Eur Food Res Technol (2010) 230:981–988
DOI 10.1007/s00217-010-1241-7

ORIGINAL PAPER

Application of a hybrid variable selection method for the classification of rapeseed oils based on ^{1}H NMR spectral analysis

Xiaojing Chen · Han Li · Di Wu · Xinxiang Lei · Xiangou Zhu · Anjiang Zhang

Received: 23 October 2009 / Revised: 3 February 2010 / Accepted: 7 February 2010 / Published online: 3 March 2010

Abstract ^{1}H nuclear magnetic resonance (NMR) spectroscopy was utilized to distinguish the brands of rapeseed oils. As there are more than four hundreds of NMR variables which can cause the discrimination model redundancy, it is necessary to do effective variable selection. Successive projections algorithm (SPA) executed on the full spectrum only improved a few correct answer rate (CAR) and Cohen's kappa coefficient (K) compared to full spectrum-least-square support vector machine (LS-SVM) model. The better results of uninformative variable elimination (UVE)-based SPA calculation show that it is necessary to do UVE before SPA. Because the cutoff threshold selection in UVE algorithm using an artificial random noise cannot obtain the optimal results, we applied simulated annealing (SA) algorithm to estimate the optimal cutoff threshold. The discrimination results show that UVE-SA did better works than conventional UVE. Only 13 variables were obtained by UVE-SA-SPA while the conventional UVE-based SPA selected 77 variables. The best 97.5% CAR and K of 0.967 result of UVE-SPA-LS-SVM model show that it is feasible to distinguish different brands of rapeseed oils using ^{1}H NMR spectra. It shows that a combination of SA, UVE, and SPA is effective method for the classification of rapeseed oils. Final result shows that all acyl chains, linolenyl and linoleyl chains, and triglycerides were most important for the classification.

Keywords ^{1}H nuclear magnetic resonance (NMR) · Uninformative variable elimination (UVE) · Successive projections algorithm (SPA) · Simulated algorithm (SA) · Variable selection

X. Chen (✉) · H. Li · X. Zhu
College of Physics and Electronic Information Engineering, Wenzhou University, 325027 Wenzhou, People's Republic of China
e-mail: chenxj10@yahoo.cn

D. Wu
College of Biosystems Engineering and Food Science, Zhejiang University, 310029 Hangzhou, People's Republic of China

X. Lei
Department of Chemistry, Wenzhou University, 325027 Wenzhou, People's Republic of China

A. Zhang
Chengdu Institute of Biology, Chinese Academy of Sciences, 610041 Chengdu, People's Republic of China

Introduction

^{1}H nuclear magnetic resonance (NMR) is the application of NMR spectroscopy with respect to hydrogen-1 nuclei within the molecules of a substance, in order to determine the structure of its molecules. In sample analytes where natural hydrogen (H) is used, practically all of the hydrogen from the analyte consists of the isotope ^{1}H (hydrogen-1; i.e., having a proton for a nucleus). Some researchers did quantitative analysis using NMR spectroscopy [1–7]. However, the calibration methods they used were almost based on the whole spectral range. They did not take into account that some spectral regions might not contain useful information about the chemical variations in the samples and should be eliminated.

Because the ^{1}H NMR instrumentations usually have a high resolution, their obtained ^{1}H NMR spectral data sets often contain hundreds of variables. Thus, with these so many variables and hundreds of samples, spectral data are too complicated to be calibrated directly. The

calibration process is time-consuming and not convenient to fulfill the high-speed features of spectroscopy. However, the calibration methods the above-mentioned researches used were almost based on the whole spectral range. They did not take into account that some spectral regions might not contain useful information about the chemical variations in the samples. Thus, it is important to select specific variables which contain useful information. More stable model with superior interpretability can be generated, and this can produce the lowest prediction error.

Developing a calibration model involves a decision on which wavelengths to be used to establish an optimal model. Some wavelengths or wavelength bands may contain useless or irrelevant information like noise and background which can worsen the predictive ability of the model. The elimination of irrelevant variables can predigest calibration modeling and improve the results in terms of accuracy and robustness. Better calibration model may be obtained by selecting characteristic information such as sample-specific or component-specific variables instead of the full spectra. Recently, both theoretical [8] and experimental evidence [9] have proved that characteristic wavelengths instead of full spectra can improve quantitative results [10, 11]. Specific regions can generate more stable models with good interpretability [12]. Thus, it is important to select specific variables which contain useful information. More stable model with superior interpretability can be generated, and this can produce the lower prediction error.

