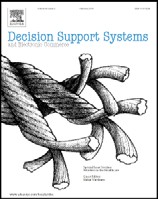
Decision Support Systems 47 (2009) 547–533

Contents lists available at [ScienceDirect](http://www.sciencedirect.com/science/journal/01679236)



Decision Support Systems

journal h omepage: [www.else v ier. com/locate/dss](http://www.elsevier.com/locate/dss)

Modeling wine preferences by data mining from physicochemical properties

Paulo Cortez [a](#_bookmark0),[⁎](#_bookmark2), António Cerdeira [b](#_bookmark1), Fernando Almeida [b](#_bookmark1), Telmo Matos [b](#_bookmark1), José Reis [a](#_bookmark0),[b](#_bookmark1)

1. Department of Information Systems/R&D Centre Algoritmi, University of Minho, 4800-058 Guimarães, Portugal
2. Viticulture Commission of the Vinho Verde Region (CVRVV), 4050-501 Porto, Portugal

# a r t i c l e i n f o

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

# a b s t r a c t

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.

© 2009 Elsevier B.V. All rights reserved.

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]](#_bookmark23). Exports of its vinho verde wine (from the northwest region) have increased by 36% from 1997 to 2007 [[8]](#_bookmark20). 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]](#_bookmark21). 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]](#_bookmark23) thus wine classiﬁcation is a difﬁcult task. Moreover, the relationships between the physicochemical and sensory analysis are complex and still not fully understood [[20]](#_bookmark23).

Advances in information technologies have made it possible to collect, store and process massive, often highly complex datasets. All

⁎ Corresponding author. Tel.: +351 253510313; fax: +351 253510300.

E-mail address: [pcortez@dsi.uminho.pt](mailto:pcortez@dsi.uminho.pt) (P. Cortez).

this data hold valuable information such as trends and patterns, which can be used to improve decision making and optimize chances of success [[28]](#_bookmark26). Data mining (DM) techniques [[33]](#_bookmark31) 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]](#_bookmark30) and later popularized in 1986 [[23]](#_bookmark23). Since then, neural networks (NNs) have become increasingly used. More recently, support vector machines (SVMs) have also been proposed [[4,26]](#_bookmark13). 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]](#_bookmark23). 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]](#_bookmark32). 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]](#_bookmark23).

When applying these DM methods, variable and model selection are critical issues. Variable selection [[14]](#_bookmark23) 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]](#_bookmark23), such as the number of NN hidden nodes or the SVM kernel parameter, in order to get good predictive accuracy (see [Section 2.3](#_bookmark6)).

The use of decision support systems by the wine industry is mainly focused on the wine production phase [[12]](#_bookmark23). Despite the potential of DM techniques to predict wine quality based on physicochemical data,

0167-9236/$ – see front matter © 2009 Elsevier B.V. All rights reserved. doi:[10.1016/j.dss.2009.05.016](http://dx.doi.org/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]](#_bookmark14). 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]](#_bookmark25), 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]](#_bookmark28), 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]](#_bookmark23) 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 characteriza- tion (e.g. Zn and Mg) was used to discriminate 54 samples into two red wine classes [[21]](#_bookmark23). 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]](#_bookmark15). 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]](#_bookmark33).

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]](#_bookmark23), i.e. by applying similar techniques to

model the consumer's preferences of niche and/or proﬁtable markets.

summer). This wine accounts for 15% of the total Portuguese production [[8]](#_bookmark20), 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-profes- sional 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 datasets1 were built with 1599 red and 4898 white examples. [Table 1](#_bookmark4) 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](#_bookmark5) 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,…, N} examples, each mapping an input vector with I

input variables (x1,…, xI ) to a given target yk. The regression

k k

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

performance is commonly measured by an error metric, such as the mean absolute deviation (MAD) [[33]](#_bookmark31):

ˆ

selection is based on sensitivity analysis [[18]](#_bookmark23), which is a computa-

MAD = XN

j y − y j = N ð1Þ

tionally efﬁcient method that measures input relevance and guides

i = 1 i i

the variable selection process. Also, we propose a parsimony search

where yk̂

is the predicted value for the k input pattern. The regression

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](#_bookmark3) presents the wine data, DM models and variable selection approach; in [Section 3](#_bookmark6), the experimental design is described and the obtained results are analyzed; ﬁnally, conclusions are drawn in [Section 4](#_bookmark10).

1. Materials and methods
   1. Wine data

error characteristic (REC) curve [[2]](#_bookmark16) 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 × 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 pi = yi,

if |yi − yî | ≤ T, else pi = y′i, where y′i denotes the closest class to yî , given

that y′i ≠ yi. 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]](#_bookmark31).

