FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE

Néstor Fredy Pérez Pérez

ISBN: 978-84-693-4053-0
Dipòsit Legal: T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Néstor Fredy Pérez Pérez

Fiabilidad de clasificación con PLS discriminante

Tesis doctoral

Dirigida por

Dr. Ricard Boqué Martí y Dr. Joan Ferré Baldrich

Departamento de

Química Analítica y Química Orgánica

Grupo de

Quimiometría, Cualimetría y Nanosensores

Universitat Rovira i Virgili

TARRAGONA
2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI

DEPARTAMENT DE QUÍMICA ANALÍTICA
I QUÍMICA ORGÀNICA

Campus Sescelades
Marcel·lí Domingo, s/n
43007 Tarragona
Tel. 34 977 55 81 37
Fax 34 977 55 95 63
e-mail: secqaqo@quimica.urv.es

El Dr. RICARD BOQUÉ MARTÍ i el Dr. JOAN FERRÉ BALDRICH, Professors Titulars del Departament de Química Analítica i Química Orgànica de la Facultat de Química de la Universitat Rovira i Virgili,

CERTIFIQUEM:

Que la present Tesi Doctoral, que porta per títol: "**Fiabilidad de clasificación con PLS discriminante**", presentada per en **Néstor Fredy Pérez Pérez** per optar al grau de Doctor en Química, ha estat realitzada sota la nostra direcció, a l'Àrea de Química Analítica del Departament de Química Analítica i Química Orgànica d'aquesta universitat, i que tots els resultats presentats són fruit d'experiències realitzades per l'esmentat doctorand.

Tarragona, març de 2010

Dr. Ricard Boqué Martí Dr. Joan Ferré Baldrich

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Agradecimientos

Al enumerar a todas aquellas personas a las que directa o indirectamente debo agradecer por completar esta nueva etapa de mi vida me harían falta nombres, mas no su recuerdo que siempre me ha acompañado y acompañará.

He de agradecer muy especialmente a mis tutores por la oportunidad de trabajar a su lado, su guía constante y conocimientos brindados. Al Prof. Joan por su paciencia y al Prof. Ricard por su interés constante y haberme llamado aquel 11 de enero.

Quiero agradecer al maestro F. Xavier Rius, por haberme recibido en su grupo, su disposición e interés por resolver las dudas burocráticas y académicas.

Agradezco a todos los compañeros de grupo, pasados y presentes, por su amistad y disposición a la hora de resolver las dudas surgidas.

A tod@s aquell@s amig@s que desde el otro lado del charco me han estado haciendo barra.

También quiero agradecer al proyecto TRACE, por el soporte financiero, y a los miembros de los grupos de trabajo WP1 y WP2, especialmente al Dr. Gerard Downey, que aportaron los datos con los que se trabajó.

Y por último, y no por ello menos importante, a toda mi familia que me ha acompañado de corazón.

A todos ellos, gracias totales.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Dedicado a mis Abuelos, Padres y Hermanita
Sin ellos, no estaría aquí

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Contenidos

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Capítulo 1

Introducción y objetivos de la tesis

1.1 El proyecto TRACE

El etiquetado de alimentos está regulado por el RD 1334/1999 y la directiva 2000/13/CE del Parlamento Europeo [1–3], que establecen la información que debe contener la etiqueta, como composición del alimento y/o valor nutricional, y en ciertas ocasiones características especiales que le dan un valor añadido, por ejemplo, su denominación de origen [4–6]. En 2005 y en cumplimiento del Reglamento 178/2002/CE del Parlamento Europeo, se sumó al etiquetado información sobre la trazabilidad del alimento, que puede considerarse como el historial de la cadena de producción del alimento. Este mecanismo de trazabilidad ha sido desarrollado por la Unión Europea (UE) dentro de los programas de seguridad alimentaria para ser implementado en la industria de alimentos [7]. No obstante, sólo parte de la información que aparece en las etiquetas puede ser verificada, como el valor nutricional, y no hay mecanismos precisos que permitan comprobar que la información de trazabilidad sea veraz. Respondiendo a ello se han desarrollado metodologías analíticas que permiten identificar el origen de ciertos alimentos, por ejemplo: identificar carnes alemanas de otras argentinas o chilenas [8], diferenciar quesos

italianos de tres regiones distintas [9] o mantequillas de diferentes países de la UE [10]. Estas metodologías analíticas tienen la ventaja que son especificas para un producto o región, y han demostrado que el análisis isotópico de bioelementos puede ser una herramienta idónea para confirmar el origen geográfico de los alimentos [11].

El proyecto "*TRAcing food Commodities in Europe* (TRACE)" [EU IP 006942], se marcó como objetivo prioritario, "*desarrollar sistemas y métodos de trazabilidad que permitan al consumidor tener una mayor confianza en la autenticidad de los alimentos de origen europeo*" [12]. Dicho desarrollo se realizó siguiendo un objetivo secundario: "*desarrollar sistemas y métodos de trazabilidad genéricos, específicos y económicos que permitieran verificar fácilmente que un alimento es auténtico desde el punto de vista de su origen geográfico, de especie y de producción del alimento*". Es decir, metodologías fáciles y de bajo coste.

El proyecto TRACE fue desarrollado por un conjunto interdisciplinario de instituciones repartidas en 13 grupos de trabajo. De ellos, tres se dedicaron al desarrollo de herramientas analíticas y otros tres al de sistemas de trazabilidad. El grupo de Quimiometría, Cualimetría y Nanosensores de la Universidad Rovira i Virgili estuvo vinculado al grupo de trabajo WP6 encargado de las especificaciones estadísticas. Su función fue desarrollar métodos multivariantes para el análisis de datos derivados de los análisis químicos. En concreto, desarrollar y utilizar métodos de reconocimiento de patrones, que permitieran hallar correlaciones entre el alimento y su origen (Figura 1-1). Esta correlación se da porque las características geológicas y climatológicas de una zona se ven reflejadas en la concentración de elementos traza e isótopos de bioelementos presentes en las plantas y animales [8–11]. Así pues, en la etapa de afianzamiento y desarrollo de las metodologías se hicieron análisis elementales e isotópicos de suelos y, además, de carnes, cereales, mieles, aceites de oliva y aguas minerales. El trabajo se realizó en coordinación con los grupos de trabajo WP1, que desarrolló y validó métodos que permitieran "localizar geográficamente el origen del alimento" y el WP2 que verificó mediante "huella dactilar y perfiles químicos" la autenticidad del alimento. Los datos analíticos suministrados por estos grupos permitieron verificar y validar las metodologías estadísticas y, además, desarrollar los programas necesarios en lenguaje *Matlab.*

Dentro de las metodologías de reconocimiento de patrones utilizadas por el WP6 se encuentran el *Discriminant Partial Least Squares* (DPLS), *Soft Independent Modelling of Class Analogy* (SIMCA), *Classification and Regression Trees* (CART), *Support Vector Machines* (SVM) y diferentes tipos de redes neuronales (*Neural Networks* (NN)). El Grupo de Quimiometría, Cualimetría y Nanosensores se centró en mejorar y establecer especificaciones de producto mediante DPLS.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

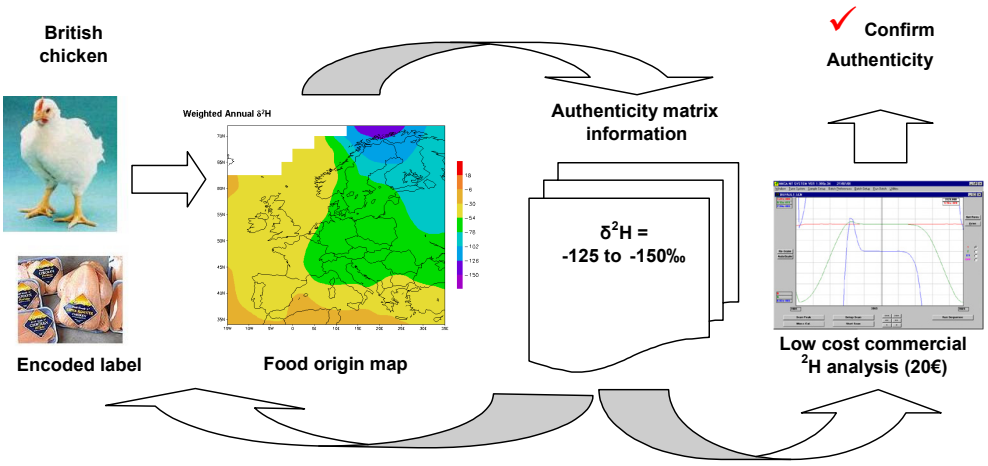

Figura 1-1: Ejemplo de verificación del origen de carne de pollo a partir del cambio en la relación isotópica de hidrogeno (δ^2H). Se contrasta la información de etiquetado de trazabilidad del pollo, que contiene la información del lugar de nacimiento y crianza, con bases de datos que contienen análisis isotópicos para pollos criados en los países de la UE. Se busca que la información contenida dentro del etiquetado coincida con los parámetros ya establecidos para pollos (en este caso británicos), verificando así su autenticidad.

1.2 Fiabilidad en la clasificación

Fiabilidad es la cualidad de fiable o probabilidad de buen funcionamiento de algo [13], y hace referencia a la seguridad, credibilidad o buenos resultados que se esperan de ese algo. La fiabilidad como probabilidad es un concepto ampliamente usado en muchas ramas de la ciencia, en donde hay un trabajo continuado para mejorar su cálculo y eliminar factores que la puedan afectar negativamente.

Una aplicación práctica de la fiabilidad podemos verla en la espectroscopia de masas, donde se halla el grado de correlación entre el espectro de una muestra desconocida y el espectro de una sustancia en la base de datos de espectros de masas [14]. La fiabilidad es la probabilidad de que ambos espectros sean el mismo. Otra aplicación es la fiabilidad de clasificación, es decir un valor de confianza en la correcta asignación de un objeto a una clase. Calcular este valor de confianza no es fácil, ya que depende del método de clasificación y de la estrategia para calcular la fiabilidad. Así podemos encontrar fiabilidades calculadas para CART, probando varias configuraciones de árboles hasta encontrar la más adecuada, y por ende la más fiable [15] o también para

DPLS, donde la estrategia de cálculo utiliza dos formas de validación cruzada junto con la selección de variables para finalmente medir la habilidad de clasificación del modelo, dándole mayor fiabilidad al método con mayor habilidad [16]. Otra aplicación es la combinación de clasificadores con la estrategia de votación. En este caso se calcula la fiabilidad de los clasificadores, lo que permite decidir si se rechaza su voto o se pondera con el valor de fiabilidad [17]. También podemos encontrar cambios en la estrategia para hallar la fiabilidad de clasificación, como encontrar regiones fiables y no fiables, en vez de un valor de fiabilidad. En este caso la seguridad en la asignación se da si el objeto cae en la zona fiable y se duda si cae en la zona no fiable [18].

Como se puede ver, el cálculo de fiabilidad en la clasificación es un tema importante y actual. No sólo basta con asignar el objeto a una clase, sino que además se debe saber la probabilidad de que esta asignación sea correcta. Aunque los métodos para calcular la fiabilidad ya referenciados han tratado de responder a ello, tienen la desventaja que basan el cálculo de fiabilidad en las respuestas de los datos de entrenamiento, dando lugar a un valor constante que, aunque correcto, sólo es válido para los datos de entrenamiento. Por ello, al asignar fiabilidades a objetos desconocidos se pueden cometer errores, ya que no se evalúa la información que éstos aportan. Por ejemplo, no se tiene en cuenta la posición del objeto desconocido en el espacio de variables ni el error con que el modelo lo predice.

1.3 Objetivos de la tesis

Dado que el desarrollo de esta tesis tuvo como marco los objetivos propuestos en el proyecto TRACE y el desarrollo de nuevos métodos quimiométricos en DPLS, se ha fijado el siguiente objetivo general:

Implementar nuevas técnicas quimiométricas en DPLS que aumenten la seguridad en la autentificación de alimentos.

Este objetivo, se concreta con los siguientes objetivos específicos:

- Desarrollar un algoritmo DPLS probabilístico que permita calcular la fiabilidad de clasificación en modelos de clasificación binarios.

- Desarrollar un programa en lenguaje *Matlab*, para el análisis y clasificación con fiabilidad de datos analíticos de alimentos empleando DPLS.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

- Aplicar la metodología probabilística a la clasificación de alimentos, partiendo de datos analíticos multivariantes, y utilizando diferentes estrategias de binarización.

- Desarrollar un método de multiclasificación por combinación de clasificadores binarios probabilísticos, que además permita obtener la fiabilidad de la clasificación multiclase.

- Aplicar la metodología de multiclasificación a la clasificación de alimentos.

- Desarrollar una metodología que permita establecer especificaciones multivariantes de alimentos.

- Aplicar la metodología de especificaciones multivariantes a la autentificación de alimentos.

1.4 Estructura de la tesis

Esta tesis está organizada en 5 capítulos.

El primer capítulo (éste) introduce el proyecto TRACE y el problema de la fiabilidad, núcleos del desarrollo de la tesis, y los objetivos que se persiguieron.

El segundo capítulo contiene los antecedentes de DPLS, el desarrollo del DPLS probabilístico (*p*-DPLS), que integra los conceptos de funciones potenciales y la probabilidad bayesiana, su posterior implementación en un programa y su aplicación a la clasificación de alimentos y suelos.

En el tercer capítulo se afronta el problema de clasificación cuando hay múltiples clases a escoger. Contiene una introducción a los problemas multiclase y la implementación de la metodología *p*-DPLS a la resolución de este problema, con el valor agregado del cálculo de la fiabilidad de clasificación multiclase. La metodología es aplicada a problemas de clasificación de aceites de oliva y suelos.

En el cuarto capítulo se estudian las especificaciones multivariantes derivadas de DPLS como método para definir la autenticidad de un alimento. Se define que es una especificación, sus tipos y como se establecen. Se termina estableciendo especificaciones multivariantes en aceites de oliva y mieles.

El quinto capítulo contiene las conclusiones de la tesis, además de sugerencias para futuras investigaciones utilizando *p*-DPLS; tanto para autentificar alimentos como otros productos y en la resolución de problemas multiclase.

Las últimas páginas contienen anexos con las notaciones y acrónimos, las contribuciones científicas y el análisis de los conjuntos de datos utilizados a lo largo de la tesis; material que consideramos enriquece este trabajo pero que no era apropiado colocar en el desarrollo de la tesis.

Referencias

1. Real Decreto 1334/1999, de 31 de julio, por el que se aprueba la norma general de etiquetado, presentación y publicidad de los productos alimenticios, BOE, 202 (24/8/1999) 31410–31418.
2. Directiva 2000/13/CE del Parlamento Europeo y del Consejo de 20 de marzo de 2000 relativa a la aproximación de las legislaciones de los Estados miembros en materia de etiquetado, presentación y publicidad de los productos alimenticios, Diario Oficial de la Unión Europea, L 109 (6/5/2000) p. 29.
3. Real Decreto 238/2000, de 18 de febrero, por el que se modifica la norma general de etiquetado, presentación y publicidad de los productos alimenticios, aprobada por el Real Decreto 1334/1999, de 31 de julio. BOE, 43 (19/2/2000) 7577–7578.
4. Reglamento (CE) No 510/2006 del Consejo de 20 de marzo de 2006 sobre la protección de las indicaciones geográficas y de las denominaciones de origen de los productos agrícolas y alimenticios. Diario Oficial de la Unión Europea, L93 (31/3/2006) p. 12–25.
5. Reglamento (CE) No 628/2008 de La Comisión de 2 de julio de 2008 que modifica el Reglamento (CE) No 1898/2006, que establece las disposiciones de aplicación del Reglamento (CE) No 510/2006 del Consejo sobre la protección de las indicaciones geográficas y de las denominaciones de origen de los productos agrícolas y alimenticios. Diario Oficial de la Unión Europea, L 173 (3/7/2008) p. 3–5.
6. Real Decreto 1069/2007, de 27 de julio, por el que se regula el procedimiento para la tramitación de las solicitudes de inscripción en el Registro comunitario de las denominaciones de origen protegidas y de las indicaciones geográficas protegidas y la oposición a ellas. BOE, 213 (5/9/2007) 36594–36596.
7. Reglamento (CE) No 178/2002 del Parlamento Europeo y del Consejo de 28 de enero de 2002 por el que se establecen los principios y los requisitos generales de la legislación alimentaria, se crea la Autoridad Europea de Seguridad Alimentaria y se fijan procedimientos relativos a la seguridad alimentaria. Diario Oficial de las Comunidades Europeas, L 31 (1/2/2002) p. 1–24.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

8. M. Boner, H. Förstel, Anal. Bioanal. Chem. 378 (2004) 301–310.
9. G. Manca, F. Camin, G. C. Coloru, A. Del Caro, D. Depentori, M. A. Franco, G. Versini, J. Agric. Food Chem. 49 (2001) 1404–1409.
10. A. Rossmann, G. Haberhauer, S. Hölzl, P. Horn, F. Pichlmayer, S. Voerkelius, Eur. Food Res. Technol. 211 (2000) 32–40.
11. A. Rossmann, Food Rev. Int. 17 (2001) 347–381.
12. Project TRACE "TRAcing food Commodities in Europa" (proyecto No FOOD-CT-2005-006942), www.trace.eu.org.
13. Real Academia Española, Diccionario de la lengua española, http://www.rae.es/rae.html (5/03/2010).
14. B. L. Atwater, D. B. Stauffer, F. W. McLafferty, Anal. Chem. 57 (1985) 899–903.
15. C. Cappelli, F. Mola, R. Siciliano, Comput. Stat. Data Anal. 38 (2002) 285–299.
16. Y. Tan, L. Shi, W. Tong, G.T. Gene Hwang, C. Wang, Comput. Biol. Chem. 28 (2004) 235–244.
17. J. Richiardi, A. Drygajlo, Reliability-Based Voting Schemes Using Modality-Independent Features in Multi-classifier Biometric Authentication. En: M. Haindl, J. Kittler, and F. Roli (Eds.), MCS 2007, LNCS 4472, p. 377–386.
18. B. Lendl, B. Karlberg, Trends Anal. Chem. 24 (2005) 488–492.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Capítulo 2

Clasificación probabilística con mínimos cuadrados parciales discriminantes (DPLS)

2.1 Introducción y revisión bibliográfica

El método de mínimos cuadrados parciales discriminantes (*Discriminant Partial Least Squares*, DPLS) es un método multivariante propuesto por *S. Wold et ál.* [1] como una aplicación de PLS al reconocimiento de patrones. Para desarrollar un modelo DPLS se parte de una matriz de datos **X** (*I*×*J*) que contiene *J* variables medidas en *I* objetos y un vector columna **y** (*I*×1) que contiene la clase a la que pertenece cada objeto indicada con codificación binaria, 0 y 1, donde el 1 indica la pertenencia del objeto a la clase de interés y el 0 su no pertenencia. En el proyecto TRACE la matriz **X** contiene medidas realizadas con técnicas tales como cromatografía de gases (CG), análisis termogravimétrico (*Thermogravimetric analysis*, TGA), espectroscopia infrarroja media (*Mid infrared*, MIR), espectroscopia infrarroja cercana (*Near infrared*, NIR), resonancia magnética nuclear (*Nuclear Magnetic Resonance*, NMR) y análisis elemental con fluorescencia de rayos X (*X-ray fluorescence,* XRF) y espectrometría de masas con fuente de plasma de acoplamiento inductivo (*Inductively Coupled Plasma Mass Spectrometry*, ICP-MS).

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2.1.1 El método DPLS clásico

El modelo DPLS se desarrolla a partir del algoritmo NIPALS/PLS1 [1,2] mostrado en la figura 2-1. En la figura 2-1, **T** (*I*×*A*) es la matriz de s*cores* de **X**; **P** (*J*×*A*) y **q** (1×*A*) son las matrices de *loadings* para **X** e **y** respectivamente; **W** (*J*×*A*) es la matriz de pesos de **X**; **B** (*J*×*A*) es la matriz de coeficientes de regresión y *A* es el número de factores fijado por el usuario.

Para un objeto con variables observadas **x**, el valor de *y* predicho por el modelo PLS es:

$$\hat{y} = \mathbf{x}^{\mathrm{T}}\mathbf{b} \tag{2-1}$$

Figura 2-1: Algoritmo NIPALS para PLS1 en lenguaje Matlab.

```
1.  Xi = X; yi = y;
2.  for a = 1:A
3.      v = yi'*Xi;
4.      w = (1/sqrt(v*v'))*v';        %Peso parcial de las variables
5.      t = Xi*w;                     %Scores parciales de X e y
6.      p = ((t'*Xi)/(t'*t))';        %Loadings parciales de X
7.      q = (yi'*t)/(t'*t);           %Loading parcial de y,
                                      %otra alternativa es
                                      %q = (yi'*t)/norm(yi'*t)

8.      np = norm(p);                 %Normalización
9.      q  = q/np;
10.     t  = t*np;
11.     w  = w*np;
12.     p  = p/np;

13.     T(:,a) = t;                   %Matrices
14.     W(:,a) = w;
15.     P(:,a) = p;
16.     q(:,a) = q;                   %Vector de loadings de y
17.     B(:,a) = W(:,1:a)*(inv(P(:,1:a)'*W(:,1:a)))*q(1,1:a)';
                                      %Calculo de los coeficientes B

18.     Xi = Xi-t*p';                 % Residual X
19.     yi = yi-t*q;                  % Residual y
20. end
```

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

donde **b** se ha calculado para un número determinado de factores A. El número óptimo A_{opt} se estima comúnmente por validación cruzada dejando fuera un objeto cada vez (*leave-one-out cross-validation*, LOOCV) o con un conjunto de prueba.

El DPLS-PLS1 es un método de clasificación binaria, es decir, que discrimina entre dos clases ($C = 2$). Aunque la codificación binaria [0,1] es la más extendida, también se ha utilizado [−1,+1] [3,4]. Cuando el problema de clasificación implica más de dos clases ($C > 2$), se puede convertir en varios problemas de dos clases (modelos binarios) cuyas respuestas se combinan para resolver el problema de C clases. Otra alternativa es utilizar el algoritmo PLS2 [5]. Éste utiliza una matriz **Y** ($I \times C$) en donde cada columna representa una clase codificando su pertenencia como 1 y su no pertenencia como 0. Los problemas multiclase se tratan en el capítulo 3.

El valor $\hat{y}$ en la ecuación 2-1 no es un número binario, 0 o 1, sino un número real cercano a éstos, siendo necesario encontrar una función de decisión que permita convertir ese valor predicho en una clase. Esta función de decisión se basa en pruebas de hipótesis [6,7]. Una de las funciones de decisión más usadas [5,6,8–10] establece un límite de decisión en el valor 0.5. Si el objeto presenta un valor $\hat{y} < 0.5$ se asigna a la clase 0 y si presenta un valor $\hat{y} > 0.5$ se asigna a la clase 1. Para eliminar la arbitrariedad del límite, éste se puede optimizar para maximizar la habilidad de clasificación [6,7]. El límite también se puede definir implícitamente si se utiliza la teoría bayesiana [11,12]. En ese caso se calcula la probabilidad *a posteriori* de $\hat{y}$ para cada clase y se asigna el objeto a la clase donde la probabilidad es mayor. En ese caso el límite de decisión es el punto donde las probabilidades *a posteriori* de pertenecer a cada clase son iguales. Si la predicción es igual al límite de decisión, se puede asignar el objeto donde su probabilidad *a priori* sea mayor. La decisión bayesiana elimina la subjetividad en la selección del límite de decisión, puesto que el objeto es asignado a una clase dependiendo sólo de la $\hat{y}$ del objeto comparada con la distribución de los valores predichos de los objetos de entrenamiento de cada clase. Es habitual considerar que los valores predichos para cada clase se distribuyen normalmente, de modo que dicha distribución se puede caracterizar fácilmente con la media y la desviación estándar de los valores predichos para los objetos de cada clase [13]. No obstante, esto no es siempre cierto, sobretodo si las clases tienen un número reducido de objetos o tienen un número muy distinto de objetos, siendo necesario desarrollar métodos más flexibles para describir la distribución de las predicciones y fijar el límite de decisión.

Otra aproximación es asignar un intervalo de predicciones, por ejemplo [0.4, 0.6], en el cual todo objeto que caiga en ese intervalo se considera dudoso y no es asignado a ninguna clase [14,15]. Con esto se pretende asignar sólo aquellos objetos cuya clasificación es más fiable, es decir, que se cuenta con la plena seguridad que

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

pertenecen a las clases, ya que se encuentran en el espacio de variables de la clase y no en los límites de éstas.

Además es importante resaltar que cualquier asignación de un objeto a una clase debe ir acompañada de una estimación de la fiabilidad de la clasificación. Intuitivamente, la asignación de un objeto cercano al límite de decisión será menos fiable que la asignación de un objeto que esté claramente en el centro de la distribución de una clase.

2.1.2 Aplicaciones de DPLS

DPLS se ha aplicado en numerosos campos. En medicina para determinar diferentes estados de cáncer oral mediante datos de fluorescencia [12]; o para identificar el tipo de lesión en laringe mediante datos de fluorescencia en vivo [11]; en psiquiatría para clasificar pacientes en diferentes estados de organización de personalidad [16]; o en medicina forense para discriminar entre suicidios y homicidios [17]. En el sector farmacéutico se pueden citar ejemplos como la identificación de comprimidos por espectroscopia NIR [18], la clasificación de compuestos con potencial farmacológico para el tratamiento del sistema nervioso central [19] o para establecer la estabilidad metabólica de potenciales medicamentos ante el citocromo humano [20]. En análisis bioquímicos se ha aplicado en la detección de sitios de fosforilación en péptidos mediante espectroscopia Raman [21] y en el análisis de extractos de hongos por fluorescencia para determinar la especie a la que pertenecen [6]. En genética se ha aplicado para clasificar clases de tumores y la etapa en que se encuentran a partir de datos de *microarrays* de genes [22–24]. Además, el DPLS se ha utilizado en la selección de variables significativas a partir de los coeficientes de regresión, siendo aplicado a la determinación del origen de suelos analizando ADN de la fauna microbiana [25] o los ácidos grasos presentes en los microorganismos [26].

En química analítica, y especialmente en el análisis de alimentos, DPLS se ha aplicado al análisis de aguas para establecer la calidad y diferencia entre aguas superficiales o subterráneas [27] o determinar el origen geográfico de las superficiales [28]. También en la diferenciación de variedades vegetales y animales como variedades de uva blanca por diferentes métodos de análisis sensorial y espectrométrico [29], variedades de jugos de manzana y si éstos son sometidos a tratamiento calórico, partiendo de datos de cromatografía de gases [7]. También se ha aplicado a la identificación de carnes de pescado de diferentes especies [30], identificación de carnes de diferentes especies animales [31], o variedades de whiskeys a partir de la calificación de características

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

dadas por panel de catadores [32,33]. En seguridad alimentaria se ha utilizado en la detección de productos adulterados, como jugos de manzana por adición de azúcares utilizando espectroscopia NIR [34], o aceites de oliva a partir de datos de espectrometría de masas [35].

En el control del cumplimento de normas para consumo humano, DPLS se ha utilizado para la detección de leche de vacas enfermas utilizando narices electrónicas [10], la determinación de la frescura de pescados según el tiempo y temperaturas de almacenamiento en frío [36]; o para establecer la presencia de pesticidas basados en carbamatos mediante CG acoplada a espectrometría masas [8]. También se ha utilizado para diferenciar calidades en los alimentos, como mayonesas con diferentes contenidos de aceites vegetales utilizando espectroscopia NIR [37]; identificar tipos de leche, según especie animal, a partir de datos de espectrometría de masas [35]; o clasificar olivas a partir de imágenes de los frutos [38]. Recientemente, DPLS se ha utilizado en la verificación del origen geográfico de productos como vinos blancos mediante espectroscopia visible y NIR [39] o de raíces de *ginseng* utilizando espectroscopia NIR y Raman [40].

2.1.3 Incertidumbre de la predicción en PLS

La incertidumbre o intervalo de incertidumbre es un concepto comúnmente asociado a la medida de propiedades físicas y químicas. Este concepto integra los errores aleatorios y sistemáticos para encontrar un intervalo donde hay una probabilidad fijada de encontrar el valor del mensurando [41]. Esta incertidumbre debe ser tomada en cuenta cuando se utilizan métodos de regresión.

La incertidumbre en la predicción de un modelo PLS se ve afectada por la incertidumbre en **X**, en **y** y por el error en la función matemática ajustada [42]. El programa "*The Unscrambler*" [43] calcula la desviación de $\hat{y}$ en modelos PLS, a partir de las varianzas de los residuales de predicción, de validación y el *leverage*. *De Vries* y *Ter Braak* [44], criticaron que el método subestima el error cuadrado medio de predicción (*Mean Square Error Prediction, **MSEP***) cuando se toma como el cuadrado de la desviación. *De Vries* y *Ter Braak* propusieron una mejora a dicho cálculo ya que el ***MSEP*** es el más apropiado para evaluar la eficiencia del modelo PLS al variar el número de factores [44,45], lo cual puede llevar a errores de optimización. *Faber* y *Kowalski* [46] criticaron tanto la expresión inicial en "*The Unscrambler*" como la corregida ya que ignoran los errores de las medidas, siendo éstos críticos en el caso de variables químicas, restringiendo el uso de estas expresiones en PLS.

Otras estrategias han planteado modelos generales para determinar la incertidumbre en regresión con PLS. *Lorber* y *Kowalski* [47] plantean una ecuación de predicción, que incluye los errores para cada uno de los términos. No obstante, calcular estos errores no es fácil ya que en muchas ocasiones no es posible establecer con exactitud los errores aleatorios y sistemáticos. *Karstang et ál.* [48], modificaron la propuesta de *Lorber* y *Kowalski* para incluir los residuales de las variables independientes; los errores de la función matemática ajustada y el error de predicción dentro del espacio de predicción, utilizando la información sobre la posición del objeto. *Faber* y *Kowalski* [46,49] propusieron una aproximación para el cálculo de la varianza de predicción en mínimos cuadrados ordinarios (*Ordinary Least Square*s, OLS), regresión por componentes principales (*Principal Component Regression,* PCR) y regresión PLS que tiene en cuenta los errores de medida en las variables independientes y dependientes. *Morsing* y *Ekman* [50] consideraron poco adecuada la propuesta de *Faber* y *Kowalski* para el cálculo de las regiones de confianza [49] porque no tiene en cuenta el sesgo; pero al estudiar detenidamente la propuesta de *Faber* y *Kowalski* se observa que sí se calcula el sesgo para la variable dependiente [51]. *Faber* y *Bro* extendieron el trabajo de *Faber* y *Kowalski* a *multiway* PLS (N-PLS), es decir, calibración con datos de segundo orden [42]. En esta extensión, *Faber* y *Bro* establecen una ecuación general para el error estándar de predicción:

$$\sigma_{PE} \equiv \left(V_{PE}\right)^{1/2} \approx \left(\mathbf{x}^{\mathrm{T}}\mathbf{V}_{\Delta\beta}\mathbf{x} + \beta^{\mathrm{T}}\mathbf{V}_{\Delta x}\beta + V_e\right)^{1/2} \tag{2-2}$$

donde σ_{PE} es el error estándar de predicción; V_{PE} es la varianza del error de predicción; $\mathbf{x}$ es la respuesta instrumental verdadera, β es el coeficiente de regresión verdadero; $\mathbf{V}_{\Delta\beta} = E[\Delta\beta\Delta\beta^{\mathrm{T}}]$ es la matriz de covarianza de los errores en los coeficientes de regresión estimados; $\mathbf{V}_{\Delta x} = E[\Delta\mathbf{x}\Delta\mathbf{x}^{\mathrm{T}}]$ es la matriz de covarianza de los errores de medida de la respuesta instrumental; y $V_e = E[e^2]$ es la varianza de los residuales. La ecuación 2-2 contiene las dos posibles contribuciones para el error de predicción. Así $\mathbf{x}^{\mathrm{T}}\mathbf{V}_{\Delta\beta}\mathbf{x}$ corresponde a la varianza en la calibración y $\beta^{\mathrm{T}}\mathbf{V}_{\Delta x}\beta+V_e$, es la varianza para la predicción del objeto desconocido. Ahora bien, si se ignoran los términos más complejos y se asume que los errores son independientes e idénticamente distribuidos (IID) se puede derivar una ecuación específica para predictores con errores homocedásticos:

$$\sigma_{PE} \approx \left[h\left(\|\beta\|^2 V_{\Delta x} + V_e + V_{\Delta y}\right) + \|\beta\|^2 V_{\Delta x} + V_e\right]^{1/2} \tag{2-3}$$

donde h es el *leverage* para el objeto desconocido con respecto al origen, $\|\cdot\|$ denota la norma euclidiana, $V_{\Delta y} = E[\Delta y^2]$ es la varianza del error de medida en el método de referencia, y $V_{\Delta x}$ es la IID simplificación de $\mathbf{V}_{\Delta x}$. El *leverage* se calcula como

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

$h_i = \mathbf{t}_i^{\mathrm{T}} \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1} \mathbf{t}_i$, donde $\mathbf{t}_i$ es el vector de *scores* para el objeto *i* con *A* factores y **T** es la matriz de *scores* de los datos de entrenamiento con *A* factores. Si asumimos que los predictores tienen un error heterocedástico, y que en ausencia de sesgo el error cuadrado medio de calibración (*Mean Square Error Calibration*, *MSEC*) es $||\beta||^2 V_{\Delta x} + V_e + V_{\Delta y}$, podemos reescribir la ecuación 2-3 como:

$$\sigma_{PE} \approx \left[(1+h)\times MSEC - V_{\Delta y}\right]^{1/2} \tag{2-4}$$

donde el *MSEC* se calcula como:

$$MSEC = \frac{\sum_{i=1}^{I}(\hat{y}_i - y_i)^2}{I - A} \tag{2-5}$$

siendo $\hat{y}$ el valor predicho e y el valor verdadero. Cuando el sesgo es significativo el *MSEC* contiene la varianza de calibración para el vector **y**, $Var[\hat{\mathbf{y}}\text{-}\mathbf{y}]$, y el sesgo de calibración, *bias*:

$$MSEC = bias^2 + Var\left[\hat{\mathbf{y}} - \mathbf{y}\right] \tag{2-6}$$

Además para datos centrados se debe sumar el termino $1/I$ al *leverage,* h (Ec. 2-4) [42] y el denominador de la Ec. 2-5 cambia por $I-A-1$. En DPLS, la variable y es un vector de 0's y 1's, y por lo tanto $V_{\Delta y} = 0$, y si tenemos en cuenta que σ_{PE} es equivalente al error estándar de predicción (*Standard Error of Prediction*, *SEP*), la Ec. 2-4 se puede reescribir como:

$$SEP = \left[(1+h)\times MSEC\right]^{1/2} \tag{2-7}$$

Otras técnicas para calcular el error estándar de predicción tienen en cuenta la propagación de la incertidumbre, como la de *Kleinknecht* que calcula el error de los *scores*, *loadings* y sus posibles combinaciones, utilizando para ello la matriz de covarianza de los pesos de la matriz de datos **X** [52]. Otra es la de *Chen et ál.* que, utilizando ensamblaje de modelos PLS, hallan la variables que más incertidumbre transmiten al modelo, a partir de la incertidumbre en los coeficientes de regresión [53]. O la de *Griffiths* y *Ellison* que utilizan un método numérico para la propagación del valor de referencia de la incertidumbre [54]. Sin embargo, la ecuación 2-7 es la que nos será de utilidad, y se utilizará para establecer la desviación estándar de las funciones potenciales.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2.1.4 Funciones potenciales

Las funciones potenciales forman parte de los llamados métodos de densidad o *kernel*, que construyen una función alrededor de un punto. Se puede utilizar cualquier función potencial. Las más usadas son el triángulo y la función gaussiana [55]. Al final una clase se describe como una función acumulativa, suma de diferentes funciones potenciales centradas en cada uno de los objetos de la clase. Un valor crítico en estas funciones, que debe ser optimizado, es el *smoothing* o amplitud de la función; por ejemplo, la desviación estándar en las gaussianas. Valores muy pequeños no permiten que se solapen las funciones que pertenecen a una misma clase, y valores muy altos aplanan demasiado la función acumulativa, solapando las clases [2]. El utilizar funciones gaussianas tiene una ventaja adicional, ya que el promedio de las funciones potenciales acumulativas es 1; permitiendo calcular valores de probabilidad que pueden usarse en la clasificación. La forma de la función acumulativa es [56]:

$$f\left(\hat{y}_{\mathrm{u}} \mid \omega_c\right) = \frac{1}{I_c} \sum_{i=1}^{I_c} g\left(\hat{y}_{\mathrm{u}}, \hat{y}_i^{\,c}\right) \tag{2-8}$$

donde $f\,(\hat{y}_{\mathrm{u}}|\omega_c)$, es la función acumulativa de densidad de probabilidad para toda $\hat{y}_{\mathrm{u}}$ que pueda pertenecer a la clase ω_c. I_c es el número de objetos de la clase ω_c. $g(\hat{y}_{\mathrm{u}}, \hat{y}_i^{\,c})$ es la función individual para cada objeto $\hat{y}_i^{\,c}$ evaluado en el punto $\hat{y}_{\mathrm{u}}$. $\hat{y}_i^{\,c}$ hace referencia a todos los objetos i de la clase ω_c y $\hat{y}_{\mathrm{u}}$ a cualquier objeto nuevo que pueda ser evaluado, pertenezca o no a la clase. En este trabajo, $\hat{y}_{\mathrm{u}}$ será la predicción $\hat{y}$ proporcionada por el modelo PLS.

2.1.5 Probabilidad bayesiana

La probabilidad bayesiana se basa en la inferencia lógica de sucesos [57]. Así, la probabilidad de un suceso E está condicionado a una proposición D. En otras palabras, $p(E|D)$ es medida de la plausibilidad de una proposición o hipótesis E, condicionada a la veracidad de la proposición D. La probabilidad bayesiana permite obtener probabilidades posteriores de sucesos desconocidos y cuantificar la incertidumbre de inferencias basadas en el análisis estadístico de los datos [58].

En clasificación se parte de un número de objetos I que son representativos de una población y resumen una serie de características que pueden ser medidas. Por ejemplo,

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

se cuenta con un conjunto de objetos que pueden pertenecer a dos clases, ω_0 y ω_1. A cada clase se le puede asociar una Función de Densidad de Probabilidad (FDP) condicionada $p(\hat{y}|\omega_c)$ (Figura 2-2), es decir, la probabilidad que tiene un objeto con valor $\hat{y}_i$ si perteneciera a ω_c y una probabilidad de clase $P(\omega_c)$ o probabilidad *a priori* de que el nuevo objeto sea de la clase ω_c. La probabilidad de que un objeto que pertenece a la clase ω_c tenga un valor menor o igual a $\hat{y}_i$, es la integral de la FDP desde menos infinito hasta $\hat{y}_i$.

$$P(\hat{y}_i \mid \omega_c) = \int_{-\infty}^{\hat{y}_i} p(\hat{y} \mid \omega_c)\, d\hat{y} \tag{2-9}$$

La probabilidad de pertenecer a la clase ω_c dado un valor observado $\hat{y}_i$ está dada por el teorema de Bayes.

$$P(\omega_c \mid \hat{y}_i) = \frac{p(\hat{y}_i \mid \omega_c) \times P(\omega_c)}{\sum_{c=1}^{C} p(\hat{y}_i \mid \omega_c) \times P(\omega_c)} \tag{2-10}$$

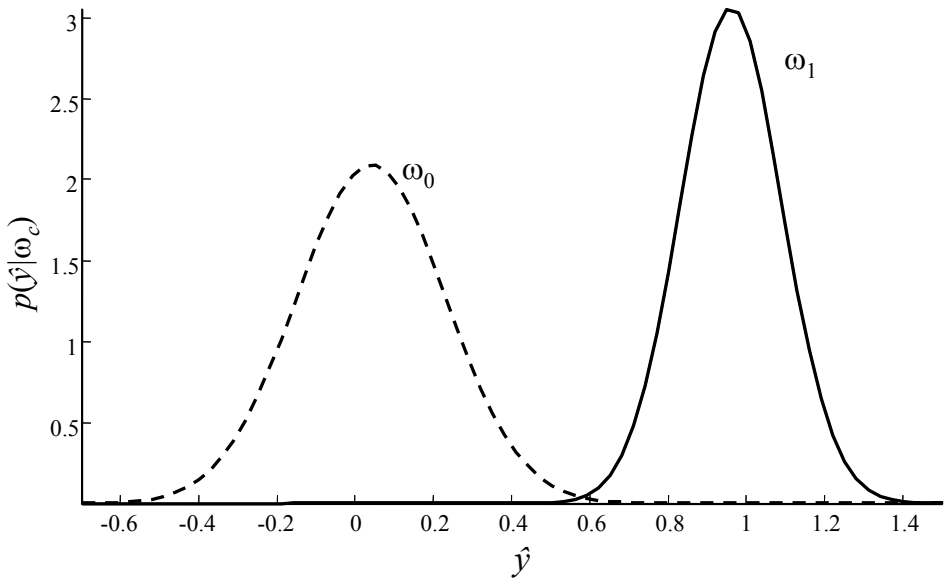

Figura 2-2: Funciones de densidad de probabilidad para las clases ω_1 y ω_2. El límite entre clases es un valor $\hat{y}$ donde $p(\hat{y}|\omega_1) = p(\hat{y}|\omega_2)$.

La decisión bayesiana busca asignar los objetos desconocidos a la clase donde se produzca el menor error posible, es decir, donde sea más probable que pertenezcan. Así, la regla de asignación de un elemento nuevo $\hat{y}_u$ a una clase ω_c es:

$$\hat{y}_u \text{ pertenece a } \omega_1 \text{ si } P(\omega_1|\hat{y}_u) > P(\omega_0|\hat{y}_u); \text{ y viceversa para la otra clase.} \qquad (2\text{-}11)$$

En el próximo apartado (2.2) se muestra como se integró el error de predicción en PLS, las funciones potenciales y el teorema de Bayes para desarrollar el algoritmo DPLS probabilístico (*p*-DPLS). Éste desarrollo se documentó en el artículo "*Calculation of the reliability of classification in discriminant partial least-squares binary classification*", *Chemometrics and Intelligent Laboratory Systems*, 95 (2009) 122–128.

Referencias

1. B. R. Kowalski "Chemometrics. Mathematics and statistics in Chemistry". D. Reidel Publishing Company. Dordrecht, Holland. 1984. p. 77–95.
2. B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. De Jong. P. J. Lewi, J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier Science B.V. AE Amsterdam, The Netherlands, 1998, p. 331–340.
3. H. Yoshida, R. Leardi, K. Funatsu, K. Varmuza, Anal. Chim. Acta. 446 (2001) 485–494.
4. L. Evans III, G. E. Collins, R. E. Shaffer, V. Michelet, J. D. Winkler, Anal. Chem. 71 (1999) 5322–5327.
5. Y. Roggo, L. Duponchel, J.-P. Huvenne, Anal. Chim. Acta. 477 (2003) 187–200.
6. C. Wittrup, J. Chemometr. 14 (2000) 765–776.
7. L. M. Reid, C. P. O'Donnell, J. D. Kelly, J. Agric. Food Chem. 52 (2004) 6891–6896.
8. M. Decker, P. V. Nielsen, H. Martens, Appl. Spectros. 59 (2005) 56–68.
9. C. Wan, P. de B. Harrington, Anal. Chim. Acta. 408 (2000) 1–12.
10. A. Eriksson, K. P. Waller, K. Svennersten-Sjaunja, Int. Dairy J. 15 (2005) 1193–1201.
11. C. Eker, R. Rydell, K. Svanberg, Laser Surg. Med. 28 (2001) 259–266.
12. C. Wang, C. Chen, C. Chlang, Photochem. Photobiol. 69 (1999) 471–477.
13. B. M. Wise N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig, R. S. Koch, "PLS Toolbox Version 3.5 for use with MATLAB™", Eigenvector Research, Inc., Manson, WA, USA, 2005, p. 185–189.
14. H. Mauser, O. Roche, M. Stahl, J. Chem. Inform. Model. 45 (2005) 1039–1046

15. L. Afzelius, C. M. Masimirembwa, A. Karlen, J. Comput. Aided Mol. Des. 16 (2002) 443–458.
16. P. Fransson, E. Sundbon, Scand. J. Psychol. 38 (1997) 95–102.
17. T. Karlsson, Forensic Sci. Int. 94 (1998) 183–200.
18. R. De Maesschalck, T. Van den Kerkhof, J. Pharmaceut. Biomed. Anal. 37 (2005) 109–114.
19. M. Adenot, R. Lana, J. Chem. Inform. Comput. Sci. 44 (2004) 239–248.
20. P. Crivori, I. Zamora, B. Speed, J. Comput. Aided Mol. Des. 18 (2004) 155–166.
21. D. Zhang, C. Ortiz, Y. Xie, Spectrochim. Acta Mol. Biomol. Spectros. 61 (2005) 471–475.
22. J-H. Cho, D. Lee, J. H. Park, K. Kim, I-B. Lee, Biotechnol. Progr. 18 (2002) 847–854.
23. C. Yoo, I. Lee, P. A. Vanrolleghem, Comput. Chem. Eng. 29 (2005) 1345–1356.
24. Y. Tan, L. Shi, W. Tong, Comput. Biol. Chem. 28 (2004) 235–244.
25. Z. Ramadan, X. Song, P. K. Hopke, Anal. Chim. Acta. 446 (2001) 233–244.
26. X. Song, P. K. Hopke, M. A. Bruns, Environ. Sci. Technol. 33 (1999) 3524–3530.
27. K. P. Singh, A. Malik, V. K. Singh, Anal. Chim. Acta. 550 (2005) 82–91.
28. G. J. Hall, K. Clow, J. Kenny, Environ. Sci. Tech. 39 (2005) 7560–7567.
29. S. Roussel, V. Bellon-Maurel, J-M. Roger, J. Food Eng. 60 (2003) 407–419.
30. D. Cozzolino, A. Chree, J. R. Scaife, J. Agric. Food Chem. 53 (2005) 4459–4463.
31. G. Downey, J. Mcelhinney, T. Fearn, Appl. Spectros. 54 (2000) 894–899.
32. K-Y. M. Lee, A. Paterson, J. R. Piggott, Food Quality and Preference. 12 (2001) 109–117.
33. D. Gonzalez-Arjona, G. Lopez-Perez, A. G. González, Talanta. 49 (1999) 189–197.
34. L. Leon, J. D. Nelly, G. Downey, Appl. Spectros. 59 (2005) 593–599.
35. B. K. Alsberg, R. Goodacre, J. J. Rowland, Anal. Chim. Acta. 348 (1997) 389–407.
36. N. Bøknæs, K. N. Jensen, C. M. Andersen, Lebensm.-Wiss. u.-Technol. 35 (2002) 628–634.
37. U. G. Indahl, N. S. Sahni, B. Kirkhus, Chemometr. Intell. Lab. Syst. 49 (1999) 19–31.
38. R. Diaz, L. Gil, C. Serrano, J. Food Eng. 61 (2004) 101–107.
39. D. Cozzolino, H. E. Smyth, M. Gishen, J. Agric. Food Chem. 51 (2003) 7703–7708.
40. Y-A. Woo, H-J. Kim, H. Chung, Analyst. 124 (1999) 1223–1226.
41. A. Maroto Sánchez, Incertidumbre en Métodos Analíticos de Rutina, Universitat Rovira i Virgili, Tarragona, España, 2002, p. 32–36.
42. N. M. Faber, R. Bro, Chemometr. Intell. Lab. Syst. 61 (2002) 133–149.
43. The Unscrambler User Manual, CAMO ASA, Oslo, Norway, 1998, p. 436.
44. S. De Vries, C. J. F. Ter Braak, Chemometr. Intell. Lab. Syst. 30 (1995) 239–245.

45. A. Höskuldsson, J. Chemometr. 2 (1988) 211–228.
46. K. Faber, B. R. Kowalski, Chemometr. Intell. Lab. Syst. 34 (1996) 283–292.
47. A. Lorber, B. R. Kowalski, J. Chemometr. 2 (1988) 93–109.
48. T. V. Karstang, J. Toft, O. M. Kvalheim, J. Chemometr. 6 (1992) 177–188.
49. K. Faber, B. R. Kowalski, J. Chemometr. 11 (1997) 181–238.
50. T. Morsing, C. Ekman, J. Chemometr. 12 (1998) 195–299.
51. N. M. Faber, J. Chemometr. 14 (2000) 363–369.
52. R. E. Kleinknecht, J. Chemometr. 10 (1996) 687–695.
53. D. Chen, W. Cai, X. Shao, Anal. Chim. Acta. 598 (2007) 19–26.
54. M. L. Griffiths, S. L. R Ellison, Chemometr. Intell. Lab. Syst. 83 (2006) 133–138
55. D. L. Massart, L. Kaufman, "The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis". John Wiley & Sons Inc. USA. 1983. p. 115–118.
56. D. Coomans, D. L. Massart, Anal. Chim. Acta. 133 (1981) 215–224.
57. P. Gregory, "Bayesian Logical Data Analysis for the Physical Sciences", Cambridge University Press, UK, 2005, p. 2–4.
58. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Rubin, Bayesian Data Analysis, Second Edition, Chapman & Hai/CRC, USA, 2004, p. 1–9.

2.2 Calculation of the reliability of classification in discriminant partial least-squares binary classification

Chemometrics and Intelligent Laboratory Systems
95 (2009) 122–128

Néstor F. Pérez, Joan Ferré, Ricard Boqué

Department of Analytical Chemistry and Organic Chemistry, Rovira and Virgili University, C/ Marcel·lí Domingo, s/n. 43007, Tarragona, Spain

A classification decision must include the degree of confidence in that decision. We have modified the binary classification method Discriminant Partial Least Squares (DPLS) to provide the reliability of the classification of an unknown object. This method, called Probabilistic Discriminant Partial Least Squares (*p*-DPLS), integrates DPLS, density methods and Bayes decision theory in order to take into account the uncertainty of the predictions in DPLS. The reliability of classification is also used to derive a new classification rule, so that an unknown object is classified in the class for which it has the highest reliability. This new methodology is tested with two data sets, the benchmark Iris data set and an Italian olive oil data set. The results show that the proposed method is comparable with other methodologies, with percentages of correct classification higher than 95%, with the advantage of providing a measurement of the reliability of classification that agrees with the distribution of the samples in the training set.

2.2.1 Introduction

Nowadays, consumers' demands on food safety can be translated into two main aspects: the consumers want to be sure that foodstuffs meet the expected healthy characteristics and that there is no alteration or falsification related to their origin. These aspects have a particular importance in organic foods [1]. For this reason, some standards have been established in order to ensure the security of consumers and producers. The European Union, through the Commission of Agriculture and Rural Development, has implemented three mechanisms of protection: Protected Designation of Origin (PDO), Protected Geographical Indication (PGI) and

Traditional Specific Guaranteed (also known as Certificate of Specific Character) [2, 3]. Related to this, the European project "Tracing the Origin of Food (TRACE)" is currently being developed within the Sixth Framework Program of the European Union [4]. Two of the main objectives of TRACE are to establish mechanisms to obtain a "fingerprint" of each food commodity (e.g. by means of multi-isotope or spectral analysis [2]) and to assure the traceability from the production site to the final consumer. A key point of the traceability is to be able to describe a correspondence between the food commodity and its geographical origin [5]. For this reason, numerous chemical/biological analyses of food and soils have been conducted for products such as meat, cereals, honey, olive oil and mineral water, thus providing databases for food and geographic areas. The large amount of data obtained for each sample analyzed makes it necessary to use methods of multivariate data analysis. Particularly, the food is classified in its geographical origin (class) using multivariate classification. However, classification methods are not only expected to assign a sample to a class of origin, but also to provide a measure of the security of such classification. The classification algorithm used in this work is Discriminant Partial Least Squares (DPLS) [6] that is based on Partial Least Squares Regression (PLSR) [7]. DPLS has already been used in the quality control of food commodities, for example to compare the time of storage of a food product [8, 9]. It has also been used to check the fulfilment of the norms for consumption [10] (e.g. detection of adulterants in natural products [11]) and to determine food authenticity, e.g. comparison of different types of meat to verify the animal species [12, 13], and the discrimination between varieties and geographical regions of origin of different commodities [14-17].

In DPLS, a PLS regression model is calculated that relates the independent variables (e.g. spectra) to an integer y that designates the class of the sample. For example, the number one (1) is used to indicate that the training sample belongs to the class of interest, and a zero (0) indicates that the sample belongs to a different class. Classification of an unknown sample is derived from the value predicted by the PLS model, $\hat{y}$. This value is a real number, not an integer, which should be ideally close to the values used to codify the class (here either 0 or 1). A cut-off value between 0 and 1 is established so that a sample is assigned to class ω_1 if the prediction is larger than the cut-off value, or assigned to class ω_0 otherwise. The simplest approach is to use an arbitrary cut-off value, such as 0.5 [10, 12]. A more advanced approach is to assume that the predictions for each class in the training set follow a Gaussian distribution. Then, the mean and the standard deviation of these predictions are used to estimate a probability density function (PDF) for each class. By combining the PDFs and the Bayes Theorem [18] the cut-off is defined as the value of $\hat{y}$ at which the *a posteriori* probability of both classes is equal [19]. This method improves the arbitrary selection of the cut-off value, which now depends on the distribution of the predictions and thus can be different from 0.5. The mentioned approaches, however, still ignore that

the PLS predictions have uncertainty and ignore that the PDFs of the predictions can be non-Gaussian. Moreover, they do not provide a measure of the reliability of the classification, which must be particular for each sample. Mauser et al. [20] and Afzelius et al. [21] proposed to establish an interval between classes ω_0 and ω_1, e.g. [0.4–0.6], so that the classification of a sample that falls in that interval is unclear. An approach for calculating the reliability in logistic regression is mentioned by Eker et al. [22], and Wang et al. [23], who used PLS scores and the Bayesian decision theory to calculate *a posteriori* probabilities and assigned a sample to the class where the *a posteriori* probability is the largest. Such a theory quantifies the tradeoffs of the classification decision, considering the probabilities and the costs of that decision [24]. However, they do not take into account the uncertainty of the PLS predictions.

In this paper we present a new classification methodology for DPLS which has the following advantages: (1) it does not assume a Gaussian distribution for the DPLS predictions of a class, (2) it takes into account the uncertainty of the predictions in PLS and (3) it provides a measure of the reliability of the classification. The new approach is illustrated with two datasets: the benchmark Iris data set, and a set of Italian olive oil samples that must be classified according to their region of origin.

2.2.2 Theoretical background

The proposed methodology for classification using DPLS is as follows:

2.2.2.1 *Calculation of the PLS model*

A training set is available which consists of a matrix $\mathbf{X}$ ($I \times J$) of J variables measured in I samples, and a dependent variable vector $\mathbf{y}$ ($I \times 1$) in which the class of each sample is coded with the integer 1 if the sample belongs to the class of interest (class ω_1) or 0 otherwise (class ω_0). Of these I samples, I_0 samples belong to class ω_0 and I_1 samples belong to class ω_1. Note that *p*-DPLS performs classification between $C = 2$ classes only (either class ω_0 or class ω_1). When the training samples belong to $C>2$ classes, one of the classes is defined as the class of interest (class ω_1) and the remaining C–1 classes are included together in class ω_0. A PLS model is calculated with this training set, using the NIPALS algorithm for PLS1 [7].

2.2.2.2 *Prediction and prediction uncertainty in PLS*

For each calibration sample i, the fitted value, $\hat{y}_i$, and its standard error of prediction, SEP_i, are calculated for a number of factors A. The prediction for sample i is:

$$\hat{y}_i = \mathbf{x}_i^{\mathrm{T}}\mathbf{b} \tag{1}$$

where $\mathbf{x}_i$ is the column vector of J variables measured for that sample and $\mathbf{b}$ is the vector of regression coefficients of the PLS model calculated with A factors. The standard error of prediction accounts for the experimental error and the fitting error of the model. It is given by [25]:

$$SEP_i = \left[(1 + h_i) \times MSEC_{bc}\right]^{1/2} \tag{2}$$

where h_i is the leverage for sample i and $MSEC_{bc}$ is the bias corrected mean squared error of calibration [26]. The leverage is computed as $h_i = \mathbf{t}_i^{\mathrm{T}}\left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{t}_i$, where $\mathbf{t}_i$ is the scores vector for sample i on A factors and $\mathbf{T}$ is the scores matrix of the training data for the A factors. The bias-corrected $MSEC$ is calculated as [27]:

$$MSEC_{bc} = \frac{\sum_{i=1}^{I}(\hat{y}_i - y_i - bias_c)^2}{I - A - 1} \tag{3}$$

where $bias_c$ is the bias corresponding to the prediction of the samples of class c, ($c = 1,\ldots, C$):

$$bias_c = \frac{\sum_{i=1}^{I_c}(\hat{y}_i - y_i)}{I_c} \tag{4}$$

Note that the bias is estimated independently for each of the C original classes of the training set (the summatory in Eq. (4) runs to I_c, which is the number of samples of class c). This method for calculating the bias is advantageous when $C>2$ classes are present in the training set, and the samples of $C-1$ classes are grouped as class ω_0. The predictions of the samples of the same class in ω_0 tend to have a characteristic bias that is different from the bias in the predictions of the samples of another class in ω_0. In

case that there are only C=2 classes, I_c becomes I_0 and I_1, the number of samples in class ω_0 and in class ω_1 respectively.

For a mean-centered model, the mean of the **y** of the calibration samples ($\bar{y}$) must be added to $\mathbf{x}_i^T\mathbf{b}$ in the Eq. (1), the term $1/I$ must be added to h_i [25] and the denominator of Eq. (3) is $I-A-2$.

2.2.2.3 *Calculation of potential functions over each PLS prediction*

Density methods [28, 29] or Kernel methods are classification methods that are based on calculating a potential function around a point. In the proposed approach, a potential function is calculated for each calibration sample i, with the shape of a Gaussian curve, which is a commonly used kernel function [28], centered at $\hat{y}_i$ and with standard deviation (smoothing parameter) equal to SEP_i:

$$g_i(\hat{y}) = \frac{1}{SEP_i\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\hat{y}-\hat{y}_i}{SEP_i}\right)^2} \quad (5)$$

where $\hat{y}$ is the variable in the abscissa axis (Fig. 1a).

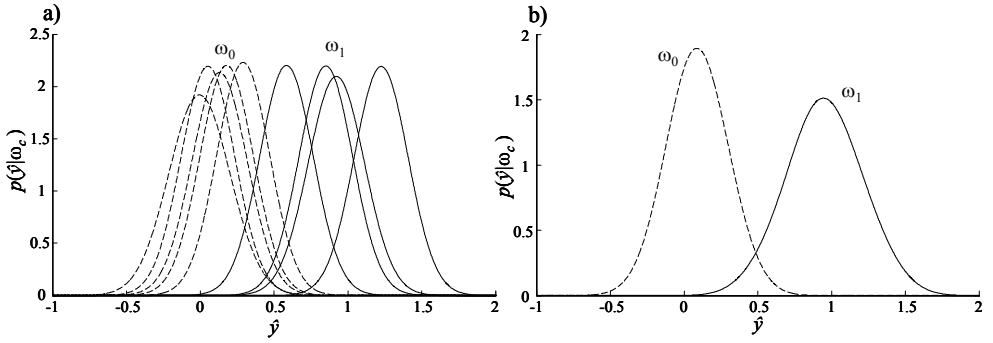

Fig. 1: Calculation of the probability density functions for classes ω_0 (– – –) and ω_1 (—). a) Individual potential functions centred at $\hat{y}_i$. b) Probability density functions calculated as the average of the functions in a).

