

SISTEMAS DE INDUCCIÓN DE REGLAS Y ÁRBOLES DE DECISIÓN APLICADOS A LA PREDICCIÓN DE INSOLVENCIAS EN EMPRESAS ASEGURADORAS

Zuleyka Díaz Martínez ^(a)

José Fernández Menéndez ^(b)

M^a Jesús Segovia Vargas ^(a)

^(a) *Departamento de Economía Financiera y Contabilidad I.*

^(b) *Departamento de Organización de Empresas.*

Universidad Complutense de Madrid.

RESUMEN

Tradicionalmente, para abordar el problema de la detección precoz de la insolvencia empresarial, se han venido utilizando métodos estadísticos que emplean ratios financieros como variables explicativas. Sin embargo, aunque la eficacia de dichos métodos ha sido sobradamente probada, presentan algunos problemas que dificultan su aplicación en el ámbito empresarial, ya que, generalmente, se trata de modelos basados en una serie de hipótesis sobre las variables explicativas que en muchos casos no se cumplen y, además, dada su complejidad, puede resultar difícil extraer conclusiones de sus resultados para un usuario poco familiarizado con la técnica.

El presente trabajo describe una investigación de carácter empírico consistente en la aplicación al sector asegurador del algoritmo de inducción de reglas y árboles de decisión *See5*, a partir de un conjunto de ratios financieros de una muestra de empresas españolas de seguros no-vida, con el objeto de comprobar su utilidad para la predicción de insolvencias en este sector. También se comparan los resultados alcanzados con los que se obtienen aplicando la metodología *Rough Set*. Estas técnicas, procedentes del campo de la Inteligencia Artificial, no presentan los problemas mencionados anteriormente.

Palabras clave: Insolvencia, sector asegurador, See5, Rough Set.

1. INTRODUCCIÓN.

La predicción de la insolvencia es uno de los temas centrales del análisis financiero que ha suscitado el interés no sólo del ámbito académico sino también de un amplio abanico de usuarios relacionados con el mundo empresarial.

En el sector del seguro, dada su gran relevancia para el conjunto de la actividad económica, la detección precoz de insolvencias o de las condiciones que pueden llevar a que ésta acaezca es una cuestión de suma importancia e interés general. En este sector la mayoría de los métodos aplicados para predecir crisis son de tipo estadístico, como el análisis univariante o el multivariante (discriminante, logit, probit, etc.) en los que los ratios contables funcionan como variables explicativas. Aunque los resultados obtenidos han sido satisfactorios, todas estas técnicas presentan el inconveniente de que parten de hipótesis más o menos restrictivas acerca de las propiedades distribucionales de las variables explicativas que, especialmente en el caso de la información contable, no se suelen cumplir.

En un intento de superar esta limitación, surge el empleo de técnicas procedentes del campo de la Inteligencia Artificial, ya que, debido a su carácter no paramétrico, no precisan de hipótesis preestablecidas sobre las variables de partida. Dentro de estas técnicas, las que más se vienen aplicando al análisis de la solvencia empresarial son los Sistemas Expertos y las Redes Neuronales Artificiales, además de otras técnicas de aparición más reciente como los Conjuntos Aproximados – “Rough Sets” en terminología anglosajona -, los Algoritmos Genéticos o los Sistemas de Inducción de Reglas y Árboles de Decisión.

Este trabajo consiste básicamente en la aplicación al sector de empresas de seguros no-vida del algoritmo de inducción de reglas y árboles de decisión desarrollado por Quinlan (1997) “See 5”, que, como veremos, proporciona un modelo de predicción de insolvencias muy sencillo y fácilmente interpretable por el usuario. También comparamos los resultados obtenidos con los que se obtienen aplicando la metodología “Rough Set” y, finalmente, extraemos nuestras conclusiones.

