Talanta 79 (2009) 1260–1264

Contents lists available at ScienceDirect

Talanta

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

Near infrared reflectance spectrometry classification of cigarettes using the successive projections algorithm for variable selection

Edilene Dantas Teles Moreira[a], Márcio José Coelho Pontes[a], Roberto Kawakami Harrop Galvão[b], Mário César Ugulino Araújo[a,*]

[a] Universidade Federal da Paraíba, Departamento de Química, Laboratório de Automação e Instrumentação em Química Analítica/Quimiometria (LAQA), Caixa Postal 5093, CEP 58051-970 – João Pessoa, PB, Brazil
[b] Instituto Tecnológico de Aeronáutica, Divisão de Engenharia Eletrônica, São José dos Campos, SP, Brazil

ARTICLE INFO

Article history:
Received 21 March 2009
Received in revised form 15 May 2009
Accepted 18 May 2009
Available online 27 May 2009

Keywords:
Cigarettes
Near infrared reflectance spectroscopy
Classification
Successive projections algorithm
Linear discriminant analysis

ABSTRACT

This paper proposes a methodology for cigarette classification employing Near Infrared Reflectance spectrometry and variable selection. For this purpose, the Successive Projections Algorithm (SPA) is employed to choose an appropriate subset of wavenumbers for a Linear Discriminant Analysis (LDA) model. The proposed methodology is applied to a set of 210 cigarettes of four different brands. For comparison, Soft Independent Modelling of Class Analogy (SIMCA) is also employed for full-spectrum classification. The resulting SPA–LDA model successfully classified all test samples with respect to their brands using only two wavenumbers (5058 and 4903 cm^{-1}). In contrast, the SIMCA models were not able to achieve 100% of classification accuracy, regardless of the significance level adopted for the *F*-test. The results obtained in this investigation suggest that the proposed methodology is a promising alternative for assessment of cigarette authenticity.

1. Introduction

Cigarette authenticity is an important matter, which involves economic aspects and consumer health issues. In fact, cigarette brands may differ in retail price, as well as in the levels of potentially hazardous substances such as nicotine and tar [1,2]. Therefore, the assessment of compliance with the cigarette label and the identification of counterfeit products are analytical problems that merit investigation.

The discrimination of cigarette types is usually carried out on the basis of visual aspect, flavour and aroma. However, such an inspection is subjective and may lead to unreliable results. As an alternative, instrumental techniques have been employed to obtain a more objective and accurate assessment of cigarette samples. Examples include gas chromatography (GC) and liquid chromatography (LC) [3–5], inductively coupled plasma mass spectrometry (ICP-MS) [6], inductively coupled plasma optical emission spectrometry (ICP-OES) [7,8], nuclear magnetic resonance (NMR) [9] and pyrolysis single-photon ionisation time-of-flight mass spectrometry (Py-SPI-TOFMS) [10]. However, these techniques are laborious and time-consuming, require harmful reagents and involve expensive equipment with high operation and/or maintenance costs. An interesting alternative to overcome such drawbacks would be the use of near infrared (NIR) spectroscopy, a technique that enables practical, fast and less dispendious analyses.

NIR spectroscopy has been successfully applied to discrimination and/or classification of various materials, including alcoholic beverages [11,12], food products [12–15], fuel samples [16–18], polymers [19] and agricultural goods [20], among others [21,22]. However, only a single paper [23] has been published on the use of NIR spectroscopy for cigarette discrimination. In that work, 142 cigarettes of two different brands were distinguished by using the Adaboost algorithm and Linear Discriminant Analysis (LDA) applied to near infrared reflectance (NIRR) measurements. Feature extraction was performed by principal component analysis (PCA) or Kernel Principal Component Analysis (KPCA).

