Using Data Mining for Wine Quality Assessment

Conference Paper · October 2009

DOI: 10.1007/978-3-642-04747-3_8 · Source: DBLP

6 authors, including:

Paulo Cortez
University of Minho
313 PUBLICATIONS **10,227** CITATIONS
SEE PROFILE

Fernando david Almeida
Universidad Técnica de Ambato (UTA)
4 PUBLICATIONS **1,263** CITATIONS
SEE PROFILE

Telmo Matos
Polytechnic Institute of Porto
18 PUBLICATIONS **1,319** CITATIONS
SEE PROFILE

José Luís Reis
University of Maia
60 PUBLICATIONS **1,427** CITATIONS
SEE PROFILE

Using Data Mining for Wine Quality Assessment

Paulo Cortez[1], Juliana Teixeira[1], António Cerdeira[2], Fernando Almeida[2], Telmo Matos[2], and José Reis[1 2]

[1] Dep. of Information Systems/Algoritmi Centre, University of Minho, 4800-058 Guimarães, Portugal,
pcortez@dsi.uminho.pt, WWW home page: http://www3.dsi.uminho.pt/pcortez
[2] Viticulture Commission of the Vinho Verde region (CVRVV), 4050-501 Porto, Portugal

Abstract. Certification and quality assessment are crucial issues within the wine industry. Currently, wine quality is mostly assessed by physicochemical (e.g alcohol levels) and sensory (e.g. human expert evaluation) tests. In this paper, we propose a data mining approach to predict wine preferences that is based on easily available analytical tests at the certification step. A large dataset is considered with white *vinho verde* samples from the Minho region of Portugal. Wine quality is modeled under a regression approach, which preserves the order of the grades. Explanatory knowledge is given in terms of a sensitivity analysis, which measures the response changes when a given input variable is varied through its domain. Three regression techniques were applied, under a computationally efficient procedure that performs simultaneous variable and model selection and that is guided by the sensitivity analysis. The support vector machine achieved promising results, outperforming the multiple regression and neural network methods. Such model is useful for understanding how physicochemical tests affect the sensory preferences. Moreover, it can support the wine expert evaluations and ultimately improve the production.
Keywords: Ordinal Regression, Sensitivity Analysis, Sensory Preferences, Support Vector Machines, Variable and Model Selection, Wine Science.

1 Introduction

Nowadays wine is increasingly enjoyed by a wider range of consumers. In particular, Portugal is a top ten wine exporting country and exports of its *vinho verde* wine (from the northwest region) have increased by 36% from 1997 to 2007 [7]. To support this growth, the 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 often assessed by physicochemical and sensory tests [9]. 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 [20], thus wine classification is a difficult task. Moreover, the relationships between the physicochemical and sensory analysis are complex and still not fully understood [16].

On the other hand, 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 [23]. Data mining (DM) techniques [26] 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. Neural networks (NNs) have become increasingly used since the introduction of the backpropagation algorithm [19]. More recently, support vector machines (SVMs) have also been proposed [3]. 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 [13]. SVMs present theoretical advantages over NNs, such as the absence of local minima in the learning phase. When applying these methods, performance highly depends on a correct variable and model selection, since simple models may fail in mapping the underlying concept and too complex ones tend to overfit the data [13][12].

The use of decision support systems by the wine industry is mainly focused on the wine production phase [10]. 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 famous "Wine" dataset was donated into the UCI repository [2]. 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 [22], 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 [24], 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. More recently, mineral characterization (e.g. Zn and Mg) was used to discriminate 54 samples into two red wine classes [17]. 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 [6]. 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 [27].

In this paper, we present a real-world application, where wine taste preferences are modeled by DM algorithms that use analytical data that are easily

available at the certification step. In contrast with previous studies, a large dataset is considered with a total of 4898 samples. Wine quality is modeled under a regression approach that preserves the order of the grades. Explanatory knowledge is given by a sensitivity analysis, which measures how the responses are affected when a given input is varied through its domain [14][6]. Variable and model selection are performed simultaneously, in a process that is guided by the sensitivity analysis. Also, we propose a parsimony search method to select the best NN and SVM parameters with a low computational effort. Finally, we show the impact of the obtained models in the wine domain.

