XVII[th] World Congress of the International Commission of Agricultural and Biosystems Engineering (CIGR)

Hosted by the Canadian Society for Bioengineering (CSBE/SCGAB)
Québec City, Canada June 13-17, 2010

APPLICATION OF NEAR INFRARED SPECTROSCOPY AND LEAST SQUARES-SUPPORT VECTOR MACHINE TO DETERMINE SOLUBLE PROTEIN IN OILSEED RAPE LEAVES

FEI LIU[1], YUN ZHAO[1], GUANGMING SUN[1], HUI FANG[1], YONG HE[1]

[1]F. Liu, College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310029, China (Corresponding author: Yong He, Email: yhe@zju.edu.cn)

CSBE101287 – Presented at Section VI: Postharvest Technology and Process Engineering Conference

ABSTRACT Soluble protein was an important parameter to indicate and monitor the growing status of oilseed rape. Near infrared (NIR) spectroscopy combined with least squares-support vector machine (LS-SVM) was investigated to determine soluble protein in oilseed rape leaves under herbicide stress. Different preprocessing methods were compared, including Savitzky-Golay smoothing (SG), standard normal variate (SNV), multiplicative scatter correction (MSC), first-derivative (1-Der), second-derivative (2-Der) and de-trending. The optimal performance was determined by the PLS model with correlation coefficients (*r*) and root mean squares error of prediction (RMSEP). Successive projections algorithm was applied to select effective wavelengths (EWs), which were used as inputs of LS-SVM. The optimal prediction results was achieved by SPA-LS-SVM model (De-trending spectra) with *r*=0.9879 and RMSEP=61.8231. The results indicated that NIR spectroscopy combined with SPA-LS-SVM was successfully applied for the determination of soluble protein in oilseed rape leaves under herbicide stress.

Keywords: Near infrared spectroscopy, oilseed rape, soluble protein, least squares-support vector machine

INTRODUCTION Oilseed rape (*Brassica napus* L.) is the most important source of edible oil in China, which is expanding rapidly as a rotation crop following rice (Zhou, 2001). In order to keep a good growing environment, some herbicide would be applied to the oilseed rape. A newly developed propyl 4-(2-(4,6-dimethoxypyrimidin -2-yloxy)benzylamino)benzoate (ZJ0273) was an ALS (acetolactate synthase)-inhibiting herbicide. Herbicide ZJ0273 also influences the growing of oilseed rape. One of the most important parameters during oilseed rape growing was soluble protein, which was traditionally determined by Bradford method (Bradford, 1976). This method was time consuming, laborious, costly and not suitable for the fast, non-destructive determination and *on field* monitoring during the various growing stages of oilseed rape.

Near infrared (NIR) spectroscopy have a large range of application fields, which have spread to agriculture, food, cosmetics and other industries for both quantitative and qualitative analysis (Yan et al., 2005). In the application of oilseed rape, NIR

spectroscopic techniques had been applied for the determination of chlorophyll of rape leaves (Fang et al., 2007), the determination of acetolactate synthase (ALS) and protein content of oilseed rape leaves using visible/near infrared (400-100 nm) spectra (Liu et al., 2008a; Liu et al., 2009a). However, up to our knowledge, there was no report about the determination of soluble protein under the herbicide ZJ0273 stress in oilseed rape using near-infrared spectroscopy within 1100-2500 nm.

The objective of this study was to study the feasibility of near infrared spectroscopy to determine the soluble protein in oilseed rape leaves under herbicide stress. Different preprocessing was compared, and successive projections algorithm (SPA) was applied for relevant variable selection. The performance of different calibrations was also compared including partial least squares (PLS) and least squares-support vector machine (LS-SVM).

