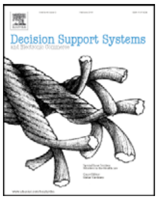
Contents lists available at ScienceDirect

Decision Support Systems

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



Modeling wine preferences by data mining from physicochemical properties

Paulo Cortez [a,*], António Cerdeira [b], Fernando Almeida [b], Telmo Matos [b], José Reis [a,b]

[a] *Department of Information Systems/R&D Centre Algoritmi, University of Minho, 4800-058 Guimarães, Portugal*
[b] *Viticulture Commission of the Vinho Verde Region (CVRVV), 4050-501 Porto, Portugal*

ARTICLE INFO

Article history:
Received 28 July 2008
Received in revised form 22 May 2009
Accepted 28 May 2009
Available online 6 June 2009

Keywords:
Sensory preferences
Regression
Variable selection
Model selection
Support vector machines
Neural networks

ABSTRACT

We propose a data mining approach to predict human wine taste preferences that is based on easily available analytical tests at the certification step. A large dataset (when compared to other studies in this domain) is considered, with white and red *vinho verde* samples (from Portugal). Three regression techniques were applied, under a computationally efficient procedure that performs simultaneous variable and model selection. The support vector machine achieved promising results, outperforming the multiple regression and neural network methods. Such model is useful to support the oenologist wine tasting evaluations and improve wine production. Furthermore, similar techniques can help in target marketing by modeling consumer tastes from niche markets.

1. Introduction

Once viewed as a luxury good, nowadays wine is increasingly enjoyed by a wider range of consumers. Portugal is a top ten wine exporting country, with 3.17% of the market share in 2005 [11]. Exports of its *vinho verde* wine (from the northwest region) have increased by 36% from 1997 to 2007 [8]. To support its growth, the wine industry is investing in new technologies for both wine making and selling processes. Wine certification and quality assessment are key elements within this context. Certification prevents the illegal adulteration of wines (to safeguard human health) and assures quality for the wine market. Quality evaluation is often part of the certification process and can be used to improve wine making (by identifying the most influential factors) and to stratify wines such as premium brands (useful for setting prices).

Wine certification is generally assessed by physicochemical and sensory tests [10]. Physicochemical laboratory tests routinely used to characterize wine include determination of density, alcohol or pH values, while sensory tests rely mainly on human experts. It should be stressed that taste is the least understood of the human senses [25] thus wine classification is a difficult task. Moreover, the relationships between the physicochemical and sensory analysis are complex and still not fully understood [20].

Advances in information technologies have made it possible to collect, store and process massive, often highly complex datasets. All this data hold valuable information such as trends and patterns, which can be used to improve decision making and optimize chances of success [28]. Data mining (DM) techniques [33] aim at extracting high-level knowledge from raw data. There are several DM algorithms, each one with its own advantages. When modeling continuous data, the linear/multiple regression (MR) is the classic approach. The backpropagation algorithm was first introduced in 1974 [32] and later popularized in 1986 [23]. Since then, neural networks (NNs) have become increasingly used. More recently, support vector machines (SVMs) have also been proposed [4,26]. Due to their higher flexibility and nonlinear learning capabilities, both NNs and SVMs are gaining an attention within the DM field, often attaining high predictive performances [16,17]. SVMs present theoretical advantages over NNs, such as the absence of local minima in the learning phase. In effect, the SVM was recently considered one of the most influential DM algorithms [34]. While the MR model is easier to interpret, it is still possible to extract knowledge from NNs and SVMs, given in terms of input variable importance [18,7].

When applying these DM methods, variable and model selection are critical issues. Variable selection [14] is useful to discard irrelevant inputs, leading to simpler models that are easier to interpret and that usually give better performances. Complex models may overfit the data, losing the capability to generalize, while a model that is too simple will present limited learning capabilities. Indeed, both NN and SVM have hyperparameters that need to be adjusted [16], such as the number of NN hidden nodes or the SVM kernel parameter, in order to get good predictive accuracy (see Section 2.3).

The use of decision support systems by the wine industry is mainly focused on the wine production phase [12]. Despite the potential of DM techniques to predict wine quality based on physicochemical data,

[*] Corresponding author. Tel.: +351 253510313; fax: +351 253510300.
E-mail address: pcortez@dsi.uminho.pt (P. Cortez).

0167-9236/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.dss.2009.05.016

their use is rather scarce and mostly considers small datasets. For example, in 1991 the "Wine" dataset was donated into the UCI repository [1]. The data contain 178 examples with measurements of 13 chemical constituents (e.g. alcohol, Mg) and the goal is to classify three cultivars from Italy. This dataset is very easy to discriminate and has been mainly used as a benchmark for new DM classifiers. In 1997 [27], a NN fed with 15 input variables (e.g. Zn and Mg levels) was used to predict six geographic wine origins. The data included 170 samples from Germany and a 100% predictive rate was reported. In 2001 [30], NNs were used to classify three sensory attributes (e.g. sweetness) of Californian wine, based on grape maturity levels and chemical analysis (e.g. titrable acidity). Only 36 examples were used and a 6% error was achieved. Several physicochemical parameters (e.g. alcohol, density) were used in [20] to characterize 56 samples of Italian wine. Yet, the authors argued that mapping these parameters with a sensory taste panel is a very difficult task and instead they used a NN fed with data taken from an electronic tongue. More recently, mineral characterization (e.g. Zn and Mg) was used to discriminate 54 samples into two red wine classes [21]. A probabilistic NN was adopted, attaining 95% accuracy. As a powerful learning tool, SVM has outperformed NN in several applications, such as predicting meat preferences [7]. Yet, in the field of wine quality only one application has been reported, where spectral measurements from 147 bottles were successfully used to predict 3 categories of rice wine age [35].

