# 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*

## 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.

*Key words:* Sensory preferences, Regression, Variable selection, Model selection, Support vector machines, Neural networks

## 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

* Corresponding author. E-mail pcortez@dsi.uminho.pt; tel.: +351 253510313; fax: +351 253510300.

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

3

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 wine evaluations, potentially improving the quality and speed of their decisions.

4

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 consumers 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

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

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).

[ insert Table 1 and Fig. 1 around here ]

*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 $k \in \{1, ..., N\}$ 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]:

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

where $\widehat{y}_k$ 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,

7

the predicted class is given by $p_i = y_i$, if $|y_i - \widehat{y}_i| \leq T$, else $p_i = y_i'$, where $y_i'$ denotes the closest class to $\widehat{y}_i$, given that $y_i' \neq y_i$. 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, 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.

## 2.3  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]. 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]:

$$\widehat{y} = w_{o,0} + \sum_{j=I+1}^{o-1} \frac{1}{1 + \exp(-\sum_{i=1}^{I} x_i w_{j,i} - w_{j,0})} \cdot w_{o,i} \qquad (2)$$

where $w_{i,j}$ 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 $\{0, 1, 2, 3, \ldots, H_{max}\}$ (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 $(H_{max})$.

In SVM regression [26], the input $x \in \Re^I$ is transformed into a high $m$-dimensional feature space, by using a nonlinear mapping $(\phi)$ 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 hyperplane, tolerating a small error $(\epsilon)$ when fitting the data, in the feature space:

$$\widehat{y} = w_0 + \sum_{i=1}^{m} w_i \phi_i(x) \tag{3}$$

The $\epsilon$-insensitive loss function sets an insensitive tube around the residuals and the tiny errors within the tube are discarded (Fig. 2).

[ insert Fig. 2 around here ]

We will adopt the popular gaussian kernel, which presents less parameters than other kernels (e.g. polynomial) [31]: $K(x, x') = exp(-\gamma||x-x'||^2), \gamma > 0$. Under this setup, the SVM performance is affected by three parameters: $\gamma$, $\epsilon$ and $C$ (a trade-off between fitting the errors and the flatness of the mapping). To reduce the search space, the first two values will be set using the heuristics [5]: $C = 3$ (for a standardized output) and $\epsilon = \widehat{\sigma}/\sqrt{N}$, where $\widehat{\sigma} = 1.5/N \times \sum_{i=1}^{N}(y_i - \widehat{y_i})^2$ and $\widehat{y}$ is the value predicted by a 3-nearest neighbor algorithm. The kernel

parameter ($\gamma$) 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 $\gamma$ 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].

## 2.4 Variable and Model Selection

Sensitivity analysis [18] 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 also be applied to other algorithms, such as SVM [7]. Let $\hat{y}_{a_j}$ denote the output obtained by holding all input variables at their average values except $x_a$, which varies through its entire range with $j \in \{1, \ldots, L\}$ levels. If a given input variable ($x_a \in \{x_1, \ldots, x_I\}$) is relevant then it should produce a high variance ($V_a$). Thus, its relative importance ($R_a$) can be given by:

$$V_a = \sum_{j=1}^{L} (\hat{y}_{a_j} - \overline{\hat{y}_{a_j}})^2 / (L-1)$$
$$R_a = V_a / \sum_{i=1}^{I} V_i \times 100 \ (\%)$$
(4)

In this work, the $R_a$ 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]. 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] has outperformed other methods (e.g. backward and genetic algorithms). Similarly

10

to [36], 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:

(1) Start with all $F = \{x_1, \ldots, x_I\}$ input variables.

(2) If there is a hyperparameter $P \in \{P_1, \ldots, P_k\}$ to tune (e.g. NN or SVM), start with $P_1$ and go through the remaining range until the generalization estimate decreases. Compute the generalization 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).

(3) After fitting the model, compute the relative importances ($R_i$) of all $x_i \in F$ variables and delete from $F$ the least relevant input. Go to step 4 if the stopping criterion is met, otherwise return to step 2.

(4) 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.

