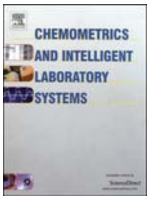
Chemometrics and Intelligent Laboratory Systems 131 (2014) 31–36

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS

Variable selection based on locally linear embedding mapping for near-infrared spectral analysis

Ruifeng Shan, Wensheng Cai, Xueguang Shao *

State Key Laboratory of Medicinal Chemical Biology, Research Center for Analytical Sciences, College of Chemistry, Nankai University, Tianjin 300071, PR China

ARTICLE INFO

Article history:
Received 30 May 2013
Received in revised form 5 December 2013
Accepted 14 December 2013
Available online 21 December 2013

Keywords:
Variable selection
Locally linear embedding
Monte Carlo cross validation
Partial least squares regression
Forward stepwise selection
Near-infrared spectroscopy

ABSTRACT

Locally linear embedding (LLE) is a nonlinear dimensionality reduction method that can preserve the relationship between samples in the mapping space. The neighbors in high dimensional space will keep their relative position in LLE space. A method based on the effect of the variables on the relative position of the samples in LLE space was proposed for variable selection in NIR spectral analysis. In the method, the spectra are mapped into LLE space with all variables at first, and then the mapping is repeated by removing a variable from the spectra. Therefore, the movement of the samples in LLE space caused by a variable can be used to evaluate the effect of the variable on the spectra. The variables that cause a large movement will be the important ones to affect the relationship of the spectra. For further selection of the informative variables specific to the target component, a forward stepwise selection is applied to the variables selected by LLE method. To validate the performance of the proposed method, it was applied to the partial least squares (PLS) modeling of three NIR spectral datasets of corn, pharmaceutical tablets and tobacco lamina samples. Results show that the proposed method can effectively select the informative variables from the NIR spectra, and build a parsimonious model by using several tens of selected variables.

1. Introduction

Multivariate calibration methods have been extensively used in the near-infrared (NIR) spectroscopic quantitative analysis. Much attention has been paid to variable selection in NIR spectral analysis for building accurate and parsimonious models. Methods based on optimization algorithms such as genetic algorithm (GA) [1–3], particle swarm optimization (PSO) [4], and interval partial least squares (iPLS) [5,6] have been applied to search the optimal subset of variables. The results of these methods demonstrated that better prediction can be obtained using the selected variables than the full spectrum. However, optimization algorithms generally need larger number of parameters and are time-consuming. Therefore, simple and efficient methods based on statistics were used for the problem. Uninformative variable elimination (UVE) and its variants [7–9], randomization test (RT) [10], Bayesian variable selection [11], successive projections algorithm (SPA) [12,13], etc. have been proposed. These methods evaluate the variables statistically and then select the variables with higher or lower statistical value. Additionally, stepwise selection method was widely used in NIR spectral analysis due to the simplicity. Competitive adaptive reweighted sampling (CARS) [14,15] selects variables in a stepwise and efficient way. In our recent works, methods based on the detection of the influential variables [16] and latent projective graph (LPG) [17] were proposed. These works proved that satisfactory PLS models can be established using several tens or even several informative variables.

Manifold learning techniques are developed for nonlinear dimensionality reduction, which discover compact representations of high dimensional data by recovering the underlying low dimensional manifold. Manifold learning algorithms such as locally linear embedding (LLE) [18], isometric mapping (Isomap) [19], Hessian LLE [20], and Laplacian Eigenmap [21] have been proposed. These methods are all based on Euclidean distance for exploiting the neighborhood information as same as locally weighted regression (LWR) and soft independent modeling of class analogy (SIMCA). Among these methods, LLE is a local method similar with LWR. The both of them need to select the nearby points and determine the weights. However, the weights in LLE are optimized by minimizing the reconstruction errors, while the weights in LWR are calculated according to the distance between the predicted data and the training data points [22]. Additionally, compared with LWR, LLE preserve the relationship between samples in mapping space and find the embedding in a noniterative way. Therefore, LLE and its extensions have become a promising techniques and used to solve the problem of dimension reduction of high dimensional data, such as face recognition [18,23], NIR spectra [24,25], gene expression [26,27], etc. Furthermore, it was also widely used in the data visualization.

In this work, a method for variable selection is proposed based on the effect of a variable on the mapping distance in LLE space. The spectra

* Corresponding author at: College of Chemistry, Nankai University, Tianjin 300071, PR China. Tel.: +86 22 23503430; fax: +86 22 23502458.
E-mail address: xshao@nankai.edu.cn (X. Shao).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.12.002

are mapped into the low dimensional space by LLE operation. According to the principle of LLE, the relationship, i.e., the relative position, between the samples in the spectral space will not change in the mapping space of LLE. However, if a variable that significantly affect the spectra is removed, the relative position of the samples in LLE space may change accordingly. Therefore, the change of the position, i.e., the movement of the samples, in LLE space caused by removal of a variable can be used to evaluate the effect of the variable on the spectra. Taking the movement calculated by the average Euclidean distance as a criterion, the variables that cause a large movement will be the important ones to the spectra.