In this paper, we present a case study for modeling taste preferences based on analytical data that are easily available at the wine certification step. Building such model is valuable not only for certification entities but also wine producers and even consumers. It can be used to support the oenologist's wine evaluations, potentially improving the quality and speed of their decisions. Moreover, measuring the impact of the physicochemical tests in the final wine quality is useful for improving the production process. Furthermore, it can help in target marketing [24], i.e. by applying similar techniques to model the consumer's preferences of niche and/or profitable markets.

The main contributions of this work are:

- We present a novel method that performs simultaneous variable and model selection for NN and SVM techniques. The variable selection is based on sensitivity analysis [18], which is a computationally efficient method that measures input relevance and guides the variable selection process. Also, we propose a parsimony search method to select the best SVM kernel parameter with a low computational effort.
- We test such approach in a real-world application, the prediction of *vinho verde* wine (from the Minho region of Portugal) taste preferences, showing its impact in this domain. In contrast with previous studies, a large dataset is considered, with a total of 4898 white and 1599 red samples. Wine preferences are modeled under a regression approach, which preserves the order of the grades, and we show how the definition of the tolerance concept is useful for accessing different performance levels. We believe that this integrated approach is valuable to support applications where ranked sensory preferences are required, for example in wine or meat quality assurance.

The paper is organized as follows: Section 2 presents the wine data, DM models and variable selection approach; in Section 3, the experimental design is described and the obtained results are analyzed; finally, conclusions are drawn in Section 4.

2. Materials and methods

2.1. Wine data

This study will consider *vinho verde*, a unique product from the Minho (northwest) region of Portugal. Medium in alcohol, is it particularly appreciated due to its freshness (specially in the summer). This wine accounts for 15% of the total Portuguese production [8], and around 10% is exported, mostly white wine. In this work, we will analyze the two most common variants, white and red (rosé is also produced), from the demarcated region of *vinho verde*. The data were collected from May/2004 to February/2007 using only protected designation of origin samples that were tested at the official certification entity (CVRVV). The CVRVV is an inter-professional organization with the goal of improving the quality and marketing of *vinho verde*. The data were recorded by a computerized system (iLab), which automatically manages the process of wine sample testing from producer requests to laboratory and sensory analysis. Each entry denotes a given test (analytical or sensory) and the final database was exported into a single sheet (.csv).

During the preprocessing stage, the database was transformed in order to include a distinct wine sample (with all tests) per row. To avoid discarding examples, only the most common physicochemical tests were selected. Since the red and white tastes are quite different, the analysis will be performed separately, thus two datasets[1] were built with 1599 red and 4898 white examples. Table 1 presents the physicochemical statistics per dataset. Regarding the preferences, each sample was evaluated by a minimum of three sensory assessors (using blind tastes), which graded the wine in a scale that ranges from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations. Fig. 1 plots the histograms of the target variables, denoting a typical normal shape distribution (i.e. with more normal grades that extreme ones).

2.2. Data mining approach and evaluation

We will adopt a regression approach, which preserves the order of the preferences. For instance, if the true grade is 3, then a model that predicts 4 is better than one that predicts 7. A regression dataset D is made up of k $\{1,\ldots, N\}$ examples, each mapping an input vector with I input variables $(x_1^k,\ldots, x_I^k)$ to a given target y_k. The regression performance is commonly measured by an error metric, such as the mean absolute deviation (MAD) [33]:

$$MAD = \sum_{i=1}^{N} |y_i - \hat{y}_i| \,/\, N \tag{1}$$

where $\hat{y}_k$ is the predicted value for the k input pattern. The regression error characteristic (REC) curve [2] is also used to compare regression models, with the ideal model presenting an area of 1.0. The curve plots the absolute error tolerance T (x-axis), versus the percentage of points correctly predicted (the accuracy) within the tolerance (y-axis).

The confusion matrix is often used for classification analysis, where a $C \times C$ matrix (C is the number of classes) is created by matching the predicted values (in columns) with the desired classes (in rows). For an ordered output, the predicted class is given by $p_i = y_i$, if $|y_i - \hat{y}_i| \le T$, else $p_i = y_i'$, where y_i' denotes the closest class to $\hat{y}_i$, given that $y_i' \ne y_i$. From the matrix, several metrics can be used to access the overall classification performance, such as the accuracy and precision (i.e. the predicted column accuracies) [33].

The holdout validation is commonly used to estimate the generalization capability of a model [19]. This method randomly partitions the data into training and test subsets. The former subset is used to fit the model (typically with 2/3 of the data), while the latter (with the remaining 1/3) is used to compute the estimate. A more robust estimation procedure is the k-fold cross-validation [9], where the data is divided into k partitions of equal size. One subset is tested each time and the remaining data are used for fitting the model. The process is repeated sequentially until all subsets have been tested. Therefore,

[1] The datasets are available at: http://www3.dsi.uminho.pt/pcortez/wine/.