## 3 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 zero mean and one standard deviation [16]. **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]. The only exception are the hyperparameters ($H$ and $\gamma$), which will be set using the procedure described in the previous section and with the search ranges of $H \in \{0, 1, \ldots, 11\}$ [36] and $\gamma \in \{2^3, 2^1, \ldots, 2^{-15}\}$ [31]. While the maximum number of searches is 12/10, in practice the parsimony approach (step 2 of Section 2.4) will reduce this number substantially.

Regarding the variable selection, we set the estimation metric to the $MAD$ value (Equation 1), as advised in [31]. 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 $L = 5$, i.e. $x_a \in \{-1.0, -0.5, \ldots, 1.0\}$ 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]. The results are summarized in Table 2. 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 average number of inputs ($\overline{I}$) and hyperparameter value ($\overline{H}$ or

$\overline{\gamma}$). The last row shows the total computational time required in seconds.

[ insert Table 2 and Fig. 3 around here ]

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). The Kappa statistic [33] measures the accuracy when compared with a random classifier (which presents a Kappa value of 0%). The higher the statistic, 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). The NN is quite similar to MR in the red wine modeling, thus similar performances were achieved. For the white data, a more complex NN model ($\overline{H} = 2.1$) 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). 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.

13

The true predictive accuracy for each class is given by the precision metric (e.g. for the grade 4 and white wine, $precision_{T=0.5}=19/(19+7+4)=63.3\%$). 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). 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.

[ insert Table 3 and Fig. 4 around here ]

The average SVM relative importance plots ($R_a$ values) of the analytical tests are shown in Fig. 4. 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 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.

14

## 4 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, 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) 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], 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], 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 minutes 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.

17

Another interesting application is target marketing [24]. 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

## References

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

[2] J. Bi and K. Bennett. Regression Error Characteristic curves. In *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, and V. Vapnik. A training algorithm for optimal margin classifiers. In *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152, NY, USA, 1992. ACM.

[5] V. Cherkassy and Y. Ma. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. *Neural Networks*, 17(1):113–126, 2004.

[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, and A. Teixeira. Lamb Meat Quality Assessment by Support Vector Machines. *Neural Processing Letters*, 24(1):41–51, 2006.

[8] CVRVV. Portuguese Wine - Vinho Verde. Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV), http://www.vinhoverde.pt, July 2008.

[9] T. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7):1895–1923, 1998.

[10] S. Ebeler. *Flavor Chemistry - Thirty Years of Progress*, chapter Linking flavour chemistry to sensory analysis of wine, pages 409–422. Kluwer Academic Publishers, 1999.

[11] FAO. FAOSTAT - Food and Agriculture Organization agriculture trade domain statistics. http://faostat.fao.org/site/535/DesktopDefault.aspx?PageID=535, July 2008.

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

[13] A. Flexer. Statistical evaluation of neural networks experiments: Minimum requirements and current practice. In *Proceedings of the 13th European Meeting on Cybernetics and Systems Research*, volume 2, pages 1005–1008, Vienna, Austria, 1996.

[14] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(7–8):1157–1182, 2003.

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

[16] T. Hastie, R. Tibshirani, and 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, and S. Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, 37(4):543–558, 2004.

[18] R. Kewley, M. Embrechts, and C. Breneman. Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks. *IEEE Transactions on Neural Networks*, 11(3):668–679, May 2000.

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

[20] A. Legin, A. Rudnitskaya, L. Luvova, Y. Vlasov, C. Natale, and 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):33–34, 2003.

[21] I. Moreno, D. González-Weller, V. Gutierrez, M. Marino, A. Cameán, a. González, and 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):263–268, 2007.

[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, http://www.R-project.org, 2008.

[23] D. Rumelhart, G. Hinton, and R. Williams. Learning Internal Representations by Error Propagation. In D. Rulmelhart and J. McClelland, editors, *Parallel*

*Distributed Processing: Explorations in the Microstructures of Cognition*, volume 1, pages 318–362, MIT Press, Cambridge MA, 1986.

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

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

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

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

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

[29] W. Venables and B. Ripley. *Modern Applied Statistics with S*. Springer, 4th edition, 2003.

[30] S. Vlassides, J. Ferrier, and 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, and X. Zhang. Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing*, 55(3):643–663, 2003.

