Food Research International 51 (2013) 924–928

Contents lists available at SciVerse ScienceDirect

Food Research International

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

Classification of cereal bars using near infrared spectroscopy and linear discriminant analysis

Anna Luiza Bizerra Brito [a], Lívia Rodrigues Brito [b], Fernanda Araújo Honorato [c], Márcio José Coelho Pontes [a], Liliana Fátima Bezerra Lira Pontes [a,*]

[a] Universidade Federal da Paraíba, Departamento de Química, João Pessoa, PB, Brazil
[b] Universidade Federal de Pernambuco, Departamento de Química Fundamental, Recife PE, Brazil
[c] Universidade Federal de Pernambuco, Departamento de Engenharia Química, Recife PE, Brazil

ARTICLE INFO

Article history:
Received 5 November 2012
Accepted 9 February 2013

Keywords:
Cereal bar
Near infrared spectroscopy
Linear discriminant analysis
Successive projection algorithm
Genetic algorithm
Stepwise formulation

ABSTRACT

This work proposes an analytical method for cereal bar classification based on the use of near infrared spectroscopy (NIRS) and supervised pattern recognition techniques. Linear discriminant analysis (LDA) is employed to build a classification model on the basis of a reduced subset of variables (wavenumbers). For the purpose of variable selection, three techniques are considered, namely successive projection algorithm (SPA), Genetic Algorithm (GA), and stepwise (SW) formulation. The methodology is validated in a case study involving the classification of 121 cereal bar samples into three different types (conventional, diet and light). The results show that the LDA/GA model is superior to the LDA/SPA and LDA/SW models with respect to classification accuracy in an independent prediction set. Some advantages of the proposed method are speed, that the analytical measurement is performed quickly (one minute or less per sample), no reagents, low sample consumption and minimum sample preparation demands. In view of the results obtained in this study the proposed method may be considered valid for use in cereal bar classification.

1. Introduction

Consumption of foods which have low calorie content has increased considerably (Bingol, Zhang, Pan, & McHugh, 2012; Carrillo, Varela, & Fiszman, 2012; Gerend, 2009; Kocken, Buijs, & Snel, 2006). With so many options on the market, however, the consumer finds it confusing to select the product they want among foods with labels such as “natural”, “light”, “diet”, “organic”, and “functional”, among others.

Light and diet foods can been found in most supermarkets and labeled as products with low fat, salt, protein, carbohydrates or sugar contents. According to the National Health Surveillance Agency (ANVISA) (Brazilian Agency of Sanitary Surveillance, 2012) these foods are specially formulated with modifications in nutrient content, suitable for use in different diets or for people with specific physiological and metabolic conditions.

The term “light” can be used for a food when the quantity of calories or nutrients is at least 25% less than the conventional product. According to ANVISA, “diet” can be applied to foods with an absence of sucrose/glucose or foods indicated for diets with restriction of some nutrients such as fat, carbohydrate, protein and sodium (Brazilian Agency of Sanitary Surveillance, 2012).

In order to complement regular food, products such as cereal bars are widely consumed as a fast snack with low caloric value. These foods were introduced nearly a decade ago, when consumers became particularly interested in health and diet, since the cereal bars have nutrients such as fibers, vitamins and minerals. Today, cereal bars are consumed worldwide, including by people on diets, or those with health problems or just as a quick snack (Farinazzi-Machado, Barbalho, Oshiiwa, Goulart, & Pessan Junior, 2012; Lobato et al., 2012; Villavicencio, Araújo, Fanaro, Rela, & Mancini-Filho, 2007; Zaveri & Drummond, 2009). Because they are easy to carry and are available on the market in different types, brands, flavors and nutritional compositions, these foods are very well adapted in the day-to-day lives of modern people.

For the consumer, the choice of the product is associated with its appearance, the description on the package and the nutritional information given. These parameters are not always effective, however, in guaranteeing a safe choice of the desired food. Incorrect information on the package or food label, such as the absence or presence of sugar or carbohydrates, can lead to incorrect choices by diabetics, for example, causing hyperglycemia and cardiovascular problems (Scott et al., 2011).

The quality control of cereal bars is performed through physical and chemical tests, such as protein content using the method described by Kjeldahl (AOAC, 1997), moisture determination (Adolfo Lutz Institute, 1985), lipid content (AOAC, 1997) as well as sensory analysis. These reference methods have the drawback of being

* Corresponding author at: Universidade Federal da Paraíba, Departamento de Química, Laboratório de Combustíveis e Materiais (LACOM), CEP 58051-970, João Pessoa, PB, Brazil. Tel./fax: +55 83 3216 7441.
E-mail address: liliana.lira@quimica.ufpb.br (L.F.B.L. Pontes).

0963-9969
http://dx.doi.org/10.1016/j.foodres.2013.02.014

destructive, tedious, time-consuming, using huge amounts of toxic chemical reagents, expensive and, in some cases, subjective.

Infrared spectroscopy (IR) is a non-destructive alternative analytical technique which allows reliable, direct and fast determination of different properties at the same time without sample pre-treatment (Pasquini, 2003). Several papers in the literature report the use of near infrared (NIR) spectroscopy to monitor the quality of foods (Riovanto, Cynkar, Berzaghi, & Cozzolino, 2011; Sinelli, Cerretani, Di Egidio, Bendini, & Casiraghi, 2010).

In chemometrics data analysis, pattern recognition methods are a powerful tool in context of food quality assessment and food composition analysis (Berrueta, Alonso-Salces, & Héberger, 2007; Cen & He, 2007). These methods have been successfully applied to classify a number of foods, including yogurt (Cruz et al., 2013), beer (Granato, Branco, Faria, & Cruz, 2011), coffee (Souto et al., 2010), and coconut oil (Rohman & Che Man, 2011), among others (Berrueta et al., 2007; Cen & He, 2007).