2.2.2.4 *Calculation of the probability density function of each class*

The potential functions of the I_0 training samples are averaged to obtain the PDF of class ω_0 (Fig. 1b):

$$p(\hat{y} \mid \omega_0) = \frac{1}{I_0} \sum_{i=1}^{I_0} g_i(\hat{y}) \tag{6}$$

Likewise, for the I_1 training samples of class ω_1:

$$p(\hat{y} \mid \omega_1) = \frac{1}{I_1} \sum_{i=1}^{I_1} g_i(\hat{y}) \tag{7}$$

2.2.2.5 *Classification of an unknown sample using bayes decision*

For an unknown sample to be classified, the prediction $\hat{y}_u$ and its standard error of prediction SEP_u are calculated with PLS for A factors following Eqs. (1)-(2). If the classes ω_0 and ω_1 have probability density functions $p(\hat{y}|\omega_0)$ and $p(\hat{y}|\omega_1)$ respectively, the probability that a sample with prediction $\hat{y}_u$ belongs to the class ω_c (either class c=0 (ω_0) or class c= 1 (ω_1)), is given by the Bayes formula [24]:

$$P(\omega_c \mid \hat{y}_u) = \frac{p(\hat{y}_u \mid \omega_c) \times P(\omega_c)}{p(\hat{y}_u)} \tag{8}$$

where the denominator is:

$$p(\hat{y}_u) = p(\hat{y}_u \mid \omega_0) \times P(\omega_0) + p(\hat{y}_u \mid \omega_1) \times P(\omega_1) \tag{9}$$

The *a priori* probabilities $P(\omega_0)$ and $P(\omega_1)$ can be calculated from the training data set after assuming that the number of samples of each class in the training set is representative of the entire population, i.e., $P(\omega_0) = I_0/I$ and $P(\omega_1) = I_1/I$. The Bayesian decision rule tries to minimize the probability of error $P(error|\ \hat{y}_u)$ that we may incur when assigning a sample to a class. The probability of a wrong decision for class ω_0 is given by the probability that the sample with prediction $\hat{y}_u$ actually belongs

to class ω_1 i.e., $P(\omega_1|\,\hat{y}_u)$. Similarly the probability of a wrong decision for class ω_1 is given by $P(\omega_0\,|\,\hat{y}_u)$. Hence, the Bayesian decision rule for assigning an unknown sample in one of two classes, ω_0 and ω_1 is:

$$\text{Decide class } \omega_0 \text{ if } P(\omega_0\,|\hat{y}_u) > P(\omega_1\,|\,\hat{y}_u); \text{ otherwise decide class } \omega_1 \qquad (10)$$

Since the denominator in Eq. 8 is not fundamental for the decision, we can rewrite the rule as:

$$\text{Decide class } \omega_0 \text{ if } p(\hat{y}_u|\omega_1)\times P(\omega_1) < p(\hat{y}_u|\omega_0)\times P(\omega_0); \text{ otherwise, decide class } \omega_1 \qquad (11)$$

Fig. 2 shows the probability density function $p(\hat{y}_u|\omega_c)$, the probability density function multiplied by the *a priori* class probability $p(\hat{y}_u|\omega_c)\times P(\omega_c)$, and the *a posteriori* probability $P(\omega_c|\hat{y}_u)$ for the classes ω_0 and ω_1. The classification rule in Eq. (11) actually compares the value functions $p(\hat{y}_u|\omega_c)\times P(\omega_c)$ at the point $\hat{y}_u$. However, in DPLS, this comparison does not take into account that $\hat{y}_u$ has uncertainty.

Eq. (8) is the usual Bayes formula that is used when the variable takes well-defined values. However, there is an interval $U_u = \{\hat{y}_{u,l} \le \hat{y}_u \le \hat{y}_{u,r}\}$, in which the true value would be (Fig. 3). Thus, $\hat{y}_{u,l} = \hat{y}_u - k{\cdot}SEP_u$ and $\hat{y}_{u,r} = \hat{y}_u + k{\cdot}SEP_u$ (being k a coverage factor), are the left and right limits of the interval. Proposed values of k are $k = 1$ and $k = 2$, which resemble the values that in a normal distribution correspond to ~ 68% and ~ 95% confidence interval, respectively. By considering the interval, the Bayes equation changes into [30].

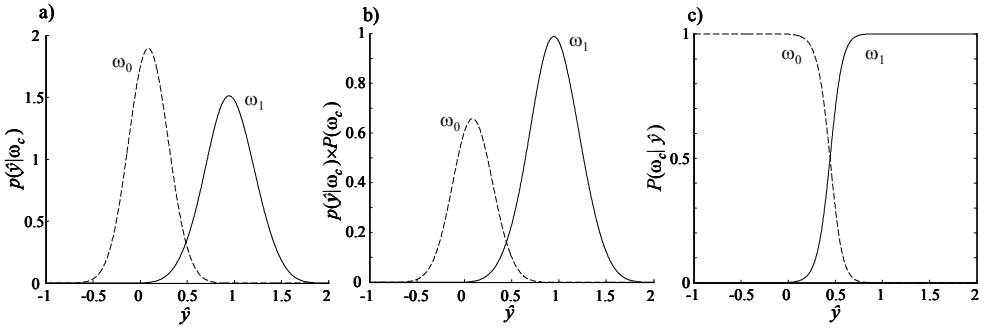

Fig. 2: (a) Probability density functions (PDF) for two classes, ω_0 (– – –) and ω_1 (—); (b) PDF multiplied by the prior probability of the class and (c) posterior probabilities.

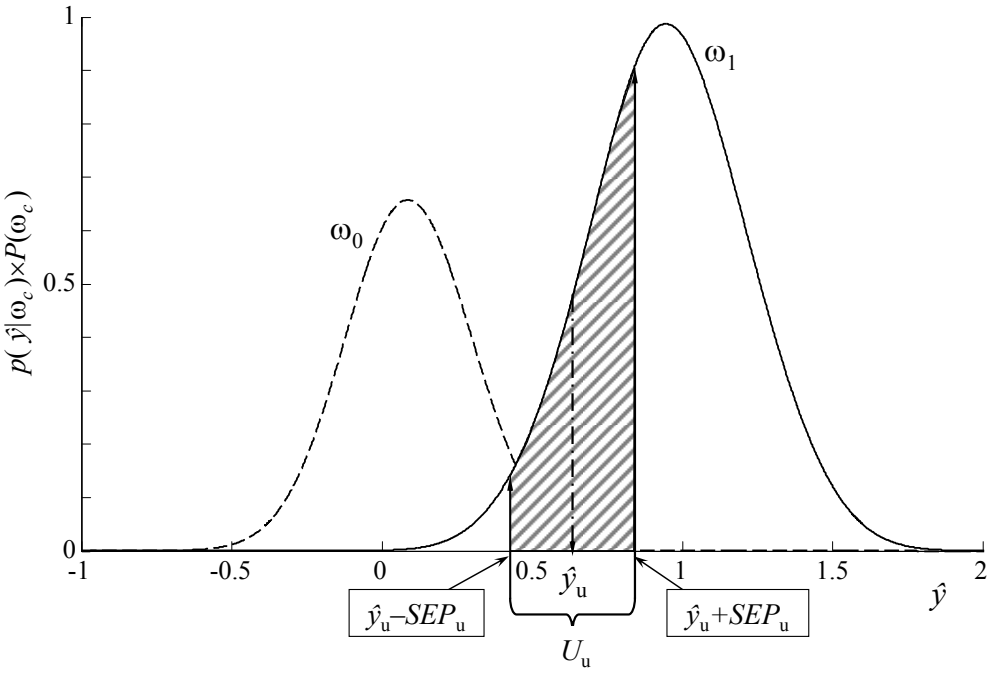

Fig. 3: Classification of a new sample and its uncertainty interval (U_U).

$$P\left\{\omega_c \mid \hat{y}_{u,l} \le \hat{y}_u \le \hat{y}_{u,r}\right\} = \frac{P\left\{\hat{y}_{u,l} \le \hat{y}_u \le \hat{y}_{u,r} \mid \omega_c\right\}}{P\left\{\hat{y}_{u,l} \le \hat{y}_u \le \hat{y}_{u,r}\right\}} P(\omega_c) = \frac{\int_{\hat{y}_{u,l}}^{\hat{y}_{u,r}} p(\hat{y}_u \mid \omega_c)\, d\hat{y}_u}{\int_{\hat{y}_{u,l}}^{\hat{y}_{u,r}} p(\hat{y}_u)\, d\hat{y}_u} P(\omega_c) \tag{12}$$

The numerator of Eq. (12) is the area under the curve $p(\hat{y}_u|\omega_c)\times P(\omega_c)$ in the interval U_u (Fig. 3):

$$Area_{u,c} = P(\omega_c)\int_{\hat{y}_{u,l}}^{\hat{y}_{u,r}} p(\hat{y}_u \mid \omega_c)\, d\hat{y}_u \tag{13}$$

The denominator of Eq. (12) can be expanded by inserting Eqs. (9) and (13):

$$\begin{aligned}\int_{\hat{y}_{u,l}}^{\hat{y}_{u,r}} p(\hat{y}_u)\,d\hat{y}_u &= P(\omega_0)\int_{\hat{y}_{u,l}}^{\hat{y}_{u,r}} p(\hat{y}_u \mid \omega_0)\,d\hat{y}_u + P(\omega_1)\int_{\hat{y}_{u,l}}^{\hat{y}_{u,r}} p(\hat{y}_u \mid \omega_1)\,d\hat{y}_u \\ &= Area_{u,0} + Area_{u,1}\end{aligned} \tag{14}$$

where $Area_{u,0}$ is the area under the curve of class ω_0 limited by the interval U_u, and $Area_{u,1}$ is the equivalent for the curve of class ω_1. Hence:

$$P\{\omega_c \mid \hat{y}_{u,l} \le \hat{y}_u \le \hat{y}_{u,r}\} = \frac{Area_{u,c}}{Area_{u,0} + Area_{u,1}} \tag{15}$$

where the subscript c (c = 0, 1) indicates either class ω_0 or class ω_1. Eq. (15) is calculated for the class ω_0, and also for the class ω_1. The classification rule is:

Decide class ω_1 if $P\{\omega_1|\hat{y}_{u,l} \le \hat{y}_u \le \hat{y}_{u,r}\} > P\{\omega_0|\hat{y}_{u,l} \le \hat{y}_u \le \hat{y}_{u,r}\}$; otherwise decide class ω_0 (16)

Since the denominator of Eq. (15) is the same both for $P\{\omega_1|\hat{y}_{u,l} \le \hat{y}_u \le \hat{y}_{u,r}\}$ and $P\{\omega_0|\hat{y}_{u,l} \le \hat{y}_u \le \hat{y}_{u,r}\}$, the decision rule can be enunciated in terms of areas as:

Decide class ω_1 if $Area_{u,1} > Area_{u,0}$; otherwise decide class ω_0 (17)

2.2.2.6 *Reliability of the classification of an unknown sample*

The classification decision is based on the actual predicted value. However, ideally it should be based on the *true* value y_u. This value is unknown but likely to be within the interval $\hat{y}_u \pm k{\cdot}SEP_u$. The difference between basing the classification decision in a single predicted value or in an interval is better seen when the prediction is in the boundary between classes, as it is shown in Fig. 3. If we base the decision only on $\hat{y}_u$ and Eq. (11), we conclude that the sample belongs to class ω_1. However, the prediction interval indicates that the true prediction could be as low as $\hat{y}_u - k{\cdot}SEP_u$ in which case the correct decision would be that the sample belongs to class ω_0. Thus, taking into account the confidence interval is a means of considering the different possible classifications for the sample. In this way, the *a posteriori* probability given by Eq. (15) gives a measure of the reliability of the classification. The higher the area is for one class and the lower for the other class, the more reliable the classification is.

2.2.3 Experimental part

2.2.3.1 *Data sets*

This new classification approach (*p*-DPLS) is illustrated with two data sets. Data set Iris [31] has 150 samples and 4 variables (length and width of both sepals and petals) of Iris Flower. There are three classes (Setosa, Versicolor and Virginica) with 50 samples each. Data set Oil [4] is composed of 226 Italian olive oil samples that are divided in 12 regions: Liguria (63), Sicilia (31), Lazio (29), Puglia (28), Umbria (18), Calabria (13), Molise (13), Veneto (9), Campania (8), Abruzzo (6), Lombardia (5) and Trentino Alto Adige (3). 700 variables were measured, corresponding to the values of absorbance in NIR spectroscopy, measured between 1100 and 2498 nm, every of 2 nm.

2.2.3.2 *Procedure and software*

The Kennard-Stone algorithm [32] applied to each class separately was used to split the data sets into a training set (with 75% of the samples) and a test set (with 25% of the samples). For dataset Iris, a model was established to classify the samples as belonging the class Virginica (class ω_1) or not (class ω_0, Setosa and Versicolor). For dataset Olive Oil, one model was developed to recognize the class Liguria (the largest class, with 63 samples, class ω_1) from the rest (class ω_0). The optimal number of factors was decided by leave-one-out cross-validation from the model that had the highest specificity (percentage of true negatives, i.e., samples that were correctly assigned to class ω_0) and sensitivity (percentage of true positives, i.e., samples that were correctly assigned to class ω_1). The classification for calculating the specificity and sensitivity was based on Eq. (17) for each sample taken out during the leave-one-out cross-validation. The proposed methodology *p*-DPLS was compared to the algorithm *PLSDTHRES* included in the PLS Toolbox 3.5 [19]. This method uses as a cut-off value the $\hat{y}$ for which the *a posteriori* probability (Eq. (8)) of the classes are equal.

All the routines were developed under Matlab environment (The MathWorks, Inc).

2.2.4 Results and discussion

2.2.4.1 *Data set Iris*

The results are reported for mean-centred data. The DPLS model of class Virginica was calculated following the proposed approach (Eqs. (1)–(7)), together with the *a priori* probability of classes $P(\omega_1) = 0.333$ and $P(\omega_0) = 0.667$. The general procedure based on the areas rule (Eq. (17)) for assigning an unknown sample is illustrated in Fig. 4a. The figure shows the product functions $p(\hat{y}_u|\omega_c)\times P(\omega_c)$, both for class ω_0 (segmented line) and for class ω_1 (solid line). The prediction for the sample Virginica124 is $\hat{y} = 0.561$ with $SEP = 0.145$. The figure also shows the intervals $\hat{y} \pm k \cdot SEP$ and the areas below the curves (Eq. (13)), with $Area_{u,0} = 0.117$ and $Area_{u,1} = 0.127$ for $k = 1$, and $Area_{u,0} = 0.225$ and $Area_{u,1} = 0.228$ for $k = 2$. These values of areas are used in the decision rule (Eq. (17)), so the sample is assigned to class ω_1, with calculated reliabilities (Eq. (15)) 52.2% (for $k = 1$) and 50.3% (for $k = 2$). Note that the sample Virginica124 is located near the boundary between classes. Hence the values of reliability obtained for class ω_1 is close to 50% indicating that although the sample has been assigned to the class ω_1, the result must be used with caution, since there is a large probability that it could belong to the class ω_0. The procedure was repeated for other five samples shown in Fig. 4. The results of classification were compared with the results of the Threshold method (*PLSDTHRES*). This method calculates $p(\omega_c|\hat{y}_u)$ as a Gaussian distribution with the mean and the standard deviation of all the predictions for each class. This distribution is then used to calculate the *a posteriori* probability, $P_T(\omega_c|\hat{y}_u)$ (Eq. (8) and Fig. 4b). The $\hat{y}$ value where $P_T(\omega_0|\hat{y}_u) = P_T(\omega_1|\hat{y}_u)$ defines the cut-off value used for assigning a sample to either class ω_0 or class ω_1. For the Threshold method, the cut-off value was found to be at $\hat{y} = 0.515$. A sample whose prediction is over 0.515 is assigned to class ω_1; otherwise it is assigned to class ω_0. Thus, the sample Virginica124 with $\hat{y} = 0.561$, is allocated by the Threshold method to the class ω_0, without taking into account that the predicted value is close to the cut-off value. Fig. 4b shows the *a posteriori* probability $P_T(\omega_1|\hat{y})$ (class ω_1) and $P_T(\omega_0|\hat{y})$ (class ω_0), and the value of $P_T(\omega_1|\hat{y})$ of the six tested samples. The figure also shows (in parenthesis) the reliability calculated by the *p*-DPLS method for the class ω_1 with $k = 1$ and $k = 2$, respectively. The classification of the six samples for the *p*-DPLS method and for the Threshold method is shown in Table 1.

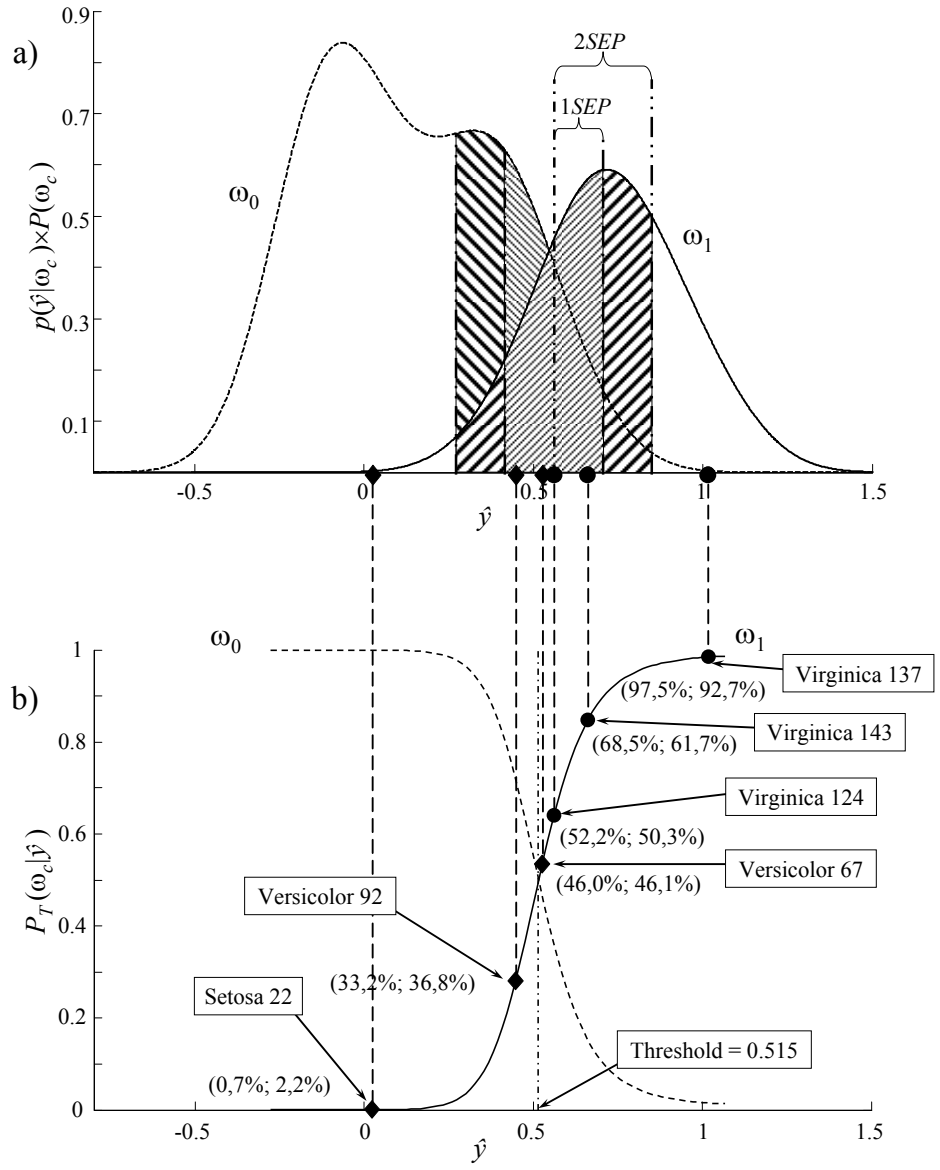

Fig. 4: Predictions of the six test samples (see text for details). a) Product functions. The areas in interval $\hat{y}_u \pm SEP_u$ and $\hat{y}_u \pm 2\times SEP_u$ for sample Virginica 124 are shown; $Area_{u,1}$ (diagonal upwards) and $Area_{u,0}$ (diagonal downwards). b) *a posteriori* probability functions for the Threshold method (continuum line for class ω_1 and dashed line for class ω_0), identifying the threshold calculated with the vertical line (0,515). The reliability of the classification for *SEP* and 2×*SEP* is also indicated in parentheses if the sample were to be classified into class ω_1. (●) samples that belong to class ω_1, (♦) samples that belong to class ω_0.

Table 1: Iris dataset. Classification of six test samples for the model with four factors.

Sample	True class	$\hat{y}$	Threshold	*p*-DPLS										
			Assigned Class	*SEP*	*SEP*					2×*SEP*				
					Area		Reliability (%)		Assigned Class	*Area*		Reliability (%)		Assigned Class
					ω_0	ω_1	ω_0	ω_1		ω_0	ω_1	ω_0	ω_1	
Virginica 137	1	1,016	1	0,148	0,002	0,082	2,5	97,5	1	0,014	0,171	7,3	92,7	1
Virginica 143	1	0,660	1	0,145	0,071	0,155	31,5	68,5	1	0,162	0,260	38,3	61,7	1
Virginica 124	1	0,561	1	0,145	0,117	0,127	47,8	52,2	1	0,225	0,228	49,7	50,3	1
Versicolor 67	0	0,525	1	0,146	0,134	0,114	54,0	46,0	0	0,248	0,212	53,9	46,1	0
Versicolor 92	0	0,447	0	0,144	0,162	0,081	66,8	33,2	0	0,290	0,169	63,2	36,8	0
Setosa 22	0	0,023	0	0,145	0,224	0,002	99,3	0,7	0	0,416	0,010	97,8	2,2	0

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

For the Virginica class, the optimal DPLS model, obtained by cross validation, was of four factors for the Threshold method, with sensitivity of ~89.5% and specificity close to 92%. For the *p*-DPLS method, the optimal model had also four factors, and the sensitivity of ~84% and specificity of ~93%. The classification performance for the test set with the Threshold method gave a sensitivity of 100% and specificity close to 96%. For the *p*-DPLS method, the sensitivity and specificity are of 100%. Note that the sensitivity of the *p*-DPLS method and the Threshold method for the test set are the same, but the specificity is higher for the *p*-DPLS method. This is due to the fact that Threshold misclassifies sample Versicolor67 (i.e. it incorrectly assigns the sample to class ω_1) because its predicted value is 0.525 (only slightly higher than the cut-off value 0.515) (Fig. 4b). However, *p*-DPLS method classifies the sample into class ω_0 because the areas for class ω_0 are 0.134 (k =1) and 0.248 (k =2) while the areas for class ω_1 are 0.114 (k =1) and 0.212 (k =2) (i.e. slightly higher for class ω_0). The reliability of such assignment is ~54%, which, again, suggests that we should be cautious with the classification decision. Note that in this case we can know that *p*-DPLS gives a correct classification. However, in a real situation in which the true class of the test sample is unknown, the low reliability might be interpreted as a need of further data about that sample before a classification decision is taken.

In order to further evaluate how the *p*-DPLS method performs, two more samples for each class were studied. The samples Virginica137 and Setosa22 are adequately classified by the model. The reliability of such classifications is above 93% for sample Virginica137 and above 97% for sample Setosa22 (for both k =1 and k = 2 confidence intervals). Note that these samples have predictions close to the ideal 0 and 1, respectively (Table 1) and there is a low overlap between the PDFs. This produces the high reliabilities for these samples. Finally, samples Versicolor92 and Virginica143 have predictions 0.447 and 0.660 respectively, in the zone of the largest overlap of the classes (near the limit of the classes). Although the *p*-DPLS assigns them to the correct class, the reliability of the classification is around 65% for each sample.

2.2.4.2 *Data set of olive oil*

A DPLS model was calculated for classifying a sample as from the Liguria region (class ω_1) or non-Liguria region (class ω_0). The optimal model was obtained for mean-centred data after removing two outliers, one sample of the Sicilia class and one sample of the Molise class. The optimal number of factors, obtained by cross validation, was 15 for the Threshold method. This model had the maximum sensitivity (close to 92%)

and maximum specificity (close to 93%). For the *p*-DPLS method, the optimal number of factors was 26, with sensitivity close to 83% and specificity close to 95%. This high value of the optimal number of factors is due to the large variability of the training set, in which class ω_0 includes 11 geographical origins.

The classification performance for the test set with the Threshold method gave a sensitivity of 87.5% and a specificity of 97.4%, which corresponds to 2 and 1 misclassified samples respectively. For the *p*-DPLS method, the sensitivity and specificity were 93.8% and 100%, respectively. This latter value of sensitivity is due to a false negative of the sample Liguria 014 (Table 2). This sample was misclassified by both classification methods. The sample seems to belong to Non-Liguria, as it is observed in the prediction plot (Fig. 5a), which made us suspect that a Non-Liguria sample was mislabelled as from Liguria. However, this hypothesis could not be checked with the information available. For this sample, the reliability for class ω_0, calculated by the *p*-DPLS method is high (around 96%).

Different from the model for the previous dataset, in which both classes studied had overlapped PDF's, the model for the Liguria class has a very low overlap of the probability product functions (Fig. 5a). For this reason, any sample whose prediction is less than around 0.25 or higher than around 0.75 is classified with reliability close to 99% in class ω_0 and class ω_1 respectively (Fig. 5b). In the interval of predictions from 0.25 to 0.75 the reliability of assignment to class ω_1 varies significantly, and samples with a very similar predicted value, may have quite a different reliability. In addition, note that, for example, sample Lazio 207 of class ω_0 has a predicted value of 0.506; in a zone where the overlap is maximum. This sample has an associated area of 0.026 for class ω_0 and of 0.021 for class ω_1 (for $k = 1$), and an area of 0.108 for class ω_0 and 0.067 for class ω_1 (for $k = 2$). It is seen that, by increasing the level of confidence, the area increases appreciably, which in turn changes the reliability of the assignment to the class ω_0, increasing from 55% ($k = 1$) to 62% ($k = 2$).

Note also that, since reliabilities are calculated as a ratio of areas, values of 50% of reliability can be obtained in the central zone both if the overlap of the PDFs is high (as in the previous dataset) but also if the overlap of the PDFs is low (as in this dataset). Some authors [20, 21] have suggested that such a zone, in which the probability density functions have a low value, should be used as a non-decision (reject) zone in which a sample is not classified in any of the two classes. Work is under way in order to adapt the *p*-DPLS method to accept a reject region.

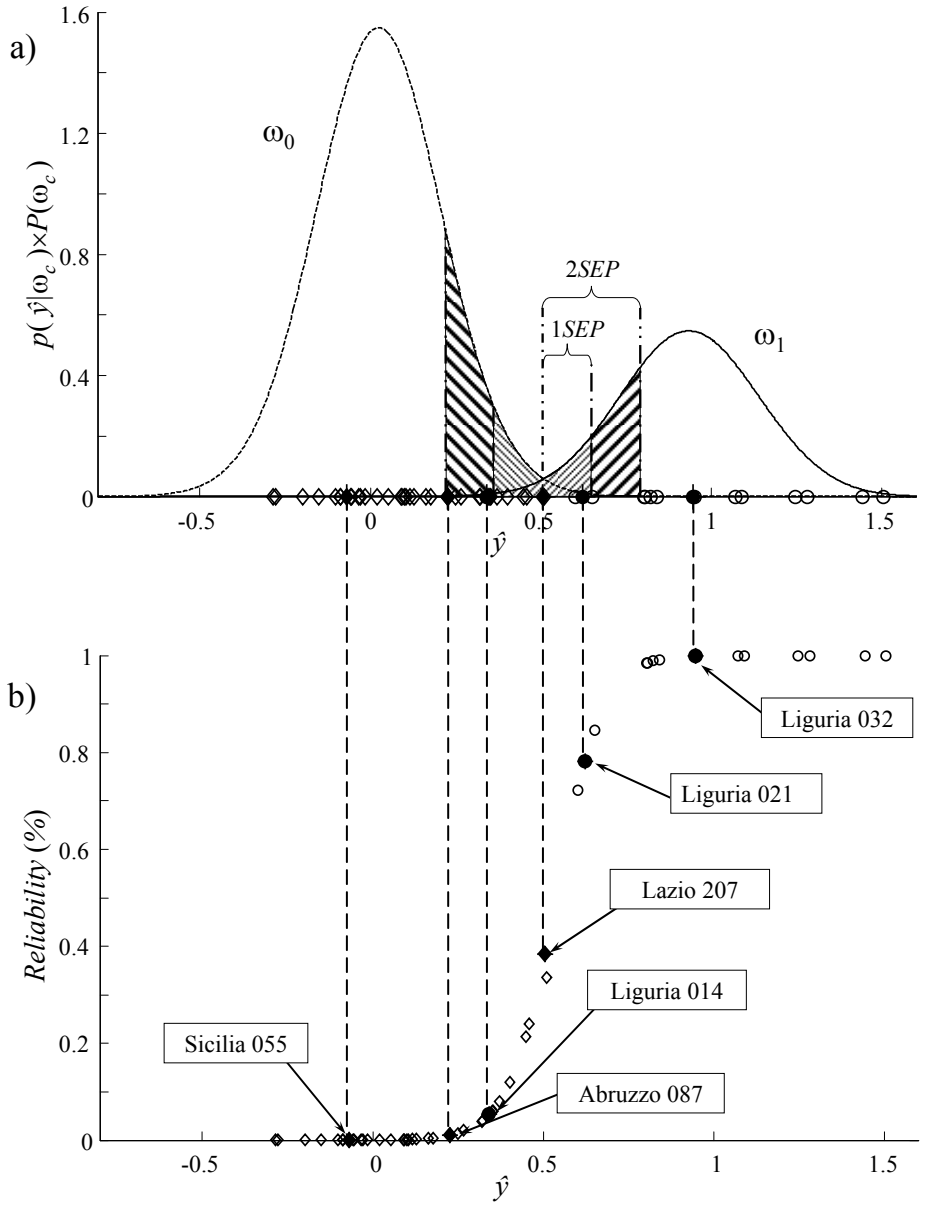

Fig. 5: Predictions of the values for the test samples. a) Product functions. The areas in the intervals $\hat{y}_u \pm SEP_u$ and $\hat{y}_u \pm 2\times SEP_u$ for sample Lazio207 are shown; $Area_{u,1}$ (diagonal upward) and $Area_{u,0}$ (diagonal down). b) $\hat{y}$ vs. reliability if the sample were to be assigned to class ω_1. The predictions for class ω_1 are indicated with ● and ○; and those of class ω_0 are indicated with ♦ and ◊. ● and ♦ indicate the samples whose results are tabulated in Table 2.

Table 2: Olive oil dataset. Classification of six test samples for the Threshold method with 15 factors, and the *p*-DPLS method with 26 factors.

Sample	True class	Threshold		*p*-DPLS										
					SEP						2×*SEP*			
						Area		Reliability (%)			*Area*		Reliability (%)		
		$\hat{y}$	Assigned Class	$\hat{y}$	*SEP*	ω_0	ω_1	ω_0	ω_1	Assigned Class	ω_0	ω_1	ω_0	ω_1	Assigned Class
Liguria 032	1	1,067	1	0,946	0,144	0	0,146	0,01	99,9	1	0	0,235	0,1	99,9	1
Liguria 021	1	0,494	0	0,623	0,142	0,006	0,052	10,2	89,8	1	0,034	0,124	21,9	78,1	1
Liguria 014	1	0,370	0	0,342	0,142	0,119	0,003	97,2	2,8	0	0,309	0,017	94,7	5,3	0
Lazio 207	0	0,423	0	0,506	0,144	0,026	0,021	55,3	44,7	0	0,108	0,067	61,7	38,3	0
Abruzzo 087	0	0,620	1	0,226	0,146	0,252	0.001	99,7	0,3	0	0,492	0,005	98,9	1,1	0
Sicilia 055	0	-0,011	0	-0,070	0,166	0,408	0	100	0	0	0,638	0	100	0	0

2.2.5 Conclusions

A methodology based on DPLS has been developed in order to correct the limitation of using an arbitrary threshold for assignment between the classes. The new method takes into account the uncertainty in the predictions and also the *a priori* probability of the classes. These concepts have been unified by constructing a PDF by averaging individual kernel functions centred in the predictions of the training set for each class. In addition, the new methodology has the added benefit of providing the reliability of the classification, which is especially interesting for those samples that are near the boundaries of the classes. The reliability of the classification is calculated from the area under the curve $p(\hat{y}_u|\omega_c)\times P(\omega_c)$ within two limits defined by the standard error of prediction of $\hat{y}_u$, where $\hat{y}_u$ is the prediction for the unknown sample calculated with the PLS model with A factors. This methodology is currently being tested in multi-class classification, in which the reliability of the classification in individual binary models plays an important role for deciding the final classification result.

Acknowledgments

The authors express their gratitude to Dr. Gerard Downey for providing the Olive oil dataset. This work was supported by the Project TRACE – "TRAcing food Commodities in Europe (EU IP 006942) from the Sixth Framework Programme of the European and by the Spanish Ministerio de Educación y Ciencia project CTQ2007-66918/BQU. The publication reflects only the authors' views and the Community nor the Spanish Ministerio are liable for any use that may be made of the information contained therein.

References

1. Y. Siderer, A. Maquet, Elke Anklam, Trends Food Sci. Tech. 16 (2005) 332–343.
2. S. Kelly, K. Heaton, J. Hoogeweff, Trends Food Sci. Tech. 16 (2005) 555–567.
3. http://ec.europa.eu/agriculture/foodqual/quali1_en.htm. Last accessed en 21st of April 2008.

4. Project TRACE – "TRAcing food Commodities in Europe" (project no. FOOD-CT-2005-006942). www.trace.eu.org.
5. Sixth Framework Programme Priority 5, Food Quality and Safety, July 2004.
6. B. R. Kowalski, Chemometrics. Mathematics and statistics in Chemistry, D. Reidel Publishing Company, Dordrecht, Holland, 1984, p.85–88.
7. B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi and J. Smeyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier Science B. V., Amsterdam, The Netherlands. 1998. p.80–82, 225–227, 331–337.
8. N. Bøknæs, K. N. Jensen, C. M. Andersen, H. Martens, Lebensm.-Wiss. u.-Technol. 35 (2002) 628–634.
9. C. Jacobsen, J. Adler-Nissen, A. S. Meyer, J. Agric. Food Chem. 47 (1999) 4917–4926.
10. Å. Eriksson, K. Persson Waller, K. Svennersten-Sjaunja, J-E. Haugen, F. Lundby, O. Lind, Int. Dairy J. 15 (2005) 1193–1201.
11. L. Leon, J. D. Nelly, G. Downey, App. Spec. 59 (2005) 593–599.
12. D. Cozzolino, A. Chree, J. R. Scaife, I. Murray, J. Agric. Food Chem. 53 (2005) 4459–4463.
13. K. Jahan, A. Paterson, J. R. Piggott, Food Res. Int. 38 (2005) 495–503.
14. D. Cozzolino, H. E. Smyth, M. Gishen, J. Agric. Food Chem. 51(2003) 7703–7708.
15. K. P. Singh, A. Malik, V. K. Singh, D. Mohan, S. Sinha, Anal. Chim. Acta. 550 (2005) 82–91.
16. G. J. Hall, K. E. Clow, J. E. Kenny, Environ. Sci. Technol. 39 (2005) 7560–7567.
17. Y. Woo, H. Kim, H. Chung, Analyst. 124 (1999) 1223–1226.
18. C. M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, NY, 2006, p. 21.
19. B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig, R. S. Koch, PLS_Toolbox Version 3.5 for use with MATLAB™, Eigenvector Research, Inc., Manson, WA, USA, 2005, p.185–189.
20. H. Mauser, O. Roche, M. Stahl, S. Müller, J. Chem. Inf. Model. 45 (2005) 1039–1046.
21. L. Afzelius, C. M. Masimirembwa, A. Karlén, T. B. Andersson, I. Zamora, J. Comput. Aided Mol. Des. 16 (2002) 443–458.
22. C. Eker, R. Rydell, K. Svanberg, S. Andersson-Engels, Laser Surg. Med. 28 (2001) 259–266.
23. C. Wang, C. Chen, C. Chlang, S. Young, S. Chow, H. K. Chiang, Photochem. Photobiol. 69 (1999) 471–477.
24. R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd ed., John Wiley & Sons, Inc., New Your, NY, USA, 2000, p.20–84.
25. N. M. Faber, R. Bro, Chemom. Intell. Lab. Syst. 61 (2002) 133–149.
26. A. Höskuldsson, J. Chemometr. 2 (1988) 211–228.

27. K. H. Esbensen, Multivariate Data Analysis – in practice, CAMO ASA, 4th ed., Oslo, Norway, 2000, p. 200–209.
28. D. Coomans, I. Broeckaert, Potential Pattern Recognition in Chemical and Medical Decision Making, Research Studies Press, Letchworth, U.K., 1986.
29. D. Coomans, D. L. Massart, Anal. Chim. Acta. 133 (1981) 215–224.
30. A. Papoulis, Probability, Random Variables and Stochastic Processes, McGraw-Hill, 3rd ed., USA, 1991, p. 63–84.
31. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/iris/ Last accessed en 21st of April 2008.
32. R.W. Kennard, L.A. Stone. Technometrics. 11 (1969) 137–148.

2.3 Corrección del sesgo en clasificación con *p*-DPLS

2.3.1 Introducción

Sesgo es el error sistemático que se comete al tomar una medida [1]. El error sistemático puede provenir de múltiples fuentes como el instrumento, el método o el analista y, a diferencia del aleatorio, se puede corregir si se conoce [2]. Otro tipo de sesgo es el error sistemático de predicción, que se toma como el promedio de las diferencias entre los valores predichos y reales de una muestra representativa de objetos [3]. Este sesgo forma parte del error cuadrado medio (*Mean Square Error*, ***MSE***) y en algunos casos es necesario corregirlo, ya que su valor puede ser superior al error aleatorio, llevando a cometer subestimaciones o sobreestimaciones en las predicciones de los modelos de regresión lineal.

La clasificación con *p*-DPLS también puede verse afectada por el sesgo de la predicción, ya que si éste sesgo es elevado se altera la decisión de clasificación del modelo, aumentando el error de clasificación. Varios métodos de clasificación han buscado corregir dicho sesgo. Por ejemplo, en redes neuronales se ha buscado un factor de corrección del sesgo mediante técnicas de *bootstrap* [4], en árboles de regresión mediante *bootstrap* han buscado las variables con sesgo y se han eliminado [5], en clasificadores binarios, como los Booleanos, en donde el error de asignación tiene un sesgo elevado cuando se tiene un conjunto de datos pequeño, lo que no permite calcular adecuadamente el error de Bayes que mide la capacidad de la variables para discriminar la clases [6] y en métodos *Kernel*, en donde la corrección del sesgo mejora las curvas de densidad y reducen el solapamiento entre clases [7].

El sesgo de la predicción en PLS se ha estudiado en diversas ocasiones. *Höskuldsson* [8] halló el sesgo para PLS llevando el coeficiente de regresión a valores de 0 o cercanos. *Faber y Kowalski* [9,10] plantearon el cálculo del sesgo teniendo en cuenta los aportes por errores en la medida de las variables dependientes e independientes, y como éstos afectan al sesgo del modelo. Sin embargo, en la mayoría de los cálculos del error de predicción en PLS se ha ignorado el sesgo en la predicción [11-15].

En DPLS se han planteado métodos particulares para calcular el sesgo, como la perturbación aleatoria de los objetos y evaluación del efecto de la perturbación [16], o aplicando *bootstrap*, lo que permite obtener un valor promedio de la desviación con respecto al valor asignado inicialmente [17].

2.3.2 El sesgo en *p*-DPLS

En PLS se considera que la magnitud y dirección de sesgo depende de la correlación de las variables independientes y que la magnitud relativa del error depende de la inter-correlación de estas variables [18]. Esto permite observar una cierta correlación entre la magnitud del error de predicción y las contribuciones del sesgo y varianza del error, siendo el error de predicción particularmente alto cuando la contribución del sesgo es elevada [9]. Así pues, ha de esperarse un comportamiento similar en DPLS, por lo que podemos suponer que las ecuaciones para el cálculo del *MSE* con corrección del sesgo [8] son igualmente aplicables a DPLS:

$$MSE = bias^2 + Var\left[\hat{\mathbf{y}} - \mathbf{y}\right] \tag{2-12}$$

donde *Var* es la varianza entre el vector predicho ($\hat{\mathbf{y}}$) por el modelo PLS y el verdadero ($\mathbf{y}$), y el *bias* es el sesgo que se calcula como:

$$bias = \frac{\sum_{i=1}^{I}(\hat{y}_i - y_i)}{I} \tag{2-13}$$

donde $\hat{y}$ es el valor predicho por el modelo PLS e y es el valor verdadero (0 ó 1) y el sumatorio se extiende a los I objetos.

Ensayos sobre el conjunto de datos *Fisher Iris* [19] muestran que el sesgo en las predicciones de los objetos en DPLS varían según la clase a la que pertenecen estos. Teniendo en cuenta que el *MSE* se utiliza en el cálculo del error de estándar de predicción (Ec. 2 apartado 2.2.2.2), presente en el cálculo de las funciones potenciales de los datos de entrenamiento y en el intervalo de incertidumbre para objetos desconocidos, se hizo necesario incluir el efecto de las clases calculando el sesgo correspondiente a las predicciones de los objetos de la clase c (c=1, 2,…, C), como:

$$bias_c = \frac{\sum_{i=1}^{I_c}(\hat{y}_i - y_i)}{I_c} \tag{2-14}$$

Téngase en cuenta que se calcula un valor de $bias_c$ para cada clase del conjunto de entrenamiento, y no un sesgo global para todas las clases del modelo binario DPLS.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2.3.3 Ejemplo ilustrativo con el conjunto de datos *Fisher Iris*

El conjunto de datos tiene tres clases: *Setosa*, *Versicolor* y *Virginica*. Se desarrollaron tres modelos DPLS codificando la clase de interés como 1 y las dos restantes como 0. En todos los casos los datos se centraron. Para hallar el número óptimo de factores se consideró el mínimo error por validación cruzada dejando fuera un objeto cada vez (LOOCV). Así el número óptimo de factores para los tres modelos según clase de interés es: dos para el modelo *Setosa vs.* restantes, tres para el *Versicolor vs.* restantes y tres para el *Virginica vs.* restantes.

La figura 2-3 muestra los errores de predicción para los objetos predichos por el modelo *Setosa vs.* restantes (clase 1: *Setosa*, clase 0: *Versicolor* y *Virginica*). Se observan errores negativos en los objetos de las clases *Setosa* y *Virginica* mientras que los objetos de la clase *Versicolor* tienen errores positivos. La tabla 2-1 muestra como la distribución de los errores afecta el cálculo de sesgo. En ella se han tabulado los resultados de calcular el sesgo siguiendo 3 premisas: calcular el sesgo con todos los objetos (fila 1, tabla 2-1), teniendo en cuenta la codificación binaria (se calculan dos sesgos, uno para los objetos de la clase "0" y otro para los objetos de clase "1" (filas 2 y 3, tabla 2-1) y

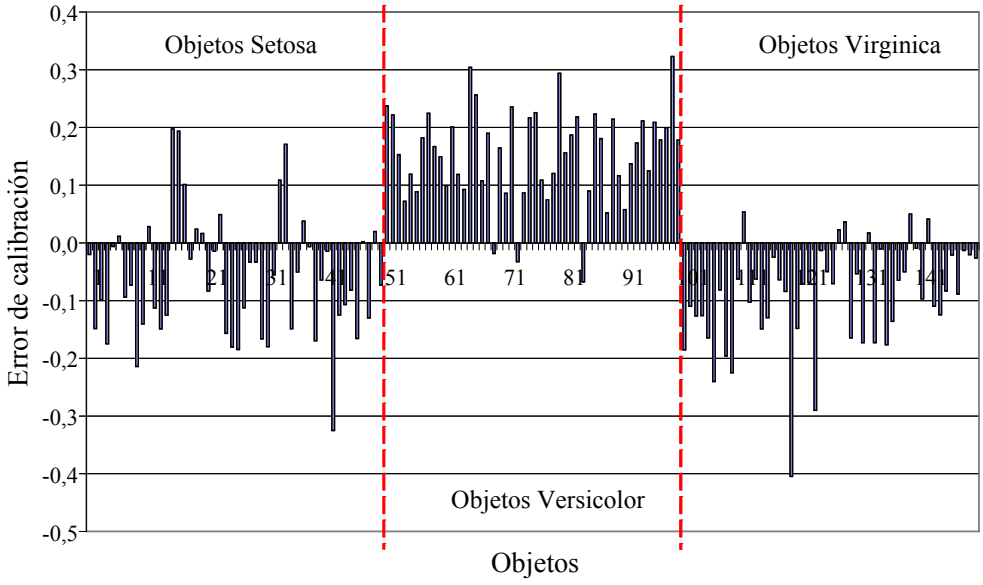

Figura 2-3: Errores de calibración para el conjunto de datos *Iris* al modelar la clase *Setosa* como de interés con DPLS.

finalmente considerando un sesgo para cada clase (se calculan tres sesgos, cada uno con sólo los objetos de cada clase, filas 4-6, tabla 2-1). Ésta última estimación del sesgo parece la más adecuada a partir de la evidencia gráfica (Figura 2-3). El cálculo se repitió para cada uno de los modelos y cada una de las clases (columnas de la tabla 2-1).

Aplicado el concepto clásico de sesgo (un sesgo global para todos los objetos), se puede decir que éste es despreciable, alrededor de 10^{-16} a 10^{-18}, debido a que se compensan los errores negativos y positivos, ocultando la tendencia por clases. Cuando se calculó el sesgo separadamente para los objetos de las clases "0" y "1", se apreció la diferencia entre clases: el sesgo es siempre negativo para la clase "1" y positivo para la clase "0"; esta diferencia es mayor en el modelo *Versicolor vs.* restantes, donde la diferencia entre sesgos es de ~0.75, contrastando con la diferencia en el modelo de *Setosa vs.* restantes que es de ~0.09; así, el sesgo es decisivo en el modelo de la clase *Versicolor vs.* restantes.

Finalmente se calculó el sesgo teniendo en cuenta las clases iniciales. En el modelo *Setosa vs.* restantes se observa un sesgo para las clases *Versicolor* y *Virginica* que no era apreciable con las dos clases reunidas (Clase "0"). Es decir, que al calcular el sesgo como una sola clase no se considera lo opuestas que son *Versicolor* y *Virginica*, perdiendo información acerca de la posición de las predicciones, de cada una de las clases, con respecto al valor binario asignado. Algo similar ocurre con el modelo *Virginica vs.* restantes, donde el sesgo entre la predicciones de clase y el valor binario asignado inicialmente fue aún mayor. Para el modelo *Versicolor vs.* restantes aunque las clases *Setosa* y *Virginica* obtuvieron un sesgo positivos, difieren en ~0.2 es suficiente para distinguir las predicciones de una clase de las de la otra.

Tabla 2-1: Sesgo calculado bajo diferentes condiciones para los tres modelos posibles del conjunto *Fisher Iris.*

Sesgo	Modelos		
	Setosa *vs.* restantes	*Versicolor* *vs.* restantes	*Virginica* *vs.* restantes
Total	$1{,}92\times10^{-18}$	$9{,}15\times10^{-16}$	$-2{,}58\times10^{-16}$
Clase "1" (interés)	-0,06	-0,50	-0,26
Clase "0" (no interés)	0,03	0,25	0,13
Clase *Setosa*	-0,06	0,15	-0,09
Clase *Versicolor*	0,15	-0,50	0,35
Clase *Virginica*	-0,09	0,35	-0,26

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Tabla 2-2: *MSEC* calculados con diferentes correcciones del sesgo.

Método para calcular el sesgo	Modelos		
	Setosa *vs.* restantes	*Versicolor* *vs.* restantes	*Virginica* *vs.* restantes
Clásico*	0,021	0,171	0,089
Binario**	0,019	0,044	0,054
Por Clases***	0,009	0,037	0,022

* El mismo sesgo para todos los objetos.
** Un sesgo para la clase de interés y otro para las que no son de interés.
*** Un sesgo para cada una de las clases originales.

Los resultados sugieren que es necesario corregir selectivamente el sesgo para los objetos dependiendo de su clase inicial. Sin embargo, aún queda por aclarar el efecto que esta corrección tiene sobre el *MSEC*. La tabla 2-2 reúne los *MSEC* considerando las posibles formas de calcular el sesgo (tabla 2-1). Al considerar un sesgo para todos los objetos se obtienen valores de *MSEC* superiores que los obtenidos al hacer la corrección del sesgo por clases. Las diferencias significativas entre *MSEC* (columnas de la tabla 2-2) se dan porque el método clásico no detecta el sesgo, y por lo tanto no lo corrige. Por el contrario con los otros métodos, binario y por clases, se observa un sesgo que es mayor cuando se toma por clases, permitiendo corregir el sesgo y obtener un *MSEC* que depende de la varianza de las predicciones y no del sesgo, como sucede con el clásico.

2.3.4 Conclusiones

El sesgo en DPLS depende tanto de las variables de la matriz de datos **X** como de las clases a que pertenezcan cada uno de los objetos de la matriz. Se confirma que es necesario utilizar la ecuación 3 del apartado 2.2.2.2 para corregir el sesgo, ya que como se observó, existe un sesgo diferente para cada clase. Además, la corrección del sesgo por clases lleva a una notable reducción del *MSEC*.

Referencias

1. A. Maroto Sánchez, Incertidumbre en Métodos Analíticos de Rutina, Universitat Rovira i Virgili, Tarragona, España, 2002, p. 32–36.
2. Comité Conjunto para las Guías en Metrología (JCGM), "Vocabulario Internacional de Metrología – Conceptos fundamentales y generales, y términos asociados", 2008, p 29–30, 49.
3. H. Martens, T. Næs, Multivariate Calibration, John Wiley & Sons, Guildford, UK, 1989, p. 246–250.
4. M. Tsujitani, T. Koshimizu, IEEE Trans. Neural Network. 11 (2000) 1394–1401.
5. W.-Y. Loh, Statistica Sinica. 12 (2002) 361–386.
6. M. Brun, D. L. Sabbagh, S. Kim, E. R. Dougherty, Bioinformatics. 19 (2003) 944–951.
7. M. L. Hazelton, Comput. Stat. Data Anal. 51 (2007) 4393–4402.
8. A. Höskuldsson, J. Chemometr. 2 (1988) 211–228.
9. K. Faber, B. R. Kowalski, J. Chemometr. 11 (1997) 181–238.
10. N. M. Faber, J. Chemometr. 14 (2000) 363–369.
11. A. Lorber, B. R. Kowalski, J. Chemometr. 2 (1988) 93–109.
12. T. V. Karstang, J. Toft, O. M. Kvalheim, J. Chemometr. 6 (1992) 177–188.
13. The Unscrambler User Manual, CAMO ASA, Oslo, Norway, 1998, p. 436.
14. S. De Vries, C. J.F. Ter Braak, Chemometr. Intell. Lab. Syst. 30 (1995) 239–245.
15. K. Faber, B. R. Kowalski, Chemometr. Intell. Lab. Syst. 34 (1996) 283–292.
16. W. J. Krzanowski, P. Jonathan, W. V. McCarthy, M. R. Thomas, Appl. Statist. 44 (1995) 101–115.
17. O. Preisner, J. A. Lopes, J. C. Menezes, Chemometr. Intell. Lab. Syst. 94 (2008) 33–42.
18. S. D. Hodges, P. G Moore, Appl. Statist. 21 (1972) 185–195.
19. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/iris/, Último acceso (7/7/2009).

2.4 Clasificación con *p*-DPLS frente a otras versiones de DPLS

2.4.1 Introducción

Un paso esencial en DPLS es hallar una función de decisión de clasificación que permita asignar un objeto a la clase a la que pertenece a partir de la predicción del modelo DPLS. La función más simple y común es colocar un límite entre clases. El más utilizado es fijar el límite en 0.5 u optimizar dicho límite hasta obtener un buen porcentaje de clasificación. Otros métodos de clasificación que utilizan DPLS establecen una función de decisión en términos de la probabilidad, como el método incluido en el "*software*" PLS *Toolbox* 3.5 [1], algoritmo PLSDTHRES o el método propuesto en esta tesis doctoral (apartado 2.2) [2], que ha sido llamado *p*-DPLS. PLSDTHRES construye dos FDPs suponiendo una distribución normal de las predicciones en cada una de las clases (ω_0 y ω_1); así, establece como límite entre clases el valor $\hat{y}$ donde las probabilidades *a posteriori* para las dos clases son iguales. *p*-DPLS, en cambio, construye FDPs como promedio de funciones gaussianas alrededor de las predicciones de cada una de las clases y establece una función de decisión que evalúa el área, limitada por la incertidumbre, bajo las FDPs de cada una de las clases, para determinar la clase con mayor probabilidad de pertenecer.

Entre estas metodologías hay dos diferencias sustanciales. La primera es cómo se construyen las FDPs. *p*-DPLS no supone una distribución normal de las predicciones, dando lugar a FDPs más acordes con la distribución real de las predicciones; mientras que PLSDTHRES supone una distribución normal de las predicciones. La segunda diferencia tiene que ver con la función de decisión. Aunque tanto PLSDTHRES como *p*-DPLS asignan el objeto a la clase más probable, PLSDTHRES no puede localizar más de un punto de corte (límite de decisión) entre las funciones de clases cuando éstas están totalmente solapadas (ver figura 2-4). Por el contrario, *p*-DPLS evalúa el valor de probabilidad de los objetos en las FDPs de las clases, asignándolos a la clase con mayor probabilidad. Para entender mejor la ventajas y desventajas de los métodos del límite en 0.5, PLSDTHRES y *p*-DPLS, compararemos los resultados obtenidos con el conjunto de datos *Fisher Iris* para la clase *Versicolor*, la más compleja de clasificar debido a que comparte características con la clase *Virginica*, en el espacio original de variables los objetos de una y otra clase están mezclados siendo difícil diferenciarlos.

2.4.2 Ejemplo ilustrativo con el conjunto de datos *Fisher Iris*

La comparación se hizo con el conjunto de datos *Fisher Iris*, que se dividido en datos de entrenamiento (75% de los objetos) y datos prueba (25% de los objetos) utilizando el algoritmo *Kennard-Stone* [3]. Se desarrolló el modelo DPLS *Versicolor* (ω_1, clase de interés) *vs. Setosa-Virginica* (ω_0, clases de no interés), se hizo validación cruzada dejando fuera un objeto cada vez (LOOCV) y se asignaron los objetos problema utilizando las tres funciones de decisión. Finalmente, el número óptimo de factores se decidió por mínimo error de tipo I en los tres casos (porcentaje de objetos de la clase de interés rechazados).

En la tabla 2-3 se han tabulado las respuestas de clasificación con las funciones de decisión: límite en 0.5, PLSDTHRES y *p*-DPLS, para los conjuntos de entrenamiento y prueba. Se ha calculado el error de tipo I (porcentaje de objetos de la clase de interés rechazados) y error de tipo II (porcentaje de objetos erróneamente asignados a la clase de interés). La peor función de decisión fue el límite en 0.5. El modelo requirió un elevado número de factores (4) y obteniendo errores de tipo I y II elevados, tanto para los objetos de entrenamiento como de prueba. En cambio, los métodos cuyas funciones de decisión se basan en probabilidad, PLSDTHRES y *p*-DPLS, obtuvieron errores de tipo I similares y reducidos, menores para *p*-DPLS. En contraste, el error de tipo II fue elevado para PLSDTHRES y reducido con *p*-DPLS. Éste comportamiento es idéntico para los objetos de entrenamiento y de prueba. Así, *p*-DPLS es el más adecuado para asignar objetos de la clase *Versicolor*, pues presenta errores de tipo I y II reducidos.

Tabla 2-3: Porcentajes de clasificación para las tres funciones de decisión a partir de predicciones del modelo DPLS *Versicolor vs. Setosa−Virginica.*

Método de clasificación	Conjunto de datos	Nº óptimo de factores	Error tipo I	Error tipo II
Límite en 0.5	Entrenamiento	4	60.5%	15.8%
	Prueba		66.7%	0.0%
PLSDTHRES	Entrenamiento	1	10.5%	50%
	Prueba		8.3%	50%
p-DPLS	Entrenamiento	1	7.9%	6.6%
	Prueba		0.0%	4.2%

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Un análisis más detallado del funcionamiento de los métodos probabilísticos se muestra en las figuras 2-4 y 2-5. En ellas se han graficado las predicciones del modelo DPLS para 1 factor (rombos para ω_0 y círculos para ω_1) y las FDPs construidas por PLSDTHRES (Figura 2-4) y por *p*-DPLS (Figura 2-5) para las dos clases (ω_1 y ω_0).

En las predicciones del DPLS, se aprecian 2 grupos bien definidos, uno para la clase *Setosa* (entre 0.16 a 0.22) y otro las clases *Versicolor* (clase de interés) y *Virginica* (entre 0.29 a 0.55). Se esperaba ver las predicciones de *Versicolor* (círculos rojos) como un sólo grupo pero realmente se encuentran solapadas por las predicciones de *Virginica* (rombos azules) (Figura 2-4). Este solapamiento entre las clases *Versicolor* y *Virginica* no fue detectado por las funciones de decisión de clasificación con límite en 0.5 ó PLSDTHRES; dado que las funciones sólo evalúan que las predicciones estén por encima o por debajo del límite. PLSDTHRES, al considerar una distribución normal de las predicciones, construye las FDPs con el promedio y desviación estándar de éstas, generando una campana estrecha para las predicciones de la clase *Versicolor* y una campana aplastada para las predicciones de las clases *Setosa* y *Virginica* (Figura 2-4), ya que se utilizaron predicciones muy dispersas y la desviación estándar fue elevada. Otra desventaja del método PLSDTHRES es la forma como establece el límite entre clases. Éste realiza una evaluación ascendente de las respuestas de las FDPs hasta que coincidan, estableciendo dicho valor como el límite entre clases. Así, el método PLSDTHRES supone que todo valor por encima del límite pertenece a la clase ω_1 (*Versicolor*) por lo que clasificó mal el 50% de los objetos de la clase ω_0 (todas las *Virginica*). Este límite por lo tanto no es el idóneo, porque se basa en los valores de las FDPs calculadas, mas no en los valores de probabilidad que tendrían las predicciones en estas FDPs. Así pues, una buena alternativa para mejorar este método seria evaluar la probabilidad que tendrían las predicciones en las FDPs y asignarla a la clase con mayor probabilidad. Algo similar ocurre con el método por límite en 0.5, y aunque este valor podría ser optimizado para mejorar la sensibilidad, el error tipo II se elevaría haciendo necesario establecer un segundo límite, aumentando la dificultad del método.

El método *p*-DPLS soluciona las desventajas de los otros métodos. Al no suponer la distribución normal pudo construir una función con dos campanas para las predicciones de la clase ω_0 (*Setosa* y *Virginica*) destacando que están divididas en dos grupos (Figura 2-5). Además, la función de decisión depende del área delimitada por la incertidumbre bajo estas curvas (Ec. 17 apartado 2.2.2.5), asignando los objetos a la clase con mayor área y por ende más probable. Además, se reduce el error de tipo II ya que *p*-DPLS puede distinguir que aunque las predicciones de los objetos *Versicolor* y *Virginica* están superpuestas corresponden a dos clases distintas, asignando correctamente los objetos *Virginica* a la clases ω_0. Finalmente, cabe aclarar que las dificultades de clasificación observadas se deben a que los modelos del tipo "uno frente al resto" no son los más adecuados para este conjunto de datos.

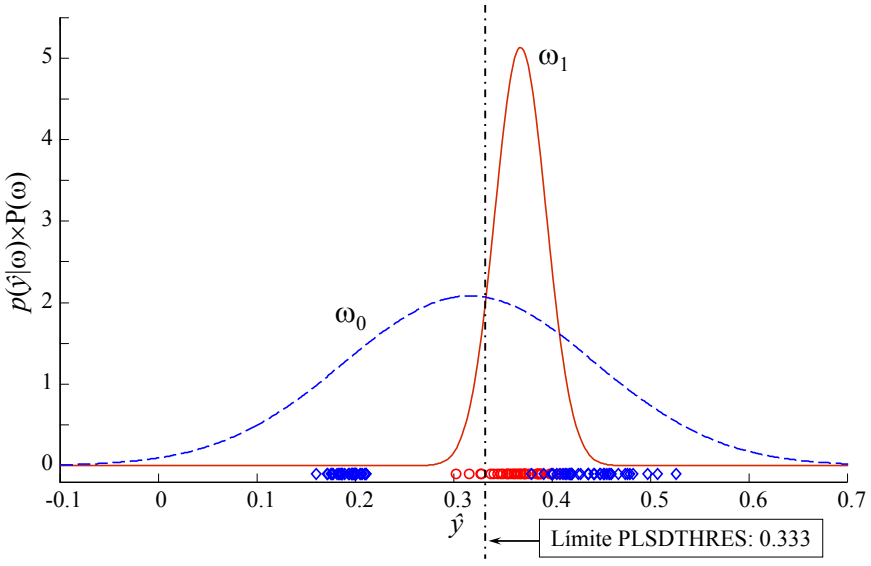

Figura 2-4: Funciones de probabilidad y límite de decisión para el método PLSDTHRES. Función para *Versicolor* (ω_1) línea sólida roja; función para *Setosa* y *Virginica* (ω_0) línea segmentada azul; predicciones para *Versicolor* círculos rojos; predicciones para *Setosa* y *Virginica* rombos azules.

