Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 91 (2008) 194–199
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Successive projections algorithm combined with uninformative variable elimination for spectral variable selection

Shengfeng Ye, Dong Wang, Shungeng Min *

College of Science, China Agricultural University, Beijing, 100094, PR China

Received 15 October 2007; received in revised form 18 November 2007; accepted 24 November 2007
Available online 4 December 2007

Abstracts

The UVE–SPA method, successive projections algorithm (SPA) combined with uninformative variable elimination (UVE) is proposed as a novel variable selection approach for multivariate calibration. UVE is used to select informative variables, and SPA is followed to select variables with minimum redundant information from the informative variables. The proposed method was applied to near-infrared (NIR) reflectance data for analysis of nicotine in tobacco lamina and NIR transmission data for active pharmaceutical ingredient (API) in single tablet. On the aspect of elimination of uninformative variables, the effect of UVE using first derivative spectra was better than that of using raw spectra. In terms of variable selection, fewer variables with better performance were selected by UVE—SPA method than by direct SPA method. For NIR spectral analysis of nicotine in tobacco lamina, the number of variables selected from 3001 spectral variables reduced from 48 by direct SPA to 35 by UVE–SPA, and the root mean squared error of prediction set (RMSEP) of the corresponding MLR models decreased from 0.174 (%, mg/mg) to 0.160. For NIR spectral analysis of API in each tablet, the number of variables selected from 650 spectral variables reduced from 46 by direct SPA to 17 by UVE–SPA, and RMSEP of the corresponding multiple linear regression (MLR) models decreased from 0.842 (%, mg/mg) to 0.473. MLR model using variables selected by UVE–SPA had better prediction performance than full-spectrum partial least-squares (PLS) model, and comparable to PLS model of UVE.

Keywords: Successive projections algorithm; Uninformative variable elimination; Variable selection; Near-infrared spectroscopy; Multivariate calibration

1. Introduction

Nowadays, many researchers make use of partial least-squares (PLS) because it is a powerful and full-spectrum based method. However, employing full spectral region does not always yield optimal results because it may include regions which contribute more noise than relevant information to models. Therefore, uninformative variable elimination (UVE) as a variable selection method, proposed by V. Centner, et al. [1], has been used to solve such problems and improve the quality of models [1–7].

PLS employs latent variables instead of real variables and provides complex models. Multiple linear regression (MLR) models are simpler and easier to interpret, but they are very affected by collinearity between variables [8]. The successive projections algorithm (SPA) proposed as a variable selection strategy by M.C.U. Araújo, et al. [9], shows the advantage of finding a small representative set of spectral variables with a minimum of collinearity. SPA has been successfully applied to select variables in UV–VIS [9–12], NIR [12–14] and ICP-AES [15] spectrometer, as well as for coefficient selection in wavelet regression models [16–18].

UVE can eliminate the variables which have no more informative variables for modeling than noise, and employing the variables selected by UVE for modeling can avoid a model over-fitting and usually improve its predictive ability. But latent variables are still required to be employed for modeling because the number of the variables selected is still too large. SPA employs simple projection operations to select variables with

* Corresponding address. Tel.: +86 10 62733075; fax: +86 10 62733075.
E-mail address: minsg@cau.edu.cn (S. Min).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.11.005

minimum of collinearity, but variables selected by SPA may be with low signal noise ratio (*S*/*N*) or useless for multivariate calibration, which can affect model precision of prediction.

In this work, successive projections algorithm combined with uninformative variable elimination called UVE–SPA method is proposed for spectral variable selection, which SPA is employed for variable selection after UVE discards uninformative variables. The proposed method was applied to two sets of NIR data for analysis of nicotine in tobacco lamina and active pharmaceutical ingredient (API) in single tablet, respectively. MLR models were developed employing the original instrument response data of spectral variables selected by UVE–SPA, and the property parameters interested could be predicted accurately using raw spectral data without any pretreatment.

