Modeling research on wheat grain in different quality categories using multilayer support vector machine

Guangyan Hui[1, a], Laijun Sun[1, b*] and Shang Gao[1, c]

[1] Key Laboratory of Electronics Engineering, College of Heilongjiang Province, Heilongjiang University, Harbin, China

[a]1271364831@qq.com, [b]slaijun@126.com, [c]771375763@qq.com

Keywords: near infrared reflectance; wheat; classification; partial least squares based dimension reduction; support vector machine

Abstract. Near Infrared Reflectance (NIR) spectroscopy is a 'green' nondestructive testing technology and it has been widely used in grain crop analysis. The experimental data were collected using 161 wheat samples from the major wheat-producing area in China. The original spectral data was represented by four characteristic variables extracted by Partial Least Squares based Dimension Reduction (PLSDR). Besides, Mahalanobis distance method, second derivative and SNV were used to preprocess spectra. A two-tier classification model based on SVM algorithm was used to achieve the classification of wheat quality. The experimental results indicated that the two-tier SVM classification model was effective in identifying the quality of wheat grain with the recognition rates of common, strong-gluten, middle-gluten and weak-gluten wheat samples being 93.3%, 87.5%, 72.7% and 92.3%, respectively, and the rejection rates of them being 90.0%, 97.4%, 100.0% and 95.2%, respectively. The model realized rapid and accurate classification of wheat, besides it could be applied to the detection system of wheat quality.

Introduction

Wheat is one of the main food crops and widely cultivated over the world. In China, wheat is classified into three categories: strong, middle and weak gluten wheat, which are given based on the wheat gluten strength and different practical uses[1]. Each type of wheat is good for producing suited food correspondingly, which can positively improve the quality of food products. The wheat which can be distinguished explicitly is called high quality wheat whose price is significantly higher than other common wheat because of its specific function. At present, growers chose the wheat varieties according to the yield of wheat or economic interests, because there was no uniform plan for the wheat varieties macroscopically nationwide. The grain storage and warehouse companies didn't store wheat according to its varieties. These two factors led that different types of wheat was blended together, which caused high quality wheat lesser. At the same time, wheat is stored to corresponding storeroom based on its categories, which will avoid the confusion of different types of wheat, and calculate price depending on wheat quality in purchasing and marketing wheat. All of those factors can promote the development of good marketing environment, and defend the benefit of farmers. Therefore, it is extremely important to classify wheat with gluten strength. Currently, the traditional physical and chemical analysis method still be applied to evaluate the types of wheat at home and abroad, which need to detect too many indicators and has complex testing process. Besides, sorts of special equipment and specialized persons are always necessary in this procedure. It can find that this method not only expends a lot of experimental materials and time, but also requires abundant manpower and financial resources. This phenomenon exists generally in the process of quantitative and qualitative analysis of food crops.

Aiming at solving the problems mentioned above, domestic and foreign researchers sought to discover simple methods to complete the quantitative and qualitative analysis with short time, and recognize real-time monitoring of the quality of crop. Moreover, some notable achievements have been acquired. Near infrared reflectance (NIR) is a technique that has been proposed as an excellent method to the traditional due to its advantages of rapidity, non-destruction, free-pollution if proper calibration and validation model is built and it has been widely used for the quantitative and qualitative analysis. B.Gaspardo et al. [2] researched and applied FT-NIR technology with PLS for the detection of fumonisins B1 and B2 in corn meal, the correlation coefficient of predicted value and real value reached 0.964, RMSEP was 0.630. In recent years, people pay more and more attentions to the quality classification based on spectroscopy technique. In Cocchi' [3] research, various wheat flour samples belonging to four different ISO classes have been analyzed by means of NIR spectroscopy and each sample of bread wheat flour has been classed to the corresponding baking category. Haiyan Zhao et al. [4] used the PLS discriminant analysis to distinguish and analyze the geographical origin of wheat, correct classification rates were above 85%. Based on the analysis above and a great number of references [5-7], it is entirely feasible to build classification model according to the requirements of actual operation for wheat classification.

