Chemometrics and Intelligent Laboratory Systems 135 (2014) 150–156

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Geographical and genotypic classification of arabica coffee using Fourier transform infrared spectroscopy and radial-basis function networks

Jade Varaschim Link [a,*], André Luis Guimarães Lemes [b], Izabele Marquetti [a], Maria Brígida dos Santos Scholz [c], Evandro Bona [a,b]

[a] Post-Graduation Program of Food Technology (PPGTA), Federal University of Technology – Paraná (UTFPR), P.O. Box 271, Via Rosalina Maria dos Santos - 1233, CEP 87301-899 Campo Mourão, PR, Brazil

[b] Food Department (DALIM), Graduation Program in Food Engineering, Federal University of Technology – Paraná (UTFPR), P.O. Box 271, Via Rosalina Maria dos Santos - 1233, CEP 87301-899 Campo Mourão, PR, Brazil

[c] IAPAR - Agronomic Institute of Paraná, Rodovia Celso Garcia Cid, km 375, CEP 86047-902 Londrina, PR, Brazil

ARTICLE INFO

Article history:
Received 11 March 2014
Received in revised form 9 April 2014
Accepted 11 April 2014
Available online 24 April 2014

Keywords:
Green coffee
Mid-infrared spectrum
RBF
Artificial neural networks

ABSTRACT

The climatic conditions of coffee cultivation give special attributes to the beverage and could increase its value. However, it is essential to prove the geographical and genotypic origin of the cultivar using reliable methods. An example of an artificial neural network (ANN) that has been used for pattern classification is the radial-basis function network (RBF). This study aimed to develop a RBF to classify the geographic and genotypic origin of arabica coffee. For this purpose, spectra obtained in the Fourier transform infrared (FTIR) were analyzed by using RBFs. In the development of networks, other methods were applied for: the choice of network parameters (sequential simplex optimization) and improve the generalization of a neural network (ensemble averaging). The optimized RBFs were able to classify the samples of arabica coffee, both geographically (100% correct classification) and genotypically (94.44%). The performance of the developed RBFs was better than the SIMCA (Soft Independent Modeling of Class Analogies) and multilayer perceptron (MLP) developed for coffee classification.

1. Introduction

Coffee comes from a tree of the genus *Coffea* and belongs to the family rubiaceae comprising more than 500 genus and 6000 species. In the genus *Coffea*, there are two main species: *Coffea arabica*, also known as arabica coffee, and *Coffea canephora* or robusta coffee [1]. Brazilian consumption of coffee increases every year, and between 2011 and 2012 the consumption was 19.975 million bags, representing an increase of 3.05% over the previous period [2]. Besides the internal market, exports of green coffee from Brazil amounted to 116.63 thousand tons in 2012, down 19.2% compared to the same period in 2011 [3].

The climatic conditions of coffee cultivation together with the genetic characteristics of the cultivars give special attributes to the beverage and could increase its value. However, it is essential to prove the geographical and genotypic origin of the cultivar by using reliable methods [4]. Thus, it becomes essential to use faster analytical techniques that have low cost and allow differentiation and certification of the composition and quality of the coffee. A methodology that combines these attributes is Fourier transform infrared (FTIR) spectroscopy. FTIR allows detection of components of greater importance in the food characterization, and is a method that quickly gives summary information about the product. Another advantage is the fact that the sample preparation generally does not require solvent extraction thereby decreasing analysis time and the environmental impact [5]. In addition, FTIR has already been used with success to authenticate mixtures of arabica and robusta coffees or adulterants addition [6–11].

Humans have excellence for learning and pattern recognition. Several statistical methods have been developed in an attempt to reproduce the human capability of pattern recognition [12,13]. Artificial neural networks (ANNs) are a set of techniques based on statistical principles, which are currently growing in food science to perform tasks of regression and pattern recognition. ANNs are techniques that can represent complex and nonlinear input–output relationships [12]. An example of an ANN that has been used to pattern classification is the radial-basis function network (RBF) [14]. In the context of a RBF, the hidden units provide a set of functions that constitute an arbitrary basis for the input patterns when they are expanded into a multidimensional hidden space [13].

This study aimed to develop a radial-basis function network to classify the geographic and genotypic origin of arabica coffee. For this purpose, spectra obtained in FTIR were analyzed by using RBFs optimized by sequential simplex. In the development of networks, other methods were applied to improve the generalization of a neural network (ensemble averaging).

* Corresponding author. Tel.: +55 46 99156494.
E-mail address: jadejvl@hotmail.com (J.V. Link).

http://dx.doi.org/10.1016/j.chemolab.2014.04.008
0169-7439/

2. Materials and methods

All mathematical and/or statistical analyses, as well as PCA (principal component analysis) such as the radial-basis function networks and the sequential simplex, were performed in MATLAB R2008b (The MathWorks Inc., Natick, USA). The spectral preprocessing was performed in IRsolution 1.5 software (Shimadzu Corporation, Kyoto, Japão).

2.1. Samples of coffee

All coffee samples are genotypes of the species *C. arabica*. Samples of modern genotypes with a wide variety of genetic background (IAPAR 59, IPR 99, IPR 105 and IPR 106) were collected in the 2010 season.

