Analytica Chimica Acta 1009 (2018) 27–38

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Nearest clusters based partial least squares discriminant analysis for the classification of spectral data

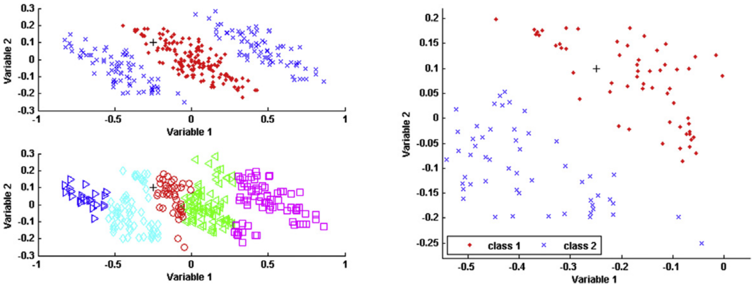
Weiran Song [a, *], Hui Wang [a], Paul Maguire [b], Omar Nibouche [a]

[a] School of Computing and Mathematics, Ulster University, BT37 0QB, Newtownabbey, Co. Antrim, UK
[b] School of Engineering, Ulster University, BT37 0QB, Newtownabbey, Co. Antrim, UK

HIGHLIGHTS

- Nearest cluster based PLS-DA for multimodal and nonlinear classification.
- Simple and distinctive structure in local space which is approximately linear separable.
- Better results achieved on 12 UCI data sets and 5 spectral data sets.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 19 July 2017
Received in revised form
18 December 2017
Accepted 15 January 2018

Keywords:
Partial least squares
Clustering
Nonlinearity
Multimodality
Spectral pattern recognition

ABSTRACT

Partial Least Squares Discriminant Analysis (PLS-DA) is one of the most effective multivariate analysis methods for spectral data analysis, which extracts latent variables and uses them to predict responses. In particular, it is an effective method for handling high-dimensional and collinear spectral data. However, PLS-DA does not explicitly address data multimodality, i.e., within-class multimodal distribution of data. In this paper, we present a novel method termed *nearest clusters based PLS-DA* (NCPLS-DA) for addressing the multimodality and nonlinearity issues explicitly and improving the performance of PLS-DA on spectral data classification. The new method applies hierarchical clustering to divide samples into clusters and calculates the corresponding centre of every cluster. For a given query point, only clusters whose centres are nearest to such a query point are used for PLS-DA. Such a method can provide a simple and effective tool for separating multimodal and nonlinear classes into clusters which are locally linear and unimodal. Experimental results on 17 datasets, including 12 UCI and 5 spectral datasets, show that NCPLS-DA can outperform 4 baseline methods, namely, PLS-DA, kernel PLS-DA, local PLS-DA and k-NN, achieving the highest classification accuracy most of the time.

1. Introduction

Spectral data analysis is used in many areas of science and engineering as a mean of exploring the constituents of matter. Absorption spectroscopy is a type of spectral analysis usually used to identify and possibly quantify particular substances in a sample. Infrared (IR) spectroscopy and ultraviolet-visible (UV-Vis) spectroscopy are two specific examples. Measurements of radiation intensity at a series of fixed wavelengths result in a spectrum that consists of a series of discrete peaks. Other types include transmission, reflectance or emission spectroscopies which, along with

* Corresponding author.
E-mail address: song-w@email.ulster.ac.uk (W. Song).

https://doi.org/10.1016/j.aca.2018.01.023

mass spectroscopy, can provide useful information on the chemical constituents of substance. Spectral data contains a molecular fingerprint of the substance of interest, which can be used to identify and/or quantify the substance. The interpretation of the fingerprint depends critically on instrumental factors. Recently, there has been an ongoing drive towards miniaturisation and field portability of such instrumentation and the development of low cost spectral-based sensors. This may significantly degrade the fingerprint data quality, thus making accurate identification problematic; in portable applications, the nature of the sample and its environment may add to the challenge. Therefore, the use of pattern recognition techniques in spectral data analysis is becoming common practice. However, the challenges in robust extraction of meaningful fingerprint data from noisy field data cannot be underestimated. These include, for example, high dimensionality [1], collinearity [2], nonlinearity [3] and a special type of nonlinearity - multimodality [4,5].