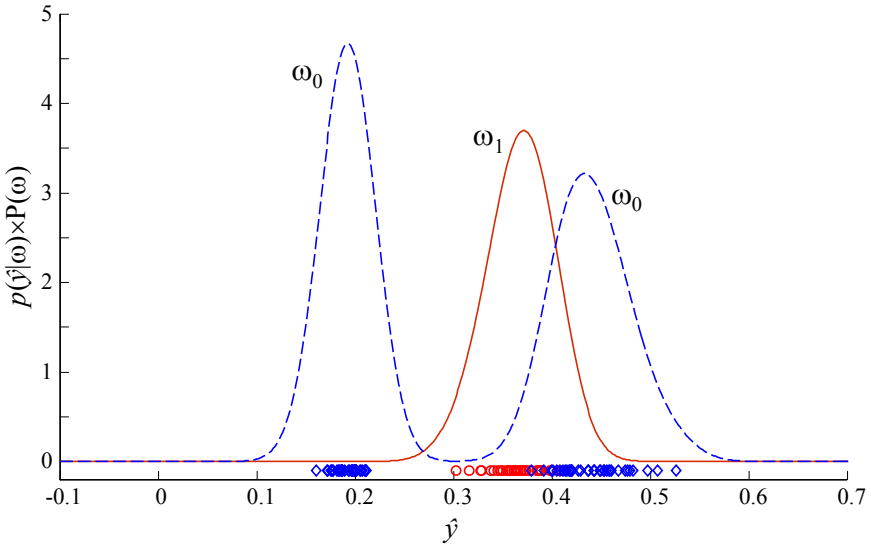

Figura 2-5: Funciones de probabilidad para el método *p*-DPLS. Función para *Versicolor* (ω_1) línea sólida roja; función para *Setosa* y *Virginica* (ω_0) línea segmentada azul; predicciones para *Versicolor* círculos rojos; predicciones para *Setosa* y *Virginica* rombos azules.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2.4.3 Conclusiones

La mayor ventaja de *p*-DPLS frente a los otros dos métodos de clasificación con DPLS estudiados es que no supone una distribución normal de las predicciones, sino que construye una función promedio que se aproxima más a la distribución de dichas predicciones, permitiéndole detectar distribuciones poco usuales de éstas. Además, al utilizar la decisión bayesiana para asignar los objetos se incrementa el porcentaje correcto de clasificación del modelo, reduciendo los errores de tipo I y II.

Referencias

1. B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Windig, R. S. Kosh, PLS_Toolbox Version 3.5 for use with MATLAB™, Eigenvector Research, Inc., Manson, WA, USA, 2005, p. 185–189.
2. N. F. Pérez, J. Ferré, R. Boqué, Chemometr. Intell. Lab. Syst. 95 (2009) 122–128.
3. R.W. Kennard, L.A. Stone. Technometrics. 11 (1969) 137–148.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2.5 Optimización de *p*-DPLS para mejorar la respuesta en errores de tipo I y II

2.5.1 Introducción

En ocasiones una clasificación debe cumplir ciertas condiciones o restricciones, como un error de tipo I o tipo II mínimo. La función de decisión de clasificación debe ser flexible y adaptarse a estos requerimientos. El método *p*-DPLS asigna el objeto a la clase con mayor probabilidad (Ec. 16 apartado 2.2.2.5); dado que este criterio incluye la incertidumbre de predicción, basa la decisión de clasificación en el área bajo la FDP asignando el objeto a la clase con mayor área (Ec. 17 apartado 2.2.2.5). Por tanto, la función de decisión de clasificación se expresa como:

$\hat{y}_u$ pertenece a la clase ω_1 si $Area_{u,1} / Area_{u,0} > r$, y viceversa para la clase ω_0 (2-15)

donde $r = 1$, si el criterio es clasificar donde el área es mayor. Sin embargo, dicho límite de decisión es flexible. Modificando la razón entre áreas a un valor menor a 1 podemos reducir el error de tipo I, a expensas de aumentar el error de tipo II. Y viceversa, si la razón es mayor a 1 se reduce el error de tipo II pero aumenta el error de tipo I. Modificar estos errores tiene aplicaciones, por ejemplo, en la autentificación de alimentos, donde un bajo error de tipo I beneficia al productor, ya que se admiten productos de menor calidad o calidad aproximada, mientras que un bajo error de tipo II protege al consumidor, ya que sólo se permiten productos que cumplan con la calidad establecida por el modelo. Para observar la respuesta de *p*-DPLS al cambio de razón entre áreas se estudió el conjunto de datos *Fisher Iris* y los modelos para *Versicolor* y *Virginica*. Al modificar la razón entre áreas a valores menores a 1, se quiere observar la respuesta de los modelos en términos de los errores de tipo I y II.

2.5.2 Resultados y discusión

Se desarrollaron los modelos *p*-DPLS *Versicolor vs. Setosa-Virginica* (ya estudiados en el apartado 2.4) y el modelo *p*-DPLS *Virginica* (ω_1, clase de interés) *vs. Setosa-Versicolor* (ω_0, clases de no interés). Se hizo validación cruzada dejando fuera un objeto cada vez (LOOCV) buscando la máxima sensibilidad (objetos de la clase de interés correctamente asignados) y por ende el minino error de tipo I.

La figura 2-6 muestra la variación de los errores de tipo I y II al modificar la razón entre áreas de las clases (ω_1 y ω_0) en el modelo *p*-DPLS *Versicolor vs. Setosa-Virginica* a diferente número de factores. Se observa que para 1 factor se obtiene un error de tipo I del 0% a una razón de 0.8 (Figura 2-6a), pero con un aumento del error de tipo II que pasa del 6.6% con una razón de 1 a 10.5% con una razón 0.8 (Figura 2-6b). Sin embargo, esta razón puede considerase óptima para mejorar el error de tipo I, ya que a razones menores el error de tipo II puede aumentar hasta el 25% sin que cambie el error de tipo I. Igualmente se puede considerar que el número óptimo de factores es 1, ya que al aumentar el número de factores se requieren razones menores a 0.4 para tener un error de tipo I menor del 5% pero con un gran coste para el error de tipo II, ya que a estas razones el error aumenta por encima del 60%.

Un comportamiento similar se observa para el modelo *p*-DPLS *Virginica vs. Setosa-Versicolor* (Figura 2-7), donde también cambia el número de factores necesarios para obtener el menor error de tipo I al cambiar la razón. Para una razón de 1 el mejor modelo es con 3 factores (error de tipo I del 18.4%), pero a una razón de 0.7 el mejor modelo es con 1 factor (error de tipo I del 2.6%). Algo similar ocurre con el error de tipo II, pero con aumento de este error; así a una razón de 0.65 con todos los factores se obtienen un error del 14.5% (Figura 2-7b), y dado que el error de tipo I sigue siendo 2.6% se puede considerar que la razón de 0.65 y 1 factor son el punto óptimo para mejorar el error de tipo I. Debe aclararse que aunque es notable la mejora en el error de tipo I (pasa de 18.4% a 2.6%), el error de tipo II empeora al pasar de 7.9% con razón de 1 a 14.5% con la razón de 0.65. En ambos modelos es notable que un intento de mejorar el error de tipo I irá acompañado de un aumento del error de tipo II; así pues, si se aplica o no esta corrección dependerá más de la utilidad que se quiera hacer del modelo y de los costes que implica reducir el error de tipo I.

2.5.3 Conclusiones

Variar la razón entre áreas a valores menores de 1, puede incidir positivamente en el error de tipo I, pero con un alto coste para el error de tipo II. Por ello la decisión de reducir uno u otro error dependerá más de la aplicación que se haga del modelo y cual de los dos errores es necesario optimizar. Se prueba así que el método *p*-DPLS es flexible ante requerimientos específicos, modificando para ello sólo la función de decisión de clasificación.

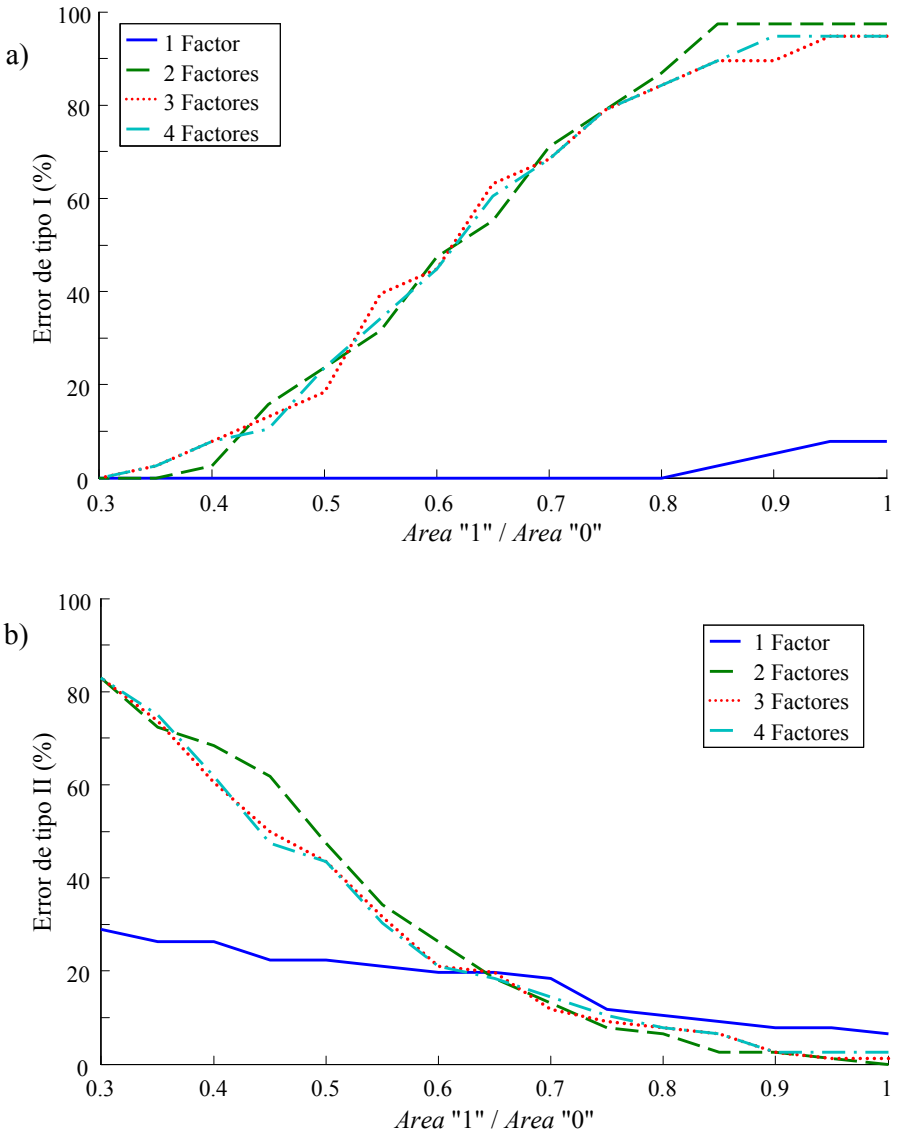

Figura 2-6: Variación del error de tipo I (a) y tipo II (b) por variación de la relación entre áreas para varios factores en el modelo *Versicolor vs. Setosa-Virginica*.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

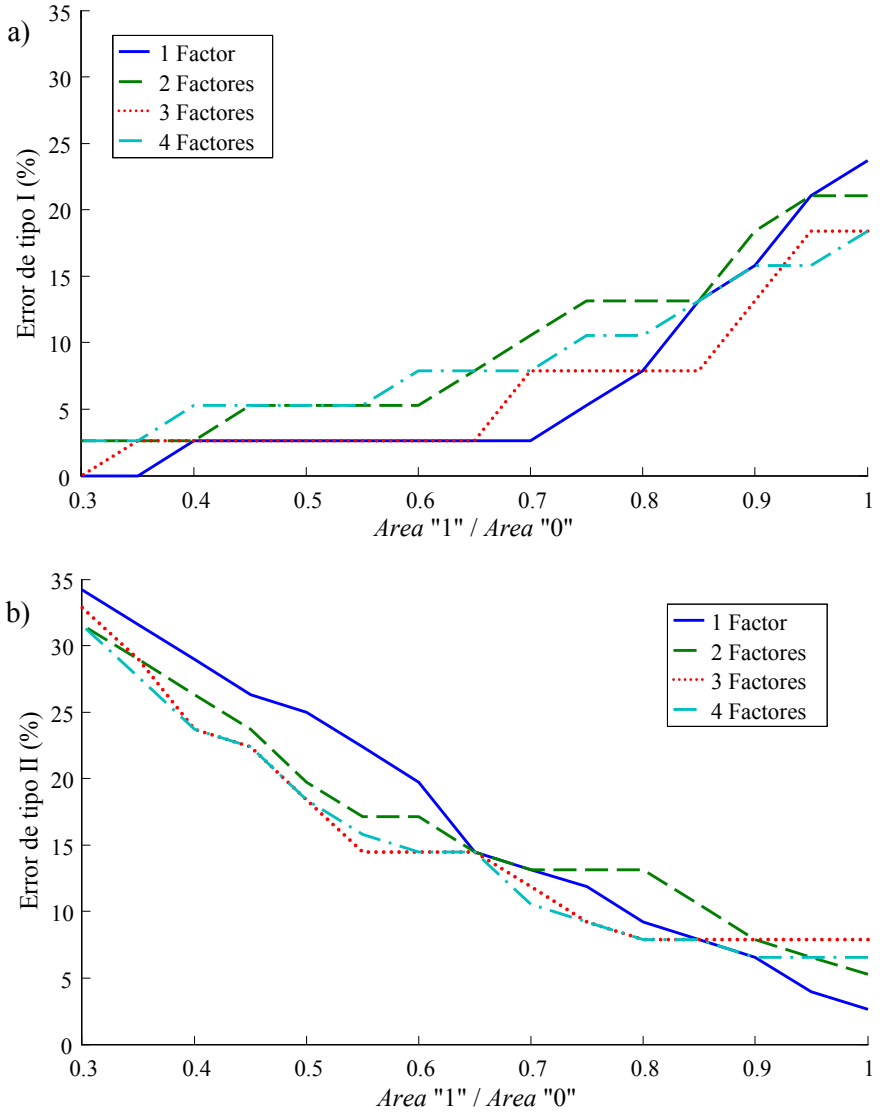

Figura 2-7: Variación del error de tipo I (a) y tipo II (b) por variación de la relación entre áreas para varios factores en el modelo *Virginica vs. Setosa-Versicolor*.

2.6 *ChemTRACE*, una interfaz en MATLAB® para *p*-DPLS

2.6.1 Introducción

Uno de lo objetivos del proyecto TRACE fue desarrollar métodos fáciles y de bajo coste para la autentificación de alimentos. El grupo de trabajo WP6 desarrolló en MATLAB® [1] la interfaz gráfica *ChemTRACE* que implementa los 5 módulos de reconocimiento de patrones desarrollados [2]. La figura 2-8 muestra la ventana principal. El módulo 1 "*Classification and Regression Trees* (CART)" fue desarrollado por S. Caetano, Y. Vander Heijden, A. Smeyers-Verbeke del Instituto de Farmacia de la Vrije Universiteit Brussel (Bruselas, Bélgica). El módulo 2 "*Soft Independent Modeling of Class Analogy* (SoftClass)" fue desarrollado por Ivana Stanimirova, Michal Daszykowski y Beata Walczak del Departamento de Quimiometría de la Universidad de Silesia (Katowice, Polonia). El módulo 3 "*Discriminant Partial Least Squares* (DPLS)" fue desarrollado por Néstor F. Pérez, Joan Ferré y Ricard Boqué del Departamento de Química Analítica y Química Orgánica de la Universidad Rovira i Virgili (Tarragona, España). El módulo 4 "*Support Vector machines* (SVM)" fue desarrollado por Bülent Üstün y Willem Melssen del Departamento de Química Analítica de Radboud University Nijmegen (Nijmegen, Holanda). Y el módulo 5 "*Kohonen/ Counter Propagation Artificial Neural Networks* (ANN)" fue desarrollado por N. Groselj y M. Novic del Instituto Nacional de Química (Liubliana, Eslovenia).

A continuación se presentan los aspectos más relevantes del módulo 3, cuyo soporte técnico se encuentra en el apartado 2.2 y algunos ejemplos partiendo del conjunto de datos de suelos europeos (detalles, anexo apartado 1) del proyecto TRACE.

2.6.2 Descripción del módulo *discriminant partial least squares* (DPLS)

El módulo *Discriminant Partial Least Squares* (DPLS) de *ChemTRACE*, se ejecuta en ambiente MATLAB 6.5 o superiores y requiere la herramienta *Statistics Toolbox* de MATLAB.

Los archivos de datos necesarios para desarrollar el modelo son la matriz de datos **X** y el vector de clases **y**. Los nombres de los objetos y los nombres de las variables se

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

pueden introducir como ficheros separados. Tanto los datos como ficheros de nombres se deben suministrar en formato ASCII.

La figura 2-9 muestra la interfaz del módulo *p*-DPLS. Los menús superiores permiten acceder a las diferentes aplicaciones y el panel *STATUS* indica el estado del análisis y advierte de operaciones no realizadas. A través de los menús se puede:

- *File*: Abrir datos o modelos, previamente analizados por el módulo *p*-DPLS, importar datos desde archivos ASCII, salvar datos y modelos y salir del módulo.
- *Task*: Abrir los paneles para calibrar o predecir con *p*-DPLS.
- *Model*: Visualizar los resultados del modelo *p*-DPLS: *scores*, *loadings*, coeficientes de regresión, predicciones, varianzas y errores.
- *Predictions*: Visualizar los resultados de predicción de objetos desconocidos.

Figura 2-8: Ventana principal de *ChemTRACE*.

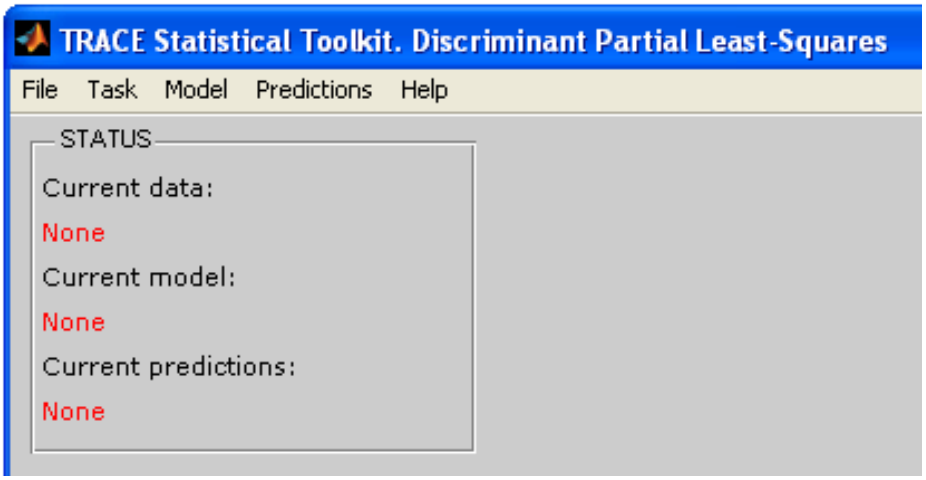

Figura 2-9: Detalle del módulo *p*-DPLS.

Una vez se han abierto o importado los datos, la interfaz muestra una gráfica de la matriz **X** y el vector **y**, cambiando el estado de los datos en el panel *STATUS* y activando los paneles de calibración y predicción (Figura 2-10). Al abrir el panel de calibración se presentan las opciones de modelado. Se puede seleccionar el pre-procesado de los datos (ninguno, centrado a la media o autoescalado), el número máximo de factores a incluir en el modelo, el método de predicción probabilística (por alturas o áreas de las funciones FDP) y por último las clases que tiene el conjunto de datos, en donde además podemos seleccionar qué clase o clases serán de interés ("1"), no interés ("0") o no se tienen en cuenta (NO) para el modelo. El método de predicción probabilística por alturas calcula el valor en la FDP de la $\hat{y}$ predicha por el modelo; mientras que la predicción probabilística por áreas calcula el área bajo la FDP en el intervalo $[\hat{y}_I, \hat{y}_D]$ ($\hat{y}_I = \hat{y} - k{\cdot}SEP$ e $\hat{y}_D = \hat{y} + k{\cdot}SEP$), para un número k de desviaciones estándar que debe ser definido por el usuario.

Establecidas las opciones de modelado se calcula y valida el modelo, cuyos resultados se pueden observar en ventanas independientes:

- Gráfica de *scores* (Figura 2-11) para diferentes combinaciones de factores y con los respectivos nombres de los objetos, para la identificación de objetos discrepantes.

- Gráficas de *loadings* (Figura 2-12) y coeficientes de regresión para diferentes factores y con el nombre de las variables.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

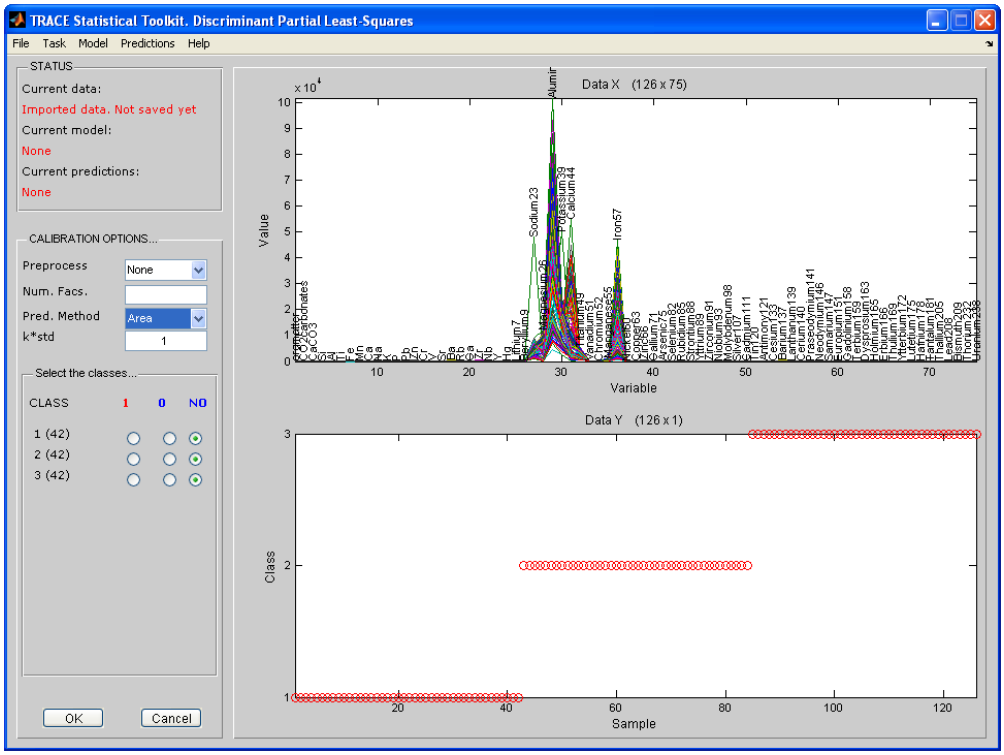

Figura 2-10: Módulo *p*-DPLS con datos y panel de calibración cargados.

- Distribución de las predicciones para cada una de las clases (Figura 2-13), de interés y de no interés. Ello permite observar la dispersión de las predicciones.

- Gráfica de "varianza y error" (Figura 2-14) que muestra la varianza explicada acumulada, el error de predicción por validación cruzada y los porcentajes de clasificación correcta con la opción seleccionada de clasificación, altura o área de la FDP. Esta última se puede usar para establecer el número óptimo de factores.

Tanto los datos de entrenamiento como el modelo obtenido se pueden guardar para usos futuros.

Una vez guardado el modelo, se importan nuevos datos desconocidos para ser clasificados. La figura 2-15 muestra la ventana con los nuevos datos. El panel de *STATUS* muestra la procedencia de los datos y el modelo que se utilizará para clasificar. En el panel de opciones de predicción se muestran los parámetros del

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

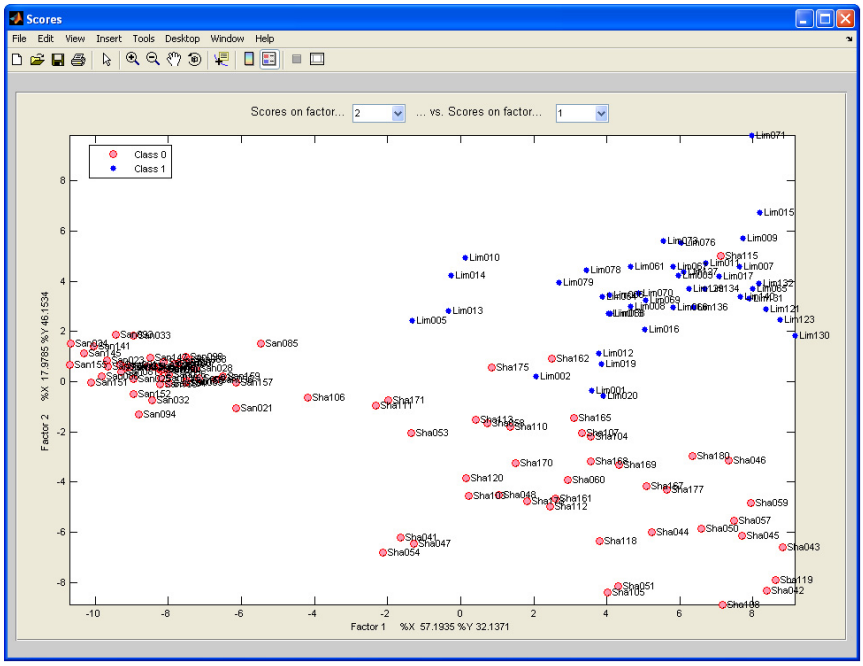

Figura 2-11: Gráfica de *scores* para los dos primeros factores con los nombres de la muestras.

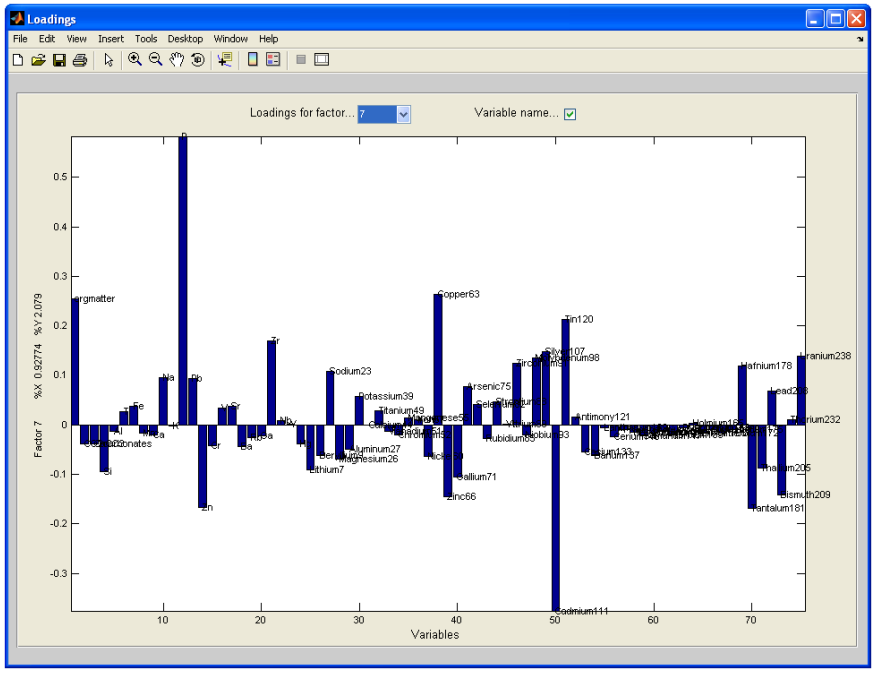

Figura 2-12: Gráfica de *loadings* para 7 factores, con los nombres de las variables.

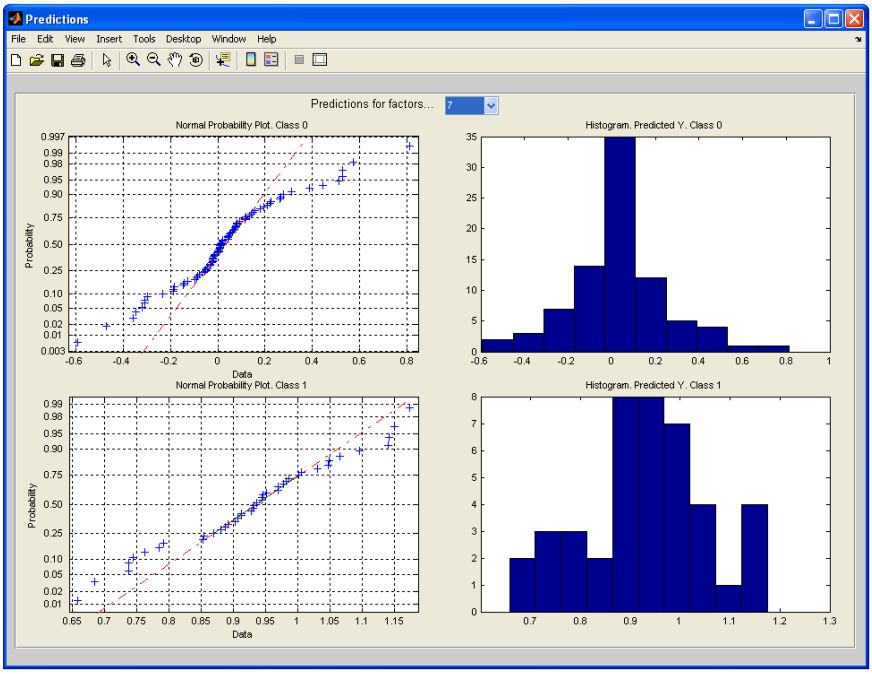

Figura 2-13: Gráficas de distribución normal para las predicciones del modelo con 7 factores.

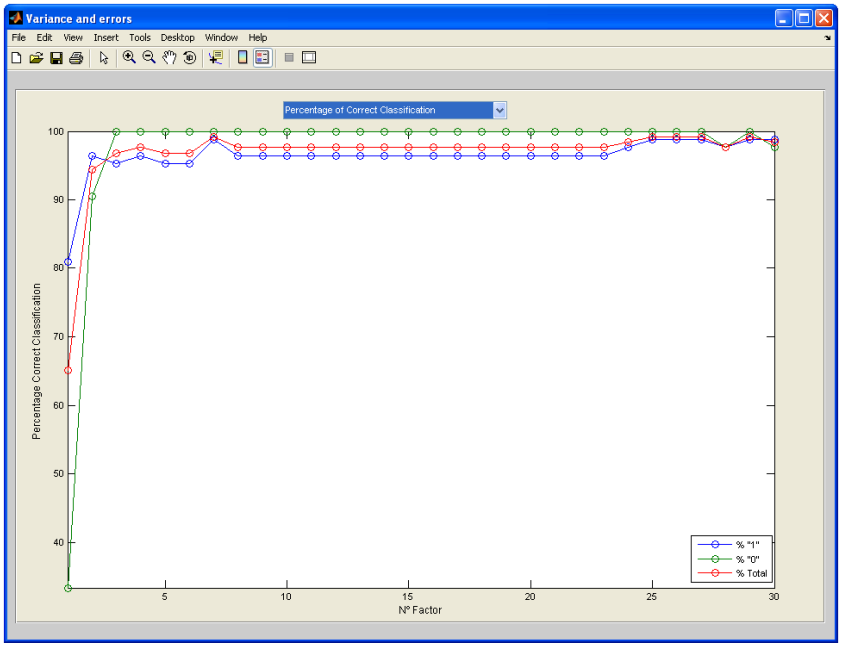

Figura 2-14: Gráfica de porcentajes correctos de clasificación por validación cruzada para los 30 factores modelados.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

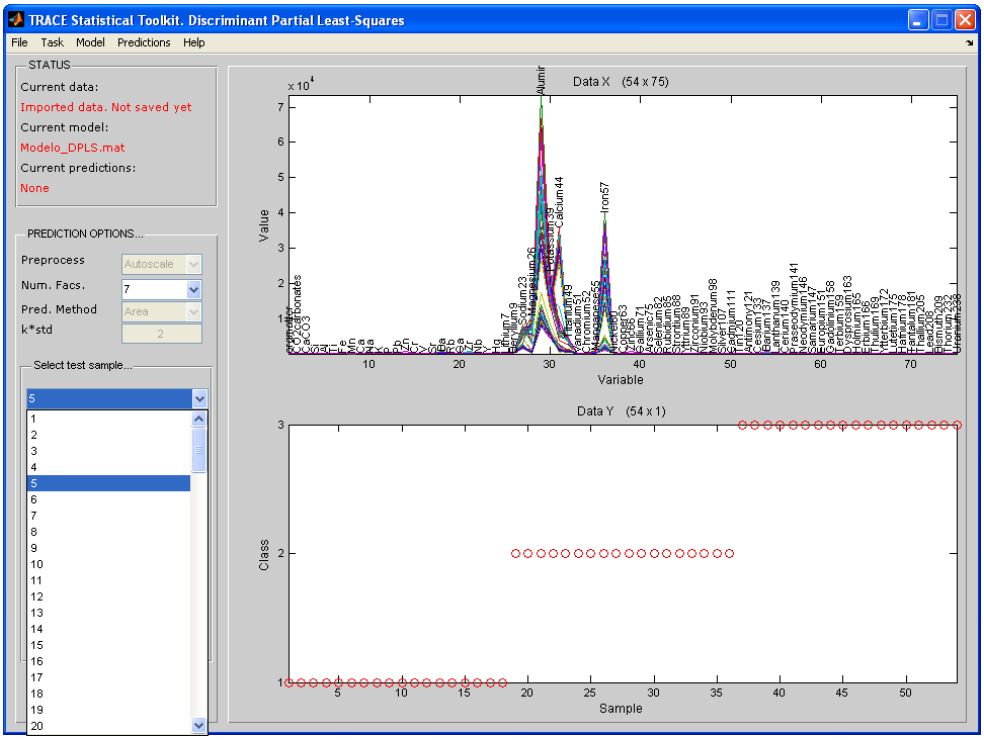

Figura 2-15: Ventana de opciones de predicción.

modelo y se puede especificar el número óptimo de factores a usar y el objeto a predecir (uno cada vez). El número óptimo de factores depende del modelo y debe ser elegido por el analista, que podrá apoyarse en las gráficas de validación cruzada o de porcentaje correcto de clasificación. Como se trata de un modelo de una clase contra las restantes (criterio establecido por TRACE), usamos como parámetro de selección para el número óptimo de factores la sensibilidad máxima, es decir, el porcentaje de clasificación más alto para la clase de interés ("1"). Cabe aclarar que, dependiendo del modelo, también se podría utilizar el mayor porcentaje de clasificación total o el mayor porcentaje de clasificación para la clase que no es de interés ("0").

Una vez hecha la predicción, el programa muestra una ventana (Figura 2-16) con las gaussianas centradas en las predicciones individuales de los datos de entrenamiento (líneas delgadas), la FDP para cada clase (líneas gruesas) y la predicción del objeto problema (línea segmentada), junto con el valor predicho, la clase a la cual se asigna y las fiabilidades para cada una de las clases.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

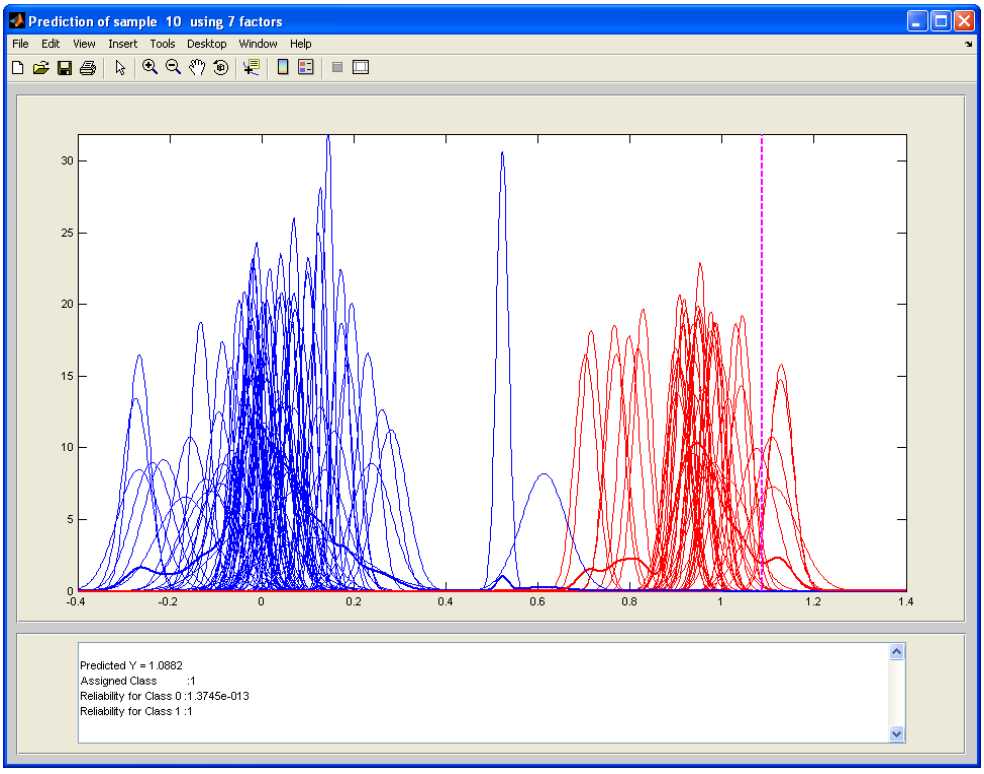

Figura 2-16: Ventana de respuesta de clasificación del método *p*-DPLS.

2.6.3 Conclusiones

Se ha desarrollado una interfaz gráfica que implementa la clasificación con *p*-DPLS. Esta interfaz permite observar la distribución de las predicciones del modelo y a su vez establecer el número óptimo de factores a partir de los criterios de altura o área. Igualmente presentan las funciones de decisión de clasificación promedio junto con las funciones individuales centradas en las predicciones, permitiendo detectar objetos con un comportamiento anómalo, y además la posición del objeto problema en el espacio de de predicciones, permitiendo entender la fiabilidad de clasificación que el modelo otorga a dicho objeto.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Referencias

1. MatLab, The MathWorks, Inc. Natwick, MA (USA), http://www.mathworks.com.
2. WP6 D6.3: Integrated Chemometrics toolbox Part 1: Users Manual (ChemTRACE) 31-12-07.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2.7 Aplicación del método de clasificación binaria *p*-DPLS a la clasificación de suelos por litologías

2.7.1 Introducción

Entiéndase como suelo: "*cuerpo natural que comprende a sólidos (minerales y materia orgánica), líquidos y gases que ocurren en la superficie de la tierra, que ocupa un espacio, y que se caracteriza por uno o ambos de los siguientes: horizontes o capas que se distinguen del material inicial como resultado de las adiciones, pérdidas, transferencias y transformaciones de energía y materia o por la habilidad de soportar plantas enraizadas en un ambiente natural*" [1]. La clasificación de suelos es compleja pues no se trata sólo de la mezcla de elementos superficiales, sino de una serie de capas que pueden llegar a los dos metros de profundidad o más. En la clasificación de suelos se tienen en cuenta tres escuelas, la climática de origen ruso; la analítica o taxonómica de la escuela norteamericana [1], y la genética que tiene en cuenta la génesis del suelo y es la oficial de la Unión Europea (UE) [2]. Esta última tiene como referencia 30 clases de suelos, basadas en características visuales y químicas. Según esta clasificación, el suelo más común en la UE es del tipo *Albeluvisols*, un suelo ácido. Fuera de la clasificación oficial se encuentran otras como la ambiental usada para clasificar cultivos mediante el ADN de la fauna microbiana [3] y para determinar el grado de contaminación con metales pesados de suelos de áreas industriales [4].

Las características del suelo influyen en la presencia de ciertos isótopos en los alimentos cultivados en éste [5]. Por tanto, la abundancia de dichos isótopos se podría utilizar para verificar la autenticidad de los alimentos [6]. Al hacer una autentificación se verifica que un alimento sea, por ejemplo, orgánico, su origen geográfico, su trazabilidad y otra información relevante que permita autentificarlo. La mayoría de los suelos de cultivo, sin importar su clasificación, tienen un componente antropogénico debido al proceso humano de labranza, fertilización, adición de materiales e irrigación, dando lugar a que en periodos relativamente largos de tiempo se origine el tipo de suelo *Anthrosols* [2]. Debido a ello, no es fácil determinar si los isótopos presentes en los alimentos son aportados por el suelo o por el material adicionado. Si dependen del material adicionado, la concentración de isótopos variaría de un ciclo a otro de cultivo. Por el contrario, si los isótopos provienen del suelo base las concentraciones de isótopos mantendrán su proporción natural.

Para verificar que el aporte isotópico proviene del suelo, se debe hallar una correlación entre los isótopos del suelo y los del alimento. Aunque es posible determinar el origen

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

geográfico de un suelo dadas las concentraciones isotópicas [7] y por ende la clase, en TRACE se ha planteado una clasificación más simple que no implique tener que identificar las 30 clases de suelos oficiales. Esta clasificación se basa en el material parental de los suelos, es decir, en las rocas constitutivas más comunes en la UE (Figura 2-17). Esta clasificación se denomina litológica y busca simplificar la correspondencia con las 30 clases de suelos, incluidas sus variaciones.

Antes de poder verificar la correspondencia isotópica entre suelo y alimentos, se debe verificar que la clasificación propuesta por litologías es posible. Para ello se utilizará el método *p*-DPLS, evaluando la sensibilidad y especificidad de clasificación para las diferentes litologías, además de observar la distribución de las predicciones y las fiabilidades.

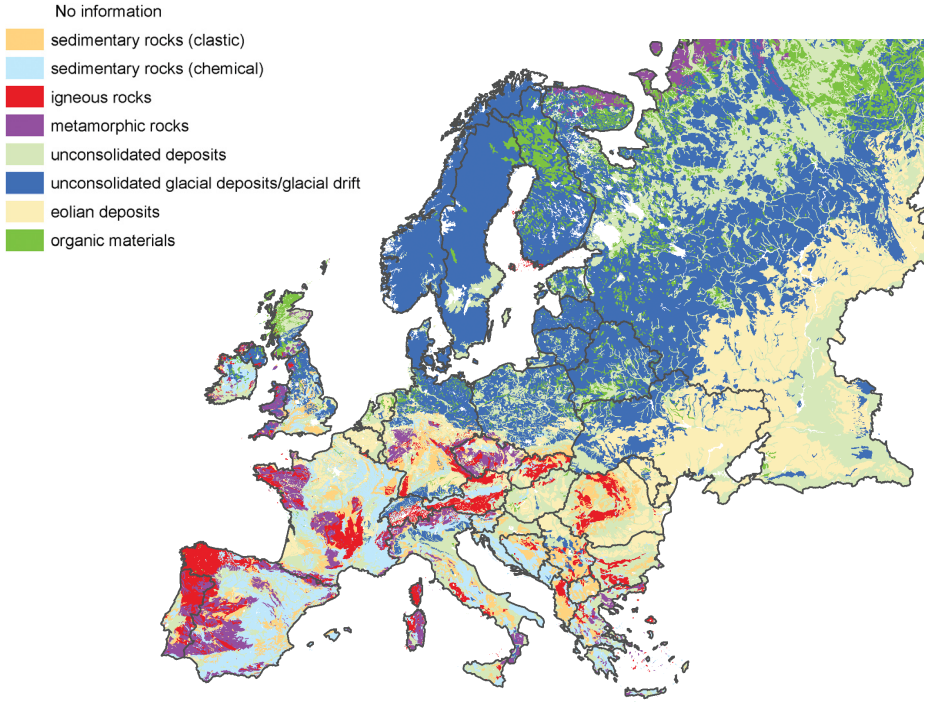

Figura 2-17: Distribución de los tipos mayoritarios de material parental en suelos a lo largo de Europa [2].

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2.7.2 Parte experimental

Para este estudio se utilizó el conjunto de datos de suelos europeos aportado por el proyecto TRACE (detalles en el anexo, apartado 1) con 180 objetos, 75 variables, concentraciones de elementos comunes y traza, y 3 clases: *caliza* (roca sedimentaria química), *arenisca* (roca sedimentaria clástica) y *esquisto* (roca metamórfica) (Figura 2-17). Se aplicó el algoritmo *Kennard-Stone* (a datos autoescalados) para dividirlo en datos de entrenamiento (70% de los datos) y datos de prueba (30% de los datos). Se desarrollaron 3 modelos del tipo "una clase contra las restantes": *caliza* (ω_1) *vs. arenisca-esquisto* (ω_0); *arenisca* (ω_1) *vs. caliza-esquisto* (ω_0) y *esquisto* (ω_1) *vs. arenisca-caliza* (ω_0). El número óptimo de factores se decidió por validación cruzada dejando fuera un objeto cada vez (LOOCV) con criterio de sensibilidad óptima (máximo porcentaje de objetos de la clase ω_1 bien asignados). Los modelos se desarrollaron con datos autoescalados y como función de decisión de clasificación se utilizó la ecuación 17 del apartado 2.2.2.5, para un intervalo incertidumbre $2\times SEP$, intervalo en el cual se calcula el área bajo las FDP de las clases.

2.7.3 Resultados y discusión

Un modelado preliminar de los datos detectó la presencia de dos objetos discrepantes, el objeto *esquisto*115, que se ubica en el espacio de la clase *caliza* (Figuras 2-20a y 2-22a), y el objeto *arenisca*085, que aunque cerca de su clase fue mal asignado por los modelos (Figuras 2-18a y 2-20a). Finalmente el conjunto de entrenamiento quedó formado por 124 objetos.

2.7.3.1 *Modelo arenisca vs. caliza-esquisto*

El número óptimo de factores para el modelo *arenisca vs. caliza-esquisto* fue 1 factor. La gráfica de *scores* del modelo (Figura 2-18a) muestra que la clase de interés *arenisca* presenta la menor dispersión, esto facilita separarla de las otras clases y contrasta con la mayor dispersión de las clases *caliza* y *esquisto*. El primer factor permite separar adecuadamente la clase *arenisca* de las otras dos clases, se obtiene cerca del 78% de

varianza explicada en *y*. La figura 2-18b muestra los coeficientes de regresión del modelo *arenisca vs. caliza-esquisto* con 1 factor. Éstos presentan un comportamiento particular, donde la mayoría de coeficientes son negativos y sólo las variables 4 (Si) y 49 (^{107}Ag) son positivas. Al revisar los valores de la variable Si, todos los objetos de la clase *arenisca* tienen un valor promedio mayor que los objetos de las clases *caliza* y *esquisto*; por contraste las variables con coeficiente negativo tienen el comportamiento opuesto, valores promedio menores para objetos de la clase *arenisca* y mayores para las clases *caliza* y *esquisto*. Es decir, que el primer factor tiene un aporte mayoritario de la variable Si. La importancia del Si para la clase *arenisca* ya se observaba en el análisis por PCA (anexo, apartado 1), donde según la gráfica de *loadings* se podía considerar al Si como único responsable de separar la clase *arenisca* de las restantes.

La habilidad de clasificación del modelo *arenisca vs. caliza-esquisto* con un factor es la mejor de los tres posibles modelos del conjunto de datos de suelos. Para los objetos de entrenamiento por LOOCV se obtuvo una sensibilidad y especificidad del 100%. Que fuera necesario un sólo factor para este modelo está supeditado a la poca dispersión que presentan los objetos de *arenisca* y a que este factor separa adecuadamente los objetos de esta clase de las restantes. Con los datos de prueba también se obtuvo una sensibilidad y especificidad del 100%.

Al calcular la fiabilidad de clasificación se debe tener en cuenta la dispersión de los objetos predichos y el ***SEP***, que determinarán cuan amplia o estrecha es la FDP. Para el modelo *arenisca vs. caliza-esquisto* con un factor, las predicciones tienen una cierta dispersión (Figura 2-19), mayor en las clases que no son de interés. Ello genera funciones de probabilidad promedio que se aproximan a la distribución normal, pero con curvas amplias asimétricas por la presencia de datos dispersos, además de estar solapadas (Figura 2-19b). Estas funciones amplias ocasionan que predicciones centradas en las clases, cerca del valor binario asignado, tengan valores de fiabilidad bajos (Figura 2-19a), por ejemplo la *arenisca*094, clase ω_1, con una predicción de 0.78 tiene una fiabilidad del 80.8%. Además, estas funciones son menos sensibles a cambios en las predicciones cuando se alejan del centro de la clase y se aproximan al límite entre clases, ya que las áreas que se tienen en cuenta para el cálculo de fiabilidad son similares. Ese es el caso de los objetos *arenisca*021 con una predicción de 0.65 y *esquisto*106 con una predicción de 0.56, que pertenecen a las clases ω_1 y ω_0, respectivamente. La *arenisca*021 tiene una fiabilidad del 63.2% de pertenecer a la clase ω_1, mientras que la *esquisto*106 tiene una fiabilidad de pertenecer a la clase ω_1 de 48.6% o lo que es lo mismo un 51.4% de pertenecer a ω_0. Esto además coincide con la proximidad que el objeto *esquisto*106 tiene con los objetos de la clase *arenisca* en la gráfica de *scores* (Figura 2-18a).

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Los datos de prueba tienen predicciones centradas en sus respectivas funciones de clase (Figura 2-19 b y c), así que las fiabilidades de los objetos *arenisca* de pertenecer a la clase ω_1 están entre un 82% a 95%, mientras que los objetos de las otras dos clases tienen fiabilidades de pertenecer a la clases ω_0 de 86% al 100%. Por lo anterior se puede considerar que el modelo *arenisca vs. caliza-esquito* es apropiado para asignar objetos nuevos a la clase *arenisca*, además de otorgar una adecuada fiabilidad de clasificación dadas las características tan definidas de esta clase y la dispersión que presentan las predicciones.

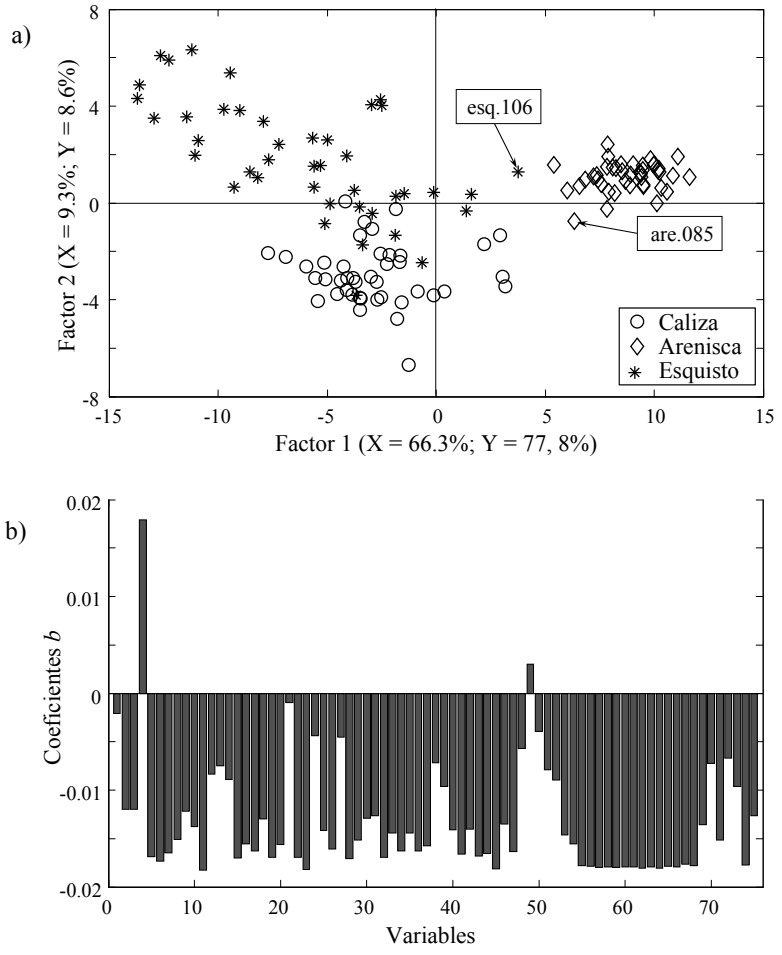

Figura 2-18: Gráficas de *scores* para los dos primeros factores (a) y coeficientes de regresión *b* con un factor (b) del modelo *arenisca vs. caliza-esquisto*.

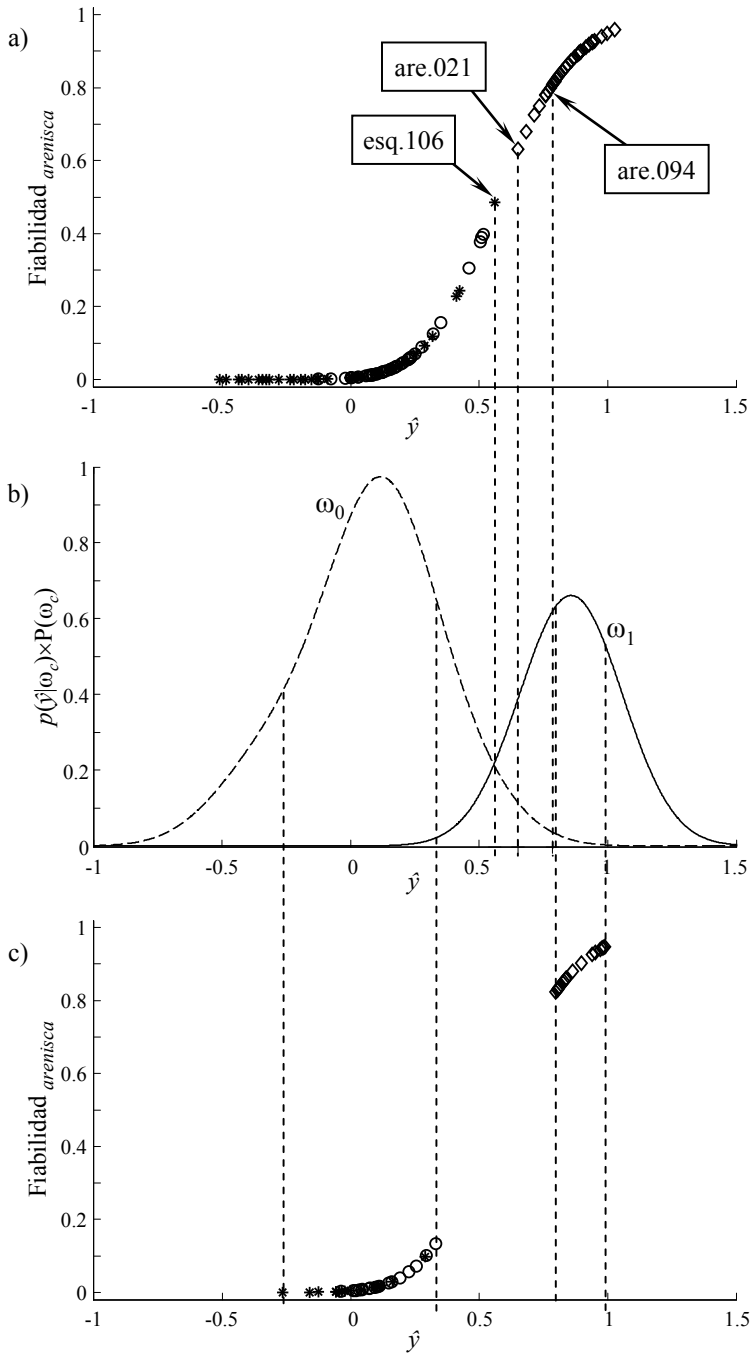

Figura 2-19: Distribución de las predicciones del modelo *arenisca vs. caliza-esquisto* con un factor para los conjuntos de entrenamiento (a) y prueba (c), y FDP para las clases *arenisca* (ω_1) y *caliza-esquisto* (ω_0) (b). *arenisca* (◊), *caliza* (○) y *esquisto* (*).

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2.7.3.2 *Modelo caliza vs. arenisca-esquisto*

El número óptimo de factores para el modelo *caliza vs. arenisca-esquisto* fue de 5. La gráfica de *scores* del modelo *caliza vs. arenisca-esquisto* (Figura 2-20a) muestra una cierta dispersión para los objetos de la clase *caliza* que se solapan con los también dispersos objetos de la clase *esquito*, haciéndolas difíciles de separar y necesitando un mayor número de factores. El primer factor permite separar adecuadamente la clase *arenisca* de las otras dos clases, algo que contrasta con el hecho de que la *arenisca* está agrupada con la clase *esquisto*, mientras que el segundo factor separa parcialmente las clases *caliza* y *esquisto*. Por ello se requieren 5 factores para discriminar la clase *caliza* de las restantes.

La figura 2-20b muestra los coeficientes de regresión del modelo *caliza vs. arenisca-esquisto*. A diferencia de lo observado en el modelo *arenisca vs. caliza-esquisto*, el modelo *caliza vs. arenisca-esquisto* presenta un comportamiento más común, con los coeficientes distribuidos indistintamente hacia valores negativos y positivos; pudiéndose distinguir cerca de 15 variables significativas, dentro de las que se incluyen carbonatos (2), $CaCO_3$ (3), Mn (8), Ca (9) y Na (10), elementos típicos de la *caliza*, es decir, que se observa una correspondencia entre las características químicas de la clase y el modelo.

La habilidad de clasificación del modelo *caliza vs. arenisca-esquisto* con 5 factores es levemente menor a la observada en el modelo *arenisca vs. caliza-esquisto*. Así, con datos de entrenamiento por LOOCV, se obtiene una sensibilidad del 100% y una especificidad del 98.8%, debido a que el objeto *esquisto*162 es mal asignado como *caliza*. Sin embargo, si se considera otro criterio como la clasificación total óptima (que maximiza tanto la sensibilidad como la especificidad) se requieren 8 factores, obteniendo una sensibilidad y especificidad del 100%. Además, un mayor número de factores mejora la predicción y reduce el *SEP* del modelo, que es significativo para la metodología *p*-DPLS, ya que valores muy elevados de éstos hacen que las áreas bajo las FDP de las clases sean similares y por tanto la fiabilidad sea muy cercana ha 50%. Por el contrario, cuando el *SEP* es pequeño las áreas bajo las FDP de clase son diferentes y permiten obtener mejores valores de fiabilidad. Con los datos de prueba se obtuvo una sensibilidad y especificidad del 100%.

Al calcular las fiabilidades para el modelo *caliza vs. arenisca-esquisto* con 5 factores se tuvieron predicciones con un menor grado de dispersión (Figura 2-21). Sin embargo, también se observaron algunos objetos que podrían ser considerados discrepantes dado que su predicción es elevada o son mal asignados. Es el caso del objeto *caliza*071 con una predicción de 1.66, que se puede considerar fuera de la campana de la clase (ω_1); o del objeto *esquisto*162 cuya predicción de 0.59 está lejos de la clase de no interés (ω_0) pero cerca del límite de la clase de interés (ω_1), siendo mal asignadas a ésta última.

Aun así, los datos anteriormente mencionados se incluyeron en el modelo *p*-DPLS y se obtuvieron FDP poco solapadas (Figura 2-21b), que permiten una buena separación de las clases.

