African Journal of Agricultural Research Vol. 6(27), pp. 5987-5991, 19 November, 2011
Available online at http://www.academicjournals.org/AJAR
DOI: 10.5897/AJAR11.746
ISSN 1991-637X ©2011 Academic Journals

Full Length Research Paper

Integration of modified uninformative variable elimination and successive projections algorithm for determination harvest time of laver by using visible and near infrared spectra

Xiaochun Guan[1], Xiaojing Chen[1]* and Jun Jiang[2]

[1]College of Physics and Electronic Engineering Information, Wenzhou University, China.
[2]College of Chemistry and Materials Engineering, Wenzhou University, China.

Accepted 26 June, 2011

In order to quickly and accurately determine the laver's harvest time, we adopt combination of modified uninformative variable elimination, successive projection algorithm and visible-near infrared spectroscopy (Vis-NIR) technology to achieve this goal; as mass of spectral data with noise cannot build a stable and efficient recognition model, the effective wavelength should be extracted from the whole spectra. Modified uninformative variable elimination (Muve) algorithm was used to eliminate uninformative variable and noise, successive projection algorithm (SPA) was used to eliminate relevant redundant information, and the remaining variables of 19 were obtained. Finally, the remaining 19 variables were used to establish recognition model using partial least squares vector machine (LS-SVM), and satisfactory prediction rate of 96.67% was obtained. Meanwhile, compared to other traditional variable selection algorithms, such as genetic algorithm (GA) and simulated annealing (SA) algorithm, the proposed algorithms have more advantages.

Key words: Laver, visible-near infrared spectroscopy (Vis-NIR), uninformative variable elimination (UVE), successive projection algorithm (SPA).

INTRODUCTION

The laver is one algae with high content of protein and salty minerals. In general, the laver's harvest time is generally divided into the first water (the first harvest time), the second water (the second harvest time), the third water (the third harvest time) and the fourth water (the fourth harvest time) in a year, its protein content varies with change of the laver's harvest time, content of protein is in the highest level at early stage of harvest, then it decreases with extension of growth time. Therefore, harvest time has a close relationship with the quality of laver (Ying et al., 2009). Although the nutritional value of laver varies with change of harvest time, but for the laver traded in the market, different kinds of laver hardly can be distinguished through their appearance and color. At present, different harvest times are generally determined subjectively, which not only requires experience, but also involves many subjective factors, the accuracy is difficult to grasp. In recent years, in order to make huge profits, some illegal factories package inferior laver with the second harvest time even the third harvest time to the superior laver with the first harvest time, in order to aim at stopping such illegal sales, there is an urgency for methods to achieve rapid detection of different harvest times of the laver.

Visible-near infrared spectroscopy (Vis-NIR) spectroscopy is widely used for rapid, low cost and non-destructive analysis in industry, such as agriculture, pharmaceuticals, food, textiles, cosmetics and polymer production (Ozturk et al., 2010; Pigorsch et al., 2010; Cen et al., 2007; Wu et al., 2009), but there are no reports for determination of the laver harvest time using the Vis-NIR. Although the visible-near infrared spectroscopy technology

*Corresponding author. E-mail: chenxj9@163.com.

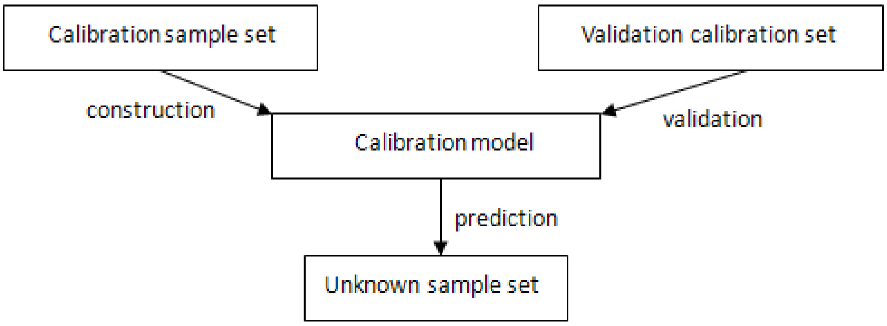

Figure 1. The course of model development.

(Vis-NIR) is an advanced method, it still has some defects: spectral data is too large, and the spectrum contains a lot of noise and uninformative variables (Ying et al., 2008), how to maximize the extractable characteristic information from the spectral variable and establish a stable and efficient model has been a technical problem. At present, the commonly used variable extraction methods, such as the load weights (Wu et al., 2008), regression coefficient method (Wu et al., 2008), and these methods are vulnerable to the subjective selection threshold, the simulated annealing (SA) algorithm (Chen et al., 2009) and genetic algorithm (GA) (Leardi et al., 2002; Hörchner et al., 1995) have large computation load. The Muve is an improved method from traditional uninformative variable elimination (UVE) algorithm (Chen et al., 2009; Chen et al., 2011; Centner et al., 1996); it can effectively eliminate noise and uninformative variables. The successive projection algorithm (SPA) (Araujo et al., 2001; Chen et al., 2010) is a kind of new method of variable extraction, it can search for the variable group with the minimum redundancy, minimize the collinearity between the variables, signify-cantly reduce the number of variables used in modeling and improve the speed and efficiency of modeling through projection analysis of vectors.

This study aims to rapidly determinate harvest times of laver using visible-near infrared spectroscopy. The characteristic information is extracted using combination of Muve and SPA and as input variable of least squares supporting vector machines (LS-SVM) to build the prediction model.

MATERIALS AND METHODS

Instrumentation and spectral acquisition

Spectra were collected using a Vis-NIR spectrometer (USB4000 Miniature Fiber Optic Spectrometer, the Ocean Optics, Inc. USA). Laver from each harvest time were fragmented and spread on a paper. An iron plate was used to make fragment's surface close to smooth. A total of ten scans were investigated each time which was considered as one sample. As the dried laver is rolled up into a strip, which is not conducive to spectral acquisition, thus use a grinder (Model: YF-110 Yongli Pharmaceutical Machinery Co., Ltd. in Ruian) to grind the laver into fragments, and then use spectra for acquisition, In the experiment, use a flat iron to try to smooth the surface of laver, in the entire process of acquisition of spectra, the spectra fiber-optic probe has always kept a distance of 4 mm from the sample.

Sample partition and model construction

The sample of laver is from Cangnan in Zhejiang at different harvest times, there are a total of three harvest times, and 35 samples are prepared for the stage of the first water, the second water and the third water respectively, a total of 105 samples. The 105 samples are divided into the calibration set of 75 and prediction set of 30 using Kennard-Stone algorithm (Macho et al., 2001), the course of model development was described in Figure 1.

RESULTS AND DISCUSSION

Reflectance spectra investigation

A low signal to noise ratio is in the spectra between 346 and 452 nm, and between 1026 and 1050 nm, therefore the spectra containing useful information was determined as being the region between 452 and 1026 nm, and a total of 3000 variables were obtained.

The visible-near infrared typical reflectance spectra of the laver in different harvest times is shown in Figure 2, as seen from Figure 2, trends of spectral curve of the laver in different harvest times are very similar, only from the spectral characteristics, it is difficult to distinguish from different harvest times. Moreover, the spectra can be observed with relatively large noise .Therefore, there is need to adopt the chemometrics for preprocessing the spectral data.

