

PHCOG MAG.: Research Article

Discrimination Of *Radix Pseudostellariae* According To Geographical Origin By FT-NIR Spectroscopy And Supervised Pattern Recognition

Han Bang-xing^{12*}, Chen Nai-fu², Yao Yong³

¹School of Pharmacy, Jiangsu University, 301, Xuefuroad, Zhenjiang, 212013, P.R. China

²Engineering Technology Research Center of Plant Cell Engineering, Anhui Province, Lu'an, 237012, P.R. China

³Xuancheng Jinqian ecological agriculture Co. Ltd. Xuancheng 242000 P.R. China

* Author to whom correspondence should be addressed: e-mail: Dr. B.X. Han (hanbx1978@sina.com)

ABSTRACT

Radix Pseudostellariae is one of the most popular Traditional Chinese Medicine (TCM) for promoting the immune system, treating asthenia after illnesses with a long history in China and some other Asian countries. Rapid discrimination of *R. Pseudostellariae* according to geographical origin is crucial to pharmacodynamic action control. FT-NIR spectroscopy and supervised pattern recognition was attempted to discriminate *R. Pseudostellariae* according to geographical origin in this work. LDA, ANN and SVM were used to construct the discrimination models based on PCA, respectively. The number of PCs and model parameters were optimized by crossvalidation in the constructing model. The performances of three discrimination models were compared. Experimental results showed that the performance of SVM model is the best among three models. The optimal SVM model was achieved when 5 PCs were used, discrimination rates being 100% in the training and 88% in prediction set. The overall results demonstrated that FT-NIR spectroscopy has a high potential to discriminate qualitatively *R. Pseudostellariae* according to geographical origins by means of an appropriate supervised pattern recognition technique.

KEYWORDS: Near infrared spectroscopy, *R. Pseudostellariae*, Pattern recognition, Geoherts.

INTRODUCTION:

Radix Pseudostellariae, the root of *Pseudostellaria heterophylla* (Miq.) Pax ex Pax et Hoffm, known as 'Taizishen', has a long history of use in Asian countries, such as China and Korea, is now classified as a traditional Chinese herbal medicine in common use(1). Studies have demonstrated its multiple pharmacological effects such as anti-oxidation, promoting the immune system, anti-depressant, anti-fatigue, treat night sweating, and asthenia after illnesses activities(2–3). But due to the different ecological factors such as soil, temperature, illumination, moisture content of different regions, the differences in composition and properties among *R. Pseudostellariae* of different geographical origins are observed(4). *R. Pseudostellariae* is widely distributed and cultivated throughout China, i.e. Anhui, Fujian, Jiangsu,

Henan, Zhejiang, Shandong and Guizhou provinces. And for this reason, *R. Pseudostellariae* of different geographical origins have often been confused, and the authenticity of *R. Pseudostellariae*-based medicine is compromised. There are also some difficulties in selecting famous-region *R. Pseudostellariae* for curing diseases. So the discrimination of *R. Pseudostellariae* according to geographical origin is still focused on at present. However, it is not easy to determine its geographical origin by external appearance evaluation. The current discrimination of *R. Pseudostellariae* is restricted to the employment of a few chemical analysis tools such as TLC, UV, GC-MS, HPLC(5–7). Chemical differentiation of *R. Pseudostellariae* is of great importance in science of TCM, especially when the origin is to be verified, thus *R. Pseudostellariae* is a complex mixture of organic as well as inorganic compounds the composition

of which is influenced by many and varying factors. In the holistic theory of TCM, TCM take effects in curing diseases as a whole. *R. Pseudostellariae* is composed of tens major components such as polysaccharides, saponins, flavones, cyclopeptides, amino acids and microelements(4–8). We cannot select only a limited number of specific components as essential screening criteria. Furthermore, these chemical analysis methods are all time-consuming, labor-intensive, expensive, and require large amounts of organic reagent. We must thus conclude *R. Pseudostellariae* cannot be discriminated and identified very well at this moment using only a conventional method. Therefore, a rapid, reliable, accurate, and non-destructive analytical method is essentially required to discriminate the different habitats for the quality control of *R. Pseudostellariae*.

Fourier transformation near-infrared (FT-NIR) spectroscopy, a nondestructive, rapid, cost-effective, and integrity-emphasized method, has important practical utility in identifying and distinguishing the TCM according to geographical origin(9–12). NIR is an optical technique that involves measurements between the visible and the mid-IR spectral region of the electromagnetic spectrum, measures overtones and combinations of fundamental vibrations from the mid-IR region: -OH, -NH, -SH and -CH(13). It presents an intriguing alternative, requiring no sample preparation while offering rapid (seconds rather than minutes), non-invasive and non-destructive sample analysis, moreover it does not require organic reagent, particularly in terms of the use of solid samples(14). Especially, this method follows the integrate principle of traditional Chinese medicine, and it does not lose original natural instinct and compatibility of TCM(9). For these reasons, NIR techniques has found widespread application in TCM and pharmaceutical sciences over the past several years(9–10,15–16). These works mentioned above show that FT-NIR spectroscopy technique has a high potential to analyze quantitatively some active components in TCM.

