Programación Matemática para las Máquinas de Vector de Apoyo

Tesis Doctoral

Belén Martín Barragán

Mathematical Programming for Support Vector Machines

UNIVERSIDAD D SEVILLA
FACULTAD DE MATEMÁTICAS
Dpto. de Estadística e Investigación Operativa

UNIVERSIDAD DE SEVILLA
FACULTAD DE MATEMÁTICAS
Dpto. de Estadística e Investigación Operativa

Programación Matemática para las Máquinas de Vector de Apoyo

Mathematical Programming for Support Vector Machines

Memoria presentada por
Belén Martín Barragán
para optar al grado de Doctor

Fdo: B. Martín Barragán

V° B° de los directores:

Dr. Emilio Carrizosa Priego
Dr. Dolores Romero Morales

Junio, 2006

AGRADECIMIENTOS

Este trabajo ha estado financiado por

- la beca de Formación de Personal Docente e Investigador de la Junta de Andalucía.
- el proyecto BFM2002-04525-C02-02 y su continuación MTM2005-09362-C03-01 del Ministerio de Ciencia y Tecnología y el fondo FEDER de la Unión Europea.
- las ayudas a la investigación que anualmente la Junta de Andalucía ha concedido al grupo de investigación *Optimización*, FQM-329.
- el proyecto *Optimization Models for Data Mining*, de la fundación METEOR, Holanda.

This work has been financially supported by:

- A four-years pre-doctoral grant by Junta de Andalucía, Spain.
- The research projects BFM2002-04525-C02-02 and MTM2005-09362-C03-01 by Ministerio de Ciencia y Tecnología, Spain, and European Regional Development Fund (ERDF), European Union.
- The research group FQM-329 of Junta de Andalucía, Spain.
- The research project *Optimization Models for Data Mining*, granted by METEOR, The Netherlands.

En primer lugar debo agradecer a mis directores Loli y Emilio por la gran ilusión con que se involucraron, desde un principio, en la difícil tarea de dirigir, no sólo una tesis, sino todo lo necesario para mi formación científica. Los dos se han entregado al máximo a esta tarea y han hecho un gran esfuerzo, muchas veces revisándose el trabajo en aviones y aeropuertos, para que este trabajo saliera adelante.

Loli ha sido para mi, no sólo una estupenda directora de tesis, sino también un modelo al que imitar, un exigente (como es ella consigo misma) modelo al que sueño parecerme un día. Conocer a Emilio ha sido para mi una suerte excepcional. Pocos doctorandos habrán aprendido tanto de su director como yo de él. Además me ha inculcado valores que son importantes en un científico, como el espíritu crítico, la rigurosidad, el afán por aprender, la conciencia de que la investigación debe ser útil a la sociedad, ...

Me gustaría recordar aquí a la gente que he conocido durante estos años en congresos y estancias: muy especialmente a Lázaro Cánovas, a quien muchos compañeros guardaremos en el recuerdo. Un par de semanas en Ankara, en el Euro Summer Institute, hicieron que ahora tenga amigos repartidos por todo el mundo. Me encantaría volver a verlos, como a Ivonne, en alguno de mis próximos viajes. María, Aranxa, Emmy, Patrick y Ken hicieron más divertida y enriquecedora mi estancia en Oxford, así como Jan-Willem es el culpable de que tenga un recuerdo imborrable de la ciudad de Maastricht.

También quiero dar las gracias a los jóvenes científicos de precarios por su (nuestra) lucha por la dignificación laboral de la carrera investigadora desde su comienzo. Sobre todo muchos ánimos para Carola, y su ilusión porque la asociación precarios-Sevilla salga adelante.

En la vida hay amigos que vienen y van, muchos de ellos permanecen a nuestro lado aunque estén lejos o reaparecen cuando el contacto parecía perdido. Por ello quiero mencionar en estas líneas a Raquel, a Pedro, a Yolanda, a María del Mar, a Mariani, y a tantos otros amigos a los que aprecio. También a mis amigas de Cartaya, con las que solía compartir noches de verano antes de involucrarme en esta aventura de la investigación.

No es habitual en mí dar las gracias a mis padres, pero sin duda se las merecen. A mi madre quiero agradecer, además de la deliciosa comida que me prepara, que esté siempre pendiente de lo que necesito y se preocupe por mi salud, mi futuro, mi alimentación, ... mi vida en general. Cuando estoy lejos de ella nunca la echo de menos porque en mi mente nunca dejo de tenerla a mi lado. A mi padre quiero agradecerle su apoyo, su comprensión y su respeto a mis decisiones. De él he heredado el constante afán por aprender cosas nuevas. Espero que mi título de doctora le haga sentirse aún más orgulloso de mí.

Por último, darle las gracias a mi hermano, que siempre me apoya en los momentos difíciles.

En Sevilla, a 12 de Junio de 2006

INTRODUCCIÓN

1. *Máquinas de vector de apoyo: un puente entre la Programación Matemática y la Minería de Datos.*

En la última década, la capacidad de almacenamiento de información digital se ha duplicado cada nueve meses. Crece, por tanto, a una velocidad muy superior a la prevista por la ley de Moore para el crecimiento de la capacidad de cálculo, [22, 32], provocando la aparición de las denominadas *fosas de datos*, [22]: datos que son almacenanados y descansan en paz, sin que nadie los reclame o los recuerde.

La constatación de la existencia de tales fosas de datos, y la consiguiente pérdida de oportunidades de avance en el conocimiento o de negocio, está provocando un enorme interés por el desarrollo de técnicas que, complementando a las previamente existentes, permitan obtener información desconocida y potencialmente útil de datos provinientes de campos tan diversos como la Bioinformática (expresión genética, ...), gestión de clientes (fuga de clientes, análisis de la cesta de la compra, ...), la banca (valoración de riesgo en créditos, detección de uso fraudulento de tarjetas de crédito, ...), Internet (clasificación de páginas web, filtrado de correo indeseado, ...), [1, 2, 3, 20, 24, 26, 27, 52].

Hablamos, usando una denominación habitual en los medios científicos, y, en particular, en las líneas editoriales de algunas de las revistas de más alto índice de impacto en nuestra área de conocimiento, de la *Minería de Datos*. Las referencias [2, 10, 27, 29, 51] pueden servir de introducción al tema.

Examinando, por ejemplo, las distintas opciones del software de código abierto `Weka`, [49], descrito en [51], se observa que uno de los pilares de la Minería de Datos, aunque bastante anterior a ésta, es la *Clasificación*. Encontramos junto a procedimientos bien conocidos en la comunidad estadística, como la regresión logística, los árboles de clasificación, los modelos bayesianos o las redes de neuronas artificiales, otros más recientes, como el

que nos ocupa en estas líneas: las *Máquinas de Vector de Apoyo* (en inglés, Support Vector Machines), que ha saltado del mundo del Aprendizaje Estadístico, [16, 47, 48] al de las aplicaciones ... pasando por el de la Programación Matemática. Véase [4, 7, 6, 15, 36, 40, 44, 46, 53] para otros métodos de clasificación que, como las Máquinas de Vector de Apoyo, usan técnicas avanzadas de Programación Matemática.