MATERIALS AND METHODS

Sample preparation One leading cultivar of oilseed rape (*Brassica napus*, cv. ZS758) was planted at the farm of Zhejiang University, Hangzhou (30° 10´N, 120° 12´E). The new developed herbicide (ZJ0273) was applied to the oilseed rape leaves. Various concentrations of ZJ0273 (0, 100, 200 and 500 mg/L) were foliar applied at 5-leaf stage at the quantity of 500 L/ha. Conventional crop management was used during the growing period. A total of 248 samples were collected by three times, and 80, 80, 88 samples were collected at each time. The samples were dried and sieved through 60-mesh. 186 samples (two thirds samples for each ZJ0273 concentration) were selected randomly for calibration set, and the remaining 62 samples were used as the validation set. The samples in calibration and validation sets were randomly changed several times to confirm the randomization. No single sample was used in both calibration set and validation set at the same time.

Spectral collection and reference method for soluble protein The reflectance spectra of each sample were obtained by the Foss NIRSystems 5000 (Foss NIRSystems, Denmark) within the wavelength region 1100-2500 nm. The resolution of instrument was 2 nm, and 700 data points were collected for each spectrum. The small round cup was used for sample container. All spectral data were stored in personal computer for further treatment. The reference method for soluble protein was measured by the method of Commassie blue according to the report of Bradford (Bradford, 1976).

Spectral preprocessing and SPA In order to achieve a good prediction performance, some preprocessing methods were applied to remove the spectral baseline shift, noise and light scatter influence (Chu et al., 2004). The reflectance spectra were firstly transformed into absorbance spectra by log(1/R) (R=reflectance). Then the absorbance spectra were transformed into ASCII format. For comparison, the following preprocessing methods were calculated, including Savitzky-Golay smoothing (SG), standard normal variate (SNV), multiplicative scatter correction (MSC), first-derivative (1-Der), second-derivative (2-Der) and de-trending. The performance was determined by the prediction results in the latter calibration stage. The pretreatments were implemented by “The Unscrambler V 9.8” (CAMO AS, Oslo, Norway).

Successive projections algorithm (SPA) was a newly proposed relevant variable selection method (Araújo et al., 2001; Galvão et al., 2008). It could use the projection procedure to select the most relevant variable with least collinearity and redundancies. The selected variable, called effective wavelengths (EWs), could be applied as input for the development of a more parsimonious model.

Partial least squares analysis Partial least squares (PLS) analysis is the most frequently applied calibration method for modeling in the application of NIR spectroscopic techniques. PLS employs the latent variables (LVs) to develop a relationship between the spectral data and soluble protein in oilseed rape leaves.

Least squares-support vector machine LS-SVM was a powerful calibration method, which could handle both linear and nonlinear problems and solve these problems in a relatively fast way using small sample database (Suykens et al., 1999). The details of LS-SVM could be found in the literatures (Suykens et al., 1999; Liu et al., 2008b). Herein, LS-SVM was applied to develop the correlation between the selected EWs by SPA and soluble protein in oilseed rape leaves. In the LS-SVM model, the inputs were the selected EWs by SPA with different preprocessing methods. The radial basis function (RBF) was recommended as kernel function. The model parameters *gamma* (γ) and *sig2* (σ^2) were determined by a two-step grid search technique. All the calculations were performed using MATLAB software v. 7.0 (The Math Works, Natick, MA, USA). The free LS-SVM v 1.5 toolbox (Suykens, Leuven, Belgium) was applied with MATLAB software v. 7.0 to develop the LS-SVM models.

The evaluation standards were correlation coefficients (r), root mean squares error of prediction (RMSEP), bias, slope and offset. The main indices in this paper were r and RMSEP. The good model should be with higher r value and lower RMSEP, absolute bias and offset values, and the slope of the regression line should be closer to 1.

RESULTS AND DISCUSSION

Spectral features of oilseed rape The raw absorbance spectra of oilseed rape leaves under herbicide ZJ0273 stress is shown in Fig. 1. As can be seen, the trends of all samples with different herbicide concentration and leaf-position were quite similar. There were many absorbance peaks which might be corresponding to chemical compositions with C-H, N-H or O-H bands. The statistics of soluble protein in oilseed rape leaves are shown in Table 1. As can be seen, the range of calibration set covered a larger range than validation set, which was thought to be helpful for the development of a stable and general model.

