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

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

THE SUCCESSIVE PROJECTIONS ALGORITHM FOR VARIABLE SELECTION ALIEN INVASIVE WEEDS CLASSIFICATION BASED ON VIC/WSNIR TECHNIQUES

PENGCHENG NIE[1,2], JIAJIA YU[1], YUN ZHAO[1],YONG HE[1]

[1] P. Nie, College of Biosystems Engineering and Food Science, Zhejiang University268 Kaixuan Road, Hangzhou, 310029, China, yhe@zju.edu.cn
[2] School of information Engineering, Nanchang Hangkong University, Nanchang 330069, China
[1] J. Yu, yjjxdg@zju.edu.cn
[1] Y. He, yhe@zju.edu.cn

CSBE101446 – Presented at Section VII: Information Systems Conference

ABSTRACT In this paper, the feasibility of visible and short-wave near-infrared spectroscopy (VIS/WNIR) techniques as means for the non-destructive and fast detection of alien invasive plants was applied. Selected sensitive bands were validated by successive projections algorithm (SPA) and soft independent models of class analogy(SIMCA) separately. The SPA- discrimination model measured and predicted Correlation coefficient is 0.926. And the SPA- SIMCA discrimination model measured and predicted Correlation coefficient is 0.931. The SPA- PLS-SVM discrimination model measured and predicted Correlation coefficient is 0.981. It was showed above indicated that the selected wavelengths validated by SPA can delegate the main characteristics of three weeds, including Veronica Persica, Veronica potita , Veronica arvensis Linn based on Vis/WNIR spectroscopy.

Keywords: Biological invasion, invasive plants, VIS/WNIR, successive projections algor ithm(SPA), LS-SVM.

INTRODUCTION Biological invasion is hot spots and difficult problems in the recently time. As the Veronica, eronica didyma Tenore is a invasive species that has strong ability of reproduction and diffusion.It has many varieties, some are precious traditonal chinese medcine while some are farmland weeds of serious damage.In the past 20 years, the invasion of alien organisms has increased at the speed of 1 to 2 kinds each year in China. Because of the strong ability to reproduce, the Invasive plants bring about serious economic, ecological, and even human health risks to the invasion region. So to find invasive plant identification techniques has a great meaning to early warning, reduce the speed of alien plants' invasion.

In this paper, We used the visible near-infrared spectroscopy rapid and non-destructive identification of three different types of Veronica . And used the recently developed methods for variable selection is successive projections algorithm (SPA) employs simple projection operations for variable selection with minimum of collinearity and

redundancy. by successive projections algorithm (SPA) and soft independent models of class analogy(SIMCA) and partial least squares (PLS) analysis, This kind of combination was helpful for the interpretation of the developed models like SPA- SIMCA. But the accuracy and prediction precision of the model would be impaired to some extent without considering the latent nonlinear relevant information in the spectral data, although SPA-MLR model performed as well as full-spectrum PLS model in some case. So, We used a new combination of SPA with least squares-support vector machine (LS-SVM) was proposed as a nonlinear calibration model for quantitative analysis using spectroscopic techniques.

MATERIALS AND METHODS In this experimentation, the Handheld Field Spec Spectrograph form Analytical Spectral Device Corporation (ASD Cor., USA) was applied, together with a 150 W halogen lamp. This spectrograph has high sensitivity range from 325 to 1075 nm with the visual angle of the spectrograph. The interval of sampling is 1.5 nm, and the sensitivity is 3.5 nm. The spectrum datum was translated into datum in the format of ASCII code. The software of ASD View Spec Pro, Unscramble V9.0 and MATLAB 7.0 software was adopted in the research.