The present paper proposes an analytical methodology for cigarette classification based on the use of NIRR spectroscopy and variable selection. For this purpose, the Successive Projections Algorithm (SPA) [17] is employed to choose an appropriate subset of wavenumbers for a Linear Discriminant Analysis (LDA) model. Recently, SPA–LDA has been successfully applied to the classification of edible vegetable oils and soil samples by using square wave voltammetry (SWV) [24] and laser-induced breakdown spectroscopy (LIBS) [25], respectively. In comparison with the approach adopted in [23], SPA–LDA provides a simpler model in the sense that the classification variables correspond to actual reflectance measurements, rather than PCA/KPCA scores.

* Corresponding author. Tel.: +55 83 3216 7438; fax: +55 83 3216 7437.
E-mail address: laqa@quimica.ufpb.br (M.C.U. Araújo).

0039-9140/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2009.05.031

The proposed methodology is applied to a set of 210 cigarettes comprising four brands of different chemical composition and retail price. For comparison, Soft Independent Modelling of Class Analogy (SIMCA) [26] is also employed. SIMCA is a well-known method for full-spectrum classification, which has been widely employed in applications involving NIR data [27–30].

2. Background

2.1. Notation

Matrices will be represented by bold capital letters, column vectors by bold lowercase letters, and scalars by italic characters. The matrix of instrumental responses will be denoted by **X**. The *n*th object in matrix **X** will be denoted by $\mathbf{x}_n$ (that is, $\boldsymbol{x}_n^T$ will correspond to the *n*th row of matrix **X**). The *k*th column of matrix **X** will be denoted by $\mathbf{x}^k$.

2.2. Linear Discriminant Analysis

The LDA classification method employs the Mahalanobis distance [31,32], which can be defined as follows. Let $x = [x_1, x_2, \ldots x_d]^T$ be an object that must be assigned to one out of *c* possible classes. In the case of NIRR data, the classification variables $x_1, x_2, \ldots x_d$ correspond to reflectance measurements acquired at *d* wavenumbers. The squared Mahalanobis distance $r^2(\mathbf{x}, \boldsymbol{\mu}_j)$ between **x** and the center of the *j*th class ($j = 1, 2, \ldots, c$) is defined as

$$r^2(\boldsymbol{x}, \boldsymbol{\mu}_j) = (\boldsymbol{x} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_j) \quad (1)$$

where $\boldsymbol{\mu}_j$ ($d \times 1$) and $\boldsymbol{\Sigma}_j$ ($d \times d$) are the mean vector and covariance matrix for the class under consideration [32]. If the true mean and covariance values for the population are unknown (which is usually the case), maximum likelihood estimates $\mathbf{m}_j$ and $\mathbf{S}_j$ may be employed in place of $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$, respectively. These estimates can be obtained from a finite set of training objects of known classification [31]. It is worth noting that LDA estimates a single pooled covariance matrix **S**, instead of using a separate estimate for each class. This regularization procedure simplifies the classification model and results in linear decision surfaces (hyperplanes) in R^d [31,33,34]. With this modification, the squared Mahalanobis distance between **x** and the center of the *j*th class is calculated as

$$r^2(\boldsymbol{x}, \boldsymbol{m}_j) = (\boldsymbol{x} - \boldsymbol{m}_j)^T \boldsymbol{S}^{-1} (\boldsymbol{x} - \boldsymbol{m}_j) \quad (2)$$

Object **x** is then assigned to the class *j* for which $r^2(\mathbf{x}, \mathbf{m}_j)$ has the smallest value.

In order to have a well-posed problem, the number of training objects must be larger than the number *d* of variables to be included in the LDA model. Otherwise, the estimated covariance matrix **S** will be singular, which prevents the calculation of the matrix inverse in Eq. (2). Therefore, the use of LDA for classification of spectral data usually requires appropriate variable selection procedures [17,33,35]. In the present work, the Successive Projections Algorithm (SPA) is adopted for this purpose.

