ATLANTIS PRESS

Advances in Engineering Research, volume 152

6th International Conference on Machinery, Materials and Computing Technology (ICMMCT 2018)

Using Consensus Strategy and Interval Partial Least Square Algorithm in Wavelet Domain for Analysis of Near-infrared Spectroscopy

Dan Peng, He Guo, Linqing Li, Yanlan Bi, Guolong Yang

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

pengdanhaut@126.com

Keywords: consensus strategy; interval partial least square algorithm; near-infrared spectroscopy; wavelet domain; selection; integration

Abstract: To improve the stability and precision performance of partial least square regression (PLS) model in near-infrared analysis application, the consensus strategy was applied in the wavelet domain. Taking the advantage of multiscale property of wavelet packet analysis, a new modelling method was developed based on the idea of the interval PLS algorithm and named as WpCo-iPLS algorithm. In WpCo-iPLS model, wavelet packet transform (WPT) algorithm was firstly adopted to split the raw spectra into a series of frequency components in wavelet domain. Then, coupled with the consensus strategy, multiple members of PLS models were established on the interval frequency components. To reduce the dependence on single model, an optimization of the weight parameters of member models was conducted. At last, a consensus model was achieved by effectively combining all the member models. To validate the WpCo-iPLS algorithm, it was applied to measure the six kinds of contents concentration of diesel samples using NIR spectra. The experimental results showed that the prediction ability and robustness of WpCo-iPLS model was stronger than that of conventional consensus algorithms, indicating that it is a promising consensus strategy for modelling using NIR spectra.

1. Introduction

Near infrared (NIR) spectroscopy is regarded as an alternative to traditional chemistry methods for evaluating the quality of food, agriculture, and petrochemical products [1-2]. Now, multivariate calibration methods are widely used in the analysis of NIR spectra, and the most frequently used ones include multiple linear regression (MLR) [3], partial least squares (PLS) [4] and so on. Thus, the successful use of NIR technique is mainly dependent on multivariate calibration. However, NIR spectra contains not only useful information but also interference information, such as noise and background, and also collinearity between wavelength variables that exists commonly. This interference information within the spectra often complicates the model, leading to inaccurate predictions in some cases. Recently, consensus modelling method introduces a new way to improve the model performance. It constructs multiple member models and then combines them to form a consensus model, which is different from traditional modelling approaches. The typical consensus models include the interval PLS algorithm (iPLS) [5], the staked interval PLS algorithm (SPLS) [6], the consensus interval PLS algorithm (CPLS) [7] and so on. These modelling methods improve the stability and precision performance of regression model to some extent in wavelength domain. However, the analyte information is more concentrated in the frequency domain and more dispersed in the wavelength domain. Therefore, it is worthwhile to explore the establishment of models in the frequency domain to further improve accuracy.

In this work, a new consensus regression algorithm (WpCo-iPLS) is proposed. In this algorithm, the wavelet packet transform (WPT) [8] decomposition, the idea of iPLS algorithm and the consensus strategy are coupled to form the prediction result. To validate the WpCo-iPLS algorithm, a real NIR spectral dataset of diesel was also analysed.

2. Principle and Method

The development of WpCo-iPLS algorithm consists of two parts including WPT decomposition and member models combination. The content of these two parts will be described in detail as below.

2.1 The Decomposition of Wavelet Packet Transform

The wavelet packet system is a generalization of wavelet transform, in which at all stages both the low-pass and high-pass bands are split. Therefore, it can be used to obtain decomposition in finer and more flexible way. WPT consists of a set of linearly combined usual wavelet functions. A wavelet packet $\psi_{j,k}^{i}(t)$ is a function with three indices where integers *i*, *j* and *k* are the modulation, scale and translation, respectively

$$\psi_{j,k}^{i}(t)=2^{j/2}\psi^{j}(2^{j}t-k) \tag{1}$$

The wavelet functions ψ^{j} can be obtained using the following recursive function as