The holdout validation is commonly used to estimate the general- ization capability of a model [[19]](#_bookmark23). 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]](#_bookmark22), 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,

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

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

Table 1

The physicochemical data statistics per wine type.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Attribute (units) | Red wine  Min | Max | Mean | White wine |  | |
| Min | Max | Mean |
| Fixed acidity (g(tartaric acid)/dm3) | 4.6 | 15.9 | 8.3 | 3.8 | 14.2 | 6.9 |
| Volatile acidity (g(acetic acid)/dm3) | 0.1 | 1.6 | 0.5 | 0.1 | 1.1 | 0.3 |
| Citric acid (g/dm3) | 0.0 | 1.0 | 0.3 | 0.0 | 1.7 | 0.3 |
| Residual sugar (g/dm3) | 0.9 | 15.5 | 2.5 | 0.6 | 65.8 | 6.4 |
| Chlorides (g(sodium chloride)/dm3) | 0.01 | 0.61 | 0.08 | 0.01 | 0.35 | 0.05 |
| Free sulfur dioxide (mg/dm3) | 1 | 72 | 14 | 2 | 289 | 35 |
| Total sulfur dioxide (mg/dm3) | 6 | 289 | 46 | 9 | 440 | 138 |
| Density (g/cm3) | 0.990 | 1.004 | 0.996 | 0.987 | 1.039 | 0.994 |
| pH | 2.7 | 4.0 | 3.3 | 2.7 | 3.8 | 3.1 |
| Sulphates (g(potassium sulphate)/dm3) | 0.3 | 2.0 | 0.7 | 0.2 | 1.1 | 0.5 |
| Alcohol (vol.%) | 8.4 | 14.9 | 10.4 | 8.0 | 14.2 | 10.4 |

under this scheme, all data are used for training and testing. However, this method requires around k times more computation, since k models are ﬁtted.

* 1. Data mining methods

We will adopt the most common NN type, the multilayer perceptron, where neurons are grouped into layers and connected by feedforward links [[3]](#_bookmark17). For regression tasks, this NN architecture is often based on one hidden layer of H hidden nodes with a logistic activation and one output node with a linear function [[16]](#_bookmark23):

hyperplane, tolerating a small error (ε) when ﬁtting the data, in the feature space:

m

yˆ = w0 + X wi /iðxÞ ð3Þ

i =1

The ε-insensitive loss function sets an insensitive tube around the residuals and the tiny errors within the tube are discarded ([Fig. 2](#_bookmark6)).

We will adopt the popular Gaussian kernel, which presents less parameters than other kernels (e.g. polynomial) [[31]](#_bookmark29): K(x, x′)= exp(−γ||x −x′||2), γ N 0. Under this setup, the SVM performance is affected by three parameters: γ, ε and C (a trade-off between ﬁtting

o − 1

yˆ = wo;0 + X

j = I + 1 1 + exp − PI

1

1 · wo;i ð2Þ

the errors and the ﬂatness of the mapping). To reduce the search space, the ﬁrst two values will be set using the heuristics [[5]](#_bookmark18): C =3 (for a

i = 1 xi wj;i − wj;0

standardized output) and e = σˆ = pﬃNﬃﬃﬃ, where σ̂= 1.5 /N ×∑N (y −ŷ )2

i =1 i i

where wi,j denotes the weight of the connection from node j to i and o the output node. The performance is sensitive to the topology choice (H). A NN with H = 0 is equivalent to the MR model. By increasing H, more complex mappings can be performed, yet an excess value of H will overﬁt the data, leading to generalization loss. A computationally efﬁcient method to set H is to search through the range {0, 1, 2, 3,…, Hmax} (i.e. from the simplest NN to more complex ones). For each H value, a NN is trained and its generalization estimate is measured (e.g. over a validation sample). The process is stopped when the generalization decreases or when H reaches the maximum value (Hmax).

and ŷ is the value predicted by a 3-nearest neighbor algorithm. The kernel parameter (γ) produces the highest impact in the SVM performance, with values that are too large or too small leading to poor predictions. A practical method to set γ is to start the search from one of the extremes and then search towards the middle of the range while the predictive estimate increases [[31]](#_bookmark29).