En el problema de clasificación, tenemos un conjunto de objetos Ω, con una partición en $\sharp(\mathcal{C})$ clases (también llamadas grupos), donde $\sharp(\cdot)$ denota el cardinal de un conjunto. El objetivo es construir una regla de clasificación para predecir la clase $c^u \in \mathcal{C}$ a la que pertenece un objeto $u \in \Omega$, por medio de su *vector predictor* x^u. Dicho vector predictor x^u toma valores en un conjunto X, normalmente un subconjunto de $\mathbb{R}^p$, y a sus componentes x_ℓ, $\ell = 1, 2, \ldots, p$, se las llama *variables predictoras.*

No toda la información acerca de los objetos de Ω está disponible, sino sólo en los objetos de un subconjunto I, llamado *muestra de aprendizaje*, donde son conocidos tanto el vector predictor como la clase a la que pertenece el objeto. La regla de clasificación se construye usando sólo la información contenida en I. Para cada $c \in \mathcal{C}$, denotaremos por I_c al conjunto de objetos de I que pertenecen a la clase c, i.e. $I_c = \{u \in I : c^u = c\}$. Además, suponemos en esta tesis, que cada clase está representada en la muestra de aprendizaje I, es decir, $I_c \neq \emptyset \forall c \in \mathcal{C}$. En general, $\mathcal{C}$ es un conjunto finito $\mathcal{C} = \{1, 2, \ldots, Q\}$. Se prestará especial atención al caso de la clasificación binaria, en el que sólo hay dos clases, $\mathcal{C} = \{-1, 1\}$. Los enfoques más utilizados para el caso multigrupo están basados en reducir el problema a una serie de problemas de clasificación para dos clases.

Este capítulo introductorio está organizado de la siguiente manera. En la Sección 2, se exponen los aspectos básicos del SVM para la clasificación binaria, y en la Sección 3, se presentan posibles extensiones para el caso multigrupo. En algunas situaciones, el SVM implica la resolución de un problema de optimización de gran tamaño. En la Sección 4, se describe brevemente una poderosa herramienta para resolverlos: la Generación de Columnas. Por último, en la Sección 5, se exponen los principales objetivos de esta tesis doctoral, y se resume el contenido de los siguientes capítulos.

2. *Clasificación binaria*

Nos centraremos ahora en el caso en el que $C = \{-1, 1\}$. El SVM propone una regla de clasificación basada en una función de puntuación f,

$$f(x) = \omega^\top x + \beta, \tag{1}$$

con $\omega \in I\!R^p$, y $\beta \in I\!R$. Con esta función de puntuación, la *regla de clasificación lineal* asigna aquellos $x \in I\!R^p$ con $f(x) > 0$ al grupo 1, y aquellos x con $f(x) < 0$ al grupo -1. En caso de empate, i.e. $f(x) = 0$, los objetos se asignarán bien aleatoriamente, o bien según un orden prefijado. A lo largo de esta tesis, siguiendo un enfoque 'en el peor caso', los empates serán considerados clasificaciones incorrectas.

La primera pregunta que nos hacemos es si existe o no (ω, β) tal que la correspondiente regla lineal clasifique correctamente a todos los objetos de I.

Definición 1: Se dice que una función de evaluación f con coeficientes (ω, β) *separa* la muestra de aprendizaje I si

$$c^u \left(\omega^\top x^u + \beta\right) > 0 \quad \forall u \in I. \tag{2}$$

Además, se dice que la muestra de aprendizaje I es separable si existe un hiperplano (ω, β) que la separa. En otro caso, se dice que I es inseparable.

En la Sección 2.1, trataremos el caso en el que I es separable, describiendo el enfoque llamado *del margen duro*. En la Sección 2.2, para el caso no separable, el enfoque del margen duro se aplica después de una transformación en los datos. En la Sección 2.3 se describe otro enfoque para el caso no separable, llamado *enfoque del margen suave*, que además es útil para evitar un fenómeno llamado sobreajuste, que ocurre cuando una baja tasa de incorrectamente clasificados en la muestra de aprendizaje I, no se generaliza a nuevos objetos en $\Omega \setminus I$. En la Sección 2.4 se presentan otras alternativas.

2.1. *El enfoque del margen duro*

Cualquier (ω, β) solución de (2) satisface que $\omega \neq 0$. En particular, (ω, β) genera un hiperplano, $\{x \in I\!R^p : \omega^\top x + \beta = 0\}$, de modo que todos los objetos en el semiespacio $\{x \in I\!R^p : \omega^\top x + \beta > 0\}$ serán asignados a la

clase 1, y todos los objetos en el semiespacio $\{x \in \mathbb{R}^p : \omega^\top x + \beta < 0\}$ serán asignados a la clase -1.

Cuando I es linealmente separable, el sistema (2) tiene infinitas soluciones, que generan infinitos hiperplanos distintos. ¿Cómo elegimos una de estas soluciones? La calidad de la clasificación, *sobre la muestra de aprendizaje*, es idéntica: todas clasifican correctamente a todos los objetos de I. Sin embargo, no todas parecen igualmente razonables. En la Figura 1 podemos ver dos hiperplanos que separan los grupos de I (círculos y cuadrados). Intuitivamente, podemos pensar que el hiperplano representado por un trazo grueso es más conveniente que el de trazo fino. En particular, este último asigna al objeto representado con '?' la clase cuadrado, cuando parece mucho más verosímil que pertenezca a la clase de los círculos.

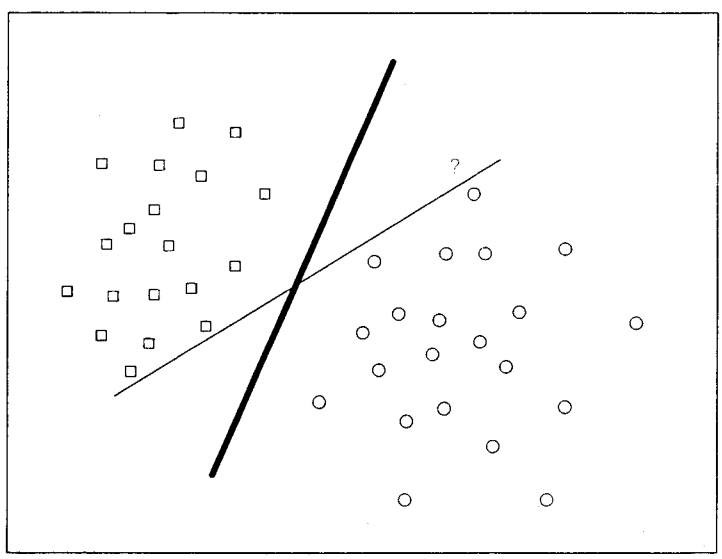

Fig. 1: ¿Dos reglas que clasifican igual de bien?

El ejemplo anterior nos indica intuitivamente la conveniencia de elegir un hiperplano que esté *alejado* de las dos clases. Las Máquinas de Vector de Apoyo se basan precisamente en este principio, como a continuación se describe. Se fija una norma $\|\cdot\|$ en $\mathbb{R}^p$ para medir las distancias (por ejemplo, la Euclídea). Para un objeto $u \in I$, la distancia entre x^u y el semiespacio en

el que quedará clasificado incorrectamente viene dada por

$$\rho^u(\omega,\beta) = \text{máx}\left\{\frac{c^u(\omega^\top x^u+\beta)}{\|\omega\|^\circ}, 0\right\},$$

e.g. [9], donde $\|\cdot\|^\circ$ denota la norma dual a $\|\cdot\|$.

Definición 2: Define el *margen del objeto* $u \in I$ como $\rho^u(\omega,\beta)$. Llamaremos *margin on the training sample* I al mínimo ρ^u :

$$\rho^I(\omega,\beta) = \min_{u\in I} \rho^u(\omega,\beta).$$

Se han obtenido cotas de la probabilidad de clasificación incorrecta para un nuevo objeto, que dependen del margen, [47, 48]. Esto proporciona un fundamento teórico para elegir el hiperplano que está más lejos de las clases. Así, el clasificador buscado es aquél que no sólo clasifica correctamente a todos los objetos de I, sino que tenga margen máximo:

$$\begin{array}{lll} \text{máx} & \rho^I(\omega,\beta) & \\ \text{s.a.:} & c^u\left(\omega^\top x^u+\beta\right) > 0 & \forall u \in I, \\ & \omega \in I\!R^p,\ \beta \in I\!R. & \end{array} \tag{3}$$

Este problema se conoce como el problema de maximización del margen duro porque exige que todos los objetos sean clasificados correctamente, mientras que su versión suave, explicada en la Sección 2.3, permite la clasificación incorrecta de algunos objetos de I.

