ON-LINE DETECTION OF CERASUS HUMILIS FRUIT BASED ON VIS/NIR SPECTROSCOPY COMBINED WITH VARIABLE SELECTION METHODS AND GA-BP MODEL

/

基于可见/近红外光谱技术结合变量选择和GA-BP模型的欧李果实在线分类检测

Wang Bin [1)], He Junlin *[1)], Zhang Shujuan *[1)], Li Lili [2)]

[1)] College of Agricultural Engineering, Shanxi Agricultural University, Taigu/China

[2)] College of Information Science and Engineering, Shanxi Agricultural University, Taigu/China

Tel: +86-0354-6288400; E-mail: hejunlin26@126.com

DOI: https://doi.org/10.35633/inmateh-63-20

Keywords: *fresh cerasus humilis fruit, detection, visible-near infrared spectroscopy, genetic algorithm, neural network, characteristic wavelengths*

ABSTRACT

In order to realize the rapid and non-destructive detection of fresh Cerasus Humilis' (CH) classification, and promote the deep-processing of post-harvest fresh fruit and improve market competitiveness, this study proposed a nonlinear identification method based on genetic algorithm (GA) optimized back propagation (BP) neural network of different varieties of fresh CH fruit. "Nongda-4", "Nongda-5", and "Nongda-7" fresh CH fruit were selected as research objects to collect their visible/near-infrared spectral data dynamically. The original spectra were preprocessed by moving smoothing (MS) and standard normal variate (SNV) methods, for the characteristic wavelengths were extracted with four dimension-reducing methods, namely principal components analysis (PCA), competitive adaptive reweighed sampling (CARS), CARS-mean impact value (CARS-MIV), and random frog (RF) algorithm. Finally, the BP prediction models were established based on full-spectrum and characteristic wavelengths. At the same time, the GA optimization was used to optimize the initial weight and threshold of the BP neural network and compared with the partial least squares' discrimination analysis (PLS-DA) linear model. Through comparing the MS (7)+SNV was proved to be the best preprocessing method, the CARS-MIV-GA-BP model had the best discriminant accuracy, the prediction set accuracy was 98.76%, of which the variety "Nongda-4" and "Nongda-5" recognition rate were 100%, the variety "Nongda-7" recognition rate was 96.29%. The results show that the GA can effectively optimize the initial weights and threshold randomization of the BP neural network, improve the discrimination accuracy of CH varieties, and the CARS-MIV algorithm can effectively reduce the number of input nodes of the BP neural network model, simplify the structure of BP neural network. This study provides a new theoretical basis for the detection of fresh CH fruit classification.

摘要

为了实现欧李鲜果分类的快速无损检测，推动采后鲜果的精深加工及提高市场竞争力。本研究提出基于遗传算法（GA）优化BP神经网络欧李鲜果品种的非线性判别方法。以产自同一地区的农大4号、农大5号和农大7号欧李果为研究对象，动态采集光谱数据。采用移动平滑法（MS）和标准正态变量（SNV）方法对原始光谱进行预处理，分别选用主成分分析（PCA）、竞争性自适应重加权算法（CARS）、竞争性自适应重加权-平均影响值算法（CARS-MIV）、随机蛙跳算法（RF）算法对光谱数据降维，将全波段和优选出的特征波长数据作为BP神经网络输入变量，采用GA优化BP神经网络的权值和阈值，建立BP、GA-BP神经网络非线性判别模型，并与偏最小二乘判别分析（PLS-DA）线性模型比较。分析得出，MS (7)+SNV为最佳预处理方法，CARS-MIV-GA-BP判别模型最佳，预测集总正确判别率为98.76%，其中"农大4号"和"农大5号"识别率均为100%，"农大7号"识别率为96.29%。研究表明，通过GA算法能有效地优化BP神经网络初始权值和阈值随机化，可提高欧李果品种判别精度，同时CARS-MIV算法可有效减少BP神经网络模型的输入节点数，简化BP神经网络结构。该研究为欧李果在线分类检测提供了新的理论基础。

INTRODUCTION

Cerasus humilis (Bge.) Sok. (CH) is a kind of Rosaceae cherry, it is not only a unique fruit plant resource in China, but also the most dwarf fruit tree in the world. CH has the characteristics of drought-resistance, cold-resistance, barren-resistance, strong root system, soil consolidation, and water conservation. CH fruit is a new kind of fruit, which contains many kinds of nutrients and mineral elements that are beneficial to the human body. The content of calcium in CH fruit is four times as high as that of Citrus Sinensis, Tangerine and Plum, more than 5 times higher than other fruits, so it is also called “fruit rich in calcium”, which is the third generation exclusive in China, and honoured as one of the three high-end fruits with American blueberry and Russian sea-buckthorn, at the same time, CH’s seed kernels are the main source of Yu Li Ren (Semen Pruni).

Therefore, as a "homology of medicine and food", CH fruit can be used as both fruit and medicinal material. With the continuous updating of CH varieties and the increase of yield year by year, the intensive processing of fresh fruits of CH has attracted more and more attention, from fresh food, fruit juice, fruit wine, fruit vinegar, fruit jam, preserved fruit to further extraction of a variety of physiologically active substances, the market prospects and its broad. However, the internal and external quality of different varieties are different, the products that can be further processed and the market price are also quite different.

Therefore, grading and sorting CH fruit according to its quality is essential in post-production processing, which is of great significance to the storage and sale of CH fruit. How to establish a rapid, non-destructive, and effective classification method of fresh CH fruit varieties has become an urgent problem.

In recent years, near-infrared spectroscopy has been widely used in agriculture and food as a rapid, non-destructive *testing method (Li et al., 2019; Du et al., 2020; Firmani et al., 2019)*. Near-infrared spectral modeling is based on multivariate statistical analysis, and an artificial neural network is one of the most widely used techniques. The back propagation (BP) network was proposed by *Rumelhart* in 1986. It is a multi-layer feedforward network trained by error back propagation theory. BP Neural Network has a strong ability of nonlinear mapping, self-learning, self-adaptation, generalization, and fault-tolerance. It is one of the most widely used neural network models, the method has been applied to improve the convergence speed and prediction accuracy of the model *(Xie et al., 2019; Yang et al., 2013; Lu et al., 2018; Wu et al., 2013)*.