[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 and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA, 2nd edition, 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, and D. Steinberg. Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37, 2008.

[35] H. Yu, H. Lin, H. Xu, Y. Ying, B. Li, and 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):307–313, 2008.

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

Table 1

The physicochemical data statistics per wine type

| Attribute (units) | Red wine | | | White wine | | |
|---|---|---|---|---|---|---|
| | Min | Max | Mean | Min | Max | Mean |
| fixed acidity ($g$(tartaric acid)$/dm^3$) | 4.6 | 15.9 | 8.3 | 3.8 | 14.2 | 6.9 |
| volatile acidity ($g$(acetic acid)$/dm^3$) | 0.1 | 1.6 | 0.5 | 0.1 | 1.1 | 0.3 |
| citric acid ($g/dm^3$) | 0.0 | 1.0 | 0.3 | 0.0 | 1.7 | 0.3 |
| residual sugar ($g/dm^3$) | 0.9 | 15.5 | 2.5 | 0.6 | 65.8 | 6.4 |
| chlorides ($g$(sodium chloride)$/dm^3$) | 0.01 | 0.61 | 0.08 | 0.01 | 0.35 | 0.05 |
| free sulfur dioxide ($mg/dm^3$) | 1 | 72 | 14 | 2 | 289 | 35 |
| total sulfur dioxide ($mg/dm^3$) | 6 | 289 | 46 | 9 | 440 | 138 |
| density ($g/cm^3$) | 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)$/dm^3$) | 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 |

Table 2

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

| | Red wine | | | White wine | | |
|---|---|---|---|---|---|---|
| | **MR** | **NN** | **SVM** | **MR** | **NN** | **SVM** |
| MAD | 0.50±0.00 | 0.51±0.00 | **0.46**±0.00$^\star$ | 0.59±0.00 | 0.58±0.00 | **0.45**±0.00$^\star$ |
| Accuracy$_{T=0.25}$ (%) | 31.2±0.2 | 31.1±0.7 | **43.2**±0.6$^\star$ | 25.6±0.1 | 26.5±0.3 | **50.3**±1.1$^\star$ |
| Accuracy$_{T=0.50}$ (%) | 59.1±0.1 | 59.1±0.3 | **62.4**±0.4$^\star$ | 51.7±0.1 | 52.6±0.3 | **64.6**±0.4$^\star$ |
| Accuracy$_{T=1.00}$ (%) | 88.6±0.1 | 88.8±0.2 | **89.0**±0.2$^\diamond$ | 84.3±0.1 | 84.7±0.1 | **86.8**±0.2$^\star$ |
| Kappa$_{T=0.5}$ (%) | 32.2±0.3 | 32.5±0.6 | **38.7**±0.7$^\star$ | 20.9±0.1 | 23.5±0.6 | **43.9**±0.4$^\star$ |
| Inputs ($\overline{I}$) | 9.2 | 9.3 | 9.8 | 9.6 | 9.3 | 10.1 |
| Model | – | $\overline{H} = 1$ | $\overline{\gamma} = 2^{0.19}$ | – | $\overline{H} = 2.1$ | $\overline{\gamma} = 2^{1.55}$ |
| Time ($s$) | **518** | 847 | 5589 | **551** | 1339 | 30674 |

$\star$ - Statistically significant under a pairwise comparison with MR and NN.

$\diamond$ - Statistically significant under a pairwise comparison with MR.

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)

