International Journal of Hybrid Information Technology
Vol. 9, No.9 (2016), pp. 35-44
http://dx.doi.org/10.14257/ijhit.2016.9.9.04

Hyperspectral Determination of Reducing Sugar in Potatoes Based on CARS

Wei Jiang[12], Junlong Fang[1]*, Shuwen Wang[1] and Runtao Wang[1]

[1]*School of Electrical Engineering and Information, Northeast Agricultural University , Harbin, 150030,China*
[2]*Department of Computing ,Harbin Finance University, Harbin, 150030,China*
jwhancg@126.com

Abstract

It usually contains a large amount of redundant information to use the hyperspectral information to create a model, which will increase the difficulty of the model analysis. Therefore, it's so important to select the characteristic wavelength in an effective and quick way. This study is proposed by using the competitive adaptive reweighed sampling (CARS) to select the characteristic wavelength for detecting the reducing sugar content in the potatoes. In that experiment a total of 238 samples are prepared. Among them, 190 samples are selected as the calibration set and 48 samples as the validation set. The performance of CARS is compared with full spectrum and classical variable extraction methods such as Monte Carlo uninformative variable elimination (MC-UVE), genetic algorithm (GA) and moving window partial least squares (MWPLS). Experimental results show that the band screened by algorithm CARS has the best effect, compared to full spectrum modeling, the wavelength of the model reduces from 203 to 33, the model validation set coefficient R^2 increases from 0.8464 to 0.8965, and the root mean square error prediction (RMSEP) decreases from 0.0758 to 0.0416. The results demonstrate that it is feasible to detect the reducing sugar content of potatoes by using CARS combined with hyperspectral imaging.

Keywords: *Hyperspectral; CARS; Potato; Partial least square*

1. Introduction

Reducing sugar content in potatoes is one of the important factors that affect the processing quality of potatoes [1]. It's significant for the deep processing of potatoes to determine reducing sugar content of potatoes in an accurate and rapid way. At present, the determination method of reducing sugar is a traditional one, such as electrochemical method and colorimetric method whose detection precision is higher. But it is difficult to popularize in the measurement of a large number of samples due to many complicated steps , a long time, high cost *etc*.[2].It has an important application prospect to discuss a rapid detection method of the Reducing sugar content of potato. Hyper spectral imaging technology is a new type of agricultural products nondestructive detection technique, which is a perfect fusion of traditional imaging techniques and spectroscopic techniques. In this way, it can obtain the spatial and continuous spectrum information of research object so that it gets the favor of researchers who are committed to studying the nondestructive testing field of the quality of agricultural products at home and abroad [3]. At present, this technique is applied to the detection of fat and moisture content of potato chips [4], and the content of potato tuber moisture, starch, protein, reducing sugar and crude fiber and potato tubers potassium *etc*. [5] However, due to the hyperspectral resolution, a large amount of collinearity and redundant data is present in the original spectral information. Therefore, it is necessary to select the

ISSN: 1738-9968 IJHIT
Copyright © 2016 SERSC

key variables before the quantitative analysis of the internal quality of agricultural products by using hyperspectral data. Studies have shown that the model that is more easily to be explained and has more stable performance can be obtained via spectrum optimization [6,7].

In this paper, the main task is to use competitive adaptive reweighed algorithm (CARS) to select the characteristic wavelength after obtaining spectral information of potato based on hyperspectral imaging technology by using potato as the research object, and also respectively create partial least squares (PLS)model by comparing with extraction methods of the full spectrum and other variables , such as Monte Carlo uninformative variable elimination (MCUVE), genetic algorithm(GA) and move window partial least squares(MWPLS) , and use the validation set to confirm the validity of the model through comprehensive comparison of various variable selection methods in the prediction results of reducing sugar content in the potato so as to obtain the optimal application of various variable selection methods of hyperspectral in quantitative analysis of the quality.