* 1. Variable and model selection

Sensitivity analysis [[18]](#_bookmark23) is a simple procedure that is applied after the training phase and analyzes the model responses when the inputs are changed. Originally proposed for NNs, this sensitivity method can

j

In SVM regression [[26]](#_bookmark24), the input xaRI is transformed into a high

also be applied to other algorithms, such as SVM [[7]](#_bookmark15). Let yâ

denote the

m-dimensional feature space, by using a nonlinear mapping (ϕ) that does not need to be explicitly known but that depends of a kernel function (K). The aim of a SVM is to ﬁnd the best linear separating

output obtained by holding all input variables at their average values except xa, which varies through its entire range with j ∈ {1,…, L} levels. If a given input variable (xa ∈ {x1,…, xI}) is relevant then it

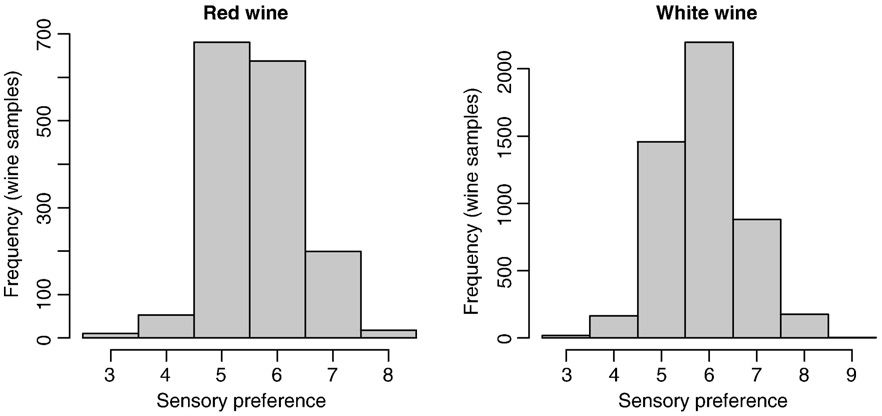


Fig. 1. The histograms for the red and white sensory preferences.

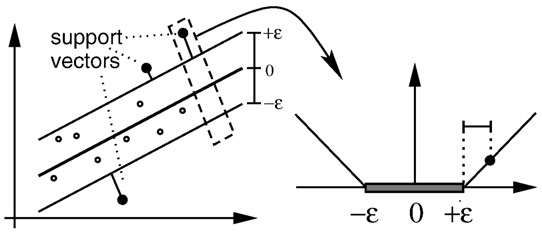


Fig. 2. Example of a linear SVM regression and the ε-insensitive loss function (adapted from [[26]](#_bookmark24)).

should produce a high variance (Va). Thus, its relative importance (Ra) can be given by:

1. Empirical results

The R environment [[22]](#_bookmark23) 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]](#_bookmark19), 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 zero mean and one standard deviation [[16]](#_bookmark23). RMiner uses the efﬁcient BFGS algorithm to train the NNs (nnet R package), while the SVM ﬁt is based on the Sequential Minimal Optimization implementation provided by LIBSVM (kernlab package). We adopted the default R suggestions [[29]](#_bookmark27). The only exception are the hyperparameters (H and γ), which will be set

using the procedure described in the previous section and with the

L 12

j

P

Va =

j = 1

I

yˆaj

−yˆa

= ðL − 1Þ

ð4Þ

search ranges of H ∈ {0, 1,…, 11} [[36]](#_bookmark34) and γ ∈ {23, 21,…, 2− 15} [[31]](#_bookmark29). While

the maximum number of searches is 12/10, in practice the parsimony

Ra = Va = Pi = 1 Vi × 100ðkÞ

In this work, the Ra values will be used to measure the importance of the inputs and also to discard irrelevant inputs, guiding the variable selection algorithm. We will adopt the popular backward selection, which starts with all variables and iteratively deletes one input until a stopping criterion is met [[14]](#_bookmark23). Yet, we guide the variable deletion (at each step) by the sensitivity analysis, in a variant that allows a reduction of the computational effort by a factor of I (when compared to the standard backward procedure) and that in [[18]](#_bookmark23) has out- performed other methods (e.g. backward and genetic algorithms). Similarly to [[36]](#_bookmark34), the variable and model selection will be performed simultaneously, i.e. in each backward iteration several models are searched, with the one that presents the best generalization estimate selected. For a given DM method, the overall procedure is depicted bellow:

approach (step 2 of [Section 4](#_bookmark10)) will reduce this number substantially.

Regarding the variable selection, we set the estimation metric to the MAD value (Eq. (1)), as advised in [[31]](#_bookmark29). To reduce the computational effort, we adopted the simpler 2/3 and 1/3 holdout split as the internal validation method. The sensitivity analysis parameter was set to L =5,

* 1. xa ∈ {− 1.0, − 0.5,…,1.0} for a standardized input. As a reasonable

balance between the pressure towards simpler models and the increase

of computational search, the stopping criterion was set to 2 iterations without any improvement or when only one input is available.

