Research on Rapid Identification Method of Buckwheat Varieties by Near-Infrared Spectroscopy Technique

Fenghua Wang[1], Ju Yang[1], Zhiyong Xi[1], and Hailong Zhu[2]

[1] Faculty of Modern Agricultural Engineering, Kunming University of Science and Technology, Kunming 650500
[2] Engineering Training Center, Kunming University of Science and Technology, Kunming 650500
wangfenghua018@163.com, {372086956,315314518,81065885}@qq.com

Abstract. In order to achieve the rapid identification of buckwheat varieties, and avoid buckwheat varieties mixtures, eight buckwheat varieties from different origins were identified by principal component analysis and support vector machines based on near-infrared spectroscopy. First, the buckwheat spectral information of the 120 samples have been collected using FieldSpec 3 spectrometer, and preprocessing through smooth + Multiplicative Scatter Correction (+MSC), a total of 120 sets were divided into 80 training sets and 40 prediction sets. After the principal component analysis, based on the binary tree support vector machine theory, the spectral information identification model of buckwheat varieties have been established and verified by LIBSVM package in MATLAB software. The results showed that the classification accuracy rate averaged 92.5% for eight different kinds of buckwheat by using near-infrared spectroscopy combined with principal component analysis and support vector machine. A new method for buckwheat varieties identification has been provided.

Keywords: Near infrared spectroscopy, Buckwheat, PCA, SVM, Variety identification.

1 Introduction

Buckwheat, mainly located in the alpine areas of the Loess plateau and the plateau mountains of Yunnan, Guizhou and Sichuan. It has a high nutritional and health value, known as the food treasures of "Medicine-Food"[1]. It also known as the triangle wheat, black wheat, belongs to the genus of dicotyledonous Polygonaceae, family of Fagopyrum Mill, mainly have two cultivar of buckwheat is Sweet Buckwheat (Fagopyrum esculentum Moench) and Tartary Buckwheat (Fagopyrum tararicum Gaerth). Therefore, the quality detection and the varieties identification of buckwheat has a very important significance in buckwheat planting, the deep processing industries and the quality improvement of buckwheat.

Near infrared spectroscopy analysis have the advantages of fast ananlysis without any pretreatment, no pollution, no damage, multi-component analysis, good reproducibility and suitable for online analysis, etc., which can be use to carry on the

D. Li and Y. Chen (Eds.): CCTA 2013, Part I, IFIP AICT 419, pp. 401–407, 2014.

qualitative or quantitative analysis by spectral data of the full spectrum or multi wavelength, and it is widely used in agricultural product quality testing and varieties identification [2-7]. Traditional identification method of buckwheat varieties is chemical method, it is cumbersome and destructive for buckwheat samples, and it is very useful to study a rapid nondestructive method for varieties identifying.

Without losing the main message, principal component analysis (PCA) can select fewer new variables to replace the original variables and resolve the overlapping bands which can not be analyzed, which is a mathematical method widely used in spectrum analysis [8]. Support vector machine (SVM) is a statistical theory learning algorithm, it has the advantages of high convergence speed and high accuracy than other classification algorithm, and it has been widely applied in the quality inspection of agricultural products and species identification [9]. However, the data indivisibility often appeared in multi-class prediction problems by using SVM. Therefore, this article uses principal component analysis combined with binary tree support vector machine to establish the NIR prediction model of different buckwheat varieties, to achieve the identification of buckwheat species.

2 Materials and Methods

2.1 Instruments and Materials

The spectral information was collected by FieldSpec3 spectrometer which made in U.S. ASD (Analytical Spectral Device) company.

Seven kinds of tartary buckwheat and a sweet buckwheat from Yunnan, Shanxi, Sichuan province in China were selected. They are Kunming tartary buckwheat, Luxi tartary buckwheat, Zhaotong tartary buckwheat No. 1 and No.2, Shanxi tartary buckwheat Y16, Xichang tartary buckwheat, Dali tartary buckwheat and Dali sweet buckwheat, a total of eight kinds of representative buckwheat varieties, each buckwheat selected 15 samples, a total of 120 samples. Each sample was air-dried, eliminating Impurities, particle size and consistent uniformity, and all samples were bagging sealing for spare use.

2.2 Acquisition of the Buckwheat Spectral Information

Buckwheat spectral data have been collected in outside, the test device was consist of computer, spectrometer, reflectance probe using the interior illuminant, calibration whiteboard, etc. and the Spectrometer selected FieldSpec 3 spectrometer, the sampling range is 350-2500 nm, spectral sampling interval is 1.4nm, and the resolution is 3nm. In the experiment, the 120 buckwheat samples were placed in the glass petri dish with the dimension of 90mm diameter, l5mm thickness respectively, put the reflectance probe close to the surface of the samples, and collected the sample spectral data using the

reflection mode. 15 repetitions and 10 times scanning for each sample, and then the average processing have been carried on using Viewspec pro software provided by ASD, and converted it to absorbance by log [l / R]. Then the spectral data have been exported as ASCII format and processed with ASD View Spec Pro, Unscramble9.7, SPSS16.0 and MATLAB software. Buckwheat near-infrared spectroscopy absorbance graph are shown in Figure 1. The abscissa is the wavelength (nm), and the vertical axis is an absorbance log [1 / R].

