ScienceDirect

Decision Support Systems
Volume 47, Issue 4, November 2009, Pages 547-553

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]

https://doi.org/10.1016/j.dss.2009.05.016
Get rights and content

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.

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, 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.

Section snippets

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 …

Empirical results

The **R** environment [22] is an open source, multiple platform (e.g. Windows, Linux) and high-level matrix programming language for statistical and data analysis. All experiments reported in this work were written in **R** and conducted in a Linux server, with an Intel dual core processor. In particular, we adopted the **RMiner** [6], a library for the **R** tool that facilitates the use of DM techniques in classification and regression tasks.

Before fitting the models, the data was first standardized to a …

Conclusions and implications

In recent years, the interest in wine has increased, leading to growth of the wine industry. As a consequence, companies are investing in new technologies to improve wine production and selling. Quality certification is a crucial step for both processes and is currently largely dependent on wine tasting by human experts. This work aims at the prediction of wine preferences from objective analytical tests that are available at the certification step. A large dataset (with 4898 white and 1599 red …

Acknowledgments

We would like to thank Cristina Lagido and the anonymous reviewers for their helpful comments. The work of P. Cortez is supported by the FCT project PTDC/EIA/64541/2006. …

Paulo Cortez has a PhD (2002) from University of Minho in Computer Science. He is a lecturer at the Department of Information Systems of the same university and a researcher at the Algoritmi Centre, with interests in the fields of: business intelligence, data mining, neural networks, evolutionary computation and forecasting. Currently, he is an associate editor of the Neural Processing Letters journal and he participated in 7 R&D projects (principal investigator in 2). His research has appeared …

…

…

Special issue articles Recommended articles

References (36)

J. Ferrer *et al.*
An optimization approach for scheduling wine grape harvest operations
International Journal of Production Economics (2008)

Z. Huang *et al.*
Credit rating analysis with support vector machines and neural networks: a market comparative study
Decision Support Systems (2004)

M. Kiang
A comparative assessment of classification methods
Decision Support Systems (2003)

I. Moreno *et al.*
Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks
Talanta (2007)

M. Shaw *et al.*
Knowledge management and data mining for marketing
Decision Support Systems (2001)

W. Wang *et al.*
Determination of the spread parameter in the Gaussian kernel for classification and regression
Neurocomputing (2003)

A. Asuncion *et al.*
UCI Machine Learning Repository(2007)

J. Bi *et al.*
Regression error characteristic curves

C. Bishop
Neural Networks for Pattern Recognition(1995)

B. Boser *et al.*
A training algorithm for optimal margin classifiers

View more references

Cited by (994)

A data-driven approach to predict the success of bank telemarketing
2014, Decision Support Systems

Citation Excerpt :
…When comparing DT, NN and SVM, several studies have shown different classification performances. For instance, SVM provided better results in Refs. [6,8], comparable NN and SVM performances were obtained in Ref. [5], while DT outperformed NN and SVM in Ref. [24]. These differences in performance emphasize the impact of the problem context and provide a strong reason to test several techniques when addressing a problem before choosing one of them [9].…

Show abstract

Optimal design of fuzzy classification systems using PSO with dynamic parameter adaptation through fuzzy logic
2013, Expert Systems with Applications

Show abstract

Using sensitivity analysis and visualization techniques to open black box data mining models

2013, Information Sciences

Citation Excerpt :

...The servo dataset corresponds to a nonlinear task related to rise time of a servomechanism, including 167 examples and four inputs (two nominal and two continuous). Finally, the wwq dataset contains 4898 wine entries and the regression goal is to predict human taste preferences, within a scale ranging from 3 – poor quality to 9 – excellent quality, based on 11 analytical continuous inputs (e.g., alcohol) [7]. All datasets are publicly available at: http://www3.dsi.uminho.pt/pcortez/data....

Show abstract

AI in marketing, consumer research and psychology: A systematic literature review and research agenda

2022, Psychology and Marketing

Using machine teaching to identify optimal training-set attacks on machine learners

2015, Proceedings of the National Conference on Artificial Intelligence

Correspondence Analysis: Theory, Practice and New Strategies

2014, Correspondence Analysis: Theory, Practice and New Strategies

View all citing articles on Scopus

Paulo Cortez has a PhD (2002) from University of Minho in Computer Science. He is a lecturer at the Department of Information Systems of the same university and a researcher at the Algoritmi Centre, with interests in the fields of: business intelligence, data mining, neural networks, evolutionary computation and forecasting. Currently, he is an associate editor of the Neural Processing Letters journal and he participated in 7 R&D projects (principal investigator in 2). His research has appeared in Journal of Heuristics, Journal of Decision Systems, Artificial Intelligence in Medicine, Neurocomputing, Neural Processing Letters, and others (see http://www.dsi.uminho.pt/~pcortez).

António Cerdeira graduated (1995) with a degree in Oenology from the University of Trás-os-Montes e Alto Douro. Currently, he is responsible for the Chemical Laboratory and for Oenological Experimentation of the Viticulture Commission of the Vinho Verde Region (CVRVV). Since 1997 he is a member of the Portuguese Group of Oenology from the OIV (International Organization of Vine and Wine) and since 2000 is the president of the ALABE — Association of Oenological Laboratories from Portugal.

Fernando Almeida has a degree in Biological Engineering (2003) from the University of Minho. Between 2003 and 2004 he participated in an R&D project in physicochemical and microbiological analysis, at the Centre of Biological Engineering of the same university. Since 2004, he is part of the sensory analysis panel of CVRVV and has been working in the accreditation of the sensory testing.

Telmo Matos has a degree in Applied Mathematics (2006) from the University of Porto. He currently works in the Information Systems Department of CVRVV.

José Reis received his MSc (2000) in Information Systems from the Portucalense University and he is currently the director of the Information Systems Department of CVRVV, and a lecturer at the IPAM and ISMAI institutes. He is also a PhD student at the Department of Information Systems of University of Minho, with research interests in the fields of personalized information systems, marketing information systems and data mining. He his the author of the book "Personalized Marketing and Information Technology".

View full text