Geométricamente, la búsqueda del clasificador de máximo margen puede verse como el problema de construir la banda de máxima anchura (las distancias medidas con la norma $\|\cdot\|$) que deja un grupo a cada lado, como se muestra en las Figuras (2)-(3) para las normas L_2 y L_∞.

Nótese que, para $\mu > 0$, $(\mu\omega, \mu\beta)$ y (ω,β) dan lugar a la misma regla de clasificación, en el sentido de que ambas clasifican un objeto a la misma clase. Además, por la Definición 2, $\rho^I(\mu\omega,\mu\beta) = \rho^I(\omega,\beta)$. Usando esta homogeneidad de la función margen, el problema de maximización del margen puede ser formulado como el siguiente problema convexo con restricciones lineales:

$$\begin{array}{lll} \text{mín} & \|\omega\|^\circ & \\ \text{s.a.:} & c^u\left(\omega^\top x^u+\beta\right) \geq 1 & \forall u \in I \\ & \omega \in I\!R^p,\ \beta \in I\!R. & \end{array} \tag{4}$$

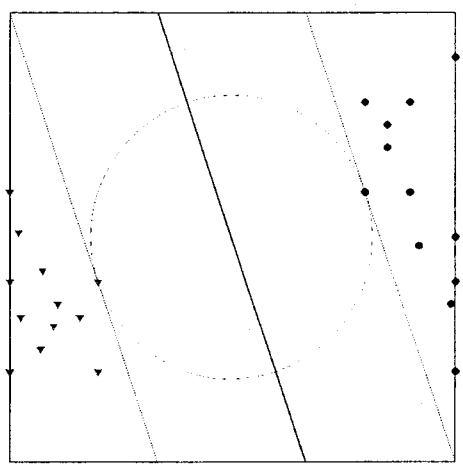

Fig. 2: Margen máximo (norma L_2)

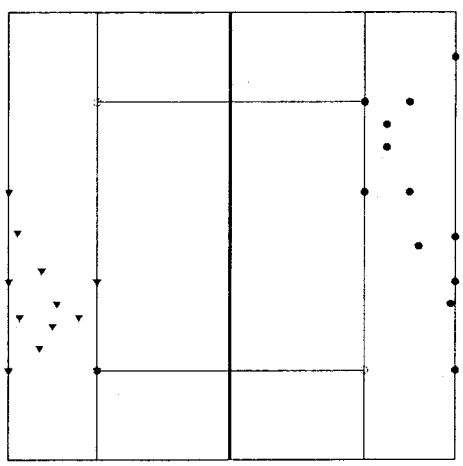

Fig. 3: Margen máximo (norma L_∞)

Si, para medir las distancias hemos usado, como en el ejemplo de la Figura 3, una norma $\|\cdot\|$ poliédrica, (i.e., cuya bola unidad es un poliedro) su norma dual $\|\cdot\|^\circ$ también es poliédrica, y por tanto, el Problema (4) puede reformularse como un problema de Programación Lineal, resoluble con optimizadores comerciales como CPLEX, [31]. Cuando la dimensión del problema de optimización es muy grande, como ocurre habitualmente en las aplicaciones de la Minería de Datos, existen técnicas avanzadas de optimización lineal, tales como la generación de columnas, brevemente descrita en la Sección 4. Además, para bases de datos en las que el número de variables predictoras p sea grande, el uso de normas poliédricas da lugar a soluciones en las que muchas de las componentes del vector ω son nulas, lo que corresponde a clasificadores que no usan todas las variables predictoras disponibles, con lo que éstas podrían descartarse para obtener así clasificadores más fácilmente interpretables o más baratos. El caso más estudiado en la literatura, no es, sin embargo, el que tiene como $\|\cdot\|$ una norma poliédrica, sino la euclídea. Entonces, el Problema (4) es equivalente al siguiente problema cuadrático convexo con restricciones lineales:

$$\begin{array}{ll} \text{mín} & \omega^\top \omega \\ \text{s.a.:} & c^u \left(\omega^\top x^u + \beta\right) \geq 1 \ \forall u \in I \\ & \omega \in I\!R^p, \ \beta \in I\!R. \end{array}$$

En [39], resultados empíricos muestran que, 'en términos de separación los SVMs basados en la norma Euclídea, la norma L_1 y la norma L_∞ tienden a ser bastante similares'.

2.2. *Inmersión en el espacio de las características*

En la Sección 2.1, hemos asumido que I era separable. Si no es el caso, el Problema (4) es infactible, por lo que deben aplicarse enfoques alternativos. Uno de estos enfoques consiste en aplicar a los datos, como preprocesamiento, una transformación $\phi : I\!R^p \to E$, donde E es, en principio, un subconjunto de $I\!R^N$, de manera que, en el nuevo espacio, la muestra de aprendizaje $\hat{I} = \{(\phi(x^u), c^u) : u \in I\}$ sea linealmente separable, [11, 18, 19, 21, 30], y pueda aplicarse, en el espacio transformado, el enfoque del margen duro descrito en la Sección 2.1. Conseguido esto, se buscan $\omega \in E$, $\beta \in I\!R$, y se construye la regla de clasificación, que estaría basada en la función de puntuación f,

$$f(x) = \omega^\top \Phi(x) + \beta, \tag{5}$$

que asigna u, como en el caso separable, asignar u, a la clase 1 si $f(x) > 0$, y a la clase -1 si $f(x) < 0$. Esta regla es lineal sobre los datos transformados, pero no lineal en el espacio original $\mathbb{R}^p$. A las distintas componentes de ϕ, se las suele conocer con el nombre de *característica* (en inglés, *feature*), mientras que al espacio E de los datos transformados, se le llama *espacio de las características* o en inglés, *feature space.*

Reescribiendo el Problema (4), pero en el espacio de las características E, en vez de en el espacio $X \subset \mathbb{R}^p$, se obtiene el siguiente problema de maximizar el margen

$$\begin{array}{ll} \text{mín} & \|\omega\|^\circ \\ \text{s.a.:} & c^u \left(\omega^\top \Phi(x^u) + \beta\right) \geq 1 \ \forall u \in I \\ & \omega \in \mathbb{R}^N, \ \beta \in \mathbb{R}. \end{array} \tag{6}$$

Para el caso en que $\|\cdot\|$ sea poliédrica y E tenga dimensión grande (pero finita), el Problema (6) se escribe como un problema lineal de gran tamaño, para cuya resolución son especialmente convenientes técnicas de generación de columnas, permitiendo, al mismo tiempo, hacer selección automática de las variables [13] y sus interacciones [14]. Además, la elección de una norma poliédrica contribuye a que la solución del Problema (6) tenga muchas componentes nulas, lo que indicaría que muchas de las características ϕ_k no son usadas por el clasificador.

Si, en cambio, usamos la norma euclídea para medir las distancias en el espacio transformado, el Problema (6) es un problema cuadrático convexo cuyo dual es

$$\begin{array}{ll} \text{máx} & \sum_{u\in I} \lambda^u - \frac{1}{2} \sum_{u,v\in I} \lambda^u \lambda^v c^u c^v \Phi(x^u)^\top \Phi(x^v) \\ \text{s.a.:} & \sum_{u\in I} c^u \lambda^u = 0 \\ & \lambda^u \geq 0 \qquad\qquad \forall u \in I. \end{array} \tag{7}$$

Definiendo el *núcleo* $K : (x, y) \in \mathbb{R}^p \times \mathbb{R}^p \longrightarrow \phi(x)^\top \phi(y) \in \mathbb{R}$, el Problema (7) se convierte en

$$\begin{array}{ll} \text{máx} & \sum_{u\in I} \lambda^u - \frac{1}{2} \sum_{u,v\in I} \lambda^u \lambda^v c^u c^v K(x^u, x^v) \\ \text{s.a.:} & \sum_{u\in I} c^u \lambda^u = 0 \\ & \lambda^u \geq 0 \qquad\qquad \forall u \in I. \end{array} \tag{8}$$

Para poder resolver el Problema (8), ni siquiera es necesario conocer ϕ, sino un algoritmo de evaluación del núcleo K que induce. De esta manera, el

espacio de las características E no necesita ser un subconjunto de $\mathbb{R}^N$ sino que, de manera más general, puede ser un espacio de Hilbert [17].

El problema de maximización resultante es cóncavo cuadrático, con tantas variables como elementos en I, y con una única restricción, lineal, junto a las de no negatividad. La dimensión de este problema es, por tanto, independiente de la dimensión p de los datos del problema original y de la dimensión de E. Esto hace de (8) una formulación especialmente atractiva en aplicaciones con no demasiados datos, pero de alta dimensionalidad, como las de, por ejemplo, [20, 52]. Para más detalles, véase, por ejemplo [17, 30].

2.3. *El enfoque del margen débil*

Una estrategia alternativa (y a veces complementaria) para abordar el caso inseparable, es la que se basa en la maximización del *margen suave*, [16, 17, 30], en la que, partiendo del problema infactible (4), se perturban sus restricciones para hacerlo factible, introduciendo una penalización en el objetivo para controlar la perturbación introducida. Así se obtiene el problema (siempre factible)

$$\begin{array}{ll} \text{mín} & \|\omega\|^\circ + C\|\xi\|^* \\ \text{s.a.:} & c^u\left(\omega^\top x^u + \beta\right) + \xi^u \geq 1, \qquad \forall u \in I \\ & \omega \in \mathbb{R}^p,\ \beta \in \mathbb{R},\ \xi \in \mathbb{R}^{\sharp(I)}, \end{array} \tag{9}$$

donde $\|\cdot\|^*$ denota una norma en $\mathbb{R}^N$, que no tiene por qué coincidir con la norma $\|\cdot\|^\circ$. $C > 0$ es una constante que se usa para equilibrar la perturbación ξ y el margen en los puntos correctamente clasificados, usualmente elegida por técnicas de validación cruzada, [35].

Por ejemplo, usando la norma poliédrica $\|\cdot\|_1$ tanto para ω como para las desviaciones $\xi = (\xi_u)_{u\in I}$, el problema relajado (9) se formula como

$$\begin{array}{ll} \text{mín} & \|\omega\|_1 + C\|\xi\|_1 \\ \text{s.a.:} & c^u\left(\omega^\top x^u + \beta\right) + \xi^u \geq 1, \qquad \forall u \in I \\ & \omega \in \mathbb{R}^p,\ \beta \in \mathbb{R},\ \xi \in \mathbb{R}^{\sharp(I)}. \end{array} \tag{10}$$

Otras elecciones de $\|\cdot\|^*$ dan lugar a diferentes problemas de maximización del margen suave.

A veces, el enfoque del margen suave, se usa incluso cuando I es separable, porque se ha demostrado empíricamente que evita el sobreajuste. Además,

dicho enfoque es habitualmente usado después de hacer una transformación de los datos, como la descrita en la Sección 2.2, dando lugar, para la norma $\|\cdot\|_1$, al siguiente problema:

$$\begin{array}{ll} \text{mín} & \|\omega\|_1 + C\|\xi\|_1 \\ \text{s.a.:} & c^u\left(\omega^\top \Phi(x^u) + \beta\right) + \xi^u \geq 1, \qquad \forall u \in I \\ & \omega \in I\!R^N,\ \beta \in I\!R,\ \xi \in I\!R^{\sharp(I)}, \end{array} \tag{11}$$

donde N denota el cardinal del conjunto de features $\mathcal{F}$, que define la inmersión $\Phi = (\phi)_{\phi\in\mathcal{F}}$. El Problema (11) se puede formular como el siguiente problema de Programación Lineal

$$\begin{array}{lrl} \text{mín} & \sum_{\phi\in\mathcal{F}}(\omega_\phi^+ + \omega_\phi^-) + C\sum_{u\in I}\xi^u & \\ \text{s.a.:} & \sum_{\phi\in\mathcal{F}}(\omega_\phi^+ - \omega_\phi^-)c^u\phi(x^u) + \beta c^u + \xi^u \geq 1 & \forall u \in I \\ & \omega_\phi^+ \geq 0 & \phi \in \mathcal{F} \\ & \omega_\phi^- \geq 0 & \phi \in \mathcal{F} \\ & \xi^u \geq 0 & \forall u \in I \\ & \beta \in I\!R. & \end{array} \tag{12}$$

2.4. *Otras alternativas*

Recientemente se han propuesto estrategias alternativas a las dos descritas previamente: en lugar de transformar el problema en otro sobre el cual las poblaciones sean separables (y poder usar, por tanto, el criterio maximin), puede utilizarse otro criterio que no requiera separabilidad. Tal es el caso, por ejemplo, del método descrito en [37, 42], en el que se propone la minimización de la suma de las distancias de cada punto a su semiespacio de clasificación correcta. El problema de optimización resultante,

$$\begin{array}{ll} \text{mín} & \sum_{u\in I} \text{máx}\left\{\frac{-c^u(\omega^\top x^u+\beta)}{\|\omega\|^\circ}, 0\right\} \\ \text{s.a.:} & \omega \in I\!R^p,\ \beta \in I\!R, \end{array}$$

goza de las buenas propiedades del hiperplano mediano, [38, 41], pero, al igual que aquél, es multimodal.

3. *Clasificación multigrupo*

La extensión del SVM al caso en que el número de grupos Q es mayor que dos no es directa. Los enfoques más usados están basados en realidad

en combinaciones de clasificadores obtenidos como solución de una serie de SVMs para dos clases. Ya en [16], que es el origen del SVM, se propone por primera vez el método *uno frente a todos*. En él, se construyen Q reglas de clasificación. Para $c \in \mathcal{C}$, la c-ésima regla de clasificación se construye mediante el SVM aplicado a la clasificación de los objetos de la clase c frente al resto de los objetos de I. La regla de clasificación final asigna un objeto a la clase a la que es asignado más veces por las Q reglas de clasificación.

Otro método basado en la misma idea es el llamado *uno frente a uno* [28], en el que se aplica el SVM a cada par de clases, ignorando los restantes objetos. Las $\frac{Q(Q-1)}{2}$ reglas de clasificación obtenidas, se combinan, por ejemplo, de la misma manera que en el método uno frente a todos. Otra posible vía para combinar esas reglas de clasificación puede verse en [43].

Otros autores, como por ejemplo [8, 25, 50], han propuesto el uso de una *función de puntuación (multiple)* $f = (f_c)_{c\in\mathcal{C}}$ con Q componentes $f_c : \mathcal{X} \longrightarrow \mathbb{R}$ de la forma (1). A continuación, un objeto $u \in \Omega$ se asignará a la clase c^* con mayor valor de la función de puntuación f_{c^*}.

$$c^* = \arg\max_{c\in\mathcal{C}} f_c(z).$$

Al igual que en el caso de dos grupos, cuando hay empates, habría ambigüedad en la clasificación de los objetos con $f(x^u) = 0$, a los que se les asignaría aleatoriamente una de las clases, o bien siguiendo un orden prefijado. Siguiendo un enfoque 'en el peor caso', los empates serán considerados clasificaciones incorrectas.

