REVISTA ESPAÑOLA DE FINANCIACIÓN Y CONTABILIDAD
Vol. XXXII, n.º 118
julio-septiembre 2003
pp. 833-869

María Bonilla
Universidad de Valencia
Ignacio Olmeda
Universidad de Alcalá
Rosa Puertas
Universidad Politécnica de Valencia

MODELOS PARAMÉTRICOS Y NO PARAMÉTRICOS EN PROBLEMAS DE *CREDIT SCORING*

Parametric and non-Parametric models in the credit scoring problems

RESUMEN

DADA la importancia creciente que esta cobrándose la actividad crediticia en la gestión diaria de los bancos, comienza a ser imprescindible la utilización de modelos de clasificación automáticos que faciliten la concesión o no del crédito solicitado con alto grado de exactitud, de manera que permita reducir la morosidad.

En el trabajo que presentamos se realiza un exhaustivo estudio de la capacidad predictiva de dos modelos paramétricos (Análisis Discriminan-

Recibido 30-01-02
Aceptado 02-07-02

ISSN 0210-2412

te y Logit) y cinco no paramétricos (Árboles de regresión, Redes Neuronales Artificiales, Algoritmo C4.5, Splines de Regresión Adaptativa Multivariante y Regresión Localmente Ponderada) en un problema de concesión de tarjetas de crédito.

PALABRAS CLAVE

Clasificación crediticia; Modelos paramétricos, Modelos no paramétricos.

ABSTRACT

Given the growing importance of credit activity in daily bank administration, the use of automatic classifying models to facilitate credit granting or refusal with a high grade of accuracy begins to be indispensable to reduce the rate of defaulters.

The paper we present carries out an exhaustive study of the predictive capacity of two parametric models (Discriminative Analysis and Logit) and five non-parametric models (Regression Trees, Neuronal Artificial Nets, Algorithm C4.5, Multivariate Adaptative Regression Splines and Locally Pondered Regression) for a credit card concession problem.

KEY WORDS

Credit scoring; Parametric models; Non parametric models.

1. INTRODUCCIÓN

El entorno cambiante del sistema financiero ha obligado al mercado crediticio a realizar una profunda transformación de sus estructuras para poder adaptarse a la creciente competencia. La globalización de los mercados es una realidad cierta que debe ser asumida por todos. La entrada de España en la Unión Monetaria Europea, el 1 de enero de 1999, junto con la reducción de los tipos de interés y unido a la implan-

tación del euro, ha intensificado el estrechamiento experimentado por los márgenes financieros con los que operan las entidades financieras españolas.

La necesidad de incrementar la cuota de mercado es una realidad actual que no precisa justificación; cuanto mayor sea el volumen de crédito concedido por una entidad, mayor será su potencial de beneficios, aunque si bien hay que decir que dicha afirmación debe ir unida a un aumento de la calidad de los mismos, pues de cualquier otro modo el resultado sería un deterioro significativo de la cuenta de resultados.

Todo ello justifica la necesidad de que las entidades incorporen calidad a sus créditos, utilizando para ello distintos modelos que faciliten y mejoren el proceso de concesión de los mismos. Entre ellos cabría citar el modelo relacional (basado en el estudio exhaustivo de la información derivada de las relaciones pasadas y presentes que el cliente ha mantenido con la entidad), el modelo económico-financiero (mediante el cual se analiza la estructura financiera de la empresa y su capacidad para generar fondos), y por último, el que constituye el centro de nuestras investigaciones, el *credit scoring*.

Se denomina *credit scoring* a todo sistema de evaluación crediticia que permite valorar de forma automática el riesgo asociado a cada solicitud de crédito. Riesgo que estará en función de la solvencia del deudor, del tipo de crédito, de los plazos, y de otras características propias del cliente y de la operación, que van a definir cada observación, es decir, cada solicitud de crédito. Unicamente, no existirá riesgo en una operación de crédito, o este sería muy reducido, cuando la entidad que los instrumenta actúe como mediadora o intermediaria, o bien cuando el crédito se conceda con la garantía del Estado.

Los créditos de clientes que no se pagan a su vencimiento no sólo generan costes financieros, sino que además producen grandes costes administrativos de gestión, por lo que las entidades financieras están prestando, cada vez más, especial atención a estas partidas que deterioran considerablemente su cuenta de resultados. Así pues, los modelos automáticos de clasificación crediticia pretenden evitar, en la medida de la posible, la concesión de créditos a clientes que posteriormente puedan resultar fallidos, lo que ocasionaría un cuantioso quebranto a la entidad emisora del mismo.

Se trata de un sistema objetivo en el que la aprobación o no del crédito solicitado no va a depender de la discrecionalidad del personal y, además, al ser un sistema automático, no precisa de mucha dedicación de tiempo y personal, permitiendo reducir costes y tiempo de tramitación.

El *credit scoring* constituye, por tanto, un problema de clasificación propiamente dicho, pues dado un conjunto de observaciones cuya pertenencia a una determinada clase es conocida *a priori*, se busca una regla que permita clasificar nuevas observaciones en dos grupos: los que con alta probabilidad podrán hacer frente a sus obligaciones crediticias, y los que, por el contrario, resultarán fallidos. Para ello se tendrá que realizar un análisis de las características personales del solicitante (profesión, edad, patrimonio...) y de las características de la operación (motivo del crédito, porcentaje financiado, ...), que permitirá inducir las reglas que posteriormente se aplicarán a nuevas solicitudes, determinando así su clasificación.

En España, la aparición de estas técnicas de evaluación automática se sitúa alrededor de 1983. Hasta la fecha se venían utilizando métodos tradicionales basados en el análisis del patrimonio y la capacidad de pago, como únicos criterios para juzgar el perfil económico, psicológico y legal del solicitante. Estos presentaban tres graves inconvenientes: el proceso de evaluación era subjetivo, el análisis de la situación se efectuaba secuencialmente, en lugar de permitir un análisis simultáneo de todas las variables y, por último, cabe señalar que el proceso era lento, lo que implicaba un elevado coste y una muy mala imagen.

Dada la coyuntura del sistema bancario español, y la creciente necesidad de poner en marcha un complejo proceso de estrategias que permita una adaptación a los cambios del entorno, resulta indispensable contar con nuevas técnicas que faciliten la correcta toma de decisiones en materia crediticia y, de este modo, se permita entre otras cuestiones: reducir el tiempo de respuesta, disminuir la tasa de morosidad, la posibilidad de una gestión masiva pero segura, y unos costes no financieros mínimos.

Muy recientemente se han comenzado a desarrollar trabajos empíricos dirigidos a la evaluación del riesgo implícito en las operaciones bancarias, y ciertamente están teniendo una gran transcendencia por su directa aplicabilidad en la gestión de créditos [Foglia *et al.*, 98; Bardos, 1998; Varetto, 1998; Liu *et al.*, 1999; Machauer y Weber, 1998; Altman, 1998]. Por ello, y debido a la creciente importancia que esta área de investigación está cobrándose en el día a día del sistema financiero, en el trabajo que presentamos nos hemos propuesto encontrar el modelo de clasificación que presente mayor potencia predictiva, realizando para ello un profundo análisis de distintas técnicas paramétricas y no paramétricas, algunas de ellas, nunca anteriormente aplicadas a este tipo de problemas financieros.

Nos proponemos estudiar la capacidad predictiva de distintos modelos estadísticos aplicados al *credit scoring*, realizando para ello un análisis comparativo, hasta la fecha el más extenso, entre dos modelos paramétricos (AD y logit) y cinco no paramétricos (CART, MARS, C4.5, RLP, y RNA). Con ello se pretende obtener una técnica que determine, con gran exactitud, la conveniencia de conceder o no el crédito solicitado.

El interés del estudio aquí realizado es doble. En primer lugar, a nuestro juicio no existe suficiente evidencia empírica que permita concluir la superioridad o inferioridad relativa de un determinado modelo frente a otros. Con demasiada frecuencia, el investigador está interesado en demostrar las mejoras de un determinado enfoque más que mostrar las ventajas o inconvenientes de cada uno de ellos.

En segundo lugar, pensamos que muchos de los estudios realizados no abordan con suficiente cuidado el dilema del aprendizaje-generalización, es decir, muestran los resultados en una situación particular para la cual el modelo en cuestión ofrece una buena capacidad predictiva. Como es lógico, esta no es la situación real a la que uno se enfrenta, en la que un determinado decisor debe elegir el modelo mas adecuado antes de disponer de las observaciones que empleará para validarlo. En el presente trabajo acometemos tal problema mediante el empleo de la validación cruzada. Finalmente, ampliamos significativamente el conjunto de herramientas empleadas en otros estudios aplicando nuevos modelos que, en nuestro conocimiento, no han sido utilizados en el presente contexto.

La estructura seguida en el desarrollo de nuestro estudio es la siguiente: en la sección segunda se explican los fundamentos teóricos de los problemas de clasificación, en la sección tercera y cuarta se realiza una breve revisión de los modelos paramétricos y no paramétricos utilizados. En la sección quinta presentamos el trabajo empírico realizado, para finalizar en la sección sexta con las principales conclusiones obtenidas y, cerrando el trabajo, las referencias utilizadas en el desarrollo del mismo.

2 FUNDAMENTOS TEÓRICOS DE LOS PROBLEMAS DE CLASIFICACIÓN

En sentido general, un problema de clasificación financiera puede ser visto como un problema de decisión en el que un sujeto, apoyándose en un conjunto de información, asigna cada observación a una categoría determinada, de manera que se minimice el coste de realizar una clasifica-

ción errónea. Se trata de un problema intrínsecamente multivariante en el que pueden diferenciarse, básicamente, dos situaciones:

1) Dado un conjunto de observaciones se pretende determinar la pertenencia de dos o más observaciones a la misma clase, no definida *a priori*. En el contexto estadístico, estos problemas de clasificación reciben el nombre de *problemas de análisis de conglomerados*.
2) Dado un conjunto de observaciones cuya pertenencia a una determinada clase es conocida *a priori*, se trata de encontrar una regla que permita clasificar nuevas observaciones para las cuales se desconoce la clase a que pertenecen. Estos problemas son denominados *problemas de clasificación* propiamente dichos, y son los que analizaremos en el presente trabajo.