2. Materials and Methods

2.1. Sample and Instrument

The potatoes with different varieties from surrounding of Harbin city, Heilongjiang Province, are chosen as the research object of the experiment. A total of 238 samples in the removal of surface defects are used for image collection. Among them, 190 samples are select randomly as the model sample set, and the other 48 as the prediction sample set. Note that the potatoes must be cleaned clearly before experiment.

The hyperspectral image acquisition system made in the HeadWall Company, USA, shown as Figure 1 is used in this experiment. The system consists of three parts, namely image acquisition unit, a light source and a sample conveying platform. The image acquisition unit includes an image spectrometer, CCD camera and lens. The light source is a fiber halogen lamp with 150W adjustable power. The slit width of hyperspectral image spectrometer is 25μm and the spectral range is from 400 to 1000nm. The spectral resolution is 1.29nm, and spacing of image acquisition band is 3nm, and the spatial resolution is 0.15mm.

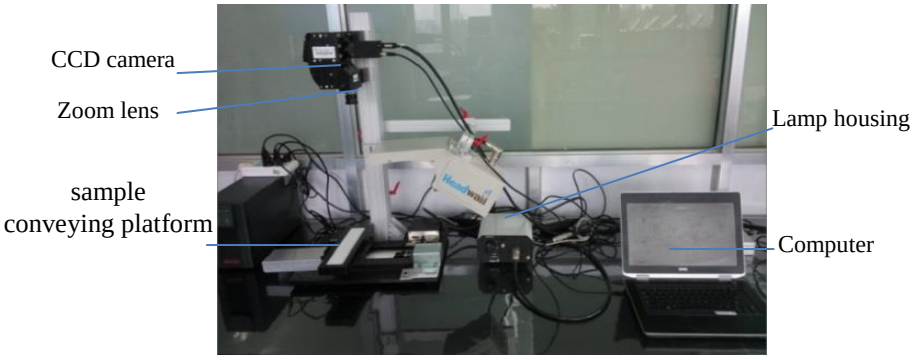


Figure 1. Hyperspectral Image System [8]

2.2. Image Acquisition

Before acquiring the hyperspectral image, the exposure time of spectral camera is set so as to ensure to get clear images according to the intensity of the light source, and the speed of conveying device need to be adjusted to avoid the distortion of image size and spatial resolution. After several debugging, the related parameters

 Copyright © 2016 SERSC

including 30 cm of object distance, 350ms of exposure time, 70W of the power for the light source are determined finally. During each experiment process, a selected potato is placed on a white background plate, and then put on the objective stage. When acquiring image, linear detector is horizontal scan in the vertical direction of the optical focal plane. As the movement of objective platform, linear detector scans the image of the entire plane so as to complete the sample image acquisition.

Under different wave band, the non-uniform distribution of the light source intensity and the influence of the dark current in the camera result in too large noise contained in light source distribution of weak band. So the hyperspectral image must be corrected. In order to reduce the interference caused by temperature change of light source on the image, a full white calibration image and a full black calibration image are acquired once when every 20 sample images are collected in this experiment. And the hyperspectral image is obtained and calibrated according to eqn. (1) [9-12].

$$I=\frac{I_s - I_d}{I_w - I_d} \quad (1)$$

In the equation, *I* is a calibrated image; *Is* is the original sample image; *Iw* is the reference image acquired from the white spectra on calibration panel; and *Id* is the dark current image obtained with a cap covering the camera lens.

2.3. Data Preprocessing

In this paper, 9 abnormal samples are removed by using Monte Carlo sampling (MCS) algorithm [13,14] . To weaken or eliminate the impact on the spectra from all kinds of aimless factors such as baseline drift, scattering, it is necessary to preprocess the spectra collected by the hyperspectral imager. [15]. The original spectra are preprocessed by using the methods of smooth 13 points, the maximum normalization, baseline correction, orthogonal signal correction and standardization respectively, and followed by comparison of influence on the calibration model with spectrum after pretreatment , as well as preprocessing method is determined by the principle of determination coefficient of aimless factor as maximum, root mean square error as minimum. In addition, in order to improve forecasting the model's ability, remove the information that has nothing to do with the sample itself as soon as possible.