To evaluate the selected models, we adopted 20 runs of the more robust 5-fold cross-validation, in a total of 20 × 5 = 100 experiments for each tested conﬁguration. Statistical conﬁdence will be given by the t-student test at the 95% conﬁdence level [[13]](#_bookmark23). The results are summarized in [Table 2](#_bookmark7). The test set errors are shown in terms of the mean and conﬁdence intervals. Three metrics are present: MAD, the classiﬁcation accuracy for different tolerances (i.e. T = 0.25, 0.5 and 1.0)

and Kappa (T = 0.5). The selected models are described in terms of the

\_ \_ \_

* + 1. Start with all F ={x1,…, xI} input variables.
    2. If there is a hyperparameter P ∈ {P1,…, Pk} to tune (e.g. NN or

SVM), start with P1 and go through the remaining range until

the generalization estimate decreases. Compute the general- ization estimate of the model by using an internal validation method. For instance, if the holdout method is used, the available data are further split into training (to ﬁt the model) and validation sets (to get the predictive estimate).

* + 1. After ﬁtting the model, compute the relative importances (Ri) of all xi ∈ F variables and delete from F the least relevant input. Go to step 4 if the stopping criterion is met, otherwise return to

step 2.

* + 1. Select the best F (and P in case of NN or SVM) values, i.e., the input variables and model that provide the best predictive estimates. Finally, retrain this conﬁguration with all available data.

average number of inputs ( I ) and hyperparameter value (H or γ). The

last row shows the total computational time required in seconds.

For both tasks and all error metrics, the SVM is the best choice. The differences are higher for small tolerances and in particular for the white wine (e.g. for T = 0.25, the SVM accuracy is almost two times better when compared to other methods). This effect is clearly visible when plotting the full REC curves ([Fig. 3](#_bookmark10)). The Kappa statistic [[33]](#_bookmark31) measures the accuracy when compared with a random classiﬁer (which presents a Kappa value of 0%). The higher the statistics, the more accurate the result. The most practical tolerance values are T = 0.5 and T = 1.0. The former tolerance rounds the regression response into the nearest class, while the latter accepts a response that is correct within one of the two closest classes (e.g. a 3.1 value can be interpreted as grade 3 or 4 but not 2 or 5). For T = 0.5, the SVM accuracy improvement is 3.3 pp for red wine (6.2 pp for Kappa), a value that increases to 12.0 pp for the white task (20.4 pp for Kappa).

Table 2

The wine modeling results (test set errors and selected models; best values in bold).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Red wine  MR | | NN | SVM | White wine |  | |
| MR | NN | SVM |
| MAD | 0.50 ± 0.00 | 0.51 ± 0.00 | 0.46 ± 0.00[a](#_bookmark8) | 0.59 ± 0.00 | 0.58 ± 0.00 | 0.45±0.00[a](#_bookmark8) |
| AccuracyT = 0.25 (%) | 31.2 ± 0.2 | 31.1 ± 0.7 | 43.2 ± 0.6[a](#_bookmark8) | 25.6 ± 0.1 | 26.5 ± 0.3 | 50.3 ± 1.1[a](#_bookmark8) |
| AccuracyT = 0.50 (%) | 59.1 ± 0.1 | 59.1 ± 0.3 | 62.4 ± 0.4[a](#_bookmark8) | 51.7 ± 0.1 | 52.6 ± 0.3 | 64.6 ± 0.4[a](#_bookmark8) |
| AccuracyT = 1.00 (%) | 88.6 ± 0.1 | 88.8 ± 0.2 | 89.0 ± 0.2[b](#_bookmark9) | 84.3 ± 0.1 | 84.7 ± 0.1 | 86.8 ± 0.2[a](#_bookmark8) |
| KappaT =\_ 0.5 (%) | 32.2 ± 0.3 | 32.5 ± 0.6 | 38.7 ± 0.7[a](#_bookmark8) | 20.9 ± 0.1 | 23.5 ± 0.6 | 43.9 ± 0.4[a](#_bookmark8) |
| Inputs (I ) | 9.2 | 9.3 | 9.8 | 9.6 | 9.3 | 10.1 |
|  |  | ― =1 | ― =20.19 |  | ― =2.1 | ― =21.55 |
|  |  |  |  |  |  |  |
| Time (s) | 518 | 847 | 5589 | 551 | 1339 | 30674 |

1. Statistically signiﬁcant under a pairwise comparison with MR and NN.
2. Statistically signiﬁcant under a pairwise comparison with MR.

