Chemometrics and Intelligent Laboratory Systems 130 (2014) 76–83

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A simple idea on applying large regression coefficient to improve the genetic algorithm-PLS for variable selection in multivariate calibration

Yong-Huan Yun [a], Dong-Sheng Cao [b], Min-Li Tan [a], Jun Yan [a], Da-Bing Ren [a], Qing-Song Xu [c], Ling Yu [d], Yi-Zeng Liang [a,*]

[a] College of Chemistry and Chemical Engineering, Central South University, Changsha 410083, PR China
[b] College of Pharmaceutical Sciences, Central South University, Changsha 410083, PR China
[c] School of Mathematics and Statistics, Central South University, Changsha 410083, PR China
[d] Shanghai Tobacco Group Co., Ltd., Shanghai 200082, PR China

ARTICLE INFO

Article history:
Received 22 July 2013
Received in revised form 30 August 2013
Accepted 15 September 2013
Available online 19 September 2013

Keywords:
Genetic algorithm
Variable selection
Partial least squares
Regression coefficient
Multivariate calibration

ABSTRACT

Genetic algorithm-based couple with partial least squares (PLS) has been successfully applied for variable selection in multivariate calibration. On the basis of the fact that a large PLS regression coefficient indicates an important variable, a new and simple idea that the structure of a proportion of chromosomes in the initial population is determined by the large regression coefficient is presented in this study. The regression coefficient is obtained by establishing the PLS modeling on the autoscaled data. With this improved approach, the modified GA-PLS method not only makes the optimization better toward the optimal solution, but also obeys the rule of the GAs. The results obtained through investigating one simulated dataset and two near infrared dataset show that the modified method has made much improvement on variable selection compared to the original GA-PLS.

1. Introduction

Variable selection techniques have been either theoretically or experimentally proved to become a critical step to obtain good prediction performance [1–5]. In recent years, a large amount of variable selection methods has been developed to be applied into multivariate calibration such as uninformative variable elimination (UVE) [6], Monte Carlo based UVE (MC-UVE) [7], competitive adaptive reweighted sampling (CARS) [8,9], latent projective graph (LPG) [10], influential variable (IV) [11], successive projection algorithm (SPA) [12], random frog [13,14] and the optimized algorithm like stepwise selection [15], forward selection [15,16], backward elimination [15–17], genetic algorithms (GAs) [18–21] and simulated annealing (SA) [1]. Mehmood et al. recently have presented a review of available methods for variable selection [22]. Zou et al. also reviewed the variable selection methods in near-infrared (NIR) spectroscopy [23].

Genetic algorithms (GAs) were first introduced by Holland in 1975 as an optimization approach with the natural principle 'survival of the fittest' on which Darwin's Evolution Theory is based [24]. Since then, GAs have been gaining widespread application into different research fields, such as multivariate calibration, optimization, quantitative structure–activity relationship/molecular modeling and other miscellaneous applications. All of these applications in chemometrics were reviewed by Niazi and Leardi [25,26]. For multivariate calibration, owing to the datasets in which each sample is described by hundreds of or thousands of variables generated by new instrumentation, one of the greatest problems is to select the optimal subset of variables that can improve the prediction performance of the predictors. Furthermore, it is very important and essential to conduct variable selection that was theoretically or experimentally proved by many papers [1-4]. Massart et al., used GAs coupled with partial least squares (PLS) and multiple linear regression (MLR) as a tool for variable selection in multivariate calibration [27]. Due to the spectral datasets that are usually of high collinearity, the latent variables method like PLS has an advantage to address this problem over MLR. Many papers about the application of GA-PLS for variable selection have been published [27-39].

GAs have five basic steps: (1) coding of variables; (2) initiation of population; (3) evaluation of the response; (4) reproduction; (5) mutation. Loop the steps3–5 until a required termination criterion is satisfied. In this paper, a simple idea that makes good use of the large PLS regression coefficient into step2 (initiation of population) was proposed to improve the performance of GA-PLS for variable selection. This modified GA-PLS method was named as GA-PLS-LRC (large regression coefficient). It was proved that a large absolute PLS regression coefficient indicates an important variable in a model obtained for autoscaled data [6, 40]. Originally, the structure

* Corresponding author. Tel.: +86 731 8830824; fax: +86 731 8830831.
E-mail address: yizeng_liang@263.net (Y.-Z. Liang).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.09.007

of each chromosome of the initial population is totally determined in a random way. In this study, some chromosomes of the initial population are composed of the variables with large PLS regression coefficient obtained by a model using all variables. The larger the PLS regression coefficient is, the more opportunity the variable has to be introduced into chromosome. In fact, it is equal to give some prior knowledge for the initial population. Other steps of GAs remain no change. It should be noted that just a proportion of the chromosomes is composed of the variables determined by the PLS regression coefficient. The variables of the left chromosomes are randomly determined. Thus, GA-PLS-LRC not only makes the optimization toward better approaching the best response, but also not affects the rule of the GAs. In order to validate this idea, one simulated dataset together with two NIR datasets was investigated. The results show that the GA-PLS-LRC is superior to the original GA-PLS proposed by Leardi [29], which indicates that this method is worthy to make more implementation.

