JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

Classification of Wines Produced in Specific Regions by UV–Visible Spectroscopy Combined with Support Vector Machines

F. JAVIER ACEVEDO,*,† JAVIER JIMÉNEZ,‡ SATURNINO MALDONADO,† ELENA DOMÍNGUEZ,‡ AND ARÁNTZAZU NARVÁEZ‡

Department of Signal Theory and Communications. University of Alcalá, 28871 Madrid, Spain, and Department of Analytical Chemistry, University of Alcalá, 28871 Madrid, Spain

Discriminating wines according to their denomination of origin using cost-effective techniques is something that attracts the attention of different industrial sectors. In search of simplicity, direct UV−visible spectrophotometric techniques and different multivariate statistical techniques are used with admissible results to characterize wine produced in specific regions. However, most of the reported classification methods do not exploit all of the statistical relations in the investigated dataset and are inherently affected by the presence of outliers. The aim of this paper is to test novel classification methods such as support vector machines as a means of improving the classification rate when UV−visible spectrophotometric methods are used to discriminate wines. The advantages of such a discrimination tool are demonstrated when classification rates are compared for a large number of Spanish red and white wines and classification rates above 96% are achieved. The proposed methodology also enables the selection of the most relevant wavelengths for sample discrimination. The proposed methodology also enables the selection of the most relevant wavelengths for sample discrimination.

KEYWORDS: Wine classification; spectrophotometry analysis; support vector machines; geographical origin; feature selection

INTRODUCTION

The European Union (EU) possesses 45% of the world's vineyards, and its wine production accounts for 60% of the global market. In addition, almost 60% of the wine produced worldwide is consumed in Europe, which makes it the leading wine exporter and at the same time the largest import market in the world. The number of countries producing and exporting wine, particularly non-European, is increasing as is the relevance of this sector within the food industry. Wine producers and exporters, regulators and consumers are all demanding analytical tools for cost-effective routine quality control (*1*). The quality of wines primarily depends on the type of grape, the climate, the soil, and the different techniques used during the cultivation and production. In search of a clearer, simpler, and more transparent policy on wine quality, the EU has established two classes of wines, those which possess a geographical indication and those which do not. Spain possesses 70 quality wines produced in specified regions ("VCPRD" on the labels), which are clearly identified and controlled by different governing bodies at both the national and regional level. Furthermore, a wine produced in a specific region with well-defined cultivation and elaboration practices verifiable by the competent bodies is awarded the Denomination of Origin (DO). This denomination guarantees the provenance indicated on the label as well as a superior quality. The purpose of this paper is to investigate a new methodology for the cost-efficient assessment of the denomination of origin of wines exemplified by a range of Spanish DOs.

Some laboratories study the denomination of origin or the authenticity of a wine using labor-intensive and costly analyses which look for specific chemical features that can be identified with a given geographical origin. Examples of these chemical features include nonvolatile acids and amino acids (*2*), phenolic compounds (*3*), the concentration of metal ions (*4*), or isotopic determination (*5*). Each type of analysis is based on instrumental techniques that, being selective and reliable, require experienced operators and are difficult to automate and implement in routine and on-site applications.

Alternatively, nonsophisticated techniques and direct measurements, when combined with multivariate analysis, have demonstrated the ability to characterize wines through rather simple procedures (*1*). Some of these statistical procedures point out that it is possible to discriminate one particular denomination of origin from others using UV−visible spectrophotometry (*6*−*7*). Research has also been done on near-infrared spectroscopy to discriminate the geographical origin of Tempranillo

* To whom correspondence should be addressed. E-mail: javier.acevedo@uah.es. Telephone: +34918856725, Fax: +34918856699.
† Department of Signal theor and communications.
‡ Department of Analytical Chemistry.

10.1021/jf070634q CCC: $37.00 © 2007 American Chemical Society
Published on Web 07/28/2007

Table 1. Description of the Samples of the White Wines Tested

denomination of origin (DO)	commercial brand	no. of samples	predominant grape(s)
La Mancha	Añadas de Oro 200	9	varietal: Airén
	Estola 2004	6	80% Airén and Chardonnay
Madrid	Vega Madroño 2004	4	Malvar and Airén
	Puerta de Alcalá 2003	7	Malvar
	Puerta de Alcalá 2004	9	Malvar
Penedés	Sant Llach 2004	9	Macabeo, Xarel.lo and Parellada
	Vall de Juy 2005	3	Macabeo, Xarel.lo and Parellada
Rioja	Viña Espolón 2004	10	varietal:Viura
	Barón de Urzande 2005	5	Viura and Verdejo
	Viña Amate 2005	4	varietal:Viura
Valdepeñas	Viña Albali 2004	9	varietal:Macabeo
	Señorío de Ojailén 2004	7	varietal: Airén