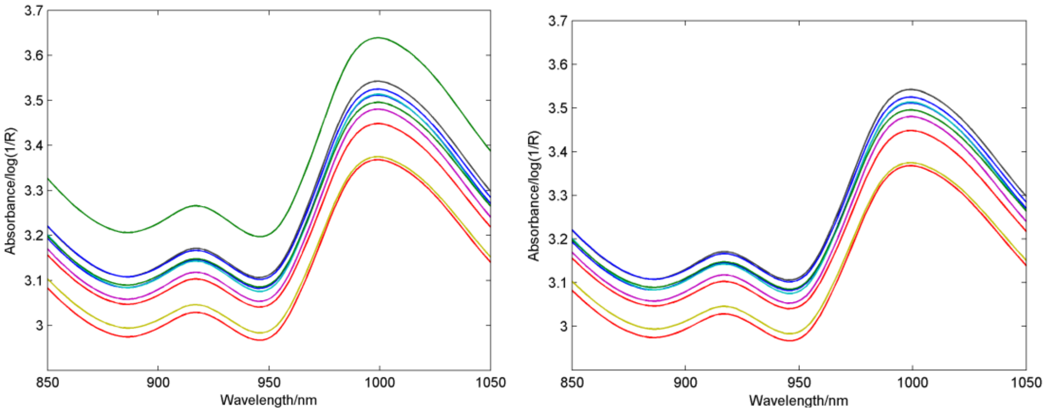
Materials and methods

Sample preparation. Samples were obtained from the Academy of Agricultural Sciences of each production region to Ministry of Agriculture of Agricultural Products Quality Safety Risk Assessment Laboratory in 2012. Some samples were winter wheat varieties from the upper and middle-lower level of Yangtze River, and others were spring wheat varieties from the northeast and northwest of China. Wheat samples may contain stones, clod, stem leaf, glume, metal and other debris which will affect the results. Hence, first of all, the impurities must be removed from each sample by the way of manual sorting before scanning, and the imperfect wheat grains also be singled out from each sample at the same time, such as shriveled and broken grains.

Acquisition of data. For the acquisition of spectra an Infratec 1241 grain quality analyzer produced by FOSS was used. The working parameters of this instrument were set as follows: the scanning temperature was controlled between 21℃ to 25℃, the range of spectra was from 850 nm to 1050 nm, scanning step was 2 nm, the number of stator was 10, and each spectrum has 100 data points. Each sample was scanned for 10 times, the average of ten times was used as the final spectra data. The classification of wheat gluten strength was based on the national standards GB/T 17320-1998" Evaluation of Wheat Quality for Specific Use", there were eight quality indicators needed to be tested. There were 41 strong-gluten samples, 25 middle-gluten samples, 45 weak-gluten samples and 50 common samples among 161 wheat samples.

Results and Discussion

Spectral data preprocessing. The spectral data collected by NIR spectrometer contains some other irrelevant information and noise besides sample information which will bring many neative effects to classification accuracy. It is necessary to preprocess the original spectra before modeling. Spectral preprocessing techniques are required to reduce or eliminate the influence of these disturbance factors in order to improve the accuracy of classification model. In this study, firstly, the Mahalanobis distance discriminant method was used to filter and remove abnormal spectra from ten scanning spectral curves of each sample. Ten scanning spectral curves of one sample were shown in Fig.1-a. From the figure, it could evidently find that there was a lager distance between this spectrum and other nine, which might be caused by impurities and other random factors. The abnormal spectral might be unsuitable for an experiment. After eliminating abnormal spectra using Mahalanobis distance method, the residual nine spectra was shown in Fig.1-b. The average value of residual spectra was used as the standard spectral data of this sample. Each sample had been made with same treatment as above, then a total of 161 sets of spectra were obtained. For those 161 sets of spectra, the next preprocessing includes second derivative and SNV transformation. Second derivative is the most common and effective pre-treatment technology which is mainly used to solve baseline drift and well weaken, and eliminate the interference of the spectra that caused by all kinds of non-target factors[8]. SNV transformation calibrates spectra with the average and standard deviation of each spectrum, which can eliminate the effects carried by scattering interference.



