Measurement 120 (2018) 92–99

Contents lists available at ScienceDirect

Measurement

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

Geographical recognition of Syrah wines by combining feature selection with Extreme Learning Machine

Nattane Luíza da Costa[b], Laura Andrea García Llobodanin[a], Márcio Dias de Lima[b,c], Inar Alves Castro[a], Rommel Barbosa[b,*]

[a] LADAF – Laboratory of Functional Foods, Department of Food and Experimental Nutrition, Faculty of Pharmaceutical Sciences, University of São Paulo, Av. Lineu Prestes, 580, B14, 05508-900 São Paulo, Brazil
[b] Instituto de Informática, Universidade Federal de Goiás, Goiânia-GO, Brazil
[c] Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Goiânia-GO, Brazil

ARTICLE INFO

Keywords:
Extreme learning machine
Artificial neural networks
Support Vector machines
Feature selection
Wine classification

ABSTRACT

Data mining techniques have been used for the classification of many types of products. In order to classify the Syrah wines from Argentina (Mendoza) and Chile (Central Valley), according to their origin, we perform two feature selection methods with the following classification algorithms: Support Vector Machines (SVM), and two types of artificial neural networks, Multilayer Perceptron (MLP) and Extreme Learning Machine (ELM), on 10-fold cross-validation. Each feature selection method has a different approach, creating also different sets of the most important features. The best model was the combination of variables peon-3-glu, malv-3-glu and pet-3-acetylglu, selected by Random Forest Importance, reaching 98.33% accuracy with ELM, outperforming SVM and MLP. The results obtained from the classifiers and feature subsets are able to confirm the importance of the anthocyanins to classify Syrah wines according to their geographic region. ELM was the best algorithm for classifying Syrah wines.

1. Introduction

Wines are beverages obtained from the alcoholic fermentation of grapes. In 2014 Argentina was considered the fifth largest wine producer in the world and Chile occupied the ninth position in the ranking carried out by the Wine Institute [24]. Among the varieties of Vitis Vinifera grapes, Syrah is one of the most used for wine production [24]. Wines of this variety are considered by the National Institute of Viticulture of Argentina [12] as ”fashionable,” representing 5.81% of all the wines produced in the country. In Chile, the production of these wines is higher, reaching 8.1% of the total amount produced in 2015 according to the Ministry of Agriculture of the Government of Chile [41].

During the last decades, the interest in classifying wines based on their grape variety and geographical origin has risen. The soil, climatic conditions, type of harvest, and production conditions contribute to the characteristics that make a wine unique. This concept is linked to the so-called ’Protected Geographical Status’ defined in European Union law for wines and other food products, which is gradually expanding internationally. The idea is to ensure that only wines originated in a certain region are allowed in the market as such, protecting the reputation of the region and assuring a quality standard. Along with this concept, the need to classify wines in an objective way has emerged in order to perform quality controls and avoid fraud.

Anthocyanins, a group of polyphenols present in wines, proved to be useful for wine classification [2,33,17,32]. They are the main substance responsible for red wine color and they also contribute (together with other groups of polyphenols) to the antioxidant properties of red wine. The Vitis Vinifera grapes generally produce 3-monoglucoside, 3-acetylglucoside and 3-p-coumarylglucoside aglycones derived from the anthocyanins: delphinidin, cyanidin, peonidin, petunidin, and malvidin [6]. The anthocyanic composition of red wines depends on the grape variety, the climatic and soil conditions and the winemaking process. This dependence enables their use for wine classification.

Statistics is the most used tool for wine classification [47]. However, in the 1990s, Fayyad described the knowledge discovery in databases (KDD), which aims to find useful information that would not be possible to find with a simple human analysis [15]. KDD uses data mining process, tools and algorithms that involve areas such as pattern recognition, machine learning, statistical and artificial intelligence [15]. Data mining techniques have been used for authentication and recognition of food, in order to find patterns and relationships providing a

[*] Corresponding author.
E-mail addresses: rommel@inf.ufg.br, rmbweb@gmail.com (R. Barbosa).

https://doi.org/10.1016/j.measurement.2018.01.052
Received 1 March 2017; Received in revised form 5 December 2017; Accepted 29 January 2018
Available online 09 February 2018

description of the data according to its chemical compounds.

Classification of Argentinean wines according to production region were performed in Azcarate et al. [3] which classified white wines from Mendoza, Rio Negro, San Juan, and Salta by their elemental profile and Linear Discriminant Analysis with a precision of 96.00%. Pisano et al. [36] classified red wines of eight varieties from Mendoza, San Juan and San Rafael using Discriminant Partial least-squares with no misclassifications. Chilean wines were classified according to their variety. Beltrán et al. [4] classified Cabernet Sauvignon, Carménère and Merlot varieties using feature extraction with quadratic discriminant analysis, linear discriminant analysis, K-nearest neighbors and probabilistic neural networks based on phenolic compounds. Gutiérrez et al. [18] classified the same varieties using a Multivariate Bayesian classifier applied to the anthocyanin profile.

As far as we know, Syrah wines from different countries have never been the exclusive focus of a classification study. Johnson et al. [25] studied the characterization of Syrah wines from different regions in Australia. Pisano et al. [36] studied the classification of Argentinean wines by geographical origin considering eight grape varieties (Aspiran, Bonarda, Cabernet Sauvignon, Malbec, Merlot, Sangiovese, Syrah, Tempranillo), this classification according to the varieties occurred considering four classes, Cabernet Sauvignon, Malbec, Merlot, and remaining varieties.

