Chemometrics and Intelligent Laboratory Systems 109 (2011) 34–41

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

A two-stage regression approach for spectroscopic quantitative analysis

F. Douak [a,b,1,2], N. Benoudjit [a,1], F. Melgani [b,*]

[a] Laboratoire d'Electronique Avancée, Université de Batna, Avenue Boukhlouf Med El Hadi, 05000 Batna, Algeria
[b] Dept. of Information Engineering and Computer Science, University of Trento, 38123 Trento, Italy

ARTICLE INFO

Article history:
Received 25 February 2011
Received in revised form 21 July 2011
Accepted 28 July 2011
Available online 7 August 2011

Keywords:
Spectrometry
Residual-based correction (RBC)
Boosting
Support vector machines (SVM)
Radial basis function neural network (RBFN)
Partial least squares regression (PLSR)
Feature selection

ABSTRACT

In this paper, we propose a two-stage regression approach, which is based on the residual correction concept. Its underlying idea is to correct any given regressor by analyzing and modeling its residual errors in the input space. We report and discuss results of experiments conducted on three different datasets in infrared spectroscopy and designed in such a way to test the proposed approach by: 1) varying the kind of adopted regression method used to approximate the chemical parameter of interest. Partial least squares regression (PLSR), support vector machines (SVM) and radial basis function neural network (RBF) methods are considered; 2) adopting or not a feature selection strategy to reduce the dimension of the space where to perform the regression task. A comparative study with another approach which exploits differently estimation errors, namely adaptive boosting for regression (AdaBoost.R), is also included. The obtained results point out that the residual-based correction approach (RBC) can improve the accuracy of the estimation process. Not all the improvements are statistically significant but, at the same time, no case of accuracy decrease has been observed.

1. Introduction

Near infrared spectroscopy (NIR) is widely used in food and pharmaceutical industries for analysis and quality control. In reflection-based NIR spectroscopy, NIR radiation is guided into the product, and some of the backscattered radiation is captured and related to variables of interest via chemometric techniques. As the backscattered radiation spectrum is affected by both the scattering and absorption properties of the product, it provides information about its physical structure as well as its chemical composition. From the NIR spectrum, quantitative and qualitative information can thus be obtained with regression and classification models, respectively [1–3].

Viewed from a statistical or data analysis perspective, the main difficulty in quantitative information extraction is to cope with the colinearity between spectral variables and the large number of variables to deal with (curse of dimensionality). Indeed, not only consecutive variables in a spectrum are highly correlated by nature, but also real applications usually concern databases with a small number of known spectra and a high number of spectral variables.

The regression problem built on the original spectral variables is thus very likely to be hard to handle. In these conditions, the most natural solution consists in reducing data dimensionality. Different methods exist in the literature, which can be distinguished either as feature selection methods or projection techniques, e.g., feature selection, mutual information, forward–backward, B-splines and genetic algorithm [4–8].

Concerning the regression problem, typically, the choice of the regression algorithm depends on the statistical distribution of the data under study and the related noise, which have a direct impact on its prediction performance [1,4]. In this context, among the linear regression methods, one can find the multiple linear regression (MLR), the principal component regression (PCR) and the partial least squares regression (PLSR) methods. The MLR is a simpler approach for calibration model creation than PCR and PLS, because it performs regression directly on the original variables while PCR carries out regression on latent variables that do not necessarily have a physical meaning and PLSR finds a projection subspace by exploiting the target variable. However, due to the colinearity between original spectral variables, overfitting problems can be encountered in MLR [8]. PCR first consists of applying a principal components analysis (PCA) to the matrix of the spectral data. Then, PCA replaces the original spectral variables (typically redundant) by principal components (linear combinations of the original variables), which contain most of the conveyed information and have the advantage of being uncorrelated [9]. The most important principal components are then used as inputs for a multiple linear regression (MLR) [10]. PLSR aims at finding linear projections which exhibit the maximum correlation with the target (output) variable; a linear regression model is then estimated in the subspace defined by the projected coordinates [11].

Linear regression has the advantage of being simple and cheap in terms of computation load, but is not reliable if the true relationship between the inputs and the output is nonlinear, unless opportune

* Corresponding author. Tel.: +39 0461 28 1573; fax: +39 0461 28 2093.
E-mail addresses: fouzi.douak@disi.unitn.it (F. Douak), nbenoudjit@gmail.com (N. Benoudjit), melgani@disi.unitn.it (F. Melgani).
[1] Tel./fax: +213 33 80 54 94.
[2] Tel.: +39 0461 28 1573; fax: +39 0461 28 2093.

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.07.007