Determination of Sugar Content in Apple by Hyperspectrum with Interval Sampling Algorithm for Sample Subset and Ga-Pls Model

Jie Chen, Na Yao, Hengyou Wang, Xiao Zhang, Haifang Lv*
College of Information Engineering, Tarim University, Alar, Xinjiang 843300, China
**Corresponding author*

Abstract

It is of vital importance of both the selection of data preprocessing and the selection of training set sample and modeling variables in the process of establishing the Partial Least Square (PLS) quantitative model of hyperspectral data. In this paper, we study the content of sugar in the sugar-sweetened apple, a characteristic fruit of Southern Xinjiang. We compare the effects of four pretreatment methods, i.e., multiple scattering correction (MSC), Savitzky-Golay (SG) smoothing, range standardization and centralized transformation on the model, and the influence of Random Sampling (RS), Kennard-Stone (KS), Sample Set Partitioning based on joint X-Y Distance (SPXY) and Spacing Selection on the model as well. By using totally 10-fold cross validations in the interaction test of the model (the interaction test of the model used 10 fold cross validation), we also give a comparison of the model established by genetic algorithm (GA) and full-wave PLS modeling. The results show that the best data preprocessing method is the combination of SG smoothing and range standardization, and the model established by the training set selected by the interval sampling method is superior to the other three methods. Compared with full-wave PLS modeling, the model established by GA-PLS has great improvement in terms of prediction speed, prediction ability and principal component selection. The correlation coefficient, RMS and prediction accuracy of the model are 0.936229, 0.919618 and 0.938575, respectively.

***Keywords**: Hyperspectrum, The training set, Range standardization, Genetic algorithm, Sampling interval*

Ⅰ. Introduction

The Fuji apple trees of Aksu grow in the sandy soil of the edge of the Taklamakan desert at the foot of Mount Tianshan, enjoying abundant sunshine all year round, being irrigated by the snow water of Tianshan Mountain. Moreover, there is large temperature difference between day and night in this region making the fructose condensed inside this kind of fruit, which is commonly known as the sugar-sweetened apple as well as the “queen of fruit” in Xinjiang. As the sugar-sweetened apples are popular, the lack of the production could not meet the consumers’ needs, inducing lots of fake products in the market. Therefore, it is imperative to study the quality of sugar-sweetened apples.

In the last decade or so, hyperspectral imaging technology has been widely used in the detection of agricultural products and food quality. In terms of information acquisition, it can realize rapid extraction of spectral information. This kind of nondestructive testing is very suitable for the internal testing of fruits and vegetables, which effectively makes up for the shortage of traditional manual testing methods. At the same time, it also ensures the correctness and accuracy of the test results[1-3].

Partial Least Square (PLS) is a common method for data processing in which useful information is extracted from complex information. However, if a more appropriate data preprocessing method is used, and a more representative training set and characteristic wavelengths are selected, the predictive reliability of the model will be enhanced. Currently, the commonly used data preprocessing methods include multiplicative scatter correction (MSC), Savitzky-Golay (SG) smoothing, centralized transformation and range standardization, etc.[4], and the commonly used selection methods of training sets include Random Sampling (RS) method, Kennard-stone (KS), Sample Set

Partitioning based on joint X-Y Distance (SPXY) and spacing partitioning method, etc.[5].

In this paper, we focus on the main quality parameter, i.e., the sugar content of Aksu Sugared Apple. The rest of this paper is organized as follows. Section 1 describes the process of GA PLS modeling. In section 2, the effects of the pre-processing methods such as MSC, SG smoothing, centralization transformation and range standardization, as well as the effects of RS, KS, SPXY and interval sampling on the division of training set and prediction set are investigated. Section 3 compares and analyzes the predicted effects of GA-PLS and full-wave PLS. Section 4 gives the conclusion remarks.

Ⅱ. Experimental Materials and Methods

The apples used in the experiment are evenly distributed in size, shape and color, and have a smooth surface without damage, being picked and kept in the cold room. Totally 160 apples are firstly taken out and placed at room temperature for 24 hours in advance before the experiment, each of which are then numbered, and finally hyperspectral images and sugar content of each apple are collected.

A. Acquisition of Hyperspectral Images

In order to avoid hyperspectral image acquisition environment and astigmatism interference imaging, the entire hyperspectral data acquisition system is placed in a customized black box when collecting hyperspectral. After each sample is released, the gate can be closed, and then the image can be collected by computer software. Figure 1 shows the hyperspectral image of the apple.

