LWT - Food Science and Technology 76 (2017) 330–336

Contents lists available at ScienceDirect

LWT - Food Science and Technology

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

Support vector machines in tandem with infrared spectroscopy for geographical classification of green arabica coffee

Evandro Bona [a, *], Izabele Marquetti [a], Jade Varaschim Link [a, b],
Gustavo Yasuo Figueiredo Makimori [a], Vinícius da Costa Arca [a],
André Luis Guimarães Lemes [a], Juliana Mendes Garcia Ferreira [a],
Maria Brígida dos Santos Scholz [c], Patrícia Valderrama [a], Ronei Jesus Poppi [d]

[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] *Federal University of Santa Catarina (UFSC), Campus Universitário – Trindade, CEP 88040-900 Florianópolis, SC, Brazil*
[c] *Agronomic Institute of Paraná (IAPAR), Rodovia Celso Garcia Cid, km 375, CEP 86047-902 Londrina, PR, Brazil*
[d] *Institute of Chemistry, University of Campinas (UNICAMP), P.O. Box 6154, CEP 13083-970 Campinas, SP, Brazil*

ARTICLE INFO

Article history:
Received 10 February 2016
Received in revised form
20 April 2016
Accepted 25 April 2016
Available online 27 April 2016

Keywords:
Machine learning
Near infrared
Mid infrared
Genetic algorithm

ABSTRACT

The coffee is an important commodity to Brazil. Species, climate, genotypes, cultivation practices and industrialization are critical to final quality of the beverage. Thus, the development of analytical methods for coffee authentication is important to ensure the origin of the bean. The purpose of this study was to develop a methodology for geographical classification of different genotypes of arabica coffee using infrared spectroscopy and support vector machines (SVM). The spectra were collected in the range of near infrared (NIRS) and mid infrared (FTIR). For the data analysis, a SVM was built using radial basis as kernel function and the one-versus-all multiclass approach. The C and γ parameters of SVM were optimized using the genetic algorithm. With the application of the NIRS-SVM approach all test samples were correctly classified with a sensitivity and specificity of 100%, while FTIR-SVM had a slightly lower performance. Therefore, it was possible to confirm that infrared spectroscopy is a fast and effective method for geographic certification with little sample preparation, and without the production of chemical wastes. Furthermore, the SVM can be a chemometric alternative in tandem with infrared spectroscopy for another classification problems.

1. Introduction

The coffee is one of the most consumed beverages in the world, and an important commodity to Brazil, which is the largest producer and exporter in the world. From November 2014 to October 2015, 36 million bags of 60 kg of coffee were exported, corresponding to US$ 6.348 billion (CeCafé, 2015). There are two main species of coffee, *Coffea arabica*, also known as arabica coffee, and *Coffea canephora* or robusta coffee (Clarke & Vitzthum, 2001). These species present a very different chemical composition and arabica coffee is known for its high quality beverage, with an intense aroma, lower caffeine content, and a less bitter taste, showing a higher aggregate price (Lashermes & Anthony, 2007; Link, Guimarães Lemes, Marquetti, dos Santos Scholz, & Bona, 2014). Among cultivars of arabica coffee, different levels of quality beverages have been found due to genetic factors and environmental conditions in the place of cultivation. Beans from regions and varieties that are known to produce high quality beverages have a great commercial value (Bertrand et al., 2012; Joët et al., 2010; Teuber, 2010). However, it is essential to prove the geographical and genotypic origin of the cultivar using a reliable method.

Different analytical techniques are often employed for coffee analysis, including chromatographic analysis (Novaes, Oigman, de Souza, Rezende, & de Aquino Neto, 2015), UV–Vis spectroscopy (Souto et al., 2015), nuclear magnetic resonance (Arana et al., 2015), and physicochemical analysis (Borsato, Pina, Spacino, Scholz, & Androcioli Filho, 2011). These are slow techniques because they require more time to prepare samples, have high costs, and

* Corresponding author.
E-mail address: ebona@utfpr.edu.br (E. Bona).

http://dx.doi.org/10.1016/j.lwt.2016.04.048

generate too much residues. To overcome these disadvantages, an alternative is employ infrared spectroscopy, which is a fast technique that requires minimum sample preparation, do not destroy samples, and allows simultaneous analyses (Downey, Briandet, Wilson, & Kemsley, 1997; Karoui, Downey, & Blecker, 2010; Terouzi et al., 2011). However, infrared spectroscopy is a technique with a high complexity data and a large amount of information. Thus, chemometric methods are required for spectra interpretation (Hiri et al., 2015; Roggo et al., 2007).

