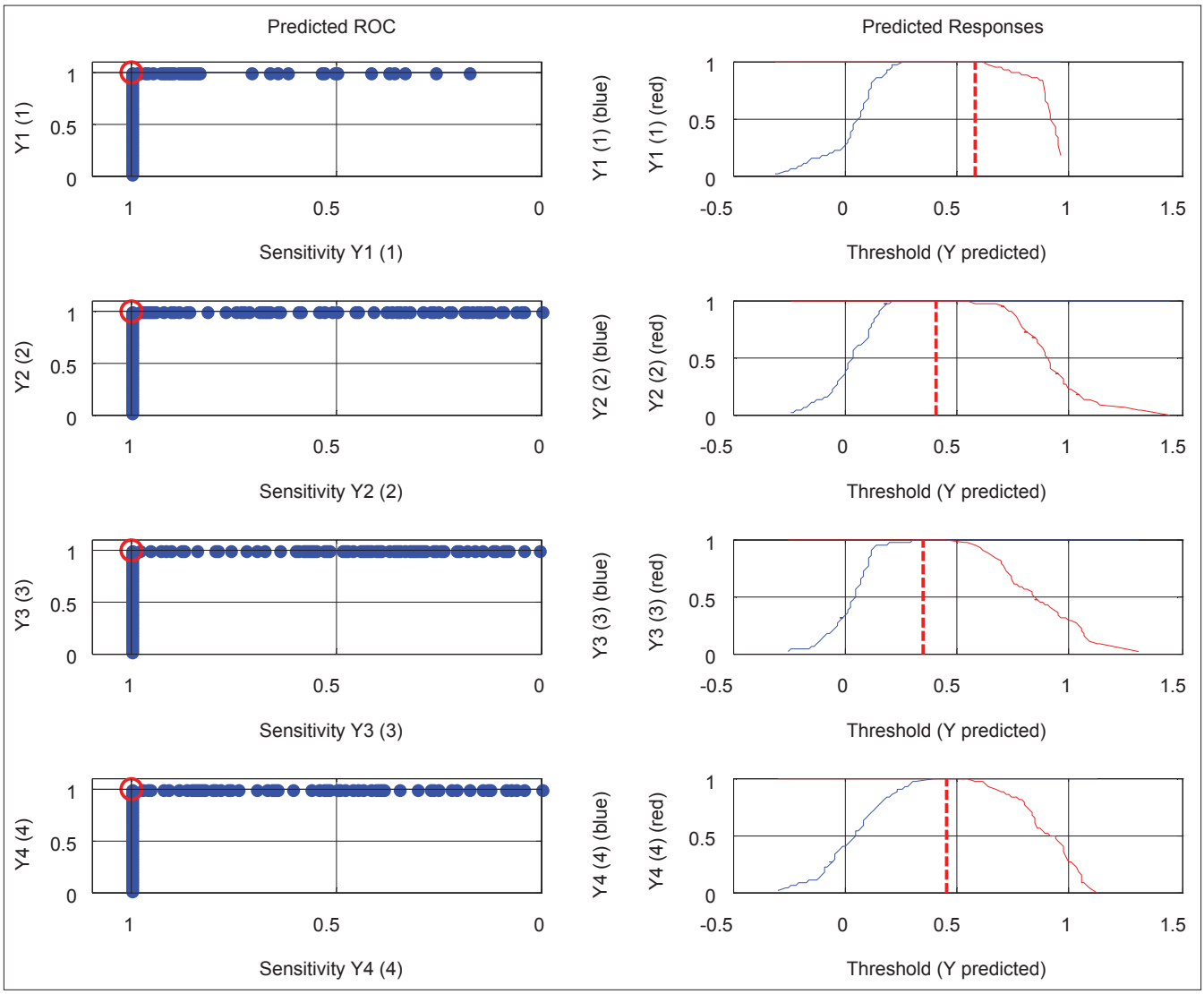

Figure 9: Specificity and sensitivity curves

with NIR, which provides a new means for rapid and noninvasive identification of chrysanthemum. The successive chrysanthemum identification requires suitable data processing and classification methods. The nonlinear classifier, by means of principal components least squares support vector machine, achieves complexity and effectiveness, which proves feasible in the research. With infrared spectrum's fingerprint resistance and pattern recognition, chrysanthemum materials can be rapidly clustering classified. This method is convenient, rapid and accurate, being suitable for the rapid identification of a large number of samples, which is reliable and practical to some degree. It will provide scientific theoretical basis for the identification of materials' authenticity and quality of genuine medicinal materials, with broad application prospect.