Table 2. Description of the Samples of the Red Wines Tested

denomination of origin (DO)	commercial brand	no. of samples	predominant grapes
La Mancha	Añadas de Oro 2004	6	varietal: Tempranillo
	Viña Alambrada 2005	3	Tempranillo and Garnacha
Madrid	Castizo 2004	4	Tempranillo and Garnacha
	Puerta de Alcalá 2004	3	varietal: Tempranillo
	Puerta de Hierro 2005	4	Tempranillo and Garnacha
	Puerta Hierro 2004	9	Tempranillo and Garnacha
Penedés	Sant Llach 2004	8	Tempranillo and Merlot
	Puig de Montlor 2004	3	Merlot and Ull de Llebre (Tempranillo)
	Val de Juy 2005	3	varietal: Tempranillo
Rioja	Viña Espolón 2004	10	Tempranillo and Garnacha
	Barón de Urzande 2005	6	80% Tempranillo and Garnacha
	Viña Amate 2005	4	70% Tempranillo and Garnacha
Valdepeñas	Viña Albali 2004	9	90% Tempranillo and Cabernet-Sauvignon
	Señorio de Ojailén 2004	7	varietal: Tempranillo
	Señorio de Ojailén Reserva 2001	7	varietal: Tempranillo
Ribera Del Duero	Dehesa de la Jara 2004	15	varietal: Tempranillo
	Vega de Nava 2004	9	varietal: Tempranillo
	Barón de Santuy 2004	7	varietal: Tempranillo
Toro	Gran Cermeño Crianza 2002	15	varietal: Tinta de Toro
Somontano	Montesierra 2005	9	Tempranillo, Moristel and Cabernet-Sauvignon
	Viñas del Vero 2005	12	Merlot and Cabernet-Sauvignon

wines from Australia and Spain (*8*). The main advantage of this approach is simplicity because the "intelligence" or, in other words, the robustness of the procedure relies on the pattern recognition system, in which the principal component analysis (PCA) and the soft independent modeling of class analogies (SIMCA) are commonly used. The first results are promising, but these methods need to be improved if they are to be compared with traditional techniques and, especially, if the challenge is to discriminate between several DOs in which differences may be more subtle than the type of grape. Other pattern recognition methods, commonly used in chemometrics and aimed at obtaining better discrimination rates, can also be applied (*9*). These include nearest neighbor (k-NN), partial least-squares discriminant analysis (PLS-DA), and artificial neural networks. However, the complexity of these statistical tools negatively affects the slightly improved classification rate because their implementation makes routine analysis rather difficult.

As was anticipated, the aim of this paper is to demonstrate how a simple and nonselective technique (UV−visible spectrophotometry) in combination with modern classification techniques, can be used to discriminate between several Spanish wine DOs, improving the classification rate when compared to other multivariate analyses. The rationale of the proposed method is to use a classifier that, once trained, requires only a few operations and gives reliable results. Support vector machines (SVM) (*10*) is a rather recent classification method that does not need a large number of samples to be trained and is not affected by the presence of outliers. SVM have been successfully applied to several classification problems, especially with the standard University of California Irvine (UCI) datasets. Two recent papers report on the use of a standard dataset containing information for wine discrimination with a high level of accuracy (*11*, *12*). However, the information contained in this dataset only discriminates between three Italian wines, and no chemical information is provided regarding the variables. We believe that the present work reinforces the advantages of SVMs for sample classification and contributes to the creation of a cost-effective quality-control tool for both regulators and industrial sectors. Finally and considering that one of the most important steps in systems that use sensors and pattern classification data processing is the so-called feature selection (*13*), this paper also attempts to select the most relevant wavelengths for classification purposes and to verify the resolution of the spectrophotometer enabling DO discrimination. A new method is proposed to cover this objective.

MATERIALS AND METHODS

Sample Preparation and Data Acquisition. A number of white and red wines from various Spanish DO was selected for this study. Each DO included different brands of wines as is summarized in **Table 1** and **Table 2** for white and red wines,

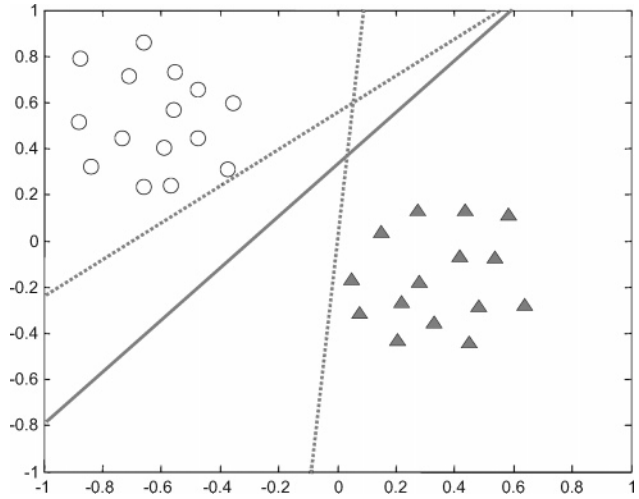

Figure 1. Training dataset and several linear planes that separate the classes.

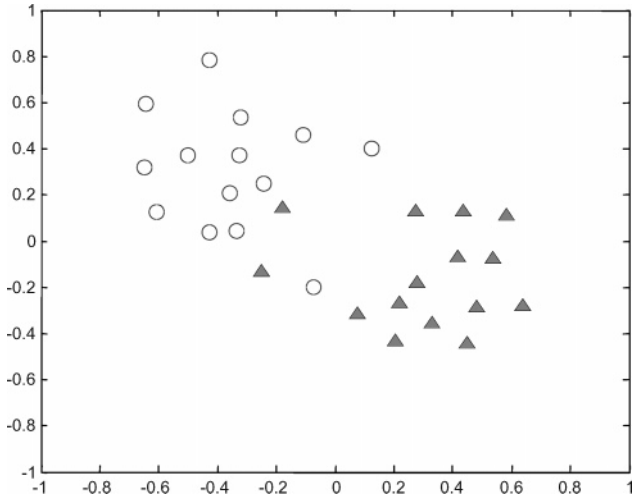

Figure 2. Dataset impossible to be separated by a linear plane without allowing training errors.