In previous studies, the efficiency of Fourier transform mid infrared spectroscopy (FTIR) was verified (Link and Lemes et al., 2014), as well as near infrared spectroscopy (NIRS) (Marquetti et al., 2016) for geographical classification of four arabica coffee genotypes. In the first work, the FTIR spectra were treated by principal component analysis (PCA) to dimensionality reduction, and the scores were used as input to an artificial neural network of radial basis functions (RBF). In the later work, the NIRS spectra were analysed using partial least square with discriminant analysis (PLS-DA). In the present study, twenty arabica coffee genotypes were classified and the spectra was collected in both FTIR and NIR. For data analysis, support vector machine (SVM) was evaluated because it is able to model nonlinear relations. SVM is a new type of machine learning based on statistical learning theory that emphasizes machine learning in the case of fewer samples (Bishop, 2006). The structural risk minimization principle derived from statistical learning theory takes this as the foundation, as compared with RBF, giving the SVM prominent advantages. First, the strong theoretical basis provides high generalization capability and avoids overfitting. In second place, the global model is capable of dealing efficiently with high-dimensional input vectors. Third, the solution is sparse and only a subset of training samples contributes to this solution, thereby reducing the workload (Argyri et al., 2013). As a nonlinear method, SVM shows advantages over PLS-DA, the latest tends to shrink the low variance directions, but can actually inflate some high variance directions. This can make the PLS a little unstable. Besides, PLS down weights noisy features, but does not throw them away; therefore a lot of noise can contaminate the predictions. In addition, highly correlated variables will tend to be chosen together, as a result, there may be much redundancy in the set of selected variables. This might indicate that the PLS method is more prone to overfitting than SVM. Overfitting is more likely to take place on high dimensional data, and infrared spectra typically show very high dimension (Liu, Yang, & Deng, 2015).

The actual objective was to develop a methodology for geographical classification of different genotypes of green arabica coffee using infrared spectroscopy in tandem with support vector machines.

2. Materials and methods

2.1. Coffee samples

This study used about 3 kg of cherry coffee, harvested between 2009 and 2010, of 20 genotypes collected at four locations (totaling 74 samples) in the coffee region of Paraná-Brazil: Paranavaí, Cornélio Procópio, Mandaguari and Londrina. The samples were placed into wooden boxes with a mesh bottom and moved eight times per day until the beans reached 11–12% moisture. After that, the samples were processed (removal of hull and parchment); the coffee beans were frozen with liquid nitrogen, ground in a mill disk (model Perten 3600) with 0.6-mm final particle size and kept at −18 °C. Before analysis, the samples were thawed and retained in a desiccator to even out moisture (Link, Lemes, et al., 2014; Marquetti et al., 2016). More information about climatic conditions, cities position and the number of samples per genotype, year and city, can be obtained in Electronic Supplementary Material (Tables 1S and 2S).

2.2. Fourier transform mid infrared spectroscopy (FTIR)

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 that, 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 370 spectra (Link, Lemes, et al., 2014).

After obtaining the spectra, normalization of the spectrum was done to eliminate effects due to minor differences among the weights of the sample used for the preparation of pellets. Next, baseline correction and smoothing of the spectrum using Savitzky-Golay algorithm was performed (Savitzky & Golay, 1964). Only the spectrum region between 1800 and 800 cm^{-1} was evaluated because it contains the most important 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 (Link, Lemes, et al., 2014).

2.3. Near infrared spectroscopy (NIRS)

Green coffee spectra were recorded using a near infrared spectroscopy NIRSystem 5000-M (Foss Tecator AB, Höganäs, Sweden). Measurements were made at room temperature (23 °C) in the wavelength range 1100–2498 nm at 2 nm intervals. The software WinISI III, version 1.50e (Foss NIRSystems/Tecator Infrasoft International, LLC, Silver Spring, MD, USA), was used to acquire the spectra. To reduce variation sources that carry no relevant information during the multivariate calibration, and considering scatter effects between samples, the multiplicative scatter correction (MSC) was applied (Marquetti et al., 2016). MSC corrects multiplicative and additive scatter effects, which are the result of differences in granules' size, morphology, and particle orientation. It uses a linear regression of spectral variables versus the average spectrum (Isaksson & Naes, 1988).

2.4. Sample selection

The spectra, both NIRS as FTIR, were split in training set (2/3) and test set (1/3) using the Kennard and Stone algorithm (Kennard & Stone, 1969). In detail, Kennard and Stone algorithm aims at selecting the most diverse set among a given set of candidate samples, to be included in the training set, according to a *maximin* criterion. At first, the distances among all pairs of samples are computed and the two most distant samples are selected to be included in the training set. Successively, for each of the remaining candidate samples, the minimum distance to all the already selected samples is computed, so that the one showing the maximum value of this minimum distance is in turn selected to be included in the training set. The whole procedure is then repeated until the desired number of training samples is selected (Westad & Marini, 2015).