2. Theory

2.1. Uninformative variable elimination

UVE is a method of variable selection based on stability analysis of regression coefficient (*b*). The main steps of UVE can be summarized as follows [1]:

(1) First PLS regression is performed on instrumental response data (**Xcal**) and property values (**y**) of calibration set, and the optimal number of PLS factors is determined.
(2) Then a noise matrix with the same size of the **Xcal** matrix are generated, whose elements are random numbers in the interval of 0.0–1.0. And the elements are multiplied with a small constant to make their influence on the model negligible.
(3) The noise matrix is appended to the original one to form an extended matrix with twice as many variables as the original one.
(4) PLS models are made on the extended matrix and **y** in manner of leave-one-out cross validation. This leads to a matrix of *b* values with as many rows as samples and one column for each variable, both original and random.
(5) The *c* value of each variable is calculated as the average of the *b* values of each column divided by the standard deviation of that column.
(6) The cut-off value is set as the maximum of absolute value *c* among the random variables. Every original variable with equal or lower absolute value of *c* is assumed to contain nothing but noise and is eliminated.

2.2. Successive projections algorithm

SPA employs simple projection operations in a vector space to obtain subsets of variables with small collinearity [9], is a forward variable selection algorithm for multivariate calibration. The principle of variable selection by SPA is that the new variable selected is the one among all the remaining variables, which has the maximum projection value on the orthogonal subspace of the previous selected variable.

The detailed description of SPA see Ref. [9], and the main procedures are summarized here. Set the maximum number of variables *N* to be selected. Starting from each variable, SPA yields *K* (total number of variables) sets of selection of *N* variables. The optimal initial variable and number of variables can be determined on the basis of the smallest root mean squared error of prediction in validation set of MLR calibration.

2.3. SPA–UVE method

The UVE–SPA method is a combination method of UVE and SPA, UVE is employed to select informative variables, and SPA is followed to select variables which have minimum redundant information from the informative variables. On aspect of spectral variable selection, the obvious advantage of the UVE–SPA lies in two aspects compared with direct SPA: (1) The first advantage is to make the association of variables and property closer; (2) The number of variables required to be sought by SPA is reduced.

In this work, the small constant of UVE is set as 1/200 of the minimum value of variable variation in **Xcal** matrix, and the maximum number of variables *N* to be selected by SPA is set as 60.

The influence of first derivative spectra on UVE and SPA for variable selection was discussed. First derivative is a widely used spectral preprocessing method, and can remove most of the influence of baseline variation. Small window of derivative will generally reduce *S*/*N* of spectra, more obviously in the spectral region with low *S*/*N*, which may be helpful for UVE to remove variables with relatively low *S*/*N* and adverse to SPA. The derivative spectra used in this article were the first derivative spectra with 9 points after smoothing by a Savitzky–Golay filter with a second-order polynomial and a 5-point window.

3. Experimental

3.1. Diffuse reflectance NIR spectra of tobacco lamina

NIR diffuse reflectance spectra of 507 tobacco lamina samples were measured on a Perkin-Elmer Spectrum ONE NTS FT-NIR spectrometer. The spectra were recorded over the wavenumber range of 10,000–4000 cm^{-1} at 8 cm^{-1} resolution and with 2 cm^{-1} interval. Each spectrum was the average of 64 scans. The concentration of nicotine was measured by continuous flow method. Before calibration, samples were split into three set, the calibration set consisted of 200 samples, the validation set consisted of 100 samples, and the prediction set consisted of the remaining 207 samples.

3.2. Transmittance NIR spectra of pharmaceutical tablets

655 transmittance spectra of pharmaceutical tablets were obtained from the web of http://software.eigenvector.com/Data/tablets/index.html. The spectra were measured on Instrument I of Foss NIRSystems 6500 spectrometer. The spectra were recorded in the region from 600 to 1898 nm in 2 nm increments. The active pharmaceutical ingredient (API) content of each