The samples and measurements. Total of 200 samples of three typical kinds of Veronica form the farm of Zhejiang university, Has 50 samples as predict samples. in order to reduce the error of operation, the uniform glass vessel (diameter: d = 65 mm, height: h = 1.4 cm) was adopted to load the Veronica which covered the bottom of the vessel, and the spectrograph was fixed 120 mm above the surface of the Veronica with the visual angle of 25cm of the spectrograph. And for each reflection spectrum, the scan number was 30 times at exactly the same position with the spectrophotometer.

spectral data preprocessing Before the calibration stage, the spectral data should be preprocessed for an optimal performance. Firstly, the transmissivity is

$$T(\%) = \text{transmission spectroscopy} / \text{background spectroscopy} \quad (1)$$

Successive projections algorithm SPA is a forward variable selection algorithm applying vector projection operations in a vector space for the selection of relevant variables with small collinearity formultivariate calibration. In the algorithm, the instrumental response data are disposed in a matrix X of dimensions ($N \times K$). such that the *k*th variable xk is corresponding to the *k*th column vector be the maximum number of selected variables used in later calibration models. Firstly, the projections are carried on the X matrix,which generate k chains of M variables each. Each element in a chain is selected in order to display the least collinearity with the previous ones. The construction of each chain starts from one of the variables x_k, k=1,. . ., K, and follows a comparison step of projections until the need relevant variables are selected. The details of these steps could be found in the previous studies. Then the selected variables, named EWs,were used as the inputs of SIMCA, and LS-SVM models.

partial least squares analysis Partial least squares (PLS) analysis is widely used in present infrared spectroscopy analysis. PLS can be performed prediction of protein of Veronica base on linear regression model. Essentially, the PLS is considered as principal

component Analysis plus Canonical Correlation Analysis and multi linear regression. PLS can also extract the latent variables. The LVs were considered as new eigenvectors representing the original spectra, it is considered that compress the original spectra data effectively. In the development of PLS model, full cross-validation was used to evaluate the quality and to prevent over-fitting appearance of calibration model. choosed the finite the LVs which can explain the variance of the original spectra data in maximum limit. Subsequently the selected LVs were employed as the inputs of least squares-support vector machine (LS-SVM).

Least squares-support vector machine Original the aim of Least squares-support vector machine (LS-SVM) is try to classify the object in linear method which can not implement in original input space. Later this method also can implement regression analysis. LS-SVM try to establish the linear model in the high dimension feature space with the lower input vector mapped to the high dimension vector using the map function $\varphi(x)$ at the case of the vector X and vector Y can not regression very well in the lower dimension space. The rationale of LS-SVM algorithm is introduced as follows. A set of training data is defined as (x_i, y) with the n-dimensional inputs $x_i \in R^n$ and the outputs $y \in R$. The map function $\varphi(x)$ maps the input vector into a high dimensional feature space. So in the high dimension the linear regression model describe as following

$$y = \omega^T \varphi(x) + b \tag{2}$$

Where $x_i \in R^n$ is the weight vector and b is the bias. A optimization problem is formulated in the principle of structural risk minimization (SRM).

$$\min J(\omega, e) = \frac{1}{2}\omega^T \omega + \frac{1}{2}\gamma \sum_{i=1}^{n} e_i^2 \tag{3}$$

With constraints,

$$y_i = \omega^T \varphi(x) + b + e_i \quad i = 1, ..., n \tag{4}$$

The LS-SVM regression model can be expressed as

$$y(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b \tag{5}$$

Where I refers to an $[n \times n]$ identity matrix, γ is the vector of reference values. $K(x_k, x_l)$ is the kernel function, which must follow Mercer's condition and perform the linear and nonlinear mapping. RBF kernel function was used in this study shown as follows:

$$K(x_k, x_l) = \exp\left(\frac{-\|x_k - x_l\|^2}{2\sigma^2}\right) \tag{6}$$

All the calculations above were performed using MATLAB 7.0 (The Math Works, Natick, USA). The free LS-SVM toolbox (LS-SVM V 1.5, Suykens, Leuven, Belgium) was applied with MATLAB 7.0 to develop the LS-SVM models.

RESULTS AND DISCUSSION As show in Fig. 1, these were the typical curves of the reflectance spectra of the three varieties Veronica. In this research,ASD View Spec Pro was applied to investigate the spectra that wavelength ranging from 400 to 1000 nm. All the achieved spectra data were averaged, and changed it to the ASCII code, That was used for building the reflectivity matrix for later use. It can be found that the spectra of eight teas are not remarkable difference in the spectral range. After comparing in detail, some tiny difference can be detected from 700 to 780 nm, which makes it possible to discriminate the difference varieties.but there were some crossovers and overlapping among these samples. Therefore, different varieties of instant Veronica with differentn internal qualities could be reflected in the Vis/NIR spectra.Sugar content is one important internal quality of instant Veronica.

