Comparative Study of Distance Discriminant Analysis And Bp Neural Network for Identification of Rapeseed Cultivars Using Visible/Near Infrared Spectra

Qiang Zou[1], Hui Fang[1], Fei Liu[1], Wenwen Kong[1], Yong He[1]

[1] College of Biosystems Engineering and Food Science, Zhejiang University,

Hangzhou 310029, China

newxfh@gmail.com

Abstract: The potential of visible/near infrared spectra as a method of nondestructive discrimination of various rapeseed cultivars was evaluated, discrimination ability of distance discriminant analysis (DDA) and BP neural network (BPNN) for identification of rapeseed cultivars was shown in this article. The spectral curves ranging from 350 to 2500 nm of rapeseed cultivars were obtained by VIS/NIR spectroscopy, and the principal component analysis (PCA) was applied to perform the clustering analysis. The first 6 principle components (PCs) extracted by PCA were employed as the inputs of DDA and BPNN, respectively, and then two different discrimination models for rapeseed cultivars were built. Forty-five samples from each species and a total of 225 samples from 5 categories rapeseed were extracted. One hundred fifty samples were elected randomly as training sets to set up the training model which was validated by the samples of prediction sets formed by the remaining 75 samples. The result error of BPNN model was set to be ±0.15, and the result indicated that no samples exceeded threshold value, therefore the distinguishing rate was 100%. The result of DDA model displayed that the recognition rate of 100% was achieved. Although the methods mentioned in the presented paper were good approaches for nondestructive discrimination of rapeseed cultivars, DDA model with prediction functions was more intuitive than BPNN and convenient to machine recognition.

Keywords: comparative study, BP neural network, distance discriminant analysis, visible/near infrared spectra，rapeseed cultivars

1 Introduction

Oilseed rape, being one kinds of major oil crops and nectar plants in China, is herbaceous cruciferous crops and is expanding rapidly as a rotation crop following rice [1]. In 2000, rapeseed was the third leading source of vegetable oil in the world, after soy and palm and is the world's second leading source of protein meal. Oil content of oilseed rape depends on rapeseed cultivars greatly, so discrimination of rapeseed cultivars is a very important part of detection for oilseed rape's quality. Randomly Amplified Polymorphic DNAs (RAPDs) were used to discriminate among 23 cultivars of oilseed rape (Brassica napus) in 1994 [2]. In 2007, Principles and advantages of particular marker techniques and application of molecular markers in rapeseed cultivar identification were discussed by Curn V [3]. Because of operational complexity and difficulty for RAPDs and molecular markers, it is very inconvenient to use it. Thus, it is interested in the possibility of on-line non-destructive sorting process for identification of rapeseed cultivars.

Nowadays, visible and near infrared (Vis/NIR) spectroscopy is widely employed as a method for measurement of quality in many fields, such as agriculture, pharmaceuticals, food, textiles, cosmetics, and polymer production industry [4]. Some reports are available on use of visible and near infrared (VIS/NIR) spectroscopy for identification of cultivar [5, 6]. Study of rapeseed using visible and infrared spectroscopy has been reported [7, 8, 9, 10]. However, these papers are just related to physical and chemical analysis, such as oil content, sinapic acid esters assessment.

The performance of a classification or discrimination depends on the separability of the classes. This suggests that the centres of clusters within the measurement space should be sufficiently separated [11]. However, study of identification with linear identification model have shown that linear discriminator do not often yield satisfactory performances [12]. Thus, comparsion of non-linear model such as BP neural network (BPNN) and linear model such as distance discriminant analysis (DDA) for discrimination of rapeseed cultivar was studied. BPNN, one kind of neural network widely used by researchers，consisting of three elements including input layer, hide layer and output layer, are computationally robust with having the ability to learn and generalize from examples to produce meaningful solutions to problems even when the input data contain errors or are incomplete [13]. DDA, which can indentify samples species on the basis of distance between the samples and training set，is a supervised learning technique of statistical pattern recognition. The same as other math models, the critical step of establishing both analysis methods is to get appropriate input variables. Because of large amounts of data, raw spectra data can not to be as inputs of model directly. And excessive input spectra data can result in lager number of iterative training [14], also it leads to the over-regression between the prediction and the training samples [15]. Principal component analysis (PCA) provides a good solution dealing with these problems for us.