Successive projections algorithm (SPA) is a novelty variable selection algorithm in order to solve the collinearity problems. It selects variables with minimally redundant. SPA employs a simple projection operation in a vector space to select subsets of variables with minimum of collinearity [13]. SPA can provide more reproducible results than genetic algorithm [14]. However, the SPA operation is time-consuming when the whole ^{1}H NMR spectra which usually have hundreds of variables were calculated. Moreover, its selected variables may be with low signal noise ratio (S/N) or useless for multivariate calibration, which can affect model precision of prediction [15]. Thus, it might be possible to improve the calibration model when SPA is followed to select variables which have minimum redundant information from the informative variables with high S/N.

Uninformative variable elimination (UVE) is another novelty variable selection method based on the stability analysis of partial least-square (PLS) regression coefficient [16]. UVE can eliminate the variables which have no more informative variables for modeling than noise. Employing the selected variables by UVE for modeling can avoid a model over-fitting and usually improve its predictive ability. In UVE process, wavelength variable whose absolute number of stability value is larger than a cutoff threshold is retained. The cutoff threshold is evaluated by the maximum of absolute stability value of an artificial random variable matrix with very small amplitude [15–17]. However, there is a stochastic way for evaluating the cutoff threshold, which is not easy to find the optimal cutoff threshold and results in a defect in the UVE algorithm. Therefore, we proposed the simulated annealing (SA) algorithm to estimate the optimal cutoff threshold of UVE, instead of using an artificial random noise.

Rapeseed (*Brassica napus*), also known as canola, is a bright yellow flowering member of the family *Brassicaceae*. Rapeseed is grown for the production of animal feed, vegetable oil for human consumption, and biodiesel. Rapeseed oil contains both omega-6 and omega-3 fatty acids in a ratio of 2:1 and is second only to flax oil in omega-3 fatty acid. Rapeseed oil's proponents claim that it is one of the most heart-healthy oils and has been reported to reduce cholesterol levels, lower serum triglyceride levels and keep platelets from sticking together. Recently, rapeseed oil consumption increases quickly in the Chinese rapeseed oil market. However, to make enormous profits, some factories produce inferior rapeseed oil that contains insufficient or superfluous nutritional contents. Others mix brands of rapeseed oil of different qualities, such as by packing a rapeseed oil belonging to a conventional brand in the packaging of a high-grade brand. These behaviors badly infringe on the rights and interests of consumers. These illegal behaviors could be avoided if a fast and accurate analytical method were in place to determine the brands and quality of rapeseed oil. Consumers conventionally judge the quality of a rapeseed oil by color and smell which are subjective and less accurate. A chemical analysis and a physical property assessment are routinely performed in the commercial trading of rapeseed oil. However, the processes of chemical methods are complex, destructive, and professional. In practice, only a small number of samples can be measured. It always takes long time to obtain the testing result for one sample from the preparation to the end [18]. The use of chemical reagents is a problem against to the sample safety and low cost. Therefore, the need exists for a rapid and nondestructive method that is suitable for screening rapeseed oil, at least for authenticity information such as brand and quality.

In this study, we studied the feasibility of the discrimination of rapeseed oil using ^{1}H NMR spectroscopy. A hybrid variable selection algorithm composed of UVE-SA and SPA was utilized to select optimal ^{1}H NMR spectral variables. UVE-SA was operated before SPA procedure to improve the discrimination results.

Theory and algorithms

Successive projections algorithm (SPA)

SPA is a forward variable selection method for multivariate calibration [13, 19, 20], its purpose is to select wavelengths whose information content is small collinearity. The main procedures are summarized here. First, the maximum number of variables N is set. Subsequently, a start vector (the first wavelength $x(j)$) is chosen, and calculate the projection of on the subspace orthogonal to the remaining wavelengths, the wavelength of higher projection is selected and becoming the new starting wavelength, so this new starting is small collinearity to the previous wavelength. This step is iterated until the number of selected wavelengths reaches the optimal number of variables (K). In the SPA, the optimal initial variable and number of variables can be determined on basis of the smallest root mean squared error of prediction in validation set of MLR calibration. The details of SPA could be found in the literatures [13, 19, 20].

