Chemometrics and Intelligent Laboratory Systems 156 (2016) 217–223

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Optimal partner wavelength combination method with application to near-infrared spectroscopic analysis☆

Tao Pan [a,*], Yun Han [a], Jiemei Chen [b,*], Lijun Yao [a,c], Jun Xie [a]

[a] Department of Optoelectronic Engineering, Jinan University, Huangpu Road West 601, Tianhe District, Guangzhou 510632, China
[b] Department of Biological Engineering, Jinan University, Huangpu Road West 601, Tianhe District, Guangzhou 510632, China
[c] State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China

ARTICLE INFO

Article history:
Received 11 October 2015
Received in revised form 11 March 2016
Accepted 31 May 2016
Available online 1 June 2016

Keywords:
Near-infrared spectroscopic analysis
Optimal partner wavelength combination
Partial least squares
Soil
Organic matter

ABSTRACT

Using binary linear regression, the optimal partner wavelength of each wavelength is selected in an initial wavelength screening region. On the basis of strategy above, a novel approach for selecting appropriate wavelengths combination, named optimal partner wavelength combination (OPWC) coupled with partial least squares (PLS), is proposed, and was successfully applied for reagent-free near-infrared spectroscopic analysis of organic matter in soil. Moving window PLS (MW-PLS), successive projections algorithm (SPA) and Monte Carlo uninformative variable elimination (MC-UVE), which are well-performed wavelength selection methods, were also conducted for comparison.

The OPWC-PLS, MW-PLS, SPA-PLS and MC-UVE-PLS methods selected 14, 210, 63, 199 wavelengths, respectively. The root-mean-square error and correlation coefficients for leave-one-out cross validation were 0.165 g kg^{-1} and 0.967 for OPWC-PLS, 0.163 g kg^{-1} and 0.968 for MW-PLS, 0.198 g kg^{-1} and 0.953 for SPA-PLS, and 0.190 g kg^{-1} and 0.956 for MC-UVE-PLS, respectively. The results indicate that OPWC-PLS and MW-PLS methods were almost the same, and were obvious better than SPA-PLS and MC-UVE-PLS methods. But the OPWC only contained 14 wavelengths, which is a high efficient approach for extracting information wavelengths and mitigating redundant wavelengths. OPWC can be also provided valuable reference for designing small dedicated spectrometers with a high signal-to-noise ratio.

OPWC can be programmed determined, which has small amount of calculation and high operating speed, and it is a deterministic search technique whose results are reproducible. We believe that OPWC has such applicability and can be applied to other fields of spectroscopic analysis.

1. Introduction

Near-infrared (NIR) spectroscopy can rapid determine samples without chemical reagents, which has been proven to be a powerful analytical tool for use in agriculture [1–5], food [6–8], environment [9], biomedicine [10–15], etc. Wavelength selection of NIR spectroscopic analysis is very important and essential for improving model prediction effect, reducing model complexity and designing small dedicated spectrometers with a high signal-to-noise ratio, especially for the complex analyte.

So far, many methods of wavelength selection have been applied in NIR spectroscopic analysis. These methods can be categorized into two classes: continuous mode and discrete mode. Continuous mode means selecting the adjacent wavelengths in a continuous waveband with good chemical interpretation. However, spectral co-linearity may appear in adjacent wavelengths that results in evaluation distortion. Moving window partial least squares (MW-PLS) has been proven an effective method to overcome co-linearity problems based on a continuous mode [2,3,7,8,9,14]. But MW-PLS is a traversal algorithm for continuous wavebands, which is time-consuming when it meets a large data set. Discrete mode usually selects the non-adjacent wavelengths, which is designed to minimize co-linearity problems. There are many effective discrete wavelength selection methods with high operating speed, such as the successive projections algorithm (SPA) [16–18], Monte Carlo uninformative variable elimination by PLS (MC-UVE-PLS) [19], competitive adaptive reweighted sampling (CARS) [20], stability competitive adaptive reweighted sampling (SCARS) [21], randomization test (RT) [22], latent projection graph (LPG) [23], and influential variables (IVs) [24] methods.