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 [7], and around 10% is exported, mostly white wine. In this work, we will analyze this common variant 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. Table 1 presents the physicochemical statistics per dataset. Regarding the preferences, each sample was evaluated by a minimum of three sensory assessors (using blind tastes), which graded the wine in a scale that ranges from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations. Fig. 1 plots the histograms of the target variable, denoting a typical normal shape distribution (i.e. with more normal grades that extreme ones).

Table 1. The physicochemical data statistics

Attribute (units)	**Min**	**Max**	**Mean**
fixed acidity (g(tartaric acid)$/dm^3$)	3.8	14.2	6.9
volatile acidity (g(acetic acid)$/dm^3$)	0.1	1.1	0.3
citric acid (g/dm^3)	0.0	1.0	0.3
residual sugar (g/dm^3)	0.6	65.8	6.4
chlorides (g(sodium chloride)$/dm^3$)	0.01	0.35	0.05
free sulfur dioxide (mg/dm^3)	2	260	35
total sulfur dioxide (mg/dm^3)	9	260	138
density (g/cm^3)	0.987	1.039	0.994
pH	2.7	3.8	3.1
sulphates (g(potassium sulphate)$/dm^3$)	0.2	1.1	0.5
alcohol (% vol.)	8.0	14.2	10.4

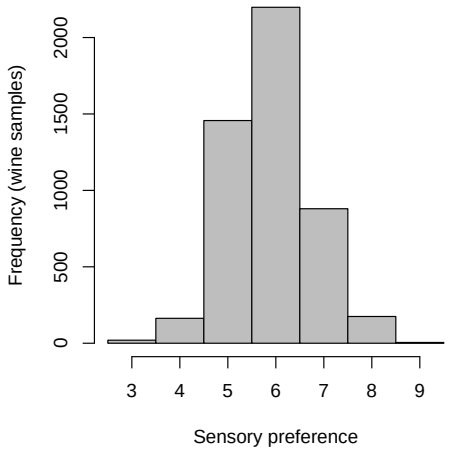

Fig. 1. The histogram for the white wine preferences

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) [26]:

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

where $\widehat{y}_k$ is the predicted value for the k input pattern. The regression error characteristic (REC) curve [1] is also used to compare regression models, with the ideal model presenting an area of 1.0. The curve plots the absolute error tolerance T (x-axis), versus the percentage of points correctly predicted (the accuracy) within the tolerance (y-axis).

The confusion matrix is often used for classification analysis, where a $C \times C$ matrix (C is the number of classes) is created by matching the predicted values (in columns) with the desired classes (in rows). For an ordered output, the predicted class is given by $p_i = y_i$, if $|y_i - \widehat{y}_i| \leq T$, 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) [26].

The holdout validation is often used to estimate the generalization capability of a model. 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 [8], 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. The validation method will be applied several runs and statistical confidence will be given by the t-student test at the 95% confidence level [11].

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 (Fig. 2). Supervised learning is achieved by an iterative adjustment of the network connection weights, called the training procedure, in order to minimize an error function. 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 [13]:

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

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 [21], the input $x \in \Re^I$ is transformed into a high 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 hyperplane in the feature space:

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

To select the best hyperplane, the ϵ-insensitive loss function is often used [21]. This function sets an insensitive tube around the residuals and the tiny errors within the tube are discarded (Fig. 2).

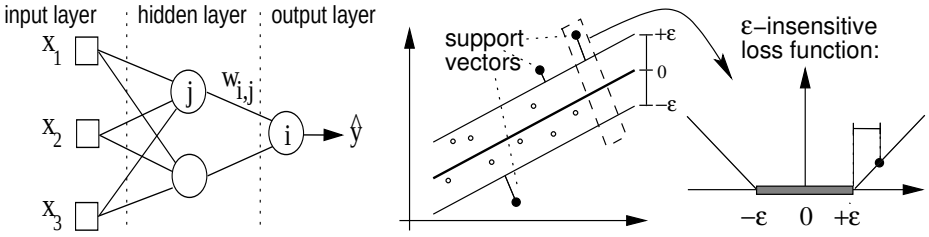

Fig. 2. Example of a multilayer perceptron with 3 inputs, 2 hidden nodes and one output (left) and a linear SVM regression (right, adapted from [21])

We will adopt the popular gaussian kernel, which presents less parameters than other kernels (e.g. polynomial) [25]: $K(x, x') = exp(-\gamma||x - x'||^2)$, $\gamma > 0$. Under this setup, the SVM performance is affected by three parameters: γ, ϵ 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 [4]: $C = 3$ (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 (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 [25].