Así pues, y basándose en un vector de variables características, los modelos de clasificación tratan de desarrollar reglas que ayuden al sujeto decisor a adoptar una postura ante la cuestión objeto de estudio, de manera que se minimice el coste del error cometido. Normalmente, como apuntan Gnanadesikan y Kettenring [1989], la alta dimensionalidad que presentan estos problemas puede ocasionar dificultades en el desarrollo de un modelo estadístico apropiado, ya que esta metodología reúne exactitud (representada por la proporción de clasificaciones correctas), velocidad en la obtención de resultados, comprensibilidad de los resultados obtenidos, y reducción del tiempo requerido para aprender la regla de clasificación.

La representación de estos problemas consta de tres elementos [Marais *et al.*, 1984]: una *función de pérdida*, que especifica el coste de cada tipo de error cometido en la clasificación; una *distribución de probabilidad conjunta*, correspondiente a las distintas categorías y características que definen la población, y la *regla de clasificación* condicional del sujeto decisor.

Resulta habitual que la función de pérdida y la distribución de probabilidad conjunta no se hallen totalmente especificadas, siendo necesario recurrir a hipótesis y determinar las variables que deberán incluirse en el modelo. La elección de las variables es una etapa difícil del proceso, siendo muy discutida en el desarrollo de los modelos de clasificación de las últimas décadas, y despertando distintas opiniones entre los investigadores (1).

(1) Ver Capon [1982] y Johnson [1989].

La finalidad de estos modelos de clasificación no es otra que la de reproducir la conducta del sujeto decisor, de manera que pueda considerarse apta para predecir en nuevas situaciones. El problema con el que nos enfrentamos, es la insuficiencia de información disponible [Zmijewski y Foster, 1996]. En ocasiones, la base de datos es tan limitada que se utiliza la misma muestra para la especificación del modelo, la estimación de sus parámetros, y el cálculo de las tasas de error, por lo que los resultados obtenidos están sesgados, produciéndose el fenómeno del sobreaprendizaje, que resulta del hecho de que el modelo «memoriza» la información que se le ha facilitado, sin ser capaz de obtener una generalización adecuada. Una forma sencilla de evitarlo consiste en contrastar el modelo con una base de datos distinta de la utilizada en su especificación, pero, como hemos comentado, ello no siempre es posible.

Como decimos, el problema que vamos a analizar en nuestro estudio es un problema de *credit scoring*. Los modelos de *credit scoring* tratan de obtener, a partir del análisis de la relación existente entre las características personales de los solicitantes (profesión, edad, patrimonio, etc.) y las características de la operación (motivo del crédito, porcentaje financiado, garantías aportadas, etc.), una regla general que permita determinar, con rapidez y fiabilidad, la probabilidad de fallido de una determinada solicitud. Por tanto, resulta imprescindible estudiar las relaciones existentes entre la información recogida de cada una de las operaciones concedidas en el pasado y los impagos observados.

Realizado este análisis, y utilizando un sistema de puntuación establecido en función de las características del cliente, se podrá determinar la probabilidad de que éste pueda afrontar sus obligaciones de pago. Así, el problema al que nos enfrentamos puede especificarse mediante la siguiente expresión:

$$P = f(x_1, x_2, \ldots, x_k) + \varepsilon \qquad [2.1]$$

donde x_i serán los atributos del sujeto, ε la perturbación aleatoria, $f(x)$ la función que determina la relación existente entre las variables utilizadas, y P la probabilidad de que el crédito resulte fallido. El objetivo principal de los modelos de clasificación se centra en estimar la función que permita ajustar con la máxima exactitud las observaciones de la muestra, de manera que el error incurrido en la predicción sea mínimo. Dependiendo de que la forma funcional de $f(x)$ sea conocida o desconocida, estaremos ante modelos paramétricos o no paramétricos. El problema que estamos analizando conlleva una decisión no estructurada, ya que no existe ningún patrón estandarizado que establezca qué variables utilizar, a lo que

se añade la dificultad de tener que especificar *a priori* una forma funcional.

Con independencia del enfoque empleado, conviene mencionar el problema de la no aleatoriedad de las muestras en problemas de *credit scoring*. La gran mayoría de los trabajos que se han desarrollado para el tratamiento de este problema han utilizado muestras truncadas, es decir, formadas únicamente por créditos concedidos, ello debido, principalmente, a la imposibilidad de obtener datos sobre los no concedidos. Los procedimientos que trabajen con muestras truncadas darán lugar a estimadores inconsistentes de los parámetros poblacionales *scoring* (2).

A pesar de esta gran limitación, y de las inherentes a cada uno de los modelos que analizaremos a continuación, los modelos estadísticos ofrecen, generalmente, buenos resultados, por lo que estas técnicas estadísticas, tanto paramétricas como no paramétricas, son consideradas herramientas de gran utilidad para la adecuada toma de decisiones en la empresa.

3. MODELOS DE CLASIFICACIÓN PARAMÉTRICOS

Los modelos paramétricos parten de una función de distribución o clasificación conocida, y reducen el problema a estimar los parámetros que mejor ajusten las observaciones de la muestra. Dichos modelos resultan muy potentes cuando el proceso generador de datos sigue la distribución propuesta, aunque pueden llegar a ser muy sensibles frente a la violación de las hipótesis de partida cuando se utilizan muestras de reducido tamaño.

Con objeto de salvar esta y otras limitaciones, se emplean los denominados modelos no paramétricos, conocidos también como métodos de distribución libre pues no se encuentran sujetos a ninguna forma funcional. Dichos modelos, como veremos en la sección siguiente, presentan pocas restricciones, por lo que en ocasiones resultan más fáciles de aplicar que los paramétricos y permiten «reconstruir» la función de clasificación en todo tipo de situaciones, incluidas aquellas en las que la función sea sencilla y conocida (por ejemplo, lineal). Ahora bien, si las variables no son de tipo cualitativo y la distribución de la muestra es normal, se ha comprobado que los métodos no paramétricos resul-

(2) Ver, por ejemplo, Gracia-Díez y Serrano [1992].

tan menos eficientes que aquellos procedimientos paramétricos que presentan como hipótesis de partida la normalidad de las variables. Yatchew [1998] realiza un profundo estudio en el que se analizan las ventajas y desventajas de la utilización de las técnicas de regresión no paramétricas.

La diferencia fundamental entre los modelos paramétricos y no paramétricos es la siguiente. Supongamos que la variable dependiente Y puede ser explicada mediante la expresión: $Y = f(x_1, x_2, \ldots, x_k) + \varepsilon$, donde x_i son las variables explicativas, ε la perturbación aleatoria y $f(x)$ la función que determina la relación existente entre las variables utilizadas. Los modelos paramétricos suponen conocida la forma funcional de $f(x)$ (por ejemplo, lineal, $f(x) = ax + b$), reduciéndose el problema a determinar los parámetros que la definen (a y b, en el caso mencionado). Por su parte, los modelos no paramétricos no tratan de encontrar los parámetros de una función conocida, sino que emplean formas funcionales flexibles que aproximen la función objetivo. Es decir, el problema consiste en calcular los parámetros de una función. Los métodos paramétricos parten de una forma funcional conocida, centrándose el problema en la estimación de los parámetros de los que depende el modelo y que permiten un mejor ajuste de los datos.

En ambos casos, es necesario estimar los parámetros de los que depende la forma funcional elegida. Sin embargo, en el caso de los modelos paramétricos, la elección de dicha forma funcional se establece *a priori*, por lo que una elección inadecuada se traducirá en un modelo que no ajuste los datos (por ejemplo, supuesta una relación lineal entre las variables, dicha función presentará un mal ajuste cuando la respuesta es, por ejemplo, cuadrática).

Dadas las características del *credit scoring*, donde es difícil suponer una relación funcional clara entre las variables del problema, los modelos paramétricos podrían parecer, *a priori*, que no poseen la flexibilidad suficiente para ajustarse a todo tipo de situaciones. Por otra parte, y en lo que respecta a su capacidad predictiva, existen algunos estudios que demuestran su inferioridad frente a los modelos no paramétricos [Tam y Kiang, 1992; Altman *et al.*, 1994]. Ambos aspectos sugieren que el análisis de la calidad predictiva de los modelos paramétricos y no paramétricos resulta relevante en el presente contexto.

De entre todos los métodos paramétricos, hemos escogido el análisis discriminante (AD) y el logit para el desarrollo de un estudio de clasificación crediticia, pues su gran capacidad predictiva en este tipo de problemas ha quedado demostrada en la literatura existente al respecto.

3.1. Análisis discriminante

El análisis discriminante [Fisher, 1936] es una técnica estadística multivariante que permite estudiar de forma simultánea el comportamiento de un conjunto de variables independientes, con objeto de clasificar un colectivo en una serie de grupos previamente determinados y excluyentes. Presenta, pues, la gran ventaja de poder contemplar conjuntamente las características que definen el perfil de cada grupo, así como las distintas interacciones que pudieran existir entre ellas.

Las variables independientes representan las características diferenciadoras de cada individuo, siendo éstas las que permiten realizar la clasificación. Indistintamente se denominan variables clasificadoras, discriminantes, predictivas, o variables explicativas.

De este modo se puede establecer que el objetivo del análisis discriminante es doble:

1. en primer lugar, obtener las mejores combinaciones lineales de variables independientes que maximicen la diferencia entre los grupos. Estas combinaciones lineales reciben el calificativo de *funciones discriminantes*,
2. y posteriormente, predecir, en base a las variables independientes, la pertenencia de un individuo a uno de los grupos establecidos *a priori*. De este modo se evalúa la potencia discriminadora del modelo.

Para el logro de estos objetivos, la muestra de observaciones se divide aleatoriamente en dos submuestras: una primera, conocida *como muestra de entrenamiento*, que se utilizará para la obtención de las funciones discriminantes, y una segunda, denominada *muestra de test*, que servirá para determinar la capacidad predictiva del modelo obtenido.

Por tanto, podemos resumir diciendo que el objetivo del análisis discriminante consiste en encontrar las combinaciones lineales de variables independientes que mejor discriminen los grupos establecidos, de manera que el error cometido sea mínimo. Para ello será necesario maximizar la diferencia entre los grupos (variabilidad entre grupos) y minimizar las diferencias en los grupos (variabilidad intragrupos), obteniendo así el vector de coeficientes de ponderación que haga máxima la discriminación.

Con objeto de asegurar la potencia discriminadora del modelo es necesario establecer fuertes hipótesis de partida que van a suponer una limitación para el análisis de cualquier problema de clasificación que se presente. Éstas son:

1. Las K variables independientes tiene una distribución normal multivariante.
2. Igualdad de la matriz de varianzas-covarianzas de las variables independientes en cada uno de los grupos.
3. El vector de medias, las matrices de covarianzas, las probabilidades *a priori*, y el coste de error son magnitudes todas ellas conocidas.
4. La muestra extraída de la población es una muestra aleatoria.