2.5. Support vector machine (SVM)

SVM is a relatively new tool, originally designed for binary classification problems involving large multidimensional data sets. One distinctive feature in SVM is to solve both linear and nonlinear classification/regression problem using the same methodology (Argyri et al., 2013). We begin our discussion of support vector machines by the binary classification problem for linearly separable patterns (Fig. 1) and only the general algorithm procedure will be reviewed, the detailed mathematical derivations can be found in Bishop (2006) and Haykin (2008). The equation of a decision surface in the form of a hyperplane that does the separation is

$$\mathbf{w}^T\mathbf{x} + b = 0 \tag{1}$$

where $\mathbf{x}$ is an input vector, $\mathbf{w}$ is an adjustable weight vector, and b is a bias. Applying the principle of empirical risk minimization (ERM), which is employed for multilayer perceptrons or radial basis function artificial neural networks, there are infinite solutions (local minimum) with classification error equal to zero (Fig. 1(a)). On the other hand, the SVM follows the principle of structural risk minimization (SRM) which is based on the theory of Vapnik—-Chervononkis (VC) dimension (Smola & Schölkopf, 2004). Using this principle, there is a single hyperplane that maximizes the margin of separation between positive and negative examples (Fig. 1(b)). In summary, the optimal hyperplane defined by Eq. (1) is unique in the sense that the optimum weight vector provides the maximum possible separation between positive and negative examples. This optimum condition is attained by minimizing the Euclidean norm of the weight vector $\mathbf{w}$ (Haykin, 2008). The larger is the margin, the better will be the generalization error of the classifier (Papadopoulou, Panagou, Mohareb, & Nychas, 2013).

The SVM project depends on the extraction of a subset of training data that serves as support vectors, and therefore represents a stable feature of the data. The dimensionality of the feature space is determined by the number of support vectors extracted from the training data (Haykin, 2008). The amount of support vectors is set automatically by the learning algorithm according to the complexity of the data (Li, Liang, & Xu, 2009). All the remaining examples in the training sample are completely irrelevant. Because of their distinct property, the support vectors play a prominent role in the operation of this class of learning machines. For linearly separable data, models developed with SRM has greater ability of generalization than those based on ERM (Haykin, 2008).

In SVM, nonlinear problems are transformed to linear ones by mapping the data into a high-dimensional feature space via a nonlinear function (kernel function) and perform a linear classification or regression in this feature space. The dimensionality of the feature space is determined by the choice of kernel function and it parameter (γ) while the complexity of the model is determined by an extra penalty parameter (C) (Argyri et al., 2013). The parameter C controls the trade-off between maximizing the margin and minimizing the training error. If C is too high, the algorithm will over fit the training data; if C is too small, then insufficient training will occur. For the RBF kernel type function, used in this application, the most important parameter is the width γ that controls the amplitude of the kernel function and hence the generalization ability of SVM (Papadopoulou et al., 2013).

As the SVM is a binary classifier, the use of a complementary technique for multiclass problems is required. There are two main types of schemes for the construction of multiclass classifiers: all-vs-all (AVA) and one-versus-all (OVA). In the OVA scheme, k SVMs are constructed separating the kth class from the rest. On the other hand, in the AVA scheme, $\binom{N}{2}$ binary classifiers that separate a couple of classes are trained. The OVA model was chosen because of the ease and speed of implementation, with the same accuracy of other multiclass schemes (Rifkin & Klautau, 2004). Thus, to apply the OVA scheme to geographical classification of arabica coffee, four SVMs were necessary (one to separate each city from the rest).

The parameters C and γ must be tuned for each problem (Haykin, 2008). In this work, the genetic algorithm (Haupt & Haupt, 2004) was applied to choose the best values based on the classifier performance, Equation (2), using 10-fold cross-validation of training samples.

$$perf = \left(\prod_{i=1}^{k} AUC_i\right)^{1/k} \tag{2}$$

The Equation (2) is the geometric mean of AUC_i [0, 1], which is the area under the receiver operating characteristic (ROC) curve for the ith SVM. The ROC curve plots the true-positive rate against the false-positive rate for a varying decision threshold (Haykin, 2008). After definition of C and γ, the generalization of the classifier was evaluated using the accuracy, sensitivity and specificity for test set. The accuracy is the proportion of both true positives and true negatives among the total number of cases examined. The sensitivity is the model ability to correctly classify the samples, relating the predicted samples to being in a class with the samples that actually are in this class. The specificity relates the predicted samples to not being in a class with the samples that actually are not in this class (Metz, 1978).

All multivariate analysis for NIRS and FTIR spectra were performed using MATLAB R2008b (The MathWorks Inc., Natick, USA)

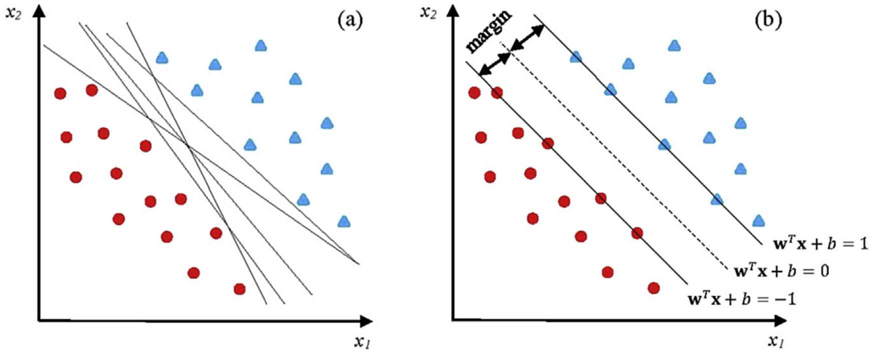
(a)
x₂
x₁
(b)
margin
$\mathbf{w}^T\mathbf{x}+b=1$
$\mathbf{w}^T\mathbf{x}+b=0$
$\mathbf{w}^T\mathbf{x}+b=-1$

Fig. 1. Two-dimensional representation of linearly separable binary classification. (a) Using the principle of empirical risk minimization, several solutions correctly classify the binary patterns. (b) For the principle of structural risk minimization, only one solution maximizes the margin.

and LIBSVM library version 3.20 (Chang & Lin, 2011) available at www.csie.ntu.edu.tw/~cjlin/libsvm/.