2. Theory and calculations

LLE is a nonlinear mapping method that computes low-dimensional, neighborhood-preserving embeddings of high-dimensional inputs [18,23]. The basic idea of LLE is to approximate each data point by a linear combination of its neighbors and to find a low dimensional configuration of data points. In LLE algorithm, each data point and its neighbors are assumed to lie on or close to a locally linear patch of a manifold. Therefore, a data point can be approximated as a linear combination of its neighbors based on the assumption of local linearity. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of N points in a high D dimensional data space R^D. The corresponding set of N points in a low d dimensional data space R^d is denoted as $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_N\}$. For each data point $\mathbf{x}_i$, find its K neighbors by using Euclidean distance at first, and then the reconstruction weights $\mathbf{w}_i$ that best reconstruct $\mathbf{x}_i$ linearly by its K nearest neighbors can be optimized by minimizing the following cost function:

$$\varepsilon(\mathbf{W}) = \sum_{i=1}^{N} \left| \mathbf{x}_i - \sum_{j=1}^{K} w_{ij} \mathbf{x}_j \right|^2 \tag{1}$$

under the constraints that each vector of reconstruction weights $\mathbf{W}$ sums to unity. It should be noted that the constrained weights are invariant to rotations, rescalings, and translations. Therefore, the optimized reconstruction weights can characterize intrinsic geometric properties of each neighborhood in the high dimensional space.

In order to preserve the local geometry of the data in low dimensional space, the embedding $\mathbf{Y}$ of $\mathbf{X}$ can be reconstructed with the weights by minimizing the embedding cost function:

$$\phi(\mathbf{Y}) = \sum_{i=1}^{N} \left| \mathbf{y}_i - \sum_{j=1}^{K} w_{ij} \mathbf{y}_j \right|^2 \tag{2}$$

and subjecting to the following constraints:

$$\begin{cases} \sum_{i=1}^{N} \mathbf{y}_i = \mathbf{0} \\ \frac{1}{N} \sum_{i=1}^{N} \mathbf{y}_i^{\mathbf{T}} \mathbf{y}_i = \mathbf{I} \end{cases} \tag{3}$$

In the calculations, a new sparse symmetric and positive semi-definite matrix $\mathbf{M}$ is constructed based on the matrix $\mathbf{W}$ [18]. Then, the constrained minimization problem can be converted to solving eigenvalue problem of the matrix $\mathbf{M}$ as calculated by

$$\mathbf{M} = (\mathbf{I} - \mathbf{W})^{\mathrm{T}} (\mathbf{I} - \mathbf{W}) \tag{4}$$

The eigenvectors of $\mathbf{M}$ associated with the smallest d nonzero eigenvalues constitute the low embedding outputs $\mathbf{Y}$.

The mapping quality is rather sensitive to the number K of the nearest neighbors, as indicated in Eq. (1). In this paper, therefore, the residual variance, defined by $1-\rho^2_{\mathbf{D_X D_Y}}$, is employed as a quantitative measure of the embedding results when the low dimensionality d is 3, and the optimal value for K is determined by [18]:

$$K_{opt} = \arg\min \left(1 - \rho^2_{\mathbf{D_X D_Y}}\right) \tag{5}$$

where $\mathbf{D_X}$ and $\mathbf{D_Y}$ are the matrices of Euclidean distances (between pairs of points) in the input data matrix $\mathbf{X}$ and the output data matrix $\mathbf{Y}$, respectively, and ρ is the standard linear correlation coefficient of $\mathbf{D_X}$ and $\mathbf{D_Y}$.

Based on the property of LLE mapping, a method for variable selection in modeling of NIR spectra is proposed. Clearly, the relative position of the samples in high dimensional space can be kept by LLE transformation. If a variable that significantly affect the spectra is removed from the spectra, the relative position of the samples would be affected in LLE space. The change of the position in LLE space caused by removal of a variable can be used to evaluate the effect of the variable on the spectra. Therefore, the method maps the full spectra into LLE space at first, and then the mapping is repeated by removing a variable from the spectra. With the data in LLE space, the movement of the samples caused by the removal of each variable can be calculated by averaging the movement (Euclidean distance) of the samples. If the movements are ranked in a descending order, a sequence indicating the significance of the variables to the spectra can be obtained.

Only the influence of the variables on the spectra is involved in the method. For building an efficient model of a component, however, the variables specific to the component are more effective. Therefore, a forward stepwise selection (FSS) is applied to the selected variables by LLE. In the calculations, PLS models with an increasing number of the variables along the ranked sequence are evaluated with the root mean squared error of cross-validation (RMSECV). When a variable makes the RMSECV smaller, the variable is accepted, otherwise, rejected. The accepted variables are taken as the final selected variables to build the PLS model of a target component. LLE-FSS-PLS is named for the method. Additionally, Monte Carlo cross-validation (MCCV) is adopted in this study. In the calculation, half of the samples in the calibration set are randomly sampled to building the model and the remaining half is used for validation. 1000 repetitions are used for calculating the RMSECV.

3. Descriptions of the datasets

3.1. Dataset 1

The dataset was downloaded from http://software.eigenvector.com/Data/Corn/index.html, which consists of NIR spectra, measured with three spectrometers, and the moisture, oil, protein and starch values of 80 corn samples. The spectra measured on mp5 NIR spectrometer and the moisture values are used in this study. Each spectrum was recorded in the wavelength range 2498–1100 nm (4003–9091 cm^{-1}) with the digitization interval 2 nm. 56 spectra were selected, by using Kennard–Stone (KS) algorithm, as calibration set and the other 24 spectra were taken as prediction set.

3.2. Dataset 2

The dataset was downloaded from the website of international diffuse reflectance conference (IDRC), http://www.idrc-chambersburg.org/shootout2002.html. It contains the spectra of 655 pharmaceutical tablets from two spectrometers (Foss NIRSystems and Multitab Spectrometers) and the spectra of each instrument were split into a calibration set with 155 spectra, a validation set with 40 spectra and a test set with 460 spectra. Transmittance mode was used and the spectral region is from 600 to 1898 nm with 2 nm increments. The assay values of the active pharmaceutical ingredient (API) were included. In this work, the