Extraction of the characteristic information using Muve and SPA

The number of input variable of LS-SVM is limited, if the

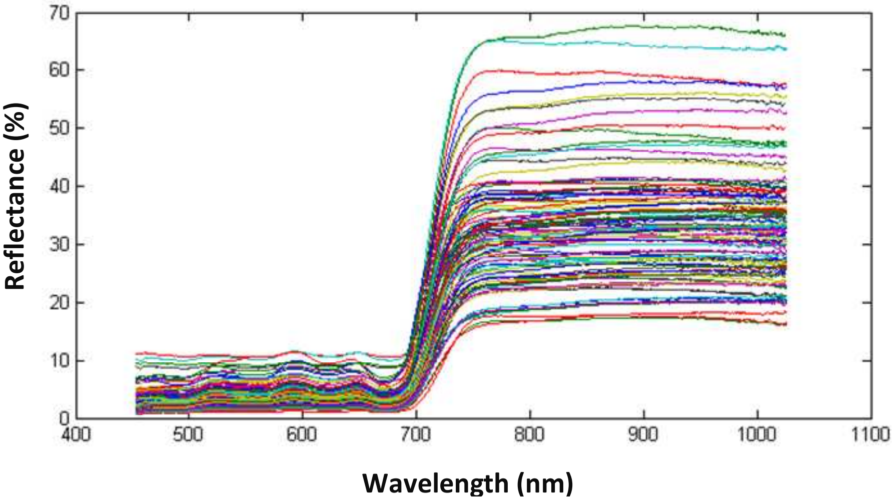

Figure 2. Original reflection spectra of three different harvest time of laver.

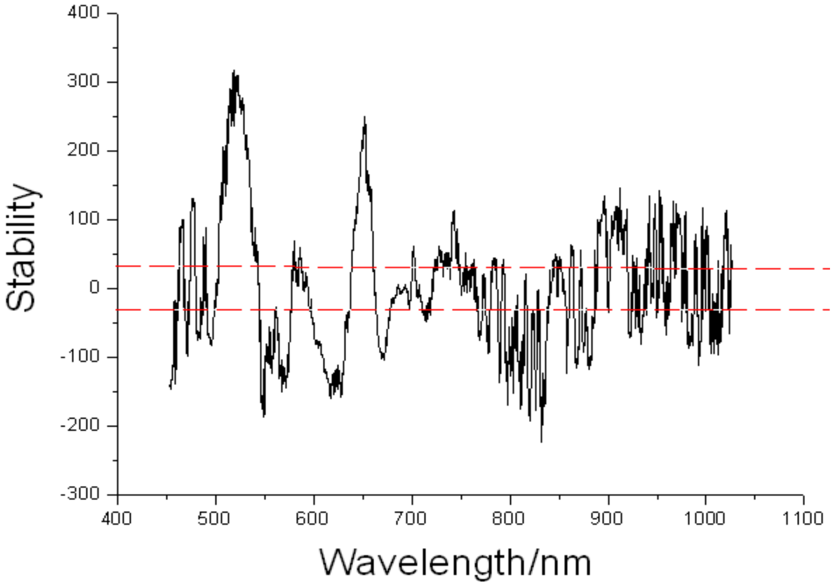

Figure 3. Stability distribution of each wavelength and cutoff threshold obtained by Muve, the two dot lines indicate the lower and upper cutoff.

whole spectra are as the input variables of LS-SVM to built model, the amount of data is too large, which affects quality of the model. Therefore, noise, redundant information and other useless information should be eliminated. Muve and SPA was used to achieve this goal here. As Muve method in our study only considered for optimal threshold, so a single cycle can satisfy it. Like the literature (Chen et al., 2009) the objective function was set as cross validation of root mean square's error (RMSECV). The number of the optimal principal components of 9 was obtained from the cross validation. Figure 3 shows the stability and the optimal threshold value, where the optimal threshold value is 35, after being processed, 992 variables were kept. Seen from Figure 3, the stability of the wavelength region 500 to 550, 590 to 640, 650 to 670, 760 to 840 and 950 to 1000 was larger than the cutoff, which means these wavelengths is relatively large contribution to the model.

Due to a lot of noise and uninformative variables was be eliminated by Muve algorithm, so that spectral variables were reduced from 3000 to 992. However, the remaining spectra still contains a lot of relevant redundant