1-(a) 1-(b)

Fig.1 Remove the anomalous spectra

Reducing data dimensionality. The transmission spectra of wheat samples ranged from 850nm to 1050nm, with 100 points. The more the wavelength points, we would to process larger amount of data and longer time would be costed. The information in sample spectra can reflect the characteristic of samples, but it was overlapping in some areas and some lacked correlation with the sample's composition, which would increase the complexity and affect the performance of the model. Consequently, it is necessary to compress the original spectral data before modeling.

Partial least squares based dimension reduction (PLSDR) is a high performance method for reducing dimensionality of complex data [9-10]. It uses a supervised mode to extract the potential characteristic variables from the original data space instead of original spectral data, which not only reduces the modeling time but also improves the accuracy of identification. In addition, the number of the best potential characteristic variable was determined by the prediction residual error sum of square (PRESS) of cross-validation method. Applying PLSDR to deal with the original spectral data, the PRESS had reached a minimum value 0.059 when the number of the potential characteristic variable took 4 judged by cross-validation method.

Multi-layer SVM model for classification. 161 wheat samples were grouped into two sets, calibration set and prediction set, they were used to build model and test the performance of model, respectively. Not only do wheat samples have difference in internal compositions, but the factors of year, region, variety and others are also different. In this study, in order to get the typical sample set for modeling, 161 wheat samples were divided by artificial selection at a ratio of 2:1. 106 representative samples screened from 161 wheat samples formed a calibration set and the remaining 55 samples were used as the prediction set. Calibration set contains 25 strong-gluten samples, 12 medium-gluten samples, 34 weak-gluten samples and 35 common samples; prediction set contains 16 strong-gluten samples, 11 medium-gluten samples, 13 weak-gluten samples and 15 common samples. The distribution of calibration set and prediction set were shown in Fig.2, it must be pointed out that the samples in calibration set are uniformly distributed throughout the entire sample space, which explained that the calibration set selected by artificial selection can positively reflect the information in the characteristic space of all samples. The spectral data of calibration set after preprocessing and reducing dimension was used as the input data of modeling. The results of physical and chemical analysis of calibration set were used as the output data of model. A model with two-layer structure was designed for classification. Firstly, the SVM1 model was established to realize the "one to one" classification about common wheat and high-quality wheat, and then SVM2 model was built to accomplish the classification of other three types of high-quality wheat which have obvious characteristics differences. In other words, two-layer structure SVM model was constructed according to the characteristic in four types of wheat, the structure of the model was shown in Fig.3.

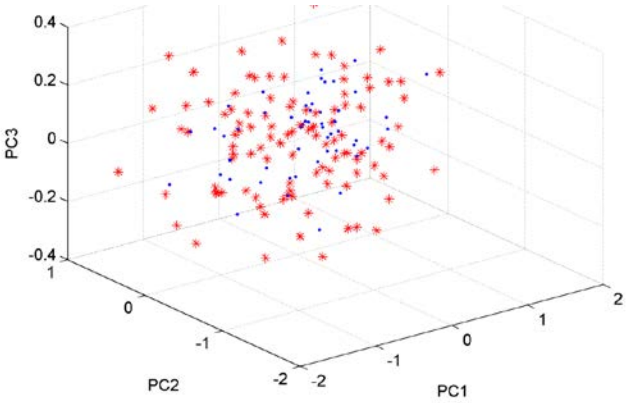


Fig.2 Distribution of calibration set and prediction set
*calibration set; ·prediction set