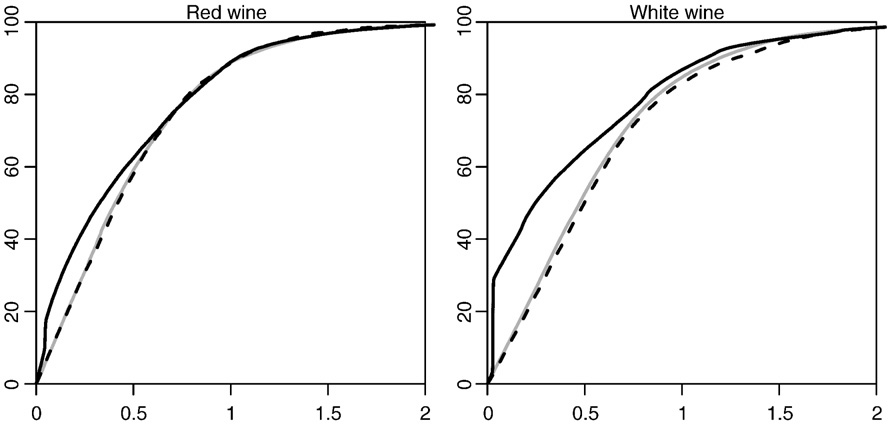


Fig. 3. The red (left) and white (right) wine average test set REC curves (SVM — solid line, NN — gray line and MR — dashed line).

The NN is quite similar to MR in the red wine modeling, thus similar performa\_nces were achieved. For the white data, a more complex NN

model (H = 2.1) was selected, slightly outperforming the MR results. Regarding the variable selection, the average number of deleted inputs ranges from 0.9 to 1.8, showing that most of the physicochemical tests used are relevant. In terms of computational effort, the SVM is the most expensive method, particularly for the larger white dataset.

A detailed analysis of the SVM classiﬁcation results is presented by the average confusion matrixes for T = 0.5 ([Table 3](#_bookmark11)). To simplify the visualization, the 3 and 9 grade predictions were omitted, since these were always empty. Most of the values are close to the diagonals (in bold), denoting a good ﬁt by the model. The true predictive accuracy for each class is given by the precision metric (e.g. for the grade 4 and white wine, precisionT = 0.5 = 19 / (19 + 7 + 4) = 63.3%). This statistic is important in practice, since in a real deployment setting the actual values are unknown and all predictions within a given column would be treated the same. For a tolerance of 0.5, the SVM red wine accuracies are around 57.7 to 67.5% in the intermediate grades (5 to 7) and very low (0%/20%) for the extreme classes (3, 8 and 4), which are less frequent ([Fig. 1](#_bookmark5)). In general, the white data results are better: 60.3/63.3% for classes 6 and 4, 67.8/72.6% for grades 7 and 5, and a surprising 85.5% for the class 8 (the exception are the 3 and 9 extremes with 0%, not shown in the table). When the tolerance is increased (T = 1.0), high accuracies ranging from 81.9 to 100% are attained for both wine types and classes 4 to 8.

The average SVM relative importance plots (Ra values) of the analytical tests are shown in [Fig. 4](#_bookmark12). It should be noted that the whole

11 inputs are shown, since in each simulation different sets of variables can be selected. In several cases, the obtained results conﬁrm the oenological theory. For instance, an increase in the alcohol (4th

Table 3

The average confusion matrixes (T = 0.5) and precision values (T = 0.5 and 1.0) for the SVM model (bold values denote accurate predictions).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Actual class | Red wine predictions  4 5 6 7 | | | | 8 | White wine predictions | | | | |
| 4 | 5 | 6 | 7 | 8 |
| 3 | 1 | 7 | 2 | 0 | 0 | 0 | 2 | 17 | 0 | 0 |
| 4 | 1 | 36 | 15 | 1 | 0 | 19 | 55 | 88 | 1 | 0 |
| 5 | 3 | 514 | 159 | 5 | 0 | 7 | 833 | 598 | 19 | 0 |
| 6 | 0 | 194 | 400 | 44 | 0 | 4 | 235 | 1812 | 144 | 3 |
| 7 | 0 | 10 | 107 | 82 | 1 | 0 | 18 | 414 | 441 | 7 |
| 8 | 0 | 0 | 10 | 8 | 0 | 0 | 3 | 71 | 43 | 59 |
| 9 |  |  |  |  |  | 0 | 1 | 3 | 2 | 0 |
| PrecisionT = 0.5 | 20.0 | 67.5 | 57.7 | 58.6 | 0.0 | 63.3 | 72.6 | 60.3 | 67.8 | 85.5 |
| (%) |  |  |  |  |  |  |  |  |  |  |
| PrecisionT = 1.0 | 93.8 | 90.9 | 86.6 | 90.2 | 100 | 90.0 | 93.3 | 81.9 | 90.3 | 96.2 |
| (%) |  |  |  |  |  |  |  |  |  |  |

and 2nd most relevant factor) tends to result in a higher quality wine. Also, the rankings are different within each wine type. For instance, the citric acid and residual sugar levels are more important in white wine, where the equilibrium between the freshness and sweet taste is more appreciated. Moreover, the volatile acidity has a negative impact, since acetic acid is the key ingredient in vinegar. The most intriguing result is the high importance of sulphates, ranked ﬁrst for both cases. Oenologically this result could be very interesting. An increase in sulphates might be related to the fermenting nutrition, which is very important to improve the wine aroma.