The performance of PLS models Using the aforementioned preprocessing methods, different PLS models were develop with different latent variables. The cross-validation was applied to avoid overfitting problem and validate the model performance during the calibration stage. The samples in validation set were used to assess the prediction performance of developed models. According to the aforementioned evaluation standards, such as r and RMSEP, the optimal preprocessing method could be determined. The prediction results by PLS models are shown in Table 2. As stated in Table 2, the optimal performance was achieved by de-trending spectra with r=9543 and RMSEP=118.2955. Then the following PLS models were 1-Der and Raw spectra based

model. This result was acceptable only considering the prediction performance. However, in the PLS models, 700 variable were employed as inputs, and there might be much collinearity and redundancy between these wavelengths. In order to develop a more parsimonious model, a relevant variable selection procedure should be performed, therefore, successive projections algorithm was recommended for such purpose.

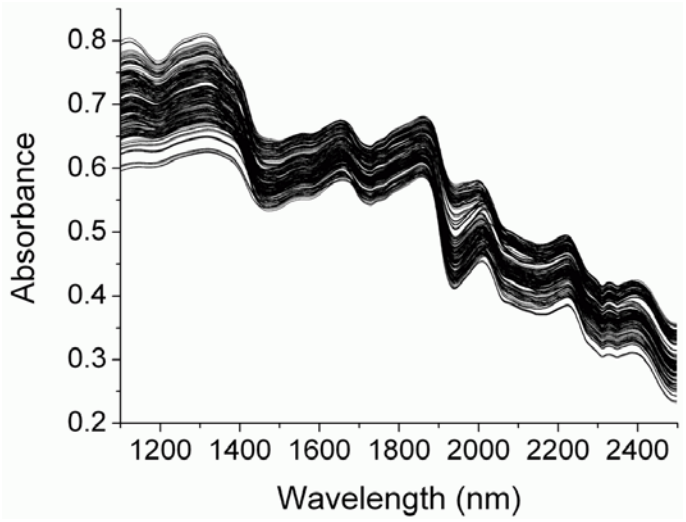

Fig. 1. The raw absorbance spectra of oilseed rape

Table 1 The statistics of soluble protein content of oilseed rape leaves (μg/g DW)

Set	No.	Range	Mean	S.D.
Cal.	186	442.885-2029.335	1300.156	400.9406
Val.	62	459.907-1955.233	115.498	395.4253
All	248	442.885-2029.335	1303.991	398.8259

EWs selected by SPA According to the developed PLS models, a better performance was achieved by de-trending spectra, then the 1-Der and Raw spectra. Hence, these three preprocessing methods were also applied in the SPA procedure. During the SPA, the maximum number of selected variable was set as 30, and cross-validation was also applied in the selection process. Different wavelengths were selected by SPA (shown in Table 3). As pointed in Table 3, the selected EWs were ranked in the order of importance. Take de-trending for instance, 1736 nm was the most important variable within all selected EWs by de-trending spectra. These EWs were applied as inputs of LS-SVM to develop SPA-LS-SVM models, which was a newly proposed combination in previous study (Liu et al., 2009b). Then the SPA-LS-SVM models could be employed for the determination of soluble protein in oilseed rape leaves.