Tan sólo bajo estas hipótesis la función discriminante obtenida será óptima. Las dos primeras hipótesis (la normalidad y de igualdad de la matriz de varianzas y covarianzas) difícilmente se verifican en muestras de carácter financiero, cuestión que no impide al análisis discriminante obtener buenas estimaciones, aunque realmente éstas no puedan considerarse óptimas.

3.2. El modelo logit

El modelo logit permite calcular la probabilidad de que un individuo pertenezca o no a uno de los grupos establecidos *a priori*. La clasificación se realizará en función del comportamiento de una serie de variables independientes características de cada individuo. Se trata de un modelo de elección binaria en el que la variable dependiente tomará valores 1 ó 0, es decir, en nuestro problema el valor dependerá de que el individuo haya hecho o no frente a sus obligaciones crediticias. Si se presentara una situación en la que el sujeto tuviera que elegir entre tres o más alternativas mutuamente excluyentes (modelos de elección múltiple), tan sólo se tendría que generalizar el proceso.

El modelo logit queda definido por la siguiente función de distribución logística obtenida a partir de la probabilidad *a posteriori* aplicada al AD mediante el teorema de Bayes,

$$P_i = P(Y = 1/X) = F(Z_i) = \frac{1}{1 + e^{-z_i}} = \frac{1}{1 + e^{-(\beta_0 + \beta X_i)}} \qquad [3.2.1]$$

en donde β_0 representa los desplazamientos laterales de la función logística, β es el vector de coeficientes que pondera las variables independientes y del que depende la dispersión de la función y X es la matriz de variables independientes.

Al igual que el modelo discriminante, el logit es un modelo multivariante paramétrico en el que existen variables categóricas tanto en el conjunto de variables explicativas como en de las variables dependien-

tes. Frente al análisis discriminante, presenta la gran ventaja de que no va a ser necesario establecer ninguna hipótesis de partida: no plantea restricciones ni con respecto a la normalidad de la distribución de variables, ni a la igualdad de matrices de varianzas-covarianzas. Ahora bien, cabe señalar que, en caso de verificarse dichas hipótesis, el modelo discriminante obtendría mejores estimadores que el logit, pues según afirma Efron [1975], «... bajo estas circunstancias, los estimadores logísticos resultan bastante menos eficientes que los de la función discriminante».

La mayoría de los problemas financieros con los que nos enfrentamos utilizan alguna variable cualitativa, imposibilitando de este modo el cumplimiento de la hipótesis de normalidad, siendo el modelo logit con los estimadores de máxima verosimilitud claramente preferible. En este sentido, Press y Wilson [1978] enumeran los distintos argumentos existentes en contra de la utilización de los estimadores de la función discriminante, presentando, asimismo, dos problemas de clasificación cuyas variables violan dicha restricción. Ambos problemas se resolvieron mediante el análisis discriminante y el logit quedando claramente demostrada la superioridad de este último.

4. MODELOS DE CLASIFICACIÓN NO PARAMÉTRICOS

El problema de clasificación crediticia que nos proponemos analizar conlleva una decisión no estructurada, ya que no existe ningún patrón estandarizado que establezca qué variables utilizar. Además resultaría difícil suponer una forma funcional establecida *a priori* (como exigen los modelos paramétricos).

Los modelos no paramétricos tratan de aproximar la función de clasificación mediante el empleo de formas funcionales flexibles, sin suponer ninguna estructura funcional *a priori*. Por tanto, tales modelos permiten «reconstruir» la función de clasificación en todo tipo de situaciones, incluidas aquellas en las que la función de clasificación es sencilla (por ejemplo, lineal). Tales modelos son, a diferencia de los paramétricos, de aplicabilidad general.

4.1. Árboles de regresión (CART)

Los árboles de decisión son una técnica no paramétrica de clasificación binaria que reúne las características del modelo clásico univariante

y las propias de los sistemas multivariantes. Permite separar las observaciones que componen la muestra asignándolas a grupos establecidos *a priori*, de forma que se minimice el coste esperado de los errores cometidos. Fue originariamente presentado por Friedman en 1977, pero sus aplicaciones a las finanzas no han sido muy numerosas, si bien cabe destacar dos estudios: el trabajo de Frydman *et al.* [1985] en el que utilizan el modelo para clasificar empresas, comparando su capacidad clasificadora con el clásico análisis discriminante, y el trabajo de Marais *et al.* [1984] que, por el contrario, lo aplican a préstamos. En ambos se ha llegado a demostrar la gran potencia que presenta este modelo como técnica de clasificación.

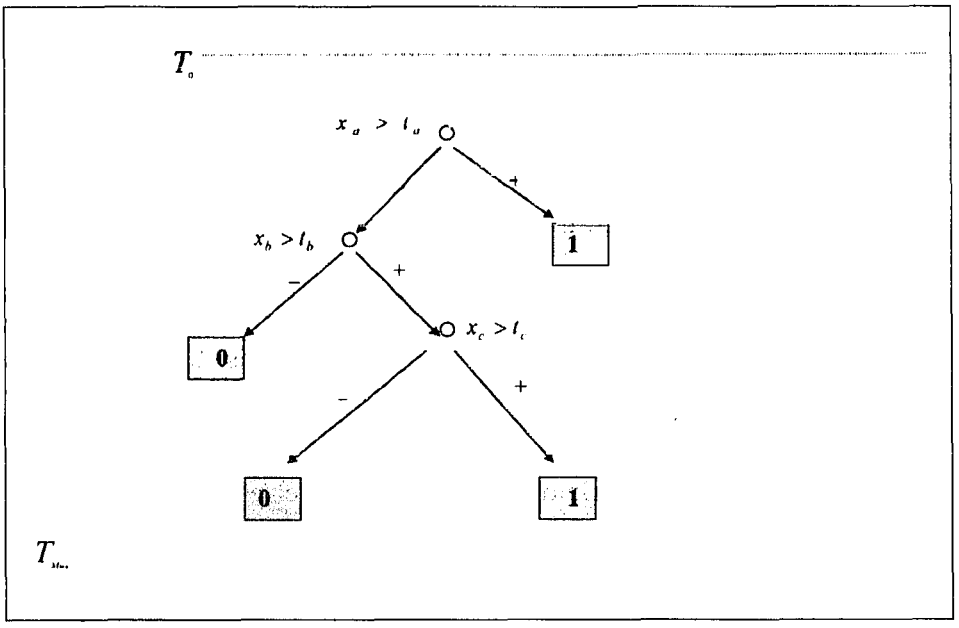
El modelo CART supone esencialmente que las observaciones a clasificar son extraídas de una distribución ϕ en $L \times X$, donde L es el espacio de categorías, y X el espacio de características. Las densidades condicionales $\phi(x|l)$ difieren al variar l, y las probabilidades marginales $\phi(l)$ son conocidas. El proceso utiliza la muestra S como conjunto de entrenamiento para la estimación no paramétrica de una regla de clasificación que permita particionar directamente el espacio X de características. Para cada l de L, el subconjunto S_l del conjunto de entrenamiento S constituye una muestra aleatoria de la distribución condicional $\phi(x|l)$ en X.

Así pues, el proceso consiste en dividir sucesivamente la muestra original en submuestras, sirviéndose para ello de reglas univariantes que buscarán aquella variable independiente que permita discriminar mejor la división. Con ello se pretende obtener grupos compuestos por observaciones que presenten un alto grado de homogeneidad, incluso superior a la existente en el grupo de procedencia (denominado *nodo madre*).

Con objeto de encontrar la mejor regla de división, el algoritmo estudiará cada una de las variables explicativas, analizando puntos de corte para, de este modo, poder elegir aquella que mayor homogeneidad aporte a los nuevos subgrupos. El proceso finaliza cuando resulte imposible realizar una nueva división que mejore la homogeneidad existente.

El modelo, como vemos en la Figura 1, se estructura como un árbol compuesto de una sucesión de nodos y ramas, que constituyen respectivamente los grupos y divisiones que se van realizando de la muestra original. Cada uno de los nodos terminales representa aquel grupo cuyo coste esperado de error sea menor, es decir, aquellos que presenten menor riesgo. El riesgo total del árbol se calcula sumando los correspondientes a cada uno de los nodos terminales.

FIGURA 1
ÁRBOLES DE CLASIFICACIÓN



FUENTE: Breiman *et al.* (1984).

En definitiva, el algoritmo de partición recursiva puede resumirse en los siguientes cuatro pasos:

1. Estudiar todas y cada una de las variables explicativas para determinar para cuál de ellas y para qué valor es posible incrementar la homogeneidad de los subgrupos. Existen diversos criterios para seleccionar la mejor división de cada nodo, todos ellos buscan siempre aquella división que reduzca más la *impureza* del nodo, definida ésta mediante la siguiente expresión,

$$i(t) = -\sum p(j/t) \cdot \log[p(j/t)] \qquad [4.1.1]$$

siendo $p(j/t)$ la proporción de la clase j en el nodo t. Como medida de la homogeneidad o impureza se utiliza una extensión del *índice de Gini* para respuestas categóricas. El algoritmo optará por aquella división que mejore la impureza, mejora que se mide comparando la que presenta el nodo de procedencia con la correspondiente a las dos regiones obtenidas en la partición.

2. El paso anterior se repite hasta que, o bien resulte imposible mejorar la situación realizando otra división, o bien el nodo obtenido presente el tamaño mínimo. En esta fase del algoritmo se obtiene el árbol binario máximo en el cual cada uno de sus nodos interiores es una división binaria del eje de características.

Ahora bien, este procedimiento, tal y como ha sido expuesto, presenta un grave problema: el *sobreaprendizaje*, para evitarlo Friedman [1977] propuso la siguiente solución: desarrollar el árbol al máximo y posteriormente ir «podándolo» eliminando, de este modo, las divisiones y, por tanto, nodos que presenten un mayor coste de complejidad, hasta encontrar el tamaño óptimo, que será aquel que minimice el coste de complejidad.

3. Seguidamente se calcula la complejidad de todos y cada uno de los subárboles podando aquellos que verifiquen la siguiente expresión,

$$R_K(T_1) = min\ R_K(T') \qquad [4.1.2]$$

siendo el coste de complejidad,

$$R_K(T) = \left[R(T) + K \cdot |T|\right] \qquad [4.1.3]$$

donde $R_K(T)$ es el coste de complejidad del árbol T para un determinado valor del parámetro K, $R(T)$ es el riesgo de errar en la clasificación (K se denomina parámetro de complejidad que penaliza la complejidad del árbol y siempre será positivo) y $|T|$ es número de nodos terminales.