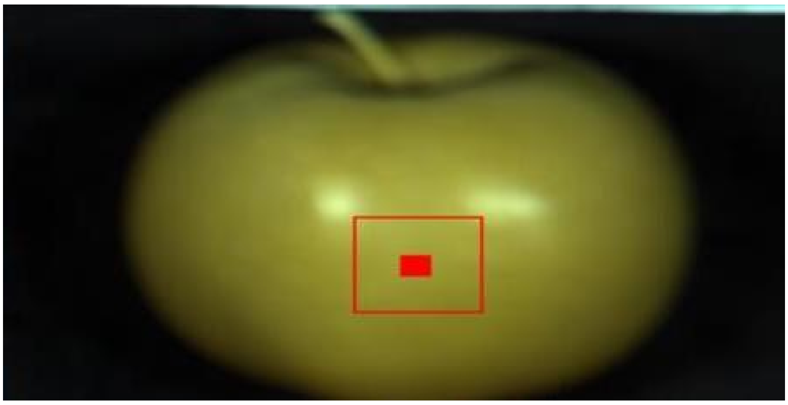

Fig.1 The Hyperspectral Image of the Apple.

Then the software ENVI4.7 is used to correct hyperspectral images and obtain spectral data. The spectral data are averaged and stored in Excel for later processing. Figure 2 shows the spectral curve.

ISSN: 0010-8189

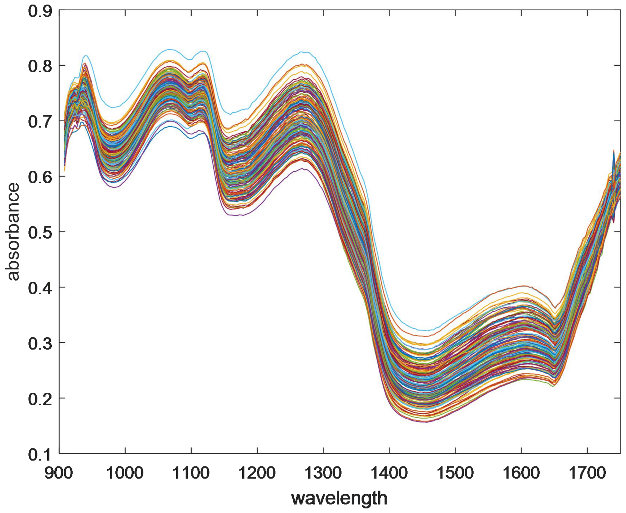

Fig.2 The Spectral Curve

B. Collection of Sugar Content Data

The sugar content is measured using the Pal-Hikari 5 portable harm-free Apple detector, which is measured at the location where spectral data are collected using the hyperspectral data acquisition system illustrated in Figure 1. The data are collected at each location for 5 times and the average value is taken.

Ⅲ. Principles and Methods

A. Data Preprocessing Method

The data preprocessing method is very important in hyperspectral analysis due to the influence of noise, baseline drift, sample inhomogeneity and light scattering during spectrum data acquisitioning.

1) SG smoothing method

The SG smoothing algorithm is an improvement over the moving smoothing algorithm, which is proposed by Savizky and Golay and is based on the least squares principle of polynomial smoothing algorithm, also known as convolution smoothing[6].

2) MSC

MSC is a commonly used data processing method in multi-wavelength calibration modeling. The spectral data obtained after the scattering correction can eliminate the spectral differences caused by different scattering levels, thus enhance the correlation between the spectrum and the data[7]. Firstly, an “ideal spectrum” of the sample to be tested is established before the method is used, i.e., the spectral change is directly linear to the content of components in the sample. The spectra of all other samples are modified using this spectrum as the standard. This includes baseline translation and offset correction. However, the spectral absorption information corresponding to the sample composition content will not be affected. Therefore, the signal-to-noise ratio of the spectrum is improved. The specific algorithm process is described as follows:

(1) Calculated mean spectrum

$$\overline{A} = \frac{\sum_{1}^{n} A_{i,j}}{n}, \qquad (1)$$

(2) Unitary linear regression

$$A_i = m_i \overline{A} + b_i, \qquad (2)$$

(3) Multiple scattering correction

$$A_{i(MSC)} = \frac{(A_i - b_i)}{m_i}, \qquad (3)$$

where *A* represents the spectral matrix and *n* the number of samples. The relative offset coefficient and shift amount obtained after the unitary linear regression of spectral A_i and average spectral *A* of each sample are expressed by m_i and b_i.

