PROBLEMAS DE CLASIFICACIÓN Y OPTIMIZACIÓN

Por el Dr. EMILIO CARRIZOSA PRIEGO,
Profesor Titular de Estadística e Investigación Operativa
y Premio del año 1998 de la Real Maestranza
de Caballería para Investigadores Jóvenes.
Conferencia pronunciada el día 12 de marzo de 2003

ABSTRACT

El desarrollo de técncias que permitan clasificar entes (seres vivos, elementos, problemas,...) en distintas categorías ha sido recurrente en diversas ramas del saber.

La creación y difusión de grandes bases de datos y la consiguiente necesidad de extraer conocimiento de las mismas han revitalizado el interés de la comunidad científica por tales técnicas.

En estas páginas se ilustra cómo la Programación Matemática puede contribuir al diseño de métodos automáticos de clasificación y profundizar en el conocimiento teórico de los mismos. Nuestra intención no es hacer una revisión completa del estado del arte en el tema, sino más bien describir someramente las aportaciones que en este campo se están realizando en el seno del grupo PAI FQM-809 y en el proyecto de investigación BFM2002-04525-C02-02 del MCYT.

Palabras clave: Programación Matemática. Análisis Discriminante.

1. INTRODUCCIÓN

Atendiendo al Diccionario de la Real Academia Española de la Lengua, *clasificación* es la acción de disponer u ordenar entes en clases.

No aclara el Diccionario qué propiedades satisfacen las *clases* en las que deben disponerse los entes en cuestión. Nosotros supondremos que estas clases constituyen una partición del conjunto y nos restringimos por consiguiente a la clasificación *nítida*, por contraposición a los métodos basados en conjuntos borrosos.

Tampoco aclara el Diccionario si las clases deben crearse (artificialmente) en el proceso mismo de clasificación o por el contrario existen previamente. De hecho ambas posibilidades han sido ampliamente estudiadas en la literatura, y proporcionan una clasificación de los problemas de clasificación.

En el primer caso hablaremos de *clasificación no supervisada*, e.g. [13, 16, 17]: debe determinarse una partición de modo que los elementos que queden en la misma clase de la partición sean homogéneos, mientras que los elementos de clases distintas tengan características dispares, donde los conceptos de homogeneidad y disparidad deberán ser definidos formalmente.

En el caso en el que las clases de la partición estén definidas previamente hablaremos de problemas de *clasificación supervisada*, e.g. [12, 15, 21], y el objetivo perseguido es construir una regla que asigne cada objeto a la clase a la que pertenece.

Problemas de los dos tipos aparecen, juntos o separados, en los más variados campos. Como ilustración, en el reciente texto [18] se describen aplicaciones en áreas tales como Marketing, Telecomunicaciones, Banca, Medicina, Farmacología, Genética, Informática, Educación, etc.

2. *PLANTEAMIENTO GENERAL*

Se considera un conjunto $\mathcal{O}^*$ de objetos, (la *población*) y dos funciones sobre $\mathcal{O}^*$,

$$G = \mathcal{O}^* \to \mathcal{G}$$
$$X = \mathcal{O}^* \to \mathcal{X}$$

$\mathcal{G}$ es un conjunto finito, que llamaremos de *etiquetas*, e induce una partición de la población $\mathcal{O}^*$ es un conjunto finito de clases. Diremos que $G(o) \in \mathcal{G}$ es la clase de la partición a la cual el objeto $o \in \mathcal{O}^*$ pertenece.

Por otro lado, $X(o) \in \mathcal{X}$ representa, en un sentido que precisaremos después, una serie de características, cuantitativas o cualitativas, del objeto o. Llamaremos por eso $\mathcal{X}$ espacio de características, y tendremos en cuenta asimismo que $\mathcal{X}$ suele venir expresado como un subconjunto de un producto finito de espacios,

$$\mathcal{X} \subset \prod_i \mathcal{X}_i, \tag{1}$$

donde cada $\mathcal{X}_i$ representa el conjunto de valores que puede tomar la i-ésima propiedad o característica objeto de estudio. Como ilustración podemos mencionar los casos clásicos, $\mathcal{X}_i = \mathbf{R}$ si la i-ésima variable es cuantitativa o $\mathcal{X}_i = \{u^i_1,...,u^i_{n_i}\}$ para variable cualitativas, a los que podemos añadir otros más recientes y motivados por las aplicaciones, como $\mathcal{X}_i = \mathbf{R} \cup \{?\}$, que modela la posibilidad de que el valor de la i-ésima propiedad sea desconocido, o $\mathcal{X}_i$ es un aespacio de series temporales, o de distribuciones de frecuencias sobre un conjunto K_i de entes (para modelar e.g. la frecuencia relativa de apariciones de cada una de las palabras de K_i en una página web).