Among the common methods for spectral data analysis are Principal Component Analysis (PCA), Support Vector Machine (SVM) and Partial Least Squares (PLS). PLS is currently the de-facto standard [6]. It is a statistical regression method that combines the features of Canonical Correlation Analysis (CCA) and Multiple Linear Regression (MLR) to predict responses based on independent variables. It searches for linear combinations of independent variables, namely *latent variables* (LV), that maximize the covariance between the latent variable and the response. PLS has proven to be a very useful method for spectral data analysis. It efficiently handles the high dimensionality and collinearity problems that widely exist in spectral data [7,8] by stably estimating regression coefficients from low-dimensional latent variables. However, it has been reported that the PLS algorithm will degrade in performance under nonlinear conditions [9–11], which is often present in spectral data for various reasons [3] and can be identified by a quantitative numerical tool (e.g. run test) with augmented partial residual plots (APaRP) [12,13]. Recent attempts to modify the PLS algorithm to handle nonlinear data have focused typically on two approaches. The first is kernel approaches which transforms the original input data into a feature space by nonlinear mapping, and then constructs a linear PLS model in the feature space [14]. The second approach is to combine locally weighted regression (LWR) [15] and PLS, namely, locally weighted PLS (LW-PLS) [16]. On one hand, this approach fills the gap that LWR cannot be used to handle the problem of ill-conditioned matrices, such as small sample size and collinearity, unless a robust variable selection is implemented. On the other hand, LW-PLS constructs a local model to enlarge the contribution of neighbouring data for a given query. As a result, the global nonlinearity can be lessened.

A particular type of nonlinearity is multimodality where the data distribution within a class is multimodal possibly due to the fact that data within a class comes from different sources or different data collection sessions [11,17,18]. For example, if we want to identify apples from other types of fruit, the apple as a class is very likely to have multiple modes in the data distribution each corresponding to a variety, since apple varieties may be quite distinct. Further still, differences within the same variety of apple from different regions may also lead to multiple modes in the data distribution. In general, it is possible that data instances within a class are more similar to data instances of a different class than to other members of its own class. Multimodality has been studied in pattern recognition; it has been shown [19,20] that modelling multimodality explicitly can significantly improve classification performance. However, multimodality has not been explicitly addressed in PLS although it has been implicitly addressed in variants of PLS. Kernel PLS-DA has been studied for analysing nonlinear chemometric data and it has been shown [21,22] to have a classification performance comparable on average to kernel Support Vector Machine (KSVM) for nonlinear chemometric data. Kernel PLS-DA uses the similarity between two data vectors as the basis to map the original data into feature space. It is not directly possibly to see the contribution of each variable with respect to the final prediction as well as to interpret the obtained kernel PLS model [14,23]. Moreover, kernel PLS-DA selects more LVs than PLS-DA on the classification of spectral data [4]. The locally weighted PLS-DA (LW-PLS-DA) has been studied for analysing nonlinear spectral data [4,24]. LW-PLS-DA can outperform standard and kernel PLS-DA [4], also the resulting model does not require all training samples to be involved in. Common weighting matrices of LW-PLS are based on the Euclidean distance or the Mahalanobis distance, which perform less well than covariance or sparse regression coefficient for many industrial processes [25–27]. Moreover, this instance-based learning approach produces multiple models for a set of queries which results in a sharp increase in computational complexity compared to the classical PLS.

This paper presents an extension of PLS-DA that explicitly addresses data nonlinearity and multimodal distributions, namely NCPLS-DA. By using hierarchical clustering, all training samples are grouped into clusters in which the clustering centres are calculated. Clusters that contain the nearest centres towards a given query are selected for PLS modelling. This strategy handles nonlinearity and multimodality by constructing linearly separable models in neighbourhoods. Thus, more accurate results can be expected. This PLS extension has been tested on a wide range of datasets including twelve UCI datasets of different data types and five spectral datasets (some being simulated with multimodality and nonlinearity, and some being publicly available).

The remainder of the paper is organized as follows: Section 2 briefly reviews PLS-DA and hierarchical clustering. The proposed method, NCPLS-DA, is presented in Section 3. Section 4 presents the experiments on UCI and spectral datasets, including datasets description, parameter settings, results and discussion. Conclusions are drawn in Section 5.

2. Related work

The standard chemometrics notation is used in this paper. Capital and lowercase letters in boldface denote matrix and vector, respectively. Table 1 lists the symbols used in this paper.

2.1. PLS-DA

PLS is a classical method in multivariate analysis that maximizes the covariance between the latent variables and the responses. It is today most widely used in chemometrics including spectral data analysis. There exist different PLS algorithms, including Nonlinear Iterative Partial Least Squares (NIPALS) [28] and SIMPLS [29], which are different in computational complexity and numerical stability. The computation time is dependent mainly on the dimensionality of data and also the number of latent variables selected, to a lesser extent [30]. The numerical stability is dependent on the numerical calculation methods used and is a factor of model precision. Theoretically, all PLS algorithms should yield the same models but in practice there are differences due to the numerical calculation methods [30]. In this paper, the SIMPLS algorithm is used, because it is faster than NIPALS and it is stable when the number of latent variables is not high [30,31].

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are mean-centred, the SIMPLS algorithm is to find a linear combination of $\boldsymbol{X}$, $\boldsymbol{t} = \boldsymbol{Xw}$, that maximizes data covariance as follows,