At the same time, combining neural network with intelligent algorithms such as genetic algorithms (GA) *(Sun et al., 2016)*, particle swarm optimization (PSO) *(Chen et al., 2016; Mohamad et al., 2018)*., krill herd algorithm (KH) *(Liu et al., 2018)* and bird swarm algorithm (BSA) *(Xiang et al., 2019)* can improve the prediction accuracy, the BP neural network optimized by GA has been applied to the prediction and discrimination of walnut shell breaking *(Zhang et al., 2014)*, leaf chlorophyll *(Chen et al., 2018)*, soil moisture *(Liang et al., 2019)*, and red bean variety *(Sun et al., 2016)*.

For example, *Zhang et al., (2017)* put forward a method of tea leaf spot recognition based on hyperspectral image technology, and optimized BP neural network modeling independent variables by GA, which improved the spot recognition rate from 85.59% to 94.17%, the establishment time was also shortened from 6.6 seconds to 1.7 seconds before optimization. *Gu et al., (2017)* used a genetic algorithm to optimize the initial parameters of the BP neural network, which can effectively improve the prediction accuracy of corn yield and the convergence speed of the network.

Tan et al., (2019) identified soybean seed varieties using hyperspectral imaging and machine learning. It is concluded that the texture feature parameters are extracted from the three images by principal component analysis, and the prediction accuracy of the established GA-BP neural network model is 92%. *Yan et al., (2020)* by using the hyperspectral imaging technique and Chemometrics method, three kinds of identification models for fresh tea, which are GA optimized BP neural network, traditional BP neural network, and support vector machines (SVM), are established. The results show that the improved BP neural network based on GA can improve the performance of the model and has 100% prediction accuracy by combining the spectral preprocessing with the multiplicative scatter correction (MSC) and the standard normal variant (SNV). The studies above show that GA can effectively improve the predictive ability of the BP neural network, but there is little application in variety identification of fresh CH fruit based on Vis/NIR spectroscopy.

“Nongda-4”, “Nongda-5”, and “Nongda-7” CH fresh fruit were selected as research objects to collect their visible/near infrared spectral data dynamically. The raw spectral data is pre-processed by different pre-processing methods, combined with various wavelength extraction methods to downscale the pre-processed spectral data, and the downscaled data is used as the input of the model to build the traditional BP neural network and GA-BP neural network classification model for prediction and discrimination, and combined with

the linear model partial least squares discriminant analysis (PLS-DA), the predictive discriminatory effects of PLS-DA method were compared.

MATERIALS AND METHODS

Research samples

Test fresh CH fruit samples (Nongda-4, Nongda-5 and Nongda-7) were picked on 1 September 2019 at the Agricultural High-Tech Industry Demonstration Zone Cerasus Humilis Planting Demonstration Base in Taigu County, Shanxi Province, China (112°29'E, 37°23'N). The fruits were cleaned and dried in the laboratory (3 hours), and then the fruits with external defects and rot diseases were removed, and 80 each of "Nongda-4", "Nongda-5" and "Nongda-7" were screened for good maturity and basic consistency in physical properties. (240 in total), and the selected samples were numbered. According to the Kennard-Stone (K-S) algorithm *(Galvão et al., 2005)*, the three experimental samples were divided in a 2:1 ratio, and "Nongda-4", "Nongda-5" and "Nongda-7" were obtained, the sample sizes of the calibration set for the three species varieties were 53, 54, and 53 (160 samples in total), and the sample sizes of the prediction set were 27, 26, and 27 (80 samples in total).

The weight and fruit diameter (maximum diameter at the equatorial part of the fruit) were weighed and measured using a high-precision electronic balance (FA1004N, Shanghai) and a vernier calliper (Mitutoyo, Japan) for each variety of fresh CH fruit samples, and SPSS17.0 was used to calculate the parameters of the three different varieties of fresh CH fruit samples, as shown in Table 1.

Table 1

Data statistics of three different varieties of CH fruit samples

Cultivar	Sample parameters	Min.	Max.	Mean	S.D.	Variable coefficient (%)
Nongda-4	**Fruit diameter [mm]**	20.72	25.15	23.02	1.09	4.74
	Weight [g]	5.92	7.26	6.57	0.47	7.15
Nongda-5	**Fruit diameter [mm]**	15.48	28.66	24.72	3.62	14.64
	Weight [g]	7.47	16.18	11.39	1.76	15.45
Nongda-7	**Fruit diameter [mm]**	21.54	28.98	25.27	1.56	6.17
	Weight [g]	10.32	17.58	13.93	1.97	14.14

S.D.=Standard deviation.

Experimental system and data acquisition

The Field Spec3 spectrometer produced by ASD (Analytical Spectral Device) and the self-developed dynamic spectral acquisition system was used to achieve the spectral acquisition of fresh CH fruit samples, the schematic diagram of the online Vis-NIR spectroscopy detection device is shown in Fig.1.

Spectral data acquisition interval is 1nm, the wavelength range is 350～2500nm, the resolution is 3.5nm, the probe field of view angle is 20°, the light source is 14.5V halogen lamp, the probe of the spectrometer is perpendicular to the upper surface of the sample, 90mm away from the upper surface of the sample, the sample is placed between the two rollers.

To minimize the errors, the spectrometer was switched on for 0.5 hours for system configuration optimization and whiteboard calibration, and after the performance test, the sample was sampled by diffuse reflection.

For data collection, three kinds of CH fruit samples were rotated at a speed of 5 r/min and spectral data of the samples were collected at 120° intervals for a total of three times, and the average value was taken as the final spectral data of the test samples. When the CH sample enters the spectral scanning range, a proximity switch is triggered, the large sprocket stops moving, the small sprocket continues to move to ensure that the CH sample rotates under the probe, and after the spectral data is collected, the large sprocket continues to move and the CH sample is transported away until the next CH sample enters the spectral scanning range and begins scanning.