Estas FDP permiten que las predicciones centradas en las clases, cerca del valor binario asignado, tenga valores de fiabilidad elevados (Figura 2-21a). Por el contrario, cuando las predicciones se alejan del centro de la clase y se aproximan al límite entre clases la fiabilidad desciende, siendo este descenso más crítico cuan más cerca se esté del límite, ya que la áreas que se tienen en cuenta para el cálculo de fiabilidad son pequeñas,

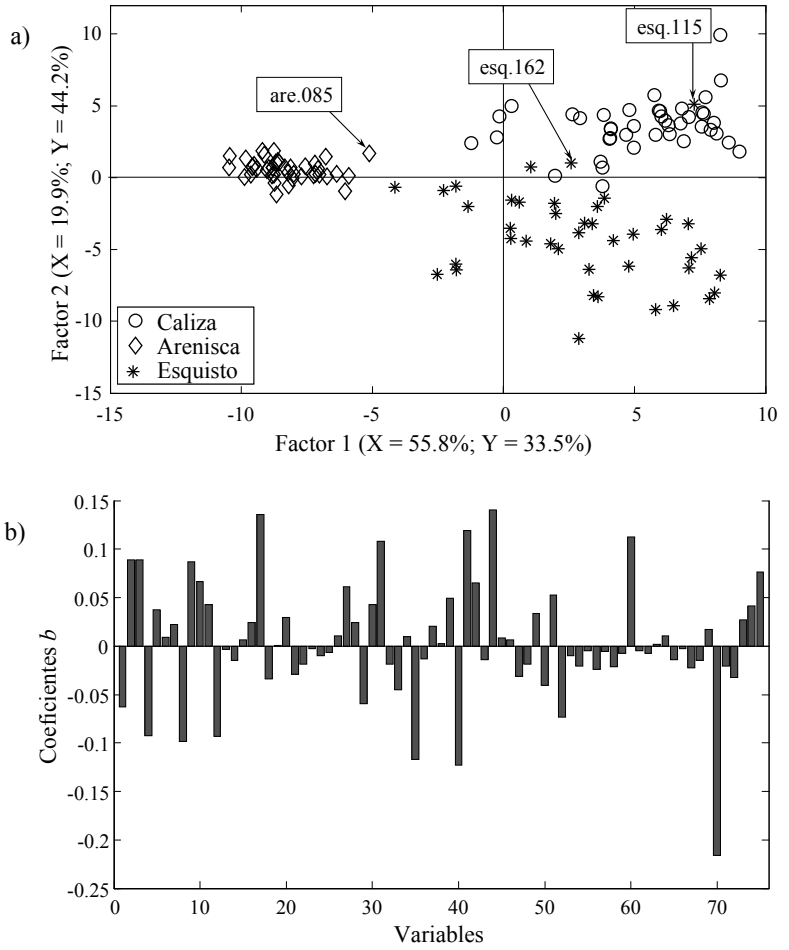

Figura 2-20: Gráficas de *scores* para los dos primeros factores (a) y coeficientes de regresión *b* con 5 factores (b) del modelo *caliza vs. arenisca-esquisto.*

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

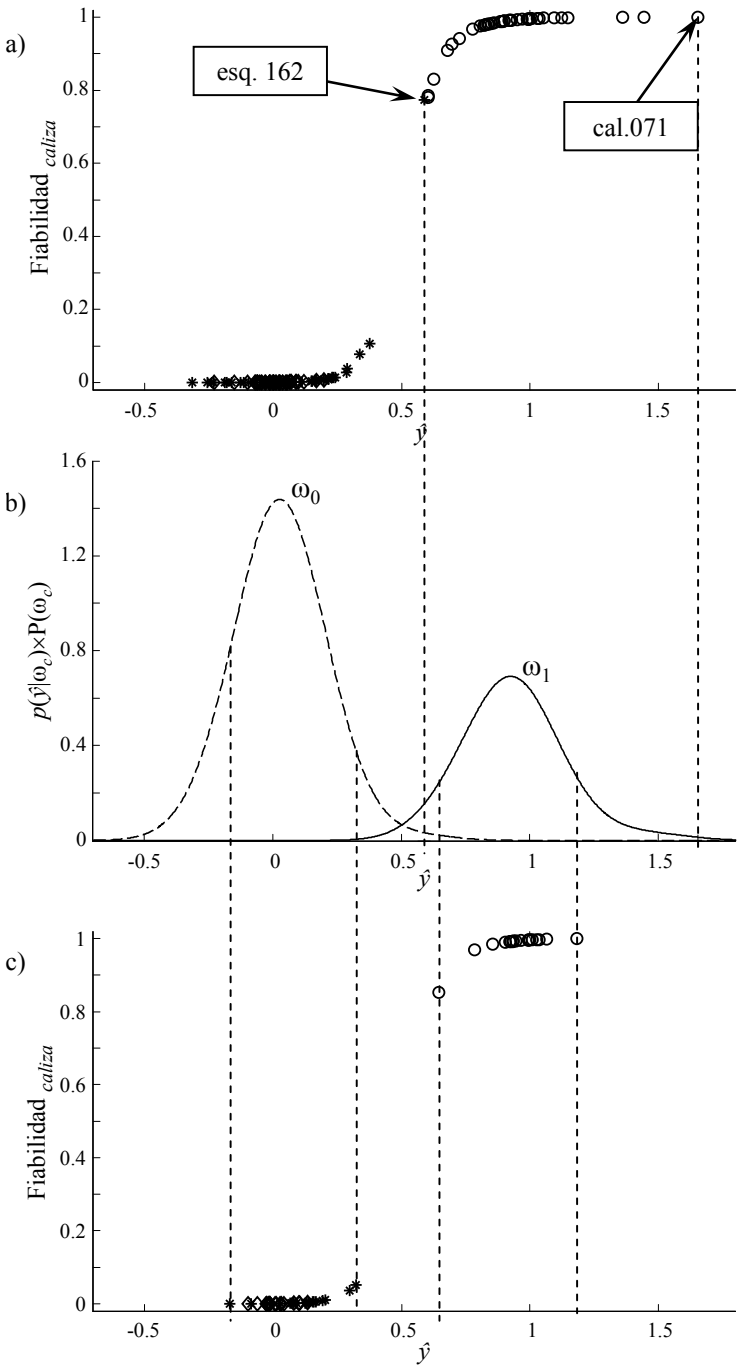

Figura 2-21: Distribución de las predicciones del modelo *caliza vs. arenisca-esquisto* con 5 factores para los conjuntos de entrenamiento (a) y prueba (c), y FDP para las clases *caliza* (ω_1) y *arenisca-esquisto* (ω_0) (b). *arenisca* (◊), *caliza* (○) y *esquisto* (*).

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

haciendo el cálculo más sensible a pequeños cambios en la predicción. Por ejemplo, los objetos *caliza*019, con una predicción de 0.68, y *caliza*005, con una predicción de 0.60, tienen fiabilidades de pertenecer a la clase *caliza* del 90.9% y 78.5%, respectivamente. Es decir, que con sólo una variación en la predicción de 0.08 se tiene una variación de fiabilidad del 12%. Cabe mencionar que las fiabilidades para los objetos *caliza* de entrenamiento calculadas con el modelo de 5 factores tienen en promedio un 97% contra un 100% del modelo con 8 factores. Igualmente el objeto *esquisto*162 tiene una fiabilidad con 5 factores de 77.4% de pertenecer a la clase ω_1, y pasa al 31.4% en el modelo con 8 factores, por lo que es asignado correctamente a la clase ω_0. Los datos de prueba tienen un comportamiento similar al observado con los de entrenamiento (Figura 2-21c). Las predicciones para las dos clases presentan una dispersión asimétrica (Figura 2-21 b y c). Así pues, el modelo *caliza vs. arenisca-esquisto* es el adecuado para la asignación de objetos nuevos a la clase *caliza*.

2.7.3.3 *Modelo esquisto vs. arenisca-caliza*

El número óptimo de factores para el modelo *esquisto vs. arenisca-caliza* fue de 5. Este número de factores lo determina la sensibilidad, si se tuviera en cuenta la especificidad se requerirían 2 factores. Las gráficas de *scores* del modelo *esquisto vs. arenisca-caliza* (Figura 2-22a) muestra una distribución similar a la observada en el modelo *caliza vs. arenisca-esquisto*, con los objetos *esquisto* y *caliza* dispersos y ligeramente solapados, aumentando la dificultad para separarlos y necesitando un mayor número de factores. El primer factor permite separar parcialmente la clase *arenisca* de las otras dos clases y el segundo factor separa parcialmente las clases *caliza* y *esquisto*, acumulando entre los dos factores el 73% de varianza explicada en *y*. La figura 2-22b muestra los coeficientes de regresión del modelo *esquisto vs. arenisca-caliza*, que presentan un comportamiento similar al observado en el modelo *caliza vs. arenisca-esquisto*, compartiendo con este algunas variables significativas pero con signos opuestos (Figuras 2-20b y 2-22b) y coincidiendo con la distribución de los objetos de las clases *caliza* y *esquisto* que se observa en las gráficas de *scores* (Figuras 2-20a y 2-22a), en donde dichos objetos se encuentran en cuadrantes opuestos.

La habilidad de clasificación del modelo *esquito vs. arenisca-caliza* con 5 factores es menor a la observada en los modelos *arenisca vs. caliza-esquisto* y *caliza vs. arenisca-esquisto*. Así, para los datos de entrenamiento por LOOCV se obtiene una sensibilidad del 95.1%, dado que los objetos *esquito*106 y *esquito*162 son mal asignados, y una especificidad del

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

100%. Los 5 factores maximizan tanto la sensibilidad como la especificidad por lo que no es posible mejorar la habilidad de clasificación aumentando el número de factores. Con los datos de prueba se obtiene una sensibilidad y especificidad del 100%.

Al calcular las fiabilidades del modelo *esquisto vs. caliza-arenisca*, éstas presentan un comportamiento similar al observado en el modelo *caliza vs. arenisca-esquisto* (Figura 2-23b) con un leve solapamiento entre las FDPs de las clases. Se observa otro comportamiento que debe considerase único de la metodología *p*-DPLS, y es que

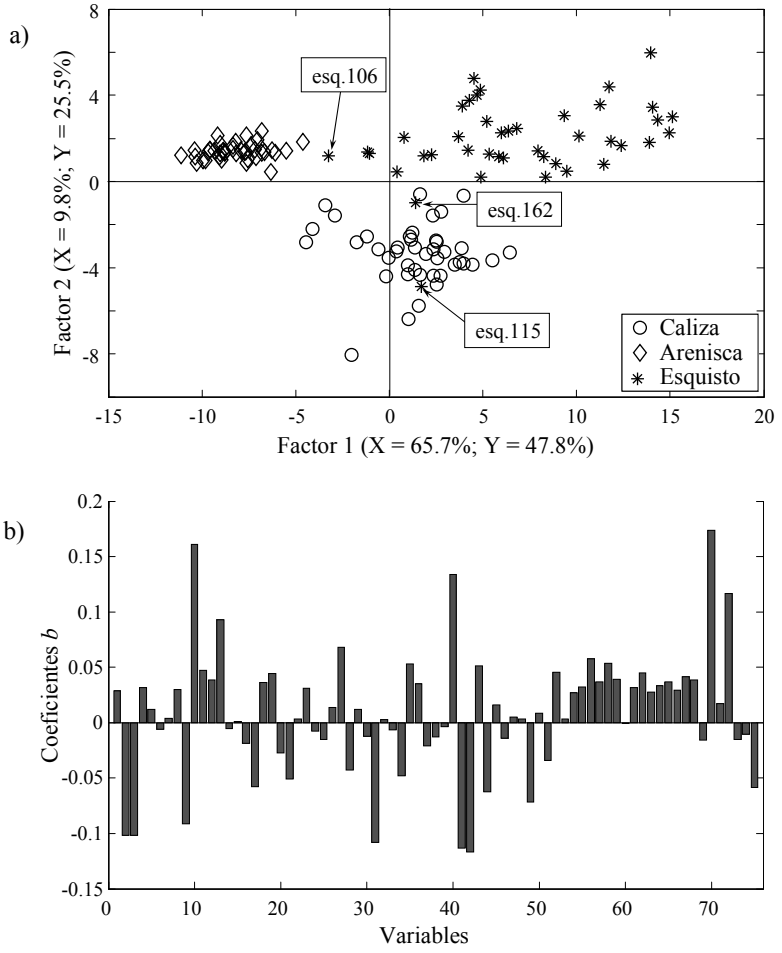

Figura 2-22: Gráficas de *scores* para los dos primeros factores (a) y coeficientes de regresión *b* con 5 factores (b) del modelo *esquisto vs. caliza-arenisca.*

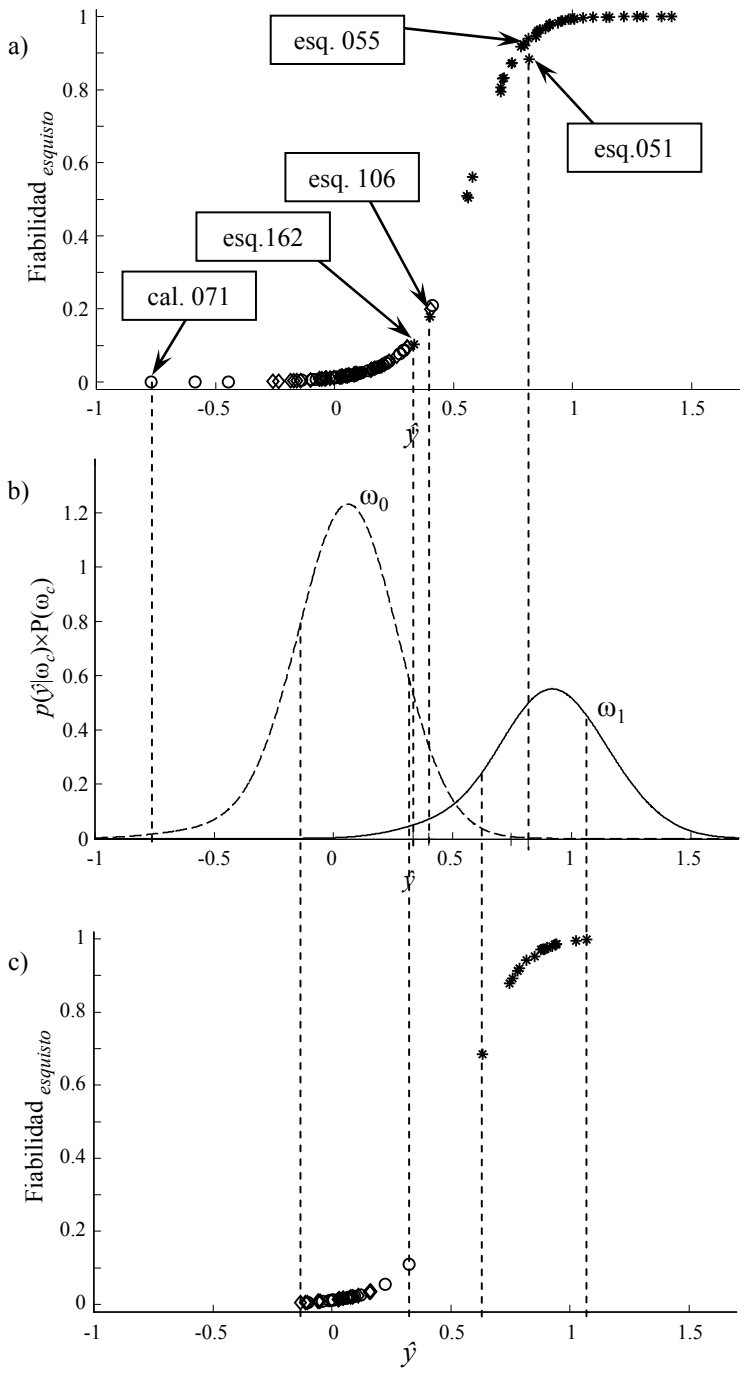

Figura 2-23: Distribución de las predicciones del modelo *esquisto vs. caliza-arenisca* con 5 factores para los conjuntos de entrenamiento (a) y prueba (c), y FDP para las clases *esquisto* (ω_1) y *caliza-arenisca* (ω_0) (b). *arenisca* (◊), *caliza* (○) y *esquisto* (*).

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

predicciones con valores similares pueden tener fiabilidades distintas. Éste es el caso de los objetos *esquisto*051 y *esquisto*055 con una predicción de 0.82 que tienen fiabilidades de pertenecer a la clase *esquisto* de 88.3% y 94.1% respectivamente. Esta variación se debe a la incertidumbre de predicción. Para la *esquisto*051 se tiene un ***SEP*** = 0.21 mientras que para *esquisto*055 el ***SEP*** = 0.17, es decir, a un menor ***SEP*** se tienen una mayor fiabilidad en la asignación, pues la relación entre áreas es más favorable a la clase *esquisto*. Nuevamente el objeto *caliza*071 presenta un valor extremo, en este caso negativo, que lo hace sospechoso pero dado el número de objetos (42) y que otros dos objeto *caliza* están cerca puede pensarse que sea parte del comportamiento de la clase *caliza*. Igualmente, los objetos *esquisto*162 y *esquisto*106 son erróneamente rechazados como *esquisto*, debido a su proximidad a la clase *caliza* (Figura 2-22a) y a la mayor dispersión de la predicciones en el límite de clases. Las predicciones para el conjunto de prueba mantiene el grado de dispersión observado en los datos de entrenamiento (Figura 2-23c). Así, el modelo *esquisto vs. caliza-arenisca* se considera adecuado para la predicción de objetos nuevos a la clase *esquisto*, además de mostrar nuevamente la importancia de incluir el error de predicción dentro de la asignación, ya que es determinante al calcular la fiabilidad de clasificación; si el error es elevado la fiabilidad será menor.

2.7.4 Conclusiones

Es posible clasificar suelos por las litologías *arenisca*, *caliza* y *esquisto* utilizando el método *p*-DPLS. Se obtienen altas sensibilidades para las clases de interés, superando el 95% para los datos de entrenamiento y el 100% para los datos de prueba. Las altas sensibilidades se deben, en primer lugar, a la óptima separación la clase *arenisca* de las restantes, y en segundo lugar, al mínimo solapamiento entre las FDP de las clases, que a su vez permiten calcular adecuadamente la fiabilidad de los objetos, aunque estos se encuentren cerca del límite entre clases. Igualmente fue posible observar la incidencia del número de factores en el cálculo de fiabilidad, que a su vez está relacionado con el grado de dispersión de la predicción y el valor del ***SEP***. Se observa la ventaja que tiene el método *p*-DPLS al utilizar el ***SEP*** como parte del criterio para calcular la fiabilidad de clasificación, si aumenta el error se reduce la fiabilidad, siendo este cálculo critico para objetos cercanos al límite entre clases.

Referencias

1. Departamento de Agricultura de los Estados Unidos, Servicio de Conservación de Recursos Naturales, "Claves para la Taxonomía de Suelos", Décima Edición, 2006.
2. European Soil Bureau Network, European Commission, "Soil Atlas of Europe", Office for Official Publications of the European Communities, Luxembourg, 2005.
3. Z. Ramadan, X.-H. Song, P. K. Hopke, M. J. Johnson, K. M. Scow, Anal. Chim. Acta. 446 (2001) 233–244.
4. L. Slavković, B. Škrbić, N. Miljević, A. Onjia, Environ. Chem. Lett. 2 (2004) 105–108.
5. F. Camin, L. Bontempo, K. Heinrich, M. Horacek, S. D. Kelly, C. Schlicht, F. Thomas, F. J. Monahan, J. Hoogewerff, A. Rossmann, Anal. Bioanal. Chem. 389 (2007) 309–320.
6. S. Kelly, K. Heaton, J. Hoogewerff, Trends Food Sci. Tech. 16 (2005) 555–567.
7. S. Dragovic, A. Onjia, Appl. Radiat. Isot. 65 (2007) 218–224.

2.8 Aplicación del método *p*-DPLS a la clasificación de mieles de origen geográfico *Córcega*

2.8.1 Introducción

La directiva 2001/110/CE define la miel como: "*sustancia natural dulce producida por la abeja Apis mellifera a partir del néctar de plantas o de secreciones de partes vivas de plantas o de excreciones de insectos chupadores presentes en las partes vivas de plantas, que las abejas recolectan, transforman combinándolas con sustancias específicas propias, depositan, deshidratan, almacenan y dejan en colmenas para que maduré*" [1]. La mayor parte de su peso está compuesto por carbohidratos de glucosa y fructuosa en aproximadamente la misma proporción [2] y el restante es agua y otras sustancias menores, como polen y proteínas [3]. La miel es un importante producto comercial, no sólo como sustituto natural del azúcar, sino por sus propiedades medicinales y terapéuticas, sirviendo como agente antibacterial, prebiótico, antioxidante y anti-mutagénico [4]; llegando ha ser autorizado como medicamento (*Medihoney*^{TM}) por Europa y Australia para el cuidado de las heridas [5].

Dichas características hacen necesario proteger al consumidor sobre la autenticidad y origen geográfico y botánico de la miel. Por ello, se han desarrollado métodos para detectar las adulteraciones por adición de azúcares, como jarabe de maíz [2], o la adición de mieles de otras procedencias o características menores [6]. Dentro de los análisis más utilizados para autentificar mieles se encuentran la detección de compuestos volátiles y el análisis de aminoácidos, para detectar el origen geográfico; y el estudio del polen, para detectar el origen botánico. Otras técnicas han sido la espectroscopia NIR con PLS para confirmar denominación de origen [6], la determinación de minerales junto con el método del vecino más cercano (*k-Nearest Neighbours*, k-NN) para determinar origen geográfico [7], o la espectroscopia de fluorescencia con PCA y análisis discriminante factorial (*Factorial Discriminant Analysis*, FDA) para reconocimiento del origen botánico [3].

El proyecto TRACE tuvo, como uno de sus objetivos, aportar métodos que permitieran confirmar la autenticidad de alimentos. En el caso de mieles se planteó la clasificación por origen geográfico a partir de datos de concentraciones de elementos comunes y elementos traza. Así, se utilizó *p*-DPLS para evaluar la posibilidad de clasificar las mieles de diferentes áreas geográficas de la Unión Europea.

2.8.2 Parte experimental

Para este estudio se utilizó el conjunto de mieles europeas proveniente del proyecto TRACE (detalles en el anexo, apartado 2) para clasificación por áreas geográficas, con 180 objetos, 26 variables de intensidad absoluta de masas de componentes por CG y 9 clases: *Córcega*, *Sicilia*, *Toscana*, *Trentino*, *Marchfeld*, *Muehlviertel*, *Carpentras*, *Limousin* y *Bavaria*. Se aplicó el algoritmo *Kennard-Stone* (a datos autoescalados) para dividirlo en datos de entrenamiento (70% de los datos) y datos de prueba (30% de los datos). Sin embargo, por requerimientos de TRACE únicamente se desarrolló el modelo "*Córcega* frente a las restantes". El número óptimo de factores se decidió por validación cruzada dejando fuera un objeto cada vez (LOOCV), con criterio de máxima sensibilidad (porcentaje de objetos de la clase ω_1 bien asignados). Como función de decisión de clasificación se utilizó la ecuación 17 del apartado 2.2.2.5, para un intervalo incertidumbre $2\times SEP$, intervalo en el cual se calcula el área bajo las FDP de las clases.

2.8.3 Resultados y discusión

Cuando se desarrolló el modelo "*Córcega* frente a las restantes" se observó, en las gráficas de *scores*, que los objetos de la clase *Carpentras* se mantenían separados de los restantes, por lo que se decidió excluir la clase del modelo. Adicionalmente dos objetos de la clase *Córcega* y uno de la *Sicilia* tenían comportamiento discrepante y fueron eliminados. Así, el conjunto de modelado terminó con 121 objetos. Que una clase tenga un comportamiento discrepante, como sucedió con la clase *Carpentras*, es común cuando se reúnen varias clases en una sola, pues se están combinando clases que pueden ser opuestas. Esta situación es uno de los inconvenientes de la estrategia de modelado de "clase de interés frente a las restantes".

La máxima sensibilidad para el modelo *p*-DPLS se obtuvo con 5 factores. Los datos de entrenamiento por LOOCV obtuvieron una sensibilidad del 97.4% (los objetos FRCOR033 y FRCOR036 fueron considerados como no *Córcega*) y una especificidad del 82.2% (8 objetos erróneamente considerados como *Córcega*), indicando un error de tipo II elevado que irá en detrimento de la clase *Córcega*, ya que se están aceptando mieles de las otras denominaciones. Con los datos de prueba se obtuvo una sensibilidad de 93.9% (los objetos FRCOR004 y FRCOR030 fueron considerado como no *Córcega*) y una especificidad del 93.8% (el objeto ITSIC118 fue considerado erróneamente como *Córcega*). La menor especificidad en el conjunto de entrenamiento

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

se puede atribuir al desbalance de datos, ya que de los 173 objetos del modelo aproximadamente el 64% corresponde a *Córcega* y, dado que los restantes objetos están distribuidos en 7 clases distintas tienen una gran variabilidad que no les permite converger en una sola clase. En la gráfica de *scores* (Figura 2-24a) se observa que con el primer factor ya se obtiene una separación parcial de los objetos de la clase *Córcega* del resto. La separación mejora si se utilizan 2 factores. Los coeficientes de regresión (Figura 2-24b) muestran que las variables 2, 4, 10, 12, 15, 16 y 21 son las que principalmente caracterizan a la clase *Córcega*.

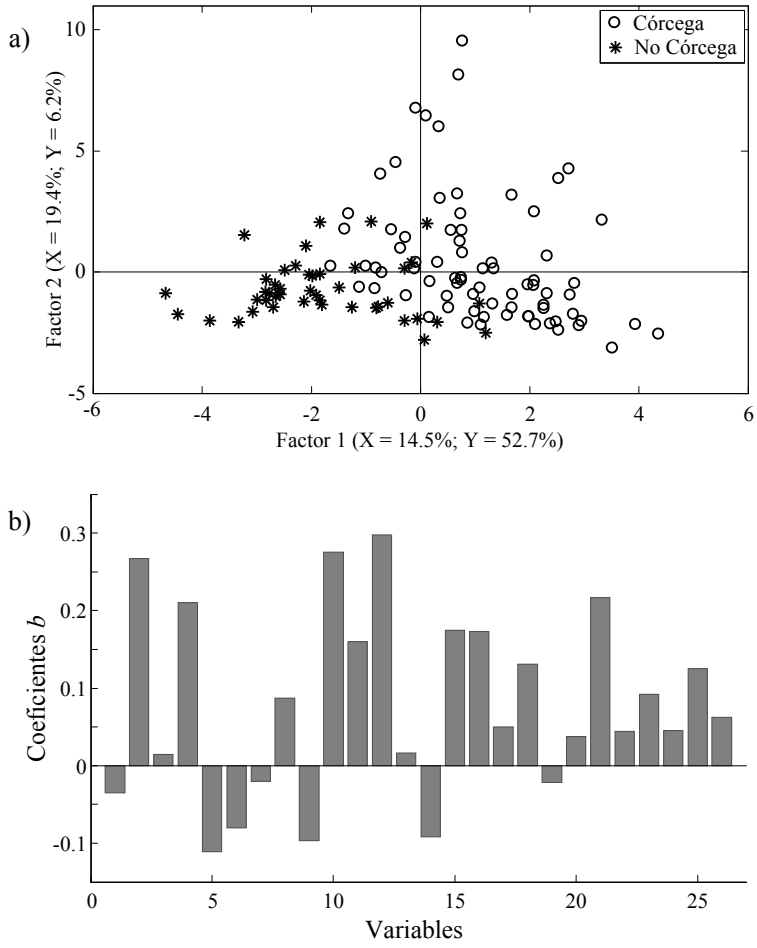

Figura 2-24: Modelo *Córcega* frente al resto. a) Gráfica de *scores* de los dos primeros factores. b) Coeficientes de regresión *b* para el modelo óptimo con 5 factores.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Al calcular las fiabilidades de clasificación del modelo *Córcega* frente al resto para los datos de entrenamiento (Figura 2-25a) y de prueba (Figura 2-25c), se observó una elevada dispersión de todas las predicciones. Aunque el 97.4% de los objetos de entrenamiento de *Córcega* fueron bien asignados, las fiabilidades estuvieron entre el 50% y el 99.8% con la mayor parte de los objetos por debajo de 95%. Algo similar ocurre con los datos de prueba cuyas fiabilidades estuvieron entre el 52%-98.3%. La baja fiabilidad de algunos objetos de entrenamiento y prueba de la clase *Córcega* se debió a la elevada dispersión en las predicciones del modelo, además, el elevado error estándar de calibración (***SEC***), en promedio 0.23, generó unas FDPs que se solapan. Este solapamiento de las FDP disminuyó la fiabilidad de clasificación de los objetos que se encontraban en la zona de solapamiento.

Las predicciones de los objetos de entrenamiento de las clases no *Córcega* obtuvieron fiabilidades entre 10% al 83% con cerca del 18% de los datos mal asignados a *Córcega*. La dispersión de las predicciones de los objetos de prueba no *Córcega* fue menor, obteniendo fiabilidades que están comprendidas entre 19% al 75%, con un objeto mal asignado a *Córcega*. La dispersión de las predicciones para los objetos no *Córcega* pudo deberse a la incompatibilidad de las clases agrupadas. A su vez, pueden existir clases que se parecen más a la clase *Córcega* que a las otras clases, por lo tanto sus objetos serán predichos cerca del límite de clases o en el espacio de la clase *Córcega*.

2.8.4 Conclusiones

Aunque es posible diferenciar geográficamente mieles de la clase *Córcega* a partir de datos de intensidad absoluta de masas de componentes de la misma, se obtiene un elevado error de tipo II que no es adecuado para diferenciar la clase *Córcega*. Podemos decir que dividir el conjunto de datos inicial en un modelo binario, en el cual se reúnen varias clases en una única clase, no es la estrategia más adecuada para el conjunto de mieles, pues se obtiene un elevado nivel de dispersión en las predicciones. Más adelante (capitulo 3) se hablará de las ventajas y desventajas de los métodos de binarización y cual o cuales podrían ser los más adecuados para este tipo de problemas en donde los objetos de las clases agrupadas son incompatibles.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

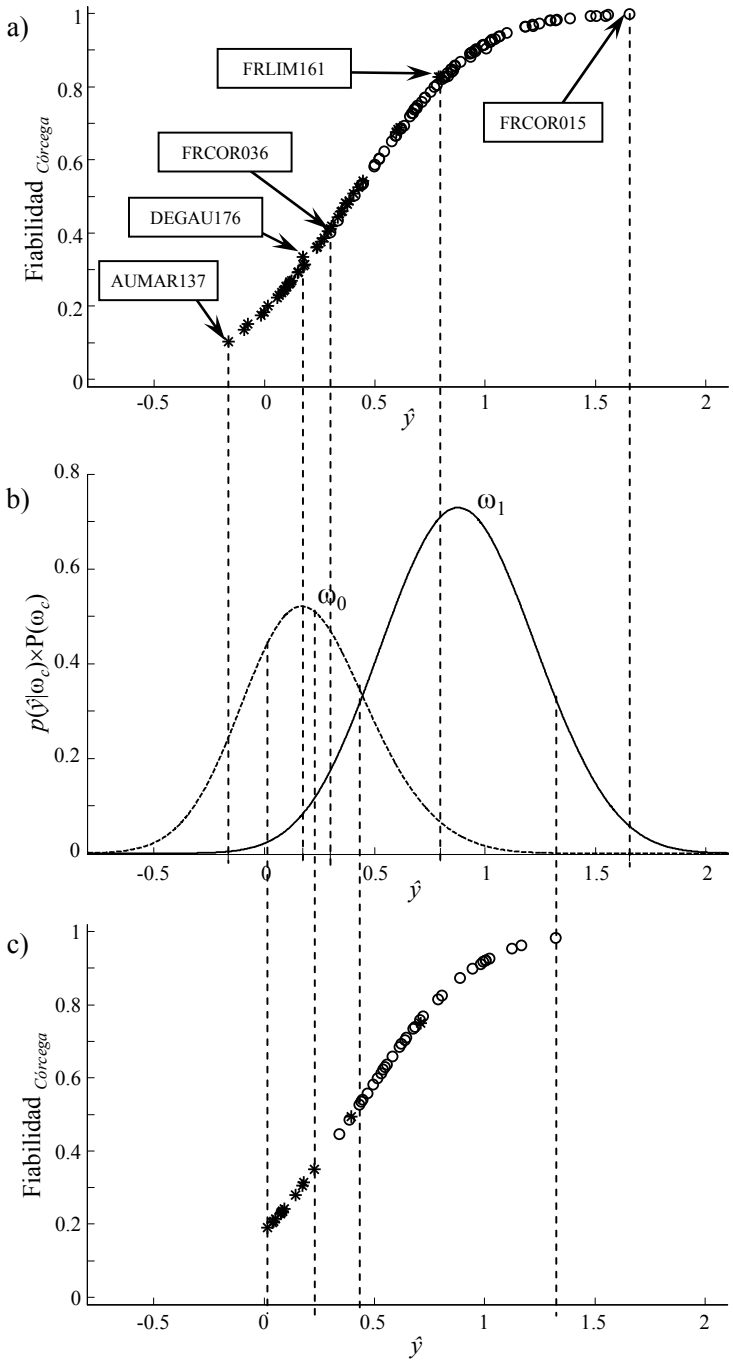

Figura 2-25: Distribución de las predicciones del modelo *Córcega* contra las restantes para 5 factores de los conjuntos de entrenamiento (a) y prueba (c), y FDP para las clases *Córcega* (ω_1) y no *Córcega* (ω_0). *Córcega* (○) y no *Córcega* (*).

Referencias

1. Directiva 2001/110/CE del Consejo de 20 de diciembre de 2001 relativa a la miel, Diario Oficial de las Comunidades Europeas. L 10 (12/1/2002) 47–52.
2. M. Lees, "Food authenticity and traceability", Woodhead Publishing Limited, Cambridge, UK, 2003.
3. R. Karoui, E. Dufour, J.-O. Bosset, J. De Baerdemaeker, Food Chem. 101 (2007) 314–323.
4. I. S. Arvanitoyannis, C. Chalhoub, P. Gotsiou, N. Lydakis-Simantiris, P. Kefalas, Crit. Rev. Food Sci. Nutr. 45 (2005) 193–203.
5. A. Simon, K. Traynor, K. Santos, G. Blaser, U. Bode, P. Molan, Evid. base Compl. Alternative Med. 6 (2009) 165–173.
6. T. Woodcock, G. Downey, C. P. O'Donnell, Food Chem. 114 (2009) 742–746.
7. M. V. Baroni, C. Arrua, M. L. Nores, P. Fayé, M. del P. Díaz, G. A. Chiabrando, D. A. Wunderlin, Food Chem. 114 (2009) 727–733.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2.9 Clasificación con *p*-DPLS de aceite de oliva de origen *Liguria* con datos de espectroscopia de ^{1}H-RMN

2.9.1 Introducción

El aceite de oliva es el jugo aceitoso del fruto del olivo (*Olea europea L*) [1]. Cultivado desde hace más de 5000 años, siempre ha estado ligado a la región mediterránea y a las culturas que la han habitado [2]. Considerado parte fundamental de la dieta mediterránea, se le considera uno de los factores responsables de las cualidades nutricionales y terapéuticas de ésta [3]. De las propiedades terapéuticas reconocidas podemos mencionar: protección contra las enfermedades cardíacas, efecto anti-cancerígeno, prevención de la perdida cognitiva y demencia con la edad, a la vez que aumento en la longevidad, entre otros [3]. Desde el punto de vista económico España, Italia y Grecia, concentran el 76% de la producción mundial de aceite de oliva y el 98% de la producción europea [4].

En el caso europeo el aceite de oliva está regulado por el reglamento Nº 2568/91/CEE [5], que establece las características que deben cumplir las diferentes clases de aceites de oliva y los métodos de análisis que permiten establecer el cumplimiento de dichas características. Sin embargo, un campo muy activo de investigación es la verificación de denominaciones de origen y trazabilidad de los aceites de oliva. En verificación de denominaciones geográficas de origen, se busca qué métodos analíticos apoyados con herramientas quimiométricas sirven para verificar la denominación de origen. Algunos de estos trabajos han utilizado espectroscopia NIR con análisis por PLS [6], PCA y DPLS [1], espectroscopia de masas combinada con *k*-NN, SIMCA [7], análisis discriminante lineal (*Linear Discriminant Analysis* (LDA)) [8], o espectroscopia FT-IR con PCA-FDA y DPLS [9]. En trazabilidad de aceites se ha aplicado el análisis de ADN [10,11], las técnicas calorimétricas [12], o los análisis fisicoquímicos y sensoriales [13]. Cabe aclarar que con estos métodos de trazabilidad también se puede verificar una denominación de origen, ya que se está verificando la procedencia del aceite, es decir, especie de la planta, época del año, tratamiento y demás características que hacen a un producto único.

Apoyando esta política el proyecto TRACE planteó utilizar la espectroscopia infrarroja media y cercana (MIR y NIR) y la resonancia magnética nuclear de protón (^{1}H-RMN) para el análisis de los aceites, apoyadas con el análisis quimiométrico de las matrices de

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

espectros. En este caso utilizamos *p*-DPLS para clasificar los aceites en denominaciones de origen geográficas de diferentes ubicaciones en la Unión Europea, además de ver las ventajas de aplicar esta metodología a este tipo de muestras.

2.9.2 Parte experimental

Para este estudio se utilizó el conjunto de datos ^{1}H-RMN de aceites de oliva del proyecto TRACE (detalles en el anexo, apartado 4) para la clasificación por regiones italianas con 478 objetos, 342 variables de desplazamientos ^{1}H-RMN (figura 2-26) y 15 clases: *Abruzzo*, *Calabria*, *Campania*, *Friuli Venezia Giulia, Lazio*, *Liguria*, *Lombardia*, *Marche*, *Molise*, *Puglia*, *Sicilia*, *Toscana*, *Trentino Alto Adige*, *Umbría* y *Veneto*. Además se observó un posible efecto por año de recolección, *2005* y *2006*. Se aplicó el algoritmo *Kennard-Stone* para dividirlo en datos de entrenamiento (70% de los datos) y datos de prueba (30% de los datos). Sin embargo, por requerimientos de TRACE sólo se desarrolló el modelo "*Liguria* frente a las restantes" y el modelo "*2005* frente a *2006*". El número óptimo de factores se decidió por validación cruzada dejando fuera un objeto cada vez (LOOCV), con criterio de máxima sensibilidad para el modelo *Liguria* frente a las restantes (porcentaje de objetos de la clase *Liguria* bien asignados) y para el modelo *2005* frente a *2006* el máximo porcentaje de clasificación total (porcentaje de objetos bien asignados para las dos clases). Los modelos se desarrollaron con datos centrados y como función de decisión de clasificación se utilizó la ecuación 17 del apartado 2.2.2.5, para un intervalo incertidumbre $2\times SEP$, intervalo en el cual se calcula el área bajo las FDP de las clases.

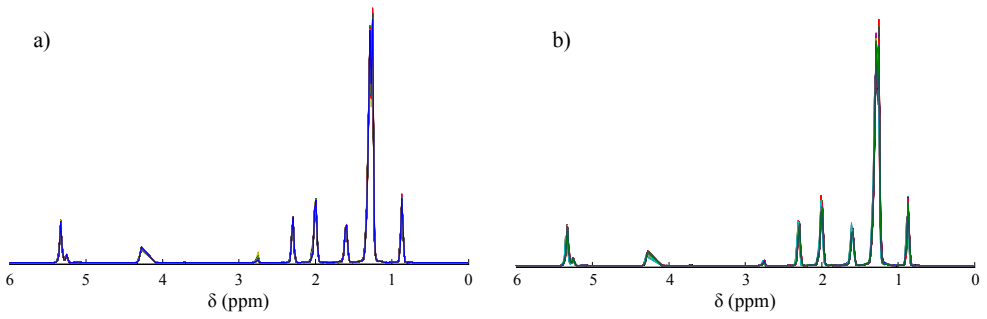

Figura 2-26: Espectros ^{1}H-RMN de los aceites de oliva para los años 2005 (a) y 2006 (b).

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2.9.3 Resultados y discusión

2.9.3.1 *Modelo Liguria frente a las restantes*

Para el modelo "*Liguria* frente a las restantes" no se tuvo en cuenta la clase *Friuli Venezia Giulia* por tener sólo dos objetos. Además, se encontraron 9 objetos discrepantes: 1 de la clase *Calabria*, 7 de la clase *Liguria* y 1 de la clase *Sicilia*. De estos objetos discrepantes, 4 coinciden con los datos discrepantes italianos detectado en el análisis por PCA (detalles ver anexo, apartado 4). Así, el modelo "*Liguria* frente a las restantes" se desarrolló con 467 muestras. El número óptimo de factores con criterio de máxima sensibilidad fue de 37. El elevado número de factores del modelo óptimo se debe, en parte, a las características de los datos. Los espectros RMN varían poco entre objetos (figura 2-26); a ello hay que sumarle que se tienen objetos de años distintos para una misma clase y pueden existir cambios de un año a otro. Otra causa puede estar en el elevado número de clases que se agrupan como no *Liguria*, en donde al haber clases incompatibles el modelo se ve forzado a modelarlas como una sola, aumentado el número de factores necesarios para poder separar la clase *Liguria* de la restantes.

Como se observa en las gráficas de *scores* (Figura 2-27a) con los dos primeros factores el modelo "*Liguria* frente a las restantes" no se puede diferenciar los objetos *Liguria* de los otros, se tiene el 19% de varianza explicada en *y* requiriendo más de 26 factores para explicar el 70%. Observando los coeficientes de regresión *b* (Figura 2-27b) el modelo "*Liguria* frente a las restantes" con 37 factores tiene cerca de 17 variables significativas, distribuidas en el espectro incluso en zonas donde no hay señal, se aprecia un ruido significativo en el modelo.

La habilidad de clasificación del modelo "*Liguria* frente a las restantes" es menor de lo esperado. Tanto con datos de entrenamiento por LOOCV como con datos de prueba se obtuvo una sensibilidad que rondó del 72%, mientras que la especificidad rondó el 98%. La baja sensibilidad, sumada al elevado número de factores utilizado indica que no es posible diferenciar la región *Liguria* de las restantes, ya que los datos *Liguria* no son homogéneos entre un año y otro y a la elevada deferencia entre el número de objetos *Liguria* (15.4% de los objetos) frente a las restantes, que además corresponden a 13 clases distintas, donde algunas de ellas pueden compartir características con la clase *Liguria*.

En la figura 2-28 se observa que las predicciones para los objetos de entrenamiento por LOOCV presentan una elevada dispersión. Así, el intervalo de predicción de los

objetos de la clase *Liguria* (0.1 a 1.4) solapa al intervalo de los objetos de las clases distintas a *Liguria* (-0.4 a 1, con la mayoría de los objetos entre -0.4 a 0.5). Si a lo anterior le sumamos un desbalance en el número de datos entre las clases, cuando se construyeron las FDP éstas fueron amplias y solapadas lo que redujo la fiabilidad de asignación. Es por ello, que el 73% de los objetos de entrenamiento de la clase *Liguria* tienen una fiabilidad del orden del 50%-90%, y el 27% de los objetos están por debajo del 50%, siendo rechazados como *Liguria* y generando el elevado error de tipo I. El caso opuesto presentan los objetos de entrenamiento de las clases diferentes a *Liguria*, donde la mayoría de estos tienen una fiabilidad del 100% al 50% de no ser *Liguria*, sólo el 2.6% de los objetos están por debajo del 50% de fiabilidad de no ser *Liguria* (bajo error de tipo II). Similar comportamiento presentan los objetos de prueba.

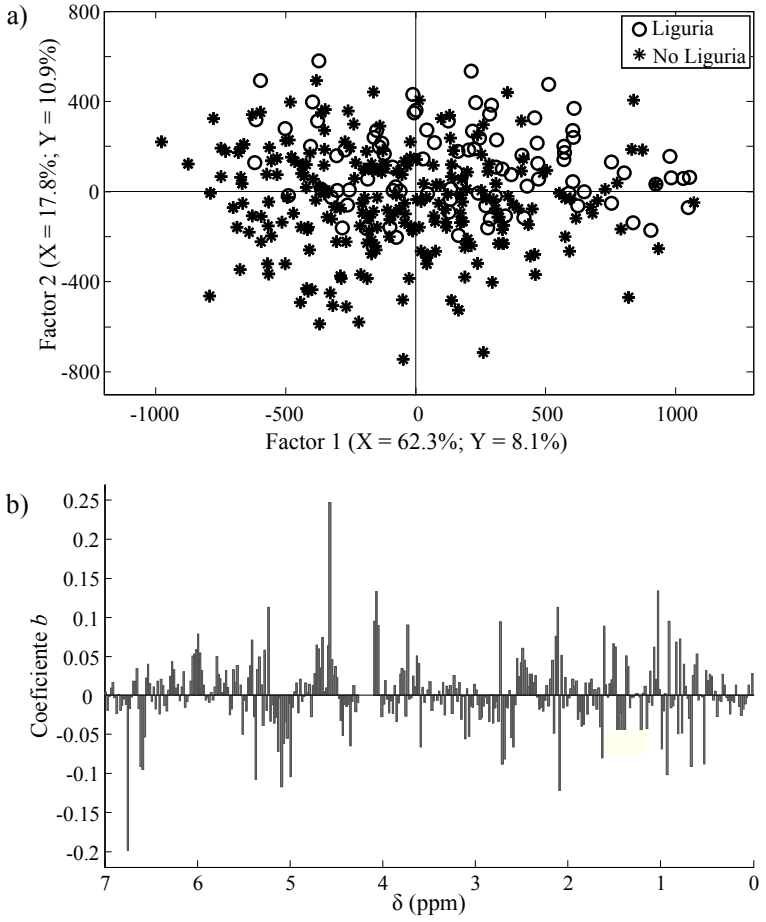

Figura 2-27: Modelo *Liguria* frente a las restantes. a) Gráfica de *scores* de los dos primeros factores. b) Coeficientes de regresión *b* para el modelo con 37 factores.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

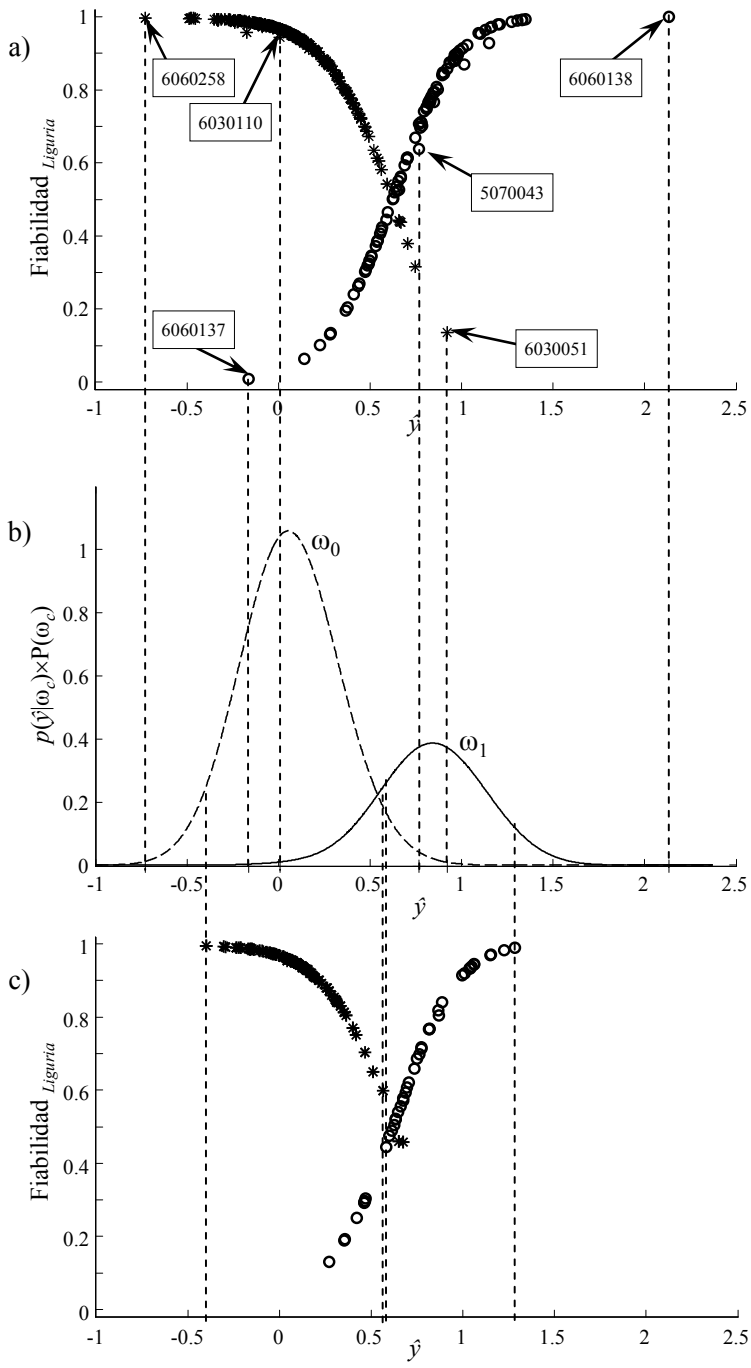

Figura 2-28: Distribución de las predicciones del modelo *Liguria* contra el resto para 37 factores de los conjuntos de entrenamiento (a) y prueba (c), y FDP para las clases *Liguria* (ω_1) y no *Liguria* (ω_0). *Liguria* (○) y diferentes a *Liguria* (*).

2.9.3.2 *Modelo 2005 frente a 2006*

Para el modelo *2005* frente a *2006* se encontraron 10 objetos discrepantes que pertenecen al *2006*. Coinciden los 4 objetos discrepantes detectados con PCA y todos los detectados en el modelo "*Liguria* frente a las restantes". El modelo *2005* frente a *2006* finalmente se desarrolló con 468 objetos, obteniendo 10 factores con el criterio de máximo porcentaje de clasificación total óptima, que asegura la máxima sensibilidad y especificidad. Como se observa en las gráficas de *scores* (Figura 2-29 a) el modelo "*2005* frente a *2006*" con el primer factor se pueden separar aceptablemente los objetos del año *2005* de los del *2006*. El segundo factor obtiene una separación parcial de estos años, limitando la mayoría de los objetos del año *2006* al cuadrante negativo de la gráfica. Así, los dos primeros factores explican más del 61% de la varianza en *y* (Figura 2-29a), requiriendo cerca 13 factores para explicar más del 90%. Es decir, que se obtuvo una buena discriminación por años a partir de los espectros de ^{1}H-RMN. Observando los coeficientes de regresión *b* (Figura 2-29 b), el modelo *2005* frente a *2006* cuenta con cerca de 15 variables significativas, que coinciden con los desplazamientos de ^{1}H más importantes del espectro, siendo significativos los que se encuentran cerca de 4.2ppm y 5.3ppm, aunque la intensidad en el espectro es baja en los coeficientes de regresión que tienen los valores más elevados.

La habilidad de clasificación del modelo "*2005* frente a *2006*" con 10 factores fue elevada. Presentó sensibilidad y especificidad cercanas al 100%, tanto para los datos de entrenamiento por LOOCV (sensibilidad y especificidad del 99.4%) como para los de prueba (sensibilidad y especificidad del 100%). Al calcular las fiabilidades de clasificación para el modelo *2005* frente a *2006* con 10 factores se obtuvieron fiabilidades promedio para objetos del *2005* del 96.7% y para los del *2006* del 97.0%. Esto se debió a la baja dispersión de las predicciones que, sumado a un *SEC* reducido, permitió generar FDP más estrechas que no se solaparon. A su vez, al reducirse el intervalo en el cual se calculó el área bajo las FDP, benefició a aquella clase con mayor área, aumentando su fiabilidad.

El hecho de poder clasificar por año y no por regiones geográficas, podría indicar que la espectroscopia ^{1}H-RMN no aporta la información necesaria para caracterizar las regiones geográficas. Buena parte de las características que delimitan una región geográfica vienen dadas por los suelos, que tienen distintos aportes isotópicos de los elementos presentes en él; al menos que algunos de estos isótopos característicos esté enlazados a un carbono y cercano a un ^{1}H, para que este último se vea afectado por su presencia, no se vería el efecto de isótopo en el espectro ^{1}H-RMN. Además, hay que tener en cuenta que los isótopos de estos elementos están en muy baja concentración

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

por lo que afectarían a muy pocos H, generando señales de baja intensidad que se confundirían con ruido. También debe tenerse en cuenta las especies de oliva cultivadas, que dan diferentes aportes de ácidos grasos y estos pueden interferir en las relaciones por áreas geográficas. Otro factor a tener en cuenta son las clases contenidas en la superclase, dado que algunas de estas pueden ser similares a *Liguria*, lo que alteraría el modelo. La clasificación por años está ligada a las condiciones climáticas, que varían las proporciones de ácidos grasos presentes en el aceite y otras sustancias ricas en hidrógeno que puedan ser registradas en el espectro.

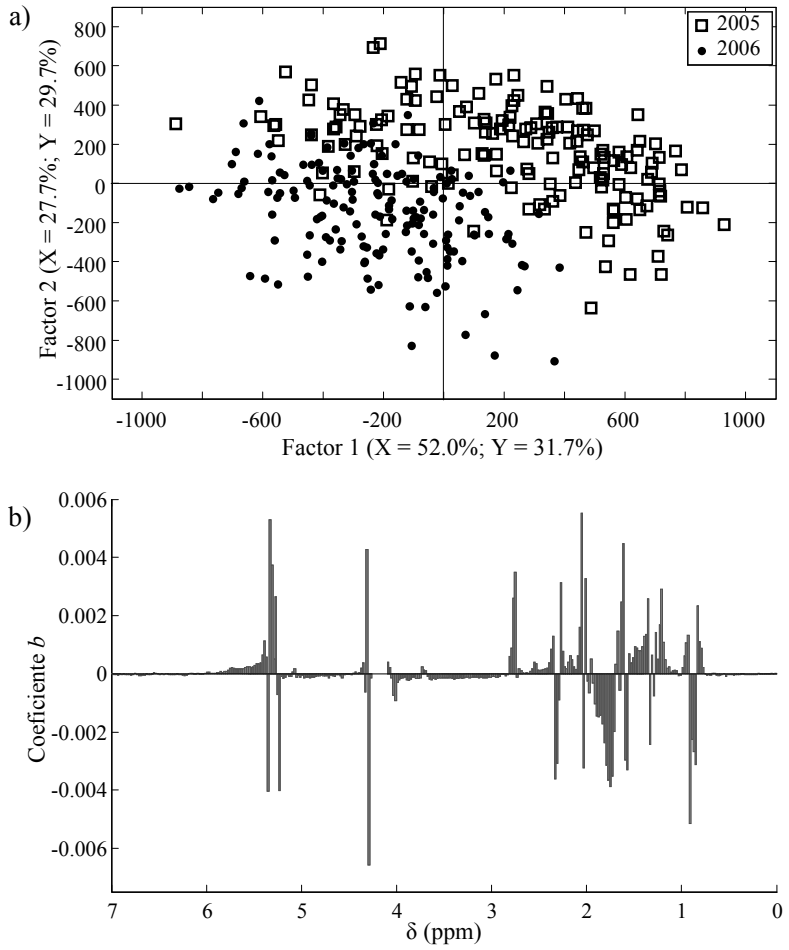

Figura 2-29: Modelo *2005* frente a *2006*. a) Gráfica de *scores* de los dos primeros factores. b) Coeficientes de regresión *b* para el modelo con 10 factores.

2.9.4 Conclusiones

La clasificación de muestras de aceite de oliva por región geográfica *Liguria* a partir de espectros de ^{1}H-RMN conlleva un elevado error de tipo I, cercano al 30%, a lo que se debe sumar una baja fiabilidad en las clasificaciones, debido al elevado solapamiento de las clases. Esto puede atribuirse al mayor número de objetos de las clases diferentes a *Liguria* y a su vez, a que se reunieron muchas clases en una sola. No obstante, se establece que la técnica *p*-DPLS puede ser utilizada en autentificación del aceite por años. Se tienen datos más balanceados, con un buen nivel de fiabilidad.

Referencias

1. Woodcock, G. Downey, C. P. O'Donnell, J. Agric. Food Chem. 56 (2008) 11520–11525.
2. http://es.wikipedia.org/wiki/Historia_del_aceite_de_oliva, última consulta 17/07/2009.
3. D. L. García-González, R. Aparicio-Ruiz, R. Aparicio, Eur. J. Lipid. Sci. Tech. 110 (2008) 602–607.
4. http://www.internationaloliveoil.org/web/aa-ingles/corp/AreasActivitie/economics/economics-oliveOilFigures.html. última consulta 17/07/2009.
5. Reglamento (CEE) No 2568/91 de la Comisión de 11 de julio de 1991 relativo a las características de los aceites de oliva y de los aceites de orujo de oliva y sobre sus métodos de análisis. Diario Oficial de la Unión Europea, 248 (5/9/1991) p. 1.
6. O. Galtier, N. Dupuy, Y. Le Dréau, D. Ollivier, C. Pinatel, J. Kister, J. Artaud, Anal. Chim. Acta. 595 (2007) 136–144.
7. S. López-Feria, S. Cárdenas, J. A. García-Mesa, M. Valcárcel, Talanta. 75 (2008) 937–943.
8. L. Vaclavik, T. Cajka, V. Hrbek, J. Hajslova, Anal. Chim. Acta. 645 (2009) 56–63.
9. S. Hennessy, G. Downey, C. P. O'Donnell, J. Agric. Food Chem. 57 (2009) 1735–1741.
10. S. Pafundo, C. Agrimonti, E. Maestri, N. Marmiroli, J. Agric. Food Chem. 55 (2007) 6052–6059.
11. P. Martins-Lopes, S. Gomes, E. Santos, H. Guedes-Pinto, J. Agric. Food Chem. 56 (2008) 11786–11791.
12. M. Angiuli, C. Ferrari, L. Lepori, E. Matteoli, G. Salvetti, E. Tombari, A. Banti, N. Minnaja, J. Therm. Anal. Cal. 84 (2006) 105–112.
13. I. S. Arvanitoyannis, A. Vlachos, Crit. Rev. Food. Sci. Nutr. 47 (2007) 441–498.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Capítulo 3

Clasificación multiclase

3.1 Introducción y revisión bibliográfica

La gran mayoría de algoritmos de clasificación son más efectivos con problemas de clasificación de dos clases (binarios, C=2), o como es el caso de *Support Vector Machines* (SVM) el algoritmo fue ideado para este propósito [1]. Sin embargo, la gran mayoría de problemas de clasificación tienen más de dos clases (C>2), por ejemplo, identificar el origen geográfico de aceites de oliva [2]. Este tipo de problemas de clasificación son conocidos como multiclase, y han sido ampliamente investigados en los últimos años en ámbitos tan diversos como el económico [3] y el médico [2,4], entre otros. En química se han utilizado para asignar alimentos a algún posible origen geográfico o en la protección de denominaciones de origen (*Protected Designations of Origin*, PDO) [2,5,6]; también se ha utilizado en la identificación de lactato deshidrogenasa (LDH) procedente de diferentes carnes de animales [7]; en la caracterización o autentificación de alimentos derivados de animales [8,9]; y en bioquímica para la identificación de diferentes tipos de tumores en humanos a partir de su expresión genética detectada mediante *microarrays* de ADN [4,10–13].

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

3.1.1 Estrategias de clasificación multiclase

Al igual que en clasificación binaria, un problema multiclase requiere una función de decisión de clasificación que tome un vector **x** de un objeto desconocido y lo asigne a alguna o ninguna de las *C* clases modeladas. Para encontrar la función de decisión existen dos estrategias. La primera halla una función directa o modelo único de clasificación, es decir, el algoritmo en un solo paso asigna el objeto a alguna de las *C* clases. Ejemplo de este tipo de estrategia son las redes neuronales (*Artificial Neural Networks*, ANN) [14] o el método de los vecinos más cercanos (*k*-NN) [8]. La segunda estrategia utiliza una función compuesta o varios modelos de clasificación de modo que el problema multiclase se divide en varios subproblemas de clasificación. Una vez resueltos se combinan sus respuestas en una única respuesta de clasificación.

Esta última estrategia tiene a su vez dos subestrategias. La primera divide el problema multiclase en modelos para cada una de las clases, es decir, que se tiene una función por clase, tal como lo hace el *Soft Independent Modelling of Class Analogy* (SIMCA) [15]. La otra subestrategia se llama binarización [15,16] y divide el problema en varios problemas binarios, permitiendo utilizar una amplia gama de algoritmos de clasificación binaria, como SVM [11] o DPLS [17], entre otros.

La problemática de utilizar varios modelos es cómo combinar sus respuestas, ya que la decisión en este caso no depende directamente del vector desconocido **x**, sino de las respuestas de los modelos menores. Por ello se requieren funciones de combinación sencillas y que minimicen el error de asignación.

3.1.2 Binarización

La binarización se utiliza ampliamente en la resolución de problemas multiclase, ya que permite utilizar algoritmos binarios que son más efectivos que la mayoría de los multiclase (CART, redes neuronales y similares). Sin embargo, presenta dificultades derivadas del proceso de división del conjunto de datos. A continuación se comentan las ventajas y desventajas de distintas estrategias de binarización.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

3.1.2.1 *Estrategias de binarización*

Se pueden distinguir tres estrategias para dividir los problemas multiclase en clasificadores binarios: "*uno-contra-todos*" (también conocida como "*uno-contra-resto*"), "*uno-contra-uno*" y "*P-contra-Q*" [17]. En todos los casos el vector original **y** (que contiene la clase a la que pertenece cada objeto) se sustituye por otro en donde las clases son codificadas como 0 y 1; asignándole 1 a la clase o clases de interés y 0 a la clase o clases que no son de interés.

- "*P-contra-Q*" ("*P-against-Q*", PAQ) divide el conjunto de datos en dos grupos, uno con P clases y otro con las Q clases remanentes. Cuando sigue una estructura jerárquica [8] PAQ transforma el problema de C clases en $C-1$ modelos binarios en donde tanto P como Q pueden tener 1 o más clases. En este caso se reúnen clases con similares características en dos macroclases [8,15] y se utiliza un modelo binario (nodo 1) para asignar el objeto a una de las dos macroclases. En un nuevo nivel la macroclase se divide en dos nuevas macroclase menores y se desarrolla un nuevo modelo binario. El procedimiento se repite hasta que los modelos binarios contengan dos clases puras. La desventaja de esta estrategia es que una mala asignación en uno de los nodos conduce a un resultado erróneo.

- "*Uno-contra-todos*" ("*one-against-all*", OAA) es similar a PAQ en donde P contiene sólo una clase, la de interés, y Q las restantes $C-1$ clases. Esta estrategia transforma el problema de C clases en C modelos binarios (un modelo para cada una de las clases) [16]. La problemática surge al combinar varias de las clases en una superclase, pues pueden reunirse clases incompatibles que una contra otra podrían ser discriminadas correctamente. Esto fuerza el modelo a considerar clases opuestas como una sola, aumentado así el error de modelado. Otra desventaja es la diferencia en el número de objetos entre clases [18], ya que la mayoría de veces la clase de interés tendrá muchos menos objetos que la superclase que reúne a las restantes. En consecuencia, el modelo tenderá a modelar mejor la superclase a expensas de modelar peor la clase de interés.

- "*Uno-contra-uno*" ("*one-against-one*", OAO) también llamada "*Round Robin*" [16] o "*Pairwise*" [19]. Esta estrategia utiliza dos clases por modelo, siendo una de ellas la de interés. En este caso se calculan $C(C-1)/2$ modelos binarios (es decir, todos los posibles pares de clases). Esta técnica elimina parte de los problemas de OAA, ya que se comparan pares de clases y no superclases. Así se minimiza el efecto de la diferencia en el número de objetos por clase y además los límites entre clases son

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

más definidos, al no mezclarse clases incompatibles. Estas características han hecho que la estrategia OAO sea muy utilizada en binarización [20].