1. 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 entries) was considered, including vinho verde samples from the northwest region of Portugal. This case study was addressed by two regression tasks, where each wine type preference is modeled in a continuous scale, from 0 (very bad) to 10 (excellent). This approach preserves the order of the classes, allowing the evaluation of distinct accuracies, according to the degree of error tolerance (T) that is accepted.

Due to advances in the data mining (DM) ﬁeld, it is possible to extract knowledge from raw data. Indeed, powerful techniques such as neural networks (NNs) and more recently support vector machines (SVMs) are emerging. While being more ﬂexible models (i.e. no a priori restriction is imposed), the performance depends on a correct setting of hyperparameters (e.g. number of hidden nodes of the NN architecture or SVM kernel parameter). On the other hand, the multiple regression (MR) is easier to interpret than NN/SVM, with most of the NN/SVM applications considering their models as black boxes. Another relevant aspect is variable selection, which leads to simpler models while often improving the predictive performance. In this study, we present an integrated and computationally efﬁcient approach to deal with these issues. Sensitivity analysis is used to extract knowledge from the NN/SVM models, given in terms of relative importance of the inputs. Simultaneous variable and model selection scheme is also proposed, where the variable selection is guided by sensitivity analysis and the model selection is based on parsimony search that starts from a reasonable value and is stopped when the generalization estimate decreases.

Encouraging results were achieved, with the SVM model providing the best performances, outperforming the NN and MR techniques,

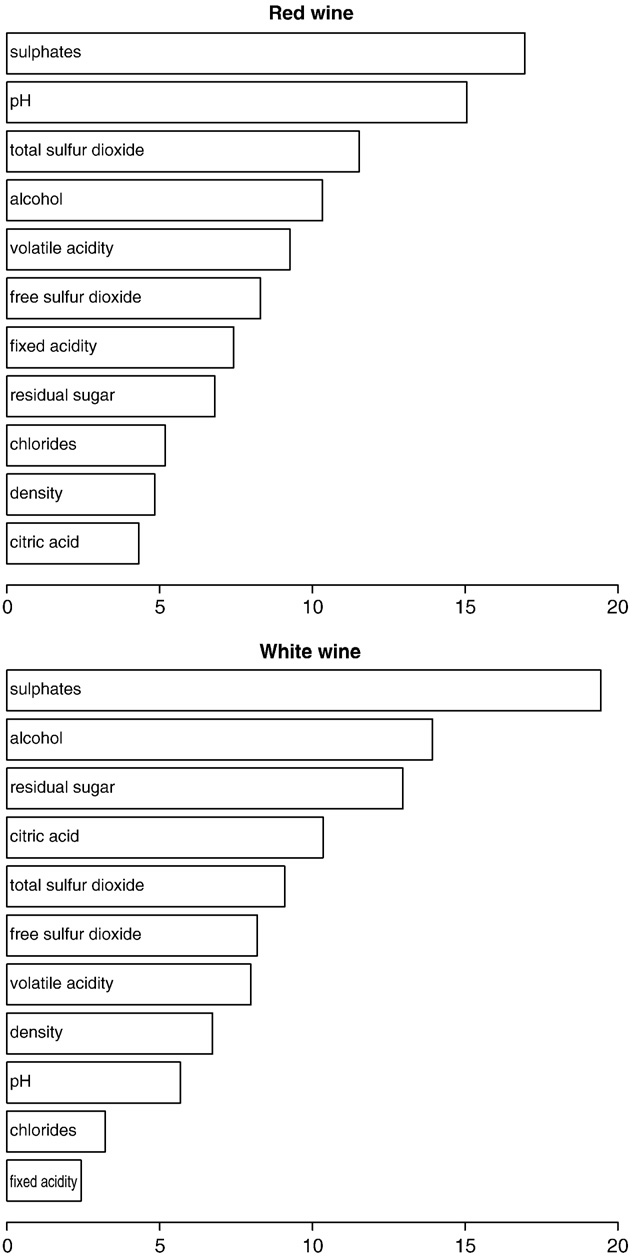


Fig. 4. The red (top) and white (bottom) wine input importances for the SVM model (in %).

particularly for white vinho verde wine, which is the most common type. When admitting only the correct classiﬁed classes (T = 0.5), the overall accuracies are 62.4% (red) and 64.6% (white). It should be noted that the datasets contain six/seven classes (from 3 to 8/9). These accuracies are much better than the ones expected by a random classiﬁer. The performance is substantially improved when the tolerance is set to accept responses that are correct within the one of the two nearest classes (T = 1.0), obtaining a global accuracy of 89.0% (red) and 86.8% (white). In particular, for both tasks the majority of the classes present an individual accuracy (precision) higher than 90%.