Una función $\pi : \mathcal{X} \to \mathcal{G}$ se dice *regla de clasificación* sobre $\mathcal{O}^*$.

Se tiene un subconjunto *finito* no vacío de objetos $\mathcal{O} \subset \mathcal{O}^*$, que llamaremos *muestra de aprendizaje*. Conociendo las funciones G, X *restringidas a* $\mathcal{O}$, se desea construir una regla de clasificación $\pi : \mathcal{X} \rightarrow \mathcal{G}$ con un doble objetivo (descriptivo y predictivo):

1. Clasificar correctamente todos los objetos de la meutra de aprendizaje, i.e., resolver la ecuación funcional

$$\pi(X(o)) = G(o) \qquad \forall o \in \mathcal{O} \tag{2}$$

2. Clasificar correctamente todos los objetos de la población, i.e., resolver la ecuación funcional

$$\pi(X(o)) = G(o) \qquad \forall o \in \mathcal{O}^* \tag{3}$$

No se asumen propiedades de inyección sobre la función X, por lo que la ecuación (2) puede no tener solución. Como solo se dispone de las imágenes de las funciones G y X sobre los objetos de la muestra de aprendizaje, en ningún caso habrá garantías de poder clasificar correctamente todos los objetos de la población.

En estas circunstancias, habrá que entender la resolución de (2) y (3) como la búsqueda de una regla de clasificación π que las viole *mínimamente.*

Para dar sentido preciso a lo anterior, introducimos una función de costo $C : \mathcal{G}^{\mathcal{O}^*} \times \mathcal{O}^* \rightarrow \mathbf{R}_+$, donde $C(\pi,o)$ representa el *costo* incurrido por clasificar en la clase $\pi(X(o))$ el objeto o.

Englobando varios modelos de función de costo existentes en la literatura, expresamos C a través de *disimilaridades* a las regiones clasificadas mediante π con la misma etiqueta: Para cada $g \in \mathcal{G}$ sea dg una disimilaridad sobre $\mathcal{X}$, i.e., $d_g : \mathcal{X} \times \mathcal{X} \rightarrow \mathbf{R}$ es una función no negativa con $d_g(x,x) = 0 \forall x \in \mathcal{X}$. No imponemos ninguna restricción sobre la disimilaridad d_g, que puede ser e.g. una métrica binaria,

$$d_g(x,z) = \begin{cases} \omega_g, \text{ si } x \neq z \\ 0, \quad \text{si } x \neq z \end{cases} \tag{4}$$

para $\omega_g > 0$, una métrica inducida por un calibrador, como se verá en la sección 3, o, en el caso de espacio de características definidos a través de (1), construido combinando disimilaridades en los $\mathcal{X}_t$, e.g. [9, 17].

Defínase la función de costo C como:

$$\begin{aligned} C(\pi,o) &= d_{G(o)}(X(o), \{x \in \mathcal{X} : \pi(x) = G(o)\}) \\ &= \inf \{d_{G(o)}(X(o),z) : \pi(z) = G(o)\} \end{aligned} \tag{5}$$

Por ejemplo, para la métrica binaria (4), la función de costo C adquiere la forma:

$$C(\pi,o) = \begin{cases} \omega_{G(o)}, & \text{si } \pi(X(o)) \neq G(o) \\ 0, & \text{si } \pi(X(o)) = G(o) \end{cases} \tag{6}$$

Cada regla de clasificación π lleva asociado el vector de costos $(C(\pi,o))_{o\in\mathcal{O}^*} \in \mathbf{R}_+^{\mathcal{O}}$, cuya o-ésima coordenada da el costo incurrido al clasificar o mediante π.