2.4. Competitive Adaptive Reweighed Sampling (CARS)

Competitive adaptive reweighed sampling (CARS) is a method of variable selection put forward according to the simulation of the law of "survival of the fittest" in the Darwin's theory of evolution [16]. Each time wavelength points with higher absolute value of regression coefficient in the PLS model can be screened through the adaptive reweighed sampling (ARS) technology, as well as the ones with lower weight can be removed, the minimal subset selected from the root mean square error of cross-validation (RMSECV) in the PLS model by Using cross validation is defined as the optimal variable subset. And the detailed algorithm as follows :

It is assumed that the spectral matrix $X_{m\times p}$ is regarded as measured sample, in which *m* is sample number and *p* is the number of variables , $Y_{m\times 1}$ represents target response vector. T is the scoring matrix of X, which is a linear combination of X and W , *w* is combination coefficient, and *c* represents the regression coefficient vector of PLS calibration model created through Y and T, e is the prediction error. And then (2) and (3) relationship establishment as follows:

Copyright © 2016 SERSC

$$T = XW \tag{2}$$

$$Y = Tc + e = XWc + e = Xb + e \tag{3}$$

In the equation, $b = Wc = [b_1, b_2, \cdots, b_p]^T$ represents a p-dimensional vector of coefficients. $|b_i|(1 \le i \le p)$, the absolute value of i-th element of b, represents the contribution of the i-th wavelength to Y, and the larger the value of the variable is, the more important it is. To evaluate the importance of each band, weight is defined as:

$$\omega_i = \frac{|b_i|}{\sum_{i=1}^{p} b_i}, i = 1,2,\cdots, p \tag{4}$$

All the weights of variables removed by CARS algorithm are set to 0. The main flow [17] is shown in Figure 2 ,and it's retention rate of variable $r_i = ae^{-ki}$

