Chemometrics and Intelligent Laboratory Systems 130 (2014) 14–19

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Support vector regression in sum space for multivariate calibration

Jiangtao Peng, Luoqing Li *

Faculty of Mathematics and Computer Science, Hubei University, Wuhan 430062, China

ARTICLE INFO

Article history:
Received 26 May 2013
Received in revised form 8 September 2013
Accepted 12 September 2013
Available online 18 September 2013

Keywords:
Support vector regression
Regularization
Sum space
Multivariate calibration
Partial least squares

ABSTRACT

In this paper, a support vector regression algorithm in the sum of reproducing kernel Hilbert spaces (SVRSS) is proposed for multivariate calibration. In SVRSS, the target regression function is represented as the sum of several single kernel decision functions, where each single kernel function with specific scale can approximate certain component of the target function. For sum spaces with two Gaussian kernels, the proposed method is compared, in terms of RMSEP, to traditional chemometric PLS calibration methods and recent promising SVR, GPR and ELM methods on a simulated data set and four real spectroscopic data sets. Experimental results demonstrate that SVR methods outperform PLS methods for spectroscopy regression problems. Moreover, SVRSS method with multi-scale kernels improves the single kernel SVR method and shows superiority over GPR and ELM methods.

1. Introduction

Multivariate calibration (MVC) is a very useful tool for extracting chemical information from spectroscopic signals by building a regression relation model between the spectra and corresponding concentrations. Traditional MVC techniques usually assume a linear spectra–concentrations relation, such as multiple linear regression (MLR), principal components regression (PCR) and partial least squares regression (PLS) [1,2]. Among these methods, PLS is most widely used in chemometrics.

PLS projects the high-dimensional predictor variables into a smaller set of uncorrelated latent variables which have a maximal covariance to the responses. It is followed by a regression step where the latent variables are used to predict the responses. PLS is especially effective in situations where the number of variables considerably exceeds the number of observations and in the presence of collinearity predictor variables [1]. However, when the data exhibits strong nonlinear behaviors, classical PLS method may not completely present the relationship between the spectra and corresponding concentrations and thus would produce large errors.

Support vector regression (SVR) method has been introduced as promising alternatives to the existing linear and nonlinear MVC approaches [3]. To describe the relation between the regressors and the dependent variables, SVR chooses a function that fits the data well in the sense of ϵ-insensitive loss cost, but is not too complex. Based on the representer theorem [4], SVR decision function can be represented as a finite linear combination of kernel products evaluated on the input samples in the training set. Thus, by choosing a nonlinear kernel, such as Gaussian kernel, SVR can easily implement nonlinear regression. SVR has exhibited good prediction performance for spectroscopy regression [3,5,6]. However, the capability of single kernel SVR is badly limited in mining abundant information from training samples. It is unsuitable to use the standard SVR to estimate complex nonlinear spectroscopy regression relations containing both the steep and smooth variations as it will either underfit the steep part or overfit the smooth part [7,8].

For spectroscopy regression problems, the regression function between the spectra and corresponding concentrations may take on non-flat characteristics as the ideal linear spectra–concentrations relation based on the Beer–Lambert law is usually corrupted by many nonlinear chemical and instrumental factors [9,10]. In this case, multi-scales kernel is more efficient than single kernel as the kernels with small and large scales can deal with the high-frequency and low-frequency components in a non-flat function, respectively.

In this paper, for multivariate calibration purpose, a support vector regression algorithm in sum space (SVRSS) is proposed. The target function of SVRSS is represented as the sum of several single kernel decision functions, where each single kernel decision function can approximate different components of the target function.

