CRITERIOS DE SELECCIÓN DE MODELO EN *CREDIT SCORING*. APLICACIÓN DEL ANÁLISIS DISCRIMINANTE BASADO EN DISTANCIAS[†]

Eva Boj[1], Mª Mercè Claramunt[2], Anna Esteve[3] y Josep Fortiana[4]

ABSTRACT

The aim of this paper is to study model selection criteria in credit scoring. Such criteria are usually derived from an error cost function which takes into account misclassification probabilities in good and bad credit risk subpopulations plus other parameters encoding context information relevant to the objective portfolio. We present a distance based classification approach to credit scoring, as an addition to the current repertoire of procedures. We illustrate both method and selection criteria with two real datasets.

KEY WORDS: Credit Risk; Credit scoring; Probability of default; Multivariate Data Analysis; Distance Based Prediction.

RESUMEN

En este trabajo estudiamos criterios de selección de modelo en *credit scoring*. Estos criterios se derivan usualmente de una función de coste del error que tiene en cuenta las probabilidades de mala clasificación en las subpoblaciones de buenos y malos riesgos de crédito y, adicionalmente, algunos parámetros con información relevante del entorno de la cartera analizada. Presentamos una metodología de análisis discriminante basado en

[†] Trabajo financiado por el Ministerio de Educación y Ciencia, proyecto número MTM2006-09920.

[1] Departamento de Matemática Económica, Financiera y Actuarial. Facultad de Economía y Empresa. Universidad de Barcelona. Avenida Diagonal 690, 08034_Barcelona. España. E-mail: evaboj@ub.edu

[2] Departamento de Matemática Económica, Financiera y Actuarial. Facultad de Economía y Empresa. Universidad de Barcelona. Avenida Diagonal 690, 08034_Barcelona. España. E-mail: mmclaramunt@ub.edu

[3] Centro de Estudios Epidemiológicos sobre las Infecciones de Transmisión Sexual y Sida de Cataluña (CEEISCAT). Hospital Universitario Hermanos Trías y Pujol. CIBER Epidemiología y Salud Pública (CIBERESP). Ctra. de Cañete, s/n. 08916_Badalona. España. E-mail: aeg.ceescat.germanstrias@gencat.net

[4] Departamento de Probabilidad, Lógica y Estadística. Facultad de Matemáticas. Universidad de Barcelona. Gran Vía de las Cortes Catalanas 595, 08007_Barcelona. España. E-mail: fortiana@ub.edu

distancias como método de *scoring* alternativo a los existentes en la literatura. E ilustramos tanto la utilización de la predicción basada en distancias como de los criterios de selección de modelo con dos conjuntos de datos reales.

PALABRAS CLAVE: Riesgo de crédito; *Credit scoring*; Probabilidad de insolvencia; Análisis estadístico multivariante; Predicción basada en distancias.

1. INTRODUCCIÓN

Las primas por riesgo de crédito de una Entidad Financiera se calculan haciendo uso de las probabilidades de insolvencia de los riesgos a partir de un modelo de *credit scoring*. La elección del modelo de *scoring* es un paso clave para la solvencia de la Entidad. En este trabajo describimos diferentes criterios de selección. El primero de ellos se basa en analizar las probabilidades de mala clasificación en las poblaciones, la de buenos y la de malos riesgos de crédito, y la global. El segundo se basa en una función de coste del error, la cuál tiene en cuenta el entorno de la cartera.

Proponemos como herramienta alternativa en el problema del *credit scoring* la utilización del *Análisis Discriminante Basado en Distancias* (ADBD). Ésta metodología es especialmente adecuada para dicho problema, ya que se trata de una metodología no paramétrica que permite de modo natural una mezcla de variables numéricas y categóricas. Por otro lado, da lugar a una relación indirecta, esencialmente no lineal, entre los predictores y la respuesta. Algunas referencias en las se utiliza también la metodología estadística de análisis discriminante en el problema del *credit scoring* son Artís *et al.* (1994), Boj *et al.* (2009a), Bonilla *et al.* (2003), Hand y Henley (1997) y Trias *et al.* (2005 y 2008).

Con dos conjuntos de datos reales de dos Entidades Financieras ilustramos la utilización de los criterios de selección de modelo y la aplicación de la predicción basada en distancias. Los métodos de la literatura con los que comparamos el ADBD son: Métodos no-paramétricos como las redes neuronales, el método de los k vecinos más próximos, el método de la estimación núcleo de la densidad y el árbol de clasificación *classification and regression trees* (CART); y Métodos paramétricos como el análisis discriminante lineal y la regresión logística.

El trabajo se estructura del siguiente modo: en el apartado 2 describimos la metodología de ADBD y explicamos cómo traspasar la información aportada por los predictores a la matriz de distancias de cada población. En el apartado 3 detallamos los criterios de selección de modelo basados en las probabilidades de mala clasificación y en una función de coste del error. En el apartado 4 ilustramos el uso de los criterios de selección de modelo y la aplicación del ADBD con dos conjuntos de datos reales.

2. ANÁLISIS DISCRIMINANTE BASADO EN DISTANCIAS

Los métodos de análisis estadístico multivariante basados en distancias son adecuados cuando tratamos con predictores de tipo mixto, es decir, una mezcla de variables cuantitativas, categóricas y/o binarias. En el problema del *credit scoring* ocurre igual que en la tarificación *a priori* en la fase de selección de variables de tarifa, que el conjunto potencial de factores de riesgo es de tipo mixto (Boj *et al*., 2000, 2001, 2004, 2009a). Recordemos, por ejemplo, que en la tarificación del seguro del automóvil teníamos como predictores: la edad y el sexo del primer conductor, la antigüedad del carné, el uso del vehículo, la zona de circulación, la potencia del vehículo, la marca y el tipo de vehículo, … Ahora en el riesgo de crédito tenemos: la duración y el importe del credito, el propósito del crédito, la edad, la situación marital y el sexo del beneficiario del crédito, … En resumen, disponemos también de un conjunto de predictores de tipo mixto. Por otro lado, es sabido que los métodos de análisis discriminante funcionan bien con variables cuantitativas o cuando se conoce la densidad de los datos, pero a menudo las variables son binarias, categóricas o mixtas, como es el caso del riesgo de crédito. Puesto que siempre es posible definir una distancia entre observaciones, también es posible dar una versión del análisis discriminante utilizando sólo distancias. A esta versión la denominamos *Análisis Discriminante Basado en Distancias* y nos referimos a Cuadras (1989, 1992), Cuadras *et al*. (1997) y Boj *et al.* (2009a,b) para un detalle teórico y práctico.