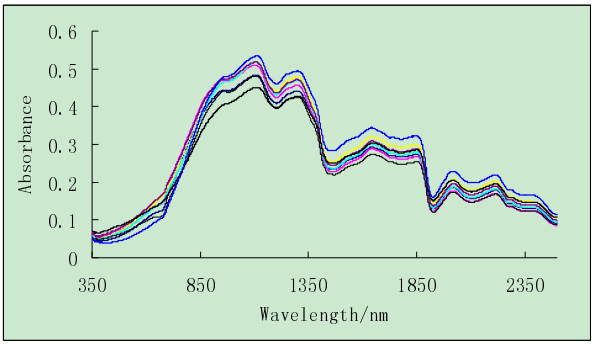

Fig. 1. Absorbance graph of buckwheat near-infrared spectroscopy

2.3 Multi-class SVM Classification Based on Binary Tree

Support vector machine is a machine learning algorithm based on the structural risk minimization (SRM) principle of statistical learning theory. For the positive and negative training samples, a hyperplane with the maximum interval has been obtained. As for the non-linear data sets, the training samples will be mapped into a high dimensional data space, to find a hyperplane in the feature space, and separate the positive and negative samples in a maximum possible [10].

Originally, SVM was designed for binary classification problems. This paper is a multi-class classification problem, using SVM classification method which based on binary tree for classification. For K types training samples, need to train K-1 support vector machine. In the first support vector machine, the first type sample as the positive training sample, and the 2,3 ..., K types of training samples as negative training samples SVM1. In the i support vector machine, the sample i as the positive training sample, and the i+1,i+2 ..., K types of training samples as the negative training samples SVMi, until the K-1 sample as the positive sample of the K-1 SVM, and the K type sample as the negative sample training SVM (K-1) [11], multi-class SVM classification process of binary tree as Figure 2:

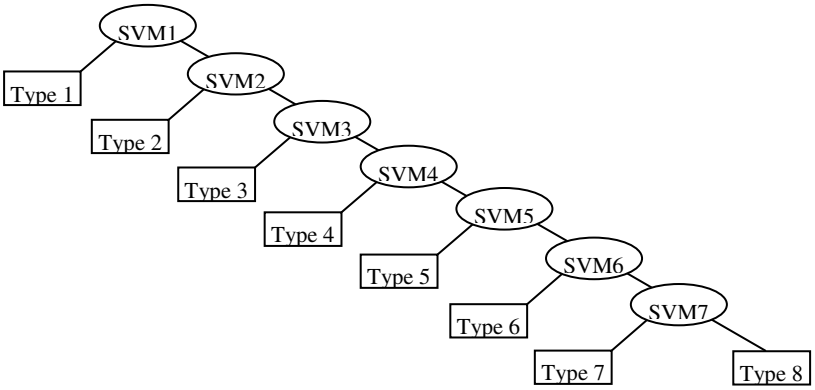

Fig. 2. SVM classification process

3 Experimental Results and Analysis

3.1 Data Preprocessing

The collected NIR spectroscopy is often influenced by high-frequency random noise, baseline drift, uneven sample and light scattering, so the spectral preprocessing is needed. Compared with different pretreatment methods, the average smoothing method was used in preprocessing of near infrared spectroscopy, the choice of the smoothing window size was 3, then multiplicative scatter correction (MSC) treatment was used, and it is good to filter high frequency noise generated by various factors.

3.2 Principal Components Extraction Based on Principal Component Analysis

The 120 samples were randomly divided into the prediction set and the validation set with 80 and 40 samples respectively. Sample spectral bands from 350 to 2500 nm, there are 2151 data in total. Computational complexity is large when using the full spectrum of computing. And due to the weak spectral information of some regional samples, the composition of the sample or the nature is lack of correlation. The principal component analysis is an effective method for data mining, it can translate the original multiple-wavelength variables into fewer new variables. These new variables are mutually independent, and it can synthetical reflect the information of original multiple-wavelength variables. [7,8]. The prediction set and the validation set were principal component analysis respectively by SPSS software, and the accumulative reliabilities of the first 7 principal components are shown in Table 1.

Table 1. Accumulative reliabilities of the first 7 PC s

Principal component	Accumulative reliabilities of prediction set/%	Accumulative reliabilities of validation set/%
PC_01	64.918	63.800
PC_02	89.035	88.092
PC_03	97.266	97.074
PC_04	98.973	98.857
PC_05	99.499	99.415
PC_06	99.865	99.836
PC_07	99.912	99.887

3.3 Buckwheat Varieties Identification Based on LIBSVM

The data arrangement have been carried on according to the multi-class SVM binary tree classification principle, and converted into a data set of .mat file for LIBSVM package. The dataset used in LIBSVM must have the attributes matrix and labels: the first column as a label, and the two to eight columns as property matrix. In 80 training sets, the 1st species tagged 1, and the rest are -1, the property matrix is representative of the principal component of each kinds of buckwheat, the attribute matrix as input, and the label as output. Based on the model, identified the No. 1 sample of the 40 samples of prediction sets; then identified the 70 training sets of the other species for the No. 2 species, and labeled No. 2 species as 1, the rest are -1 for modeling, and identified the 35 training sets for the No. 2 species; Until the SVM training for No. 7 sample, identified eight buckwheat varieties from different origin.