2.4 Input Relevance and Variable/Model Selection

Sensitivity analysis [14] is a simple procedure that is applied after the training phase and analyzes the model responses when the inputs are changed. Origi-

nally proposed for NNs, this sensitivity method can also be applied to other algorithms, such as SVM [6]. Let $\widehat{y}_{a_j}$ 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:

$$\begin{array}{l} V_a = \sum_{j=1}^{L} (\widehat{y}_{a_j} - \overline{\widehat{y}_{a_j}})^2/(L-1) \\ R_a = V_a / \sum_{i=1}^{I} V_i \times 100\,(\%) \end{array} \quad (4)$$

The R_a values will be used to measure the relevance of the inputs. For a more detailed input influence analysis, in this work we propose the Variable Effect Characteristic (VEC) curve. For a given a attribute, the VEC plots the x_{a_j} values (x-axis) versus the $\widehat{y}_{a_j}$ predictions (y-axis) (see Section 3.3).

The sensitivity analysis will be also used to discard irrelevant inputs, guiding the variable selection algorithm. We will adopt a backward selection scheme, which starts with all variables and iteratively deletes one input until a stopping criterion is met [12]. The difference, when compared to the standard backward selection, is that 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 and that in [14] has outperformed other methods (e.g. backward and genetic algorithms). Similarly to [28], 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:

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

3.1 Experimental setup

All experiments reported in this work were written in **R** [18] and conducted in a Linux server, with an Intel dual core processor. **R** is an open source, multiple

platform (e.g. Windows, Linux) and high-level matrix programming language for statistical and data analysis. In particular, we adopted the **RMiner** [5], 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 [13]. **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). The the hyperparameters (H and γ) will be set using the procedure described in the previous section and with the search ranges of $H \in \{0, 1, \ldots, 11\}$ [28] and $\gamma \in \{2^3, 2^1, \ldots, 2^{-15}\}$ [25]. 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 (Eq. 1), as advised in [25]. 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 $L = 6$, i.e. $x_a \in \{-1.0, -0.6, \ldots, 1.0\}$ 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.

3.2 Predictive Knowledge

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. The results are summarized in Table 2. The test set errors are shown in terms of the mean and 95% 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.

Table 2. The wine modeling results (test set errors and selected models; best values are in **bold**; underline denotes a statistical significance when compared with MR and NN)

	MR	**NN**	**SVM**
MAD	0.59±0.00	0.58±0.00	**0.45**±0.00
$\text{Accuracy}_{T=0.25}$ (%)	25.6±0.1	26.5±0.3	**50.2**±1.1
$\text{Accuracy}_{T=0.50}$ (%)	51.7±0.1	52.6±0.3	**64.3**±0.4
$\text{Accuracy}_{T=1.00}$ (%)	84.3±0.1	84.7±0.1	**86.8**±0.2
$\text{Kappa}_{T=0.5}$ (%)	20.9±0.1	23.5±0.6	**43.4**±0.4
Inputs ($\overline{I}$)	9.6	9.3	10.0
Model	–	$\overline{H} = 2.1$	$\overline{\gamma} = 2^{0.7}$
Time (s)	**551**	1339	34644

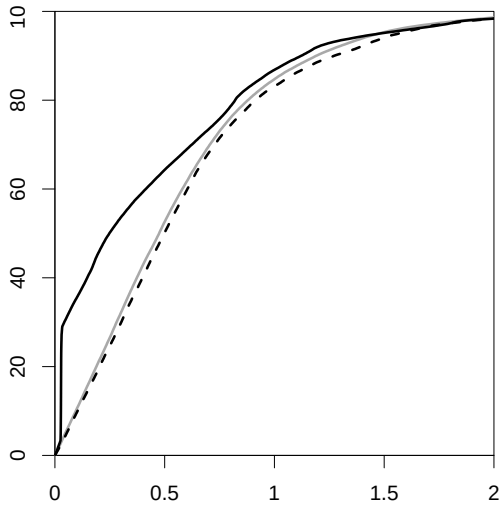

Fig. 3. The average test set REC curves (SVM – solid line, NN - gray line and MR – dashed line)

For all error metrics, the SVM is the best choice. The differences are higher for small tolerances (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 [26] 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 11.7

pp (19.9 pp for Kappa). The NN model slightly outperforms the MR results. Regarding the variable selection, the average number of deleted inputs ranges from 1.0 to 1.7, showing that most of the physicochemical tests used are relevant. In terms of computational effort, the SVM is the most expensive method.