The objective of this study was to investigate the potential utilization of VIS/NIR spectroscopy combined with chemometric or statistical methods to discriminat the rapeseed culvers with different attributes. PCA was used to extract principal components (PCs) from a large number of spectra data. Then, BPNN and DDA model were established, respectively. And the quality of two discriminant model was assessed in this paper.

2 Material and methods

2.1 Instrument and system set-up

In this research, the ASD FieldSpe Pro FRTM Spectrometer from ASD (Analytical Spectral Device) was applied. This spectrograph has high sensitivity range from 350 to 2500 nm. The interval of sampling is 1.4 nm and the sensitivity is 3.5 nm range from 350 to 1000nm. From 1000 to 2500 nm, the interval of sampling is 2nm. The software of ASD View Spec Pro, Unscramble V9.6 (CAMO, PROCESS, AS, OSLO, Norway), DPS (data procession system for practical statistics) and SAS (Statistical Analysis System) were used in this study.

2.2 Samples and measurements

In this study, a total of 225 samples of five kinds of rapeseed, as showed in Table 1, were from farm of Zhejiang University, Hangzhou ($30^0$10'N, $120^0$12'E). In order to reduce the error of operation and Environmental differences, the uniform petri dish (diamitor: d = 65 mm, height: h = 1.4 mm) was chosen to bloom rapeseed which covered the bottom of petri dish.

2.3 Pretreatment of original spectral data

The pretreatment methods, such as multiplicative scatter correction (MSC), Savitzky- Golay (SG) smoothing, Standard normal variate (SNV) and so on, usually are used to reduce the error which

contains in original spectral data. Chu, Yuan and Lu [16] uesd the smoothing method of Savitzky Golay (SG) with segment size 3 and default polynomial order zero to decrease the noise. It had been proved that many high frequency could be eliminated [17]. The standard normal variate (SNV) was used to used to remove the multiplicative interferences of scatter, the change of light distance, and particle size [18]. Both pretreatment methods were used in this study. The pre-process and calculations were carried out using a statistical software package named Unscrambler V9.6 for multivariate calibration. To avoid the low signal-noise ratio, the first 100 wavelength values and the last 200 wavelength values were removed, only the wavelength ranging from 450 to 2300 nm was used in this investigation [19, 20].

Table 1. The varieties of the sample in the research

Varieties	Class	From	Number of samples
Qingza2(QZ2)	Qing Za	Zhejiang University	45
Qingza3(QZ3)	Qing Za	Zhejiang University	45
Qingza5(QZ5)	Qing Za	Zhejiang University	45
Qingyou14(QY14)	Qing You	Zhejiang University	45
Qingyou241(QY241)	Qing You	Zhejiang University	45

2.4 Method

Two kinds of methods for clustering, visualization of VIS/NIR spectra and classification the different varieties rapeseed, by PCA, BPNN and DDA model, were showed and compared with each other. The features of spectra data after pretreatment were visualized by principal component analysis in PCs space, then the PCs closely correlative with the categories of these samples, were used as the inputs of an BPNN model and DDA model, respectively, for classification of rapeseed cultivars.

3 Results and discussion

3.1 The reflectance spectra data of rapeseed cultivars

All the achieved spectra data were averaged, and changed to the ASCII code, which was used for building the reflectivity matrix for later use. Fig.1 shows the typical curves of the reflectance spectra of each variety of rapeseed. The trend of all spectra curves was similar, but the differences of wavelength values ranging 450mm from 1000mm among the different rapeseed cultivars were existed after comparing in detail. The mainly reason might be differences in apparent color of different rapeseed cultivars in visible spectrum. At the same time, it can be found that there was not remarkable difference in the other spectral range, some tiny difference can be detected, which make it possible to discriminate the different varieties.

3.2 Data visualization by PCA

The spectra data matrix after pretreatment contained 1851×225 numbers, it was difficult to use it as input of model. Thus, PCA was used to extract PCs to enhance the fetures of spectra data and reduce

dimensionality of matrix. 20 PCs were extracted from 225 spectra curves of rapeseed. Table 2 listed the accumulative reliabilities of PC1, PC2, PC3, …, PC10. It could be find that the accumlative reliabillities of the first 6 principal components was 99.817% and the rest of accumlative reliabillities just was 0.183%. it meant that, 6 PCs contained all the informatins of variables, to some extent. Thus, 6 principal components could provide good discrimination of varieties.