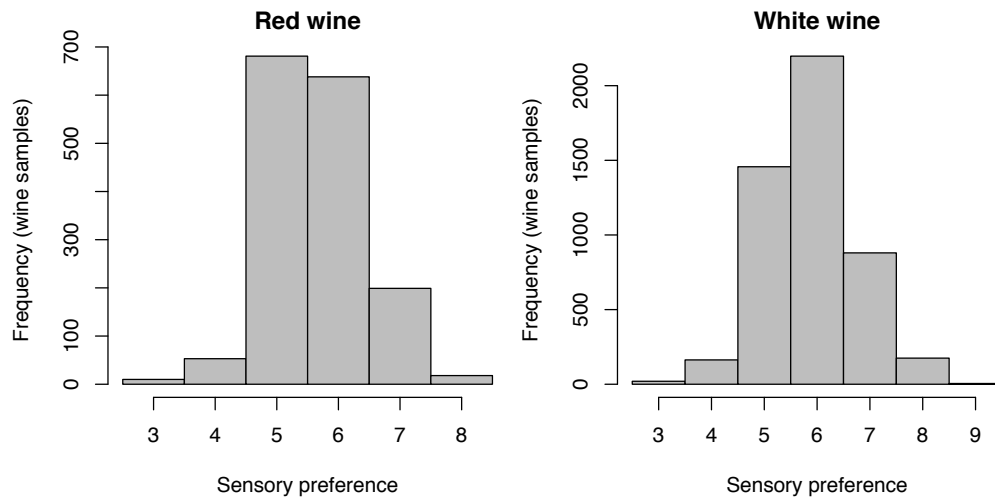| Actual Class | Red wine predictions | | | | | White wine predictions | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **4** | **5** | **6** | **7** | **8** | **4** | **5** | **6** | **7** | **8** |
| **3** | 1 | 7 | 2 | 0 | 0 | 0 | 2 | 17 | 0 | 0 |
| **4** | **1** | 36 | 15 | 1 | 0 | **19** | 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 |
| **Precision**$_{T=0.5}$ | 20.0% | 67.5% | 57.7% | 58.6% | 0.0% | 63.3% | 72.6% | 60.3% | 67.8% | 85.5% |
| **Precision**$_{T=1.0}$ | 93.8% | 90.9% | 86.6% | 90.2% | 100% | 90.0% | 93.3% | 81.9% | 90.3% | 96.2% |

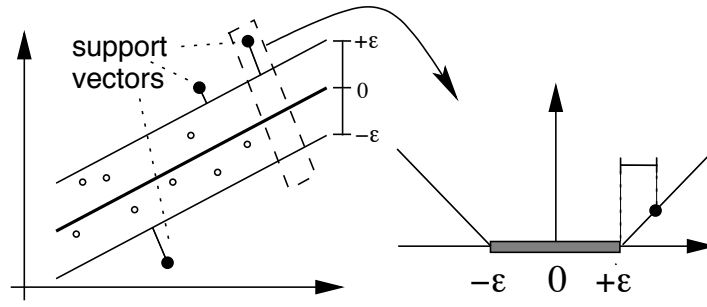Fig. 1. The histograms for the red and white sensory preferences