3. Results and discussion

The mean spectra for each city are shown in Fig. 2. The FTIR spectra, Fig. 2(a), show a clear separation between the cities based on bands 1744 cm^{-1} (esters); 1285 cm^{-1} (chlorogenic acids); the regions from 1650 cm^{-1}–1600 cm^{-1} (caffeine carbonyl); 1200 cm^{-1} to 900 cm^{-1} and 1500 cm^{-1} to 1400 cm^{-1} (carbohydrates) (Link, Lemes, et al., 2014). On the other hand, for NIRS spectra, Fig. 2(b), the differences between the cities are less visible. The major differences are noted in the region between 1900 and 2400 nm which is related to lipids, water, caffeine, chlorogenic acids, trigonelline, proteins, amino acids, sugars, and carbohydrates (Marquetti et al., 2016).

However, for principal component analysis the result was opposite (Fig. 3). The NIRS spectra (Fig. 3(b)) lead to a higher segmentation of the samples when comparing with FTIR spectra (Fig. 3(a)). For both spectral ranges, two principal components were sufficient to represent more than 95% of the variance. However, the large number of overlapping classes indicate that the classification problem is not trivial. The interaction between the genetic variability of arabica and the cultivation conditions affects both the chemical composition, as the physicochemical characteristics of the coffee beans (Scholz, Figueiredo, Silva, & Kitzberger, 2011).

The samples were automatically selected using the Kennard and Stone algorithm, as shown in Table 1. The algorithm was set up to selected 2/3 of total samples to training set, after choosing all samples; it was observed that for each city this proportion was approximately maintained. In Fig. 1S, Electronic Supplementary Material, the splitting of the available data is shown in score units of the first two principal components.

After the selection of the samples, the SVM parameters (C and γ) were adjusted through genetic algorithm. The parameters of the genetic algorithm were selected through some preliminary tests. For FTIR the optimized values were $C = 3.40 \times 10^5$ e $\gamma = 5.14 \times 10^{-4}$; and for NIRS the optimized values were $C = 4.28 \times 10^6$ e $\gamma = 1.14 \times 10^{-2}$. The main difference when changing parameters involves the smoothness or complexity of classification boundaries. For high γ, the decision function is very spiky, for high C, the margins are narrower, and the number of support vectors is lower. The consequence of this is that the shape of the surface can be more complex (Brereton & Lloyd, 2010). Furthermore, when the parameter C is assigned a large value, the implication is that the designer of the support vector machine has high confidence in the quality of the training sample. Conversely, when C is assigned a small value, the training sample is considered to be noisy, and less emphasis should therefore be placed on it (Haykin, 2008). Thus, for geographical classification of arabica coffee, the NIRS spectra were less noisy, decision boundaries were more spiky and complex.

Table 2 exhibits the performance of SVM using FTIR spectra. The number of support vectors ranges from 34 to 55, in conceptual terms, the support vectors are those data points that lie closest to the optimal hyperplane and are therefore the most difficult to classify (Haykin, 2008). Therefore, the easier problem results in a smaller number of support vector being required by the SVM classifier (Brudzewski, Osowski, & Markiewicz, 2004; Pardo & Sberveglieri, 2005).

The best accuracy for test set was obtained for Londrina and Paranavaí, and the worst accuracy was achieved for Cornélio

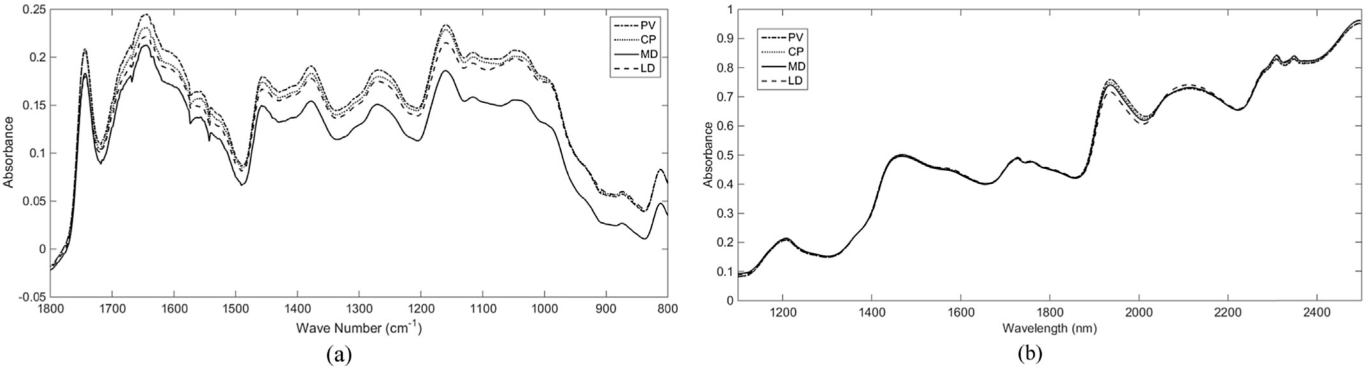

Fig. 2. The mean infrared spectra for each city: (a) FTIR and (b) NIRS.