Conventional uninformative variable elimination (UVE)

The UVE method was put forward in reference [16]. In the conventional uninformative variable elimination method, PLS regression is performed on instrumental response data X and property values (y) of calibration, the optimal latent variable number is calculated firstly, and then, a noise matrix $N(n \times p)$ with very small amplitude (e.g. 10^{-10}) is generated and append to the X matrix, forming an extended matrix $Z(n \times 2p)$ (with twice as many variables as the X matrix). Finally,the PLS model is computed on the matrix Z, and the regression coefficient matrix $b = [b_1, \ldots, b_p]$ of model is calculated through a leave-one-out validation [16]; the reliability of each variable is quantitatively measured according to its stability. The stability of variable j can be calculated as:

$$s_j = \text{mean}(\beta_j)/\text{std}(\beta_j) \quad (1)$$

In this equation, mean(β_j) and std(β_j) are the mean and standard deviation of the regression coefficients of variable j. The larger the absolute stability, the more important the corresponding variables is. It is obvious that any variables whose stability is less than that of noise variables should be known as uninformative and be eliminated. Usually the cutoff threshold is calculated as:

$$\text{cutoff} = k \times \max(\text{abs}(S_{\text{noise}})) \quad (2)$$

In the definition, k is an arbitrary value, in our work, we used $k = 0.9$. The detailed descriptions of UVE are given in Ref. [16, 17].

Simulated annealing (SA) algorithm

SA algorithm is a simulation of the annealing process used for metals, its potential as a general combinatorial search method was put forward by Kirkpatrick et al. [21]; it was originally developed as a simulation model for physical annealing process, and hence, it is referred to as simulated annealing. SA algorithm belongs to the class of iterative improvement strategies, which allows occasional worsening moves so that these can prevent the algorithm freezing in local optimum. The acceptance criterion of the worsening s move is determined by probability:

$$p(\Delta F) = \exp\left(\frac{-\Delta F}{T}\right) \quad (3)$$

where ΔF is the change in the energy value from one point to the next, T is temperature (control parameter), this equation is popularly referred to as the Metropolis criterion [22]. In the SA algorithm, if temperature T is lowered sufficiently, no further changes in the solution space are possible. The detailed descriptions of SA algorithm are given in Ref. [21, 23, 24]. Therefore, due to the merit of SA algorithm, it was employed to search for the optimal cutoff threshold for UVE.

Simulated algorithm-based uninformative variable elimination

On the combination of the UVE method and SA algorithms (UVE-SA), a new method is developed for variable selection in NMR modeling. In the UVE-SA method, the stability of each variable was calculated in PLS model as the first step. Then, instead of adding artificial random noise variables to the original data matrix for the estimate the cutoff threshold, SA algorithm is employed to select optimal cutoff threshold. Finally, according to obtained cutoff threshold, the variables are selected for the further calculation. On the aspect of cutoff threshold value selection, the obvious advantage of the UVE-SA method excludes the shortcoming of selecting the cutoff threshold value experientially.

Chemometric calibration methods

LS-SVM is an optimized algorithm based on the standard support vector machine. As giving a good performance under general smoothness assumptions on handling the nonlinear relationships between the spectra and target attributes, RBF kernel was used in this study. Grid-search technique was applied to find out the optimal parameter values which include regularization parameter gam (γ) and the RBF kernel function parameter sig^2 (σ^2). In this study, these parameters were optimized with values of γ in the

range of 2^{-1}–2^{10} and σ^2 in the range of 2–2^{15} with adequate increments. These ranges were chosen from previous studies where the magnitude of parameters to be optimized was established. For each combination of γ and σ^2 parameters, the root mean square error of cross-validation (RMSECV) was calculated and the optimum parameters were selected when produced smaller RMSECV. The details of LS-SVM description could be found in the literature [25].

Model evaluation standard

In this study, the performances of all the established spectral models were evaluated in terms of two parameters, namely correct answer rate (CAR) and Cohen's kappa coefficient. CAR is the common used simple percent calculation. Cohen's kappa coefficient (K) is a statistical measure of inter-rater agreement for qualitative (categorical) items [26]. It is generally thought to be a more robust measure than simple percent agreement calculation. In this study, K was calculated between referenced brand and classified brand. Landis and Koch gave the following interpretation: $K < 0.00$ means no agreement, K between 0.00 and 0.20 means slight agreement, K between 0.21 and 0.40 means fair agreement, K between 0.41 and 0.60 means moderate agreement, K between 0.61 and 0.80 means substantial agreement, and K between 0.81 and 1.00 means almost perfect agreement [26].

