Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Original papers

Classification of geographic origin of rice by data mining and inductively coupled plasma mass spectrometry

Camila Maione [a], Bruno Lemos Batista [b], Andres Dobal Campiglia [c], Fernando Barbosa Jr [d], Rommel Melgaço Barbosa [a,*]

[a] Instituto de Informática, Universidade Federal de Goiás, Goiânia, Goiás, Brazil
[b] Centre for Natural Sciences and Humanities, Federal University of ABC, Santo André, São Paulo, Brazil
[c] Department of Chemistry, University of Central Florida, Orlando, USA
[d] Department of Clinical Analyses, Toxicology and Food Sciences, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, São Paulo, Brazil

ARTICLE INFO

Article history:
Received 7 July 2015
Received in revised form 3 October 2015
Accepted 16 November 2015
Available online 29 December 2015

Keywords:
Data mining
ICP-MS
Rice
Classification
Support vector machines
F-score

ABSTRACT

Rice is one of the most consumed cereals in the world and the main food product in the diet of the Brazilian population. Brazil itself is among the ten largest producers of rice, and most of the harvest comes from the South and Midwest regions. This paper presents a data mining study of samples of rice obtained from producers in Goiás (Midwest region) and Rio Grande do Sul (South region), and builds classification models capable of predicting the geographical origin of a rice sample based on its chemical components. We use three popular classification techniques, support vector machines, random forests and neural networks, along with the F-score formula which measures the relative importance of the input variables. We achieved very good performances for the SVM, RF and MLP models with 93.66%, 93.83% and 90% prediction accuracy, respectively, on the 10-fold cross validation. The F-score shows that Cd(cadmium), Rb(rubidium), Mg(magnesium) and K(potassium) are the four most relevant components for prediction.

1. Introduction

Rice is one of the most consumed cereal grains in the world. Brazil is among the ten largest rice producers and, according to the Brazilian Agricultural Research Corporation (Embrapa), 11.26 million tons of rice were harvested in 2009/2010. In 2001, Brazilian production accounted for 1.8% of the world total and about 50% in Latin America. In 2005, Brazil exported around 272,000 tons of rice. Today, only 5% of national production is exported. The Brazilian population spends about 22% of its budget on food, and rice is the main product of a typical family's groceries. The annual consumption is on average 25 kilograms per inhabitant.

Most of the rice produced in Brazil comes from the South and Midwest regions, respectively. The production comes from two cropping systems: irrigated and rainfed. The state of Rio Grande do Sul is the largest producer of irrigated rice. Rainfed rice is produced in uplands and concentrated in the Midwest region (Mato Grosso and Goiás), Northeast (Piauí and Maranhão) and North (Pará and Rondônia). The chemical composition of rice grains changes depending on weather conditions, cultivation, pre-storage, storage and processing system (Embrapa, 2005; Miranda et al., 2007).

The classification of food samples based on their chemical composition is an interesting problem and provides useful information for a variety of purposes, such as recognition of geographical origin and authenticity, the characteristics of a product, quality control for companies, preservation, and category differentiation (Barbosa et al., 2014, 2015). Moreover, the analysis of chemical components in rice samples have been approached by a lot of recent studies aiming for different purposes (Marini et al., 2004; Suzuki et al., 2008; Saruta et al., 2013; Cheajesadagul et al., 2013).

A simple method for determination of geographical origin of polished rice Koshihikari, using multiple element and stable isotope analyses, is proposed by Suzuki et al. (2008). A total of 14 samples obtained from Australia, Japan and USA were used in the study. Elements and isotopic compositions extracted from the samples were C, N, $\sigma^1$3C, $\sigma^1$3N and $\sigma^1$3O. The presented method finds a pentagonal radar plot capable of clearly distinguishing the cultivated area based on the elemental and isotopic compositions, proving these components useful for rapid and routine discrimination of the geographical origin of the polished rice. However, only

* Corresponding author.
E-mail address: rommel@inf.ufg.br (R.M. Barbosa).

http://dx.doi.org/10.1016/j.compag.2015.11.009
0168-1699/© 2015 Elsevier B.V. All rights reserved.

one sample was collected from Australia and USA, which is not an adequate sample for assessing the natural variability of rice samples from these two countries.

In another recent research project, rice geographic origin determination is achieved using radar plots generated by high resolution inductively coupled plasma mass spectometry (HR-ICP-MS) multi-element fingerprinting, combined with multivariate statistical analysis, principal component analysis (PCA) and discriminant analysis (DA) (Cheajesadagul et al., 2013). The study was carried out with 31 samples from Thai jasmine white rice and 5 foreign rice samples (collected from France, India, Italy, Japan and Pakistan). Isotopes of 21 elements were monitored by HR-ICP-MS and used as descriptors for the rice samples, and all techniques used were able to create clear separation of Thai jasmine rice from the foreign rice samples. Particularly, DA achieved 100% prediction performance for all rice samples using 12 variables, and 90.32% prediction performance for Thai jasmine rice samples using 7 variables.