Supervised pattern recognition refers to techniques in which a priori knowledge about the category membership of samples is used for classification. The classification model is developed on a training set of samples with categories. The model performance is evaluated by means of some samples from a prediction set by comparing their categories predicted with their own true categories. FT-NIR spectroscopy combined with supervised pattern recognition is also used to tackle classification problem(17). Recently, NIR spectroscopy technique has been applied in discrimination of *Fritillary*, *Ganoderma lucidum*, trace element in Italian virgin olive oils according to geographical origin(9–10,18–19). However, No studies have been reported on the applications of FT-

NIR spectroscopy to the identification of the cultivation origins of *R. Pseudostellariae* until now. In present studies, a rapid method for classifying the different geographical origin of *R. Pseudostellariae* samples was first studied by FT-NIR spectroscopy combined with pattern recognition techniques.

In this work, three well-known supervised pattern recognition algorithms were attempted to develop the discrimination models: Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), and Support Vector Machine (SVM). Among them, LDA is linear method, both ANN and SVM are two non-linear methods. Principal component analysis (PCA) was conducted on the NIR data to extract some principal components (PCs) as the inputs of the supervised pattern recognition model. Three spectral preprocessing methods. Standard Normal Variate Transformation (SNV), Multiplicative Scatter Correction (MSC), first-derivative and second-derivative were applied comparatively. The number of PCs was optimized by cross-validation.

MATERIALS AND METHODS

Materials.

All of the samples were collected from the local drug shops. But their fresh roots were collected from four provinces of PR China (i.e. Jiangsu, Henan, Fujian and Guizhou Province), Except sample from Anhui Province was provided by Xuancheng Jinqian Ecological Agriculture Co., Ltd., (Anhui, China). All the samples were dried in a forced-draught oven from Shanghai Jinghong Pharmacy Machine Co. (Shanghai, China) at 100°C for about 10h upon acquisition. Considered the heterogeneities of samples, *R. Pseudostellariae* materials were crushed into powder by a pulverizer made in Wuyi Yili Pharmacy Machine Co. (Zhejiang, China) and controlled below 80 mesh before spectra collection. and these powders sieved were used as for further analysis. All of the roots identified by Professor De-qun Wang of Anhui College of TCM. Voucher specimens are deposited in the Pharmacognosy Laboratory, School of Pharmaceutical, Jiangsu University.

Spectra collection.

The NIR spectra were scanned on a Antaris II Near-infrared spectrophotometer (Thermo Electron Co., USA) with an integrating sphere. The NIR measurements were performed within the region 4000-10,000cm⁻¹. Each spectrum was the average spectrum of 32 scans. and the raw data were measured in 3.856cm⁻¹ intervals, which

resulted in 1557 variables. About 1.0 g of the sample in powder form was individually filled in a glass sample cup. Each sample spectrum was collected three times. The mean of three spectra which were collected from the same sample was used as the further analysis. The temperature was kept around 25 °C, while the humidity was kept at an ambient level in the laboratory. All spectra were recorded as $\log(1/R)$, where R is the relative reflectance.

Spectra preprocessing.

Figure 1a shows the raw spectral profile of *R. Pseudostellariae*. NIR spectra are affected by both the concentration of the chemical constituents and the physical properties of the analyzed product, and the latter properties account for the majority of the variance among spectra while the variance due to chemical composition is considered to be small(20). It is necessary to perform mathematical pre-treatments to reduce the systematic noise, such as baseline variation, light scattering, path length differences and so on. In this study, three spectral preprocessing methods were applied comparatively, and they were Standard Normal Variate Transformation (SNV), Multiplicative Scatter Correction (MSC), first-derivative and second-derivative. SNV is a mathematical transformation method of the $\log(1/R)$ spectra used to remove slope variation and to correct for scatter effects. MSC was used to modify the additive and multiplicative effects in the spectra. First and second derivatives eliminate baseline drifts and small spectral differences are enhanced(21). Compared with results obtained by three preprocessing methods, SNV preprocessing method is as good as MSC, and much better than first and second derivatives. *R. Pseudostellariae* roots are particle solids that bring to easily scatter light in spectra collection. SNV spectral preprocessing methods can remove slope variation and correct light scatter

because of different particle sizes. Therefore, SNV spectral preprocessing method was used in this work. The NIR spectra after SNV preprocessing are showed in Figure 1b.