3) Centralization transformation

A centralization transformation is a transformation of variables minus their mean. It's just a matter of shifting the data. After translation, the center of all data is (0,0), which can make different features having the same scale.

4) Range standardization

Range standardization is also called deviation standardization. It is a linear transformation of the original data that maps the resulting value to the interval [0, 1]. The conversion function is given by:

$$x^* = \frac{x - \min}{\max - \min} \qquad (4)$$

where *x* represents data, the symbol min represents the minimum value, and max the maximum value. Range standardization can realize the standardization of positive and negative indexes, and reduce the influence caused by different types, dimensions, ranges and orders of magnitude of each measurement index.

B. Training Set Selection Method

1) RS

RS method is the random selection method. A specified number of samples are selected as training sets by random functions. The principle of this method is simple, while the training set selected is different each time due to its randomness, so the model established is not widely applicable[5].

2) KS

All samples are treated as training set candidate samples in KS. The training set is selected from the candidate samples in turn[5-8]. First, the two samples with the longest Euclidean distance are selected into the training set.

Then, the selected samples with the maximum and minimum distance are found by calculating the Euclidean distance between each remaining sample and each known sample in the training set, and put into the training set, and so on, until the required number of samples is reached. The Euclidean distance formula is

$$d_x\left(p,q\right)=\sqrt{\sum_{j=1}^{N}\left(x_p\left(j\right)-x_q(j)\right)^2};\quad p,q\in[1,N] \qquad (5)$$

In the above equation, x_p and x_q represent the spectra of the two samples, respectively, N represents the number of samples. The advantage of this method is that the samples in the training set can be evenly distributed according to the spatial distance. The disadvantage is that it requires a large amount of calculation which contains data conversion and calculation of space distance between two samples[8].

3) SPXY

SPXY is developed based on KS algorithm. It takes into account both the x and y variables when calculating the distance between samples[9]. Therefore, in addition to KS European distance formula, it also includes the following two expressions:

$$d_y\left(\mathrm{p,q}\right)=\sqrt{(y_p-y_q)^2}=\left|y_p-y_q\right|;p,q\in[1,N] \qquad (6)$$

$$.\,d_{xy}\left(\mathrm{p,q}\right)=\frac{d_x\left(\mathrm{p,q}\right)}{\max d_x\left(\mathrm{p,q}\right)}+\frac{d_y\left(\mathrm{p,q}\right)}{\max d_y\left(\mathrm{p,q}\right)};p,q\in[1,N] \qquad (7)$$

where x represents spectral data and y represents quality parameters.

Its modeling and forecasting capabilities are improved by including more space.

4) Interval sampling method

The interval sampling method sorts the quality parameters and spectral data first. Then the training set and test set are selected according to the ratio column of training set and test set. If, for example, the ratio between the training set and the test set is listed as 3:1, then the training process is like the following: Originally, the number of samples is divided by 4 for grouping. Next, the first three samples of each group are taken as the training set, and the last sample is taken as the test set, and so on. This kind of sampling method increases the ergodicity of the training set, so the model performance index will be better.

C. Modeling Method

1) GA - PLS model

GA-PLS model selects featured variables in the full spectrum region by using genetic algorithm (GA). After that, PLS method is used for modeling, and the principal component is selected. The model is evaluated by the 10-fold cross validation method.

2) Full-wave PLS model

The full-wave PLS model does not screen the spectral regions, and each component is involved in the modeling. The principal component is still selected by the 10-fold cross-validation method in the established model.

Ⅳ. Results and Discussion

A. Comparison of Pretreatment Methods

There are a total of 160 apple samples in this study. Four pretreatment methods including multiple scattering correction (MSC), SG smoothing, centralization transformation and range standardization are used, respectively. GA-PLS is used to model and evaluate the training samples under the premise of SPXY method. There are three evaluation indexes used in this paper, namely correlation coefficient *R*, root mean square error prediction *MSE* and *prediction precision*. The evaluation indexes of the model established by the data processed by the four pretreatment methods are shown in Table 1.

Table 1 Comparison of Pretreatment Methods

Methods	R	MSE	PRECISION
Range standardization	0.882334	0.845670	0.942739
s-g smoothing	0.852626	0.872517	0.946330
MSC	0.850269	0.803504	0.945671
Centralization transformation	0.716266	1.172702	0.920963
Range standardization+MSC	0.866482	0.777387	0.944672
Range standardization+SG smoothing	0.895814	0.79745	0.946345

According to the analysis of the results in Table 1, different pretreatment methods have a significant impact on the reliability of the established model. The best pretreatment is the combination of range standardization and SG smoothing. The correlation of the established model is 0.895814. Figure 3 shows the preprocessed image.

