

Aproximación a un problema financiero mediante redes neuronales con funciones base radiales y máquinas de soporte vectorial

Xavier Parra, Cecilio Angulo
ESAII - GREC

Universitat Politècnica de Catalunya
Av. Víctor Balaguer s/n. Vilanova
{xparra,cangulo}@esaii.upc.es

Núria Agell, Xari Rovira
ESADE - GREC

Universitat Ramon Llull
Av. Pedralbes 62. Barcelona
{agell,rovira}@esade.es

Resumen

En el artículo que se presenta se utilizan técnicas conexionistas de aprendizaje para reproducir el proceso de evaluación en la predicción del riesgo de crédito de las empresas. Como este proceso no se basa en información únicamente cuantitativa, sino que además tiene una fuerte dependencia del conocimiento de los expertos, parece razonable la utilización de estos métodos. Las técnicas de aprendizaje aplicadas son, en primer lugar, redes neuronales con funciones base radiales, y, en segundo lugar, máquinas de soporte vectorial. En ambos casos ha sido necesario adaptar la información disponible, es decir los datos financieros de la empresa, para mejorar su funcionamiento.

1 Introducción

El rating es una opinión cualificada sobre la calidad o riesgo de crédito de un activo financiero o de un emisor, y se considera actualmente una guía muy importante para los mercados financieros y para los inversores de estos mercados [1]. El presente artículo pretende modelizar el proceso que siguen las agencias que analizan y valoran el riesgo de crédito de las empresas para asignarles un rating, utilizando para ello determinados sistemas de aprendizaje conexionistas. Las dos agencias de rating más importantes son Moody's y Standard & Poor's. Estas agencias clasifican a las empresas según su nivel de riesgo, utilizando información cuantitativa y cualitativa. La simbología que hoy en día utilizan ambas agencias es extremadamente similar. En concreto, las etiquetas que utiliza Standard &

Poor's, que son las que se han considerado en este artículo, son: AAA, AA, A, BBB, BB, B, CCC, CC. De izquierda a derecha esta clasificación va de “muy alta” a “muy baja” calidad de crédito, es decir, de mucha a muy poca capacidad de la empresa para afrontar la deuda. Puesto que la etiqueta que se asigna a una empresa es el diagnóstico de la agencia y refleja su probabilidad de quiebra, esta predicción requiere un sólido conocimiento de los parámetros que indican su situación y de las relaciones que existen entre ellos, junto con los factores que pueden alterarlos. Los procesos usados por estas agencias son muy complejos.

Las técnicas de decisión involucradas no están basadas en modelos puramente numéricos. Por un lado, utilizan la información dada por los datos financieros y los valores que les influyen, por otro, analizan el

sector y el país o países donde se presenta la empresa, y tienen en cuenta las posibilidades de crecimiento del negocio y de su posición competitiva. Finalmente, los analistas hacen una evaluación global abstracta basada en su propia experiencia para determinar el rating.

Identificar y reproducir el razonamiento humano, es uno de los objetivos principales de la Inteligencia Artificial. Así pues el uso de procedimientos propios de esta disciplina, como son las técnicas conexionistas utilizadas en este trabajo, parece adecuado para reproducir estos procesos de razonamiento realizados por las agencias de rating, aún más teniendo en cuenta que la clasificación final viene dada en una escala ordinal. En primer lugar, en la sección 2, enfocamos el problema utilizando redes neuronales con funciones base radiales (*RNFBR*). Introducida la arquitectura de dichas redes neuronales, se adaptan las variables de entrada para el proceso de aprendizaje, teniendo en cuenta la importancia de los órdenes de magnitud de los factores que intervienen el cálculo del rating. En segundo lugar, en la sección 3, se introducen las máquinas de soporte vectorial (*SV*), para abordar el mismo problema. El algoritmo de aprendizaje utilizado, el *K-SVCR*, es una extensión de los métodos de vectores soporte que permite controlar el grado de generalización de las clasificaciones en una escala ordinal. Finalmente se discuten los resultados obtenidos en cada caso llegando a unas conclusiones generales y sugiriendo posibles alternativas.

2 Aproximación mediante redes neuronales con funciones base radiales

2.1 Introducción

La razón principal que motiva la aproximación conexionista propuesta en el presente trabajo es la enorme dependencia que existe de la solución del problema con respecto al conocimiento del experto. No obstante, una de las decisiones clave que influye decisivamente en la calidad de las soluciones que se pueden obtener la constituye la manera en que se representan las variables de entrada y de salida del problema de aprendizaje en una implementación mediante redes neuronales artificiales. Además, esto es especialmente importante cuando se dispone de información cualitativa para realizar el en-

trenamiento.

La presente aplicación muestra la manera en que técnicas propias de las redes neuronales artificiales combinadas con técnicas propias del razonamiento cualitativo (en concreto, con el cálculo de los órdenes de magnitud [8]) pueden resultar útiles en el campo de las finanzas [6].

En un primer apartado se presenta una breve introducción a la arquitectura neuronal utilizada durante el desarrollo. Inmediatamente después se establece el proceso a seguir para preparar las escalas de referencia de las distintas variables cualitativas que deben operarse para aportar una solución al problema. La aplicación efectiva de la red neuronal artificial para evaluar el riesgo crediticio de una empresa se presenta en el apartado que precede a la discusión de los resultados que se derivan del trabajo desarrollado.

2.2 Arquitectura de las redes neuronales con funciones base radiales

Las redes neuronales artificiales con funciones base radiales resultan especialmente indicadas para la resolución del problema planteado dado que se caracterizan por ser clasificadores universales [9]. Este tipo de redes se han asociado tradicionalmente con una arquitectura simple de tres capas [2] (véase Figura 1), en que cada capa de la arquitectura está completamente conectada con la capa inmediatamente consecutiva. La capa oculta se compone de un conjunto de nodos que se caracterizan por tener asociadas unas funciones de activación de tipo radial, denominadas funciones base radiales. Dichas funciones radiales reciben como entrada todos y cada uno de los atributos de los patrones, y se caracterizan por estar centradas en un punto del espacio de entrada. Las salidas de estas funciones se combinan linealmente mediante unas ponderaciones para generar la salida de la red neuronal. Una característica importante de este tipo de funciones radiales es que generan una respuesta local (en oposición a la respuesta global característica de la función sigmoide) puesto que su salida solamente depende de la distancia que existe entre la entrada y el centro de cada función radial.

Las funciones radiales de la capa oculta presentan una estructura que se puede representar de la siguiente manera:

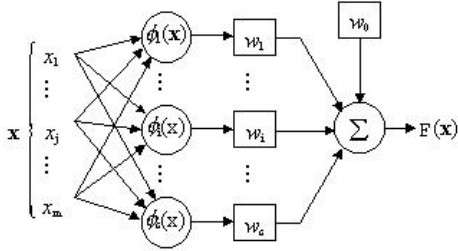


Figura 1: Arquitectura de las redes neuronales artificiales con funciones base radiales.

$$\phi_i(\mathbf{x}) = \varphi((\mathbf{x} - \mathbf{c}_i)^T \mathbf{R}^{-1}(\mathbf{x} - \mathbf{c}_i)), \quad (1)$$

donde ϕ es la función radial utilizada, $\{\mathbf{c}_i \mid i = 1, 2, \dots, c\}$ es el conjunto de centros de función radial y \mathbf{R} es una métrica. El término $(\mathbf{x} - \mathbf{c}_i)^T \mathbf{R}^{-1}(\mathbf{x} - \mathbf{c}_i)$ representa la distancia desde la entrada \mathbf{x} al centro \mathbf{c} en la métrica definida por \mathbf{R} . Existen diversos tipos de funciones radiales que tradicionalmente suelen utilizarse, aunque la función radial gaussiana es la utilizada de manera habitual, combinada con la métrica euclídea. En este caso, la salida de la red neuronal con funciones base radiales es:

$$F(x) = w_0 + \sum_{i=1}^c w_i \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}_i\|^2}{r^2}\right) \quad (2)$$

donde c es el número de funciones base utilizadas, $\{\mathbf{c}_i \mid i = 1, 2, \dots, c\}$ son los pesos sinápticos, $\|\cdot\|$ denota la norma euclídea y r es el radio de la función radial.