respectively. Overall, 82 white wine samples and a total of 153 red wine samples were scanned. The spectrophotometric measurements were obtained using a Shimadzu UV−vis (UV-160A) spectrophotometer. The absorbance spectra were recorded in duplicate, with a working range from 200 to 800 nm and with a step resolution of 10 nm, but to be sure that this low resolution had no influence on the classification results, the spectra were also recorded with a step resolution of 1 nm. All samples were tested immediately after the wine bottles were opened to prevent oxidation reactions. Prior to these measurements, the samples were filtrated through 0.45 μm membranes to remove any possible solids. Filtrated wines were then diluted with water in a 1:6 ratio. Most wines were young since wines aged in oak barrels may include additional characteristics that could cover up the original properties.

Support Vector Machines. Presently, the use of SVM for classification purposes is gaining attention in chemical systems. SVM is a binary discriminant classification tool that is based on the statistical learning theory (SLT). To visualize how SVM work, focus on **Figure 1**. The question that should be asked is, what is the best plane separating both classes? All of the planes drawn separate the training set, which is to say that they minimize the empirical risk. However, to find the best plane, we have to think that the classifier will have to discriminate those samples in the future. The answer to this question is provided in ref *10*, describing how the best plane is the one that maximizes the separation margin. Thus, in **Figure 1** the best plane is drawn with a continuous line. The maximization in the margin implies that only a few samples of the training set, namely those samples that are in the boundaries of the margin zone, influence the decision function. The samples that have an influence on the decision function are called support vectors. From the training phase, a decision function is obtained in such a way that new samples, spectrophotometric signals in our case, are classified according to:

$$f_x = \text{sign}\left(\sum_{i=1}^{l} \alpha_i y_i \langle s_i, x\rangle + b\right) \quad (1)$$

where l is the number of training samples, y_i is the class of the training sample i (+1,−1), and s_i are the training samples. The α_i and b values are obtained as a result of the training phase. For those training patterns that are not support vectors, the α_i value is zero, whereas for support vectors this value is greater than zero. The b parameter is also obtained in the training phase in such a way that the function decision for support vectors is equal to one. The above explanation is only valid for cases in which the training data are separable by a hyperplane, but in practice, most of the chemometric experiments tend to be more like **Figure 2**, in which it is not possible to find a linear hyperplane that separates the samples. In this situation, SVM theory also searches for the hyperplane that maximizes the margin while allowing a number of training errors. The number of errors allowed is determined by a regularization parameter C that has to be fixed before training. Thus, when a training dataset is evaluated, it is divided into those samples that have no influence on the margin and are correctly classified with an $\alpha_i = 0$ value, those samples that are at the limits of the margin with a value between $0 < \alpha_i < C$, and those samples that are in the margin zone or are poorly classified with a value of $\alpha_i = C$ and which are usually called second-type support vectors. In this case, the decision function is:

$$f_x = \text{sign}\left(\sum_{i=1}^{Ns} \alpha_i y_i \langle s_i, x\rangle + b\right) \quad (2)$$

where Ns is the number of training samples with α_i greater than zero or support vectors. The SVM theory with lineal hyperplanes has been described, but there are many problems where it is interesting to have decision functions based on nonlinear boundaries. The SVM theory can be then extended using kernel functions (*14*), in which the decision function is:

$$f_x = \text{sign}\left(\sum_{i=1}^{Ns} \alpha_i y_i K(s_i, x) + b\right) \quad (3)$$

Among all the advantages that SVM provide, we can highlight the fact that having a maximum margin allows better future samples to be classified once the model has been built. Moreover, there are very few parameters to tune or select a priori. One of the typical disadvantages when comparing SVM to other methods is that SVM were designed to solve two-class problems and extending them to multiclass problems is still an open research field. There are also other disadvantages when comparing them with nondiscriminant methods like SIMCA such as that this method can classify a sample into several classes or none of them. However, in this paper the modification described in ref *15* was used which adds a posterior stage which

makes it more probable to belong to a single class. Therefore, with this method the advantages provided by nondiscriminant functions are able to be kept.

The method selected was to use linear kernels as the result was worse with nonlinear kernels and the number of operations required in the test phase is several orders of magnitude greater. The software used was a modification of ref *16* developed in our research group to adapt the abovementioned posterior probability stage.

Other Pattern Recognition Methods. One of the primary goals of this paper is to compare SVM with those traditional pattern recognition methods employed in the chemometric field such us SIMCA, PLS-DA, NN-MLP, and k-NN. The software used to train and test SIMCA and k-NN was developed by our research group and written in Matlab language. Before making this comparison spectrophotometric signals were mean-centered and preprocessed to have ± 1 variance.

k-Nearest neighbor was used with several nearest-neighbor values, finally selecting a total of three neighbors. The SIMCA method was developed with a first stage where a PCA was done which reduces the X-space to two components. After that, a confidence interval of 95% is selected to train the system. The NN-MLP software used was the Matlab Neural Network Toolbox. Several parameters have to be tuned in order for a neural network to be developed. In our case, after testing several combinations, we selected an NN-MLP with 40 neurons with sigmoidal function in the hidden layer and a linear function in the output layer. The PLS-DA software used was developed by eigenvector.