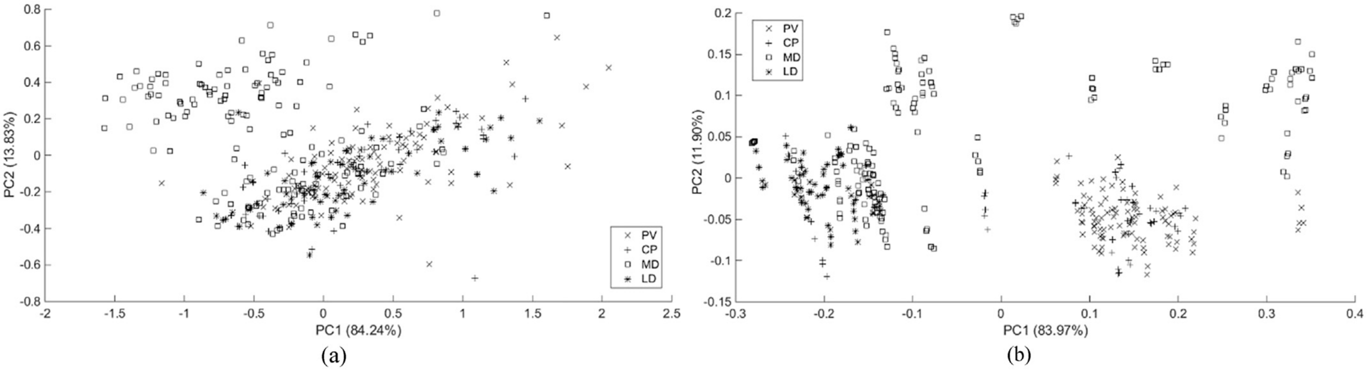

Fig. 3. Principal component analysis for infrared spectra: (a) FTIR and (b) NIRS.

Table 1
Sample selection using Kennard and Stone algorithm.

Spectra	City	Samples		
		Training	Test	Total
FTIR	Paranavaí	68 (68%)	32 (32%)	100
	Cornélio Procópio	31 (62%)	19 (38%)	50
	Mandaguari	100 (69%)	45 (31%)	145
	Londrina	44 (64%)	25 (36%)	69
	Total	243 (67%)	121 (33%)	364[a]
NIRS	Paranavaí	66 (66%)	34 (34%)	100
	Cornélio Procópio	36 (72%)	14 (28%)	50
	Mandaguari	96 (64%)	54 (36%)	150
	Londrina	49 (70%)	21 (30%)	70
	Total	247 (67%)	123 (33%)	370

[a] Some spectra were lost due to experimental errors.

Procópio. Both in training and in the test, some samples of Mandaguari were classified as belonging to Cornélio Procópio. Both cities have similar altitude that could explain the same level of adaptation of genotypes at these locations. In Link, Lemes, et al. (2014) the PCA-RBF approach in tandem with FTIR achieved 100% of correct classification for test set (accuracy, sensitivity and specificity were not reported). The present work reached 95.87% of correct classification for test set, but greater number of samples and genotypes were used, whereas in the first work only four genotypes were evaluated. Furthermore, the SVM does not require the use of PCA to reduce the dimensionality of the input space and has a smaller number of parameters to be optimized, when compared to a RBF.

The graphical representation of SVM output for geographical classification using FTIR spectra is shown in Fig. 4. In this figure, it is possible to observe the false positives and false negatives for each city. It is noted also that the SVM for CP, which had the lowest accuracy (Table 2), has the narrowest margin.

For NIRS spectra, SVM required less support vectors for each city: 15 for Paranavaí, 24 for Cornélio Procópio, 25 for Mandaguari, and 10 for Londrina. As previously mentioned, based on the number of support vectors, it can be concluded that the classification of arabica coffee was easier with the use of NIRS spectra than FTIR spectra. It also points out that the classification performance using NIRS spectra was perfect for both training set, as for test set. The accuracy, sensitivity and specificity were equal to 1.0 for all cities, so no false positive or false negative was observed (Fig. 5). It also highlights an increase in the margin, especially for samples of Cornélio Procópio. The worst accuracy using the FTIR spectra can be the result of noise generated by the preparation of pellets that, even with experimental care, tend to be heterogeneous. To eliminate this effect, the use of an attenuated total reflectance (ATR) accessory could be interesting. ATR is a versatile and powerful technique for easy infrared sampling. It is useful to sample the surface of smooth

Table 2
SVM performance for geographical classification using FTIR spectra.