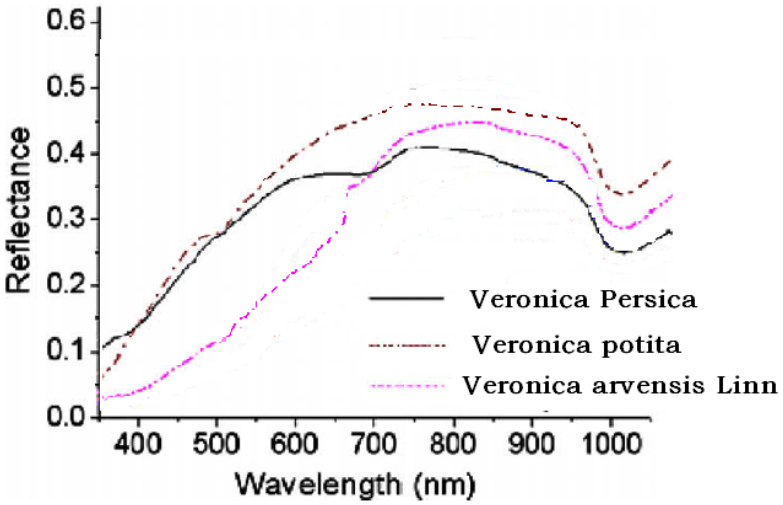

FIG.1 Vis／NIR absorbance spectra of three different Veronica

Successive projections algorithm was commonly used in multivariate technique to reduce dimension, it has applied to eliminate the repetitive information under the precondition of no information loss. In this paper, Successive projections algorithm was used to cluster the different three varieties of 150 Veronica samples, As shown in figure 2. the SPA- discrimination model measured and predicted Correlation coefficient is 0.926. And the SPA- SIMCA discrimination model measured and predicted Correlation coefficient is 0.931.As is shown in figure 4. the SPA- LS-SVM discrimination model measured and predicted Correlation coefficient is 0.981.was shown in figure 3.

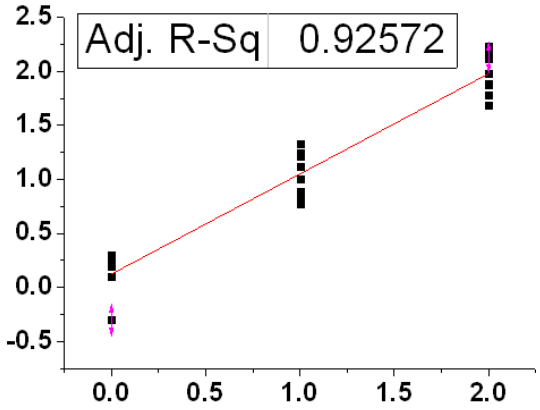

Figure 2. Correlation coefficient between measured and predicted valuesof 50 unknown Veronicas by using SPA model

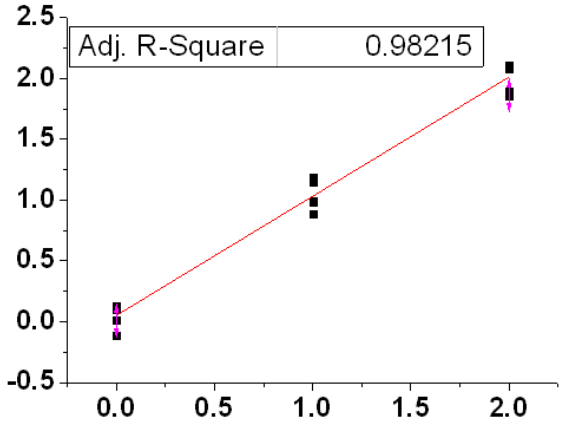

Figure 3. Correlation coefficient between measured and predicted values of 50 unknown Veronicas by using SPA- SIMCA model