2. The algorithm

Given a training set $S = \{(x_1, y_1),\ldots,(x_n, y_n)\}$, the problem of learning is to choose a function that best approximates the supervisor's response [11]. The problem of approximating a function from finite sparse data is usually ill-posed, and classical regularization strategy

* Corresponding author. Tel.: +86 18971655636.
E-mail addresses: lilq@live.cn, humcli@gmail.com (L. Li).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.09.005

can be used to solve the problem by choosing a function that fits the training data well, but is not too complex (with small norm):

$$f = \arg\min_{f\in\mathcal{H}} \sum_{i=1}^{n} V(y_i, f(\mathrm{x}_i)) + \frac{\lambda}{2}\|f\|_K^2 \tag{1}$$

where $\|f\|_K^2$ is a norm in reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ defined by the positive definite function K, λ is the regularization parameter, and $V(\cdot,\cdot)$ is a loss function. When V is chosen as Vapnik's ϵ-insensitive loss,

$$V(y_i, f(\mathrm{x}_i)) = |y_i - f(\mathrm{x}_i)|_\epsilon = max(0, |y_i - f(\mathrm{x}_i)| - \epsilon) \tag{2}$$

the regularization problem (1) corresponds to support vector regression (SVR).

Based on the representation theorem [4], the solution of Eq. (1) is

$$f(\mathrm{x}) = \sum_{i=1}^{n} \alpha_i K(\mathrm{x}, \mathrm{x}_i). \tag{3}$$

Assume that there are m Mercer kernels $K_1,\ldots, K_m$. The function f in the sum spaces RKHSs induced by these kernels can be represented as: $f = \sum_{t=1}^{m} f_t$, $f_t \in \mathcal{H}_{K_t}$, which is the minimizer of the following optimization problem:

$$\min_{f_t\in\mathcal{H}_{K_t}} \sum_{i=1}^{n} \left| y_i - \sum_{t=1}^{m} f_t(\mathrm{x}_i) \right|_\epsilon + \frac{1}{2}\sum_{t=1}^{m} \lambda_t \|f_t\|_{K_t}^2. \tag{4}$$

According to Eq. (3), f_t can be represented as $f_t = K_t\alpha_t$, $\alpha_t = (\alpha_{t,1}, \ldots, \alpha_{t,n})^{\mathrm{T}}$. We can rewrite the optimization problem (4) using slack variables $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^*$:

$$\min_{\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\xi}^*} \sum_{i=1}^{n} (\xi_i + \xi_i^*) + \frac{1}{2}\sum_{t=1}^{m} \lambda_t \alpha_t^{\mathrm{T}} K_t \alpha_t \tag{5}$$

under the constraints

$$\begin{aligned} & y_i - \sum_{t=1}^{m} K_{t,x_i}\alpha_t \le \epsilon + \xi_i^* \\ & \sum_{t=1}^{m} K_{t,x_i}\alpha_t - y_i \le \epsilon + \xi_i \\ & \xi_i,\ \ \xi_i^* \ge 0,\ \ \ i = 1, \ldots, n. \end{aligned}$$

A Lagrange functional is constructed to solve the above problem

$$\begin{aligned} L = \sum_{i=1}^{n} (\xi_i + \xi_i^*) & - \sum_{i=1}^{n} \beta_i \left(y_i - \sum_{t=1}^{m} K_{t,x_i}\alpha_t + \epsilon + \xi_i \right) \\ & - \sum_{i=1}^{n} \beta_i^* \left(\sum_{t=1}^{m} K_{t,x_i}\alpha_t - y_i + \epsilon + \xi_i^* \right) \\ & - \sum_{i=1}^{n} (\gamma_i \xi_i + \gamma_i^* \xi_i^*) + \frac{1}{2}\sum_{t=1}^{m} \lambda_t \alpha_t^{\mathrm{T}} K_t \alpha_t. \end{aligned}$$