Prestaremos especial atención a esta función de puntuación múltiple, ya que, en el Capítulo 5, se usa un modelo basado en dicha función. Cada componente de dicha función de puntuación Denotemos por $W = (\omega^1, \ldots, \omega^Q)$ y $b = (\beta^1, \ldots, \beta^Q)$, a los coeficientes $\omega^c \in \mathbb{R}^p$, y $\beta^c \in \mathbb{R}$ de la función de puntuación f_c, componente c-ésima de f,

$$f_c(x) = \sum_{k=1}^{p} \omega_k^c x_k + \beta^c. \tag{13}$$

A continuación, extendemos al caso multigrupo el concepto de separabilidad, definido en la Sección 2.

Definición 3: Una función de puntuación $f = (f_c)_{c\in\mathcal{C}}$ de la forma (13) se dice que *separa* I si

$$f_{c^u}(x^u) > f_j(x^u) \quad \forall j \neq c^u, \forall u \in I.$$

Además, decimos que I es *separable* si existe una función $f = (f_c)_{c \in \mathcal{C}}$, de la forma (13), que separa I.

Nótese que, para la clasificación binaria, $\mathcal{C} = \{-1, 1\}$, este concepto de separabilidad es equivalente al concepto dado en la Sección 2, Definición 1, como puede verse en la siguiente propiedad.

Propiedad 4: Sea $\mathcal{C} = \{-1, 1\}$ y la muestra de aprendizaje I. Las dos condiciones siguientes son equivalentes

- Existe una función de puntuación múltiple $\hat{f} = (f_{-1}, f_1)$ que separa I.
- Existe una función de puntuación f de la forma (1) que separa I.

Demostración. Dada una función de puntuación múltiple $\hat{f} = (f_{-1}, f_1)$ con coeficientes $(\omega^{-1}, \beta^{-1})$ y (ω^1, β^1), la función de puntuación f, definida por los coeficientes $\omega = \omega^1 - \omega^{-1}$ y $\beta = \beta^1 - \beta^{-1}$, separa I.

A la inversa, dado (ω, β), tal que satisface (2), fijando $\omega^1 = \omega$, $\beta^1 = \beta$, $\omega^{-1} = 0$ y $\beta^{-1} = 0$, tenemos una función de puntuación múltiple que separa I. □

Al igual que en la Sección 2, exploramos primero las formulaciones para el margen duro (Sección 3.1), y más adelante, en la Sección 3.2, las extendemos al enfoque del margen suave, en el que se permite la clasificación incorrecta de algunos objetos de I.

3.1. *El enfoque del margen duro*

Supongamos que I es separable, lo que asegura la existencia de al menos una función de puntuación f que separa I. La unicidad de dicha función f nunca se da. De hecho, es fácil ver que, dado $(\hat{\omega}, \hat{\beta}) \in \mathbb{R}^{p+1}$ las reglas de clasificación obtenidas con los coeficientes (W, b) y $(\tilde{W}, \tilde{b})$ con $\tilde{\omega}^c = \lambda \omega^c + \hat{\omega}$ y $\tilde{\beta}^c = \lambda \beta^c + \hat{\beta}$, para todo $c \in \mathcal{C}$, son equivalentes para todo $\lambda > 0$, en el sentido de que asignan los objetos a las mismas clases.

Además, hay también más de una función de puntuación que separa I y no son equivalentes. Por ejemplo, dada una función de puntuación f que separa I, sea ε cualquier valor que satisface

$$0 < \varepsilon < \min_{u \in I} \min_{j \neq c^u} \{ f_{c^u}(x^u) - f_j(x^u) \} .$$

Entonces, la función $f^\varepsilon = (f_1 + \varepsilon, f_2, \ldots, f_Q)$ también separa I. Necesitamos un criterio para elegir una de dichas funciones de puntuación múltiple. En el caso de la clasificación binaria, siguiendo los trabajos de Vapnik [47], se presentó el criterio de maximización del margen. A continuación, presentamos una extensión del concepto de margen para el caso multigrupo.

Definición 5: Sea $\|\cdot\|^\circ$ una norma en $\mathbb{R}^{pQ}$. El *margen de un objeto* u con respecto a la función de puntuación (W, b) con $W \neq 0$, se define como

$$\rho^u(W, b) = \min_{j \in \mathcal{C}, j \neq c^u} \frac{f_{c^u}(x^u) - f_j(x^u)}{\|W\|^\circ}.$$

Así mismo, el *margen de la función de puntuación* (W, b) con respecto a la muestra de aprendizaje I viene dado por el mínimo

$$\rho^I(W, b) = \min_{u \in I} \rho^u(W, b).$$

Como en el caso de la clasificación binaria, una elección habitual de la norma $\|\cdot\|^\circ$, es la norma Euclídea, pero otras normas podrían ser útiles. Por ejemplo, en el Capítulo 5, usamos la norma L_1, que nos permite formular el problema del máximo margen como un problema de Programación Lineal, resoluble con software comercial ya existente.

A continuación, consideramos el problema de maximización del margen

$$\begin{array}{ll} \text{máx} & \rho^I(W, b) \\ \text{s.t:} & (W, b) \in \mathbb{R}^{pQ} \times \mathbb{R}^Q. \end{array} \tag{14}$$

Denotemos por $\theta(W, b)$ la cantidad

$$\min_{u \in I} \min_{j \in \mathcal{C}, j \neq c^u} \left(f_{c^u}(x^u) - f_j(x^u)\right),$$

que es el numerador del margen $\rho^I(W, b)$. Nótese que para todo $\lambda > 0$, las soluciones (W, b) y $(\lambda W, \lambda b)$ dan lugar a la misma regla de clasificación, y, siguiendo la Definición 5, también dan lugar al mismo margen. Usando esta propiedad, el Problema (14) puede ser formulado como el siguiente problema de optimización convexo:

$$\begin{array}{ll} \text{máx} & \theta(W, b) \\ \text{s.t:} & \|W\|^\circ \leq 1, \\ & (W, b) \in \mathbb{R}^{pQ} \times \mathbb{R}^Q, \end{array} \tag{15}$$

en el sentido de que cualquier solución óptima del Problema (15) es también óptima para el Problema (14), y para cualquier solución óptima (W^*, b^*) del Problema (14), se tiene que

$$(\hat{W}, \hat{b}) = \frac{1}{\|W^*\|^\circ}(W^*, b^*)$$

es también una solución óptima de (15).

La Formulación (15) se deriva del uso de la restricción de normalización $\|W\|^\circ = 1$, i.e. se normaliza el denominador del margen $\rho^I(W, b)$ y se maximiza su numerador, $\theta(W, b)$. Otra formulación equivalente consiste en normalizar el numerador y minimizar $\|W\|^\circ$, dando lugar al problema

$$\begin{array}{ll} \text{mín} & \|W\|^\circ \\ \text{s.a.:} & \theta(W, b) \geq 1 \\ & (W, b) \in I\!R^{pQ} \times I\!R^Q. \end{array} \tag{16}$$

A lo largo de esta tesis doctoral, usaremos la formulación cuyas propiedades nos sean más útiles.

3.2. *El caso no separable*

En las Secciones 2.2 y 2.3, se han presentado dos enfoques diferentes para la clasificación binaria en el caso no separable: transformar los datos a un espacio de dimensión superior, y la maximización del margen suave, que permite que algunos objetos queden incorrectamente clasificados por la regla de clasificación. Los dos enfoques se pueden extender al caso no separable multigrupo.