Fig. 2. Example of a linear SVM regression and the $\epsilon$-insensitive loss function (adapted from [26])
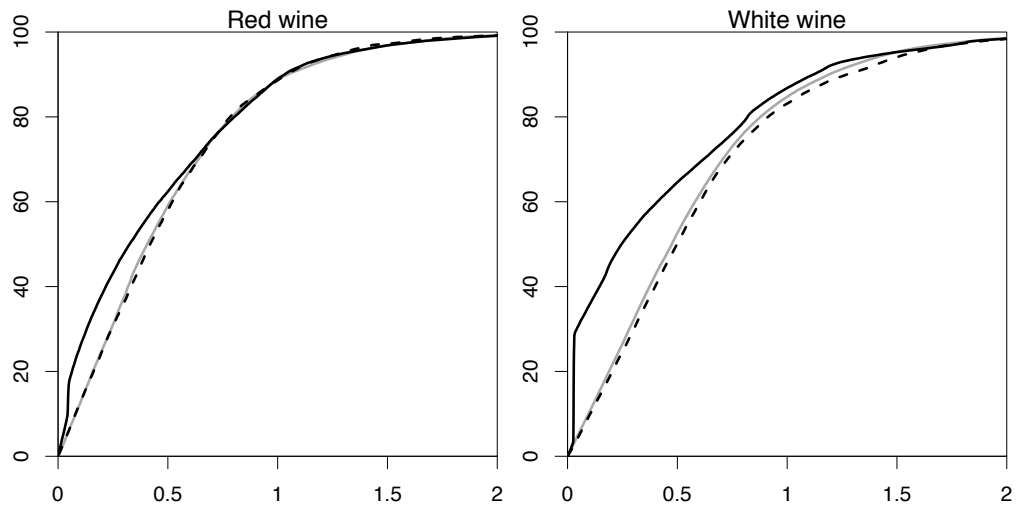
Fig. 3. The red (left) and white (right) wine average test set REC curves (SVM – solid line, NN - gray line and MR – dashed line).
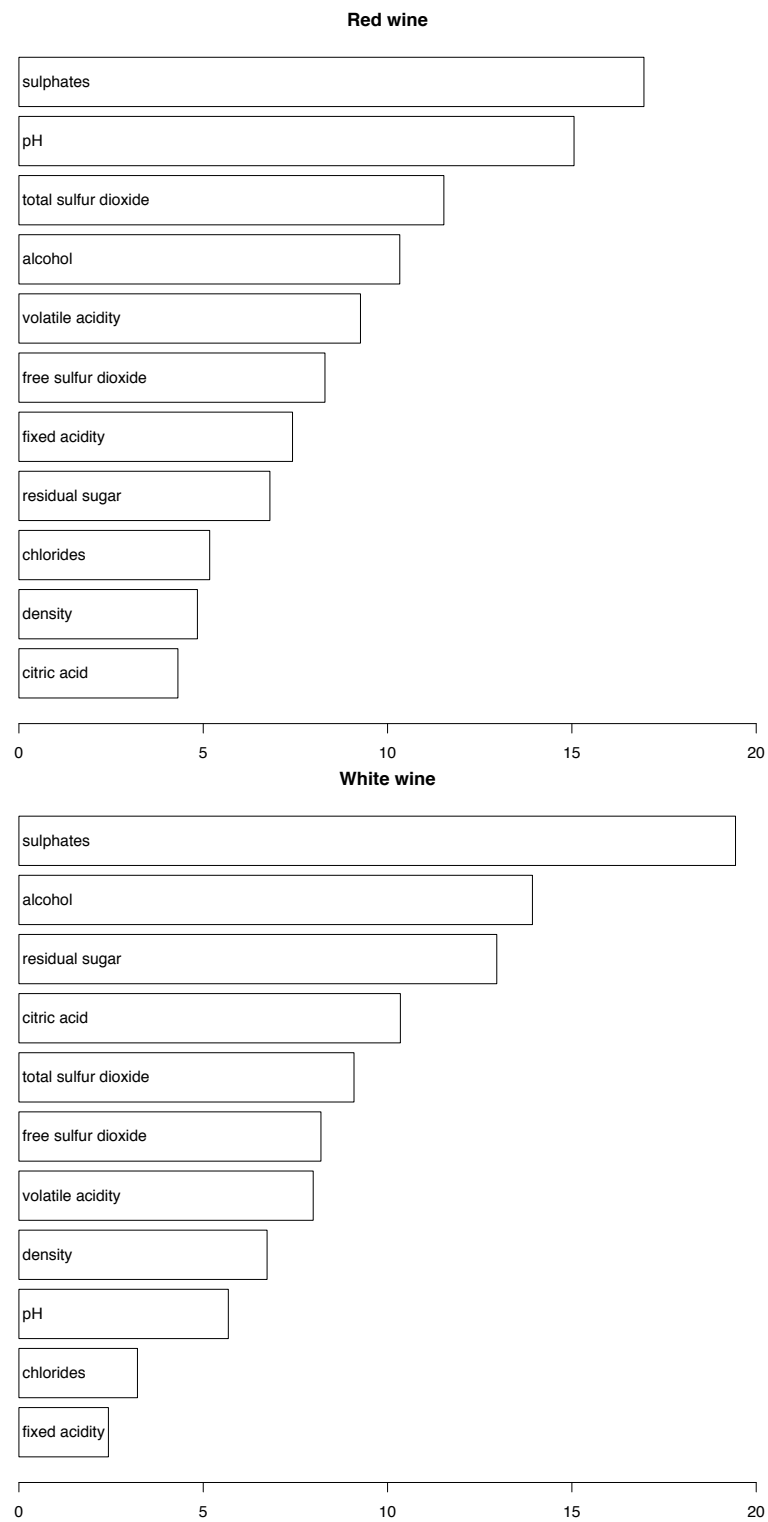
Fig. 4. The red (top) and white (bottom) wine input importances for the SVM model (in %).