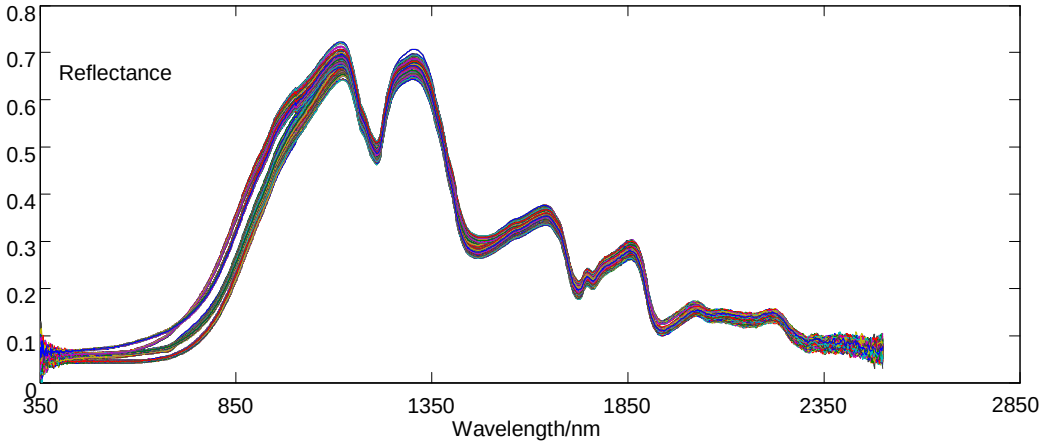

Fig. 1. VIS/NIR reflectance spectra of five varieties of rapeseed

Table 2. PCs and accumulative reliabilities

PC	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Ar(%)	91.213	97.224	98.905	99.658	99.758	99.817	99.838	99.854	99.865	99.871

PC: Principal components; Ar: accumulative reliabilities.

The PCA scores graph shown in Fig.2, which was organized according to the number of the rapeseed, was built using PC1 with PC2. It was easy to draw that the rapeseed was closely clustered and the differences among most kinds of rapeseed were displayed except the QZ3 and QZ5, which were mixed with each other. QZ varieties and QY varieties were separated by CP2-axis. The reason could be that the differences of their variety characteristics were apparent. Although the principal component analysis can qualitatively distinguish between different varieties of rapeseed, but could not give a quantitative identification model.

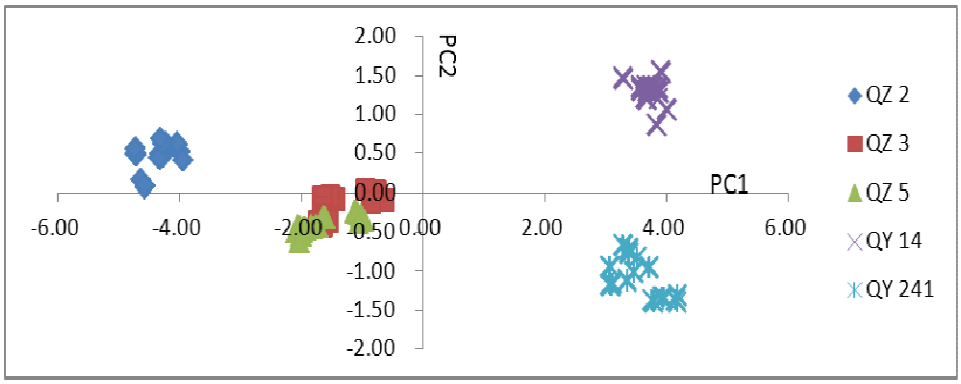

Fig. 2. Scatter plot (PC1 ×PC2) of 225 rapeseed samples

3.3 Classification by BPNN model

One hundred fifty samples from 225 samples were elected randomly as training sets to set up the training model which was validated by the samples of prediction sets formed by the remaining 75 samples. The first six PCs, which can explain the 99.817% of variables, were elected as inputs to build BP neural network model. The numbers from 1 to 5 stood for varieties of five kinds of rapeseed. So, the output value 1 was QZ2 rapeseed, 2 stood for QZ3 rapeseed, 3 was denoted as QZ5 rapeseed, QY14 rapeseed was stood for by 4 and 5 was denote as QY241 rapeseed.