Después de aplicar una inmersión $\Phi : I\!R^p \to E$, con $\Phi = (\phi_1, \phi_2, \ldots, \phi_N)$ a los datos de I, la función de puntuación f se expresa como

$$f_c(x) = \sum_{k=1}^{N} \omega_k^c \phi_k(x) + \beta^c = (\omega^c)^\top \Phi(x) + \beta^c, \tag{17}$$

con $\omega^c \in I\!R^N$, y $\beta^c \in I\!R$, $\forall c \in \mathcal{C}$.

Al igual que en el caso de la clasificación binaria, se puede desarrollar una versión del margen suave para el Problema (14), de la cual se puede esperar que evite el sobreajuste, al permitir que algunos objetos de I queden incorrectamente clasificados.

Una función de puntuación f que no separa I, sería infactible para el Problema (16) porque la restricción $\theta(W,b) \geq 1$ no se cumpliría. Dicha restricción puede reescribirse como el conjunto de restricciones

$$(\omega^{c^u})^\top \Phi(x^u) + \beta^{c^u} - (\omega^j)^\top \Phi(x^u) - \beta^j \geq 1, \quad \forall u \in I, \forall j \in \mathcal{C}, j \neq c^u.$$

Entonces, para permitir que algunos objetos queden incorrectamente clasificados, relajamos dichas restricciones añadiendo las perturbaciones ξ_u^j, que serán luego penalizadas en la función objetivo. Sea ξ el vector de todas las perturbaciones ξ_u^j. Entonces, la versión del Problema (16) para la maximización del margen suave, se puede formular como

$$\begin{array}{ll} \text{mín} & \|W\|^\circ + C\|\xi\|^* \\ \text{s.a.:} & (\omega^{c^u})^\top \Phi(x^u) + \beta^{c^u} - (\omega^j)^\top \Phi(x^u) - \beta^j + \xi_u^j \geq 1, \\ & \qquad\qquad \forall u \in I, \forall j \in \mathcal{C}, j \neq c^u \\ & (W,b) \in I\!R^{NQ} \times I\!R^Q, \end{array} \tag{18}$$

donde $\|\cdot\|^*$ es una norma dada, que no necesita ser la misma que la norma $\|\cdot\|$ usada para los coeficientes de W. C es una constante dada, cuyo propósito es equilibrar las perturbaciones ξ_u^j y el margen. Una elección común para la norma $\|\cdot\|^*$ es, por ejemplo, la norma L_1, $\|\xi\|_1 = \sum_{u \in I, j \in \mathcal{C}, j \neq c^u} \xi_u^j$, [50].

En principio, la Formulación (18) usa, para cada objeto $u \in I$, $Q-1$ perturbaciones, una para cada restricción. Para una función de puntuación definida por (W,b), si un objeto u es incorrectamente clasificado, la cantidad

$$f_{c^u}(x^u) - f_j(x^u) = (\omega^{c^u})^\top \Phi(x^u) + \beta^{c^u} - (\omega^j)^\top \Phi(x^u) - \beta^j$$

puede verse como una medida lo alejado que está el objeto u de ser correctamente clasificado. De esta manera, puede parecer más lógico usar sólo una perturbación η_u por objeto u, lo que puede hacerse, en el marco de la Formulación (18), por medio de la elección de la norma $\|\xi\|^* = \sum_{u \in I} \text{máx}_{j \in \mathcal{C}, j \neq c^u} \xi_u^j$, que da lugar al siguiente problema

$$\begin{array}{ll} \text{mín} & \|W\|^\circ + C\|\eta\|_1 \\ \text{s.a.:} & (\omega^{c^u})^\top \Phi(x^u) + \beta^{c^u} - (\omega^j)^\top \Phi(x^u) - \beta^j + \eta_u \geq 1, \\ & \qquad\qquad \forall u \in I, \forall j \in \mathcal{C}, j \neq c^u \\ & (W,b) \in I\!R^{NQ} \times I\!R^Q, \end{array} \tag{19}$$

donde η denota el vector $(\eta_u)_{u \in I}$ de una perturbación por objeto.

4. *Problemas lineales de gran tamaño*

Cuando la inmersión Φ tiene una dimensión alta (finita), el Problema (12) llega a tener muchas variables de decisión, por lo que proponemos el uso de una conocida técnica de Programación Matemática llamada Generación de Columnas, propuesta inicialmente para el problema de corte [23]. En esta sección, describimos brevemente dicha herramienta, que será usada en los Capítulos 2 y 3.

Cuando un problema de Programación Lineal (P) tiene un alto número de variables de decisión, en vez de resolverlo directamente, la técnica de Generación de Columnas, resuelve una serie de problemas reducidos donde se considera sólo un subconjunto V del conjunto de variables de decisión $\mathcal{V}$. Las restantes variables de decisión de $\mathcal{V}$ son añadidas iterativamente a V sólo cuando se necesitan.

Para $V \subset \mathcal{V}$, sea *el Problema Maestro* (P-V) una modificación del Problema (P), el cual estamos interesados en resolver, en la que se han fijado a cero todas las variables de decisión que no están en V. Empezamos con un conjunto inicial V, por ejemplo, extraído aleatoriamente del conjunto $\mathcal{V}$. Una vez que hemos generado dicho conjunto inicial, resolvemos el Problema Maestro (P-V). El siguiente paso es comprobar si la solución obtenida es también óptima para el Problema (P), y, en caso contrario, generar una nueva variable de decisión v en $\mathcal{V} \setminus V$ tal que la solución del nuevo problema (P-$(V \cup \{v\})$) sea mejor, para el Problema (P), que la anterior. Entonces, la variable de decisión v se añade al subconjunto V, y se resuelve de nuevo el Problema Maestro (P-V). Este proceso se repite hasta que no se encuentra ninguna variable de decisión prometedora.

En cada paso, nos gustaría idealmente encontrar qué variable de decisión es la más prometedora, en el sentido de que, al añadirla a V dará lugar a la máxima mejora en la función objetivo del Problema (P). El problema auxiliar de encontrar la variable de decisión más prometedora, llamado en inglés *pricing problem*, está relacionado con el problema de encontrar la restricción más violada de la formulación dual del Problema (P). A veces, para acelerar el algoritmo de generación de columnas, se añaden a V varias variables de decisión, en vez de sólo una.

A continuación presentamos un esquema resumen de un algoritmo de generación de columnas típico.

CG-resumen: Resumen del algoritmo de generación de columnas

Paso 0. Generar un conjunto inicial de variables de decisión V.

Paso 1. Resolver el problema auxiliar para generar una nueva variable de decisión v.

Paso 2. Si no se encuentra ninguna variable de decisión prometedora, FIN: se ha encontrado la solución óptima del Problema (P). En caso contrario, añadir v al conjunto V, resolver el Problema Maestro (P-V), e ir al Paso 1.

Cuando, en vez de usar un algoritmo exacto para resolver el problema auxiliar, se usa un heurístico, entonces, el algoritmo de Generación de Columnas genera un resultado sin garantías de optimalidad.

5. *Contenido de la tesis*

Las reglas de clasificación basadas en el margen, han mostrado ser muy eficientes en las aplicaciones de la Minería de Datos. A pesar de los avances obtenidos en los últimos años, hay aún muchos aspectos (de modelización, numéricos, algorítmicos, etc) a explorar.