Software.

All algorithms were implemented in Matlab V7.0 (Mathworks, USA) under Windows XP in data processing. Result Software (Antaris IISystem, Thermo Electron Co., USA) was used in NIR spectral data acquisition.

RESULTS AND DISCUSSION

Principal component analysis.

PCA is often the first step of the data analysis in order to detect patterns in the measured data. Although PCA can only be used as an unsupervised pattern recognition method, this behavior can indicate data trends in a visualizing dimensional space. To visualize the cluster trends of these samples, a scatter plot was obtained using the top three principal components (i.e. PC1, PC2, PC3) issued from PCA.

Figure 2. shows a 3D plot constructed by PC1, PC2, and PC3, and *R. Pseudostellariae* sample is labeled according to its geographical origin (i.e., Anhui, Jiangsu, Henan, Fujian and Guizhou Province). All *R. Pseudostellariae* samples appear clustered along the three principal components axes, confirming the presence of five groups. PC1 can explain 82.2% of the variance, PC2 can explain 9.8% of the variance, and PC3 can explain 3.8% of the variance. The total accumulative contribution rate of variance from PC1, PC2, and PC3 is 95.8%. Therefore, the 3D representation of the PC1, PC2, and PC3 scores for the 300 samples can explain 95.8% raw spectral information from all samples.

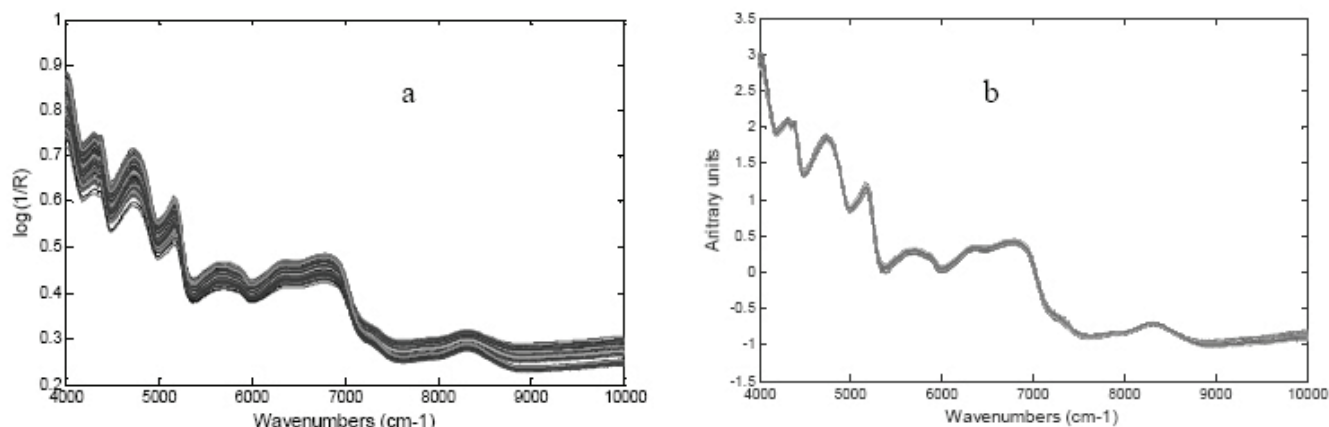


Figure 1. NIR spectra before and after SNV was applied.