In this study, a novel approach for selecting a combination of appropriate wavelengths, named optimal partner wavelength combination (OPWC) coupled with partial least squares (PLS), is proposed. Based on binary linear regression (BLR), the optimal partner wavelength of each wavelength is selected in initial wavelength screening region. A wavelength subset, called partner wavelength subset (PWS), is

☆ Selected paper from 15th Chemometrics in Analytical Chemistry Conference, 22-26 June 2015, Changsha, China

* Corresponding authors.
E-mail addresses: tpan@jnu.edu.cn (T. Pan), tchjm@jnu.edu.cn (J. Chen).

http://dx.doi.org/10.1016/j.chemolab.2016.05.022
0169-7439/

determined. The same procedure is performed repeatedly until PWS stop shrinking after limited times, i.e. the last obtained wavelength subset is just own PWS, which is called OPWC. The leave-one-out cross validation (LOOCV) based on PLS model is performed to evaluate the prediction capability of OPWC.

Given that soil is complex system with multiple components, the NIR spectroscopic analysis of major components in soil has to mitigate noise disturbance through the selection of appropriate wavelengths [1–5]. In this study, NIR spectroscopic analysis of organic matter (OM) in soil was taken as example to evaluate the performance of the proposed OPWC-PLS. OPWC was performed for appropriate wavelength selection. MW-PLS, SPA-PLS and MC-UVE-PLS, which are well-performed wavelength selection methods, were also conducted for comparison.

2. Materials and methods

2.1. Experimental materials, instruments, and measurement methods

A total of 114 soil samples of the same type were collected and then ground after drying. The samples were sifted by using a 0.25 mm soil sifter. The OM content of each sample was measured by using the standard potassium dichromate ($K_2Cr_2O_7$) oxidation soil analysis method [25]. The obtained values were used as the reference values for spectroscopic analysis. The OM values for all samples ranged from 1.24 g kg^{-1} to 4.70 g kg^{-1}, and the mean value and standard deviation were 2.685 and 0.653 g kg^{-1}, respectively.

The spectroscopy instrument was an XDS Rapid Content™ Grating Spectrometer (FOSS, Denmark) equipped with a diffuse reflection accessory and a round sample cell. The scanning scope of the spectrum spanned 780–2498 nm with a 2 nm wavelength gap; wavebands of 780–1100 nm as well as 1100–2498 nm were adopted for Si and PbS detection, respectively. Every sample was measured thrice, and the mean value of the three measurements was used for modeling. The spectra were measured at 25 ± 1 °C and 46 ± 1%RH relative humidity.

2.2. Leave-one-out cross validation based on PLS model

All n samples ($n = 114$) were performed leave-one-out cross validation (LOOCV) based on PLS model; and the model parameters, such as waveband combination, number of PLS factor (F), etc., were optimized according to prediction effect. The specific procedure was the follows.

First, every one sample was left out from all n samples, and the calibration PLS model was constructed via the remaining $n-1$ samples to calculate prediction value of the left out sample. Based on the same process, the prediction values of all n samples were calculated. The actual and predicted values for the ith sample were denoted as C_i, $\tilde{C}_i$, respectively, $i = 1, 2, \cdots, n$. The mean actual and mean predicted values of all samples were denoted as C_{Ave}, $\tilde{C}_{\mathrm{Ave}}$, respectively. The prediction accuracy was evaluated by the root-mean-square (RMS) error and the correlation coefficients for LOOCV, which were denoted as SECV and R_{CV}, respectively. The calculation formulas were as the follows:

$$\mathrm{SECV} = \sqrt{\frac{\sum_{i=1}^{n}\left(\tilde{C}_i - C_i\right)^2}{n}}, \tag{1}$$

$$R_{CV} = \frac{\sum_{i=1}^{n}(C_i - C_{\mathrm{Ave}})\left(\tilde{C}_i - \tilde{C}_{\mathrm{Ave}}\right)}{\sqrt{\sum_{i=1}^{n}(C_i - C_{\mathrm{Ave}})^2\left(\tilde{C}_i - \tilde{C}_{\mathrm{Ave}}\right)^2}}, \tag{2}$$

where, the smaller SECV value shows the prediction precision is higher, the bigger R_{CV} value shows the prediction correlation is higher, they synthetically reflect the prediction effect of LOOCV-PLS model in a wavelength combination. The model parameters were selected to achieve minimum SECV.