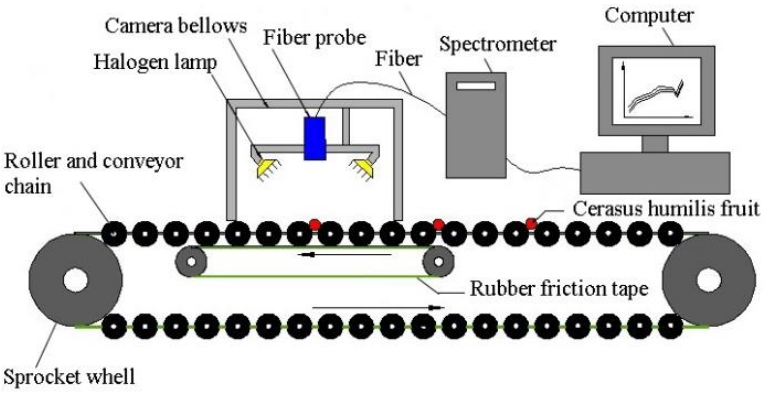

Fig. 1 - Schematic diagram of online Vis-NIR spectroscopy detection device

Data analysis and models establishment

Spectral data preprocessing

Due to the influence of the external environment and random errors during the spectral acquisition process, there are random noise and baseline drift in the spectral data, and the noise and background interference can be removed by appropriate spectral preprocessing methods, which is expected to improve the performance of the discriminant model. In this paper, moving smoothing (MS) and standard normal variate (SNV) are used to remove the noise and background interference.

Effective variable selection algorithms

There is a large amount of data redundancy in the original spectral information collected, the strong correlation among the wavelengths, and a large number of wavelength dimensions, which can affect the accuracy and prediction speed of the calibration model prediction classification. In order to solve these problems, several methods of extracting characteristic wavelengths are usually used to downscale the original spectral matrix. In this paper, we choose the feature extraction method of principal components analysis (PCA), feature selection method of competitive adaptive reweighed sampling (CARS), mean impact value (MIV), random frog (RF) and their combination algorithms are used to downscale the spectral data, screen the characteristic wavelengths that can reflect all the spectral information, and optimize the input variables of the model.

BP neural network

BP neural network is a kind of multi-layer feedforward network trained by back propagation of errors algorithm, with local search capability, it is a very mature regression analysis method, it can establish a non-linear model for classification and prediction, it is the most widely used back propagation artificial neural network, BP neural network contains input layer, implicit layer and output layer, if the input layer has X neurons, the output layer has Y neurons, there are P neurons in the hidden layer, and the weights are denoted by W. Then the basic structure of the BP neural network is shown in Fig.2, where X_1, X_2, ..., X_n are the input values, Y_1, Y_2, ..., Y_m are the predicted output values, and W_{ij} and W_{jk} are the weights of the BPNN.

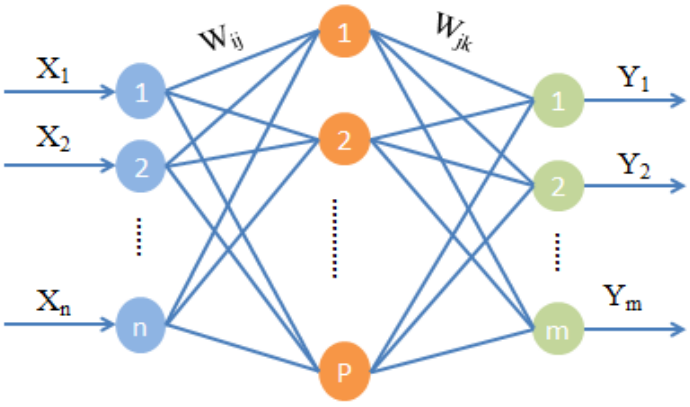

Fig. 2 - The basic principle of BP neural network

For the general pattern recognition problem, a three-layer network can be a good solution. In this study, a three-layer BP neural network is used, i.e., input layer, implicit layer and output layer, in which the number of nodes in the input layer is determined by the input data dimension, and the number of nodes in the output layer is determined by the sample characteristics, assuming that the number of neurons in the input layer is N_1 and the number of neurons in the implicit layer is N_2, and there is an approximate relationship between them: $N_2=2N_1+1$. The transfer function of the implicit layer neurons of the neural network adopts S-type tangent The function tansig( ), the transfer function of the output layer neurons uses the S-type logarithmic function logsig( ), this is because the output mode is 0-1, which exactly satisfies the output requirements of the network.

However, there are some shortcomings in the BP neural network, such as the network structure, the choice of initial connection weights and thresholds, which have a great influence on the network training, and the weights and thresholds are usually randomly initialized to random numbers in the interval of [-0.5,0.5]. To address these shortcomings this paper uses a genetic algorithm to optimize the BP neural network.

Genetic algorithms

In 1962, Professor Holland of the University of Michigan proposed the GA. The algorithm is a parallel random search that simulates the natural genetic mechanism and biological evolution, with excellent global search capabilities. According to the principle of the selection function, through a series of selection, intersection and variation screening of the population, the variables of the input population can be trained and optimized, and then continuously optimized according to the principle of "survival of the fittest", and finally better adapted. The GA can effectively optimize the BP neural network by randomizing the initial weights and thresholds, easily falling into local extremes and slow convergence, and locate the ideal search space for it. Searching for the optimal individual through selection, crossover and variation operations and assigning the optimized weights and thresholds parameters to BP neural networks not only exerts the powerful nonlinear mapping ability of neural networks, but also gives the neural networks faster convergence and stronger learning ability, which enables the optimized BP neural networks to make better sample predictions.

The idea of GA optimization of BP neural network can be divided into: (1) BP neural network structure determination; (2) GA optimization weights and thresholds; (3) BP neural network training and prediction. The flow of the GA to optimize the BP neural network is shown in Fig.3.

