Spectrochimica Acta Part B 143 (2018) 12–17

Contents lists available at ScienceDirect

Spectrochimica Acta Part B

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

Automatic variable selection method and a comparison for quantitative analysis in laser-induced breakdown spectroscopy

Fajie Duan [a,b], Xiao Fu [a,*], Jiajia Jiang [a], Tingting Huang [a], Ling Ma [a], Cong Zhang [a]

[a] State Key Lab of Precision Measuring Technology & Instruments, Tianjin University, Tianjin 300072, China
[b] Electronic Engineering Department, Tsinghua University, Beijing 100084, China

ARTICLE INFO

Article history:
Received 25 October 2017
Received in revised form 19 February 2018
Accepted 19 February 2018
Available online 21 February 2018

Keywords:
LIBS
Quantitative analysis
Variable selection
Genetic algorithm
Successive projections algorithm

ABSTRACT

In this work, an automatic variable selection method for quantitative analysis of soil samples using laser-induced breakdown spectroscopy (LIBS) is proposed, which is based on full spectrum correction (FSC) and modified iterative predictor weighting-partial least squares (mIPW-PLS). The method features automatic selection without artificial processes. To illustrate the feasibility and effectiveness of the method, a comparison with genetic algorithm (GA) and successive projections algorithm (SPA) for different elements (copper, barium and chromium) detection in soil was implemented. The experimental results showed that all the three methods could accomplish variable selection effectively, among which FSC-mIPW-PLS required significantly shorter computation time (12 s approximately for 40,000 initial variables) than the others. Moreover, improved quantification models were got with variable selection approaches. The root mean square errors of prediction (RMSEP) of models utilizing the new method were 27.47 (copper), 37.15 (barium) and 39.70 (chromium) mg/kg, which showed comparable prediction effect with GA and SPA.

1. Introduction

Laser-induced breakdown spectroscopy (LIBS), as one of the most promising techniques, has demonstrated its tremendous potential in quantitative analysis of various samples [1–3]. Regarding to the chemometric processing of LIBS data in quantification, researchers have done great works on many meaningful issues such as noise reduction [4], normalization [5,6], background subtraction [7], outliers filtering [8], etc. Among them, variable selection draws plenty of attentions. LIBS spectra usually consist of huge numbers of atomic spectral lines due to the wide detection range and high resolution of the increasingly advanced spectrometer as well as the complex components of the samples. This may be beneficial for the application of multivariate analysis, which features more robust and accurate model than univariate analysis as is proved by many researchers [9–11]. However, too many variables along with the increase of number of samples may lead to considerable computational workload and complicated models. Therefore it is necessary to reduce the dimensions of the original spectrum and select proper number of variables for modeling. Several variable selection methods have been reported in different applications. Forina et al. proposed an iterative predictor weighting-partial least squares (IPW-PLS) method based on the cyclic iteration of PLS regression for the elimination of useless predictors in multivariate regression problems [12,13]. Chen et al. proposed a modified IPW-PLS (mIPW-PLS) method by defining a hard threshold and using continuous wavelet transform (CWT) [14]. Jouan-Rimbaud et al. studied the performance of genetic algorithms (GA) in variable selection and a good model was obtained [15]. Araújo et al. developed successive projections algorithm (SPA) as a novel variable selection strategy for multivariate calibration [16].

It is interesting to find that few variable selection methods were first proposed for the application of atomic spectrum (LIBS typically). The reason may be that the atomic spectrum of LIBS usually contains tens of thousands of informative variables, which would result in extensive calculation for classical variable selection methods such as GA, SPA etc. Pontes et al. use a data compression procedure in the wavelet domain to reduce the computational workload involved in the variable selection process and compared effects of methods SPA, GA and a stepwise formulation (SW) [17]. However, the wavelets utilized for compression and decomposition levels have to be tested beforehand to find out the best parameter, which may be different under other circumstances. In our early research, we proposed a fast variable selection method by combining interval PLS and mIPW-PLS with a defined correction factor [18]. Nevertheless, the number of intervals still has to be determined in advance and may change in different cases. In other words, some artificial processes are inevitable in existing variable selection methods, making them unadapted for various situations.

In this work, we present an automatic variable selection method used for LIBS technology, which gets rid of artificial process and keeps adaptable for quantitative analysis of different elements in soils. To illustrate the effectiveness and superiority of this method, we implement a

* Corresponding author.
E-mail address: fuxiao215@tju.edu.cn (X. Fu).

https://doi.org/10.1016/j.sab.2018.02.010
0584-8547/

comparison with conventional methods (GA and SPA). The computation time and the prediction effect are considered and compared in the results of variable selection and quantification.

2. Theory and mehod

Researchers have already found that modeling in quantitative analysis could be faster, more robust and accurate with variables selected properly [19–21]. To achieve this goal, methods with intelligent algorithm are preferable to manual selection methods because of their higher efficiency and better capability of selection. In this section, conventional variable selection methods including GA and SPA are described previously as comparisons. Then a novel automatic variable selection method is presented with detailed procedure.

2.1. Genetic algorithm

The genetic algorithm (GA) is a classical approach to intelligent searching and optimization, which was put forward in 1970s by Holland [22]. Then thanks to the large amounts of contribution made by Goldberg et al. [23], GA attracted enormous attention and was successfully applied for artificial intelligence and automation.