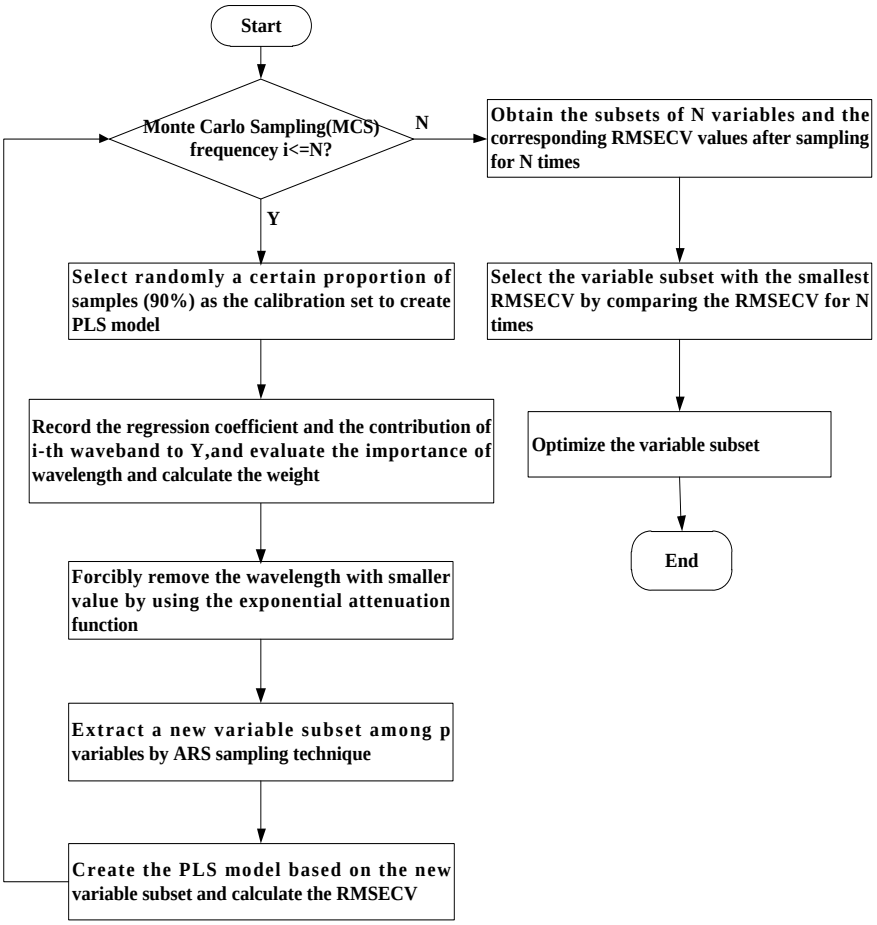


Figure 2. Flow Chart of CARS Algorithm

Among them, both a and k represent different constants respectively when samples concentrate all of p variables involved in modeling to conduct MCS sampling at the first time and just 2 variables involved in it at the N-th time. Namely, $r_1 = 1$ and $r_N = 2/p$, and then:

$$a = (\frac{p}{2})^{1/(N-1)} \tag{5}$$

 Copyright © 2016 SERSC

$$k = \frac{\ln(p/2)}{N-1} \tag{6}$$

In this paper, the number of variables p is set to 203, and the MC sample frequency is set to 200, so the values of constant a and k are 1.0235 and 0.0232 respectively.

2.5. Model Evaluation

First, parameters of the model is optimized by using cross validation method, and then the model is evaluated according to the decision coefficient, the root mean square error (RMSEC), root mean square error of prediction (RMSEP) and root mean square error of cross validation(RMSECV). Higher the model's decision coefficient, smaller the RMSEP and RMSECV, and stronger the model's forecasting ability.

3 Results and Discussion

In this paper, the content of reducing sugar content in fresh potato is used as the modeling object, and the Matlab R2013a is used to analyze the 203 bands with the range from 400 to 1000nm.

3.1. Selection of Preprocessing Method

It is shown in Figure3 that interest region of original spectra of the sample is accompanied with wavelength range from 400 to 1000nm, and each spectrum contains 203 bands. It is easy to find that the sample spectral curve trend is the similar, and no obvious abnormal sample is found through observing Figure 3. It is also found that there is a larger scatter and baseline drift in the spectral region due to the rough surface of the potato and the astigmatism of the environment *etc*. Therefore, the preprocessing methods such as smoothing 13 points, maximum normalization, baseline correction, orthogonal signal correction and standardization, are used to create the PLS model respectively before the spectrum is analyzed further, and it can be found that the effect of smoothing pretreatment is the best after comparison and analysis.

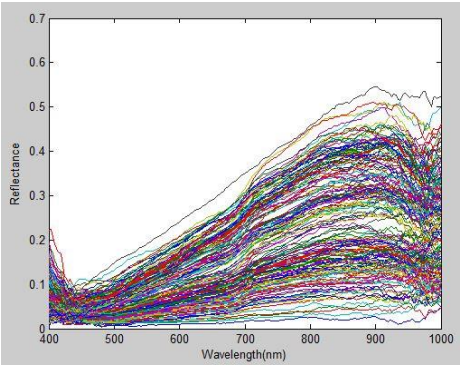


Figure 3. Spectra of Samples

It is shown in Table 1 that diverse results of PLS modeling are forecasted according to different preprocessing methods for reducing sugar content of potato. It can be found from Table 1 that performance of PLS model which is created through 13 point smoothing preprocessing is the best, whose decision coefficient R^2 and the root mean square error (RMSEP) are 0.8516 and 0.0671 respectively. With the

Copyright © 2016 SERSC

increase of smooth points, the performance of PLS model is gradually decreased. After the maximum normalized and orthogonal signal correction, the performance of PLS model is slightly decreased, while it is the worst after the pretreatment of baseline correction, and the RMSEP is 0.0791.

