NIR Spectroscopy Identification of Persimmon Varieties Based on PCA-SVM

Shujuan Zhang, Dengfei Jie, Haihong Zhang
zsujuan@263.net

Abstract. In order to achieve non-destructive measurement of varieties in persimmon, a fast discrimination method based on Vis / NIRS spectroscopy was put forward. A Field Spec 3 spectroradiometer was used for collecting 22 sample spectra data of the three kinds of persimmon separately. Then principal component analysis (PCA) was used to process the spectral data after pretreatment. The near infrared fingerprint of persimmon was acquired by principal component analysis(PCA) ，and support vector machine (SVM) methods were used to further identify the persimmon separately. The result of PCA indicated that the score map made by the scores of PCI，PC2 and PC3 was used, and 8 principal components (PCs) were selected as the input of support vector machine (SVM) based on the reliabilities of PCs of 99．888 ％.51 persimmon samples were used for calibration and the remaining 15 persimmon samples were used for validation. A one-against- all multi-class SVM model was built, and the result showed that SVM possessing with the RBF kernel function has the best identification capabilities with the accuracy of 100％. This research indicated that the mixed algorithm method of principal component analysis(PCA) and support vector Machine(SVM) has a good identification effect, and can work as a new method for quick, efficient and correct identification of persimmon separately.

Keywords: Vis-NIR Spectroscopy, support vector machine (SVM), Persimmon, Principal component analysis (PCA)

1 Introduction

Visible_Near Infrared Diffuse Reflectance (Vis_NIR) Spectroscopy analysis technology can be used for qualitative and quantitative data analysis with full band or multi-wavelength spectrum. With many advantages of the spectra technology, such as dispense with pre-treatment, fast, pollution-free, no damage, multi-component analyzing synchronously, good reproducibility and online analysis, it has been widely used in quality detecting studies of agricultural products[1-10]. At present, some scholars have researched the Varieties of some kinds of fruit by near-infrared spectroscopy, for example, Peach [2] 、Soy sauce [3]、Yogurt[4]and Juice [5]. But there is no research paper to study the varieties of persimmon.

Persimmon, as a characteristic fruit of our country, has many characteristics, such as artistic contour, sweet, succulent and rich nutrition. It contains a lot of carotene, vitamin C, glucose, fructose, calcium, phosphorus, iron and other minerals. So persimmon has very high culinary and medicinal properties. It can be used to eat as the fresh food, to make wine and to make vinegar. It also can be processed into persimmons biscuits, dried persimmons and persimmon cake. Therefore, the persimmon enjoys the “holy in the fruit” of reputation. In recent years, with the rural industrial structure adjustment, the persimmon has become an important source of increasing income for farmers in some areas. However, as the case stands, the domestic and foreign markets of fresh persimmon sales have not developed fully. As native products, the storage and grading of persimmon can not meet the market demands. Therefore, the developments and research of detecting technology for the quality of persimmon are important， In particular, Quality detection of persimmon for Simple, rapid, nondestructive is very necessary. Initial establishment of persimmon varieties of forecast model was calculated, using Near-Infrared Spectroscopy, based on principal component analysis and support vector machine combined data mining approach.

2. Materials and Methods

2.1 Software and analysis equipment

The experiment equipment is composed by computer, spectrometer, and halogen lamp and correction whiteboard. The spectrum sampling has been collected by diffuse reflectance methods using the FieldSpec3 spectrometer made by America ASD (Analytical Spectral Device) company. The interval of spectral sampling is 1nm, the range of sampling is 350 ~ 2500nm, the times of scanning are 30, the resolving power is 3nm and the view angle of probe field is 25°. The halogen light of 14.5V is used to match with the spectrometer. Spectral data is exported into the form of ASCII code for processing and the analysis software is the ASD View Spec Pro, Unscramble V9.7, Matlab7.3 and DPS（Data Procession System for Practical Statistics）.

2.2. Sample source and spectral data acquisition

Three kinds of persimmon including “covered persimmon”, “Cynanchum persimmon” and “red persimmon” have been purchased from the market. The 22 samples have been collected for each kind of persimmon and the total are 66 samples. All the samples are randomly divided into modeling sets including 51 samples and prediction sets including 15 samples. The spectral data are collected by spectrometer located at the top of persimmon. The distance between the top of persimmon and spectrometer is 50mm, the view angle of probe field is 25° and the time for scanning each sample is 10.

2.3. Spectral data preprocessing

In order to remove the high-frequency random noise, baseline drift, sample asymmetry and the impact of light scattering, the move average smoothing method with 9 smoothing points was used for spectral pre-processing, which can remove the high frequency noise effectively. Then the data was processed by MSC (Multiplicative Scatter Correction) [6]. All of the pretreatment process carried out in the Unscrambler V9.6. The typical near-infrared spectroscopy curve of three kinds of persimmon is shown in Fig.1

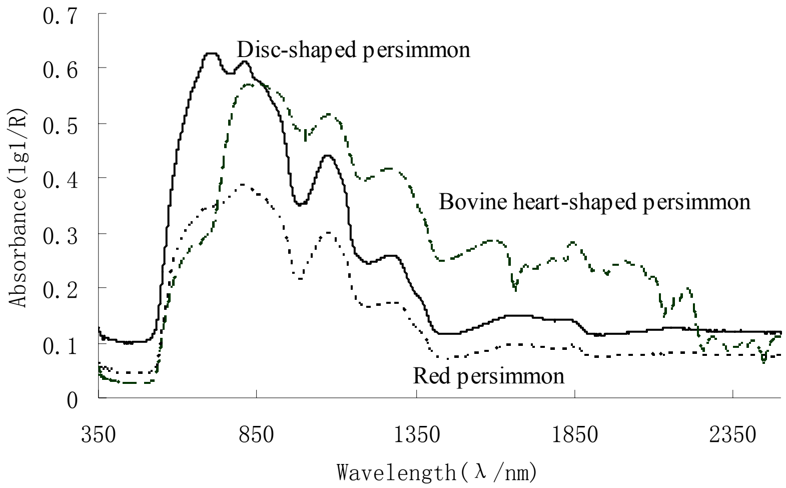

Fig. 1. Visible_Near infrared reflectance spectroscopy of three different varieties of persimmon

2.4. Principal component analysis of three varieties of persimmon

Spectral curve has a larger noise form the first side to the end. Spectrum of 350 ~ 2050nm band has been selected for eliminating noise. PCA (principal component analysis) was used in this study. In the Fig.2, X-axis samples the PC1 (first principal component scores), Y-axis samples the PC2 (second principal component scores), Z-axis samples the PC3 (third principal component scores). Classification performance of persimmon is shown from the Fig.2, and the characteristics curve of three different varieties of persimmon was described. However, the distinction wasn't obvious between the edges of the sample. To improve prediction accuracy, the prediction model of three persimmons was established using SVM and PCA in this experiment.

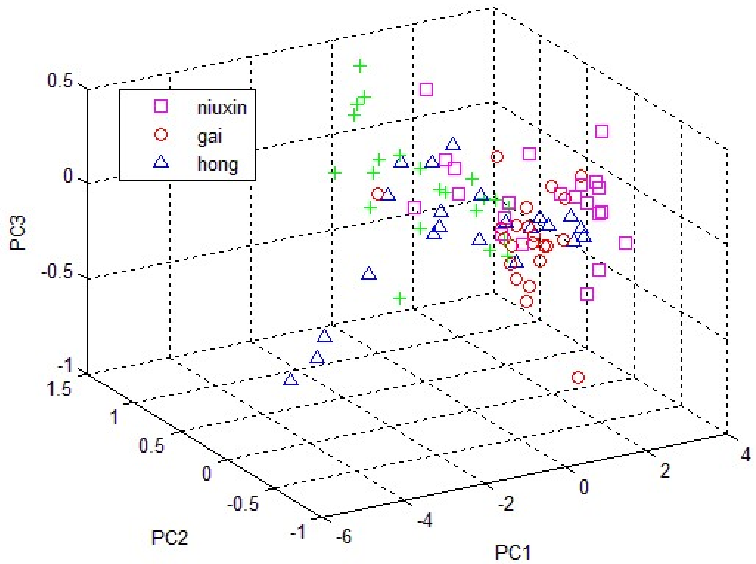

Fig.2. Principal Component scores scatter plot (PC1×PC2×PC3) of persimmon

3.Test results and analysis

3.1. Principal component analysis

Sample spectral bands have total 2150 points from 350 ~ 2500nm.Using full spectrum, computation is large, spectral information is weak in some of the regional samples and correlation is lack between the composition of the sample and the nature. The purpose of PCA is data reduction. Under the premise of more variable spectral information without loss of participate, the original variables are replace using a smaller number of new options, and many information of overlapping chemical are excluded[11]. Therefore, the Unscramble V9.6 software is used, principal component of three persimmons were analyzed. The accumulated credibility of first 8 principal components of was got in Table 1.It has reached 99.888%. So the original wavelength can be interpreted by the 8 principal components.

Table.1 Accumulative reliabilities of the first 8 Principal components

PCs	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Accumulative reliabilities /%	84.988	95.515	97.742	99.301	99.550	99.722	99.821	99.888

3.2. Support Vector Machine Identification of persimmon

SVM is a new pattern recognition method. The principle risk minimization of structural is uses, Training error and generalization are both, and many advantages are expressed in solving small sample, nonlinear, high dimension and local minimum problems. The 51 samples of the principal component scores as SVM training set, the remaining 15 samples were taken as the prediction set of SVM, and the radial basis function (RBF) is used as a core function to identify varieties of persimmon. Optimal parameter is found using Grid-search function. Namely, main parameters of gamma (γ) and the RBF of kernel function in the sig2 (σ2) are found. The appropriate gamma (γ) can increase the ability of SVM model. The parameter sig2 (σ2) can control the error of regression function, and directly affect the noise immunity of the SVM model. Generalization ability, as learning and prediction of the model, are largely determined by those two parameters. The SVM is realized by the Matlab 7.3. Classification performance of the RBF kernel function was listed in the Table 2.