The superiority of SVM over NN is probably due to the differences in the training phase. The SVM algorithm guarantees an optimum ﬁt, while NN training may fall into a local minimum. Also, the SVM cost function ([Fig. 2](#_bookmark6)) gives a linear penalty to large errors. In contrast, the NN algorithm minimizes the sum of squared errors. Thus, the SVM is expected to be less sensitive to outliers and this effect results in a higher accuracy for low error tolerances. As argued in [[15]](#_bookmark23), it is difﬁcult to compare DM methods in a fair way, with data analysts tending to

favor models that they know better. We adopted the default suggestions of the R tool [[29]](#_bookmark27), except for the hyperparameters (which were set using a grid search). Since the default settings are more commonly used, this seems a reasonable assumption for the comparison. Nevertheless, different NN results could be achieved if different hidden node and/or minimization cost functions were used. Under the tested setup, the SVM algorithm provided the best results while requiring more computation. Yet, the SVM ﬁtting can still be achieved within a reasonable time with current processors. For example, one run of the 5-fold cross-validation testing takes around

26 min for the larger white dataset, which covers a three-year collection period.

The result of this work is important for the wine industry. At the certiﬁcation phase and by Portuguese law, the sensory analysis has to be performed by human tasters. Yet, the evaluations are based in the experience and knowledge of the experts, which are prone to subjective factors. The proposed data-driven approach is based on objective tests and thus it can be integrated into a decision support system, aiding the speed and quality of the oenologist performance. For instance, the expert could repeat the tasting only if her/his grade is far from the one predicted by the DM model. In effect, within this domain the T = 1.0 distance is accepted as a good quality control process and, as shown in this study, high accuracies were achieved for this tolerance. The model could also be used to improve the training of oenology students. Furthermore, the relative importance of the inputs brought interesting insights regarding the impact of the analytical tests. Since some variables can be controlled in the production process this information can be used to improve the wine quality. For instance, alcohol concentration can be increased or decreased by monitoring the grape sugar concentration prior to the harvest. Also, the residual sugar in wine could be raised by suspending the sugar fermentation carried out by yeasts. Moreover, the volatile acidity produced during the malolactic fermentation in red wine depends on the lactic bacteria control activity. Another interesting application is target marketing [[24]](#_bookmark23). Speciﬁc consumer preferences from niche and/or proﬁtable markets (e.g. for a particular country) could be measured during promotion campaigns (e.g. free wine tastings at supermarkets) and modeled using similar DM techniques, aiming at the design of brands that match these market needs.

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.

References

[1] A. Asuncion, D. Newman, UCI Machine Learning Repository, University of California, Irvine, 2007 [http://www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/%7Emlearn/MLRepository.html).

[2] J. Bi, K. Bennett, Regression error characteristic curves, Proceedings of 20th Int.

Conf. on Machine Learning (ICML), Washington DC, USA, 2003.

[3] C. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995. [4] B. Boser, I. Guyon, V. Vapnik, A training algorithm for optimal margin classiﬁers, COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning

Theory, ACM, NY, USA, 1992, pp. 144–152.

[5] V. Cherkassy, Y. Ma, Practical selection of SVM parameters and noise estimation for SVM regression, Neural Networks 17 (1) (2004) 113–126.

[6] P. Cortez. RMiner: Data mining with neural networks and support vector machines using R. In R. Rajesh (Ed.), Introduction to Advanced Scientiﬁc Softwares and Toolboxes, In press.

[7] P. Cortez, M. Portelinha, S. Rodrigues, V. Cadavez, A. Teixeira, Lamb meat quality assessment by support vector machines, Neural Processing Letters 24 (1) (2006) 41–51.

[8] CVRVV. Portuguese Wine — Vinho Verde. Comissão de Viticultura da Região dos

Vinhos Verdes (CVRVV), [http://www.vinhoverde.pt](http://www.vinhoverde.pt/), July 2008.

[9] T. Dietterich, Approximate statistical tests for comparing supervised classiﬁcation learning algorithms, Neural Computation 10 (7) (1998) 1895–1923.

[10] S. Ebeler, Flavor Chemistry — Thirty Years of Progress, Kluwer Academic

Publishers, 1999, pp. 409–422, chapter Linking ﬂavour chemistry to sensory analysis of wine.

[11] FAO, FAOSTAT — Food and Agriculture Organization Agriculture Trade Domain Statistics, July 2008 <http://faostat.fao.org/site/535/DesktopDefault.aspx?PageID=535>.

[12] J. Ferrer, A. MacCawley, S. Maturana, S. Toloza, J. Vera, An optimization approach for scheduling wine grape harvest operations, International Journal of Production Economics 112 (2) (2008) 985–999.

