Feature extraction and classification of Chilean wines

N.H. Beltrán [a], M.A. Duarte-Mermoud [a,*], M.A. Bustos [a], S.A. Salah [a], E.A. Loyola [b], A.I. Peña-Neira [b], J.W. Jalocha [b]

[a] *Department of Electrical Engineering, University of Chile, Av. Tupper 2007, Casilla 412-3, Santiago, Chile*
[b] *Department of Enology and Agroindustry, University of Chile, Av. Santa Rosa 11315, Santiago, Chile*

Abstract

In this work, results of Chilean wine classification by means of feature extraction and Bayesian and neural network classification are presented. The classification is made based on the information contained in phenolic compound chromatograms obtained from an HPLC-DAD. The objective of this study is to classify different Cabernet Sauvignon, Merlot and Carménère samples from different years, valleys and vineyards of Chile. Different feature extraction techniques including the discrete Fourier transform, the Wavelet transform, the class profiles and the Fisher transformation are analyzed together with several classification methods such as quadratic discriminant analysis, linear discriminant analysis, K-nearest neighbors and probabilistic neural networks. In order to compare the results, cross validation and re-sampling techniques were used.

Keywords: Wine classification; Pattern recognition; Statistical classification; Bayesian classification; Wavelet transform; Fisher transform; Probabilistic neural networks; K-nearest neighbors

1. Introduction

During the last years, the Chilean wine industry has experienced a sustained growth, becoming one of the most important industries in the Chilean economy, reaching exportations of US$ 570 millions on 2000, US$ 590 on 2001 and US$ 610 on 2002. This growth is due to the incorporation of technology in this industry to compete in the international market.

In this work, we present the results of Chilean red wine classification, considering the varieties Cabernet Sauvignon, Merlot and Carménère from different valleys, years and vineyards. The classification is based on the information contained in phenolic compound chromatograms obtained from an HPLC-DAD.

In most of previous wine classification papers the concentration of specific compounds are the main variables on which the classification is based. Typically this concentrations are obtained from liquid and gas chromatography or other techniques, and correspond to characteristics such as major acids (Cabezudo, Herraiz, & De Gorostiza, 1983; Etievant, Schlich, Cantagrel, Bertrand, & Bouvier, 1989), anthocyanins (Aires de Sousa, 1996; Berente, García, Reichenbacher, & Danzer, 2000), free amino acids (Vasconcelos & Chaves das Neves, 1989), biogenics amines (Csomos, Heberger, & Simon-Sarkadi, 2002), isotropic ratios (Kosir, Kocjancic, Ogrinc, & Kidric, 2001), aromas (Weber et al., 1999; García, Reichenbacher, & Danzer, 1998), phenolic composition (García-Parrilla, González, Heredia, & Troncoso, 1997), color (Almela, Javaloy, Fernández-Lopez, & López-Roca, 1996; Ortiz-Fernández, Herrero-Gutiérrez, Sánchez-Pastor, Sarabia, & Iñiguez-Crespo, 1995) etc. Classification has then been done using directly the information provided by the sensors using a wide

* Corresponding author. Tel.: +56 2 678 4213; fax: +56 2 672 0162.
E-mail address: mduartem@cec.uchile.cl (M.A. Duarte-Mermoud).

variety of methods ranging from statistical (Csomos et al., 2002; Kallithraka et al., 2001; Kaufmann, 1997; Tzouros & Arvanitoyannis, 2001) to artificial neural networks and genetic algorithms (Aires de Sousa, 1996; Almela et al., 1996; Beltrán et al., 2005; Ortiz-Fernández et al., 1995; Sun, Danzer, & Thiel, 1997).

In a chromatogram the compound concentration depends on the area under the peak that appears at the time when the column liberates it, known as retention time. This time depends mainly on factors like the temperature gradient applied to the sample, column aging and compound type. The commonly used methodology in order to identify specific compounds is to establish experimental conditions and then associate a specific compound to a retention time using standard chromatographic patterns. This approach, besides of requiring a previous identification of the analyzed compounds, needs to identify which of them are the most important for a specific wine characterization, which is an open problem.

In this work, a different approach is presented, which does not requires previous compound identification, because the classification is made using the whole information contained in the chromatogram signal, instead of only the areas of some interesting peaks. The difficulty of this approach is that commonly, the chromatographic information is characterized for having a huge data volume, making a direct approach with classification techniques, like Discriminant Analysis or Neural Networks classifiers, almost impossible, because of the denominated course of dimensionality (Fukunaga & Hayes, 1989).

Nevertheless using signal analysis tools and feature extraction techniques before the classification task, it is possible to reduce the dimension of the data and to obtain wine classification rates of about 95%.

2. Description of the experimental data

In this study 172 Chilean red wine liquid chromatograms were analyzed. Chromatograms were obtained from confident samples of 80 Cabernet Sauvignon, 35 Merlot and 57 Carménère wines, cultivated in Maipo, Rapel, Curicó, Maule and Itata valleys in the central zone of Chile, between the years 2000 and 2001.

The information contained in the chromatograms, corresponds to phenolic compounds of small molecular weight obtained through a high performance liquid chromatogram (HPLC) attached with an aligned photodiode detector (DAD) (Peña-Neira, Hernández, García-Vallejo, Estrella, & Suarez, 2000). The equipment used in this study is a liquid chromatograph Merk-Hitachi model L-4200 UV–Vis Detector with internal pump, and a thermostat column holder. The column used is a Novapack C18, of 300 mm length and 3.9 mm of internal diameter. To separate different phenolic compounds the following solutions were used as solvents:

(A) 98% H_2O, 2% acetic acid,
(B) 78% H_2O, 20% acetonitrile 2% acetic acid.

The gradient used in this tests was, 0–55 min, 100% A at 1 ml/min; 55–57 min, 20% A and 80% B at 1 ml/min; 57–90 min, 10% A and 90% B at 1.2 ml/min.

Each digitalized chromatogram has a length of 6751 points and some peaks can be identified as an specific phenolic compound. These compounds have been widely studied and identified by chemical investigators and agronomic researchers (Alamo, 2002; Muñoz, 2002; Peña-Neira et al., 2000). Figs. 1–3 show typical profiles of the Chilean Merlot, Cabernet Sauvignon and Carménère red wines obtained from an HPLC-DAD suitably normalized.

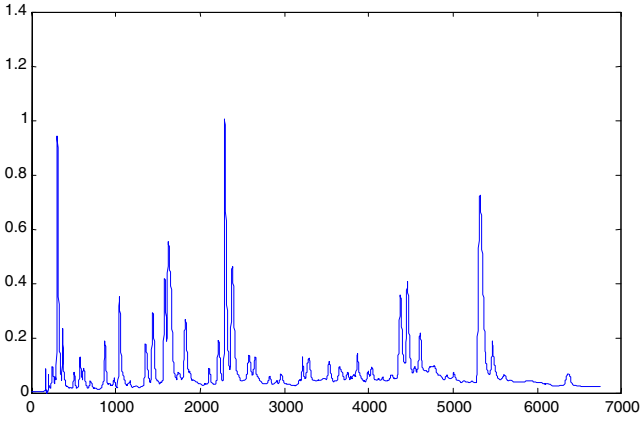

Fig. 1. Typical Chilean Merlot red wine phenolic normalized chromatogram.

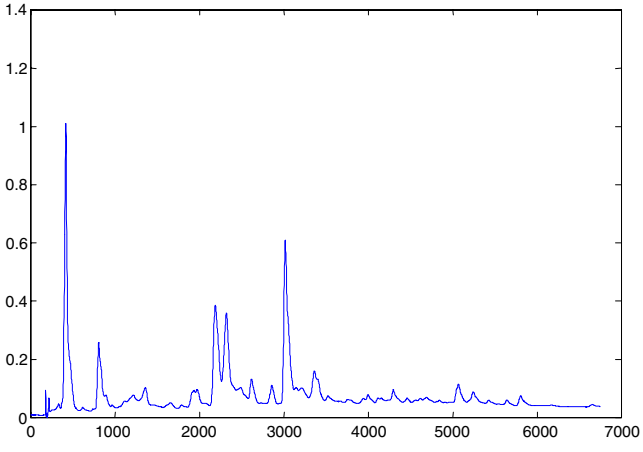

Fig. 2. Typical Chilean Cabernet Sauvignon red wine phenolic normalized chromatogram.

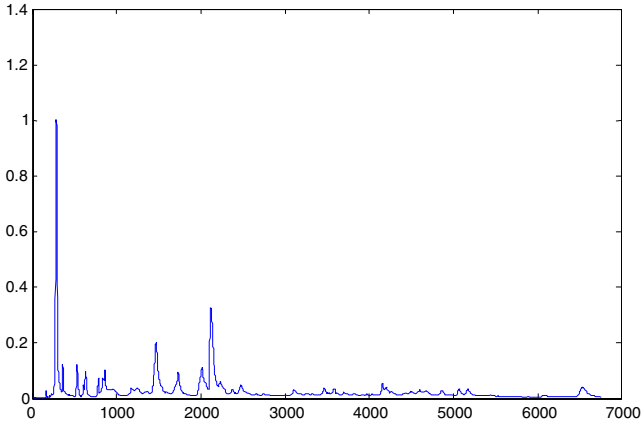

Fig. 3. Typical Chilean Carménère red wine phenolic normalized chromatogram.

In order to use efficiently the information of the chromatograms, and to reduce the data dimension, signal analysis techniques were first applied. Using the Shannon sampling theorem (Middleton & Goodwin, 1990) it can be found that the Nyquist frequency of the data is 0.1227 Hz, or equivalently a critical sample period of 4 s can be used without loosing information contained in the data. Resampling the data with this sample time allows a reduction of the length of the original chromatograms (6751 points) into one fifth of its original length (1350 points), maintaining the entire original information. Even though this reduction is significant, the dimension is still too high for using multivariate statistical techniques directly.

Another important issue is that chromatograms of the same variety present different amplitudes because different volumes of wine were used in the preparation of the samples for HPLC injection. To overcome this difficulty, a normalization procedure of the chromatogram amplitudes into $[0,1]$ was used. For this purpose the following transformation was used

$$\tilde{y} = \frac{y - y_{\min}}{y_{\max} - y_{\min}}$$

where y represents the original signal amplitude, $y_{\min}$ the minimum amplitude of the original signal, $y_{\max}$ the maximum amplitude of the original signal and $\tilde{y}$ corresponds to the mapped signal into the interval $[0,1]$.

Finally, it is important to point out that the first 5 min of each chromatogram (375 points) were disregarded, since it contain information concerning the effluents used to perform the HPLC and they are not compounds present in the wine.

3. Methodology

In this work several statistical classifiers and one probabilistic neural network (PNN) classifier were analyzed and compared. Even thought these classifiers are not the state of the art in pattern recognition, were used in this study because of their simplicity and robustness.

Two approaches for the recognition system design were explored:

(1) Parametric approach: in this case a specific functional form of the probability density is assumed (typically Multivariate Normal), but with unknown mean and covariance matrix. The classifiers studied here are the linear discriminant analysis (LDA) (Fukunaga, 1990; Webb, 2002) and quadratic discriminant analysis (QDA) (Fukunaga, 1990; Ripley, 1995; Webb, 2002).
(2) Non-parametric approach: No specific assumptions on the form of the probability density functions are made in this case. The classifiers used in this work are the PNN (Ripley, 1995; Specht, 1990) and the well-known and popular K-nearest neighbors (KNN) rule (Ripley, 1995; Webb, 2002).

Though data volume was reduced using Shannon theorem, the dimension of the data is still high. With this in mind, a feature extraction stage was used before the classification stage. In Fig. 4 is presented a block diagram, illustrating the data processing and classification stages, used in the classification system. Computational implementation of the classification system was made in MATLAB 6.1, using the version 0.3 of the "Discriminant Analysis Toolbox" (Kiefte, 1999) and the Mathworks Wavelet, Neural Network Toolbox and Signal Processing Toolboxes.

In what follows, we will briefly describe the feature extraction and classifiers used in this work.

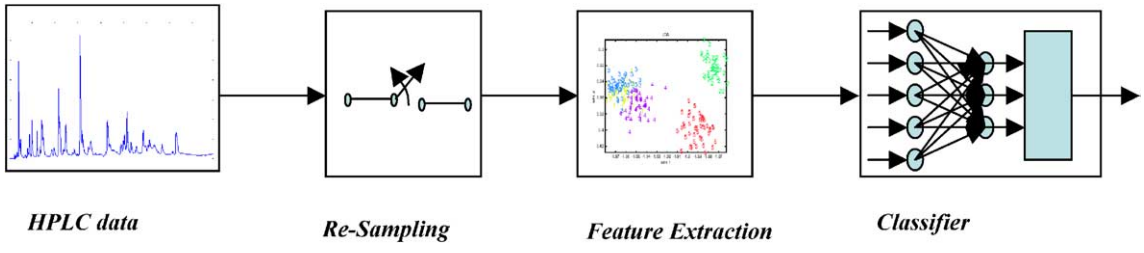

Fig. 4. Block diagram of the classification system.

3.1. Feature extraction algorithms used in the study

The main objective in the design of a pattern recognition system is to achieve the best classification rate of new patterns with respect to patterns used to train the system. This quality is known as generalization. In parametric methods one of the key factors in achieving a good generalization is to maintain the classifier complexity as low as possible, which means to reduce the number of classifier parameters. For example, weights between two layers in a neural network, or the matrix size in a linear classifier. This is achieved by maintaining a low data dimension, because these parameters directly depend on this factor. The same problem appears at the non-parametric approach, where in order to achieve a good generalization, is necessary that the number of training patterns, N, be greater than the dimension of the classification space, d. The problems associated with the relation of N and d are known as the course of dimensionality (Fukunaga & Hayes, 1989). One way to overcome this difficulties is to keep the ratio d/N as low as possible.

The different techniques used in this study for dimension reduction are presented in the next paragraphs.

3.1.1. Fourier transform (Middleton & Goodwin, 1990)

The Fourier Transform is a mathematic tool that allows the representation of a time function $f(t)$ in the frequency domain as a function $F(\omega)$. In this domain the explicit harmonic content of a signal and its distribution in terms of the frequency ω is revealed. If the temporal variable t is continuous (belong to the real numbers) then the Fourier transform is continuous (CFT) but if t belongs to a discrete set of values (typically natural union zero) we talk about discrete Fourier transform (DFT), and its definition is given by

$$F(k) = \sum_{n=0}^{N-1} f(nT) e^{-\frac{j2\pi nk}{N}}$$

$$f(nT) = \frac{1}{N} \sum_{k=0}^{N-1} F(k) e^{\frac{j2\pi nk}{N}}$$

In the communication area (signal transmission), the propriety of data compression of the DFT is well known because it allows representing a time signal into a reduced number of data points, the coefficients $f(nt)$ of the Fourier series. This propriety is to be exploited in this study by representing the whole chromatogram by the correspondent coefficients of its Fourier transform.

3.1.2. Wavelet transform (Theodoridis & Koutroumbas, 1999)

A time function $f(t) \in L^2$ (square integral) can be expressed in term of some basis functions $\psi_{j,k}(t)$ and $\phi_k(t)$ by means of the following decomposition.

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \phi_k(t) + \sum_{j=0}^{\infty} \sum_{k=-\infty}^{\infty} d_{j,k} \psi_{j,k}(t), \quad j,k \in \boldsymbol{Z}$$

where $\{\psi_{j,k}(t)\}$ and $\{\phi_k(t)\}$ are orthonormal set of basis functions of L^2 which satisfy some proprieties and are denominated wavelet functions. The coefficients c_k and $d_{j,k}$ corresponds to the discrete wavelet transform coefficients of the function $f(t)$ known as approximation and decomposition coefficients respectively, defined as follow:

$$d_{j,k} = < f(t), \psi_{j,k}(t) > = \int_{-\infty}^{\infty} f^*(t) \psi_{j,k}(t)\, \mathrm{d}t$$

$$c_k = < f(t), \phi_k(t) > = \int_{-\infty}^{\infty} f^*(t) \phi_k(t)\, \mathrm{d}t$$

The more commonly used basis are the so called Haar wavelets whose definition are given by

$$\psi_{j,k}(t) = 2^{j/2} \psi(2^j t - k) \quad \text{with } j,k \in \boldsymbol{Z}$$

$$\phi_k(t) = \phi(t-k)$$

where $\psi_{j,k}(t)$, denominated as mother wavelet, and $\phi_k(t)$ the scale function, are defined as

$$\psi(t) = \phi(2t) - \phi(2t-1)$$

$$\phi(t) = \begin{cases} 1 & \text{if } 0 \leqslant t \leqslant 1 \\ 0 & \text{other} \end{cases}$$

The sub index j is denominated decomposition level and the sub index k is known as the shifting time. In this study we employ the Haar wavelet keeping fixed the j index and considering as features the scale coefficients of the wavelet transform.

3.1.3. Class profiles

We also include in this study as feature extraction method, the distance and correlation coefficient of a test pattern to a typical class profile ω_i defined as representative of the class W_i. The class profiles ω_i were generated by minimizing the following cost function

$$\min J(\omega_1, \ldots, \omega_c) = \sum_{k=1}^{C} \left[\lambda \sum_{i \in C_k} \|X_i - \omega_k\| - (1-\lambda) \sum_{i \notin C_k} \|X_i - \omega_k\| \right] \quad \text{with } 0 < \lambda \leqslant 1$$

where C is the number of classes ($C = 3$). The main idea behind this cost function is to find the vectors (class profile) which are closest to all the data points of its own class which at the same time are far apart form the data points of the other classes. The definition of the residuals and correlation coefficients for a pattern X are presented in the Table 1.

The residuals and the correlation coefficients were calculated in the Fourier, Wavelet and time domains.

Table 1
Class profile residuals and correlation coefficients

Residuals	Correlation coefficients
$e_1 = \|\omega_1 - X\|^2$	$\rho_1 = \frac{E(X\omega_1)}{\mathrm{Var}(X)\mathrm{Var}(\omega_1)}$
$e_2 = \|\omega_2 - X\|^2$	$\rho_2 = \frac{E(X\omega_2)}{\mathrm{Var}(X)\mathrm{Var}(\omega_2)}$
$e_3 = \|\omega_3 - X\|^2$	$\rho_3 = \frac{E(X\omega_3)}{\mathrm{Var}(X)\mathrm{Var}(\omega_3)}$

The free parameter λ was chosen using a cross-validation round with a reduced number of training examples and fixed in this study at the value $\lambda = 0.75$.

3.1.4. Fisher transform (Fukunaga, 1990)

The objective of the linear discriminant analysis (LDA) is to perform dimensionality reduction while preserving as much as possible of the class discriminatory information. The dimension reduction is based on finding a linear transform $Z = MX$ such that the means of the projected variables are mapped as far apart as possible while keeping as low as possible the dispersion of the points of the class around each class mean. Fukunaga (1990) and Fukunaga and Hayes (1989) proposed the following criteria to obtain the parameters M for the multi-class problem.

$$\max_M J(M) = \mathrm{Tr}\left\{\left(MS_W M^{\mathrm{T}}\right)^{-1}\left(MS_B M^{\mathrm{T}}\right)\right\}$$

with

$$S_W = \sum_{k=1}^{C} S_k$$

$$S_k = \sum_{n\in C_k}(X_n - \hat{\mu}_k)(X_n - \hat{\mu}_k)^{\mathrm{T}}$$

$$S_B = \sum_{k=1}^{C} N_k(\mu_k - \mu)(\mu_k - \mu)^{\mathrm{T}}$$

where μ_k is the mean of the class k, μ the mean of the entire population, S_k the intra-class scatter matrix, S_B the inter-class scatter matrix, S_W the population scatter matrix, N_k the number of training patterns of the class W_k and C is the number of classes.

The solution of this problem is given by the $C-1$ eigenvectors of the matrix $S_W^{-1}S_B$ associated to the non zero eigenvalues.

3.2. Classifiers considered in this study

In what follows, a brief description of the classification algorithms used in the study is provided. For further information the reader can refer to the cited bibliography.

3.2.1. Quadratic discriminant analysis (QDA) (Fukunaga, 1990)

Among the rules used to assign objects to one of several classes, the Bayes minimum error rule (Fukunaga, 1990) is theoretically optimal, in the sense that a test pattern X, is classified into the class with the maximum a posteriori probability (MAP), or in other words, with the minimum probability of misclassification. Applying the Bayes Theorem, this rule can be expressed as: Assign the object X to the class W_k if

$$p(X|W_k)P(W_k) > p(X|W_j)P(W_j) \quad \text{for all } j \neq k \tag{1}$$

Here $p(X|W_k)$ are the class probability densities, and $P(W_k)$ is the a priori probability of the class W_k.

The densities $p(X|W_k)$ are usually unknown and have to be estimated from training samples. Quadratic discriminant analysis assumes that the distribution of the data is multivariate normal. Substituting the expression of the multivariate normal distribution and taking the logarithm in both sizes in (1) leads to the following classification index $c_k(X)$ (discriminant function quadratic in X)

$$c_k(X) = (X-\mu_k)^{\mathrm{T}}\sum_k^{-1}(X-\mu_k) + \ln\left(\left|\sum_k\right|\right) - 2\ln(P(W_k)) \tag{2}$$

where $\sum_k$ is the population covariance matrix of the class W_k and μ_k is the respective class mean. QDA assign a test object X to the class that minimizes c_k. The population parameters, $\sum_k$ and μ_k are also unknown and are replaced by their sample estimates $\hat{\mu}_k$ and $\hat{\sum}_k$

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{j\in C_k} X_j \tag{3}$$

$$\hat{\Sigma}_k = \frac{1}{n_k - 1}\sum\nolimits_{j\in C_k}^{n_k}(X_j - \hat{\mu}_k)^2 \tag{4}$$

3.2.2. Linear discriminant analysis (LDA) (Fukunaga, 1990; Webb, 2002)

As in QDA, LDA also assumes that the class populations follow a multivariate normal distribution. However LDA makes the extra assumption in the sense that the classes have identical covariance matrices:

$$\sum\nolimits_1 = \sum_2 = \cdots \sum_C = \sum$$

Under this assumption the quadratic term in the classification index (2) is the same for all classes, and therefore can be omitted. The classification index (discriminant function) can be now rewritten as

$$c_k(X) = -2X^{\mathrm{T}}\sum^{-1}\mu_k + \mu_k\sum^{-1}\mu_k - 2\log(P(W_k)) \tag{5}$$

Because the classification index is linear with respect to X the decision boundaries generated by LDA are linear and that is way Linear Discriminant Analysis is used.

As in QDA the population parameters are estimated from training data using equations (3), (4). Since fewer parameters need to be estimated (only one covariance matrix instead of C) LDA has been widely used in high dimensional pattern classification problems with good results.

3.2.3. K-nearest neighbors method (KNN) (Jain, Duin, & Mao, 2000; Webb, 2002)

One of the most popular non-parametric methods is the KNN classifier. The KNN method classifies a test objet into the class that results in the largest number of neighbors among the k nearest neighbors of the test pattern. The simplicity and the fact that KNN does not make any assumptions about the class distributions are the main strong advantages of this method. However, as all non-parametric methods, it requires a large number of training data in order to perform well in high dimensional spaces.

3.2.4. Probabilistic neural networks (PNN) (Ripley, 1995; Specht, 1990)

Probabilistic neural networks (PNN) are a class of neural networks, which combine some of the best attributes of statistical pattern recognition and feed-forward neural networks. PNNs are the Neural Network implementation of kernel discriminant analysis and were introduced into the neural network literature in the late 1980's (Specht, 1990). PNNs have very fast training times and produce outputs with Bayes a posteriori probabilities. These useful features are obtained at the expense of larger memory requirements and slower execution speed for prediction of unknown patterns if compared with conventional neural networks.

The probabilistic neural network is a three-layer, feed-forward neural network with one pass training algorithm, used for classification and mapping tasks.

Table 2
Results of the classification schemes

Extraction method	Number of features	Classifier							
		LDA		QDA		KNN		PNN	
		Average correct classification rate (%)	Variance (%)	Average correct classification rate (%)	Variance (%)	Average correct classification rate (%)	Variance (%)	Average correct classification rate (%)	Variance (%)
TFi	2	83.72	2.69	82.56	2.82	80.23	2.81	84.30	2.21
RT	3	62.21	0.26	64.53	0.75	61.63	0.54	65.12	0.16
CT	3	81.40	0.41	83.14	0.94	82.56	0.62	86.05	0.25
TFo	480	81.98	2.23	84.30	2.83	79.07	2.45	82.56	2.25
RF	3	64.53	1.85	65.70	1.45	64.53	1.65	68.02	1.44
CF	3	66.28	1.17	68.60	1.97	69.77	1.24	73.26	1.52
TW	43	90.12	0.89	87.21	1.05	88.37	0.91	91.86	0.62
RW	3	77.91	0.77	88.95	0.92	79.07	0.87	82.56	0.55
CW	3	79.07	1.25	78.49	1.44	77.33	1.32	80.81	0.91
TW + RW	46	80.23	0.74	78.49	0.88	82.56	0.95	86.05	0.58
TW + CW	46	89.53	0.98	90.70	1.18	84.88	1.15	88.37	0.68
TW + RW + CW	49	88.95	1.12	90.12	1.25	86.05	1.22	89.53	1.05
TW + RT	43	87.79	0.53	91.86	0.68	88.37	0.74	91.86	0.47
TW + CT	43	91.86	0.84	93.02	0.96	91.28	0.88	94.77	0.72
TW + RT + CT	49	88.95	0.71	88.37	0.88	88.37	0.95	91.86	0.71
TFo + RF	483	83.14	1.58	82.56	1.72	80.23	1.26	83.72	1.48
TFo + CF	483	85.47	1.27	84.88	1.44	81.98	1.32	85.47	1.15
TFo + RF + CF	486	87.79	1.43	87.21	1.85	83.14	1.51	86.63	1.25
TFo + RT	483	80.23	1.64	79.65	1.42	77.33	1.78	80.81	1.24
TFo + CT	483	79.65	1.36	80.81	1.51	75.00	1.65	78.49	1.45
TFo + RT + CT	486	89.53	1.95	90.70	2.45	82.56	1.84	86.05	1.63
TW + RF	46	91.86	0.78	91.28	1.06	91.28	0.65	93.02	0.70
TW + CF	46	91.28	0.51	90.70	0.84	88.37	0.95	91.86	0.59
TW + RF + CF	49	91.28	0.99	90.70	1.64	89.53	1.12	91.28	0.92
TFo + RW	483	87.79	1.67	90.12	2.31	88.95	1.62	92.44	1.54
TFo + CW	483	88.37	1.92	86.63	2.44	85.47	1.98	88.95	1.84
TFo + RW + CW	486	86.63	1.75	84.88	2.15	84.88	1.95	88.37	1.90

TFi: Fisher transform, RT: Residuals in time domain, CT: Correlation coeffs. in time domain, TFo: Fourier transform, RF: Residuals in frequency domain, CF: Correlation coeffs. in frequency domain, TW: Wavelet transform, RW: Residuals in frequency domain, CW: Correlation coeffs. in frequency domain.

Table 3
Results of the hypothesis test of McNemar (PNN)

T	TFi	RT	CT	TFo	RF	CF	TW	RW	CW	TW + RW	TW + CW	TW + RW + CW	TW + RT	TW + CT	TW + RT + CT	TFo + RF	TFo + CF	TFo + RF + CF	TFo + RT	TFo + CT	TFo + RT + CT	TW + RF	TW + CF	TW + RF + CF	TFo + RW	TFo + CW	TFo + RW + CW
TFi		5.10	4.95	3.48	4.99	4.77	4.25	4.6	4.12	3.97	4.56	4.66	5.22	6.48	4.62	2.21	3.92	4.15	3.97	4.05	5.42	5.36	5.21	5.22	4.87	4.25	4.22
RT			4.51	4.42	3.93	3.91	4.84	5.46	4.26	3.92	5.68	4.07	5.24	5.42	5.49	4.65	4.86	4.24	4.48	4.36	4.98	5.43	6.15	5.27	5.22	4.83	4.48
CT				4.05	4.21	4.32	5.22	4.05	4.56	0.52	4.21	5.22	5.62	6.21	5.42	4.22	3.52	2.21	4.22	5.21	1.05	5.42	5.22	5.32	5.42	4.23	4.12
TFo					5.21	4.84	5.58	0.05	2.54	4.58	4.22	4.22	5.42	6.55	5.42	3.42	3.98	4.83	4.36	5.15	5.22	4.07	5.42	5.22	4.86	4.84	4.77
RF						5.52	6.25	5.42	5.25	5.05	5.25	5.42	6.05	6.95	5.05	4.22	4.83	4.84	4.42	4.05	4.42	5.88	5.62	5.44	5.95	4.46	4.37
CF							6.48	4.62	4.55	5.25	5.72	6.02	4.92	6.22	5.22	4.25	4.22	4.32	4.26	4.05	4.38	5.22	4.05	4.62	5.02	5.44	5.82
TW								4.65	4.36	4.86	4.22	4.05	0.1	4.05	0.05	4.22	4.65	4.87	3.89	5.24	4.42	3.72	0.05	1.27	3.22	4.28	5.05
RW									4.05	4.22	5.05	4.65	4.75	6.22	5.22	4.25	4.22	4.32	4.26	4.05	4.38	5.22	4.05	4.62	5.02	5.44	5.82
CW										4.58	4.22	4.22	5.42	6.55	5.42	4.25	4.42	4.83	0.05	5.15	5.22	4.07	5.42	5.22	4.86	4.84	4.77
TW + RW											4.35	5.54	5.26	6.22	5.27	4.22	3.22	1.88	4.52	5.21	2.44	5.52	5.44	5.37	5.48	4.63	4.57
TW + CW												1.22	4.25	5.25	4.22	4.26	4.65	4.27	4.22	5.24	4.42	4.72	0.05	4.27	5.22	1.05	3.24
TW + RW + CW													4.92	6.22	5.22	4.25	4.22	4.32	4.26	4.05	4.38	5.22	4.05	4.62	4.82	2.52	2.27
TW + RT														4.05	1.05	0.22	4.22	4.83	4.36	5.15	5.22	4.07	5.62	5.44	3.77	5.05	4.28
TW + CT															5.02	5.22	5.65	4.87	5.89	6.24	4.42	4.88	5.35	4.46	3.92	5.37	5.84
TW + RT + CT																5.22	4.67	4.83	4.36	5.15	5.22	4.07	5.62	5.44	3.77	5.05	4.28
TFo + RF																	4.02	5.05	4.29	5.22	5.19	4.35	5.32	5.62	4.72	4.84	4.17
TFo + CF																		1.28	4.37	5.47	3.28	4.49	5.17	5.28	5.56	4.33	4.29
TFo + RF + CF																			4.22	5.21	1.05	5.42	5.22	5.32	5.42	4.23	4.12
TFo + RT																				5.35	4.22	4.82	4.37	5.82	5.27	5.36	5.22
TFo + CT																					5.22	4.47	5.28	5.22	4.56	4.24	4.37
TFo + RT + CT																						5.42	5.17	5.28	5.42	4.23	4.29
TW + RF																							4.01	3.94	3.47	4.88	5.40
TW + CF																								0.87	0.98	5.28	5.37
TW + RF + CF																									0.78	4.89	5.28
TFo + RW																										5.44	5.36
TFo + CW																											0.25
TFo + RW + CW																											

Unlike other ANNs, like the multilayer perceptron neural network, it is based on well-established statistical principles derived from Bayes decision theory and non-parametric kernel based estimators of probability density functions.

The PNN uses Parzen (or Parzen-like) probability distribution function estimators that approach the true underlying parent density (Theodoridis & Koutroumbas, 1999). An advantage of the PNN is that it is guaranteed to approach the Bayes optimal decision surface provided that the class probability density functions are smooth and continuous. The PNN operates by using spherical Gaussian radial basis functions centered at each training vector. The PNN probability density function estimate of an unknown vector belonging to a given class can be expressed as

$$\hat{p}_i(X) = \frac{1}{(2\pi)^{d/2}\sigma^d N_i} \sum_{j=1}^{N_i} e^{\frac{-(X-X_{ij})^{\mathrm{T}}(X-X_{ij})}{2\sigma^2}}$$

where i is the class number, j is the pattern number, X_{ij} is the jth training vector from class i, X is a test vector, N_i is the number of training vectors in class i, d is the dimension of vector X, σ is the smoothing factor (the standard deviation), and $\hat{p}_i(X)$ is the ith class probability density function estimate.

Classification decisions are consequently made in accordance with the Bayes decision rule, assigning the pattern X to the class W_k if

$$\hat{p}_k(X) > \hat{p}_i(X) \quad \text{for all } i \neq k \tag{6}$$

4. Classification results

In this section, we present the classification results obtained using the five feature extraction methods described in Section 3.1 together with the four classifiers presented in Section 3.2. The classification results are summarized in Table 2 for the 27 cases studied.

For each case the leave-one-out (LOO) validation procedure was used (Theodoridis & Koutroumbas, 1999). In this sense one sample was left out and the classification system was trained using the 171 remaining samples. Then the sample left out was presented to the classifier to determine to which class belongs. This classification procedure is the best method to be used in cases where the amount of information is low, as in our case. For each combination of feature extraction and classifier 172 tests were performed, computing the average correct classification rate and standard deviation resulting the values shown in Table 2.

Analyzing the results given in Table 2, the best results are obtained when using as feature extraction method the combination of Wavelet Transform of the resampled chromatogram together with the computation of the correlation coefficients in the time domain and the PNN classifier, reaching correct classification rates of 94.77% on average. This can be explained in part because of compression propriety of the Wavelet Transform that allows representing a chromatogram of 1351 points with a reduced profile of 43 points.

Beside, the information given by the Wavelet Transform result orthogonal with that contained in the correlation coefficient between the patterns and the typical profile of each class, providing additional information for classification purposes.

The PNN classifier uses the potential of a neural network to expand the capacity of the parametric statistical classifiers (LDA and QDA) providing slightly higher classification results.

To compare the error rate of the different feature extraction schemes studied in this work, the Hypothesis Test of McNemar (Fleiss, 1981) developed using as classifier the PNN. The results are shown in Table 3, from it can be observed that the method of using wavelet transform together with correlation coefficients and PNN gives statistical significant difference with respect to all other method tested.

The Hypothesis Test also shows that there is no a statistical significant difference between the correct classification rate of LDA and QDA when using the wavelet transform coefficients and correlation coefficients as inputs (Results not shown here for the sake of space).

5. Conclusions

Results obtained in this work are the first in classification of Chilean wines using pattern recognition techniques. The best classification scheme combining Wavelet transform, correlation coefficients and residuals, together with PNN, shows an appropriated performance, obtaining a 95% of correct classification between Cabernet Sauvignon, Merlot and Carménère wines coming from different Chilean valleys and from different harvesting years.

The methodology proposed in this work allows a significant reduction in the setup and calibration procedures times associated to the classification of HPLC data signals.

The main difficulty of this approach is the high data dimension of the patterns (6751 points in chromatogram profiles) which requires the application of feature extraction and signal analysis techniques in order to reduce the classifier input space dimension.

From all the feature extraction and signal analysis methods applied in this study, the Wavelet transform, resulted the most efficient extraction method, allowing a 10% classification rate increase over the alternative of classify directly on the original data space. Moreover, if this information is combined with class profile infor-

mation, a 15% improvement can be achieved in the classification.

Discrete Fourier transform as feature extraction method, does not improve in a significant way the system performance, mainly because even when the input space dimension of the classifier is reduced almost to a half, the quotient d/N is still larger than one.

Results obtained with class profiles were not significantly better than the others feature extraction methods presented, but combined with other techniques, specifically with the Wavelet transform coefficients, can produce high classification rates.

The results obtained here are promising. As future work and in order to improve wine classifier performance, it is suggested to explore the use of Support Vector Machines (SVM) as classifier and extraction methods based on Kernel Fisher and its non linear variants.

Acknowledgements

Results presented in this work were supported by CONYCIT- Chile, under the grant FONDEF D01-1016, "Chilean Red Wine Classification by means of Intelligent Instrumentation".

References

Aires de Sousa, J. (1996). Verifying wine origin: A neural network approach. *American Journal of Enology and Viticulture, 47*(4), 410–414.

Alamo, V.S., (2002). Caracterización de la composición fenólica de vinos comerciales Merlot y Sauvignon Blanc de la vendimia 2002, provenientes de cinco valles de Chile. (In Spanish). Agronomic Engineering Thesis, Faculty of Agronomical Sciences, University of Chile.