2. LAS TÉCNICAS DE CLASIFICACIÓN.

Dentro del conjunto de técnicas a las que podemos recurrir si pretendemos analizar la información contenida en cualquier base de datos de tipo contable, destacan las de clasificación, en las que, a partir de un conjunto de variables explicativas, se trata de predecir la pertenencia de la empresa considerada a una serie de grupos o categorías

mutuamente excluyentes. Los sistemas clasificadores proceden fundamentalmente de la Estadística o de la Inteligencia Artificial y, entre otras, una manera de diferenciarlos radica en la forma en la que se construyen. Así, podemos entender la construcción de un sistema clasificador con n variables explicativas como la división del espacio n -dimensional que forman esas variables en una serie de regiones, cada una de las cuales se asigna a una de las categorías previamente definidas. Esta división puede ser realizada de dos formas:

- Definiendo una o varias hipersuperficies separadoras. Algunas técnicas estadísticas muy empleadas en el análisis de la solvencia empresarial siguen este enfoque (discriminante, logit, probit). Además, se han desarrollado en los últimos años modelos basados en redes neuronales artificiales, que permiten la definición de hipersuperficies separadoras muy complejas y tienen un carácter estrictamente no paramétrico.

- Realizando particiones sucesivas en el espacio de las variables explicativas, empleando una sola variable en cada partición. Dentro de este segundo tipo de sistemas clasificadores los Sistemas Expertos son la rama de la Inteligencia Artificial más empleada en la gestión empresarial, si bien sus posibilidades se han visto enriquecidas con los nuevos enfoques que han aportado otras técnicas de aparición más reciente, como los algoritmos de inducción de reglas y árboles de decisión, entre los que se encuentra el *See5*.

3. EL ALGORITMO DE INDUCCIÓN DE REGLAS Y ÁRBOLES DE DECISIÓN *See5*.

Encontrado dentro de las técnicas de Aprendizaje Automático (*Machine Learning*), el algoritmo *See5*¹ permite construir automáticamente a partir de un conjunto de datos de entrenamiento un árbol de clasificación. Para inferir el árbol, el algoritmo realiza particiones binarias sucesivas en el espacio de las variables explicativas, de forma que en cada partición se escoge la variable que aporta más información en función de una medida de *entropía* o cantidad de información. El árbol así construido consta del mínimo número de atributos (variables) que se requieren para la clasificación correcta de los ejemplos dados, con lo que es claro el alto poder explicativo de esta técnica.

¹ Este algoritmo constituye una extensión de los algoritmos ID3 y C4.5. Una descripción detallada puede verse en Quinlan (1993 y 1997).

También se pueden elaborar, a partir del árbol, reglas de clasificación fácilmente interpretables, que definen las características que más diferencian a las distintas categorías establecidas inicialmente.

Este tipo de sistemas clasificadores presentan la ventaja, frente a las técnicas estadísticas, de que tienen un carácter estrictamente no paramétrico. Además, aunque no alcanzan el poder predictivo de las redes neuronales, sus resultados son mucho más fácilmente interpretables que los modelos de “caja negra” suministrados por aquéllas.

4. ANÁLISIS EMPÍRICO.

4.1. Metodología.

En cuanto a la muestra de empresas seleccionada para el análisis, ésta es la utilizada para la aplicación de la metodología Rough Set a la predicción de crisis empresariales en seguros no-vida en el trabajo de Segovia (2003). Consta de 36 empresas no fracasadas y 36 empresas fracasadas (en adelante, “buenas” y “malas”), emparejadas por tamaño y tipo de negocio, eliminando así el efecto de estas variables en el estudio, y escogiendo como criterio de selección de las empresas fracasadas el hecho de haber sido intervenidas por la Comisión Liquidadora de Entidades Aseguradoras (CLEA), por entender que se trata de una medida objetivamente determinable de las empresas que fracasan.

Una vez tomada la muestra, nos situamos en el periodo anterior al de la insolvencia para tratar de determinar qué indicios de este suceso nos proporcionan los datos de las cuentas anuales en forma de ratios. El éxito o fracaso de una empresa será entendido como una variable dependiente que deberá ser explicada por un conjunto de ratios financieros que actuarán como variables independientes. Así que de cada una de las empresas se han obtenido las cuentas anuales del año previo a la quiebra y, a partir de dicha información, se han calculado una serie de ratios, unos populares en la literatura contable para medir la solvencia empresarial y otros específicos del sector asegurador. En la Tabla I se presentan los 21 ratios seleccionados. Dadas las peculiaridades de la muestra, los ratios 15 y 16 han sido eliminados en el análisis posterior por tomar valores carentes de sentido económico, utilizando finalmente como variables independientes los 19 ratios restantes.

De las 72 empresas de que consta la muestra original, hemos utilizado únicamente 27 empresas de cada una de las submuestras para la elaboración de los modelos,

reservando las 9 restantes para poder comprobar la validez de los mismos aplicándolos a empresas cuyos datos no hubieran sido utilizados en dicha elaboración. En consecuencia, tendremos una muestra de *entrenamiento* para obtener los árboles y reglas de decisión formada por 54 empresas y una muestra de *validación* para verificar su capacidad predictiva formada por 18 empresas. La selección de la muestra de validación se ha efectuado aleatoriamente, tomando las empresas numeradas en Segovia (2003) de la 19 a la 27 para las “malas” y de la 119 a la 127 para las “buenas” como submuestras de validación.

El algoritmo se aplicó utilizando el programa *See5* de *Rulequest Research*.

Tabla I: Ratios empleados

| Ratio | Definición |
|--------------|---|
| R1 | Fondo de Maniobra / Activo Total |
| R2 | Beneficio antes de Impuestos(BAI)/ Capitales propios |
| R3 | Ingresos Financieros/ Total Inversiones |
| R4 | BAI*/ Pasivo Total BAI* = BAI+ Amortizaciones + Provisiones + Resultados Extraordinarios |
| R5 | Total Primas adquiridas de seguro directo / Capitales propios |
| R6 | Total Primas adquiridas de negocio neto / Capitales propios |
| R7 | Total Primas adquiridas de seguro directo / Capitales propios + Provisiones Técnicas |
| R8 | Total Primas adquiridas de negocio neto /Capitales propios + Provisiones Técnicas |
| R9 | Capitales Propios / Pasivo Total |
| R10 | Provisiones Técnicas / Capitales Propios |
| R11 | Gastos Técnicos de seguro directo / Capitales propios |
| R12 | Gastos Técnicos de negocio neto / Capitales propios |
| R13 | Gastos Técnicos de seguro directo / Capitales propios + Prov. Técnicas |
| R14 | Gastos Técnicos de negocio neto / Capitales propios + Provisiones Técnicas |
| R15 | Ratio Combinado 1 = Ratio Siniestralidad de seguro directo (RSD)+ Ratio de Gastos (RG) RSD = Gastos Técnicos de seguro directo/ Total Primas adquiridas de seguro directo RG = Comisiones y otros gastos de explotación/ Otros ingresos explotación |
| R16 | Ratio Combinado 2 = Ratio Siniestralidad de negocio neto (RSN)+ Ratio de Gastos (RG) RSN = Gastos Técnicos de negocio neto/ Total Primas adquiridas de negocio neto RG = Comisiones y otros gastos de explotación/ Otros ingresos explotación |
| R17 | (Gastos Técnicos de seguro directo + Comisiones y otros gastos de Explotación)/ Total Primas adquiridas de seguro directo |
| R18 | (Gastos Técnicos de negocio neto + Comisiones y otros gastos de Explotación)/ Total Primas adquiridas de negocio neto |
| R19 | Provisiones Técnicas de reaseguro cedido / Provisiones Técnicas |
| R20 | RSD = Gastos Técnicos de seguro directo/ Total Primas adquiridas de seguro directo |
| R21 | RSN = Gastos Técnicos de negocio neto/ Total Primas adquiridas de negocio neto |

4.2. Resultados.

A continuación se presenta el árbol de decisión obtenido aplicando el algoritmo *See5* a nuestra muestra, así como la evaluación del mismo tanto para la muestra utilizada en su elaboración como para la muestra de validación.

```

R13 > 0.68:
:...R9 <= 0.59: mala (14)
:   R9 > 0.59:
:     :...R17 <= 0.99: mala (3)
:       R17 > 0.99: buena (3)
R13 <= 0.68:
:...R1 > 0.29: buena (20/2)
:   R1 <= 0.29:
:     :...R2 > 0.04: mala (3)
:       R2 <= 0.04:
:         :...R6 > 0.64: buena (3)
:           R6 <= 0.64:
:             :...R9 <= 0.85: mala (4)
:               R9 > 0.85: buena (4/1)

```

Evaluation on training data (54 cases):

```

      Decision Tree
-----
Size      Errors
      8      3(5.6%)  <<

(a)  (b)  <-classified as
----  ----
  27      (a): class buena
   3  24  (b): class mala

```

Evaluation on test data (18 cases):

```

      Decision Tree
-----
Size      Errors
      8      5(27.8%) <<

(a)  (b)  <-classified as
----  ----
   7      (a): class buena
   3      (b): class mala

```

Como se puede observar, en el árbol aparecen únicamente 6 de los 19 ratios iniciales, lo que indica que 13 de los ratios empleados no aportan información relevante para clasificar las empresas como “buenas” o “malas”. Como ya se ha mencionado, el árbol nos proporciona el menor número de atributos (ratios) necesarios para alcanzar el objetivo deseado. Nuestro árbol se leería del modo siguiente:

- Si el ratio R13 es mayor de 0,68 y además el ratio R9 es menor o igual de 0,59, la empresa será “mala”, siendo 14 el número de empresas de la muestra que verifican este hecho.

- Si el ratio R13 es mayor de 0,68 y el ratio R9 es mayor de 0,59 y el ratio R17 menor o igual de 0,99, la empresa será “mala”, cumpliendo estas condiciones 3 empresas.

Y así continuaríamos descendiendo por el árbol, hasta completar un total de 8 hojas. Obsérvese que al final de cada hoja aparece un valor (n) o (n/m): n representa el número de empresas en la muestra que se clasifican de acuerdo a las condiciones que nos llevan hasta esa hoja y m el número de empresas mal clasificadas.

La evaluación de este árbol de decisión construido con la muestra de entrenamiento (54 empresas) indica que el árbol consta de 8 ramas y comete un total de 3 errores (5,6%), lo que supone un porcentaje de aciertos del 94,4%. También se muestra una *matriz de confusión* que señala el tipo de errores cometidos.

Por último, para comprobar la capacidad predictiva del árbol, se clasifican de acuerdo con éste las 18 empresas de la muestra de validación, obteniendo un porcentaje de clasificaciones correctas del 72,2%.

Aunque no en este caso, en ocasiones puede resultar difícil interpretar un árbol de decisión. See5 permite solventar este problema derivando, a partir del árbol, un conjunto de reglas más simples de la forma *si* (condiciones) - *entonces* (decisión). Las reglas que se obtienen a partir del árbol anterior son las siguientes:

Rules:

```
Rule 1: (20/2, lift 1.7)
  R1 > 0.29
  R13 <= 0.68
  -> class buena [0.864]

Rule 2: (12/1, lift 1.7)
  R2 <= 0.04
  R6 > 0.64
  R13 <= 0.68
  -> class buena [0.857]

Rule 3: (7/1, lift 1.6)
  R9 > 0.85
  -> class buena [0.778]

Rule 4: (14, lift 1.9)
  R9 <= 0.59
  R13 > 0.68
  -> class mala [0.938]

Rule 5: (7, lift 1.8)
  R13 > 0.68
  R17 <= 0.99
  -> class mala [0.889]
```

```
Rule 6: (26/6, lift 1.5)
  R1 <= 0.29
  -> class mala [0.750]

Default class: buena
```

Cada regla consiste en:

- Una serie de estadísticas $(n, \text{lift } x)$ o $(n/m, \text{lift } x)$; n y m representan lo mismo que en el árbol y x es el resultado de dividir la precisión estimada de la regla entre la frecuencia relativa de la clase predicha en la muestra de entrenamiento. La precisión de la regla se estima mediante el denominado ratio de Laplace $(n-m+1)/(n+2)$.
- Una o más condiciones que deben ser satisfechas para que la regla sea aplicable.
- La clase predicha por la regla.
- Un valor entre 0 y 1 que indica el nivel de confianza con el que ha sido hecha la predicción.

También existe una clase por defecto (en este caso “buena”) para cuando ninguna de las reglas sea aplicable.

El número de errores cometidos al clasificar mediante estas reglas y el tipo de los mismos coinciden, tanto con la muestra de entrenamiento como con la de validación, con los de las clasificaciones hechas con el árbol (aunque no siempre tiene por qué ser así).

A pesar de que los resultados que hemos obtenidos son satisfactorios, es posible mejorarlos recurriendo a la opción que incorpora *See5* de *adaptive boosting*, basado en el trabajo de Freund y Schapire (1997). Muy brevemente, la idea consiste en generar varios clasificadores (árboles o conjuntos de reglas) en vez de sólo uno. Como primer paso, se construye un único árbol (o conjunto de reglas) del mismo modo que acabamos de ver, que cometerá algunos errores en la clasificación (3 en nuestro caso). Estos errores serán el foco de atención al construir el segundo clasificador en aras de corregirlos. En consecuencia, el segundo clasificador generalmente será diferente al primero y también cometerá errores que serán el foco de atención durante la construcción del tercer clasificador. Este proceso continúa para un número predeterminado de iteraciones o *trials*. Mediante este procedimiento, se consigue obtener un clasificador verdaderamente preciso. Así, partiendo de nuestro primer árbol de decisión los resultados que alcanzamos realizando 18 iteraciones son:

- Con la muestra de entrenamiento, el 100% de clasificaciones correctas, como podemos observar en la *matriz de confusión*:

| (a) | (b) | <-classified as |
|-----|-----|-------------------------------------|
| 27 | 27 | (a): class buena (b): class mala |

- Con la muestra de validación, el 83,3% de clasificaciones correctas:

| (a) | (b) | <-classified as |
|-----|-----|------------------|
| 7 | 2 | (a): class buena |
| 1 | 8 | (b): class mala |

4.3. Comparación con el enfoque Rough Set.

La Teoría Rough Set fue propuesta por Pawlak a comienzos de los ochenta. Es un enfoque que se encuadra también dentro de las aplicaciones de la Inteligencia Artificial. Brevemente, el enfoque Rough Set consiste en descubrir dependencias entre atributos en una tabla de información y reducir el conjunto de los mismos eliminando aquéllos que no son esenciales para caracterizar el conocimiento. Un *reducto* se define como el mínimo subconjunto de atributos que asegura la misma calidad de clasificación que el conjunto de todos ellos. De la tabla de información reducida pueden derivarse reglas de decisión en forma de sentencias lógicas (*si <condiciones> entonces <decisión>*). Los principales conceptos de esta teoría se exponen en Segovia (2003).

De la aplicación de la metodología Rough Set² a la muestra de entrenamiento se obtuvieron 229 reductos (conjunto de ratios que clasifican a las 54 empresas en “buenas” y “malas” igual que si tomásemos en consideración los 19 ratios originales), de los cuales se seleccionó el formado por los ratios R3, R4, R9, R14 y R17 (Segovia, 2003). Hemos de señalar que para la aplicación de este método previamente se discretizaron los valores continuos de los ratios.

Finalmente, una vez seleccionado el reducto, se derivaron 27 reglas de decisión, que constituyen un algoritmo de clasificación que puede ser utilizado para evaluar

² El análisis Rough Set de la información se desarrolló utilizando el programa ROSE (Predki *et al.*, 1998 y Predki y Wilk, 1999).

cualquier otra empresa. El poder predictivo de este algoritmo se contrastó con la muestra de validación formada por 18 empresas.