3.1.3 Métodos de combinación

Dentro de los métodos de combinación más utilizados se encuentran aquellos que siguen el principio de "el ganador se lo lleva todo" (*winner-takes-all*) [15,19] y el "*Error-Correcting Output Code*" (ECOC) [21].

- El ganador se lo lleva todo (*winner-takes-all*) es la más simple y usada de las técnicas de combinación de respuestas de clasificadores. Métodos de combinación como Voto Mayoritario [22], Voto Mayoritario Ponderado [22], *Behaviour Knowledge Space* [22], *Dempster-Shafer* [23] o *Naïve Bayes* [24] son de este tipo. Esta técnica de combinación asigna un objeto a una clase, basándose en el valor más alto resultado de combinar las decisiones de los clasificadores base:

$$F(\mathbf{x}) = \underset{c\in\{1,\ldots,C\}}{\arg\max}\, f_c(\mathbf{x}) \tag{3-1}$$

en donde $F(\mathbf{x})$ es la función de decisión del método de combinación para el vector $\mathbf{x}$; $f_c(\mathbf{x})$, es la función de decisión de los clasificadores base para una clase c partiendo de un vector $\mathbf{x}$ y C es el número total de clases. Por ejemplo, una función común para combinar modelos binarios usando la estrategia OAO es la sumar la respuesta de los clasificadores base a una clase c:

$$f_c(\mathbf{x}) = \sum_{k}^{K} f_{c,k}(\mathbf{x})$$

en donde $f_{c,k}(\mathbf{x})$ es la función de decisión para una clase c en el clasificador base k para el vector $\mathbf{x}$, y K es el número total de clasificadores base. Es decir, que la función evalúa la respuesta de cada uno de los clasificadores, suma las respuestas para una misma clase c, y asignan el objeto a la clase con la mayor respuesta. Por ello, los métodos mencionados anteriormente son considerados *winner-takes-all*, ya que cada uno halla una propiedad (por ejemplo: votos en voto mayoritario o probabilidad en *Naïve Bayes*), y asignan el objeto a la clase donde la propiedad sea mayor.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

- *Error-Correcting Output Code* (ECOC) fue propuesto por *Dietterich y Bakini* [21] y asigna un código binario, llamado "*codeword*", a cada una de las clases. Estos *codeword* son trasladados como filas a una matriz de confusión, cuyas columnas contienen la codificación de las clases en cada uno de los modelos binarios, con lo que hay tantas columnas como modelos se utilicen. La mayor dificultad del método radica en establecer los *codeword*, ya que deben contener los términos (modelos) suficientemente para diferenciar las clases. Si el código es corto pueden darse inconsistencias, como que el código de un objeto desconocido sea similar a más de una clase. Para encontrar estos códigos se han desarrollado algoritmos que permiten construir la matriz más adecuada para diferenciar correctamente las clases [25]. ECOC utiliza comúnmente la distancia de *Hamming* [21] para calcular las diferencias entre el código del objeto problema, calculado con los modelos binarios, y los códigos de las clases. El objeto se asigna a la clase donde la distancia sea menor. ECOC es adecuado para problemas con un número elevado de clases, ya que permite optimizar la cantidad de modelos binarios requeridos.

La búsqueda de técnicas de combinación efectivas ha llevado a plantear sistemas tan complejos como la metodología "*Decision pathway modeling*" [17] que, a través de una red de modelos binarios (todas las posibles combinaciones binarias entre clases y superclases), hace la asignación del objeto con el mínimo error. Sin embargo, resolver esta red resulta más complejo que otras alternativas de clasificación, por ejemplo: para un problema con 4 clases se requieren 10 modelos o para 5 clases 25 modelos. Otra técnica para la resolución de problemas multiclase está basada en gráficos acíclicos dirigidos (*Directed Acyclic Graph*, DAG). Algunas técnicas que aplican SVM para resolver problemas multiclase utilizan DAG [26]. Dada su flexibilidad, los DAG son ideales para resolver problemas multiclases en donde otros métodos o estrategias de clasificación no obtienen buenos resultados.

En el siguiente apartado se establece la metodología para resolver problemas multiclase partiendo de modelos *p*-DPLS. Se desarrolla un nuevo método de combinación de modelos binarios *p*-DPLS a partir del producto de fiabilidades para una misma clase y se calcula la fiabilidad de multiclasificación dividiendo el producto de una clase sobre la suma de productos de todas las clases presentes en el problema multiclase. La metodología *p*-DPLS multiclase y sus resultados se documentaron en el artículo "*Multi-class classification with probabilistic discriminant partial least squares (p-DPLS)*", *Analytica Chimica Acta* 664 (2010) 27–33.

Referencias

1. N. García-Pedrajas, D. Ortiz-Boyer, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1001–1006.
2. H. S. Tapp, M. Defernez, E. K. Kemsley, J. Agric. Food Chem. 51 (2003) 6110–6115.
3. D. Feldman, S. Gross, J. R. Estate Finance Econ. 30 (2005) 369–396.
4. Y. Tan, L. Shi, W. Tong, G. T. G. Hwang, C. Wang, Comput. Biol. Chem. 28 (2004) 235–244.
5. F. Marini, A. L. Magrì, R. Bucci, F. Balestrieri, D. Marini, Chemometr. Intell. Lab. Syst. 80 (2006) 140–149.
6. L. Pillonel, U. Bütikofer, H. Schlichtherle-Cerny, R. Tabacchib J. O. Bosset, Int. Dairy. J. 15 (2005) 557–562.
7. D. Bylund, J. Samskog, K. E. Markides, S. P. Jacobsson, J. Am. Soc. Mass Spectrom. 14 (2003) 236–240.
8. B. K. Alsberg, R. Goodacre, J. J. Rowland, D. B. Kell, Anal. Chim. Acta. 348 (1997) 389–407.
9. J. M. Herrero-Martinez, E. F. Simo-alfonso, G. Ramis-Ramos, C. Gelfi, P. G. Righetti, Electrophoresis. 21 (2000) 633–640.
10. C. H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, T. Golub, Bioinformatics, 17, Suppl. 1 (2001) S316–S322.
11. A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, Bioinformatics. 21 (2005) 631–643.
12. T. Y. Yang, Comput. Stat. Data Anal. 53 (2009) 756–765.
13. T. Kawamura, H. Mutoh, Y. Tomita, R. Kato, H. Honda, J. Biosci. Bioeng. 106 (2008) 442–448.
14. R. Anand, K. Merota, C. K. Mohan, S Ranka, IEEE Trans. Neural Network. 6 (1995) 117–124.
15. M. Pardo, G. Sberveglieri, A. Tarino, F. Maulli, G. Valentini, Anal. Chim. Acta. 446 (2001) 223–232.
16. J. Fürnkranz, J. Mach. Learn. Res. 2 (2002) 721–747.
17. A. J. Myles, S. D. Brown, J. Chemometr. 18 (2004) 286–293.
18. X.–M. Zhao, X. Li, L. Chen, K. Aihara, Proteins. 70 (2008) 1125–1132.
19. U. Kreßel, "Pairwise classification and support vector machines", En: B. Schölkopf, C. J. C. Burges, A. J. Smola (Eds.), "Advances in Kernel Methods-Support Vector Learning", MIT Press, Cambridge, MA, 1999, p. 255–268.

20. J. H. Oh, Y. B. Kim, P. Gurnani, K. P. Rosenblatt, J. X. Gao, Bioinformatics. 24 (2008) 1812–1818.
21. T. G. Dietterich, G. Bakiri, J. Artif. Intell. Res. 2 (1995) 263–286.
22. L. I. Kuncheva; "Combining Pattern Classifiers: Methods and Algorithms", A Wiley-Interscience publication, 2004, p. 112–125.
23. M. Reformat and R. R. Yager, Soft. Comput. 12 (2008) 543–558.
24. L. I. Kuncheva, Pattern Recogn. Lett. 27 (2006) 830–837.
25. L. I. Kuncheva, Pattern Recogn. Lett. 26 (2005) 83–90.
26. B. Fei, J. Liu, IEEE Trans. Neural Network. 17 (2006) 696–704.

3.2 Multi-class classification with probabilistic discriminant partial least squares (*p*-DPLS)

Analytica Chimica Acta 664 (2010) 27–33

Néstor F. Pérez, Joan Ferré*, Ricard Boqué

Department of Analytical Chemistry and Organic Chemistry, Rovira and Virgili University. C/ Marcel·lí Domingo, s/n. 43007. Tarragona, Spain

This work describes multi-classification based on binary probabilistic discriminant partial least squares (*p*-DPLS) models, developed with the strategy one-against-one and the principle of winner-takes-all. The multi-classification problem is split into binary classification problems with *p*-DPLS models. The results of these models are combined to obtain the final classification result. The classification criterion uses the specific characteristics of an object (position in the multivariate space and prediction uncertainty) to estimate the reliability of the classification, so that the object is assigned to the class with the highest reliability. This new methodology is tested with the well-known Iris data set and a data set of Italian olive oils. When compared with CART and SIMCA, the proposed method has better average performance of classification, besides giving a statistic that evaluates the reliability of classification. For the olive oil set the average percentage of correct classification for the training set was close to 84% with *p*-DPLS against 75% with CART and 100% with SIMCA, while for the test set the average was close to 94% with *p*-DPLS as against 50% with CART and 62% with SIMCA.

3.2.1 Introduction

In multi-class classification problems we have an $I \times J$ set $\mathbf{X}$ of J observed variables in I training objects, a vector $\mathbf{y}$ that codifies the class c ($c = 1,\ldots, C$; with $C>2$) of each object and a vector $\mathbf{x}$ of variables measured for the unknown object that must be assigned to one (or none) of the C possible classes. Examples of multi-class classification problems are the assignation of food commodities to one out of several possible origins [1,2] and the identification of different tumour types from microarray gene expression data [3,4].

ISBN:978-84-693-4053-0/DL:T.990-2010

A multi-class classification problem is solved by using an adequate classifier decision function that maps **x** onto a class label [5]. One approach is to use a single classification function like in *k*-Nearest Neighbours (*k*-NN) [6], or Artificial Neural Networks (ANN) [7]. In these cases, the classification of an object in one of the *C* classes is done in one step. Another approach is to divide the multi-class problem into *K* smaller classification problems, each one with its own decision rule, and then combine the output of the *K* individual classifications to obtain the final result. This can be done not only by using single-class models, such as the Soft Independent Modelling of Class Analogy (SIMCA) method [1], but also by using binary classification methods that have to decide between two classes or superclasses. The latter approach is known as dichotomization or binarization [1] and has the advantage that a wide range of binary classification methods, such as Support Vector Machines [5] and discriminant partial least squares (DPLS) [8] can be used.

There are three possible ways of splitting the classes for binary classifiers: one-against-all (where "all" means "the rest"), one-against-one and P-against-Q [9]. In all cases the original vector **y** is replaced by another one that codifies with a "1" the objects that belong to the class or classes of interest, and with a "0" the objects that do not belong to the classes of interest. The strategy P-against-Q (PAQ) first splits the data into two groups, one with P classes, and one with the remaining Q classes. At the next level, the classes in P are also split into two groups, and a binary model that discriminate between them is calculated. The division is also done for the classes in Q. The split continues at the successive levels until models only discriminate between two classes. This procedure can solve a classification problem of *C* classes using *C*−1 binary models [6]. The drawback of hierarchical PAQ is that an error of allocation in one node results in the object being misclassified. The strategy one-against-all (OAA) is similar to PAQ, in which P only contains one class, and Q contains the remaining *C*−1 classes. In this case, the problem is solved either hierarchically (which involves *C*−1 models) or by simultaneous combination of *C* binary models [5]. A weakness of OAA is that the number of objects of the class of interest can be imbalanced with respect to the other super-class that contains the rest of the objects. Moreover, incompatible classes (that could be correctly discriminated if they were modelled one against the others) are grouped together, thus forcing the model to consider opposite classes as a unique super-class. Finally, the strategy one-against-one (OAO) involves $C(C-1)/2$ binary models, each model discriminating only between two classes [5], so models for all pairs of classes are calculated. OAO overcomes some of the problems of OAA. Since the models compare only two classes and not groups of classes, the effect of having a different number of objects per class is minimized, and the boundaries between classes are clearer [10] (classes that may be incompatible are not grouped into

the same class). These characteristics have made of OAO a strategy of binarization widely used [11] and it has been the one applied in this paper.

To classify an unknown object, the outputs of the binary models are combined with an appropriate decision function $F(\mathbf{x})$ to obtain the final classification result. The search for combination functions that minimize the allocation error gave rise to the simple and widely used principle winner-takes-all [1,10] or more complex methods such as the Error-Correcting Output Code (ECOC) [12,13]. The winner-takes-all principle assigns the object to the class with the highest score obtained from the results of the base classifiers. An often used winner-takes-all strategy is the majority vote [14], in which each base classifier decides the class of the object, and the object is assigned to the class with the largest number of votes. A limitation of combining outputs by simple vote is that every vote counts equal and it ignores the reliability that the model attributes to each vote, considering reliability as the probability that the vote or classification is reliable. Although some methods like ECOC or the weighted majority vote calculate the reliability of classification [15,16], their disadvantage is that they assign the same reliability of classification to any unknown object, without taking into account the position of the object in the multivariate space. In other words, they calculate an average reliability value during the validation step, and this value is assigned to any new object to be classified. In this work another winner-takes-all strategy is proposed, in which a reliability, that is particular for each object, is derived from each binary model, and the object is assigned to the class with the largest combined reliability.

Probabilistic discriminant partial least squares (*p*-DPLS) [17] is a binary classification method that allocates an object to the class with the highest reliability. In this work we extend the binary *p*-DPLS [17] to resolve multi-class problems. Basically, $C(C-1)/2$ binary *p*-DPLS models are calculated following the strategy OAO; each k binary model provides one measure of the reliability of the classification of the unknown object in each of the two classes, c and c', that are being modelled: $R_{c,k}$ and $R_{c',k}$ [17]. This measure is calculated taking into account the standard error of prediction of the DPLS model and is particular for the object. The output reliabilities of all the binary models are combined for each class c, thus giving the reliability of the classification in each class in the multi-class model. Then, using the principle of winner-takes-all, an object is allocated to the class with the highest reliability. The proposed method is illustrated with two datasets: the Fisher's Iris data set and a set of Italian olive oils that must be classified according to their origin.

Anal. Chim. Acta. 664 (2010) 27–33

3.2.2 Multiclass classification for *p*-DPLS

3.2.2.1 *Two-class classification*

Binary classification using *p*-DPLS has been described elsewhere [17]. For a model k that must discriminate between two classes, ω_0 and ω_1, it first regresses **X** on **y** using PLS, where **X** contains objects that belong to both classes and **y** contains a 0 when the object belongs to class ω_0 and a 1 when the object belongs to class ω_1. The model is then used to predict the training set and the fitted y's, $\hat{y}$, are used to estimate a probability density function (PDF) for each class. For an unknown object, its predicted value, $\hat{y}_k$, and the confidence interval around the prediction, $\hat{y}_{k,l} \leq \hat{y}_k \leq \hat{y}_{k,r}$ are calculated using the PLS model. In that interval, the area below the PDF for class c is given by:

$$Area_{c,k} = P(\omega_c) \int_{\hat{y}_{k,l}}^{\hat{y}_{k,r}} p(\hat{y}_k \mid \omega_c)\, d\hat{y}_k \tag{1}$$

where $P(\omega_c)$ is the *a priori* class probability in the binary model, calculated as the number of objects of one class divided by the total number of objects of the binary model (I_c/I), and $p(\hat{y}_k|\omega_c)$ is the conditional probability in the binary model k; that is, the probability that an object with prediction $\hat{y}_k$ belongs to class c ("0" or "1"). Then, the model k delivers the *a posteriori* probability of the object to belong to class c dependent of the prediction, $\hat{y}_k$:

$$R_{c,k} = P\left\{\omega_c \mid \hat{y}_{k,l} \leq \hat{y}_k \leq \hat{y}_{k,r}\right\} = \frac{Area_{c,k}}{Area_{0,k} + Area_{1,k}} \tag{2}$$

$R_{c,k}$ is used as the reliability of classification. Note that, for a given object, there are two reliability values, one to belong to class ω_0 and one to belong to class ω_1. In binary classification, the object is assigned to the class for which it has the highest reliability. In multiclass problems, these reliabilities are combined for the different models, as explained below.

3.2.2.2 *Multi-class classification*

The procedure for classifying an unknown object is described here for three classes ($C = 3$) (see also Fig. 1):

1. Calculate the three possible OAO binary *p*-DPLS models by defining **X** that contains only the objects of class ω_1 and class ω_2 (model k=1), only objects of class ω_1 and class ω_3 (model k=2) and only the objects of class ω_2 and class ω_3 (model k=3). The reliability of classification of the object, $R_{c,k}$, is calculated for each class c in each binary model k with its own number of optimal factors. In this way, model 1 provides the reliability that the unknown object belongs to class 1 ($R_{1,1}$) and also the reliability that the object belongs to class 2 ($R_{2,1}$); model 2 provides the reliability that the unknown object belongs to class 1 ($R_{1,2}$) and the reliability that the object belongs to class 3 ($R_{3,2}$); and similarly for model 3.

2. Calculate the combined reliability of classification of the object in class c as

$$\Gamma_c = \frac{\prod R_{c,k}}{\sum_{c=1}^{C} \prod R_{c,k}} \tag{3}$$

   where the product of the reliabilities in the numerator only takes into account the reliability from the models that considered class c. That is to say, for class $c = 1$, the numerator is the reliability to belong to class ω_1 when the object was classified with the model of class ω_1 *vs.* class ω_2 and the reliability to belong to class ω_1 when the object was classified with the model of class ω_1 *vs.* class ω_3. The denominator of eq. (3) normalizes the result so that the reliability is equivalent to a probability.

3. Following the winner-takes-all principle, the object is assigned to the class that has the highest reliability, Γ_c. The classification decision function is then:

$$F(\mathbf{x}) = \arg\max_{c \in \{1,..,C\}} \Gamma_c(\mathbf{x}) \tag{4}$$

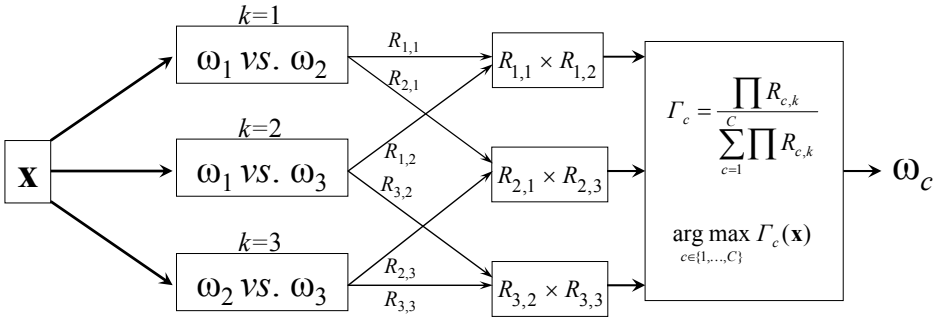

Fig. 1: Calculation of the reliability of classification for a classification problem with C= 3 classes.

3.2.3 Experimental part

3.2.3.1 *Data sets*

The multi-class classification approach is illustrated with two data sets. Data set Iris [18] (Fig. 2) has 150 flowers and 4 variables (length and width of both sepals and petals). There are three classes (Setosa, Versicolor and Virginica) with 50 flowers each. Data set Olive Oil [19] contains 166 Italian olive oils that belong to 5 different regions: Liguria (63), Sicilia (28), Lazio (29), Puglia (28) and Umbria (18). Seven hundred variables were measured, corresponding to the values of absorbance in NIR spectroscopy, measured between 1100 and 2498 nm, with a spectral window of 2 nm.

3.2.3.2 *Procedure and software*

The Kennard–Stone algorithm [20] applied to each class separately was used to split the data set into a training set (with 70% of the objects) and a test set (with 30% of the objects). The data were mean-centered before each *p*-DPLS model was calculated. The strategy OAO was applied and the values of $R_{c,k}$ for each object in each class in the binary models were estimated. The optimal number of factors was selected by leave-one-out cross-validation (LOOCV) from each binary model and taken as the number that provided the model with the highest accuracy of classification, i.e. the number of

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

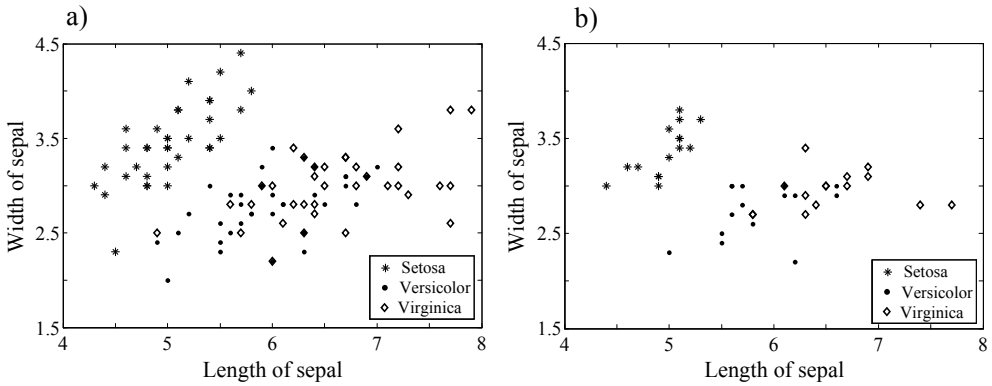

Fig. 2: Training data (a) and test data (b) for the Iris dataset. The variables length of sepal and width of sepal produce the largest overlap between classes.

objects correctly classified in each class, divided by the total number of training objects. This ensures the best results of classification for the two classes in each binary model. The proposed methodology was compared with the CART algorithm using the Gini index as splitting criterion [21] and with the SIMCA algorithm. CART is a simple algorithm for multi-classification that allocates objects in one step. SIMCA is a classification algorithm that calculates, for each class, one model based on principal components (PCs); a new object can be assigned to one class, to several classes or to none of them. An alternative approach is to assign the object to the closest class [22].

Calculations for the DPLS models were done with in-house Matlab (The MathWorks, Inc.) subroutines, CART calculations were run using the Chemometrics Toolkit developed within the TRACE project [19], and SIMCA was run using the PLS_Toolbox 3.5 for MATLAB [22].

3.2.4 Results and discussion

3.2.4.1 *Multi-class classification: iris data set*

Figs 3 and 4 show the predictions ($\hat{y}$) of the three binary models (Virginica *vs.* Versicolor, Virginica *vs.* Setosa and Versicolor *vs.* Setosa) for the training set (LOOCV) and the test set, respectively. The binary models were obtained following the

procedure proposed by *Pérez et al* [17]. The optimal number of factors in each model was: one for the binary model Versicolor–Setosa; one for the model Virginica–Setosa and two for the model Virginica–Versicolor.

Fig. 3 shows that, as expected, the LOOCV predictions of the objects of class ω_1 are around the reference value 1 and the predictions of objects of class ω_0 are around 0. It can also be observed that the predictions of the model Virginica *vs.* Versicolor overlap. This is caused by the partial overlap of these two classes in the original variable space (Fig. 2a).

Fig. 4 shows the predictions of the test set (that includes objects from the three classes) in the three binary models. Each binary model is expected to classify well the objects of the two classes that were modelled, and to provide extreme $\hat{y}$ values for the objects of the third (non-modelled) class, so that these objects would be detected as prediction outliers. This is clearly the case for the Virginica–Versicolor model, in which the (unknown) Setosa objects have extreme predictions, and the predictions for the Versicolor and Virginica objects are distinct enough to enable a good classification. A similar behaviour is observed for the Versicolor–Setosa model although Virginica objects do not have such extreme predictions. The difficulty arises in the model Virginica–Setosa, where the Versicolor test objects have predictions between the Virginica and Setosa classes. In this case, both the Virginica and Versicolor objects would be classified in the Virginica class if only that binary model was considered.

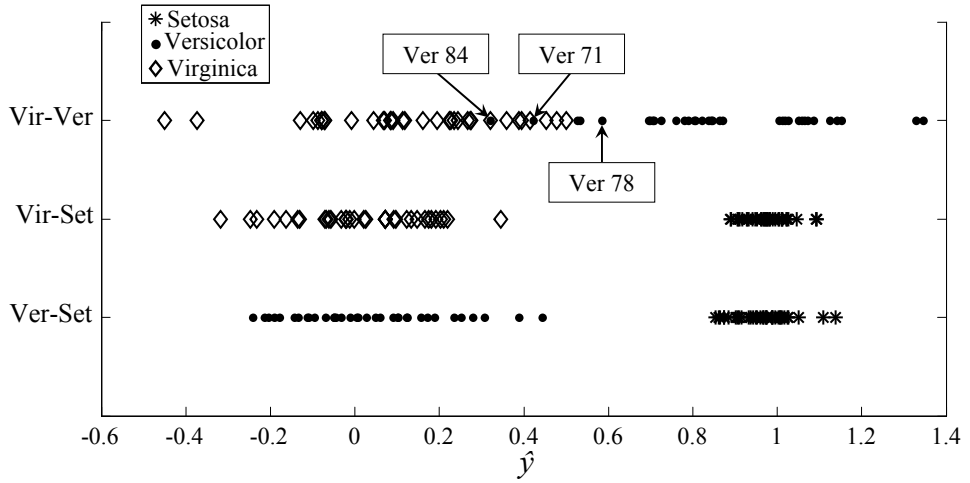

Fig. 3: Iris dataset. Predictions by LOOCV of the three optimal binary DPLS models: Virginica–Versicolor, Virginica–Setosa and Versicolor–Setosa.

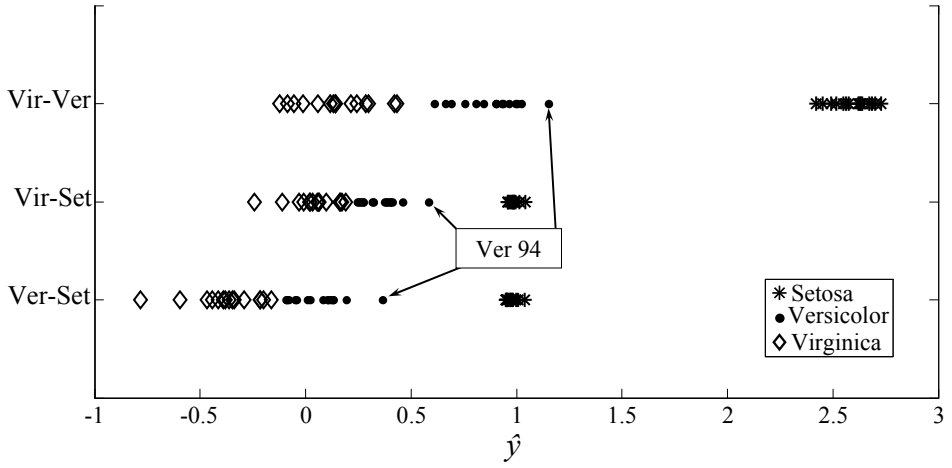

Fig. 4: Iris dataset. Predictions of the test set for the three optimal binary DPLS models: Virginica–Versicolor, Virginica–Setosa and Versicolor–Setosa.

The application of the multi-class model, by combining the results of the three binary models is commented next. The optimal multi-class model is the combination of the optimal binary models (one-factor model Versicolor–Setosa; one-factor model Virginica–Setosa and two-factors model Virginica–Versicolor) and is used to calculate the reliability, Γ_c, of the multi-class system using eq. (3), for the test set.

Fig. 5 shows the combined reliability of the classification for the 45 test objects in each of the three classes. Fig. 5a shows the combined reliability to belong to class Setosa, obtained from the reliabilities calculated from the models Versicolor–Setosa and Virginica–Setosa. As expected, all the Setosa objects (objects 1–15) have a reliability of 100% to belong to class Setosa (Fig. 5a) and 0% reliability to belong to the other two classes (Fig. 5b and 5c), so they are correctly assigned to the class Setosa. For the Versicolor objects (objects 16–30), the multi-class model indicates that there is a 0% reliability of belonging to the Setosa class (Fig. 5a), but gives a range of reliabilities between 60% and 100% for class Versicolor (Fig. 5b). In any case, all these objects have a higher reliability to belong to the Versicolor class than to belong to the Setosa class (Fig. 5a) or the Virginica class (Fig. 5c), so they are all correctly assigned to the class Versicolor. The objects of the Virginica class (objects 31–45) behave like the Versicolor objects. The model gives a reliability of 0% for these objects to belong to Setosa (Fig. 5a), reliabilities lower than 50% to belong to the Versicolor class and reliabilities higher than 50% to belong to the Virginica class (Fig. 5c).

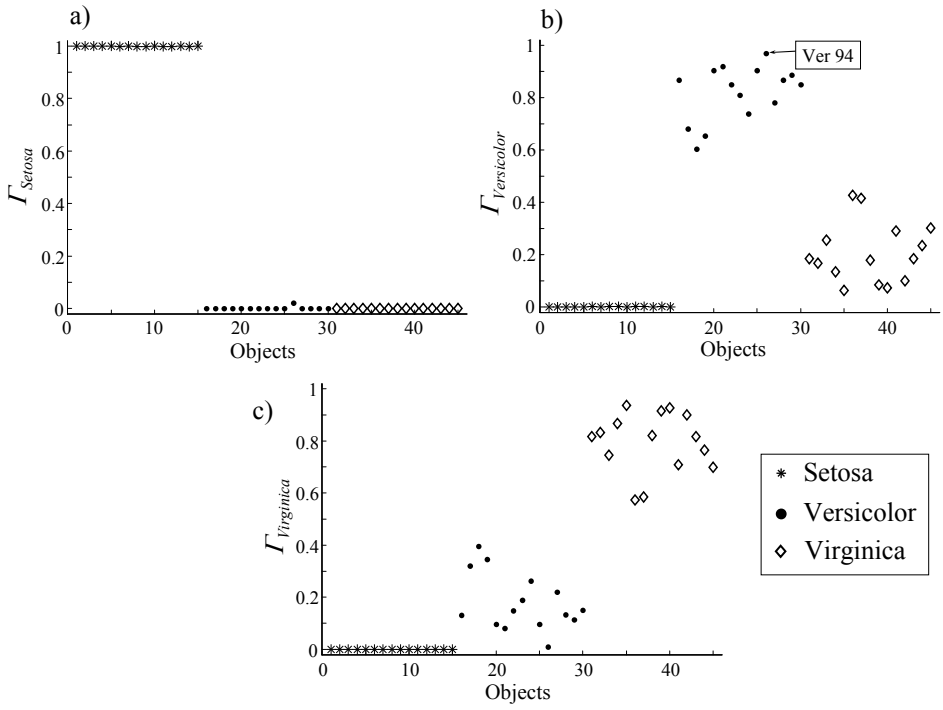

Fig. 5: Iris dataset. Multiclass reliability Γ_c, for the test objects: (a) reliability for class Setosa; (b) reliability for class Versicolor and (c) reliability for class Virginica.

To better illustrate the procedure, Table 1 shows the results for the Versicolor 94 object. The Versicolor–Setosa model indicates that the object belongs to Versicolor with a reliability of 0.97 and that it belongs to Setosa with a reliability of 0.03. So, by considering this model only, the object would be assigned to Versicolor (Fig. 4). Moreover, the Virginica–Versicolor model also indicates that it is more probable that the object belongs to Versicolor than to Virginica, by assigning a higher reliability to the Versicolor class. When the object is submitted to the Virginica–Setosa model (Fig. 4) the object is undetected by outlier diagnostics and the model assigns the object to Setosa with a reliability of 0.76. Hence, according to two of the three OAO models, the object belongs to Versicolor, while according to the Virginica–Setosa model, the object is of Setosa class. Similar ambiguous classification results are common when multiple binary OAO classification models are calculated. The criterion proposed in eqs. (3) and (4), that uses the product of reliabilities followed by normalization, gives the reliability of the multi-class model, which is 97% for Versicolor, and of 2% for Setosa and 1% for Virginica. So, following the winner-takes-all, the object is assigned to the Versicolor class.

Table 1: Classification results when the Versicolor 94 object is submitted to the three binary DPLS models: Versicolor–Setosa, Virginica–Setosa and Virginica–Versicolor.

Class	Reliability (R)			ΠR	Γ	Assigned class
	Ver – Set	Vir – Set	Vir – Ver			
Setosa	0.03	0.76	–	0.02	0.02	No
Versicolor	0.97	–	0.96	0.93	0.97	Yes
Virginica	–	0.24	0.04	0.01	0.01	No

Note that the proposed criterion (eq. (3) and (4)) does not ignore the result of the Virginica–Setosa model, but evaluates the probability for each class by combining the outputs of the base classifiers, i.e., using the responses of the base classifiers even though these are not correct. So, when a class has a high probability, the classifiers that model that class assign a high reliability; and the other way round, when the class has a low probability, the classifiers assign a low reliability. Thus, if any of the classifiers incorrectly assigns a high reliability to a given class, this is minimized by the other classifiers that assign a low reliability to the class, which is the case of the Versicolor object in the classifier Virginica–Setosa.

The results of combining the three OAO binary models with the multi-class *p*-DPLS methodology are shown in Table 2. This methodology classified correctly 100% of the objects of the Setosa and Virginica classes. This was to be expected because of the good separation of the Setosa objects from the other objects, already observed in the initial data (Fig. 2). This separation was even better observed in the plot of the predictions (Fig. 3). More interestingly, the results were excellent for the Virginica objects, which are partially overlapped with the Versicolor objects in the original variable space (Fig. 2); the *p*-DPLS predictions overlap much less (Fig. 3 and 4). Approximately 94% of the Versicolor objects were correctly assigned. Only two objects were wrongly assigned to the Virginica class but with reliabilities of 59% and 70%, respectively. These low reliabilities can warn the experimenter about the classification decision. For the test data set, the multi-class model correctly allocates the objects of the three classes with 100% reliability.

Table 2: Iris dataset. Classification performance for the multi-class *p*-DPLS, CART and SIMCA models.

Real class	Data set	Assigned class								
		Setosa			Versicolor			Virginica		
		p-DPLS	CART	SIMCA	*p*-DPLS	CART	SIMCA	*p*-DPLS	CART	SIMCA
Setosa	Training	100%	100%	100%	0%	0%	0%	0%	0%	0%
	Test	100%	100%	100%	0%	0%	0%	0%	0%	0%
Versicolor	Training	0%	0%	0%	94.3%	91.4%	94.3%	5.7%	8.6%	5.7%
	Test	0%	0%	0%	100%	100%	93.3%	0%	0%	6.7%
Virginica	Training	0%	0%	0%	0%	0%	0%	100%	100%	100%
	Test	0%	0%	0%	0%	0%	0%	100%	100%	100%

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

The results of the multi-class *p*-DPLS method were similar to the ones obtained with the optimal CART model with five terminal nodes (Table 2) and also similar to the results for the SIMCA models using three PCs for Setosa, two PCs for Versicolor and two PCs for Virginica (Table 2). The percentage of correct classification is 100% for Setosa and Virginica classes, both for the training data and the test data. For the Versicolor class the percentage of correct classification for the training data with both *p*-DPLS and SIMCA (~ 94%, two objects, Versicolor 71 and Versicolor 84, wrongly classified as Virginica) was higher than with CART (~ 91%, three objects wrongly classified as Virginica). Versicolor 78 was also wrongly classified by CART but correctly with *p*-DPLS, although with a very low reliability (57.4%) because this object is close to the boundary between classes (Fig. 3). In addition, SIMCA also misclassified the object Versicolor 73, that, again, is correctly classified by *p*-DPLS but with a very low (52%) reliability. For the test data CART classified correctly 100% of the samples of each class. Similarly, SIMCA classified correctly 100% of the Setosa and Virginica samples, and 93.3% of the Versicolor samples. The object Versicolor 69 was wrongly assigned to Virginica by SIMCA but it was correctly assigned by the *p*-DPLS although also with a low reliability (65%).

3.2.4.2 *Multi-class classification: olive oil data set*

For the olive oil training set, 10 binary models OAO were calculated and validated by LOOCV. The optimal number of factors was selected considering the total accuracy of classification, i.e. the objects correctly classified in the two classes of each binary model, divided by the number of training objects of each binary model. The optimal number of factors was: 16 for Liguria *vs.* Sicilia, 23 for Liguria *vs.* Puglia, 16 for Liguria *vs.* Lazio, 14 for Liguria *vs.* Umbria, 17 for Sicilia *vs.* Puglia, 14 Sicilia *vs.* Lazio, 8 Sicilia *vs.* Umbria, 15 for Puglia *vs.* Lazio, 6 for Puglia *vs.* Umbria and 5 for Lazio *vs.* Umbria. The high number of factors needed in some models is due to the very similar spectra between the classes.

Table 3 shows that the best classifications by the multi-class approach are for the training objects of the Liguria class, followed by classes Lazio, Umbria and Sicilia and finally Puglia. The results for the test set improve between 5% and 15% in relation to the training set, except for class Liguria where it is a 3% lower. It is noteworthy that the reliabilities (not shown) for all the objects wrongly assigned are above 74%, except for the Umbria class, whose training objects are wrongly assigned to Liguria and Puglia with just a 51% reliability. The low reliability of the objects wrongly classified of the class Umbria suggests that they were near the limits of the class.

Table 3. Olive oil dataset: Classification performance for the multi-class *p*-DPLS method. Training (LOOCV) and test data set.

Assigned class	Data set	Real class				
		Liguria	Sicilia	Puglia	Lazio	Umbria
Liguria	Training	97.7%	0.0%	15.0%	0.0%	7.7%
	Test	94.7%	0.0%	12.5%	0.0%	0.0%
Sicilia	Training	0.0%	84.2%	10.0%	0.0%	0.0%
	Test	0.0%	100%	0.0%	0.0%	0.0%
Puglia	Training	2.3%	5.3%	70.0%	15.0%	7.7%
	Test	5.3%	0.0%	75.0%	0.0%	0.0%
Lazio	Training	0.0%	5.3%	5.0%	85.0%	0.0%
	Test	0.0%	0.0%	12.5%	100%	0.0%
Umbria	Training	0.0%	5.3%	0.0%	0.0%	84.6%
	Test	0.0%	0.0%	0.0%	0.0%	100%

Table 4 presents the values of reliability of each binary model for two samples of the test set, an oil of the Liguria class wrongly assigned to Puglia and an oil of the Sicilia class assigned correctly. For the sample Liguria 014, only two of the four models involving the class Liguria (Lig-Laz and Lig-Umb) predicted the oil as Liguria; the other two assigned the oil to either Puglia or Sicilia (although the last one with a very low reliability, 58%). However, four models assigned the oil to Puglia with 100% reliability and other two models to Lazio. The assignments to Puglia class are significant because the model Liguria–Puglia (wrongly) identifies with a 100% of probability that the oil is from Puglia. The same happens with the Sicilia–Puglia model, which assigns the oil to Puglia with a reliability of 99%. So, when the multi-class *p*-DPLS combines the binary models, it gives higher significance to those that contain Puglia and diminish the significance of the others.

For oil Sicilia 134, the four models involving the class Sicilia predicted the oil as belonging to Sicilia, three of them with a 100% reliability and the other with 91% reliability. Of the rest of the models, three of them assign the oil to Umbria (although one with 54% of reliability) and the other two to Puglia. Thus, the combined results lower the significance of the Puglia and Umbria results and give all the importance to the Sicilia class.

Table 4. Olive oil dataset. Reliability of classification for individual binary *p*-DPLS models and combined reliability in the multi-class model for test objects Liguria 014 and Sicilia 134.

Reliability (*R*)	Class									
	Liguria 014					Sicilia 134				
	Liguria	Sicilia	Puglia	Lazio	Umbria	Liguria	Sicilia	Puglia	Lazio	Umbria
Lig-Sic	0.42	0.58	–	–	–	0.00	1.00	–	–	–
Lig-Pug	0.00	–	1.00	–	–	0.00	–	1.00	–	–
Lig-Laz	0.97	–	–	0.03	–	0.01	–	–	0.99	–
Lig-Umb	0.82	–	–	–	0.18	0.00	–	–	–	1.00
Sic-Pug	–	0.01	0.99	–	–	–	1.00	0.00	–	–
Sic-Laz	–	0.00	–	1.00	–	–	1.00	–	0.00	–
Sic-Umb	–	0.99	–	–	0.01	–	0.91	–	–	0.09
Pug-Laz	–	–	0.99	0.01	–	–	–	0.99	0.01	–
Pug-Umb	–	–	1.00	–	0.00	–	–	0.06	–	0.94
Laz-Umb	–	–	–	1.00	0.00	–	–	–	0.46	0.54
ΠR	0.00	0.00	0.98	0.00	0.00	0.00	0.91	0.00	0.00	0.05
Γ	0.00	0.00	1.00	0.00	0.00	0.00	0.95	0.00	0.00	0.05
Assigned class	No	No	Yes	No	No	No	Yes	No	No	No

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T-990-2010

The optimal CART model for olive oil data contained 19 terminal nodes. By comparing the results of the CART model (Table 5) with the results of the multi-class *p*-DPLS method (Table 3), it is seen that *p*-DPLS and CART classified correctly between 93% and 97% of the training objects of class Liguria. However, for the test objects *p*-DPLS gave the best classification results (94.7% versus 57.9% in CART). In contrast, the Lazio training objects are better classified by CART than by *p*-DPLS, while the test objects are better classified by *p*-DPLS than by CART. The higher difference in classification between CART and *p*-DPLS is observed in Sicilia and Umbria classes, where training data are better classified by *p*-DPLS (above 84%) than by CART (below 58%). The difference increases with the test data, where *p*-DPLS gets 100% correct classifications against a maximum of 40% for CART. For the Puglia class, both *p*-DPLS and CART do not reach more than 75% of correct classification for training data, but for the test data *p*-DPLS provided a better percentage of classification than CART. This significant difference in the classification performances can be attributed to the lack of variables with a high discriminative power. This makes it more difficult for CART to separate the classes at each node. In contrast, *p*-DPLS finds significant factors with a greater capacity for discrimination between classes. These factors, when applied to unknown objects, allow a better classification of them.

Table 5: Olive oil dataset: classification performance for the CART model.

Assigned class	Data set	Real class				
		Liguria	Sicilia	Puglia	Lazio	Umbria
Liguria	Training	93.2%	36.8%	0%	0%	46.2%
	Test	57.9%	22.2%	25.0%	33.3%	40.0%
Sicilia	Training	0%	57.9%	0%	0%	0%
	Test	10.5%	33.3%	0%	0%	20.0%
Puglia	Training	2.3%	0%	75.0%	5.0%	0%
	Test	10.5%	11.1%	50.0%	0%	0%
Lazio	Training	4.5%	5.3%	25.0%	95.0%	0%
	Test	15.8%	11.1%	25.0%	66.7%	0%
Umbria	Training	0%	0%	0%	0%	53.8%
	Test	5.3%	22.2%	0%	0%	40.0%

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

The optimal models for SIMCA have 20 factors for the Liguria class, 10 factors for Sicilia, 12 factors for Puglia, 14 factors for Lazio and 9 factors for Umbria, found by the minimum prediction error sum of squares (PRESS). Unlike *p*-DPLS, SIMCA correctly assigned 100% of the training data to their respective classes (Table 6). This was not the case for the test objects: objects of the classes Liguria and Umbria, both in *p*-DPLS and in SIMCA obtain the same percentage of classification, 94.7% and 100% respectively. However, while *p*-DPLS wrongly assigns Liguria014 to Puglia, SIMCA wrongly assigns Liguria010 to Sicilia, while this sample is assigned correctly by *p*-DPLS with 99.0% reliability. For the other three classes, Sicilia, Puglia and Lazio, *p*-DPLS performs better than SIMCA. The Lazio objects are of specially interesting, since *p*-DPLS correctly assigned 100% of the them, while SIMCA misclassified all of them. It is also relevant the 75% of correct classifications in Puglia class by *p*-DPLS while SIMCA only classified correctly 25% of them. The Sicilia class is the class that *p*-DPLS and SIMCA classified with the most similar percentages (100% and 89%, respectively).

When comparing the results for the class Liguria (Table 7), it was observed that there was no correspondence between the objects wrongly classified by the three methods. The training object that *p*-DPLS wrongly assigned to Puglia with a 100% of reliability

Table 6: Olive oil dataset: classification performance for the SIMCA model.

Assigned class	Data set	Real class				
		Liguria	Sicilia	Puglia	Lazio	Umbria
Liguria	Training	100%	0%	0%	0%	0%
	Test	94.7%	11.1%	62.5%	100%	0%
Sicilia	Training	0%	100%	0%	0%	0%
	Test	5.3%	88.9%	12.5%	0%	0%
Puglia	Training	0%	0%	100%	0%	0%
	Test	0%	0%	25.0%	0%	0%
Lazio	Training	0%	0%	0%	100%	0%
	Test	0%	0%	0%	0%	0%
Umbria	Training	0%	0%	0%	0%	100%
	Test	0%	0%	0%	0%	100%

was correctly assigned to Liguria by CART and SIMCA. Reviewing the structure of the classification tree in CART, the node that assigns this object decides between Liguria and Puglia and in this case it is finally assigned to Liguria. In contrast, three training Liguria oils were wrongly assigned by CART to Puglia and Lazio, whereas *p*-DPLS assigned them correctly to Liguria with 100% reliability. A similar pattern was observed with the test data, where the *p*-DPLS allocated an oil of Liguria to Puglia with a 100% of reliability, while the CART and SIMCA models classified it correctly to Liguria. The CART model wrongly allocated eight objects of the Liguria class to the other four classes (two oils to Sicilia, two to Puglia, three to Lazio and one to Umbria); while the proposed *p*-DPLS method allocated them correctly, seven of them with more than 99% reliability and one with 70% reliability. The object wrongly assigned by SIMCA was also correctly assigned by *p*-DPLS with a 99% of reliability.

Table 7: Olive oil data set: classification results of representative oils of the Liguria class.

	Class and sample number	Assigned class		Multi-Class *p*-DPLS	
		CART	SIMCA	Assigned class	Reliability (%)
Training	Liguria 016	Puglia	Liguria	Liguria	100
	Liguria 025	Lazio	Liguria	Liguria	100
	Liguria 028	Lazio	Liguria	Liguria	100
	Liguria 081	Liguria	Liguria	Puglia	100
Test	Liguria 004	Sicilia	Liguria	Liguria	100
	Liguria 008	Lazio	Liguria	Liguria	100
	Liguria 010	Umbria	Sicilia	Liguria	99
	Liguria 014	Liguria	Liguria	Puglia	100
	Liguria 015	Lazio	Liguria	Liguria	70
	Liguria 020	Puglia	Liguria	Liguria	100
	Liguria 022	Lazio	Liguria	Liguria	100
	Liguria 038	Puglia	Liguria	Liguria	100
	Liguria 040	Sicilia	Liguria	Liguria	100

3.2.5 Conclusions

We have proposed a methodology for multi-class classification problems based on combining the outputs of binary classification *p*-DPLS models. This method calculates the reliability of multi-class classification as the product of *a posteriori* probability of the binary models that contain the class of interest. For the tested datasets, the method gave better results than CART, because the reduction of variables into significant factors takes into account all the data variability, and not only the variables that allow the best separation at each of the nodes in the CART model. *p*-DPLS also have better results with the test set than SIMCA, possibility because DPLS maximizes the variance–covariance between the spectra and the class number, while SIMCA considers the spectral information of each class separately. Some of these qualities were observed in the results of classification of the olive oil data set, where the multi-class *p*-DPLS model had better classification performance for the test set, from 75 to 100%, than CART and SIMCA.

Acknowledgments

The authors express their gratitude to Dr. Gerard Downey (Teagasc, Ireland) for providing the Olive oil dataset. This work was supported by the Project TRACE – "TRAcing food Commodities in Europe (EU IP 006942)" – from the Sixth Framework Programme of the European and by the Spanish Ministerio de Educación y Ciencia project CTQ2007-66918/BQU. This paper reflects only the authors' views and neither the Community nor the Spanish Ministerio are liable for any use that may be made of the information contained therein.

References

1. M. Pardo, G. Sberveglieri, A. Tarino, F. Maulli, G. Valentini, Anal. Chim. Acta. 446 (2001) 223–232.
2. H. S. Tapp, M. Defernez, E. K. Kemsley, J. Agric. Food Chem. 51 (2003) 6110–6115.
3. C. H. Yeang, S. Ramaswamy, P. Tamayo, S. Mukherjee, R. M. Rifkin, M. Angelo, M. Reich, E. Lander, J. Mesirov, T. Golub, Bioinformatics. 17, Suppl. 1 (2001) S316–S322.

4. A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, S. Levy, Bioinformatics. 21 (2005) 631–643.
5. N. Garcia-Pedrajas, D. Ortiz-Boyer, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1001–1006.
6. B. K. Alsberg, R. Goodacre, J. J. Rowland, D. B. Kell, Anal. Chim. Acta. 348 (1997) 389–407.
7. Z. Ramadan, X.-H. Song, P. K. Hopke, M. J. Johnson, K. M. Scow, Anal. Chim. Acta. 446 (2001) 233–244.
8. G. Downey, P. McIntyre, A. N. Davies, J. Agric. Food Chem. 50 (2002) 5520–5525.
9. A. J. Myles, S. D. Brown, J. Chemometr. 18 (2004) 286–293.
10. U. Kreßel, Pairwise classification and support vector machines, in: B. Schölkopf, C. J. C. Burges, A. J. Smola (Eds.), Advances in Kernel Methods-Support Vector Learning, Cambridge, MA: MIT Press, 1999, p. 255–268.
11. J. H. Oh, Y. B. Kim, P. Gurnani, K. P. Rosenblatt, J. X. Gao, Bioinformatics. 24 (2008) 1812–1818.
12. T. G. Dietterich, G. Bakiri, J. Artif. Intell. Res. 2 (1995) 263–286.
13. G. Ou, Y. L. Murphey, Pattern Recogn. 40 (2007) 4–18.
14. L. I. Kuncheva, "Combining Pattern Classifiers: Methods and Algorithms", A Wiley-Interscience publication, 2004, pp: 112–125.
15. T. Windeatt, R. Ghaderi, Inform. Fusion. 4 (2003) 11–21.
16. P. He, K.-T. Fang, Y.-Z. Liangc, B.-Y. Li, Anal. Chim. Acta. 543 (2005) 181–191.
17. N. F. Pérez, J. Ferré, R. Boqué. Chemometr. Intell. Lab. Syst. 95 (2009) 122–128.
18. ftp://ftp.ics.uci.edu/pub/machine-learning-databases/iris/ Last accessed on 21 January 2009.
19. Project TRACE – "TRAcing food Commodities in Europe" (project no. FOOD-CT-2005-006942). www.trace.eu.org.
20. R. W. Kennard, L. A. Stone, Technometrics 11 (1969) 137–148.
21. S. Caetano, B. Üstün, S. Hennessy, J. Smeyers-Verbeke, W. Melssen, G. Downey, L. Buydens, Y. Vander Heyden, J. Chemometr. 21 (2007) 324–334.
22. B. M. Wise, N. B. Gallagher, R. Bro, J. M. Shaver, W. Winding , R. S. Koch , PLS_Toolbox Version 3.5 for use with MATLABTM, Eigenvector Research, Inc., Manson, WA, USA, 2005, pp 180–184.

3.3 Multiclasificación de suelos europeos por litologías

3.3.1 Introducción

En el apartado 2.7 se aplicó el método de clasificación binaria *p*-DPLS al conjunto de datos de suelos europeos utilizando la estrategia de binarización "uno contra todos" (OAA). Aunque los resultados de clasificación fueron buenos (se superó el 95% de clasificación correcta) al igual que las fiabilidades (entre el 60% y 100% para las tres clases), se tiene el inconveniente de que pueden haber inconsistencias entre modelos. La estrategia no considera la posibilidad de que hayan objetos mal asignados por los tres modelos o que los objetos sean asignados a más de una clase. Por ello, se retomó este conjunto de datos y se aplicó el método propuesto *p*-DPLS multiclase (apartado 3.2). Con ello se pretende estudiar qué ventajas presenta la combinación de los modelos binarios OAO y obtener mejoras tanto en el porcentaje de clasificación como de fiabilidad de clasificación.

3.3.2 Parte experimental

Para este estudio se utilizó el conjunto de datos de suelos europeos (detalles en el anexo, apartado 1) con 180 objetos, 75 variables, concentraciones de elementos comunes y traza, y 3 clases: *arenisca* (roca sedimentaria clástica), *caliza* (roca sedimentaria química) y *esquisto* (roca metamórfica). Se aplicó el algoritmo *Kennard-Stone* (a datos autoescalados) para dividirlo en datos de entrenamiento (70% de los datos) y datos de prueba (30% de los datos). Se desarrollaron tres modelos siguiendo la estrategia "uno contra uno": *arenisca* (ω_1) *vs*. *caliza* (ω_2); *arenisca* (ω_1) *vs*. *esquisto* (ω_3) y *caliza* (ω_2) *vs*. *esquisto* (ω_3). El número óptimo de factores se decidió por validación cruzada dejando fuera un objeto cada vez (LOOCV), con criterio de máxima clasificación total (objetos de las dos clases bien asignados sobre el total de objetos considerados en el modelo binario). Los modelos se desarrollaron con datos autoescalados y como función de decisión de clasificación se utilizó la ecuación 17 del apartado 2.2.2.5, para un intervalo incertidumbre $2\times SEP$, intervalo en el cual se calcula el área bajo las FDP de las clases.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

3.3.3 Resultados y discusión

Un modelado preliminar detectó que los objetos *arenisca*085 y *esquisto*115 son discrepantes. El objeto *arenisca*085 no fue asignado a ninguna de la clases por *p*-DPLS multiclase, y el objeto *esquisto*115 fue erróneamente considerado *caliza*. Estos objetos también se consideraron discrepantes en el apartado 2.7. Los modelos binarios óptimos, siguiendo la estrategia OAO fueron: para *arenisca vs. caliza* y *arenisca vs. esquisto*, 2 factores, y *caliza vs. esquisto* 7 factores. Con los modelos OAO únicamente se tuvo en cuenta el criterio de selección de máxima clasificación total, dado que se trata de modelos binarios puros y este criterio asegura obtener los mejores resultados para las dos clases.

Los gráficos de *scores* (Figura 3-1 a, c y e) muestran el comportamiento de las clases una frente a la otra en los modelos binarios. En el modelo *arenisca vs. caliza* (Figura 3-1a) con sólo un factor se pueden separar adecuadamente las clases y se explica cerca del 90% de varianza en *y*. Similar con el modelo *arenisca vs. esquisto* (Figura 3-1c) que separa adecuadamente las clases, pero con el 84% de varianza explicada en *y*. Aunque en el modelo *caliza vs. esquisto* (Figura 3-1e) se observa una separación de las clase con los dos primeros factores (el primer factor separa parcialmente las clases) se tienen un objeto *esquisto* en la zona de la clase *caliza* que puede afectar el modelo. Si compramos estas gráficas de *scores* con las de los modelos OAA (apartado 2.7) se observa una mejor separación de las clases *caliza* y *esquisto*.

Las gráficas de coeficientes de regresión muestran qué variables son determinantes para separar los pares de clases. En el modelo *arenisca vs. caliza* se pueden observar 15 variables determinantes (Figura 3-1b) principalmente la materia orgánica (1), carbonatos (2), $CaCO_3$ (3), Ca (9), Na (10), ^{181}Ta (70) y ^{208}Pb (72), siendo 3 de ellas (carbonatos, $CaCO_3$ y Ca) características de la *caliza*. El modelo *arenisca vs. esquisto* tiene 17 variables determinantes (Figura 3-1d), Na (10), ^{23}Na, ^{91}Zr y 13 de ellas agrupadas entre las variables 55 a 68 (isótopos traza). Estas 13 variables agrupadas se repite en el modelo *arenisca vs. caliza*, por lo que dichos isótopos pueden ser característicos de la clase *arenisca*. El modelo *caliza vs. esquisto* tiene 10 variables determinantes (Figura 3-1f), ^{71}Ga (40), ^{151}Eu (60), ^{181}Ta (70), las dos últimas fueron determinantes en el modelo *arenisca vs. caliza*, dado que la *caliza* es la clase común podemos atribuir estos elementos a dicha clase.

La tabla 3-1 muestra los porcentajes de clasificación cuando se utilizó el método *p*-DPLS multiclase. Para los datos de entrenamiento (LOOCV) se obtuvo un 100% de clasificación para las clases *arenisca* y *caliza* y un 97.6% para la clase *esquisto* (el objeto *esquisto*106 fue asignado erróneamente a *arenisca*). Con los datos de prueba se obtuvo un

100% de clasificación para todas las clases. Estos porcentajes son comparables con los observados al utilizar *p*-DPLS con estrategia OAA (apartado 2.7), aunque en ésta el porcentaje para la clase *esquisto* fue del 95.1%. La mejora entre una estrategia y otra, OAA frente a OAO, está en que se discriminan mejor las clases *caliza* y *esquisto* con la estrategia OAO. Además, se sabe que el objeto *esquisto*106 es erróneamente asignado a la clase *arenisca*, coincidiendo con la proximidad del objeto a esta clase observada en la gráfica de *scores* del modelo *arenisca vs. esquisto* (Figura 3-1c).

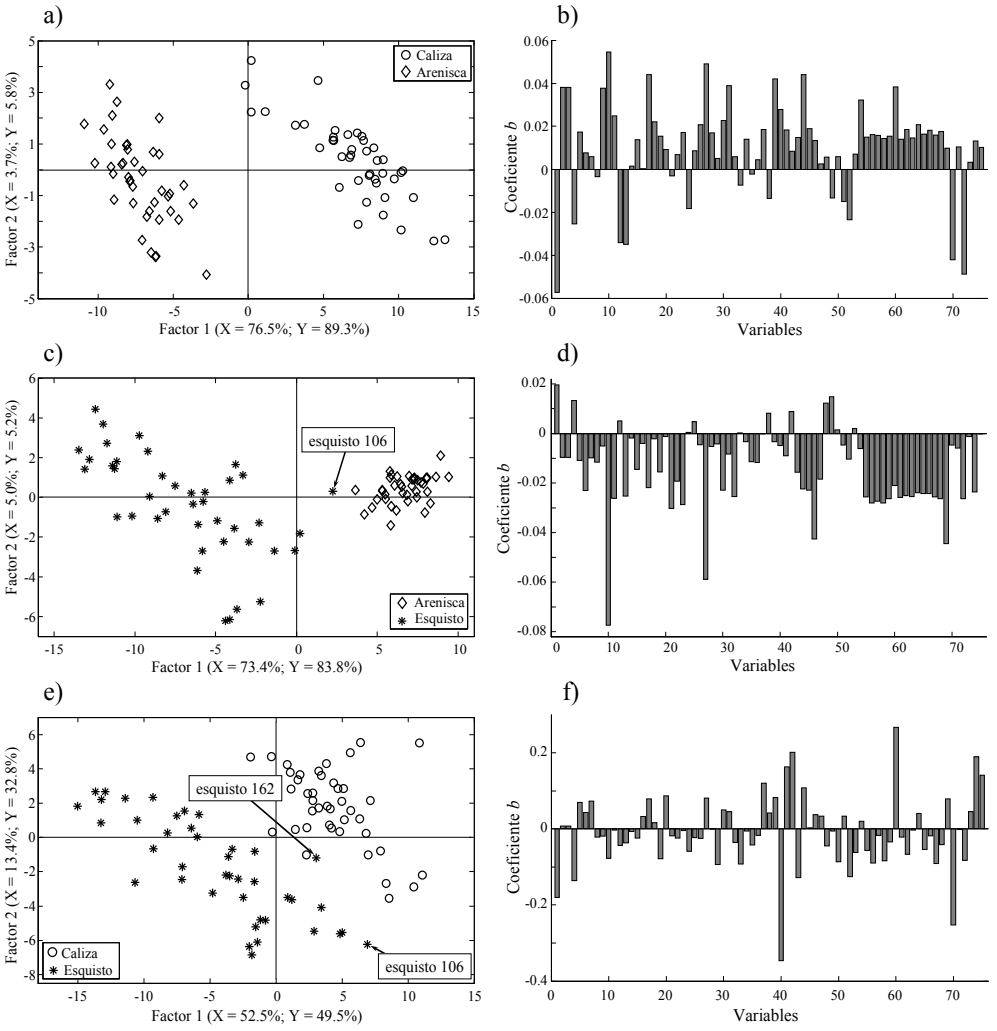

Figura 3-1: Gráficas de *scores* (dos primeros factores) y coeficientes *b* de los modelos binarios óptimos de la estrategia OAO, a) y b) *arenisca vs. caliza*. c) y d) *arenisca vs. esquisto*. e) y f) *caliza vs. esquisto*.

Tabla 3-1: Porcentajes de clasificación *p*-DPLS multiclase con estrategia OAO para el conjunto de datos de suelos.