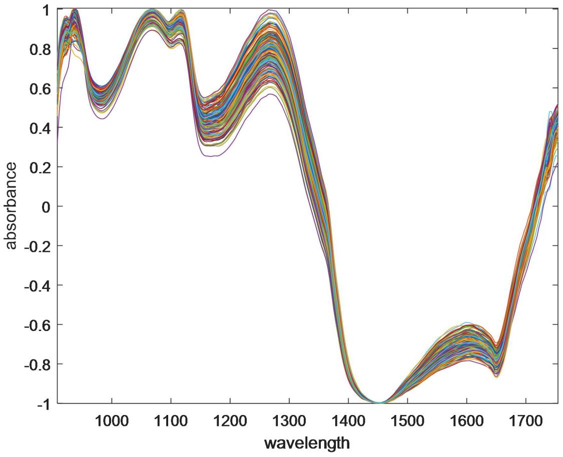

Fig.3 The Preprocessed Image

B. Comparison of Training Set Selection Methods

A total of 160 apple samples are taken, 75% of which are used as training sets and 25% as test sets. RS, KS, SPXY and interval sampling are used to select the training set. After that, GA-PLS is used to establish the model for the four methods and make prediction and evaluation. The results are shown in Table 2. As can be seen from Table 2, the correlation coefficient of interval sampling method is 0.936229, which is the largest. The correlation coefficient of SPXY sample selection method is 0.895814, which is the second largest one. Because of its randomness, RS method generates different evaluation indexes each time. So the average of 100 outcomes is taken as the last result, with the correlation coefficient 0.886556. The model established by KS method has the least correlation (0.71561).

Table 2 the Model Indexes Obtained by Different Training Set Selection Methods

	R	MSE	PRECISION
RS	0.886556	0.847847	0.951865
KS	0.71561	1.267949	0.924516
SPXY	0.895814	0.79745	0.946345
Sampling interval	0.936229	0.919618	0.938575

C. Comparison of Modeling Methods

Range standardization and SG smoothing are used for data preprocessing. Interval sampling method is used for *training set selection. GA-PLS method and full-wave PLS* model are established respectively. The evaluation indexes of the model are shown in Table 3. The model correlation established by full-wave PLS is 0.929385 with 18 principal components.

122 characteristic wavelengths are screened by GA-PLS for modeling. The correlation coefficient of the established model is 0.936229, with 12 principal components. Compared with the model established by full-wave PLS, the accuracy is improved, and the model is also very effective in simplification. Figure 4 and 5 are scatter plots of predicted and measured values of the two modeling methods, respectively. Figure 6 and 7 are the comparison between the predicted and measured values of the test set samples of the two modeling methods, respectively.

Table 3 the Evaluation Indexes of the Model

	R	MSE	PRECISION	Number of Wavelengths	Number of principal components
GA-PLS	0.9362	0.9196	0.938575	122	12
Full-wave -PLS	0.9294	0.9488	0.938096	~	18