The present study aims to find a model able to discriminate Syrah wine from Argentina and Chile. The use of feature selection allows a better understanding of the data, reduces processing time and improves the classification performance [8]. Thus, we combined two methods of feature selection with three classifiers: Support Vector Machines (SVM) and two types of neural networks, Multilayer Perceptron (MLP) and Extreme Learning Machine (ELM). The results from SVM and MLP were compared with those obtained from the novel learning algorithm for single hidden layer feedforward neural networks (SLFNs) proposed by Huang et al. [23], ELM.

The most popular algorithms, SLFNs with back-propagation (BP), consume an enormous amount of computational time due to the need to optimize all the parameters from the neural network [23]. ELM does not require such computational time because their parameters are attributed randomly. This technique is rapid and has a good performance generalization. Some applications using ELM demonstrate satisfactory classification performances in problems of diagnosis of Parkinson's disease [10], breast mass classification in digital mammography [49], classification of coffee, meats, olive oil and fruit [51], prediction of daily dew point temperature [34], speaker recognition [27] and estimation of building energy consumption [35]. In addition, ELM has lower computational time, better performance, and generalization ability than the conventional classifiers in applications of pattern recognition, forecasting and diagnosis, image processing and other areas [13].

The main contributions and the remainder of this paper can be summarized as follows: (1) Syrah wine classification in two different countries, Argentina and Chile; (2) the use of feature selection to identify the variables that best discriminate Argentineans from Chilean wines; (3) a comparison of SVM, MLP and ELM in a new type of problem, wine classification.

2. Material and methods

The data analysis of Chilean and Argentine wines occurred as follows: (1) we extracted values related to characteristics of wines and organized them in a database; (2) we performed the steps of data preprocessing taking samples with missing values and normalizing them on a scale from zero to one; (3) we applied the feature selection techniques and combined them with classifiers using 10-fold cross-validation; (4) we interpreted the data and reported results. All tests were performed with the R software [37]. R is a free software environment for statistical computing and graphics that contains a wide variety of packages to perform statistical analysis. In our experiments we used the FSelector package [39] for feature selection, the caret package [26] for data classification and the ggplot2 package [48] to visualize some of the results. The following is a brief description of the data and each of the techniques used.

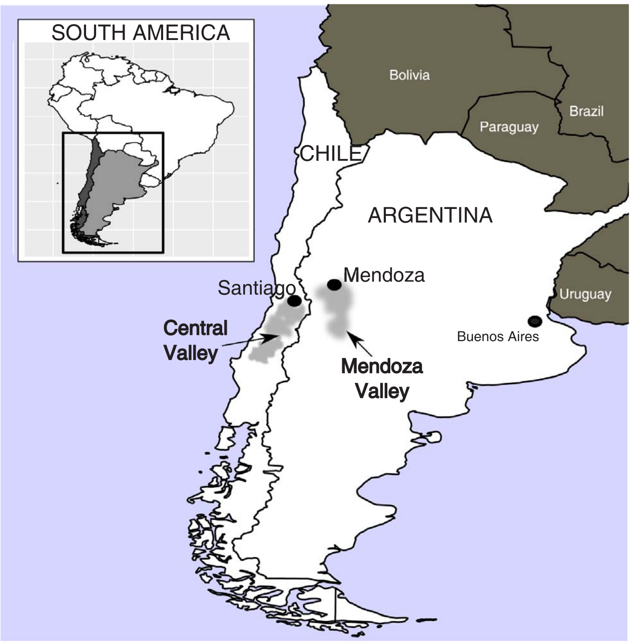

Fig. 1. Map of regions of the Central Valley (Chile) and Mendoza (Argentina).

2.1. Wine samples

Syrah wine samples were obtained from local markets and wine distributors in the city of São Paulo (Brazil). All the wines are monovarietal (at least 75% of Syrah variety), from 2009 and 2010 vintages, bottled in 750 mL bottles and with retail prices of 1–50 United States dollars (USD). Twenty-six samples were from Argentina (Mendoza) and 11 from Chile (Central Valley) which were described by 20 features. Fig. 1 shows the location of Mendoza and Central Valley, the main wine producing regions of Argentina and Chile, respectively.

The features describe components as the color of wine (L, a., b.), total polyphenols (TPI), total anthocyanins (TA), antioxidant activity by oxygen radical absorbance capacity (ORAC) and free radical scavenging activity (DPPH), and anthocyanins (cyan-3-glu, delph-3-acetylglu, delph-3-glu, malv-3- (coum) glu, malv-3-acetylglu, malv-3-glu, peon-3-(coum) glu, peon- 3-acetylglu, peon-3-glu, pet-3- (coum) glu, pet-3-acetylglu, pet-3-glu and vitisin A). Analyses were performed as described by Llobodanin et al. [30], as follows.

2.2. Colour determination

The analysis of color was performed by measuring the transmittance in a ColorQuest XE colorimeter (Hunter Associates Laboratory, Inc., Reston, USA) using the CIE 1964 standard observer (10° visual field) and the CIE standard illuminant D65 as references. The three CIELAB coordinates a^* (red-green; $+a^*,-a^*$), b^* (yellow-blue; $+b^*,-b^*$) and lightness L (white-black, 0–100) were determined using the software EasyMatch QC (Hunter Associates Laboratory, Inc., Reston, USA). The analyses were performed in triplicate.

2.3. Total polyphenols

Total polyphenols were determined by the Folin–Ciocalteu colorimetric method [45], adapted for measurement with a microplate