En esta tesis doctoral, presentamos algunas propuestas en las que las herramientas de la Programación Matemática se usan para obtener clasificadores que tengan algunas propiedades interesantes. En aplicaciones prácticas, el principal objetivo es obtener clasificadores con un bajo porcentaje de objetos mal clasificados, pero el hecho de que, además, sean fácilmente interpretables, o baratos, o útiles para detectar el papel que juegan las distintas variables y sus interacciones, podría ser también de gran interés. Por ejemplo, en análisis de ADN, la interpretabilidad de las reglas de clasificación obtenidas es uno de los aspectos que influye en la elección de un método de predicción, [45]; así, a veces, se prefieren modelos fácilmente interpretables, que puedan proporcionar algún conocimiento médico más allá de la predicción. En algunos campos, tan diversos como diagnóstico de cáncer o valoración del crédito (en inglés *credit scoring*), médicos o prestamistas podrían encontrar importante explicar fácilmente la regla de clasificación usada, y detectar qué combinaciones de variables son críticas en la predicción. Otras veces, la importancia o coste de clasificar correctamente un objeto o individuo varía dependiendo de la clase a la que éste pertenece. De esta manera,

en muchas aplicaciones, tienen alto interés, métodos que tengan en cuenta dicha importancia o coste de clasificación incorrecta que depende de la clase.

Nuestros objetivos principales en este trabajo son:

1. Modelar, con técnicas basadas en SVM, problemas de clasificación que incorporen costes (de clasificación incorrecta, de medición de la variable, de obtención de las características).

2. Detectar automáticamente transformaciones no lineales de los datos, e interacciones relevantes entre las variables, por medio de SVM.

3. Analizar empíricamente el comportamiento de los métodos propuestos, en bases de datos reales habituales en la literatura de Minería de Datos.

En el Capítulo 2, basado en nuesto artículo [13], se propone un modelo basado en SVM, en el que, a la vez que goza de una buena tasa de clasificación correcta, detecta automáticamente las variables más importantes y los valores de dichas variables que son críticos para la clasificación. El método supone la optimización de un problema lineal de gran escala, para el que se usa la técnica conocida como Generación de Columnas, descrita en la Sección 4. Además el clasificador propuesto es poco sensible frente a la presencia de outliers. En el Capítulo 3, extendemos dicho método a uno que, aparte de detectar las variables relevantes, también detecta las interacciones entre ellas. La habilidad clasificadora del método propuesto es comparable al estándar SVM para diferentes kernels, y claramente mejor que los Árboles de Clasificación. Dichos resultados son la base de nuestro artículo [14].

El problema de incorporar costes de clasificación incorrecta basados en la clase se analiza en el Capítulo 4. Basado en nuestro artículo [11], en dicho capítulo proponemos un modelo en el que, para la función de puntuación f, se define el margen de una clase c independientemente del margen de la otra clase, y se analiza el problema de la maximización simultánea de ambos márgenes.

En muchas aplicaciones prácticas, es importante que el clasificador construido sea barato y rápido de aplicar a nuevos individuos. Por ejemplo, en un sondeo de KDnuggets, [33], dirigido tanto al mundo profesional como al académico, se eligió a 'tratar con datos no equilibrados y sensibles a los costes'

como uno de los temas más importantes en la Minería de Datos, donde sensible a los costes significa datos 'que tienen diferente coste de adquisición', [34]. Esta situación se ilustra en [34] con el ejemplo siguiente:

> *En el problema del diagnóstico médico, podemos usar datos de un análisis de sangre, o de un análisis de líquido cefalorraquídeo, pero el análisis de sangre es mucho más barato (y fácil) de hacer que un análisis de líquido cefalorraquídeo. Tomar decisiones en tales casos, (realmente en todos los casos) requiere combinar técnicas exactas con otras medidas del coste de obtener los datos.*

En el Capítulo 5, basado en nuestro artículo [12], proponemos, para el caso de la clasificación multigrupo, un problema de optimización biobjetivo que tiene en cuenta tanto los costes de medición (económico, de riesgo para el paciente, computacional, requerimientos de almacenaje), como también la tasa de clasificación correcta. De nuevo, dentro de un marco basado en el SVM, la tasa de clasificación incorrecta se trata a través de la maximización del margen. Presentamos un problema biobjetivo entero mixto, para el que se obtienen las soluciones Pareto-ópitmas. Dichas soluciones Pareto-óptimas corresponden a diferentes reglas de clasificación, entre las cuales el usuario eligirá aquella que corresponde a un apropiado compromiso entre el coste de clasificación y la tasa de clasificación incorrecta esperada.

Los resultados de esta tesis doctoral, muestran el gran potencial que las herramientas de la Programación Matemática, y en particular, la Programación Lineal y la Programación Entera Mixta, tienen a la hora de desarrollar potentes variantes de las Máquinas de Vector de Apoyo.

CONCLUSIONES Y CONSIDERACIONES FINALES

El problema de clasificación es una de las principales tareas en la Minería de Datos. En esta tesis doctoral, hemos mostrado el gran potencial de las herramientas de la Programación Matemática, y en particular de la Programación Lineal, para modelar y resolver problemas de clasificación donde el poder clasificador de la regla de clasificación obtenida no es el único objetivo, sino que otras propiedades son también deseables.

Cuando el SVM se formula como un problema de Programación Lineal, técnicas avanzadas, tales como Generación de Columnas, son útiles para resolver problemas que, de otro, modo serían difíciles de atacar. Por ejemplo, cuando se tienen en cuenta todos los umbrales y todas las interacciones entre variables, el problema de Programación Lineal es demasiado grande para resolverlo con técnicas estándar. Además, en los algoritmos de Generación de Columnas propuestos en los Capítulos 2 y 3, las columnas corresponden a características (definidas por una variable con su umbral, o por un conjunto de variables que interactúan). Así, el problema auxiliar que genera las columnas realmente genera dichas características, de manera que no es necesario generarlas todas en lo que sería una tediosa fase de preprocesamiento. El estudio de otros algoritmos para resolver el problema auxiliar, de manera que utilicen información estadística de las variables podría ser interesante y podría acelerar el proceso de generación de columnas. Por ejemplo, para generar nuevas columnas podríamos usar aquellas medidas (tales como la entropía, la ganancia de información, el índice de Gini) que se usan con éxito en los Árboles de Clasificación.

Podríamos además añadir restricciones a las características, de manera que se permiten sólo algunas de ellas. Por ejemplo, en [5], se propone el uso de un conjunto de características (llamadas patrones positivos y patrones negativos) que son similares a las presentadas en el Capítulo 3, pero imponiendo, además, ciertas condiciones sobre cómo clasifican los objetos de I cuando no están combinadas con otras características. En [5], se proponen varios enfoques para combinar dichas características (o patrones), uno de los

cuales es la maximización del margen. Sin embargo, en el método propuesto en [5], se debe generar todos los posibles patrones antes de resolver el problema de maximización. Una extensión del algoritmo de Generación de Columnas propuesto en los Capítulos 2 y 3, permitiría generar esos patrones sólo cuando se necesitan.

En el Capítulo 4, formulamos el problema de la maximización simultánea del margen en cada clase, definido independientemente de la otra clase. Este problema es útil para encontrar clasificadores en los que la importancia de cometer un error depende no es igual para las dos clases. La optimización multicriterio nos permite analizar dicha situación y caracterizar los clasificadores que no pueden ser mejorados simultaneamente en las dos clases. En el caso Euclídeo, este resultado nos da un nuevo fundamento teórico para el uso de las curvas ROC. Generalizar este resultado a otras normas y al caso multigrupo es, en nuestra opinión, una interesante línea abierta.

Por último, el desarrollo de las herramientas de la Programación Lineal y de la Programación Entera Mixta para resolver problemas multicriterio permite encontrar un conjunto reducido de buenos clasificadores, en el sentido de que no pueden ser mejorados simultáneamente en todas las propiedades consideradas. Un ejemplo de esta situación ha sido presentado en el Capítulo 5, en el que se buscaban clasificadores baratos con buena calidad de clasificación.