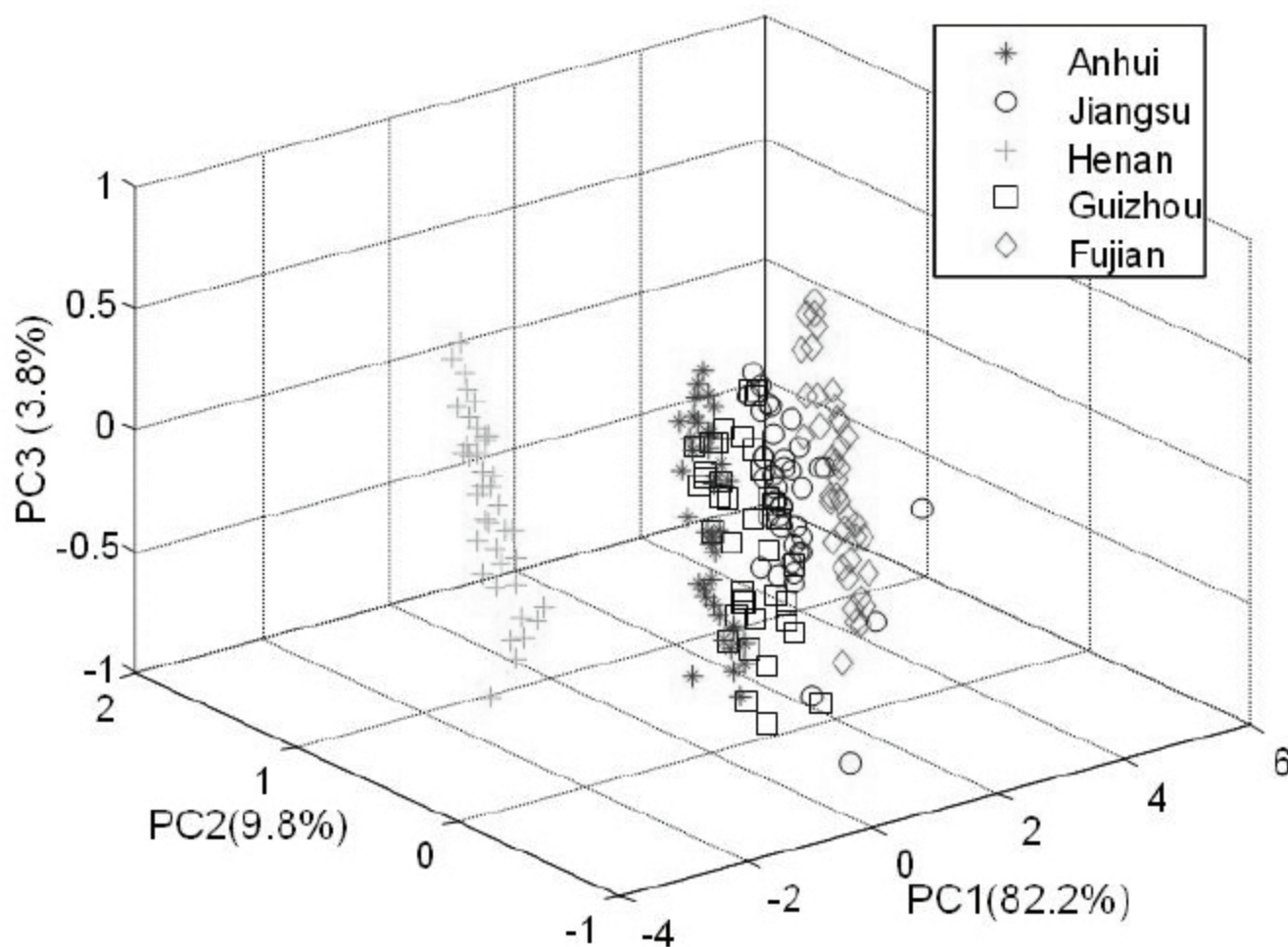


Figure 2. The first three PCs distribution of *R. Pseudostellariae* from different habitations.

Figure 2. shows that there is a separation of five groups in the 3D space represented by the first three principal components. Such good classification in this 3D space could be explained by the chemical background of *R. Pseudostellariae* and PCA methods. *R. Pseudostellariae* can exhibit considerable differences in their own chemical characteristics according to different geographical origins. The differences from chemical characteristics of *R. Pseudostellariae* can be reasonably differentiated in the NIR spectroscopy. Therefore, NIR spectroscopy data can exhibit the cluster trend of *R. Pseudostellariae* samples according to geographical origins by means of PCA.

Discrimination model of supervised pattern recognition.

Geometrical exploration of 3D plot by PCA only gives the cluster trend of samples. Moreover, it is not perfect,

the lack of definite index describing the same differences will lower the credibility of the results. Therefore, actual discrimination of *R. Pseudostellariae* according to geographical origin by means of NIR spectra data and supervised pattern recognition were utilized in the following studies.

In this work, all 300 samples were divided into two subsets. One of subset was called the training set that was used to build model, and other was called the prediction set that was used to test the model reliability. The training set contained 200 samples, and the remaining 100 samples constituted the prediction set. Before developed discrimination model, principal components vectors were extracted by PCA, as the inputs of model. Three supervised pattern recognition algorithms (LDA, ANN, and SVM) were attempted to develop the discrimination model, respectively. The number of PCs was optimized by cross-validation.