Decisión Multiclase	Conjunto de datos	Clases Reales		
		Caliza	Arenisca	Esquisto
Caliza	Entrenamiento	100%	0%	0%
	Prueba	100%	0%	0%
Arenisca	Entrenamiento	0%	100%	2.4%
	Prueba	0%	100%	0%
Esquisto	Entrenamiento	0%	0%	97.6%
	Prueba	0%	0%	100%

La fiabilidad de clasificación multiclase de pertenecer a cada una de las clases para los objetos de entrenamiento (LOOCV) se muestra en la figura 3-2. Se observa que todos los objetos de la clase *caliza* (objetos 1 a 42) obtuvieron un 100% de fiabilidad de pertenecer a su respectiva clase (Figura 3-2b). Por contraste, tanto los objetos de la clase *arenisca* (objetos 43 a 84) como *esquisto* (objetos 85 a 126) obtuvieron fiabilidades entre el 70% al 100% (Figura 3-2 a y c). Cabe destacar el objeto *esquisto* 106, que es erróneamente asignado a *arenisca* con el 85.5%, y el *esquisto* 162 que tiene un 63.8% de fiabilidad de pertenecer a *esquisto*. Los objetos de prueba obtuvieron una mayor fiabilidad. Los objetos de la clase *caliza* y *esquisto* presentaron una fiabilidad del 100% de pertenecer a su respectiva clase, mientras que los objetos de la clase *arenisca* presentan un intervalo de fiabilidades de pertenecer a su respectiva clase de 93% al 100%. En general, las fiabilidades fueron elevadas y, en el caso de los objetos con fiabilidades bajas, éstas se corresponden con la ubicación vista en el espacio de *scores* (Figura 3-1), así la fiabilidad es menor cuando más cerca se está del límite entre clases.

Al comparar las fiabilidades del método *p*-DPLS multiclase con las fiabilidades obtenidas utilizando la estrategia OAA (apartado 2.7), se observa que las mayores diferencias las presentan los objetos de la clase *arenisca*; pues con la estrategia OAA se obtuvieron fiabilidades del 60 al 100% con los datos de entrenamiento y del 80% al 100% con los datos de prueba. Por el contrario, con el método *p*-DPLS multiclase las fiabilidades fueron del 86% al 100% en objetos de entrenamiento y del 93% al 100% con los de prueba, con la mayoría de los objetos cerca del 100%. Algo similar ocurrió con las clases *caliza* y *esquisto,* donde con la estrategia OAA se obtuvieron fiabilidades entre el 60% al 100% para los objetos cercanos al límite entre clases. En contraste, con

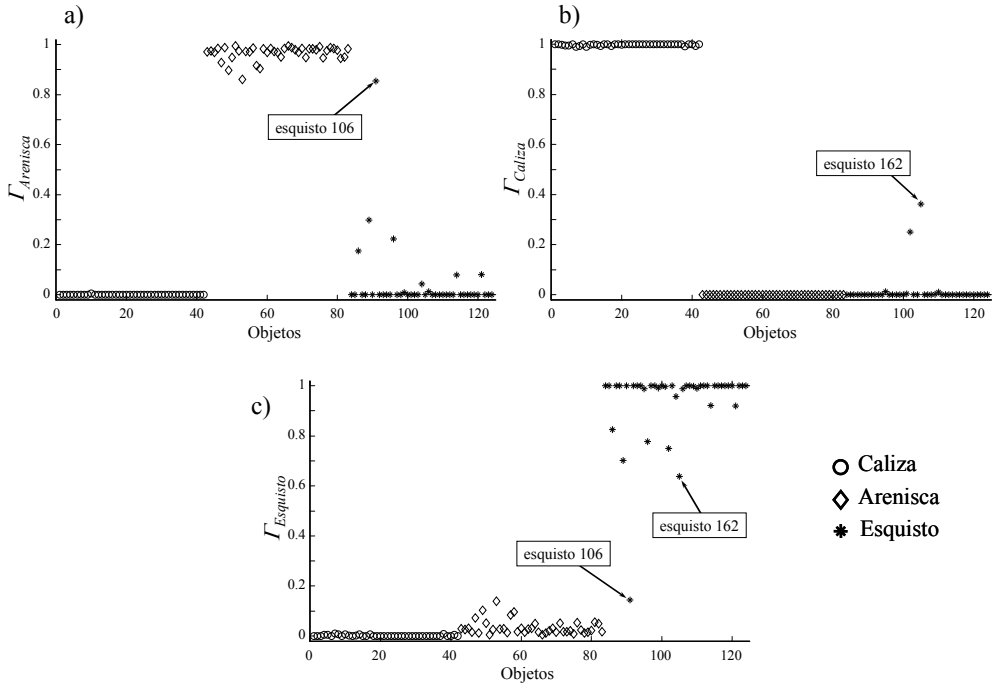

Figura 3-2: Gráfica de fiabilidades multiclase, Γ_c, para los objetos de entrenamiento (LOOCV). Fiabilidad de pertenecer a la clase: a) *arenisca*, b) *caliza*, c) *esquisto*.

p-DPLS multiclase las fiabilidades fueron del orden del 100%. Este aumento en la fiabilidad de clasificación se debió, por un lado, a utilizar la estrategia OAO, que compara únicamente dos clases, mejorando los límites y reduciendo el *SEP*. Así, cada modelo binario puede dar mayores fiabilidades de clasificación para los objetos. Por otro lado, el aumento de la fiabilidad se debe a la estrategia de combinación, que permite calcular la fiabilidad multiclase, minimizando el efecto de las asignaciones erróneas de los modelos binarios.

3.3.4 Conclusiones

La estrategia de combinación *p*-DPLS multiclase se muestra más adecuada para evaluar varias clases simultáneamente. Aunque se obtuvieron resultados de clasificación similares a los de la estrategia OAA, se mejoran las fiabilidades de clasificación, siendo

estas últimas determinantes para definir qué método es mejor. Hay que tener en cuenta que no importa si el porcentaje de clasificación es elevado, cercano al 100%, ya que si la fiabilidad de clasificación es baja, cercana al 60%, dichas clasificaciones son dudosas, siendo este el caso de los objetos *arenisca* con estrategia OAA. Se busca que tanto los porcentajes de clasificación como las fiabilidades sean cercanos al 100%; dado que fiabilidades elevadas validarán porcentajes de clasificación elevados.

3.4 Métodos de combinación alternativos para modelos *p*-DPLS

En el apartado 3-2 se propuso un nuevo método para resolver problemas multiclase mediante la combinación de modelos binarios *p*-DPLS que tenía como objetivo calcular la fiabilidad de clasificación multiclase, dado que los modelos binarios ya aportan la fiabilidad. Sin embargo, existen otros métodos de combinación ya probados, que aunque no suministran la fiabilidad de clasificación son adecuados para combinar modelos binarios *p*-DPLS, ya sea porque se ajustan a los requerimientos del modelo *p*-PLS o porque el modelo *p*-PLS puede hacer un aporte significativo al método de combinación.

3.4.1 Combinación de modelos *p*-DPLS a partir de la suma de probabilidades de modelos binarios

Este método de combinación de modelos binarios fue propuesto por *Huang et ál.* [1], y asigna un objeto a la clase cuya suma de probabilidades sea mayor:

$$F(\mathbf{x}) = \arg\max_{c\in\{1,..,C\}} p_c \tag{3-2}$$

donde p_c es la suma de probabilidades de que el objeto pertenezca a la clase c, calculadas a partir de los modelos binarios que modelaron esta clase. Si aplicamos esta ecuación con *p*-DPLS, se tiene una ecuación que utiliza las fiabilidades de clasificación del objeto en cada una de las clases y para cada uno de los modelos binarios como un equivalente de probabilidad:

$$F(\mathbf{x}) = \arg\max_{c\in\{1,..,C\}} \sum_{k}^{K} R_{c,k} \tag{3-3}$$

donde $R_{c,k}$, es la fiabilidad de clasificación para un objeto desconocido en una clase c calculada por el modelo k. Como se observa, se trata de un método sencillo que combina los modelos *p*-DPLS óptimos (se deben validar primero los modelos binarios) siguiendo el principio del "*winner-takes-all*"; además de tener en cuenta la posición del objeto desconocido dentro del espacio de variables, pues depende de las respuestas previas de los clasificadores base.

3.4.2 Combinación de modelos *p*-DPLS, tomando la fiabilidad como constante de ponderación en el voto mayoritario ponderado (WMV)

El Voto Mayoritario Ponderado (*Weighted Majority Vote,* WMV) [2] es una evolución del Voto Mayoritario e incluye la medida de fiabilidad del clasificador en la regla de combinación. Esta medida de fiabilidad cambia de acuerdo a la habilidad del clasificador para asignar los objetos a sus respectivas clases. Así, la ecuación de combinación por WMV es:

$$F(\mathbf{x}) = \max_{c=1}^{C} \sum_{k=1}^{K} l_k d_{c,k} \tag{3-4}$$

donde l_k es el coeficiente de ponderación para el clasificador k y $d_{c,k}$ es el voto que el modelo k da al objeto si perteneciera a la clase c (1 si pertenece o 0 si no pertenece) y C es número total de clases. Así, un objeto se clasifica en la clase donde la suma de los votos ponderados es mayor.

Este método asigna más peso a los clasificadores más eficientes, que presentan mejores porcentajes de clasificación, e ignora los clasificadores inadecuados. Dado que la fiabilidad del clasificador depende de los datos de entrenamiento, ésta será constante para todo objeto nuevo. La fiabilidad de clasificación puede ser igual al porcentaje de clasificación, o puede calcularse utilizando métodos de remuestreo. En este último caso se desarrollan varios modelos, cada uno es un remuestreo distinto, y la fiabilidad se calcula como el promedio de los porcentajes de clasificación de cada uno de los modelos desarrollados. Sin embargo, cuando se quiere clasificar un objeto se pueden presentar dos casos: 1) tener un clasificador confiable pero un objeto poco confiable, por estar cerca del límite entre clases; y 2) tener un clasificador poco confiable pero un objeto confiable, porque está ubicado en el espacio de variables de la clase. En el primer caso se dará demasiado peso al objeto poco confiable, es decir, que se confiaría en una asignación que debería tratarse con cuidado dada la cercanía al límite entre clases. En el segundo caso se le resta peso al objeto que sí es confiable, es decir, que le restaríamos confianza a la asignación cuando ésta es totalmente confiable dada la posición del objeto en el espacio de variables.

Para combinar modelos binarios *p*-DPLS utilizando el método de voto mayoritario ponderado se puede utilizar la fiabilidad de clasificación, $R_{c,k}$, como coeficiente de ponderación del objeto. Así, podemos reescribir la ecuación 3-4 como:

$$F(\mathbf{x}) = \max_{c=1}^{C} \sum_{k=1}^{K} R_{c,k} d_{c,k}(\mathrm{x}) \tag{3-5}$$

Utilizar la fiabilidad como coeficiente de ponderación tiene la ventaja que es fácil de calcular y que es única para cada objeto, eliminado el problema de que sea constante para todos los objetos nuevos. Esto es así ya que el coeficiente de ponderación depende de la posición del objeto nuevo en el espacio de variables y no de los resultados de los datos de entrenamiento. Al final el método WMV combina las respuestas, votos y fiabilidades, $R_{c,k}$, de los modelos binarios *p*-DPLS óptimos.

3.4.3 Combinación de modelos *p*-DPLS por funciones de densidad de probabilidad multivariante (FDPMV)

Esta metodología es una evolución del método *p*-DPLS del espacio de funciones de probabilidad bidimensional al multidimensional para resolver problemas multiclase. Al igual que en el apartado 2.2, los objetos se asignan a la clase con mayor probabilidad *a posteriori*, $P(\omega_c|\hat{y})$, calculada con la ecuación de Bayes, en donde la probabilidad condicionada para una clase, $p(\hat{y}|\omega_c)$, se calcula de la combinación multivariante de las predicciones ($\hat{y}$) y el error estándar de predicción (*SEP*) de lo modelos binarios *p*-DPLS con técnica de binarización OAO.

El uso de este método se ilustra para un problema con tres clases (C = 3), aunque el procedimiento es extrapolable para un mayor número de clases.

1. Se calculan y validan los tres posibles modelos binarios *p*-DPLS, utilizando la estrategia OAO, siguiendo la metodología propuesta por *Pérez et ál.* [4]. Se obtienen así los modelos: ω_1 *vs.* ω_2; ω_1 *vs.* ω_2 y ω_2 *vs.* ω_3.

2. Se calculan los valores $\hat{y}$ y *SEP*, para todos los factores de los modelos del paso anterior; obteniendo así las matrices $\hat{\mathbf{Y}}$ ($I \times A_{max}$) y **SEP** ($I \times A_{max}$), una por cada modelo binario, para I objetos de entrenamiento por A_{max} número de factores máximo calculado en los modelos binarios.

3. Se desarrolla la función de densidad de probabilidad normal multivariante (FDPNM). Ésta se construye como el promedio de funciones gaussianas multivariantes individuales $g_n(y_1, \ldots, y_K)$, centradas en el vector de predicciones

DPLS, $\hat{\mathbf{y}}_i$ ($\hat{\mathbf{y}}_i = [\hat{y}_{i,1}, \ldots, \hat{y}_{i,K}]$), de los I_c objetos de calibración (es decir I_c filas de la matriz $\hat{\mathbf{Y}}_c$). Así, la probabilidad condicionada se expresa como:

$$p(\hat{\mathbf{y}} \mid \omega_c) = \frac{1}{N_c} \sum_{i=1}^{I_c} g_i(\hat{\mathbf{y}}) \tag{3-6}$$

donde $g_i(\hat{\mathbf{y}})$ es:

$$g_i(\hat{\mathbf{y}}) = \frac{1}{(2\pi)^{K/2} |\boldsymbol{\Sigma}_i|^{1/2}} e^{\left[-\frac{1}{2}(\hat{\mathbf{y}}-\hat{\mathbf{y}}_i)^T \boldsymbol{\Sigma}_i^{-1} (\hat{\mathbf{y}}-\hat{\mathbf{y}}_i)\right]} \tag{3-7}$$

$\hat{\mathbf{y}}$ es un vector de variables en el espacio de K dimensiones ($\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_K]$); $\boldsymbol{\Sigma}_i$ es una matriz simétrica cuya diagonal principal son los elementos del vector **SEP** (**SEP** = [SEP_1, …, SEP_K]) elevados al cuadrado y el resto de elementos iguales a 0 [3]:

$$\Sigma_i = \begin{pmatrix} SEP_{1,1}^2 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & SEP_{K,K}^2 \end{pmatrix}_i \tag{3-8}$$

Así se obtiene una FDPNM para cada una de las clases, ω_c; en el caso de tener más clases se obtendrían C FDPNMs. Dichas FDPNMs son independientes del número de modelos binarios utilizados.

4. Se valida el modelo multivariante por validación cruzada dejando fuera un objeto cada vez (LOOCV). Para validar el modelo multiclase se debe hallar qué combinación de factores óptimos, A_{opt}, en cada uno de los modelos binarios, k, presenta los mejores resultados de clasificación total multiclase, es decir, objetos bien clasificados a las C clases divididos por el total de objetos de entrenamiento. Téngase en cuenta que deben estudiarse todas las posibles combinaciones de factores entre los modelos, es decir A_{max} factores elevado a la K modelos (A_{max}^K). A diferencia de otros métodos de combinación, que utilizan las respuestas de los modelos binarios óptimos, en este caso se debe encontrar qué combinación de factores da los mejores resultados, ya que se están combinando respuestas parciales de lo modelos binarios (los valores de $\hat{y}$ y *SEP*), y no la respuesta final, como lo es la fiabilidad.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

5. Para asignar un objeto desconocido con el método multivariante:

   5.1. Se obtiene el vector $\hat{\mathbf{y}}_u$, que contiene las predicciones del objeto desconocido dadas por los modelos binarios *p*-DPLS, cada uno con su respectivo número óptimo de factores hallado en el paso anterior, $\hat{\mathbf{y}}_u = [\hat{y}_{u,1}, \hat{y}_{u,2}\, \hat{y}_{u,3}]$.

   5.2. Se estima la probabilidad condicionada del objeto en cada una de las clases, $p(\hat{\mathbf{y}}_u|\omega_c)$, y en cada uno de los modelos óptimos; estas probabilidades se evalúan con la ecuación de Bayes para hallar la probabilidad *a posteriori* o probabilidad que tiene el objeto de pertenecer a alguna de las tres clases:

$$P(\omega_c \mid \hat{\mathbf{y}}_u) = \frac{p(\hat{\mathbf{y}}_u \mid \omega_c) \times P(\omega_c)}{p(\hat{\mathbf{y}}_u)} \tag{3-9}$$

   donde el denominador es:

$$p(\hat{\mathbf{y}}_u) = \sum p(\hat{\mathbf{y}}_u \mid \omega_c) \times P(\omega_c) \tag{3-10}$$

   La probabilidad *a priori*, $P(\omega_c)$, se puede calcular como $P(\omega_c) = I_c/I$, asumiendo que el número de objetos para cada una de las clases es representativo del total de la población.

   5.3. El objeto es asignado utilizando la función de decisión bayesiana:

   Se decide a la clase ω_c si $P(\omega_c|\ \hat{y}_u)$ es máxima ($\max\ (P(\omega_c\,|\hat{y}_u))$ (3-11)

   Es decir, el objeto es asignado a la clase con una mayor probabilidad *a posteriori*.

Aunque el planteamiento es similar al *p*-DPLS binario, incluir la incertidumbre de predicción del objeto desconocido como criterio para calcular la fiabilidad de clasificación no es fácil; ya que no se tiene una incertidumbre sino *K* incertidumbres, una por cada modelo. Así, al calcular la incertidumbre alrededor de un objeto problema, a diferencia de *p*-DPLS en el que se obtiene un área bajo la curva de la FDP, para el *p*-DPLS FDPNM se obtiene un "volumen" bajo la superficie de la FDP multivariante de cada una de las clases. Este "volumen" de incertidumbre no tiene la misma interpretación que tiene el área en *p*-DPLS, pues se debe tener en cuenta la dirección de las superficies de cada una de las clases, para determinar qué secciones de esas incertidumbres se solapan.

3.4.4 Ejemplo ilustrativo de las tres alternativas de combinación con el conjunto de datos *Fisher Iris*

Se retomaron los conjuntos de datos *Fisher Iris* de entrenamiento y prueba del apartado 3.2.3; para comparar las ventajas y desventajas de los tres métodos de combinación anteriormente mencionados con el método propuesto, *p*-DPLS multiclase, que combina modelos *p*-DPLS hallando el producto de fiabilidades de clasificación de éstos. Se calcularon y validaron los tres posibles modelos *p*-DPLS con datos centrados y siguiendo la estrategia OAO: *Setosa vs. Versicolor*, *Setosa vs. Virginica* y *Versicolor vs. Virginica*. Todos los modelos *p*-DPLS se calcularon siguiendo la metodología propuesta por *Pérez et ál.* [4]. El número óptimo de factores se halló por validación cruzada dejando fuera un objeto cada vez (LOOCV), con criterio de máxima clasificación total, es decir, objetos bien clasificados a las dos clases, dividido por el total de objetos de entrenamiento. Como intervalo de incertidumbre para el cálculo por áreas se utilizó $2\times SEP$. Se hallaron las asignaciones (equivalentes al voto $d_{c,k}$ para el WMV) y las fiabilidades de clasificación, $R_{c,k}$, de los objetos en cada una de las clases y en cada uno de los modelos con su respectivo número óptimo de factores, tanto para los resultados de validación cruzada como para los del conjunto de prueba. Como funciones de decisión de clasificación multiclase se utilizaron las siguientes ecuaciones: 4 del apartado 3.2 para *p*-DPLS multiclase, 3-3 para el método de suma de probabilidades (apartado 3.6.1), 3-5 para voto mayoritario ponderado (apartado 3.6.2) y 3-11 para FDPNM (apartado 3.6.3).

La primera diferencia entre métodos de combinación es el número óptimo de factores que se deben tener en cuenta en cada uno de los modelos binarios. Por un lado, los métodos de suma de probabilidades, voto ponderado y *p*-DPLS multiclase, utilizaron las respuestas de los modelos binarios óptimos (1 factor para *Setosa vs. Versicolor* y *Setosa vs. Virginica* y 2 factores para *Versicolor vs. Virginica*.); es decir, primero se optimizaron los modelos binarios antes de poder combinar las respuestas de estos. En contraste, *p*-DPLS FDPNM halla los modelos binarios óptimos a partir del modelo multiclase, dado que utiliza respuestas parciales de éstos y no la final; es decir, que el método *p*-DPLS FDPNM optimiza qué combinación de factores de los modelos binarios da la mejor respuesta de clasificación. En este caso utiliza los modelos *Setosa vs. Versicolor* con 1 factor, *Setosa vs. Virginica* con 2 factores y *Versicolor vs. Virginica* con 3 factores.

Los resultados de los tres métodos de combinación alternativos y el *p*-DPLS multiclase son similares (Tabla 3-2); así, con los datos de entrenamiento se obtiene un 100% de correcta clasificación para la clase *Setosa* y 94.3% de correcta clasificación para la clase *Versicolor* (con los mismos dos objetos mal asignados por los cuatro métodos). Sólo con la clase *Virginica* cambian los resultados con los datos de entrenamiento, en donde

el método de *p*-DPLS FDPNM obtiene un 97.1% de correcta clasificación, asigna mal un objetos a *Versicolor* mientras que los otros tres métodos lo clasifican bien obteniendo un 100% de correcta clasificación. Con los datos de prueba los cuatro métodos obtienen el 100% de correcta clasificación para las tres clases.

3.4.5 Discusión

Los métodos de combinación por suma de probabilidades y el voto mayoritario ponderado son parecidos; dada la semejanza de las ecuaciones de decisión de clasificación (Ec. 3-3 y 3-5). No obstante, el método de suma de probabilidades (Ec. 3-3) no preasigna los objetos a alguna de las clases del modelo binario. Esta es una ventaja ya que le permite utilizar las fiabilidades que los modelos binarios dan a un objeto en cada una de las clases. Por el contrario, el voto mayoritario ponderado

Tabla 3-2: Respuestas de clasificación de los 4 métodos de combinación para el conjunto de datos de entrenamiento (LOOCV) de *Fisher Iris*.

Decisión Multiclase	Método de combinación	Clase reales		
		Setosa	*Versicolor*	*Virginica*
Setosa	*p*-DPLS multiclase	100%	0%	0%
	$\Sigma R_{c,k}$	100%	0%	0%
	WMV	100%	0%	0%
	FDPMV	100%	0%	0%
Versicolor	*p*-DPLS multiclase	0%	94.3%	0%
	$\Sigma R_{c,k}$	0%	94.3%	0%
	WMV	0%	94.3%	0%
	FDPMV	0%	94.3%	2.9%
Virginica	*p*-DPLS multiclase	0%	5.7%	100%
	$\Sigma R_{c,k}$	0%	5.7%	100%
	WMV	0%	5.7%	100%
	FDPMV	0%	5.7%	97.1%

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

elimina una de las fiabilidades al multiplicarla por 0 (recuérdese que este método multiplica el voto por la fiabilidad, 1 si pertenece y 0 si no pertenece), ya que multiplica el coeficiente de ponderación por el voto (Ec. 3-5), para convertirlo en una probabilidad hacia una clase. Así, cuando se suman los votos ponderados de una misma clase pueden darse los mismos resultados que para el método de suma de probabilidades.

La similitud entre los métodos de suma de probabilidades y el voto mayoritario ponderado no es crítica cuando la fiabilidad es cercana al 100% o al 0%. El voto mayoritario ponderado multiplicaría 1 por una fiabilidad del 100% y 0 por una fiabilidad del 0%. Sin embargo, la multiplicación sí es crítica cuando la fiabilidad es cercana al 50%. Ello se observa en la tabla 3-3 para un ejemplo improvisado de un objeto totalmente desconocido del cual queremos saber a que clase *Iris* pertenece. En la tabla 3-3 se pueden ver los valores de fiabilidad, $R_{c,k}$, y voto ponderado que cada modelo da al objeto. Así, si la asignación la hiciéramos por voto ponderado se asignaría a la clase *Setosa*; sin embargo, si la asignación se hace por suma de probabilidades se asignaría a *Versicolor*, pues la suma de probabilidades para esta clase es mayor que para *Setosa*.

Los resultados para el objeto improvisado ponen de manifiesto una de las mayores desventajas de estos métodos; no tener una certeza de la asignación que éstos hacen; más si se tiene en cuenta que métodos distintos dan resultados distintos. Otra desventaja es que los modelos binarios pueden asignar a más de una clase un objeto, por ejemplo, los objetos de la clase *Setosa* son asignados por el modelo *Versicolor vs. Virginica* a *Versicolor* con una alta fiabilidad.

Tabla 3-3: Respuestas para un objeto nuevo y desconocido evaluado con los modelos del conjunto de datos *Iris*.

Método / Modelo	*Setosa*		*Versicolor*		*Virginica*	
	$\Sigma R_{c,k}$	WMV	$\Sigma R_{c,k}$	WMV	$\Sigma R_{c,k}$	WMV
Setosa vs. Versicolor	0.51	0.51	0.49	0	–	–
Setosa vs. Virginica	0.60	0.60	–	–	0.40	0
Versicolor vs. Virginica	–	–	0.80	0.80	0.20	0
Total	1.11	1.11	1.29	0.80	0.60	0
Asignación	NO	SI	SI	NO	NO	NO

Estas desventajas, asignar un objeto a más de una clase o no tener una fiabilidad en la asignación, no son un problema exclusivo de conjunto *Iris* sino que pueden presentarse en cualquier otro conjunto de datos, ya que la asignación de los objetos depende de la posición de éstos en el espacio original de variables. Así, los objetos de clases no modeladas que estén cerca del espacio de variables de las clases modeladas, por ejemplo los objetos de *Setosa* en el modelo *Versicolor vs. Virginica*, serán erróneamente tomados como objetos de alguna de las clases modeladas y serán asignados a una de ellas, induciendo a un error en el modelo multiclase.

Estas desventajas son resueltas por los métodos *p*-DPLS multiclase y *p*-DPLS FDPNM. *p*-DPLS multiclase, al igual que la suma de probabilidades y el voto ponderado, utiliza las fiabilidades de clasificación $R_{c,k}$. Sin embargo, puede evaluar si la información entregada por los modelos binarios es confiable. Así, la alta fiabilidad de pertenecer a la clase *Versicolor* que el modelo *Versicolor vs. Virginica* otorga a los objetos de la clase *Setosa* se puede anular con la baja fiabilidad que el modelo *Setosa vs. Versicolor* da de pertenecer a *Versicolor*. Además, y dado que puede calcular la fiabilidad de clasificación multiclase, puede aclarar dudas como la registrada con el objeto improvisado (Tabla 3-3). Así, *p*-DPLS multiclase asigna el objeto improvisado a la clase *Versicolor* pero con un 50.4% de fiabilidad, es decir, que aunque la respuesta más probable es la clase *Versicolor* no se debe confiar mucho en esta asignación, pudiendo indicar que dicho objeto no debe pertenecer a la especie *Iris*, aunque inicialmente algunas de las características lo hicieran suponer.

p-DPLS FDPNM aprovecha mejor la información suministrada por lo modelos binarios, $\hat{y}$ y *SEP*; permitiéndole optimizar la combinación de modelos binarios que dan la mejor respuesta de clasificación. A diferencia de los otros tres métodos de combinación que utilizan la fiabilidad de clasificación, $R_{c,k}$, para asignar objetos desconocidos, *p*-DPLS FDPNM utiliza la $\hat{y}$ para calcular la probabilidad *a posteriori* de pertenencia a alguna de las clases modeladas. Aunque esta probabilidad *a posteriori* podría ser usada como un indicativo de la fiabilidad de clasificación multiclase, no incluye el *SEP* del objeto desconocido, que permite una mejor diferenciación de objetos cercanos al límite entre clases. De ahí el error de asignación que se observa para la clase *Virginica* (Tabla 3-2), donde el *p*-DPLS FDPNM asigna mal un objeto de *Virginica* a la clase *Versicolor*.

En general los cuatro métodos de combinación de modelos binarios *p*-DPLS se basan en el principio del "ganador se lo lleva todo", asignando siempre los objetos a la clase donde haya mayor probabilidad. Sin embargo, los métodos de suma de probabilidades y voto ponderado son más sensibles a las ambigüedades que se presentan entre modelos binarios OAO, pues este tipo de modelado no tiene en cuenta las relaciones

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

que puedan darse entre las clases modeladas y las que se dejan fuera del modelo. Aunque *p*-DPLS FDPNM puede eliminar esta ambigüedad, al tener en cuenta las relaciones entre clases, es más complejo de desarrollar y comete un mayor error de clasificación al no tener en cuenta el *SEP* de los objetos. Así pues, *p*-DPLS multiclase es la mejor alternativa, dado que es sencilla y puede eliminar las ambigüedades observadas en los modelos con estrategia OAO.

3.4.6 Conclusiones

Se ha comprobado la posibilidad de utilizar el método de suma de fiabilidades para combinar resultados de *p*-DPLS de modelos binarios, con resultados de multiclasificación que superan el 94%. Sin embargo, al intentar calcular la fiabilidad de predicción multiclase se observa que las ambigüedades de asignación de los modelos binarios, como asignaciones a más de una clase, afectan el cálculo de la fiabilidad, ya que el método no diferencia qué resultados de los modelos binarios pueden ser rechazados.

Es viable utilizar la fiabilidad como constante de ponderación en el voto mayoritario ponderado, obteniendo resultados comparables con otros métodos de combinación. Sin embargo, no es posible calcular la fiabilidad de la multiclasificación.

La combinación de modelos binarios utilizando el método *p*-DPLS FDPNM es factible, y dado que éste hace una combinación multivariante de respuestas preliminares de los modelos binarios, $\hat{y}$ y *SEP*, proporciona una ventaja sobre otros métodos de combinación que dependen de optimizar primero los modelos binarios. En otras palabras, el método *p*-DPLS FDPNM asigna un objeto desconocido partiendo del vector **x** de variables y no de las respuestas de los clasificadores binarios. Sin embargo, a diferencia del *p*-DPLS, en donde podemos utilizar el intervalo de incertidumbre para calcular la fiabilidad, en el caso multivariante no es posible calcularlo, dado que no se tiene un solo intervalo de incertidumbre, sino tantos intervalos como modelos haya.

El cálculo de fiabilidad multiclase obtenido por el método de combinación propuesto *p*-DPLS multiclase (apartado 3.2), demuestra su utilidad al permitir definir si la asignación puede ser aceptada o no. Es decir, se puede tomar como criterio que asegure no sólo que la asignación es correcta, sino si se correlaciona con un conjunto de datos determinado, y no que se trate de un objeto ajeno a las clases modeladas.

Referencias

1. K. Huang, Z. Xu, I. King, M. R. Lyu, Z. Zhou, "A Novel Discriminative Naïve Bayesian Network for Classification". En: A.Mittal, A. Kassim (Ed.) "Bayesian Network Technologies. Application and Graphical Models", IGI Publishing, Hershey, NY, USA, 2007. p. 1–13.
2. L. I. Kuncheva, "Combining Pattern Classifiers. Methods and Algorithms", John Wiley & Sons, Inc., Hoboken, NJ, USA, 2004, p. 123–125.
3. R. A. Johnson, D. W. Wichern, "Applied Multivariate Statistical Analysis", 5ª Edición, Prentice Hall, Upper Saddle River, NJ, USA, 2002, p. 149–167.
4. N. F. Pérez, J. Ferré, R. Boqué, Chemometr. Intell. Lab. Syst. 95 (2009) 122–128.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Capítulo 4

Especificaciones multivariantes

4.1 Introducción y revisión bibliográfica

Los consumidores consideran las especificaciones de un producto como una forma de evaluar si un alimento cumple con las condiciones óptimas de consumo (manufacturación, nutricional y de salubridad) [1]. Los continuos cambios en los hábitos de consumo así como en los sistemas de producción hacen necesarias nuevas especificaciones que permitan verificar la autenticidad de los alimentos [2], en otras palabras, es necesario incluir nuevos parámetros que garanticen el origen y las condiciones de producción (como en el caso de los alimentos orgánicos) y prevenir así falsificaciones o engaños con disminución en la calidad del producto. Dichas medidas son también demandadas por los productores que buscan reducir la competencia desleal y dar un valor añadido al producto. En respuesta a ello, la Unión Europea, dentro del programa de seguridad alimentaria [3], ha fortalecido los sistemas de protección existentes: Denominación de Origen Protegida (DOP), Indicación Geográfica Protegida (IGP) [4] y Especialidad Tradicional Garantizada (ETG) [5]. En este marco el proyecto TRACE ha querido desarrollar métodos analíticos que permitan otorgar una "huella digital" a los alimentos e identificar productos falsificados [6].

Dado que muchos de estos métodos proveen la información del producto como un vector de variables, es necesario adaptar las especificaciones a este contexto multivariante.

4.1.1 ¿Que es una especificación?

El real diccionario de la lengua española define especificación como [7]: "*Información proporcionada por el fabricante de un producto, la cual describe sus componentes, características y funcionamiento*". Una definición más amplia la proporciona la *American Society for Testing and Materials International* (ASTM International) que define especificación como [8]: "*conjunto explícito de requerimientos que deben ser satisfechos por un material, producto o servicio*", y además aclara que: "*los ejemplos de especificaciones incluyen, mas no se limitan a, los requerimientos de: propiedades físicas, mecánicas o químicas y a criterios de seguridad, calidad y desempeño. Una especificación identifica los métodos de prueba que permiten determinar que cada uno de los requerimientos es satisfecho*". De esta manera, el concepto de especificación es ampliamente usado en áreas de la ingeniería, la industria y el comercio buscando establecer parámetros y normas que garanticen los mínimos requerimientos de calidad en el desarrollo, construcción, manufacturación o prestación de servicios. La importancia de las especificaciones es tal que en la mayoría de los casos recae sobre entes gubernamentales y organismos internacionales el regular, mediante leyes y normas, estas especificaciones [9,10], las cuales deben ser supervisadas por los organismos competentes [11,12].

Una especificación debe ser sencilla y fácil de entender. Dentro de las entidades que se encargan de desarrollar y establecer estas especificaciones o normas se encuentran la "*International Organization for Standardization* (ISO)", a la cual se encuentran adheridos la gran mayoría de países, el "*Comité Européen de Normalisation* (CEN)", que asocia a los países europeos, o la ASTM International norteamericana. Además, están presentes otras asociaciones o entidades que dictan normas más particulares; en el campo de la química encontramos la *International Union of Pure and Applied Chemistry* (IUPAC) que también unifica el "leguaje químico"; en el farmacéutico diversas farmacopeas, como la europea y la británica, y en el alimentario entidades de estándares alimentarios, como el *Codex Alimentarius*, desarrollado en conjunto por la Organización para la Agricultura y la Alimentación (*Food and Agriculture Organization*, FAO) y la Organización Mundial de la Salud (OMS).

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

4.1.2 ¿Como se define una especificación?

Al definir una especificación deben tenerse en cuenta dos aspectos. El primero corresponde a los parámetros o características de la especificación. En este caso podemos encontrar dos situaciones: especificaciones que deben ser satisfechas por un producto (determinadas *a priori*), como la tolerancia de materiales o las dimensiones de manufacturación, y las especificaciones derivadas de observaciones o investigaciones, como requerimientos nutricionales mínimos en alimentos [10] o máximos permitidos de pesticidas, metales pesados y contaminantes en general [11]. En este último también se incluyen las especificaciones de producto que hacen referencia a la constitución, el origen y las características (por ejemplo, el etiquetado de los alimentos procesados). Téngase en cuenta que una especificación estará delimitada por un máximo, un mínimo o un intervalo de valores que debe ser cumplido para que el producto se considere dentro de la "especificación", asegurando la idoneidad del mismo.

El segundo aspecto corresponde a la reglamentación de la especificación. Tanto la ISO como la ASTM consideran que las especificaciones son documentos técnicos que deben cumplir una serie de pasos antes de ser consideradas como un estándar [13]. Para ello se requiere que sean apoyadas en consenso como normas, tengan el suficiente desarrollo técnico y no entren en contradicción con otras normas ya existentes.

4.1.3 Tipos de especificación

Dependiendo de la forma como son definidas, podemos considerar dos tipos generales de especificaciones: univariantes y multivariantes. La forma más usual de especificación es la univariante, que se define mediante una variable como el peso, la dimensión, la densidad. A este tipo también pertenecen las variables cualimétricas, como las derivadas de un análisis sensorial [14]. En las especificaciones multivariantes se toman en cuenta varias variables, dando lugar a un vector de variables. Estas pueden ser analizadas cada una por separado, mediante un análisis multi-univariante, o en conjunto, utilizando técnicas multivariantes que tienen en cuenta las relaciones entre dichas variables. El análisis multivariante de especificaciones ha tomado gran relevancia en los últimos años, impulsado por los modernos métodos de análisis que proporcionan grandes cantidades de datos [15,16].

El desarrollo de especificaciones ha estado ligado al control de calidad, compartiendo con éste los métodos que permiten verificar el cumplimiento de las especificaciones.

Así, para verificar que una o más especificaciones univariantes cumplen lo estipulado, se utiliza el control estadístico de calidad (*Statistical Quality Control* (SQC)). Mediante herramientas SQC univariantes se verifica que los valores de las variables satisfagan una gráfica con valores límite (por ejemplo la gráfica de *Shewhart*) [17]. En el caso de las especificaciones multivariantes se utilizan herramientas SQC multivariantes (*Multivariate Statistical Quality Control* (MSQC)) tales como el estadístico *Hotelling* T^2 [18].

4.1.4 Especificaciones univariantes y especificaciones multivariantes

Las especificaciones univariantes son las más extendidas tanto, en el sector privado como en el gubernamental, y la mayoría de las normas vigentes se basan en ellas. Un claro ejemplo es la directiva 80/778/CEE de la Unión Europea sobre aguas destinadas al consumo humano, que en su anexo 1 establece 56 variables a ser medidas, señalando sus niveles y procedimientos de medida [19]. Por el contrario, las especificaciones multivariantes son apenas mencionadas. Uno de los pocos ejemplos es la norma ASTM E1790–04, para el análisis cualitativo de líquidos y sólidos mediante espectroscopia de infrarrojo cercano (NIR) [20]. La preferencia por especificaciones univariantes también se observa en el análisis de especificaciones multivariantes como multi-univariantes (una por una); lo que desliga las variables y lleva a conclusiones erróneas sobre la calidad del producto [15]. Ello contrasta, con el amplio desarrollo que ha tenido el control estadístico multivariante de procesos (*Multivariate Statistical Process Control* (MSPC)), el cual proporciona muchos de los conceptos y herramientas que pueden ser utilizados para definir especificaciones multivariantes.

Inicialmente mencionado en econometría [21], el concepto de especificaciones multivariantes fue estudiado y aplicado a la industria química por *De Smet*, *Duchesne* y *MacGregor*, que consideran que es la mejor forma de asegurar la óptima calidad de las materias primas en los procesos industriales [22]. Cuando se usa más de una variable para definir un producto y se analizan una por una (análisis multi-univariante) no se tienen en cuenta las posibles correlaciones que se puedan dar entre variables, lo que conlleva a una errónea definición de los límites del producto [15,18,22]. Un ejemplo se muestra en la figura 4-1, donde se han graficado dos variables, X_1 y X_2, con sus correspondientes límites univariantes. Estos límites crean un "rectángulo" donde se concentran la mayoría de los objetos. Sin embargo, la forma que mejor se amolda a la distribución de los objetos es una elipse. Esta elipse acepta dos objetos rechazados por la variable X_1 y rechaza dos objetos que estaban dentro de las especificaciones univariantes. Dichos objetos, aún cumpliendo las especificaciones univariantes,

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

seguramente no cumplirían los requisitos de calidad. De esta forma, un correcto análisis del conjunto de datos del que se definirán las especificaciones nos permite obtener unos límites de especificación más acordes y reducir los errores de tipo I y II, es decir, rechazar objetos que están dentro de la especificación y aceptar objetos que están fuera de la especificación, respectivamente. Así pues, las especificaciones multivariantes pueden ayudar a resolver muchas de las desventajas del sistema univariante.

Aunque las especificaciones univariantes ya están consolidadas, son varias las iniciativas que buscan fomentar las especificaciones multivariantes. Una de ellas parte del proyecto TRACE, que busca sean una alternativa a las especificaciones actuales de producto en el ámbito de la autentificación y trazabilidad alimentaria, ya sea a partir de las variables originales o por reducción de variables mediante métodos como PCA o PLS. Además, dentro del proyecto TRACE se han desarrollado especificaciones derivadas a partir métodos de clasificación como CART, SIMCA o ANN, entre otros [23,24].

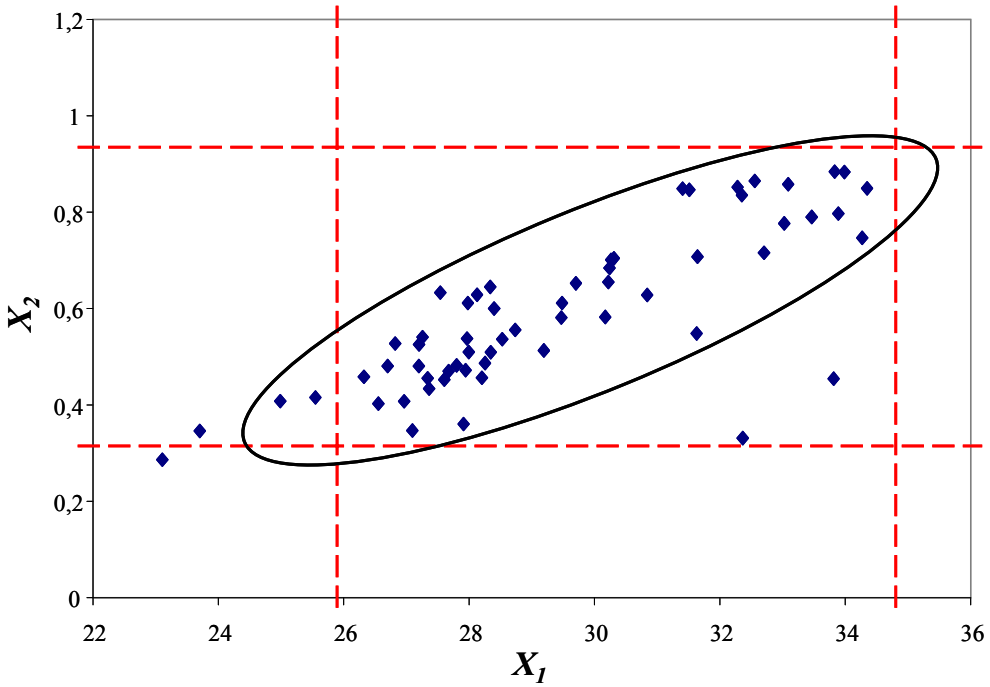

Figura 4-1: Gráfica de control multivariante para las variables X_1 y X_2. Las líneas segmentadas representan los límites univariantes y la elipse el límite multivariante más aproximado a la distribución de los datos.

4.1.5 Implementando especificaciones multivariantes

Para establecer especificaciones multivariantes *De Smet*, *Duchesne* y *MacGregor* [22] proponen tres pasos: 1) adquirir un conjunto de datos adecuado; 2) desarrollar las especificaciones multivariantes, y 3) implementar las especificaciones multivariantes. Para adquirir el conjunto de datos debemos tener claro para qué material queremos establecer las especificaciones. Así, para definir las especificaciones de un producto terminado solamente se necesita una matriz de datos. Por el contrario, para especificaciones de materias primas pueden ser necesarias tres matrices de datos: las propiedades de las materias primas, las variables de proceso y las propiedades del producto terminado, ya que las características del producto terminado dependen tanto de las variables del proceso como de las propiedades de las materias primas. En el paso dos debe establecerse el modelo matemático que definirá la región de especificación. Dicho modelo puede establecerse en el espacio original de las variables, en el espacio reducido (mediante PCA o PLS) de variables o en el espacio de una propiedad predicha por el modelo. La validación de dicho modelo es fundamental en este paso. Finalmente, el paso de implementación consiste en aplicar el modelo a un objeto problema y decidir si cumple o no con los límites de la especificación, idealmente con una indicación del nivel de confianza.

El paso más crítico es el segundo, en donde para obtener especificaciones fiables y minimizar los errores de tipo I y II, es necesario detectar y eliminar los datos discrepantes [15,25]. Una vez definido el conjunto de datos adecuado podemos desarrollar la región de especificaciones multivariantes y monitorizar objetos nuevos con los estadísticos multivariantes Chi-cuadrado (χ^2) o T^2. χ^2 permite calcular la distancia del objeto nuevo al centro del conjunto de datos cuando conocemos la matriz de varianzas; si la matriz de varianzas no es conocida se utiliza el T^2 o estadístico *Hotelling* [26]. Mediante la monitorización se observa si los objetos están por debajo del límite superior de control (*Upper Control Limit* (UCL)), a un determinado nivel de confianza (habitualmente del 95% o 99%) y una distribución normal de los datos. Otros métodos de monitorización son las versiones multivariantes de las gráficas de medias móviles con ponderación exponencial (*Exponentially Weighted Moving Averge* (EWMA)) y de suma acumulativa (*Cumulative Sum* (CUSUM)) [18].

Cuando el número de variables se incrementa sustancialmente es difícil definir especificaciones a partir de las variables, debido a la correlación entre variables y al ruido, lo cual hace perder calidad a los modelos. En estos casos se requiere reducir la dimensionalidad de los datos mediante PCA o PLS [22]. PCA se utiliza cuando se tiene

una única matriz de datos, por ejemplo, propiedades del producto. Así, el estadístico T^2 se aplica sobre los *scores* de PCA, y se complementa con el estadístico Q o error cuadrado de predicción (*SPE*) que toma en cuenta los residuales [15]. Cuando se dispone de diferentes matrices (por ejemplo propiedades de las materias primas y variables de calidad del producto terminado) es preferible reducir la dimensionalidad mediante PLS o DPLS ya que maximiza la covarianza entre las dos matrices de datos (materias primas – producto terminado) [22]. Ello es particularmente útil cuando se quieren establecer, por ejemplo, las especificaciones de materias primas a partir de unos parámetros de calidad del producto terminado. En este caso se aplican los estadísticos T^2 y *SPE*, que pueden ser complementados con el estadístico *DModX*, que mide la distancia del objeto al modelo [26].

Finalmente y en cualquiera de los casos, los límites de especificación se seleccionan optimizando los parámetros estadísticos que permitan obtener los menores errores de tipo I y II [22]. Los errores de tipo I se cometen cuando objetos que cumplen con la especificación son rechazados y los errores de tipo II se cometen cuando objetos que están fueran de la especificación son aceptados. En este punto cabe aclarar que dependiendo del ámbito de aplicación de las especificaciones es necesario priorizar uno u otro error. Por ejemplo, en el caso de especificaciones de origen, al minimizar el error de tipo I se da preferencia al productor, ya que el objetivo es que ningún producto que pertenezca a una determinada denominación de origen sea rechazado, aún a costa de aceptar como pertenecientes a esa denominación productos que en realidad no lo son. Por el contrario, al minimizar el error de tipo II se está dando preferencia al consumidor, ya que el objetivo es prevenir fraudes o engaños, y se persigue que ningún producto que no sea de una determinada denominación de origen sea aceptado como perteneciente a tal denominación, aún a costa de rechazar productos que sí pertenecen.

En el siguiente apartado se muestra la metodología para establecer especificaciones multivariantes de un producto alimenticio. Las especificaciones se fijan a partir de los *scores* y las predicciones de un modelo binario DPLS, con estrategia de binarización "uno contra todos" para un producto de interés. Se muestra el procedimiento para fijar los límites de los estadísticos *Hotelling* T^2 y *SPE*, a partir de los *scores* de los objetos del producto de interés, y se plantea como nuevo parámetro las predicciones del modelo, para los objetos del producto de interés. Un objeto será considerado dentro de la especificación si cumple los dos estadísticos y la predicción simultáneamente. La metodología es finalmente utilizada para fijar las especificaciones del aceite de oliva de la región italiana de *Liguria*.

Referencias

1. M. R. Hubbard, Statistical quality control for the food industry, Third Edition, Kluwer Academic / Plenum Plublishers, USA, 2003, pp: 253–276.
2. S. Kelly, K. Heaton, J. Hoogewerff, Trends Food Sci. Tech. 16 (2005) 555–567.
3. Reglamento (CE) No 178/2002 del Parlamento Europeo y del Consejo de 28 de enero de 2002 por el que se establecen los principios y los requisitos generales de la legislación alimentaria, se crea la Autoridad Europea de Seguridad Alimentaria y se fijan procedimientos relativos a la seguridad alimentaria. Diario Oficial de las Comunidades Europeas, L 31 (1/2/2002) 1–24.
4. Reglamento (CE) No 510/2006 del Consejo de 20 de marzo de 2006 sobre la protección de las indicaciones geográficas y de las denominaciones de origen de los productos agrícolas y alimenticios. Diario Oficial de la Unión Europea, L93 (31/3/2006) 12–25.
5. Reglamento (CE) No 509/2006 del Consejo de 20 de marzo de 2006 sobre las especialidades tradicionales garantizadas de los productos agrícolas y alimenticios. Diario Oficial de la Unión Europea L 93 (31/3/2006) 1–11.
6. Project TRACE – "TRAcing food Commodities in Europe" (Project no. FOOD-CT-2005-006942). www.trace.eu.org.
7. Real Academia Española, Diccionario de la lengua española, http://www.rae.es/rae.html (12/03/2010).
8. ASTM International, "Regulations Governing ASTM Technical Committees", March 2010. http://www.astm.org/COMMIT/Regs.pdf (25/03/2010).
9. D. L. Flumignan, G. C. Anaia, F. de O. Ferreira, A. G. Tininis, J. E. de Oliveira, Chromatographia, 65 (2007) 617–623.
10. J. M. Betz, K. D. Fisher, L. G. Saldanha, P. M. Coates, Anal. Bioanal. Chem. 389 (2007) 19–25.
11. M. L. Weiner, W.F. Salminen, P. R. Larson, R. A. Barter, J. L. Kranetz, G. S. Simon, Food Chem. Toxicol. 39 (2001) 759–786.
12. J. J. Chen, Y. Tsong, J. Biopharm. Stat., 7 (2007) 259–270.
13. International Organization for Standardization y International Electrotechnical Commission, ISO/IEC Directives, Part 1. Procedures for the technical work, Sixth edition, 2008. http://www.iec.ch/tiss/iec/Directives-Part1-Ed6.pdf (12/03/2010).
14. M. Muñoz, Food Quality and Preference 13 (2002) 329–339.
15. L. H. Chiang, L. F. Colegrove, Chemometr. Intell. Lab. Syst. 88 (2007) 143–153.
16. D. M. Ennis, J. Bi, J. Food Qual. 23 (2000) 541–552.
17. D C. Montgomery, G C. Runger, Applied Statistics and Probability for Engineers, Third Edition, John Wiley & Sons, Inc., USA, 2003, pp: 595–648.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

18. T. Kourti, J. F. MacGregor, Chemometr. Intell. Lab. Syst. 28 (1995) 3–21.
19. Directiva 80/778/CEE del Consejo, de 15 de julio de 1980, relativa a la calidad de las aguas destinadas al consumo humano, Diario Oficial, L 229 (30/08/1980) 11–29.
20. Designation: E 1790 – 04 Standard Practice for Near Infrared Qualitative Analysis, ASTM International.
21. M. Guidolin, A. Timmermann, J. Econometrics, 131 (2006) 285–208.
22. C. Duchesne, J. F. MacGregor, J. Qual. Tech. 36 (2004) 78–94.
23. Project TRACE – "TRAcing food Commodities in Europe" (Project no. FOOD-CT-2005-006942), "Work Package 6, deliverable 6.6, Setting Specifications and Compliance Assessment", 31-08-2008.
24. Marjana Novič, Neva Grošelj, Anal. Chim. Acta. 649 (2009) 68–74.
25. M-J. Bruwer, J. F. MacGregor, W. M. Bourg Jr., Food Quality and Preference, 18 (2007) 890–900.
26. C. Wikström, C. Albano, L. Eriksson, H. Fridén, E. Johansson, Å. Nordahl, S. Rännar, M. Sandberg, N. Kettaneh-Wold, S. Wold, Chemometr. Intell. Lab. 42 (1998) 221–231.

4.2 Establishment of multivariate specifications for food commodities with Discriminant Partial Least Squares

Submitted to Talanta

Néstor F. Pérez, Ricard Boqué*, Joan Ferré

Department of Analytical Chemistry and Organic Chemistry, Rovira and Virgili University. C/ Marcel·lí Domingo, s/n. 43007. Tarragona (Spain)

A novel method for establishing multivariate specifications of food commodities is proposed. The specifications are established for discriminant partial least squares (DPLS) by setting limits on the predictions of the DPLS model together with Hotelling T^2 and square error of prediction (*SPE*). These limits can be tuned depending on whether type I error (i.e. a correct sample is declared out-of-specification) or type II error (i.e. an out-of-specification sample is declared within specifications) need to be minimized. The methodology is illustrated with a set of NIR spectra of Italian olive oils, corresponding to five regions and the class Liguria is the class of interest. The results demonstrate the possibility of establishing multivariate specification for olive oils from the Liguria region on the basis of spectral data obtaining type I and type II errors lower than 5%.

4.2.1 Introduction

According to the American Society for Testing and Materials (ASTM) [1], specification is "an explicit set of requirements to be satisfied by a material, product, system or service." The document also states that "Examples of specifications include, but are not limited to requirements for: physical, mechanical, or chemical properties, and safety, quality, or performance criteria. A specification identifies the test methods for determining whether each of the requirements is satisfied". Specifications have a large importance in engineering, manufacturing and trade, and the governments must ensure proper development or provision of services to establish the minimum requirements needed to ensure the quality and adequacy of the item or service provided. These quality requirements are of such importance that sometimes are regulated by laws or standards [2,3] and overseen by competent agencies [4].

Specifications can be derived in different ways. First, specifications can be a set of parameters or characteristics that the user defines that products must satisfy, such as the tolerances of materials. Specifications can also be derived from observations or researches, such as the minimum nutritional requirements in foods [3], or the maximum levels permitted of pesticides, the heavy metals and contaminants in general [4] and also the product specifications that refer to the constitution, origin and/or characteristics of the product (e.g. the specifications of a protected designation of origin of a food commodity).

Consumers feel product specifications as the way to evaluate whether a food has the optimal conditions for consumption (manufacturing, nutrition and health) [5]. Lately, it became necessary to include specifications that ensure the authenticity of food [6]; i.e. parameters that guarantee the origin and the production conditions (e.g. organic food) and that there are not counterfeits. This requirement is also demanded by the producers because it ensures that there is no unfair competition and because it adds value to the product. In response, the European Union has set up three mechanisms of protection: Protected Designation of Origin (PDO), Protected Geographical Indication (PGI) and Traditional Specific Guaranteed [6,7]. Within this context, the primary objective of the European project "Tracing the Origin of Food (TRACE)" was to develop analytical methodologies to find a fingerprint for different food commodities and to identify counterfeit products [8].

A product specification can be either univariate or multivariate. Univariate specifications are the most commonly used and are defined by one or more individual variables (e.g. mass, length, density, etc.). However, most specifications are multivariate by nature, that is, several variables must be measured. Also, many analytical methodologies provide the information of the product as a vector of measured variables (i.e. mass spectrum); thus, the specifications must be adapted to this multivariate context. The variables can be analyzed either separately, without taking into account the relationship between them, or using multivariate analysis, that takes into account the correlations between variables. Treating multivariate specifications as multi-univariate has been reported to lead to erroneous conclusions about the quality of the product [9]. A clear example is shown in Fig. 1, where two variables, X_1 and X_2, and their corresponding univariate limits, are plotted. The limits create a rectangle that frames most of the points. However, it is more efficient if we frame the points within an ellipse, a bivariate limit. The ellipse fits better the distribution of the points, two "bad" objects are rejected, and two objects rejected by the variable X_1 are accepted. In this way, the joint analysis of the specifications can refine the specification limits and reduce the type I and type II errors. A Type I error is committed when a sample that comply with the specification is rejected, while a type II error is committed when a sample that does not meet the specification is accepted. In general, producers will require low type I errors, because they will not be satisfied if a complying product is

said to be out of the specifications. On the contrary, consumers would like to be protected against out-of-specification products, and so want low type II errors. Multivariate analysis of specifications has become more important in recent years, also because of the vast amount of data generated by the analytical methods [9,10]. An example is the work by Novič and Grošelj [11], who established product specifications based on classification models with neural networks.

To verify one or more univariate specifications statistical quality control (SQC) tools are commonly used. Univariate SQC verifies if the variable is within limit values (e.g. Shewhart chart) [12]. For multivariate specifications, multivariate SQC (MSQC) tools are used, such as the Hotelling's T^2 statistics [13]. When the number of variables is large, such as those generated by spectral methods of analysis, principal component analysis (PCA) or partial least squares (PLS) regression are used to reduce the number of variables, so that the multivariate control limits or specification limits are defined using the significant PCA or PLS factors. PLS has the advantage over PCA that provides a control both on the input variables (e.g. the raw materials), and of the output variables (quality of the final product), and it has been applied in the control of chemical industrial processes [13].

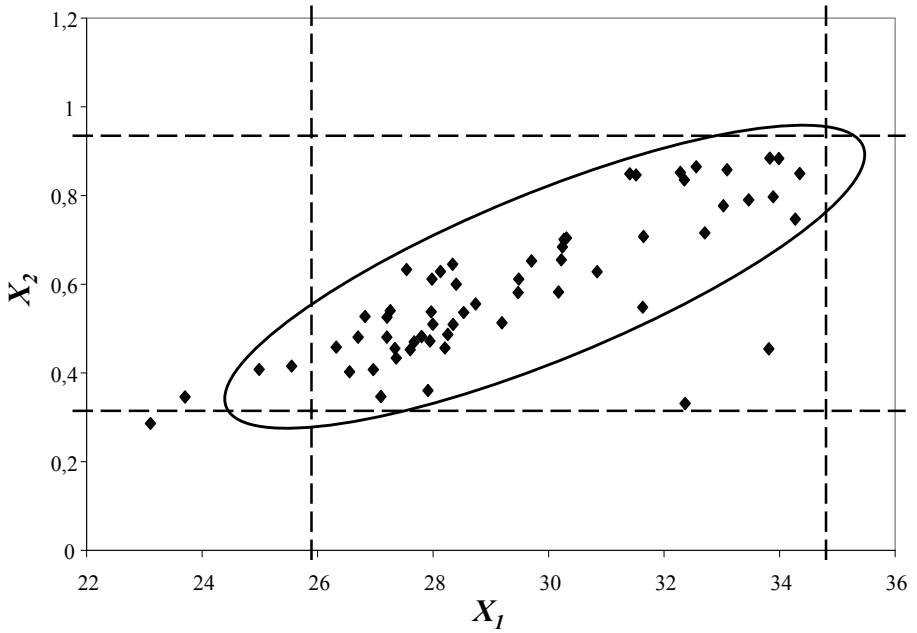

Fig. 1: Multivariate control charts for variables X_1 and X_2. Dotted lines represent the univariate specification limits, and the ellipse the bivariate specification limits.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Multivariate specifications will largely depend on the data analysis methods used. In this paper we present a procedure for establishing product specifications from discriminant PLS (DPLS) (PLS applied to classification) [14]. DPLS binary models are derived with the strategy "one against all" [15], thus obtaining as many models as classes are being modeled ($C = K$). For each of the models, the scores, the x-residuals and the predicted $\hat{y}$ of the samples of the class of interest are used to establish the boundaries of Hotelling T^2, Q (or squared prediction error (SPE)) statistics and also the PLS prediction $\hat{y}$. A product that meets the specification will be within the statistical limits. This procedure is illustrated with the Olive Oil data set, a dataset generated within the TRACE project.

4.2.2 Multivariate specifications

Unlike univariate specifications, which are used to define many of the regulations now in observance, multivariate specifications are scarcely mentioned in the literature. This contrasts with the extensive references to Multivariate Statistical Process Control (MSPC), which shares many concepts and analytical tools with multivariate specifications. Initially referred in econometrics [16], multivariate specifications were first studied and applied in the chemical industry by *De Smet*, *Duchesne* and *MacGregor* to ensure optimal raw materials in industrial processes [17].

To establish multivariate specifications *De Smet*, *Duchesne* and *MacGregor* proposed three steps: 1) acquire an adequate set of data, 2) develop the multivariate specifications, and 3) implement the multivariate specifications. When acquiring the data set, we must consider the item for which we want to establish specifications. Thus, for specifications of a final product only a single data matrix is needed. On the contrary, for specifications of raw materials to be input in a process, three data matrices may be required: properties of raw materials, process variables and properties of the finished product, because the characteristics of the finished product depend both on the variables of the process and on the raw materials. In step two, besides studying the possibility of reducing the dimensionality of the data by PCA or PLS, the possible correlations between matrices must be taken into account. Finally, the implementation requires an appropriate pretreatment of the problem item and to establish that the object meets the specification limits.