Supongamos que disponemos de un conjunto de n individuos, pertenecientes a g grupos conocidos $\Omega = \Omega_1 \cup \cdots \cup \Omega_g$ de tamaños $n_1, \ldots, n_g$, siendo el total de individuos $n = n_1 + \cdots + n_g$. Sean $\delta_1, \ldots, \delta_g$, g funciones de distancia con la *propiedad euclídea* en el sentido del Escalado Métrico Multidimensional (ver Borg y Groenen, 2005). Estas funciones pueden o no coincidir para cada población. A partir de los

predictores observados $\mathbf{Z}=\left(Z_1,\ldots,Z_p\right)$ calculamos las matrices de distancias euclídeas entre las muestras de cada población:

$$\Delta_\alpha^{(2)}=\left(\delta_{ij}^2\left(\alpha\right)\right) \text{ de tamaño } n_\alpha \times n_\alpha \text{ para } \alpha=1,\ldots,g\,.$$

Las estimaciones de las variabilidades geométricas son:

$$\hat{V}_\alpha=\frac{1}{2n_\alpha^2}\sum_{i=1}^{n_\alpha}\delta_{ij}^2\left(\alpha\right) \text{ para } \alpha=1,\ldots,g\,.$$

Sea ω un nuevo individuo a clasificar en una de las g poblaciones, y sean $\delta_i^{(2)}\left(\alpha\right)$ para $i=1,\ldots,n_\alpha$ y para $\alpha=1,\ldots,g$ las distancias al cuadrado de este nuevo individuo a los n_α individuos de la población Ω_α, calculadas a partir de los predictores originales $\mathbf{Z}$. Las estimaciones de las funciones de proximidad son:

$$\hat{f}_\alpha\left(\omega\right)=\frac{1}{n_\alpha}\sum_{i=1}^{n_\alpha}\delta_i^2\left(\alpha\right)-\hat{V}_\alpha \text{ para } \alpha=1,\ldots,g\,.$$

La regla basada en distancias consiste en asignar al nuevo individuo ω a la población Ω_α tal que

$$\hat{f}_\alpha\left(\omega\right)=\min_{1\le\beta\le g}\left\{\hat{f}_\beta\left(\omega\right)\right\}.$$

Es de especial interés que esta regla sólo depende de distancias entre observaciones y clasifica a ω en la población más próxima. Finalmente, las probabilidades de que el individuo ω pertenezca a la población α, las estimamos como:

$$\hat{\pi}_\alpha\left(\omega\right)=\frac{e^{-\hat{f}_\alpha(\omega)}}{\sum_{i=1}^{g}e^{-\hat{f}_i(\omega)}} \text{ con } \alpha=1,\ldots,g\,.$$

2.1. MÉTRICAS Y CONJUNTOS DE VARIABLES

Tal y como veremos en la aplicación del apartado 4.1, en riesgo de crédito, al igual que ocurre en la tarificación *a priori* de los seguros no vida, los factores de riesgo pueden agruparse en conjuntos de variables. Recordemos cómo, en el seguro del automóvil podíamos tener por ejemplo los siguientes conjuntos (Boj *et al*., 2004):
Factores relativos al vehículo asegurado: valor, antigüedad, categoría, clase, tipo, marca, modelo, número de plazas, potencia, peso, o relación potencia / peso, color, etc.
Factores relativos al conductor: edad, sexo, antigüedad del carné, estado civil, profesión, número de hijos, posibilidad de conductores ocasionales, resultado de la experiencia en el pasado, etc.
Factores relativos a la circulación: zona de circulación, uso del vehículo, kilómetros anuales, etc.

Para aplicar el ADBD debemos traspasar la información de los predictores de tipo mixto a las matrices de distancias de cada población, la de buenos y malos riesgos. Es decir, tenemos que calcular $\Delta_{\alpha}^{(2)}$ de tamaño $n_{\alpha} \times n_{\alpha}$ (siendo $\alpha = 1,2$). Supongamos que construimos *b* conjuntos de predictores. Para cada conjunto podemos calcular la matriz de distancias euclídea asociada $\Delta_{\alpha}^{(2)}{}_{[s]}$ con $s = 1,\ldots,b$ de tamaño $n_{\alpha} \times n_{\alpha}$. En el ejemplo anterior, del seguro del automóvil, disponíamos de $b = 3$ conjuntos de variables, por lo tanto podríamos calcular tres matrices de distancias para cada población. En una primera aproximación, podemos construir la matriz de una población como la **suma pitagórica** de las matrices de los diferentes conjuntos, asumiendo implícitamente independencia entre predictores:

$$\Delta_{\alpha}^{(2)} = \sum_{s=1}^{b} \Delta_{\alpha}^{(2)}{}_{[s]} \,. \tag{1}$$

Una alternativa, en la que además podemos ponderar cada uno de los conjuntos formados *a priori*, es la utilización de **familias de métricas adaptativas dependientes de parámetros** (ver Esteve, 2003 para mayor detalle). Estas familias se pueden obtener como combinación lineal convexa de las diferentes matrices de los conjuntos:

$$\Delta_{\alpha}^{(2)}(\lambda) = \sum_{s=1}^{b} \lambda_s \Delta_{\alpha}^{(2)}{}_{[s]}, \tag{2}$$

cumpliendo los parámetros $\sum_{i=1}^{b} \lambda_i = 1$. Y de una manera más completa, pero también más compleja de cálculo, podríamos utilizar el repertorio de distancias:

$$G_{\alpha}(C) = \sum_{s=1}^{b} G_{\alpha[s]} + \sum_{s \neq l} G_{\alpha[s]}^{1/2} C_l^s G_{\alpha[l]}^{1/2}, \tag{3}$$

siendo $G_{\alpha[s]}$ la matriz de productos escalares de la métrica asociada a la distancia $\Delta_{\alpha}^{(2)}{}_{[s]}$ y C_l^s matrices de parámetros de tamaño $n_{\alpha} \times n_{\alpha}$. La familia (3) nos permite incluir relaciones de dependencia entre los conjuntos de variables.