In the field of variable selection, GA could be utilized to search for latent variables spontaneously [24–26]. In detail, a population of binary strings (i.e. chromosomes) is created randomly from all the variables. Each position (i.e. gene) of the binary string corresponds to one specific variable, which is coded as "1" if the variable is selected and "0" if not. According to the algorithm, the offspring generations are formed with crossover and mutation operation. The probability of a given chromosome being selected is proportional to its fitness, which is related to the evaluation response (in this work refers to the root mean square error of cross validation). In each generation, the population size is kept constant and new individuals take place of old ones except the best one to avoid missing optimal solutions. The evolutionary process will come to an end when a determined number of cycles are reached.

GA could find out latent variables through random searching and combination. As an effective intelligent algorithm, it has many benefits. For instance, GA features global searching capability, which could guarantee avoidance of the local extremum. However, the drewbacks of GA also exist. On one hand, variables selected may be different to some extent in repeated experiments because of the random searching procedure adopted in GA. On another hand, the process of variable selection will be rather time-consuming when the size of population is large.

2.2. Successive projections algorithm

The successive projections algorithm (SPA) was firstly proposed by Araújo et al. for simultaneous analysis of complexes [16]. Afterwards, several researchers presented improvements of SPA such as a cost function associated to the average risk of misclassification [27], combination with uninformative variable elimination (UVE) [28] or new criteria for selection of robust variables for classification [29].

SPA is a forward selection method, which can be used to select a small representative set of spectral variables with a minimum of collinearity. The main process of SPA can be summarized as follows: firstly, a parameter N should be set as the maximum number of variables to be selected; secondly, starting from each variable, SPA calculates the projection of the initial variable on the subspace by iteration and gets J (total number of variables) sets of selection of N variables [28].The function of the projection is defined as [16].

$$\mathbf{Px}_j = \mathbf{x}_j - \left(\mathbf{x}_j^{\mathrm{T}}\mathbf{x}_{k(n-1)}\right)\mathbf{x}_{k(n-1)}\left(\mathbf{x}_{k(n-1)}^{\mathrm{T}}\mathbf{x}_{k(n-1)}\right)^{-1}$$

where $\mathbf{P}$ is the projection operator, $\mathbf{x}_j$ is the jth column (corresponding to jth variable) of calibration set $\mathbf{X}$, $\mathbf{x}_{k(n\text{-}1)}$ refers to the column selected at the (n-1)th iteration of SPA. Thirdly the optimal set and selected variables could be determined by the assessment of the root mean square error of cross-validation (RMSECV) in multiple linear regression (MLR).

SPA employs operations in a vector space to select variables for either quantitative or qualitative analysis. The advantage of SPA is that the selected variables feature the most information but the smallest collinearity. However, the disadvantages of SPA also exist. In the first place, the number of variables to be selected should be no more than the number of samples in the calibration set. In the second place, when the total number of the variables is huge, the iterations of projection calculation and circulations of every each initial variable will still bring about considerable workload.

2.3. Automatic variable selection method based on full spectrum correction and modified iterative predictor weighting-partial least squares (FSC-mIPW-PLS)

To select variables efficiently and automatically considering the enormous amount of spectral data in LIBS experiment, in this work an automatic variable selection method is developed based on improvement of our previous research [18]. The new method employs two defined full spectrum correction factors combined with mIPW-PLS to realize automatic selection. The detailed procedure could be decomposed into the following steps:

(1) Correlation coefficient ρ_i of each variable i and the reference concentration is treated as the first correction factor and calculated as

$$\rho_i = \frac{Cov(\mathbf{X}_i, \mathbf{Y})}{\sigma_{\mathbf{X}i} \cdot \sigma_{\mathbf{Y}}}$$

where $Cov(\boldsymbol{X}_i, \boldsymbol{Y})$ is the covariance of $\boldsymbol{X}_i$ and $\boldsymbol{Y}$, $\sigma_{\boldsymbol{X}i}$ and $\sigma_{\boldsymbol{Y}}$ are the standard deviation of $\boldsymbol{X}_i$ and $\boldsymbol{Y}$. All the variables are multiplied by the vector $\boldsymbol{R} = [\rho_1, \rho_2, \ldots, \rho_N]$.

(2) PLS regression of all the variables after (1) is implemented and the second correction factor k_i of each variable i is defined as

$$k_i = \frac{b_i^2}{\sum_{i=1}^{m} b_i^2}$$

where b_i is PLS regression coefficient of the corresponding variable i, and m is the total number of variables. All the variables are multiplied by the vector $\boldsymbol{K} = [k_1, k_2, \ldots, k_N]$.

(3) PLS regression of variables in the current IPW cycle is computed and the importance z_j of each variable j is given by [12]

$$z_j = \frac{|b_j|\sigma_j}{\sum_{j=1}^{n}|b_j|\sigma_j}$$

where σ_j and b_j are the standard deviation and PLS regression coefficient of the corresponding variable j, and n is the number of variables in current cycle. The hard threshold Thr of the current IPW cycle is calculated as [14].

$$Thr = \frac{\sqrt[v]{2\ \log_2(n)}}{\sqrt{n}}$$

where σ is the standard deviation of all variables included in current cycle, and n is the number of variables in current cycle. If any importance z_j is smaller than Thr, the corresponding variable j will be removed.