		Paranavaí (PV)	Cornélio Procópio (CP)	Mandaquari (MD)	Londrina (LD)
Support vectors		54	42	55	34
Train	Accuracy	1.0000	0.9959	0.9959	1.0000
	False positive[a]	0	1 (MD)	0	0
	False negative	0	0	1	0
	Sensitivity	1.0000	1.0000	0.9900	1.0000
	Specificity	1.0000	0.9953	1.0000	1.0000
Test	Accuracy	0.9835	0.9504	0.9587	0.9835
	False positive[a]	1 (CP)	6 (MD)	0	0
	False negative	1	0	5	2
	Sensitivity	0.9688	1.0000	0.8889	0.9200
	Specificity	0.9888	0.9412	1.0000	1.0000

[a] In parentheses is the geographical origin of the sample misclassified.

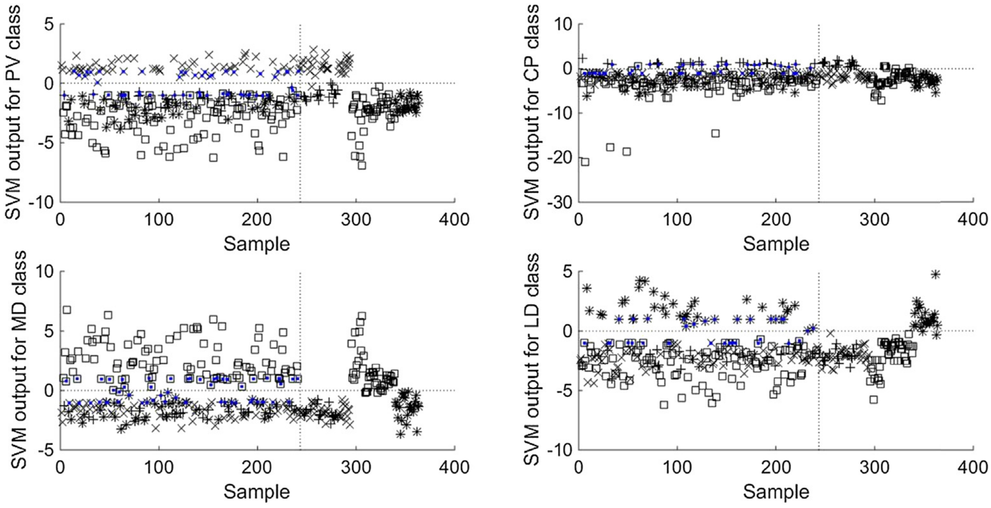

Fig. 4. FTIR-SVM output for each city: (×) Paranavaí, (+) Cornélio Procópio, (□) Mandaguari, (✳) Londrina. The vertical dotted line separates the training samples of the test samples. The training samples selected as support vectors are marked with a blue dot. For interpretation of the references to the color in this figure legend, the reader is referred to the web version of this article.

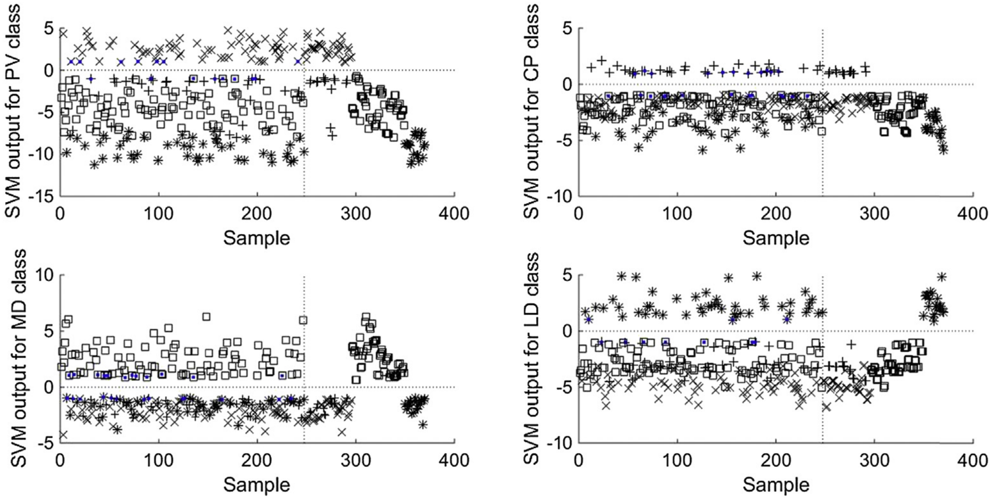

Fig. 5. NIRS-SVM output for each city: (×) Paranavaí, (+) Cornélio Procópio, (□) Mandaguari, (✳) Londrina. The vertical dotted line separates the training samples of the test samples. The training samples selected as support vectors are marked with a blue dot. For interpretation of the references to the color in this figure legend, the reader is referred to the web version of this article.

materials that are either too thick or too opaque for transmission measurements. ATR is a nondestructive method; in addition, little or no sample preparation is needed, and it allows fast and simple sampling regardless of the state of the food system (Hiri et al., 2015; Karoui et al., 2010; Terouzi et al., 2011).