La búsqueda de reglas de clasificación con costos de clasificación pequeños *sobre* $\mathcal{O}$ puede formalizarse como el problema de optimización multiobjetivo.

$$\min_{\pi\in\Pi}(C(\pi,o))_{o\in\mathcal{O}} \tag{7}$$

donde la minimización hay que entenderla en $\mathbf{R}^{\mathcal{O}}$ con respecto al orden parcial natural, y $\Pi \subset \mathcal{G}^{\mathcal{O}}$ es el conjunto de reglas de clasificación permitidas.

En la práctica sustituiremos el problema multiobjetivo (7) por alguna escalarización del mismo, resolviendo en lugar de (7) el problema escalar

$$\min_{\pi\in\Pi} \varphi((C(\pi,o))_{o\in\mathcal{O}}) \tag{8}$$

para una cierta $\varphi \in \mathbf{R}_+^{\mathcal{O}}$.

En particular, con una función de agregación lineal se obtiene

$$\min_{\pi\in\Pi} \sum_{o\in\mathcal{O}} \omega_o C(\pi,o), \tag{9}$$

para unos ciertos ω_o positivos con suma 1, que convierten (9) en el problema de minimizar el costo en un elemento seleccionado aleatoriamente (con probabilidades ω_o) de $\mathcal{O}$.

Obviamente, (9) no es la única escalarización posible o razonable de (7). Como ejemplo, también podemos seguir el enfoque minimax,

$$\min_{\pi\in\Pi} \max_{o\in\mathcal{O}} C(\pi,o), \tag{10}$$

Obsérvese que en la expresión anterior, al igual que en todas las que siguen, los errores se consideran solo sobre la muestra de aprendizaje. El comportamiento sobre $\mathcal{O}^* \setminus \mathcal{O}$ de la regla de clasificación así obtenida es, en principio, imprevisible: los datos disponibles no son, en muchas ocasiones, experimentales, sino observacionales.

Pueden seguirse dos estrategias de aproximación al problema sobre $\mathcal{O}^*$: el enfoque minimax, en el que se explora el peor caso, y se derivan cotas para el error de la "peor" extensión, e.g. [12]; como alternativa, podemos seguir un enfoque muestral, suponiendo que $\mathcal{O}$ es una muestra aletoria de la población $\mathcal{O}^*$, donde el mecanismo de generación aleatoria es conocido, e.g. es un muestreo aleatorio simple, o un muestreo estratificado con probabilidades de generación en cada estrato (grupo) conocidas o estimadas.

3. CLASIFICADORES LINEALES

Consideraremos en esta sección el caso más simple en el que el espacio $\mathcal{X}$ en el que se representan las características de los objetos medidas por X es $\mathbf{R}^p$, que sólo hay dos etiquetas, $\mathcal{G} = \{g_1, g_2\}$, y que el epsacio Π de reglas de clasificación permitidas se reduce al conjunto de reglas definidas por un hiperplano: Cualquier hiperplano en $\mathbf{R}^p$ divide a éste en dos regiones (un semiespacio cerrado y su complementario) que podemos identificar con las regiones en las que la regla de clasificación toma respectivamente los valores g_1 y g_2.

Las hipótesis anteriores son, evidentemente, restrictivas. No obstante, la construcción de clasificadores lineales para dos grupos puede usarse como subrutina para la construcción de clasificadores no lineales para un número arbitrario $\mathcal{G}$ de grupos. En efecto, la separación de varios grupos puede realizarse mediante un proceso secuencial en el que en cada iteración se separan dos grupos, constituidos por varios grupos. El lector puede encontrar e.g. en [4] más detalles y referencias sobre la implantación de estos métodos.

Para la construcción de clasificadores no lineales, una estrategia muy exitosa en las aplicaciones prácticas se basa en realizar una inmersión (no lineal) φ de $\mathbf{R}_p$ en un espacio $\mathbf{R}_q$ de mayor dimensión. Se consideran entonces las reglas lineales $\bar{\pi} \in \mathcal{G}^{\mathbf{R}^q}$, o, que dan lugar a reglas no lineales $\pi \in \mathcal{G}^{\mathbf{R}^q}$ canónicamente:

$$\pi(x) = \bar{\pi}\,(\varphi(x)), \tag{11}$$

como se describe e.g. en [10].