$$\psi^{2j}(t)=\sqrt{2}\sum_{k=-\infty}^{\infty}h(k)\psi^{i}(2t-k) \tag{2}$$

$$\psi^{2j+1}(t)=\sqrt{2}\sum_{k=-\infty}^{\infty}g(k)\psi^{i}(2t-k) \tag{3}$$

where the discrete filters *h(k)* and *g(k)* are the quadrature mirror filters associated with the scaling function and the mother wavelet function. Unlike WT, the WPT contains a complete decomposition at each level. The recursive function between the *j*th and the (*j*+1)th level components of signal s(t) are

$$s_{j}^{i}(t)=s_{j+1}^{2i-1}(t)+s_{j+1}^{2i}(t) \tag{4}$$

$$s_{j+1}^{2i\text{-}1}(t)=H\cdot s_{j}^{i}(t),\ \ s_{j+1}^{2i}(t)=G\cdot s_{j}^{i}(t) \tag{5}$$

where *H* and *G* are the filtering-decimation operators related to the discrete filters *h(k)* and *g(k)*. After *P* levels decomposition, the original spectral signal *s(t)* can be expressed as

$$s(t)=\sum_{i=0}^{2^{P}-1}s_{i}^{P}(t) \tag{6}$$

It should be noted that these (2^{P}) components in the *P*th level, where the analyte signal resides in, don't overlap with each other. This is the reason that it is possible to analyse the raw signal.

2.2 Interval partial least squares algorithm

iPLS is a band selection method proposed by Norgaard [5]. The main idea of this approach is to split the full spectrum into n disjoint intervals with equal width. For each interval, a local PLS model (member model) is constructed. Because each spectral interval may contain different analyte information, as well as noise, the prediction precision of the member model is also different from each other. In iPLS model, the member model with the best prediction precision is selected as the output model to provide the prediction result.

Conventional iPLS can extract the spectral wavelengths, which is highly relevant to the analyte property, to improve the stability of the regression model and increase the interpretability of the relationship between the interval spectrum and analyte property.

2.3 WpCo-iPLS algorithm

Similar to iPLS, WpCo-iPLS algorithm also splits the full spectrum matrix into n disjoint

intervals with equal width. However, this split is applied in frequency domain, not the wavelength domain. Taking the advantage of WPT transform, the member PLS model is established on wpt component in this algorithm.

In WpCo-iPLS, the aim of consensus strategy is that multiple member models will effectively identify and encode more aspects of the relationship between independent and dependent variables than a single model, which can take the advantage of reducing dependence on single sample to obtain prediction precision and stability by combining all the member models.

Let X be an $m \times n$ spectral matrix with n wavelengths in m samples, and Y be an $m \times u$ analyte matrix with u analyte properties in m samples. The combinational model can be described as

$$y_{WpCo} = \sum_{i=1}^{L} w_k \hat{y}_i \tag{7}$$

where y_{WpCo} is the prediction result of WpCo-iPLS model, $\hat{y}_i$ is the prediction result from the PLS member model developed on the ith interval WPT components, w_k is the weight of the PLS member model developed on the ith interval WPT components and L is the number of member models. Here, the interval WPT components is the summation of a series of WPT components. As for the member models, the combination problem is to find the appropriate w_k to provide the smaller prediction error. This problem can be illustrated below:

$$w_k = ARG \min(y - y_{WpCo})^2 \tag{8}$$

where y is the true value of the analyte property. Thus, it is clear that the aim of WpCo-iPLS algorithm is to minimize the prediction error by a trade-off between each member model. To provide the appropriate performance, the value of w_k is computed like the CPLS algorithm [7] as

$$w_k = \frac{1/\sigma_k^2}{\sum_{i=1}^{L} 1/\sigma_i^2} \approx \frac{1/RMSECV_k^2}{\sum_{i=1}^{L} 1/RMSECV_i^2} \tag{9}$$

where σ_k is the variance of the random error of PLS member model on the kth interval WPT components, and $RMSECV_k$ is the root mean square error of cross validation of PLS member model on the kth interval WPT components. Here, the $RMSECV_k$ can be calculated as

$$RMSECV_k = \sqrt{\sum_{i=1}^{m} (y_i - w_k \hat{y}_i)^2 / m} \tag{10}$$

where y_i is the true value of analyte property.