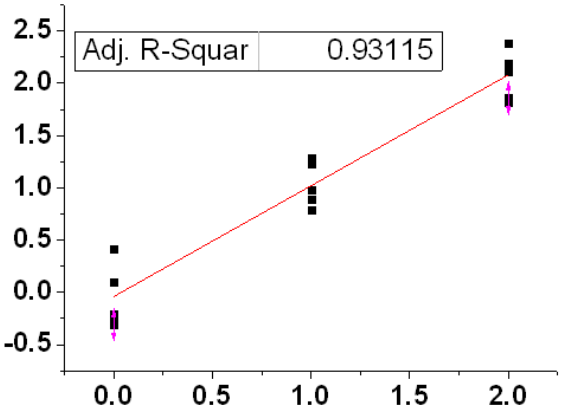

Figure 4. Correlation coefficient between measured and predicted values of 50 unknown Veronicas by using SPA-LS-SVM model

CONCLUSION In this work, It has been demonstrated that it is possible to develop a nondestructive technique for discrimination of Veronica. And provide a new pattern recognition method. the SPA combined with PLS and SVM was used to establish the discrimination model and showed an excellent prediction performance. In this research, The SPA- discrimination model measured and predicted Correlation coefficient is 0.926. And the SPA- SIMCA discrimination model measured and predicted Correlation coefficient is 0.931. The SPA- LS-SVM discrimination model measured and predicted Correlation coefficient is 0.981. All data showed above indicated that the selected wavelengths validated by SPA can delegate the main characteristics of three weeds,including Veronica Persica, Veronica potita , Veronica arvensis Linn based on Vis/WNIR spectroscopy.

Acknowledgements. This work was supported by the 863 National High Technology Research and Development Program of China (2007AA10A210)

REFERENCES

D.C. Albach and M.W. Chase, Paraphyly of Veronica (Veroniceae; Scrophulariaceae): evidence from the internal transcribed spacer (ITS) sequences of nuclear ribosomal DNA, J. Plant Res. 2001,114 (1): 9–18.

D.C. Albach and M.W. Chase, Incongruence in Veroniceae (Plantaginaceae): evidence from two plastid and a nuclear ribosomal DNA region, Mol. Phylogenet. 2004,32 (11): 183–197.

Barnes, R. J., M. S. Dhanoa, and S. J. Lister. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Appl. Spectrosc. 1989, 43(5): 772-777.

Woo, Y. A., H. J. Kim, K. R. Ze, and H. Chung. Near-infrared (NIR) spectroscopy for the non-destructive and fast determination of geographical origin of Angelicae gigantis Radix. J. Pharm. Biomed. Anal. 2005,36(5): 955-959

Van Gestel, T., J. A. K. Suykens, B. Baesens, S. Viaene, J.Vanthienen, G. Dedene, B. de Moor, and J. Vandewalle. Benchmarking least squares support vector machine classifiers.Mach. Learn. 2004,54(1): 5-32.

Tsenkova, R., S. Atanassova, K. Toyoda, Y. Ozaki, K. Itoh, and T.Fearn. Near-infrared spectroscopy for dairy management:Measurement of unhomogenized milk composition. J. Dairy Sci. 1999,82(11): 2344-2351.

Fassio, A., and D. Cozzolino. Non-destructive prediction of chemical composition in sunflower seeds by near-infrared spectroscopy. Ind. Crop. Prod. 2004,20(3): 321-329.

Li, J. Z., H. X. Liu, X. J. Yao, M. C. Liu, Z. D. Hu, and B. T. Fan. Structure-activity relationship study of oxindole-based inhibitors of cyclin-dependent kinases based on least-squares support vector machines. Anal. Chim. Acta. 2007, 581(2): 333-342.

Jacobsen, S., I. Sondergaard, B. Moller, T. Deslerc, and L. Munck. A chemometric evaluation of the underlying physical and chemical patterns that support near-infrared spectroscopy of barley seeds as a tool for explorative classification of endosperm genes and gene combinations. J. Cereal Sci. 2005,42(3): 281-299.

He, Y., Feng, S. J., Deng, X. F., & Li, X. L. Study on lossless discrimination of varieties of yogurt using the Visible/NIR-spectroscopy. Food Research International. 2006,39(6), 645–650.

Suykens, J. A. K., J. Vandewalle. Least squares support vector machine classifiers. Neural Process. Lett. 1999,(3): 293-300.

Qi, X. M., Zhang, L. D., & Du, X. L.. Quantitative analysis using NIR by building PLS-BP model. Spectroscopy and Spectral Analysis,2003.23(5), 870–872.