CLASIFICACIÓN DE VINOS CHILENOS USANDO CROMATOGRAFÍA LÍQUIDA

• *M.A. Duarte-Mermoud*
Doctor en Ingeniería Electrica, Director Instituto de Investigación y Postgrado FINARQ

• *G. Ceballos-Benavides*
M. Sc. Ingeniería mención Eléctrica.

En este trabajo se presentan resultados de aplicar técnicas de extracción de características y clasificadores (estadísticos y neuronales) para determinar la cepa principal de vinos tintos chilenos del valle central. Haciendo uso de la información contenida en cromatogramas de compuestos fenólicos provenientes de un HPLC-DAD (High Performance Liquid Chromatograms), convenientemente procesada, se logra identificar la variedad del vino tinto bajo análisis. Se presentan una serie de métodos que permiten clasificar adecuadamente las cepas Cabernet Sauvignon, Merlot y Carmenere, de diferentes valles, años y viñas chilenas. Como primera etapa, con la finalidad de reducir el número de puntos del cromatograma y realizar el procesamiento de la información más rápido, se usan diferentes métodos de extracción de características del cromatograma de la muestra en proceso, tales como Transformada Discreta de Fourier, Transformación de Fischer y Perfiles Tipo por clases. Una vez reducida la cantidad de información inicialmente entregada por el HPLC, se procede a utilizar varios métodos de clasificación de patrones tales como Análisis Discriminante Lineal (LDA), Análisis Discriminante Cuadrático (QDA) y Redes Neuronales Probabilísticas (PNN)), los cuales son comparados y combinados entre sí, obteniéndose tasas de clasificación correcta de alrededor del 90 %.

1. Introducción

En los últimos años la industria del vino chileno ha experimentado un notable crecimiento, convirtiéndose en uno de los sectores industriales más dinámicos de su economía. Dado el crecimiento del sector surge la necesidad de incorporar tecnología para poder competir de mejor manera en los mercados internacionales.

En este trabajo se presentan resultados de clasificación de la variedad de vinos chilenos de las cepas Cabernet Sauvignon, Merlot y Carmenere a partir de cromatogramas de compuestos fenólicos provenientes de mediciones realizadas por un cromatógrafo de líquidos de alta eficacia con detector de fotodiodos alineados (HPLC-DAD) y procesadas mediante métodos estadísticos y redes neuronales, enfatizando comparaciones de los distintos métodos de extracción de características con fines de clasificación.

La mayoría de los trabajos previos de clasificación de vinos utilizan como variables de clasificación las concentraciones de compuestos específicos obtenidas a partir de cromatografías líquidas HPLC o cromatografía de gases [1-4]. En un cromatograma, la concentración de un compuesto depende del área del máximo (peak) que aparece en el tiempo en que la columna lo libera. Este tiempo depende de factores como el gradiente de temperatura aplicado a la muestra, envejecimiento de la columna, tipo de compuesto, etc. La metodología comúnmente utilizada es fijar las condiciones experimentales y posteriormente asociar al área un compuesto específico utilizando cromatogramas patrones. Este enfoque, además de requerir una identificación previa de los compuestos bajo análisis, necesita identificar cuáles de ellos son los más importantes para caracterizar un tipo de vino, problema que sigue abierto.

En este trabajo se presenta un enfoque distinto que no requiere la identificación previa de los compuestos presentes en el cromatograma, debido a que la clasificación se realiza utilizando toda la información contenida en el cromatograma y no solo las áreas de algunos máximos interesantes. La dificultad de este enfoque es que normalmente la información resultante de las cromatografías se caracteriza por tener un gran volumen de datos, por lo que abordar el problema directamente con técnicas de clasificación como Análisis Discriminante o Redes Neuronales, resulta complejo. Sin embargo, utilizando herramientas de análisis de señales y técnicas de extracción de características para procesar los cromatogramas, se logró realizar clasificación de cepas de vinos tintos chilenos con una certeza del orden del 90%.

Este trabajo está dividido en 6 secciones. En la Sección 2 se describe la información experimental utilizada en el estudio, mientras que la Sección 3 está dedicada a explicar la metodología empleada. En la Sección 4 se entrega una breve descripción de los métodos utilizados y los resultados obtenidos se presentan en la Sección 5. Por último, en la Sección 6 se plantean las principales conclusiones del estudio.

2. Información Experimental

En el estudio se utilizaron datos correspondientes a 172 cromatogramas de vinos tintos chilenos correspondientes a 80 vinos Cabernet Sauvignon, 35 vinos Merlot y 57 vinos Carmenere, cultivados en los valles del Maipo, Rapel, Curicó, Maule e Itata de la zona central de Chile, entre los años 2000 y 2001. Los cromatogramas corresponden a la de compuestos fenólicos de pequeño peso molecular, obtenidos mediante un análisis por cromatografía líquida de alta eficacia (HPLC) acoplada a un detector de fotodiodos alineados (DAD) [15]. El equipo utilizado es un cromatógrafo de líquidos Merck-Hitachi, modelo L-4200 UV-Vis Detector con bomba y portacolumna Thermostat. La columna utilizada es una Novapack C18, de 300 mm de longitud y 3,9 mm de diámetro interno. Para la separación de los diferentes compuestos fenólicos, en el equipo se utilizaron como solventes las siguientes soluciones : A) 98% H2O y 2% ácido acético; B) 78% H2O, 20% acetonitrilo y 2% ácido acético; C) 100% acetonitrilo. El gradiente utilizado fue: 0-55 min., 100% de A (flujo de 1 ml/min); 55-57 min., 20% de A y 80% de B (flujo de 1 ml /min); 57-90 min., 10% de A y 90% de B (flujo de 1,2 ml /min). Cada cromatograma consta de 6751 puntos y cada máximo (peak) presente corresponde a un compuesto fenólico específico. Estos compuestos han sido mayoritariamente estudiados e identificados por investigadores químicos y agrónomos activos en el área [15-17].

Cada perfil fenólico es una señal en el tiempo de 90 minutos de duración, muestreada a una tasa de 800 [ms], que corresponden a 6751 puntos en total. En la Figura 1 se presenta un perfil típico normalizado de un vino tinto Merlot entregado por el HPLC-DAD.

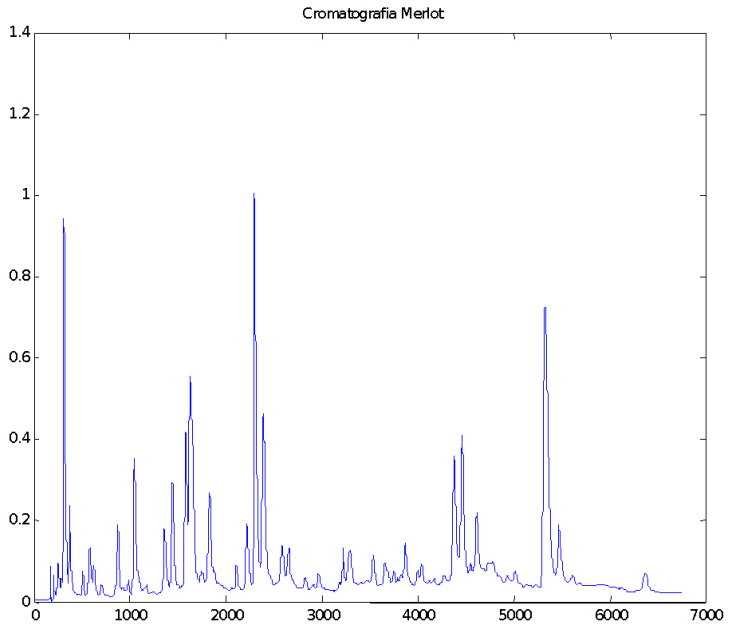

Figura 1

Cromatograma fenólico normalizado típico de un vino chileno Merlot.

La información contenida en los primeros 5 minutos del cromatograma corresponden a efluentes que se usan en el HPLC para obtener la cromatografía de líquidos y no contienen información de compuestos presentes en el vino, de modo que fueron descartados los primeros 375 puntos de cada cromatograma.

Con el propósito de aprovechar eficientemente la información contenida en los cromatogramas y reducir la dimensión de los datos, se les aplicó a éstos técnicas de análisis de señales. Aplicando el Teorema del Muestreo (Teorema de Shannon) [5], se observa que la frecuencia de Nyquist de los datos corresponde a una frecuencia igual a 0,1227 [Hz], con lo que el periodo crítico de muestreo es aproximadamente igual a 4 [s]. Esta primera observación permite concluir que la información original puede ser remuestreada, sin perder información esencial, cada 4 [s]. Como consecuencia de este procesamiento, el largo de los perfiles adquiridos mediante HPLC-DAD se reduce desde 6751 a 1350 puntos, que si bien es significativamente menor a la dimensión original, aún es muy elevada para efectos de análisis multivariable.

Por otra parte, las cromatografías de una misma variedad presentan amplitudes distintas, debido a los distintos volúmenes de vino utilizados al preparar las muestras para inyectarlas al HPLC. Para normalizar la amplitud de las cromatografías resultantes en estas mediciones al intervalo [0;1] se utilizó la siguiente transformación para los datos:

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

donde y representa la amplitud de la señal original, y_{min} , y_{max} representan la amplitud mínima y máxima respectivamente e $\tilde{y}$ corresponde a la señal mapeada en el intervalo [0, 1]

3. Metodología Empleada

En este estudio se utilizaron clasificadores estadísticos y un clasificador neuronal de la clase PNN (Probabilistic Neural Networks). Si bien estos clasificadores no corresponden al estado del arte en clasificación, se utilizaron principalmente debido a su simplicidad y robustez. Además, este estudio correspondió a una etapa inicial de una investigación de más largo plazo, tendiente a determinar la mejor manera de identificar la cepa de los vinos chilenos en base a información de tipo físico, químico y organoléptico.

En el diseño del sistema de clasificación se emplearon dos enfoques.

1. Enfoque paramétrico. En este caso se supone una distribución normal multivariable de las clases, con media y matriz de covarianza desconocidas. Los clasificadores estudiados corresponden a los denominados análisis discriminante lineal (LDA) y análisis discriminante cuadrático (QDA) [7-9, 13].

2. Enfoque no-paramétrico. En este caso se supone que la distribución de las clases es desconocida. El reconocedor utilizado en este estudio corresponde a una red neuronal probabilística (PNN) [9].

Aún cuando el volumen de datos fue reducido empleando el Teorema de Shannon, la dimensión de ellos sigue siendo grande (1350). Por esta razón y en ambos enfoques, se utilizó una etapa de extracción de características previa al reconocedor.

En la Figura 2 se presenta un diagrama en bloques que muestra el procesamiento realizado a la información, previamente a ser ingresada al sistema de clasificación.

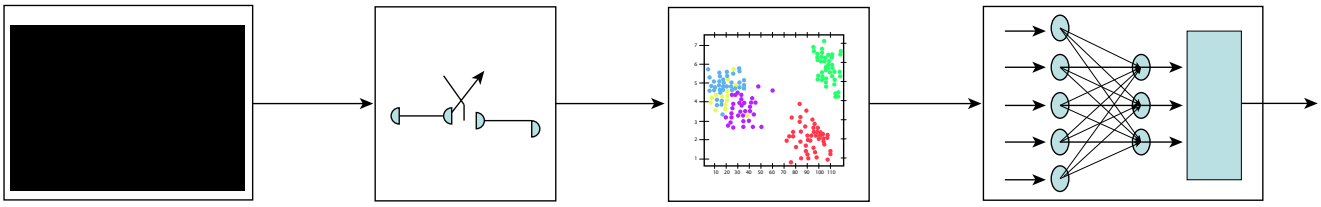

Figura 2

Diagrama en bloques del sistema de clasificación utilizado.

La implementación computacional del clasificador se realizó en MATLAB utilizando las "Discriminant Analysis Toolbox" [11], Neural Network Toolbox y Signal Processing Toolbox.

4.0 Descripción de los Métodos

A continuación, y en beneficio del espacio, se realiza una muy breve descripción de los algoritmos de extracción de características y métodos de clasificación empleados en el estudio.

4.1 Descripción de los Algoritmos de Extracción de Características

El principal objetivo del diseño de un clasificador es lograr clasificar elementos u objetos en relación a referencias o patrones con las que ha sido entrenado. Esta cualidad es conocida como generalización. En los métodos paramétricos, uno de los factores claves para obtener una buena generalización es mantener la complejidad del clasificador lo más baja posible, lo que se traduce en disminuir el número de parámetros del clasificador, como por ejemplo, los pesos entre dos capas de una red neuronal MLP (perceptron multicapa) o el tamaño de las matrices de un clasificador lineal. Esto se logra manteniendo baja la dimensión de los datos, ya que el número de pesos o parámetros del modelo depende directamente de este factor. En el enfoque no-paramétrico se tiene el mismo problema, observándose que para que el clasificador pueda generalizar bien, es necesario que el número de patrones de entrenamiento N sea mayor que la dimensión del espacio *d* de las muestras, de modo de lograr poblar el espacio de manera adecuada. Este último efecto se conoce como la *maldición de la dimensión* [10] y la forma de atacarlo es manteniendo la relación d/N lo más pequeña posible, lo que se logra nuevamente reduciendo el número de entradas al clasificador.

Debido a que la dimensión de los datos utilizados en este trabajo *d*, aún al remuestrear las cromatografías, sigue siendo mayor que N, es necesario disminuir la dimensionalidad de los datos debido a los problemas que ello genera. Con este fin se utilizaron herramientas que permiten disminuir la dimensión de los patrones y al mismo tiempo no perder información valiosa para la clasificación. Ellas se describen brevemente a continuación. Para mayores detalles el lector puede consultar las referencias citadas.

- **Transformada de Fourier (Tfo)** [5].

La Transformada de Fourier es una transformación matemática que permite representar una función del tiempo *f(t)*, en el dominio de la frecuencia, como una función *F(ω)*. En este dominio se pone de manifiesto el contenido armónico de la señal del tiempo y cómo se distribuye en términos de la frecuencia ω. Si la variable temporal t es continua la Transformada de Fourier es continua (TFC) y si t pertenece a un conjunto discreto de valores (típicamente los números naturales unión cero) se habla de la Transformada de Fourier Discreta (TFD)

Su definición está dada por la siguiente relación para el caso de funciones de tiempo discreto

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

$$(2) \qquad f(nT) = \frac{1}{N} \sum_{k=0}^{N-1} F(k) e^{\frac{j2\pi nk}{N}}$$

En el ámbito de las comunicaciones (transmisión de señales) es conocida y utilizada la propiedad de compresión de la Transformada de Fourier, ya que es capaz de representar una señal del tiempo por un número reducido de datos (los coeficientes *f(nT)* de la serie), sin perder información alguna.

En este estudio se logró determinar que una buena representación en frecuencia de la señal del tiempo remuestrada de 1350 puntos, lo constituía una descomposición que conteniendo 480 coeficientes, considerando el espectro de la señal sólo en lado positivo de las frecuencias.

- **Transformada de Fisher (Tfi)** [7]

El objetivo de la Transformación de Fisher es lograr obtener una representación de los datos en un espacio de menor dimensión, conservando la información útil para la clasificación. La idea es encontrar una transformación lineal de la forma *Z=MX*, en donde se busca que las medias de las nuevas variables *Z* de cada clase estén lo más separadas posibles y la dispersión de cada clase en torno a su media sea la menor posible. Fukunaga [7,10] propuso la siguiente función de costo o criterio para determinar *M* en el caso de elementos *X* pertenecientes a una de C clases.

(9)

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

donde

$$(10) \qquad S_w = \sum_{k=1}^{C} S_k$$

$$(11) \qquad S_k = \sum_{n \in C_k} (X_n - \hat{\mu}_k)(X_n - \hat{\mu}_k)^T$$

$$(12) \qquad S_B = \sum_{k=1}^{C} N_k (\mu_k - \mu)(\mu_k - \mu)^T$$

donde μ_k es la media de la clase *k*, μ es la media de toda la población, S_k es la matriz de dispersión intra-clases, S_B es la matriz de dispersión inter-clases y S_w es la matriz de dispersión de toda la población.

La solución a este problema está dada por la matriz *M* formada por los (C-1) vectores propios asociados a valores propios distintos de cero de la matriz $S_W^{-1}S_B$. Además, los valores propios representan el poder discriminante de las direcciones asociadas a sus vectores propios, llamadas componentes principales.

En este estudio la dimensión de la matriz M es de 2 x 1350, lo cual reduce dramáticamente el número de características a analizar en el espacio transformado de Fisher, desde 1350 a sólo 2.

- **Perfiles Tipo**

Además de los métodos anteriores de extracción de características se generaron los denominados Perfiles Tipo, los cuales resultan de minimizar la distancia entre el perfil tipo deseado y los elementos de la clase que representa, pero procurando además que estén lo más lejos posible de los elementos de las clases vecinas. En términos estrictos los perfiles tipo se obtienen mediante la solución de

(13) $$\min J(\omega_1,...,\omega_c)=\sum_{k=1}^{C}\left[\lambda\sum_{i\in C_k}\|X_i-\omega_k\|-(1-\lambda)\sum_{i\notin C_k}\|X_i-\omega_k\|\right]$$

con $0<\lambda\le 1$

Una vez obtenidos los 3 perfiles tipo representativos de cada clase, ω_1, ω_2 y ω_3, se generan dos conjuntos de 3 características para cada elemento (patrón) nuevo X que se desea clasificar, los que dominaremos residuos y coeficientes de correlación, calculados como se indica en la Tabla 1.

Residuos (R)	Coeficientes de Correlación (C)
$e_1=\|\omega_1-X\|^2$	$\rho_1=\frac{E(X\omega_1)}{Var(X)Var(\omega_1)}$
$e_2=\|\omega_2-X\|^2$	$\rho_2=\frac{E(X\omega_2)}{Var(X)Var(\omega_2)}$
$e_3=\|\omega_3-X\|^2$	$\rho_3=\frac{E(X\omega_3)}{Var(X)Var(\omega_3)}$

Tabla1.- Definición de Residuos y Coeficientes de Correlación con respecto a Perfiles Tipo.

Por ejemplo, e_1 representa la distancia Euclidiana entre el patrón desconocido *X* y el perfil tipo de la clase 1, ω_1 . El índice ρ_1 representa cuán correlacionado está el patrón desconocido *X* con el perfil tipo de la clase 1, ω_1 .

Los residuos y coeficientes de correlación fueron aplicados en los distintos espacios resultantes de las transformaciones descritas anteriormente (tanto en el dominio del tiempo como en el de la frecuencia), como una etapa de extracción complementaria. Reduciendo el número de características a analizar desde 1350 a 3, en el caso de trabajar en el dominio del tiempo y desde 480 a 3, en el caso de trabajar en el dominio de la frecuencia. Para este estudio se utilizó $\lambda=0.75$ y fue determinado empíricamente.

4.2. Descripción de los Algoritmos de Clasificación

En esta Sección se describen brevemente los algoritmos de clasificación usados en el estudio. Para mayores detalles el lector puede consultar las referencias citadas.

- **Análisis Discriminante Cuadrático (QDA) [7] [12]**

En el ambiente de reconocimiento de patrones existe una gran variedad de criterios o reglas para asignar un objeto (patrón) a una de entre *C* clases. De todos estos criterios, la regla de mínimo error de Bayes es el óptimo teórico en el sentido que minimiza la probabilidad de realizar la asignación de manera incorrecta. Esta consiste en que dado un patrón desconocido *X*, se calculan las probabilidades a posteriori de que este patrón pertenezca a cada una de las *C* clases , $P(W_j/X)$ y éste es asignado a la clase con la máxima probabilidad a posteriori. Es decir el patrón *X* es asignado a la clase j si y sólo si

(14) $$P(W_i/X)\ge P(W_k/X)\forall k\ne j$$

En este enfoque, las probabilidades $P(W_j/X)$ son calculadas utilizando el Teorema de Bayes

(15) $$P(W_j/X)=\frac{p(X/W_j)P(W_j)}{P(X)}$$

donde $p(X/W_j)$ corresponde a las densidad de probabilidad de la clase W_j y $P(X)$ la probabilidad total de *X*.

Dado que para todas las clases *P(X)* es constante y si se supone que $P(W_i)=P(W_j)\ \forall i,j\in\{1,..,C\}$, es decir que todas las clases tienen igual probabilidad a priori, basta comparar las densidades *p (X/W$_i$)*.

La distribuciones de probabilidades *p (X/W$_i$)* son usualmente desconocidas y deben ser estimadas a partir de las muestras de identificación o entrenamiento. El análisis discriminante cuadrático supone que la distribución de los datos sigue una distribución normal multivariable. Si se sustituye la expresión de una distribución normal multivariable y se toma el logaritmo natural en ambos lados de la regla de Bayes (15) se obtienen los siguientes índices de clasificación.

$$C_k=(X-\mu_k)^T\Sigma_k^{-1}(X-\mu_k)+\ln(|\Sigma_k|)-2\ln(P(W_k))$$

donde Σ_k es la matriz de covarianza de la población de la clase W_k y μ_k corresponde a la media de la clase *k*. El método QDA asigna el patrón desconocido a la clase *i* que obtenga el menor C_k . En la práctica la matriz Σ_k y las medias de las clases μ_k son desconocidas, por lo que se reemplazan por los siguientes estimadores

$$\hat{\mu}_k=\frac{1}{n_k}\sum_{i=C_k}^{n_k}X_i \qquad \hat{\Sigma}_k=\frac{1}{n_k-1}\sum_{i=C_k}^{n_k}(X_i-\hat{\mu}_k)^2$$

- **Análisis Discriminante Lineal (LDA) [7, 8, 12]**

Al igual que QDA, el método LDA supone que la población sigue una distribución normal multivariable. La diferencia es que LDA realiza una hipótesis extra suponiendo que las matrices de covarianza de las clases son iguales, o sea

$$\Sigma_1=\Sigma_2=...\Sigma_C=\Sigma$$

Bajo esta hipótesis los índices de clasificación C_k se simplifican a:

$$C_k=2X^T\Sigma^{-1}\mu_k+\mu_k\Sigma^{-1}\mu_k-2\log(P(W_k))$$

Dado que los términos cuadráticos desaparecen sólo hay que estimar los parámetros de una matriz de covarianza y las medias de cada clase. El método LDA necesita un menor número de patrones de entrenamiento que el esquema QDA.

- **Probabilistic Neural Networks (PNN) [9]**

Las redes neuronales probabilísticas (PNN por sus sigla en inglés) son una clase de redes neuronales que combinan las cualidades

de clasificadores estadísticos y redes neuronales prealimentadas. Las PNN son la implementación neuronal de análisis discriminante mediante estimadores del tipo kernel. Las principales ventajas de las PNN son la rapidez de su algoritmo de aprendizaje y su cualidad de aproximar arbitrariamente bien la regla de Bayes para cualquier densidad de probabilidad que describa las clases, siempre que ésta sea continua y suave y se tengan suficientes patrones de entrenamiento.

La estimación de las densidades de probabilidades de cada clase se realiza utilizando funciones de base radial centradas en cada patrón de entrenamiento.

5. Resultados Obtenidos

En este estudio se comparó el desempeño de los tres clasificadores presentados en la Sección 4.2, al utilizar las distintas técnicas de extracción de características descritas en la Sección 4.1. Los resultados del empleo de los métodos ya descritos, utilizando la base de datos de vinos chilenos de la Sección 2, se encuentran resumidos en la Tabla 2.

Dado que la cantidad de datos utilizados en este estudio es reducida, el proceso de validación se efectuó mediante validación cruzada leave-one-out (LOO). Este procedimiento consiste en ir entrenando el sistema dejando fuera una muestra, la cual es usada luego con fines de validación [6]. Los valores indicados en la Tabla 2 corresponden al valor medio y la varianza de los 172 experimentos de validación realizados en LOO para cada método de extracción/ clasificación.

		LDA		QDA		PNN	
Método Extracción	**Número de características**	**Promedio Clasificación Correcta**	**Varianza Clasificación Correcta**	**Promedio Clasificación Correcta**	**Varianza Clasificación Correcta**	**Promedio Clasificación Correcta**	**Varianza Clasificación Correcta**
Tfi	2	83,72%	2.69 %	82,56%	2.82 %	84,30%	2.21 %
RT	3	62,21%	0.26 %	64,53%	0.75 %	65,12%	0.16 %
CT	3	81,40%	0.41 %	83,14%	0.94 %	86,05%	0.25 %
Tfo	480	81,98%	2.23 %	84,30%	2.83 %	82,56%	2.25 %
RF	3	64,53%	1.85 %	65,70%	1.45 %	68,02%	1.44 %
CF	3	66,28%	1.17 %	68,60%	1.97 %	73,26%	1.52 %
TFo+RF	483	83,14%	1.58 %	82,56%	1.72 %	83,72%	1.48 %
TFo+CF	483	85,47%	1.27 %	84,88%	1.44 %	85,47%	1.15 %
TFo+RF+CF	486	87,79%	1.43 %	87,21%	1.85 %	86,63%	1.25 %
Tfo+RT	483	80,23%	1.64 %	79,65%	1.42 %	80,81%	1.24 %
TFo+CT	483	79,65%	1.36 %	80,81%	1.51 %	78,49%	1.45 %
TFo+RT+CT	486	89,53%	1.95 %	90,70%	2.45 %	86,05%	1.63 %

TFi: Transformada Fisher.
RT: Residuos en el tiempo.
CT: Coef. de Correlación en el tiempo
TFo: Transformada Discreta de Fourier.
RF: Residuos en frecuencia.
CF: Coef. de Correlación en frecuencia.

Tabla 2.- Resumen de los resultados de clasificación para los diferentes esquemas estudiados.

La extracción mediante Transformada de Fourier, en conjunto con los coeficientes de correlación y residuos con respecto a los perfiles tipo de cada clase, en el dominio del tiempo, y usando clasificación mediante discriminación cuadrática, resultó ser el esquema más exitoso, según puede apreciarse en la Tabla 2, lográndose porcentajes promedios de clasificación correcta de 90,7 %. Esto se explica ya que Transformada de Fourier resulta ortogonal con la información contenida en los coeficientes de correlación y residuos (en el tiempo) con respecto a los perfiles tipo, por lo que éstos aportan importante información adicional para la clasificación, elevando la tasa de clasificación correcta desde 84,30% (TFo sola) hasta 90,70% , para el clasificador QDA.

Para comparar la tasa de error de los clasificadores entre ellos, se utilizó el Test de Hipótesis de McNemar [14]. Para esta metodología, si el índice $t > 3.841$ entonces la probabilidad de que el desempeño de los 2 clasificadores en comparación sea igual, es menor a 0.05.

En la Tabla 3 se observa una diferencia en el desempeño obtenido con QDA al clasificar utilizando los coeficientes de la Transformada Fourier en conjunto con los residuos y coeficientes de correlación en el tiempo (TFo+RT+CT), que es estadísticamente significativa frente a los otros métodos de extracción, exceptuando aquellos que usan CT, TFo+CF y TFo+RF+CF, como métodos de extracción de características.

Es importante destacar aquí que la tasa de clasificación promedio de 90.7% alcanzada por el mejor sistema de clasificación aquí estudiado, significa que sólo 16 de las 172 muestras que conforman la base de estudio fueron mal clasificadas.

6. Conclusiones

Los resultados obtenidos en este trabajo son los primeros en clasificar vinos Chilenos utilizando técnicas de extracción y clasificación, en base a información de compuestos fenólicos de bajo peso molecular.

t	Tf	RT	CT	TFo	RF	CF	TFo+RF	TFo+CF	TFo+RF+CF	TFo+RT	TFo+CT	TFo+RT+CT
TFi		3,10	0,00	0,00	2,50	1,82	0,00	0,01	0,14	0,07	0,03	0,48
RT			3,34	3,52	0,01	0,13	3,12	3,73	4,80	2,09	2,43	6,33
CT				0,00	2,66	2,00	0,00	0,00	0,10	0,11	0,05	0,41
TFo					2,85	2,17	0,00	0,00	0,07	0,15	0,07	0,34
RF						0,03	2,55	3,03	4,13	1,59	1,89	5,50
CF							1,85	2,26	3,16	1,07	1,29	4,42
TFo+RF								0,01	0,14	0,07	0,03	0,48
TFo+CF									0,05	0,19	0,10	0,28
TFo+RF+CF										0,50	0,35	0,07
TFo+RT											0,00	1,05
TFo+CT												0,82
TFo+RT+CT												

Tabla 3.- Test de Hipótesis de McNemar para el clasificador basado en QDA.

El método de clasificación basado en QDA, utilizado en combinación con técnicas de extracción de características, basadas en la Transformada de Fourier en conjunto con los residuos y coeficientes de correlación con respecto a perfiles tipo, en el dominio del tiempo, muestra un desempeño bastante adecuado, lográndose un promedio de 90,7 % de clasificación correcta del tipo de cepa entre vinos Cabernet Sauvignon, Merlot y Carménère, provenientes de distintos valles de Chile y que se caracterizaron por pertenecer a años de cosecha distintos.

La principal dificultad encontrada en este trabajo se relaciona con la alta dimensionalidad de los datos (6751 puntos en los perfiles de cromatografía) lo que requiere de la aplicación de técnicas para disminuir la dimensión del espacio de entrada al clasificador.