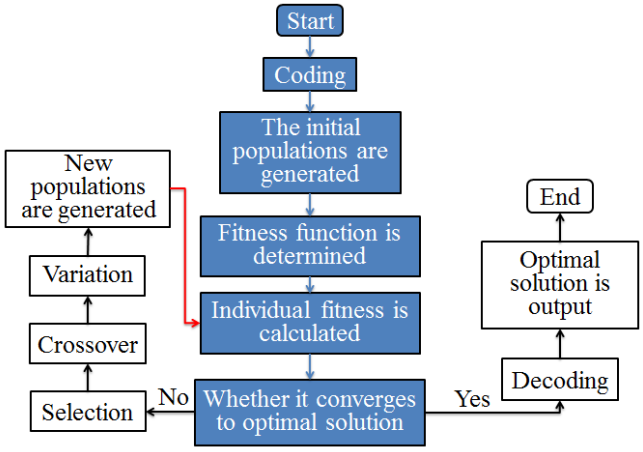

Fig. 3 - The optimization process of genetic algorithm

RESULTS

Spectral analysis of samples

The original spectra of three cultivars of CH fruit are shown in Fig.4(a). There were some crossovers and overlapping among these spectra, however, the trends of spectra were quite similar. Fig.4(b) shows the averaged reflectance spectra of three cultivars of CH fruit in the spectral region of 350-2500 nm, in the spectral region of 900-2500 nm, the average reflectivity of "Nongda-5" and "Nongda-7" is lower than that of "Nongda-4". In the visible spectral region of 500-700 nm, there were some differences that might have a direct correlation with the peel colour variances due to different CH fruit cultivars, there is an obvious peak around 680 nm which may be related to chlorophyll absorbance of CH's peel. In the near-infrared spectral

region, there are three obvious absorption peaks around 980 nm, 1190 nm and 1450 nm, the first two were assigned to second overtones of band O-H and band C-H, respectively *(Hideyuki et al., 2011)*.

The third may be associated with the first overtone of bond O-H. Three CH species have distinct absorption peaks at 1660nm, which is the second overtone of band C-H. It can be observed that the original spectra have obvious sample inhomogeneity, high-frequency random noise and light scattering. Therefore, the preprocessing of moving smoothing (7 points) transformation was filtered for random noise and high-frequency noise, standard normal variate transformation (SNV) was applied for light scatter correction. Reflectance spectra after the preprocessing is shown in Fig.4(c). From Fig.4(c), it can be seen that there was large noise at the beginning part (350-450nm) of spectra and the end of the spectral curve (2400-2500nm), which will directly affect the accuracy of the experiment. Therefore, only the spectral ranges of 450–2400nm were used for this study.

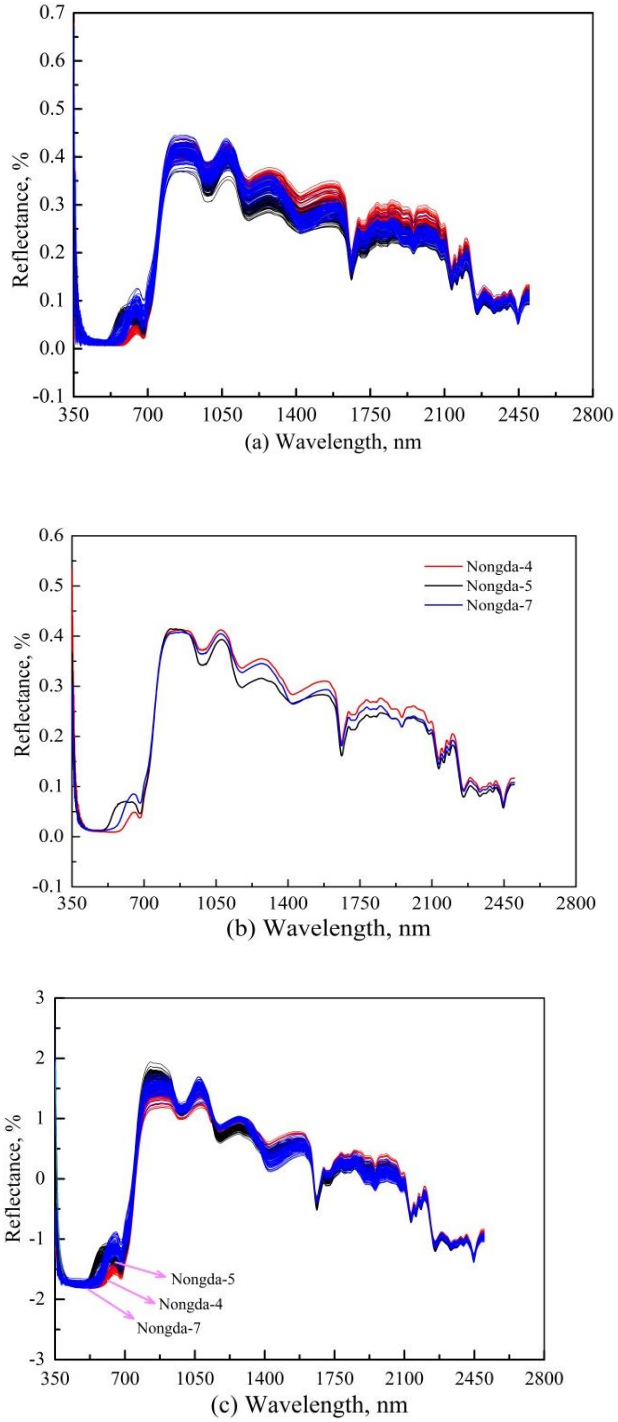

Fig. 4 - (a) Original spectra of 240 samples of three cerasus humilis varieties; (b) Average spectra of each cultivar; (c) Spectra after pre-treatment by moving average（7 points）and SNV

Selection of effective variables

Extraction of principal components

The main purpose of PCA is to extract variables and achieve the classification of the original large number of spectral variables into a few comprehensive indicators, which not only eliminates the overlap between the numerous information and extracts the most representative subset of variables but can also characterize the main features of the original data, as shown in Table 2.

Table 2

The accumulated contribution rate of the first 9 principal components

No. of principal components	PC 01	PC 02	PC 03	PC 04	PC 05	PC 06	PC 07	PC 08	PC 09
Cumulative contribution [%]	67.05	90.14	95.84	97.68	98.71	99.24	99.51	99.66	99.78

From Table 2, we can see that the cumulative contribution of the first six principal components has reached 99.24%, which contains more than 99% of the feature information of the spectral data, the cumulative contribution does not change much as the number of principal components increases, and the first six principal components do not affect each other. The first six principal components were selected to represent the main information of the original visible/near-infrared spectra. A total of 240 samples of the three species of CH were clustered by PCA, and the scatter plots of the first three principal components are shown in Fig.5.