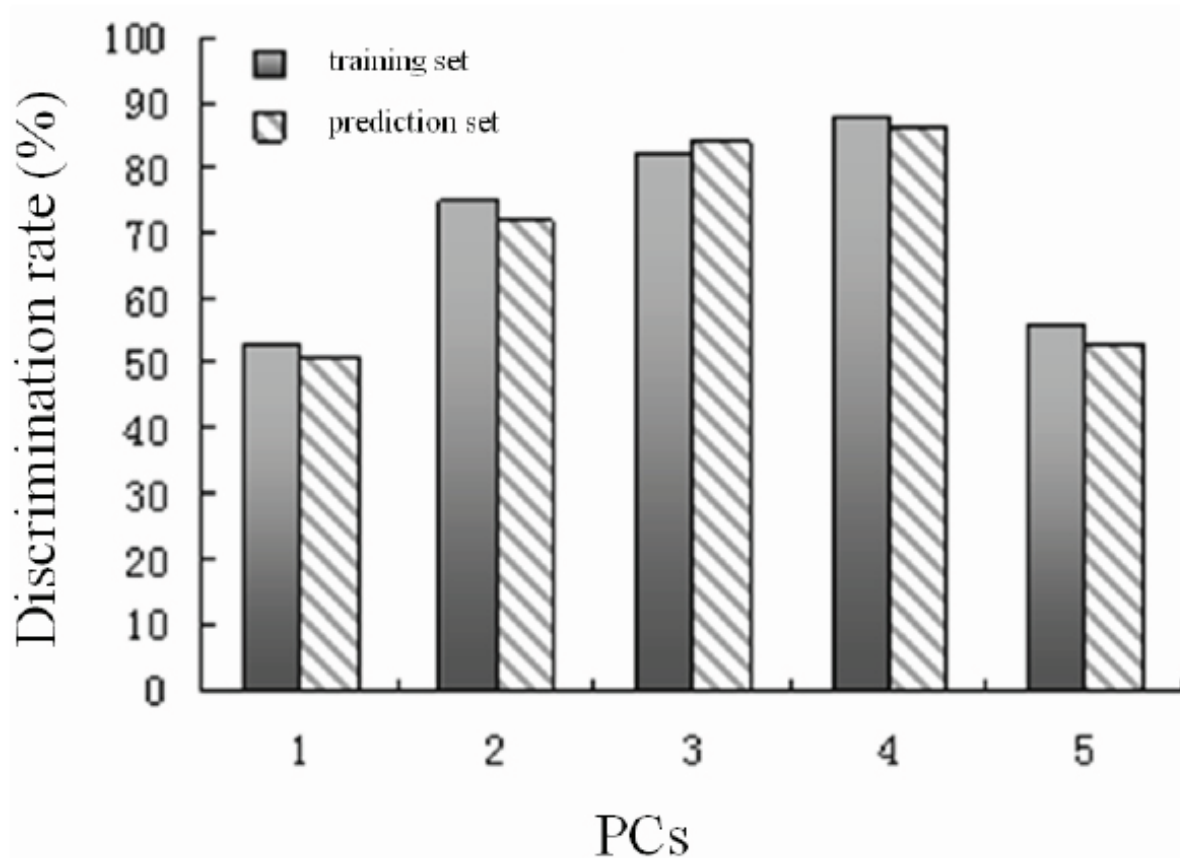


Figure 3. Recognition rate of *R. Pseudostellariae* from different habitations was developed with LDA on the condition of different principal component scores

Linear Discriminant Analysis.

Linear Discriminant Analysis (LDA) is a linear and parametric method with discriminating character. LDA focus on finding optimal boundaries between classes. The number of principal component factors is crucial to the performance of the LDA discrimination model. The discrimination rates by cross-validation were used to optimize the number of PCs.

Figure 3. shows the discrimination rates of LDA model according to different PCs by cross-validation. The optimal number of PCs is according to the highest discrimination rates by cross-validation. As shown in Figure 3, the optimal LDA model is achieved when PCs = 4. The discrimination rate is 88% in the training set and 86% in the prediction set, respectively.

Artificial Neural Networks.

The linear model did not provide a complete solution to the classification problem relatively, Therefore, non-linear

approach such as artificial neural networks (ANN) was used in this work. ANNs are widely used for discrimination. Many researches proved that ANN is a effectual and powerful model in discrimination(22). After the first simple neural network was developed by McCulloch and Pitts in 1943(23), many types of ANN have been proposed. The Back Propagation Artificial Neural Network (BP-ANN) is the most widely used model among ANN models and is used in this study. As an important supervised pattern recognition method, many parameters exert to some extent certain influence on the performance of BP-ANN models. These parameters include the number of neurons in the middle layer, scale functions, learning rate factor, momentum factors, and initial weights.

In our modeling, a three hidden layers BP-ANN was used. These parameters of BP-ANN models were optimized by cross-validation as follows: the number of neurons in the hidden layer was set to 3, the learning rate factor and momentum factor were set to 0.1, the initial weight was set to 0.3, and the scale function was set as

Table 1: Recognition rate of *R. Pseudostellariae* from different habitations was developed with BP-ANN on the condition of different principal component scores

| PCs | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
|---|-----|-----|-----|-----|-----|-----|
| discrimination rate of training set (%) | 51 | 80 | 82 | 97 | 97 | 91 |
| discrimination rate of prediction set(%) | 52 | 81 | 80 | 97 | 98 | 90 |

‘tan h’ function. It is crucial to select the appropriate number of PCs in constructing an ANN model. Table1 shows the discrimination rates of ANN model according to the number of PCs by cross-validation. The optimal ANN model is obtained when 5 PCs are used. The discrimination rate of this BP-ANN model is 97% in the training set and 98% in the prediction set.