Experiment and calculation

Rapeseed oil samples

In the present work, four brands of rapeseed oil were prepared for the experiment. All of these brands are popular in Chinese markets. Finally, 120 samples of rapeseed oil samples were obtained. Each brand has thirty samples. In order to obtain a 2:1 division of calibration/prediction spectra, the four samples of every six samples are selected into the calibration set. Finally, the calibration set contains 80 samples, and other 40 samples constitute the prediction set.

^{1}H NMR spectra measurement

Hundred microliter of the bulk oil was dissolved in 600 μL of deuterated chloroform, shaken in a vortex, and placed in a 5-mm NMR capillary. The ^{1}H NMR spectra were recorded on a Bruker AVANCE 300 spectrometer operating at 300.13 MHz for the proton nucleus at 300 K. The experiments were carried out to obtain ^{1}H NMR spectra with the following acquisition parameters: time domain, 32 K; 90° pulse width, 11 μs, spectral width, 12 ppm; relaxation delay, 2 s. Sixteen scans and four dummy scans were accumulated for each free induction decay. Baseline correction was performed carefully by applying a polynomial fourth order function in order to achieve a quantitative evaluation of all signals of interest. The spectra were acquired without spinning the NMR tube in order to avoid artificial signals, such as spinning sidebands of the first or higher order. The ^{1}H NMR spectra (δ 0.5–5.5) were divided into regions with equal width of 0.004 ppm using AMIX (v. 3.8, Bruker Biospin) after phase and baseline corrections.

Determination of SA algorithm parameters

In this work, initial temperature T_I of the SA algorithm was 100 and the termination temperature T_s was 0. Student's t-distribution was employed to generate a new solution in the SA algorithm. The random disturbance can be regarded as the jumping of the optimal model. The metropolis criterion was used to determine whether a new point was acceptable or not by calculating the difference of function values at the current point and the new point [27]. The annealing schedule behaves exponentially, which updates the current temperature based on the following formula:

$$T_{n+1} = 0.95^k T_n \tag{4}$$

where, the parameter k is the number of evaluations of the objective function.

In general, there are two stopping rules; The first one works when the number of temperature transitions satisfies the temperature termination rules, while the second rule takes effects when the neighbor solution is not improved after a certain period [27]. In our strategy, the algorithm stopped when the average change on the value of the fitness function at the current point was less than 10^{-6} after 300 iterations. In this work, all calculation was executed in MATLAB 7.6 (The Math Works, Natick, USA).

Results and discussion

Overview of the spectra

The typical ^{1}H NMR spectrum of rapeseed oil is shown in Fig. 1. It shows some typical peaks of rapeseed oil. Peak 1 was CH=CH which was attributed to all unsaturated fatty acids. Peak 2 was CH–OCOR, and peak 3 was CH_2–OCOR. Both of them were attributed to triglycerides. Peak 4 was CH=CH–CH_2–CH=CH which was attributed to linolenyl and linoleyl chains. Peak 5 was CH_2-COOH, peak 7 was CH_2-CH_2COOH, and peak 8 was $(CH_2)_n$. All of them were attributed to all acyl chains. Peak 6 was CH_2–CH=CH