The detail of WpCo-iPLS algorithm can be summarized as following:

Step1. Perform a WPT decomposition on matrix X by P level to get a series of frequency components as $\{ X_0^P, X_1^P, , \ldots, X_{2^P-1}^P \}$.

Step2. Construct member PLS models with *lv* latent variables for all the interval WPT components and compute their prediction performance. At the same time, save all the vectors of PLS models for unknown sample prediction. The ith interval WPT components ($X_i^{P,interval}$) can obtained as

$$X_i^{P,interval} = \begin{cases} \sum_{j=(i-1)\times \mathrm{int}(2^P/L)}^{i\times \mathrm{int}(2^P/L)-1} X_j^P, & 1 \le i < L \\ \sum_{j=(L-1)\times \mathrm{int}(2^P/L)}^{2^P-1} X_j^P, & i = L \end{cases} \tag{11}$$

Step3. Calculate weight parameters w_k for each member model.

Step4. With the parameters in *Step3*, develop the consensus interval partial least squares model. The performance of WpCo-iPLS model can be evaluated as

$$RMSECV_k = \sqrt{\sum_{i=1}^{m} (y_i - \sum_{j=0}^{2^P-1} w_k \hat{y}_{i,j})^2 / m} \tag{12}$$

where $\hat{y}_{i,j}$ is the corresponding prediction results of the *j*th member model.

For an unknown sample, the WpCo-iPLS model can be applied as following:

Step1. Perform a WPT decomposition on matrix $\tilde{X}$ by *P* level to get a series of frequency components as $\{\tilde{X}_0^P, \tilde{X}_1^P, , \ldots, \tilde{X}_{2^P-1}^P\}$.

Step2. Predict the unknown sample using the consensus model. Here, the weighted average of the prediction results is achieved as the final prediction result.

$$\tilde{y} = \sum_{i=0}^{2^P-1} w_i \sum_{j=1}^{lv} b_{j,i} t_{j,i} q_{j,i}^T \tag{13}$$

where subscript '*j*, *i*' means the corresponding PLS vector of *i*th PC in the *j*th member model, and the *i*th PC, t_i, p_i, u_i and q_i are the corresponding column vectors of *T*, *P*, *U*, and *Q*, respectively.

3. Experiments

NIR spectra data were downloaded from http://www.eigenvector.com/data/SWRI/index.html. This data set consists of 784 samples of diesel on Southwest Research Institute. The wavelength range is 750-1550nm at 2 nm intervals. In Figure 1, the spectra scanned by m5spec were depicted. The objective is to develop the consensus regression model for parameters boiling point at 50% recovery (bp50), cetane Number (CN), density (d4052), freezing temperature of the fuel, total aromatics and viscosity measurement. Because some properties have missing values, the number of samples for above 6 properties are 395, 381, 395, 395, 395 and 395. Here, these diesel samples were split into a calibration set including 70% samples and a prediction set including 30% samples.

All computation was developed in Matlab v2017a using the PLS Toolbox v8.1. A leave-one-out cross-validation procedure was applied to the calibration set.

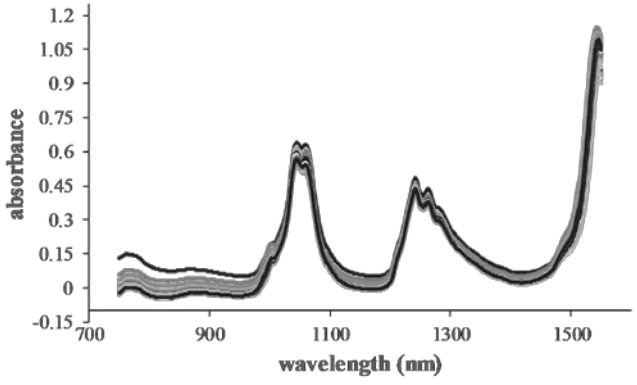

Figure 1 This caption has one line so it is centred

4. Results and Discussions

Interval number is a very important parameter for consensus model. For a specific decomposition level P and interval number L, total 2^P WPT components generate and L member models need to be established based on interval WPT components defined in Equation (11). Conventional iPLS chooses the best member model, and SPLS, CPLS and WpCo-iPLS combine all the member models to develop a new model using weighted strategy. In WpCo-iPLS model, different decomposition level leads to different division methods of the full spectrum, which leads to different interval WPT components. If an interval WPT components contains more analyte information and less noise, the prediction ability of the corresponding member PLS model should be better.