Almela, L., Javaloy, S., Fernández-Lopez, J. A., & López-Roca, J. M. (1996). Varietal classification of young red wines in terms of chemical and colour parameters. *Journal of the Science of Food and Agriculture, 70*(2), 173–180.

Beltrán, N. H., Duarte-Mermoud, M. A., Salah, S. A., Bustos, M. A., Peña-Neira, A. I., Loyola, E. A., et al. (2005). Feature selection algorithms using Chilean wine chromatograms as examples. *Journal of Food Engineering, 67*(4), 483–490.

Berente, B., García, D. D. L., Reichenbacher, M., & Danzer, K. (2000). Method development for the determination of anthocyanins in red wines by high-performance liquid chromatography and classification of German red wines by means of multivariate statistical methods. *Journal of Chromatography A, 871*(1–2), 95–103.

Cabezudo, M. D., Herraiz, M., & De Gorostiza, E. F. (1983). On the main analytical characteristics for solving enological problems. *Process Biochemistry, 18*, 17–23.

Csomos, E., Heberger, K., & Simon-Sarkadi, L. (2002). Principal component analysis of biogenic amines and polyphenols in Hungarian wines. *Journal of Agricultural and Food Chemistry, 50*(13), 3768–3774.

Etievant, P., Schlich, P., Cantagrel, R., Bertrand, A., & Bouvier, J. C. (1989). Varietal and geographic classification of French red wines in terms of mayor acids. *Journal of the Science of Food and Agriculture, 46*(4), 421–438.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (Second ed.). NY, USA: John Wiley.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. NY, USA: Academic Press Inc.

Fukunaga, K., & Hayes, R. R. (1989). Effects of sample size in classifier design. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*(8), 873–885.

García, D. D., Reichenbacher, M., & Danzer, K. (1998). Classification of wines by means of multivariate data analysis using the SPME/CGC-chromatograms of volatile aroma compounds. *Vitis, 37*(4), 181–188.

García-Parrilla, M. C., González, G. A., Heredia, F. J., & Troncoso, A. M. (1997). Differentiation of wine vinegars based on phenolic composition. *Journal of Agricultural and Food Chemistry, 45*(9), 3492.

Jain, A. K., Duin, R. P. W., & Mao, J. C. (2000). Statistical pattern recognition. A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(1), 4–37.

Kallithraka, S., Arvanitoyannis, I. S., Kefalas, P., El-Zajouli, A., Soufleros, E., & Psarra, E. (2001). Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin. *Food Chemistry, 73*(4), 501–514.

Kaufmann, A. (1997). Multivariate statistics as a classification tool in the food laboratory. *Journal of AOAC International, 80*(3), 665–675.

Kiefte, M. (1999). *Discriminant analysis toolbox for MATLAB*. Canada: University of Alberta Edmonton.

Kosir, I. J., Kocjancic, M., Ogrinc, N., & Kidric, J. (2001). Use of SNIF-NMR and IRMS in combination with chemometric methods for the determination of chaptalisation and geographical origin of wines (the example of Slovenian wines). *Analytica Chimica Acta, 429*(2), 195–206.

Middleton, R.H., Goodwin, G.C., (1990). Digital control and estimation. A unified approach. Prentice Hall Int. Ed.

Muñoz, L.P., (2002). Caracterización de la composición fenólica de vinos comerciales Cabernet Sauvignon y Chardonnay de la vendimia 2002, provenientes de cinco valles de Chile. (In Spanish). Agronomic Engineering Thesis, Faculty of Agronomical Sciences, University of Chile.

Ortiz-Fernández, M. C., Herrero-Gutiérrez, A., Sánchez-Pastor, M. S., Sarabia, L. A., & Iñiguez-Crespo, M. (1995). The UNEQ, PLS and MLF neural network methods in the modelling and prediction of the colour of young red wines from the Denomination of Origin 'Rioja'. *Chemometrics and Intelligent Laboratory Systems, 28*(2), 273–285.

Peña-Neira, A., Hernández, T., García-Vallejo, C., Estrella, I., & Suarez, J. A. (2000). A survey of phenolic compounds in Spanish wines of different geographical origins. *European Food Research Technology, 210*(6), 445–448.

Ripley, B. D. (1995). *Pattern recognition and neural networks*. Cambridge, UK: Cambridge University Press.

Specht, D. F. (1990). Probabilistic neural networks. *Neural Networks, 3*(1), 109–118.

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

Theodoridis, S., & Koutroumbas, K. (1999). *Pattern recognition*. NY, USA: Academic Press.

Tzouros, N. E., & Arvanitoyannis, I. S. (2001). Agricultural produces. Synopsis of employed quality control methods for the authentication of foods and application of chemometrics for the classification of foods according to their variety or geographical origin. *Critical Reviews in Food Science and Nutrition, 41*(4), 287–319.

Vasconcelos, A. M. P., & Chaves das Neves, H. J. (1989). Characterization of elementary wines of Vitis Vinifera varieties by pattern recognition of free amino acid profiles. *Journal of Agricultural and Food Chemestry, 37*, 931–937.

Webb, A. (2002). *Statistical pattern recognition* (Second ed.). NY, USA: John Wiley & Sons.

Weber, J., Beeg, M., Bartzsch, C., Feller, K. H., García, D. D., Reichenbacher, M., et al. (1999). Improvement of the chemometric variety characterization of wines by improving the detection limit for aroma compounds. *Journal of high resolution chromatography, 22*(6), 322–326.