Support vector machine.

Support vector machine (SVM) is a supervised learning technique, based on the statistical learning theory, proposed

by Vapnik and Chervonenkis(24), have been successfully applied for mid and near infrared classification tasks, such as material identification(25) and food discrimination(26). The SVM is originated from the classification of two-class problems, in which SVM can be considered to create a ‘optimal’ boundary (hyperplane) of two classes in a vector space independently on the probabilistic distributions between two sets of data for classification. In case the linear boundary in the low dimension input space would not be enough to separate two classes properly, it is possible to create a hyperplane that allows linear separation in the higher dimension feature space. The readers can get more information from the references and tutorials about SVM in detail(27–28).

SVM was attempted in this work. Optimization of parameters is the key step in SVM as their combined values determine the boundary complexity and thus the classification performance(29). There are several classical kernel functions: Gaussian kernel function (is also called RBF kernel function), Polynomial kernel function,

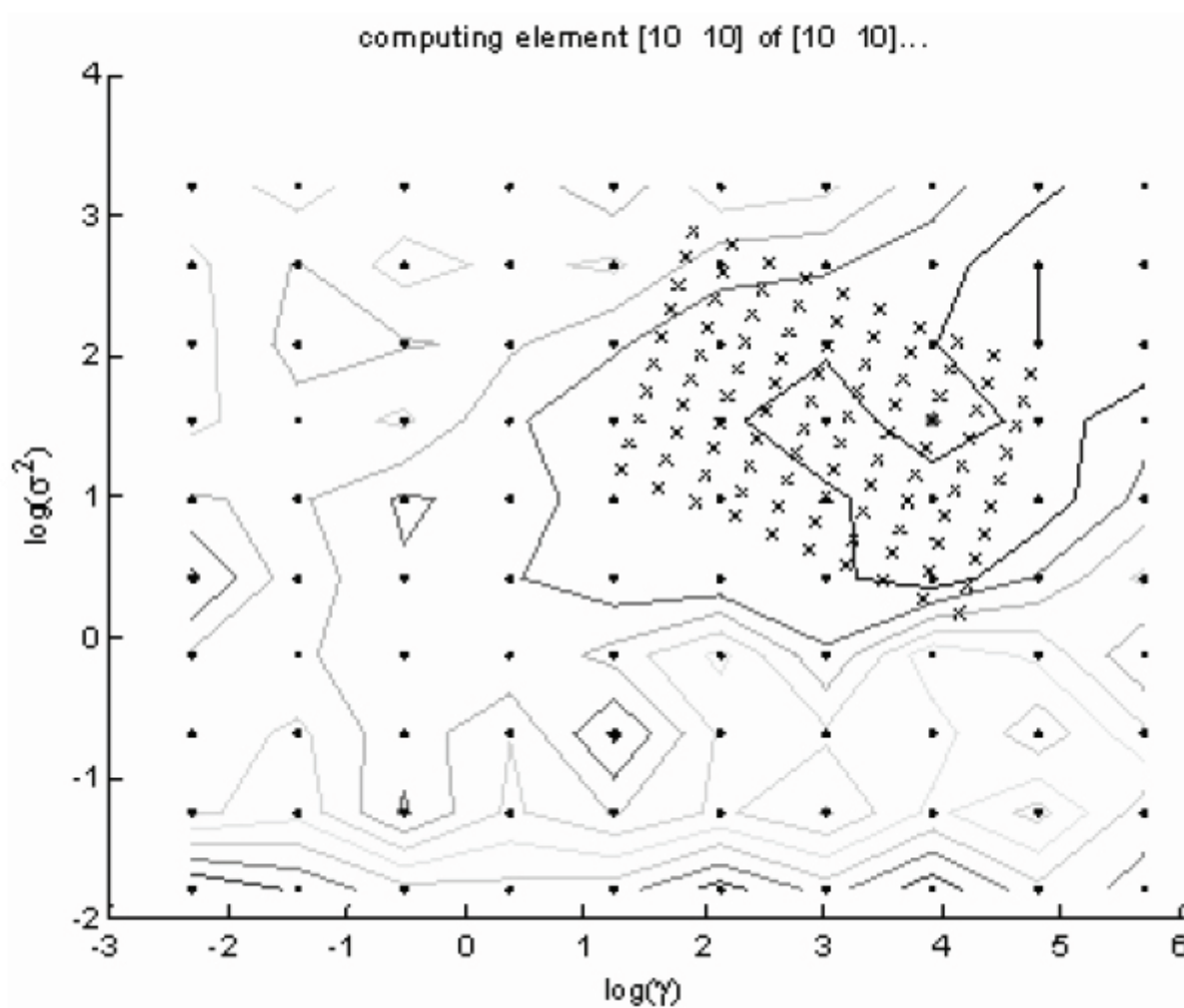


Figure 4. Contour plot of the optimization parameters C and γ of the model using RBF kernel