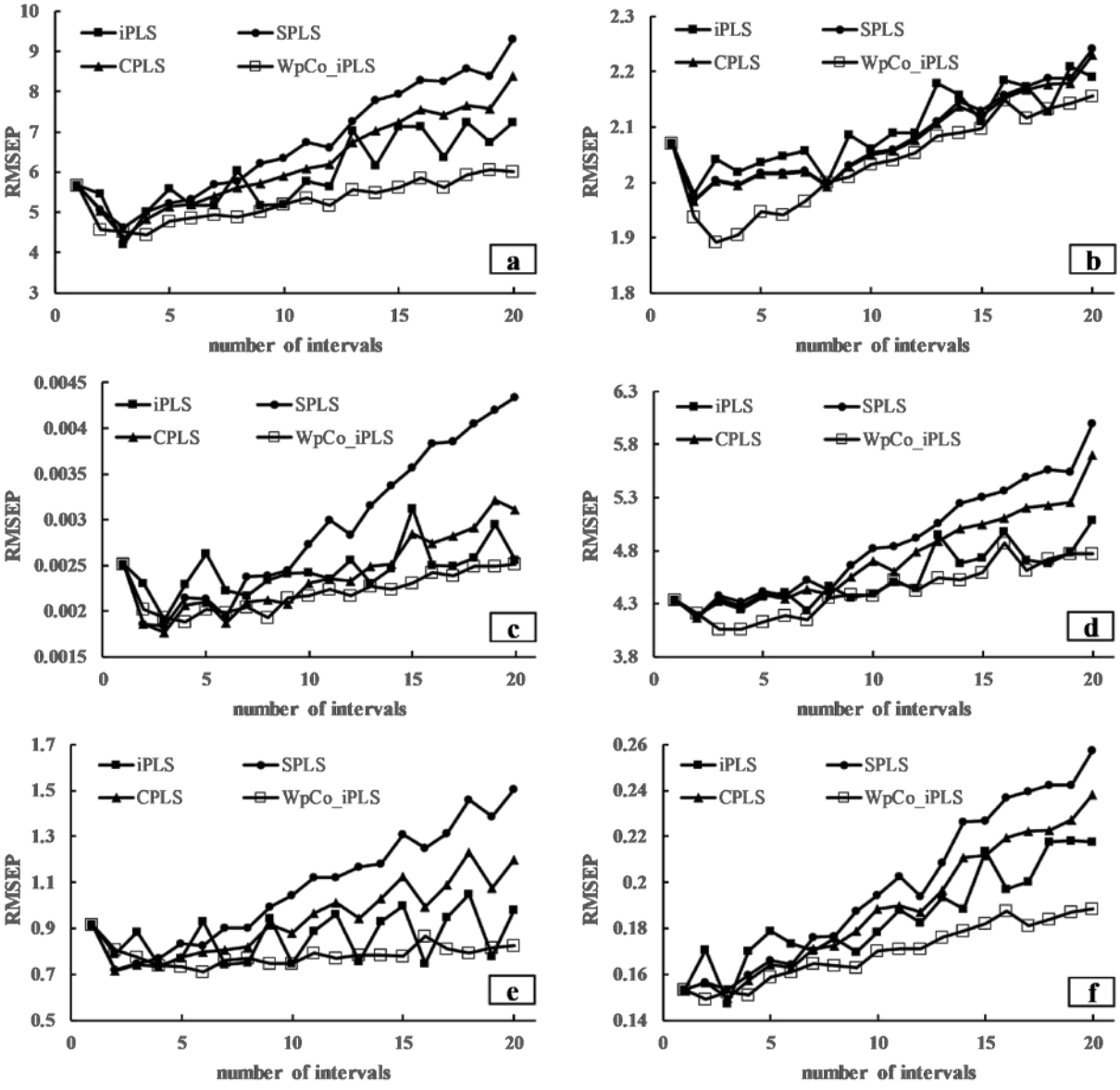

Figure 2 Result of iPLS, SPLS, CPLS and WpCo-iPLS based on dataset: (a) bp50, (b) cetane Number, (c) density-d4052, (d) freezing temperature of the fuel, (e) total aromatics and (f) viscosity measurement