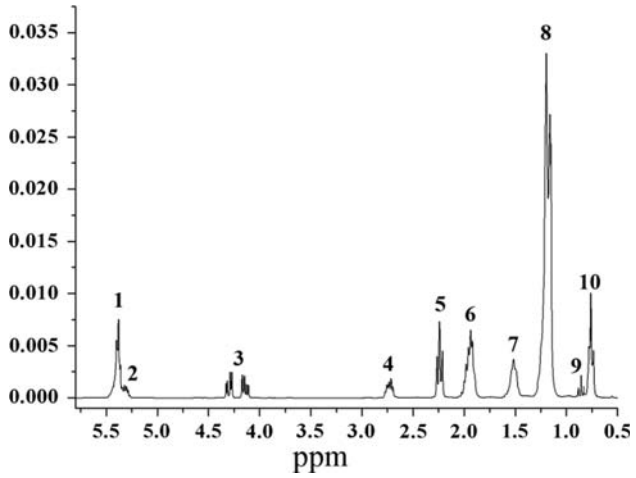

Fig. 1 Typical ^{1}H NMR spectrum of rapeseed oil

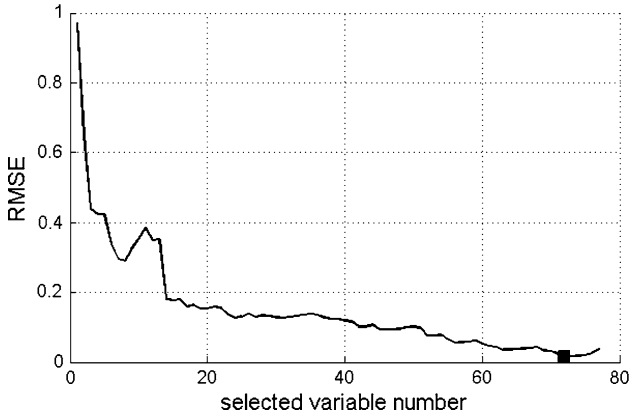

Fig. 2 RMSE scree plot of SPA operated based on the whole ^{1}H NMR spectra for the discrimination of rapeseed oil brands

which was all unsaturated acyl chains. Peak 9 was CH=CH–CH_2–CH_3 which was attributed to linolenyl chain. Peak 10 was $CH_2CH_2CH_2$–CH_3 which was attributed to all acyl chains except linolenyl.

Full-spectra analysis

From Fig. 1, it could be seen that there are no other peaks except peak 1 to 10. Therefore, spectra chemical shifts whose values are close to zero were eliminated before data analysis. The remaining chemical shifts have 557 variables. An LS-SVM-based discrimination model was established based on these variables. A good performance of 85.0% CAR and K of 0.800 was obtained. The K values show the agreement between referenced brand and classified brand is good. Each four brands of rapeseed oils can be mostly classified. However, the CAR is not enough high for the industry application. It can be seen that there are more than five hundreds of variables calculated in the full-spectra model, where some wavelengths which contain irrelevant information were considered. These wavelengths can worsen the predictive ability of the model. Variable selections can let the model more interpretable. More simple calibration model may be obtained by selecting characteristic information such as sample-specific or component-specific variables instead of the full spectra.

SPA calculation directly based on the full spectrum

SPA was carried out on selecting effective variables from the full spectra. Figure 2 shows the RMSE scree plot obtained by SPA based on the whole spectra. The solid square shows the selected variable numbers. As can be seen, a sharp fall is shown in the starting part of the RMSE curve as the numbers of selected variables were from one to fourteen. Then, the trends of RMSE curves become marginal with further increasing number of selected variables. The curve tends to level off after the determination of selected variables by the SPA cutoff threshold procedure by F-test criterion with $\alpha = 0.25$ [15]. Finally, 72 variables (RMSE = 0.016072) were selected.

The selected 72 variables were set as the input variables of LS-SVM model for the discrimination. After the variable selection using SPA, the variable numbers were much reduced (72 vs. 460). CAR of 87.5% and K of 0.833 were obtained. However, based on the selected variables, the performance of SPA-LS-SVM model only little increased compared to full spectrum-LS-SVM model. The reason of little improvement might be because the SPA process operated on the whole spectra caused the selected variables with low S/N [15]. Moreover, the SPA operation based on the whole spectra with hundreds of variables is time-consuming. Thus, it might be possible to reduce the SPA calculation time and improve the SPA performance by eliminating uninformative variables before SPA.

Uninformative variable elimination process

In the process of UVE, different LV numbers of the PLS model in UVE process were compared. The numbers of LVs were calculated from one to thirty. Full cross-validation was used in UVE to prevent overfitting problems. The smallest RMSECV values were obtained based on fifteen LVs. Figure 3 shows the stability of each wavelength variables based on the fifteen LVs. Wavelength variables are at the left of the vertical line, while random variables are at the right side. Two horizontal lines show the lower and upper cutoff thresholds. The variables whose stability is within the cutoff threshold lines should be treated as uninformative and be eliminated. On the basis of optimal LVs, 132 variables were selected from 460 full-spectral variables.