4. El cuarto y último paso consiste encontrar todos los valores críticos de K, y utilizar la técnica de validación cruzada para cada uno de ellos con objeto de estimar $R(T(K))$, eligiendo aquella estructura que presente mejor valor estimado de $R(T(K))$.

Por tanto, el principal problema con el que se enfrenta este modelo es la complejidad de su estructura que, como ya hemos indicado, fácilmente puede desembocar en el sobreaprendizaje del modelo. De ahí que no sólo se persiga crear conjuntos homogéneos con bajo riesgo, sino que también se pretenda obtener aquella estructura que presente una complejidad óptima. Bajo este doble objetivo resulta necesario penalizar la excesiva complejidad del árbol.

Una ventaja del modelo CART, así como de otros modelos que generan árboles de decisión (como el C4.5) es que permite interpretar de manera sencilla los resultados obtenidos, de manera que éstos puedan ser analizados posteriormente. Por ejemplo, el árbol representado en la Figura 1 puede ser trasladado de manera inmediata al siguiente conjunto de reglas:

1. «si la variable x_a es mayor que un valor t_a entoces conceder el crédito»;
2. «si la variable t_a no es mayor que un valor t_a, pero la variable x_b es mayor que t_b y la variable x_c es mayor que t_c, entonces conceder el crédito», y
3. «en otro caso rechazar el crédito».

En lo que respecta a la complejidad del modelo CART, ésta es medida por el numero de nodos, por lo que éste es el parámetro a determinar para una correcta identifiación del modelo.

4.2. El algoritmo C4.5

El algoritmo C4.5 [Quinlan, 1993] es un modelo de clasificación basado en el aprendizaje inductivo (3). Se trata de una versión actualizada del algoritmo original ID3 propuesto por Quinlan [1983], cuya aplicabilidad en distintas áreas de conocimiento ha quedado demostrada en numerosos trabajos desarrollados recientemente (4).

Su filosofía es muy similar a la del modelo CART [Breiman *et al.*, 1984]. El C4.5 presenta igualmente una estructura en forma árbol compuesto por *hojas y nodos de decisión* que irán ramificando el conjunto de observaciones en subarboles cada vez más homogéneos. Cada *hoja* representa un nodo terminal en el que no se realiza ningún test, y cuyo objetivo final es intentar que contengan un número significativo de elementos pertenecientes tan sólo a una de las clases establecida *a priori*. Los *nodos de decisión* simbolizan reglas de decisión aplicadas a los atributos que caracterizan cada una de las observaciones que componen la muestra, permitiendo, de este modo, la división y clasificación de las observaciones.

La construcción del algoritmo pasa por dos etapas: el desarrollo de un árbol capaz de clasificar correctamente la muestra de observaciones presentada y la simplificación del mismo que permitirá eliminar el sobreaprendizaje, aumentando de esta forma su capacidad predictiva.

La primera fase consiste en aplicar sucesivos test a las variables independientes que definen el conjunto de observaciones, originando las subsiguientes divisiones de la muestra original. El proceso finaliza cuando

(3) Entendiendo por aprendizaje inductivo «... el proceso de adquisición de conocimiento mediante la extracción de inferencias inductivas sobre hechos proporcionados por un profesor o por el entorno» [Michalski, 1983]. Este decir, consiste en deducir una regla general a partir de las características más relevantes de las observaciones estudiadas.

(4) Ver Hansen *et al.* [1992] y Kattan *et al.* [1993], entre otros.

cada una de las hojas contenga tan sólo elementos de una misma clase, es decir, cuando no sea posible incrementar la homogeneidad de los subgrupos.

Llegado a este punto, y pasando así a la segunda fase, resulta imprescindible simplificar el modelo, pues un modelo sobreparametrizado es incapaz de obtener una generalización adecuada. Para evitarlo, una vez desarrollada su estructura, se utilizará el método de poda descrito en el modelo CART con las peculiaridades que comentaremos a continuación.

La diferencia fundamental entre estos dos algoritmos radica en la regla de partición utilizada; mientras que el modelo CART emplea como criterio de optimalidad el error cuadrático medio o número de clasificaciones incorrectas, el C4.5 utiliza reglas basadas en la *maximización de la ganancia de información* inducida por una determinada partición, es decir, minimiza la entropía de la partición en lugar de minimizar el coste inherente a la clasificación errónea,

$$gan(X) = Info(T) - Info_x(T) \qquad [4.2.1]$$

siendo $Info(T)$ la información necesaria para identificar el grupo de pertenencia de cada observación de la muestra T, e $Info_x(T)$ la información obtenida al realizar una partición utilizando para ello un test x cualquiera aplicado sobre una de las variables independientes de la muestra.

A pesar de que la representación del modelo es idéntica (un árbol de decisión), el algoritmo C4.5 y el CART no conducen necesariamente a los mismos resultados, puesto que los procedimientos de particionamiento y poda son diferentes. Por este motivo, generalmente resulta útil emplear ambos modelos. El parámetro que determina la flexibilidad del C4.5 es el umbral de ganancia de información, por lo que el valor de este último será el parámetro a determinar mediante el procedimiento de validación cruzada, que explicaremos en la sección 4.6.

4.3. Regresión localmente ponderada (RLP)

La *regresión localmente ponderada* analiza los problemas de clasificación ajustando una curva localmente a los datos. Este tipo de ajuste proporciona una estimación de la variable respuesta con menor variabilidad que la respuesta realmente observada, por ello el resultado de este procedimiento se denomina alisado. El alisado multivariante es una simple extensión del método de alisado univariante, introducido por Cleveland [1979].

El *procedimiento de regresión localmente ponderada* permite ajustar una superficie de regresión a los datos a través de un alisado multivariante. De esta forma, la variable dependiente es alisada como una función de las variables independientes de manera móvil, similar a como se calcula una media móvil de una serie temporal. El esquema básico es el siguiente: llamamos y_i (siendo $i = 1, 2, \ldots, n$) a los valores de la variable dependiente y $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ (siendo $i = 1, 2, \ldots, n$) a las n observaciones de las p variables independientes. Suponemos que el proceso generador de datos es $y_i = g(x_i) + \varepsilon_i$. Consideramos que ε_i son variables aleatorias independientes que se distribuyen según una normal con media 0 y varianza constante σ^2. En el esquema básico, también supondríamos que g pertenece a una clase de funciones paramétricas, tales como las polinomiales, en nuestro caso únicamente consideramos que g es una función alisada de variables independientes, puesto que, mediante el ajuste local, podemos estimar una amplia tipología de funciones alisadas, efectivamente, mucho mayor de las que razonablemente podríamos esperar estimar a partir de cualquier clase específica de funciones paramétricas.

La regresión localmente ponderada proporciona una estimación $\hat{g}(x)$ de la superficie de la regresión para cualquier valor de x en el espacio p-dimensional de variables independientes. Dado un determinado valor de q, que es un entero comprendido entre 1 y n ($1 \leq q \leq n$), la estimación de g en x utiliza las q observaciones cuyos valores son los más cercanos a x. Es decir, definimos una vecindad en el espacio de variables independientes, cuyos puntos se ponderan en función de su distancia respecto de x. Los puntos próximos a x poseen pesos elevados, mientras que, por el contrario, a los puntos alejados de x se les asignan ponderaciones bajas. Sobre la base de estos pesos asignados, aplicando mínimos cuadrados ponderados, se ajusta una función lineal o cuadrática de las variables independientes a la variable dependiente. La estimación $\hat{g}(x)$ se corresponde con el valor de esta función ajustada en x. Como este proceso debe repetirse para cada valor de x para el cuál se desea obtener la estimación $\hat{g}(x)$, la regresión localmente ponderada se considera una técnica computacionalmente intensiva.

En definitiva, el método RLP consiste en construir la función alisada $g(x)$ en cada punto del siguiente modo:

1. Tomar un ejemplo x y buscar los q «vecinos» más próximos a dicho punto, constituyendo así una vecindad $N(x)$ (en términos de similitud de los atributos de ambos ejemplos, tomando para ello la distancia euclídea de los atributos). El número de vecinos q se

especifica como un porcentaje, f, de la cantidad total de observaciones n, $f = q/n$.

2. Calcular la distancia máxima entre x y cualquier punto del vecindario.
3. Asignar pesos a cada uno de los puntos de la vecindad $N(x)$ a través de la función de ponderación tri-cúbica. Estos puntos se ponderan en función de su distancia respecto de x. Los puntos próximos a x poseen pesos elevados; por el contrario, a los puntos alejados de x se les asignan ponderaciones bajas.
4. Sobre la base de estos pesos asignados se ajusta, mediante mínimos cuadrados ponderados, una función (lineal o cuadrática) $g(x)$ sobre el vecindario $N(x)$. Con ello obtiene el valor ajustado $\hat{g}(x)$.
5. Repetir este procedimiento para cada valor de la variable predictora para el cual se desea obtener una estimación $\hat{g}(x)$.

Resulta sencillo ver que la RLP no es más que una ponderación no lineal de los ejemplos más parecidos a la observación cuya clasificación tratamos de establecer.

4.4. Splines de regresión adaptativa multivariante (MARS)

Los *splines de regresión adaptativa multivariante*, MARS [Friedman, 1991], consisten en un algoritmo basado en las ideas de particionamiento recursivo [Morgan y Sonquist, 1963] y regresión multietapa que emplea funciones tipo spline para lograr la aproximación a una función de regresión arbitraria. El procedimiento consiste en particionar el dominio de definición de la función en diferentes regiones, ajustando en cada una de ellas una función tipo spline.

Un *spline* cúbico univariante con umbrales $k_1, k_2, \ldots, k_s$ es un polinomio cúbico definido sobre los *intervalos* $(-\infty, k_1)$, (k_1, k_2), ..., (k_s, ∞), cuya segunda derivada es continua en todos los puntos. Fijados los umbrales, las funciones $1, x, x^2, x^3, (x-k_1)^3+, \ldots, (x-k_s)^3+$, donde $(x-k_i)^3+$ es la parte positiva de $(x-k_i)^3$, constituyen una base del espacio vectorial de todas las funciones spline cúbicas, a estas funciones se les denomina *funciones base.*

Los splines cúbicos multivariantes también forman un espacio vectorial. En n dimensiones, cada función base es el producto de n funciones base univariantes, una para cada coordenada, es decir, un spline multivariante tiene la forma,

$$B(x) = B(x_1, x_2, \ldots, x_n) = \prod_{v=1}^{n} B_v(x_v) \qquad [4.4.1]$$

donde B_v es una función base univariante para la v-ésima coordenada. Un spline multivariante incluye, por tanto, todas las posibles interacciones que se producen entre las variables introduciendo productos cruzados de las funciones spline univariantes.