Definamos $\mathcal{H} \stackrel{\text{def}}{=} (\mathbf{R}^d \setminus \{0\}) \times \mathbf{R}$. Para cada par $\sigma \stackrel{\text{def}}{=} (\mu,\beta) \in \mathcal{H}$, consideremos los semiespacios e hiperplanos

$$H^{\#}(\sigma) = H\#(\mu,\beta) \stackrel{\text{def}}{=} \{x \in \mathbf{R}^d : \langle u;x\rangle \,\#\, \beta\}$$

con $\# \in \{\leq, <, =, \geq, >\}$.

Para cada σ se obtiene la regla de clasificación π_σ,

$$\pi_\sigma(x) = \begin{cases} g_1, & \text{si } x \in H^{\leq}(\sigma) \\ g_2, & \text{si } x \in H^{>}(\sigma) \end{cases} \tag{12}$$

Siguiendo a [19], en la definición de la función de costos C de (4), imponemos que las disimilaridades d_g vienen inducidas por calibraciones en $\mathbf{R}^d$, e.g. [22]. En otras palabras, se tiene un conjunto compacto conveso $B \subset \mathbf{R}^p$ conteniendo al origen en su interior de modo que

$$d(x,y) = \gamma(y-x) \forall x,y \tag{13}$$

donde γ es el calibrador con bola unidad B,

$$\gamma(x) = \min \{t \geq 0 : x \in tB\} \tag{14}$$

Se tiene entonces, e.g. [24],

$$\begin{aligned} d(x,H^{\leq}(u,\beta)) &= \frac{(\langle u;x\rangle - \beta)^+}{\gamma^o(-u)} \\ d(x,H^{>}(u,\beta)) &= \frac{(\beta - \langle u;x\rangle)^+}{\gamma^o(-u)} \end{aligned} \tag{15}$$

donde γ^o es el calibrador dual de γ, $y(s)+ = \max\{s,0\}$.

De esta forma, el problema multiobjetivo (7) puede reformularse como

$$\min_{u\neq 0,\beta} \left(\left(\frac{(\langle X(o);u\rangle - \beta)^+}{\gamma^o(-u)} \right)_{o\in O_1} , \left(\frac{(\beta - \langle X(o);u\rangle)^+}{\gamma^o(-u)} \right)_{o\in O_2} \right), \tag{16}$$

donde $O_i = \{o \in \mathcal{O} : G(o) = g_i\}$, $i = 1,2$.

Este problema guarda una cierta analogía formal con el problema multiobjetivo abordado en [5] en el que se minimizaban simultaneamente las distancias *verticales* en un problema de regresión. Sin embargo, mientras que en el caso de la regresión vertical el problema multiobjetivo era poliédrico (lineal a trozos y convexo), en este caso, el objetivo asociado a cada $o \in \mathcal{O}$ es el cociente de dos funciones convexas homogéneas, aparentemente sin propiedad de convexidad alguna. Una descripción de las soluciones eficientes de (16), como la obtenida en [5], sigue siendo un problema abierto.

El grado de conocimiento sobre el problema es mayor en las escalarizaciones de (7). En efecto, el problema (9) se convierte en particualr en

$$\min_{u\neq 0,\beta} \sum_{o\in O_1} \left(\omega_o \frac{(\langle X(o);u\rangle - \beta)^+}{\gamma^o(-u)} \right) + \sum_{o\in O_2} \left(\omega_o \frac{(\beta - \langle X(o);u\rangle)^+}{\gamma^o(-u)} \right), \tag{17}$$

introducido en [19].

Este problema guarda gran paralelismo, no solo formal, con el problema de determinación de un *hiperplano mediano* en $\mathbf{R}^p$: un hiperplano que minimiza la suma de las distancias (medidas con el calibrador γ) a un conjunto de puntos dados de dimensión máxima, e.g. [23, 26]. Para el problema de determinación de hiperplanos medianos, en [24] se descompone el espacio en regiones poliédricas en las cuales la función a minimizar es cuasicóncava o suma de dos cuasicóncavas; se tiene, [7], que existe una solución en una cara cero- o uno-dimensional respectivamente de alguno de los politopos anteriores, que se corresponden a hiperplanos que pasan por p o $p-1$ puntos afínmente independientes.