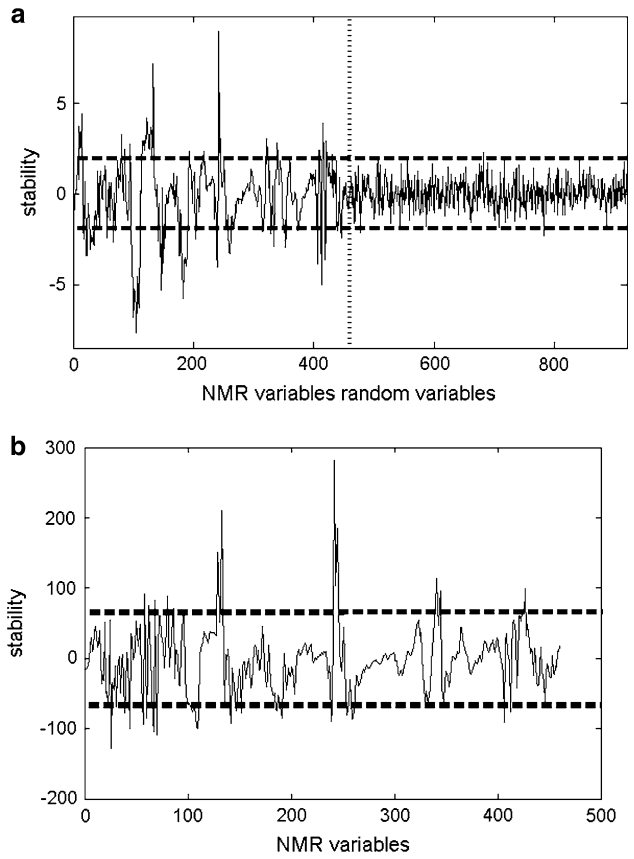

Fig. 3 Stability of each variable in the conventional UVE (**a**) and UVE-SA (**b**). Two *horizontal lines* indicate the *lower* and *upper* cutoff threshold

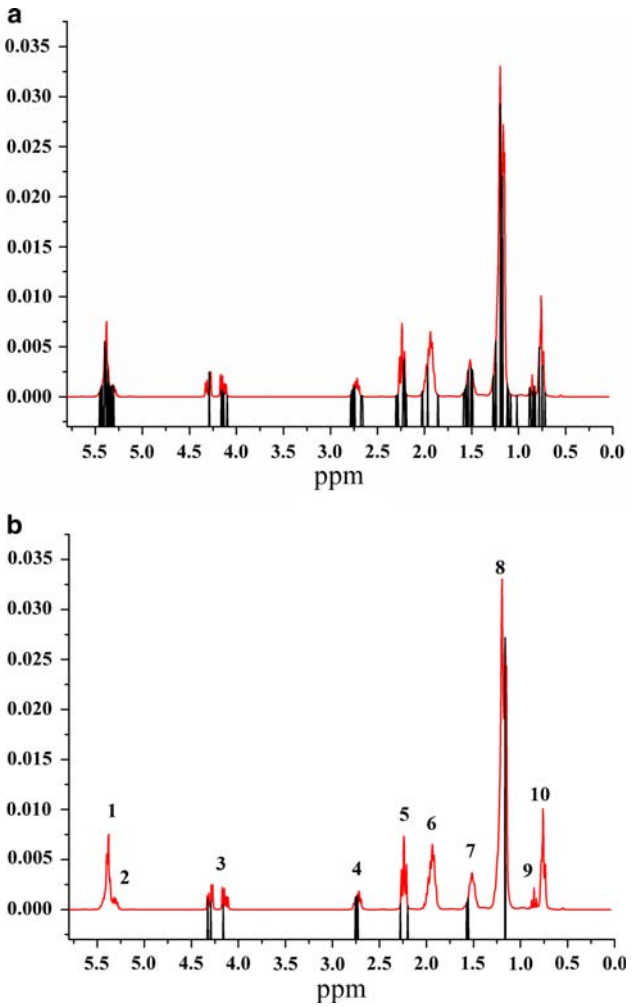

Fig. 4 ^{1}H NMR plot of 77 variables by UVE-SPA (**a**) and 13 variables by UVE-SA-SPA (**b**) for the discrimination of rapeseed oil brands. *Black columns* represent selected wavelength variables

The selected variables by UVE were set as input variables of LS-SVM. Although there are more than three hundreds of variables were eliminated, a good discrimination result of 90.0% CAR and K of 0.867 was obtained. It could be seen that UVE can much eliminated uninformative variables. Therefore, after UVE process, the informative variables can be remained and the model's discrimination ability was increased.

However, as mentioned earlier, the cutoff threshold selection using an artificial random noise cannot obtain the optimal results and results in a defect in the UVE algorithm. Therefore, we applied SA algorithm to estimate the optimal cutoff threshold, instead of using an artificial random noise.

