Functional Dimension Reduction for Chemometrics

Tuomas Kärnä
Helsinki University of Technology, Finland

Amaury Lendasse
Helsinki University of Technology, Finland

INTRODUCTION

High dimensional data are becoming more and more common in data analysis. This is especially true in fields that are related to spectrometric data, such as chemometrics. Due to development of more accurate spectrometers one can obtain spectra of thousands of data points. Such a high dimensional data are problematic in machine learning due to increased computational time and the curse of dimensionality (Haykin, 1999; Verleysen & François, 2005; Bengio, Delalleau, & Le Roux, 2006).

It is therefore advisable to reduce the dimensionality of the data. In the case of chemometrics, the spectra are usually rather smooth and low on noise, so function fitting is a convenient tool for ***dimensionality reduction***. The fitting is obtained by fixing a set of basis functions and computing the fitting weights according to the least squares error criterion.

This article describes a unsupervised method for finding a good function basis that is specifically built to suit the data set at hand. The basis consists of a set of Gaussian functions that are optimized for an accurate fitting. The obtained weights are further scaled using a Delta Test (DT) to improve the prediction performance. Least Squares Support Vector Machine (LS-SVM) model is used for estimation.

BACKGROUND

The approach where multivariate data are treated as functions instead of traditional discrete vectors is called Functional Data Analysis (FDA) (Ramsay & Silverman, 1997). A crucial part of FDA is the choice of basis functions which allows the functional representation. Commonly used bases are B-splines (Alsberg & Kvalheim, 1993), Fourier series or wavelets (Shao, Leung, & Chau, 2003). However, it is appealing to build a problem-specific basis that employs the statistical properties of the data at hand.

In literature, there are examples of finding the optimal set of basis functions that minimize the fitting error, such as Functional Principal Component Analysis (Ramsay et al., 1997). The basis functions obtained by Functional PCA usually have global support (i.e. they are non-zero throughout the data interval). Thus these functions are not good for encoding spatial information of the data. The spatial information, however, may play a major role in many fields, such as spectroscopy. For example, often the measured spectra contain spikes at certain wavelengths that correspond to certain substances in the sample. Therefore these areas are bound to be relevant for estimating the quantity of these substances.

We propose that locally supported functions, such as Gaussian functions, can be used to encode this sort of spatial information. In addition, variable selection can be used to select the relevant functions from the irrelevant ones. Selecting important variables directly on the raw data is often difficult due to high dimensionality of data; computational cost of variable selection methods, such as Forward-Backward Selection (Benoudjit, Cools, Meurens, & Verleysen, 2004; Rossi, Lendasse, François, Wertz, & Verleysen, 2006), grows exponentially with the number of variables. Therefore, wisely placed Gaussian functions are proposed as a tool for encoding spatial information while reducing data dimensionality so that other more powerful information processing tools become feasible. Delta Test (DT) (Jones, 2004) based scaling of variables is suggested for improving the prediction performance.

A typical problem in chemometrics deals with predicting some chemical quantity directly from measured spectrum. Due to additivity of absorption spectra, the problem is assumed to be linear and therefore linear

models, such as Partial Least Squares (Härdle, Liang, & Gao, 2000) have been widely used for the prediction task. However, it has been shown that the additivity assumption is not always true and environmental conditions may further introduce more non-linearity to the data (Wülfert, Kok, & Smilde, 1998). We therefore propose that in order to address a general prediction problem, a non-linear method should be used. LS-SVM is a relatively fast and reliable non-linear model which has been applied to chemometrics as well (Chauchard, Cogdill, Roussel, Roger, & Bellon-Maurel, 2004).

USING GAUSSIAN BASIS WITH SPECTOMETRIC DATA

Consider a problem where the goal is to estimate a certain quantity $p \in \Re$ from a measured absorption spectrum X based on the set of N training examples $(X_j, p_j)_{j=1}^{N}$. In practice, the ***spectrometric data*** X_j is a set of discretized measurements $(x_i^j, y_i^j)_{i=1}^{m}$ where $x_i^j \in [a,b] \subset \Re$ stand for the observation wavelength and $y_i^j \in \Re$ is the response.

Adopting the FDA framework (Ramsay et al., 1997), our goal is to build a prediction model F so that $\hat{p} = \mathrm{F}(X)$. Here, the argument X is a real-world spectrum, i.e. a continuous function that maps wavelengths to responses. Without much loss of generality it can be assumed that X belongs to $L_2([a, b])$, the space of square integrable functions on the interval [a,b]. However, since the spectrum X is unknown and infinite dimensional it is impossible to build the model F(X) in practice. Therefore X must be approximated with a q dimensional representation $\omega = \mathrm{P}(X), \mathrm{P} : L_2 \to \Re^q$, and our prediction model becomes $\hat{p} = \mathrm{F}(\omega)$. Naturally, in order to obtain ***dimensionality reduction***, we require that q is smaller than the number of points in the spectra.

Figure 1 presents a graph of the overall prediction method. Gaussian fitting is used for the approximation of X. The obtained vectors $\boldsymbol{\omega}$ are further scaled by a diagonal matrix A before the final LS-SVM modeling. The following sections explain these steps in greater detail.

Gaussian Fitting: Approximating Spectral Function *X*

Because the space $L_2([a, b])$ is infinite dimensional function space, it is necessary to consider some finite dimensional subspace $V \subset L_2([a, b])$ in order to obtain a feasible ***function approximation***. We define V by a set of ***Gaussian functions***

$$\varphi_k(x) = e^{-\|x-t_k\|^2/\sigma_k^2},\ k = 1,\ldots,q, \qquad (1)$$

where t_k is the center and σ_k is the width parameter. The set $\varphi_k(x)$ spans a q dimensional normed vector space and we can write $V = \mathrm{span}\{\varphi_k(x)\}$. A natural choice for the norm is the L_2 norm:

$$\left\|\hat{f}\right\|_V = \left(\int_a^b \hat{f}(x)^2\,dx\right)^{1/2}.$$

Now X can be approximated using the basis representation $\hat{X}(x) = \omega^T \phi(x)$, where

$$\phi(x) = [\varphi_1(x), \varphi_2(x), \ldots, \varphi_q(x)]^T.$$

The weights $\boldsymbol{\omega}$ are chosen to minimize the square error:

Figure 1. Outline of the prediction method