With the goal of providing knowledge and aid for farmer's management and decision-making, Saruta et al. (2013) developed predictive models for yield and protein content of brown rice from regions in Fukuoka Prefecture, Japan, based on support vector machines (SVMs). The models predicted three classes (low, middle and high) of yield and protein content of brown rice using variables which represent the growth and nutrition conditions after the heading stage and the meteorological environment after the late spikelet initiation stage. The models achieved good performance, but little information on the variables' behavior, correlation and variance was offered, which could be better explained if a proper feature selection method were used.

Italian rice classification based on its physical characteristics is approached by Marini et al. (2004). A counter propagation artificial neural network (CP-ANN) is used for classification, along with PCA in order to find the most suitable clustering representation of the 1779 rice samples of 11 different varieties in their dataset. All the samples were obtained from the northern region of Italy, and the eight extracted variables described information regarding the shape, uniformity and processing quality of the grains, instead of using chemical components to describe the rice samples. The CP-ANN model presented prediction performance between 91 and 99% of the test set samples.

The present project brings a data mining study for classification of rice samples from Goiás and Rio Grande do Sul, states from Brazil's Midwest and South regions, respectively, based on their chemical components. We use three popular data mining techniques (support vector machines, random forest and neural network) along with F-score (Chen and Lin, 2006) for variable evaluation and selection. The main objectives of the study are:

- to provide a classification model capable of predicting in which Brazilian region (Midwest or South) a determined rice sample was produced;
- to confirm authentication of rice samples;
- to help understand which chemical characteristics are significant in distinguishing rice samples from the two regions.

In comparison to the research studies in the recent literature listed above, our study presents the following advantages:

- We analyze rice samples obtained from different regions in the same country.
- We employ data mining methods such as classification models and feature selectors, which can capture hidden patterns and variance inside the data better than simple observation and traditional statistical methods.
- We work with a balanced dataset, i.e., similar number of samples obtained from each region. In an unbalanced dataset, a classifier might ignore the importance of the minority class because its representation inside the dataset is not strong enough and the classifier is biased toward the majority class. Consequently, the examples that belong to the minority class are misclassified more often than those belonging to the majority class (Sáez et al., 2015).
- We use a feature selection method to rank the descriptor variables and access chemical component subsets that are capable of discriminating the rice samples and provide good prediction results. It is useful in order to understand how the chemical component concentrations vary among the rice samples from the two regions, which components are most useful to discriminate them, which components are homogeneous, which components are most frequent in the rice sample from a determined region, and which components are most useful in order to discriminate the samples from the two regions.

2. Materials and methods

In this section we detail our adopted methodology, including description of our data, identification of variables, data preparation and a brief explanation about supervised machine learning and data mining techniques employed in the analysis.

2.1. The rice dataset

Our dataset includes 31 white rice samples of Oriza Sativa variety, which is the most common rice in Brazil. These samples were obtained from the Midwest and South regions in Brazil, which use the irrigated and rainfed systems, respectively, to cultivate rice. We have 12 samples obtained from Goiás state (Midwest region) and 19 samples obtained from Rio Grande do Sul state (South region).

In order to obtain its chemical components, each sample went through the following decomposition process. Before digestion, samples (15 g) were separated by quartering and were divided into three plastic tubes. The contents of each tube were ground for 3 min in a ball mill (TECNAL TE 350, Brazil) and sifted in a 106 μm sieve (BERTEL, Brazil). Then, samples were digested in closed vessels using a microwave oven decomposition system (MILESTONE ETHOS D, Italy) according to the process used by Batista et al. (2010). Determination of elements in rice samples was carried out with an inductively coupled plasma mass spectrometer equipped with a reaction cell (DRC-ICP-MS ELAN DRCII, PerkinElmer, SCIEX, Norwalk, CT, USA) operating with high-purity argon (99.999%, Praxaair, Brazil). Instrumental settings and operative conditions are reported in (Batista et al., 2010). In order to verify the accuracy and precision of the proposed method, the Standard Reference Material SRM 1568a Rice Flour (from the National Institute of Standards and Technology -NIST) was digested and analyzed. Obtained values are in good agreement with target values.

This process resulted in a total of 21 components found in each rice sample, the same components used by Cheajesadagul et al. (2013), which are: copper (Cu), zinc (Zn), magnesium (Mg), boron (B), phosphorus (P), molybdenum (Mo), arsenic (As), lead (Pb), cadmium (Cd), manganese (Mn), selenium (Se), cobalt (Co), chromium (Cr), barium (Ba), rubidium (Rb), iron (Fe), potassium (K), calcium (Ca), lanthanum (La) and cerium (Ce). The concentrations found for each element were set as our descriptive variables. Table 2 shows the chemical components identified along with their median, minimum and maximum concentration for the samples from Rio Grande do Sul and Goiás.