3. CRITERIOS DE SELECCIÓN DE MODELO EN *CREDIT SCORING*

En este apartado detallamos el cálculo de dos criterios de selección de modelo en *credit scoring*, las probabilidades de mala clasificación y el coste del error. En el apartado 4 ilustraremos cómo estos dos criterios nos ayudarán a decidir un modelo para nuestra cartera.

3.1. PROBABILIDADES DE MALA CLASIFICACIÓN

Vamos explicar y comentar cómo se calculan las probabilidades de mala clasificación en una técnica de predicción discriminativa, tanto para cada población como global. Para ello es necesario calcular la **matriz de confusión** que se define del siguiente modo:

		Estimada		
		Buenos riesgos	Malos riesgos	Total
Real	Buenos riesgos	n_{11}	n_{21}	$n_{11}+n_{21}$
	Malos riesgos	n_{12}	n_{22}	$n_{12}+n_{22}$
	Total	$n_{11}+n_{12}$	$n_{21}+n_{22}$	n

En esta matriz, las filas representan la clasificación real y las columnas la clasificación predicha. Explicamos el significado de los elementos con un ejemplo. Escogemos la matriz resultante de aplicar en el apartado 4.1.1 el ADBD a los datos alemanes de crédito con λ = [0.16 0.05 0.32 0.47]. En este ejemplo n = 1000 individuos, de los cuales 700 han sido buenos riesgos y 300 malos. La matriz de confusión resultante es:

		Estimada		
		Buenos riesgos	Malos riesgos	Total
Real	Buenos riesgos	394	306	700
	Malos riesgos	73	227	300
	Total	467	533	1000

Probabilidades de mala clasificación:

Para cada grupo:

- El de buenos riesgos de crédito

$$\frac{n_{21}}{n_{11}+n_{21}} = \frac{306}{700} = 0.437$$

- El de malos riesgos de crédito

$$\frac{n_{12}}{n_{12}+n_{22}} = \frac{73}{300} = 0.243$$

Probabilidad *global*:

$$\frac{n_{21}+n_{12}}{n}=\frac{306+73}{1000}=0.379$$

En este ejemplo, la probabilidad de clasificar mal a un buen riesgo es de 0.437, la de clasificar mal a un mal riesgo es de 0.243, y la probabilidad global de clasificar mal a un individuo es de 0.379. En general, una Compañía Financiera podría decidir que la probabilidad global es una buena estimación de cuánto se va a equivocar con una técnica predictiva determinada. Pero hay que tener en cuenta las probabilidades de cada una de las poblaciones, la de buenos y malos riesgos. La probabilidad de equivocarse en la población de malos riesgos, es decir de conceder créditos a malos riesgos, es realmente importante. Si esta probabilidad es elevada, significará que nos equivocaremos a menudo concediendo crédito a malos riesgos. El coste de conceder un crédito que quedará impagado es mucho mayor que el de rechazar a un buen cliente cuyo coste es cero. En el ejemplo, la probabilidad más pequeña es la de clasificar mal a un mal riesgo, lo que es de interés. Sin embargo, tampoco es bueno clasificar mal a todos los buenos riesgos, ya que si no concedemos créditos a buenos clientes, en términos esperados no podremos compensar las pérdidas de los siniestros. Por todo ello, debemos elegir una técnica predictiva que mantenga un equilibrio entre las tres probabilidades.

3.2. COSTE DEL ERROR

En este apartado consideramos los **costes del error en *credit scoring*** y su impacto en la selección de modelos. Puesto que no es posible saber el coste futuro de una Compañía Financiera, y las probabilidades *a priori* de buenos y malos riesgos no están disponibles para una cartera concreta, queremos enfatizar que este criterio aplicado a los datos del apartado 4 servirá sólo a modo de propuesta ilustrativa.

En general, en *credit scoring* el coste de conceder un crédito a un candidato con mal riesgo de crédito, al que llamaremos C_{12}, es significativamente mayor que el coste de denegar un crédito a un candidato con buen riesgo de crédito, al que llamaremos C_{21}. En esta situación es adecuado tener en cuenta el coste:

$$\text{Coste} = C_{12}\pi_2 \frac{n_{12}}{n_{11}+n_{12}} + C_{21}\pi_1 \frac{n_{21}}{n_{21}+n_{22}}, \qquad (4)$$

en lugar de la probabilidad global de mala clasificación de una metodología (ver Frydman *et al.*, 1985). Para ilustrar esta función de coste utilizaremos las siguientes estimaciones:

- Respecto de los costes relativos, utilizaremos los propuestos por el Dr. Hans Hofmann (que fue quien recopiló y cedió los datos alemanes del apartado 4 en el repositorio *Statlog*). Éstos son: $C_{12} = 5$ y $C_{21} = 1$. También son utilizados para estos datos por Frydman *et al.* (1985) y West (2000).