Dado el modelo de regresión:

$$y_i = g(x_{i1}, x_{i2}, \ldots, x_{in}) + \varepsilon_i \qquad j = 1, 2, \ldots, N \qquad [4.4.2]$$

donde n es el número de atributos, N es el número de casos, g es la función de clasificación desconocida y ε_i un término de error con media cero. El procedimiento empleado por el modelo MARS para estimar g puede resumirse en tres etapas.

En primer lugar, se emplea un algoritmo «hacia adelante» para seleccionar las funciones base y los puntos de truncado. Seguidamente, se emplea un algoritmo «hacia atrás» para eliminar funciones base, hasta que el mejor conjunto de las mismas es encontrado, el propósito de esta fase consiste en reducir el grado de complejidad del modelo, aumentando su capacidad de generalización. Finalmente es utilizado un suavizado que proporciona a la aproximación obtenida el grado de continuidad deseable en las fronteras de las particiones. Se trata de un procedimiento adaptativo en el sentido de que la selección de las funciones base es guiada por los datos y, por tanto, específica al problema en cuestión. Puesto que sólo un número reducido de funciones base son empleadas, este procedimiento permite reducir significativamente la alta dimensionalidad inherente a algunos problemas.

El procedimiento hacia adelante comienza con una función constante, $B_0(x) = 1$. Seguidamente se van añadiendo funciones univariantes en la forma de funciones lineales truncadas $(x\text{-}k)_+$, $(x\text{-}k)_-$. Las funciones base que recogen las interacciones son creadas multiplicando una función base existente por una función lineal truncada que involucre otra variable. Tanto la función base originaria como la nueva función base creadas son empleadas (a diferencia de otros algoritmos de particionamiento recursivo, que utilizan sólo la nueva función obtenida a partir de la anterior). Una vez que M-1 funciones base han sido elegidas, la m-ésima función base es seleccionada de entre las que pueden crearse a partir de las M-1 anteriores de manera que minimice un criterio de pérdida cuadrática:

$$P(\hat{g}_M) = \frac{1}{N} \sum_{j=1}^{N} [y_i - \hat{g}_M(x_i)] \qquad [4.4.3]$$

El algoritmo se detiene cuando la aproximación construida incluye un número máximo de funciones fijadas por el usuario.

La aproximación final proporcionada por MARS tiene la forma,

$$\hat{g}_M(x) = a_0 + \sum_{m=1}^{M} a_m \left\{ \prod_{l=1}^{L_m} [s_{l,m}(x_{v(l,m)} - k_{l,m})]_+ \right\} \qquad [4.4.4]$$

donde la fórmula entre corchetes es la m-ésima función base, M es el número de funciones base linealmente independientes, L_m es el número de divisiones o funciones lineales truncadas multiplicadas en la m-ésima función base, a_m es el coeficiente de la m-ésima función base, $x_{v(l,m)}$ es la variable predictora correspondiente a la l-ésima función lineal truncada en la m-ésima función base, $k_{l,m}$ es el umbral correspondiente a $x_{v(l,m)}$ y $s_{l,m}$ es +1 o –1. Es habitual restringir el grado máximo de interacción a tres, $L_m = 3$, mientras que el número de umbrales es automáticamente seleccionado por MARS, dividiendo el rango de definición de cada una de las variables en subintervalos.

En el procedimiento hacia atrás, MARS emplea una versión modificada del criterior de validación cruzada generalizada de Craven y Wahba [1979], dado por la expresión:

$$CVG(M) = \frac{\frac{1}{N}\sum_{i=1}^{N} [y_i - \hat{f}_M(x_i)]^2}{[1 - [C(M)^*]/N]^2} \qquad [4.4.5]$$

donde $C(M)$ es un término que penaliza la complejidad del modelo, por ejemplo, medida por el número de funciones base (es decir, $C(M) = M$). Según este criterio, un modelo más sencillo puede ser preferido frente a otro más complejo aunque aquél muestre un peor ajuste, en el caso de que el número de parámetros sea significativamente menor.

Finalmente, y puesto que la aproximación mediante splines lineales produce discontinuidades en las derivadas en las fronteras de los subintervalos (k_s, $k_s + 1$), MARS emplea un suavizado consistente en reemplazar las funciones base truncadas lineales por funciones cúbicas que coinciden con la aproximación lineal por tramos en los extremos de los subintervalos. Este suavizado asegura la continuidad de la primera y segunda derivada de la aproximación en todos los puntos excepto en los extremos del intervalo.

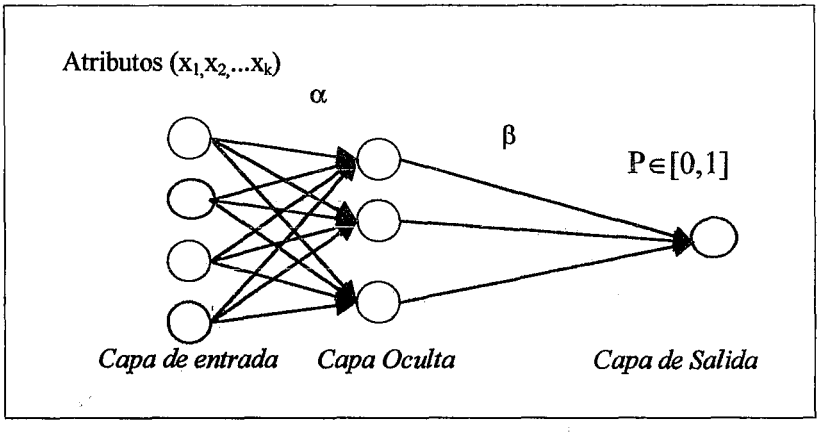
La flexibilidad del modelo MARS queda determinada por dos parámetros: el número de funciones base y el grado de interacción entre las variables. La determinación del modelo optimo consistirá, por tanto, en establecer cual de las combinaciones de ambos parámetros proporciona menores errores de validación cruzada.

4.5. Redes neuronales artificiales

Las *redes neuronales artificiales* (RNA) tratan de emular el sistema nervioso, de forma que son capaces de reproducir algunas de las principales tareas que desarrolla el cerebro humano, al reflejar las características fundamentales de comportamiento del mismo. Lo que realmente intentan modelizar las redes neuronales es una de las estructuras fisiológicas de soporte del cerebro, la neurona y los grupos estructurados e interconectados de varias de ellas, conocidos como redes de neuronas. De este modo, construyen sistemas que presentan un cierto grado de inteligencia.

Una RNA está formada, como vemos en la Figura 2, por un conjunto de procesadores simples altamente interconectados denominados *nodos* o *neuronas*, los cuales se organizan en capas que permiten el procesamiento de información.

Figura 2
REDES NEURONALES ARTIFICIALES



Los nodos o elementos de proceso operan a modo de procesadores simples cuya finalidad consiste en dar respuesta a una determinada señal de entrada. Cada nodo, al igual que ocurre en una neurona biológica, recibe de otros nodos vecinos múltiples entradas que transformará, mediante sencillos cálculos internos, en un sólo valor de salida siendo éste enviado al resto de nodos, y constituyendo así la entrada de éstos. Las conexiones de entrada llevan asociadas un peso que determina cuantitativamente el efecto que producen unos elementos sobre otros. Por tanto, la señal de entrada total a cada una de las q neuronas de la capa inter-

media se calculará sumando los valores de entrada ponderados por sus pesos correspondientes,

$$y_i = \sum_{i=1}^{k} x_i w_{ji} \qquad [4.5.1]$$

donde x_i representa el valor del atributo $i, (i = 1, 2, \ldots, k)$, w_{ji} las ponderaciones asociadas a la conexión entre la neurona de entrada i y la intermedia j, e y_i la señal total de entrada a la neurona j. Posteriormente, a dicha entrada se le aplica una función denominada *función de activación* (en nuestro caso ha sido la sigmoidal), obteniendo de esta forma el valor de salida de cada nodo intermedio, $F(y_j)$, que, a su vez, será transmitido a la neurona de salida a través de la conexión ponderada correspondiente. Así pues, la solución de la red vendrá dada por la siguiente expresión,

$$y = \sum_{j=1}^{q} \beta_i F(y_j) \qquad [4.5.2]$$

donde y es la salida de la red y β_i son los asociados a las conexiones entre la capa intermedia y la de salida, y F la función de activación.

Al igual que en el cerebro biológico, en una RNA el «conocimiento» se encuentra almacenado en los pesos, de ahí que el aprendizaje sea el proceso por el cual la red, a partir de una serie de patrones-ejemplo, modifica sus pesos hasta obtener una regla general que le permita realizar correctamente una tarea determinada.

El entrenamiento o aprendizaje permite a la red autoadaptarse, es decir, durante tal proceso, los nodos de las capas intermedias aprenden a reconocer la relación existente entre un conjunto total de entradas dadas como ejemplo y sus salidas correspondientes. El procedimiento de entrenamiento puede ser visto como un problema de minimización multimodal, al que son aplicables una diversidad de algoritmos. De todos ellos, sin duda, el procedimiento más empleado (y que nosotros también utilizaremos) es el algoritmo de retropropagación de errores, basado en el algoritmo de aproximación estocástica de Robbins-Monro [1951]. Finalizado el entrenamiento, la red habrá encontrado una representación interna que le permita, cuando se le presente una nueva entrada (aunque ésta presente ruido o esté incompleta), proporcionar una salida, siempre y cuando la nueva entrada sea parecida a la presentada durante el aprendizaje.

Las principales características que hacen atractiva la utilización de las RNA en el tratamiento de los problemas de clasificación financiera son su capacidad de generalización, es decir, de aprender a partir de observa-

ciones reales, y la tolerancia a fallos, debido a que el conocimiento se halla distribuido entre los pesos de las conexiones.

Como principal desventaja, hay que señalar que las RNA adolecen del mismo problema de sobreaprendizaje, común a los modelos no paramétricos. En este caso, el problema es todavía más grave debido a que la carga computacional para el cálculo de los valores de los parámetros es muy superior a la de los otros modelos. Por otra parte, puesto que se trata de un problema de optimización multimodal, es previsible que los algoritmos basados en el descenso por gradiente (como el empleado en la retropropagación de errores) produzcan soluciones locales pero no globalmente óptimas. Con objeto de evitar este problema, para cada una de las configuraciones empleadas (cada una de ellas con diferente número de neuronas, que determinan el grado de flexibilidad) realizamos cinco simulaciones, agregando las predicciones obtenidas.