With errors comparison, the best 3-layer neural network structure was determined to build BPNN model after adjusting by many times [21, 22, 23]. The optimal nodes of the hidden layer of neural network was 8 and the maximum number of iterations was 2000. The parameter of sigmoid and the least training speed were set as default values of 0.9 and 0.1, respectively. The residual error was set as 0.00001. The threshold value of prediction error was ±0.15. If error exceeded the threshold, the prediction result was wrong. The prediction error of each sample was shown in Fig.3 for the training set and prediction set. The result showed that the errors were close to 0. Identification result for training set and the prediction set was shown in Table 3, and the recognition ratio of 100% was achieved.

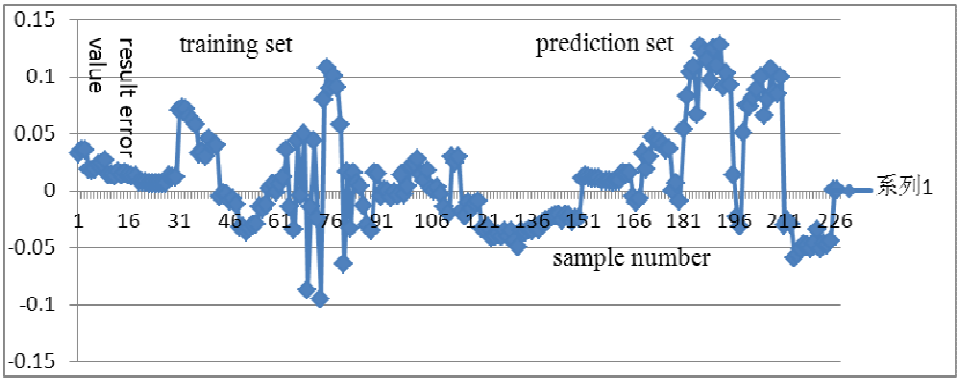

Fig.3. The result error value of prediction results of training set and prediction set by BPNN model.

Table 3. Classification and prediction rate of BPNN model

Class	Classification			Prediction		
	NO.	False NO.	Accuracy rate(%)	NO.	False NO.	Accuracy rate(%)
QZ2	30	0	100	15	0	100
QZ3	30	0	100	15	0	100
QZ5	30	0	100	15	0	100
QY14	30	0	100	15	0	100
QY241	30	0	100	15	0	100
Total	150	0	100	75	0	100

3.4 Clustering with DDA model

DDA, a linear discriminant method, is a multivariate technique, which can allocate new objects to populations previously defined threshold error of recognition and separate objects from distinct populations [24]. In this study, there were five varieties of rapeseed that were classified through distance discriminant analysis. PCs, inputs of BP neural network model, were used as inputs of DDA model. SAS statistical software was used to calculate the spectra data. The discriminant functions of DDA model achieved by SAS were listed in Table 4.

Table 4. Discriminant functions of DDA model

Varieties	Functions
QZ2	$Y_1(x)$=-201.44725-66.17473PC1+204.59991PC2-5.21388PC3+264.62153PC4 -109.86786PC5+314.48234PC6
QZ3	$Y_2(x)$=-290.19263-59.33282PC1+413.00733PC2-685.86068PC3+412.97573PC4 +71.88543PC5+257.70740PC6
QZ5	$Y_3(x)$=-119.06113+0.61504PC1-154.57826PC2-1.45941PC3-362.24558PC4 +737.24938PC5-76.96774PC6
QY14	$Y_4(x)$=-267.17728-8.88928PC1+422.19211PC2-491.43925PC3+285.38541PC4 +93.56473PC5+427.92977PC6
QY241	$Y_5(x)$=-918.38635+135.18458PC1-868.94013PC2+1112PC3-607.09393PC4 -59.34903PC5-931.26578PC6

The more information applied by functions. The greater the absolute value of coefficients was, the greater the contribution of this variable on the function was, and vice versa. For example, the absolute value of PC6 was maximum and the absolute value of PC3 was minimum in discriminant function for QZ2, so variable PC6 make the greatest contribution for function and PC3's was the smallest.

