Analytica Chimica Acta 1000 (2018) 109–122

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Novel near-infrared spectrum analysis tool: Synergy adaptive moving window model based on immune clone algorithm

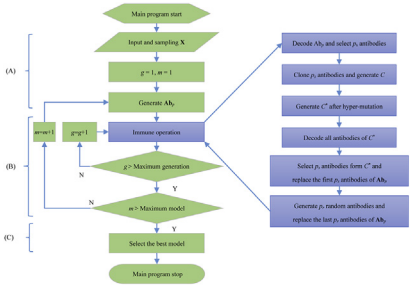
Shenghao Wang [a, b], Yuyan Zhang [b], Fuyi Cao [b], Zhenying Pei [b], Xuewei Gao [b], Xu Zhang [b], Yong Zhao [a, *]

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

HIGHLIGHTS

- SA-MWM-ICA is first introduced into spectrum analysis field as a new optimization strategy.
- SA-MWM-ICA is obviously better than usually methods, especially for spectra with relatively complex structure.
- SA-MWM-ICA can take full advantage of the useful information throughout the spectral range.
- SA-MWM-ICA can converge in fewer generation.
- The parameter setting requirements of SA-MWPLS-ICA are easier to operate.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 18 September 2017
Received in revised form
13 November 2017
Accepted 16 November 2017
Available online 22 November 2017

Keywords:
Multivariate calibration
Pre-processing methods
Selection of spectral variables
Immune clonal algorithm
Spectral analysis
Partial least squares regression

ABSTRACT

This paper presents a novel spectrum analysis tool named synergy adaptive moving window modeling based on immune clone algorithm (SA-MWM-ICA) considering the tedious and inconvenient labor involved in the selection of pre-processing methods and spectral variables by prior experience. In this work, immune clone algorithm is first introduced into the spectrum analysis field as a new optimization strategy, covering the shortage of the relative traditional methods. Based on the working principle of the human immune system, the performance of the quantitative model is regarded as antigen, and a special vector corresponding to the above mentioned antigen is regarded as antibody. The antibody contains a pre-processing method optimization region which is created by 11 decimal digits, and a spectrum variable optimization region which is formed by some moving windows with changeable width and position. A set of original antibodies are created by modeling with this algorithm. After calculating the affinity of these antibodies, those with high affinity will be selected to clone. The regulation for cloning is that the higher the affinity, the more copies will be. In the next step, another import operation named hyper-mutation is applied to the antibodies after cloning. Moreover, the regulation for hyper-mutation is that the lower the affinity, the more possibility will be. Several antibodies with high affinity will be created on the basis of these steps. Groups of simulated dataset, gasoline near-infrared spectra dataset, and soil near-infrared spectra dataset are employed to verify and illustrate the performance of SA-MWM-ICA. Analysis results show that the performance of the quantitative models adopted by SA-MWM-ICA are better especially for structures with relatively complex spectra than traditional models such as partial least squares (PLS), moving window PLS (MWPLS), genetic algorithm PLS (GAPLS), and pretreatment

* Corresponding author.
E-mail address: 116112023@qq.com (Y. Zhao).

https://doi.org/10.1016/j.aca.2017.11.028

method classification and adjustable parameter changeable size moving window PLS (CA-CSMWPLS). The selected pre-processing methods and spectrum variables are easily explained. The proposed method will converge in few generations and can be used not only for near-infrared spectroscopy analysis but also for other similar spectral analysis, such as infrared spectroscopy.

1. Introduction

Since 1990, near-infrared spectral analysis has become an indispensable tool in industrial and agricultural production because of its excellent properties, such as convenient instrumentation, low cost, and suitability for online analysis, the use of this technology has reached an all-time high in the agriculture, pharmaceutical, petrochemical, environment monitoring, and food industries [1–6]. With the improvement in physical devices and the rapid development of chemometrics, this technology has achieved great success in various applications. In chemometrics, principal component regression (PCR) [7] and partial least squares regression (PLSR) [8] are two classical algorithms that overcome the limitations of severe multicollinearity, overlapping of analyte information spectrum, and interference of spectral noise.