El algoritmo de aprendizaje de las redes neuronales con funciones base radiales es un proceso incremental y evolutivo. El fundamento matemático sobre el que se basa dicho proceso es la denominada *selección de subconjuntos* y consiste en comparar los distintos modelos que se obtienen de escoger, de entre un mismo conjunto de elementos candidatos, distintos subconjuntos de elementos. Habitualmente, la determinación del mejor subconjunto de elementos resulta computacionalmente intratable por lo que debe recurrirse a la heurística para intentar determinar una pequeña e interesante fracción de entre el espacio de todos los subconjuntos posibles. No obstante, debemos tener presente que el uso de dicha heurística no garantiza que las soluciones obtenidas incluyan el menor número de elementos que se necesitan para reducir el error de aproximación hasta un determinado valor.

El método heurístico denominado *selección progresiva* es uno de los métodos disponibles y que con mayor frecuencia suele utilizarse para entrenar las redes neuronales con funciones base radiales [4]. De acuerdo con este método, el subconjunto que debe determinarse durante el aprendizaje es el subconjunto de centros de funciones radiales que permite localizar dichas funciones en el espacio de entrada del problema. El método parte de un subconjunto vacío de centros al cuál se le va añadiendo, a cada paso del algoritmo, un nuevo centro. El centro de función radial se selecciona de entre el conjunto de todos los patrones de entrada del problema y, en concreto, es aquel que permite conseguir una mayor reducción del error de aproximación. El proceso de aprendizaje continúa hasta que algún criterio de selección del modelo deja de decrecer (por ejemplo, validación cruzada generalizada o inferencia bayesiana).

2.3 Adaptación cualitativa de las entradas

El problema de la extracción de conocimiento a partir de valores cualitativos representados a partir de referencias heterogéneas no es inusual. Son diversas las técnicas propias del razonamiento cualitativo que se suelen utilizar para manejar este tipo de referencias. No obstante, cuando los valores deben ser procesados con una red neuronal artificial, no resulta evidente el modo en que dichos valores deben prepararse para poder tener en consideración el conocimiento que los expertos pueden poseer sobre ellos.

En general, el rendimiento que se obtiene cuando se trabaja con redes neuronales artificiales tiene una dependencia significativa de la representación del problema que se utiliza, es decir, del modo en que se representan las entradas y las salidas del problema. Además, cuando el tipo de red neuronal utilizado es una red neuronal con funciones base radiales dicha dependencia de la representación resulta más crítica dado que, como podemos observar en el apartado anterior, la salida de la función radial es una función directa de una distancia definida en el espacio de entrada. Dependiendo del tipo de problema, pueden haber distintos tipos de variables que deben representarse. Desgraciadamente, no existe un único método para representar todos y cada uno de los posibles tipos de variables. Sin embargo, resulta común utili-

zar alguna de las siguientes recomendaciones:

- *Atributos reales* suelen transformarse a través de una función lineal que permite trasladar un valor original al rango $0 \dots 1$ ó $-1 \dots +1$, de manera que se consiga una distribución más uniforme en dicho rango. También es frecuente transformar linealmente los valores de manera que su media resulte cero y su desviación estándar uno.
- *Atributos nominales* con m valores distintos suelen representarse una codificación 1-de- m , o bien directamente en código binario.
- *Atributos ordinales* con m valores distintos suelen representarse con $m - 1$ variables de las cuales las primeras k toman un valor igual a uno para representar que el valor k -ésimo atributo mientras que el resto de variables toman un valor igual a cero.

Como ya se ha indicado con anterioridad, la resolución de numerosos problemas financieros parte del conocimiento preciso de diversos índices y valores que, de algún modo, indican situaciones financieras distintas. Normalmente, dichos índices y valores se representan a través de un valor real aunque, con frecuencia, de donde realmente el experto extrae información no es directamente de dicho valor numérico, sino de alguna representación más cualitativo de éste. Por ejemplo, frente a un valor concreto observado por el experto resulta más probable que piense en términos de *bueno*, *malo*, *muy bueno*, etc., que en cualquier otro modo. Así, por un lado tenemos una información financiera representada a partir de valores numéricos reales y por otro lado tenemos el conocimiento del experto que acostumbra a tratar cualitativamente dicha información. Resulta interesante establecer o determinar alguna manera de combinar tanto la representación cuantitativa del problema como la representación cualitativa del experto, con objeto de preparar las variables para el proceso de aprendizaje.

Supongamos que, para cada una de las variables involucradas en el problema, el experto establece un conjunto de marcas formado por número impar n de valores $L = \{l_1, l_2, \dots, l_i, \dots, l_n\}$ que nos permiten describir cualitativamente la variable (dichos valores determinan intervalos en la recta real). De entre el conjunto de marcas podemos identificar un valor central l_i característico para cada una de

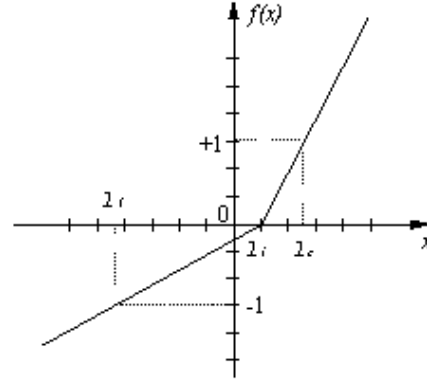


Figura 2: Discretización de la recta real a partir de las marcas del experto.

las variables del problema. La preparación de las variables del problema para utilizarlas durante el entrenamiento de las redes neuronales artificiales con funciones base radiales se concreta en dos pasos:

1. La transformación de la marca central l_i a cero, a través de una traslación:

$$t_i : \mathbb{R} \rightarrow \mathbb{R}, t_i = x - l_i.$$

2. La transformación de los valores a través de una función $f : \mathbb{R} \rightarrow \mathbb{R}$,

$$f(x) = \begin{cases} +s \cdot t_i(x)/l_r & \text{si } t_i(x) \geq 0 \\ -s \cdot t_i(x)/l_l & \text{si } t_i(x) < 0 \end{cases} \quad (3)$$

donde s es el signo de la transformación que determina el experto y l_r y l_l denotan respectivamente las marcas derecha e izquierda, con relación a la marca central l_i .

Una vez se han realizado los dos pasos descritos, todos los valores de las variables pasan a estar representados en un rango de valores similar, aunque para ello se realizan discretizaciones distintas para cada variable dado que el conjunto de marcas es distinto para cada una de las variables (véase Figura 2).

2.4 Experimentos y resultados

Inicialmente, la base de datos de partida constaba de 495 patrones. Cada uno de los patrones de entrada estaba constituido por 12 atributos cuantitativos, 1 atributo cualitativo (el sector de producción de la empresa) y 1 salida. Los patrones que presentaban algún atributo con

valor desconocido fueron eliminados de la base de datos. A partir de las recomendaciones de los expertos y debido a la especial peculiaridad de unos de los sectores de actividad (el sector tecnológico), las compañías tecnológicas también fueron eliminadas del conjunto de patrones original. El siguiente paso fue identificar y seleccionar las variables que, según el criterio del experto, eran más importantes y relevantes cuando lo que se pretende es evaluar al riesgo crediticio. Después de realizar todas las eliminaciones y selecciones descritas, el tamaño de la base de datos pasó de 495 a 244 patrones y el espacio de entrada se redujo de 12 a 5 atributos por patrón. Las 5 variables de entrada resultantes eran de tipo real mientras que la variable de salida era de tipo nominal con 6 clases diferentes {AAA, AA, A, BBB, BB, B}, y fue representada siguiendo una codificación 1-de-6.

Llegados a este punto, existían al menos dos opciones: (1) entrenar una única red neuronal con 5 atributos por patrón y 6 salidas por patrón o (2) entrenar 6 redes neuronales distintas con 5 atributos por patrón y 1 salida por patrón. La primera de las opciones no es demasiado apropiada para el problema que nos planteamos dado el reducido número de patrones disponibles en la base de datos para realizar el entrenamiento. La segunda opción es más eficiente desde el punto de vista de rendimiento y optimización de los recursos. Aunque el tamaño final de la arquitectura neuronal resultante del aprendizaje probablemente será menor para la red neuronal individual y su aprendizaje será más rápido, también es cierto que su capacidad de generalización será menor. Para el problema planteado se considera que la capacidad de generalización es más prioritaria que el tamaño de la red neuronal o la velocidad de aprendizaje, por lo que se ha optado por la segunda de las opciones descritas anteriormente. Los experimentos se han realizado considerando que el problema inicial de clasificar un patrón en 1 clase entre 6 clases distintas se ha transformado en 6 problemas diferentes que consisten en clasificar un patrón en 1 única clase. Cada una de las redes neuronales obtenidas para cada uno de los 6 nuevos problemas estará especializada en determinar si un patrón es o no es de la clase sobre la que la red neuronal ha sido entrenada.

Las simulaciones realizadas se han llevado a cabo siguiendo las reglas estándar PROBEN1 [10]. El conjunto de datos disponible se ha ordenado a partir del nombre de la compañía en

orden ascendente. Posteriormente, dicho conjunto de datos se ha dividido en tres subconjuntos: conjunto de entrenamiento, conjunto de validación y conjunto de test. Las distribuciones de los patrones del conjunto de datos se ha realizado secuencialmente siguiendo unas proporciones iguales al 50%, 25% y 25%, respectivamente. En la Tabla 1 podemos observar la distribución de los patrones en cada uno de los subconjuntos. Resulta interesante observar que para la clase AAA no se dispone de patrones en el conjunto de test y que para la clase B no se dispone de patrones ni en el conjunto de entrenamiento ni en el conjunto de validación.

Rating	AAA	AA	A	BBB
Training	5	18	53	41
Validación	2	10	28	18
Test	0	7	23	27
Total	7	35	104	86

Rating	BB	B	Total
Training	5	0	122
Validación	3	0	61
Test	3	1	61
Total	11	1	244

Tabla 1: Distribución de patrones sobre subconjuntos de datos.

Con el objetivo de poder estudiar y analizar el efecto que tiene la adaptación cualitativa de los atributos de entrada del problema sobre la capacidad de generalización de la red neuronal, se han llevado a cabo dos tipos de entrenamiento. En el primer tipo de entrenamiento (denominado *entrenamiento clásico*) se ha utilizado una representación de los valores reales de entrada a partir de la transformación lineal que fija su media a cero y su desviación estándar a uno. Para éste primer tipo de entrenamiento no se ha tenido en consideración el conocimiento del experto para adaptar los atributos de entrada de los patrones. En el segundo tipo de entrenamiento (denominado *entrenamiento experto*) se ha utilizado la transformación descrita en el apartado 2.3, es decir, se ha utilizado la información relativa a los órdenes de magnitud facilitada por el experto (véase Tabla 2).

Inicialmente, las redes neuronales se han entrenado sobre el conjunto de entrenamiento mientras que el conjunto de validación se ha utilizado para ajustar el valor del radio de las

	l_l	l_i	l_r	s
V_1	1	4	10	+1
V_2	1	2	8	+1
V_3	0.02	0.07	0.15	+1
V_4	0.2	1	10	-1
V_5	0.0	0.1	0.3	+1

Tabla 2: Marcas y signos de experto.

Entrenamiento <i>clásico</i>			
	r	CA_{va}	CA_{te}
AAA	1.055	96.7%	100.0%
AA	11.2	82.0%	88.5%
A	0.831	73.7	50.8%
BBB	80.6	63.9%	57.4%
BB	5.11	95.1%	95.1%
B	111.1	100.0%	98.4%

Entrenamiento <i>experto</i>			
	r	CA_{va}	CA_{te}
AAA	51.2	98.4%	100.0%
AA	68.1	85.2%	90.2%
A	4.7	70.5	59.0%
BBB	32.5	75.4%	59.0%
BB	19.9	95.1%	95.1%
B	111.1	100.0%	98.4%

Tabla 3: Anchura de la función radial (r) y precisión en la clasificación sobre el conjunto de validación (CA_{va}) y el conjunto de test (CA_{te}).

funciones radiales (r). El radio de las funciones radiales se ha ajustado a partir de la información obtenida después de realizar 4000 simulaciones para cada clase. Los valores de radios que se han utilizado van desde 0,0001 hasta 0,1 con incrementos de 0,0001, desde 0,101 hasta 1,1 con incrementos de 0,001, desde 1,1 hasta 11,1 con incrementos de 0,01 y desde 11,2 hasta 111,1 con incrementos de 0,1. El valor final del radio (véase Tabla 3) se ha seleccionado de entre los 4000 valores posibles a partir de la aplicación de los siguientes criterios:

1. Seleccionar el radio que maximiza la precisión de la clasificación para el conjunto de validación.
2. Seleccionar el radio que, de acuerdo con el criterio (1), genera la red neuronal de menor tamaño.

	<i>Clásico</i>		<i>Experto</i>	
Correcto	13	21.3%	18	29.5%
Incorrecto	19	31.2%	18	29.5%
No Clasificado	29	47.5	25	41.0%

Tabla 4: Clasificación final sobre el conjunto de datos de test.

3. Seleccionar el radio que, de acuerdo con el criterio (2), minimiza el error cuadrático medio para el conjunto de validación.
4. Seleccionar el radio que, de acuerdo con el criterio (3), minimiza el error cuadrático medio para el conjunto de entrenamiento.
5. Seleccionar el radio que, de acuerdo con el criterio (4), maximiza la precisión de la clasificación para el conjunto de entrenamiento.

Una vez se ha determinado el radio de las funciones radiales para cada uno de los seis problemas de clasificación, las redes neuronales se entrenan sobre el conjunto de entrenamiento y validación mientras que el conjunto de test se utiliza para determinar la capacidad de generalización de la solución final. Como podemos observar en la Tabla 3, la precisión de la clasificación para el entrenamiento *experto* es mejor, o en el peor de los casos igual, que la obtenida a partir del entrenamiento *clásico*. Dado que la única diferencia que existe entre ambos tipos de entrenamiento es la utilización de las marcas determinadas por el experto para la adaptación de los valores de entrada, parece razonable suponer que el uso de dicha información durante el entrenamiento es útil. No obstante, el problema inicial no consistía en realizar seis clasificaciones independientes sino una única clasificación. Además, cada una de las seis redes neuronales con funciones base radiales entrenadas únicamente determina a su salida, para cada patrón de entrada, si dicho patrón pertenece o no a la clase sobre la que se ha realizado el entrenamiento. Así, la salida de dichas redes neuronales puede interpretarse como: *Si, el patrón es de la clase o No, el patrón no es de la clase*.

Desgraciadamente, si combinamos la clasificación de cada una de las seis redes neuronales el resultado que obtenemos no nos garantiza que sólo una de las seis clases esté activa, es decir, podemos encontrarnos con que más de

una de las salidas de las seis redes neuronales de una clasificación positiva dé un mismo patrón. Indudablemente, esto significa que un determinado patrón puede clasificarse correctamente o incorrectamente, e incluso puede llegar a no ser clasificado. La Tabla 4 recoge la triple clasificación que podemos obtener sobre el conjunto de test y, como podemos observar, el entrenamiento *experto* es nuevamente mejor que el entrenamiento *clásico*. A la vez, el entrenamiento *experto* permite obtener un menor porcentaje de indeterminación en la clasificación (41.0% frente al 47.5% del entrenamiento *clásico*) y lo mismo puede aplicarse al número de patrones clasificados de manera incorrecta (29.5% frente a 31.2%).

2.5 Discusión

El trabajo presentado muestra diversas estrategias que se pueden seguir para sintetizar información cualitativa que tenemos asociada a distintos atributos, cada uno de los cuales está descrito cualitativamente de manera distinta. El uso de la información facilitada por expertos durante la adaptación y preparación de dichos atributos para realizar el entrenamiento de redes neuronales se ha mostrado útil dado que permite aumentar el grado de generalización obtenido con dichas redes neuronales. El método de adaptación de variables propuesto se ha aplicado a un problema financiero relacionado con la predicción del riesgo crediticio. No obstante, la aproximación desarrollada también puede aplicarse a problemas propios de otras áreas siempre y cuando se disponga de variables descritas a través de órdenes de magnitud. Las limitaciones del método presentado no pueden evaluarse con profundidad hasta que la implementación sea completa y esté suficientemente validada.

3 Aproximación mediante máquinas con SV

3.1 Introducción

El problema de predicción de variables de escala ordinal se conoce como regresión ordinal y es complementario a los problemas estándares de aprendizaje de clasificación y regresión métrica. El estudio presentado en esta sección surge como una aproximación para resolver el problema de tipo económico planteado cuando aquello

que se pretende realizar es una clasificación de usuarios de forma que los diferentes grupos tienen un grado de calidad que es metrizable. El algoritmo de aprendizaje utilizado, el K -SVCR, es una extensión de los métodos de Vectores Soporte (SV) utilizados en el área de Aprendizaje Máquina que permite controlar el grado de generalización de la regresión ordinal.

Los problemas de regresión ordinal pueden ser expresados dentro del dominio de las máquinas de aprendizaje como: Sea

$$\mathcal{S} = \{(\mathbf{x}_p, y_p)\}_{p=1}^{\ell} \sim P_{XY}^{\ell}, \quad (4)$$

un conjunto independiente idénticamente distribuido (i.i.d.) de datos y

$$\mathcal{H} = \{h(\cdot) : X \rightarrow Y\}, \quad (5)$$

un conjunto de funciones, entonces un algoritmo de aprendizaje selecciona una función h^{ℓ} tal que el funcional de riesgo $R(h^{\ell})$ definido sobre la función de pérdida

$$l : Y \times Y \rightarrow \mathbb{R}, \quad (6)$$

es mínimo.

Usualmente se ha considerado que, o bien Y es un conjunto finito no ordenado y entonces el problema a solucionar es un problema de clasificación, y la 0 – 1 función de pérdida resulta adecuada, o bien Y es un espacio métrico y por tanto se habla de estimar una regresión, hecho que obliga a considerar una función de pérdida que tenga en cuenta la estructura métrica, por ejemplo la norma 2 dentro del espacio de los números reales.

En regresión ordinal el algoritmo de aprendizaje induce una ordenación, igualmente que en el caso de la regresión métrica, pero sobre un conjunto Y finito, como en el caso de la clasificación. La existencia de esta escala ordinal lleva a problemas en la definición de cuál sería la función de pérdida mas acertada para esta tipología.

De inicio, se presenta un modelo independiente de distribución (i.d.) para regresión ordinal que es una modificación de las ideas presentadas en [7] sobre una función de pérdida que actúa sobre pares de rangos. A continuación, es mostrada la máquina de aprendizaje a utilizar, la K -SVCR, sobre la cual se detalla el nuevo algoritmo para regresión ordinal. Por último se presenta la aplicación desarrollada sobre el problema financiero que ha motivado el presente estudio.

3.2 Formulación para Regresión Ordinal

Se considera un espacio de entrada $X \subset \mathbb{R}^n$ con elementos de la forma $\mathbf{x} = (x_1, \dots, x_n)^T \in \mathbb{R}^n$ y un espacio de salida $Y = \{r_1, \dots, r_q\}$ con una ordenación

$$r_q \succ_Y r_{q-1} \succ_Y \dots \succ_Y r_1, \quad (7)$$

donde el símbolo \succ_Y podría traducirse por “es preferido a”. Dado el i.i.d. conjunto de datos \mathcal{S} , se considerará que cada función del espacio de modelos \mathcal{H} induce un orden \succ_X sobre los elementos del espacio de entrada siguiendo la regla

$$\mathbf{x}_i \succ_X \mathbf{x}_j \Leftrightarrow h(\mathbf{x}_i) \succ_Y h(\mathbf{x}_j). \quad (8)$$

Continuando las notaciones, un modelo i.d. de regresión ordinal debe hallar la función $h_{pref}^* \in \mathcal{H}$, induciendo una ordenación sobre el espacio de entrada X , que cometa el mínimo número de inversiones de orden sobre los pares de objetos $(\mathbf{x}_1, \mathbf{x}_2)$. Se puede definir la probabilidad de realizar una inversión de orden a partir de un funcional de riesgo $R_{pref}(h)$ que utilice la esperanza estadística

$$\mathbf{E}[\ell_{pref}(h(\mathbf{x}_1), h(\mathbf{x}_2), y_1, y_2)], \quad (9)$$

sobre una función de pérdida definida por

$$\ell_{pref}(\hat{y}_1, \hat{y}_2, y_1, y_2) = \begin{cases} 1 & \text{si } y_1 \succ_Y y_2 \\ & \wedge \neg(\hat{y}_1 \succ_Y \hat{y}_2) \\ 1 & \text{si } y_2 \succ_Y y_1 \\ & \wedge \neg(\hat{y}_2 \succ_Y \hat{y}_1) \\ 0 & \text{en otro caso.} \end{cases} \quad (10)$$

Aplicando el principio de minimización del riesgo empírico (ERM), la función h^ℓ hallada es aquella que hace mínimo el riesgo empírico $R_{emp}(h; \mathcal{S})$ definido como

$$\frac{1}{\ell^2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \ell_{pref}(h(\mathbf{x}_i), h(\mathbf{x}_j), y_i, y_j), \quad (11)$$

que se observa está basado en parejas de objetos. Utilizando las abreviaciones $\mathbf{x}^{(1)}$ y $\mathbf{x}^{(2)}$ para denotar el primer y segundo objeto del par, se puede crear un nuevo conjunto de entrenamiento $\mathcal{S}^* \subset X \times X \times \{-1, 0, +1\}$ de la forma

$$\begin{aligned} \mathcal{S}^* &= \left\{ \left(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)} \right), z_i \right\}_{i=1}^{\ell^2} \\ z_i &= \text{sign}(y_1 \ominus y_2), \end{aligned} \quad (12)$$

donde \ominus es la operación diferencia de rango. El nuevo conjunto \mathcal{S}^* puede partirse en dos subconjuntos

$$\begin{aligned} \mathcal{S}^0 &= \left\{ \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \right), z \right\} \in \mathcal{S}^* : z = 0 \\ \mathcal{S}^1 &= \left\{ \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)} \right), z \right\} \in \mathcal{S}^* : |z| = 1 \end{aligned} \quad (13)$$

tales que $\mathcal{S}^* = \mathcal{S}^0 \cup \mathcal{S}^1$, con $\mathcal{S}^0 \cap \mathcal{S}^1 = \emptyset$.

En el trabajo [7] es considerado sólo el conjunto \mathcal{S}^1 , ya que las máquinas de Soporte Vectorial para Clasificación (SVC) estándares trabajan únicamente sobre dicotomías $+1/-1$. Puesto que la cardinalidad de \mathcal{S}^1 es mucho mayor que la de \mathcal{S}^0 , queda asegurada una reducida pérdida de información durante el entrenamiento, pero esta realidad también provoca que el conjunto de entrenamiento sea muy numeroso y, por tanto, sólo se pueda trabajar sobre un subconjunto reducido de estos nuevos patrones. Así, si se define como ℓ_i el número de elementos de cada uno de los q rangos r_i , considerando por ejemplo una condición usual como que todos los rangos tengan igual número de representantes, $\ell_i = \ell_j = \ell, \forall i \neq j$, se obtiene como cardinalidad de los nuevos subconjuntos

$$\begin{aligned} \mathbf{s}_0 &= \#\mathcal{S}^0 = \sum_{i=1}^q \ell_i^2 = q\tilde{\ell}^2 \\ \mathbf{s}_1 &= \#\mathcal{S}^1 = (q-1)q\tilde{\ell}^2. \end{aligned} \quad (14)$$

En cambio, se verá como el proceso de aprendizaje basado en las K -SVCRs, aunque parta de un conjunto de aprendizaje más elevado, $\#\mathcal{S}^* = \mathbf{s}_0 + \mathbf{s}_1$, acabará trabajando sobre un conjunto de patrones más reducido y significativo.

3.3 Máquina K -SVCR

Dado un conjunto de entrenamiento i.i.d.

$$\mathcal{T} = \{(\mathbf{t}_p, z_p)\}_{p=1}^{\ell} \sim P_{TZ}^{\ell}, \quad (15)$$

donde el espacio de salida Z es $\{-1, +1\}$, los clasificadores basados en Vectores Soporte (SV) [3], [5], [11], [12], también denominados de margen amplio, hallan una función de decisión no lineal en la forma¹

$$\begin{aligned} h(\mathbf{t}) &= \text{sign}(\langle \mathbf{w}, \mathbf{t} \rangle_{\mathcal{F}} + b) = \\ &= \text{sign}(k(\mathbf{w}, \mathbf{t}) + b), \end{aligned} \quad (16)$$

¹Con objeto de reducir la complejidad notacional, mediante $\langle \mathbf{w}, \mathbf{t} \rangle_{\mathcal{F}}$ se pretende representar $\langle \mathbf{w}, \phi(\mathbf{t}) \rangle_{\mathcal{F}}$, haciendo implícita la inserción no lineal en el espacio de características.

tal que el espacio original de patrones T ha sido introducido dentro de otro espacio de dimensión mucho mayor y dotado de producto interno \mathcal{F} , denominado espacio de características, vía una aplicación no lineal,

$$\phi : T \subset \mathbb{R}^n \rightarrow \mathcal{F}_{\langle \cdot, \cdot \rangle},$$

con objeto de poder hallar la función de decisión como una solución lineal en este espacio de características, con la sola restricción que el núcleo k ha de cumplir el teorema de Mercer.

Continuando la teoría de vectores soporte, se ha desarrollado una nueva máquina de aprendizaje, la K -SVCR, con la intención de extender las SVCs para tareas multi-clase. La función de decisión a ser hallada h_{pref}^* toma la forma

$$h(\mathbf{t}_p) = \begin{cases} +1 & , p = 1, \dots, \ell_1 \\ -1 & , p = \ell_1 + 1, \dots, \ell_1 + \ell_2 \\ 0 & , p = \ell_1 + \ell_2 + 1, \dots, \ell \end{cases} \quad (17)$$

donde, sin pérdida de generalidad, se ha supuesto que los primeros ℓ_1 i ℓ_2 patrones ($\ell_{12} = \ell_1 + \ell_2$) corresponden a las dos clases a ser separadas, y los otros patrones ($\ell_3 = \ell - \ell_{12}$) pertenecen a cualquier otra clase – se etiquetarán con 0 –. Para controlar el ancho de la zona de etiquetado 0, se hará servir la función ε -insensitiva de Vapnik utilizada en la extensión de los vectores soporte al problema de regresión métrica y que se define por

$$|z - h(\mathbf{t})|_\varepsilon \stackrel{def}{=} \max \{0, |z - h(\mathbf{t})| - \varepsilon\}. \quad (18)$$

Obviamente, no existe, en general, ningún hiperplano separador cumpliendo las restricciones (17) en el espacio de entrada $T \subset \mathbb{R}^n$, y por tanto no tiene sentido buscar una solución lineal del problema en este espacio. En cambio, si insertamos este espacio vía una función no lineal en un espacio con una dimensión suficientemente grande, la capacidad del hiperplano para cumplir las restricciones se incrementa y será posible hallar una solución. Por ejemplo, cuando se soluciona el problema de programación cuadrática (QP) que lleva a la solución SVC es muy usual formular el problema añadiendo la restricción $b = 0$, que es equivalente a requerir que el hiperplano contenga el origen, reduciendo en uno el número de grados de libertad. El requerimiento de la máquina de aprendizaje K -SVCR es mayor, ya que obliga al hiperplano óptimo a contener todos los ℓ_3 patrones de aprendizaje con etiqueta 0.

El problema de optimización restringida asociado al método K -SVCR está definido en su caso general como: dado $0 \leq \delta < 1$ *a priori*, hallar los valores óptimos de los parámetros que hacen mínimo el funcional

$$J(\mathbf{w}, \xi, \varphi^{(*)}) = \frac{1}{2} \|\mathbf{w}\|_{\mathcal{F}}^2 + C \sum_{i=1}^{\ell} \xi_i \quad (19)$$

$$+ D \sum_{i=1}^{\ell} (\varphi_i + \varphi_i^*),$$

restringido a

$$z_i \cdot (\langle \mathbf{w}, \mathbf{t}_i \rangle_{\mathcal{F}} + b) \geq 1 - \xi_i, i = 1, \dots, \ell_{12} \quad (20)$$

$$\begin{aligned} \langle \mathbf{w}, \mathbf{t}_i \rangle_{\mathcal{F}} + b &\leq \delta + \varphi_i \\ -(\langle \mathbf{w}, \mathbf{t}_i \rangle_{\mathcal{F}} + b) &\leq \delta + \varphi_i^*, i = \ell_{12} + 1, \dots, \ell. \end{aligned} \quad (21)$$

Una solución al problema definido por (19) con restricciones (20)-(21) puede hallarse solucionando el problema dual de optimización de Wolfe: dado $0 \leq \delta < 1$ *a priori*, hallar los valores óptimos de los parámetros que hacen mínimo el funcional

$$L(\gamma) = \frac{1}{2} \gamma^T \cdot \mathbf{H} \cdot \gamma + \mathbf{c}^T \cdot \gamma, \quad (22)$$

donde

$$\begin{aligned} \gamma^T &= (\gamma_1, \dots, \gamma_{\ell}, \gamma_{\ell+1}, \dots, \gamma_{\ell+\ell_3}) \\ \mathbf{c}^T &= \left(\frac{-1}{z_1}, \dots, \frac{-1}{z_{\ell_{12}}}, \delta, \dots, \delta \right) \\ \gamma^T, \mathbf{c}^T &\in \mathbb{R}^{\ell_{12} + \ell_3 + \ell_3}, \end{aligned} \quad (23)$$

y la matriz \mathbf{H} se describe como

$$\begin{pmatrix} (k(\mathbf{t}_i, \mathbf{t}_j)) & -(k(\mathbf{t}_i, \mathbf{t}_j)) & (k(\mathbf{t}_i, \mathbf{t}_j)) \\ -(k(\mathbf{t}_i, \mathbf{t}_j)) & (k(\mathbf{t}_i, \mathbf{t}_j)) & -(k(\mathbf{t}_i, \mathbf{t}_j)) \\ (k(\mathbf{t}_i, \mathbf{t}_j)) & -(k(\mathbf{t}_i, \mathbf{t}_j)) & (k(\mathbf{t}_i, \mathbf{t}_j)) \end{pmatrix}$$

$$\mathbf{H} = \mathbf{H}^T \in \mathcal{M}(\mathbb{R}^{\ell_{12} + \ell_3 + \ell_3}, \mathbb{R}^{\ell_{12} + \ell_3 + \ell_3}),$$

restringido a

$$\begin{aligned} 0 &\leq \gamma_i \cdot z_i \leq C, \quad i = 1, \dots, \ell_{12} \\ 0 &\leq \gamma_i \leq D, \quad i = \ell_{12} + 1, \dots, \ell + \ell_3 \end{aligned} \quad (24)$$

$$\sum_{i=1}^{\ell_{12}} \gamma_i = \sum_{i=\ell_{12}+1}^{\ell} \gamma_i - \sum_{i=\ell+1}^{\ell+\ell_3} \gamma_i. \quad (25)$$

El hiperplano de decisión puede escribirse

$$h(\mathbf{t}) = \text{sign}(g(\mathbf{t})) \cdot |g(\mathbf{t})|_\delta, \quad (26)$$

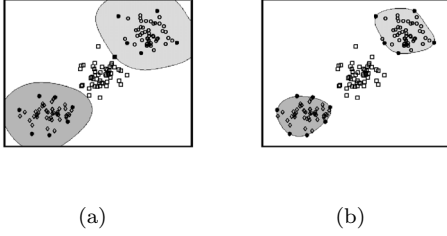


Figura 3: Clasificación K -SVCR multi-clase con diferente grado de ε -insensitividad.

donde

$$g(\mathbf{t}) = \sum_{i=1}^{SV} \nu_i \langle \mathbf{t}_i, \mathbf{t} \rangle_{\mathcal{F}} + b, \quad (27)$$

$$\begin{aligned} \nu_i &= \gamma_i, \quad i = 1, \dots, \ell_{12} \\ \nu_i &= \gamma_{i+\ell_3} - \gamma_i, \quad i = \ell_{12} + 1, \dots, \ell, \end{aligned} \quad (28)$$

y el término b es calculado de (20), (21) sobre los vectores soporte en función de los parámetros γ_i . Puede observarse que la restricción (25) puede traducirse por

$$\sum_{i=1}^{SV} \nu_i = 0. \quad (29)$$

La nueva metodología es consistente con la SVC bi-clase estándar, ya que si $K = 2$ entonces $\ell_3 = 0$ y los problemas de optimización son equivalentes. De hecho, aunque $K > 2$, si $\delta = 0$ la función de decisión (16) es la misma que (26). Esta imposición, sin embargo, implica la no generalización dentro de la clase con etiqueta 0, el incremento en el número de vectores soporte entre los patrones de entrenamiento etiquetados 0 [11] y un coste computacional mas elevado.

Como ilustración del funcionamiento del nuevo algoritmo, en la figura 3.3 puede observarse una multi-clasificación realizada en \mathbb{R}^2 con datos separables artificiales con insensitvidades $\delta = 0,05$ y $\delta = 0,95$ para el caso de las gráficas superior e inferior, respectivamente. A destacar como la amplitud de la zona de no clasificación está en relación directa con el grado de insensitividad.

3.4 K -SVCR para Regresión Ordinal

Siguiendo la exposición de [7] y utilizando la teoría expuesta en las secciones anteriores, se

derivará un nuevo algoritmo para regresión ordinal. Se asumirá que para un conjunto de modelos \mathcal{H} definido como en la formulación para regresión lineal de la Sección 2 existe un conjunto \mathcal{U} de funciones lineales sobre \mathcal{F} definidas desde el espacio origen X hacia \mathbb{R} tal que para toda función $h \in \mathcal{H}$ existe una función $U \in \mathcal{U}$ (y viceversa) cumpliendo

$$h(\mathbf{x}) = r_i \iff U(\mathbf{x}) \in [\theta(r_{i-1}), \theta(r_i)], \quad (30)$$

donde

$$U(\mathbf{x}) = \langle \mathbf{w}^T, \mathbf{x} \rangle_{\mathcal{F}}, \quad (31)$$

con $\theta(r_0) = -\infty$ i $\theta(r_q) = \infty$. Esta función es denominada usualmente función utilidad y no incurre en error sobre el i -ésimo ejemplo del conjunto de entrenamiento \mathcal{S}^* si, y sólo si, se cumplen las dos dobles implicaciones

$$z_i \langle \mathbf{w}^T, \mathbf{x}_i^{(1)} \rangle_{\mathcal{F}} \mathbf{w}^T > z_i \langle \mathbf{w}^T, \mathbf{x}_i^{(2)} \rangle_{\mathcal{F}} \quad (32)$$

$$\iff z_i \langle \mathbf{w}^T, \mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)} \rangle_{\mathcal{F}} > 0.$$

Asumiendo un margen finito entre los vectores n -dimensionales $\mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}$ de clases $z_i = +1$ y $z_i = -1$, se definen las restricciones a cumplir por el QP problema como

$$z_i \cdot (\langle \mathbf{w}, \mathbf{x}_i \rangle_{\mathcal{F}} + b) \geq 1 - \xi_i, i = 1, \dots, s_1 \quad (33)$$

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x}_i \rangle_{\mathcal{F}} + b &\leq \delta + \varphi_i \\ -(\langle \mathbf{w}, \mathbf{x}_i \rangle_{\mathcal{F}} + b) &\leq \delta + \varphi_i^*, i = 1, \dots, s_0, \end{aligned} \quad (34)$$

con $\delta < 0,25$ para asegurar que $|\theta(r_{i-1}), \theta(r_i)| < 0,5$ y los parámetros $\xi_i, \varphi_i, \varphi_i^* < 0,125$ para asegurar que no existe intersección entre los intervalos definidos por la función utilidad. El vector de pesos \mathbf{w}^ℓ que maximiza el margen puede ser determinado minimizando la norma cuadrada de la ecuación (19) definida en el método K -SVCR.

Definiendo $\mathbf{t}_i = \mathbf{x}_i^{(1)} - \mathbf{x}_i^{(2)}$, se llega al problema dual de optimización de Wolfe similar a aquel solucionado en el método K -SVCR (22)–(25). En este caso, la función solución hallada, lineal sobre el espacio \mathcal{F} , se expresa como

$$U(\mathbf{x}; \mathbf{w}^\ell) = \sum_{i=1}^{SV} \nu_i \langle \mathbf{t}_i, \mathbf{x} \rangle_{\mathcal{F}}, \quad (35)$$

donde

$$\mathbf{w}^\ell = \sum_{i=1}^{SV} \nu_i \langle \mathbf{t}_i, \cdot \rangle_{\mathcal{F}}. \quad (36)$$

Esta nueva formulación del problema de regresión lineal centra el entrenamiento sobre pares de patrones que pertenecen al mismo rango, estableciendo que la anchura de la función utilidad sobre cada rango dentro del espacio \mathcal{F} es como máximo 0,5. En cambio, el desarrollo comparado de [7] sólo tiene en consideración los pares de patrones que poseen rangos diferentes, rehusando toda la información de las parejas de patrones de igual rango.

El conjunto de restricciones (34) garantiza la inserción de cada rango dentro de una bola acotada. El conjunto de restricciones (33) sólo es útil para distribuir estos intervalos sobre una recta en el espacio \mathcal{F} . Por tanto sólo es necesario considerar, en su caso más reducido, un par de patrones por cada par de rangos diferentes y así con la nueva formulación el número de restricciones se puede reducir a $2s_0 + q(q-1)$ desigualdades. Si se considera, tal como se hizo en la Sección 2, que todos los rangos vienen representados por un mismo número de patrones, ℓ , puede observarse que la nueva formulación trabaja sobre $2q\ell^2 + q(q-1)$ desigualdades, mientras que aquella definida en [7] lo hace sobre $(q-1)q\ell^2$ desigualdades.

El único punto crítico de la presente formulación es aquel que hace referencia a la estimación de los valores frontera $\theta(r_i)$. Para hallarlos es necesario asegurar que los pares de patrones en rangos diferentes en consideración durante el entrenamiento tienen un valor en la variable artificial $\xi_i = 0$, o sea, se ha de estar seguro que los elementos considerados no tienen 'ruido' o distorsión. Aun así, la restricción de crear variables artificiales menores que 0,125 aseguraría la no intersección de los intervalos, pero no evitaría la propagación del error del patrón de entrenamiento considerado sobre la definición del intervalo definitorio del rango. Una vez asegurada esta condición, la estimación del rango $\theta(r_i)$ viene dada por

$$\theta(r_i) = \frac{U(\mathbf{x}_1; \mathbf{w}^\ell) + U(\mathbf{x}_2; \mathbf{w}^\ell)}{2}. \quad (37)$$

Al acabar el entrenamiento de la máquina de aprendizaje y tras ser obtenidos los valores frontera de los rangos, los nuevos elementos a clasificar se etiquetarían siguiendo la ecuación (30).

3.5 Problema Financiero de Riesgo Crediticio

El resultado que se persigue en esta aplicación es la creación de un sistema clasificador que permita realizar una ordenación sobre el riesgo crediticio de diferentes empresas en función de una serie de ratios económicas, tal como ya se expuso en la Sección 1. En una primera aproximación al problema deben determinarse las variables del espacio de entrada que resulten representativas para mostrar el estado financiero y de negocio de las empresas por lo que se hace necesaria la presencia de un experto que ayude en esta decisión, aunque, como era de esperar, ya existe un conjunto amplio de indicadores estándares que facilitan la selección². Para esta aplicación se decidió tomar como entradas los 5 indicadores siguientes: la cobertura de intereses el año anterior y las ratios de endeudamiento, de rendimiento sobre activo, de deuda a corto sobre largo y de activo circulante sobre total. Por otra parte, se han separado las empresas por sectores de producción, en total se han considerado 7 sectores, debido a la especificidad que muestran éstas respecto a su área de negocio para así, además de crear una máquina de aprendizaje sobre la globalidad de los datos, desarrollar para cada sector una máquina de aprendizaje específica.

Por lo que se refiere al conjunto de entrenamiento, sólo puede disponerse de un conjunto de empresas reducido de las cuáles se tengan todos los datos de entrada necesarios y su rating crediticio. Aunque en principio esto resulta un contratiempo para poder realizar un adecuado entrenamiento, es bien conocida la gran capacidad de las máquinas de soporte vectorial para generalizar sobre un número pequeño de datos en entornos de gran dimensionalidad, por lo que resultan una elección del todo adecuada. Además, la máquina K -SVCR para regresión ordinal presentada en la Sección anterior se veía obligada a trabajar sobre un número de representantes pequeño si se quería evitar la explosión en la dimensión del conjunto de entrenamiento modificado sobre pares. De esta forma, se tomaron como patrones de aprendizaje aquellas empresas de las que se dispone de todos los datos y como patrones de test las empresas de las que carecemos de la información de salida, su riesgo crediticio, pero cuyo eva-

²Al respecto, destacar que incluso una SVC puede servirnos como selector de ítems representativos.

luación aproximada no resulta demasiado complicada en una primera valoración para un experto. Esta decisión ha sido tomada teniendo en cuenta el número tan reducido de datos disponibles, hecho que motiva no reservar en un principio datos reales para la fase de test³.

Los rangos de salida son 4, elegidos de forma que se intenta minimizar la cantidad de información perdida en esta reducción del número de etiquetas, que se corresponderían con las calificaciones ordenadas de mayor a menor seguridad en el retorno del crédito {'AAA', 'AA'}, 'A', 'BBB' y {'BB', 'B'}, con sus respectivas modificaciones en la etiqueta mediante los signos + y -.

Las máquinas *K-SVCR* en regresión ordinal serán ahora utilizadas sobre la aplicación financiera. Un estudio de todos los sectores en conjunto fue desaconsejado desde el punto de vista experto por lo que se realizará una ordenación de las empresas ordenadas por sectores. Una vez realizado el entrenamiento se obtienen las divisiones intervalares de la recta real.

En cuanto a los parámetros de cada tipo de núcleo y el factor de insensitividad, se ha determinado entrenar toda una serie de máquinas con diferentes elecciones de estos parámetros. Para el caso de las funciones gaussianas, se han entrenado máquinas para 20 posibles valores del parámetro de varianza de forma que $\sigma \in [0.1, 1]$ y para cada valor de varianza 10 máquinas con parámetro de insensitividad $\delta \in [0.09, 0.45]$ ⁴, lo que conforma un total de 100 máquinas. Si bien para valores elevados de δ y pequeños de σ la generalización será escasa, mientras que en condiciones inversas la máquina dará como posible salida válida el conjunto completo de clases debido a la gran indefinición de los núcleos clasificadores, se considera que como promedio la clase que resulte adecuada será aquella que más votaciones obtendrá sobre el global de todas las máquinas entrenadas.

Sobre los resultados obtenidos con las máquinas *K-SVCR* en regresión ordinal pueden realizarse las siguientes observaciones.

- El análisis de porcentajes en las respues-

³Durante la elaboración del estudio continuó el proceso de búsqueda de la información no disponible y en algún caso se obtuvo recompensa. Sin embargo, estos datos se ignoraron durante el proceso de entrenamiento de las máquinas para conseguir que el test fuera completamente 'a ciegas'.

⁴Recuérdese que el parámetro de insensitividad debe ser menor que 0.50 para asegurar la no intersección entre intervalos.

tas resulta en la emisión de una predicción muy similar a aquella emitida por SVC estándares, por lo que puede entenderse que ambas técnicas permiten un nivel de aprendizaje similar.

- En el caso de la regresión ordinal el aprendizaje se ha realizado también sobre la preferencia de orden por lo que pudiera pensarse que es posible además de determinar el r  ting general, indicar la tendencia de esa empresa o su situaci  n dentro del r  ting pudiendo realizarse comparaciones de situaci  n financiera entre dos empresas de igual r  ting crediticio. Como bien puede intuirse, el uso de t  cnicas de l  gica difusa o de an  lisis intervalar podr  an acabar de determinar la calificaci  n final de la empresa.
- La regresi  n ordinal es menos dependiente de los par  metros relacionados con el aprendizaje de las m  quinas pues los porcentajes obtenidos poseen una menor desviaci  n que en el caso de multclasificaci  n est  ndar.
- Destacar finalmente que el estudio en ordenaci  n emite unos datos que pueden entenderse como m  s congruentes con el tipo de problema analizado. Las variaciones en los porcentajes entre rangos siguen una l  nea con s  lo una zona de m  ximo, mientras que en los resultados de multclasificaci  n est  ndares pueden existir dos rangos con niveles de porcentaje similares aunque no sean rangos continuos. Por ejemplo los datos obtenidos sobre un patr  n en multclasificaci  n est  ndar puede arrojar un empate entre los rangos 1 y 4. Este tipo de inestabilidades no se producen en ning  n caso cuando se utiliza la formulaci  n en ordenaci  n aunque el tipo de n  cleos utilizado es el mismo.

Por completitud, se ha realizado tambi  n una evaluaci  n del aprendizaje de la nueva m  quina en ordenaci  n reservando parte de la informaci  n como conjunto de test, aunque ello suponga perder informaci  n cr  tica. Se han realizado 50 experimentaciones diferentes extrayendo de forma aleatoria y sin considerar el rango al que pertenecen un 75% de los patrones del conjunto de aprendizaje original para elaborar el conjunto de aprendizaje modificado en pares y reservando el 25% para realizar la evaluaci  n.

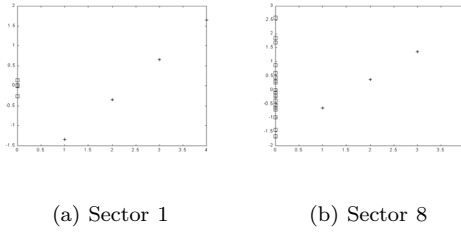


Figura 4: Resultados de la ordenación sobre dos sectores de producción.

sector	%error	SV	# patrones
sector 1	58.33%	141.04	559
sector 2	61.67%	139.60	601
sector 3	70.33%	49.30	121
sector 4	52.86%	61.76	195
sector 6	50.67%	39.60	127
sector 7	45.00%	77.56	507
sector 8	46.00%	19.96	37

Tabla 5: Porcentaje de error cometido sobre test y número de vectores soporte necesarios.

El conjunto modificado se construyó de nuevo considerando todas las combinaciones posibles entre elementos de igual rango y sólo una diferenciación de entre todas las posibles entre dos rangos distintos cualesquiera. Tras seleccionar el tipo de núcleo, una función gaussiana, y fijar una varianza σ e insensitividad δ única para todas las máquinas se procedió al entrenamiento. En la Figura 3.5 pueden observarse gráficamente algunos de los resultados del entrenamiento. Las marcas cuadradas situadas con valor de eje de abscisas 0 corresponden a las salidas de los patrones de test junto a los patrones sin rango conocido.

En la Tabla 5 se recogen los porcentajes de error sobre el conjunto de test. Tal como era de esperar el porcentaje de error cometido es muy elevado debido a la escasez de información disponible en la fase de entrenamiento, aunque en la mayoría de casos el nivel de error siempre está muy por debajo del 75% que se obtendría al trabajar sobre 4 clases en caso de no existir entrenamiento alguno.

4 Conclusiones

En el artículo presentado se han desarrollado dos tipos de técnicas conexionistas de aprendizaje con el objeto de emular el proceso de evaluación de un grupo de expertos en la predicción del riesgo de crédito de empresas.

Las técnicas de aprendizaje aplicadas han sido redes neuronales con funciones base radiales y máquinas de soporte vectorial. En ambos casos ha sido necesario adaptar la información disponible, es decir los datos financieros de la empresa, para mejorar su funcionamiento. En la primera aproximación del problema se han aplicado estrategias diversas que se pueden utilizar para sintetizar información cualitativa asociada a distintos atributos, cada uno de los cuales está descrito cualitativamente de manera distinta. Los resultados obtenidos en las simulaciones fruto de la aplicación de esta metodología dan muestra de su utilidad dado que permiten aumentar el grado de generalización obtenido con dichas redes neuronales.

En segunda instancia se presenta una aplicación de la máquina K -SVCR para tratar el problema de predicción de riesgo crediticio. La nueva formulación define el problema sobre pares de objetos y su relación de orden en el espacio de salida Y . La motivación original de este estudio es crear máquinas de aprendizaje capaces de establecer un rango sobre el riesgo crediticio empresarial.

La nueva línea de aplicación financiera está en una fase de trabajo no concluyente, pero ha permitido avanzar en la investigación teórica que actualmente se está implementando sobre la aplicación. Las limitaciones de los métodos presentados no pueden evaluarse con profundidad hasta que la implementación sea completa y esté suficientemente validada. Uno de los aspectos necesarios para mejorar el análisis es la ampliación del conjunto de empresas disponibles para plantear la extracción de conocimiento con suficientes garantías de éxito. Así, no sólo sería posible el análisis único en función del sector de actividad o de la valoración de los expertos, sino que existirían suficientes datos como para realizar una evaluación más afinada.

Referencias

- [1] N. Agell et al., Homogenising references in orders of magnitude spaces: An application to credit risk prediction, *14th Interna-*

tional Workshop on Qualitative Reasoning, Morelia, Mexico, 2000.

- [2] D.S. Broomhead y D. Lowe, Multivariate functional interpolation and adaptive network, *Complex Systems*, 2 (1988) 321–355.
- [3] C. Burges, A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discovery*, 2 (1998), 1–47.
- [4] S. Chen et al., Orthogonal least squares learning for radial basis function networks, *Transactions on Neural Networks*, 2 (1991) 302–309
- [5] C. Cortes y V. Vapnik, Support vector networks, *Machine Learning*, 20 (1995), 273–297.
- [6] S. Goonatilake y P. Treleaven, *Intelligent systems for finance and business*, John Wiley & Sons, 1996.
- [7] R. Herbrich, T. Graepel y K. Obermayer, Large margin rank boundaries for ordinal regression, *Advances in Large Margin Classifiers*, A.J. Smola, P. Bartlett, B. Schölkopf y D. Schuurmans, ed. MIT Press (2000), 281–296.
- [8] N. Piera, Current trends in qualitative reasoning and applications, *Monografía CIM-NE*, 33, International Centre for Numerical Methods in Engineering, Barcelona, 1995.
- [9] T. Poggio y F. Girosi, Networks for approximation and learning, *Proceedings IEEE*, 78 (1990) 1481–1497.
- [10] L. Prechelt, PROBEN1: Set of neural network benchmark problems and benchmarking rules, Technical Report 21/94, University of Karlsruhe, 1994.
- [11] A. Smola, *Learning with kernels*, Ph.D. thesis, Department Computer Science, Technical University Berlin, 1998.
- [12] V. Vapnik, *The nature of statistical learning theory*, Springer Verlag New York, 1995.