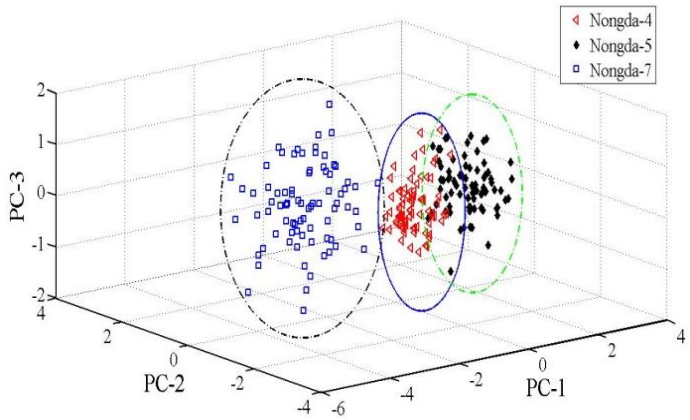

Fig. 5 - Scatter plots of cerasus humilis samples of the first three principal components

From Fig.5, it can be seen that there is a clear difference in characteristics between Nongda-4, Nongda-5 and Nongda-7, indicating that the first three principal components have a better clustering effect on the three species of CH. Nongda-4 and Nongda-5 had better aggregation, with some samples intersecting between them, with almost no overlap, and could be relatively distinguished; Nongda-7 had 80 samples less aggregated than the other two, but had no overlap with the other two, and was closely located on one side. The analysis showed that the first three principal components had a good clustering effect on the three species of CH, which provided a basis for qualitatively differentiating the different varieties of CH.

Competitive adaptive reweighted sampling (CARS)

CARS is a variable selection algorithm that employs the simple but effective principle “survival of the fittest” originating in Darwin's Evolution Theory. The adaptive reweighted sampling (ARS) technique is used to build the PLS model, and the wavelengths with larger absolute values of the regression coefficients are selected from the built model, the wavelength points with smaller weights are removed, and cross-validation selected the optimal subset of variables with the smallest root mean squares error of cross-validation (RMSECV) in the PLS model.

The process of CARS screening characteristic wavelengths is shown in Fig.6. The number of Monte Carlo sampling runs was set to 100 and the final variable number to be selected was determined by 10-fold cross-validation. Fig.6(a), (b) and (c) show the changing trend of the number of the sampled variables, RMSECV values and the regression coefficient path of each variable with the increase of Monte Carlo sampling runs in terms of running one CARS.

As shown in Fig.6(a), the number of wavelengths gradually decreases and finally stabilizes as the number of sampling runs gradually increases, which verified the rough and fine selection during wavelength screening. Fig.6(b) shows that the cross-validation RMSECV gradually decreases when the number of sampling runs increases to 67, and then shows an increasing trend; when RMSECV gradually decreases, it means that the useless information in the spectral information was eliminated; when RMSECV increases, it means that the valid info among spectral info was eliminated. In Fig.6(c), each curve represents the changing trend of the regression coefficient of each spectral variable with the number of sampling, when the position of the line of "*" indicated the runs were 67, RMSECV was minimized (RMSECV=0.2678). The 20 characteristic wavelengths selected by CARS are 455, 483, 575, 647, 648, 649, 650, 670, 671, 918, 919, 1105, 1106, 1322, 1324, 1325, 1326, 1327, 1916 and 1917nm, respectively.

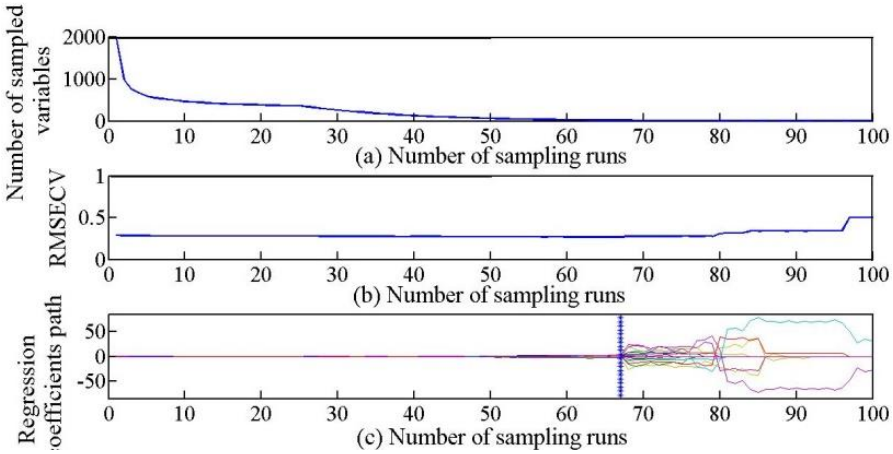

Fig. 6 - Process of selecting wavelength variables by the CARS method. (a) Number of Sampled variables; (b) The changes of RMSECV; (c) Paths of Regression coefficients

Effective variable selection by CARS-MIV

The MIV is an important index for evaluating the influence of input variables on output variables in neural networks *(Zhu et al., 2019)*. The number of Monte Carlo sampling times and the number of cross-validation groups are set randomly when the CARS algorithm is used to extract the characteristic wavelength, which makes the regression coefficient of the selected variables change with the random sampling times, the importance of characteristic variables can't be reflected comprehensively, which affects the robustness of prediction model. To reduce this effect, this study uses the MIV algorithm for secondary selection of the characteristic wavelengths extracted by the CARS algorithm to obtain the corresponding MIV values for 20 characteristic wavelengths, as shown in Fig.7(a). The MIV numerical values are arranged in descending order to analyse the cumulative MIV contribution rate of 20 feature wavelengths, as shown in Fig.7(b).