Generally, the number of intervals affects the performance of consensus model in two ways. On one hand, if the interval number is smaller, each interval WPT components contains broader band spectrum. Broader band spectrum often contains more analyte information, but it also brings lots of interference information to deteriorate the performance of model. On the other hand, if the interval number is bigger, each interval WPT components contains narrow band spectrum. Some WPT components may not contain any analyte information, so the corresponding member model may

greatly reduce the accuracy of the model. Therefore, the decomposition level is set to 9 (meaning that 512 WPT components) and the maximum number of intervals is set to 20, here. As for each member PLS model, the maximum latent variable (*LV*) number is set to 10, since too large *LV* number will result into local PLS over fitting.

To test the effectiveness of algorithms, grid search technique is used to find the optimal interval number and *LV* number. The range of interval number and *LV* number is set first. Then, for a specific interval number *L*, the optimal *LV* number is the one which gives the best RMSEP. After grid search, the performances of these four algorithms changes with the interval number, as shown in Figure 2(a)-(f). Generally, the bigger the interval number is, the more difficult it is to combine the member models. It can be seen from Figure 2 that the RMSEP change bigger with the increasing of interval number, and the RMSEP of models with many interval number is even larger than that of conventional PLS model. This is due to that bigger interval number leads to less analyte information and poor precision of member models. Then, the performance of the corresponding consensus model has to decreases. In Figure 2, it is also clear that for most interval numbers, the RMSEP values of WpCo-iPLS model are smaller than those of iPLS, SPLS and CPLS. This results can be explained by the fact that the analyte information and noise is more concentrated in the frequency domain than in the wavelength domain. Consequently, based on the multiscale property of WPT analysis, the performance of WpCo-iPLS model will be greatly improved by a small number of well-performing member models, which provide the greater weight value.

5. Conclusions

This paper proposed a new consensus algorithm for regression model on the WPT domain. A new WpCo-iPLS is developed by combining the WPT analysis, the consensus algorithm and the idea of weighted iPLS. Compared to the conventional consensus algorithms built on wavelength domain, the WpCo-iPLS can take full advantage of multiscale property of WPT decomposition in frequency domain and also can make a good compromise between precision performance and difficulty. This algorithm was successfully applied to analyze the spectra of diesel samples for contents measurement. Experimental results indicate that this algorithm can effectively improve the prediction ability of the NIR regression models and the prediction ability is stronger especially in the case in which the member models are complicated.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (Grant No. 31601537).

References

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

[2] M. Mahmoudi, S. Sant, B. Wang, S. Laurent, T. Sen. (2011) Superparamagnetic iron oxide nanoparticles (SPIONs): development, surface modification and applications in chemotherapy. Advanced Drug Delivery Reviews, 63, 24-46.

[3] T. Lemos, J. H. Kalivas. (2017) Leveraging multiple linear regression for wavelength selection. Chemometrics and Intelligent Laboratory Systems, 168, 121-127.

[4] P. Geladi, B.R. Kowalski. (1986) Partial least-squares regression: a tutorial. Analytica Chimica Acta, 185, 1-17.

[5] A. Borin, R.J. Poppi. (2005) Application of mid infrared spectroscopy and iPLS for the quantification of contaminants in lubricating oil. Vibrational Spectroscopy, 37, 27-32.

[6] W. Ni, S.D. Brown, R. Man. (2009) Stacked partial least squares regression analysis for spectral calibration and prediction. Journal of Chemometrics, 23, 505-517.

[7] G.L. Ji, G.Z. Huang, Z.J. Yang, X.H. Wu, X.J. Chen, M.S. Yuan. (2015) Using consensus interval partial least square in near infrared spectra analysis. Chemometrics and Intelligent Laboratory Systems, 144, 56-62.

[8] B. Jawerth, W. Swedens. (1994) An overview of wavelet based multiresolution analyses. SIAM Review, 39, 377-412.