Weighted Wavelet Packet Domain Regression for Analysis of Near-infrared Spectroscopy at Different Temperatures

Dan Peng, Jingyun Wang, Huanhuan Hou

College of Grain Oil and Food Science, Henan University of Technology, Zhengzhou, China, 450001

pengdanhaut@126.com

Keywords: N-way partial least square regression; weighted strategy; discrete wavelet packet transform; prediction accuracy

Abstract. To efficiently make use of the temperature information in near-infrared (NIR) spectra, a new hybrid algorithm named as WP-WNPLS is proposed to improve the prediction ability of partial least square (PLS) based regression model. In WP-WNPLS, the discrete wavelet packet transform (DWPT) was firstly applied to decompose the 3-D NIR spectra into a series of frequency components. In each frequency component, a sub-model was obtained through using N-way PLS (NPLS) regression. Then, the weighted strategy was employed to take the advantage of multi-scale properties, and all the sub-models were mixed together to build the final weighted-prediction model. To validate the WP-WNPLS algorithm, it was applied to measure the fat concentration of milk using NIR spectra at different temperatures. The experimental results showed that the prediction ability of model obtained was superior to that obtained using conventional PLS algorithm, and the root mean square error of prediction can improve by up to 18.1%, indicating that it is a promising tool for NIR spectra regression model development.

Introduction

In the past years, near-infrared (NIR) spectroscopy has become a very useful analytical tool in the measurement of the composition of analyte mixture because of its rapidity and simplicity [1]. To efficiently use NIR technique, the multivariate model becomes the key problem. Thus, an appropriate calibration model is a matter of primary importance for analysis of unknown samples using NIR spectroscopy. However, the prediction results are sensitive to the variation of external conditions such as temperature, pressure and so on. Therefore, most of the existing NIR-based calibration models can not meet the requirements of practical application [2]. As well known to us, the temperature is the most important factor during measurement. The condition of temperature can change not only the absorbency, but also the waveform of spectra [3]. If the sample containing large amount of water, the variation of temperature can lead to the nonlinear variation of absorption in the second overtone of O-H bond [4]. Thus, it is very important to develop the calibration method for analyzing the sample at different temperatures. To solve this problem, the temperature can be taken as a parameter for inputting, but the dimension of NIR spectra becomes higher. In this paper, a novel hybrid algorithm named as WP-WNPLS is proposed for taking advantage of multi-scale properties of 3-D NIR spectra and useful information in temperature dimension during the process of analysis model construction. Experimental results show that the WP-WNPLS algorithm can effectively improve the model precision in 3-D NIR spectra analysis.

Principle and Method

N-way Partial Least Squares (NPLS) Algorithm

Assume $\underline{X}$ is a $m\times n\times l$ data matrix with m samples (first order), n measurements in the second order and l measurements in the third order. In addition, assume Y is the $m\times p$ concentration matrix with p calibration properties in m samples. NPLS [5] algorithm is the high dimensional version of PLS. In this paper, the emphasis is on the 3-D space. The method of applying NPLS algorithm to 3-D space can be described as:

Step1. Unfold $\underline{X}$ as a two dimensional matrix $\underline{X}_2$ ($m\times n\cdot l$). The element of $\underline{X}_2$ can be arranged as

$$\underline{X}(i,j,k)=\underline{X}_2(i,(j-1)\times n+k) \tag{1}$$

Step2. Perform centralization algorithm on $\underline{X}_2$, and then fold it according to (11) to obtain the centered matrix $\underline{X}_{center}$.

Step3. Let u be equal to a column of Y.

Step4. Let f (latent variable) be equal to 1.

Step5. Compute matrix Z as

$$Z(j,k)=\sum_{i=1}^{m}u(i,1)\cdot\underline{X}_{center}(i,j,k) \tag{2}$$

Perform SVD decomposition on Z as $(w_f^J,w_f^K)=SVD(Z)$.

Step6. Calculate the vector H as

$$H(i,1)=\sum_{j=1}^{n}\sum_{k=1}^{l}\underline{X}(i,j,k)\cdot w_f^J(j,1)\cdot w_f^K(k,1) \tag{3}$$

Step7. Calculate vector q_f as

$$q_f=Y^TH/\sqrt{Y^TH\cdot(Y^TH)^T} \tag{4}$$

Renew vector u as $u=Y\cdot q_f$. If vector u is not convergent, algorithm switches to Step5.