Table 5. Classification and prediction rate of DDA model

Class	Classification				Prediction			
	NO.	False NO.	PP(%)	Accuracy rate(%)	NO.	False NO.	PP(%)	Accuracy rate(%)
QZ2	30	0	100	100	15	0	100	100
QZ3	30	0	100	100	15	0	100	100
QZ5	30	0	100	100	15	0	100	100
QY14	30	0	100	100	15	0	100	100
QY241	30	0	100	100	15	0	100	100
Total	150	0	—	100	75	0	—	100

PP: posterior probabilities.

The rule for discriminating the varieties of prediction samples is that: the probability for a sample with G possible varieties belonging to each category is pi (i = 1, 2,…, G). Assume that pS is maximum among pi (i = 1, 2,…, G), it means pS = max{p1, p2, …, pG}, ($1 \leq S \leq G$). And then, it is concluded that this sample belongs to S variety. In this work, the posterior probability of prediction samples was calculated. Each sample was classified accurately and their posterior probabilities for correct cultivars

were 1.00. The accuracy rate was 100% and result was displayed in Table 5. Thus, the stepwise discriminant analysis model was realistic, correct and valuable.

3.5 Methods comparison

There were two methods for discrimination of rapeseed cultivars using VIS/NIR spectroscopy, BPNN and DDA. Two methods are ordinary used in many fields, such as food, chemical industry, and so forth. BPNN could be used for prediction the content of a component of a substance, while DDA is generally applied for identification of category of unknown samples based on samples with known types. 225 samples from 5 rapeseed categories were analyzed for comparison of two analysis methods.

Because of generalization ability and fault tolerance of BPNN model, a very few error data for samples has little effect on rules between inputs and outputs. But BPNN model for identification of rapeseed cultivars wasted a plenty of training time, this is its shortcoming. For obtaining the rules of inputs and outputs we wish, model parameters, such as residual error, least training speed and maximum number of iterations must be adjusted repeatedly, whereas for DDA model PCs can be finished using SAS statistical software without adjusting time.

Although the recognition ratios of 100% for two methods were acquired, the result of DDA model with SAS was more credible. Recognition ratio of BPNN model depended largely on the threshold of error of prediction value. In this study, the threshold was set to be ±0.15, so recognition ratio of 100% was achieved. If the threshold was set to be ±0.1, the recognition ratio maybe reduce. The reliability of recognition rate without giving authentication probability using BPNN model should be analyzed again. While posterior probabilities, 1.00 for all samples including training set and prediction set, were got by discriminant analysis using SAS and it meant that its recognition ratio's reliability was good.

Finally, although the methods mentioned in this study were good approaches for non-destructive identification of rapeseed cultivars, DDA with prediction functions was more intuitive than BPNN and more convenient to machine recognition.

4 Conclusion

A rapid and non-destructive approach for identification of rapeseed cultivars was presented. And rapeseed with different varieties was classified by two methods, chemometrics and statistical methods, using VIS/NIR spectroscopy. In this study, quantitative analysis for the varieties of rapeseed was made, by combining with SG smoothing, SNV, PCA, BPNN and DDA, relative between reflectance spectra and rapeseed cultivars was built. The BPNN and DDA model displayed an excellent data prediction performance, and the recognition rate of 100% was achieved using two methods. BPNN model should be adjusted repeatedly, compared with DDA, it was not convenient to analyze the spectra data for identification. So DDA was a better method for discrimination of varieties and discriminant analysis with prediction functions was more intuitive than BPNN for identification of rapeseed cultivars.

Acknowledgements

Funding for this research was provided by Science and Technology Department of Zhejiang Province (Project No. 2009C12002), National Natural Science Foundation (Project No. 60802038), National

High Technology Research and Development Program of China (863 Program, Project No. 2006AA10Z234).

