Research Article

Journal of CHEMOMETRICS

Received: 11 July 2011, Revised: 23 February 2012, Accepted: 29 February 2012, Published online in Wiley Online Library: 19 April 2012

(wileyonlinelibrary.com) DOI: 10.1002/cem.2443

Active learning for spectroscopic data regression

Fouzi Douak[a,b], Farid Melgani[a]*, Naif Alajlan[c], Edoardo Pasolli[a], Yakoub Bazi[c] and Nabil Benoudjit[b]

In this work, we introduce an active learning approach for the estimation of chemical concentrations from spectroscopic data. Its main objective is to opportunely collect training samples in such a way as to minimize the error of the regression process while minimizing the number of training samples used, and thus to reduce the costs related to training sample collection. In particular, we propose two different active learning strategies developed for regression approaches based on partial least squares regression, ridge regression, kernel ridge regression, and support vector regression. The first strategy uses a pool of regressors in order to select the samples with the greatest disagreements among the different regressors of the pool, while the second one is based on adding samples that are distant from the current training samples in the feature space. For support vector regression, a specific strategy based on the selection of the samples distant from the support vectors is proposed. Experimental results on three different real data sets are reported and discussed. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: active learning; chemical component concentration estimation; partial least squares regression (PLSR); ridge regression (RR); kernel ridge regression (KRR); support vector regression (SVR); spectroscopy

1. INTRODUCTION

Spectroscopy is an important technology for product analysis and quality control in different chemical fields. For example, it has been applied successfully in the pharmaceutical [1], [2], food [3], and textile industries [4]. Chemical analysis by spectroscopy relies on the fast acquisition of a large number of spectral data, which can be analyzed in order to yield accurate estimations of the concentration of the chemical component of interest in a given product.

From a methodological point of view, the problem of concentration estimation can be viewed as an inverse modeling issue in which it is necessary to define a model that relates the acquired observations to the concentration of interest. Typically, the model is estimated by adopting supervised regression techniques, which require the availability of a set of training samples. By training samples, we mean pairs of spectral data acquired by the spectrometer and measurements of the concentration to be estimated. In the literature, two main approaches of regression have been proposed. The first is based on linear models, appreciated for their simplicity, such as multiple linear regression, principal component regression, ridge regression (RR), and partial least squares regression (PLSR) [5]. The second approach makes use of nonlinear models. They are characterized by greater computational complexity, but they can give better performances when a strong nonlinearity between the acquired spectral data and the concentrations to be estimated is present. In this context, two state-of-the-art methods are radial basis functions neural network and support vector regression (SVR) [6], [7].

In the aforementioned works, the regression process is undertaken by assuming that the training set is composed of a sufficient number of samples in order to obtain reliable and accurate estimations. However, from a practical point of view, the process of collecting training samples is not trivial, because the concentration measurements associated with the spectral data have to be performed manually by human experts and thus are subject to errors and costs in terms of time and money. For this reason, the number of available training samples is typically limited and performances can be consequently affected owing to data scarcity. A solution to this problem is given by semi-supervised approaches [8], in which the unlabeled samples are exploited during the design of the regression model in order to compensate for the deficit in labeled samples. By unlabeled samples, we mean samples whose spectral values are known, but for which the corresponding concentration values are unknown. Such samples exhibit the advantage that they are available at zero cost from the data under analysis. In the chemometrics literature, a few regression works have been proposed in this context [9], [10].

In the data classification context, another solution to the problem of training sample collection is given by the active learning approach [11]. Starting from a small training set, additional samples are selected from a large amount of unlabeled data. These samples are labeled by the expert and added to the training set. The process is iterated until a stop criterion is reached. Active learning strategies have been applied successfully in different fields in classification [11–15].

* *Correspondence to: Farid Melgani, Department of Information Engineering and Computer Science, University of Trento, Via Sommarive, 14, I-38123, Trento, Italy.*
E-mail: melgani@disi.unitn.it

a *F. Douak, F. Melgani, E. Pasolli*
Department of Information Engineering and Computer Science, University of Trento, Via Sommarive, 14, I-38123, Trento, Italy

b *F. Douak, N. Benoudjit*
Laboratoire d'Electronique Avancée, Université de Batna, Avenue Boukhlouf Med El Hadi, 05000 Batna, Algeria

c *N. Alajlan, Y. Bazi*
ALISR Laboratory, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia