INVESTIGACIONES ECONÓMICAS. vol. XXXI (2), 2007, 221-262

EL ANÁLISIS DE PREFERENCIAS: UN NUEVO ENFOQUE PARA EL ESTUDIO DE LA RENTABILIDAD EMPRESARIAL

PEDRO LORCA
JAVIER DE ANDRÉS
JORGE DÍEZ
JUAN JOSÉ DEL COZ
ANTONIO BAHAMONDE
Universidad de Oviedo

En el presente trabajo se construye un modelo para la determinación de los factores influyentes en la rentabilidad futura de una empresa a través de un enfoque basado en preferencias. Con este planteamiento se consiguen superar las limitaciones de los modelos de regresión y de los sistemas de clasificación. Los resultados obtenidos en una muestra de 1.745 empresas indican que es más efectivo adoptar estrategias de aumento del margen, principalmente a través de subidas de precios, que incrementar la rotación de los activos. Es destacable que el modelo construido funciona con una precisión muy superior a otros basados en regresiones.

Palabras clave: Ratios, rentabilidad financiera, análisis de preferencias.

(*JEL* M41)

1. Introducción

Para el estudio y pronóstico de la rentabilidad empresarial, así como de otros aspectos económicos, como la solvencia, los analistas e investigadores se han servido tradicionalmente de una variedad de modelos econométricos y computacionales. Esos sistemas pueden agruparse en dos clases principales: los modelos de regresión y los sistemas clasificadores.

Los autores agradecen los comentarios de los evaluadores anónimos y de Antonio Cabrales, codirector de Investigaciones Económicas. Todos los errores que subsistan en el trabajo son de nuestra exclusiva responsabilidad.

Los modelos de regresión son útiles para pronosticar el nivel futuro de una determinada variable continua sobre la base de los niveles presentes y pasados de una serie de indicadores. Dentro de este bloque de métodos se encuadran técnicas muy diversas, que van desde el tradicional modelo de regresión mínimo-cuadrática a desarrollos más recientes basados en la computación neuronal o los conjuntos borrosos (ver, por ejemplo, Kosko, 1992, 1994).

Los sistemas clasificadores, en cambio, son utilizados en el análisis económico y financiero para pronosticar, tomando como base un conjunto de variables financieras, a qué categoría pertenecerá una empresa en el futuro de entre un conjunto cerrado de posibles estados, que por lo general no suele ser muy numeroso. Los sistemas más populares dentro de este grupo son las técnicas estadísticas, las redes neuronales clasificadoras y los sistemas de inducción de árboles y reglas de decisión. Entre las técnicas estadísticas se encuentran el análisis discriminante o los modelos de variable respuesta cualitativa, ampliamente utilizados a partir del trabajo seminal de Altman (1968), principalmente en el estudio de los procesos de insolvencia empresarial. Respecto a las redes neuronales, desde finales de los 80 los avances conseguidos en el desarrollo de múltiples arquitecturas de red posibilitan que este tipo de sistemas pueda aplicarse al estudio de problemas económicos. En este sentido, han tenido especial repercusión los trabajos de Bell *et al.* (1990), Odom y Sharda (1992), o Serrano Cinca y Martín del Brío (1993). Por último, los sistemas de inducción de árboles y reglas de decisión, basados en una división continua, que permiten la construcción de modelos con altas capacidades semánticas, se han popularizado con los trabajos pioneros de Marais *et al.* (1984) y de Frydman *et al.* (1985).

Cuando el fenómeno económico que se intenta modelar puede ser representado mediante una variable continua, en un principio lo adecuado es recurrir a técnicas de regresión. Sin embargo, en ocasiones esto conduce a modelos con muy bajo poder explicativo, debido principalmente a que no es fácil determinar la forma funcional de la relación entre las distintas variables económico-financieras. Como se explicará más adelante, eso es lo que ocurre en el caso de la rentabilidad empresarial.

Una posible solución cuando la regresión no conduce a los resultados deseados es convertir la variable continua estudiada en una variable discreta con el fin de posibilitar el uso de sistemas clasificadores. En este proceso se produce una pérdida de información, por lo que po-

demos afirmar que este enfoque supone la persecución de un objetivo de segundo orden. Por ello, los modelos construidos a través de esta metodología tienen un menor poder para explicar el fenómeno que se modela y, por tanto, una menor utilidad para los analistas y los investigadores. Este inconveniente fue apuntado ya por los primeros estudiosos que abordaron el tema de la insolvencia empresarial (véase, por ejemplo, Eisenbeis, 1977).

Para superar los inconvenientes de estas dos alternativas, los desarrollos que recientemente se han producido en el campo de las ciencias de la computación ofrecen la posibilidad de acudir a una solución intermedia, como es el análisis y pronóstico de preferencias (Tesauro, 1989; Utgoff y Clouse, 1991; Cohen *et al.* 1999; Herbrich *et al.,* 1999; Joachims, 2002; Bahamonde *et al.,* 2004; Díez *et al.,* 2004). Los modelos basados en el análisis de preferencias tienen un mayor contenido informativo que los obtenidos a través de técnicas de clasificación, pues conducen a una ordenación de los individuos, mientras que con los sistemas clasificadores sólo se pronostica la pertenencia a un intervalo o categoría concretos. Además, se caracterizan por ser estrictamente no paramétricos, pues no es necesario conocer la forma funcional de la relación entre las variables que intervienen en el modelo. Por ello, resultan *a priori* más adecuados que los modelos de regresión para el estudio de problemas en los cuales la función que relaciona a las variables es cambiante o difícil de determinar, como es el caso de la rentabilidad empresarial.

En el presente artículo se pretende construir un modelo para la determinación de los factores influyentes en la rentabilidad financiera futura de una empresa a través de un enfoque basado en preferencias. El análisis de los factores que determinan el desempeño futuro de las organizaciones es un tópico que ha recibido considerable atención por parte de la literatura empresarial. Sin embargo, es un tema que dista de estar cerrado pues no se ha llegado a conclusiones unánimes. Además, la metodología empleada por la mayor parte de los estudios presenta las carencias que anteriormente se mencionaron, consistentes en suponer una forma funcional determinada para la relación entre las variables o el descarte arbitrario de observaciones. Por todo ello creemos que el enfoque de preferencias puede contribuir a arrojar luz sobre el fenómeno de la rentabilidad.

La implementación del enfoque de preferencias que proponemos recurre a un algoritmo basado en las máquinas de vectores soporte (Vapnik,

1998) adaptado al aprendizaje de preferencias (Herbrich *et al.,* 1999; Díez *et al.,* 2004). Las máquinas de vectores soporte o SVM (*Support Vector Machines*) constituyen el método de aprendizaje más importante de la actualidad. Su metodología de aprendizaje es diferente a la de otros métodos de aprendizaje tradicionales: los modelos sintetizados por las máquinas de vectores soporte no persiguen minimizar el error sobre los datos de entrenamiento, su objetivo es construir modelos estructuralmente fiables, que ofrezcan menos riesgos al valorar casos no vistos durante el entrenamiento.

El tramo de empresas estudiado corresponde a las de menor dimensión. Esta elección se debe a que son las que menos atención han recibido en la literatura previa, y a que nos permite disponer, en cada uno de los sectores de actividad estudiados, de un número de empresas suficiente para realizar comparaciones fiables entre la metodología propuesta y otras basadas en técnicas econométricas tradicionales.

El trabajo se estructura de la siguiente manera: en el Epígrafe 2 se expone el marco teórico elegido para el estudio de la rentabilidad empresarial. El Epígrafe 3 ofrece una breve revisión de los trabajos previos que han abordado de forma empírica el estudio de la rentabilidad empresarial. El cuarto está dedicado a la explicación del funcionamiento de los algoritmos utilizados. El quinto detalla la metodología empleada en el análisis empírico de los datos, incluyendo la selección de la muestra de empresas analizadas, las variables consideradas y las diferentes formas de aplicación del modelo de análisis propuesto. Los principales resultados obtenidos se comentan en el Epígrafe 6, mientras que el último se dedica al resumen y conclusiones.

2. La elección de un marco teórico para la rentabilidad empresarial

En la literatura empresarial existen dos paradigmas que proveen explicaciones alternativas sobre los factores que afectan a la rentabilidad de una empresa. Son, respectivamente, la perspectiva de la organización industrial y la perspectiva basada en los recursos.

La perspectiva de la organización industrial supone que la rentabilidad de la firma depende de las características estructurales de la industria en la que esa firma opera (Scherer y Ross, 1990). Por lo tanto, el papel de la gerencia puede ser obviado pues su conducta está restringida por las peculiaridades de la industria, que determinan tanto la conducta de

los gerentes como la rentabilidad que cada una de las compañías alcanza. Entre las características estructurales que más influencia ejercen en el desempeño se encuentran la concentración industrial, las barreras a la entrada, especialmente aquellas determinadas por la inversión publicitaria y el crecimiento de la industria (Buzell y Gale, 1987).

La perspectiva basada en los recursos supone, sin embargo, que la rentabilidad está determinada por las capacidades y los recursos propios de cada empresa y no por las peculiaridades de la industria a la que esta pertenezca (Barney, 1991; Peteraf, 1993). La distribución de estos recursos y capacidades es heterogénea a lo largo de cada sector de actividad, y la movilidad de los mismos no es perfecta y está restringida frente a los competidores, lo que explica las diferencias entre el desempeño de las distintas compañías (Amit y Schoemaker, 1993). Entre las causas que explican esta distribución heterogénea y restricción a la movilidad destaca el hecho de que los activos y capacidades estratégicas no se pueden desarrollar fácilmente en el corto plazo, su enajenación no suele ser posible, y en el caso de que lo sea el precio que el adquirente habrá de pagar será igual al valor actual de las rentas que genere su utilización (Teece *et al.,* 1997).

Habida cuenta de la existencia de estos dos paradigmas en competencia, se ha desarrollado una línea de investigación que tiene por objeto determinar cuál de ellos constituye un mejor modelo para la explicación de la rentabilidad empresarial. Las conclusiones de estos estudios no son unánimes, pues algunos trabajos concluyen que la perspectiva de la organización industrial tiene mayor poder explicativo mientras que otros autores aportan evidencia de que la perspectiva basada en los recursos constituye un modelo más adecuado. No obstante, debe tenerse en cuenta que los trabajos que anteponen la perspectiva de la organización industrial (véanse, entre otros, Schmalensee, 1985; McGahan y Porter, 1997) estudiaban exclusivamente empresas de elevada dimensión, y que la totalidad de los trabajos que consideraron una muestra exclusiva de empresas no cotizadas de tamaño pequeño o mediano (Claver *et al.,* 2002; Caloghirou *et al.,* 2004) evidenciaron la superioridad de la perspectiva basada en los recursos.

Debido a los anteriores argumentos y dado que, como se indicó en la introducción, la muestra a analizar incluye empresas de no muy elevada dimensión, en el presente estudio se ha optado por seguir la Perspectiva Basada en los Recursos.

3. Estudios previos sobre la rentabilidad empresarial

En los primeros trabajos en los que se trataba de medir el grado de relación entre la tasa de retorno contable y alguna(s) característica(s) de la empresa se optó por el empleo de técnicas de regresión, considerando una *ratio* de rentabilidad como variable dependiente y uno o varios factores como variables independientes. Dentro de esta línea podemos enmarcar, entre otros, los trabajos de Gort (1963) y Harris (1976), y en nuestro país los de Maravall (1976), Suárez Suárez (1977), Bueno y Lamothe (1983), y Lafuente Félez y Salas Fumás (1983). En todos ellos se intentaba medir el grado de relación entre la rentabilidad y el tamaño empresarial, para probar la existencia de economías de escala.

Estos estudios concluyeron que la dimensión empresarial por sí sola es una variable con un poder muy limitado para explicar el nivel de rentabilidad alcanzado, pues la bondad del ajuste de las funciones fue relativamente baja, y además aparecieron con frecuencia signos inestables en los coeficientes regresores. El trabajo de Petitbó (1982), en el que se introducían como variables explicativas, además del tamaño, parámetros indicativos de la concentración sectorial y las barreras de entrada a la industria, también evidenció las dificultades que existen para explicar la rentabilidad a través de una función.

Las razones que motivaron los pobres resultados de los trabajos que utilizaron el análisis de regresión pueden resumirse en las siguientes:

— Sensibilidad de las estimaciones ante el nivel de homogeneidad de las actividades productivas desarrolladas por las empresas incluidas en la muestra. Ésta era muy reducida, pues en todos los trabajos se partía de bases de datos que incluían únicamente empresas de gran tamaño, por lo que no se pudo llevar a cabo un proceso de desagregación sectorial, ya que ello hubiera significado trabajar con un número muy reducido de empresas en cada rama de actividad considerada, lo que hubiera perjudicado la validez estadística del proceso inferencial.

— Inexistencia de acuerdo sobre la forma funcional a emplear para la regresión, pues éste es un punto que no aclara la teoría económica.

La conclusión que se extrajo es que debe renunciarse a explicar la rentabilidad mediante una función matemática. Por ello, en trabajos posteriores a los antes reseñados los investigadores se plantean perse-

guir un objetivo más modesto como es el de tratar de determinar las variables que diferencian a las empresas más rentables de las menos rentables. Para lograr este propósito es necesario transformar la variable rentabilidad en una variable discreta, a fin de poder aplicar un modelo clasificador multivariante, la mayoría de las veces se trata del análisis discriminante. Este enfoque ya se había aplicado con éxito para el estudio de otras variables, como la *ratio* PER de las compañías cotizadas americanas (Walter, 1959) o la eficiencia de la banca comercial (Haslem y Longbrake, 1971).

Esta metodología, no obstante, presenta una serie de inconvenientes. En primer lugar, como ya se apuntó en la introducción, la transformación de variables continuas en discretas para formar clases implica una pérdida de contenido informativo, pues se trata de una transformación no biunívoca. Además, el procedimiento conlleva *(per se)* un alto grado de subjetividad debido a que el investigador lo establece arbitrariamente ya que, para la formación de categorías, no existen reglas comúnmente aceptadas. Así, algunos autores establecen según su criterio uno o más puntos de corte para formar las clases (Gillingham, 1980; Chaganti y Chaganti, 1983; Fernández Álvarez y García Olalla, 1986; González Pérez, 1996; Weir, 1996). Este procedimiento tiene el inconveniente adicional de que se induce una baja separabilidad entre las clases, debido a que entre las empresas próximas a los puntos de corte por uno y otro lado no es previsible que haya diferencias, y sin embargo son artificialmente incluidas en clases diferentes.

Por ello, y con el propósito de llegar a una identificación más precisa de las características que diferencian a las empresas según su rentabilidad, otros autores emplean procedimientos que implican el descarte de las empresas de rentabilidad intermedia, conformando dos grupos constituidos respectivamente por las compañías de más alta y más baja rentabilidad. Para la eliminación de las firmas de rentabilidad intermedia los distintos investigadores han recurrido a métodos basados estrictamente en su criterio subjetivo (Woo, 1983), y también a la fijación de intervalos construidos a partir de estadísticos descriptivos, como la media aritmética (Arraiza Antón y Lafuente Félez, 1984; Fernández Sánchez *et al.,* 1996), o los distintos valores percentiles (De Andrés, 2001; De Andrés *et al.,* 2005).

Cuadro 1
Estudios empíricos sobre las variables que más diferencian a las empresas según su rentabilidad

Autor(es)	Fte. de datos empleada	Rentabilidad	Variables consideradas	Variables con mayor influencia en la rentabilidad
Gillingham (1980)	Encuesta y cuentas anuales de empresas canadienses de la industria del cuero y británicas de industria de la lana.	Financiera.	62 variables descriptoras del diseño organizacional, posición financiera y capacidades gerenciales.	• Eficiencia y productividad.· • Tasa de crecimiento.· • Edad media del equipo directivo (mayor en las empresas no rentables).· • Concentración de las ventas (las compañías no rentables concentran sus ventas en pocos clientes).
Chaganti y Chaganti (1983)	Encuesta realizada a empresas de menos de 100 empleados de la provincia de Saskatchewan (Canadá), durante 1980.	Financiera.	32 variables descriptoras de la organización de la empresa y de sus operaciones y estrategias de producto y de mercado.	• Eficacia en la gestión de la tesorería. • Costes de producción. • Porcentaje de ventas en el mercado local y amplitud de la línea de productos ofertada a los clientes (mayores en las empresas más rentables).
Woo (1983)	Empresas líderes en sus respectivos mercados. Base de datos del proyecto PIMS 1972-1975.	Económica.	47 variables indicativas de la estabilidad en el mercado, características de la demanda, estrategia competitiva y organización.	Las empresas menos rentables se caracterizan por:· • Operar en mercados fragmentados o regionales.· • Comercializar productos no innovadores, y que necesitan continuamente servicios auxiliares y/o asistencia profesional.
Arraiza Antón y Lafuente Félez (1984)	Publicación *Las grandes empresas industriales españolas 1979-1980*. Datos de 1980.	Financiera.	Ratios resultantes de la descomposición de la rentabilidad financiera en sus componentes multiplicativos.	Las empresas más rentables se caracterizan por tener:· • Mayor ratio de fondos propios entre pasivo total.· • Menor ratio de exportaciones entre cifra de ventas. • Mayor ratio de valor añadido entre personal.
Fdez. Álvarez y García Olalla (1991)	Base de datos de la Central de Balances Anual del Banco de España. Datos de 1985 y 1986.	Financiera.	Ratios definidores de la posición económico-financiera de la empresa.	Las empresas más rentables se caracterizan por tener mayor: • Capacidad de autofinanciación / Valor de la producción.· • Reservas / Capital social. Y por tener menor: • Activo fijo / Recursos permanentes • Gastos financieros / Valor de la producción.
Fdez. Sánchez *et al.* (1996)	Muestra de 81 empresas cotizadas en la Bolsa de Madrid. Datos de 1990, 1991 y 1992.	Económica.	Ratios definidores de la posición económico-financiera de la empresa	Las empresas más rentables se caracterizan por tener mayores:· • Ventas / Activo total neto.· • Recursos generados / Deuda. Y por tener menores:· • Gastos financieros / Deuda.

Cuadro 1
Estudios empíricos sobre las variables que más diferencian a las empresas según su rentabilidad *(Continuación)*

Autor(es)	Fte. de datos empleada	Rentabilidad	Variables consideradas	Variables con mayor influencia en la rentabilidad
González Pérez (1996)	Cuentas depositadas en el Registro Mercantil de Santa Cruz de Tenerife. Datos de 1991 y 1992.	Financiera.	Factores resultantes de aplicar el método de extracción de componentes principales a una batería de ratios financieros.	Las empresas más rentables se caracterizan por tener mayores valores para los factores: • Efecto impositivo.· • Márgenes beneficiarios.· • Rotación de activos.
Weir (1996)	Encuesta y estados financieros de empresas británicas de gran tamaño. Periodo 1990-1992.	Financiera.	Variables indicativas de la forma organizativa, relación de agencia y últimas reorganizaciones efectuadas.	Las empresas más rentables se caracterizan por:· • Tener menos consejeros no ejecutivos.· • Separar lo procesos de gestión y control de las decisiones.· • No haber efectuado reorganizaciones recientes.
De Andrés (2001)	Cuentas depositadas en el Registro Mercantil de Asturias. Años 1994 y 1995.	Económica.	Ratios definidores de la posición económico-financiera de la empresa.	Las empresas más rentables tienen las siguientes características:· • Mayor coste aparente de la financiación ajena.· • Mayor solvencia a corto plazo.· • Menor periodo de pago a proveedores y acreedores.
Goddard *et al.* (2005)	Base de datos Amadeus. Empresas de Bélgica, Francia, Italia, España y Reino Unido. Años 1993 a 2001.	Económica	Ratios definidores de la posición económico-financiera de la empresa. ·	Las empresas más rentables tienen: • Menor tamaño, pero mayor cuota de mercado.· • Menor apalancamiento.· • Mayor liquidez.

En el Cuadro 1 se muestra un resumen de los principales resultados obtenidos por cada uno de los autores, así como la fuente de datos que emplearon y las variables que consideraron a la hora de analizar su posible influencia en el fenómeno de la rentabilidad empresarial.

Debe notarse que no es posible la comparación entre los resultados de los diferentes estudios debido a que en algunos se consideran solamente datos incluidos en la información financiera elaborada por las empresas, mientras que otros, al estar realizados a partir de encuestas, incluyen un rango de variables más amplio, principalmente organizativas y relativas al entorno. Además, mientras algunos estudios definen la rentabilidad empresarial en su vertiente de rentabilidad de los capitales propios, es decir, considerando el beneficio repartible en relación a los recursos invertidos, otros consideran como variable objetivo la rentabilidad económica, es decir, el cociente entre el resultado de las operaciones y el total de activo de la firma. Por último, debe indicarse que también tienen influencia en la disparidad de los resultados aspectos tales como el intervalo temporal que se considere, el entorno geográfico objeto de análisis y el tamaño de las empresas que formen la muestra.

La conclusión que se puede extraer de todo lo anterior es que es necesaria la validación y/o recálculo de los modelos antes de proceder a su extrapolación a un ámbito geográfico y temporal distinto de aquél que se utilizó para su estimación. Además, si bien todos estos trabajos constituyen una aproximación válida al fenómeno de la rentabilidad, por cuanto ofrecen modelos fiables que permiten conocer qué variables están más relacionadas con la pertenencia a los tramos de alta o de baja rentabilidad, su utilidad práctica es relativamente escasa. Ello es debido a que en el *mundo real* los usuarios de la información contable han de tomar decisiones de inversión en un entorno en el que disponen de una cantidad limitada de recursos para invertir, y por ello es necesario establecer un *ranking* a fin de seleccionar las mejores compañías, y no solamente clasificar a éstas entre más rentables y menos rentables sin ofrecer ninguna información adicional.

Hasta fechas recientes, la obtención de tales *rankings* pasaba por el empleo de modelos de regresión que, como se comentó, conducen a resultados poco fiables. Sin embargo, el desarrollo de algoritmos para representar las preferencias pone a disposición de los investigadores y analistas una potente herramienta para superar estos problemas. El

siguiente epígrafe está dedicado a la descripción de los algoritmos que se han utilizado en el presente trabajo.

4. Algoritmos para aprender funciones de *ranking* u ordenación

4.1 Las máquinas de vectores soporte

Las máquinas de vectores soporte, conocidas bajo las siglas SVM (*Support Vector Machines*) fueron introducidas por el grupo liderado por Vapnik a partir de sus trabajos sobre las teorías del aprendizaje estadístico (Boser *et al.,* 1992; Vapnik, 1998). Desde entonces las SVM han ganado un merecido reconocimiento gracias a los sólidos principios teóricos en los que se fundamenta su diseño y al estupendo rendimiento que ofrecen en una gran variedad de aplicaciones reales. Buena parte de su popularidad radica en el hecho de que son capaces de producir buenos modelos para múltiples tipos de aplicaciones prácticas.

¿Qué hace diferentes las SVM de otros métodos de aprendizaje? ¿Dónde reside su éxito? La mayor diferencia entre las máquinas de vectores soporte y otros métodos tradicionales de aprendizaje es que las SVM no siguen el principio de Minimización del Riesgo Empírico (ERM, *Empirical Risk Minimization*) que consiste en construir modelos que cometen pocos errores sobre los datos de entrenamiento, esperando que se comporten de igual forma ante datos futuros. Lo que pretenden las SVM es buscar modelos confiables, es decir, que produzcan predicciones en las que se pueda tener mucha confianza, aún a costa de cometer ciertos errores sobre los datos de entrenamiento. Ese principio recibe el nombre de Minimización del Riesgo Estructural (SRM, *Structural Risk Minimization*). Las SVM buscan un modelo que estructuralmente tenga poco riesgo de cometer errores ante datos futuros, aunque puedan cometer más errores sobre los datos de entrenamiento.

El planteamiento original de las máquinas de vectores soporte se centró en resolver problemas de clasificación binaria, donde el objetivo es generar un modelo capaz de separar los objetos de dos clases. Se parte de un conjunto de entrenamiento

$$S = \{(x_i, y_i) : i = 1, ..., n\} \qquad [1]$$

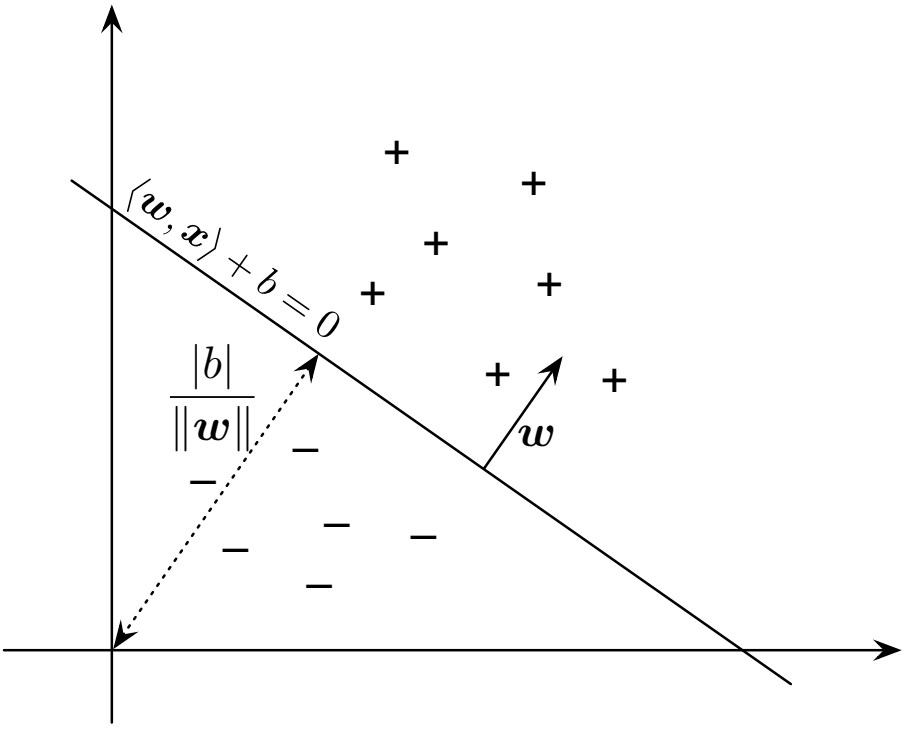
donde los objetos $x_i \in X$ y la clase $y_i \in \{+1, -1\}$. Si suponemos que $X \subseteq \Re^d$ esta clasificación binaria puede realizarse mediante una función lineal $f : X \to \Re$ que asigne valores positivos a los objetos de la

clase +1 y valores negativos a los de la clase −1. Para ello simplemente debemos encontrar un hiperplano definido por un vector w, que se llama vector de *pesos* o vector director, y un término independiente b de forma que la ecuación

$$\langle w, x \rangle + b = 0 \qquad [2]$$

divida el espacio de entrada en dos partes, una para cada clase (véase Gráfico 1).

GRÁFICO 1

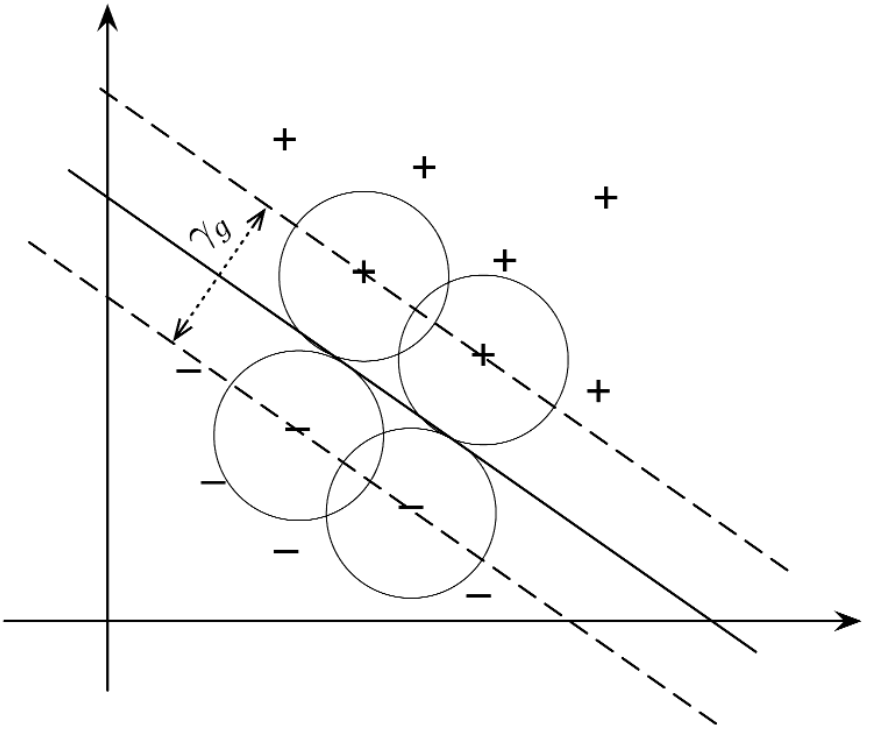


Hiperplano de separación. Dos clases linealmente separables se pueden clasificar mediante un hiperplano que divida el espacio de entrada en dos regiones

Si ambas clases son linealmente separables, es decir, si es posible encontrar un hiperplano que separe los ejemplos de cada una de las clases sin errores, el problema planteado tiene múltiples soluciones posibles, ya que en general habría muchos hiperplanos capaces de separar correctamente los ejemplos. En ese caso, ¿cuál escoger?, ¿cuál será capaz de generalizar mejor de acuerdo al conjunto de entrenamiento utilizado? Intuitivamente, parece claro que no todos los hiperplanos que consiguen separar los ejemplos sin cometer errores son igual de buenos.

Parece evidente que el hiperplano que esté más alejado de los ejemplos de ambas clases, es decir, que defina una frontera más ancha, es más resistente al ruido que puedan tener los ejemplos de entrenamiento y es menos probable que cometa errores ante datos futuros. Esa separación entre el hiperplano y cada una de las dos clases es lo que se denomina margen y es el concepto fundamental de las máquinas de vectores soporte: obtienen el hiperplano de separación de mayor margen (véase Gráfico 2).

Gráfico 2



El hiperplano de margen geométrico máximo es el que está más distanciado de ambas clases. Esa separación permite que aún los ejemplos más cercanos al hiperplano separador, los situados justo en el margen (líneas punteadas) de cada clase, podrían ser clasificados bien aunque tuvieran mucho ruido (circunferencias).

La gran ventaja de las SVM frente a otros métodos radica en el hecho de que la búsqueda del hiperplano de margen máximo puede realizarse resolviendo un problema de optimización sin óptimos locales, en el que se busca minimizar la norma del hiperplano sujeto a que los ejemplos estén bien clasificados. El hiperplano resultante será una combinación

lineal de los ejemplos que están justo en el margen (por ejemplo, en el Gráfico 2 los cuatro ejemplos rodeados por circunferencias) que son los denominados *vectores soporte* y que dan nombre al método. El vector w que define el hiperplano se obtiene mediante la expresión:

$$w = \sum_{j \in VS} y_j \alpha_j x_j \qquad [3]$$

donde los coeficientes α_j son obtenidos al resolver el problema de optimización. El vector w sólo depende del conjunto de vectores soporte VS. Una vez obtenido w se calcula el término independiente b más apropiado. Con ambos elementos ya podemos clasificar cualquier ejemplo x sin más que calcular el signo al aplicarle la función f:

$$f(x) = \langle w, x \rangle + b = \sum_{j \in VS} y_j \alpha_j \langle x_j, x \rangle + b \qquad [4]$$

Como se observa, la función f depende de productos escalares entre los vectores soporte y x.

La otra gran ventaja de las SVM es que el hiperplano de separación no tiene que ser necesariamente lineal. Mediante el uso de funciones *kernel* es fácil trasladar los ejemplos del espacio de entrada a un espacio de mayor dimensión, formado en algunos *kernels* por relaciones entre los atributos originales. Las funciones *kernel* permiten calcular de forma sencilla el producto escalar de dos vectores en un espacio de dimensión mayor, sin necesidad de obtener la representación de esos vectores en ese espacio. Al aumentar la dimensión del espacio donde se representan los ejemplos es más probable poder separarlos linealmente. La solución lineal obtenida en ese espacio de mayor dimensión, se convertirá en no lineal al trasladarla al espacio de entrada original mediante el uso de la misma función *kernel* k que empleó SVM para obtener el hiperplano de separación:

$$f(x) = \sum_{j \in VS} y_j \alpha_j k \langle x_j, x \rangle + b \qquad [5]$$

Para el lector interesado en la formulación completa y el desarrollo matemático de las máquinas de vectores soporte, así como en el uso de funciones *kernel*, se recomienda la lectura de Cristianini y Shawe-Taylor (2000) o Vapnik (1998).

4.2 Máquinas de vectores soporte para preferencias

Aunque en su origen las SVM se diseñaron para resolver solamente problemas de clasificación binaria, en la actualidad su aplicación se

ha extendido a tareas de regresión, multiclasificación, agrupamiento o, como en el caso que aquí tratamos, tareas en las que se pretende ordenar un conjunto de elementos.

Entre las distintas formas que existen para aprender a ordenar objetos (Utgoff y Clouse, 1991; Cohen *et al.* 1999), la que aquí utilizaremos consistirá en generar un modelo capaz de asignar un valor real a cada objeto de forma que los objetos preferibles alcancen valores más altos (Herbrich *et al.,* 1999; Joachims, 2002; Bahamonde *et al.,* 2004; Díez *et al.,* 2004). En nuestro caso, pretendemos aprender una función $f : Empresas \rightarrow \Re$; la función f asignará valores más altos a aquellas empresas que se prevé obtengan una mayor rentabilidad. A pesar de lo que puede sugerir su formulación, la función f en ningún caso permitirá predecir el valor exacto que alcanzará la rentabilidad de cada una de las empresas en el futuro, sino que es una función construida únicamente a los efectos de poder realizar ordenaciones. Es decir, que tal y como se indicaba en la introducción, el objetivo será un poco más modesto: f debe asignar, dadas dos empresas u y v de las que sabemos que v *es preferible a* (más rentable que) u, un mayor valor a v que a u:

$$preferencia\,(v,u) \Longleftrightarrow f\,(v) > f\,(u) \qquad [6]$$

Es evidente que con una función como f seremos capaces de ordenar subconjuntos de empresas. A una función como ésta se le llama función de *preferencias*, de *ranking* o función de *utilidad*. Este planteamiento, aparentemente, sugiere que el aprender a ordenar se puede resolver como un problema de regresión. Sin embargo, veremos que no es estrictamente necesario hacerlo así y que puede lograrse también mediante la resolución de un problema de clasificación binaria. Una vez sintetizada la función de preferencias será posible usarla como un método para valorar empresas; valores que serán coherentes con las preferencias, en términos de rentabilidad, de las que se aprendió la función.

Lo que pretendemos es aprender a ordenar objetos, en nuestro caso, pronosticar entre dos empresas cuál de ellas tendrá mayor rentabilidad. La información que debemos usar para entrenar un sistema de esas características es precisamente de ese tipo: pares de empresas en las que una de ellas es mejor (más rentable) que la otra. Por ejemplo, si sabemos que de tres empresas u_1, u_2, u_3, u_1 es la más rentable y u_3 es

la menos rentable, podemos expresar esta relación entre ellas mediante tres preferencias:

$$preferencia\,(u_1, u_2)\,,\ preferencia\,(u_1, u_3)\,,\ preferencia\,(u_2, u_3)$$

A cada uno de los enunciados $preferencia(v, u)$ se le llama *juicio de preferencia*.

- *Funciones lineales*

Nuestro punto de partida para aprender una función de *ranking*, preferencias o utilidad es un *conjunto de juicios de preferencias,* donde partiremos de pares de empresas en las que una de ellas es más rentable que la otra:

$$JP = \{preferencia\,(v_1, u_i) : i = 1, ..., n\}\,. \qquad [7]$$

En nuestro caso, las empresas se representan mediante d atributos numéricos, es decir, son vectores en $\Re^d$. La clave para buscar una función *lineal* $f : \Re^d \to \Re$ que cumpla, en la medida de lo posible, la ecuación [6] está en extender esta ecuación de la forma siguiente:

$$preferencia\,(v, u) \Longleftrightarrow f\,(v) > f\,(u) \Longleftrightarrow f\,(v - u) > 0. \qquad [8]$$

La última expresión constituye una especificación de las restricciones que ha de cumplir la función buscada. Leyendo las versiones positivas y las negativas (negando el predicado *preferencia*) se tiene que la función f ha de tomar:

— valores positivos en las diferencias entre objetos donde el primero es preferible al segundo (pares *mejor-peor*), y

— valores negativos en las diferencias entre objetos donde el primero *no* es preferible al segundo (pares *peor-mejor*).

Leída de esta manera la especificación de una función de *ranking* es evidente que se puede obtener usando un algoritmo de clasificación binaria que asigne valores positivos a los objetos de una clase y negativos a los de la otra, como sucede con los SVM tradicionales. Por tanto, el conjunto de entrenamiento

$$E = \{(v_i - u_i; +1)\,, (u_i - v_i, -1) : i = 1, ..., n\} \qquad [9]$$

permite obtener la función lineal f buscada. Es importante destacar que su carácter lineal hace que $f : \Re^d \rightarrow \Re$ tenga asociado un vector w en este caso sin término independiente b, ya que los ejemplos son simétricos respecto del origen. El valor de la función en cada punto es:

$$f(z) = \langle w, z \rangle = \sum_{j-1}^{d} w_j z_j \qquad [10]$$

Por tanto, el valor viene dado por el producto escalar de las coordenadas del punto por el vector de pesos. Intuitivamente, cada componente de z está ponderada por la correspondiente de w para determinar la calificación final con la que se harán comparaciones para ordenar conjuntos de objetos. Una ventaja atractiva del caso lineal es que nos permite esta interpretación intuitiva de las funciones de valoración.

El carácter lineal de la función permite dar una interesante interpretación geométrica. Si w es el vector de pesos que representa a f, este vector determina un hiperplano que pasa por el origen: el que forman los vectores x perpendiculares a w, aquellos cuyo producto escalar $< w, x >= 0$. La distancia de un punto cualquiera z a este hiperplano es:

$$\text{distancia } (\langle w, x \rangle = 0; z) = \frac{\langle w, z \rangle}{\|w\|} = \frac{f(z)}{\|w\|} \qquad [11]$$

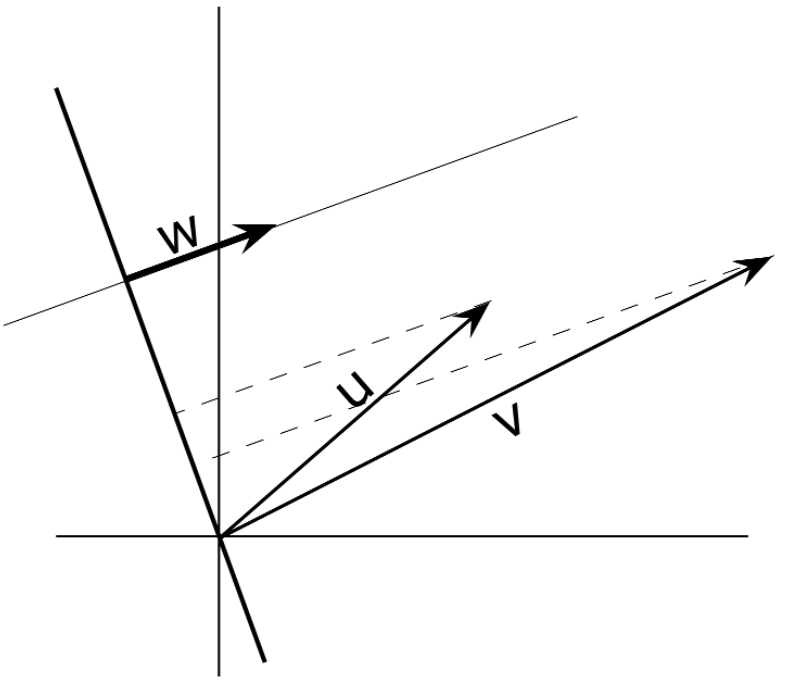
Se puede entonces decir que, salvo un factor de escala (la norma del vector de pesos), la calificación de un objeto es igual a la distancia a un hiperplano que pasa por el origen. Por tanto, un juicio de preferencias como *preferencia*(v, u) indica que el hiperplano que determina la función de calificación ha de estar más lejos de v que de u. O lo que es lo mismo, el vector director del hiperplano (el w) ha de formar un ángulo de coseno positivo (desde -90º a +90º) con el vector diferencia $v - u$. Esto aparece ilustrado en el Gráfico 3.

- Funciones no lineales

Por supuesto, no siempre será posible encontrar una función lineal que determine una línea de separación entre las diferencias positivas y las negativas. En términos de preferencias, a veces las preferencias por un objeto no serán proporcionales a las componentes de un vector que las describa. No en todos los casos se cumple que se prefiera *cuanto más* (*o menos*) *mejor*. Por ejemplo, el apalancamiento financiero hace que el nivel de endeudamiento de una empresa suela tener un punto óptimo en el que se maximiza la rentabilidad; niveles de endeudamiento menores o mayores hacen decrecer la rentabilidad. Es decir, para

aprender a ordenar objetos se debe contemplar la posibilidad de construir funciones de preferencia no lineales. En este apartado se verá cómo se puede hacer esto utilizando funciones *kernel*.

GRÁFICO 3



Se busca un vector *w* tal que el hiperplano *<w,x> = 0* se encuentre más lejano de los vectores preferibles. En el gráfico *v* es mejor que *u* (simbólicamente, *v > u*). La calificación de cada objeto es proporcional a la distancia del vector que lo representa al hiperplano perpendicular a *w*.

Sea X un espacio de objetos de entrada que se representan en un espacio vectorial usando la transformación $\phi : X \rightarrow \Re^m$. Por lo visto en el apartado anterior, se puede construir un separador lineal si representamos cada juicio de preferencias por las diferencias positiva y negativa de sus vectores asociados:

$$preferencia\left(x^{(1)}, x^{(2)}\right)$$
$$\rightarrow \left\{\left(\phi\left(x^{(1)}\right) - \phi\left(x^{(2)}\right); +1\right), \left(\phi\left(x^{(2)}\right) - \phi\left(x^{(1)}\right); -1\right)\right\}. \quad [12]$$

Como ahora no se podrán manejar las componentes de los vectores que representan a los objetos (los $\phi(x)$), será necesario utilizar los productos escalares mediante la función *kernel* correspondiente. Pero es importante destacar que, desde el punto de vista formal, los objetos

que se representan en un espacio con producto escalar ($\Re^m$ en este caso) no son objetos individuales, sino pares de objetos que aparecen mencionados en los juicios de preferencias. Es decir, que si se tiene que describir el conjunto de entrenamiento de una manera explícita, éste estará formado, para cada juicio de preferencias, por dos entradas: con la clase $+1$ la secuencia de los objetos mejor-peor, y con la clase -1 la secuencia de objetos peor-mejor. En símbolos:

$$\textit{preferencia}\left(x^{(1)}, x^{(2)}\right) \rightarrow \left\{\left(x^{(1)}, x^{(2)}; +1\right), \left(x^{(2)}, x^{(1)}; -1\right)\right\}. \quad [13]$$

Desde este punto de vista, las entradas de este conjunto de entrenamiento se representan en un espacio de vectorial usando la transformación:

$$\psi : X \times X \rightarrow \Re^m, \psi\left(x^{(1)}, x^{(2)}\right) = \phi\left(x^{(1)}\right) - \phi\left(x^{(2)}\right). \quad [14]$$

El *kernel* asociado a esta transformación del espacio de entrada en el de características es el siguiente:

$$\begin{aligned} K\left(x^{(1)},x^{(2)},x^{(3)},x^{(4)}\right) &= \left\langle \psi\left(x^{(1)},x^{(2)}\right), \psi\left(x^{(3)},x^{(4)}\right)\right\rangle = \left\langle \left(\phi\left(x^{(1)}\right) - \phi\left(x^{(2)}\right)\right), \left(\phi\left(x^{(3)}\right) - \phi\left(x^{(4)}\right)\right)\right\rangle \\ &= \left\langle \phi\left(x^{(1)}\right), \phi\left(x^{(3)}\right)\right\rangle - \left\langle \phi\left(x^{(1)}\right), \phi\left(x^{(4)}\right)\right\rangle - \left\langle \phi\left(x^{(2)}\right), \phi\left(x^{(3)}\right)\right\rangle + \left\langle \phi\left(x^{(2)}\right), \phi\left(x^{(4)}\right)\right\rangle \\ &= K\left(x^{(1)},x^{(3)}\right) - K\left(x^{(1)},x^{(4)}\right) - K\left(x^{(2)},x^{(3)}\right) + K\left(x^{(2)},x^{(4)}\right) \end{aligned} \quad [15]$$

donde $K(x^{(i)}, x^{(j)})$ es el *kernel* asociado a la transformación ϕ de los objetos individuales en un espacio de características adecuado. Al *kernel* K se le llama el *Kernel de Herbrich* asociado al *kernel* K (Herbrich *et al.* 1999).

El caso lineal del apartado anterior es, por supuesto, un caso particular de este último en el que el *kernel* K es, simplemente, el producto escalar de las componentes de los vectores que describen los objetos. Si los objetos están descritos por los vectores cuyas componentes son los monomios de hasta un cierto grado, K es por ejemplo un *kernel* polinómico como el que se usará en las pruebas experimentales.

La valoración de los objetos se hacía directamente con la función de separación $f : X \rightarrow \Re$, pero ahora no se conocen las coordenadas en el espacio vectorial ($\Re^d$) y, por tanto, no se conoce esta función f. El

truco del kernel utilizado ahora con el *Kernel de Herbrich* conduce a una función de separación $f : X \times X \to \Re$ dada por

$$F(x,y) = \sum_{s \in VS} \alpha_s z_s \left\langle \phi\left(x_s^{(1)}\right) - \phi\left(x_s^{(2)}\right), \phi(x) - \phi(y) \right\rangle = \\ = \sum_{s \in VS} \alpha_s z_s K\left(x_s^{(1)}, x_s^{(2)}, x, y\right) \quad [16]$$

donde VS es el conjunto de índices de los vectores soporte, z_s son las clases y α_s los coeficientes que calcula SVM. Esta función, teniendo en cuenta cómo está construido el *kernel*, cumple que

$$F(x,0) > F(y,0) \Longleftrightarrow F(x,y) > 0, \quad [17]$$

con lo cual se puede definir como función de valoración de los objetos

$$f : X \to \Re, f(x) = F(x,0). \quad [18]$$

Esta definición cumple dos propiedades interesantes:

1. en el caso lineal esta función f coincide con la definida en el apartado anterior;
2. es coherente con las preferencias en la medida en que lo sea F ya que esta función se aprendió con las restricciones

$$preferencia(v,u) \Longleftrightarrow F(v,u) > 0. \quad [19]$$

La expresión de f se puede simplificar ya que del desarrollo completo de $F(x,0)$ se puede eliminar la suma de una constante (en el caso lineal es 0) que resulta irrelevante al hacer comparaciones. Por tanto, en términos prácticos, la función general de valoración f, tanto para el caso lineal como para el no lineal, queda como sigue:

$$f(x) = \sum_{s \in VS} \alpha_s z_s \left\langle \phi\left(x_s^{(1)}\right) - \phi\left(x_s^{(2)}\right), \phi(x) \right\rangle = \\ = \sum_{s \in VS} \alpha_s z_s \left(K\left(x_s^{(1)}, x\right) - K\left(x_s^{(2)}, x\right)\right) \quad [20]$$

5. Metodología

Una vez realizada la descripción de los algoritmos para el aprendizaje de preferencias, el resto del trabajo se dedica a exponer las pruebas empíricas que se han llevado a cabo con el fin de determinar la idoneidad de dichas técnicas para la previsión de la rentabilidad empresarial. A ello se dedican el presente Epígrafe, en el que se describe la metodología empleada, y el Epígrafe 5, en el que se ofrece el comentario de los resultados obtenidos.

5.1 La base de datos

La información para llevar a cabo la investigación se ha tomado de los estados financieros de las empresas comerciales e industriales españolas. Conforme a la legislación española se exige que las sociedades que limitan su responsabilidad depositen las cuentas anuales en el Registro Mercantil. Esta información se recoge por *Bureau van Dijk* y por *Informa* en la base *Sistema de Análisis de Balances Ibéricos* (SABI). Los estados financieros analizados corresponden a los años 1999, 2000, 2001 y 2002.

Como se indicó en la introducción, la necesidad de disponer de una base de datos amplia nos llevó a centrar el estudio en las empresas de menor dimensión. Por ello, se eliminaron de la base las empresas con más de 250 trabajadores. Para la fijación de este límite se tuvo en cuenta que las firmas que lo superan generalmente trabajan en varios sectores de actividad, por lo que en ausencia de información segmentada se podría distorsionar el análisis. Esto es especialmente importante en nuestro estudio puesto que, como se verá, la adscripción sectorial es una variable fundamental en alguna de las pruebas realizadas. Además, se eliminaron las empresas con menos de 100 empleados. De esta forma se garantiza una cierta calidad en los estados contables analizados, dada la obligatoriedad de auditoría para la gran mayoría de las que superan dicho tamaño.

Tras esta identificación de las empresas objeto de estudio, se aplicaron una serie de filtros para garantizar: 1) la calidad de la información financiera, y 2) que la muestra seleccionada represente adecuadamente la actividad económica de cada sector. Por ello se eliminaron, además de las empresas cuya información es insuficiente para calcular alguna de las variables utilizadas en el análisis, las que no hubieran tenido actividad o se hubieran creado en los años objeto de estudio. Con el fin de garantizar un número mínimo de empresas por sector que evite

distorsiones en las comparaciones intrasectoriales, se eliminaron también las empresas incluidas en una división de la CNAE-93 (2 dígitos) con menos de 10 empresas.

Después de estos recortes la base de datos que finalmente se analizó cuenta con 1.745 empresas, cuyo desglose sectorial aparece en el Cuadro 2.

CUADRO 2
Empresas en la muestra según rama de actividad

Nº	Nombre	Empresas
01	Agricultura, ganadería, caza y actividades de los servicios relacionados con las mismas	16
14	Extracción de minerales no metálicos ni energéticos	15
15	Industria de productos alimenticios y bebidas	128
17	Industria textil	66
18	Industria de la confección y de la peletería	14
19	Preparación, curtido y acabado del cuero; fabricación de artículos de marroquinería y viaje; artículos de guarnicionería, talabartería y zapatería	11
20	Industria de la madera y del corcho, excepto muebles; cestería y espartería	26
21	Industria del papel	34
22	Edición, artes gráficas y reproducción de soportes grabados	58
24	Industria química	68
25	Fabricación de productos de caucho y materias plásticas	61
26	Fabricación de otros productos minerales no metálicos	93
27	Metalurgia	27
28	Fabricación de productos metálicos, excepto maquinaria y equipo	101
29	Industria de la construcción de maquinaria y equipo mecánico	63
31	Fabricación de maquinaria y material eléctrico	34
33	Fabricación de equipo e instrumentos médico-quirúrgicos, de precisión, óptica y relojería	14
34	Fabricación de vehículos de motor, remolques y semirremolques	47
35	Fabricación de otro material de transporte	21
36	Fabricación de muebles; otras industrias manufactureras	30
41	Captación, depuración y distribución de agua	14
45	Construcción	190
50	Venta, mantenimiento y reparación de vehículos de motor, motocicletas y ciclomotores; venta al por menor de vehículos de motor	55
51	Comercio al por mayor e intermediarios del comercio, excepto de vehículos de motor y motocicletas	178
52	Comercio al por menor, excepto comercio de vehículos de motor, motocicletas y ciclomotores; reparación de efectos personales y enseres domésticos	55
55	Hostelería	89
60	Transporte terrestre; transporte por tuberías	59
63	Actividades anexas a los transportes; actividades de agencias de viajes	27
70	Actividades inmobiliarias	18
72	Actividades informáticas	17
74	Otras actividades empresariales	44
85	Actividades sanitarias y veterinarias, servicio social	34
90	Actividades de saneamiento publico	13
92	Actividades recreativas, culturales y deportivas	25
	Total	1.745

5.2 La variable a predecir

La mayor parte de los trabajos han considerado como variable a predecir la tasa de retorno contable, que es la tasa de rentabilidad de las inversiones calculada a partir de los estados financieros de la empresa, en sus distintas variantes (rentabilidad económica o rentabilidad financiera). Si bien es comúnmente aceptado que este cociente no es el mejor indicador de la eficacia en la gestión, pues es inferior a otras medidas, como el valor de mercado de la empresa o la tasa interna de retorno (véase, por ejemplo, Whittington, 1980; Foster, 1998), las limitaciones en la información disponible, especialmente cuando se pretende analizar empresas no cotizadas en bolsa, hacen que sea la mejor medida que se puede obtener (Kay, 1976; Horowitz, 1984; Martin 2002, entre otros).

En el presente estudio se tomó la rentabilidad financiera, puesto que ha sido la variable objetivo en la mayor parte de los estudios previos, como se puede ver en el Cuadro 1. Este *ratio* se define como el cociente entre el beneficio neto de la empresa y el patrimonio neto. Diversos autores (Kelly y Tippet, 1991; Brief y Lawson, 1992; Martin, 2002, entre otros) indican que, pese a sus limitaciones, este *ratio* proporciona una medida adecuada del desempeño de la empresa.

En el presente estudio se ha considerado la rentabilidad media para los años 2001 y 2002. De este modo se eliminan, al menos parcialmente, las distorsiones indeseables en las cifras contables causadas por cambios coyunturales en el entorno de la empresa.

En el Cuadro 3 se ofrecen algunas estadísticas descriptivas de la rentabilidad financiera de las empresas que forman la muestra. A la vista de dicha información se observa claramente que la distribución de la variable a predecir es relativamente simétrica y extremadamente leptocúrtica. Esto significa que la mayoría de las observaciones están muy próximas a la media y que la varianza de la distribución se debe sólo a unas pocas observaciones que están muy por encima o por debajo de la media.

CUADRO 3
Información descriptiva de la rentabilidad de las empresas analizadas

1er cuart.	Mediana	3er cuart.	Media	Desv. Típ.	Asimetría	Curtosis
0,042	0,104	0,187	0,120	0,369	3,899	85,231

Estas evidencias preliminares proporcionan una razón adicional de los pobres resultados de técnicas como el análisis de regresión que tratan de predecir el valor absoluto de la rentabilidad de las firmas. Si la mayoría de las observaciones están concentradas en torno a la media, predecir la media para todas las empresas provoca bajas tasas de error, pero esto es de poca utilidad para el análisis económico.

5.3 Los predictores

De acuerdo con la perspectiva basada en los recursos éstos se pueden clasificar en tres categorías: tangibles, intangibles y capacidades (Fahy, 2000). Uno de los mayores problemas para identificar los recursos es que los sistemas de información empresarial ofrecen una visión fragmentada de los mismos (Grant, 1991). No obstante, la primera categoría de recursos (los tangibles) se caracteriza por ser transparente y, por tanto, relativamente fácil de medir (Hall, 1989). Por ello en el presente trabajo se han tomado únicamente recursos que se puedan extraer de los datos procedentes de los estados financieros. Si bien en un principio esto puede parecer una limitación, la elección de esta fuente de información ha venido motivada por dos razones:

a) Permite disponer de un número de empresas en la muestra más elevado que otros procedimientos, como el envío de encuestas. Este aspecto es importante para llegar a conclusiones robustas.

b) En muchas de las situaciones del mundo real en que se persigue pronosticar la rentabilidad futura de una empresa la principal información con que cuentan los analistas es la procedente de las cuentas anuales por ser éstas documentos de acceso público a través de los registros mercantiles.

Dado que la finalidad del trabajo es determinar los factores influyentes en la rentabilidad futura acudiendo a la información financiera que la empresa publica, es preciso seleccionar, de entre los múltiples indicadores que se pueden construir con los datos contenidos en las cuentas anuales, aquéllos que se considerarán como posibles predictores.

En el presente trabajo se ha tomado un conjunto de 27 *ratios* y variables que recogen la situación económico-financiera de la empresa (ver Cuadro 4).

CUADRO 4
Variables financieras

V1	*Total Activo*
V2	*Importe Neto Cifra Negocios*
V3	$\frac{\textit{Activo Circulante}}{\textit{Pasivo Circulante}}$
V4	$\frac{\textit{Tesorería}}{\textit{Pasivo Circulante}}$
V5	$\frac{\textit{Activo Circulante - Existencias}}{\textit{Pasivo Circulante}}$
V6	$\frac{\textit{Total Pasivo - Fondos Propios}}{\textit{Total Pasivo}}$
V7	$\frac{\textit{Fondos Propios}}{\textit{Fondos Propios + Deudas l/p}}$
V8	$\frac{\textit{Fondos Propios}}{\textit{Deudas l/p}}$
V9	$\frac{\textit{Fondos Propios}}{\textit{Activo Total}}$
V10	$\frac{\textit{Importe Neto Cifra Negocios}}{\textit{Total Activo}}$
V11	$\frac{\textit{Ingresos Explotación - Consumo Mercaderías - Otros Gastos Explotación}}{\textit{Gastos Personal}}$
V12	$\frac{\textit{Resultados Antes Impuestos}}{\textit{Ingresos Explotación}}$
V13	$\frac{\textit{Resultados Antes Impuestos}}{\textit{Número Empleados}}$
V14	*Crecimiento importe neto cifra negocios*
V15	*Crecimiento valor añadido*
V16	$\frac{\textit{Ingresos Explotación}}{\textit{Número Empleados}}$
V17	$\frac{\textit{Gastos Personal}}{\textit{Ingresos Explotación}}$
V18	$\frac{\textit{Gastos Personal}}{\textit{Número Empleados}}$
V19	$\frac{\textit{Gastos Financieros + Variación Provisión Inversiones Financieras}}{\textit{Importe Neto Cifra Negocios}}$
V20	$\frac{\textit{Resultados Explotación}}{\textit{Gastos Financieros}}$
V21	$\frac{\textit{Total Activo}}{\textit{Número Empleados}}$
V22	$\left(\frac{\textit{Deudores}}{\textit{Importe Neto Cifra Negocios}}\right)$ x 360
V23	$\left(\frac{\textit{Pasivo Líquido}}{\textit{Consumo Mercaderías y Materias Primas}}\right)$ x 360
V24	$\left(\frac{\textit{Acreedores Comerciales}}{\textit{Ingresos Explotación}}\right)$ x 360
V25	$\left(\frac{\textit{Fondos Propios + Deudas l/p - Inmovilizado}}{\textit{Importe Neto Cifra Ventas}}\right)$ x 360
V26	$\frac{\textit{Resultado Ejercicio}}{\textit{Activo Total}}$
V27	$\frac{\textit{Resultado Ejercicio}}{\textit{Fondos Propios}}$

En el proceso de selección de los mismos se han tenido en cuenta tanto la metodología como los resultados de las investigaciones previas que se reseñaron en el Epígrafe 3. Así, se han incluido indicadores de tamaño, como V1 y V2, ya que aunque la mayoría de los estudios previos no evidenciaran una relación significativa entre tamaño y rentabilidad, pudiera ser que este resultado se debiera a las limitaciones de la metodología empleada, y máxime cuando en ciertas investigaciones más recientes (Goddard *et al.,* 2005) el tamaño resultó ser un factor relevante.

Figuran también indicadores de liquidez: V3, V4 y V5 (Chaganti y Chaganti, 1983; De Andrés, 2001; Goddard, 2005), endeudamiento: V6, V7, V8 y V9 (Arraiza Antón y Lafuente Félez, 1984; Goddard, 2005), rotación de activos: V10 (Fernández Sánchez *et al.,* 1996; González Pérez, 1996), eficiencia, productividad e intensidad en la utilización de la mano de obra: V11, V13, V16, V17 y V18 (Gillingham, 1980; Chaganti y Chaganti, 1983; Arraiza Antón y Lafuente Félez, 1984), margen: V12 (González Perez, 1996), crecimiento: V14 y V15 (Gillingham, 1980), coste de la financiación: V19, V20 (Fernández Álvarez y García Olalla, 1991; Fernández Sánchez *et al.,* 1996; De Andrés, 2001).

A priori, la rotación de activos, margen, eficiencia y productividad deberían contribuir a una mayor rentabilidad, mientras que los costes deberían estar negativamente relacionados con la misma. Para las variables de crecimiento y liquidez la relación teórica con la rentabilidad es menos clara, pues no están directamente relacionados con ninguno de los componentes de la misma. Sin embargo, niveles altos de crecimiento y liquidez suelen ser señal de que la empresa cuenta con capacidades y recursos que en un corto plazo se van a traducir en mayores niveles de rentabilidad, como así se ha comprobado en las antes citadas investigaciones empíricas que encontraron relevantes estas variables. En el caso del endeudamiento de la empresa, su efecto sobre la rentabilidad puede ser positivo o negativo dependiendo del coste de la financiación, pudiéndose afirmar en general que para cada empresa existe un nivel óptimo de utilización del apalancamiento financiero que maximiza la rentabilidad de la empresa.

Además de los anteriores indicadores, y a pesar de que otras investigaciones anteriores no las consideraron, se ha creído procedente incluir variables que caracterizan la gestión del ciclo de explotación de la empresa (V22, V23, V24 y V25). Ello se debe a que, *a priori,* una buena gestión del ciclo de explotación, reflejada en periodos de maduración

más cortos y una mayor rotación del capital circulante, debe traducirse en una mayor rentabilidad futura. Este argumento ha sido corroborado por algunas investigaciones previas no directamente dirigidas a determinar las causas de la rentabilidad pero sí las de otras magnitudes que guardan relación con la misma, como el valor de mercado de la firma o el valor añadido generado (ver, por ejemplo, Johnson y Soenen, 2003).

Por último, se han considerado como posibles predictores de la rentabilidad futura los indicadores de la rentabilidad actual, tanto económica como financiera. Son las variables V26 y V27. Su inclusión se debe a que podría ocurrir que el resto de variables fueran únicamente *proxies* de la rentabilidad pasada y no factores con una incidencia específica sobre la rentabilidad futura.

En la especificación final de los predictores, y por las mismas razones que para la variable a predecir también en este caso se ha considerado el promedio de dos años. Pero, debido a la naturaleza predictiva del trabajo, se han tomado los años 1999 y 2000.

El cálculo de la matriz de correlaciones entre las variables puso de relieve la presencia de una fuerte correlación entre algunos pares de variables. Debido a ello, se procedió a reducir la dimensión del número de variables que inicialmente se habían considerado en el análisis, identificando los factores principales a través de la técnica del análisis factorial. El objetivo de esta técnica es transformar un conjunto de K variables intercorrelacionadas en un conjunto de W componentes o factores incorrelacionados entre sí, siendo $W < K$. Cada uno de los W factores es una combinación lineal de las K variables.

Con carácter previo al análisis factorial se determinó si era adecuada la realización de una reducción de dimensiones. Para ello se utilizó la medida de adecuación muestral de Kaiser-Meyer-Olkin que dio un valor de 0,598, valor considerado como tolerable por Hair *et al.* (1995). Asimismo, el test de Barlett rechazó la hipótesis de que la matriz de correlaciones sea la unidad. Para la extracción de los componentes principales se tomaron aquellos factores cuyo valor propio (varianza explicada) fuera superior a uno, es decir, aquéllos que incorporan más varianza que una variable original (criterio de Kaiser). Además, se exigió que los factores seleccionados explicasen un 80% de la varianza total que aporta la matriz de datos. Por ello fue preciso eliminar las variables con menor *comunalidad* (variables cuya varianza está peor recogida por los factores cuyo valor propio excede de 1).

Por último, para simplificar la interpretación de los factores se procedió a rotarlos con el criterio *varimax*. Las correlaciones entre los nuevos factores rotados y las variables permiten interpretar el significado de los componentes retenidos. La matriz de covarianza de los factores extraídos mostró la inexistencia de correlación entre los mismos.

La matriz de componentes rotados (ver Cuadro 5) recoge las correlaciones entre los factores rotados y las variables iniciales, de modo que se puede interpretar el contenido de los ocho factores identificados. Es decir, como alternativa al conjunto de 27 variables utilizamos el conjunto de factores identificados no directamente observables. Dadas las correlaciones expresadas en el Cuadro 5 podemos identificar a cada uno de los factores con las siguientes dimensiones de la empresa 1) liquidez a corto plazo, 2) margen de beneficio, 3) tamaño, 4) endeudamiento, 5) crecimiento, 6) costes por empleado, 7) ciclo de explotación, y 8) rotación de activos. Por lo tanto, estos 8 factores serán los predictores definitivos que se consideran para explicar la rentabilidad futura (Rentabilidad futura $= f(F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_8)$).

CUADRO 5
Matriz de componentes rotados

	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
V1	0,222	0,247	0,695	-0,101	-0,018	0,012	0,014	0,379
V2	-0,063	0,055	**0,839**	0,015	0,008	-0,033	0,010	-0,281
V3	**0,891**	0,104	0,066	0,190	-0,018	0,008	0,128	0,032
V4	**0,810**	0,168	-0,007	0,099	-0,023	0,001	0,020	-0,033
V5	**0,929**	0,121	0,038	0,120	-0,026	0,018	0,054	-0,017
V6	-0,510	-0,264	0,092	**-0,732**	0,136	0,022	-0,037	0,006
V7	0,034	0,090	0,051	**0,898**	-0,028	-0,002	0,036	-0,044
V10	-0,198	-0,141	0,275	0,027	0,095	-0,034	0,116	**-0,760**
V11	-0,068	0,010	0,644	0,017	0,014	0,587	0,002	-0,187
V12	0,254	**0,825**	-0,091	0,198	0,006	-0,039	0,082	-0,001
V13	0,118	**0,824**	0,247	0,051	0,028	0,203	0,037	0,070
V14	-0,027	0,096	-0,027	-0,092	**0,900**	-0,013	-0,016	0,025
V15	0,066	**0,732**	0,353	-0,016	0,082	0,031	0,009	0,191
V16	-0,047	0,013	0,031	-0,048	**0,912**	0,010	0,005	0,051
V18	-0,014	0,055	-0,096	-0,014	-0,003	**0,890**	-0,010	-0,030
V21	0,112	0,104	0,502	-0,035	-0,010	0,655	0,019	0,295
V22	0,076	-0,034	0,025	-0,037	-0,001	0,018	**-0,918**	0,121
V23	-0,256	-0,116	0,108	-0,023	0,199	-0,010	0,064	0,699
V25	0,389	0,096	0,063	0,043	-0,015	0,023	**0,832**	0,119
V26	0,122	**0,720**	-0,137	0,308	0,046	-0,030	0,028	-0,240

Las correlaciones superiores a 0,700 aparecen en negrita.

También se ha llevado a cabo un análisis descriptivo de los factores identificados, que indica la existencia de elevados niveles de asimetría y de *leptocurtosis* en las distribuciones de frecuencias de la mayoría de ellos, lo cual corrobora los resultados previos sobre la distribución estadística de los indicadores financieros (Lau *et al.,* 1995; Martikainen *et al.,* 1995, entre otros). Esto obliga a rechazar las hipótesis de normalidad y ofrece una razón adicional para cuestionar las investigaciones empíricas previas que tratan de explicar la rentabilidad usando técnicas de clasificación paramétricas o semiparamétricas como el análisis discriminante lineal o la regresión logística.

5.4 Las pruebas realizadas

Como se ha indicado previamente, el principal objetivo del presente trabajo es analizar la eficiencia de las máquinas de vectores soporte cuando se utilizan para prever qué compañía, entre dos dadas, alcanzará un mayor nivel de rentabilidad financiera en el futuro, sobre la base de un conjunto de variables que describen la posición actual de la empresa. Además, se han realizado pruebas con objeto de ver si, en el problema que nos ocupa, el aprendizaje a partir de comparaciones o preferencias obtiene un conocimiento más útil que el que se obtiene a partir de modelos basados en regresión.

Con el fin de aproximar el problema a las condiciones a las que el analista financiero puede enfrentarse, se han llevado a cabo diferentes pruebas empíricas. Estas pruebas son las siguientes:

— Prueba 1 (P1): Cada empresa se compara con el resto de empresas incluidas en la muestra. Esta prueba trata de replicar una decisión de inversión que consiste en encontrar las empresas más adecuadas para invertir de forma rentable sin tener en cuenta la adscripción sectorial.

— Prueba 2 (P2): Cada empresa se compara únicamente con las empresas de la misma división de la CNAE-93 (2 dígitos).

— Prueba 3 (P3): Cada empresa se compara sólo con las empresas del mismo sector de cuatro dígitos de la CNAE-93. Las pruebas 2 y 3 tratan de replicar la decisión de inversión cuando el analista se ve forzado a elegir una empresa de una cierta rama de actividad que se puede definir de un modo amplio (divisiones de

la CNAE-93) o usando una clasificación más específica (sectores de 4 dígitos de la CNAE-93).

— PRUEBA 4 (P4): Las empresas de la muestra se dividen en cuatro grupos determinados en función de los cuartiles de la distribución de la rentabilidad financiera a la que pertenecen. Cada empresa se compara sólo con diez empresas seleccionadas aleatoriamente de los otros tres grupos que no son aquél al que pertenece la empresa.

— PRUEBA 5 (P5): Del conjunto de cuatro grupos definidos en la prueba P4, se eliminan los dos centrales. De los dos restantes, cada empresa se compara con otras 10 seleccionadas aleatoriamente del grupo al que no pertenece la empresa. Tanto esta prueba como la anterior no pretenden replicar decisiones de inversión reales a que los analistas deban hacer frente, sino que están orientadas a la comparación del modelo de análisis que proponemos con otros métodos que fueron los que inspiraron la metodología de las investigaciones previas sobre el tema. En estos trabajos, como se indicó en el Epígrafe 3, se *facilitó* la tarea del modelo, pues al sistema se le proporcionaba una clasificación de las empresas según su rentabilidad en grupos, descartándose en algunos casos las empresas de rentabilidad intermedia y analizando sólo las más y las menos rentables. Por lo tanto, pretendemos comparar el comportamiento de nuestro sistema tanto en unas condiciones reales donde cada empresa puede ser comparada con cualquier otra como en otras situaciones donde el pronóstico es *a priori* más sencillo, pues o existe una información previa acerca del cuartil al que pertenece la empresa (Prueba 4) o además de esto se han eliminado las empresas de los dos cuartiles intermedios, y el pronóstico se realiza comparando sólo las de los cuartiles extremos (Prueba 5).

Para evaluar las cinco pruebas anteriormente descritas se ha seguido el siguiente método. Cuando el número de juicios de preferencia posible es muy elevado, lo que ocurre en las pruebas en las que se comparan empresas de distintos sectores o divisiones de la CNAE-93, se han realizado experimentos de entrenamiento y test, ya que pueden generarse conjuntos de test muy grandes que dan una estimación adecuada del nivel de error cometido por los sistemas. Así, en las pruebas 1, 4 y 5 se han generado conjuntos de entrenamiento en los que cada empresa

se compara solamente con 3 empresas (elegidas al azar), mientras en los conjuntos de test cada empresa se compara con otras 30. Obtenemos así conjuntos de test con más de 30.000 juicios de preferencias. Por otra parte, en las pruebas en la que se comparan empresas de la misma división o sector de la CNAE-93 (pruebas 2 y 3) el número de comparaciones posibles es mucho menor. Por este motivo, en estos casos se han realizado experimentos del tipo *hold-out* con 5 repeticiones, es decir, se han separado unos juicios de preferencias para entrenar y el resto se ha utilizado para evaluar lo aprendido, repitiendo este proceso 5 veces.

6. Resultados experimentales

Para interpretar los resultados de la regresión utilizamos el *error relativo medio* (erm), que se calcula a partir del *error absoluto medio* (eam) de la función f obtenida por el método de regresión utilizado. Siento T_t un conjunto de prueba, podría expresarse matemáticamente de la siguiente manera:

$$eam\,(f) = \frac{1}{|T_t|} \sum_{x \in T_t} |f\,(x) - clase\,(x)| \quad erm\,(f) = 100.\frac{eam\,(f)}{eam\,(media)} \qquad [21]$$

donde *media* es el predictor constante que siempre devuelve el valor medio de todas las clases y $clase(x)$ es la clase de la empresa x. El *erm* trata de medir la mejora relativa de un método de regresión frente al método más trivial (*media*). Por tanto, si el *erm* de una función f es 100, el *eam* entre las clases predichas y las reales es el mismo que el dado por el predictor que siempre predice la media de las clases. Si el *erm* de f es 0, quiere decir que la predicción efectuada por f es perfecta. Asumimos que el predictor *media* no es perfecto, es decir, asumimos que la clase no es constante.

Con el fin de analizar el comportamiento de la regresión, hemos utilizado dos métodos: una regresión lineal simple y un reputado algoritmo de regresión no lineal conocido como $M5'$ (Quinlan, 1992; Wang y Witten, 1997).

Dado que la mayoría de las empresas tienen una rentabilidad muy parecida, hemos decidido crear dos conjuntos de entrenamiento para los algoritmos de regresión: 1) *cjto1745*, con todas las empresas y 2) *cjto873*, eliminando las empresas pertenecientes al segundo y tercer cuartil. En ambos casos, el número que aparece en el nombre indica el

número de empresas que forman la muestra. La necesidad del conjunto *cjto873* surge por el hecho de que un número bastante grande de empresas del conjunto *cjto1745*, las de los cuartiles centrales, tienen casi la misma rentabilidad. Eso hace que la regresión tienda a generar una función constante que asigna siempre la misma rentabilidad para cualquier empresa. Si tenemos como objetivo obtener funciones capaces de ordenar las empresas según su rentabilidad, esa función constante no sirve para este fin. Al eliminar las empresas de los cuartiles centrales y obtener el conjunto *cjto873*, estamos suministrando a la regresión la información sobre los dos extremos en cuanto a rentabilidad: las empresas más y menos rentables, de forma que estos algoritmos puedan ver más fácilmente la diferencia entre ambos tipos de empresas.

CUADRO 6
Resultados en erm obtenido por los algoritmos de regresión

	Reg. Lineal		Reg. no Lineal	
	cjto1745	cjto873	cjto1745	cjto873
Prueba 1	94,71%	97,58%	100%	96,03%
Prueba 2	93,52%	94,15%	100%	92,03%
Prueba 3	95,46%	96,57%	100%	95,24%
Prueba 4	94,78%	97,79%	100%	96,55%
Prueba 5	90,58%	89,71%	100%	83,13%
Media	93,81%	95,16%	100%	92,60%

En el Cuadro 6 se pueden apreciar los resultados obtenidos por los algoritmos de regresión al aplicar su modelo a las empresas incluidas en los conjuntos de test de cada prueba. Es importante indicar que todas las empresas que aparecen en los conjuntos de test de cada prueba están incluidas en *cjto1745*; por tanto, cuando se aprende a partir de dicho conjunto el modelo evalúa empresas que ya ha visto en el entrenamiento y por tanto sale beneficiado. Aún así, los mejores resultados se obtienen con el método de regresión no lineal aprendiendo a partir del conjunto que tiene únicamente las empresas pertenecientes a los cuartiles extremos. En el caso de la regresión no lineal nos encontramos con que tiene un *erm* de 100% si utilizamos el modelo aprendido a partir de todas las empresas. La causa de este comportamiento es que, como se comento anteriormente, el algoritmo tiene dificultades para aprender un buen modelo con este conjunto y opta por otorgar la misma valoración (la media de la rentabilidad) a todas las empresas. En general todos estos resultados son bastante malos puesto que están muy

cercanos al 100%, es decir, no se alejan demasiado de los resultados obtenidos por el predictor media.

A pesar de los malos resultados obtenidos por los modelos de regresión al tratar de predecir la rentabilidad exacta de una empresa, vamos a estudiar si son capaces de al menos ordenar las empresas de acuerdo a su rentabilidad y compararemos la ordenación que producen con la de SVM. Los resultados pueden verse en el Cuadro 7. Vemos que la regresión lineal presenta unos resultados muy pobres (falla siempre más del 60% de las comparaciones). La regresión no lineal aprendiendo a partir de todas las empresas falla todas las comparaciones; esto se debe al hecho de que, como ya se ha comentado, otorga la misma valoración a todas las empresas, lo cual, no es útil para compararlas. Sin embargo, cuando se aprende a partir del conjunto con empresas más distanciadas en rentabilidad los resultados mejoran considerablemente situándose el error en torno al 33% (falla 1 de cada 3 comparaciones) en todas las pruebas excepto en la 5, en la que el error se sitúa en un 21% (falla 1 de cada 5). Esto no es sorprendente, ya que la prueba 5 sólo compara empresas pertenecientes a los cuartiles extremos.

CUADRO 7
Porcentaje de comparaciones falladas

	SVM	Reg. Lineal		Reg. no Lineal	
		cjto1745	cjto873	cjto1745	cjto873
Prueba 1	20,85%	71,00%	70,62%	100%	34,57%
Prueba 2	13,69%	68,41%	68,03%	100%	32,43%
Prueba 3	14,60%	62,41%	61,77%	100%	34,01%
Prueba 4	11,97%	70,90%	70,40%	100%	31,98%
Prueba 5	0,82%	75,45%	74,69%	100%	20,90%
Media	12,39%	69,63%	69,10%	100%	30,78%

En el Cuadro 7, también se puede ver el porcentaje de comparaciones fallado por el algoritmo que aprende una función de *ranking*, el SVM con un *kernel* polinómico de grado 6 adaptado para trabajar con juicios de preferencias. Los resultados obtenidos con este método son mucho mejores que los obtenidos con los algoritmos de regresión. Cuando se comparan todas las empresas se falla un 20,85% (1 de cada 5 comparaciones), si se comparan las empresas de la misma división o sector de cuatro dígitos de la CNAE-93 el error baja al 13,69% y al 14,60%, respectivamente (fallan 1 de cada 7), si se comparan por cuartiles se falla el 11,97% de las comparaciones (1 de cada 8) y si se comparan

solo las empresas pertenecientes a los cuartiles extremos se tiene un error inferior al 1% (falla 1 de cada 122 comparaciones).

Otra fuente relevante de información es conocer el grado de importancia de las variables implicadas en el aprendizaje. Para ello hemos utilizado un algoritmo basado en la ortogonalización de Gram-Schmidt (para más explicación, puede verse Quevedo *et al.,* 2005). Se trata de un algoritmo iterativo que en cada iteración realiza tres pasos:

1) Entre todos los factores disponibles se busca el que mejor explique la variable objetivo. Siendo más explícitos, se trata de calcular para cada factor el ángulo que forma respecto a la variable objetivo. El factor que forme un ángulo menor, esto es que maximice $cos2(factor, var\_objetivo)$, será el más relevante,

2) Se proyecta la variable objetivo sobre el espacio de todos los factores seleccionados en el paso 1.

3) Si la norma de esa proyección es muy próxima a cero, se considera que el espacio formado por los factores seleccionados es capaz de explicar la variable objetivo. En ese caso el algoritmo se detiene. En otro caso, se vuelve al paso 1.

En nuestro caso, no se emplean solamente los factores originales, sino que se incluyen además potencias de los mismos (hasta grado 6). El orden de importancia de cada factor para cada prueba puede verse en el Cuadro 8.

CUADRO 8
Factores en orden de importancia para cada prueba

Prueba 1	Prueba 2	Prueba 3	Prueba 4	Prueba 5
2	2	2	2	2
5	8	8	5	5
4	4	5	4	4
1	1	1	1	1
8	7	4	8	8
7	6	7	3	7
3	5	6	7	6
	3	3		3

1. Liquidez a corto plazo, 2. Margen de beneficio, 3. Tamaño, 4. Endeudamiento, 5. Crecimiento, 6. Costes por empleado, 7. Ciclo de explotación, 8. Rotación de activos.

Se pueden apreciar varios hechos interesantes:

1) El margen de beneficio es el factor más importante. Este resultado se encuentra en línea con el alcanzado, también para las empresas españolas, por otros autores como Fernández Álvarez y García Olalla (1991) o González Pérez (1996). De ahí la importancia que para las empresas tiene el conseguir mejorar sus márgenes como base de una rentabilidad futura.

2) La rotación es el segundo factor en importancia siempre que se compare la rentabilidad futura de empresas pertenecientes al mismo sector de actividad. Ello es un reflejo de que la rotación es un factor fuertemente influido por las características específicas de cada sector de actividad, ya que ciertos sectores presentan un fuerte inmovilizado o una baja tasa de ocupación lo que tiene incidencia sobre el ratio de rotación de activos. Por ello, no se puede utilizar para comparar empresas de diferentes sectores, pero sí para comparar unidades productivas encuadradas en la misma rama.

3) La rotación siempre aparece por detrás del margen (con independencia de la adscripción sectorial), lo que indica que aunque la rotación es importante, lo que conduce a las empresas a una elevada rentabilidad futura es una estrategia basada en aumentos del margen.

4) Al contrario que lo que concluyen muchos de los trabajos previos que se revisaron en el Cuadro 1, las variables relacionadas con los costes operativos (F6) no tienen excesivo peso.

5) Si la mayor rentabilidad se logra con una estrategia basada en aumentos en el margen y al mismo tiempo se obtiene evidencia de que no hay diferencias en costes eso quiere decir que este margen se logra a través de un aumento de ingresos, es decir, de estrategias basadas en la diferenciación del producto u otras similares que permitan que el precio de venta sea mayor.

6) El hecho de que el precio sea el factor más importante ofrece una explicación de porqué una metodología no paramétrica ofrece mejores resultados que los métodos lineales que se emplearon en investigaciones anteriores. A partir de un cierto nivel, los aumentos de precios se traducen en un aumento de la rentabilidad

cada vez menor, pues se produce una reducción en las unidades vendidas, llegando incluso a invertirse el efecto, y creándose por lo tanto una relación no lineal entre margen y rentabilidad.

7) El endeudamiento es un factor importante, si bien no tanto como la definición de una estrategia de diferenciación que conduzca a aumentar el margen. Esto también arroja luz sobre el porqué la metodología propuesta supera a las regresiones, incluso a las no paramétricas. El endeudamiento de una empresa debe mantenerse dentro de unos niveles óptimos que permitan aprovechar las ventajas del apalancamiento financiero. Niveles superiores o inferiores conducen a pérdidas de rentabilidad, bien vía penalización a través de un coste de la financiación más alto (exceso de deuda), bien a través de una menor remuneración de los capitales propios (insuficiente aprovechamiento de las fuentes de financiación ajena). Por lo tanto, una metodología basada en *kernels*, que identifica núcleos formados por empresas con un comportamiento óptimos, parece más adecuado que la definición de funciones, lineales o no lineales.

8) Aunque menos que el margen, la rotación y el endeudamiento, los factores de crecimiento (F_5) y liquidez (F_1) también tienen importancia. En este punto se corroboran los resultados de investigaciones anteriores (Gillingham, 1980; De Andrés, 2001).

9) El factor tamaño (F_3) se encuentra siempre entre los peores. Dada la metodología no lineal y no paramétrica que se ha empleado en el presente estudio, queda confirmado que la escasa significación del tamaño se debe a la inexistencia de economías de escala (en el conjunto de empresas estudiadas) y no a limitaciones en la metodología.

7. Sumario y conclusiones

Tradicionalmente, para la previsión de fenómenos económicos se han propuesto dos enfoques diferentes: los modelos de regresión y los sistemas clasificadores. Sin embargo, para ciertas tareas, como el análisis de rentabilidad, ambos enfoques llevan a resultados insatisfactorios. El análisis de regresión ofrece modelos con bajo poder explicativo, mientras que las técnicas de clasificación no proporcionan al analista la información que necesita para la valoración de las inversiones.

Para superar los inconvenientes señalados, los investigadores en el campo de la Inteligencia Artificial han desarrollado algoritmos cuyo objetivo es pronosticar, para cada par de individuos cuál tendrá una mejor posición en una determinada variable partiendo de un conjunto de indicadores. Este enfoque, basado en comparaciones, tiene una ventaja fundamental para la toma de decisiones de inversión por parte del analista financiero, puesto que permite ordenar los individuos. Esta posibilidad es de gran interés cuando se trata de gestionar unos recursos limitados, puesto que habrá que seleccionar las mejores opciones con el fin de maximizar la rentabilidad. El presente artículo comprueba empíricamente la validez de un modelo basado en el enfoque de preferencias.

En la presente investigación se ha definido una tarea que consiste en construir un modelo para la determinación de los factores influyentes en la rentabilidad financiera futura de una empresa a través de un enfoque basado en preferencias. En el diseño de la investigación se han planteado tanto pruebas que intentan replicar las condiciones reales en que trabajan los analista financieros como tests más sencillos basados en la metodología de las investigaciones previas sobre la rentabilidad empresarial, a fin de realizar una comparación exhaustiva de los resultados obtenidos con los que nos ofrecen las técnicas de regresión.

Los resultados obtenidos muestran que para lograr una mayor rentabilidad futura es más efectivo adoptar las estrategias dirigidas a aumentar el margen comercial que aquellas otras que persiguen incrementar la rotación. Estos aumentos en el margen se consiguen principalmente a través de aumentos en los precios, dado que el análisis no muestra diferencias significativas entre las estructuras de costes de las empresas más y menos rentables. Por otra parte, y corroborando los resultados de estudios anteriores, encontramos que el apalancamiento y los factores de crecimiento y liquidez contribuyen a predecir la rentabilidad futura. Sin embargo, el tamaño ejerce muy poca influencia, con lo que queda comprobado que la escasa relevancia de esta variable en la literatura previa no se debía a limitaciones metodológicas.

En relación a la comparación entre el modelo de análisis propuesto y los sistemas basados en regresiones, se puede señalar que los sistemas de aprendizaje de preferencias aportan ventajas sobre los enfoques tradicionales puesto que en todas las pruebas las tasas de error obtenidas son sensiblemente inferiores a los modelos de regresión tanto lineal como no paramétrica. Ello se debe a que la naturaleza de alguno de los

factores más influyentes, como el margen y el apalancamiento financiero, implica que lo que conduce a una alta rentabilidad es un conjunto de valores situado en torno a un óptimo. Un sistema como el que se propone, basado en *kernels*, detecta mejor esos valores óptimos que un modelo de regresión, más dirigido a la identificación de relaciones monotónicas de dependencia (lineales o no lineales).

Como cierre, señalar que la aplicación de esta metodología ofrece múltiples posibilidades en el ámbito económico-financiero. Así, pueden ser una útil herramienta para los estudios sobre solvencia empresarial y predicción de quiebra partiendo de una serie de indicadores económicos y financieros.

Referencias

Altman, E.I. (1968): "Financial ratios, discriminant analysis and the prediction of the corporate bankruptcy", *Journal of Finance* 230, pp. 589-609.

Amit, R. y P. Schoemaker (1993): "Strategic assets and organizational rent", *Strategic Management Journal* 14, pp. 33-46.

Arraiza Antón, C. y A. Lafuente Félez (1984): "Caracterización de la gran empresa industrial española según su rentabilidad", *Información Comercial Española,* julio, pp. 127-139.

Bahamonde, A., G.F. Bayón, J. Díez, J.R. Quevedo, O. Luaces, J.J. del Coz, J. Alonso, y F. Goyache (2004): "Feature subset selection for learning preferences: A case study", en *Proceedings of the International Conference on Machine Learning (ICML '04),* Ban, Alberta (Canada).

Barney, J. (1991): "Firm resources and sustained competitive advantage", *Journal of Management* 17, pp. 99-120.

Bell, T.B., G.S. Ribar y J.R. Verchio (1990): "Neural nets versus logistic regression: a comparison of each model's ability to predict commercial bank failures", *Deloitte & Touche/University of Kansas Simposium of Auditing Problems,* pp. 29-53.

Boser, B.E., I. Guyon, y V. Vapnik (1992): "A training algorithm for optimal margin classifiers", en *Computational Learning Theory,* pp. 144-152.

Brief, R.P. y R.A. Lawson (1992): "The role of the accounting rate of return in financial statement analysis", *Accounting Review* 67, pp. 411-426.

Bueno, E. y P. Lamothe (1983): "Tamaño y rentabilidad de la gran empresa española. Un análisis empírico de su relación basado en un método multicriterio", Comunicación presentada al II Congreso AECA, Santa Cruz de Tenerife.

Buzell, R.D. y B.T. Gale (1987), *The PIMS Principles: Linking Strategy to Performance* Free Press, New York, USA.

Caloghirou, Y., A. Protogerou, Y. Spanos y L. Papagiannakis (2004): "Industry versus firm-specific effects on performance: contrasting SMEs and large-sized firms", *European Management Journal* 22, pp. 231-243.

Chaganti, R. y R. Chaganti (1983): "A profile of profitable and not-so-profitable small businesses", *Journal of Small Business Management,* July, pp. 43-51.

Claver, E., J. Molina y J. Tarí (2002): "Firm and industry effects on profitability: a Spanish empirical análisis", *European Management Journal* 20, pp. 321-328.

Cohen, W.W., R.E. Shapire y Y. Singer (1999): "Learning to order things", *Journal of Artificial Intelligence Research* 10, pp. 243-270.

Cristianini N. y J. Shawe-Taylor (2000), *An Introduction to Support Vector Machines and other Kernel-based Learning Methods,* Cambridge University Press.

De Andrés, J. (2001): "Statistical techniques vs. SEE5 algorithm. An application to a small business environment", *International Journal of Digital Accounting Research* 1, pp. 153-178.

De Andrés, J., M. Landajo, y P. Lorca (2005): "Forecasting business profitability by using classification techniques: a comparative analysis based on a Spanish case", *European Journal of Operational Research* 167, pp. 518-542.

Díez, J., G.F. Bayón, J.R. Quevedo, J.J. del Coz, O. Luaces, J. Alonso y A. Bahamonde (2004): "Discovering relevancies in very difficult regression problems: applications to sensory data analysis", *Proceedings of the European Conference on Artificial Intelligence (ECAI 2004),* pp. 993-994.

Eisenbeis, R.A. (1977): "Pitfalls in the application of discriminant analysis in business, finance, and economics", *Journal of Finance* 32, pp. 875-900.

Fahy, J. (2000): "The resource-based view of the firm: some stumbling-blocks on the road to understanding sustainable competitive advantage", *Journal of European Industrial Training* 24, pp. 94-104.

Fernández Alvarez, A.I. y M. García Olalla (1991): "Análisis del comportamiento económico-financiero de los sectores empresariales en España", *ESIC-Market,* abril-junio, pp. 113-128.

Fernández Sánchez, E.; J.M. Montes Peón y C.J. Vázquez Ordás (1996): "Caracterización económico-financiera de la gran empresa industrial española según su rentabilidad", *Revista Española de Financiación y Contabilidad* 87, pp. 343-359.

Fiechter, C. y S. Rogers (2000): "Learning subjective functions with large margins", *Proceedings of the 17th International Conference on Machine Learning,* Stanford, California, USA. Morgan Kaufmann, pp. 287-294.

Foster, G. (1998), *Financial Statements Analysis (2nd ed)* Pearson Education, New York, USA.

Frydman, H.; E.I. Altman y D.L. Kao (1985): "Introducing recursive partitioning for financial classification: the case of financial distress", *Journal of Finance* 40, pp. 269-291.

Gillingham, D.W. (1980): "A comparison between the attribute profiles of profitable and unprofitable companies in the United Kingdom and Canada", *International Review* 20, pp. 64-73.

Goddard, J., M. Tavakoli y J.O.S. Wilson (2005): "Determinants of profitability in European manufacturing and services: evidence from a dynamic panel model", *Applied Financial Economics* 15, pp. 1269-1282.

González Pérez, A.L. (1996), *La rentabilidad empresarial. Evaluación empírica de sus factores determinantes.* Colegio de Registradores de la Propiedad y Mercantiles de España, Madrid.

Gort, M. (1963): "Analysis of stability and change in market shares", *Journal of Political Economy* 71, pp. 51-63.

Grant, R.M. (1991): "The Resource-Based Theory of Competitive Advantage: Implications for Strategy Formulation", *California Management Review* 33, pp. 114-135.

Hall, R. (1989): "The management of intellectual assets: a new corporate perspective", *Journal of General Management* 15, pp. 53-68.

Harris, M.N. (1976): "Entry and barriers to entry", *Industrial Organization Review* 3, pp. 165-175.

Haslem, J.A. y W.A. Longbrake (1971): "A discriminant analysis of commercial bank profitability", *Quarterly Review of Economics & Business* 11, pp 39-46.

Herbrich, R.; T. Graepel y K. Obermayer (1999): "Support Vector Learning for ordinal regression", *Proceedings of the Ninth International Conference on Artificial Neural Networks* Edinburgh, UK, pp. 97-102.

Horowitz, I. (1984): "The misuse of accounting rates of return: a comment", *American Economic Review* 74, June, pp. 492-493.

Izan, H.Y. (1984): "Corporate distress in Australia", *Journal of Banking and Finance* 8, pp. 303-320.

Joachims, T. (2002): "Optimizing search engines using clickthrough data", *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD).*

Johnson, R. y L. Soenen (2003): "Indicators of successful companies", *European Management Journal* 21, pp. 364-369.

Kay, J. (1976): "Accountants, too, could be happy in a golden age: the accountant's rate of profit and the internal rate of return", *Oxford Economic Papers,* November, pp. 447-460.

Kelly, G. y M. Tippet (1991): "Economic and accounting rates of return: a statistical model", *Accounting & Business Research* 21, pp. 321-329.

Kosko, B. (1992), *Neural Networks and Fuzzy Systems. A Dynamical Systems Approach to Machine Intelligence* Prentice-Hall, Englewood Cliffs, New Jersey, USA.

Kosko, B. (1994): "Fuzzy systems as universal approximators", *IEEE Transactions on Computers* 43, pp. 1329-1333.

Lafuente Félez, A. y V. Salas Fumás (1983): "Concentración y resultados de las empresas en la economía española", *Cuadernos Económicos del ICE* 22/23, pp. 7-21.

Lau, H.S., A. Hing-Ling y D.W. Gribbin (1995): "On modelling cross sectional distributions of financial ratios", *Journal of Business Finance & Accounting* 22, pp. 521-549.

Marais, M.L., J.M. Patell y M.A. Wolfson (1984): "The experimental design of classification models: an application of recursive partitioning and bootstrap", *Journal of Accounting Research* 22, pp. 87-118.

Maravall, F. (1976), *Crecimiento, dimensión y concentración de las empresas industriales españolas,* Fundación del INI. Serie E, 7.

Martikainen, T., J. Perttunen, P. Yli-Olli y A. Gunasekaran (1995): "Financial ratio distribution irregularities: implications for ratio classification", *European Journal of Operational Research* 80, pp. 34-44.

Martin, D. (1977): "Early warning of bank failure: a logit regression approach", *Journal of Banking & Finance* 1, pp. 249-276.

Martin, S. (2002), *Advanced Industrial Economics* Blackwell Publishers, Oxford, UK.

McGahan, A. y M. Porter (1997): "How much does industry matter, really?", *Strategic Management Journal* 18, pp. 15-30.

Odom, M.D. y R. Sharda (1992): "A neural network model for bankruptcy prediction", en *Neural networks in Finance and Investing*, Probus Publishing, pp. 163-168.

Peteraf, M. (1993): "The cornerstones of competitive advantage: a resource-based view", *Strategic Management Journal* 14, pp. 179-191.

Petitbó, A. (1982), *Aproximació a L'estudi dels Elements Explicatius de la Rendibilitat de las Grans Empreses Industrials Espanyoles,* Papers de Seminari 20, pp. 36-49, Centre d'Estudis de Planificació.

Platt, H.D. y M.B. Platt (1990): "Development of a class of stable predictive variables: the case of bankruptcy prediction", *Journal of Business Finance and Accounting* 17, pp. 31-51.

Quevedo, J.R., G.F. Bayón, O. Luaces y A. Bahamonde (2005): "Ensembling an algebraic procedure for ranking features by their usefulness when learning consumers' preferences", Technical Report. Centro de Inteligencia Artificial. Universidad de Oviedo.

Quinlan, J.R. (1992): "Learning with continuous classes", en *Procedings of the Fifth Australian Joint Conference on Artificial Intelligence,* World Scientific, Singapore, pp. 343-348.

Scherer, F.M. y D. Ross (1990), *Industrial Market Structure and Economic Performance (3rd ed.)* Houghton Mifflin, Boston, Massachussets, USA.

Schmalensee, R. (1985): "Do markets differ much?", *American Economic Review* 75, pp. 341-351.

Serrano Cinca, C. y B. Martín del Brío (1993): "Predicción de la quiebra bancaria mediante el empleo de redes neuronales artificiales", *Revista Española de Financiación y Contabilidad* 22, pp. 153-176.

Suárez Suárez, A. (1977): "La rentabilidad y el tamaño de las empresas españolas", *Económicas y Empresariales* 5, pp. 56-63.

Teece, D.J., G. Pisano y A. Shuen (1997): "Dynamic capabilities and strategic management", *Strategic Management Journal* 18, pp. 509-533.

Tesauro, G. (1989): "Connectionist learning of expert preferences by comparison training", *Advances in Neural Information Processing Systems, Proceedings of the NIPS'88,* MIT Press, Massachussets, USA, pp. 99-106.

Utgoff, P. y S. Saxena. (1987): "Learning a preference predicate", *Proceedings of the Fourth International Workshop on Machine Learning,* Morgan Kaufmann, Irvine, California, USA, pp. 115-121.

Vapnik, V. (1998), *Statistical Learning Theory,* John Wiley, New York, USA.

Walter, J.E. (1959): "A discriminant function for earnings-price ratios of large industrial corporations", *Review of Economics & Statistics* 41, pp. 44-52.

Wang Y. y I.H. Witten (1997): "Induction of model trees for predicting continuous classes", en *European Conference on Machine Learning,* pp. 128-137.

Weir, C. (1996): "Internal organization and firm performance: an analysis of large UK firms under conditions of economic uncertainty", *Applied Economics* 28, pp. 473-481.

Whittington, G. (1980): "Some basic properties of accounting ratios", *Journal of Business Finance & Accounting* 7, pp. 219-232.

Woo, C.Y. (1983): "Evaluation of the strategies and performance of low ROI market share leaders", *Strategic Management Journal* 4, pp. 123-135.

Abstract

In this paper we estimate a model for the determination of the factors influencing the future profitability of a firm. We use a preference-based machine learning system. Under this approach it is possible to overcome the drawbacks of regression models and classification techniques. The main results estimated from a database of 1.745 firms indicate that strategies based on profit margin increases involving price raises are more effective than those based on increasing the turnover of assets. It is noticeable that our model is much more accurate than those based on regression techniques.

Keywords: ratios, financial profitability, preference analysis.

Recepción del original, mayo de 2004
Versión final, enero de 2007