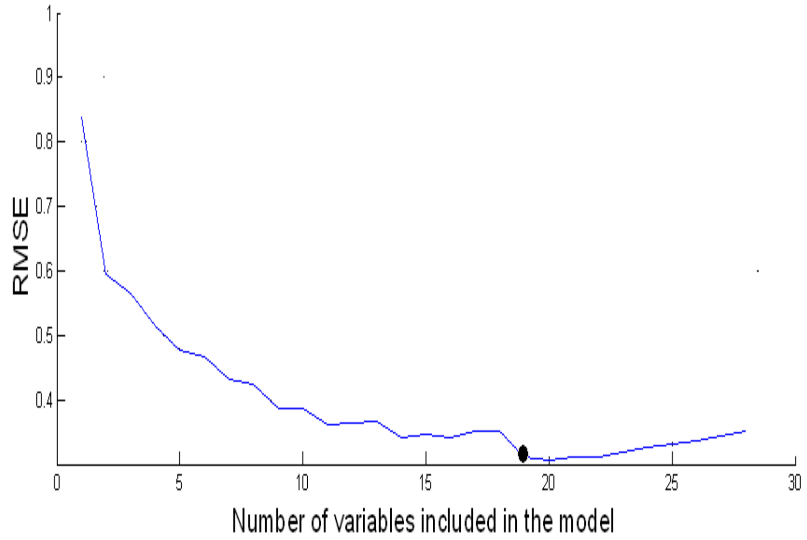

Figure 4. RMSE plot of number of selected variables by SPA.

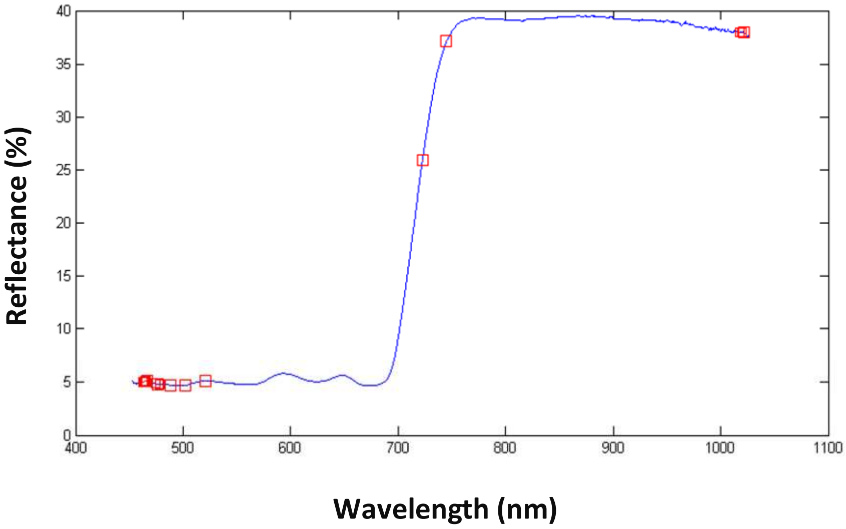

Figure 5. Nineteen selected wavelengths, hollow box represent selected wavelength.

information that affects the LS-SVM's quality of the modeling. Therefore, SPA was used to deal with the remaining 992 variables. The distribution of the RMSE with different wavelengths by SPA is shown in Figure 4, and black solid dots indicate the number of selected wavelengths. As seen from Figure 4, RMSE generally has a trend of descending, when it descends to 19 wavelengths, the trend of RMSE curve changes little, so it can determine the optimal wavelength with the number of 19. Distribution of 19 wavelengths obtained by Muve and SPA is shown in Figure 5, in which hollow box indicates the selected wavelength, the selected wavelength mainly concentrates on the wavelength region 452 to 550 and 1000 to 1026 nm. The selected 19 wavelengths are 462.2, 463.04, 463.25, 463.89, 464.73, 465.36,466, 466.84, 474.01, 476.33, 478.43, 488.1, 502.34, 520.71, 722.74, 745.07, 1017.7, 1019.6 and 1021.7 nm respectively. With these 19 wavelengths as the input variable to establish Muve and SPA-LS-SVM model, the prediction rate of 96.67% was obtained.

Here, the traditional variable selection methods were employed to compare the prediction results. The prediction results using Muve, SPA, GA, SA and Muve and SPA are showed in Table 1. Seen from Table 1, the results of Muve and SPA-LS-SVM are best. Also it can be seen that after Muve processing, recognition rate was not

Table 1. Prediction results of three different harvest time of laver by LS-SVM models by using different variable selection methods.