- Por otro lado, requerimos las probabilidades *a priori* de buenos, π_1, y de malos riesgos, π_2. Hemos considerado una buena estimación (ajustada a los datos reales en estudio) la propuesta por West (2000), quien también analiza las dos carteras que nosotros trabajamos en el apartado 4. En el citado trabajo, West propone estimar la probabilidad de los malos riesgos entre dos cotas: $\pi_2 = 0.144$ y $\pi_2 = 0.249$. Ambas cotas suponen dos escenarios, uno peor que otro. De este modo es posible averiguar entre qué valores podría oscilar el coste (4) si se dieran las dos situaciones. Las cotas están calculadas mediante el cociente de unos ratios obtenidos por Gopinathan y O'Donnell (1998) y Jensen (1992) a partir de experiencia real, y divididos por la media de las probabilidades estimadas en West (2000). Nos referimos a West (2000) para un mayor detalle.

Finalmente, el significado de los ratios $\frac{n_{12}}{n_{11}+n_{12}}$ y $\frac{n_{21}}{n_{21}+n_{22}}$ es el siguiente:

$\frac{n_{12}}{n_{11}+n_{12}}$: proporción de malos riesgos que son concedidos (ratio de falso positivo)

$\frac{n_{21}}{n_{21}+n_{22}}$: proporción de buenos riesgos que son denegados (ratio de falso negativo)

Con todo ello, los costes que aplicaremos en el apartado 4 para ilustrar el uso de (4) en la selección de modelo en *credit scoring* serán:

$$\text{Coste}(0.144) = 5\times(0.144)\times\frac{n_{12}}{n_{11}+n_{12}} + 1\times(1-0.144)\times\frac{n_{21}}{n_{21}+n_{22}} \quad (5)$$

$$\text{Coste}(0.249) = 5\times(0.249)\times\frac{n_{12}}{n_{11}+n_{12}} + 1\times(1-0.249)\times\frac{n_{21}}{n_{21}+n_{22}} \quad (6)$$

Que en el ejemplo, los costes (5) y (6) son:

$$\text{Coste}(0.144) = 5\times(0.144)\times\frac{73}{467} + 1\times(1-0.144)\times\frac{306}{533} =$$
$$= 5\times(0.144)\times 0.156 + 1\times(1-0.144)\times 0.574 = 0.604$$

$$\text{Coste}(0.249) = 5\times(0.249)\times 0.156 + 1\times(1-0.249)\times 0.574 = 0.625.$$

4. APLICACIONES

En este apartado, aplicamos la metodología de ADBD a dos conjuntos de datos reales de riesgo de crédito. Con el objetivo de establecer criterios de selección de modelos, comparamos los resultados de las probabilidades de mala clasificación y de los costes explicados en el apartado anterior, con los de otras metodologías de *credit scoring*.

Los datos han sido descargados gratuitamente del repositorio *Statlog*. Ambos conjuntos son carteras de Entidades Financieras, los primeros de una Financiera alemana y los segundos de una australiana. Los datos alemanes están descritos y pueden descargarse en la dirección electrónica

http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data),

y los australianos en

http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval).

4.1. DATOS DE RIESGO DE CRÉDITO ALEMANES

4.1.1. DESCRIPCIÓN DE LA BASE DE DATOS Y TRATAMIENTO

Estos datos clasifican a un conjunto de individuos como buenos o malos riesgos en función de una serie de predictores de tipo mixto. La cartera contiene datos cedidos en fecha 17-11-1994. En total contiene n = 1000 individuos, de los cuales 700 han sido buenos riesgos y 300 malos. Los factores potenciales de riesgo considerados son p = 20, de los cuales 7 son continuos, 11 categóricos y 2 binarios.

Para la aplicación del ADBD consideramos g = 2 poblaciones, la de buenos riesgos y la de malos riesgos. Puesto que conocemos la descripción de los predictores, construimos b = 4 conjuntos de variables en función de su significado. Esta agrupación previa es usual en el problema del riesgo de crédito (ver por ejemplo, Artís *et al.*, 1994) . Cabe notar que no existe un único criterio de agrupación, siempre dependerá de los factores disponibles y de la decisión del experto que establezca los conjuntos.

Respecto de la función de distancias en el cálculo de las matrices de distancias al cuadrado de cada conjunto de predictores, $\Delta_{\alpha}^{(2)}{}_{[s]}$ con $s = 1,\ldots,b$, utilizamos el índice de similaridad de Gower (Gower, 1971, Boj *et al*., 2004, 2007, 2009b). Esta función de distancias permite el tratamiento adecuado de datos de tipo mixto. Posteriormente utilizamos la familia paramétrica de distancias convexa (2), utilizando varias combinaciones *ad-hoc* de parámetros λ. Estas combinaciones nos permiten ponderar *a priori* los conjuntos de variables construidos, que en nuestro caso son:

Conjunto 1. Características del crédito

En este conjunto hemos considerado todas las características que hacen referencia al crédito. En total tenemos dos variables continuas y una categórica nominal. Los factores incluidos son:

- **Factor 2**- Duración en meses (numérica)
- **Factor 5**- Importe del crédito (numérica)
- **Factor 4**- Propósito (categórica nominal con 11 niveles)
  [1. car (new); 2. car (used); 3. furniture/equipment; 4. radio/television; 5. domestic appliances; 6. repairs; 7. education;

8. (vacation - does not exist?); 9. retraining; 10. business; 11. others]

Conjunto 2. Características sociales del creditor (beneficiario del crédito)

En este conjunto hemos considerado todas las variables que hacen referencia a características sociales del creditor o beneficiario del crédito. En total disponemos de dos variables continuas, una categórica nominal y dos binarias. Los factores incluidos son:

- **Factor 11**- Residencia actual desde (numérica)
- **Factor 13**- Edad en años (numérica)
- **Factor 9**- Situación personal y sexo (categórica nominal con 5 niveles)
  [1. male:divorced/separated; 2. female:divorced/separated/married; 3. male:single; 4. male:married/widowed; 5. female:single]
- **Factor 19**- Teléfono (binaria)
  [1. none; 2. yes, registered under the customers name]
- **Factor 20**- Trabajador extranjero (binaria)
  [1. yes; 2. no]