Table 2. Results of PLS model with different spectrum

Sample number	Real value	Predicted value	Bias
1	1	0.9608	-0.0392
2	1	1.0566	0.0566
3	1	1.0131	0.0131
4	1	0.9829	-0.0171
5	1	1.0191	0.0191
6	2	2.0263	0.0263
7	2	1.9957	-0.0043
8	2	2.0493	0.0493
9	2	1.9331	-0.0669
10	2	1.9269	-0.0731
11	3	2.9930	-0.0070
12	3	3.0981	0.0981
13	3	3.0272	0.0272
14	3	2.9040	-0.0960
15	3	2.9606	-0.0394

From Table 2, the forecast error of the SVM model was less than ± 0.1, and the recognition rate has reached to 100%. Compared with the conventional model of discriminate analysis, the SVM model of near infrared spectral classification was more accurate.

4. Conclusion

Identification of persimmon varieties were studied based on Near Infrared Spectroscopy visible. On the basis of Visible-near infrared spectroscopy and spectral data pretreatment, the variety identification model of persimmon was established using PCA and SVM. The Principal Components of PCA as input set, the Gaussian radial basis (RBF) kernel function of SVM as choose, the model of varieties of persimmon has been established. Experimental results show that the model prediction error of unknown varieties of persimmon was less than ± 0.1, and the recognition rate has reached to 100%. The model was accurate, predicted better results, and improved the speed and accuracy. New way of persimmon detection was provided for future spectroscopy.

Acknowledgements

Funding for this research was provided by Shanxi Provincial Department of Science and technology (NO.2007031109-2).

References

1. Zhao Jiewen, Zhang Haidong, Liu Muhua. Non-destructive determination of sugar contents of apples using near infrared diffuse reflectance [J]. Transactions of the Chinese Society of Agricultural Engineering, 2005,21(3):162～165.
2 .Li Xiaoli, Hu Xingyue, He Yong. New approach of discrimination of varieties of juicy peach by Near Infrared spectra based on PCA and MDA model [J]. Journal of Infrared and millimeter waves,2006,25(6):417～420.
3. Tong Xiaoxing, Bao Yidan, He Yong. Study on Fast Discrimination of Brands of Soy Sauce Using Near Infrared Spectra [J]. Spectroscopy and Spectral Analysis, 2008, 8(3):597-601.
4 .He Yong, Feng Shuijuan, Li Xiaoli, et al. Study on Fast Discrimination of Varieties of Acidophilous Milk Using Near Infrared Spectra [J]. Spectroscopy and Spectral Analysis ,2006,26(11):2021～2023.
5..Zhang Haihong, Zhang Shujuan, Jie Dengfei, et al. Research on Varieties of Sea Buckthorn Juice by Near-Infrared Diffuse Reflectance Spectroscopy Based on the PCR and PLS [J]. Journal of Shanxi Agricultural University(Nature Science Edition), 2010,30(1):46～48.
6. Li Guifeng, Zhao Guojian, Wang Xiangdong, et al. Nondestructive measurement and fingerprint analysis of apple texture quality based on NIR spectra [J]. Transactions of the Chinese Society of Agricultural Engineering, ,2008,24 (6):169～173.
7. Zhang Shujuan, Wang Fenghua, Zhang Haihong, et al. Detection of the Fresh Jujube Varieties and SSC by NIR Spectroscopy [J]. Transactions of the Chinese Society for Agricultural Machinery, 2009,40(4):139～142.
8. Li Xiaoli, He Yong, Qiu Zhengjun. Application PCA-ANN Method to Fast Discrimination of Tea Varieties Using Visible/Near Infrared Spectroscopy [J]. Spectroscopy and Spectral Analysis, 2007,27(2):279～282.

9. Liu Yande, Sun Xudong, Chen Xingmiao. Research on the Soluble Solids Content of Pear Internal Quality Index by Near-Infrared Diffuse Reflectance Spectroscopy [J]. Spectroscopy and Spectral Analysis, 2008,28(4):797～800.
10. Wang Gang, Zhu Shiping, Kan Jianquan, et al. Nondestructive Detection of Volatile Oil Content in Zanthoxylum Bungeagum Maxim by Near Infrared Spectroscopy [J]. Transactions of the Chinese Society for Agricultural Machinery, 2008,39(3):79～85.
11. He Yong, Li Xiaoli, Shao Yongni. Discrimination of Varieties of Apple Using Near Infrared Spectra Based on Principal Component Analysis and Artificial Neural Network Model [J]. Spectroscopy and Spectral Analysis, 2006, 26(5): 850～853.
12. Li Guozheng, Wang Meng, Zeng Huajun. An Introduction to Support Vector Machines and other Kernel-Based Learning Methods [M]. Beijing: Publishing House of Electronics industry, 2004, 3.
13. Deng Naiyang, Tian Yingjie. New Method of Data Minning_Support Vector Machine[M]. Beijing: Science Press, 2004, 6.