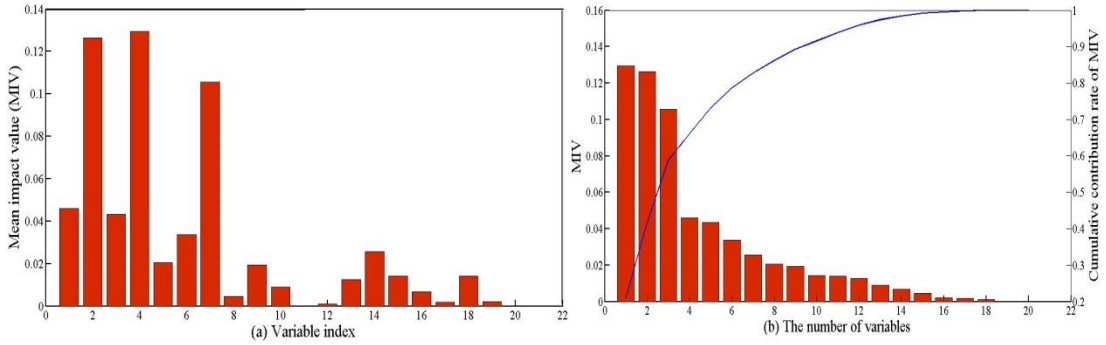

Fig. 7 - (a) Mean impact value of selected variables; (b) cumulative contribution rate of MIV

It can be found from Fig.7(a) that the 11th, 12th, 17th and 20th feature wavelengths correspond to smaller MIV values, which will affect the prediction accuracy of the calibration model. As shown in Fig. 7(b), the curve stands for the cumulative contribution of n(1≤n≤20) characteristic wavelengths, combined with Figure 8(a), the number of characteristic wavelengths after secondary selection of the MIV algorithm is 16.

Effective variable selection by RF

The RF method is similar to the reversible jump Markov chain Monte Carlo algorithm, it can iterate on multidimensional data variables and calculate the weight of each variable. The higher the value, the greater the probability of being selected, and the more important the corresponding wavelength. Therefore, the selection probability of all variables can be sorted, and the variables with higher probability can be selected as the characteristic wavelength.

In order to reduce the influence of random factors, it is necessary to run multiple times and calculate the results. In this study, the RF algorithm was run 2000 times, the selected threshold value was 0.4, and the top 10 wavelengths above the threshold were selected as characteristic wavelengths, which are 1011, 2326, 2397, 979, 2426, 2341, 2327, 2270, 1380, and 1056nm in descending order according to the probability of being selected, respectively. Most of the characteristic wavelengths were concentrated in the range of 2200-2450, which may be related to the region of the combined NIR spectral bands of C-H, N-H, and O-H bonds insoluble solids of CH. The process of filtering characteristic wavelengths using the RF algorithm is shown in Fig.8.

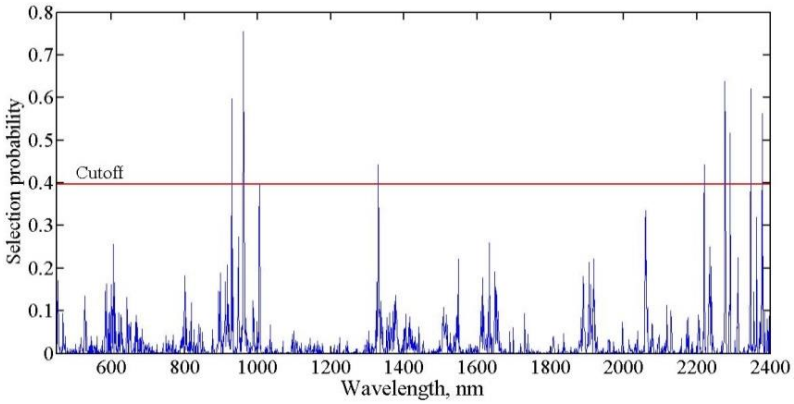

Fig. 8 - Effective variable selection by RF

Optimization of BP discriminant model

The weights and thresholds of traditional BP neural networks are generally random numbers in the interval of [-0.5,0.5], which makes the models have poor generalization ability and low prediction accuracy. In this study, GA is used to optimize the weights and thresholds of the BP neural network. After repeated tests, the initial population size is 40, the maximum number of genetic algebra is 50, the number of a binary number of the variable is 10, the crossover probability is 0.7, and the variation probability is 0.01.

In order to compare the modeling performance of BP before and after GA optimization, based on the spectral data in the range of 450-2400 nm, the full spectrum (FS) data and the preferred characteristic wavelength data are taken as input variables, where the number of input nodes of the BP neural network is determined by the respective input data dimension and the number of output nodes is determined by the variety number of CH. That is, the number of output nodes of the model is all 3, “Nongda-4”, “Nongda-5” and “Nongda-7” are represented by 100, 010 and 001, respectively, and different discriminant models are established. The discriminant results of the BP model before and after GA optimization is shown in Table 3.

Table 3

Comparison of discriminant results of BP model before and after GA optimization

Variable selection methods	No. of wavelength	BP				GA-BP			
		Prediction set accuracy [%]	Number of errors			Prediction set accuracy [%]	Number of errors		
			A	B	C		A	B	C
FS	1951	82.53	4	4	6	84.99	3	4	5
PCA	6	90.03	3	2	3	91.26	3	2	2
CARS	20	85.04	4	3	5	93.73	1	2	2
CARS-MIV	16	88.79	3	2	4	98.76	0	0	1
RF	10	83.76	4	4	5	86.22	3	4	4

Notes: A=Nongda-4, B=Nongda-5, C=Nongda-7

Looking at Table 3, the accuracy of the prediction set was increased using the feature extraction method of PCA, feature selection method of CARS, CARS-MIV and RF, compared with the FS-BP neural network and FS-GA-BP neural network models. Among the pre-optimized BP neural network models, the FS-BP model has a low discrimination accuracy of 82.53% and the PCA-BP identification model performed best, with a recognition rate of 90.03%.

The GA-BP neural network model built with FS data as input to the BP neural network model improved the recognition accuracy of the prediction set from 82.53% to 84.99%.

The recognition accuracy of the GA-BP neural network model based on the first six principal components extracted by PCA was improved from 90.03% to 91.26%.

The recognition accuracy of the GA-BP neural network model established for the 20 feature wavelengths screened by CARS improved from 85.04% to 93.73% for the prediction set. The recognition accuracy of prediction set by the GA-BP neural network model based on 16 characteristic wavelengths screened by CARS-MIV is improved from 88.79% to 98.76%.