Conjunto 3. Características económicas del creditor (beneficiario del crédito)

En este conjunto hemos considerado todas las variables que hacen referencia a características económicas del beneficiario del crédito. En total disponemos de cinco variables cuantitativas y cuatro categóricas nominales. Los factores finalmente incluidos en este conjunto y su tratamiento han sido:

- **Factor 1**-Situación actual de la cuenta corriente (categórica ordinal). Este factor merece un tratamiento especial. En la base de datos ha sido codificado como una discretización en clases de una variable cuantitativa originariamente:
  [1. ... < 0 DM
  2. 0 <= ... < 200 DM
  3. ... >= 200 DM / salary assignments for at least 1 year
  4. no checking account]

Puesto que los datos numéricos reales no los podemos recuperar, hemos decidido utilizar las marcas de clase de los intervalos, [–50, 100, 250, 0] (ver Boj *et al*., 2004 para otras aplicaciones con las mismas características). Adicionalmente, puesto que hemos codificado un 0 para la clase 4, que se corresponde con "no tener cuenta corriente", hemos añadido una variable binaria para indicar dichos ceros, que de hecho pasan a ser datos faltantes en la similaridad de Gower.

- **Factor 6**- Cuenta de ahorros (categórica ordinal). A esta variable le ocurre lo mismo que a la anterior, por ello hemos aplicado el mismo tratamiento. Para este factor las clases son:

  [1. ... < 100 DM
  2. 100 <= ... < 500 DM
  3. 500 <= ... < 1000 DM
  4. ... >= 1000 DM
  5. unknown/ no savings account]

  Y las marcas de clase utilizadas han sido [50, 300, 750, 1250, 0]. La clase con indicador de cero ha sido la 5, que se corresponde con "no tener cuenta de ahorros".

- **Factor 7**- Presenta empleo desde (categórica ordinal). De nuevo tenemos una variable obtenida como discretización de una numérica, además de tener una clase a indicar a parte. Las clases son:

  [1. unemployed
  2. ... < 1 year
  3. 1 <= ... < 4 years
  4. 4 <= ... < 7 years
  5. ... >= 7 years]

  Las marcas de clase utilizadas han sido [0,0.5,2.5,5.5,8.5], y la clase con indicador de cero ha sido la 1, que se corresponde con "estar desempleado".

- **Factor 7b**- Indicador de empleo (binaria). Hemos pensado que el hecho de no disponer de empleo, considerado en el Factor 7 anterior, era un factor de riesgo muy importante por su significado. Por ello, hemos decidido crear una variable binaria que indica si se dispone o no de empleo. También indicaría la clase de "estar o no desmpleado" pero tendría el tratamiento de variable binaria adicional en el índice de similaridad de Gower y no de indicador de dato faltante de la variable 7.

- **Factor 8**- Cuota del crédito en porcentaje de la renta disponible (numérica)
- **Factor 18**- Número de personas mantenidas (numérica)
- **Factor 12**- Propiedades (categórica nominal con 4 niveles)
  [1. real estate; 2. if not 1.: building society savings agreement/life insurance; 3. if not 1./2.: car or other, not in Factor 6; 4. unknown/no property]
- **Factor 14**- Otros planes periódicos (categórica nominal con 3 niveles)
  [1. bank; 2. stores; 3. none]
- **Factor 15**- Vivienda (categórica nominal con 3 niveles)
  [1. rent; 2. own; 3. for free]
- **Factor 17**- Trabajo (categórica nominal con 4 niveles)
  [1. unemployed/unskilled - non-resident; 2. unskilled – resident; 3. skilled employee/official; 4. management/self-employed/highly qualified employee/ officer]

Conjunto 4. Relación del creditor (beneficiario del crédito) con el banco

En este conjunto hemos considerado todas las características que hacían referencia a la relación del beneficiario del crédito con el banco. En total disponemos de una variable continua y dos categóricas nominales. Los factores incluidos en este conjunto son:

- **Factor 16**- Número de créditos activos en este banco (numérica)
- **Factor 3**- Historial crediticio (categórica nominal con 5 niveles)
  [1. no credits taken/ all credits paid back duly; 2. all credits at this bank paid back duly; 3. existing credits paid back duly till now; 4. delay in paying off in the past; 5. critical account/ other credits existing (not at this bank)]
- **Factor 10**- Otras personas en el crédito (categórica nominal con 3 niveles)
  [1. none; 2. co-applicant; 3. guarantor]

A continuación detallamos las matrices de confusión resultantes de aplicar diferentes métricas con la metodología de ADBD.

Caso 1: Suponemos que todos los predictores corresponden a un único conjunto de variables, $b = 1$, así que la utilización de la similaridad de

Gower se corresponde con el caso de **suma pitagórica de distancias euclídeas**, y la matriz de confusión resultante es**:**

$$\begin{bmatrix} 544 & 156 \\ 188 & 112 \end{bmatrix}.$$

Caso 2: Suponemos la **familia convexa de distancias paramétricas** (2). Fijamos diferentes pesos *a priori* y para cada uno detallamos su matriz de confusión:

1) λ = [0.25 0.25 0.25 0.25]: $\begin{bmatrix} 455 & 245 \\ 86 & 214 \end{bmatrix}$.

2) λ = [0.16 0.05 0.32 0.47]: $\begin{bmatrix} 394 & 306 \\ 73 & 227 \end{bmatrix}$.

3) λ = [0.14 0.05 0.36 0.45]: $\begin{bmatrix} 407 & 293 \\ 76 & 224 \end{bmatrix}$.

4) λ = [0.10 0.10 0.40 0.40]: $\begin{bmatrix} 420 & 280 \\ 81 & 219 \end{bmatrix}$.

5) λ = [0.40 0.40 0.10 0.10]: $\begin{bmatrix} 461 & 239 \\ 106 & 194 \end{bmatrix}$.