Step8. Calculate vector b_f as $b_f=(H^TH)^{-1}H^Tu$.

Step9. Renew $\underline{X}_{center}$ and Y as

$$\underline{X}_{center}(i,:,:)=\underline{X}_{center}(i,:,:)-H_f(i,1)\cdot w^J\cdot(w^K)^T \tag{5}$$

$$Y=Y-H_f\cdot b_f\cdot q_f^T \tag{6}$$

where $\underline{X}_{center}(i,:,:)$ denotes the two dimensional matrix with the first order index equaling to i. Let variable f increase by one, and the algorithm continues from Step5 until the proper description of Y is obtained.

WP-WNPLS Algorithm

The aim of WP-WNPLS (wavelet packet transform-weighted n-way partial least squares) is that multi-scale properties of NIR spectra will effectively identify and encode more aspects of the relationship between independent and dependent variables, which can take the advantage of reducing dependence on the entire spectra to obtain prediction precision and stability by processing every frequency component for information extraction. Here, each frequency component can generate a sub-model, and the prediction model can be built based on the combination of the sub-models. WPNOSC-NPLS algorithm consists of two parts: the calibration procedure and the prediction procedure. The calibration procedure can be described as:

Step1. Unfold the spectra of calibration set ($\underline{X}$) according to (1), and perform discrete wavelet packet transform (DWPT) algorithm [6] by L levels, getting 2^L frequency components. Then fold these components to get a series of matrix $\{\underline{X}_n\}$($m\times n\cdot l$) with the form of $\underline{X}$ using the inverse of (1).

Step2. Perform NPLS algorithm on each matrix in $\{\underline{X}_n\}$, getting the optimum $\left\{f_n\right\}$ and saving all $\left\{H_{f,n}\right\}$, $\left\{b_{f,n}\right\}$ and $\left\{q_{f,n}\right\}$.

Step3. Calculate the error of each sub-model in DWPT domain as

$$e_i=Y-\sum_{j=1}^{f_i}H_{j,i}\cdot b_{j,i}\cdot(q_{j,i})^T \tag{7}$$

The weighted value of each sub-model for further improving precision is defined as

$$d_i=(e_i^Te_i)^{-1}/\sum_{j=0}^{2^L-1}(e_j^Te_j)^{-1} \tag{8}$$

To predict unknown samples with spectra $\underline{X}^{un}$, the prediction procedure can be described as:

Step4. Perform unfolding, DWPT decomposition and folding operations on $\underline{X}^{un}$, obtaining the corresponding matrix set $\left\{\underline{X}_n^{un}\right\}$.

Step5. Using the vectors $q_{f,i}$, $H_{f,i,}$, $b_{f,i}$ and d_i computed in Step2 and Step3, the prediction values of unknown samples can be obtained through the weighted model as

$$Y^{un} = \sum_{i=0}^{2^L-1} d_i \sum_{j=1}^{f_i} H_{j,i} \cdot b_{j,i} \cdot (q_{j,i})^T \tag{9}$$

Experiments

Homogenized milk samples were supplied by Tianjin University of Science & Technology. Concentrations of fat were determined by Rose-Gottlied method. A total of 120 samples were split into two sets, one used as calibration set including 80 samples and the other with 40 samples used as validation set. NIR transmission spectra (1100nm to 2300nm) were collected by a Spectrum GX FT-IR Spectrometer (Perkin-Elmer, USA), operating at 2nm resolution. With the help of temperature controller, the spectra at 25℃, 30℃, 35℃ and 40℃ were collected. Therefore, the dimensions of spectra matrix is 120×600×4. All algorithms were performed in Matlab V2010b (MathWorks). DWPT algorithm was achieved using the wavelet toolbox for Matlab with the ‘db4’ mother wavelet for 9 levels. The program of NPLS was developed by the N-way toolbox for Matlab. Performances of the developed models were evaluated by the squared correction coefficient (R^2), the root mean square error of calibration (RMSEC) and the root mean error of prediction (RMSEP).

Results and Discussions

Effect of Single Dimension on the Calibration models

According to most of liquid samples, the optical parameters vary significantly with the variation of temperature. Consequently, the collected spectra of the same sample at different temperatures are also different from each other. In other words, the spectra sets at different temperatures contain the specified information about the tempperature, and the useful information related to analyte also may be located at differnet wave lengths. If the prediction model of one temperature is applied to predict the sample with spectra of another temperature, the prediction precision has to be deteriorated. To the contrary, if all the information at differnet temperatures can be appropriately inputed to the prediction model, the prediction result should have more precision and stability. To study the effect of temperature on prediction models, the prediction results of fat concentration using the spectra of different temperatures based on the PLS regression are presented in Table Ⅰ.