Even with a larger number of genotypes this study has a better accuracy than that obtained by Marquetti et al. (2016) which geographically classified four genotypes using NIRS-PLS-DA approach. As the SVM is a nonlinear method, a similar or better performance than the one obtained by linear methods is expected (Haykin, 2008).

4. Conclusions

The study confirmed that infrared spectroscopy is a fast and effective method for geographic certification of different arabica coffee genotypes. Furthermore, it is a nondestructive analysis with little sample preparation and no production of chemical wastes. It was also observed that NIRS attained a superior performance when compared to FTIR. Therefore, the SVM can be a chemometric alternative in tandem with infrared spectroscopy for different classification problems.

Acknowledgments

The authors thank CAPES, CNPq (307483/2015-0) and Fundação Araucária (383/2014) for their financial support and scholarships.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.lwt.2016.04.048.

References

Arana, V. A., Medina, J., Alarcon, R., Moreno, E., Heintz, L., Schäfer, H., et al. (2015). Coffee's country of origin determined by NMR: the Colombian case. *Food Chemistry, 175*, 500–506. http://dx.doi.org/10.1016/j.foodchem.2014.11.160.

Argyri, A. A., Jarvis, R. M., Wedge, D., Xu, Y., Panagou, E. Z., Goodacre, R., et al. (2013). A comparison of Raman and FT-IR spectroscopy for the prediction of meat spoilage. *Food Control, 29*(2), 461–470. http://dx.doi.org/10.1016/j.foodcont.2012.05.040.

Bertrand, B., Boulanger, R., Dussert, S., Ribeyre, F., Berthiot, L., Descroix, F., et al. (2012). Climatic factors directly impact the volatile organic compound fingerprint in green Arabica coffee bean as well as coffee beverage quality. *Food Chemistry, 135*(4), 2575–2583. http://dx.doi.org/10.1016/j.foodchem.2012.06.060.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (1st ed.). New York: Springer.

Borsato, D., Pina, M. V. R., Spacino, K. R., Scholz, M. B. D. S., & Androcioli Filho, A. (2011). Application of artificial neural networks in the geographical identification of coffee samples. *European Food Research and Technology, 233*(3), 533–543. http://dx.doi.org/10.1007/s00217-011-1548-z.

Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *The Analyst, 135*(2), 230–267. http://dx.doi.org/10.1039/b918972f.

Brudzewski, K., Osowki, S., & Markiewicz, T. (2004). Classification of milk by means of an electronic nose and SVM neural network. *Sensors and Actuators, B: Chemical, 98*(2–3), 291–298. http://dx.doi.org/10.1016/j.snb.2003.10.028.

CeCafé, (Conselho dos Exportadores de Café do Brasil). (2015). *Resumo das exportações brasileiras de café: Outubro de 2015*. Retrieved January 13, 2016, from www.cncafe.com.br.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a libray for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), 1–27. http://doi.org/10.1145/1961189.1961199.

Clarke, R. J., & Vitzthum, O. G. (Eds.). (2001). *Coffe: Recent developments*. London: Blackwell Science.

Downey, G., Briandet, R., Wilson, R. H., & Kemsley, E. K. (1997). Near- and mid-infrared spectroscopies in food authentication: coffee varietal identification. *Journal of Agricultural and Food Chemistry, 45*(11), 4357–4361. http://dx.doi.org/10.1021/jf970337t.

Haupt, R. L., & Haupt, S. E. (2004). *Practical genetic algorithms* (2nd ed.). New Jersey: John Wiley & Sons. http://dx.doi.org/10.1198/jasa.2005.s45.

Haykin, S. (2008). *Neural networks and learning machines* (3rd ed.). New York: Prentice Hall.

Hiri, A., De Luca, M., Ioele, G., Balouki, A., El Basbassi, E. M., Kzaiber, F., et al. (2015). Chemometric classification of citrus juices of Moroccan cultivars by infrared spectroscopy. *Czech Journal of Food Sciences, 33*(2), 137–142. http://dx.doi.org/10.17221/284/2014-CJFS.

Isaksson, T., & Naes, T. (1988). The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR spectroscopy. *Applied Spectroscopy, 42*(7), 1273–1284.

Joët, T., Laffargue, A., Descroix, F., Doulbeau, S., Bertrand, B., De Kochko, A., et al. (2010). Influence of environmental factors, wet processing and their interactions on the biochemical composition of green Arabica coffee beans. *Food Chemistry, 118*(3), 693–701. http://dx.doi.org/10.1016/j.foodchem.2009.05.048.

Karoui, R., Downey, G., & Blecker, C. (2010). Mid-infrared spectroscopy coupled with chemometrics: a tool for the analysis of intact food systems and the exploration of their molecular structure-quality relationships-A review. *Chemical Reviews, 110*(10), 6144–6168. http://dx.doi.org/10.1021/cr100090k.