4.1.2. ANÁLISIS COMPARATIVO DE MODELOS

En las Tablas 1 y 3 se encuentran resumidas las probabilidades de mala clasificación para cada grupo, buenos y malos riesgos, y la probabilidad global, para los datos de riesgo de crédito Alemanes y Australianos, utilizando diferentes metodologías de *credit scoring*. En las Tablas 2 y 4 hemos resumido el coste del error asociado, (4), a cada base de datos y cada metodología, suponiendo $\pi_2 = 0.144$ y $\pi_2 = 0.249$. Respecto de las metodologías que aparecen en las tablas, que serán con las que vamos a comparar nuestro método de ADBD, cabe notar que los resultados numéricos están extraídos de West (2000) y que para obtener información de los procesos de estimación y las hipótesis en que se basan debemos consultar la referencia citada. A continuación las listamos en dos clases:

Metodologías no-paramétricas:

- Redes neuronales, cuyas siglas se refieren a (consultar West, 2000 para mayor detalle):
    - *Mixture of experts* (MOE)
    - *Radial basis function* (RBF)
    - *Multi-layer perceptron* (MLP)
    - *Learning vector quantization* (LVQ)
    - *Fuzzy adaptive resonance* (FAR)
- Método de los *k* vecinos más próximos
- Método de la estimación núcleo de la densidad
- Árbol de clasificación *classification and regression trees* (CART)

Metodologías paramétricas:

- Análisis discriminante lineal
- Regresión logística

Si nos centramos en la probabilidad de mala clasificación de los malos riesgos de la Tabla 1, la metodología con menor probabilidad es la del ADBD con λ = [0.16 0.05 0.32 0.47], que da 0.243. Esta ponderación en la distancia convexa da mucha importancia al conjunto 4 (Relación del beneficiario del crédito con el banco), seguida del Conjunto 3 (Características económicas del beneficiario del crédito), del conjunto 1 (Características del crédito) y finalmente del Conjunto 2 (Características sociales del beneficiario del crédito). Esta ponderación es la que *ad-hoc* da una interpretación más intuitiva de los conjuntos formados. Este hecho ratifica que el correcto uso de la métrica conlleva a mejores resultados. La siguiente técnica que ofrece una menor probabilidad es la de análisis discriminante lineal, con 0.266.

Si nos centramos en la probabilidad global de la Tabla 1, la técnica que menor error de clasificación conlleva es la de regresión logística con 0.237, sin embargo hay que notar que la probabilidad de los malos riesgos es muy elevada, de 0.513. La siguiente técnica es la red neuronal MOE con una probabilidad de 0.243. Observamos que algunas de las técnicas que tienen también una probabilidad global pequeña tienen también una probabilidad de mala clasificación elevada, haciendo que la global disminuya a costa de clasificar mal a los buenos riesgos.

	Probabilidades estimadas de mala clasificación		
	Buenos riesgos	Malos riesgos	Global
Modelos no-paramétricos			
ADBD (Suma pitagórica)	0.223	0.627	0.344
ADBD (λ = [0.25 0.25 0.25 0.25])	0.350	0.287	0.331
ADBD (λ = [0.16 0.05 0.32 0.47])	0.437	**0.243**	0.379
ADBD (λ = [0.14 0.05 0.36 0.45])	0.419	0.253	0.369
ADBD (λ = [0.10 0.10 0.40 0.40])	0.400	0.270	0.361
ADBD (λ = [0.40 0.40 0.10 0.10])	0.341	0.353	0.345
Red neuronal MOE	0.142	0.477	0.243
Red neuronal RBF	0.134	0.529	0.254
Red neuronal MLP	0.135	0.575	0.267
Red neuronal LVQ	0.249	0.481	0.316
Red neuronal FAR	0.403	0.488	0.427
***K* vecinos próximos**	0.225	0.553	0.324
Estimación núcleo de la densidad	0.155	0.630	0.308
Árbol de clasificación CART	0.206	0.545	0.304
Modelos paramétricos			
Análisis discriminante lineal	0.277	0.266	0.274
Regresión logística	0.118	0.513	**0.237**

Tabla 1. Probabilidades de mala clasificación para cada grupo, buenos y malos riesgos, y global para los datos de riesgo de crédito Alemanes utilizando diferentes metodologías de *credit scoring*.

	Costes estimados	
	$\pi_2 = 0.144$	$\pi_2 = 0.249$
Modelos no-paramétricos		
ADBD (Suma pitagórica)	0.683	0.756
ADBD (λ = [0.25 0.25 0.25 0.25])	0.562	0.591
ADBD (λ = [0.16 0.05 0.32 0.47])	0.604	0.625
ADBD (λ = [0.14 0.05 0.36 0.45])	0.598	0.621
ADBD (λ = [0.10 0.10 0.40 0.40])	0.596	0.622
ADBD (λ = [0.40 0.40 0.10 0.10])	0.607	0.647
Red neuronal MOE	0.432	0.653
Red neuronal RBF	0.469	0.707
Red neuronal MLP	0.483	0.758
Red neuronal LVQ	0.501	0.714
Red neuronal FAR	0.668	0.942
***K* vecinos próximos**	0.592	0.858
Estimación núcleo de la densidad	0.587	0.901
Árbol de casificación CART	0.569	0.834
Modelos paramétricos		
Análisis discriminante lineal	0.429	**0.540**
Regresión logística	**0.471**	0.728

Tabla 2. Coste del error suponiendo $\pi_2 = 0.144$ y $\pi_2 = 0.249$ para los datos de riesgo de crédito Alemanes utilizando diferentes metodologías de *credit scoring*.

Observamos en la Tabla 2, que los costes varían mucho en función de la metodología. Cabe notar, que aunque los costes del ADBD no son los más pequeños, sí que ofrecen unos intervalos de variación menores en general. Eso significa que al enfrentarse a diferentes escenarios, el coste previsto es bastante estable, a diferencia del resto de métodos, en los que una variación de la probabilidad *a priori* de malos riesgos de la cartera supone un resultado muy diferente.