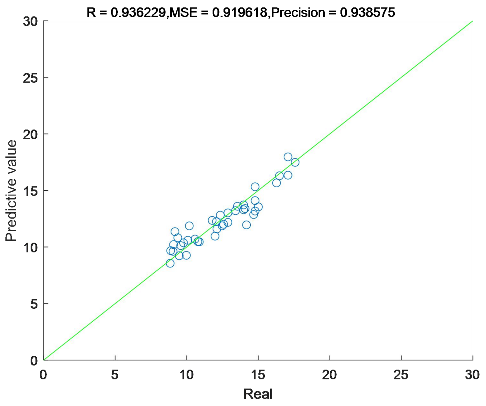

Fig.4 Scatter Plots of Predicted and Measured Values of the Ga-Pls

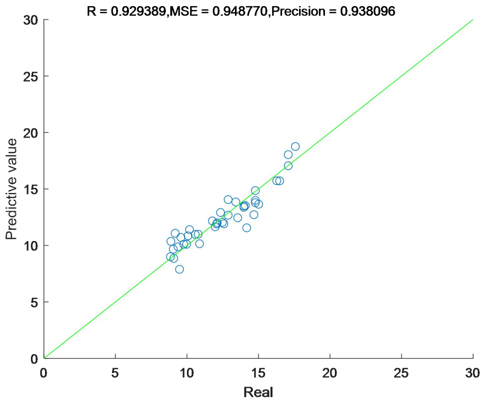

Fig.5 Scatter Plots of Predicted and Measured Values of the Pls

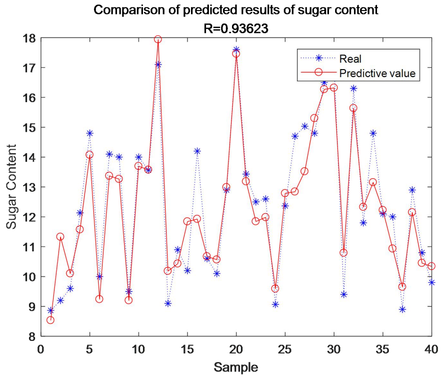

Fig.6 The Comparison between the Predicted and Measured Values of the Test Set Samples of the Ga-Pls

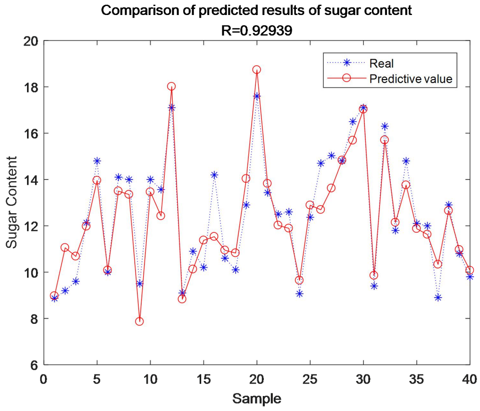

Fig.7 The Comparison between the Predicted and Measured Values of the Test Set Samples of the Pls

Ⅴ. Conclusion

In this paper, hyperspectrum is used to predict the sugar content of Akesu apples. Four data preprocessing methods, four training set selection methods and two modeling methods are discussed. The experimental results show that the range standardization and SG smoothing pretreatment have a good effect on hyperspectral data processing. The predictive ability of the training set selected by the interval sampling method is superior to the other three methods. The model established by GA-PLS method is not only improved in model prediction speed, but also greatly improved in prediction ability and principal component selection. Therefore, it is necessary to select appropriate data preprocessing method, training set selection method and feature wavelength screening when hyperspectral combination with PLS is used to determine apple saccharinity. And it can effectively improve the extrapolation ability of the model.

References

[1] Peng Y, Lu R. Analysis of spatially resolved hyperspectral scattering images for assessing apple fruit firmness and soluble solids content [J]. Postharvest Biology & Technology, 2008, 48(1):52-62.

[2] Wu J , Peng Y , Jiang F , et al. Hyperspectral Scattering Profiles for Prediction of Beef Tenderness[J]. Transactions of the Chinese Society for Agricultural Machinery, 2009, 40(12):135-138+150.

[3] Tao F, Peng Y, Gomes C L, et al. A comparative study for improving prediction of total viable count in beef based on hyperspectral scattering characteristics [J]. Journal of Food Engineering, 2015, 162(oct.):38-47.

[4] Jian Y, Jiyu G, Qibing Z . Predicting bruise susceptibility in apples using Vis/SWNIR technique combined with ensemble learning [J]. International Journal of Agricultural & Biological Engineering, 2017, 10(5):144-153.

[5] Zhu X. F, Pan Y, Zhang J.S, etc. The training sample of TM scale wheat planting area measurement precision influence study (Ⅰ) - classification accuracy response relationship between the training sample and classification method research [J].Journal of Remote Sensing, 2007, 11(6):826-837.

[6] Liu G.S, Guo H.S, PAN T, et al. Vis-NIR spectral pattern recognition combined with SG smoothing for transgenic sugarcane breeding screening [J].Spectroscopy and Spectral Analysis, 2014, 34(010):2701-2706.

[7] Lu Y.J, Qu Y.L, Song M. Correction of multiple Scattering in near infrared Correlation Spectra [J].Spectroscopy and spectral analysis,2007,27(5):877-880.

[8] Wu J.Z, Wang Y.M, Zhang X.C, et al. Study on sample selection method of calibration set in near infrared spectroscopy [J].Journal of Agricultural Machinery, 2006(04):86-88+107.

[9] Galvo R K H , Mário César Ugulino Araujo, Gledson Emídio José, et al. A method for calibration and validation subset partitioning[J]. Talanta, 2005, 67(4):736-740.