The most critical step is the second one where, in order to obtain robust specifications, it is necessary to detect and remove outliers [9,18]. Having defined the data set, we can define the region of multivariate specifications and monitor new objects with

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

multivariate χ^2 or T^2 statistics. χ^2 or chi-square calculates the distance of a new object to the center of a data set when the covariance matrix is known. If the covariance matrix is not known and has to be estimated then the Hotelling T^2 statistic is applied [19]. The monitoring seeks that the samples are lower than the upper control limit (UCL), calculated usually at a 95% or 99% confidence level, since it is assumed that the data follow a normal distribution. Other possible methods of monitoring are the multivariate versions of EWMA and CUSUM charts [19]. However, when the number of variables is substantially increased, it is more difficult to use those monitoring methods [19]. Therefore, the dimensionality of the data set should be reduced with PCA or PLS [17]. PCA is used when only a single data matrix (e.g. quality properties of the product) is considered, so the T^2 statistic is applied to the scores of PCA, and is complemented by the Q (or *SPE*) statistic that takes into account the residuals, that is, the information not modelled by PCA [9]. When data are in different (correlated) matrices (e.g. composition of raw materials and properties of finished products) the dimensionality reduction is done with PLS, since it maximizes the covariance between the two data matrices [17]. Once the scores have been obtained we can apply the T^2 and the *SPE* statistics. Finally, the limits of the specifications are selected to produce the smallest type I and/or type II errors [17].

4.2.2.1 *Multivariate specifications in DPLS*

The way multivariate specifications are used depends on the multivariate analysis method used. It is therefore necessary to study the advantages and disadvantages of using DPLS [14] to define specifications. Take as an example a product with three classes ($C = 3$), e.g. production sites. In this case the set of data will consist of the data matrix **X**, with I objects (representing the products) and J measured variables, and a vector **y** that encodes the classes to which each object belongs. When developing the DPLS models the binarization strategy "one against all" (the class of interest (coded 1) is modelled against the rest (coded 0)) is used. However, to establish multivariate specifications only the scores of the I_c objects of class ω_c (class of interest) are used [15], i.e. for the model ω_1 *vs.* ω_2-ω_3, only the I_1 scores of the class ω_1 are used. Thus, the total number of models developed is equal to the number of classes ($K = C = 3$).

Once the J variables have been reduced to an optimal number of significant factors A, using DPLS, we can monitor new objects using the Hotelling's T^2 [20] statistic:

$$T^2 = \left(\mathbf{t} - \bar{\mathbf{t}}_c\right)^T \mathbf{S}_c^{-1} \left(\mathbf{t} - \bar{\mathbf{t}}_c\right) \quad (1)$$

where **t** is the vector of scores for a new object, $\bar{\mathbf{t}}_c$ and $\mathbf{S}_c$ are the mean vector and covariance matrix, respectively, for the scores of the training samples of class ω_c, these are calculated as:

$$\bar{\mathbf{t}}_c = \frac{1}{I_c}\sum_{i=1}^{I_c}\mathbf{t}_i \tag{2}$$

$$\mathbf{S}_c = \frac{1}{I_c - 1}\sum_{i=1}^{I_c}\left(\mathbf{t}_i - \bar{\mathbf{t}}_c\right)\left(\mathbf{t}_i - \bar{\mathbf{t}}_c\right)^T \tag{3}$$

where I_c is the number of objects in class ω_c, and $\mathbf{t}_i$ is the ith vector of scores for objects in class ω_c. Since Hotelling's T^2 monitoring assumes a normal distribution of the data, the upper control limit (UCL) is calculated as:

$$T^2_{\mathrm{UCL}} = \frac{(I_c - 1)(I_c + 1)A}{I_c(I_c - A)}F_\alpha(A, I_c - A) \tag{4}$$

where $F_\alpha(A, I_c - A)$ is the upper 100α% critical point of the F distribution with A and $I_c - A$ degrees of freedom [13].

Another statistic used to monitor new objects is the squared prediction error (*SPE*), or *Q*. *SPE* provides, in the space of the original variables, the squared difference between the actual and predicted values:

$$SPE = \sum_{J=1}^{J}\left(x_j - \hat{x}_j\right)^2 \tag{5}$$

The upper control limits for *SPE* are based on a χ^2 distribution approximation:

$$SPE_{\mathrm{UCL}} = (v/2m)\chi^2_{1-\alpha}\left(2m^2/v\right) \tag{6}$$

where, v and m are the variance and mean value, respectively, of the *SPE* values of the training objects and $\chi^2_{1\text{-}\alpha}$ is a weighted chi-square distribution ($g\chi^2_h$) with the weight g and h degrees of freedom [21].

In addition, different from other methods, in DPLS the predictions for the class of interest (ideally values around 1) can be used to complement the T^2 and *SPE* statistics. For the predictions of the class of interest it is necessary to establish a lower and an

upper limit. Since the predictions are not necessarily normally distributed, the percentiles from the distribution of objects of the class of interest are used. Thus, the limits at a given confidence level are established from the percentage of objects that are within those limits (for example 95% of the data for a confidence level of 95%). Thus, the multivariate specification is defined by three limits: T^2, *SPE* and $\hat{y}$. The object is within specifications if it fulfils these three requirements simultaneously. That is:

$$T^2_i < \lim T^2_{\mathrm{UCL},\alpha}$$

and

$$SPE_i < \lim SPE_{\mathrm{UCL},\alpha} \qquad (7)$$

and

$$\lim_{\mathrm{low},\alpha} \hat{y} < \hat{y}_i < \lim_{\mathrm{up},\alpha} \hat{y}$$

Although the commonly accepted limits in MSQC are built for α values of 5% or 1%; when defining multivariate specifications these limits must be based on the behaviour of the training data, i.e. we must find a limit that allows a balance between type I and type II errors. This makes it necessary to optimize the limits of T^2, *SPE* and $\hat{y}$.

To apply the specifications to new objects a series of steps must be followed. First, the object must be pretreated in the same way as the training objects (i.e. log transformed, mean centered, autoscaled, and so on). Second, the scores, the x-residuals, and the $\hat{y}$ predicted have to be calculated; and third, Hotelling T^2, *SPE* and $\hat{y}$ statistics have to be monitored to verify that the new objects are within the product specifications.

4.2.3 Experimental part

4.2.3.1 *Data sets*

The establishment of multivariate specifications is illustrated with the data set Olive Oil [8], which contains 166 Italian olive oils belonging to 5 different regions: Liguria (63), Sicilia (28), Lazio (29), Puglia (28) and Umbria (18). 700 variables were measured,

corresponding to the values of absorbance in the near infrared region, measured between 1100 and 2498 nm, with a spectral window of 2 nm.

4.2.3.2 *Procedure and software*

The Kennard-Stone algorithm [22] applied to each class separately was used to split the data set into a training set (with 70% of the objects) and a test set (with 30% of the objects). The olive oil data were mean-centered before the DPLS models were calculated. The procedure is illustrated with the oils from the Liguria class, but it can be extrapolated to other classes. The DPLS models are developed as the class of interest against the other classes, that is, Liguria *vs.* Sicilia, Lazio, Puglia and Umbria. The optimal number of factors was selected by leave-one-out cross validation (LOOCV), and three criteria where tested: minimum type I error, minimum type II error and minimum overall error.

All calculations were done using in-house made Matlab (The Math Works, Inc) subroutines.

4.2.4 Results and discussion

4.2.4.1 *Multivariate specifications. Italian olive oil data set*

Fig. 2 shows the mean centered training data set. No important differences were observed among the objects of the class Liguria (solid line) and the rest (dotted line), except for the objects Liguria025 and Umbria184 that have the most extreme values.

Fig. 3 shows the scores for the first two factors of the DPLS model Liguria *vs.* others (94.6% of cumulative variance explained in **X** and 4% in *y*). At a first glance there is no clear separation between Liguria class and Non Liguria class groups and the objects are evenly spread in the factor space. As it is to be expected from Fig. 2, objects Liguria025 and Umbria184 have the most extreme scores. These objects, however, were not removed from the dataset since the distance to the other objects was not appreciably large.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

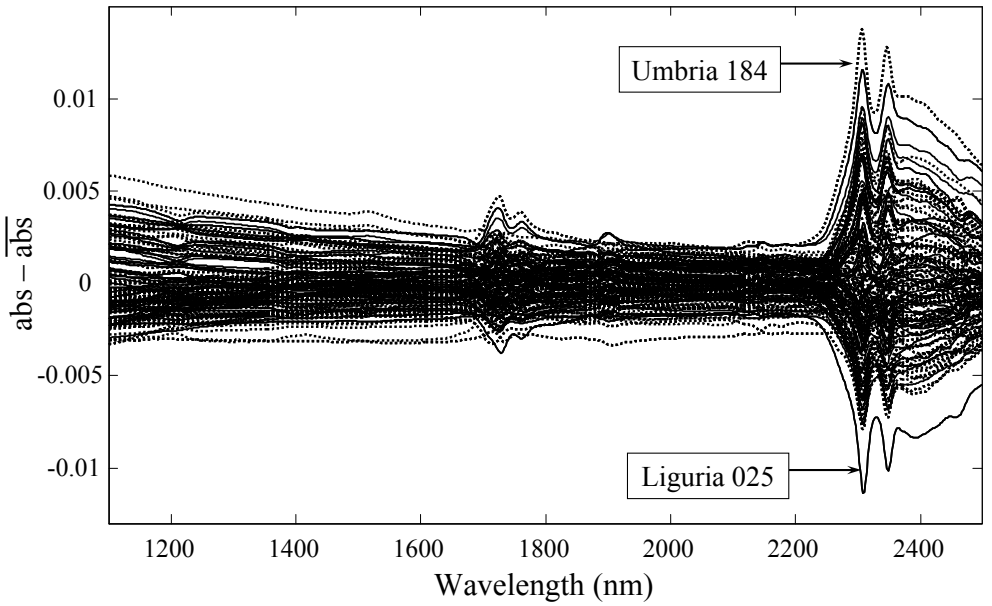

Fig. 2: Olive oil data. Mean-centered training set. Ligurian oils (solid lines) and non Ligurian oils (dotted lines).

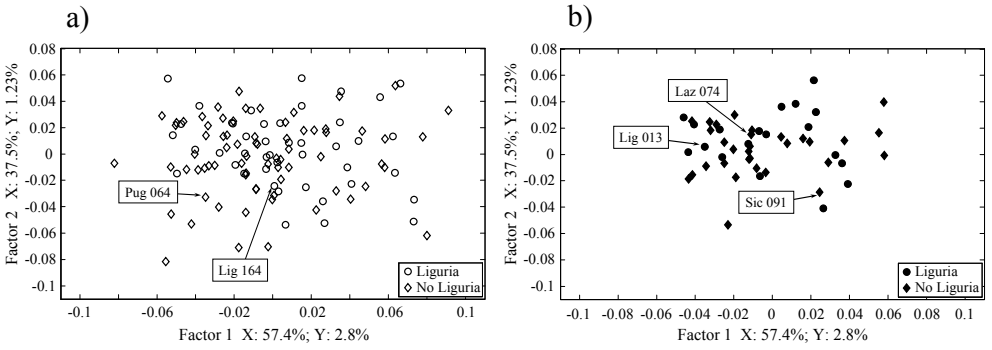

Fig.3: DPLS scores plots for the first two factors. (a) Training set and (b) test set..

After the DPLS had been calculated, the specifications for the class Liguria were established by defining limits for Hotelling T^2, *SPE* and $\hat{y}$ at a confidence level of 99%. Fig. 4 shows the values of Hotelling T^2 versus $\hat{y}$ and Fig. 5 shows the values of *SPE* versus $\hat{y}$, for the DPLS model with 15 factors. By first considering $\hat{y}$, all the Liguria training objects are within the two $\hat{y}$ limits (0% of type I error). However, four non Liguria objects are also within the limits, equivalent to a 5.5% of type II error. This

indicates that the specification of the Liguria class can be almost defined by the two limits of the $\hat{y}$ of the PLS model. By considering the limit value of the Hotelling T^2 statistic only, all Liguria oils fall below that specification limit (0% type I error) but many non Liguria oils are also below the limit (high type II error). The type II error is even higher, almost 100%, for *SPE* (Fig. 5), since most non Liguria oils are below the *SPE* limit. This is because DPLS, unlike PCA, models the two classes simultaneously and when an object is predicted, even if it is not Liguria, it has a low *SPE*. Note that $\hat{y}$ statistic enabled rejection of one non-Ligurian object that was inside the Hotelling T^2 limit. When the limits for Hotelling T^2, *SPE* and $\hat{y}$ are considered together, the type I error was 11.4% and the type II error was 9.7%.

The type I and type II errors depend on the number of factors in the PLS model. For specifications that protect the producer, the optimal number of factors is the one that minimizes the type I error. To protect the consumer, the optimal number of factors is the one that minimizes the type II error. Fig. 6a shows the variation of type I error with respect to number of factors when the theoretical limits of Hotelling T^2, *SPE* and $\hat{y}$ are set at a confidence level of 95%. By considering Hotelling T^2 only, the calculated cross-validation type I error is almost constant at 5% for any number of factors, which

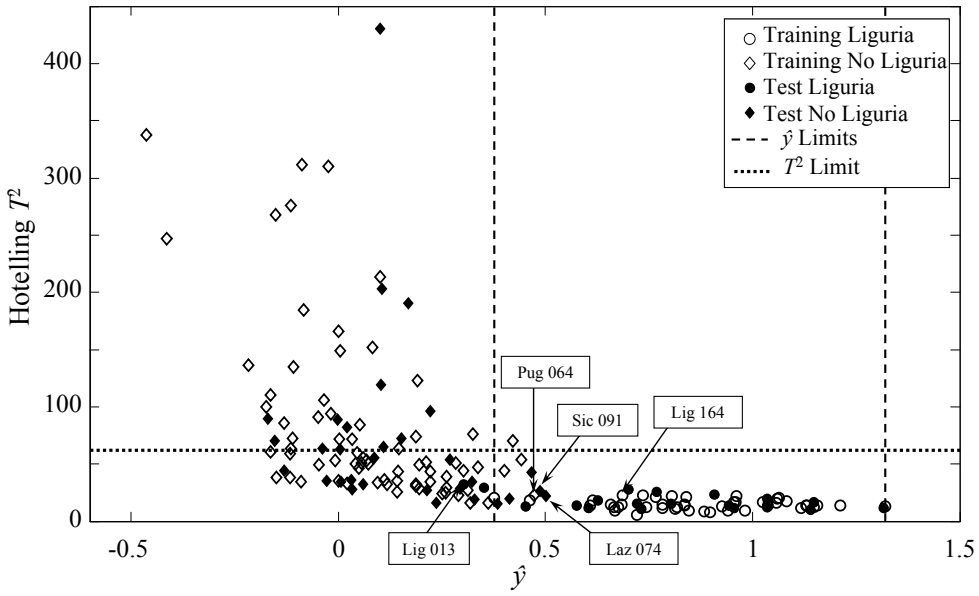

Fig. 4: Hotelling T^2 *vs.* $\hat{y}$ for training and test objects, showing the limits with 15 factors and confidence level of 99%. Some of the objects that have a particular behaviour are indicated.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

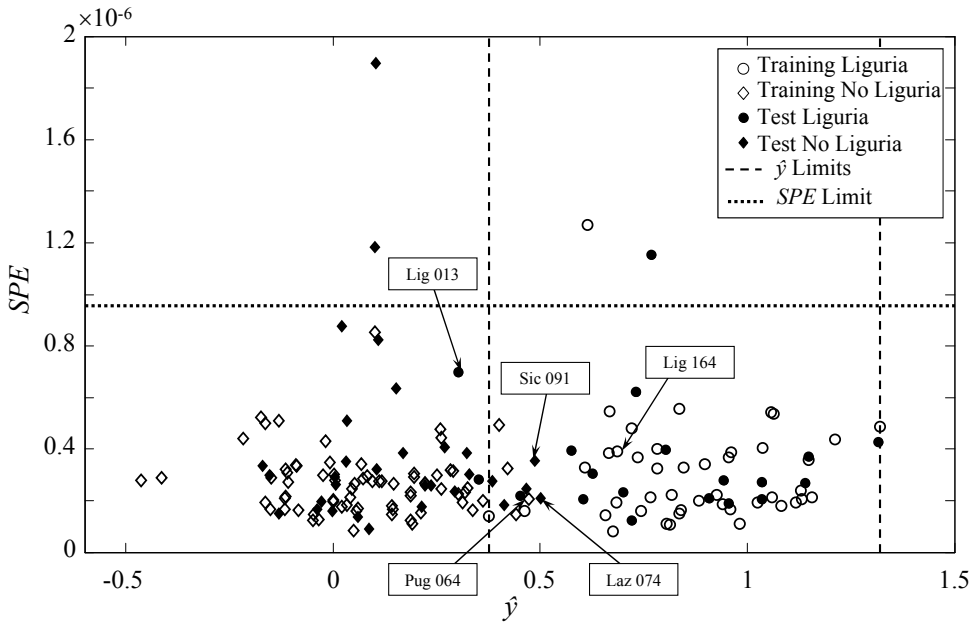

Fig. 5: *SPE vs.* $\hat{y}$ for the training and test objects, showing the limits with 15 factors and confidence level of 99%. Some of the objects that have a particular behaviour are indicated.

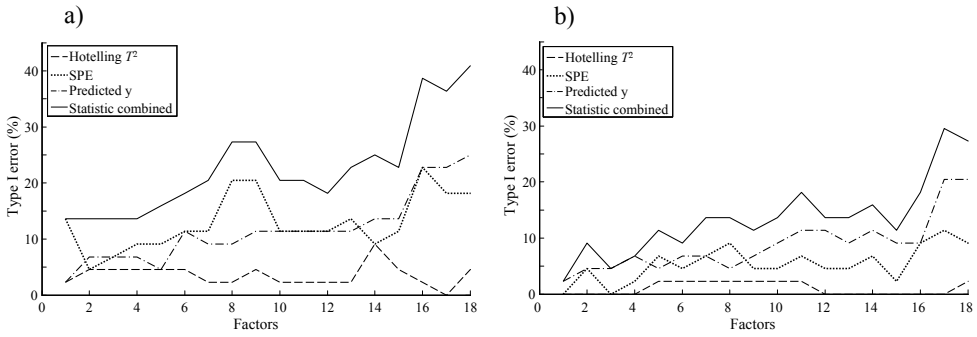

Fig. 6: Variation of the type I error of the Hotelling T^2, *SPE* and $\hat{y}$ statistics and the three statistics combined, assuming a confidence level of 95% (a) and 99% (b).

agrees with the theoretical value. Even for the model with 17 factors the type I error for this statistic can be as low as 0% although this is likely an overfitted model (Fig. 6a). For the *SPE* statistic only, the theoretical 5% type I error is obtained only for the model with 2 factors, and the error increases when more factors are included in the model. The reason for this increase is that the residuals of the training samples

decrease when so does the number of factors, thus making the limit of the *SPE* statistic lower. This makes that the *SPE* value for the cross-validated samples more easily exceeds the *SPE* limit, so more samples are rejected. For the predicted $\hat{y}$, the models from 1 to 5 factors maintain the type I error around the theoretical value of 5% and the error increases for models with more than 5 factors, for similar reasons than for the *SPE* described above (Fig. 6a). Since the objects that are declared out-of-specification by each statistic are not necessarily the same, the type I error of the combined use of the three statics is almost the sum of the type I errors of each statistic. Hence, a 13.6% type I error is obtained for the models with 1 to 4 factors, and then increases the more factors are added to the model. A global type I error of 5% can be obtained by increasing the confidence level of each statistic to 99%. Fig. 6b shows the variation of type I error with respect to number of factors, when the limits of the statistics are set to a theoretical confidence level of 99%. In general, the type I error for Hotelling T^2, *SPE* and $\hat{y}$ statistics decrease up to 0% in some cases. The combined type I error is 4.5% for the PLS model with 3 factors, close to the desired 5% for setting the specification.

To establish specifications that protect the consumer the type II error should be minimized. Fig. 7 shows the variation of type II error with the number of factors. In general, a large number of factors are required to obtain type II errors lower than 10%. For example, for a confidence limit of 95%, a model with 22 factors is required for the Hotelling T^2 statistic and a model with 15 factors for the $\hat{y}$ statistic. The *SPE* statistic required 36 factors, although this can be considered casual because the *SPE* of all objects (both Liguria and non Liguria) was so large that they all were considered out of specification. By combining the three statistics, 15 factors are required in the model (Fig. 7a). Moreover, when the confidence level is increased from 95% to 99%, more factors are required to have minimal type II error, 14 factors with 95% against 16 for

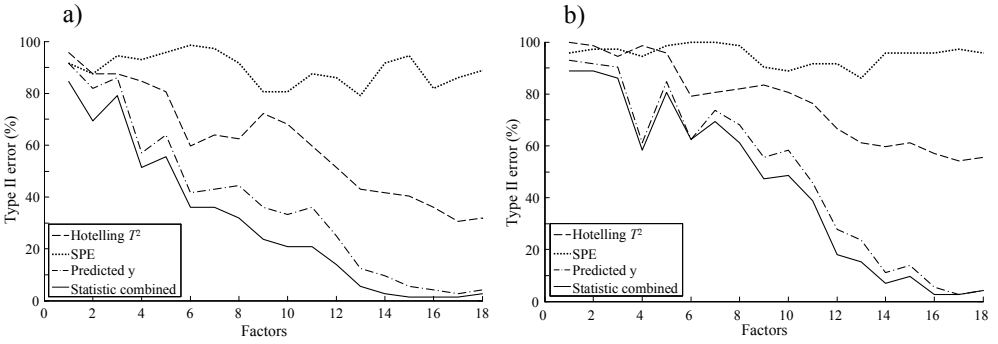

Fig. 7: Variation of the type II error of the Hotelling T^2, SPE and $\hat{y}$ statistics and the three statistics combined, assuming a confidence level of 95% (a) and 99% (b).

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

99% (Fig. 7). This behaviour is opposite to what it was observed with the type I error. Note that the limits on $\hat{y}$ have the largest effect in reducing the type II error. For example, for the model with 14 factors, the combined use of only T^2 and *SPE* statistics gives a type II error of 37.5%, while combining T^2, *SPE* and $\hat{y}$ the error decreases down to 2.8% (Fig. 7).

Given the behaviour observed with type I and II errors, the specification limits for individual statistics (Hotelling T^2, *SPE* and $\hat{y}$) should be established at a confidence level of 99% in order to obtain a combined type I error of 5%. The specification limits for Liguria class are then:

a) For specifications that protect the producer:

Object *i* is within specification if $T^2_i < 13.8$ and $SPE_i < 3.30\times10^{-4}$ and $-0.054 < \hat{y}_i < 1.02$, for a DPLS model with 3 factors. (8)

b) For specifications that protect the consumer:

Object *i* is within specification if $T^2_i < 68.2$ and $SPE_i < 6.52\times10^{-7}$ and $0.51 < \hat{y}_i < 1.31$, for a PLS model with 16 factors. (9)

c) For specifications with balanced error:

Object *i* is within specification if $T^2_i < 62.0$ and $SPE_i < 9.57\times10^{-7}$ and $0.38 < \hat{y}_i < 1.32$, for a PLS model with 15 factors. (10)

Also note that the training objects Liguria164 and Puglia 064 were found out of specification and within specification, respectively, in the cross-validation step. However, in the calibration step, which defines the specifications, both objects are within specifications. The object Liguria164 is near the limit of 99% in the three specifications types. The object has a $T^2 = 51.9$, $SPE = 6.71\times10^{-7}$ and $\hat{y} = 0.380$. For the error balanced specification, $\hat{y}$ is out of the limits (limit of 0.412), while the T^2 and *SPE* values are within the limits ($T^2_{\text{limit}} = 63.4$ and $SPE_{\text{limit}} = 8.73\times10^{-7}$). Since we require the object to simultaneously satisfy the three limits, the object is declared out of specifications. In contrast, the object Puglia064 is far from the three specification types. For example, for error balanced specifications the object has a T^2 of 18.6 (limit of 62.0), a *SPE* of 2.61×10^{-7} (limit of 9.14×10^{-7}) and a $\hat{y}$ of 0.631 against lower and upper limits of 0.370 and 1.28. Since the object is within the three boundaries it is considered to meet specifications.

Table 1: Error performance from Liguria olive oil test data set, specification limits with confidence level of 99%.

Type of specification	Optimal Factors	Type I Error	Type II Error
Protect producer	3	0%	87.1%
Protect consumer	16	26.3%	0%
Balanced error	15	15.8%	16.1%

The defined specifications were checked against a test set. Table 1 shows the percentages of classification for test objects when the limits are set with a theoretical confidence level of 99%. For the specification that protects the producer and the specification that protects the consumer the results are good, with errors lower than for the training set. For example, for specifications that protect the producer, the type I error was 0%, against 4.5% for the training set. For specifications that protect the consumer, the type II error was 0%, against 2.8% for the training set. On the contrary, with the specifications for balanced error, the test data produces larger errors than the training data (type I error of 11.4% and type II of 9.7%). Thus we consider that the specifications defined for the class Liguria are appropriate, although the balanced-error specification has a higher type I and II error because the objects have to simultaneously comply with the three boundaries.

Three test objects require a particular analysis: the object Liguria013 that is out of specification, and the objects Sicilia091 and Lazio074, that are within specifications. For the error-balanced specifications, Liguria013 is within the limits of T^2 (Fig. 4) and *SPE* (Fig. 5) but outside of (although close to) the limits of $\hat{y}$ (Fig. 5). This also occurs for the specification that protects the consumer, so the object is finally declared out of specification in the two cases. Objects Sicilia091 and Lazio074 are within the three limits both of the specification that protects the producer and the balanced-error specification, so they are within specifications.

4.2.5 Conclusions

Multivariate specifications based on T^2, *SPE* and predicted $\hat{y}$ have been established for the NIR spectra of olive oils from the *Liguria* region. Adding limits on $\hat{y}$, together with

the commonly used T^2 and *SPE* statistics, improves the definition of the specification and reduces the number of factors needed in the DPLS model. This optimal number of factors depends on the type of specification (either specification that protect the producer, specification that protect the consumer or specification that provides a balance of type I and type II errors). Note that, in order to reach a general confidence level close to 95%, type I and II error of 5%, the individual confidence levels for T^2 and *SPE* and $\hat{y}$ had to be set to 99%.

Acknowledgments

The authors express their gratitude to Dr. Gerard Downey (Teagasc, Ireland) for providing the Olive oil dataset. This work was supported by the project TRACE – "TRAcing food Commodities in Europe (EU IP 006942)" – from the Sixth Framework Programme of the European and by the Spanish Ministerio de Educación y Ciencia project CTQ2007-66918/BQU. This paper reflects only the authors' views and neither the Community nor the Spanish Ministerio are liable for any use that may be made of the information contained therein.

References

1. http://www.astm.org/COMMIT/Regs.pdf (25/11/2009). ASTM International, "Regulations Governing ASTM Technical Committees", October 2009.
2. D. L. Flumignan, G. C. Anaia, F. de O. Ferreira, A. G. Tininis and J. E. de Oliveira, Chromatographia. 65, 9/10 (2007) 617–623.
3. J. M. Betz, K. D. Fisher, L. G. Saldanha, and P. M. Coates, Anal. Bioanal. Chem. 389 (2007) 19–25.
4. M.L. Weiner, W.F. Salminen, P.R. Larson, R.A. Barter, J. L. Kranetz, G. S. Simon, Food Chem. Toxicol. 39 (2001) 759–786.
5. M. R. Hubbard, Statistical quality control for the food industry, Third Edition, Kluwer Academic / Plenum Plublishers, USA, 2003, pp. 253–276.
6. S. Kelly, K. Heaton, J. Hoogewerff, Trends Food Sci. Tech. 16 (2005) 555–567.
7. http://ec.europa.eu/agriculture/foodqual/quali1_en.htm. Last accessed in 21st of January 2010.

8. Project TRACE – "TRAcing food Commodities in Europe" (Project no. FOOD-CT-2005-006942). www.trace.eu.org.
9. L. H. Chiang, L. F. Colegrove, Chemometr. Intell. Lab. Syst. 88 (2007) 143–153.
10. D. M. Ennis, J. Bi, J. Food Qual. 23 (2000) 541–552.
11. M. Novič, N. Grošelj, Anal. Chim. Acta. 649 (2009) 68–74.
12. D C. Montgomery, G C. Runger, Applied Statistics and Probability for Engineers, Third Edition, John Wiley & Sons, Inc., USA, 2003, pp. 595–648.
13. T. Kourti, J. F. MacGregor, Chemometr. Intell. Lab. Syst. 28 (1995) 3–21.
14. B. R. Kowalski, Chemometrics, Mathematics and Statistics in Chemistry, D. Reidel Publishing Company, Dordrecht, Holland, 1984, pp. 85–88.
15. G. Ou, Y. L. Murphey, Pattern Recogn. 40 (2007) 4–18.
16. M. Guidolin, A. Timmermann, J. Econometrics. 131 (2006) 285–208.
17. C. Duchesne, J. F. MacGregor, J. Qual. Tech. 36, 1 (2004) 78–94.
18. M-J. Bruwer, J. F. MacGregor, W. M. Bourg Jr., Food Quality and Preference. 18 (2007) 890–900.
19. C. Wikström, C. Albano, L. Eriksson, H. Fridén, E. Johansson, Å. Nordahl, S. Rännar, M. Sandberg, N. Kettaneh-Wold, S. Wold, Chemometr. Intell. Lab. Syst. 42 (1998) 221–231.
20. R. A. Johnson, D. W. Wichern, Applied multivariate statistical analysis, Fifth edition, Prentice Hall, USA, 2002, pp: 210–252.
21. A. Nijhuis, S. de Jong, B. G. M. Vandeginste, Chemometr. Intell. Lab. Syst. 38 (1997) 51–62.
22. R. W. Kennard, L. A. Stone, Technometrics, 11 (1969) 137–148.

4.3 Especificaciones multivariantes para mieles de la región *Córcega*

4.3.1 Introducción

La comercialización de miel en el territorio europeo esta reglamentada en la directiva 2001/110/CE que establece la características mínimas que deben tener las mieles para ser consideradas naturales [1], ya sean destinadas al consumidor final o para uso industrial. Esta directiva en su anexo II establecen 8 parámetros mínimos a ser considerados y sus correspondientes límites superiores. Sin embargo, estos 8 parámetros no son suficientes para certificar la autenticidad de las mieles, pues se trata de parámetros genéricos que aseguran que la miel es natural y no ha sufrido procesos que la degraden o alteren. Para autenticar una miel se requieren análisis que brinden una mayor información sobre la miel, como estudio del polen o análisis por espectroscopia infrarroja, y en este caso, análisis mediante CG-MS. Se busca establecer si la información suministrada por este análisis es suficiente para implementar la especificación de producto para la clase *Córcega*.

4.3.2 Parte experimental

Para este estudio se utilizó el conjunto de datos de mieles europeas proveniente del proyecto TRACE (detalles en el anexo, apartado 2), con 180 muestras, 26 variables de intensidad absoluta de masas de componentes de la miel y 9 clases según región: *Córcega*, *Sicilia*, *Toscana*, *Trentino*, *Marchfeld*, *Muehlviertel*, *Carpentras*, *Limosín* y *Bavaria*. Se aplicó el algoritmo *Kennard-Stone* a datos autoescalados para dividirlo en datos de entrenamiento (70% de los datos) y datos de prueba (30% de los datos). Se desarrolló el modelo DPLS *Córcega vs.* Restantes. El número óptimo de factores se decidió por validación cruzada dejando fuera un objeto cada vez (LOOCV), con criterios de mínimo error de tipo I (protección al productor), tipo II (protección al consumidor) y total (balance del error). En los tres casos se consideró un nivel de confianza del 99% para los límites de los estadísticos, *Hotelling* T^2 y ***SPE***, y las predicciones, $\hat{y}$.

4.3.3 Resultados y discusión

Para establecer las especificaciones de la clase *Córcega* se definió el conjunto de datos adecuado. Así, basándonos en el análisis por PCA (anexo, apartado 2) y *p*-DPLS (apartado 2.8.3) se eliminaron los objetos discrepante de la clase *Carpentras*, además de dos objetos de la clase *Córcega* y uno de la *Sicilia*. El conjunto de datos para establecer especificaciones quedó definido por 170 mieles. Al establecer las especificaciones se tuvo en cuenta que las futuros objetos debían cumplir con tres estadísticos simultáneamente, el *Hotelling* T^2, el *SPE* y la *y* predicha; además de tener en cuenta el destino de la especificación, protección al productor, protección del consumidor o equidad entre los dos.

La figura 4-2 muestra como evolucionaron los errores de tipo I y II con LOOCV al variar el número de factores para los límites fijados a un nivel de confianza del 99%. Se observa que con pocos factores hay un elevado error de tipo I y que éste va en aumento con el número de factores (figura 4-2a). Se consideró un nivel de confianza del 99%, que nos aseguraría rechazar pocos objetos *Córcega*. No obstante, la elevada dispersión que presentaron los objetos *Córcega* ocasionó un aumento del *SPE* y una mayor dispersión en las $\hat{y}$ predichas por el modelo, por ende mayor error en los estadísticos individuales. Así, al combinar los estadísticos el error de tipo I se elevó. Similar para el error de tipo II, que no descendió a más del 13.3% con 14 factores. en este caso se tienen dos causas por las cuales se aceptaron objetos diferentes a la clase *Córcega*. Primero, se tienen 7 clases reunidas en una sola, y segundo, los objetos de estas clases sólo representan el 37% del total de objetos. Es decir, se tienen pocos objetos por clase, lo que no permitió distinguirlas correctamente de la clase *Córcega*; por

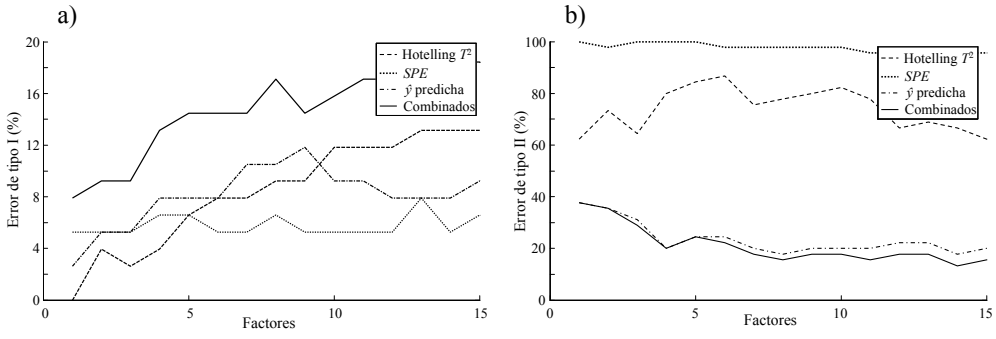

Figura 4-2: Variación del error del error de tipo I (a) y II (b) para los estadísticos *Hotelling* T^2, *SPE* e $\hat{y}$ asumiendo un nivel de confianza del 99%.

consiguiente estos objetos podían tener valores de *Hotelling* T^2 por debajo del límite fijado para el estadístico, aumentando así el error del tipo II. Sin embargo, al combinar los estadísticos y la predicción, la $\hat{y}$ permitió rechazar muchos de los objetos diferentes a *Córcega* que fueron aceptados por los otros dos estadísticos, reduciendo el error de tipo II (figura 4-2b).

De esta manera, los límites de especificación para la clase *Córcega* se fijaron como:

a) Para especificaciones que protegen al producto:

Un objeto i esta dentro de las especificaciones si $T^2_i < 10.1$ y $SPE_i < 78.9$ y $0.342 < \hat{y}_i < 1.35$, para un modelo DPLS con 2 factores (4-12)

b) Para especificaciones que protegen al consumidor:

Un objeto i esta dentro de las especificaciones si $T^2_i < 40.9$ y $SPE_i < 17.4$ y $0.368 < \hat{y}_i < 1.46$, para un modelo DPLS con 14 factores (4-13)

c) Para especificaciones con balance de errores:

Un objeto i esta dentro de las especificaciones si $T^2_i < 22.4$ y $SPE_i < 32.3$ y $0.364 < \hat{y}_i < 1.46$, para un modelo DPLS con 7 factores (4-14)

La tabla 4-1 muestra lo porcentajes de clasificación de los datos de entrenamiento (LOOCV) y prueba a un nivel de confianza del 99% para la tres especificaciones. En general el error con los datos de entrenamiento fue elevado, siendo el menor de ellos para especificaciones que protegen al productor con un 9.2% (7 objetos *Córcega* rechazados). Aunque mejor que el error de tipo II del 13.3% para especificaciones que protegen al consumidor (6 objetos erróneamente considerados *Córcega*). Para el balance de errores los porcentajes de error de tipo I y II son aún mayores. No obstante, al evaluar los objetos de prueba se produjo una notable reducción del error. Así, para especificaciones que protegen al productor el error de tipo I fue del 0%, para especificaciones que protegen al consumidor el error de tipo II fue del 6.2% y para balance de errores los errores del tipo I y II fueron 3% y 6.2%, respectivamente. Es decir, que tanto el modelo DPLS como las especificaciones derivadas de éste (Ecs. 4-7, 4-8 y 4-9) son adecuadas para verificar si objetos nuevos cumplen con las respectivas especificaciones.

Las figuras 4-3 y 4-4 muestran la distribución de mieles de entrenamiento y prueba de acuerdo a los valores de sus estadísticos, *Hotelling* T^2 y *SPE*, y las predicciones $\hat{y}$ calculados para el modelo DPLS con 7 factores. Tanto en la gráfica de *Hotelling* T^2 *vs.* $\hat{y}$

(Figura 4-3) como en la *SPE vs.* $\hat{y}$ (Figura 4-4) se observa que los límites de especificación para $\hat{y}$ son los que tienen mayor peso a la hora de verificar las especificaciones. El caso opuesto se observa con los límites por *Hotelling* T^2 ya que la mayoría de mieles están por debajo del límite. Igual para el límite por *SPE*, todas la mieles están por debajo del límite, únicamente un objeto *Córcega* está por encima de éste límite. Si las especificaciones de mieles para la región *Córcega* dependieran de un único estadístico, $\hat{y}$ sería la mejor opción para establecerlas, ya que presenta los menores errores de tipo I y II, mientras que el *SPE* sería la peor opción, pues tiene un elevado error de tipo II. Otro aspecto a tener en cuenta es que las mieles *Córcega* fuera de especificación en *Hotelling* T^2 y *SPE* no son las mismas. Así, al verificar las especificaciones en los tres estadísticos, las diferentes respuestas de éstos sobre qué mieles *Córcega* están fuera de las especificaciones, aumentan el número de mieles consideradas fuera de las especificaciones y por consiguiente que se eleve el error de tipo I.

Que cada una de las mieles de entrenamiento o prueba deban cumplir con los tres límites de especificación simultáneamente es determinante en el elevado error de tipo I. Sin embargo, es también esta condición, cumplir con los tres límites, la que permite reducir el error de tipo II; ya que como se observó *Hotelling* T^2 y *SPE* presentan elevados errores de tipo II. Hay que destacar que algunos objetos como FRCOR004 y ITTRE127 fueron declarados erróneamente fuera y dentro de las especificaciones, respectivamente; pero se encuentran cerca del límite de especificaciones, y estos objetos en particular cerca del límite inferior de $\hat{y}$. En otras ocasiones los objetos pueden estar cerca de los límites para alguno de los estadísticos, como el objeto ITSIC118 que esta cerca del límite de *Hotelling* T^2, pero el *SPE* no deja dudas que debe ser declarado fuera de especificación.

Tabla 4-1: Porcentajes de error con datos de entrenamiento (LOOCV) y prueba de las mieles *Córcega* para especificación con un nivel de confianza del 99%.

Tipo de especificación	Conjunto de datos	Factores óptimos	Error de tipo I	Error de tipo II
Protección del productor	Entrenamiento	2	9.2%	35.6%
	Prueba		0%	12.5%
Protección del consumidor	Entrenamiento	14	18.4%	13.3%
	Prueba		3.0%	6.2%
Balance de errores	Entrenamiento	7	14.5%	17.8%
	Prueba		3.0%	6.2%

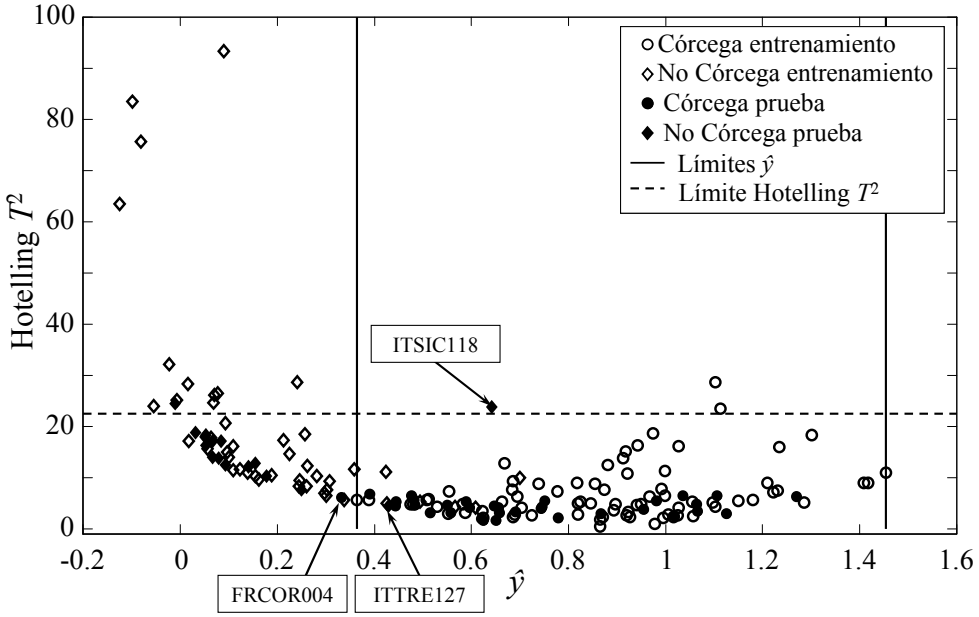

Figura 4-3: Gráfica de *Hotelling* T^2 *vs.* $\hat{y}$ para los datos de entrenamiento y prueba de mieles europeas, modelo DPLS con 7 factores.

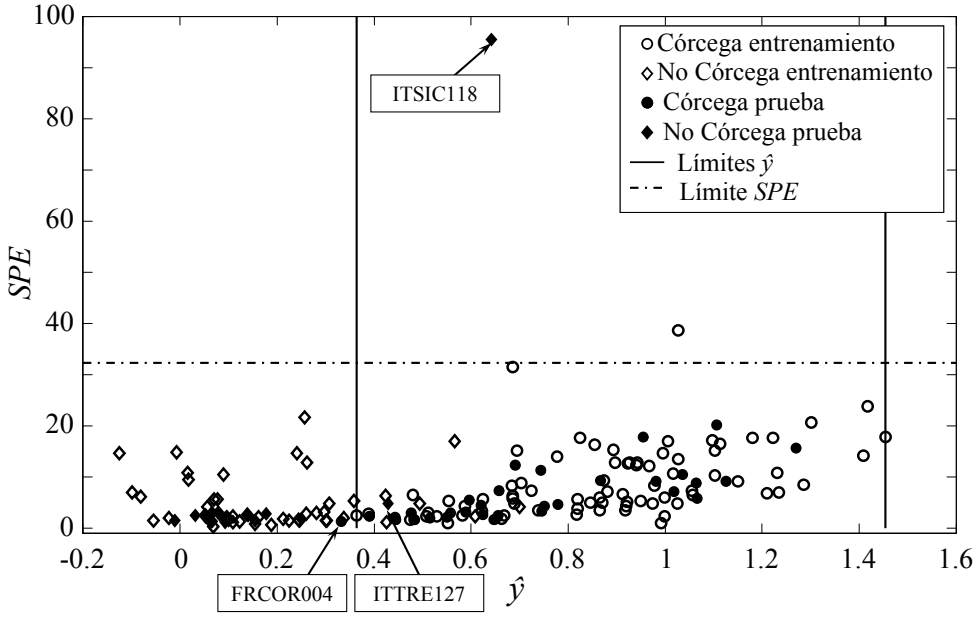

Figura 4-4: Gráfica de *Hotelling* T^2 *vs.* $\hat{y}$ para los datos de entrenamiento y prueba de mieles europeas, modelo DPLS con 7 factores.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

4.3.4 Conclusiones

Es posible establecer especificaciones multivariantes para mieles de la región *Córcega* con un nivel de confianza del 99% en los estadísticos individuales y la predicción. Aunque, con los objetos de entrenamiento se obtuvo errores elevados, los objetos de prueba demostraron la factibilidad de las especificaciones al obtener errores cercanos al 5%, que consideramos equivalentes a un nivel de confianza del 95% para los estadísticos combinados con la predicción. No obstante, se puede atribuir parte del error al desbalance del número de mieles, en donde la clase *Córcega* tiene cerca del 60% de los objetos, y las otras 7 clases presentes no más del 10% de los objetos cada una, lo cual no permite caracterizarlas correctamente y por ende aumenta la posibilidad de confundirlas con mieles *Córcega*. Por ello los elevados errores de tipo II. Se observa que la predicción $\hat{y}$ es determinante para reducir el error de tipo II al rechazar objetos que los otros dos estadísticos han aceptado.

Referencias

1. Directiva 2001/110/CE del Consejo de 20 de diciembre de 2001 relativa a la miel, Diario Oficial de las Comunidades Europeas, L 10 (12/1/2002) 47–52.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Capítulo 5

Conclusiones

5.1. Conclusiones generales

El presente trabajo, enmarcado dentro del proyecto europeo TRACE de seguridad alimentaria, aplicó DPLS a la autentificación de alimentos a partir de datos multivariantes. Como resultado del trabajo, se desarrolló el método *p*-DPLS de clasificación para sistemas biclase y multiclase que, además, proporciona la fiabilidad de la clasificación. El algoritmo *p*-DPLS fue implementado en lenguaje *Matlab*®. Adicionalmente, DPLS se utilizó para derivar especificaciones multivariantes de alimentos como un nuevo criterio para la autentificación de los mismos.

1. El desarrollo del algoritmo *p*-DPLS implicó combinar el DPLS clásico, funciones potenciales, incertidumbre de predicción y decisión bayesiana. De este desarrollo se concluyó que:

- La combinación de funciones potenciales individuales permite obtener funciones de densidad de probabilidad (FDP) que representan mejor la distribución de las predicciones de cada clase que utilizar una función gaussiana calculada a partir de

la media y la desviación estándar de las predicciones de cada clase. Esta ventaja es más notable cuando el número de objetos de una clase es pequeño.

- Utilizar el error estándar de calibración (*SEC*) para cada función potencial individual permite considerar el comportamiento de cada objeto en el espacio multivariante del modelo DPLS. Objetos con un *SEC* pequeño generan funciones potenciales estrechas y dan más importancia al valor de la predicción que objetos con un *SEC* grande.
- Se propuso el cálculo de la fiabilidad de clasificación de un objeto desconocido utilizando su error estándar de predicción (*SEP*) para calcular un área bajo las FDP de las clases, y la ecuación de Bayes. Dicha fiabilidad incluye información tanto de los datos de entrenamiento, representada en las FDP de clase, como del objeto desconocido, representada en el *SEP* y la *y* predicha ($\hat{y}$). La fiabilidad es mayor con FDP estrechas y bajos *SEP*. Por el contrario, se obtienen fiabilidades bajas, cercanas al 50% cuando se tienen FDP amplias, altos *SEP* y predicciones cercanas al límite de solapamiento entre clases.

Además, se encontró que la predicción del modelo DPLS, a partir de la cual se decide la clase asignada, está afectada por un sesgo que depende de la clase a la cual pertenecen los objetos del modelo. Por ello se estableció el cálculo del error cuadrado medio de calibración (*MSEC*) teniendo en cuenta el sesgo para cada clase. Así, se obtiene un *MSEC* con el sesgo corregido que depende de la dispersión de las predicciones de cada una de las clases y no de la posición de las predicciones con respecto al valor binario asignado, que es lo que ocasiona el sesgo. Con este *MSEC* corregido se pueden obtener *SEP* bajos que mejoran la fiabilidad de clasificación de los objetos.

2. El programa en *Matlab* desarrollado a partir de *p*-DPLS e incluido en *ChemTRACE* se presentó en diversas reuniones del proyecto TRACE, recibiendo una buena valoración tanto desde el punto de vista técnico como de presentación gráfica, adoptándose la interfaz de ingreso para los demás módulos de *ChemTRACE*. Igualmente, el programa fue validado y utilizado por otros miembros del proyecto.

3. Resolver problemas multiclase con el método *p*-DPLS binario y la estrategia de binarización "uno contra todos" presenta ciertas limitaciones, debido al desbalance en el número de objetos entre la clase de interés y la superclase que engloba las clases restantes, además de presentar inconsistencias de asignación (por ejemplo, modelos binarios distintos puede acabar asignando un objeto a clases distintas). Se planteó una metodología que utiliza la fiabilidad de clasificación de modelos binarios *p*-DPLS, siguiendo la estrategia de binarización "uno contra uno". Con esta metodología:

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

- Se calculó la fiabilidad de multiclasificación como el producto de fiabilidades parciales de un objeto a una de las clases sobre la suma del producto de fiabilidades para todas las clases. Este método subsana las incongruencias debidas a la estrategia de binarización, como asignar el objeto a más de una clase, mejorando a su vez el nivel de confianza sobre la asignación.
- El método *p*-DPLS multiclase resultó ser más efectivo al asignar objetos desconocidos que los métodos CART y SIMCA.
- Se obtuvo una fiabilidad de clasificación acorde a la distribución de las clases en el espacio original de variables. Cuando un par de clases estaban solapadas se mejoró el porcentaje de clasificación con respecto a los modelos "uno contra todos", pero manteniendo un intervalo de fiabilidad bajo. Cuando las clases estaban bien definidas el método *p*-DPLS multiclase mejoró tanto el porcentaje de clasificación como la fiabilidad.

Además, se demostró que es posible resolver problemas muticlase con *p*-DPLS utilizando otros métodos de combinación de resultados de clasificadores base. Aunque con dichos métodos no se calculó la fiabilidad de multiclasificación.

4. Se ha propuesto una metodología para establecer especificaciones multivariantes de alimentos a partir los estadísticos *Hotelling* T^2 y ***SPE*** basados en los *scores* de DPLS y la predicción del modelo DPLS ($\hat{y}$).

- El incorporar la predicción del modelo DPLS ($\hat{y}$) para establecer las especificaciones permitió reducir notablemente el error de tipo II (considerar productos dentro de la especificación cuando están fuera de ella), ya que deja fuera de especificación aquellos objetos erróneamente aceptados por los otros estadísticos.
- Para establecer un nivel de confianza de 5% en la especificación, que es la combinación de los dos estadísticos y la predicción, es necesario establecer niveles de confianza menores en los estadísticos y la predicción, un 1%, dado que cada uno de ellos declara fuera de especificación objetos distintos.
- El número de factores del modelo PLS afecta a los errores de tipo I, de tipo II, o su combinación. Su optimización permite establecer especificaciones para diferentes situaciones, como protección del productor (reducción del error de tipo I), protección del consumidor (reducción del error de tipo II) o cuando se busca un balance entre los dos tipos de error.

5. *p*-DPLS se aplicó a la clasificación de suelos europeos en tres grandes litologías (*arenisca*, *caliza* y *esquisto*). Los resultados de aplicar el método *p*-DPLS fueron:

- Los modelos binarios *p*-DPLS con estrategia "uno contra todos" presentaron porcentajes de clasificación, para las tres litologías, superiores al 95%, con fiabilidades para las clases *caliza* y *esquisto* cercanas al 100%. Para la clase *arenisca* sólo se alcanzó una fiabilidad promedio del 80%. Además, se observaron fiabilidades diferentes para objetos con predicciones similares, ya que la fiabilidad depende del *SEP* (menor fiabilidad a mayor *SEP*). La amplitud de las FDP de las clases fue crítica a la hora de hallar la fiabilidad; así, FDP amplias implicaron fiabilidades promedio menores que las de FDP estrechas.
- Para el método *p*-DPLS multiclase con modelos "uno contra uno" se obtuvieron porcentajes de clasificación superiores al 95%, con fiabilidades de clasificación para las tres litologías superiores al 95%, con excepción de algunos objetos de *arenisca* y *esquisto* que presentaron valores menores. También se concluyó que el sitio de recolección y los laboratorios que realizaron los análisis no influyen en la clasificación.
- Se observó que la estrategia de binarización "uno contra uno" es superior a la "uno contra todos", porque mejora la separación de las clases y, por ende, la fiabilidad de clasificación.

6. Al aplicar *p*-DPLS para discriminar mieles de la región de *Córcega* de otras mieles europeas los resultados fueron:

- Para *p*-DPLS binario se obtuvo una sensibilidad superior al 97% pero el error de tipo II fue elevado, debido a que la estrategia de binarización "uno contra todos" genera un desbalance entre el número de objetos de la clase *Córcega* y las restantes clases reunidas en una sola clase. Además, algunas de las otras clases eran similares a *Córcega* y se clasificaron erróneamente como *Córcega*, elevando el error de tipo II. Este comportamiento dio origen a unas FDP amplias que se solapaban, y produjo que la fiabilidad promedio de clasificación de la clase *Córcega* fuera menor del 90% y para las clases diferentes a *Córcega* inferior al 80%.
- Al establecer la especificación multivariante de la región *Córcega* se obtuvo más error de tipo I del esperado, según el modelo *p*-DPLS biclase. Ello se debió a que las especificaciones se basan en los estadísticos *Hotelling* T^2, ***SPE*** y la predicción $\hat{y}$, que en conjunto declaran más objetos fuera de especificación que cualquiera de ellos solos, más si se tiene en cuenta que los límites se fijan sólo con las respuestas de los objetos de la clase de interés. Por contraste, el error de tipo II mantuvo un porcentaje cercano o menor al de clasificación. Las especificaciones se consideraron adecuadas ya que los errores obtenidos con los objetos de prueba fueron cercanos al 5%, mostrando que las especificaciones son apropiadas para objetos nuevos.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

7. El conjunto de datos de aceites de oliva europeos fue analizado por dos métodos espectroscópicos, lo que permitió observar la variabilidad de la información y determinar qué método sería más adecuado para análisis futuros. Al utilizar *p*-DPLS para clasificar la región *Liguria* de las restantes, partiendo de los espectros NIR, los resultados fueron:

- Para *p*-DPLS binario se obtuvieron porcentajes de sensibilidad y especificidad superiores al 90% y fiabilidades cercanas al 100%, siendo la mayoría de objetos mal clasificados o con baja fiabilidad los cercanos al límite entre clases.
- Cuando se aplicó el método *p*-DPLS multiclase para clasificar 5 clases de aceites italianos, fue posible diferenciarlas con sensibilidades superiores al 75% y fiabilidades cercanas al 100%. Las bajas clasificaciones se debieron a la poca diferencia entre clases. Aun así, el conjunto de prueba obtuvo porcentajes cercanos al 100%. Dichos resultados se compararon con los métodos de clasificación CART y SIMCA y se observó que el *p*-DPLS multiclase es más apropiado para clasificar objetos desconocidos.

Al aplicar *p*-DPLS al conjunto de espectros ^{1}H-RMN de aceites de oliva europeos, los resultados fueron:

- Para *p*-DPLS biclase se observó una baja sensibilidad (72%) que contrastó con una especificidad del 98%. Igualmente, la fiabilidad de clasificación para la región *Liguria* fue en promedio inferior al 70% mientras que para los objetos diferentes a *Liguria* estuvo alrededor del 90%. Estos pobres resultados, sumado a un número óptimo de factores del modelo *p*-DPLS muy elevado, hizo concluir que *p*-DPLS no es adecuado para clasificar la clase *Liguria* de las restantes con este tipo de datos.
- Por contraste, los espectros de RMN permitieron diferenciar años de recolección con *p*-DPLS, dando sensibilidades cercanas al 100% y fiabilidades que superaron el 96%.

Al comparar las respuestas de ambas técnicas espectroscópicas se consideró que los espectros de RMN no permiten diferenciar regiones, al contrario que los espectros NIR, con los que se obtienen buenos porcentajes de clasificación y fiabilidades elevadas. No obstante, el conjunto de espectros RMN es más apropiado para identificar años de recolección, siendo relevante a la hora de verificar la trazabilidad del alimento.

5.2. Futuras investigaciones

Dado que *p*-DPLS es un algoritmo nuevo aún se debe probar con más conjuntos de datos. Es necesario también mejorar su robustez y la detección de objetos discrepantes, tanto para el conjunto de objetos de entrenamiento como de prueba. Para la detección de los objetos discrepantes se sugiere partir de FDP de cada una de las clases y definir los límites superiores e inferiores de estas FDP. Así pues, un objeto será discrepante si está fuera de estos límites. Esta capacidad de detección es necesaria cuando se utiliza la estrategia de binarización "uno contra uno" para resolver problemas multiclase, ya que es necesario predecir objetos de clases que no se han modelado en los modelos binarios, pero sí forman parte del problema multiclase. Igualmente, queda pendiente mejorar el cálculo de la fiabilidad.

Otro campo en el que *p*-DPLS puede tener una notable aplicación es en la resolución de problemas multiclase. En esta tesis se propuso una metodología para aplicarlo, pero también quedó abierta la posibilidad de utilizar funciones de densidad de probabilidad multivariante. Sin embargo, aún se debe establecer cómo calcular la fiabilidad a partir de funciones de densidad multivariante, dado que se construyen funciones en un espacio de 3 o más dimensiones, dependiendo del número de modelos utilizado, y por ello se requiere analizar cómo los solapamientos de clases influyen en el cálculo de fiabilidad.