El mismo tipo de descomposición del espacio es válido para el problema (17), para el que se prueba en [25] que, bajo débiles condiciones, existe una regla de clasificación óptima, generada por un hiperplano que contiene a $p-1$ puntos afínmente independientes de $\mathcal{O}$ cuando γ es un calibrador arbitrario, incrementándose este número a p cuando γ es una norma.

Una estrategia de resolución es válida pra la escalarización (10), que en este caso se convierte en

$$\min_{u\neq 0,\beta} \max \left(\max_{o\in\mathcal{O}_1} \left(\frac{(\langle X(o);u\rangle - \beta)^+}{\gamma^o(-u)} \right), \max_{o\in\mathcal{O}_2} \left(\frac{(\beta - \langle X(o);u\rangle)^+}{\gamma^o(-u)} \right) \right). \quad (18)$$

Un *hiperplano centro* es un hiperplano que minimiza la máxima de las distancias a un conjunto de puntos, e.g. [26, 27]. En [8] se realiza una linealización parcial de los objetivos, transformando el problema (localmente) en uno de minimizar una función cuasicóncava. De esta manera llegan a probar que, cuando los conjuntos $\mathcal{O}_1$, $\mathcal{O}_2$ no son linealmente separables y de dimensión máxima, existe un hiperplano óptimo tal que el número de puntos sobre él o a distancia igual al valor óptimo, es al menos $p+1$. En el caso en que el calibrador sea una norma, además podemos garantizar que estos $p+1$ puntos son afínmente independientes.

4. MODELOS BASADOS EN PROTOTIPOS

Los métodos descritos previamente explotan fuertemente el hecho de que el espacio $\mathcal{X}$ de características, con la representación (1), es $\mathbf{R}^p$. Cuando no es ese el caso, otros métodos, que no estén basados en la esctructura de espacio vectorial de $\mathbf{R}^p$, pueden ser más apropiados.

Uno de estos métodos es el basado en la generación de prototipos para las distintas clases $g \in \mathcal{G}$, como se ha descrio en [6, 20].

Se tiene una disimilaridad $d : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$. De cada clase $g \in \mathcal{G}$ se tiene un conjunto $\mathcal{P}_g \subset \mathcal{O}$ de candidatos a *prototipos* (representantes) de la clase g.

Como espacio de reglas de clasificación Π se considera el siguiente: cada subconjunto $P \subset \cup_g \mathcal{P}_g$ de k prototipos con todas las clases representadas, se identifica con la regla de clasificación π_p al vecino más próximo, [11],

$$\pi(x) = G(\arg\min \{ d(x,X(o)) : o \in \cup_g \mathcal{P}_g \}), \quad (19)$$

donde los empates se rompen mediante un criterio preestablecido.

Para la estructura de costes (6) inducida por la métrica binaria (4), en [6,20] se demuestra que el problema (9) es NP-duro, aunque es posible formularlo como un problema de programación lineal en números enteros, resoluble exactamente para problemas de tamaño muy reducido, o aproximadamente mediante heurísticas para problemas de mayor tamaño.

5. CLASIFICACIÓN Y REGRESIÓN

Es posible acercarse a los problemas de clasificación desde la regresión y la estimación paramétrica. Tal es el caso de multitud de métodos basados en hipótesis distribucionales (e.g. normalidad) de los datos, como ocurre en el clásico y potente método de Fisher, [14], o en (generalizaciones de) la regresión logística, que describimos.

Suponemos que, si un objeto tiene características x, entonces éste proviene del grupo $g \in G$ con una probabilidad $P_g(\vartheta,x)$ donde $\vartheta = (\vartheta_g)_{g\in G}$ es un vector de parámetros.