Through several tests the LIBSVM model parameters have been obtained, and the set as follows: Set Model Type: C-SVC, model parameter settings is -S=0 -C=1, Kernel function type: polynomial kernel function: (gamma*u'*v + coef0)^degree, Therefore,-t = 2

1) To establish the model by using svmtrain:

model = svmtrain(train_label, train_data, '-s 0 -t 2 -c 1 -g 0.1');

Input: [train_data training set attribute matrix, the size is n * m, n represents the number of samples, m represents the number of attributes (dimensions), the data type is double; train_label is the training set labels, the size is n * 1, n represents the number of samples, and the data type is double]

2) To make predictions by using svmpredict:

[predictlabel,accuracy]=svmpredict(testdatalabel,testdata,model)

Input: [test_data is the test set attribute matrix, the size is N * m, N represents the number of the test sample set, m represents the number of attributes (dimensions), the data type is double; test_label is the test set labels, the size is N * 1, N represents the number of samples, data type is double]

In this paper, eight kinds of buckwheat samples were classified, and 7 support vector machine were needed to be trained: training set recognition rate of the first support vector machine is 98.75%, and the recognition rate of prediction set is 100%.

Accuracy = 98.75% (79/80) (classification)，

Accuracy1= 100% (5/5) (classification) ；

The remaining 7 types buckwheat prediction identification were: 100%、80%、100%、80%、80%、100% and 100%。The average recognition rate of the eight buckwheat varieties is 92.5%.

4 Conclusion

In this paper eight varieties of buckwheat from different origins were identified using near-infrared spectroscopy combined with principal component analysis and binary tree multi-class SVM classification methods.The buckwheat varieties spectral information identification model have been established by LIBSVM package in MATLAB softweare. The average accuracy rate of eight different types of buckwheat is 92.5%. the results showed that near infrared spectroscopy combined with principal component analysis and support vector machine algorithm for buckwheat spectral information modeling can be achieved to identify buckwheat species.

Acknowledgment. Funds for this research was provided by Yunnan Province application basis surface research projects (Item number: 2010ZC028), Kunming University of Science and Technology Analysis and Testing Fund (Item Number: 2011432, 2011436).

References

1. Kayashita, J., Shimaoka, I., Nakajoh, M., et al.: Consumption of a buckwheat protein extract retards 7,12-Dimethylbenz[alpha] anthracene-induced mammary carcinogenesis in rats. Bioscience Biotechnology Biochemistry 63(10), 1837–1839 (1999)
2. Wang, D., Ma, Z., Pan, L., et al.: Research on the Quantitative Determination of Lime in Wheat Flour by Near-Infrared Spectroscopy. Spectroscopy and Spectral Analysis 33(1), 69–73 (2013)
3. Zhang, Q., Zhang, J.: Region Selecting Methods of Near Infrared Wavelength Based on Wavelet Transform. Transactions of the Chinese society for Agricultural Machinery 41(2), 138–142 (2010)
4. Zhou, Z., Zhang, Y., He, Y., et al.: Method for Rapid Discrimination of Varieties of Rice using Visible NIR Spectroscopy. Transactions of the Chinese Society of Agricultural Engineering 25(8), 131–135 (2009)
5. Su, X., Zhang, X., Jiao, B., et al.: Determination of Geographical Origin of Navel Orange by Near Infrared spectroscopy. Transactions of the Chinese Society of Agricultural Engineering 28(15), 240–245 (2012)
6. Xi, Z., Wang, F.: Research Progress of Grain Quality Nondestructive Testing Methods. Science and Technology of Food Industry (15), 394–396, 400 (2012)
7. Wang, F., Yang, J., Zhu, H., Xi, Z.: Research of Determination Method of Starch and Protein Content in Buckwheat by Mid-Infrared Spectroscopy. In: Li, D., Chen, Y. (eds.) Computer and Computing Technologies in Agriculture VI, Part II. IFIP AICT, vol. 393, pp. 248–254. Springer, Heidelberg (2013)
8. Zhang, H., Zhang, S., Wang, F., et al.: Study on Fast Discrimination of Seabuckthorn Juice Varieties Using Visible-Nir Spectroscopy. Acta Optica Sinica 30(2), 574–578 (2010)

9. Wang, F., Zhu, H.G.Z.: Progress of Near-infrared Spectral Data Modeling Method. Agricultural Engineering 1(1), 56–61 (2011)
10. Zhang, T., Ding, Y.: Protein Structural Class Prediction with Binary Tree-Based Support Vector Machines. Journal of Biomedical Engineering 25(4), 921–924 (2008)
11. Lv, X., Li, L., Cao, W.: SVM Multi-class Classification based on Binary tree. Information Technology (4), 1–3 (2008)