LS-SVM models As stated above, the SPA-LS-SVM models were developed using the selected EWs by SPA, RBF kernel, and combination of (γ, σ^2) determined by two-step grid search technique. The search region of (γ, σ^2) was set as 10^{-3}-10^5, which was determined according to experience and previous literature (Liu et al., 2007; Liu et al., 2008a; Liu et al., 2008b; Liu et al., 2009b). The SPA-LS-SVM models were developed using the calibration set, and the optimal (γ, σ^2) were (4.5×10^3, 19.2), (46.2, 19.8) and (24.7, 2.0) for raw, 1-Der and de-trending spectra based SPA-LS-SVM models, respectively. The prediction results for validation set are shown in Table 2. As can be

seen, SPA-LS-SVM (de-trending) model was optimal and slightly outperformed the other two LS-SVM models (raw and 1-Der). The prediction results by SPA-LS-SVM (de-trending) were r=0.9879 and RMSEP=61.8231. After comparison, it could be concluded

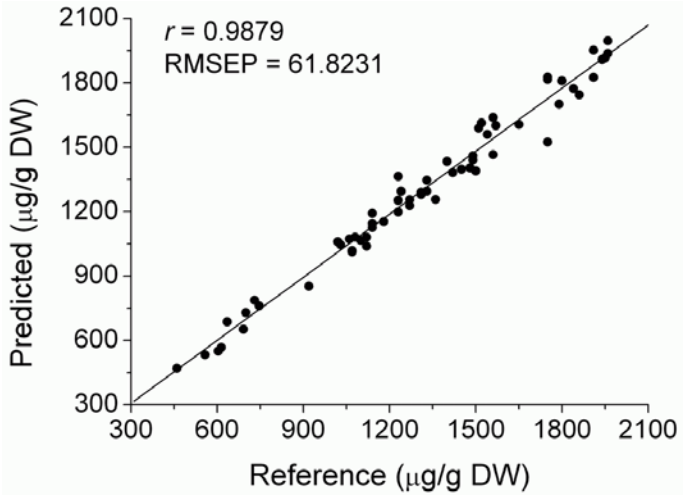

Fig. 2 Reference vs predicted values of soluble protein by SPA-LS-SVM (de-trending)

Table 2 The prediction results of soluble protein content by PLS and SPA-LS-SVM models

Preprocessing	LVs/EWs	*r*	RMSEP	Bias	Slope	Offset
PLS						
Raw	10/-	0.9465	127.7280	-9.3343	0.9293	83.6486
SG+SNV	8/-	0.9410	133.4913	-5.9434	0.9178	102.2002
MSC	8/-	0.9460	127.9245	-3.5027	0.9273	92.1758
1-Der	8/-	0.9541	119.2503	-14.9960	0.9438	58.9913
2-Der	6/-	0.9331	141.8678	9.8734	0.8989	142.8278
De-trending	9/-	0.9543	118.2855	-12.5918	0.9344	73.7385
SPA-LS-SVM						
Raw	-/16	0.9869	67.1786	-23.6942	0.9920	-14.1200
1-Der	-/12	0.9747	90.9994	-22.9597	0.9850	-4.2530
De-trending	-/11	0.9879	61.8231	-13.9826	0.9790	12.6000

Table 3 The selected EWs by SPA

Preprocessing	No.	Selected EWs (nm)
Raw	16	2210, 2176, 1228, 1728, 2432, 1188, 1648, 1594, 1784, 1686, 2312, 2026, 2498, 1328, 2336, 2244
1-Der	12	2194, 2374, 1706, 1768, 2330, 2108, 1596, 2326, 1202, 2352, 1282, 2486
De-trending	11	1736, 2094, 1782, 1606, 2052, 2312, 1436, 2250, 1114, 2400, 1154

that de-trending preprocessing was the most suitable one for such specific study in the determination of soluble protein in oilseed rape leaves. The reference vs predicted values of soluble protein are shown in Fig. 2. It was worth noting that all developed SPA-LS-SVM models showed a better performance than that of all developed PLS models. This result indicated the following two advantages of SPA-LS-SVM calibration method. One was that SPA was a powerful approach for the selection of relevant variables as EWs. The other one was that LS-SVM method considered both the linear and nonlinear information in the spectral data, whereas PLS only take advantage of the linear information. The results indicated that NIR spectroscopy combined with SPA-LS-SVM

could be applied for the determination of soluble protein in oilseed rape leaves under herbicide stress. This study proposed a new approach for the *on- field* monitoring and detecting the growing status of oilseed rape.

