Chemometrics and Intelligent Laboratory Systems 152 (2016) 118–124

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

An improved changeable size moving window partial least square applied for molecular spectroscopy

Yong Zhao [a], Shenghao Wang [a,b,*], Zhi Li [b], Zhenying Pei [b], Fuyi Cao [b]

[a] College of Information Science and Engineering, Northeastern University, Shenyang 110819, China
[b] Centre of Simulation, Shenyang Institute of Engineering, Shenyang 110136, China

ARTICLE INFO

Article history:
Received 25 September 2015
Received in revised form 12 January 2016
Accepted 1 February 2016
Available online 7 February 2016

Keywords:
Spectra analysis
Genetic algorithm
Partial least square
Pretreatment method
Wavelength variable
Optimization selection

ABSTRACT

When analyzing molecular spectra, optimizing the pretreatment method and the wavelength variable is always an important issue. However, currently there are unsatisfied phenomena that select the same type of pretreatment method multiple times in some results generated by previous common optimizing algorithms. Additionally, the parameters and calculation priorities of the pretreatment methods cannot be optimized. To solve those problems, an improved changeable size moving window partial least square (CSMWPLS) named pretreatment method classification and adjustable parameter changeable size moving window partial least square (CA-CSMWPLS) is presented. With regard to the chromosome construction of CA-CSMWPLS, there is a region for pretreatment method optimization and another one for wavelength variable optimization. In the former, the common pretreatment methods are classified into four different types such as smoothing, derivation, correction, and standardization, and the parameters and calculation priorities of pretreatment methods serve as genes of the CA-CSMWPLS chromosome. In the latter, there are changeable size moving windows that consist of window position genes and window width genes. Moreover, a scale factor η is designed for reducing model complexity in CA-CSMWPLS fitness function and a peculiar coding and a decoding rule are adopted in this algorithm. After testing a group of corn and a group of gasoline spectra with CA-CSMWPLS, the model accuracy was significantly improved, for the root mean square error cross validation (RMSECV) and root mean square error prediction (RMSEP) of the corn spectra were 0.0028 and 0.0032, and those of gasoline were 0.165 and 0.170, respectively. Furthermore, the optimized pretreatment methods and wavelength variables are more reasonable, the model complexity is smaller, and the model robustness is stronger than other relative methods.

1. Introduction

When some materials are exposed under special external photons, the molecular spectra including ultraviolet, near infrared, infrared, and Raman spectroscopy are generated because of the atom energy transition level. This technology has been applied in many industries based on modern chemical analysis instruments whose output data usually contain multi-co-linearity. Principle component regression (PCR) and partial least square regression (PLSR) are generally accepted methods to solve this problem [1]. Although those methods can be directly used to analyze a full spectrum after projecting useful information into high score loading vectors to filter noise of low score ones, some documents show that the model performance could be enhanced by reasonably selecting wavelength variables [2–5]. Until now, there are two different selecting methods: discrete single wavelength selection and successive interval wavelength selection. Knowledge-based selection [6], correlation coefficient selection [7], successive projections algorithm [8], uninformative variable elimination [9], genetic algorithm (GA) [4–5], and simulated annealing [10] belong to the first one. Interval partial least square (iPLS) [11], backward interval partial least square (biPLS) [12], synergy interval partial least square (siPLS) [13], moving window partial least square (MWPLS) [14], CSMWPLS, searching combination moving window partial least square (SCMWPLS) [15], and interval random frog (IRF) [16] belong to the second one. The latter one and its derivatives are becoming more and more popular than the former because the continuity of the absorption bands of molecular spectra is considered capable of avoiding random system error and the influence of unconcerned ingredients.

During the process of modeling, spectra data pretreatment methods selection is another important mission, especially for some complex materials or some materials with weak information. There are about four classes of common pretreatment methods, such as smoothing, derivation, correction, and standardization [17]. Wavelength selection and pretreatment method selection, to some extent, influence each other. A specular chromosome encoding was proposed by Devos. In this algorithm, the first $p \times 5$ binary value bits denote p pretreatment methods, and the following q ones denote q consecutive wavelengths. Although this method was applied in corn, pork, and sugar beet with good results,

* Corresponding author at: Northeastern University, Shenyang 110819, China. Tel.: +86 24 31975293.
E-mail address: wangyiwzzy@163.com (S. Wang).

http://dx.doi.org/10.1016/j.chemolab.2016.02.001
0169-7439/© 2016 Elsevier B.V. All rights reserved.

to more reasonably explain the optimized chromosomes, a new technology needs to be explored that only selects one pretreatment method [18–19]. Another kind of GA combined with ant colony optimization, whose one bit denotes one category pretreatment method with sample set partitioning based on joint X–Y distances (SPXY) [20] and Kennard–Stone (KS) sampling, was presented by Allegrini and was successfully applied in corn near infrared (NIR) analysis [21].

The previous methods used for co-optimizing pretreatment methods and wavelength variables still have some disadvantages: (1) one kind of pretreatment method exists in the same individual too many times, which is difficult to explain with chemometrics, (2) the priority of pretreatment cannot be optimized, and (3) the parameters cannot be changed, which limits the export of better result.

To solve the above problem, improved CSMWPLS, named CA-CSMWPLS, results are presented in this paper. The performance of the spectra model has been enhanced by its special chromosome structure, coding, and decoding rules after testing two different groups of near infrared spectra data.

2. Theory

2.1. The algorithm flow

Except for genetic selection, duplication, crossover, and mutation, operations like traditional CSMWPLS derivations based on GA, CA-CSMWPLS's chromosome is divided into pretreatment methods, and wavelength variable regions that are constructed with non-negative decimal integers to reduce computation complexity, as well as to flexibly add and delete pretreatment method and its parameters. The algorithm flow is as follows:

Step 1: Main program initialization.

