Talanta 86 (2011) 316–323

Contents lists available at SciVerse ScienceDirect

Talanta

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

^{1}H NMR variable selection approaches for classification. A case study: The determination of adulterated foodstuffs

Carolina V. Di Anibal, M. Pilar Callao, Itziar Ruisánchez *

Department of Analytical and Organic Chemistry, Rovira i Virgili University, Marcel·lí Domingo s/n, 43007, Tarragona, Spain

ARTICLE INFO

Article history:
Received 16 June 2011
Received in revised form 5 September 2011
Accepted 12 September 2011
Available online 16 September 2011

Keywords:
Variable selection
Nuclear magnetic resonance
Food adulteration
Sudan dyes
PLS-DA

ABSTRACT

Whenever dealing with large amount of data as is the case of a NMR spectrum, carrying out a variable selection before applying a multivariate technique is necessary. This work applies various variable selection techniques to extract relevant information from ^{1}H NMR spectral data. Three approaches have been chosen, because each is based on very different foundations. The first method, called Xdiff, is based on calculating the normalized differences between the mean spectrum of a class considered to be the reference and the spectra of each sample. The second approach is the interval Partial Least Squares method (iPLS), which investigates the influential zones of the spectra that contains the most discriminating predictors calculating local PLS-DA models on narrow intervals. The last one is Genetic Algorithms (GAs) which finds the optimal variables from a random initial subset of variables by means of an iterative process. The performance of each variable selection strategy is determined by the classification results obtained when multiclass Partial Least Squares-Discriminant Analysis is applied. This study has been applied to NMR spectra of culinary spices that might be adulterated with banned dyes such as Sudan dyes (I–IV). The three techniques give neither the same number nor the same selected variables, but they do select a common zone from the spectra containing the most discriminating variables. All three techniques give satisfactory classification and prediction results, being higher than 95% with iPLS and GA and around 89% with Xdiff, therefore the three variable selection techniques are suitable to be used with NMR data in the determination of food adulteration with Sudan dyes as well as the specific type of adulterant used (I–IV).

1. Introduction

In food industry, the addition of some colorants to some products is a common practice, as colorants enhance its visual aesthetics and promote sales. Up to now four Sudan (I–IV) dyes have been detected in certain food products, as culinary spices destined to human consumption, although they are normally used for coloring plastics and other synthetic materials. The current European legal framework on colors in food establishes that Sudan dyes are not included in the list of authorized colorants, as these dyes have potential carcinogenic effects [1] and even more, Sudan I may also have genotoxic effects. Therefore Sudan dyes are banned to be used as additive in food matrices for human consumption.

Previous studies have demonstrated that Proton Nuclear Magnetic Resonance Spectroscopy (^{1}H NMR) is a well suited analytical technique capable of detecting these four Sudan dyes when they are as adulterants in culinary spices [2]. NMR is a technique that generates a specific profile of the sample studied. Recent breakthroughs in NMR technology have led to measurements with increased sensitivity, resolution and reproducibility, thereby contributing to the production of high quality data [3]. These improvements in data quality, coupled with multivariate techniques, have given rise to a well-known rapid screening method [4] which has been demonstrated to be an efficient method for food screening, discrimination and characterization [2,5–7].

Because of the large amount of data obtained from a ^{1}H NMR spectrum, carrying out a variable selection before applying a multivariate technique is common practice. There are many potential benefits of variable selection: facilitating data visualization and data understanding, reducing the variables/samples ratio, eliminating noisy variables as well as redundant information, among others. All these advantages are important when chemometrics methods want to be applied. Many methods are found in the bibliography for variable selection in NMR in classification problems. Some methods include reducing noisy variables or variables of low intensity or even bucketing, with the consequent reduction of data. Other methods are supervised with the aim to find which variables are the most discriminatory in order to achieve the best discrimination when working with different groups of samples or classes.

* Corresponding author. Tel.: +34 977558490; fax: +34 977558446.
E-mail address: itziar.ruisanchez@urv.cat (I. Ruisánchez).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.09.019

Some examples of both mentioned types found in literature include stepwise discriminant analysis [8,9], supervised variable selection methods [10], self organizing maps (SOMs) [11], PLS weight coefficients [12], wavelet transform [13], univariate selection based on the maximum intensity differences [14], interval Partial Least Squares (iPLS) [15,16] and Genetic Algorithms (GAs) [17,18]. This wide variety of variable selection techniques implies that choosing the most appropriate one for a specific problem is not an easy task.

The aim of this work is to study the ability of three supervised techniques for selecting variables when working with NMR spectral data in the Sudan dyes classification problem, as it is well known that choosing the optimal variable selection technique is problem dependent. The three approaches studied have been chosen because each is based on a very different principle. Our work focuses on the region and number of selected variables, the number and type of misclassified samples and the overall performance of the classification process obtained with each technique.