Variable selection methods	Number of selected variables	Calibration set (n=75) correct rate (%)	Prediction set (n=30) correct rate (%)
None	3000	57.33	53.33
Muve	992	66.67	63.33
SPA	16	80.00	76.6
GA	23	86.67	86.67
SA	22	81.33	83.33
Muve&SPA	19	97.33	96.67

highly raised, while there was higher increasing of prediction rate after processing of GA, SA and SPA. Poor prediction results for several classical methods are that the noise or redundant information limited the capability of these methods. From the results, it can be concluded that Muve can effectively eliminate the spectral noise and uninformative variables and improve search capability of SPA.

Conclusions

Muve and SPA is used to select the characteristic information of spectra, the selected wavelength is used as the input variable to establish LS-SVM model to determine different laver’s harvest times. Moreover, the selected 19 wavelengths are able to reflect spectral information from the whole wavelength; the results prediction rate is up to 96.67%. The results show that it is feasible for identification of the laver harvest times by combination of Vis-NIR, Muve and SPA, at the same time, it is demonstrated that Muve and SPA is an effective variable selection method.

ACKNOWLEDGEMENT

This study was supported by Natural Scientific Research Fund of Zhejiang Provincial (ZJNSFC NO.3110289).

REFERENCES

Araujo MCU, Saldanha TCB, Galvão RKH, Yoneyama T, Chame HC, Visani V (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom. Intell. Lab. Syst., 57: 65-73.

Cen H, Bao Y, He Y (2007). Fast Discrimination of Varieties of Bayberry Juice Based on Spectroscopy Technology, *Spectrosc.Spectr.* Anal. 27: 503-506.

Centner V, Massart DL (1996), Elimination of Uninformative Variables for Multivariate Calibration. Anal. Chem., 68: 3851-3858.

Chen X, Lei X (2009). Application of a Hybrid Variable Selection Method for Determination of Carbohydrate Content in Soy Milk Powder Using Visible and Near Infrared Spectroscopy. J. Agric. Food Chem., 57: 334-340.

Chen X, Li H, Wu D, Lei X, Zhu X, Zhang A (2010). Application of a hybrid variable selection method for the classification of rapeseed oils based on ^{1}H NMR spectral analysis, Eur. Food Res. Technol., 230: 981-988.

Chen X, Wu D, He Y (2011) An integration of modified uninformative variable elimination and wavelet packet transform for variable selection. Spectroscopy, 26: 42-27.

Chen X, Wu D, He Y (2009). Detecting the quality of glycerol monolaurate: A method for using Fourier transform infrared spectroscopy with wavelet transform and modified uninformative variable elimination. Anal. Chim. Acta, 638: 16-22.

Hörchner U, Kalivas JH (1995). Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection, Anal. Chim. Acta 311: 1-13.

Leardi R, Seasholtz MB, Pell RJ (2002). Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. Anal. Chim. Acta, 461: 189-200.

Macho S, Rius A, Callao MP (2001). Monitoring ethylene content in heterophasic copolymers by near-infrared spectroscopy: Standardisation of the calibration model. Anal. Chim. Acta, 445: 213-220.

Ozturk B, Yalcin A, Ozdemir D (2010) Determination of olive oil adulteration with vegetable oils by near infrared spectroscopy coupled with multivariate calibration, J. Near Infrared Spectrosc., 18: 191-201

Pigorsch E (2010). Classification of offset and flexo printed newspapers by near infrared spectroscopy. J. Near Infrared Spectrosc., 18: 225-229.

Wu D, He Y, Feng S (2008). Study on infrared spectroscopy technique for fast measurement of protein content in mild powder based on LS-SVM. J. Food Eng., 84: 124-131.

Wu D, He Y, Shi J, Feng S (2009). Exploring Near and Midinfrared Spectroscopy to Predict Trace Iron and Zinc Contents in Powdered Milk. J. Agric. Food Chem., 57: 1697-1704.

Ying M, Shi W, Pan F (2009). Analysis of nutrition on harvest time of laver. J. Zhejiang Agric. Sci., 6: 1227-1228.

Ying Y, Liu Y (2008) Nondestructive measurement of internal quality in pear using genetic algorithms and FT-NIR spectroscopy. J. Food Eng., 84: 206-213.