Si suponemos un escenario en que la probabilidad *a priori* de la población de malos riesgos es 0.144, la metodología con un menor coste es la de la regresión logística, que a su vez recordemos que era la que tenía una probabilidad de mala clasificación global también menor. En un escenario peor, suponiendo que la probabilidad *a priori* es de 0.249, la metodología con coste menor es la de análisis discriminante lineal. Observamos también que los siguientes costes más bajos en este escenario peor, se obtienen con ADBD.

4.2. DATOS DE RIESGO DE CRÉDITO AUSTRALIANOS

4.2.1. DESCRIPCIÓN DE LA BASE DE DATOS Y TRATAMIENTO

Estos datos hacen referencia al riesgo asociado a tarjetas de crédito de una Entidad Financiera. Para mantener la confidencialidad de los datos, el autor no cedió los nombres de los factores de riesgo ni lo que significan sus clases y valores. Este conjunto de datos es de especial interés porque el conjunto de predictores es de tipo mixto, mezcla de continuos, nominales con un número reducido de clases, nominales con un gran número de clases y binarios. Además, el número de datos faltantes es reducido. Para las variables continuas, los datos faltantes han sido re-emplazados por la media de las variable correspondiente, y para las variables categóricas y binarias, éstos han sido re-emplazados por la moda.

En total contiene $n = 690$ individuos, de los cuales 307 han sido buenos riesgos y 383 malos. Los factores potenciales de riesgo considerados son $p = 14$, de los cuales 6 son continuos, 4 categóricos y 4 binarios. Para la aplicación del ADBD consideramos de nuevo $g = 2$ poblaciones, la de buenos riesgos y la de malos riesgos. Puesto que no conocemos la descripción de los predictores, no podemos construir conjuntos en función de su significado, así que trataremos con un único conjunto, $b = 1$. También utilizaremos el índice de similaridad de Gower (Gower, 1971).

Cabe notar que, si quisiéramos ponderar *a priori* cada uno de los factores en el modelo de ADBD, una posibilidad sería utilizar la familia de métricas convexas del tipo (2), aunque en este caso, puesto que no conocemos el significado de las variables, no tiene mucho sentido. Por otro lado, también se podría pensar en agrupar las variables en función de la información que aportan, juntando en conjuntos las que aportan una información similar. Esta agrupación podría hacerse con alguna metodología de *cluster* y no se basaría en el significado, sino en incluir en un conjunto la misma información. Por ejemplo, si se utilizara una métrica que tuviera en cuenta las correlaciones entre variables como (3), y que además fuera capaz de tratar de forma adecuada datos de tipo mixto, ajustaríamos un modo muy adecuado, sin incorporar información redundante en el modelo.

La matriz de confusión resultante del ADBD para estos datos utilizando el índice de similaridad de Gower entendido como **suma pitagórica de las distancias euclídeas** aportadas por cada uno de los 14 predictores es:

$$\begin{bmatrix} 278 & 29 \\ 62 & 321 \end{bmatrix}.$$

4.2.2. ANÁLISIS COMPARATIVO DE MODELOS

Con estos datos observamos en la Tabla 3 que la metodología con menor probabilidad de mala clasificación de los malos riesgos es el CART, con 0.120, seguida de la red neuronal MOE, con 0.124. Si nos centramos en la probabilidad global la técnica con menor probabilidad es la regresión logística, con 0.127, seguida del ADBD, con 0.132.

	Probabilidades estimadas de mala clasificación		
	Buenos riesgos	Malos riesgos	Global
Modelos no-paramétricos			
ADBD (Suma pitagórica)	0.094	0.162	0.132
Red neuronal MOE	0.145	0.124	0.133
Red neuronal RBF	0.131	0.127	0.128
Red neuronal MLP	0.154	0.132	0.141
Red neuronal LVQ	0.171	0.171	0.170
Red neuronal FAR	0.256	0.238	0.246
***K* vecinos próximos**	0.153	0.133	0.142
Estimación núcleo de la densidad	0.185	0.151	0.166
Árbol de casificación CART	0.192	**0.120**	0.156
Modelos paramétricos			
Análisis discriminante lineal	0.078	0.190	0.140
Regresión logística	0.110	0.140	**0.127**

Tabla 3. Probabilidades de mala clasificación para cada grupo, buenos y malos riesgos, y global para los datos de crédito Australianos utilizando diferentes metodologías de *credit scoring*.

En la Tabla 4 volvemos a observar al igual que en la Tabla 2, que aunque no tenemos los menores costes con ADBD, el intervalo de variación no es muy grande. Casi todos los métodos ofrecen unos costes similares. Sólo resaltar que la metodología con mejor resultado es la red neuronal MOE, empatada cuando la probabilidad *a priori* es de 0.144 con la regresión logística, y por la red neuronal MLP cuando la probabilidad es de 0.249.

	Costes estimados	
	$\pi_2 = 0.144$	$\pi_2 = 0.249$
Modelos no-paramétricos		
ADBD (Suma pitagórica)	0.202	0.289
Red neuronal MOE	**0.196**	**0.243**
Red neuronal RBF	0.194	0.245
Red neuronal MLP	0.198	**0.243**
Red neuronal LVQ	0.237	0.300
Red neuronal FAR	0.319	0.388
***K* vecinos próximos**	0.227	0.281
Estimación núcleo de la densidad	0.267	0.328
Árbol de casificación CART	0.251	0.294
Modelos paramétricos		
Análisis discriminante lineal	0.204	0.296
Regresión logística	**0.196**	0.258

Tabla 4. Coste del error suponiendo $\pi_2 = 0.144$ y $\pi_2 = 0.249$ para los datos de riesgo de crédito Alemanes utilizando diferentes metodologías de *credit scoring*.