The first technique is based on the computation of new variables called Xdiff, which are the normalized differences between the mean spectrum of a class considered to be the reference and the spectra of each sample. The hypothesis is that variables having different intensities will be reinforced in the new x_{diff} values, thus enabling differentiation between the classes. The second approach is the interval Partial Least Squares method (iPLS) [19], which investigates the influential zones of the spectra that contains the most discriminating predictors, and calculates local PLS-DA models in narrow intervals. The third variable selection technique used is the well known Genetic Algorithms (GAs) [20,21], which can find the optimal variables from a random initial subset of variables by means of an iterative process. The performance of each variable selection strategy is determined from the classification results obtained when Partial Least Squares-Discriminant Analysis is applied.

2. Data analysis methods

2.1. Variable selection

2.1.1. Xdiff method

This variable selection method [2,14] is applied to a multiclass problem and it is based on calculating the x_{diff} values in accordance with Eq. (1)

$$x_{diff,ij} = \frac{\left|x_{ij} - \overline{x_i}\right|}{\sigma_i} \tag{1}$$

where x_{ij} is the ith variable for the jth sample and $\overline{x_i}$ and σ_i are the mean and standard deviation, respectively, calculated from each ith variable obtained from a reference class. As we are dealing with an adulteration problem, among our predefined five classes, we have set the unadulterated class as the reference one.

The **Xdiff** matrix is calculated for all five classes. A threshold value was defined from the x_{diff} values of the reference class through a visual inspection in a way that most of the x_{diff} values are kept below. Therefore, only those original variables that correspond to x_{diff} values higher than the prefixed threshold are selected. Several threshold values around the first prefixed one are checked, retaining the one which gives the best PLS-DA classification results.

2.1.2. Interval Partial Least Squares method

Interval PLS (iPLS) develops local PLS-DA models on equidistant subintervals of the full-spectrum region and the prediction performance of these local models and the global (full-spectrum) model is compared, mainly by means of the validation parameter RMSECV (root mean squared error of cross-validation, Eq. (2)):

$$\mathrm{RMSECV} = \frac{\sqrt{\left(\sum \hat{y}_i - y_i\right)^2}}{n} \quad i = 1, \ldots, n \tag{2}$$

with y_i as the true class assignation value for sample i, $\hat{y}_i$ as the predicted class assignation value from cross-validation and n as the number of samples.iPLS provides an overall picture of the relevant information in different spectral subdivisions, thereby removing non-relevant information from other regions.

2.1.3. Genetic Algorithms (GAs) method

The GA theory is explained in detail elsewhere [20,21], so we will limit ourselves to show the procedure followed in the present study depicted in Fig. 1 which involves an iterative process. As for running the GA algorithm, the adequate number of input variables has to be not far away from 200, the original NMR variables have to be reduced, so the average of "n" consecutive variables is obtained. To decide the optimal window size n, the Principal Components Analysis (PCA) score plots of both, the original and the mean of n variables, are compared until a similar distribution of samples is kept. The next step is to apply the GA algorithm to the mean variables, with a previous step in which a randomization test is applied to check whether the dataset is adequate to run the algorithm, in order to avoid the overfitting problem commonly found in a GA-based feature selection [22]. Therefore, the mean variables selected by GA are expanded to the n consecutive original ones used to obtain the mean variables. Finally, PLS-DA is applied to those original ones.

All this procedure is done iteratively until a subset of variables giving the optimal classification results comparing the previous and last iteration is found. It has to be remarked that the second (and so on) iterations start with the previous subset of "expanded original variables", so a lower n value is looked for.

2.2. Classification method: Partial Least Squares-Discriminant Analysis

PLS-DA is a regression technique adapted to a supervised classification task [23]. A PLS regression model is calculated, which relates the independent variables (e.g. spectra) to a binary "y" vector which has as many values as classes in order to designate the class of the sample. For example, a vector [1,0,0,0,0] means that of five possible classes, the sample belongs to class 1, and so on. Classification of an unknown sample is derived from the value predicted by the PLS model, $\hat{y}$. Ideally, this value should be close to the values used to codify the class (here either 0 or 1). A threshold value for each pre-defined class is defined between 0 and 1 so that a sample is assigned to the class for which its prediction is larger than the threshold value. Typically, a normal distribution fits $\hat{y}$ values and the threshold value is estimated using Bayes' rule [24]. The optimal number of latent variables (LVs) was chosen to minimize the root mean square-cross validation prediction error (RMSECV) for all the classes. This number is selected through a compromise between the optimal values for each class.

Both, class assignment and the percentage of predicted probability are considered for the evaluation of the multiclass PLS-DA classification results. In this study, the selected variables are autoscaled before running the multiclass PLS-DA algorithm.

2.3. Training and test set

In order to avoid overoptimistic results, the dataset is divided into training and test set, using the training set to select the most