Selection of kernel function and Linear kernel function. As a non-linear kernel function, RBF kernel function have more capable of handling non-linear relationship between the signals in response to the characteristics and the results than that of Linear kernel function, moreover, it has a terse structure of the function which can reduce the complexity of the process in training. In general, RBF kernel function is the optimal choice, without prior experienced knowledge. To obtain a good performance of SVM model, some parameters in the nuclear function (regularization paramete C and γ) have to be optimized too.

Figure 4 is contour plot of the optimization parameters C and γ of the model using RBF kernel. It can be found that the optimal SVM model is achieved when the optimized parameter values of C and γ^2 , respectively, 800 and 2.5. After two parameters of SVM model were determined, the number of PCs was also optimized by crossvalidation, and the optimal number of PCs was also determined according to the highest discrimination rate by cross-validation. The optimal SVM model is achieved when PCs = 5. The discrimination rates of this optimal SVM model are 100% in the training and 88% in the prediction sets.

CONCLUSIONS

The results described in this research open the possibility of discriminating *R. Pseudostellariae* according to their geographic origin using FT-NIR spectroscopy and supervised pattern recognition, such as LDA, ANN, and SVM models. Table 2 shows the discrimination results from LDA, ANN and SVM models in the training and prediction sets. As shown in Table 2, discrimination rates of LDA model are 88% in the training set and 86% in the prediction set when the PCs = 4, and discrimination rates of ANN model are 97% in the training set and 98% in the prediction set when PCs = 5. Compared with the performances of LDA and ANN, Discrimination rates of SVM model are 100% in the training set and 88% in the prediction set when PCs=5. Seen from total discrimination results in the training sets, the SVM model is the best. But, Seen from total discrimination results in the prediction sets, the linear models are superior to the non-linear model, the ANN model is the best.

In generally, non-linear method is stronger than linear method in the level of self-learning and self-adjust. Thus the results of this research showing that the linear models are superior to the non-linear model accroding to discrimination results in the prediction sets. The season is possibly that the parameters in the nuclear function (regularization paramete C and γ) have to be optimized further so as to construct the better SVM model.

Table 2 Comparison of the identification results from three models

| Models | discrimination rate of training set (%) | discrimination rate of prediction set(%) |
|---------|---|---|
| LDA | 88 | 86 |
| BP-ANN | 97 | 98 |
| RBF-SVM | 100 | 88 |

It can be concluded that FT-NIR spectroscopy technique combined with pattern recognition has high potential to discriminate other TCM according to geographical origin. But, further research will be devoted to ultimately removing the misclassifications. Multi-identification is expected to be the simplest solution to overcome this limitation because the identification probability is high enough to classify.

ACKNOWLEDGEMENTS

This work has been financially supported by the Natural and Science Foundation of Anhui Educational Committee of China (Grant No. KJ2008B330). Innovation Fund for small and medium-sized enterprises of Anhui Province of China (Grant No. cz3401122)

REFERENCES

1. The Pharmacopoeia Committee of People's Republic of China. *Chinese Pharmacopoeia*, (Beijing, 2005) 46.
2. Bauer A. W., Kirby W.M.W., Sherris J.C. Antibiotic susceptibility testing by a standardized single disc method. *Am J Clin Pathol.* **45**:493–496 (1996).
3. Wong C.K., Leung K.N., Fung K.P. The immunostimulating activities of anti-tumor polysaccharides from *Pseudostellaria heterophylla*. *Immunopharmacology.* **28**:47–54(1994).
4. Liu X.H., Tan X.H., Zeng Y.P. Comparison of quality of *Radix Pseudostellariae* from different habitats. *Research and practice of Chinese medicine.* **22**:36(2007).
5. Xu X.Q., Li Q.L., Yuan J.D. Determination of Three Kinds of Chloroacetanilide Herbicides in *Radix Pseudostellariae* by Accelerated Solvent Extraction and Gas Chromatography-Mass Spectrometry. *Chin J Anal Chem.* **35**:206–210 (2007).
6. Han C., Chen J.H., Kang H.N. Determination of PseudostellarinB (cyclic peptide) in Taizishen (*Pseudostellaria heterophylla* (Miq.) Pax) by RP-HPLC. *Chinese Journal of Analysis Laboratory.* **26**:42–45 (2007).
7. Chen Y.Y., Ding Y., Wang W. Determination of Polysaccharide in *Radix Pseudostellariae* Extract by Size-Exclusion High-Performance Liquid Chromatography. *Tsinghua science and technology.* **12**:389–393(2007).
8. Tan N.H., Zhou J. A new cyclopeptid from *Pseudostellaria heterophylla*. *Acta Botanica Yunnanica*, **17**:60–64 (1995).
9. Chen Y., Xie M.Y., Yan Y. Discrimination of *Ganoderma lucidum* according to geographical origin with near infrared diffuse reflectance spectroscopy and pattern recognition techniques. *Anal Chim Acta.* **618**:121–130 (2008).
10. Hua R., Sun S.Q., Zhou Q. Discrimination of *Fritillary* according to geographical origin with Fourier transform infrared spectroscopy and two-dimensional correlation IR spectroscopy. *J Pharmaceut Biomed.* **33**:199–209 (2003).
11. Woo Y.A., Kim H.J., Cho J.H. Discrimination of herbal medicines according to geographical origin with near infrared reflectance spectroscopy and pattern recognition techniques. *Pharmaceut Biomed.* **21**:407–413(1999).