5. CONCLUSIONES

En este trabajo describimos diferentes criterios de selección de modelo en *credit scoring*. El primer criterio se basa en analizar las probabilidades de mala clasificación en las poblaciones, la de buenos y la de malos riesgos de crédito, y la global. Una Compañía Financiera puede decidir que la probabilidad global es una buena estimación de cuánto se va a equivocar con una técnica predictiva determinada. Pero hay que tener en cuenta las probabilidades de cada una de las poblaciones. La probabilidad de equivocarse en conceder créditos a malos riesgos es realmente importante. El coste de conceder un crédito que quedará impagado es mucho mayor que el de rechazar a un buen cliente cuyo coste es cero. Tampoco es bueno clasificar mal a todos los buenos riesgos, ya que si no concedemos créditos a buenos clientes, en términos esperados no podremos compensar las pérdidas de los siniestros. Por todo ello, debemos elegir una técnica predictiva que

mantenga un equilibrio entre las tres probabilidades. El segundo criterio se basa en una función de coste del error, la cuál tiene en cuenta el entorno de la cartera en estudio.

Proponemos la utilización del ADBD como método de *scoring* alternativo a los existentes en la literatura. Con dos conjuntos de datos reales de dos Entidades Financieras ilustramos la utilización de los criterios de selección de modelo y el funcionamiento de la predicción basada en distancias. Los métodos de la literatura con los que comparamos el ADBD son: las redes neuronales, el método de los *k* vecinos más próximos, el método de la estimación núcleo de la densidad, el árbol de clasificación *classification and regression trees* (CART), el análisis discriminante lineal y la regresión logística.

Concluyendo que, no existe un método óptimo para todas las carteras, sino que conviene estudiar el entorno basándose en la experiencia reciente. Cada metodología tiene ventajas e inconvenientes (en sus hipótesis, en el tiempo computacional, etc) y un estudio de selección de modelo ayudará en la toma de decisiones.

BIBLIOGRAFÍA

Artís, M, Guillén, M. and J. M. Martínez (1994). A model for credit scoring: an application of discriminant anlysis. *QÜESTIIÓ* 18:3, 385–395.

Boj, E., Claramunt, M. M. y J. Fortiana (2000). Una alternativa en la selección de los factores de riesgo a utilizar en el cálculo de primas. *Anales del Instituto de Actuarios Españoles*, Tercera Época 6, pp. 11–35.

Boj, E., Claramunt, M. M. y J. Fortiana (2001). Herramientas estadísticas para el estudio de perfiles de riesgo. *Anales del Instituto de Actuarios Españoles*, Tercera Época 7, pp. 59–89.

Boj, E., Claramunt, M. M. y J. Fortiana (2004). *Análisis multivariante aplicado a la selección de factores de riesgo en la tarificación*. Cuadernos de la Fundación MAPFRE, 88. Fundación MAPFRE Estudios, Madrid.

Boj, E., Claramunt, M. M. and J. Fortiana (2007). Selection of predictors in distance-based regression. *Communications in Statistics–Simulation and Computation* 36:1, 87–98.

Boj, E., Claramunt, M. M., Esteve, A. y J. Fortiana (2009a). Credit Scoring basado en distancias: coeficientes de influencia de los predictores. En: Heras, A. y otros (2009). *Investigaciones en Seguros y Gestión de riesgos: RIESGO 2009*, pp. 15–22. Cuadernos de la Fundación MAPFRE, 136. Fundación MAPFRE Estudios, Madrid.

Boj, E., Claramunt, M. M., Grané, A. and J. Fortiana (2009b). Projection error term in Gower's interpolation. *Journal of Statistical Planning and Inference* 139, 1867–878.

Bonilla, M., Olmeda, I. y R. Puertas (2003). Modelos paramétricos y no paramétricos en problemas de credit scoring. *Revista Española de Financiación y Contabilidad* 32:118, 833–869.

Borg, I. and P.J.F Groenen (2005). *Modern Multidimensional Scaling: Theory and Applications*. Second ed. Springer, New York.

Cuadras, C.M. (1989). Distance analysis in discrimination and classification using both continuous and categorical variables. In: Y. Dodge (Ed.), *Statistical Data Analysis and Inference*, pp. 459–473. Elsevier Science Publishers B. V. (North.Holland), Amsterdam.

Cuadras, C.M. (1992). Some examples of distance based discrimination. *Biometrical Letters* 29, 3–20.

Cuadras, C. M., Fortiana, J. and F. Oliva (1997). The proximity of an individual to a population with applications in discriminant analysis. *Journal of Classification* 14, 117–136.

Esteve, A. (2003). *Distancias estadísticas y relaciones de dependencia entre conjuntos de variables*. Tesis Doctoral. Universidad de Barcelona.

Frydman, H.E., Altman, E.I. and D. Kao (1985). Introducing recursive partitioning for financial classification: the case of financial distress. *Journal of Finance* 40:1, 269–291.

Gopinathan, K. and D. O'Donnell (1998). Just in time risk management. *Credit World* 87:2, 10–12.

Gower J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* 27, 857–874.

Hand, D. and W. Henley (1997). Statistical classification in consumer credit scoring: a review. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 160:3, 523–541.

Jensen, H.L. (1992). Using neural networks for credit scoring. *Managerial Finance* 18, 15–26.

Trias, R., Carrascosa, F., Fernández, D., Parés, Ll. y G. Nebot (2005). *Riesgo de Créditos: Conceptos para su medición, Basilea II, Herramientas de Apoyo a la Gestión*. AIS Group - Financial Decisions. www.ais-int.com

Trias, R., Carrascosa, F., Fernández, D., Parés, Ll. y G. Nebot (2008). *El método RDF (Risk Dynamics into the Future). El nuevo estándar de stress testing de riesgo de crédito*. AIS Group - Financial Decisions. www.ais-int.com

West, D. (2000). Neural network credit scoring models. *Computer & Operations Research* 27, 1131–1152.