A detailed analysis of the SVM classification results is presented by the average confusion matrix 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. The true predictive accuracy for each class is given by the precision metric (e.g. for the grade 4, precision$_{T=0.5}$=18/(18+6+4)=64.3%). 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 accuracies are 60.1/64.3% for classes 6 and 4, 67.1/72.3% for grades 7 and 5, and a surprising 86.6% 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 are obtained, ranging from 82.0 to 96.2%.

Table 3. The average confusion matrix ($T = 0.5$) and precision values ($T = 0.5$ and 1.0) for the SVM model (**bold** denotes accurate predictions)

Actual	**White wine predictions**				
Class	**4**	**5**	**6**	**7**	**8**
3	0	3	17	1	0
4	**18**	53	91	1	0
5	6	**832**	598	21	0
6	4	241	**1806**	144	3
7	0	20	418	**436**	6
8	0	2	71	45	**58**
9	0	0	2	2	0
Precision$_{T=0.5}$	64.3%	72.3%	60.1%	67.1%	86.6%
Precision$_{T=1.0}$	89.7%	93.4%	82.0%	90.1%	96.2%

3.3 Explanatory Knowledge

The relative importances of the SVM input variables, given in terms of the mean and 95% confidence intervals of the R_a values, 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. A more detailed analysis will be given to sixth most relevant analytical tests (Fig. 5). For a given input, each plot shows the histogram (frequency values are shown at the right of the y-axis) and the VEC curves ($\widehat{y}_{a_j}$ values, shown at the left of the y-axis) when the analytical test values (x-axis) are changed through their domain. For a given test, we built a VEC curve with $L = 6$ points (the sensitivity levels). Since 100 experiments we performed, we

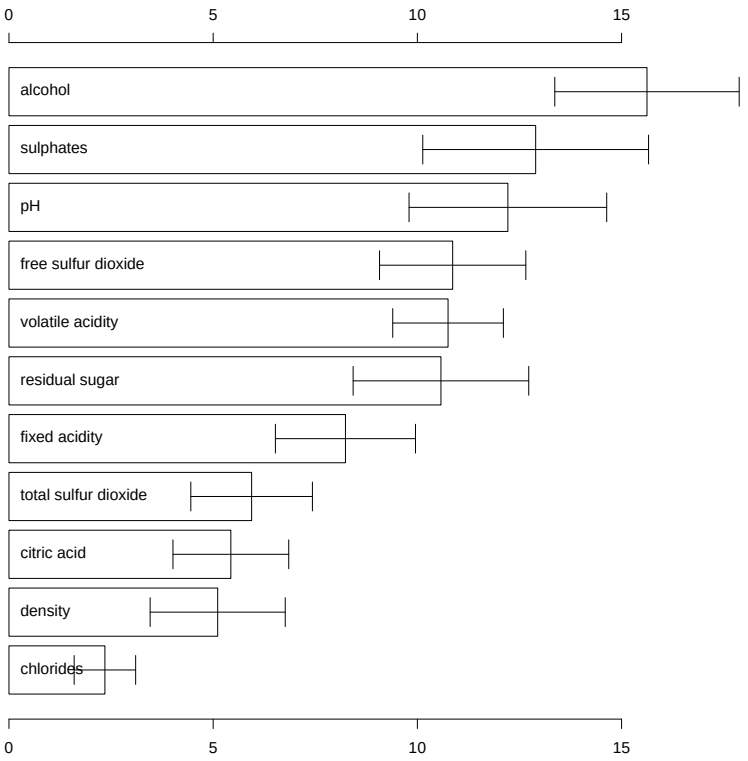

Fig. 4. The relative input importances for the SVM model (in %; bars denote the average value while the whiskers show the 95% confidence intervals)

performed a vertical averaging (with the respective 95% confidence intervals) of the 100 curves.