Kennard, R. W., & Stone, L. A. (1969). Computer aided design of experiments. *Technometrics, 11*(1), 137–148.

Lashermes, P., & Anthony, F. (2007). Coffee. In C. Kole (Ed.), *Technical Crops: Genome mapping and molecular breeding in plants* (pp. 109–118). Berlin: Springer Berlin Heidelberg.

Li, H., Liang, Y., & Xu, Q. (2009). Support vector machines and its applications in chemistry. *Chemometrics and Intelligent Laboratory Systems, 95*(2), 188–198. http://dx.doi.org/10.1016/j.chemolab.2008.10.007.

Link, J. V., Guimarães Lemes, A. L., Marquetti, I., dos Santos Scholz, M. B., & Bona, E.

(2014). Geographical and genotypic segmentation of arabica coffee using self-organizing maps. *Food Research International, 59*, 1–7. http://dx.doi.org/10.1016/j.foodres.2014.01.063.

Link, J. V., Lemes, A. L. G., Marquetti, I., dos Santos Scholz, M. B., & Bona, E. (2014). Geographical and genotypic classification of arabica coffee using Fourier transform infrared spectroscopy and radial-basis function networks. *Chemometrics and Intelligent Laboratory Systems, 135*, 150–156. http://dx.doi.org/10.1016/j.chemolab.2014.04.008.

Liu, C., Yang, S. X., & Deng, L. (2015). A comparative study for least angle regression on NIR spectra analysis to determine internal qualities of navel oranges. *Expert Systems with Applications, 42*(22), 8497–8503. http://dx.doi.org/10.1016/j.eswa.2015.07.005.

Marquetti, I., Link, J. V., Lemes, A. L. G., Scholz, M. B. dos S., Valderrama, P., & Bona, E. (2016). Partial least square with discriminant analysis and near infrared spectroscopy for evaluation of geographic and genotypic origin of arabica coffee. *Computers and Electronics in Agriculture, 121*, 313–319. http://dx.doi.org/10.1016/j.compag.2015.12.018.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8*(4), 283–298. http://dx.doi.org/10.1016/S0001-2998(78)80014-2.

Novaes, F. J. M., Oigman, S. S., de Souza, R. O. M. A., Rezende, C. M., & de Aquino Neto, F. R. (2015). New approaches on the analyses of thermolabile coffee diterpenes by gas chromatography and its relationship with cup quality. *Talanta, 139*, 159–166. http://dx.doi.org/10.1016/j.talanta.2014.12.025.

Papadopoulou, O. S., Panagou, E. Z., Mohareb, F. R., & Nychas, G.-J. E. (2013). Sensory and microbiological quality assessment of beef fillets using a portable electronic nose in tandem with support vector machine analysis. *Food Research International, 50*(1), 241–249. http://dx.doi.org/10.1016/j.foodres.2012.10.020.

Pardo, M., & Sberveglieri, G. (2005). Classification of electronic nose data with support vector machines. *Sensors and Actuators, B: Chemical, 107*(2), 730–737. http://dx.doi.org/10.1016/j.snb.2004.12.005.

Rifkin, R., & Klautau, A. (2004). In defense of one-vs-all classification. *The Journal of Machine Learning Research, 5*, 101–141. Retrieved from http://dl.acm.org/citation.cfm?id=1005336.

Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A., & Jent, N. (2007). A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis, 44*(3), 683–700. http://dx.doi.org/10.1016/j.jpba.2007.03.023.

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627–1639.

Scholz, M. B. dos S., Figueiredo, V. R. G., Silva, J. V. N., & Kitzberger, C. S. G. (2011). Características físico-químicas de grãos verdes e torrados de cultivares de café (Coffea arabica L.) do IAPAR. *Coffee Science, 6*(3), 245–255.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199–222. http://dx.doi.org/10.1023/B:STCO.0000035301.49549.88.

Souto, U. T. D. C. P., Barbosa, M. F., Dantas, H. V., de Pontes, A. S., Lyra, W. D. S., Diniz, P. H. G. D., et al. (2015). Identification of adulteration in ground roasted coffees using UV–Vis spectroscopy and SPA-LDA. *LWT – Food Science and Technology, 63*(2), 1037–1041. http://dx.doi.org/10.1016/j.lwt.2015.04.003.

Terouzi, W., De Luca, M., Bolli, A., Oussama, A., Patumi, M., Ioele, G., et al. (2011). A discriminant method for classification of Moroccan olive varieties by using direct FT-IR analysis of the mesocarp section. *Vibrational Spectroscopy, 56*(2), 123–128. http://dx.doi.org/10.1016/j.vibspec.2011.01.004.

Teuber, R. (2010). Geographical indications of origin as a tool of product differentiation: the case of coffee. *Journal of International Food & Agribusiness Marketing, 22*(3–4), 277–298. http://dx.doi.org/10.1080/08974431003641612.

Westad, F., & Marini, F. (2015). Validation of chemometric models – a tutorial. *Analytica Chimica Acta, 893*, 14–24. http://dx.doi.org/10.1016/j.aca.2015.06.056.