2. Theory and method

2.1. Genetic algorithm-PLS

There are many different versions of genetic algorithms that perform reproduction, crossover, etc. in different way. The algorithm we have used for this study is specifically devoted to the problem of variable selection. The GA-PLS proposed by Leardi was used for this work [29]. The whole steps are introduced as follows:

Step1. Define the parameters of the GA-PLS: all the parameters are listed in Table 1. The parameters were defined according to the Ref. [28, 29]. The details of the settings required are fully covered in the references.

Step2. Initiation of population: each chromosome in the population is row vector containing as many genes as there are variables, each gene being coded as 1 if the corresponding variable was selected and 0 if not. The structure of each chromosome is determined in a totally random way. Of note, each chromosome would be checked to avoid having the same structure in the population.

Step3. Evaluation of the response: based on the variables selected by each chromosome, a number of subset data could be extracted from the full data. PLS regression method is applied into each subset to evaluate the response that is the cross-validated explained variance (CVEV, %). The larger the value of the cross-validated explained variance, the better the chromosome.

Step4. Crossover and mutation: in order to generate two new chromosomes as offspring, one pair of chromosome of the existing population is randomly selected to carry out crossover and mutation approach as well as to evaluate the cross-validated explained variance of the new offspring. At this step, two new chromosomes should be also checked to avoid containing the same variables.

Step5. Update the population by comparing the CVEV of the two new chromosomes with the one of the existing chromosomes of the current population. The updating rule is that each chromosome of the new offspring would survive if it is better than the worst chromosome which would be discarded later.

Step6. Go back to **Step4** when the amount of the evaluations does not satisfy the criterion of the entrance of backward selection. When it is satisfied, backward selection is conducted to choose the best subset of the population.

Step7. If the criterion of the final termination is reached, the whole evolution process of GA has ended. If not, go back to **Step4**. As we can see the parameters from the Table 1, the amount of evaluations is set to 200. Thus, only if the evaluations reach 200, the GA run is terminated.

It is noted that from **Step1** to **Step7** is one run of GA-PLS. The final model is selected from the results of a number of independent and very short runs, which can reduce the risk of overfitting [28].

Step8 After processing the predefined runs, the selection frequency of each variable could be obtained. Rank the variables by the selection frequencies, and then choose the optimal subset with the maximum CVEV according to the ranking of variables.

Fig. 1 briefly shows the flowchart of GAs.

2.2. Genetic algorithm-PLS guided by large regression coefficient

PLS is a multivariate calibration method for modeling the relationship between chemical measured variables $\mathbf{X}$ ($n_{\text{objects}} \times p_{\text{variables}}$) and properties of interest $\mathbf{y}$ ($n_{\text{objects}} \times 1$). The relationship is as follows:

$$\mathbf{y} = \mathbf{X} \times \mathbf{b} + \mathbf{e} \tag{1}$$

Where $\mathbf{b}$ is the vector of PLS regression coefficients and $\mathbf{e}$ is the vector of residuals. The regression coefficients cannot be used directly to select which variables are the most important for modeling because a large regression coefficient may indicate a variable with small absolute value and a large variance.

To address this problem, the data should be autoscaled (standardization) as follows:

$$\mathbf{X}_{\text{autoscaling},i} = (\mathbf{X}_i - \mu)/\sigma \quad i = 1, 2, \ldots, p \tag{2}$$

Where $\mathbf{X}_i$ represents the i column vector of $\mathbf{X}$. The μ is the mean value of each column of $\mathbf{X}$, and σ is standard deviation value of each column of $\mathbf{X}$. With this pretreatment, the $\mathbf{X}$ has been weighted with the inverse of the standard deviation to guarantee that each variable has the same variances. The formula (1) is transferred as follows:

$$\mathbf{y} = \mathbf{X}_{\text{autoscaling}} \times \mathbf{b}_w + \mathbf{e}_w \tag{3}$$

Now, it can be said that a large absolute b_w indicates an important variable of $\mathbf{X}$.

Based on the knowledge that important variables can be identified by the regression coefficient, we think that if the structure of the chromosome is determined by the important variables in the initial population, GA-PLS can perform better for variable selection with a better starting response (the max CVEV of the initial population) and ending response (the max CVEV of the final population). Thus, in this study, we propose the modified GA-PLS method guided by large regression coefficient, called GA-PLS-LRC. It should be noted that GA-PLS-LRC just modifies the **Step2** of GA-PLS in Section 2.1.

To maintain the random selection of GA-PLS, just a proportion of chromosomes is depending on important variables, and the rest obeys

Table 1
Parameters of the GA-PLS.

Population size: **30** chromosomes
On average, **5** variables per chromosome in the original population
Response: cross-validated explained variance % (**5** deletion groups; the number of PLS components is determined by cross-validation)
Maximum number of variables selected in the same chromosome: **30**
Probability of cross-over: **50**%
Probability of mutation: **1**%
Maximum number of PLS components: **15**
Number of runs: **100**
The amount of evaluations: **200**
Backward elimination after every **100th** evaluation and at the end (if the number of evaluations is not a multiple of 100)