2.3. Proposed OPWC-PLS method

By finding the optimal partner wavelength of each wavelength in a wavelength screening region based on BLR, a wavelength subset, called partner wavelength subset (PWS), is determined. According to obtained correspondence, the optimal partner wavelengths of all wavelengths in the PWS are also combined, i.e. a new PWS is obtained. The same procedure is performed repeatedly until PWS stop shrinking after limited times. Namely, the last obtained wavelength subset is just own PWS, which is called OPWC. The specific steps are as follows:

2.3.1. Step 1

A wavelength region (Δ) is set as the wavelength screening region, which could be set as entire or partial region according to the physical and chemical characteristics of the measured object and the instrument properties.

2.3.2. Step 2

Set $\Delta = \{\lambda_1, \lambda_2, \cdots, \lambda_N\}$, N is the number of wavelengths of Δ. For each λ_i of Δ and for other wavelength λ_k in Δ, the LOOCV-BLR analysis based on the wavelength combination (λ_i, λ_k) is performed. According to minimum SECV(λ_i, λ_k), the optimal partner wavelength of λ_i is determined and denoted as $f(\lambda_i)$, and goes as follows

$$\mathrm{SECV}(\lambda_i, f(\lambda_i)) = \min_{\substack{k=1,2,\cdots,N \\ k\neq i}} \mathrm{SECV}(\lambda_i, \lambda_k) \tag{3}$$

In fact, f is a projection from Δ to Δ. The $f(\Delta)$ is called partner wavelength subset ($\mathrm{PWS}^{(1)}$) of Δ, and the number of wavelengths of $\mathrm{PWS}^{(1)}$ is denoted as $N^{(1)}$. Theoretically, each wavelength λ_i corresponds unique optimal partner wavelength$f(\lambda_i)$, but different wavelengths could correspond the same optimal partner wavelength. Therefore, it is possible that some wavelength is not the optimal partner wavelength of any wavelength. If some λ is not the optimal partner wavelength of any wavelength, then λ is not belong to $\mathrm{PWS}^{(1)}$, and $N^{(1)} < N$.

2.3.3. Step 3

Based on the projection f defined in Step 2, the partner wavelength subset ($\mathrm{PWS}^{(2)}$) of $\mathrm{PWS}^{(1)}$ is also determined. The same procedure is performed repeatedly. In fact, it can be interesting proved that PWS stop shrinking after limited times of the same procedure (Considering the limitation of article length, its rigorous mathematical proof were omitted). Assuming that, after s-times projections, the PWS stop shrinking, $N^{(s)} = N^{(s-1)}$. And

$$\mathrm{PWS}^{(s)} = \{^{(s)}(\Delta) = \{\overbrace{\{\cdots\{}^{s}(\Delta) \tag{4}$$

is called optimal partner wavelength combination (OPWC). In OPWC, each wavelength is optimal partner wavelength of some other wavelength. Theoretically, OPWC can form several loops according to the attribution direction for partner wavelengths, such as,

$\lambda_1^{(s)} \rightarrow \lambda_2^{(s)} \rightarrow \lambda_1^{(s)}$,

$\lambda_1^{(s)} \rightarrow \lambda_2^{(s)} \rightarrow \lambda_3^{(s)} \rightarrow \lambda_1^{(s)}$,

$\lambda_1^{(s)} \rightarrow \lambda_2^{(s)} \rightarrow \lambda_3^{(s)} \rightarrow \lambda_4^{(s)} \rightarrow \lambda_1^{(s)}$,

……

where "→" denotes attribution direction for partner wavelengths, e.g. $\lambda_1^{(s)} \rightarrow \lambda_2^{(s)}$ denotes $\lambda_2^{(s)}$ is the optimal partner wavelength of $\lambda_1^{(s)}$.