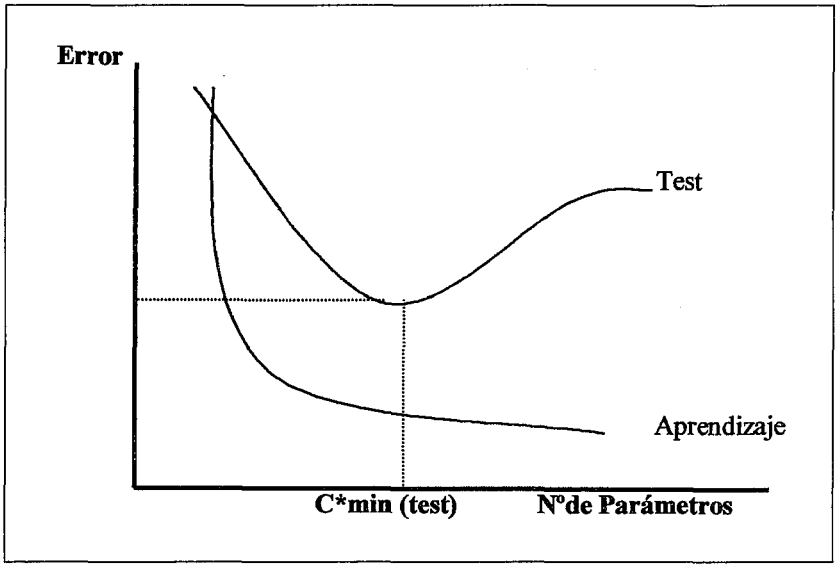
4.6. Problema del sobreaprendizaje: método de validacion cruzada

Un problema que, como hemos podido comprobar en nuestra exposición anterior, resulta común en todos los métodos no paramétricos, es el «sobreaprendizaje»: el modelo memoriza las observaciones de la muestra siendo incapaz de extraer las características más importantes, lo que le impedirá «generalizar adecuadamente», clasificando incorrectamente los casos no contemplados con anterioridad.

La aparición de este fenómeno puede atribuirse fundamentalmente a dos causas. En primer lugar a la sobreparametrización, el modelo presenta una estructura más compleja de la necesaria para tratar el problema en cuestión. En segundo lugar a la escasez de datos que impide al modelo extraer en la fase de entrenamiento las características más relevantes de la muestra, y posteriormente, en la fase de test, verificar la capacidad predictiva del modelo con otra muestra de datos distinta a la utilizada en el entrenamiento. En la Figura 3 se presentan, gráficamente, ambos fenómenos.

Supongamos que disponemos de un conjunto de observaciones y lo dividimos en dos: un conjunto «de entrenamiento», que servirá para ajustar el modelo, y un conjunto «de test» que será empleado para validarlo. En el eje de abscisas hemos representado el número de parámetros de un determinado modelo (siendo el modelo más complejo, es decir, el de mayor número de parámetros, el más alejado del origen), y en el eje de ordenadas el error cometido sobre los conjuntos de aprendizaje y test.

FIGURA 3
PROCESO DE SOBREAPRENDIZAJE DE LOS MODELOS



Cuando la estructura del modelo es muy simple, éste es incapaz de capturar la relación subyacente entre los atributos y la variable respuesta, por lo que cometerá un elevado porcentaje de fallos tanto sobre el conjunto de entrenamiento como sobre el de test. A medida que el número de parámetros aumenta, va adquiriendo suficiente potencia o flexibilidad, lo que le permitirá «aprender» la relación existente entre las variables independientes y dependiente, relación que debe verificarse sobre ambos conjuntos, por lo que el error cometido irá decreciendo.

Si incrementamos sucesivamente la complejidad, el error a lo largo del conjunto de entrenamiento seguirá disminuyendo progresivamente, es decir, el modelo se irá acomodando a las características peculiares de los ejemplos propios de dicho conjunto que no tienen porque estar presentes en el de test. Por este motivo, llegados a un punto, C^*, el error incurrido sobre el conjunto de test, que es el que determina la potencia predictiva del modelo, se incrementará considerablemente. Por consiguiente, a partir de C^* la estructura es tan compleja que el modelo ha «memorizado» la muestra, lo que se traduce en una débil capacidad de generalización.

Con objeto de evitar este gran problema se viene utilizando, entre otros (5), el método de *validación cruzada* propuesto por Stone [1974] que, como veremos cuando presentemos los resultados obtenidos por los distintos modelos, emplearemos para elegir la estructura idónea de cada uno de los modelos no paramétricos, es decir, aquella que facilite la obtención de una adecuada generalización del problema que estamos analizando.

El proceso es el siguiente: el conjunto de entrenamiento se divide aleatoriamente en diez particiones distintas, de manera que cada una de ellas conserve la misma proporción de fallidos y no fallidos existente en el conjunto total. Seguidamente, por rotación, un conjunto de nueve particiones se utiliza para estimar el modelo con un número de parámetros determinado, y la décima partición para contrastar su capacidad predictiva. Este proceso se repite diez veces, de forma que cada modelo, utilizando distintas estructuras, va a ser entrenado y testeado con diez pares distintos de conjuntos de entrenamiento y test, siendo la estructura óptima aquélla que minimice el error de predicción a lo largo de los diez conjuntos de test (este error se denomina error de validación cruzada, EVC).

Puesto que el EVC es un estimador insesgado del error de predicción [Stone, 1974], el modelo seleccionado tendrá también una capacidad de generalización óptima cuando sea empleado con observaciones no presentes en el conjunto de entrenamiento. Por último, elegida la estructura óptima, C^*, se utilizará toda la muestra para reentrenar el modelo, de manera que se entrenará y testeará con los conjuntos totales para obtener el error de predicción.

En concreto, en nuestro análisis disponemos de una muestra de 690 observaciones, de las cuales 90 hemos reservado, como conjunto de test, para testear la capacidad generalizadora del modelo (error de predicción), y las otras 600 observaciones se han utilizado como conjunto de entrenamiento para elegir aquel modelo cuya estructura presente el menor EVC.

Las 600 observaciones destinadas al entrenamiento se han dividido a su vez en conjuntos de entrenamiento y test, representando el conjunto de test el 10% de la muestra (60 observaciones). Con objeto de obtener el EVC, las observaciones de estas dos submuestras se han combinado de tal forma que disponemos de 10 pares no solapados de conjuntos de entrenamiento y test formados por 540 y 60 observaciones respectivamente.

(5) Existen diferentes alternativas; ver, por ejemplo, Olmeda [1993].

Cada uno de los modelos no paramétricos (CART, C4.5, RLP, MARS y RNA) ha sido entrenado y testeado con estos diez pares conjuntos utilizando distintas estructuras, para, de este modo, poder determinar la estructura óptima de cada uno de ellos, que no será otra que aquella que presente el menor EVC, calculado éste como una media de los errores cometidos a lo largo de los diez conjuntos de test validados.

Los parámetros que determinan la complejidad de cada uno de los modelos son los siguientes: para el modelo CART el número de nodos, para las RNA el número de neuronas y ciclos de entrenamiento, para el C4.5 el umbral de ganancia de información, para la RLP el número de vecinos próximos, y para el modelo MARS el número de funciones base así como el grado de interacción de las variables. La selección de la estructura óptima de cada uno de estos modelos consiste, por tanto, en determinar el valor óptimo de tales parámetros.

Para finalizar, cada modelo elegido será entrenado y testeado con la muestra total (600 observaciones de entrenamiento y 90 de test) con objeto de obtener el error de predicción que nos permitirá comparar la potencia predictiva de los distintos modelos.

5. ANÁLISIS DE LOS RESULTADOS (6)

La base de datos utilizada ha sido obtenida a partir de un estudio previo realizado por Quilan (1987), está formada por 690 observaciones sobre 14 características crediticias de individuos demandantes de una tarjeta de crédito, así como de los respectivos comportamientos posteriores a la concesión. Todas las variables se hayan codificadas con objeto de garantizar la confidencialidad, por lo que resulta imposible valorar los costes relativos de error, así como establecer probabilidades *a priori* sobre los clientes. El 55% de las observaciones corresponde a individuos fallidos y el 45% representa a los no fallidos, denotando con «1» cuando el individuo al que se le concedió la tarjeta tuvo la solvencia esperada por el prestamista, y «0» en caso contrario.

Previo al desarrollo de los modelos, y mediante el contraste de Kolmogorov-Smirnov, rechazamos la hipótesis de normalidad de las variables para un nivel de significatividad del 5%, cuestión que era de esperar pues se trata de variables que definen características muy dispares de los individuos, siendo algunas de ellas cualitativas, Tabla 1.

(6) Todas las figuras y tablas de resultados se adjuntan en el Anexo.

La no normalidad de las variables tiene una importancia no trivial en el análisis discriminante. Sin embargo, la violación de esta hipótesis de partida ha resultado frecuente en numerosos trabajos en los que se ha utilizado esta técnica, despertando diversas opiniones al respecto; existen autores que aconsejan ignorar su incumplimiento, mientras que otros opinan que se debería transformar el modelo lineal en cuadrático. Considerando los malos resultados que se obtienen, generalmente, con el empleo de funciones discriminantes cuadráticas (7), hemos optado por la primera opción señalada.

Asimismo, calculando el coeficiente de correlación entre las variables hemos podido desechar la existencia de multicolinealiedad, que de existir podría tener importantes efectos en los resultados de cualquier proceso de regresión, porque limita el tamaño del coeficiente de determinación y dificulta la estimación de la contribución de cada variable independiente.

Como ya hemos indicado, el conjunto total se ha dividido en dos subconjuntos: el conjunto de entrenamiento, sobre el que se estiman los modelos, formado por 600 observaciones, y el conjunto de test, que emplearemos para determinar la capacidad clasificatoria de los modelos, constituido por las 90 observaciones restantes. Para determinar la estructura óptima en los modelos no paramétricos, utilizamos el procedimiento de validación cruzada descrito en la sección anterior.

Empezaremos analizando los resultados de los modelos paramétricos. En la Tabla 2 podemos comprobar que el AD y el logit obtienen exactamente el mismo error de predicción sobre el conjunto de test, 12,22%, mientras que sobre el conjunto de entrenamiento el logit supera al AD. Por lo que aunque, globalmente, el modelo logit resulta ser superior dentro de la muestra (12,50 frente a 14,17% de errores), en términos predictivos, la potencia de ambos modelos resulta ser idéntida (12,22%).