Table 1. Results of PLS Regression of Different Pretreatment Methods

pretreatment	R^2	RMSEC	R^2	RMSEP
The original spectrum	0.8254	0.0689	0.8160	0.0729
Smooth 13 points	0.8801	0.0573	0.8516	0.0671
Maximum normalization	0.8241	0.0692	0.8128	0.0732
Baseline correction	0.8202	0.0696	0.8069	0.0791
Orthogonal signal correction	0.8245	0.0691	0.8139	0.0746
Standardization	0.8621	0.0637	0.8400	0.0685

3.2. Key Variable Selection

3.2.1. CARS Variable Selection

The full spectral variables of reducing sugar content of fresh potato are screened for many times by using CARS algorithm, finally 33 wavelength points are selected ,and the results are shown in Figure4.During screening, the Monte Carlo sampling number is set to 200.

With the increase of the number of sampling ,(a), (b) and (c) represent the changing of the number of variables, the cross validation RMSECV and the regression coefficient of each variable respectively in the algorithm running.

It can be seen in Figure4 (a),Under the influence of exponential decay function, the number of variables selected declines from fast to slow with the increase of the number of samples, which reflects the algorithm can be used to elect rough and select variables in the variable selection, and improve the efficiency of the algorithm greatly.

It also can be seen in Figure 4 (b), with the increase of the number of samples the cross validation RMSECV value of the single PLS model is decreased first and then increased, it reaches minimum when the number of sampling is 28. It demonstrates that a large number of unrelated information is removed from the prediction of reducing sugar content in the high spectra, and RMSECV begins to increase after the 43 sampling, which proves that some key information is removed so that the model's performance is becoming poor.

The PLS model is created by using CARS algorithm to screen the spectral data of each band, compared with the full band model. By Table 2, both RMSECV and RMSEP obtained through modeling after CARS variable selection are better than the ones obtained through full band, and the model quality is improved obviously. In addition, the number of bands is reduced from 203 to 33, and the number of variables of the model is reduced significantly.

 Copyright © 2016 SERSC

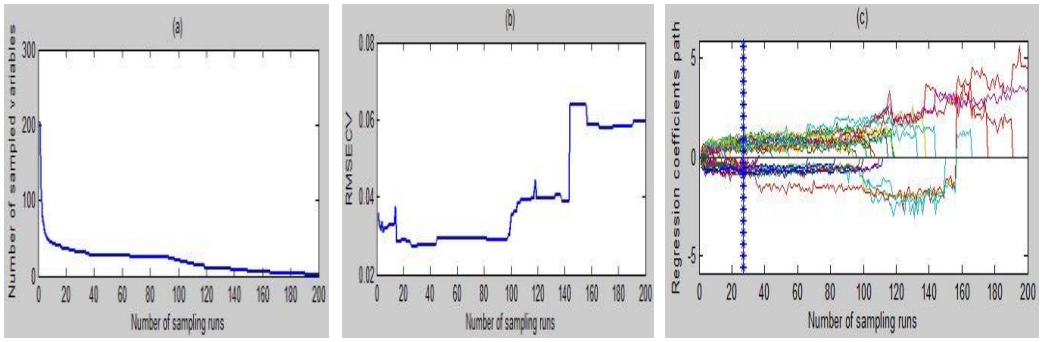
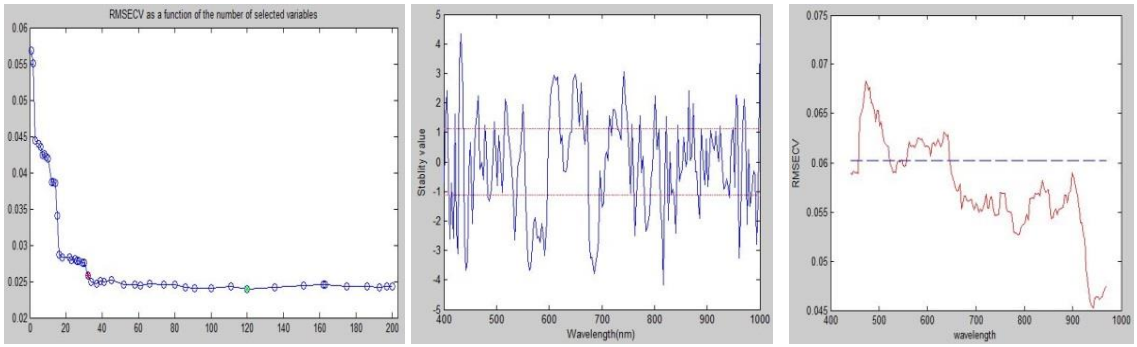