About 3 kg of cherry coffee of 4 genotypes of coffee were collected (totaling 18 samples) (Table 1) at four locations in the coffee region of Paraná-Brazil: Paranavaí, Cornélio Procópio, Mandaguari and Londrina and were transported to Agronomic Institute of Paraná (IAPAR) in Londrina-Paraná. The samples were immediately placed into wooden boxes with a mesh bottom and moved eight times per day until the beans reached 11–12% moisture and then the samples were processed (removal of hull and parchment). The coffee beans were frozen with liquid nitrogen and were ground in a mill disk (model Perten 3600) with 0.6-mm final particle size [15]. The samples provided by IAPAR were kept frozen at −18 °C and thawed for analysis. Fig. 1 presents the location of the cities where the coffees were cultivated and the distance (in straight line) between the most distant cities (Paranavaí and Cornélio Procópio).

2.2. Fourier transform infrared spectroscopy

Pellets were prepared by adding about 100 mg of dry KBr (FTIR grade – Sigma-Aldrich) and approximately 1 mg of finely ground sample. The mixture was compressed in a hydraulic press (Bovenau, P15 ST) using a mold (ICL, ICL's Macro/Micro KBr dye) employing about 35 MPa pressure to produce a transparent pellet. Before the analysis of each sample, the FTIR (Shimadzu FTIR - 8300) was programmed to perform a background spectrum of the air, which was used to subtract the influence of air components in the spectrum. After this, the pastille was positioned on the instrument shaft and the spectra were obtained in the range 4000 to 400 cm^{-1}. Accumulated scans ($n = 32$) were used to form the final spectrum and five repetitions (pellets) were performed for each sample, totalizing 90 spectra. Seventy-two (80%) of these spectra were used as training samples for artificial neural networks and 18 spectra (20%) were used as test samples (one repetition of each coffee studied).

2.3. Preprocessing

After obtaining the spectra, transformations consisting of several stages were performed by using IRsolution 1.5 (Shimadzu Corporation, 2008). First, normalization of the spectrum was done (largest band obtained absorbance 1 and the lower absorbance 0) to eliminate effects due to minor differences among the masses of the sample used for the preparation of pellets. After, baseline correction, smoothing of the spectrum (17 smoothing points), and a band related to CO_2 concentration was removed to disregard CO_2 variations between samples. For the PCA and the ANN, the spectral region between 3750 and 750 cm^{-1} was used, excluding noises that were beyond this region. The spectrum region between 1900 and 800 cm^{-1} was also evaluated because it contains absorption bands due to carbonyl axial symmetric deformation (esters, aldehydes, and ketones), methylene angular symmetric deformation, and angular and axial symmetric deformations of C–O (esters and alcohols). Therefore, this region contains the fingerprint information for discrimination of different coffee samples [7,10,11, 16].

After preprocessing, principal component analysis (PCA) was used to reduce the data dimensionality. The PCA has the objective to produce a lower dimensional description of the rows and columns of a multivariate data matrix and the principle idea of reducing the dimension of the dataset is achieved through linear combinations. Low dimensional linear combinations are often easier to interpret and serve as an intermediate step in a more complex data analysis [17]. PCA was performed on correlation matrix of normalized data as well as the 1st and 2nd derivative spectra. The application of the 1st and 2nd derivative of the raw spectral data is a procedure that can separate spectral differences and minimize the effect of slopes caused at baseline spectra. By applying the derivation operations, the information across the different wavenumbers are usually pronounced (the noise become steep too) [18]. The number of principal components used was one of the parameters optimized by sequential simplex.

2.4. Normalization

The input vectors (principal components chosen) were normalized before being fed into the neural network so that the activation functions of artificial neurons are not easily saturated (overflow error) [13]. The preprocessing methods used were: minimum and maximum (minimax), transformation to a range between −1 and 1; autoscale, input vector with zero mean and unit variance; unit vector, Euclidean norm equal to 1 [19].

2.5. Radial-basis function networks

The architecture of a neural network RBF (Fig. 2), involves three layers with entirely different roles. The input neurons connect the network to its environment. The second layer, the only hidden layer in the network, provides a set of radial functions that form an arbitrary basis for input patterns. The hidden space is of high dimensionality, allowing a linear separation of groups (Cover's theorem on the separability of patterns). The output layer is linear, supplying the response of the network to the active pattern applied to the input layer [13].

The number of neurons in the input layer is determined by the number of principal components used, and it is a parameter that was determined by the sequential simplex. In the hidden layer are φ_{jk} radial-basis functions, where j represents the class number and k represents the kth radial basis function. The number of radial bases used was also a parameter determined by the sequential simplex. The number of hidden nodes associated with each class is dependent on the complexity of the patterns to be separated [14].

There is a large class of radial-basis functions, which includes the following functions that are of particular interest in the study of RBF

Table 1
Studied coffee genotypes.

Genotype	City	Samples
IPR 99	Cornélio Procópio	1
	Paranavaí	1
	Mandaguari	1
	Londrina	1
IPR 105	Cornélio Procópio	1
	Paranavaí	1
	Mandaguari	1
	Londrina	1
IPR 106	Cornélio Procópio	1
	Paranavaí	1
	Mandaguari	1
	Londrina	1
IAPAR 59 (IA 59)	Cornélio Procópio	2
	Paranavaí	2
	Mandaguari	1
	Londrina	1