This study just conducts the identification research on the four best chrysanthemums. In the future, different mathematic models will be employed to establish pattern recognition database of all chrysanthemum for medicine, so as to distinguish the modes of chrysanthemum for medicine, providing new ideas and methods for the modernization of TCM identification.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No 30901972).

REFERENCES

1. Han BX, Peng HS, Huang LQ. Research advances of Dao-di herbs in China. Chin J Nat 2012;33:281-5.
2. Huang LQ, Guo LP, Hu J, Shao AJ. Molecular mechanism and genetic basis of geoherbs. Zhongguo Zhong Yao Za Zhi 2008;33:2303-8.
3. The Pharmacopoeia Comittee of People's Republic of China. Beijing: Chinese Pharmacopoeia; 2010.
4. Jing DL, Liu W, Xing ZX, Xu Y. A comparative quality analysis of Chrysanthemum morifolium from five defferent production areas. Chin J Mod Appl Pharm 2007;24:467-9.
5. Dong L, Wang J, Deng C, Shen X. Gas chromatography-mass spectrometry following pressurized hot water extraction and solid-phase microextraction for quantification of eucalyptol, camphor, and borneol in Chrysanthemum flowers. J Sep Sci 2007;30:86-9.
6. Yang W, Glover BJ, Rao GY, Yang J. Molecular evidence for multiple polyploidization and lineage recombination in the Chrysanthemum indicum polyploidy complex (Asteraceae). New Phytol 2006;171:875-86.
7. Hung QQ, Pan RL, Wei JH, Wu YW, Zhang LD. Determination of baicalin and total flavonoids in Radix scutellariae by near infrared diffuse reflectance spectroscopy. Guang Pu Xue Yu Guang Pu Fen Xi 2009;29:2425-8.
8. Rodriguez-Saona LE, Khambaty FM, Fry FS, Dubois J, Calvey EM. Detection and identification of bacteria in a juice matrix with Fourier transform-near infrared spectroscopy and multivariate analysis. J Food Prot 2004;67:2555-9.
9. Han BX, Chen NF, Yao Y. Discrimination of Radix Pseudostellariae according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition. Pharmacogn Mag 2009;20:279-86.
10. Chen Y, Xie MY, Yan Y, Zhu SB, Nie SP, Li C, *et al*. Discrimination of Ganoderma lucidum according to geographical origin with near infrared diffuse reflectance spectroscopy and pattern recognition techniques. Anal Chim Acta 2008;618:121-30.
11. Hua R, Sun SQ, Zhou Q, Noda I, Wang BQ. Discrimination of Fritillary according to geographical origin with Fourier transform infrared spectroscopy and two-dimensional correlation IR spectroscopy. J Pharm Biomed Anal 2003;33:199-209.
12. Woo YA, Kim HJ, Ze KR, Chung H. Near-infrared (NIR) spectroscopy for the non-destructive and fast determination of geographical origin of Angelicae gigantis Radix. J Pharm Biomed Anal 2005;36:955-9.
13. Yan H, Han BX, Wu QY, Jiang MZ, Gui ZZ. Rapid detection of Rosa laevigata polysaccharide content by near-infrared spectroscopy. Spectrochim Acta A Mol Biomol Spectrosc 2011;79:179-84.

Cite this article as: Han B, Yan H, Chen C, Yao H, Dai J, Chen N. A rapid identification of four medicinal chrysanthemum varieties with near infrared spectroscopy. Phcog Mag 2014;10:353-8.

Source of Support: Nil, **Conflict of Interest:** None declared.