In several cases, the obtained results confirm the oenological theory. For instance, an increase in the alcohol (the most relevant factor) tends to result in a higher quality wine. Fig. 5 shows that this is true between the range from 9 to 13 % (which is related to most samples). In addition, the volatile acidity has a negative impact within the range that corresponds to the majority of the examples. This outcome was expected, since acetic acid is the key ingredient in vinegar. Moreover, residual sugar levels are important in white wine, where the equilibrium between the freshness and sweet taste is more appreciated. The most intriguing result is the high importance of sulphates, ranked second. 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, in an effect that occurs within the range 0.4 to 0.7 that contains most of the samples.

4 Conclusions

Due to the increase in the interest in wine, companies are investing in new technologies to improve their production and selling processes. Quality certification is a crucial step for both processes and is currently 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 entries) was considered, including white *vinho verde* samples from the northwest region of Portugal. This case study was addressed by a regression tasks, where wine 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. SVM kernel parameter) and the input variables used by the model. In this study, we present an integrated and computationally efficient approach that simultaneously addresses both issues. Sensitivity analysis is used to extract knowledge from the NN/SVM models, given in terms of the effect on the responses when one input is varied, leading to the proposed Variable Effect Characteristic (VEC) curves, and relative importance of the inputs (measured by the variance of the response changes). The 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. The overall accuracies are 64.3% ($T = 0.5$) and 86.8% ($T = 1.0$). It should be noted that the datasets contain six/seven classes (from 3 to 8/9) and these accuracies are much better than the ones expected by a random classifier. While requiring more computation, 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.

The result of this research is relevant to the wine science domain, helping in the understanding of how physicochemical characterization affects the final quality. In addition, this work can have an impact in 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. In future work, we intend to model preferences from niche and/or profitable markets (e.g. for a particular country by providing free wine tastings at supermarkets), aiming at the design of brands that match these market needs. We will also test other DM algorithms that specifically build rankers, such as regression trees [15].

References

1. J. Bi and K. Bennett. Regression Error Characteristic curves. In *Proceedings of 20th Int. Conf. on Machine Learning (ICML)*, Washington DC, USA, 2003.
2. C. Blake and C. Merz. UCI Repository of Machine Learning Databases, 1998.
3. 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.
4. V. Cherkassy and Y. Ma. Practical Selection of SVM Parameters and Noise Estimation for SVM Regression. *Neural Networks*, 17(1):113–126, 2004.
5. 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.
6. 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.
7. CVRVV. Portuguese Wine - Vinho Verde. Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV), http://www.vinhoverde.pt, July 2008.
8. T. Dietterich. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, 10(7):1895–1923, 1998.
9. S. Ebeler. *Flavor Chemistry - Thirty Years of Progress*, chapter Linking flavour chemistry to sensory analysis of wine, pages 409–422. Kluwer Academic Publishers, 1999.
10. J. Ferrer, A. MacCawley, S. Maturana, S. Toloza, and J. Vera. An optimization approach for scheduling wine grape harvest operations. *Production Economics*, pages 985–999, 2008.
11. 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.
12. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
13. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, NY, USA, 2001.

14. 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.
15. S. Kramer, G. Widmer, B. Pfahringer, and M. De Groeve. Prediction of Ordinal Classes Using Regression Trees. *Fundamenta Informaticae*, 47(1):1–13, 2001.
16. 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*, pages 33–34, 2003.
17. 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*, 72:263–268, 2007.
18. 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.
19. 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.
20. D. Smith and R. Margolskee. Making sense of taste. *Scientific American*, 284:26–33, 2001.
21. A. Smola and B. Scholkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.
22. 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*, 359:143–149, 1997.
23. E. Turban, R. Sharda, J. Aronson, and D. King. *Business Intelligence, A Managerial Approach*. Prentice-Hall, 2007.
24. 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.
25. W. Wang, Z. Xu, W. Lu, and X. Zhang. Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing*, 55:643–663, 2003.
26. I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, CA, 2005.
27. 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*, 56:307–313, 2008.
28. M. Yu, M. Shanker, G. Zhang, and M. Hung. Modeling consumer situational choice of long distance communication with neural networks. *Decision Support Systems*, 44:899–908, 2008.