Otra problemática observada se relaciona con el número de clases que es necesario discriminar. El método de binarización se ve limitado por el número de clases, ya que al aumentar el número de clases el número de modelos binarios crece de forma significativa. A este respecto quedaría pendiente la adaptación de métodos ya existentes, como las matrices *ECOC*, que permitan aplicar el algoritmo *p*-DPLS y a su vez calcular la fiabilidad de clasificación. Otro camino para resolver estos problemas multiclase es la clasificación por niveles. Esta clasificación se basa en modelos binarios representados como nodos bidireccionales, unidos entre sí y que siguen una única dirección hasta poder discriminar cada una de las clases, o nodo terminal (de forma similar a las estructuras de CART). Esta configuración permite aplicar los conceptos de redes bayesianas y establecer la propagación de la fiabilidad entre los nodos de la red para hallar la fiabilidad final de cada una de las clases, asignando así un objeto a la clase más probable.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

ANEXOS

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Notación y acrónimos

ANN	Artificial Neural Networks
ASTM	American Society for Testing and Materials
CART	Classification and Regression Trees
CE	Comisión Europea
CEN	Comité Européen de Normalisation
CG	Cromatografía de Gases
CUSUM	Cumulative Sum
DAG	Directed Acyclic Graph
DOP	Denominación de Origen Protegida
DPLS	Discriminant Partial Least Squares
ECOC	Error-Correcting Output Code
ETG	Especialidad Tradicional Garantizada
EWMA	Exponentially Weighted Moving Average
FAO	Food and Agriculture Organization
FDP	Función de Densidad de Probabilidad
FDPNM	Función de Densidad de Probabilidad Normal Multivariante
ICP-MS	Inductively Coupled Plasma Mass Spectrometry
IGP	Indicación Geográfica Protegida
IID	Independientes e Idénticamente Distribuidos
ISO	International Organization for Standardization
IUPAC	International Union of Pure and Applied Chemistry
k-NN	k-Nearest Neighbours
LDA	Linear Discriminant Analysis
LDH	Lactato Deshidrogenasa
LOOCV	Leave-One-Out Cross-Validation
MIR	Mid Infrared
MSE	Mean Square Error
MSEC	Mean Square Error of Calibration
$MSEC_{bc}$	Bias Corrected Mean Squared Error of Calibration
MSEP	Mean Square Error of Prediction
MSPC	Multivariate Statistical Process Control
MSQC	Multivariate Statistical Quality Control
N-PLS	Multiway PLS
NIPALS	Nonlinear Iterative Partial Least Squares
NIR	Near Infrared
OAA	One-against-All

OAO	One-against-One
OLS	Ordinary Least Squares
OMS	Organización Mundial de la Salud
p-DPLS	Probabilistic DPLS
PAQ	P-against-Q
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares Regression
RD	Real Decreto
RMN	Resonancia Magnética Nuclear
SEC	Standard Error of Calibration
SEP	Standard Error of Prediction
SIMCA	Soft Independent Modelling of Class Analogy
SPE	Square Prediction Error
SQC	Statistical Quality Control
SVM	Support Vector Machines
TGA	Thermogravimetric Analysis
TRACE	TRAcing food Commodities in Europe
UCL	Upper Control Limit
UE	Unión Europea
WMV	Weighted Majority Vote
XRF	X-ray Fluorescence

A	Número de factores
A_{opt}	Número de factores óptimos
arg max	Selección de la respuesta con mayor argumento
$\mathbf{B}$	Matriz de coeficientes de regresión
$\mathbf{b}$	Vector de coeficientes de regresión
bias	Sesgo de la predicción
$bias_c$	Sesgo de predicción para los objetos de la clase c
C	Número de clases en la matriz de datos
c	La clase c de la matriz de datos
$d_{c,k}$	Voto que un modelo k da al objeto si perteneciera a la clase c
e	Residuales
$f(y_i\|\omega_c)$	Función de densidad de probabilidad para toda y_i que pueda pertenecer a la clase ω_c
$f_c(\mathbf{x})$	Función de decisión del clasificador base para una clase c
$g(y_u,y_i^c)$	Función individual para los elementos $y_i{}^c$ evaluado en el punto y_u
h	Leverage
I	Número de objetos en un conjunto de datos

J	Número de variables de un objeto
K	Número de modelos binarios
k	El modelo k del conjunto de modelos K
l_k	Coeficiente de ponderación para el clasificador k
$\mathbf{P}$	Matriz de loadings de $\mathbf{X}$
$P(\omega_c)$	Probabilidad de clase o *a priori*
$P(\omega_c \mid y_u)$	Probabilidad de pertenecer a la clase ω_c dado un valor observado y_u o probabilidad *a posteriori*
p_c	Sumatoria de probabilidades de que un objeto pertenezca a la clase c
$p(E\|D)$	Probabilidad de un suceso E condicionado a una proposición D
$p(y_i\|\omega_c)$	Probabilidad condicionada de un objeto y_i que pueda pertenecer a la clase ω_c
$\mathbf{q}$	Vector de loadings de $\mathbf{y}$
$R_{c,k}$	Fiabilidad de clasificación a la clase c por el modelo k
$\mathbf{S}_c$	Matriz de covarianzas de los *scores*
$\mathbf{T}$	Matriz de *scores* de $\mathbf{X}$
T^2	Estadístico T^2 de *Hotelling*
$\bar{\mathbf{t}}_c$	Vector de medias de los *scores*
$\mathbf{t}_i$	Vector de *scores* del objeto i
$\mathbf{V}_{\Delta x}$	Matriz de covarianza de los errores de medida de la respuesta instrumental
$\mathbf{V}_{\Delta\beta}$	Matriz de covarianza de los errores en los coeficientes de regresión estimados
V_e	Varianza de los residuales
V_{PE}	Varianza del error de predicción
$V_{\Delta y}$	Varianza del error de medida en el método de referencia
Var	Varianza del vector $\mathbf{y}$
$\mathbf{W}$	Matriz de pesos de $\mathbf{X}$
$\mathbf{X}$	Matriz de datos de los objetos
$\mathbf{x}$	Vector de datos de un objeto
$\mathbf{y}$	Vector que identifica las clases de los objetos
$\hat{y}$	Valor de y predicho por el modelo
β	Coeficiente de regresión verdadero
Γ_c	Fiabilidad de multiclasificación a la clase c
$\Delta\mathbf{x}$	Error de medida de la respuesta instrumental
$\Delta\beta$	Error de los coeficientes de regresión estimados
σ_{PE}	Error estándar de predicción
$\|\|\cdot\|\|$	Norma euclidiana

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Contribuciones científicas

Artículos

1. N. F. Pérez, J. Ferré, R. Boqué.
   Calculation of the reliability of classification in discriminant partial least-squares binary classification.
   Chemometrics and Intelligent Laboratory Systems 95 (2009) 122–128.
   (Capítulo 2)

2. N. F. Pérez, J. Ferré, R. Boqué.
   Multi-class classification with probabilistic discriminant partial least squares (p-DPLS).
   Analytica Chimica Acta 664 (2010) 27–33.
   (Capítulo 3)

3. N. F. Pérez, R.Boqué, J. Ferré
   Establishment of multivariate specifications for food commodities with Discriminant Partial Least Squares.
   Enviado a Talanta.
   (Capítulo 4)

Comunicaciones en congresos

1. N. F. Pérez, R. Boqué, J. Ferré.
   Probabilistic discriminant partial least-squares regression for classifying food commodities.
   1er Workshop de la Xarxa Catalana de Quimiometría, Barcelona (España), 2005.
   Póster

2. N. F. Pérez, R. Boqué, J. Ferré
   Calculation of the Reliability of Classification for Discriminant Partial Least Squares (DPLS).
   TRACE 2nd Annual Meeting, Praga (Republica Checa), 2006.
   Póster

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

3. N. F. Pérez, R. Boqué, J. Ferré
   Calculation of the Reliability of Classification for Discriminant Partial Least Squares (DPLS).
   10th International Conference on Chemometrics in Analytical Chemistry CAC-2006, Campinas (Brasil).
   Póster

4. N. F. Pérez, R. Boqué, J. Ferré.
   Multi–class classification with Discriminant Partial Least Squares.
   VI Colloquium Chemometricum Mediterraneum, Saint Maximin (Francia), 2007.
   Póster.

5. B. Vandeginste, N. F. Pérez, R. Boqué, J. Ferré, Y. Vander Heyden, S. Caetano, A. Durand, M. Novic, N. Groselj, B. Walczak, I. Stanimirova, L. Buydens, B. Üstün
   ChemTRACE. Chemometrics Toolbox for Food Traceability
   TRACE, 4th Annual Meeting and Conference, Lost without TRACE New approaches for tracing the origin of food, Torremolinos (España), 2008.
   Póster.

Otras contribuciones

1. N. F. Pérez, J. Ferré, R. Boqué
   Módulo "Discriminant Partial Least Square (DPLS)" para el programa "ChenmTRACE". Version 1.0, Workpackage 6, 2007
   Programa en lenguaje MatLab.

2. N. F. Pérez, J. Ferré, R. Boqué
   Module 1, Discriminant Partial Least Squares. En: WP6 D6.3: Integrated Chemometrics toolbox. Part 1: Users Manual (ChemTRACE). Vrije Universiteit Brussel and Radboud Universiteit Nijmegen, December 2007.

3. N. F. Pérez, J. Ferré, R. Boqué
   Module 1, Discriminant Partial Least Squares. En: WP6 D6.3: Integrated Chemometrics toolbox. Part 2: Technical Documentation (ChemTRACE). Vrije Universiteit Brussel and Radboud Universiteit Nijmegen, December 2007.

4. N. F. Pérez, J. Ferré, R. Boqué
Module 1, Discriminant Partial Least Squares. En: WP6 D6.3: Integrated Chemometrics toolbox Part 3. Worked out examples. Vrije Universiteit Brussel and Radboud Universiteit Nijmegen, December 2007.

5. N. F. Pérez, J. Ferré, R. Boqué
3.3 Can we attribute class probabilities? En: D6.4 Results of Mineral water data analysis. Vrije Universiteit Brussel. June 2006.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Conjuntos de datos del proyecto TRACE

1. Conjunto de datos de suelos europeos

Fecha: 09/03/2005
Suministrado por: Dr. *Bernard Vandeginste*
Nº de objetos: 180
Nº de variables: 87
Nº de clases: 3

Características

Conjunto de datos de suelos recolectados por tres grupos de muestreo en nueve zonas de la Unión Europea y para tres litologías diferentes: *caliza*, *arenisca* y *esquisto* (siendo las clases a modelar), con 60 objetos cada una. Las variables se derivan del TGA para compuestos con contenido de carbono (matriz orgánica, carbonatos y carbonato de calcio); y para el resto de variables del análisis elemental con XRF e ICP- MS, indicándose los elementos mayoritarios en porcentaje y los elementos traza en ppm (tabla A-1). Los valores no reportados se muestran como 0 (datos faltantes) y las lecturas por debajo del límite de detección como -999 (datos censurados). En la figura A-1 se representan las variables en diagramas de caja.

Análisis de los datos

Se eliminaron aquellas variables con más de un 5% de datos faltantes o censurados: Mg (10), S (14), As (15), Cu (16), Pb (17), Zn (18), Ni (19), Sn (22), Sc (30), La (31), Nd (32), Th (33), U (34) y ^{11}B (38) (tabla A-2). Finalmente el conjunto quedó definido por 75 variables. Para las 75 variables se corrigieron los datos censurados con 1/4 del menor valor observado en la variable correspondiente, y los faltantes con la rutina ***mdcheck*** de *PLS_toolbox*, que los predice a partir de PCA. Los valores negativos generados por la rutina se sustituyeron por un cero.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

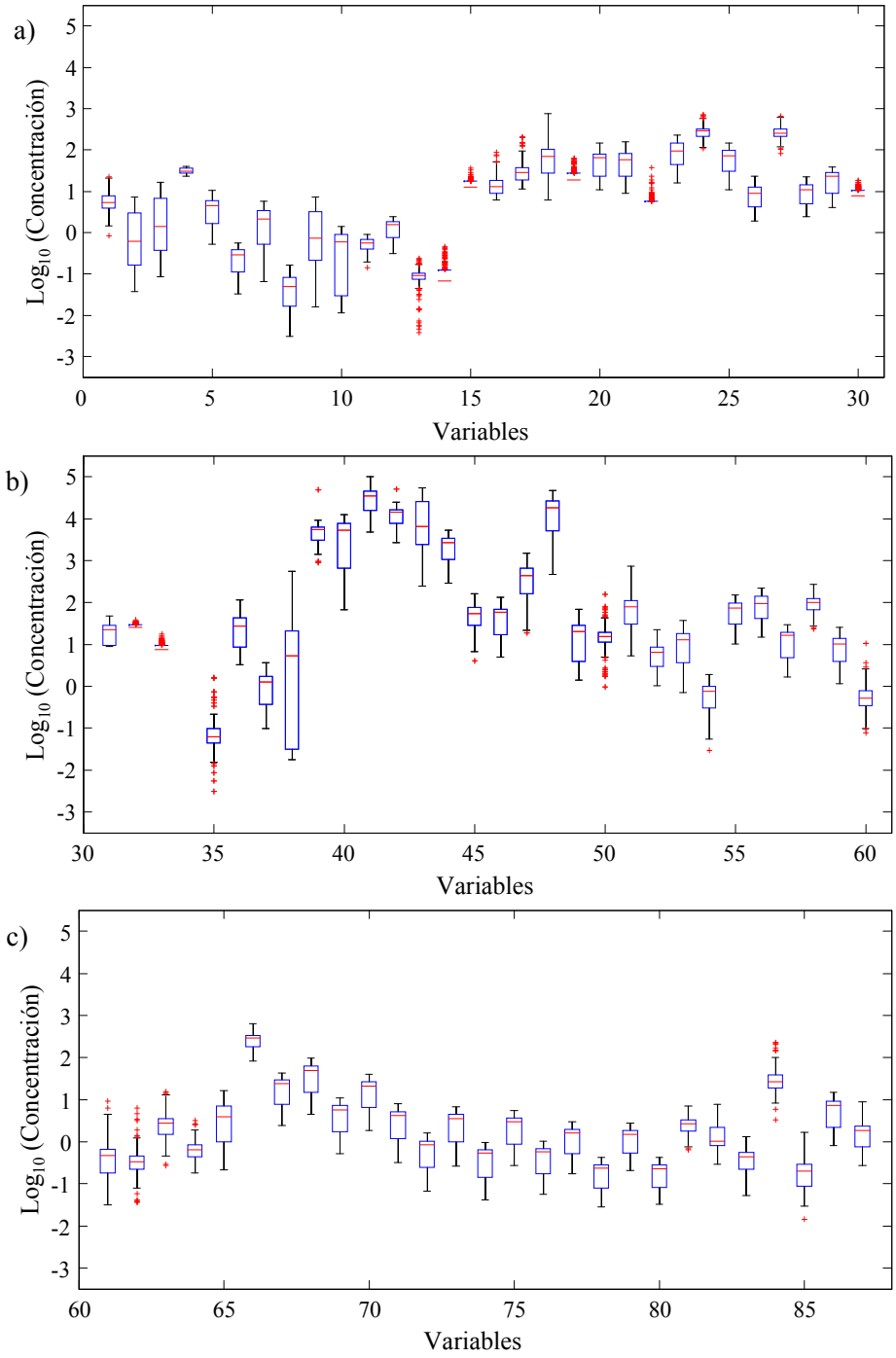

Figura A-1: Variables de suelos transformadas a su logaritmo en base 10 y representadas en diagramas de caja.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Tabla A-1: Variables del conjunto de datos de suelos.

Nº	Variable	Nº	Variable	Nº	Variable	Nº	Variable
1	Mat. Org.	23	Sr	45	^{51}V	67	^{139}La
2	Carbonatos	24	Ba	46	^{52}C	68	^{140}Ce
3	$CaCO_3$	25	Rb	47	^{55}Mn	69	^{141}Pr
4	Si	26	Ga	48	^{57}Fe	70	^{146}Nd
5	Al	27	Zr	49	^{60}Ni	71	^{147}Sm
6	Ti	28	Nb	50	^{63}Cu	72	^{151}Eu
7	Fe	29	Y	51	^{66}Zi	73	^{158}Gd
8	Mn	30	Sc	52	^{71}Ga	74	^{159}Tb
9	Ca	31	La	53	^{75}As	75	^{163}Dy
10	Mg	32	Nd	54	^{82}Se	76	^{165}Ho
11	Na	33	Th	55	^{85}Rb	77	^{166}Er
12	K	34	U	56	^{88}Sr	78	^{169}Tm
13	P	35	Hg	57	^{89}Y	79	^{172}Yb
14	S	36	^{7}Li	58	^{91}Zr	80	^{175}Lu
15	As	37	^{9}Be	59	^{93}Nb	81	^{178}Hf
16	Cu	38	^{11}B	60	^{98}Mo	82	^{181}Ta
17	Pb	39	^{23}Na	61	^{107}Ag	83	^{205}Tl
18	Zn	40	^{26}Mg	62	^{111}Cd	84	^{208}Pb
19	Ni	41	^{27}Al	63	^{120}Sn	85	^{209}Bi
20	Cr	42	^{39}K	64	^{121}Sb	86	^{232}Th
21	V	43	^{44}Ca	65	^{133}Cs	87	^{238}U
22	Sn	44	^{49}Ti	66	^{137}Ba		

Como se observa en la figura A-1 existen elevadas diferencias entre las concentraciones de algunas variables. Para equiparar el efecto de las 75 variables el conjunto de datos se autoescaló (figura A-2). El análisis por PCA permite observar las clases preestablecidas *caliza*, *arenisca* y *esquisto* (Figura A-3). El PC_1 permite separar la clase *arenisca* de las otras dos, mientras que el PC_2 permite separar las clases *caliza* y *esquisto*. Con el PC_3 no es clara la posibilidad de separar alguna clase. Además, se pueden observar 2 objetos discrepantes de la clase *esquisto*, 051 y 056; sin embargo, los objetos no fueron eliminados dado que el conjunto de datos se utilizó en DPLS y el comportamiento de los objetos cambia.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Tabla A-2: Número y porcentaje de objetos faltantes y censurados para las variables eliminadas del conjunto de datos de suelos.

Variable	Datos Faltantes (%)	Datos Censurados (%)
Mg	13 (7.2)	15 (8.3)
S	16 (8.9)	34 (18.9)
As	0	52 (28.9)
Cu	5 (2.8)	14 (7.8)
Pd	0	11 (6.1)
Zn	2 (1.1)	11 (6.1)
Ni	27 (15.0)	32 (17.8)
Sn	13 (7.2)	112 (62.2)
Sc	1 (0.6)	73 (40.6)
La	0	40 (22.2)
Nd	0	46 (25.6)
Th	7 (3.9)	60 (33.3)
U	1 (0.6)	175 (97.2)
^{11}B	44 (24.4)	0

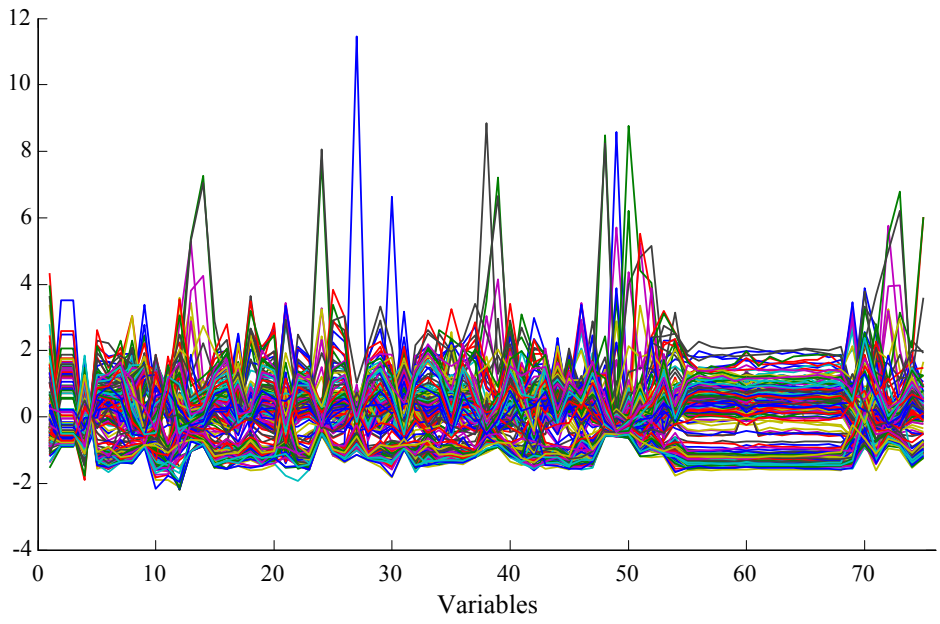

Figura A-2: Datos autoescalados de suelos europeos.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

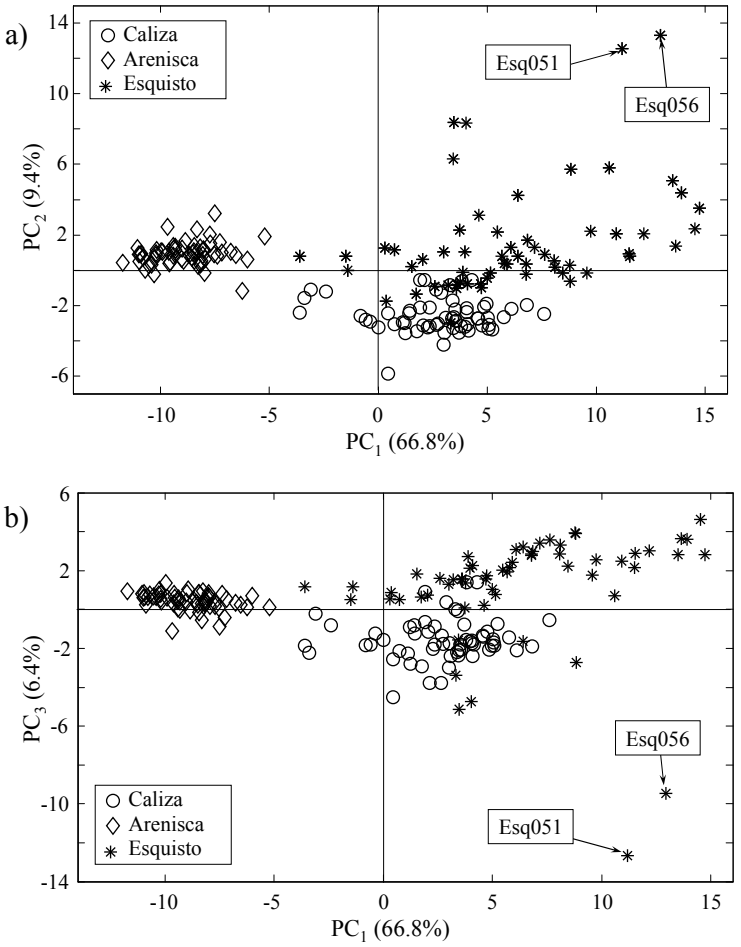

Figura A-3: Gráficas de *scores* para los tres primeros PCs del conjunto de datos de suelos europeos.

La figura A-4 muestra la distribución de los *loadings* para los tres primeros PCs. Al comparar estas gráficas con las de *scores* (Figura A-3) se establece qué variables influyen sobre cada clase. Así la clase *arenisca* tiene valores muy altos de la variable Si, indicando que el componente mayoritario puede ser *cuarcita*. La clase *caliza* se ve influenciada por carbonatos, además del *calcio* y el *sodio*; correspondiendo con su origen. Finamente la clase *esquisto* parece tener valores más elevados de materia orgánica y algunos elementos como ^{181}Ta y ^{98}Mo. Los otros elementos parecen influir por igual a las clases *caliza* y *esquisto*.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

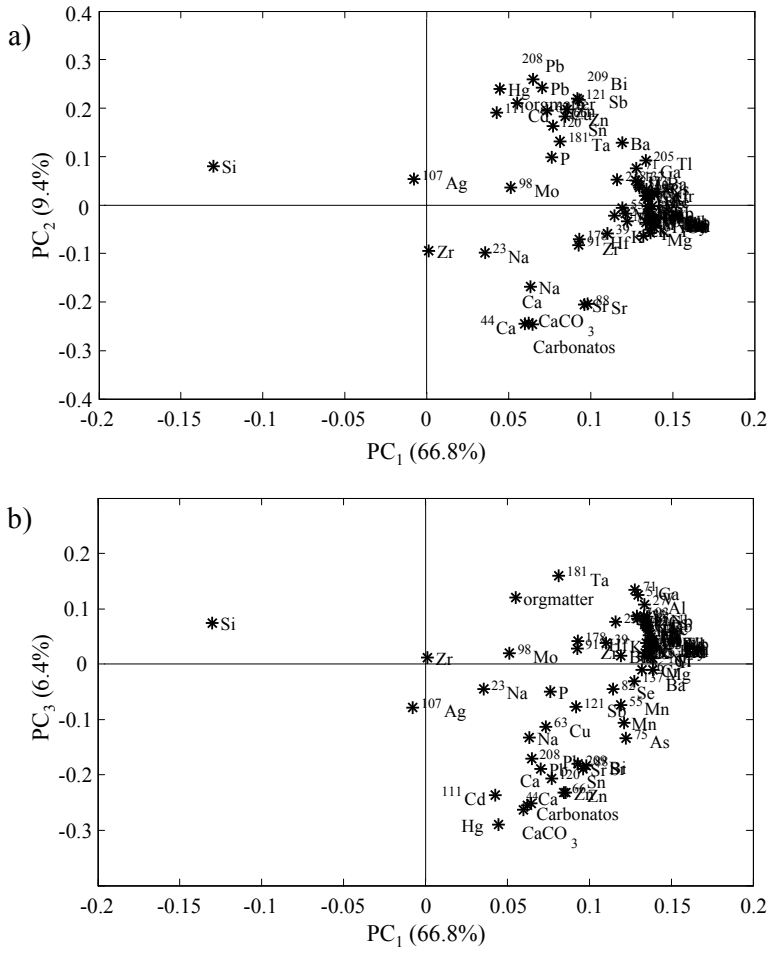

Figura A-4: Gráficas de *loadings* para los primeros tres PCs del conjunto de datos de suelos europeos.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

2. Conjunto de datos de mieles europeas

Fecha: 12/01/2007
Suministrado por: Dra. *Jana Hajslova*
Nº de objetos: 182
Nº de variables: 26
Nº de clases: 5 países o 10 regiones.

Características

Conjunto de datos de mieles europeas recolectadas en varios países de la UE, con presencia mayoritaria de mieles francesas. Se recolectaron 182 muestras en 5 países (*Francia*, *Italia*, *Austria*, *Irlanda* y *Alemania*) y 10 regiones (*Córcega*, *Sicilia*, *Toscana*, *Trentino*, *Marchfeld*, *Muehlviertel*, *Carpentras*, *Limousin*, *Bavaria* y *Galway*). Las mieles se analizaron por CG-MS hallando 26 componentes (Figura A-5). Los datos se suministraron en intensidad relativa e intensidad absoluta de los componentes. Desde TRACE se pidió priorizar la clasificación de mieles *Córcega* de las restantes regiones.

Análisis de los datos

Las figuras A-6 y A-7 muestran el número de mieles por países y regiones respectivamente. En ellas se observa que la mayor parte de las mieles corresponden a *Francia* (70.9%) y a *Córcega* (61.0%), y que *Irlanda*, representada por la región *Galway*,

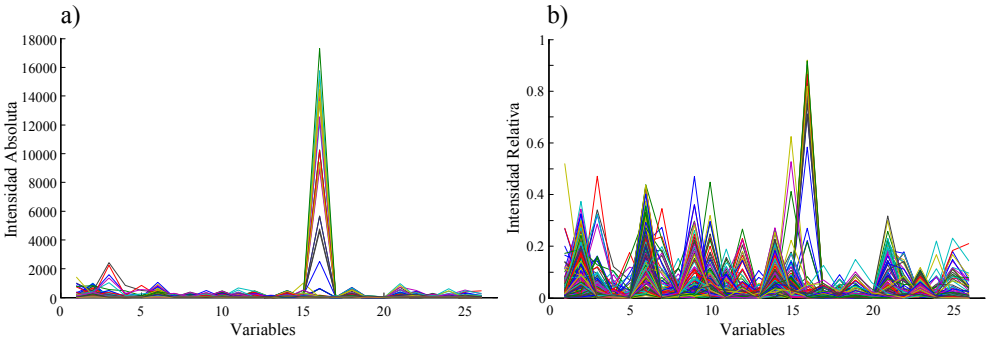

Figura A-5: Datos en intensidad absoluta (a) e intensidad relativa (b) del análisis por CG-MS de mieles europeas.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

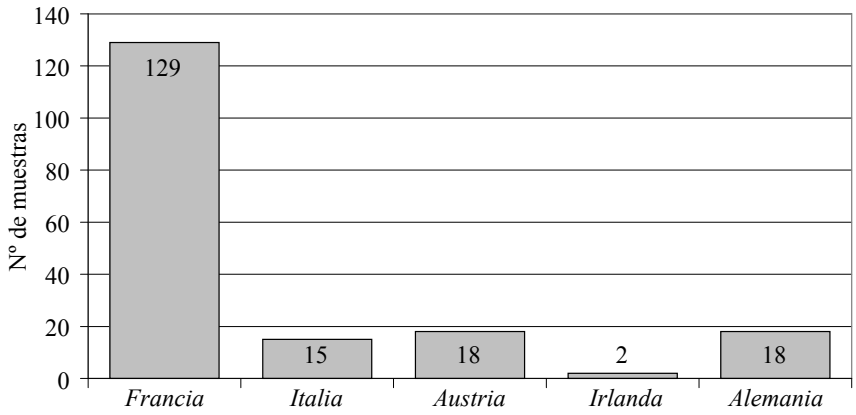

Figura A-6: Número de mieles por países.

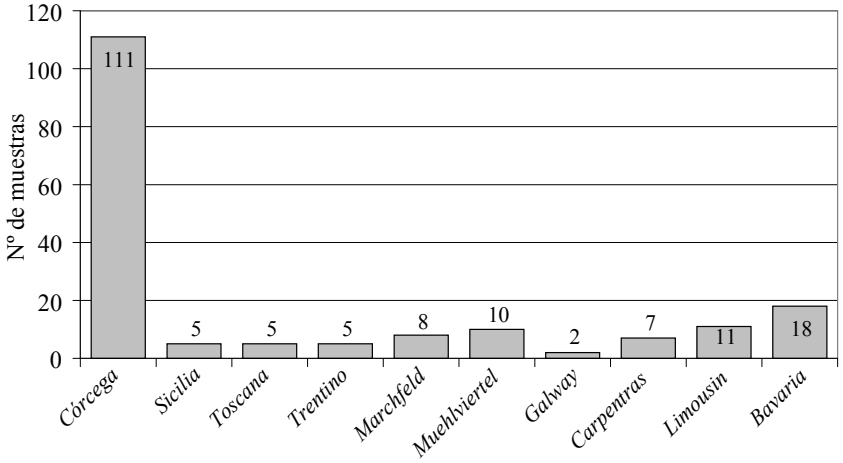

Figura A-7: Número de mieles por regiones.

tiene sólo dos mieles (1.1%). El desbalance en el número de mieles por clase puede ser desfavorable para obtener modelos que clasifiquen las clases, dado que en el mejor de los casos se cuenta con 18 mieles o menos por clase.

Dada la disparidad de las variables, incluso en intensidad relativa (Figura A-5b), los datos se autoescalaron. La figura A-8 muestra los datos autoescalados que presentan pequeñas variaciones en los valores de intensidades absolutas e intensidad relativa, por lo que se esperaría obtener resultados similares con uno u otro conjunto de datos.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Las figuras A-9 y A-10 muestras las gráficas de *scores* y *loadings*, respectivamente, para los dos primeros PCs de los datos de intensidad relativa y absoluta de mieles. Estos primeros dos PCs no tienen más del 38% y 36% de varianza acumulada y se requieren 13 PCs para acumular un 90% de la varianza. Este bajo porcentaje de varianza explicada en los PCs causa la elevada dispersión de los objetos en los gráficos de *scores* al igual que la baja discriminación entre clases. Sin embargo, la baja discriminación también se puede atribuir al desbalance en el número de mieles francesas con respeto a las otras clases. Si se observan las gráficas de *loadings* la mayor parte de los componentes están correlacionados con las mieles francesas. Sólo las variables 17, 19 y 20 parecen estar correlacionadas con mieles alemanas, que al corresponder a una única región (*Alemania-Bavaria*) son más homogéneas y por lo tanto con características mejor definidas. Sin embargo, nuevamente se ha de tener en cuenta que sólo se explica alrededor del 35% de las x con los dos primeros PCs.

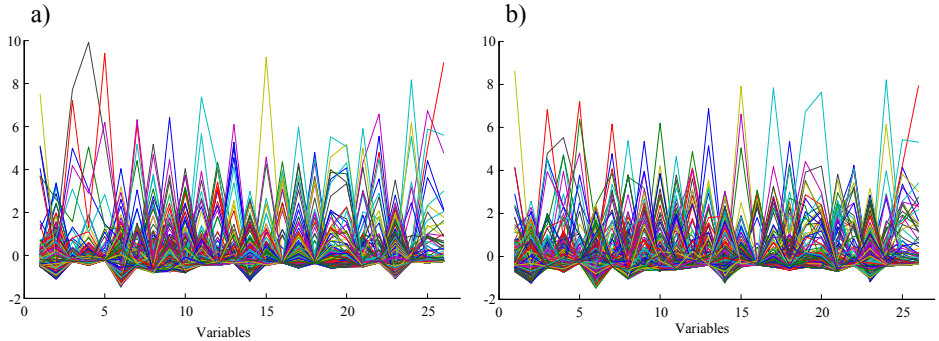

Figura A-8: Datos autoescalados de intensidad absoluta (a) e intensidad relativa (b) del análisis por CG-MS de mieles europeas.

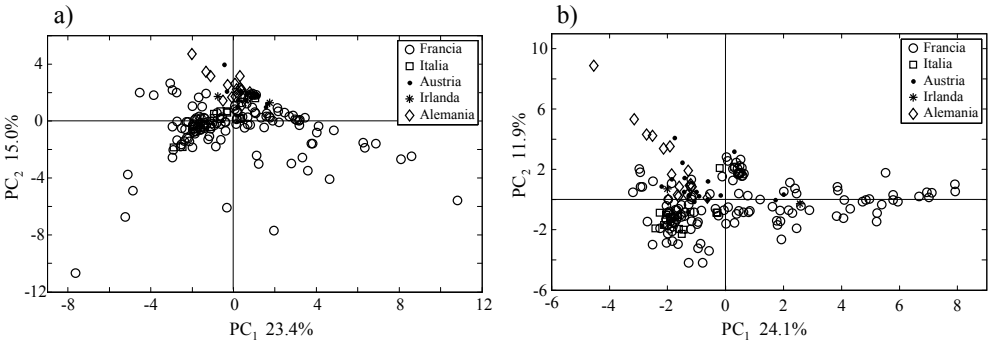

Figura A-9: Gráficas de *scores* para los dos primeros PCs de datos autoescalados de intensidad absoluta (a) e intensidad relativa (b) del conjunto de mieles europeas.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

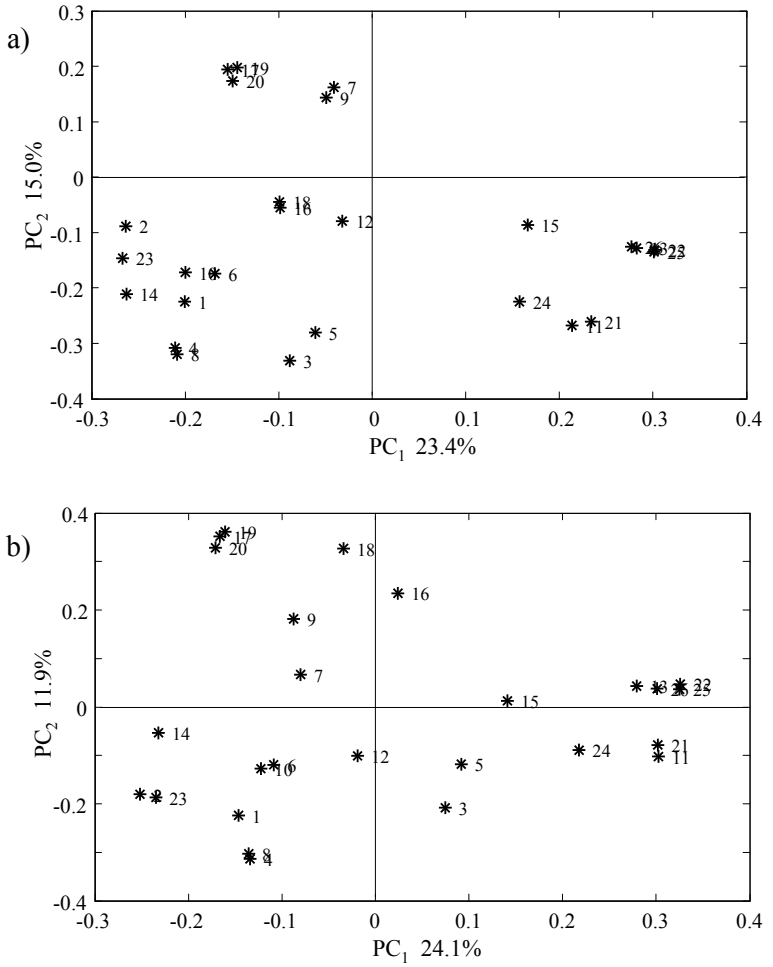

Figura A-10: Gráficas de *loadings* para los dos primeros de PCs de datos autoescalados de intensidad absoluta (a) e intensidad relativa (b) del conjunto de mieles europeas.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

3. Conjunto de datos NIR de aceites de oliva europeos

Fecha: 09/03/2006
Suministrado por: Dr. *Gerard Downey*
Nº de objetos: 316
Nº de variables: 1050
Nº de clases: Existe la posibilidad de clasificar por países y región.

Características

Conjunto de datos de aceites de oliva analizados por espectroscopia NIR de 400 a 2498 nm con una ventana espectral de 2 nm, generando 1050 variables (Figura A-11). Los aceites fueron recolectadas en 5 países europeos: *España*, *Francia*, *Grecia*, *Italia* y *Turquía*, y 22 regiones: *Andalucía*, *Extremadura*, *Cataluña*, *Aragón*, *Castilla La Mancha*, *Comunidad Valenciana*, *Liguria*, *Lombardia*, *Sicilia*, *Calabria*, *Puglia*, *Lazio*, *Trentino Alto Adige*, *Abruzzo*, *Campania*, *Veneto*, *Izmir*, *Peloponneso*, *Creta*, *Islas Baleares*, *Umbría Provence-Alpes-Côte d'Azur* y *Molise*. Además, se suministran los códigos de aceite y las DOP.

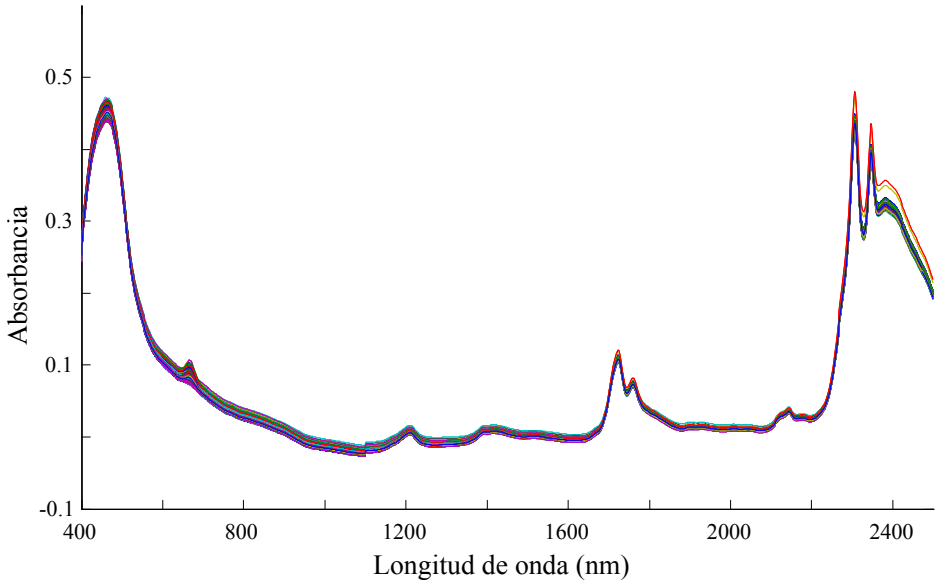

Figura A-11: Espectros NIR de los aceites de oliva.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Análisis de los datos

Las figuras A-12 y A-13 muestran el número de aceites de oliva por países y regiones. La mayoría de éstos son italianos (71.5%), seguidos de los españoles (13.3%), griegos (7.9%) y un 7.3% que corresponden a *Francia* y *Turquía*. En cuanto al número de aceites por regiones la mayoría (19.9%) pertenecen a *Liguria*. Algunas regiones con más del 6% de aceites son las italianas además de *Andalucía* y *Creta*. Debido al desbalance en el número de aceites de oliva por país y región, TRACE recomendó centrarse en discriminar la región *Liguria* de todas las restantes y de las regiones italianas.

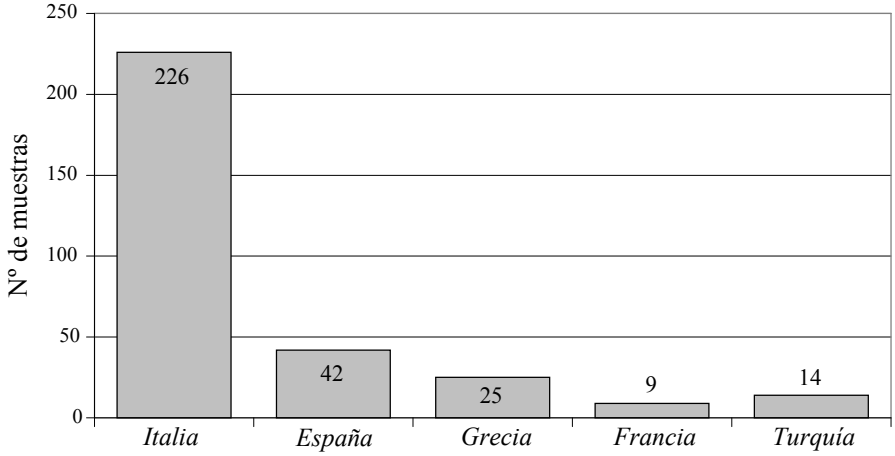

Figura A-12: Número de aceites de oliva por países.

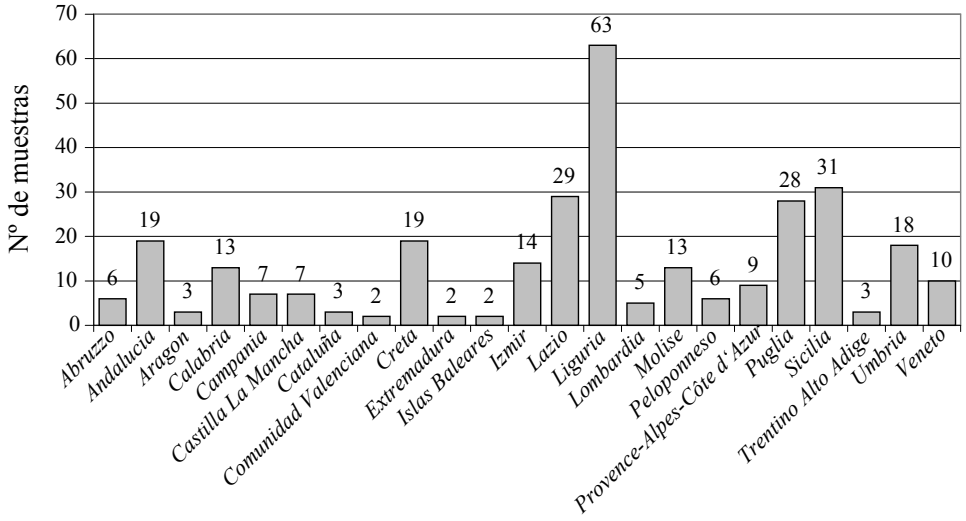

Figura A-13: Número de aceites de oliva por regiones.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Para reducir el número de variables y ajustar los espectros a la región de medida en espectroscopia NIR, se eliminaron 350 variables, de 400 nm a 1098 nm. Así, los análisis posteriores se realizaran con los datos entre 1100 nm a 2498 nm (700 variables). Los datos se centraron a la media (Figura A-14). Aunque para la mayoría de aceites de oliva las variaciones son mínimas, dos aceites, 5050062 y 5050115, presentan valores discrepantes. Estos mismos aceites ya eran diferenciables de los restantes por sus espectros (Figura A-11).

La figura A-15 muestra la gráfica de *scores* para los dos primeros PCs. Lo más destacable es que explican más del 96% de la varianza, aunque no se logren diferenciar las clases. La falta de diferenciación de las clases podemos atribuirla al desbalance en el número de aceites de oliva por clase, dado que cerca del 70% de éstos pertenecen a *Italia*. Además, la dispersión de las muestras aceites y la presencia de aceites de oliva discrepantes dificultan hallar una relación entre las variables y la distribución de los aceites.

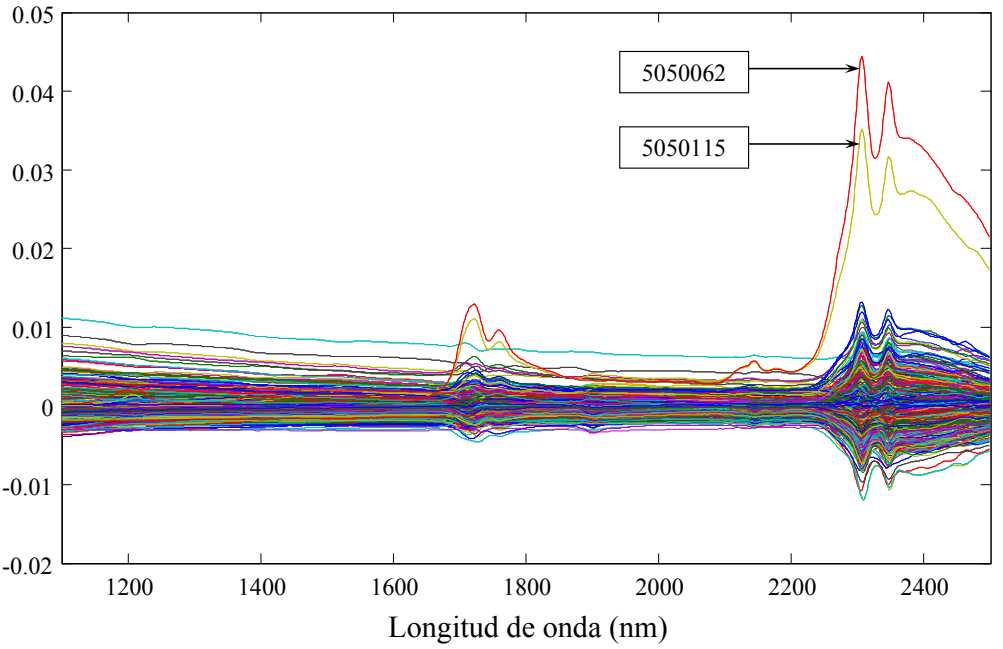

Figura A-14: Espectros de aceites de oliva centrados a la media, señalando dos posibles aceites discrepantes.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

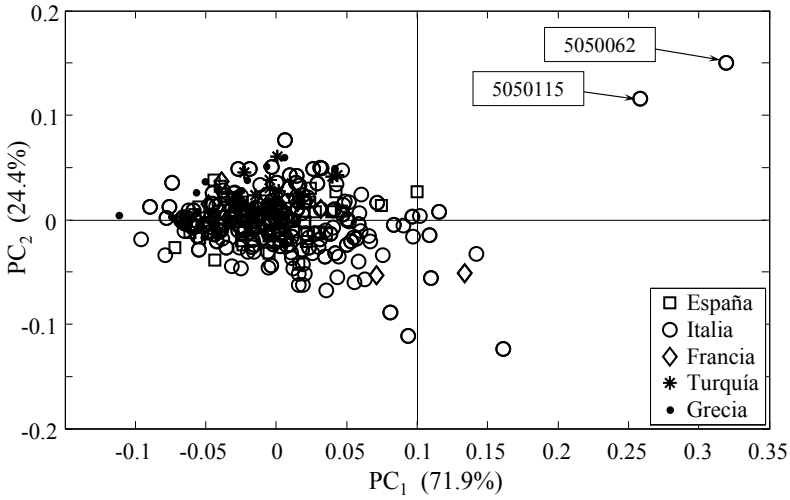

Figura A-15: Gráfica de *scores* para los dos primeros PCs del conjunto de datos de aceites de oliva.

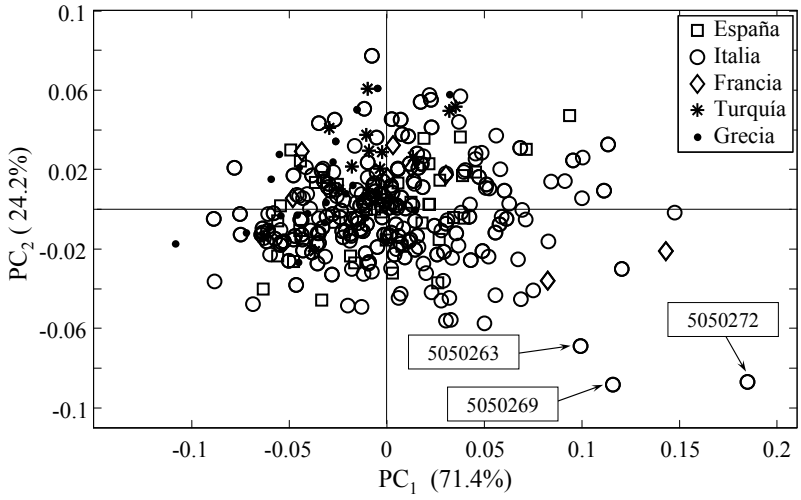

Figura A-16: Gráfica de *scores* para los dos primeros PCs del conjunto de datos de aceites de oliva eliminando las aceites 5050062 y 5050115.

La figura A-16 muestra las gráfica de *scores* después de eliminar los aceites discrepantes 5050062 y 5050115. En los *scores* persiste la mezcla de muestras de aceites sin poder discriminar entre clases y aparecen otros tres posibles aceites discrepantes, pero dada la dispersión de los aceites no se aconseja eliminarlos.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

4. Conjunto de datos ^{1}H-RMN de aceites de oliva europeos

Fecha: 12/09/2007
Suministrado por: Dr. *José Manuel Moreno Rojas*
Nº de objetos: 316 (2005) y 352 (2006)
Nº de variables: 342
Nº de clases: Existe la posibilidad de clasificar por países, región, denominación de origen protegida y año.

Características

Conjunto de datos de aceites de oliva europeos recolectados durante los años *2005* (son los mismos que en el conjunto de datos NIR) y *2006* siendo analizados por ^{1}H-RMN obteniendo 342 variables (Figura A-17). Además se suministró el código de aceite, el país, la región, la provincia, la comunidad, la DOP y el año. Los aceites de oliva de *2005* se pueden clasificar en 5 países, 23 regiones y 39 denominaciones de origen. Los aceites de oliva de *2006* se pueden clasificar en 5 países, 24 regiones y 29 denominaciones de origen. Al reunir los dos años se puede clasificar 6 países, 32 regiones y 48 denominaciones de origen protegidas. En la figura A-17 se muestran los espectros de ^{1}H-RMN para los aceites de oliva de los años *2005* y *2006*. No se observan diferencias apreciables ni para los aceites de un mismo año ni para los aceites entre años, a excepción del desplazamiento cercano a 4 ppm donde en el *2006* hay diferencias entre aceites que en el *2005* son mínimas.

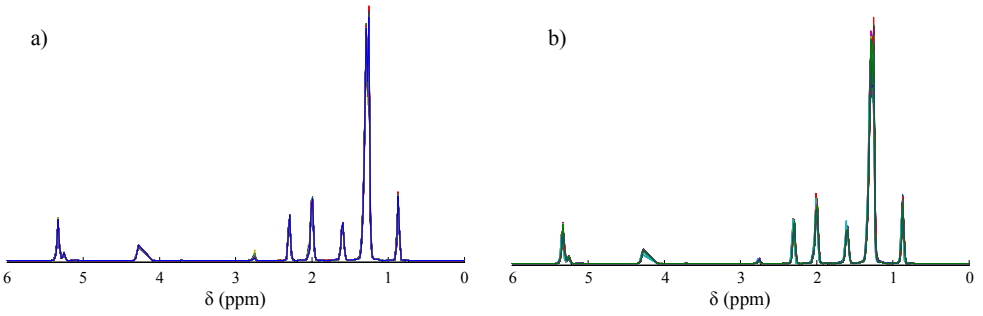

Figura A-17: Espectros de ^{1}H-RMN de los aceites de oliva para los años *2005* (a) y *2006* (b).

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Análisis de los datos

Las figuras A-18 y A-19 muestran el número de aceites de oliva por países y regiones, respectivamente, para el año 2006 (para el año 2005 consultar el conjunto de datos NIR Fig. A-12 y A-13). Los aceites son mayoritariamente italianos (71.6%) seguidos de *Grecia* (13.1%) y *España* (10.8%). El 4.5% restante lo representan aceites de *Francia* y *Chipre*.

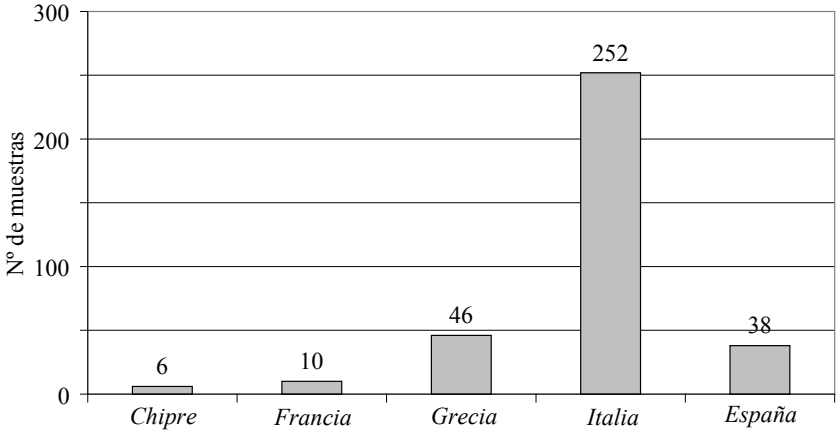

Figura A-18: Número de aceites de oliva por países de *2006*.

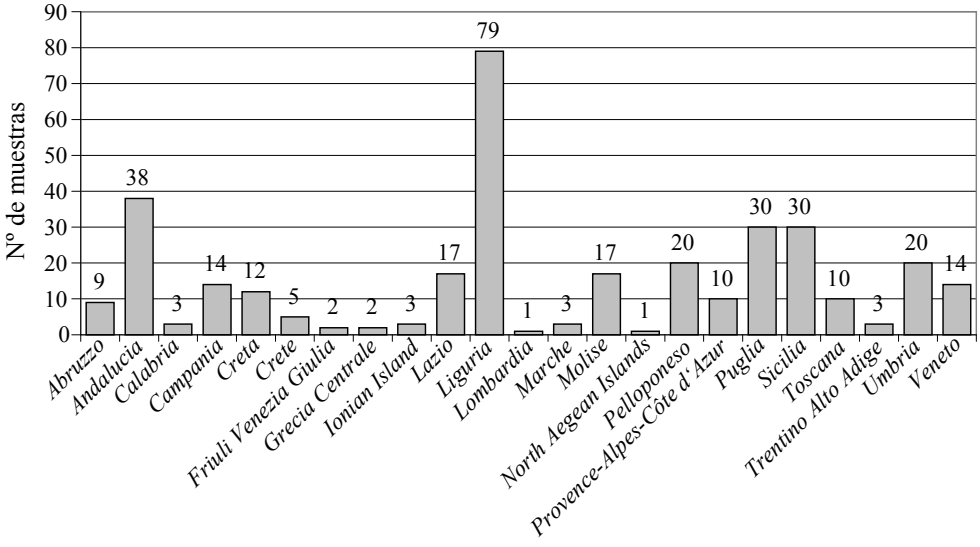

Figura A-19: Número de aceites de oliva por regiones de *2006*.

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

Para los aceites por regiones, la mayoría pertenecen a *Liguria* (22.4%). Otras regiones con más de un 6% de los aceites son italianas, además de *Andalucía* con el 10.8% de los aceites. Debido al desbalance en el número de aceites por clase, TRACE aconsejó centrarse en discriminar la región *Liguria* de las restantes y de las regiones italianas, y a su vez, verificar la posibilidad de discriminar los aceites de oliva por años. Sin embargo, algunas de las regiones sólo tienen uno o dos objetos por lo que no permitiría dividirlas en datos de entrenamiento y prueba.

Los datos ^{1}H-RMN se centraron a la media para realizar el análisis por PCA, para cada año y en conjunto. La figura A-20 muestra los datos centrados para los años *2005* y *2006*. A diferencia de los datos en bruto, en donde no se observaba ninguna diferencia entre objetos o entre años (Figura A-17), con los datos centrados de *2005* se observa un objeto que se diferencia en el desplazamiento de 2 a 3 ppm, y en el año *2006* una serie de objetos destacan sobre los restantes, llegando a duplicar el valor con respecto a los objetos restantes. Entre años, se observa que los valores del *2006* duplican a los del *2005*, los valores del *2005* están entre -1000 a 1000 mientras que los de *2006* entre -2500 y 1500. La gráfica de *scores* de *2005* (Figura A-21a) muestra que el objeto 5050161 (no se consideró en el conjunto de datos NIR) de *España* puede ser considerado discrepante, y corresponde con el observado en los datos centrados (Figura A-20a). Algo similar se observa en los *scores* de *2006* (Figura A-21c) donde se tienen 4 o 6 posibles objetos discrepantes, un objeto griego (6060071), un objeto francés (6060282) y cuatro objetos italianos (6030041, 6060148, 6060177 y 6060376). Estos objetos se corresponden con los que tienen el mayor valor en datos centrados (Figura A-20b). Para los *scores* combinando los años (Figura A-21e) nuevamente aparecen los datos considerados como discrepantes en el *2006*, mas no el del *2005*.

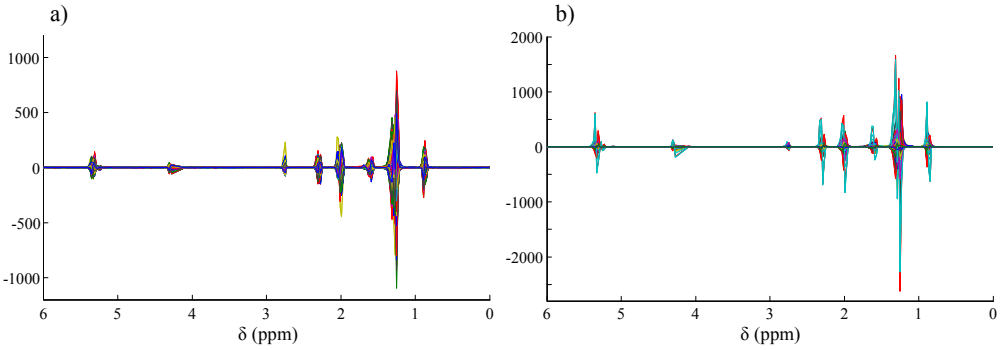

Figura A-20: Espectros ^{1}H-RMN centrados a la media de los aceites de oliva para los años *2005* (a) y *2006* (b).

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

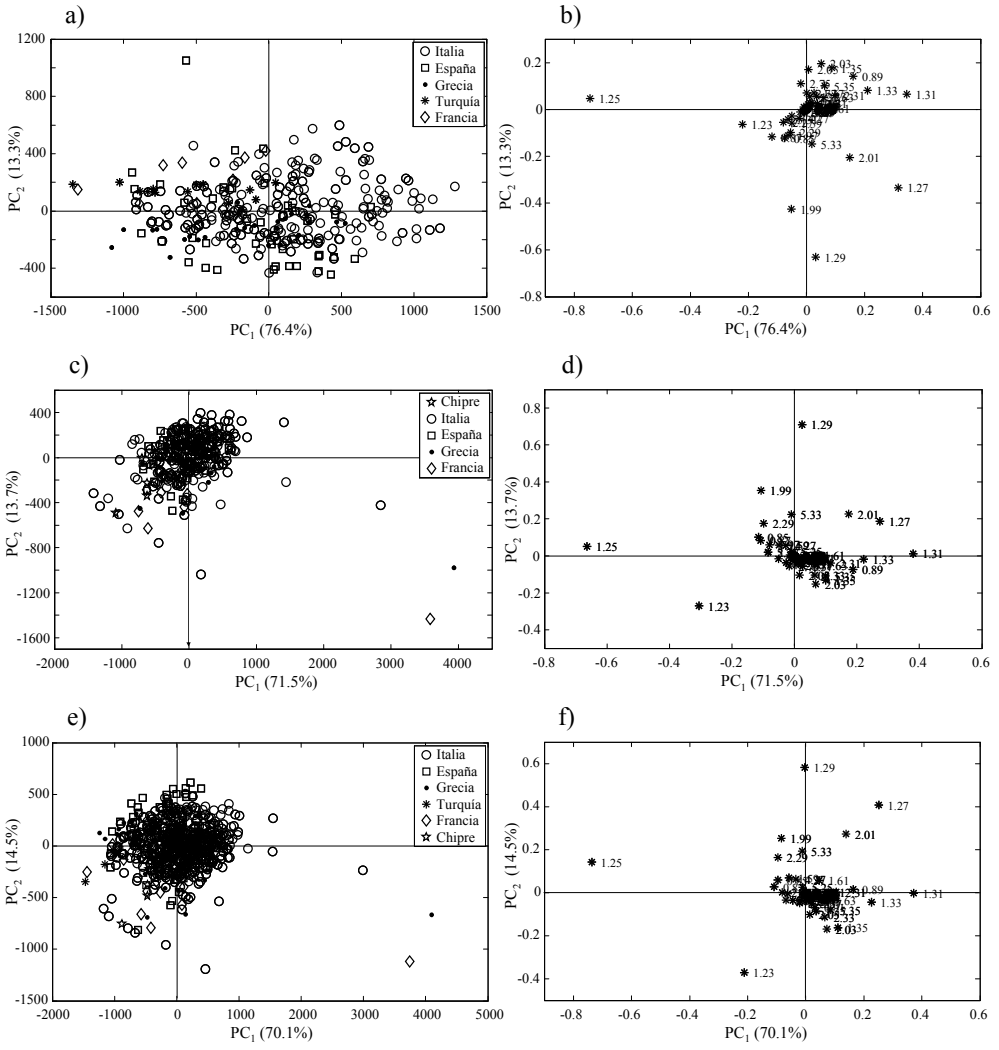

Figura A-21: Gráficas de *scores* y *loadings* para los dos primeros PCs de los espectros ^{1}H-RMN para los años *2005*(a y b), *2006* (c y d) y *2005* más *2006* (e y f).

En cuanto a la presencia de agrupamientos, e indistintamente de la presencia de los datos discrepantes, en ninguno de los tres casos se observan grupos definidos, ya que la mayoría de objetos distintos de *Italia* se dispersan dentro de los datos italianos. Aun así, en los tres casos los 2 primeros PCs tienen más del 84% de varianza explicada, necesitándose 4 PCs para explicar más del 96%. La relación entre los desplazamientos (variables) con la distribución de los objetos no es fácil de observar. Las tres gráficas de

UNIVERSITAT ROVIRA I VIRGILI
FIABILIDAD DE CLASIFICACIÓN CON PLS DISCRIMINANTE
Néstor Fredy Pérez Pérez
ISBN:978-84-693-4053-0/DL:T.990-2010

loadings presentan una configuración similar (Figura A-21b, d y f) con el desplazamiento de 1.25 ppm al extremo izquierdo y sólo los desplazamiento de 1.23, 1.29, 1.99, 1.27, 1.31 y 2.01 que pueden diferenciarse del resto. Aunque para los tres casos se puede decir que existen datos discrepantes no se eliminan hasta que sean verificados por modelos DPLS, que serán los que finalmente se usen con estos datos.