SA-based UVE process

In UVE-SA method, the designed fitness function guides the PLS model to obtain the optimal cutoff threshold. Finally, the optimal cutoff threshold value was obtained as 75, and the best function value is 0.8682. Figure 3 shows the stability obtained by the UVE (a) and UVE-SA (b). In the Fig. 3, the cutoff threshold is shown by the dot lines. Variables with their stability within the cutoff threshold are eliminated. With a comparison of Fig. 3a, b, it can be seen that the two stability curves are similar when the amplitudes are not considered. The different amplitudes are because of the extended noisy matrix which was used for the stability calculation in the conventional UVE method. On the contrary, the spectral data matrix is the only matrix for the stability calculation in UVE-SA.

Further variable selection using SPA

After UVE analysis, variables with no more information for modeling than noise were eliminated. Then, SPA was operated based on the variables selected by conventional UVE and UVE-SA, respectively. Finally, UVE-SPA obtained 77 variables (RMSE = 0.00096138) based on RMSE, and UVE-SA-SPA obtained 13 variables Fig. 4 shows the selected variables by UVE-SPA (a) and UVE-

SA-SPA (b). Columns represent the selected wavelengths. The curve shows the original spectrum. The selected effective variables by UVE-SA-SPA in Fig. 4b show that peaks 3, 4, 5, 7, and 8 are more important for the discrimination, which shows that all acyl chains, triglycerides, and linolenyl and linoleyl chains are the main components for the discrimination.

The selected variables by UVE-SPA and UVE-SA-SPA were separately set as input variables of LS-SVM. UVE-SPA obtained 92.5% CAR and *K* of 0.900, and UVE-SA-SPA obtained 97.5% CAR and *K* of 0.967. Their discrimination results are improved compared to the full spectrum-LS-SVM model and SPA-LS-SVM model. Moreover, the SPA calculation operated on the UVE selected variables is simpler than on the full spectrum, as fewer variables were considered. Therefore, it shows that it is necessary to operate UVE before SPA, which can both reduce the calculation time and increase the model's performance.

Comparing the results of UVE-LS-SVM and UVE-SA-LS-SVM and their further SPA calculations, it can be seen that UVE-SA did better works than conventional UVE. When SA was added for UVE calculation, UVE-SA obtained *K* of 0.933 which is higher than conventional UVE of 0.867 *K*. UVE-SA-SPA also better than UVE-SPA on both CAR and *K*. Moreover, only 13 variables were obtained by UVE-SA-SPA while the conventional UVE-based SPA selected 77 variables. It shows that SPA can do more effective variable selection according to SA-based UVE. The best CAR of 97.5% and *K* of 0.967 were obtained by UVE-SA-SPA-LS-SVM, and are suitable for the industrial application. The results show that SPA can do better discrimination based on UVE, and SA is helpful for UVE to selection more informative variables.

Analysis of final selected variables

After the variable selection of UVE-SA-SPA, there were 13 variables remained, which were from peaks of 3, 4, 5, 7, and 8. In order to find out the most important variables, the performances of each variable, all the combination of each two variables, and all the combination of each three variables were analyzed. When only one variable was used to establish LS-SVM model, their CARs were between 27.5% and 65.0%, which were not good. Therefore, it was not possible to do the classification by only considering one variable. When two variables were used, the best CAR of 87.5% was obtained by the combination of variables 1 and 11 and the combination of variables 8 and 11, respectively. When three variables were considered, the best CAR of 95.0% was obtained by the combination of variables 1, 8, and 11. The results show that variables of 1, 8, and 11 were more important for the classification. Specifically, variable 1 was at peak 8 which was $(CH_2)_n$ and attributed to all acyl chains, variable 8 was at peak 4 which was CH=CH–CH_2–CH=CH and was attributed to linolenyl and linoleyl chains, and variable 11 was at peak 3 which was CH_2–OCOR and attributed to triglycerides. As a conclusion, all acyl chains, linolenyl and linoleyl chains, and triglycerides were most important for the classification.