Imponemos una forma separable, salvo constantes multiplicativas, para P_g, i.e.

$$P_g(\vartheta,x) \propto f_g(\vartheta_g,x) \tag{20}$$

dando lugar a

$$P_g(\vartheta,x) = \frac{f_g(\vartheta_g,x)}{\sum_{g\in G} f_h(\vartheta_g,x)}\,. \tag{21}$$

En caso particular más notorio es el de la regresión logística, en el que se supone $\mathcal{X} = \mathbf{R}^p$ y se hace en (20)

$$f_g(\vartheta,x) = e^{\langle\vartheta_g,x\rangle}\,, \tag{22}$$

donde $\langle\cdot;\cdot\rangle$ denota el producto escalar usual.

El conjunto de objetos $\mathcal{O}$ se entiende entonces como una realización del experimento aleatorio regido por la ley anterior, de modo que su verosimilitud conjunta $L(\vartheta;\mathcal{X}))$ viene dada por

$$L(\vartheta;\mathcal{O}) = \prod_{o\in\mathcal{G}} P_{G(o)}\,(\vartheta,X(o)), \tag{23}$$

donde los empates, caso de existir, se rompen con un mecanismo prefijado.

El espacio de reglas de clasificación considerado Π viene parametrizado a través de ϑ : para cada ϑ se define la regla de asignación al grupo más probable π_ϑ como

$$\pi_\vartheta(x) = \arg\max_{g\in G} P_g(\vartheta,x) \tag{24}$$

Como ϑ es desconocido, se propone la regla de máxima verosimilitud $\hat{\pi} = \pi_{\hat{\vartheta}}$, identificada con un estimador de máxima verosimilitud para ϑ obtenido maximizando $L(\cdot,\mathcal{O})$ como se definió en (23):

$$\max_{\vartheta\in\Theta} \sum_{o\in\mathcal{O}} \log P_{G(o)}\,(\vartheta,X(o)). \tag{25}$$

Definiendo la función de costo C como

$$C(\pi,o) = -\log P_{G(o)}(\pi,X(o)), \tag{26}$$

el problema de determinación de la regla de máxima verosimilitud (25) aparece como el problema de determinación de la regla de mínimo costo, como se definió en (8).

El problema (25), en general multimodal, deberá ser resuelto numéricamente. Mientras que un óptimo local puede ser obtenido con los procedimientos usuales de búsqueda local, para la optimización global serán necesarias técnicas más sofisticadas. En particular, bajo débiles hipótesis adicionales en el modelo (20), podemos utilizar métodos de optimización d.c., como los descritos en [1,2,3]. Para ello, suponemos que las funciones $\vartheta_g \in \Theta_g \to \log(f_g(\vartheta_g,x))$ son *d.c.* para cada $x \in \mathcal{X}$, esto es,

$$\log(f_g(\vartheta_g,x)) = \alpha_g(\vartheta_g,x) - \beta_g(\vartheta_g,x)), \tag{27}$$

donde, para cada $x \in \mathcal{X}$, las funciones $\alpha_g(\cdot,x)$, $\beta_g(\cdot,x)$ son convexas (y nosoros suponemos que conocidas). Entonces, sencillas manipulaciones algebraicas nos llevan a poder expresar la función objetivo de (25) como $\Psi_+(\vartheta) - \Psi_-(\vartheta)$, donde Ψ_+ y Ψ_- son las funciones *convexas*

$$\Psi_+(\vartheta) = \sum_{o\in\mathcal{O}} \left(\alpha_{G(o)}(\vartheta_{G(o)}, X(o)) - \log(\sum_{h\in G} f_h(\vartheta_h, X(o))) + \sum_{h\in G} \alpha_h(\vartheta_h, X(o)) \right)$$

$$\Psi_-(\vartheta) = \sum_{o\in\mathcal{O}} \left(\log(\sum_{h\in G} f_h(\vartheta_h, X(o))) + \sum_{h\in G} \alpha_h(\vartheta_h, X(o)) \right).$$

Para una estructura poliédrica de ϑ, el problema (25) puede resolverse e.g. por aproximación exterior.

REFERENCES

[1] BLANQUERO, R. *Localización de servicios en el plano mediante técnicas de optimización d.c.* Tesis Doctoral. Universidad de Sevilla, 1999.