BIBLIOGRAFíA

[1] S. Alexe, E. Blackstone, P. Hammer, H. Ishwaran, M. Lauer, and C.E. Pothier Snader. Coronary risk prediction by logical analysis of data. *Annals of Operations Research*, 119:15–42, 2003.

[2] C. Apte. The big (data) dig. *OR/MS Today*, February 2003.

[3] C. Apte, B. Liu, E.P.D. Pednault, and P. Smyth. Business applications of data mining. *Communications of the ACM*, 45:49–53, 2002.

[4] K.P. Bennet and O.L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software*, 1:23–24, 1992.

[5] E. Boros, P. L. Hammer, T. Ibaraki, and A. Kogan. A logical analysis of numerical data. *Mathematical Programming*, 79:163–190, 1997.

[6] P.S. Bradley, U.M. Fayyad, and O.L. Mangasarian. Mathematical programming for data mining: formulations and challenges. *INFORMS Journal on Computing*, 11(3):217–238, 1999.

[7] P.S. Bradley, O. Mangasarian, and D. Musicant. Optimization methods in massive datasets. In J. Abello, P.M. Pardalos, and M.G.C. Resende, editors, *Handbook of Massive Datasets*, pages 439–472. Kluwer Academic Publishers, 2002.

[8] E. Bredensteiner and K. Bennet. Multicategory classification by support vector machines. *Computational Optimization and Applications*, 12:53–79, 1999.

[9] E. Carrizosa and J. Fliege. Generalized goal programming: Polynomial methods and applications. *Mathematical Programming*, 93:281–303, 2002.

[10] E. Carrizosa and B. Martín-Barragán. Problemas de clasificación: una mirada desde la localización. In Blas Pelegrín, editor, *Avances en localización de servicios y sus aplicaciones*. Servicio de Publicaciones de la Universidad de Murcia, 2005.

[11] E. Carrizosa and B. Martín-Barragán. Two-group classification via a biobjective margin maximization model. *European Journal of Operational Research*, 2006. In press.

[12] E. Carrizosa, B. Martín-Barragán, and D. Romero Morales. A biobjective model to select features with good classification quality and low cost. In *Proceedings of the Fourth IEEE ICML*, pages 339–342. IEEE Publications, 2004.

[13] E. Carrizosa, B. Martín-Barragán, and D. Romero Morales. A column generation approach for support vector machines. Technical report, Optimization Online, 2006. `http://www.optimization-online.org/DB_HTML/2006/04/1359.html`.

[14] E. Carrizosa, B. Martín-Barragán, and D. Romero Morales. Detecting relevant variables and interactions for classification in support vector machines. Technical report, Optimization Online, 2006. `http://www.optimization-online.org/DB_HTML/2006/05/1385.html`.

[15] E. Carrizosa and F Plastria. Optimal expected-distance separating halfspace. Technical Report MOSI/7, Vrije Universiteit Brussel, 2004.

[16] C. Cortes and V. Vapnik. Support-vector network. *Machine Learning*, 1:113–141, 1995.

[17] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines*. Cambridge University Press, 2000.

[18] A. Demiriz, K. P. Bennett, and J. Shawe-Taylor. Linear programming boosting via column generation. *Machine Learning*, 46(1):225–254, 2002.

[19] A.P. Duarte Silva and A. Stam. Second order mathematical programming formulations for discriminant analysis. *European Journal of Operational Research*, 72:4–22, 1994.

[20] B. Efron, R. Tibshirani, J. Storey, and V. Tusher. Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

[21] J.E. Falk and V.E. Karlov. Robust separation of finite sets via quadratics. *Computers and Operations Research*, 28:537–561, 2001.

[22] U. Fayyad and R. Uthurusamy. Evolving data mining into solutions for insight. *Communications of the ACM*, 45:28–31, 2002.

[23] P.C. Gilmore and R.E. Gomory. A linear programming approach to the cutting-stock problem. *Operations Research*, 9:849–859, 1961.

[24] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[25] Y. Guermeur. Combining discriminant models with multi-class SVMs. *Pattern Analysis and Applications*, 5:168–179, 2002.

[26] J. Han, R.B. Altman, V. Kumar, H. Mannila, and D. Pregibon. Emerging scientific applications in data mining. *Communications of the ACM*, 45:54–58, 2002.

[27] H. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining.* The MIT Press, 2001.

[28] T. Hastie and R. Tibshirani. Classification by pairwise coupling. *The Annals of Statistics*, 26(2):451–471, 1998.

[29] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer, 2001.

[30] R. Herbrich. *Learning Theory Classifiers. Theory and Algorithms.* MIT Press, 2002.

[31] ILOG CPLEX 8.1 User's Manual. `http://www.pcs.cnu.edu/~riedl/software/cplex81/doc/userman/onlinedoc`.

[32] Informe de intel sobre la ley de moore. http://www.intel.com/technology/mooreslaw/index.htm.

[33] KDnuggets. http://www.kdnuggets.com/about/index.html.

[34] KDnuggets : Polls : Important data mining topics. http://www.kdnuggets.com/polls/2005/important_data_mining_topics.htm.

[35] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 1137–1143. Morgan Kaufmann, 1995.

[36] O.L. Mangasarian. Mathematical programming in data mining. *Data Mining and Knowledge Discovery*, 42(1):183–201, 1997.

[37] O.L. Mangasarian. Arbitrary-norm separating plane. *Operations Research Letters*, 24:15–23, 1999.

[38] H. Martini and A. Schöbel. Median and center hyperplanes in minkowski spaces –a unifying approach. *Discrete Mathematics*, 241:407–426, 2001.

[39] J.P. Pedroso and N. Murata. Support vector machines with different norms: motivation, formulations and results. *Pattern recognition letters*, 22:1263–1272, 2001.

[40] S. Piramuthu. Evaluating feature selection methods for learning in data mining applications. *European Journal of Operational Research*, 156:483–494, 2004.

[41] F. Plastria and E. Carrizosa. Gauge distances and median hyperplanes. *Journal of Optimization Theory and Applications*, 110:173–182, 2001.

[42] F. Plastria and E. Carrizosa. Optimal distance separating halfspace. Technical report, BEIF/124, Vrije Universiteit Brussel, 2002.

[43] J.C. Platt, N. Cristianini, and J. Shawe-Taylor. Large margin DAGs for multiclass classification. *Advances in Neural Information Processing Systems*, 12:547–553, 2000.

[44] A.M. Rubinov, A.M. Bagirovand, N.V. Soukhoroukova, and J. Yearwood. Unsupervised and supervised data classification via nonsmooth and global optimization. *TOP*, 11(1):1–93, 2003.

[45] D.K. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics Supplement*, 32:502–508, 2002.

[46] A. Stam. Nontraditional approaches to statistical classification: Some perspectives on l_p-norm methods. *Annals of Operations Research*, 74:1–36, 1997.

[47] V. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, 1995.

[48] V. Vapnik. *Statistical learning theory.* Wiley, 1998.

[49] Weka 3: Data mining software in java. `http://www.cs.waikato.ac.nz/ml/weka/`.

[50] J. Weston and Watkins. Multi-class support vector machines. In *Proceedings of ESANN99, Brussels, D. Facto Press*, 1999.

[51] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques.* Morgan Kaufmann, 2005.

[52] D. Xie, S.B. Singh, E.M. Fluder, and T. Schlick. Principal component analysis combined with truncated-Newton minimization for dimensionality reduction of chemical databases. *Mathematical Programming*, 95(1):161–185, 2003.

[53] C. Zopounidis and M. Doumpos. Multicriteria classification and sorting methods. *European Journal of Operational Research*, 138:229–246, 2002.