Conclusion

^{1}H NMR spectroscopy was successfully utilized for the discrimination of rapeseed oil. The result 97.5% CAR and *K* of 0.967 by UVE-SPA-LS-SVM model shows that it is feasible to distinguish different brands of rapeseed oils using ^{1}H NMR spectra. SPA executed on the full spectrum was time-consuming and only improved a few CAR compared to full spectrum-LS-SVM model. UVE was used to eliminate uninformative variables and improve SPA's performance. The better results of UVE-based SPA calculation show that it is necessary to do UVE before SPA, which can both reduce the calculation time and increase the model's performance. Because the cutoff threshold selection in UVE algorithm using an artificial random noise cannot obtain the optimal results and results in a defect, we applied SA algorithm to estimate the optimal cutoff threshold, instead of using an artificial random noise. Finally, SPA obtained 77 and 13 effective variables, respectively, based on conventional UVE and UVE-SA. Comparing the results of UVE-LS-SVM and UVE-SA-LS-SVM and their further SPA calculations, it can be seen that UVE-SA did better works than conventional UVE. Only 13 variables were obtained by UVE-SA-SPA while the conventional UVE-based SPA selected 77 variables. It shows that SA algorithm is helpful for UVE to select more informative variables and SPA can do more effective variable selection according to SA-based UVE. Final result shows that all acyl chains, linolenyl and linoleyl chains, and triglycerides were most important for the classification.

Acknowledgments This work was supported by NSFC (30900129), ZJNSFC (Y2080331), ZJST (2009C34014), and WZST (H20080052).

References

1. Garcia-Gonzalez DL, Mannina L, D'Imperio M, Segre AL, Aparicio R (2004) Eur Food Res Technol 219(5):545–548
2. Lachenmeier DW, Frank W, Humpfer E, Schäfer H, Keller S, Mörtter M, Spraul M (2005) Eur Food Res Technol 220(2):215–221
3. Torbica A, Jovanovic O, Pajin B (2006) Eur Food Res Technol 222(3–4):385–391
4. Molina VD, Uribe UN, Murgich J (2007) Energ Fuel 21(3):1674–1680

5. Wooten JB, Kalengamaliro NE, Axelson DE (2009) Phytochemistry 70(7):940–951
6. Monteiro MR, Ambrozin ARP, Liao LM, Boffo EF, Pereira ER, Ferreira AG (2009) J Am Oil Chem Soc 86(6):581–585
7. de Peinder P, Visser T, Petrauskas DD, Salvatori F, Soulimani F, Weckhuysen BM (2009) Energ Fuel 23:2164–2168
8. Abrahamsson C, Johansson J, Sparen A, Lindgren F (2003) Chemometr Intell Lab 69(1–2):3–12
9. Kalivas JH, Roberts N, Sutter JM (1989) Anal Chem 61(18):2024–2030
10. Kleynen O, Leemans V, Destain MF (2003) Postharvest Biol Tec 30(3):221–232
11. Norgaard L, Saudland A, Wagner J, Nielsen JP, Munck L, Engelsen SB (2000) Appl Spectrosc 54(3):413–419
12. Borin A, Poppi RJ (2005) Vib Spectrosc 37(1):27–32
13. Araujo MCU, Saldanha TCB, Galvao RKH, Yoneyama T, Chame HC, Visani V (2001) Chemometr Intell Lab 57(2):65–73
14. Breitkreitz MC, Raimundo IM, Rohwedder JJR, Pasquini C, Dantas HA, Jose GE, Araujo MCU (2003) Analyst 128(9):1204–1207
15. Ye SF, Wang D, Min SG (2008) Chemometr Intell Lab 91(2):194–199
16. Centner V, Massart DL, deNoord OE, deJong S, Vandeginste BM, Sterna C (1996) Anal Chem 68(21):3851–3858
17. Han QJ, Wu HL, Cai CB, Xu L, Yu RQ (2008) Anal Chim Acta 612(2):121–125
18. Oliver J, Palou A (2000) J Chromatogr A 881(1–2):543–555
19. Galvao RKH, Araujo MCU, Fragoso WD, Silva EC, Jose GE, Soares SFC, Paiva HM (2008) Chemometr Intell Lab 92(1):83–91
20. Galvao RKH, Pimentel MF, Araujo MCU, Yoneyama T, Visani V (2001) Anal Chim Acta 443(1):107–115
21. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Science 220(4598):671–680
22. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) J Chem Phys 21(6):1087–1092
23. Suman B (2004) Comput Chem Eng 28(9):1849–1871
24. Horchner U, Kalivas JH (1995) Anal Chim Acta 311(1):1–13
25. Wu D, Yang HQ, Chen XJ, He Y, Li XL (2008) J Food Eng 88(4):474–483
26. Landis JR, Kock GG (1977) Biometrics 33(1):159–174
27. Ghazanfari M, Alizadeh S, Fathian M, Koulouriotis DE (2007) Appl Math Comput 192(1):56–68