Chemometrics and Intelligent Laboratory Systems 156 (2016) 7–13

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Strategy for constructing calibration sets based on a derivative spectra information space consensus

Zhigang Li [a,*], Jiemin Liu [a], Peng Shan [a], Silong Peng [b], Jiangtao Lv [a], Zhenhe Ma [a]

[a] College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning 110819, China
[b] Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

ARTICLE INFO

Article history:
Received 24 November 2015
Received in revised form 8 May 2016
Accepted 13 May 2016
Available online 19 May 2016

Keywords:
Consensus selection
Calibration set
Multivariate regression model
Partial least squares (PLS)
Derivative spectra

ABSTRACT

Constructing an excellent calibration set is crucial to ensuring accurate multivariate calibration of spectra data. The purpose of this paper is to present an improved Kennard–Stone (KS) calibration set construction strategy based on different derivative spectra information spaces, termed Consensus Kennard–Stone (CKS). The core idea is to make full use of different derivative spectra information spaces when constructing the calibration set using a consensus selection method as well as to improve the prediction performance of the multivariate regression model. The experimental results from two public spectra datasets indicate that the proposed CKS strategy can use a more appropriate subset of samples for constructing the calibration set in the multivariate regression model and has superior predictive performance compared with the existing classic sample-selection KS strategies.

1. Introduction

Multivariate regression model have been widely used in the spectra quantitative analysis and chemical analysis fields [1–4]. The most commonly used multivariate calibration technique is partial least squares (PLS) and the performance of the PLS model heavily depends on the calibration set samples. An appropriate calibration set that includes a more representative sample should be constructed from a pool of samples.

Several approaches have addressed the problem of constructing a representative subset as a calibration set for the PLS model from a large dataset [5–11]. Among them, the Kennard–Stone (KS) algorithm is well known in the field of chemometrics [5]. The KS algorithm is aimed at separating the samples in the calibration and test sets. It selects representative samples to uniformly cover the spectra data space by maximizing the Euclidean distances between the instrumental response vectors of the selected samples. However, the KS algorithm is currently only used in a single spectra information space, which is mainly in the original or smooth spectra space. As a result, the implementation of the KS algorithm relies heavily on the quality of the spectra. In fact, useful information can be thoroughly mined from the samples by combining the multiple spectra information space, especially the derivative spectra space. Therefore, the inclusion of different spectra information spaces in constructing the calibration set process often results in a more reasonable distribution of the calibration subset, improving the quality of the multivariate calibration. This topic merits further study.

An improved Kennard–Stone calibration construction strategy based on derivative spectra information space, termed Consensus Kennard–Stone (CKS), is proposed. The core idea is to fully use different spectra information spaces when constructing the calibration set via the consensus selection method as well as improve the model performance. Two different public spectra datasets are used to evaluate the proposed CKS strategy. The performances of the PLS model based on KS and the performance of PLS model based on CKS are compared in our experiment. We first introduce the method of obtaining derivative spectra information space, the classic KS algorithm and proposed CKS strategy in Section 2. Section 3 introduces the sample data sets used in the experiments and the calculation method. The results and discussion are presented in Section 4. A brief conclusion is presented in Section 5.

2. Materials and methods

2.1. The method of obtaining derivative spectra information space

The derivative spectra method is one of the most widely used approaches for data analysis because of its ability to extract more analytical information from the spectra. Additionally, this approach provides a better resolution of the spectral overlap by magnifying

* Corresponding author at: College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning, 110819, China.
E-mail address: lizgqhd@163.com (Z. Li).

http://dx.doi.org/10.1016/j.chemolab.2016.05.007
0169-7439/

small differences between the spectral curves [12–15]. Calculating derivatives of the spectra data with the Savitzky–Golay (SG) algorithm is a landmark development. As a preliminary preprocessing step, it can effectively resolve overlapping signals, suppress unwanted spectral features and enhance signal properties [16–17]. However, it has some limitations and caveats whereby its usefulness heavily depends on proper selection of various parameters, such as the polynomial order and window size.

A new derivative spectra estimator (DSE) based on the singular perturbation technique was designed according to the description in a previously published paper [18]. The DSE was designed as follows:

$$\begin{cases} \dot{x}_1(v) = x_2(v) \\ \dot{x}_2(v) = x_3(v) \\ \dot{x}_3(v) = -\dfrac{P_3^3}{\varepsilon^3}(x_1(v) - u(v)) - \dfrac{P_3^2}{\varepsilon^2}x_2(v) - \dfrac{P_3^1}{\varepsilon}x_1(v) \\ y(v) = x_1(v) \end{cases} \tag{1}$$

$u(v)$represents the spectra signals from the instrument. $P_n^m = \frac{n!}{(n-m)!}$, $m = 1,2,3.n = 1,2,3$. x_1, x_2 and x_3 are the zero-order (smoothing), first-order and second-order derivative signals of $u(v)$, respectively. The parameter ε is the system perturbation parameter, and $\varepsilon > 0$. Currently, the parameter ε selection procession is performed by a trial-and-error method. Considering that RMSEP and RMSECV are suitable as evaluation indicators for the model performance, εcorresponding to lower RMSEP and RMSECV was selected. Additionally, there should be no significant differences between the performance of the calibration and test sets for successful modelling. That is, RMSEP and RMSECV are as close as possible. In the future, we will focus on designing a suitable optimizing function for automatic parameter selection.