[13] A. Flexer, Statistical evaluation of neural networks experiments: minimum requirements and current practice, Proceedings of the 13th European Meeting on Cybernetics and Systems Research, vol 2, 1996, Austrian Society for Cybernetic Studies, Vienna, Austria, pp. 1005–1008.

[14] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, Journal of

Machine Learning Research 3 (7–8) (2003) 1157–1182.

[15] D. Hand, Classiﬁer technology and the illusion of progress, Statistical Science 21 (1) (2006) 1–15.

[16] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer-Verlag, NY, USA, 2001.

[17] Z. Huang, H. Chen, C. Hsu, W. Chen, S. Wu, Credit rating analysis with support vector machines and neural networks: a market comparative study, Decision Support Systems 37 (4) (2004) 543–558.

[18] R. Kewley, M. Embrechts, C. Breneman, Data strip mining for the virtual design of pharmaceuticals with neural networks, IEEE Transactions on Neural Networks 11 (3) (May 2000) 668–679.

[19] M. Kiang, A comparative assessment of classiﬁcation methods, Decision Support Systems 35 (4) (2003) 441–454.

[20] A. Legin, A. Rudnitskaya, L. Luvova, Y. Vlasov, C. Natale, A. D'Amico, Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception, Analytica Chimica Acta 484 (1) (2003) 33–34.

[21] I. Moreno, D. González-Weller, V. Gutierrez, M. Marino, A. Cameán, a. González, A. Hardisson, 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 72 (1) (2007) 263–268.

[22] R Development Core Team, R: a language and environment for statistical computing,

R Foundation for Statistical Computing, Vienna, Austria, ISBN: 3-900051-00-3, 2008,

[http://www.R-project.org](http://www.R-project.org/).

[23] D. Rumelhart, G. Hinton, R. Williams, in: D. Rulmelhart, J. McClelland (Eds.), Learning internal representations by error propagation, Parallel Distributed Processing: Explorations in the Microstructures of Cognition, vol 1, MIT Press, Cambridge MA, 1986, pp. 318–362.

[24] M. Shaw, C. Subramaniam, G. Tan, M. Welge, Knowledge management and data mining for marketing, Decision Support Systems 31 (1) (2001) 127–137.

[25] D. Smith, R. Margolskee, Making sense of taste, Scientiﬁc American, Special issue 16 (3) (2006) 84–92.

[26] A. Smola, B. Schölkopf, A tutorial on support vector regression, Statistics and Computing 14 (2004) 199–222.

[27] L. Sun, K. Danzer, G. Thiel, Classiﬁcation of wine samples by means of artiﬁcial neural networks and discrimination analytical methods, Fresenius' Journal of Analytical Chemistry 359 (2) (1997) 143–149.

[28] E. Turban, R. Sharda, J. Aronson, D. King, Business Intelligence, A Managerial Approach, Prentice-Hall, 2007.

[29] W. Venables, B. Ripley, Modern Applied Statistics with S4th edition, Springer,

2003.

[30] S. Vlassides, J. Ferrier, D. Block, Using historical data for bioprocess optimization: modeling wine characteristics using artiﬁcial neural networks and archived process information, Biotechnology and Bioengineering 73 (1) (2001).

[31] W. Wang, Z. Xu, W. Lu, X. Zhang, Determination of the spread parameter in the Gaussian kernel for classiﬁcation and regression, Neurocomputing 55 (3) (2003) 643–663.

[32] P. Werbos. Beyond regression: new tools for prediction and analysis in the behavioral sciences. PhD thesis, Harvard University, Cambridge, MA, 1974.

[33] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and

Techniques with Java Implementations2nd edition, Morgan Kaufmann, San Francisco, CA, 2005.

[34] X. Wu, V. Kumar, J. Quinlan, J. Gosh, Q. Yang, H. Motoda, G. MacLachlan, A. Ng, B. Liu, P. Yu, Z. Zhou, M. Steinbach, D. Hand, D. Steinberg, Top 10 algorithms in data mining, Knowledge and Information Systems 14 (1) (2008) 1–37.

[35] H. Yu, H. Lin, H. Xu, Y. Ying, B. Li, X. Pan, Prediction of enological parameters and discrimination of rice wine age using least-squares support vector machines and near infrared spectroscopy, Agricultural and Food Chemistry 56 (2) (2008) 307–313.

[36] M. Yu, M. Shanker, G. Zhang, M. Hung, Modeling consumer situational choice of long distance communication with neural networks, Decision Support Systems 44 (4) (2008) 899–908.

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, evolu- tionary 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 Med- icine, Neurocomputing, Neural Processing Letters, and others (see [http://www.dsi.uminho.pt/~pcortez](http://www.dsi.uminho.pt/%7Epcortez)).

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".