Discrimination Of *Radix Pseudostellariae* According To Geographical Origin By FT-NIR Spectroscopy

12. Woo Y.A., Kim H.J., Ze K.R. Near-infrared (NIR) spectroscopy for the non-destructive and fast determination of geographical origin of *Angelicae gigantis* Radix. *J Pharmaceut Biomed.* **36**:955–959(2005).
13. Rodriguez-Saona, L.E. Khambaty, F.M. Detection and identification of bacteria in a juice matrix with Fourier transform-near infrared spectroscopy and multivariate analysis. *J Food Protect.* **67**:2555–2559 (2004).
14. Kim Y., Singh M., Kays E.S. ear-infrared spectroscopic analysis of macronutrients and energy in homogenized meals. *Food Chem.* **105**:1248–1255 (2007).
15. Sakamoto T., Fujimaki Y., Hiyama Y. Studies on the influence of uniformity of particle size of powder, tapping and sample replacement for diffusion reflectance quantitative NIR spectrometric analysis. *Pharmazie.* **62**:841–846 (2007).
16. Sakamoto T., Fujimaki Y., Hiyama Y. NIR spectroscopic investigation of two fluoroquinolones, levofloxacin and ofloxacin, and their tablets for qualitative identification of commercial products on the market. *Pharmazie.* **63**:628–632(2008).
17. Petri J., Kaunzinger A., Niemoller A. Quality control of tablets by Near Infrared (NIR)-Spectroscopy. *Pharmazie.* **60**:743–746 (2005).
18. Roggo Y., Duponchel L., Huvenne J.P. Comparison of supervised pattern recognition methods with McNemar's statistical test: Application to qualitative analysis of sugar beet by near-infrared spectroscopy. *Anal Chim Acta.* **477**:187–200(2003).
19. Benincasa Cinzia, Lewis John, Perri Enzo. Determination of trace element in Italian virgin olive oils and their characterization according to geographical origin by statistical analysis. *Anal Chim Acta.* **585**:366–370 (2007).
20. Tigabu M., Oden PC. Multivariate classification of sound and insectinfested seeds of a tropical multipurpose tree, *Cordia africana*, with near infrared reflectance spectroscopy. *J Near Infrared Spectrosc.* **10**:45–51 (2002).
21. He Y., Li X.L., Deng X.F. Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model. *J Food Eng.* **79**:1238–1244 (2007).
22. Pardo M., Niederjaufner G., Benussi G. Data preprocessing enhances the classification of different brands of Espresso coffee with an electronic nose. *Sens Actuators B.* **69**:397–403 (2000).
23. McCulloch, W.S., Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics.* **5**:115–133 (1943).
24. Vapnik V.N., Chervonenkis A.Y. Theory Probab. *Appl.* **16**:264–266(1971).
25. Langeron Y., Doussot M., Hewson DJ. Classifying NIR spectra of textile products with kernel methods *Eng Appl Artif Intell.* **20**:415–427 (2007).
26. De La Haba M.J., Fernández Pierna J.A, Fumière O. Discrimination of fish bones from other animal bones in the sedimented fraction of compound feeds by near infrared microscopy. *J Near Infrared Spectrosc.* **15**:81–88 (2007).
27. Scholkopf B, Platt J, Shawe-Taylor J. Estimating the support of a high-dimensional distribution *Neural Comput.* **13**:1443–1447 (2001).
28. Scholkopf B, Smola A., Williamson R. New support vector algorithms. *Neural Comput.* **12**:1207–1245(2000).
29. Devos Olivier, Ruckebusch Cyril, Durand Alexandra. Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometr Intell Lab.* **96**:27–33 (2009).