In order to compare the different classification methods mentioned with the proposal of using SVM, a performance measure is needed. This required the definition of the accuracy (ρ_i) which was calculated as follows:

$$\rho_i = \frac{1}{d}\sum_{k=1}^{d} \delta\,(k)$$

$$\delta\,(k) = \begin{cases} 1 & \hat{y}_{(k)} = y_{(k)} \\ 0 & \hat{y}_{(k)} \neq y_{(k)} \end{cases} \qquad (4)$$

where d is the length of the test set, $y_{(k)}$ is the real DO of the wine k, and $\hat{y}_{(k)}$ is the estimated label found by the classifier.

It is very common in chemometrics to compare classifiers using the cross-validation procedure since the number of samples is reduced compared to other pattern recognition problems. For cross validation, each dataset, red and white wines, was divided into five sections. In every step, four sections were used as training patterns, and the remaining samples were used to test the classifier and to obtain the accuracy of each DO. The criterion used to make the sections was a random composition with an agent making sure that in every section at least one sample of each class was present. The ρ_i calculated is then the mean of the accuracies obtained in the cross-validation steps. The global performance of each classifier, the total accuracy, ρ, is finally calculated as the mean of the accuracies obtained for each DO. False negatives are another important measurement when comparing different classifiers. This measurement is defined as the percentage of samples of one class that are classified as any of the other classes.

On the other hand, we would like to study the influence of the brand in a DO. The second experiment done consisted of building two datasets per each wine, white and red. One dataset was used to train or calibrate the classifier, and the other, to test the accuracy and false negatives. The test dataset was composed of a randomly selected brand per each DO, and the training set was composed of the rest of the brands.

Feature Selection. The use of all variables for classification purposes is not an adequate strategy because it produces the so-called "curse of dimensionality". If variables are not selected, the results can be negatively affected because of the limitations caused by handling features that are not relevant. Thus, selecting the key variables, wavelengths in this paper, becomes one of the most crucial steps in chemometric methods based on pattern recognition. Moreover, feature selection also helps to decide on the most adequate technical conditions. Increasing the resolution of the wavelengths scanned, in an effort to obtain better results, will not necessarily improve classification accuracy.

Formally, the feature selection problem consists of finding a subset Cd of d discrete variables within the whole set H of D available wavelengths which minimizes an adopted criterion. When the criterion is based on the results given by the classifier used, it is said that the feature selection method is a closed-loop or a wrapper approach, whereas if it is based on another type of measurement, it is said that is an open-loop or a filter approach. An example of the filter approach is the commonly used Fischer criterion selection (*17*). The problem with methods based on this approach is that they are not coherent with the classifier used. Wrapper approach methods are known to be more reliable than filter approach methods although the former require more operations in the training stage. Since the critical part of our design is the test phase and not the training phase, this paper follows a wrapper approach. The optimal feature selection method is the exhaustive search of all possible combinations (presence/no presence). However, if the problem needs to take a large number of features (wavelengths), the number of combinations makes this method unfeasible.

Some of the most common suboptimal search methods include sequential forward selection (SFS) and sequential backward selection (SBS) (*18*). Although both simple, these methods do not take into account the possible relationships between features and suffer from the so-called "nesting-effect". Genetic algorithms (GA) are also used for feature selection (*19*) using the relationship between proximal features as a way of guiding the search. If this relationship is not present in the problem under study, or it is unknown, the use of GA for feature selection requires a significant number of generations to find an optimal value. The oscillating search method (*20*) avoids the nesting effect and yields good results, but the final number of features must be known a priori. This paper proposes a new method aimed at suppressing this requirement. The procedure can be summarized in the following steps:

1. Initialization. The first step of the algorithm is to execute the SFS procedure. As a result of this execution only a few wavelengths of the spectra are taken into account to discriminate wines. A new classifier is trained with only those selected wavelengths. This classifier has an obtained error E_d. From this initial step there is a set of selected wavelengths and a set of discarded wavelengths.

2. Down-swing. From the set of selected wavelengths, remove the one that achieves the minimum error E_{d-o}, that is, the error of the classifier built using all of the selected wavelengths except for the one that was removed, called w_1. Add the best wavelength from the discarded set, labeled w_2, and calculate the new E_{dn}, that is, the error of the classifier built using all of the selected wavelengths except for the one that was removed and with the best wavelength from the discarded set added. If the error E_{d-o} is smaller than either E_d or E_{dn}, reduce the feature

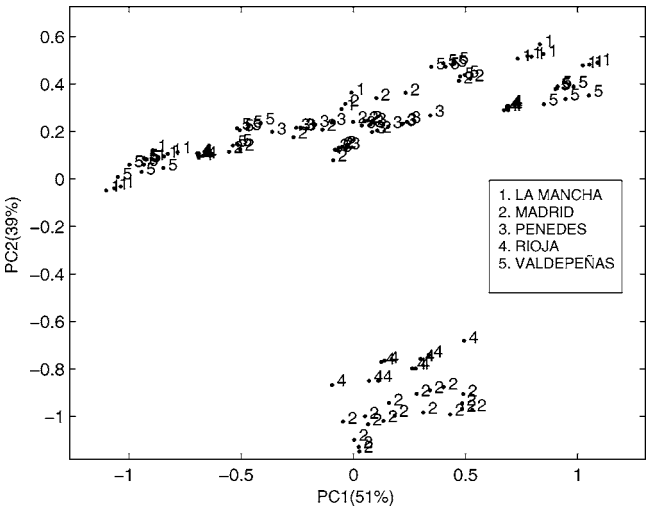

Figure 3. PCA scores on the wavelength variables obtained from 82 samples of Spanish white wines from five specific regions.