Figure 4. Key Variables Screening by CARS

3.2.2. Other Variable Selection Algorithm

Genetic algorithm (GA) is an optimization method [18] simulating the genetic and evolutionary processes in the biological community based on the theory of biological evolution of Darwin's, namely survival of the fittest.

In this study, the PLS model of reducing sugar content is created according to setting the number of genetic iterations to 200, the initial population size to 50, the crossover rate to 50%, and the mutation rate to 0.5%, choosing "F=RMSE" as the fitness function , as shown in Table 2. And then the number of optimal variables is determined according to the number of variables selected and RMSECV. Figure5 (a) indicates the corresponding relationship between RMSECV and the number of variables. When the number of variables is 119, the minimum RMSECV=0.0249 is obtained.



(a) Variables screening by GA (b)Variables screening by MC-UVE (c) Variables screening by MWPLS

Figure 5. Key Variables Selection Results. The Iteration Number of GA and MC-UVE Is 200.The Window Size of MWPLS is Fixed at 27

MC-UVE is also a relatively new variable selection method, based on the analysis of the PLS regression coefficient c algorithm, which is used to eliminate variables that cannot provided effective information. MC sampling technique is used to sample for N times by using this method, extract a percentage of the samples from a sample set as the modeling set for PLS modeling every time, and select variables by evaluating the stability of each variable. UVE is used to select the 203 wavelength points, as shown in Figure 5 (b), and the dotted line indicates the stability of the line. Finally, 95 wavelengths are selected by the UVE variable selection method, in this way, PLS model is created, and the forecast results are shown in Table 2.

Copyright © 2016 SERSC

The hyperspectral data of potato contains the target information of reducing sugar, it is also interfered by instrument noise or non target, such as starch, cellulose, protein, *etc*.. Hyperspectral information corresponding to reducing sugar is complex, ranging different Width of interval. Moving window partial least squares method, in a certain cross validation of the root mean square error level, can obtain a number of Intervals of hyperspectral information which reducing sugar content corresponding to.In this paper, we use the moving window partial least squares (MWPLS) method to locate the information section of the calibration set of potato samples by setting the window width to 27 and the upper limit of PLS extracted to 15, and the result is shown in Figure 5 (c).The inverted peak curves figure is formed by each variable point RMSECV changing with the position of the window, in which the dotted lines indicate it is more appropriate that root mean square error of cross validation is 0.0603 when the full spectrum contains 12 principal components. From Figure 5, we can see that RMSECV value is smaller when the wavelength range from 450 to 470nm, from 520 to 560nm, from 730 to 810nm, from 860 to 890nm and from 910 to 980nm. And then the PLS model is created through the above 106 feature spectral variables combined into a new data set as the optimal wavelength combination for variable selection. The results are shown in Table 2.

It can be found in Table 2 that prediction results from the GA-PLS model (r^2_{pre} and RMSEP are 0.8521 and 0.0683 respectively) are higher than the MC-UVE-PLS and MW-PLS models, and also better than the full PLS model. Further variable selection is helpful to improve the performance of the model. Compared with the CARS-PLS model, the prediction performance of the both is almost the same. However, the GA-PLS model uses more 72.24% variables (33 and 119), and the r^2pre has only increased by 0.08% compared with the CARS-PLS model. Therefore, the wavelength selection of CARS is the strongest among the above four variable selection methods.