[2] BLANQUERO, R. y E. CARRIZOSA. "On covering methods for D.C. optimization", *Journal of Global Optimization* 18, 265-274, 2000.

[3] BLANQUERO, R., E. CARRIZOSA y E. CONDE. "Finding GM-estimators with Global-Optimization techniques", *Journal of Global Optimization* 21, 223-237, 2001.

[4] BOCK, H.H. "Classification methodology", en *Handbook of data mining and knowledge discovery*, W. Klösgen and J.M. Zytkow (Eds.), Oxford, Oxford University Press, 258-267, 2002.

[5] CARRIZOSA, E., E. CONDE, F.R. FERNÁNDEZ, M. MUÑOZ y J. PUERTO. "Pareto-Optimality in Linear Regression", *Journal of Mathematical Analysis and Applications* 190, 129-141, 1995.

[6] CARRIZOSA, E., B. MARTÍN-BARRAGÁN, F. PLASTRIA y D. ROMERO-MORALES. "A Dissimilarity-based approach for Classification". METEOR Research Memorandum RM/02/027, Universiteit Maastricht, 2002.

[7] CARRIZOSA, E. y F. PLASTRIA. "Dominators for Multiple-objective Quasiconvex Maximization Problems", *Journal of Global Optimization*, Vol. 18, N° 1, pp. 35-58, 2000.

[8] CARRIZOSA, E. y F. PLASTRIA. Gauge-distances and center hyperplanes. En preparación, 2003.

[9] CHAUDHURI, D., C.A. MURTHY y B.B. CHAUDURY. "A modified metric to compute distance", *Patterns Recognition* 25, 667-677, 1992.

[10] CRISTIANINI, N. y SHAWE-TAYLOR. *An introduction to Support Vector Machines*, Cambridge, Cambridge University Press, 2000.

[11] DASARATHY, B.V. *Nearest neighbor (NN) norms - NN Patern Classification techniques*, Los Alamitos, IEEE Computer Society Press, 1991.

[12] DEVROYE, L., L. GYÖRFY y G. LUGOSY. *A probabilistic theory of Pattern Recognition*, Springer, New York, 1997.

[13] EVERITT, B.S. *Cluster Analysis*, London, Edward Arnold, 1992.

[14] FISHER, R. "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, 7, 179-188, 1936.

[15] HAND, D.J. *Construction and assessment of classification rules*, New York, Wiley, 1997.

[16] JAIN, A.K. y R.C. DUBES. *Algorithms for clustering data*, Englewoods Clifs, Prentice-Hall, 1988.

[17] KAUFMAN, L. y P.J. ROUSSEEUW. *Finding groups in data: An introductio to Cluster Analysis*, New York, Wiley, 1990.

[18] KLÖSGEN, W. y J.M. ZYTKOW. *Handbook of data mining and knowledge discovery*, Oxford, Oxford Unviersity Press, 20002.

[19] MANGASARIAN, O.L. "Arbitray-Norm Separating Plane", *Operations Research Letters*, 24, 15-23, 1999.

[20] MARTÍN-BARRAGÁN, B. *Análisis discriminante. Métodos basados en la Programación Matemática.* Trabajo de investigación DEA. Universidad de Sevilla, 2002.

[21] McLACHLAN, G.J. *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.

[22] MICHELOT, C. "The mathematics of continuous location", *Studies in Locational Analysis*, 5, 59-83, 1993.

[23] NORBACK, J.P. y J.G. MORRIS. "Fitting hyperplanes by minimizing orthogonal deviations", *Mathematical Programming* 19, 102-105, 1980.

[24] PLASTRIA, F. y E. CARRIZOSA. "Gauge-distances and median hyperplanes", *Journal of Optimization Theory and Applications*, 110, 173-182, 2001.

[25] PLASTRIA, F. y E. CARRIZOSA. "Optimal distance separating hyperplanes". *Working Paper.* Report BEIF/124, Vrije Universiteit Brussel, 2002.

[26] SCHÖBEL, A. *Locating lines and hyperplanes*, Kluwer Academic Press, Dordrecth, 1998.

[27] STILL, G. and STRENG, M. (1997). "The Ghebyshev hyperplane optimization problem", *Journal of Global Optimization* 11, 361-376.