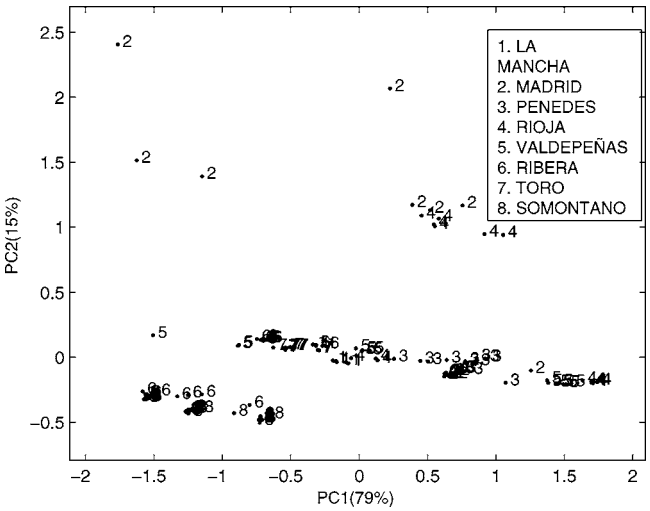

Figure 4. PCA scores on the wavelength variables obtained from 153 samples of Spanish red wines from eight specific regions.

set by removing the worst wavelength from the selected set, so that the number of selected features is reduced to one.

3. Look for changes. If E_{dn} is smaller than E_d, change the wavelengths in such a way that w_1 now belongs to the set of discarded wavelengths and w_2 belongs to the selected set and mark that changes were made in this step.

4. Up-swing. Add the best wavelength from the discarded set, labeled w_3 and calculate the new E_{d+o}, that is, the error of the classifier built using all of the wavelengths from the selected set plus w_3. Then, remove the worst wavelength, w_4, from the set of selected wavelengths and calculate the new E_{dn}. If the error E_{d+o} is smaller than E_d or E_{dn}, increase the set of selected features by adding w_3 and go to Step 2.

5. Look for changes. If the E_{dn} is smaller than E_d, change the wavelengths in such a way that w_1 now belongs to the discarded set and w_2 belongs to the selected set, and mark that changes were made in this step.

6. If changes were made to the set of selected wavelengths, repeat from Step 2.

7. If no changes were made, the algorithm is repeated from Step 2, but instead of removing and adding one wavelength each time, now pairs of wavelengths are considered. When pairs of wavelengths do not produce changes in the above steps, the search can be increased to consider groups of n wavelengths. The algorithm stops when the number n reaches a previous fixed value. In our experiments, the maximum value allowed for n was four.

RESULTS AND DISCUSSION

Comparing SVM to Other Classifiers. The first experiment was composed of two different datasets created by separating the white wines from the red wines. Every sample was labeled according to its DO. The first comparison was done with principal component analysis (PCA) which, although not a classification method itself, is widely used in chemometrics due to the graphical concept the resulting score and loading plot offers. **Figure 3** and **Figure 4** show the plots obtained for white and red wines, respectively, with the PCA method. Both figures clearly show the limitations of PCA in discriminating this problem, that is to say, in separating wine samples from different specific regions.

Further comparisons with other classification methods were made with the estimation of the accuracy and false negatives for each DO and with a cross-validation procedure. The results of these comparisons are presented in **Table 3** and **Table 4** for white and red wines, respectively. The comparison of the average accuracy shows that SVM performs better than other classifiers for discriminating the specific regions within the two sets of wines, white and red. Further comparisons of the accuracies obtained with SVM and other classifiers for each DO within its dataset also show the former to performance better. With the exception of three cases, white Valdepeñas and red Somontano, both with k-NN, and red Ribera del Duero with SIMCA, the discrimination performance by SVM for each DO within the corresponding dataset is always higher. The comparison also shows that neural networks classify the DO fairly well, but the number of operations needed in the test phase is several times higher than with SVM. It is worth mentioning that k-NN, SIMCA, and PLS-DA, although extensively used in chemometrics, do not provide results as good as SVM, at

Table 3. Comparison of Accuracies and False Negatives in Percentage Obtained with Different Classifying Methods with 82 Samples of White Wines and Cross-validation Method

	SVM		NN-MLP		kNN		SIMCA		PLS-DA	
DO	ρ	FN	ρ	FN	ρ	FN	ρ	FN	ρ	FN
La Mancha	93.33	0.00	93.33	0.00	93.33	0.00	86.67	0.00	93.33	0.00
Madrid	96.00	2.18	96.00	2.18	86.00	0.00	93.33	19.08	96.00	0.00
Penedés	100	1.33	100	1.33	100	2.33	70.00	0.00	100	1.33
Rioja	100	0.00	98.53	0.00	100	0.00	85.00	0.00	93.33	3.09
Valdepeñas	93.33	1.13	93.33	1.13	100	3.77	80.00	2.52	93.33	0.00
average	96.53	0.93	96.24	0.93	95.87	1.22	83.00	4.32	95.20	0.88

Table 4. Comparison of Accuracies and False Negatives in Percentage Obtained with Different Classifying Methods with 153 Samples of Red Wines and Cross-Validation Method