Table 2. Performance of PLSR Model

Method	Number of variables	Number of factor	Calibration set		Validation set	
			r^2_{cal}	RMSEC	r^2_{pre}	RMSEP
Full -PLS	203	9	0.8516	0.0729	0.8464	0.0758
CARS-PLS	33	8	0.8610	0.0625	0.8513	0.0651
MCUVE-PLS	95	8	0.8541	0.0632	0.8441	0.0732
MW-PLS	106	12	0.8599	0.0637	0.8514	0.0672
GA-PLS	119	7	0.8612	0.0618	0.8521	0.0683

3.3. Model Validation

The prediction model of the potato reducing sugar content of CARS-PLS was used to predict the prediction sample which are not included in the model. Figure 6 shows the prediction of the content of reducing sugar of the prediction samples, the solid lines are regression lines of refrence values and predicted values. In Figure 6, the regression line and the straight line with the slope equal to 1 are very close, the expression of the regression line is y=0.85x+0.02, suggesting that hyperspectral combined with CARS selection can effectively predict the content of reducing sugar of potatoes.

 Copyright © 2016 SERSC

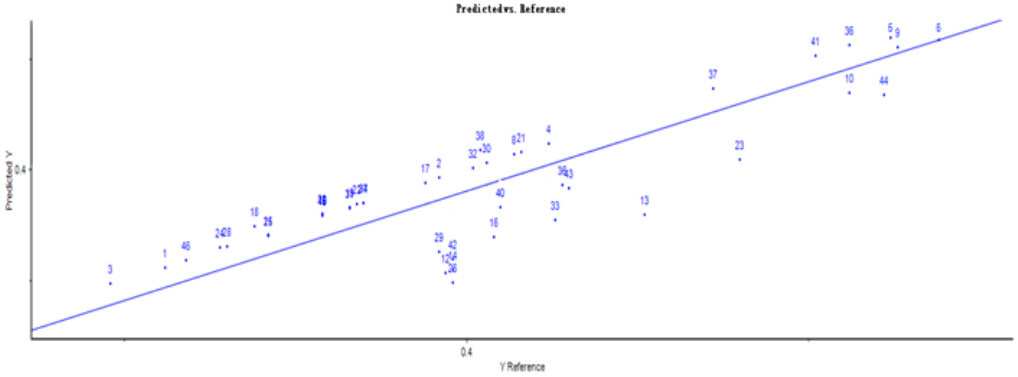


Figure 6. Predicted Results of Reducing Sugar in Potato By CARS-PLS in Prediction Set

4. Conclusions

Firstly, the content of reducing sugar in potato was predicted by using hyperspectral imaging technology combined with competitive adaptive reweighed sampling algorithm. The results show that the performance of the key variables obtained by CARS algorithm is better than that of the full-spectrum PLSR model. At the same time, the CARS algorithm is superior than them on the variable selection, compared with MWPLS, GA and MC-UVE. In the end, the 203 variables of the original spectrum are reduced to 33, the r^2pre and RMSEP of the PLSR model are 0.8513 and 0.0651 respectively, and the results are better than those of other variables. Moreover, CARS can carry on the effective quantitative analysis to the reducing sugar content of potato.

Acknowledgements

This work is financially supported by the Natural Science Foundation of Heilongjiang Province (Grants Nos. C2015006) , Harbin science and technology innovation talents research special fund(Grants Nos. RC2015QN009056) ,Heilongjiang province postdoctoral research fund(Grants Nos. LEB-013d22).