References

1. Fei Liu, Fan Zhang, Zonglai Jin, et al. Determination of acetolactate synthase activity and protein content of oilseed rape (Brassica napus L.) leaves using visible/near-infrared spectroscopy. analytica chimica acta. 629, (1, 2), 56--65 (2008)
2. Mailer RJ, Scarth R, Fristensky B. Discrimination among cultivars of rapeseed (brassica-napus L.) using DNA polymorphisms amplified from arbitrary primers. Theoretical and Applied Genetics. 87(6), 697--704 (1994)
3. Curn V, Zaludova J. Fingerprinting of Oilseed Rape Cultivars. Advances in Botanical Research: Incorporating Advances in Plant Pathology, 45, 155--179 (2007)
4. Y.L. Yan, L.L. Zhao, D.H. Han, et al. The Foundation and Application of Near Infrared Spectroscopy Analysis. China Light Industry Press, Beijing (2005)
5. He, Y., Li, X. L., Shao, Y. N. Discrimination of varieties of apple using near infrared spectra based on principal component analysis and artificial neural network model. Spectroscopy and Spectral Analysis. 26(5), 850--853 (2006)
6. Romdhane K., Éric D., Laurent P., et al. The potential of combined infrared and fluorescence spectroscopies as a method of determination of the geographic origin of Emmental cheeses. International Dairy Journal. 15(3), 287--298 (2005)
7. Huang M, He Y, Cen HY, et al. Rapeseed nitrogen status estimation of Vis-NIR spectra based on partial least square and BP neural network. 2007 IEEE international conference on control and automation. (1--7), 2387--2391 (2007)
8. GAN L., SUN X.L.,JIN L., et al., Establishment of math models of NIRS analysis for oil and protein contents in seed of Brassica napus. Scientia Agricultura Sinica. 36(12), 1609--1613 (2003)
9. DING X.X.， LI P.W.， LI G.M. et al. Analysis of erucic acid and glucosinolate in intact rapeseed by fourier transform near infrared diffuse reflectance spectroscopy. Chinese journal of oil crop sciences. 26 (3), 76--79 (2004)
10. Velasco, L., Matthaus, B., Mollers, C. Nondestructive assessment of sinapic acid esters in Brassica species: I. Analysis by near infrared reflectance spectroscopy. Crop Science. 38(6), 1645~1650 (1998)
11. Younes C., Dominique B., Yvette D., et al. Identification of Seeds by Colour Imaging :Comparison of Discriminant Analysis and Artificial Neural Network. Journal of the science of food and agriculture. 71(4), 433--441 (1996)
12. Kim, J., Mowat, A., Poole, P. Linear and non-linear pattern recognition models for classification of fruit from visible-near infrared spectra. Chemometrics and Intelligent Laboratory Systems. 51(2), 201--216. (2000)
13. M. Nasseri, K. Asghari, M.J. Abedini. Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network. Expert Systems with Applications. 35(3), 1415--1421 (2008)
14. Tang YF, Zhang ZY, Fan GQ, Zhu HJ, et al. Identification of official rhubarb samples based on IR spectra and neural networks. Spectroscopy and spectral analysis. 25(5), 715--718 (2005)
15. Y.N. Shao, Y. He, Y.Y Wang. A new approach to discriminate varieties of tobacco using vis/near infrared spectra. Eur Food Res Technol. 224(5), 591--596 (2007)
16. Chu, X. L., Yuan, et al. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique. Progress in Chemistry. 16(4), 528--542 (2004)

17. Y. He , X.L. Li, X.F. Deng. Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model. Journal of Food Engineering. 79 (4), 1238--1242. (2007)
18. Barnes R., Dhanoa M., Lister J. Standard normal variable transformation and de-trending of near infrared diffuse reflectance spectra. Applied Spectroscopy. 43(5), 772--777. (1989)
19. Bao Y.D., Wu Y.P., He Y., Optimal mix forecasting method based on BP neutral network and its application. Journal of Agricultural Mechanization Research. 3, 162--164 (2004)
20. Qi X.M., Zhang L.D., Du X.L.. Quantitative analysis using NIR by building PLS-BP model. Spectroscopy and Spectral Analysis. 23(5), 870--872 (2003)
21. Yin Q., Su X.Z., Xu, Z. A. et al. Analysis on the ultra-spectral characteristics of water environmental parameters about lake. Journal of Infrared and Millimeter Waves. 23(6), 427--435 (2004)
22. Zhao C., Qu H.B., Cheng Y.Y.. A new approach to the fast measurement of content of amino acids in cordyceps sinensis by ANN-NIR. Spectroscopy and Spectral Analysis. 24(1), 50--54 (2004)
23. LI X.L., TANG Y.M., HE Y. et al. Discrimination of Varieties of Paddy Based on Vis/NIR Spectroscopy Combined with Chemometrics. Spectroscopy and Spectral Analysis. 28(3), 578--581 (2008)
24. F.A. Molfetta, A.T. Bruni, K.M. Honório, et al., A structure–activity relationship study of quinone compounds with trypanocidal activity. European Journal of Medicinal Chemistry. 40(4), 329--338 (2005)