	SVM		NN		kNN		SIMCA		PLS-DA	
DO	ρ	FN	ρ	FN	ρ	FN	ρ	FN	ρ	FN
La Mancha	100	0.00	100	0.00	93.33	5.88	62.22	0.00	0.00	0.00
Madrid	96.00	0.00	96.00	0.00	75.00	2.88	95.00	15.52	87.00	2.56
Penedes	100	0.47	100	0.47	91.67	0.00	34.17	0.82	85.00	0.47
Rioja	100	1.78	100	1.78	80.00	3.15	76.67	0.79	80.00	1.26
Valdepeñas	93.33	0.68	93.33	0.68	86.67	0.72	93.33	0.00	86.67	0.72
Ribera Del Duero	97.14	0.44	96.67	0.44	89.29	2.24	97.50	2.27	97.14	1.31
Toro	100	0.00	100	0.00	95.00	0.00	80.00	0.00	100	6.38
Somontano	96.67	0.00	93.33	0.00	100	1.42	93.33	0.00	100	3.79
average	97.89	0.42	97.42	0.42	88.87	2.04	79.03	2.42	79.48	2.06

Table 5. Comparison of Accuracies and False Negatives in Percentage Obtained for White Wines with Different Classifying Methods with Brands Not Included in the Training Phase

	SVM		NN-MLP		kNN		SIMCA		PLS-DA	
DO	ρ	FN	ρ	FN	ρ	FN	ρ	FN	ρ	FN
La Mancha	66.67	11.11	50.00	16.67	0.00	33.33	0.00	33.33	33.33	22.22
Madrid	50.00	10.00	25.00	15.00	0.00	20.00	0.00	20.00	25.00	15.00
Penedés	100	0.00	100	0.00	100	0.00	100	0.00	100	0.00
Rioja	100	0.00	100	0.00	75.00	5.00	75.00	5.00	100	0.00
Valdepeñas	71.43	11.76	57.14	17.65	42.86	23.53	42.86	23.53	57.14	17.65
average	77.62	6.58	66.43	9.86	43.57	16.37	43.57	16.37	63.09	10.97

Table 6. Comparison of Accuracies and False Negatives in Percentage Obtained for Red Wines with Different Classifying Methods with Brands Not Included in the Training Phase

	SVM		NN		kNN		SIMCA		PLS-DA	
DO	ρ	FN	ρ	FN	ρ	FN	ρ	FN	ρ	FN
La Mancha	66.67	2.08	66.67	2.08	0.00	6.25	0.00	6.25	0.00	6.25
Madrid	66.67	2.08	66.67	2.08	33.33	4.17	33.33	4.17	66.67	2.08
Penedes	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00
Rioja	75.00	2.13	75.00	2.13	25.00	6.38	50.00	4.26	50.00	4.26
Valdepeñas	71.43	4.55	57.14	6.82	28.57	11.36	28.57	11.36	42.86	9.09
Ribera Del Duero	71.43	4.55	71.43	4.55	57.14	6.82	0.00	15.91	0.00	15.91
Toro	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00	100.00	0.00
Somontano	77.78	4.76	66.67	7.14	33.33	14.29	22.22	16.67	33.33	14.29
average	78.62	2.52	75.45	3.10	47.17	6.16	41.77	7.33	49.11	6.48

least for this particular problem. The classification of white wines seems to be more challenging than that of red wines. Particularly, white wines from La Mancha, Madrid, and Valdepeñas could hardly ever be discriminated. This misclassification may be due to the geographical proximity of these three regions which may conceal the intrinsic characteristics of the different grapes. For instance, La Mancha white wines are primarily elaborated with Airén grapes, while among the white wines from Valdepeñas used in this study one is elaborated with Macabeo and the other one with Airén grapes. The other two white wines with DO are clearly discriminated. Rioja and Penedes constitute rather different geographical locations, and both use Macabeo or Viura grapes in the elaboration of white wines, but Penedes white wines also include Xarelo and Parellada grapes. The results for eight different DO of red wines are shown in **Table 4**. In this case, the geographical proximity of three specific regions (La Mancha, Madrid, and Valdepeñas) and the predominant use of Tempranillo to elaborate these red wines do not seem to negatively affect the robustness of UV−visible spectrophotometry and SVM in classifying red wines according to their specific regions. It is not surprising that red wines from the Toro region are almost perfectly classified irrespective of the classifier used (with the exception of SIMCA). These wines are well-recognized because of their dark color, nearly black, being elaborated with "tinta de toro" (Toro ink) grape.

Although presented results point out that UV−visible spectrophotometry combined with SVM can be considered as a simple and reliable method to prevent fraud in DO, it should be considered that different brands inside a DO can produce wines with significant differences. As it was mentioned in the Materials and Methods section, the second experiment built two datasets, keeping out a brand per each DO. **Table 5** and **Table 6** show the results obtained whith this second experiments. It is interesting to see how the conclusions abovementioned with the cross-validation procedure can be applied to the results obtained with this second experiment. SVM with a linear kernel is again the best classifier for red and white wines, but the accuracy is quite worse than in the previous experiment. This is due to the fact that different brands provide different characteristics although all the brands are in specific region, as it has been reported in ref *22*. However, this experiment is also interesting to demonstrate that there are some common factors within a DO and SVM performs better than the rest of classifiers tested.

It should be mentioned that both experiments were repeated with a 1-nm step resolution with quite similar results. As one of the targets of the paper was to search for nonexpensive devices, the 10-nm resolution is presented.