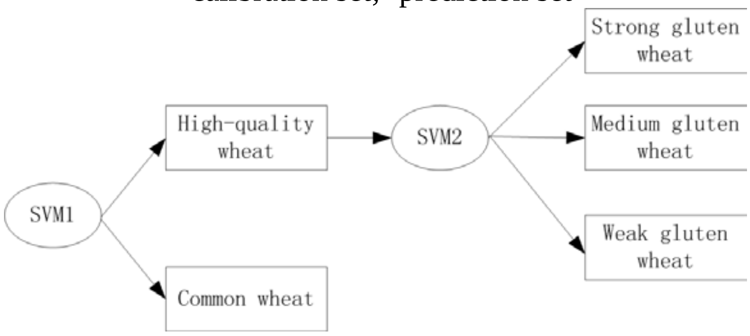


Fig.3 Structure of two-tier SVM classification model

The classification results of prediction set output from the two-tier SVM classification model were shown in Table 1. For the first-tier classification model, the recognition rate and rejection rate of 40 high-quality samples were 90% and 93.3%, respectively; for the 15 common wheat samples were 93.3% and 90%, respectively. Considering the results, it was concluded that the first-tier classification model generated good identification abilities for high-quality wheat and common wheat. The second-tier of the classification model was still established by SVM with 71 high-quality wheat samples (25 strong-gluten samples, 12 middle-gluten samples and 34 weak-gluten wheat samples). 37 high-quality wheat samples identified by the first-tier model were put into the second-tier model for further identification. The second-tier model had a good recognition effect for strong-gluten, middle-gluten and weak-gluten wheat samples with the recognition rates being 100%, 80.0% and 100% and rejection rates being 95.7%, 100% and 95.8%, respectively.

Table 1 Results of SVM classification model

Model	Class	Num of samples	Num of recognition	Num of rejection	Recognition rate (%)	Rejection rate (%)
First-tier	High	40	36	14	90.0	93.3
	common	15	14	36	93.3	90.0
Second-tier	strong	14	14	22	100.0	95.7
	middle	10	8	27	80.0	100.0
	weak	13	13	23	100.0	95.8
two-tier	common	15	14	36	93.3	90.0
	strong	16	14	38	87.5	97.4
	middle	11	8	29	72.7	100.0
	weak	13	12	26	92.3	95.2

For the two-tier SVM classification model, 55 samples in prediction set have been identified correctly, the recognition rates of common, strong-gluten, middle-gluten and weak-gluten wheat samples were 93.3%, 87.5%, 72.7% and 92.3%, respectively, the rejection rates of them were 90.0%, 97.4%, 100.0% and 95.2%, respectively. As can be seen from the results of classification, only the recognition result of medium-gluten wheat samples was a little low while the other three types were better.

Summary

This study analyzed the feasibility of NIR spectroscopy technology for classifying wheat according to gluten strength. The test results indicated that it was feasible and the model built in this research could achieve a higher accuracy for forecasting comparing to the traditional methods. Based on the introduction of NIR spectroscopy and SVM, the multi-layer classification model was proposed to recognize the classification of wheat with gluten strength. 106 wheat samples were used to build model and other 55 wheat samples were used to test the effectiveness of the model. The results of the entire two-tier SVM classification model showed that only the recognition rate of medium gluten wheat samples was a little low. According to the testing results, the model has higher identification ability and reliability for four different types of wheat. The idea of multi-layer classification modeling in this research can be used for the qualitative analysis of other substances.

References

[1] National Standards of P. R. China “Wheat varieties for specific end-uses” GB/T 17320-1998, Beijing: national standard publishing house (1998).

[2] Gaspardo. B., Del. Zotto. S. and Torelli. Food Chemistry. 135: 1608-1612, (2012).

[3] Marina Cocchi, Maria Corbellini, Giorgia Foca, Mara Lucisano, M. Ambrogina Pagani, Lorenzo Tassi, Alessandro Ulrici. Analytica Chimica Acta. 2005. 544(1):100-107.

[4] Haiyan Zhao, Boli Guo and Yimin Wei. Food Chemistry. 138(2): 1902-1907, (2013).

[5] Marina Cocchi, Caterina Durante and Giorgia Foca. Talanta. 68(5): 1505-1511, (2006).

[6] Sindhuja Sankaran, Ashish Mishra and Joe Mari Maja. 77(2): 127-134, (2011).

[7] Giorgia Foca, Marina Cocchi and Mario Li Vigni. Chemometrics and Intelligent Laboratory Systems. 99: 91-100, (2009).

[8] Raffaele Vitale, Marta Bevilacqua and Remo Bucci. Chemometrics and Intelligent Laboratory Systems. 121: 90-99, (2013).

[9] Mohammad Goodarzi, Sandeep Sharma and Herman Ramon. Trends in Analytical Chemistry. 67: 147-158, (2015).

[10] Kabir Yunus Peerbhay, Onisimo Mutanga and Riyad Ismail. ISPRS Journal of Photogramm etry an d Remote Sensing. 79: 19-28, (2013).