Millet Origin Identification Model Based on Near-infrared Spectroscopy

Penghe LYU, Dongfeng YANG *

College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163700, China

Abstract [**Objectives**] This study was conducted to clarify the difference of millet from different producing areas in near-infrared spectroscopy (NIRS) modeling. [**Methods**] Millet samples from six different regions were collected for NIRS analysis, and an origin identification model based on BP neural network was established. The competitive adaptive reweighted sampling (CARS) algorithm was used to extract characteristic wavelength variables, and a CARS-BP model was established on this basis. Finally, the CARS-BP model was compared with support vector machine (SVM), partial least squares discriminant analysis (PLS) and KNN models. [**Results**] The characteristic wavelengths were extracted by CARS, and the number of variables was reduced from 1 845 to 130. The discrimination accuracy of the CARS-BP model for the samples from six producing areas reached 98.1%, which was better than SVM, PSL and KNN models. [**Conclusions**] NIRS can quickly and accurately identify the origin of millet, providing a new method and way for the origin identification and quality evaluation of millet.

Key words Millet; Identification of origin; CARS-BP model; NIR

DOI:10.19759/j.cnki.2164 – 4993.2024.03.009

Millet is an important food crop widely planted in northern China, as well as one of the traditional staple foods in northern China[1]. Because it contains a lot of protein, carbohydrates, dietary fiber and other nutrients, it has been favored by consumers, and has the effects of lowering blood sugar, improving digestion and promoting sleep[1]. Although there is no obvious difference in appearance between different varieties and different producing areas, their taste and nutritional value are different[2]. The research results obtained by Liang *et al.* [3] showed that geographical factors had a great influence on the nutritional quality of millet, and they affected protein, fat and dietary fiber in millet, while variety factors mainly affected protein and fat contents. The research results obtained by Feng *et al.* [4] showed that there were differences in nutritional components (such as protein, fat and carbohydrates) among different millet varieties in Shanxi. Therefore, it can be seen that different producing areas and varieties may lead to quality differences of millet. It is very important for consumers to identify different varieties and producing areas of millet. At present, the identification methods of millet origin mainly include morphological identification, genetic methods[5], Raman spectroscopy[6], liquid chromatography and chemical analysis. However, the morphological identification method has the disadvantages of strong subjectivity and large error, while other methods have the disadvantages of high cost, long time, and destructive and cumbersome operation. Therefore, it is necessary to establish a rapid, accurate and simple method for identifying millet varieties and producing areas. Near-infrared spectroscopy (NIRS), as a modern instrumental analysis method, has the advantages of rapid detection, simple treatment, no damage to samples and no need for chemical reagents, and has been widely used in crop origin identification and quality evaluation[7 – 8]. This study aimed to collect data of millet samples from six different regions by using NIRS, and to establish an origin identification method based on BP neural network model. In order to extract appropriate characteristic wavelength variables, we adopted competitive adaptive reweighted sampling (CARS) algorithm to realize rapid identification analysis of millet from different producing areas.

Materials and Methods

Experimental instruments and samples

In order to extract suitable characteristic wavelength variables, we adopted CARS to test six kinds of experimental samples from five provinces in China by using Fourier transform infrared spectrometer, Tango model of German Bruker company. The test conditions were as follows: scanning range 3 950 – 11 550 cm^{-1}, scanning times 32, resolution 8 cm^{-1}, and 1 845 data points collected for each spectrum. The experimental samples were all purchased on the spot. Table 1 shows information on varieties and producing areas.

Table 1 Varieties and origin information of experimental millet

Sample	No.	Source	Sampling quantity // g
Chifeng 035	CF	Chifeng City, Inner Mongolia	450
Zhaodong Xiaomi	ZD	Zhaodong City, Heilongjiang Province	450
Mapo Jingu 958	MPJG	Jining City, Shandong Province	450
Dongfangliang 010	DFL	Datong City, Shanxi Province	450
Chaoyang 535	CY	Chaoyang City, Liaoning Province	450
Qinzhou 2	QZ	Changzhi City, Shanxi province	450

Spectral measurement and data processing

The spectral data of millet samples were collected by integrating sphere diffuse reflectance. Specifically, a 450 g of sample

Received: March 16, 2024 Accepted: May 19, 2024

Penghe LYU (1998 –), male, P. R. China, mater, devoted to research about pattern recognition and near-infrared spectroscopy technique.

* Corresponding author. Dongfeng YANG (1977 –), female, P. R. China, associate professor, PhD, devoted to research about application of pattern recognition in agriculture.

was divided into 30 equal parts, each of which was measured for 3 times (the sample was turned and shaken evenly before each measurement), and a total of 540 spectra were collected. These data were divided by Kennard-Stone algorithm according to the ratio of 4 : 1 into training set samples and test set samples, of which the training set samples were used to establish millet origin identification models, and the test set samples were used to verify the prediction ability of the models to millet samples.