Feature Selection. As was previously mentioned, the selection of the most relevant wavelengths has a double function for discrimination purposes. On the one hand, the curse of dimensionality is avoided, facilitating classification by removing those features that only add noise. But on the other hand, it allows us to verify that the wavelength resolution is sufficient enough to discriminate between wines. Once this selection is made, further variable resolutions can be tested in an effort to obtain better classification accuracy. The selection method applied in this paper is based on the wrapper approach, meaning that a classifier method has to be selected first. SVM with linear kernel was chosen since its results were better when compared with other classification techniques. The method used to evaluate the performance of a combination of features was a 5-cross validation obtained when training and testing only takes into account those features under consideration.

Since the selection of the most important wavelengths is made using the training set itself, we have separated three sets to test red wines and three different sets to do the same for white wines. The result of this exercise for red wines, plotted as global accuracy, is depicted in **Figure 5**. We start with the minimum number of features obtained from the proposed algorithm, and then the best feature of all the remaining sets is added. This figure shows that the global accuracy of the independent test sets does not improve when more wavelengths are added. In some cases, training the classifier with a larger number of variables negatively affects the resulting accuracy. This is due to the previously mentioned curse of dimensionality, reflecting

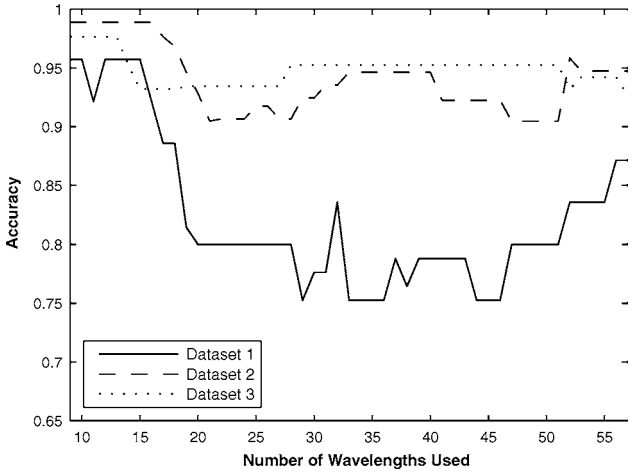

Figure 5. Global accuracy versus the number of wavelengths used for the classification of red wines.

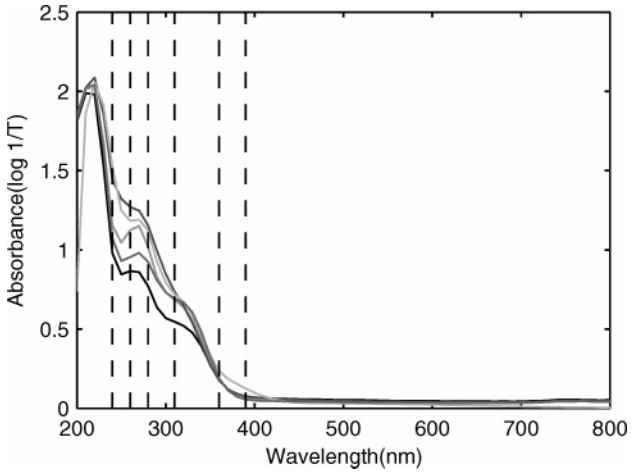

Figure 6. Selected wavelengths for white wines and representative spectra of each class

.

Table 7. Most Relevant Wavelengths after Feature Selection

wine	wavelengths selected (nm)
white	240, 260, 280, 310, 340, 360, 390
red	290, 330, 360, 490, 540, 610, 650, 760, 800

the difficulty of a classifier to work with a high number of features including wavelengths that only introduce noise into the system. **Table 7** summarizes the wavelengths obtained for each wine. In **Figure 6**, these relevant wavelengths are shown with several representative spectra for each class of white wine where it is possible to notice that the proposed wavelength selection method chose those wavelengths which were most important in separating the different wines. As expected, the relevant wavelengths in the discrimination of white wines fall within the range of 240−400 nm. Similar conclusions have recently been reported (*8*) and linked to the presence of esters from hydroxycinnamic acids. This phenomenon was equally expected to be applicable to red wines. The range of key variables, 290−800 nm, also includes the visible and NIR wavelengths which most likely reflect the presence of anthocyans, their derivatives, and/or other phenolic compounds, an expected result since the concentration and profile of wine anthocyanins is affected by the variety of grape and the vinification technique used (*21*).

In conclusion, the novelty of this paper rests on the use of UV−visible spectrophotometry combined with SVM as a reliable analytical tool in order to discriminate the wines produced in the different specific regions of Spain. The reliability of the proposed methodology is demonstrated and further validated by calculating the resulting accuracy and by comparing it with other frequently used chemometric classifiers. SIMCA, k-NN, and PLS-DA seem to require more selective techniques/variables than SVM if a large variety of wines is to be discriminated according to the specific region in which they are produced. Both SVM and NN-MPL can solve this problem using UV−visible spectral data, although SVM is preferred due to its inherent simplicity when compared to neural networks. However, it is important to take into account that each brand in the same DO can present significant differences. In order to prevent fraud it is necessary to include in the training set an important number of different brands. A second contribution of this paper consists of proposing a new method for selecting the most important wavelengths that affect the classification rate. The results obtained recommend the suppression of nonrelevant wavelengths in search of better accuracy. The importance of these conclusions relies on the design of real-time systems, where we can focus on simpler devices.

Future work will discuss the research being done on the combination of other simple instrumental techniques based on electrochemical data, e.g., voltammetry, to improve the ability to discriminate in this particular problem. Important research is also being done to improve the proposed method by means of parametric analysis of the spectra. Thus, instead of analyzing raw spectra, measurements can be obtained such as first- and second-order derivatives to obtain simpler classifiers.