Podemos concluir que aun a pesar de que las variables que definen la muestra no verifican todas las hipótesis necesarias para la aplicación del AD, este modelo ha resultado ser tan potente, en términos predictivos, como el modelo logit en el problema de clasificación que estamos analizando.

A continuación pasamos a analizar los resultados obtenidos mediante los modelos no paramétricos. De todos ellos presentaremos la tabla de errores de validación cruzada, así como su representación gráfica. En cada una de las figuras siguientes situamos en el eje de abcisas un índice que representa una determinada parametrización de cada uno de los mo-

(7) Ver Wagner *et al.* [1983].

delos (creciente, cuanto más alejada del origen), y en el de ordenadas el porcentaje de error medio cometido a lo largo de los diez conjuntos de test del procedimiento de validación cruzada.

Siguiendo el mismo orden que en la sección 4, comenzaremos con el modelo CART. En primer lugar hemos obtenido el error de validación cruzada para el árbol de clasificación sin realizar ninguna poda (Tabla 3), y comprobamos que existe un alto grado de sobreaprendizaje. A lo largo de los diez conjuntos de entrenamiento el árbol presenta una estructura tan compleja que ha sido capaz de memorizar las observaciones del conjunto de aprendizaje, sin poder destacar sus características principales, puesto que ante nuevas observaciones (conjunto de test) el error medio de test es muy superior al del aprendizaje (del 6,66 al 17,33%).

Para eliminar esta sobreparametrización evidente simplificamos la estructura del modelo «podando» aquellas ramas que presenten un mayor coste de complejidad. El parámetro de poda lo representaremos por k, de manera que un mayor k implica una menor complejidad en la estructura final del árbol. Utilizando distintos valores de k, se comprueba que: para valores pequeños no se elimina el sobreaprendizaje, para k superior a 35 la poda resulta excesiva, y para $0{,}15 \leq k \leq 0{,}35$ el error de validación cruzada se mantiene constante. El árbol de estructura óptima es aquel que presenta una menor complejidad que, como podemos comprobar en la Figura 4, se obtiene para $k = 35$.

Para la estructura óptima determinada por el procedimiento de validación cruzada, el modelo CART tiene exactamente el mismo porcentaje (14,66%) de error medio sobre los diez conjuntos de entrenamiento (de 540 observaciones) que sobre los diez de test (de 60 observaciones). Si reestimamos ahora tal estructura óptima sobre el conjunto de entrenamiento total (600 observaciones) comprobamos que el error de entrenamiento es el mismo que antes (14,66%) mientras que el cometido sobre el conjunto de test (90 observaciones) es inferior al esperado (13,33%).

Para el modelo C4.5 (Tabla 4 y Figura 5) podemos observar que para valores $c < 0{,}1$ (donde c es el parámetro que mide la complejidad del modelo) existe cierto grado de sobreaprendizaje, alcanzando el óptimo para un valor de $c = 0{,}2$ (aunque el EVC es similar al obtenido con $c = 0{,}1$ el principio de parsimonia aconseja elegir siempre aquel cuya estructura sea más sencilla). Por tanto, para el modelo C4.5 con $c = 0{,}2$ no existe sobreaprendizaje y el EVC es mínimo. Para este modelo, el error de entrenamiento sobre el total de observaciones es del 14,7%, muy similar al esperado (14,1%), mientras que el error sobre el conjunto de test es del 13,3%, sensiblemente inferior al esperado (14,8%).

La construcción de modelos de RLP presentó un problema adicional respecto al resto de los modelos: el programa estadístico utilizado no admite más de ocho variables independientes, por lo que, previo a desarrollar el modelo, fue necesario realizar un proceso de selección de variables con objeto de determinar las que van a utilizarse. El criterio seguido fue la selección paso a paso (*stepwise*) para el logit.

Este procedimiento comienza con una ecuación que no contiene variables de predicción, y en cada paso entrará o saldrá aquella variable que produzca una mayor reducción en el valor de la suma de los cuadrados de los errores, pudiendo, asimismo, eliminar una variable cuya inclusión se llevó a cabo en una etapa anterior. Para evaluar y comparar los distintos modelos utilizamos el estadístico Cp de Mallows (8).

Siguiendo este criterio las variables elegidas han sido las correspondientes a los números: 5, 6, 9, 11, 14 y 15, utilizando, por tanto, éstas en el desarrollo del RLP. Además, con objeto de elegir la estructura óptima del modelo y evitar problemas de sobreparametrización se han calculado los EVC para distintos tamaños de vecindario. En concreto, primero sin establecer la proporción del mismo, y posteriormente para f = 0,2, 0,4, 0,6, 0,8 y 1. En la Figura 6 se facilitan los resultados del EVC. Como puede comprobase, el EVC mínimo se alcanza para un vecindario de f = 0,8 (el 13,0%), obteniéndose un error de predicción, para la estructura óptima, del 12,3%.

Los resultados obtenidos con el modelo MARS son mostrados en la Figura 7 y Tabla 6. Realizamos los análisis considerando de 4 a 24 funciones base, y un grado de interacción igual a 2 y 3 (en todos los casos los modelos con un grado de interacción igual a 2 resultaron superiores, por lo que la tabla se refiere a esta estructura). Nuevamente, para la estructura óptima, el error de entrenamiento medio para los conjuntos de validación cruzada es muy similar al error cometido sobre el conjunto de entrenamiento total (el 11,52 y 11,17%, respectivamente), sin embargo, el error sobre el conjunto de test difiere del esperado (12,22 y 13,33%, respectivamente).

Por último, pasamos a analizar las Redes Neuronales. Como mencionamos, a diferencia de los otros modelos no paramétricos, las RNA presentan problemas adicionales relacionados con el algoritmo de optimiza-

(8) $C_p = p + (n - p) \cdot \left[\frac{\hat{S}_R^2(p) - \hat{S}_R^2(k+1)}{\hat{S}_R^2(k+1)} \right]$, donde $\hat{S}_R^2(k+1)$ es la varianza residual del modelo con k variable; $\hat{S}_R^2(p)$ es la del modelo con $p - 1$ variables y p parámetros, y n es el número total de datos.

ción empleado: dos redes idénticas entrenadas a partir de valores iniciales de los parámetros distintos pueden producir resultados diferentes. Por este motivo, realizamos cinco simulaciones con cada conjunto de entrenamiento y sus correspondientes conjuntos de test. Finalmente, elegida la estructura óptima, con el propósito de reducir la varianza de las predicciones, re-entrenamos un conjunto de 25 redes idénticas y agregamos las predicciones de las mismas.

En la Tabla 7 mostramos los EVC que se han obtenido dependiendo del número de neuronas utilizado en la capa oculta, los cuales representamos en la Figura 7. Como podemos observar, tanto en la Figura 7 y como en la Tabla 7, la estructura óptima consiste en una red con 10 nodos intermedios. Para esta estructura, el error medio sobre los conjuntos de entrenamiento de validación cruzada es muy similar al del conjunto de entrenamiento completo (12,2 y 12,33%, respectivamente), mientras que los errores de test difieren, de nuevo, de forma sensible (13,33 y 10%, respectivamente). Como conclusión, nuestros resultados sugieren que el EVC sobreestima el error de predicción, siendo este efecto más grave en unos modelos que en otros.

En lo que respecta a la comparación entre modelos, presentamos en la Tabla 8 los errores de validación cruzada correspondientes a la estructura óptima de cada modelo, así como su correspondiente error de predicción. Las diferencias en los errores de validación cruzada no parecen ser muy significativas entre los modelos, salvo en el caso de los modelos CART y C4.5, que parecen resultar inferiores en términos predictivos (ambos con una estructura de árboles de regresión). Esta situación se mantiene para los errores de predicción real: CART y C4.5 resultan ser nuevamente los que obtienen peores resultados, AD, Logit, RLP y MARS resultan comparables, mientras que las RNA ofrecen ventajas aparentemente significativas.

6. CONCLUSIONES

En este trabajo hemos realizado un análisis comparativo, en términos de su capacidad predictiva, de distintos modelos estadísticos en un problema de concesión de tarjetas de crédito. Como principal conclusión, los modelos no paramétricos no dominan de forma sistemática a los paramétricos, lo que contradice en cierta medida algunos resultados de la literatura.

Adicionalmente, hemos comprobado que el procedimiento más generalmente utilizado para la identificación de modelos no paramétricos, el

procedimiento de validación cruzada, aunque nos ha permitido obtener la estructura óptima, no ha resultado adecuado en el problema en cuestión. En todas las ocasiones el error esperado ha sido superior al real, lo que induce a pensar que (salvo que este sesgo sea sistemático) es necesario desarrollar procedimientos más exactos.

Las RNA han resultado ser el modelo de mayor capacidad predictiva, superando a todos los demás modelos. Hemos de señalar, sin embargo, que la escasez de datos dificulta severamente una adecuada comparación entre los modelos, por lo que no es posible asegurar definitivamente si esta aparente mejora es o no estadísticamente significativa.

Finalmente, en lo que respecta al proceso de toma de decisiones, es posible que un método que combine las predicciones de los modelos individuales podría resultar más adecuado en el problema que estamos analizando [Olmeda y Fernández, 1997; Kumar y Olmeda, 1999].

REFERENCIAS

ALTMAN, E. [1998]: «The Importance and Subtlety of Credit Rating Migration», *Journal of Banking and Finance,* vol. 22: 1231-1247.

ALTMAN, E.; MARCO, G., y VARETTO, F. [1994]: «Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks», *Journal of Banking and Finance,* vol. 18: 505-529.

BARDOS, M. [1998]: «Detecting the Risk of Company Failure at the Banque de France», *Journal of Banking and Finance,* vol. 22: 1405-1419.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R., y STONE, C. [1984]: *Classification and Regresion Trees,* Wadsworth & Brooks.

CLEVELAND, W. S. [1979]: «Robust Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting», *Journal of the American Statistical Association,* vol. 83: pp. 596-610.

CRAVEN, P., y WAHBA, G. [1979]: «Smoothing Noisy Data with Spline Functions. Estimating the Correct Degree of Smoothing by the Method of Generalized Cross - Validation», *Numerische Mathematik,* vol. 31: 317-403.

EFRON, B. [1975]: «The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis», *Journal of the American Statistical Association,* vol. 70: 892-898.

FISHER, R. A. [1936]: «The Use of Multiple Measurements in Taxonomic Problems», *Annals of Eugenics,* vol. 7: 179-188.