Results and Analysis

Spectral analysis

In this study, we collected the near-infrared spectra of millet samples from different producing areas, and these spectra were analyzed through the information of stretching vibration and sum frequency absorption of hydrogen-containing chemical bonds in the samples. We paid special attention to extracting the absorption peak data of bonding of hydrogen groups and hydroxyl groups with metal cations in millet in the range of 4 000 – 12 000 cm^{-1}, so as to facilitate comparative analysis. Fig. 1 shows the near infrared spectra of millet from different producing areas. At the wavelengths of 8 210, 6 846, 5 182, 4 737 and 4 366 cm^{-1}, we found five significant absorption peaks. The absorption peaks were more frequent in the low wave number part of the spectra, and the absorbance increased with the decrease of wave number. The difference in peak intensity of millet samples from different producing areas might be due to the differences in moisture, fiber and starch content, but the overall similarity was high. Therefore, it is necessary to further establish a discrimination model.

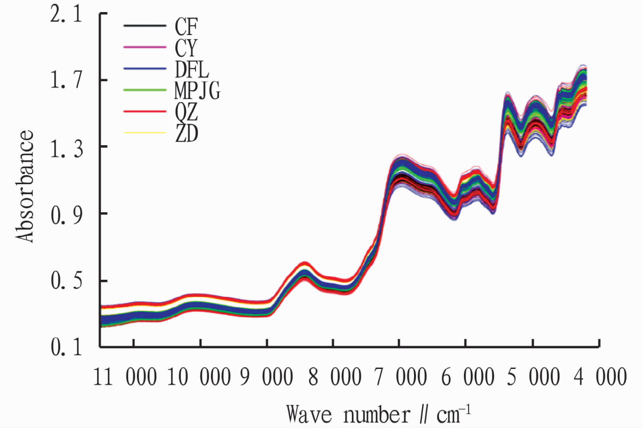

Fig. 1 Near-infrared spectra of millet

Competitive adaptive reweighted sampling (CARS)

CARS is a feature selection method. Its basic idea is to improve the discrimination and robustness of features by adaptively adjusting sample weights according to the importance of features. Specifically, the method first calculates the importance of each feature by using the Relief algorithm, and then initializes sample weights, and sets the initial weight of each sample as an equal value. Next, the sample weights are re-adjusted according to the importance of features, and the above three steps are iterated until the root-mean-squares error of cross validation (RMSECV) is minimum. As shown in Fig. 2, when the number of iterations was 18, the RMSECV reached the lowest point, and the number of selected wavelength variables was reduced from 1 845 to 130, reaching the optimal value. It greatly shortened the operation time of the model and further improved the prediction accuracy of CARS-BP model.

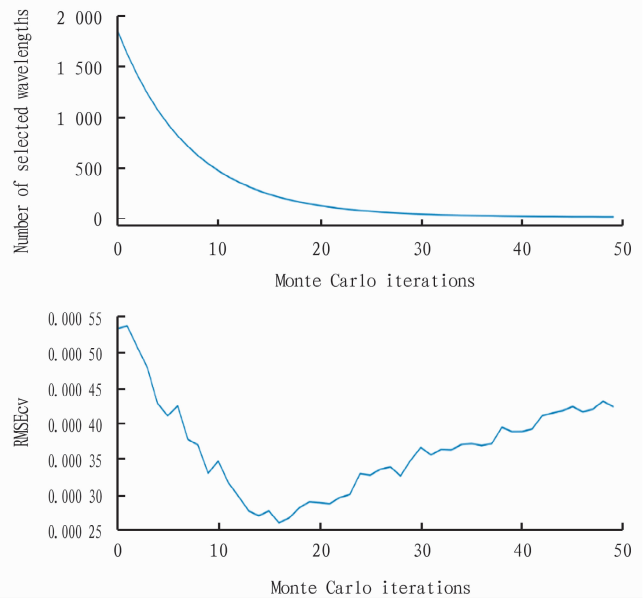

Fig. 2 Number of selected wavelengths, RMSECV and changes with Monte Carlo iterations

Construction of CARE-BP neural network model

BP neural network includes an input layer, a hidden layer and an output layer. The number of neurons, learning rate, activation function and other parameters of the hidden layer need to be determined through repeated experiments and adjustment. Through CARS algorithm, 130 characteristic bands were extracted from the original data as the input of BP neural network. Next, the divided 432 training sets were used to train the BP neural network, and the maximum number of iterations was set to 1 000, that is, the weights and offset values were constantly adjusted through the back propagation algorithm, so as to make the output result of the neural network as close as possible to the real label. After the training was completed, the trained CARE-BP neural network model was evaluated by using the test set. According to Fig. 3, the accuracy of the prediction set samples was as high as 98.1%.