References

[1] H. Zhu, Y. Shi and Q. Zhang, "Applying 3,5-dinitrosalicylic Acid(DNS) Method to Analyzing the Content of Potato Reducing Sugar", Chineese Potato, vol. 19, no. 5, **(2005)**, pp. 266-269.
[2] C. Wang, Y. Chen and Y. Shi, "Research progress of Influence the quality for Fried potato", Chineese Potato, vol. 3, **(2003)**, pp. 23~24.
[3] B. M. Nicolai, K. Beullens and E. Bobelyn, "Nondestructive measurement of fruit and vegetable quality by means of NIR spectroscopy:A review", Postharvest Biology and Technology, vol. 46, no. 2, **(2007)**, pp. 99-108.
[4] S. Cecilia and R. Luis, "Application of NIR and MIR spectroscopy in quality control of potato chips", Journal of Food Composition and Analysis, vol. 22, no. 6, **(2009)**, pp. 596-605.
[5] C. E. Larsen, R. Trip and C.R. Johnson, "Methods for procedures related to the electrophysiology of the heart", U.S. Patent 5, vol. 529, no, 067, **(1995)**.
[6] M. R. Balabin and S. V. Smirnov, Analytica Chimica Acta, vol. 692, no. 1, **(2011)**, pp. 63.
[7] X. B. Zou, J. W. Zhao and J. W. Malcolm, Analytica Chinica Acta, vol. 667, no. 1,**(2011)**, pp. 14.
[8] W. Jiang, J. Fang, S. Wang and R. Wang, "Detection of Starch Content in Potato Based on Hyperspectral Imaging Technique.Signal Processing", Image Processing and Pattern Recognition, vol. 8, no. 12, **(2015)**.

Copyright © 2016 SERSC

[9] Z. Zhou, X. Li and H. Gao, "Comparison of Different Variable Selection Methods on Potato Dry Matter Detection by Hyperspectral Imaging Technology", Transactions of the Chinese Society for Agricultural Machinery, vol. 43, no. 2, **(2012)**, pp. 128-133.
[10] S. Min, N. Li and M. Zhang, "Outlier Diagnosis and Calibration Model Optimization for Near Infrared Spectroscopy Analysis", Spectroscopy and Spectral Analysis, vol. 24, no. 10, **(2004)**, pp. 1205-1209.
[11] W. Lu, H. Yuan and G. Xu, "Modern analysis technique of near infrared spectroscop", Beijing: China Petrochemical Press, **(2005)**.
[12] D. Cao, Y. Liang and Q. Xu, "A new strategy of outlier detection for QSAR/QSPR", Journal of Computational Chemistry, vol. 31, no. 3, **(2010)**, pp. 592-601.
[13] G. ElMasry, N Wang and C. Vigneault, "Detecting chilling injury in Red Delicious apple using hyperspectral imaging and neural networks", Postharvest Biol.Technol, **(2009)**, vol. 52, no 1, pp. 1-8.
[14] H. D. Li, Q. S. Xu and Y. Z. Liang, Analytica Chimica Acta, vol. 648, **(2009)**, pp. 77-84.
[15] B. Zhan, J. Ni and J. Li, "Hyperspectral Technology Combined with CARS Algorithm to Quantitatively Determine the SSC in Korla Fragrant Pear", Spectroscopy and Spectral Analysis, vol. 34, no. 10, **(2014)**, pp. 2752-2757.
[16] S. Min, N. Li and M. Zhang, "Outlier Diagnosis and Calibration Model Optimization for Near Infrared Spectroscopy Analysis", Spectroscopy and Spectral Analysis, vol. 24, no. 10, **(2004)**, pp. 1205-1209.
[17] Q. Kong, Z. Su and W. Shen, "Research of Straw Biomass Based of NIR by Wavelength Selection of IPLS-SPA", Spectroscopy and Spectral Analysis, vol. 35, no. 5, **(2015)**, pp. 1233-1238.
[18] G. Liu, H. Jiang and C. Mei, "Transaction of the Chinese Society of Agricultural Engineering, vol.29, no.1", **(2013)**, pp. 218.

Authors

Wei Jiang, she was born in 1980, in China. PhD, Lecturer. Research Interests: Electrical Engineering, Science and Technology of Computer.

E-mail: jwhancg@126.com

Corresponding author: Junlong Fang, PhD, professor, Research Interests: Electrical Engineering. Email:58525638@qq.com,

Mailing address: 59 Mucai Street, Harbin, 150030,China,

Tel:13945697595

 Copyright © 2016 SERSC