CONCLUSION Near infrared spectroscopy combined with SPA-LS-SVM was successfully performed to determine the soluble protein in oilseed rape leaves when they were under the stress of herbicide ZJ0273. The optimal preprocessing method for such specific study was de-trending. SPA was a powerful way for effective wavelengths selection, and the new Spa-LS-SVM (de-trending) achieved the best prediction results with r=0.9879 and RMSEP=61.8231. The results indicated that NIR combined with SPA-LS-SVM could be applied for soluble protein detection, and the results would be helpful for further studies about other physiological parameters and *on-field* monitoring of growing status of oilseed rape.

Acknowledgements. This work was supported by the 863 National High Technology Research and Development Program of China (2007AA10A210), Natural Science Foundation of China (60802038), Zhejiang Innovation Program for Graduates (YK2008014), Zhejiang Provincial Natural Science Foundation of China (Z3090295), and Agricultural Science and Technology Achievements Transformation Fund Programs (2009GB23600517).

REFERENCES

Araújo, M.C.U., T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, and V. Visani. 2001. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. Chemom. Intell. Lab. Syst. 57: 65-73.

Bradford M.M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal. Biochem. 72: 248-254.

Chu, X.L., H.F. Yuan, and W.Z. Lu. 2004. Progress and application of spectral data pretreatment and wavelength selection methods in NIR analytical technique. Prog Chem. 16: 528-542.

Fang,H., H.Y. Song, F. Cao, Y. He, and Z.J. Qiu. 2007. Study on the relationship between spectral properties of oilseed rape leaves and their chlorophyll content. Spectrosc. Spectr. Anal. 27: 1731-1734.

Galvão, R.K.H., M.C.U. Araújo, W.D. Fragoso, E.C. Silva, G.E. José, S.F.C. Soares, and H.M. Paiva. 2008. A variable elimination method to improve the parsimony of MLR models using the successive projections algorithm. Chemom. Intell. Lab. Syst. 92: 83-91.

Liu, F., and Y. He. 2007. Use of visible and near infrared spectroscopy and least squares-support vector machine to determine soluble solids content and pH of cola beverage. J. Agric. Food Chem. 55: 8883-8888.

Liu, F., F. Zhang, Z.L. Jin, Y. He, H. Fang, Q.F. Ye, and W.J. Zhou. 2008a. Determination of acetolactate synthase activity and protein content of oilseed rape (*Brassica napus* L.) leaves using visible/near infrared spectroscopy. Anal. Chim. Acta 629: 56-65.

Liu, F., H, Fang, F, Zhang, Z.L. Jin, W.J. Zhou, and Y. He. 2009a. Nondestructive determination of acetolactate synthase in oilseed rape leaves using visible and near infrared spectroscopy. Chinese J. Anal. Chem. 37: 67-71.

Liu, F., Y. He, and L. Wang. 2008b. Comparison of calibrations for the determination of soluble solids content and pH of rice vinegars using visible and short-wave near

infrared spectroscopy. Anal. Chim. Acta 610: 196-204.
Liu, F., Y.H. Jiang, and Y. He. 2009b. Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: A case study to determine soluble solids content of beer. Anal. Chim. Acta 635: 45-52.
Suykens, J.A.K. and J. Vandewalle. 1999. Least squares support vector machine classifiers. Neural Process. Lett. 9: 293-300.
Yan, Y.L., L.L. Zhao, D.H. Han, and S.M. Yang. 2005. The Foundation and Application of Near Infrared Spectroscopy Analysis. Beijing: China Light Industry Press.
Zhou, W.J. 2001.Oilseed Rape, Cultivation of Crops. Hangzhou: Zhejiang University Press.