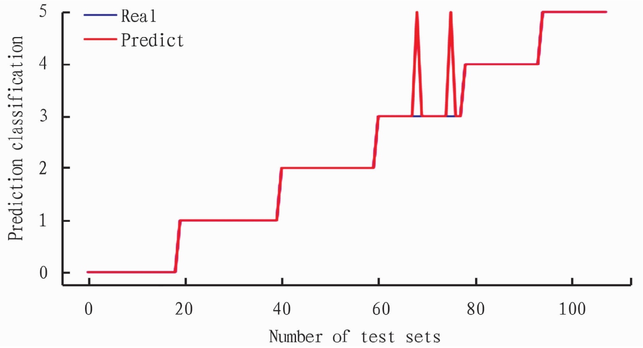

Fig. 3 Prediction classification of the CARE-BP model test set

Comparative analysis of models

In order to further illustrate the effect of CARE-BP neural network model in identifying the origin of millet, 130 characteristic wavelengths were used as input variables to construct full-spectrum BP neural network, support vector machine (SVM), partial least squares discriminant analysis (PLS) and K-nearest neighbor algorithm (KNN) models, respectively. Table 2 shows the identification results of millet from different producing areas using different models. As can be seen from the table, the identification effect of the CARE-BP model was better than other four models. CARE-BP model not only has strong generalization ability when the number of samples is small, but also can be applied to the analysis of complex nonlinear spectra, so it is an effective method to identification of millet from different places.

Table 2 Comparison of different modeling results

Model	Number of variables	Rate of identification//%
Full spectrum BP	1 845	90.2
CARE-BP	130	98.1
SVM	130	97.5
PLS	130	93.1
KNN	130	95.9

Conclusions and Discussion

In this study, the BP algorithm based on NIRS was adopted to effectively distinguish millet from different producing areas. In order to simplify the model and eliminate redundant spectral variables, the CARS method was applied to extract characteristic wavelengths, and a CARS-BP neural network model was constructed. Compared with other three classification models (SVM, PLS and KNN), this model showed obvious advantages, and the discrimination accuracy was as high as 98.1%. The research results showed that the CARE-BP neural network model showed high accuracy and stability in feature extraction and classification tasks. Compared with the traditional methods of sensory evaluation and physical and chemical tests, NIRS combined with CARS-BP model can quickly and accurately identify the origin of millet, providing a new method for the authenticity identification and quality evaluation of millet.

References

[1] LI X, WANG HH, SHEN Q. Studies on the quality characteristics of different varieties of millet[J]. Journal of Chinese Institute of Food Science and Technology, 2017, 17(7): 248 –254. (in Chinese).

[2] TIAN X, CHE Q, YAN WM, *et al.* Discrimination of millet varieties and producing areas based on infrared spectroscopy[J]. Spectroscopy and Spectral Analysis, 2022, 42(6): 1841 –1847. (in Chinese).

[3] LIANG KH, ZHU DZ, SUN JM. Study on variety and regional factors related to nutrient quality in millet[J]. The Food Industry, 2017, 38(4): 192 –196. (in Chinese).

[4] FENG NH, HOU DH, YANG CY, *et al.* Discrimination of millet varieties and producing areas based on infrared spectroscopy[J]. Science and Technology of Food Industry, 2020, 41(8): 224 –229. (in Chinese).

[5] SI CJ. Application of system genetics method based on protein interaction in identification of plant functional genes[D]. Wuhan: Huazhong Agricultural University, 2021. (in Chinese).

[6] SHA M, LI LC, HUANG JL, *et al.* Effect of data processing method on identification of rice from different geographical origins by Raman spectroscopy[J]. Journal of Chinese Institute of Food Science and Technology, 2021, 21(5): 369 –376. (in Chinese).

[7] WANG Y, LI Y, YE HZ, *et al.* Geographical origin discrimination of *Campsis grandiflora* by near-infrared spectroscopy coupled with support vector machine[J]. Journal of Fuzhou University: Natural Science Edition, 2022, 50(4): 568 –573. (in Chinese).

[8] YANG J, MA X, GUAN H, *et al.* A recognition method of corn varieties based on spectral technology and deep learning model[J]. Infrared Physics & Technology, 2023(128): 104533.

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU

(Continued from page 30)

[6] HUANG C, GUO M. Optimization of purifying agent combination for detection of pyrethroid pesticide residues in cucumbers by orthogonal test[J]. Chinese Journal of Food Hygiene, 2012, 24(5): 438 –440. (in Chinese).

[7] KANG WJ. Application of gas chromatograph in detection of organophosphorus pesticide residues in vegetables[J]. Hunan Agricultural Sciences, 2010(1): 76 –78. (in Chinese).

[8] WANG DQ, HAN MH. Differences in stability of nine pesticides during the process of gas chromatography column[J]. Journal of Zhejiang Agricultural Sciences, 2010(1): 113 –115. (in Chinese).

[9] FRENICH AG, BOLANOS PP, VIDAL JLM. Multiresidue analysis of pesticides in animal liver by gas chromatography using triple quadrupole tandem mass spectrometry[J]. Journal of Chromatography, A, 2007, 1153: 194 –202.

[10] LIU GP, HUANG C, XUE RX, *et al.* Determination of 14 organophosphorus pesticides and 7 pyrethroids pesticides in five kinds of food by GPC and GC-MS/MS [J]. Chinese Journal of Food Hygiene, 2014, 26(4): 366 –372. (in Chinese).

Editor: Yingzhi GUANG Proofreader: Xinxiu ZHU