Optimizing the spectral variables and spectral data preprocessing methods has become important step to strengthen the prediction accuracy of near-infrared spectrum (NIR) and improve the robustness of the model [9]. As the most primitive and simplest optimization method, single variable selection, which is based on the spectral variables and correlation coefficients of the analyte, can directly obtain the preferred variables by hand or by computer. Since the 1990s, with the rapid development of chemometrics, a number of multivariate selection methods have been developed; these methods include interactive variable selection [10,11], the successive projections algorithm [12], basic interval partial least squares [13], forward interval partial least squares (FiPLS) [14], backward interval partial least squares (BiPLS) [15], synergy interval partial least squares (SiPLS) [13], genetic algorithm based on interval partial least squares (GA-iPLS) [14], moving window partial least squares (MWPLS) [16], and variable moving window partial least squares (CSMWPLS) combined with moving window partial least squares (SCMWPLS) [17]. These methods are mainly used to select spectral variables. The selection of spectral preprocessing method is also an important task. Although common near-infrared spectral pretreatment methods only include scattering correction, spectral derivative, spectral smoothing, and spectral standardization [18], the actual work usually requires the combination of the parameters of these preprocessing methods and their calculation priorities. Selecting the best result in a variety of combinations is a significant challenge for any spectral analyst. The design of experiments (DoE) [19,20] can help the analyst to rapidly and objectively screen a large number of parameter settings for the calibration model. The optimal calibration model is then validated. Recently, a work using DoE and a special form of PLS even can optimize wavelength variables together with other model parameters such as preprocessing methods [21].

The emergence of artificial intelligence, as a new way, has solved the limitations of the above existing selection methods; this technology has been applied to genetic algorithm [22], ant colony algorithm [23], and particle swarm algorithm [24]. References [25] and [26] proposed a parallel genetic algorithm that can simultaneously optimize the spectral variables and spectral data pretreatment method by using quantitative models of corn moisture content, pork fat content, beet sucrose content, and heavy oil SARA (saturates, aromatics, resins and asphaltenes). An assembly method composed of ant colony algorithm and genetic algorithm was used to analyze a set of simulated data, sugar cane sugar content, and corn starch content [27]. In our previous work, we proposed a novel interval integer genetic algorithm [28] and a pretreatment classification method with adjustable parameter changeable size moving window partial least square (CA-CSMWPLS) [29] to analyze the near-infrared spectra of some substances; both methods can optimize the spectral variables and the parameters of the pretreatment method.

The artificial immune system was established in the 1990s, and clone selection algorithm is an important branch of this system [30]. In comparison with traditional evolutionary algorithms, clonal selection algorithm based on the cloning and replication of immune cells and fitness maturity principle is a novel optimization algorithm; this algorithm, similar to genetic algorithm, exhibits selection and duplication operation. Clonal selection technology in optimization problem is compared with the GA algorithm: first, the main advantage of the former is that its convergence rate is better than the latter; second, the individual diversity is more abundant, the solution cannot only obtain the global optimal solution but can also search the local optima, while the latter often only focus on the global optimal solution [31,32]. In addition, the spectral analysis technique using the immune optimization algorithm has not been reported yet. Therefore, this work establishes a novel algorithm named synergy adaptive moving window model based on immune clone algorithm (SA-MWM-ICA) based on the clonal selection technique and the CA-CSMWPLS algorithm. The word *modeling* in SA-MWM-ICA denotes a kind of modeling method such as PLS, support vector machine regression (SVR) or artificial neural network (ANN). The model exhibits satisfactory results when used to analyze three different kinds of data set.

This paper is arranged into the following. The second part will introduce the basic principles of SA-MWM-ICA. The third part describes the data sets and related experimental equipment used in this paper. The fourth part studies the important parameters of SA-MWM-ICA and analyzes three different datasets. The fifth part shows the conclusion of the paper.

2. Theory

2.1. Notations

Assume that the property of matter need to be analyzed is $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$, and the spectral data is $\mathbf{X}$ which has n samples and has l_a variables. $\mathbf{X}$ is divided into calibration set $\mathbf{X}_c$ with n_1 samples, validation set $\mathbf{X}_v$ with n_2 samples, and test set $\mathbf{X}_t$ with n_3 samples, that is $\mathbf{X} = \mathbf{X_c} \cup \mathbf{X_v} \cup \mathbf{X_t}$, $n = n_1 + n_2 + n_3$, and $\boldsymbol{y} = \boldsymbol{y_c} \cup \boldsymbol{y_v} \cup \boldsymbol{y_t}$. Given that PLS is a very classical method, it is adopted as a model strategy needed to be optimized. In this manner, SA-MWM-ICA is also called SA-MWPLS-ICA in the next section. Additionally, $\mathbf{X}$ and $\boldsymbol{y}$ are pretreated with mean center.