The recognition accuracy of the GA-BP neural network model established for the 10 feature wavelengths screened by RF improved from 83.76% to 86.22% for the prediction set.

It is known that the CARS-MIV-GA-BP discriminant model is the best and the RF-GA-BP discriminant model is the worst. This is related to the fact that the characteristic wavelength extracted by the CARS-MIV algorithm contains information on the CH fruit visible and NIR spectral bands, whereas the characteristic wavelength extracted by the RF algorithm contains information only on the NIR spectral bands. The correct recognition rate of “Nongda-4”, “Nongda-5” and “Nongda-7” by CARS-MIV-GA-BP was 100%, 100% and 96.29%, respectively, and the total correct recognition rate was 98.76%. The results show that the characteristic wavelengths filtered by the CARS-MIV algorithm can effectively respond to the full-band spectral information, CARS-MIV-GA-BP identification model performed best, with a recognition rate of 98.76%. Fig.9 shows the GA optimized BP model of CARS-MIV screening variables, and Fig.10 shows the discriminant result of the CARS-MIV-GA-BP model.

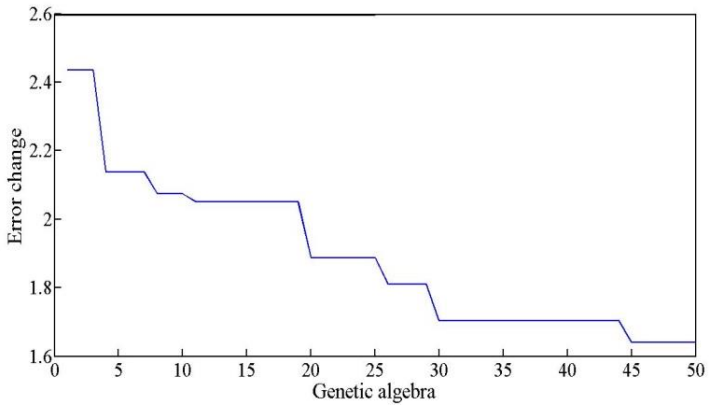

Fig. 9 - GA optimized BP model of CARS-MIV screening variables

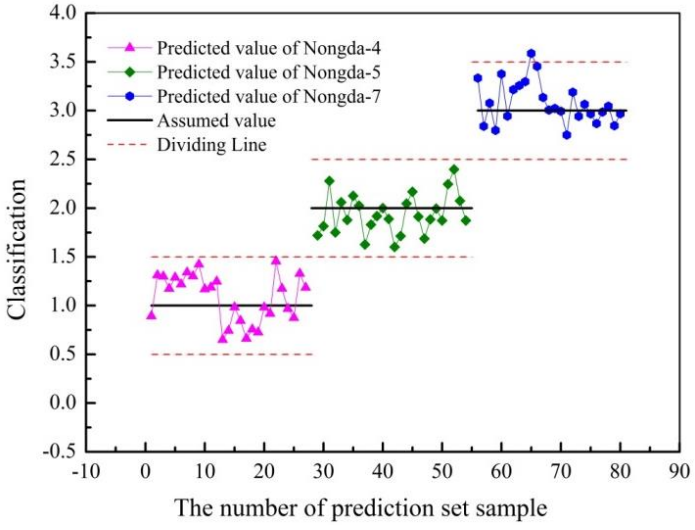

Fig. 10 - The discriminant result of CARS-MIV-GA-BP model

Comparison of optimum GA-BP model and PLS-DA discriminant model

PLS-DA is a multivariate statistical analysis method for discriminant analysis, which can reduce the influence of multicollinearity among variables, and has the characteristics of stable modeling performance and good prediction effect. In order to verify the prediction and discrimination effect of the CARS-MIV-GA-BP model built in this experiment, the characteristic wavelength variable extracted by CARS-MIV Algorithm is used as input of the PLS-DA linear model, and the prediction and discrimination model of CARS-MIVP-LS-DA is established, the discriminant results of the prediction set of CH samples are shown in Table 4.

Table 4

Discrimination of three cerasus humilis varieties by CARS-MIV-PLS-DA models

Cultivar	No. of predicted samples	No. of false positives	Correct discrimination rate [%]	Total [%]
Nongda-4	27	2	92.59%	
Nongda-5	26	1	96.15%	93.78%
Nongda-7	27	2	92.59%	

It can be seen from Table 4 that the CARS-MIV-PLS-DA model has a good discriminant effect on the sample of prediction set. The correct classification rate of “Nongda-4”, “Nongda-5” and “Nongda-7” are 92.59%, 96.15% and 92.59%, respectively, the total correct recognition rate was 93.78%. However, the correct recognition rate of “Nongda-4”, “Nongda-5” and “Nongda-7” by CARS-MIV-GA-BP were 100%, 100% and 96.29%, respectively, and the total correct recognition rate was 98.76%. Thus, the CARS-MIV-GA-BP discriminant model achieved better discriminatory results for CH fruit samples in this study.

CONCLUSIONS

In this study, the dynamic detection of classification discriminatory models of three fresh CH fruit samples was established based on Vis/NIR spectroscopy and GA-BP neural network, the original spectra were pre-processed using different kinds of pre-processing methods, Four different variable selection methods, PCA, CARS, CARS-MIV and RF, were used for variable optimization of the whole spectrum data, and the classification model established by different variable selection methods was discussed, and compared with traditional BP neural network and PLS-DA model.

The results show that, moving smoothing (7 points) and SNV were used to process original Vis/NIR spectral data, the GA-BP neural network based on 16 characteristic wavelengths extracted by the CARS-MIV algorithm has the best discrimination effect on prediction set, the CARS-MIV-GA-BP model total correct discrimination rate of prediction set was 98.76%, among them, the identification rates of “Nongda-4”, “Nongda-5” and “Nongda-7” were 100%, 100% and 96.29%, respectively.

Therefore, the on-line detection of fresh CH fruit based on Vis/NIR spectroscopy combined with variable selection methods and the GA-BP model is an effective method. Also, in later studies, more intelligent algorithms can be selected to optimize the model and further improve the accuracy and versatility of the model.

ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (31271973); Key Research and Development Program of Shanxi Province, China (201703D221029-1)

REFERENCES

[1] Chen C. L., Jin Y., Cao Y. L., et al., (2018), Analysis of chlorophyll contents in maize leaf based on GA-BP neural network hyperspectral inversion model, *Journal of Shenyang Agricultural University,* Vol.49, Issue 5, pp.626-632, Shenyang/China;

[2] Chen X., Wang H. Y., Kong D. D., et al., (2016), Quality prediction model of pellet feed basing on BP network using PSO parameters optimization method, *Transactions of the Chinese Society of Agricultural Engineering,* Vol.32, Issue 14, pp.306-314, Beijing/China;

[3] Du X. B., He J. L., He Y. Q., et al., (2020), Design and experiment of eccentric swing combing device for cerasus humilis, *INMATEH-Agricultural Engineering,* Vol.60, Issue 1, pp.89-98, Bucharest/ Romania;

[4] Firmani P. , Bucci R., Marini F., et al., (2019), Authentication of “Avola almonds” by near infrared (NIR) spectroscopy and chemometrics, *Journal of Food Composition and Analysis,* Vol.82, pp.103235, San Diego /United States;
[5] Gu J., Yin G. H., Huang P. F., et al., (2017), An improved back propagation neural network prediction model for subsurface drip irrigation system, *Computers & Electrical Engineering,* Vol.60, pp.58-65, Oxford/England;
[6] Galvão R. K. H., Araujo M. C. U., José G. E., et al., (2005), A method for calibration and validation subset partitioning, *Talanta,* Vol.67, Issue 4, pp.736-740, Amsterdam/Netherlands;
[7] Hideyuki S., Pitiporn R., Yukihiro O. , (2011), Kernel analysis of partial least squares (PLS) regression models, *Applied spectroscopy,* Vol.65, Issue 5, pp.549-556, New York/United States;
[8] Li X., Liu Y. D., Ouyang A. G., et al., (2019), A general model for judging and brix detection of grapefruit variety based on near infrared, *Chinese Journal of Luminescence,* Vol.40, Issue 6, pp.808-814, Changchun/China;
[9] Liang Y., Ren C., Wang H. et al., (2019), Research on soil moisture inversion method based on GA-BP neural network model, *International Journal of Remote Sensing,* Vol.40, Issue 5-6, pp.2087-2103, Oxford/England;
[10] Liu D., Li S., Fu Q., et al., (2018), Comprehensive evaluation method of groundwater quality based on BP network optimized by Krill Herd algorithm, *Transactions of the Chinese Society for Agricultural Machinery,* Vol.49, Issue 9, pp.275-284, Beijing/China;
[11] Lu M. Y., Yang K., Song P. F., et al., (2018), The study of classification modeling method for near infrared spectroscopy of tobacco leaves based on convolution neural network, *Spectroscopy and Spectral Analysis,* Vol.38, Issue 12, pp.3724-3728, Beijing/China;
[12] Mohamad E. T., Armaghani D. J., Momeni E., et al., (2016), Rock strength estimation: a PSO-based BP approach, *Neural Computing and Applications,* Vol.30, Issue 5, pp.1635-1646, New York/USA;
[13] Sun J., Lu X. Z., Zhang X. D., et al., (2019), Identification of red bean variety with probabilistic GA-PNN based on hyperspectral imaging, *Transactions of the Chinese Society for Agricultural Machinery,* Vol.47, Issue 6, pp.215-221, Beijing/China;
[14] Sun J., Lu X. Z., Zhang X. D., et al., (2016), Identification of red bean variety with probabilistic GA-PNN based on hyperspectral imaging, *Transactions of the Chinese Society for Agricultural Machinery,* Vol.47, Issue 6, pp.215-221, Beijing/China;
[15] Tan K., Wang R., Li M., et al., (2019), Discriminating soybean seed varieties using hyperspectral imaging and machine learning, *Journal of Computational Methods in Sciences and Engineering,* Vol.19, Issue 4, pp.1001-1015, Amsterdam/Netherlands;
[16] Wu J., Huang F. R., Huang C., et al., (2013), Study on near infrared spectroscopy of transgenic soybean identification based on principal component analysis and neural network, *Spectroscopy and Spectral Analysis,* Vol.33, Issue 6, pp.1537-1541, Beijing/China;
[17] Xiang L., Deng Z. Q., Hu A. J., (2019), Forecasting short-term wind speed based on IEWT-LSSVM model optimized by bird swarm algorithm, *IEEE Access,* Vol.7, pp.59333-59345, New York/USA;
[18] Xie H., Chen Z. G., Zhang Q. H., (2019), Rapid discrimination of japonica rice seeds based on near infrared spectroscopy, *Spectroscopy and Spectral Analysis,* Vol.39, Issue10, pp.3267-3272, Beijing/China;
[19] Yan L., Pang L., Wang H., et al., (2020), Recognition of different Longjing fresh tea varieties using hyperspectral imaging technology and chemometrics, *Journal of Food Process Engineering*, Vol.43, Issue 4, pp.1-9, New York/United States;
[20] Yang D. F., Zhu H. D., (2013), Recognition of soybean varieties based on near infrared transmittance spectroscopy and BP neural network, *Soybean Science,* Vol.32, Issue 2, pp.249-253, Haerbin/China;
[21] Zhang H., Ma Y., Li Y., et al., (2014), Rupture energy prediction model for walnut shell breaking based on genetic BP neural network, *Transactions of the Chinese Society of Agricultural Engineering,* Vol.30, Issue 18, pp.78-84, Beijing/China;
[22] Zhang S. T., Wang Z. Y., Zou X. G., et al., (2017), Recognition of tea disease spot based on hyperspectral image and genetic optimization neural network, *Transactions of the Chinese Society of Agricultural Engineering,* Vol.33, Issue 22, pp.200-207, Beijing/China;
[23] Zhu X. L., Li G. H., Zhang M., (2019), Prediction of soluble solid content of Korla pears based on CARS-MIV, *Spectroscopy and Spectral Analysis,* Vol.39, Issue11, pp.3547-3552, Beijing/China.