Minimization with respect to α_t implies

$$\hat{\alpha}_t = (\beta^* - \beta)/\lambda_t$$

where the Lagrange multipliers β^* and β can be obtained by solving the following quadratic programming (QP) problem:

$$\begin{aligned} \min_{\theta}\ & \frac{1}{2}\theta^{\mathrm{T}} H \theta - \theta^{\mathrm{T}} \mathrm{c} \\ \text{s.t.}\ & 0 \le \theta \le 1 \end{aligned} \tag{6}$$

with

$$\theta = \begin{bmatrix} \beta^* \\ \beta \end{bmatrix}, \quad \mathrm{c} = \begin{bmatrix} \epsilon - \mathrm{y} \\ \epsilon + \mathrm{y} \end{bmatrix}, \quad H = \begin{bmatrix} \sum_{t=1}^{m} K_t/\lambda_t & -\sum_{t=1}^{m} K_t/\lambda_t \\ -\sum_{t=1}^{m} K_t/\lambda_t & \sum_{t=1}^{m} K_t/\lambda_t \end{bmatrix}.$$

The decision function of SVR in sum space (SVRSS) is

$$f(\mathrm{x}) = \sum_{t=1}^{m} f_t(\mathrm{x}) = \sum_{t=1}^{m}\sum_{i=1}^{n} \hat{\alpha}_{t,i} K_t(\mathrm{x}, \mathrm{x}_i).$$

3. Experimental

To evaluate the performance of the proposed method, a simulated data set and four real spectroscopic data sets are used.

3.1. Data sets

Data set 1 consists of NIR spectra from 310 pharmaceutical tablet samples with a relative active substance content (%, w/w) in the range of 4.6–9.8% [12,13]. The transmittance spectra have 404 variables collected in the range of 7400–10507 cm^{-1}. The 310 NIR spectra are divided into 150 calibration samples, 80 validation samples and 80 prediction samples based on the SPXY (Sample set Partitioning based on joint x–y distances) algorithm [14].

Data set 2 is from the Software Shootout at the IDRC98 containing NIR spectra of 141 fescue grass powdered samples with specified carbon, nitrogen and sulphur contents ranging from 29.6% to 40.9%, 1.1% to 6.6% and 0.3% to 1.7%, respectively. The related chemical values are the average of the blind duplicates determined on a LECO CNS-2000 carbon, nitrogen and sulphur analyzer [12]. The 141 grass NIR spectra are divided into 71 calibration samples, 35 validation samples and 35 prediction samples based on the SPXY algorithm.

Data set 3 consists of 32 marzipan FTIR spectra with traditional moisture and sugar contents ranging from 7 to 19%, and 33 to 68%, respectively. The spectra in the region 6500–650 cm^{-1} have been recorded with Perkin Elmer System 2000, equipped with the horizontal ATR sampling accessory (ZnSe cell) [12,15]. The 32 marzipan IR spectra are divided into 24 calibration samples and 8 prediction samples.

Data set 4 consists of NIR transmittance spectra of meat samples [16]. The spectra have been recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850–1050 nm. For each meat sample, the data consists of a 100 channel spectrum of absorbances and the contents of moisture (water), fat and protein. The three contents, measured in percent, are determined by analytic chemistry. The data contain 129 calibration samples, 43 validation samples and 43 prediction samples. The spectra are normalized according to the standard normal variate (SNV) method.

3.2. Methods

The proposed SVRSS algorithm is compared with SVR, PLS, power PLS (PPLS) [17], Gaussian process regression (GPR) [18,19] and extreme learning machine (ELM) [20,10]. PPLS improves PLS by taking powers of correlations and standard deviations in computing the PLS loading weights, which adds flexibility to the modeling and provides better predictions [17]. A Gaussian process is a collection of random variables, any finite number of which has a joint Gaussian distribution [18]. In GPR, assume that the latent function is a Gaussian process, based on the Bayesian inference, by conditioning the joint Gaussian prior distribution on the observations, the prediction function values corresponding to test inputs can be sampled from the joint posterior distribution [18]. GPR has exhibited good performance on spectroscopic data as its flexibility in the parameterization of covariance function [19]. ELM is a new learning algorithm for the single hidden layer feedforward neural networks.