A continuación se presentan los porcentajes de clasificaciones correctas obtenidos con los dos métodos:

See5

| Clasificaciones correctas | Muestra de entrenamiento | Muestra de validación |
|----------------------------------|---------------------------------|------------------------------|
| Empresas “buenas” | 100% | 77,78% |
| Empresas “malas” | 100% | 88,89% |
| Total | 100% | 83,33% |

Rough Set

| Clasificaciones correctas | Muestra de entrenamiento | Muestra de validación |
|----------------------------------|---------------------------------|------------------------------|
| Empresas “buenas” | 100% | 77,78% |
| Empresas “malas” | 100% | 77,78% |
| Total | 100% | 77,78% |

Tal y como puede observarse en estas tablas, los resultados de ambas metodologías muestran su capacidad para responder de manera eficiente al problema de la predicción del fracaso empresarial, siendo alternativas muy fiables a las técnicas estadísticas tradicionales, y más aún en el caso del algoritmo See5, que obtiene un porcentaje de aciertos superior con la muestra de validación clasificando correctamente el 88,89% de las empresas “malas” frente al 77,78% del Rough Set, lo cual es importante teniendo en cuenta que precisamente lo que nos interesa captar es la insolvencia.

5. CONCLUSIONES.

Los dos métodos que hemos aplicado son estrictamente no paramétricos, lo que les convierte en superiores a las técnicas estadísticas en el sentido de que se adecúan más a la información contable, que suele presentar datos interrelacionados, incompletos, adulterados o erróneos; ofrecen productos muy sencillos entendibles fácilmente por el analista humano, ya sea en forma de árboles o reglas de decisión, realizando una clasificación de las empresas entre solventes e insolventes que permite determinar la importancia de cada variable en el proceso de asignación. Además, dan buenos resultados incluso cuando se trabaja con escaso número de datos, aspecto éste importante en las aplicaciones al ámbito financiero.

Sin embargo, cabe señalar algunas ventajas que el algoritmo See5 presenta frente al Rough Set: éste trabaja mejor con datos discretos, bien sean variables cualitativas o variables cuantitativas previamente discretizadas, mientras que el See5 acepta atributos de tipo discreto o continuo, sin ningún tipo de limitación; el subconjunto de variables relevantes se derivan automáticamente con See5, por contra, la metodología Rough Set proporciona muchos subconjuntos, entre los cuales habrá de seleccionar el analista para derivar las reglas de decisión; además, aunque no se ha visto aquí, See5 posee una ventaja muy importante desde el punto de vista económico, ya que permite considerar distintos costes de clasificación errónea, mientras que Rough Set sólo computa el número global de errores sin distinguir si se trata de clasificar una empresa sana como fracasada o clasificar una fracasada como sana, error este último que resultaría mucho más grave.

6. REFERENCIAS BIBLIOGRÁFICAS.

- FREUND, Y. y SCHAPIRE, R.E. (1997). “A decision-theoretic generalization of on-line learning and an application to boosting”. *Journal of Computer and System Sciences*, vol. 55(1), pp. 119-139.
- PAWLAK, Z. (1991). *Rough Sets. Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, Boston, London.
- PREDKI, B., SLOWINSKI, R., STEFANOWSKI, J., SUSMAGA, R. y WILK, S. (1998). “ROSE – Software Implementation of the Rough Set Theory”, en POLKOWSKI, L. y SKOWRON, A. (Eds.): *Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence*, vol. 1424. Springer-Verlag, Berlin, pp. 605-608.
- PREDKI, B. y WILK, S. (1999). “Rough Set Based Data Exploration Using ROSE System”, en RAS, Z.W. y SKOWRON, A. (Eds.): *Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence*, vol. 1609, Springer-Verlag, Berlin, pp. 172-180.
- QUINLAN, J.R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, Inc., California.
- QUINLAN, J.R. (1997). *See5* (en Internet, <http://www.rulequest.com/see5-info.html>).
- SEGOVIA, M.J. (2003). *Predicción de crisis empresariales en seguros no vida mediante la metodología Rough Set*, Tesis Doctoral, Universidad Complutense de Madrid.
- SEGOVIA, M.J., GIL, J.A., HERAS, A. y VILAR, J.L. (2003) “La metodología Rough Set frente al Análisis Discriminante en los problemas de clasificación multiatributo”, comunicación presentada a las XI Jornadas ASEPUMA, Oviedo.