FOGLIA, A.; LAVIOLA, S., y MARULLO REEDTZ, P. [1998]: «Multiple Banking Relations and the Fragility of Corporate Borrowers», *Journal of Banking and Finance,* vol. 22: 1441-1456.

FRIEDMAN, J. H. [1977]: «A Recursive Partitioning Decisión Rule for Nonparametric Classification», *IEEE Transactions on Computers*: 404-509.

— [1991]: «Multivariate Adaptive Regression Splines (with discussion)», *The Annals of Statistics*, vol. 19: pp. 1-141.

FRYDMAN, H.; ALTMAN, E., y KAO, D. [1985]: «Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress», *The Journal of Finance*: 269-291.

GNANADESIKAN, R., y KETTENRING, J. R. [1989]: «Discriminant Analysis and Clustering», *Statistical Science*: 34-69.

GRACIA-DÍEZ, M., y SERRANO, G. [1992]: «Algunos aspectos sobre el análisis empírico de *credit scoring*», *Estadística Española*, vol. 34: 261-283.

HANSEN, J.; MCDONALD, J., y STICE, J. [1992]: «Artificial Intelligence and Generalized Qualitative-Response Models: An Empirical Test on Two Audit Decision-Making Domains», *Decision Sciences*, vol. 23: 708-723.

KATTAN, M.; ADAMS, D., y PARKS, M. [1993]: «A Comparison of Machine Learning with Human Judgement», *Journal of Management Information Systems*, vol. 9: 37-57.

KUMAR, A., y OLMEDA, I. [1999]: «A Study of Composite or Hybrid Classifiers for Knowledge Discovery», *INFORMS Journal of Computing*, vol. 11: 267-277.

LIU, P.; SEYYED, F., y SMITH, S. [1999]: «The Independent Impact of Credit Rating Changes-The Case of Moody's Rating Refinement of Yield Premiums», *Journal of Business Finance & Accounting*, 337-465.

MACHAUER, A., y WEBER, M. [1998]: «Bank Behavior Based on Internal Credit Ratings of Borrowers», *Journal of Banking and Finance*, vol. 22: 1355-1383.

MARAIS, M.L; PATELL, J., y WOLFSON, M. [1984]: «The Experimental Design of Classification Models: An Application of Recursive Partitioning and Bootstrapping to Commercial Bank Loan Classifications», *Journal of Accounting Research*: 87-114.

MICHALSKI, R. [1983]: «A Theory and Methodology of Inductive Learning», en R. S. MICHALSKI, J. G. CARBONELL y T. M. MITCHELL (eds.), *Machine Learning: An Artificial Intelligence Approach*, Tioga Publishing Company, Palo Alto, CA.

OLMEDA, I., y BARBE-ROMERO, S. [1993]: *Redes neuronales artificiales: Fundamentos y aplicaciones*, Servicio de Publicaciones de la Universidad de Alcalá de Henares, Madrid.

OLMEDA, I., y FERNÁNDEZ, E. [1997]: «Hybrid Classifiers for Financial Multicriteria Decision Making: The Case of Bankruptcy Prediction», *Computational Economics*: 1-19.

PRESS, J., y WILSON, S. [1978]: «Choosing Between Logistic Regression and Discriminant Analysis», *Journal of the American Statistical Association*, vol. 73: 699-705.

QUINLAN, J. [1983]: «Learning Effiecient Classification Procedures and their Application to Chess End Games», en R. S. MICHALSKI, J. G. CARBONELL y T. M. MICHELL (eds.), *Machine Learning: An Artificial Intelligence Approach*, Tioga Publishing Company, Palo Alto, CA.

QUINLAN, J. [1987]: «Symplifying Decision Trees», *International Journal of Machine Studies*, vol. 27: pp. 221-234.

— [1993]: *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, California.

ROBBINS, H., y MONRO, S. [1951]: «A Stochastic Approximation Method», *The Annals of Mathematical Statistics*, vol. 22: 400-407.

STONE, M. [1974]: «Cross-validatory Choice and Assessment of Statistical Predictions», *Journal of the Royal Statistical Society*, vol. 36: 11-144.

TAM, K., y KIANG, M. [1992]: «Managerial Applications of Neural Networks: The Case of Bank Failure Predictions», *Management Science*, vol. 38: 926-947.

VARETTO, F. [1998]: «Genetic Algorithms Applications in the Analysis of Insolvency Risk», *Journal of Banking & Finance*, vol. 22: 1421-1439.

WAGNER, G.; REICHERT, A., y CHO, C. [1983]: «Conceptual Issues in Credit Scoring Models», *Credit World*, vol. 71 (May/June) 22-25 (part 1) (July/August) 22-28,41 (part 2).

YATCHEW, A. [1998]: «Nonparametric Regression Techniques in Economics», *Journal of Economic Literature*, vol. 16: 669-721.

ZMIEWSKI, M., y FOSTER, B. [1996]: «Credit-Scoring Speeds Small Business Loan Processing», *The Journal of Lending & Credit Risk Management*, 42-56.

ANEXO

TABLAS Y FIGURAS

TABLA 1

CONTRASTE DE KOLMOGOROV-SMIRNOV

Variables	$\alpha = 0,05$
Variable A	0,4326
Variable B	0,1027
Variable C	0,4545
Variable D	0,4701
Variable E	0,1401
Variable X	0,3302
Variable G	0,4868

Variables	$\alpha = 0,05$
Variable H	0,3531
Variable I	0,3786
Variable J	0,3108
Variable K	0,3629
Variable L	0,5113
Variable M	0,1426
Variable N	0,4226

TABLA 2
ERRORES DE PREDICCIÓN

Modelos	*Entrenamiento*	*Test*
AD	14,17%	12,22%
Logit	12,50%	12,22%

TABLA 3
ESTRUCTURAS DEL MODELO CART

	Sin poda		$k = 1$		$k = 10$		$k = 15$		$k = 20$		$k = 25$		$k = 30$		$k = 35$	
Muestras	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test
E.V.C.	6,6	17,1	6,6	17,1	12,8	16,8	14,6	14,6	14,6	14,6	14,6	14,6	14,6	14,6	14,6	14,6
Error de predicción															14,6	13,3

COMPLEJIDAD ←

TABLA 4
ESTRUCTURAS DEL MODELO C4.5

	$c = 0,01$		$c = 0,05$		$c = 0,1$		$c = 0,2$		$c = 0,5$		$c = 1$		$c = 5$		$c = 25$	
Muestras	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test
E.V.C.	7,1	18,5	7,2	18	14,7	14,7	14,1	14,8	14,2	15,2	14,1	15,2	11,5	16	5	17,6
Error de predicción							14,7	13,3								

COMPLEJIDAD ←

TABLA 5
ESTRUCTURAS DEL MODELO RLP

	Sin f		$f = 0,2$		$f = 0,4$		$f = 0,6$		$f = 0,8$		$f = 1$	
Muestras	Entren.	Test	Entren.	Test	Entren.	Test	Entren.	Test	Entren.	Test	Entren.	Test
E.V.C.	12,5	13,5	13,2	15,7	12,7	13,5	12,5	13,8	12,4	13,0	14,1	14,2
Error de predicción									12,3	12,3		

COMPLEJIDAD ←

TABLA 6
ESTRUCTURAS DEL MODELO MARS

N.º f. base	*4*		*8*		*12*		*16*		*20*		*24*	
Muestras	Entren.	Test	Entren.	Test	Entren.	Test	Entren.	Test	Entren.	Test	Entren.	Test
E.V.C.	13,83	14,33	13,02	14,50	12,00	14,17	11,52	13,33	11,30	13,33	11,37	13,50
Error de predicción							11,17	12,22				

COMPLEJIDAD ←

TABLA 7
ESTRUCTURAS DEL MODELO RNA

N.º nodos	2		4		6		8		10		12	
Muestras	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test	Entre.	Test
E.V.C.	12,46	14,17	12,69	13,43	12,15	13,86	11,59	13,73	12,2	13,3	12,31	13,87
Error de predicción									12,33	10,0		

COMPLEJIDAD ←

TABLA 8
RESUMEN

	ERRORES DE VALIDACIÓN CRUZADA		*ERRORES DE PREDICCIÓN*	
Modelos	*Entrenamiento*	*Test*	*Entrenamiento*	*Test*
AD	14,17%	13,99%	14,17%	12,22%
Logit	12,18%	14,12%	12,50%	12,22%
CART	14,60%	14,60%	14,60%	13,30%
RLP	12,45%	13,05%	12,33%	12,35%
RNA	12,20%	13,30%	12,33%	10,00%
C4.5	14,10%	14,80%	14,70%	13,30%
MARS	11,52%	13,33%	11,17%	12,22%

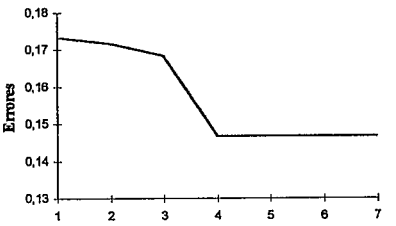


Figura 4

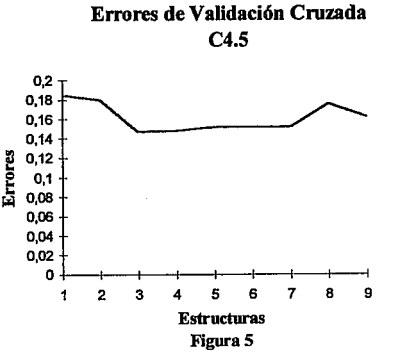


Figura 5

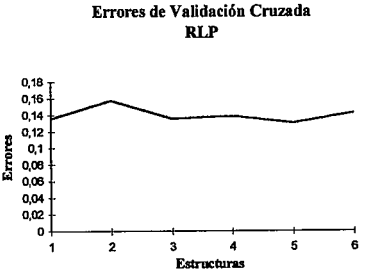


Figura 6

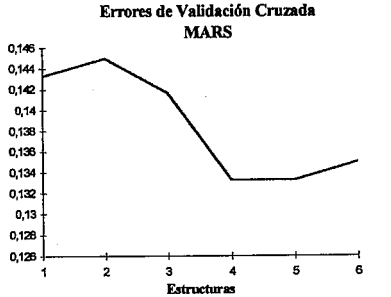


Figura 7

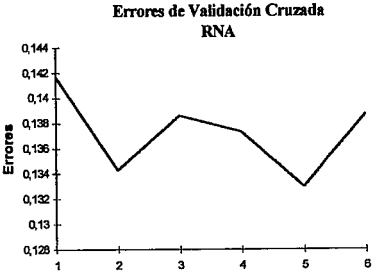


Figura8