This step includes inputting the spectra data and the analyte ingredient; subgrouping all of the samples, based on SPXY or other methods, into modeling samples and testing samples with proportion 4:1 or 3:1; setting the window number, the range of window width, the elite number of GA E, the number of completed independent runtime $i = 0$, the maximum loop number of independent runtime L, the population of GA, the generation of GA, the number of completed GA operate time $g = 0$, and the maximum loop number of GA operate G.

Step 2: Population initialization.

For accelerating the convergence speed of CA-CSMWPLS and increasing the probability of selecting valuable information, 80% of CA-CSMWPLS individuals' window positions are randomly distributed in the useful regions, which are obtained by MWPLS, and the rest of the individuals are randomly initialized in the full spectra region. Here, the valuable region is determined by a threshold.

Step 3: Model evaluation.

First, pretreatment method and wavelength variable information corresponding to pretreatment method and wavelength variable region are obtained after decoding individuals. Second, based on leave-one-out cross validation with the above method and wavelength variable information, the performance of the best PLSR is calculated by the testing samples after calculating any individual's fitness. Here, the optimal latent variable number selected for each individual is the first local minimum RMSECV when constructing the PLS regression model in GA operation.

Step 4: Genetic operation.

The first E highest fitness individuals (the elite of GA) are directly pulled out into the next loop, and the rest of them are dealt with in a genetic operation. Let $g = g + 1$, if $g \le G$, then go back to Step 3, else go to Step 5.

Step 5: Output result.

Let $i = i + 1$, if $i \le L$, then go back to Step 2, else output L time results and the optimal one. The optimal result can be obtained with Eq. (1). Here, C is the criterion for selecting the optimal result (the smaller the better), Var is the number of selected variable number, RMSECV is acquired by modeling samples, RMSEP is acquired by testing samples, and Q is the selected latent variable number after optimizing with CA-CSMWPLS; Ω_1, Ω_2,Ω_3, and Ω_4 are scale factors of the above four variables.

$$C = \mathrm{Var} \times \Omega_1 + \mathrm{RMSEC}V \times \Omega_2 + \mathrm{RMSEP} \times \Omega_3 + Q \times \Omega_4 \tag{1}$$

2.2. Coding rule

In our work, smoothing denotes Savitzky–Golay smoothing (SGS) with two parameters, derivation denotes Savitzky–Golay derivation (SGD) with three parameters, correction is classified into standard normal vitiate (SNV) and multiplicative scatter correction (MSC), and standardization is classified into mean center (MC) and autoscale. Thus, the pretreatment method region is defined by the first 11 genes for CA-CSMWPLS, as shown in Fig. 1. Here, $O_x \in [0–4]$ denote the priorities of SGS, SGD, correction, and standardization; $K_1 \in [2–5]$ and $K_2 \in [2–5]$ denote the polynomial degree of the SGS and SGD; $F_1 \in [3–15]$ and $F_2 \in [3–15]$ denote the polynomial window width; $N \in [1–2]$ denotes the order of derivation; and $P_1 \in [1–2]$ and $P_2 \in [1–2]$ denote which method to use for correction and standardization.

The wavelength variable region is defined by the rest of the other genes, W_j and L_j denote the position and window width of the jth window, respectively, and the window numbers depend on the actual situation. For example, if W_1 is 1100 and L_1 is 200, then the first window covers 1100–1299 nm, as shown in Fig. 1.

Finally, an individual chromosome with $11 + n \times 2$ genes represented with non-negative decimal integers can be generated. Here n is the window number.

2.3. Decoding rule

After genetic operations, CA-CSMWPLS chromosome is decoded by the following rule: (1) a larger O_x value correlates to a higher corresponding pre-processing method priority; (2) if $O_x = 0$, the pre-processing method x is abandoned for modeling; (3) if there are two or more than two equal O_x, only the leftmost O_x is used for calculating; (4) F_1 and F_2 are converted into $F^*{}_1 = F_1 \times 2 + 1$ and $F^*{}_2 = F_2 \times 2 + 1$ to ensure that these values are odd numbers; (5) if $P_1 = 1$, the correction mode is MSC, else it is SNV; (6) if $P_2 = 1$, standardization mode is centralization, else it is autoscale; (7) if some windows overlap, the overlapping wavelength variables are calculated only once; (8) if some windows overstep the boundary of spectra, the exceeded wavelength variable is deleted from the input data set.

2.4. Fitness function

RMSECV, always considered as a criterion of PLSR, is acquired from cross validation, which includes random subsets K-fold cross validation (RSCV), contiguous blocks K-fold cross validation (CBCV), and venetian blinds K-fold cross validation (VBCV). For RSCV, during the ith iteration, an RMSECV is obtained when a combination of samples is constructed; however, during the $(i + 1)$th iteration, an RMSECV* is obtained when another combination of samples is constructed. Because of the randomness of RSCV, RMSECV* is not always smaller then RMSECV, and there may be unstable RMSECV in optimal operation when the $K > 1$. As for CBCV and VBCV, there are some combinations of certain samples not employed to model [22]. So when $K = 1$, leave-one-out cross validation is adopted in this paper. Moreover, with respect to the parsimonious and robust performance model [23], a smaller latent variable performs better. The fitness function is defined by Eq. (2):

$$f = \mathrm{RMSECV} + \eta \times Q = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} + \eta \times Q, \quad \eta \in \left[\frac{\mathrm{RMSECV}}{Q^*} - \theta, \frac{\mathrm{RMSECV}}{Q^*} + \theta\right] \tag{2}$$