2.2. Kennard–Stone (KS) algorithm

The classic KS algorithm focuses on selecting a representative subset from many samples. KS considers the instrumental response (X) in subset selection. To ensure a uniform distribution of the subset along the X data space, the KS algorithm follows a stepwise procedure in which new selections are taken in the regions of the space that are far from the samples already assigned to the subset. The closest sample to the mean of the data set can be considered the most representative one, and it is include as the first subset sample. For this purpose, the algorithm employs the Euclidean distances dx(m, n) between the x-vectors of each pair (m, n) of samples, which are calculated as

$$d_X(m,n) = \sqrt{\sum_{j=1}^{J}[X_m(j) - X_n(j)]^2} \qquad m,n \in [1,N] \tag{2}$$

Wherein, $x_m(j)$ and $x_n(j)$ are the instrumental responses at the jth wavelength or wave-number for samples m and n, respectively. j denotes the number of wavelengths or wave-numbers in the spectra. The procedure is repeated until a specified number of samples are achieved.

2.3. Consensus selection strategy based on derivative spectra information space

Construction of the calibration set is the key factor for the PLS models in quantitative analysis fields. The sample set should be as extensive and well distributed as possible. The calibration set should encompass the expected range of concentrations for the sample components. Consensus (ensemble) strategies have been widely used in multivariate regression modelling to improve the model performance [19–21]. However, this strategy rarely considers the process of constructing the calibration set. As a result, it is meaningful to explore construction of the calibration set via a consensus strategy. In this paper, to construct an excellent calibration set that preserves the data set information, an improved KS calibration construction strategy based on the derivative spectra space, CKS, is proposed. All samples are divided into the initial calibration set and initial validation set in the respective derivative spectra space using the KS method. CKS finds the common parts of three initial calibration data sets, which are generated from respective derivative spectra information

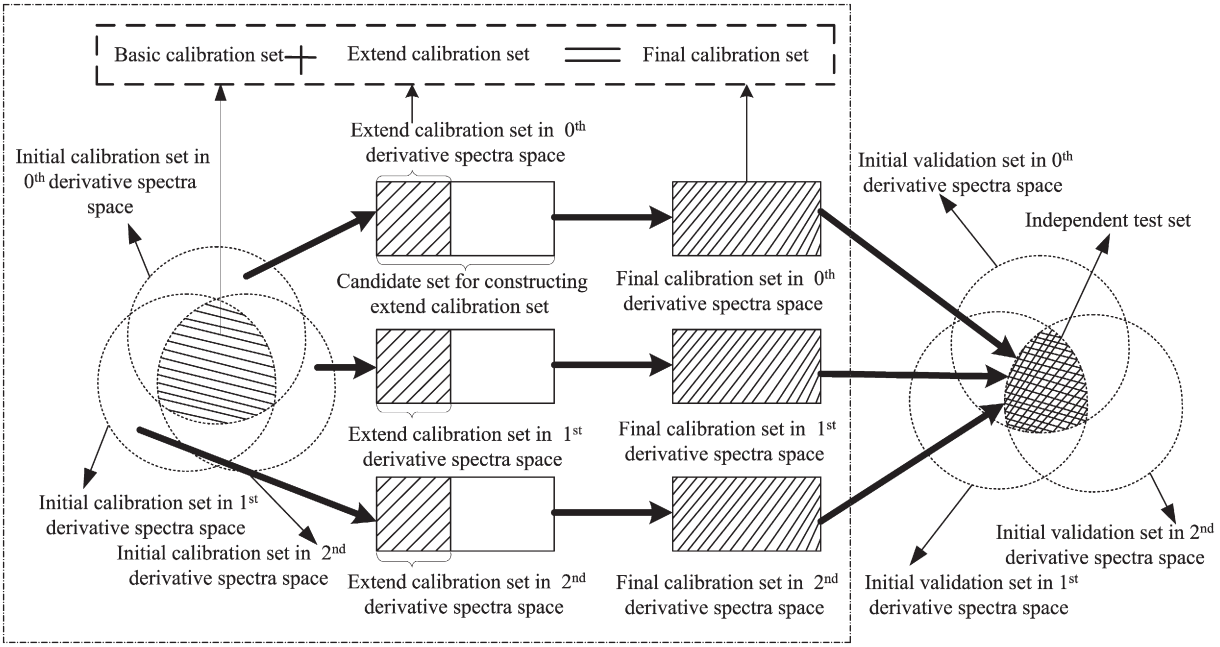

Fig. 1. Sketch map of the CKS sample selection strategy.