2.3. Successive Projections Algorithm

The Successive Projections Algorithm [36,37] was originally proposed by Araújo et al. [38] in the context of multivariate calibration. In SPA, variable selection is formulated as a constrained combinatorial optimization problem, in which subsets of variables are tested and compared with respect to the performance of the resulting model. The optimization is said to be constrained because the search for an optimum is restricted to certain subsets of variables. Such subsets are formed according to a sequence of projection operations involving the matrix **X** of instrumental responses, as follows.

Suppose that the available *x*-data are disposed in a matrix **X** of dimensions ($N \times K$) such that the *k*th variable x_k is associated to the *k*th column vector $\mathbf{x}^k \in \Re^N$. The column vectors are assumed to be mean-centered. Starting from each variable x_k, $k = 1, \ldots, K$, the following sequence of projection operations is carried out [39].

Step 1 (initialization). Let

$$\boldsymbol{z}^1 = \boldsymbol{x}^k$$
$$i = 1$$
$$\boldsymbol{x}^{j,i} = \boldsymbol{x}^j, \quad j = 1, \ldots, K$$
$$SEL(1, k) = k$$

Let *M* be the largest number of variables to be included in a subset, as specified by the analyst.

Step 2. Calculate the matrix $\mathbf{P}^i$ of projection onto the subspace orthogonal to $\mathbf{z}^i$ as

$$\boldsymbol{P}^i = \boldsymbol{I} - \frac{\boldsymbol{z}^i (\boldsymbol{z}^i)^T}{(\boldsymbol{z}^i)^T \boldsymbol{z}^i} \quad (3)$$

where **I** is an identity matrix of appropriate dimensions.

Step 3. Calculate the projected vectors $\mathbf{x}^{j,i+1}$ as

$$\boldsymbol{x}^{j,i+1} = \boldsymbol{P}^i \boldsymbol{x}^{j,i} \quad (4)$$

for all $j = 1, \ldots, K$.

Step 4. Determine the index j^* of the largest projected vector and store this index in matrix **SEL**:

$$j^* = arg \max_{j=1,\ldots,K} ||x^{j,i+1}|| \quad (5)$$

$$SEL(i + 1, k) = j^* \quad (6)$$

Step 5. Let $\mathbf{z}^{i+1} = \mathbf{x}^{j^*,\, i+1}$

Step 6. Let $i = i + 1$. If $i < M$ return to Step 2.

After these operations are completed, a total of $K \times M$ subsets of variables will be considered in the search for the optimum solution. For each value of *k* (ranging from 1 to *K*), and for each value of *i* (ranging from 1 to *M*), a subset of *i* variables is defined by the indexes $SEL(1, k), SEL(2, k), \ldots, SEL(i, k)$.

In a subsequent paper [17], SPA was adapted for use in classification problems. As in the original formulation [38], candidate subsets of variables are formed as the result of projection operations carried out on the matrix of instrumental responses for the training data. However, prior to these operations, the objects belonging to the same class are centered in the mean of the class. The resulting subsets of variables are then compared in terms of a cost function *G* calculated for a given validation data set as

$$G = \frac{1}{N_v} \sum_{n=1}^{N_v} g_n, \quad (7)$$

where g_n is defined as

$$g_n = \frac{r^2(\boldsymbol{x}_n, \boldsymbol{m}_{I(n)})}{\min_{I(m) \neq I(n)} r^2(\boldsymbol{x}_n, \boldsymbol{m}_{I(m)})}. \quad (8)$$

where $I(n)$ is the index of the true class for the *n*th validation object $\mathbf{x}_n$. In Eq. (8), the numerator $r^2(\mathbf{x}_n, \mathbf{m}_{I(n)})$ is the squared Mahalanobis distance between $\mathbf{x}_n$ and the center of its true class, whereas the denominator corresponds to the squared Mahalanobis distance between $\mathbf{x}_n$ and the center of the closest wrong class. The cost function *G* can be interpreted as an average risk of misclassification of the validation data.