ACKNOWLEDGMENT

José Mª Mínguez is acknowledged for technical advice on the wines used in this work.

LITERATURE CITED

(1) Arvanitoyannis, I. S.; Katsota, M. N.; Psarra, E. P.; Soufleros, E. H.; Kallithraka, S. Application of quality control methods for assessing wine authenticity: Use of multivariate analysis (chemometrics). *Trends Food Sci. Technol.* **1999**, *10*, 321−336.

(2) Maarse, H.; Tas, A.; Schaefer, J. Classification of wines according to type and region based on their composition. *Z. Lebensm.-Unters.-Forsch.* **1987**, *184*, 198−203.

(3) Kallithraka, S.; Tsoutsouras, S.; Tzourou, E.; Lanaridis, P. Principal phenolic compounds in Greek red wines. *Food Chem.* **2006**, *99*, 784−793.

(4) Baxter, J. M.; Crews, M. E.; Dennis, J.; Goodall, I.; Anderson, D. The determination of the authenticity of wine from its trace elements composition. *Food Chem.* **1997**, *60*, 443−450.

(5) Martin, G. J.; Guillou, C.; Martin, M. L. Natural factors of isotope fractionation and the characterization of wines. *J. Agric. Food Chem.* **1998**, *36*, 316−322.

(6) Urbano, M.; Luque de Castro, M. D.; Pérez, P. M.; García-Olmo, J; Gómez-Nieto, M. A.; Ultraviolet-visible spectroscopy and pattern recognition methods for differentiation and classification of wines. *Food Chem.* **2006**, *97*, 166−175.

(7) Urbano, M.; Luque de Castro, M. D.; Pérez, P. M.; Gómez-Nieto, M. A.; Study of spectral analytical data using fingerprint and scaled similarity measurements. *Anal. Bioanal. Chem.* **2005**, *381*, 953−963.

(8) Liu, L.; Cozzolino, D.; Cynkar, W. U.; Gishen, M.; Colb, C. B. Geographic classification of Spanish and Australian Tempranillo red wines by visible and near-infrared spectroscopy combined with multivariate analysis. *J. Agric. Food Chem.* **2006**, *54*, 6754−6759.

(9) Esbensen, K. H. *Multivariate data analysis: In practice*; Camo Process AS: Oslo, 2002.
(10) Vapnik, N. V. *The Nature of Statistical Learning Theory*; Springer-Verlag: Berlin, 2000.
(11) Kim, Y. G.; Jang, M. S.; Cho, K. S.; Park, G. T. Performance comparison between backpropagation, neuro-fuzzy network, and SVM. *Lect. Notes Comput. Sci.* **2006**, *3967*, 438−446.
(12) Zeng, F. Z.; Qiu, Z. D.; Yue, J. H.; Li, X. Q. Multiclass classification based on the analytical center of version space. *Chin. J. Electron.* **2005**, *14*, 83−86.
(13) Beltrán, N. H.; Duarte-Mermoud, M. A.; Salah, S. A.; Bustos, M. A.; Peña-Neira, A. I.; Loyola, E. A.; Jalocha, J. W. Feature selection algorithms using Chilean wine chromatograms as examples. *J. Food Eng.* **2005**, *67*, 483−490.
(14) Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discuss.* **1998**, *2*, 121−167.
(15) Platt, J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers*; Smola, A., Schölkopf, B., Schuurmans, D., Eds.; MIT Press: Cambridge, 1999; Vol. 1, pp 61−74.
(16) Chang C.; Lin C. *LIBSVM*: A library for support vector machines; 2001. http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
(17) Duda, R.; Hart, P.; Stork, D. *Pattern Classification*, 2nd ed.; Willey-Interscience: New York, 2001.
(18) Bishop, C. *Neural Networks for Pattern Recognition*; Oxford University Press: Oxford, 1995.
(19) Raymer, M. L.; Punch, W. F.; Goodman, E. D.; Kuhn, L. A.; Jain, A. K. Dimensionality reduction using genetic algorithms. *IEEE Trans. Evol. Comput.* **2000**, *4*, 164−171.
(20) Somol, P.; Pudil, P. Oscillating search algorithms for feature selection. In *Proceedings. 15th International Conference on Pattern Recognition*; IEEE Computer Society: Washington, DC, 2000; Vol. 2, pp 406−409.
(21) Barbosa-Garcia, O.; Ramos-Ortíz, G.; Maldonado, J. L.; Pichardo-Molina, J. L.; Meneses-Nava, M. A.; Landgrave, J. E. A.; Cervantes-Martínez, J. UV-vis absorption spectroscopy and multivariate analysis as a method to discriminate tequila. *Spectrochim. Acta, Part A* **2007**, *66*, 129−134.
(22) Gomez, M.; Heredia, F. Effect of the maceration technique on the relationships between anthocyanin composition and objective color Syrah wines. *J. Agric. Food Chem.* **2004**, *52*, 5117−5123.

Received for review March 5, 2007. Revised manuscript received June 6, 2007. Accepted June 27, 2007. Financial support from the Community of Madrid is acknowledged (ref UAH-CAM/2005/031 and GR/MAT/0916/2004). A. Narváez acknowledges the Spanish Ministry of Education and Science for a research contract within the Ramón y Cajal program.

JF070634Q