Supervised pattern recognition methods essentially differ in the way they define classification rules. Basically, they can be divided into discriminating and class-modelling methods (Roggo, Duponchel, & Huvenne, 2003). Linear discriminant analysis (LDA) and partial least square-discriminant analysis (PLS-DA) are examples of discriminating techniques, whereas the soft independent modelling of class analogy (SIMCA) is a class-modelling method. The modeling strategies among these methods are substantially different. LDA and PLS-DA focus on the dissimilarity between classes and the samples must be classified in a particular training classe (Roggo et al., 2003; Vaid, Burl, & Lewis, 2001). The SIMCA method, however, considers each class separately and performs outlier tests to decide whether a new object belongs to a certain class, to all classes or does not belong to any class. The SIMCA method is frequently used in data sets with high dimensionality (Brereton, 2009) such as spectroscopic data. However, the testing procedure adopted by SIMCA has the disadvantage that one has to set a confidence level, α. If the data are normally distributed, α % (e.g. 5%) of objects belonging to the class will be considered as not belonging to it. This misclassification problem can be avoided when the LDA method is employed. The LDA method employs linear decision boundaries, which are defined in order to maximize the ratio of between-class to within-class dispersion (Fisher, 1936). Its has been successfully applied to a number of classification problems (Gambarra Neto et al., 2009; Gori, Maggio, Cerretani, Nocetti, & Caboni, 2012; Riovanto et al., 2011; Sinelli et al., 2010; Souto et al., 2010). When compared with SIMCA and PLS-DA, the LDA method has the disadvantage that the number of training samples must be larger than the number of variables included in the LDA model. Therefore, procedures based on the selection of each variable are required for the classification of spectral data. The successive projections algorithm (SPA) (Pontes et al., 2005; Pontes, Pereira, Pimentel, Vasconcelos, & Silva; Silva, Pontes, Pimentel, & Pontes, 2012; Silva, Borba, et al., 2012), genetic algorithm (GA) (Pontes et al., 2005) and stepwise (SW) formulation (Caneca et al., 2006) methods have been adopted for this purpose in different classification problems.

In general, the literature reports only on works involving the characterization of cereal bars as well as new food products being developed for the market by conventional methods (Egert et al., 2012; Fonseca, Santo, Souza, & Pereira, 2011; Heenan et al., 2012; Santos et al., 2011). Explicit studies regarding the classification of cereal bars with respect to type (conventional, diet and light) using NIR spectrometry and multivariate classification with wavenumber selection have not been found in the specialized literature. Thus, development of rapid and accurate methodologies is important for the economy and public health in order to identify non-conformity in cereal bar samples.

In the present work, an analytical method to classify cereal bar samples according to type (diet, light, conventional) using NIR spectroscopy and pattern recognition technique is proposed. For this, LDA is employed to build a classification model on the basis of a reduced subset of spectral variables selected using three different techniques: successive projection algorithm (SPA) (Moreira, Pontes, Galvão, & Araújo, 2009; Pontes et al., 2005), the genetic algorithm (GA) (Pontes et al., 2005), and a stepwise (SW) formulation (Caneca et al., 2006). The results obtained by these three methods (LDA/SPA, LDA/GA, and LDA/SW) have been assessed in terms of classification results in a set of samples not used in the model-building process (prediction samples).

2. Material and methods

2.1. Samples

One hundred and twenty-one samples of three different types of cereal bars were analyzed: Diet (35); Conventional (44) and Light (42). All samples were crushed and sieved with 20 mesh particle size.

2.2. NIR spectra measurements

A Spectrum 400 FT–IR/FT–NIR spectrophotometer (Perkin Elmer) equipped with accessory reflectance (NIRA) was employed to obtain NIR spectra in the range of 10,000 a 4000 cm^{-1}. All spectra were recorded with an average of 16 scans, and a spectral resolution of 8 cm^{-1}. The background spectra were obtained using the Spectralon standard. Temperature was controlled at 23 ± 1 °C throughout the spectral acquisition process.

2.3. Chemometric procedure

Raw spectra and some pre-processing strategies such as baseline correction, standard normal variate (SNV) (Barnes, Dhanoa, & Lister, 1989), smoothing and the Savitzky–Golay first derivative (Savitzky & Golay, 1964), second-order polynomial (7, 11 and 15 window points), were evaluated in terms of overall classification errors. Detection and elimination of outliers were carried out using score, residual and leverage plots. For each type, the Kennard–Stone (KS) algorithm (Kennard & Stone, 1969) was applied in order to divide the samples into training, validation and prediction subsets. Table 1 presents the number of samples in each set. The validation method employed in this study is based on the test set, thereby the training and validation samples were used in the modelling procedures (including variable selection for LDA) whereas the prediction samples were only used in the final evaluation of the classification models. All spectral data were mean-centered before modeling procedures.

An exploratory analysis was initially performed in order to observe the existence of natural groupings. For this purpose, the principal component analysis (PCA) (Souza et al., 2011) was applied to the overall spectral data set.

The present work adopts the stepwise (SW) formulation developed by Caneca et al. (2006). In order to decide which wavenumbers were to be discarded, seven threshold values (0.30, 0.50, 0.70, 0.80, 0.90, 0.95, and 0.99) of multiple correlation coefficients were tested in the LDA/SW model. The best threshold was selected on the basis of the classification errors in the validation set.

The SPA algorithm utilized in this paper (Pontes et al., 2005) employs, as the cost function, the average risk G of misclassification by

Table 1
Number of training, validation and prediction samples in each cereal bar type.

Type	Sets		
	Training	Validation	Prediction
Diet	17	09	09
Conventional	22	11	11
Light	20	11	11
Total	59	31	31