Table I. Prediction results using different PLS models

Prediction Models	LVs	RMSEP of Fat Content (%)
PLS-using 25℃ spectra	8	0.149
PLS-using 30℃ spectra	8	0.144
PLS-using 35℃ spectra	7	0.159
PLS-using 40℃ spectra	7	0.119
NPLS	9	0.127

From Table I, it can be seen that the calibration models of different temperatures had different precision. The PLS model at 40℃ can achieve the best results. This is probably because that the fat becomes liquid when the temperature of milk is more than 37℃, leading to the weaker scattering effect during measurement.

The Prediction Result of WP-WNPLS

As for WP-WNPLS algorithm in this paper, the 3D spectra matrix including information of wavelength and temperature were used as the input data. After frequency components computed by the DWPT algorithm, the prediction model for fat concentration can be constructed by the weighted NPLS algorithm. To further investigate the prediction ability of the models, the combination of NPLS, Unfold-PLS and weighted strategy were tested, and the best RMSEP for validation set are illustrated in Table II.

Table I. Prediction results using different PLS models

Prediction Models	LVs	RMSEP of Fat Content (%)
Unfold-PLS	9	0.162
Weighted-Unfold-PLS	8	0.143
Weighted-NPLS	7	0.118
WP-WNPLS	6-9	0.104

In N-way models, the 3D matrix is not simply viewed as the set of 2D matrix because the relationships of the 2D matrix are introduced during modeling. Thus, compared with Table I, it was clear that the high dimensional model can improve the prediction ability of PLS models. Also, the results of NPLS are better than these of unfold-PLS. Furthermore, it can be seen that the weighted strategy in wavelet domain can effectively make use of the multi-scale property of spectra, leading to precision improvement. Therefore, the WP-WNPLS-based model can achieve the best results.

Conclusions

In this paper, a new algorithm named as WP-WNPLS is proposed to develop the multivariate regression model using 3-D NIR spectra. In this algorithm, the weight strategy was applied to improve the performance of conventional NPLS algorithm. Through the decomposition by DWPT, a satisfied weighted-based model was developed for predicting the analyte in milk sample. Experimental results show that the WP-WNPLS algorithm can significantly improve the prediction ability of the NPLS-based model, indicating that we must pay more attention to the multi-scale property of NIR spectra during model construction.

Acknowledgement

In this paper, the research was sponsored by the Science and Technology Research Project of Department of Education of Henan Province (No. 14A550016).

References

[1] M. Blanco, M. Alcalá, J.M. González, E. Torras. Near Infrared Spectroscopy in the Study of Polymorphic Transformations [J]. Analytica Chimica Acta, 2006 567 (2) 262–268.

[2] Ulf G. Indahl, Narinder S. Sahni, Bente Kirkhus, Tormod Næs. Multivariate Strategies for Classification Based on NIR-Spectra-With Application to Mayonnaise [J]. Chemometrics and Intelligent Laboratory Systems, 1999 49 (1) 19–31.

[3] L.G. Thygesen, S.O. Lundqvist. NIR Measurement of Moisture Content in Wood under Unstable Temperature Conditions. Part 1. Thermal Effects in Near Infrared Spectra of Wood [J]. Journal of Near Infrared Spectroscopy, 2000 8 (3) 183–189.

[4] Venyaminov SYu, Prendergast FG. Water (H_2O and D_2O) Molar Absorptivity in the 1000-4000 cm^{-1} Range and Quantitative Infrared Spectroscopy of Aqueous Solutions [J]. Analytical Biochemistry, 1997 248 (2) 234–245.

[5] Kunwar P. Singh, Amrita Malik, Nikita Basant, Puneet Saxena. Multi-way Partial Least Squares Modeling of Water Quality Data [J]. Analytica Chimica Acta, 2007 584 (2) 385–396.

[6] Young-Ho Seo, Hyun-Jun Choi, Dong-Wook Kim. Digital Hologram Encryption Using Discrete Wavelet Packet Transform [J]. Optics Communications, 2009 282 (3) 367–377.