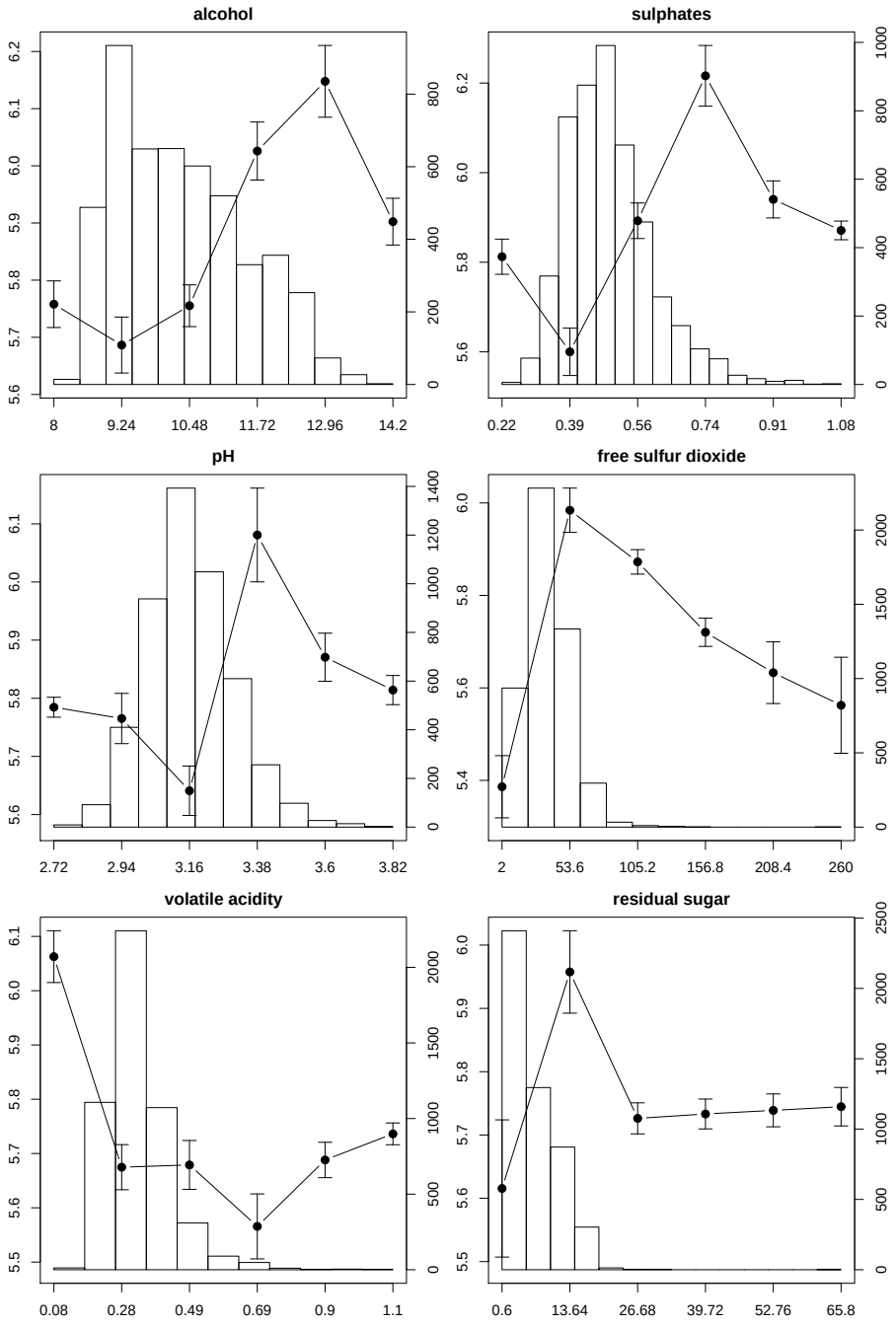

Fig. 5. The vertical averaging of the VEC curves (points and whiskers) and histogram (in bars) for the SVM model and the sixth most relevant physicochemical tests