La Transformada Discreta de Fourier resultó ser un buen método de extracción de características, permitiendo aumentar la tasa de clasificación con respecto a métodos en el dominio del tiempo. Además, si esta información se combina con la información proveniente de los perfiles tipo (coeficientes de correlación y residuos) se logra obtener mejoras del orden del 10% con respecto a métodos en el dominio del tiempo.

Los residuos obtenidos con los perfiles tipo (tanto en el dominio del tiempo como en el de la frecuencia) no resultaron en general adecuados como espacio de clasificación, aunque disminuyen notablemente la dimensión de los datos a sólo 3. No obstante, al combinarlos con la información de la Transformada de Fourier se logró mejorar las tasas de clasificación basadas en Transformada Fourier sola, del orden de un 5%.

Los resultados obtenidos son prometedores considerándose como trabajo futuro, utilizar Support Vector Machines y extracción basada en Kernel Fisher y Wavelets para mejorar el desempeño del clasificador de vinos.

Agradecimientos

Los resultados obtenidos en este trabajo fueron financiados por CONICYT-Chile, a través del proyecto FONDEF D01-1016, "Identificación varietal de vinos Chilenos mediante instrumentación Inteligente".

Referencias

[1] Cabezudo M.D., M.Herraiz and Gorostiza de E.F., "On the main analytical characteristics for solving enological problems", Process Biochemistry, vol. 18, August 1983, pp. 17-23.
[2] Etievant P. and Schlich P. "Varietal and geographic classification of French red wines in terms of Mayor Acids", Journal of the Science of Food and Agriculture, vol. 46, 1989, pp. 421-438.
[3] J. Aires-de-Sousa, "Verifying wine origin: A neural network approach", American Journal of Enology and Viticulture, vol. 47, No. 4, 1996, pp. 410-414.
[4] Vasconcelos A.M.P.and das Neves H.J, "Characterization of elementary wines of Vitis Vinifera varieties by pattern recognition of Free Amino Acid profiles", Journal of Agricultural and Food Chemestry, vol. 37, 1989, pp. 931-937.
[5] Middleton R.H. and G.C. Goodwin, Digital control and estimation. A unified approach. Prentice Hall Int. Ed, 1990.
[6] Theodoridis S.and Koutroumbas K., Pattern recognition, Academic Press, 1999.
[7] Fukunaga K., Introduction to statistical pattern recognition. Academic Press Inc, 1990.
[8] Webb A.R., Copsey K.D. Statistical pattern recognition. John Wiley & Sons, Third Edition, 2011.
[9] Ripley B.D., Pattern recognition and neural networks. Cambridge University Press, 2014.
[10] Fukunaga K. and Hayes R.R., "Effects of sample size in classifier design". IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 11, No. 8, 1989, pp. 873-885.
[11] Kiefte M., Discriminant analysis toolbox. University of Alberta Edmonton, Canada, 2000.
[12] Aeberhard S., de Vel O. and Coomans D., "Comparative analysis of statistical pattern recognition methods in high dimensional settings". Pattern Recognition , vol. 27, 1994, pp.1065-1077.
[13] Jain A.K., Duin R.P.W. and Mao J., "Statistical pattern recognition: A review". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, No. 1, 2000, pp. 4-37.
[14] Fleiss J. L., Statistical methods for rates and proportions. John Wiley, Third Edition. 2003.
[15] Peña-Neira A.I., Hernández T., García-Vallejo C., Estrella I. and Suarez J., "A survey of phenolic compounds in Spanish wines of different geographical origins". Eur. Food. Res. Technol., vol. 210, 2000, pp. 445-448.
[16] Alamo V.S. Caracterización de la composición fenólica de vinos comerciales Merlot y Sauvignon Blanc de la vendimia 2002, provenientes de cinco valles de Chile. Memoria de Ingeniero Agrónomo, Facultad de Ciencias Agronómicas, Universidad de Chile, 2002.
[17] Muñoz L.P. Caracterización de la composición fenólica de vinos comerciales Cabernet Sauvignon y Chardonnay de la vendimia 2002, provenientes de cinco valles de Chile. Memoria de Ingeniero Agrónomo, Facultad de Ciencias Agronómicas, Universidad de Chile, 2002.